



Evaluation of the internal Centres of Excellence (CoE) Åbo Akademi University

Questionnaire

Evaluated time span: Jan 1st, 2010 – May 31st, 2012

Deadline for CoE answers: October 15th, 2012

Recipient of CoE answers: registrator@abo.fi

Name of the CoE	Optimization and Systems Engineering
Department or equivalent	Department of Chemical Engineering
Address	Biskopsgatan 8, FIN-20500 ÅBO, FINLAND
Phone	+358-2-2154460
Internet home page (CoE)	http://www.abo.fi/ose

Principal Investigator of the CoE	Prof. Tapio Westerlund
Phone	+358-2-2154460
Email	tapio.westerlund@abo.fi

1 ORGANIZATION OF THE CoE

The steering group of this CoE currently consists of the following members:

- Tapio Westerlund, project leader, Professor of process design
- Göran Högnäs, Professor of mathematics
- Hannu Toivonen, Professor of automation and computer science
- Jukka Corander, Professor of statistics

The steering group has convened 3-5 times per year to discuss how the project is evolving, to make strategic decisions and to plan the future of the CoE in OSE. Within the university, the OSE group is connected to the Department of Chemical Engineering, Department of Natural Sciences and Department of Information Technologies. All members of the steering group have broad national and international contact networks.

Also playing a strategic role in the CoE is the International scientific panel consisting of the following members:

- Professor Ignacio Grossmann, Carnegie Mellon University, Pittsburgh, USA
- Professor Christodoulos Floudas, Princeton University, USA
- Professor Efstratios Pistikopoulos, Imperial College London, UK
- Professor Leo Liberti, École Polytechnique, Paris, France
- Professor Montaz Ali, University of the Witwatersrand, Johannesburg, South Africa

The OSE group organizes an annual seminar with one of the members of the International scientific panel as keynote speaker. The seminars are an opportunity for the PhD students and post-doctoral researchers to present their latest research findings and work as a catalyst for international collaboration. As a result, after each seminar one of the PhD students have had the opportunity to visit the university of the keynote speaker for an extended research visit.

News and information about the activities in the CoE can be found from the webpage <http://www.abo.fi/ose>.

1.1 Senior researchers of the CoE

The following senior researchers are active within the group, either as supervisors to the PhD students or through research collaborations on one or more of the research topics.

Name	Gender (M/F)	Academic Degree	Position
Corander, Jukka	M	Ph.D.	Professor of Statistics
Glader, Christer	M	Ph.D.	Post-doctoral researcher / university lecturer
Högnäs, Göran	M	Ph.D.	Professor of Applied Mathematics
Kurula, Mikael	M	Ph.D.	Post-doctoral researcher
Jansson, Fredrik	M	Ph.D.	Post-doctoral researcher
Lastusilta, Toni	M	Ph.D.	Post-doctoral researcher
Lundell, Andreas	M	Ph.D.	Post-doctoral researcher
Östermark, Ralf	M	Ph.D.	Professor of Financial Accounting and Optimization Systems
Pörn, Ray	M	Ph.D.	Post-doctoral researcher / lecturer
Salminen, Paavo	M	Ph.D.	Professor of Mathematics
Staffans, Olof	M	Ph.D.	Professor
Toivonen, Hannu	M	Ph.D.	Professor of Automation and Computer Science
Westerlund, Tapio	M	Ph.D.	Professor of Process Design

2 SELF ASSESSMENT

2.1 How well has the goals of the CoE been achieved so far and what can be expected during 2013-2014?

The research has followed the initial research plan and the goals have, at this stage, been completely fulfilled. In this CoE a very high priority has been put on post-gradual educational aspects. Eight PhD students, each following scheduled work-plans connected to the main research

topics of the CoE have been funded. During the years 2013 three PhD students are expected to defend their theses, in 2014 three more, and the final two in 2015. Additionally, several PhD students are closely connected to this CoE, but funded by other sources. The number of publications for the reported period is about 90 including 70 peer-reviewed journal papers, a monograph, four book chapters and 16 refereed articles in conference proceedings. The complete list is given in Section 4.1.

The research in the CoE is focused on two main blocks: *Theory and methods* as well as *Applications in science and engineering*. The first block contains four work-packages (WP1: *Optimization*; WP2: *Systems engineering*; WP3: *Systems theory*; WP4: *Mathematical statistics*). The second application block contains three work-packages (WP5: *Physics and material science*; WP6: *Biotechnology and -medicine*; WP7: *Engineering*). Below are short summaries of the main results achieved in each work-package during the period 2010-2012.

2.1.1 WP1: Optimization

Within this work-package, the focus has been on mixed-integer global optimization, model representations and generic methods and techniques for global optimization of certain classes of large scale mixed-integer nonlinear problems. Some of the most important results during the reported period, include the creation of a rigorous method for solving nonsmooth convex mixed-integer nonlinear programming (MINLP) problems as well as new reformulation techniques in global optimization. In a paper *On the generalization of ECP and OA methods to nonsmooth convex MINLP problems* by Eronen et al., published in the journal *Optimization*, a cutting-plane method for nonsmooth problems was presented for the first time. With respect to reformulation techniques, the focus has been on handling signomial and twice-differentiable (C^2) constraints. A complete algorithm for solving signomial problems was published in the book chapter *Global optimization of mixed-integer signomial programming problems* by Lundell and Westerlund 2012. For twice-differentiable constraints, a method using an α BB type underestimator was presented in the articles *A reformulation framework for global optimization* by Lundell et al. in *Journal of Global Optimization*, and *Global optimization of C^2 constraints by convex reformulations* by Skjäl et al. in *Chemical Engineering Transactions*.

2.1.2 WP2: Systems engineering

The research in this work-package has focused on two different problem areas. The first one is connected to system identification techniques. One specific problem studied consists of system identification involving sparse variable representations, where it is required that most of the variable values are zero. These problems can be solved efficiently and with good accuracy by using an l_1 cost for imposing sparsity. We have applied the approach to identify systems which switch their dynamics between a number of different models. The problem has been solved using support vector regression, for which the dual problem turns out to be a sparse optimization problem. The use of support vector regression has the benefit that it allows identification of nonlinear systems, which has not previously been adequately addressed in switching system identification. Another challenging system identification problem that we have managed to solve using sparse optimization consists of identification of systems subject to unknown outliers, offsets and/or trends. The second research area in this work package is concerned with linear periodically time-varying systems. Here, our most important contribution during the reported period consists of the analysis of a state-space Fourier-series representation of such systems allowing for the derivation of a number of new results, such as, computation of the optimal time-invariant Kalman-filter state estimator for time-periodic systems. Part of the results described above has been preliminary presented in conference proceedings, and journal publications are presently under review or in preparation.

2.1.3 WP3: Systems theory

The main contribution of this work-package is the development of an extensive new theory for linear systems, the so-called state/signal systems theory. The theory unifies the theories of a few seemingly different classes of systems into one general framework. The discrete-time theory has been worked out by Staffans and Arov in a series of around ten extensive articles, starting from 2005, and the continuous-time analogue has been researched in co-operation with Kurula in a series of five journal papers. Kurula's thesis of 2010 is mainly concerned with continuous-time state/signal systems. The main developments during the period 2010-2012 are the following: A general realization theory for state/signal systems in the spirit of de Branges and Rovnyak has been worked out by Arov, Staffans, and Kurula. As a by-product of the state/signal theory new results concerning the somewhat older class of well-posed input/state/output systems have been achieved. The general theory now extends to some new physical systems governed by partial differential equations, see the paper *A physically motivated class of scattering passive linear systems* by Staffans and Weiss in *SIAM Journal on Control and Optimization*. The state/signal approach has also inspired Opmeer and Staffans to find

new results in optimal control, see the papers *Optimal input-output stabilization of infinite-dimensional discrete time-invariant linear systems by output injection* and *Coprime factorization and optimal control on the doubly infinite discrete time axis*, both published in *SIAM Journal on Control and Optimization*. Notable work in progress is a monograph by Arov and Staffans that will cover the recent advances in state/signal systems theory. Ball, Kurula, and Staffans are about to submit a comprehensive paper detailing the intricacies of continuous-time realization theory of operator-valued irrational transfer functions. This article will be presented at a workshop in Leiden, the Netherlands, in February 2013.

2.1.4 WP4: Mathematical statistics

This work-package focuses on theoretical and computational developments of a wide variety of novel model classes for statistical description of complex systems, cluster analysis, predictive classification, inference on population genealogies, bacterial recombination and metagenomics. Our major recent breakthroughs include: generalized context-specific undirected graphical models and context-specific Bayesian networks (one paper by Nyman et al. submitted to *Journal of the American Statistical Association* and one paper by Pensar et al. submitted to *Data Mining Knowledge Discovery*), asymptotic and small-sample properties of various predictive classification models (Corander et al. in the paper *Have I seen you before? Principles of Bayesian predictive classification revisited* in the journal *Statistics and Computing*, the paper *Optimal Viterbi Bayesian predictive classification for data from finite alphabets* in *Journal of Statistical Planning and Inference* by, as well as a paper by Corander et al. submitted to *Journal of Multivariate Analysis*), theory for learning genetic population structures using Minimum Description Length (MDL) with Chow-Liu trees (*Learning genetic population structures using minimization of stochastic complexity* by Corander et al. in the journal *Entropy*), new family of diffusion-based approximations to inferring population genealogies which are scalable to several orders of magnitude larger data sets than with state-of-the-art coalescence based models (*Inference on Population Histories by Approximating Infinite Alleles Diffusion* by Sirén et al. in *Molecular Biology and Evolution*), model and algorithm for inferring recombination in large-scale bacterial genome data which is currently the only available model-based approach that is computationally scalable for tens to hundreds of genomes (*Detection of recombination events in bacterial genomes from large population samples* by Marttinen et al. in *Nucleic Acids Research*), Bayesian clustering model for bacterial metagenomics that has substantially increased species-resolution (*Bayesian estimation of bacterial community composition from 454 sequencing data* by Cheng et al. *Nucleic Acids Research*), Approximate Bayesian Computation (ABC) for inference of population dynamics from cross-sectional meta-population samples (a paper by Numminen et al. 2012 submitted to the journal *Biometrics*).

2.1.5 WP5: Physics and material science

The focus in this applied work-package has been on the creation of mathematical models for minimizing the energy of electron configurations in disordered materials. Models for Coulomb glasses, where the conduction electrons are localized to impurity sites and the electrons strongly interact with each other, have been studied. The physical applications are connected, for example, to lightly doped semiconductors at low temperatures. During the reported period, we have created binary quadratic formulations for this complex large-scale problem. Different techniques have been used for its solution, including semidefinite programming, linear and convex reformulation techniques. Results are given, for example, in the paper *The Coulomb glass – modeling and computational experience with a large scale 0-1 QP problem* by Pörn et al. in *Computer Aided Chemical Engineering*, as well as in the conference proceedings *A Mixed Integer Quadratic Reformulation of the Quadratic Assignment Problem with Rank-1 Matrix* by Nissfolk et al.

2.1.6 WP6: Biotechnology and -medicine

The application in this work-package is connected to peptide docking, and to obtain equilibrium structures for macromolecule-ligand complexes. The peptide docking problem can be interpreted as a global optimization problem, where the total conformation of energy should be minimized. The problem is complex and very demanding, requiring very sophisticated solution methods. In the so-called ECEPP/3 (empirical conformational energy program or peptides) potential model, the contributions to the total conformation energy are the electrostatic, nonbounded, hydrogen bounded, torsional, cysteine loop-closing and cysteine torsional energies. The energy minimization problem is a highly nonconvex optimization problem describing the peptide chain. During the reported period, the efforts have been concentrated on creating tight convex underestimators for smooth, twice-differentiable functions. Some of the main results have been published by Skjäl et al. in the paper *A Generalization of the Classical α BB Convex Underestimation via Diagonal and Nondiagonal Quadratic Terms* in *Journal of Optimization Theory and Applications*, and in a conference article *A generalization of classical α BB underestimation to include bilinear terms*. The research has been performed in close cooperation with Professor Christodoulos Floudas at Princeton University, USA.

2.1.7 WP7: Engineering

Within this work-package some applications connected to the solution of complex production networks and quadratic assignment problems have been studied. In the area of production networks, we have considered scheduling problems in the manufacturing industry, where flexible multi-stage production processes are used. One case-study is connected to the food industry and include a plant with several units that can be run either in parallel or stand-alone. The flexibility creates great opportunities for improving the efficiency of the process. The large number of raw materials and products to be scheduled and rescheduled over weekly and monthly horizons introduces thousands of variables and constraints. Different techniques have been used when formulating and solving this large-scale problem. Some of the results can be found in the paper *Improving convergence in a complex production planning problem with two hybrid methods* by Nyberg and Björk published in the *Journal of Information, Intelligence and Knowledge*.

With respect to the quadratic assignment problems, we initially expressed the goal “to create solution methods able to solve larger instances of quadratic assignment problems to global optimality than ever before”. We are happy to report that this very ambitious goal has, at least partially, been fulfilled. The problem library QAPLIB, maintained at the University of Pennsylvania, USA (<http://www.seas.upenn.edu/qaplib>) contains a number of problem instances of which some of the most challenging are still unsolved. In the paper *A New Exact Discrete Linear Reformulation of the Quadratic Assignment Problem* by Nyberg and Westerlund, published in *European Journal of Operational Research*, we were, in fact, able to solve five of the previously unsolved instances.

2.2 Have there been unexpected obstacles during the project?

No major obstacles have been encountered during the reported period.

2.3 Benchmark the CoE with leading research groups in the same scientific area

Optimization and systems engineering is an exciting research area where the diffusion of the theoretical research into many applied areas is proceeding very rapidly. In the past few years, the research efforts in the area have experienced a tremendous growth. In many sub areas, the number of publications in scientific journals have increased drastically. From a bibliometric analysis it is, for example, found that the number of

annually published papers with the keyword MINLP (mixed-integer nonlinear programming) has increased from about 50 in the year 1990 to about 200 in 2000 and to over 1000 a decade later. Many prominent universities with research clusters within this field, such as Carnegie Mellon University (CMU), Princeton University (PU) and Imperial College London (ICL), have been deeply involved in this development.

Thanks to the funding provided for the CoE in OSE, we have been able to bring together the kernel of the expertise in the fields of optimization, applied mathematics, statistics, systems theory and systems engineering at ÅAU. The OSE group has obtained a valuable critical mass and this CoE is today an attractive platform for theoretical research in this area in Finland. Due to the CoE funding, the group has been able to further strengthen the research collaboration with many leading research groups in the world, such as the Center for Advanced Process Decision-making at CMU, the Computer-Aided Systems Laboratory at PU and the Center for Process Systems Engineering at ICL. The OSE group is already producing top-level research, measured on both a national and international scale, and we think the group has the potential of becoming one of the major research groups for the future development of certain areas of the OSE field. Although much has been accomplished during the funding period of the CoE, a long-term commitment is needed to ultimately reach this goal.

2.4 What would be achieved by an unchanged CoE funding for 2013–2014?

If the funding continues as expected for the last two years of the project, all of the PhD students in the project will be able to graduate. It will also allow the group to continue to strengthen research collaborations for achieving the long-term goal discussed in Section 2.3.

2.5 Additional funding

Current external projects, their contributed funding and funding sources are listed in Section 7. In addition, the CoE in OSE has applied for status as a national Center of Excellence from the Academy of Finland for the time-period of 2014-2019. Funding decisions about the National CoEs will be made in November 2012 and in spring 2013. The national CoE application has been done in cooperation with the Department of Mathematics and Statistics at University of Turku, Finland, as well as Princeton University, USA.

Professor Westerlund and Dr. Lundell have also applied for funding for an Academy of Finland project titled *Nonsmooth mixed-integer global optimization* for the duration of September 2013-August 2017, also in cooperation with the Department of Mathematics and Statistics at University of Turku. Professor Salminen and Professor Högnäs have further applied for an Academy of Finland project titled *Markovian models; theory, analysis and applications* for the same duration. Finally, Professor Corander is applying to be an Academy Professor at the Academy of Finland.

2.6 Strategy for the CoE after the funding period (“exit plan”)

The CoE has applied for a status as a national CoE with funding from the Academy of Finland. If the Academy project proposal is accepted, the OSE research group will have guaranteed funding until 2019. If it is not, we will apply for another internal CoE period at ÅAU. Thanks to the collaboration with the Optimization Group at the Department of Mathematics and Statistics at University of Turku, a research cluster in optimization and systems engineering has been formed in Turku, and it will be further tightened.

3 PERSONNEL EMPLOYED USING THE CoE FUNDING

3.1 Senior (postdoctoral) researchers

Name	Gender (M/F)	Academic Degree	Position	Period of employment	Estimated percentage of salary
Kurula, Mikael	M	Ph.D.	Post doc researcher	1.8.2010-31.3.2011	100 %
Lundell, Andreas	M	Ph.D.	Post doc researcher	1.1.2010-	100 %
Pörn, Ray	M	Ph.D.	Post doc researcher	1.8-31.8.2011	100 %

3.2 Doctoral students

Name	Gender (M/F)	Academic Degree	Registration for current PhD studies (start date)	Estimated dissertation year	Period of employment with CoE funding	Estimated percentage of salary
Anckar, Andreas	M	M.Sc.	1.1.2012	2015	1.1-30.6.2012	100 %
Nissfolk, Otto	M	M.Sc.	20.10.2010	2014	1.11.2010-31.12.2013	100 %
Nyberg, Axel	M	M.Sc.	23.8.2010	2013	1.1 -31.12.2013	100 %
Nyberg, Mikael	M	M.Sc.	27.5.2010	2013	1.8.2010-31.7.2013	100 %
Nyman, Henrik	M	M.Sc.	1.1.2010	2014	1.1.2010-31.12.2012	100 %
Pensar, Johan	M	M.Sc.	1.1.2012	2015	1.7-31.12.2012	100 %
Shirdel, Amir	M	M.Sc.	1.4.2011	2014	1.4.2011-31.3.2014	100 %
Skjäl, Anders	M	M.Sc.	31.8.2010	2013	1.9.2010-31.8.2013	100 %

3.3 Other personnel

Name	Gender (M/F)	Academic Degree	Position	Period of employment	Estimated percentage of salary
Fougstedt, Annika	F	M.A.	Secretary	1.1.2010-	10 %
Lundqvist, Kurt	M	Lic.Tech.	Laboratory engineer	1.1.2010-	10 %
Westerlund, Tapio	M	Ph.D.	Project leader	1.1.2010-	10 %

4 RESEARCH RESULTS OF THE CoE

4.1 Scientific publications and other outputs during the period

During the reported period, the number of publications for the OSE group was about 90 including 70 peer-reviewed journal papers, a monograph, four book chapters and 16 refereed articles in conference proceedings. As a sample of the research performed, ten publications have been selected. These have been marked with an asterisk in the list below, and are summarized in the Appendix.

4.1.1 Papers in refereed journals

Authors	Title	Publication year	Journal/Publisher etc.	Additional comments
* Arov, D., M. Kurula, and O. Staffans	Canonical state/signal shift realizations of passive continuous time behaviors	2011	<i>Complex Analysis and Operator Theory</i> 5 (2), pp. 331–402. issn: 1661-8254. doi: 10.1007/s11785-010-0128-8.	

Arov, D. and O. Staffans	Canonical conservative state/signal shift realizations of passive discrete time behaviors	2010	<i>Journal of Functional Analysis</i> 259.12, pp. 3265–3327. issn: 0022-1236. doi: 10.1016/j.jfa.2010.05.019.	
Arov, D. Z. and O. J. Staffans	Symmetries in special classes of passive state/signal systems	2012	<i>Journal of Functional Analysis</i> , 262 (2012), pp. 5021–5097	Available online
Beaumont, M. A., R. Nielsen, C. Robert, J. Hey, O. Gaggiotti, L. Knowles, A. Estoup, M. Panchal, J. Corander, M. Hickerson, S. A. Sisson, N. Fagundes, L. Chikhi, P. Beerli, R. Vitalis, J.-M. Cornuet, J. Huelsenbeck, M. Foll, Z. Yang, F. Rousset, D. Balding, and L. Excoffier.	In defence of model-based inference in phylogeography	2010	<i>Molecular Ecology</i> 19.3, pp. 436–446. issn: 1365-294X. doi: 10.1111/j.1365-294X.2009.04515.x.	
Cheng, L., A. Walker, and J. Corander	Bayesian estimation of bacterial community composition from 454 sequencing data	2012	<i>Nucleic Acids Research</i> 40.12, pp. 5240–5249. doi: 10.1093/nar/gks227.	
Cheng, L., T. Connor, D. Aanensen, B. Spratt, and J. Corander.	Bayesian semi-supervised classification of bacterial samples using MLST databases	2011	<i>BMC Bioinformatics</i> 12.1. issn: 1471-2105. doi: 10.1186/1471-2105-12-302.	
* Christensen, S., P. Salminen, and B. Q. Ta	Optimal stopping of strong Markov processes	2012	<i>Stochastic Processes and their Application</i> .	Accepted for publication
Connor, T. R., J. Corander, and W. P. Hanage	“Population subdivision and the detection of recombination in non-typable <i>Haemophilus influenzae</i>	2012	<i>Microbiology</i> . doi: 10.1099/mic.0.063073-0.	

Corander, J., T. R. Connor, C. A. O'Dwyer, J. S. Kroll, and W. P. Hanage	Population structure in the Neisseria, and the biological significance of fuzzy species	2011	<i>Journal of The Royal Society Interface</i> . doi: 10.1098/rsif.2011.0601.	
* Corander, J., Y. Cui, T. Koski, and J. Sirén	Have I seen you before? Principles of Bayesian predictive classification revisited	2012	<i>Statistics and Computing</i> issn: 0960-3174. doi: 10.1007/s11222-011-9291-7.	Available online
Corander, J., J. Xiong, Y. Cui, and T. Koski	Optimal Viterbi Bayesian predictive classification for data from finite alphabets	2012	<i>Journal of Statistical Planning and Inference</i> . issn: 0378-3758. doi: 10.1016/j.jspi.2012.07.013.	Available online
Corander, J., M. Gyllenberg, and T. Koski	Learning genetic population structures using minimization of stochastic complexity	2010	<i>Entropy</i> 12.5, pp. 1102–1124. issn: 1099-4300. doi: 10.3390/e12051102.	
* Eronen, V.-P., M. M. Mäkelä, and T. Westerlund	On the generalization of ECP and OA methods to nonsmooth convex MINLP problems	2012	<i>Optimization</i> doi: 10.1080/02331934.2012.712118.	Available online
Gauriot, R., L. Gunaratnam, R. Moroni, T. Reinikainen, and J. Corander	Statistical Challenges in the Quantification of Gunshot Residue Evidence	2012	<i>Journal of Forensic Sciences</i>	Accepted for publication
Haan, C. de, R. Kivistö, M. Hakkinen, J. Corander, and M.-L. Hänninen	Multilocus sequence types of Finnish bovine <i>Campylobacter jejuni</i> isolates and their attribution to human infections	2010	<i>BMC Microbiology</i> 10.1, p. 200. issn: 1471-2180. doi: 10.1186/1471-2180-10-200.	
Haan, C. de, K. Lampén, J. Corander, and M.-L. Hänninen	Multilocus sequence types of environmental <i>Campylobacter jejuni</i> isolates and their similarities to those of human, poultry and bovine <i>C. jejuni</i> isolates	2012	<i>Zoonoses and Public Health</i>	In press

Högnäs, G. and L. Allen	Special issue on Stochastic Difference Equations	2012	<i>Journal of Difference Equations and Applications</i>	Editors
Högnäs, G. and B. Jung	Analysis of a stochastic difference equation: Exit times and invariant distributions	2010	<i>Fasciculi Mathematici</i> 44, pp. 69–74.	
Jenkins, H. E., R. B. Aylward, A. Gasasira, C. Donnelly, M. Mwanza, J. Corander, S. Garnier, C. Chauvin, E. Abanida, M. A. Pate, F. Adu, M. Baba, and N. Grassly	Implications of a circulating vaccine-derived poliovirus in Nigeria	2010	<i>New England Journal of Medicine</i> 362.25, pp. 2360–2369. doi: 10.1056/NEJMoa0910074.	
Jern, P., L. Westberg, A. Johansson, L. Jonsson, J. Corander, N.K. Sandnabba, and P. Santtila	Are single nucleotide polymorphisms in the oxytocin and vasopressin 1A/1B receptor genes likely candidates for variation in ejaculatory function?	2012	<i>British Journal of Urology International</i> doi: 10.1111/j.1464-410X.2012.11419.x.	Accepted for publication
Johansson, A., P. Santtila, J. Corander, K. Alanko, P. Jern, B. von der Pahlen, and N. K. Sandnabba	Genetic effects on anger control and their interaction with alcohol intoxication: A self-report study	2010	<i>Biological Psychology</i> 85.2, pp. 291–298. issn: 0301-0511. doi: 10.1016/j.biopsycho.2010.07.016	
Johansson, A., P. Santtila, J. Corander, P. Jern, B. von der Pahlen, M. Varjonen, and K. Sandnabba	Controlling anger in self-reported sober and alcohol intoxicated states: Moderating effects of trait anger and alcohol consumption	2011	<i>Scandinavian Journal of Psychology</i> 52.4 (2011), pp. 382–388. issn: 1467-9450. doi: 10.1111/j.1467-9450.2011.00880.x.	
Karakka, T. and P. Salminen	On fractional Ornstein-Uhlenbeck processes	2011	<i>Communications on Stochastic Analysis</i> 5.1, pp. 121–133.	

Karelahti, J., P. Vainiomäki, and T. Westerlund	Large scale production planning in the stainless steel industry	2011	<i>Industrial and Engineering Chemistry Research</i> 50.9, pp. 4893–4906. doi: 10.1021/ie101376b.	
Konstantopoulou, T., A. Kyprianou, and P. Salminen.	On the excursions of reflected local time processes and stochastic fluid queues	2011	<i>Journal of Applied Probability Spec. Vol. 48A</i>	Accepted for publication
Kurula, M.	On passive and conservative state/signal systems	2010	<i>Integral Equations and Operator Theory</i> 67 (3), pp. 377–424. issn: 0378-620X. doi: 10.1007/s00020-010-1787-6.	
Kurula, M. and O. Staffans	Connections between classical and generalised trajectories of a state/signal system	2011	<i>Complex Analysis and Operator Theory</i> 5 (2), pp. 403–422. issn: 1661-8254. doi: 10.1007/s11785-010-0129-7	
Kurula, M. and O. Staffans	Well-posed state/signal systems in continuous time	2010	<i>Complex Analysis and Operator Theory</i> 4 (2), pp. 319–390. issn: 1661-8254. doi: 10.1007/s11785-009-0021-5.	
Kurula, M., H. J. Zwart, A. J. van der Schaft, and J. Behrndt	Dirac structures and their composition on Hilbert spaces	2010	<i>Journal of Mathematical Analysis and Applications</i> 372.2, pp. 402–422. issn: 0022-247X. doi: 10.1016/j.jmaa.2010.07.004.	
Laajala, T. D., J. Corander, N.M. Saarinen, K. Mäkelä, S. Savolainen, M.I. Suominen, E. Alhoniemi, S. Mäkelä, M. Poutanen, and T. Aittokallio	Improved statistical modeling of tumor growth and treatment effect in pre-clinical animal studies with highly heterogeneous responses in vivo	2012	<i>Clinical Cancer Research</i>	Accepted for publication

Lastusilta, T., L. G. Papageorgiou, and T. Westerlund	A comparative study of solving the problem of module identification in a complex network	2011	<i>Chemical Engineering Transactions</i> 24 (2011), pp. 319–324. doi: 10.3303/CET1124054.	
* Lundell, A., A. Skjäl, and T. Westerlund	A reformulation framework for global optimization	2012	<i>Journal of Global Optimization</i> , pp. 1–27. issn: 0925-5001. doi: 10.1007/s10898-012-9877-4.	
Marttinen, P. and J. Corander	Efficient Bayesian approach for multilocus association mapping including gene-gene interactions	2010	<i>BMC Bioinformatics</i> 11.1. issn: 1471-2105. doi: 10.1186/1471-2105-11-443.	
Marttinen, P., W. P. Hanage, N. J. Croucher, T. R. Connor, S. R. Harris, S. D. Bentley, and J. Corander	Detection of recombination events in bacterial genomes from large population samples	2012	<i>Nucleic Acids Research</i> 40.1, e6. doi: 10.1093/nar/gkr928.	
Medvedev, A. and H. T. Toivonen	Directional sensitivity of continuous least-squares state estimators	2010	<i>Systems and Control Letters</i> 59.9, pp. 571–577. issn: 0167-6911. doi: 10.1016/j.sysconle.2010.07.001.	
Mølgaard, B., T. Hussein, J. Corander, and K. Hämeri	Forecasting size-fractionated particle number concentrations in the urban atmosphere	2012	<i>Atmospheric Environment</i> 46.0, pp. 155–163. issn: 1352-2310. doi: 10.1016/j.atmosenv.2011.10.004.	
Moroni, R., L. Aalberg, T. Reinikainen, and J. Corander	Bayesian adaptive approach to estimating sample sizes for seizures of illicit drugs	2010	<i>Journal of Forensic Sciences</i>	Accepted for publication
Moroni, R., P. Blomstedt, L. Wilhelm, T. Reinikainen, E. Sippola, and J. Corander	Statistical modelling of measurement errors in gas chromatographic analyses of blood alcohol content	2010	<i>Forensic Science International</i> 202.1-3, pp. 71–74. issn: 0379-0738. doi: 10.1016/j.forsciint.2010.04.029.	

Nasution, D. and R. Östermark	The impact of social pressures, locus of control, and professional commitment on auditors' judgment: Indonesian evidence	2012	<i>Asian Review of Accounting</i> 20, pp. 163–178.	
* Nyberg, A. and T.Westerlund	A new exact discrete linear reformulation of the quadratic assignment problem	2012	<i>European Journal of Operational Research</i> 220.2, pp. 314–319. issn: 0377-2217. doi: 10.1016/j.ejor.2012.02.010.	
Nyberg, M. and K.-M. Björk	Improving convergence in a complex production planning problem with two hybrid methods	2012	<i>Journal of Information, Intelligence and Knowledge</i> 3	
Nyman, H., T. Talonen, A. Roine, M. Hupa, and J. Corander	Statistical Approach to Quality Control of Large Thermodynamic Database	2012	<i>Metallurgical and Materials Transactions B</i> 43 (5), pp. 1113–1118. issn: 1073-5615. doi: 10.1007/s11663-012-9679-6.	
Opmeer, M. and O. Staffans	Optimal input-output stabilization of infinite-dimensional discrete time-invariant linear systems by output injection	2010	<i>SIAM Journal on Control Optimization</i> 48.8, pp. 5084–5107. doi: 10.1137/090762233.	
Opmeer, M. and O. Staffans	Coprime factorization and optimal control on the doubly infinite discrete time axis	2012	<i>SIAM Journal on Control Optimization</i> 50, pp. 266–285. doi: 10.1137/110823742.	
Orozco-terWengel, P., J. Corander, and C. Schlötterer	Genealogical lineage sorting leads to significant, but incorrect Bayesian multilocus inference of population structure	2011	<i>Molecular Ecology</i> 6, pp. 1108–1021. issn: 1365-294X. doi: 10.1111/j.1365-294X.2010.04990.x.	

Östermark, R.	A short note on the Growth Potential of risky fund investments	2011	<i>Journal of Derivatives & Hedge Funds</i> 17, pp. 281–290. doi: 10.1057/jdhf.2011.22.	
Östermark, R.	Concurrent processing of heteroskedastic vector-valued mixture density models	2010	<i>Journal of Applied Statistics</i> 37.10, pp. 1637–1659. doi: 10.1080/02664760903121236.	
Östermark, R.	Genetic hybrid tuning of VARMAX and state space algorithms	2010	<i>Soft Computing — A Fusion of Foundations, Methodologies and Applications</i> 14 (1), pp. 91–99. issn: 1432-7643. doi: 10.1007/s00500-008-0393-x.	
Östermark, R.	Hedging with options and cardinality constraints in multi-period portfolio management systems	2011	<i>Kybernetes</i> 40.5, pp. 703–718. doi: 10.1108/03684921111142269.	
Östermark, R.	Incorporating asset growth potential and bear market safety switches in international portfolio decisions	2012	<i>Applied Soft Computing</i> 12.8, pp. 2538–2549. issn: 1568-4946. doi: 10.1016/j.asoc.2012.03.052.	
Östman, F. and H. T. Toivonen	Adaptive cylinder balancing of internal combustion engines	2011	<i>Control Systems Technology, IEEE Transactions on</i> 19.4, pp. 782–791. issn: 1063-6536. doi: 10.1109/TCST.2010.2052925.	
Östman, F. and H. T. Toivonen	Torsional system parameter identification of internal combustion engines under normal operation	2011	<i>Mechanical Systems and Signal Processing</i> 25.4, pp. 1146–1158. issn: 0888-3270. doi: 10.1016/j.ymssp.2010.11.001.	

Pitkänen, L.K., M. Tamminen, A. Hynninen, A. Karkman, J. Corander, A. Kotilainen, and M. Virta	Fish farming affects the abundance and diversity of the mercury resistance gene <i>merA</i> in marine sediments	2011	<i>Microbes and Environments</i> 26.3, pp. 205–211. doi: 10.1264/jsme2.ME11119.	
Robb-Post, C., J. Stamp, M. Brännback, A. Carsrud, and R. Östermark	Do gooders versus good doers: an empirical examination of growth versus efficiency in social entrepreneurship	2011	<i>Frontiers of Entrepreneurship Research</i> 31.	
Salminen, P.	Optimal stopping, Appell polynomials, and Wiener-Hopf factorization	2011	<i>Stochastics An International Journal of Probability and Stochastic Processes</i> 83.4–6, pp. 611–622. doi: 10.1080/17442508.2011.552722.	
Salminen, P. and M. Yor	On hitting times of affine boundaries by reflecting Brownian motion and Bessel processes	2011	<i>Periodica Mathematica Hungarica</i> 62(1), pp. 75–101.	
Sihvonen, L.M., K. Jalkanen, E. Huovinen, S. Toivonen, J. Corander, M. Kuusi, M. Skurnik, A. Siitonen, and K. Haukka	Clinical isolates of <i>Yersinia enterocolitica</i> Biotype 1A represent two phylogenetic lineages with differing pathogenicity-related properties	2012	<i>BMC Microbiology</i>	Accepted for publication
Sirén, J., W.P. Hanage, and J. Corander	Inference on Population Histories by Approximating Infinite Alleles Diffusion	2012	<i>Molecular Biology and Evolution</i>	Accepted for publication
Sirén, J., P. Marttinen, and J. Corander	Reconstructing population histories from single nucleotide polymorphism data	2011	<i>Molecular Biology and Evolution</i> 28.1, pp. 673–683. doi: 10.1093/molbev/msq236.	

* Skjäl, A., T.Westerlund, R. Misener, and C. A. Floudas	A generalization of the classical aBB convex underestimation via diagonal and non-diagonal quadratic terms	2012	<i>Journal of Optimization Theory and Applications</i> 154 (2). doi: 10.1007/s10957-012-0033-6, pp. 462–490. issn: 0022-3239.	
Smýkal, P., N. Bacová-Kerteszová, R. Kalendar, J. Corander, A. Schulman, and M. Pavelek	Genetic diversity of cultivated flax (<i>Linum usitatissimum</i> L.) germplasm assessed by retrotransposon-based markers	2011	<i>TAG Theoretical and Applied Genetics</i> , pp. 1–13. issn: 0040-5752. doi: 10.1007/s00122-011-1539-2.	
Smýkal, P., G. Kenicer, A. J. Flavell, J. Corander, O. Kosterin, R. J. Redden, R. Ford, C. J. Coyne, N. Maxted, M. J. Ambrose, and N. T.H. Ellis	Phylogeny, phylogeography and genetic diversity of the <i>Pisum</i> genus	2010	<i>Plant Genetic Resources</i> , pp. 1–15. doi: 10.1017/S147926211000033X.	Available online
Staffans, O. J.	On scattering passive system nodes and maximal scattering dissipative operators	2012	<i>Proceedings of the American Mathematical Society</i> , doi: 10.1090/S0002-9939-2012-11887-8.	Available online
* Staffans, O. J. and G. Weiss	A physically motivated class of scattering passive linear systems	2012	<i>SIAM Journal on Control and Optimization</i> 50.5 , pp. 3083–3112. doi: 10.1137/110846403.	
Still, C. and T. Westerlund	A linear programming-based optimization algorithm for solving nonlinear programming problems	2010	<i>European Journal of Operational Research</i> 200.3, pp. 658–670. issn: 0377-2217. doi: 10.1016/j.ejor.2009.01.033.	
Sunnåker, M., A. G. Busetto, E. Numminen, J. Corander, M. Foll, and C. Dessimoz	Approximate Bayesian computation	2012	<i>PLoS Computational Biology</i>	Accepted for publication

Tamminen, M., A. Karkman, J. Corander, L. Paulin, and M. Virta	Differences in bacterial community composition in Baltic Sea sediment in response to fish farming	2011	<i>Aquaculture</i> 313.1–4, pp. 15–23. issn: 0044-8486. doi: 10.1016/j.aquaculture.2011.01.020.	
* Toivonen, H. T., A. Meinander, T. Asaoka, M. Westerlund, F. Pettersson, A. Mikhailov, J. E. Eriksson, and H. Saxén	Modeling reveals that dynamic regulation of c-FLIP levels determines cell-to-cell distribution of CDd95-mediated apoptosis	2011	<i>Journal of Biological Chemistry</i> 286.21, pp. 18375–18382. doi: 10.1074/jbc.M110.177097.	
Westerlund, T., A. Lundell, and J. Westerlund	Some notes on convex relaxations	2011	<i>AIDIC Conference Series</i> . Ed. by S. Pierucci. Vol. 10. Aidic Servizi Srl, 2011, pp. 383–392. doi: 0.3303/ACOS1110042.	
Willems, R.J. L., J. Top, W. van Schaik, H. Leavis, M. Bonten, J. Sirén, W. P. Hanage, and J. Corander	Restricted gene flow among hospital subpopulations of <i>Enterococcus Faecium</i>	2012	<i>mBio</i> 3	Accepted for publication

4.1.2 Papers in *refereed* edited volumes and conference proceedings

Authors	Title	Publication year	Publisher, Conference etc.	Additional comments
Arov, D., M. Kurula, and O. Staffans	Passive behaviors and their passive state/signal realizations in continuous time	2010	<i>Proceedings of the 19th International Symposium on Mathematical Theory of Networks and Systems – MTNS 2010</i>	
Corander, J., Y. Cui, and Koski T.	Inductive inference and partition exchangeability in classification	2011	<i>Springer Lecture Notes in Artificial Intelligence (LNAI)</i>	

Corander, J., T. Koski, T. Pavlenko, and A. Tillander	Bayesian Block-diagonal Predictive Classifier for Gaussian data	2012	<i>Advances in Intelligent and Soft Computing</i>	Accepted for publication
Hietarinta, L. and H. T. Toivonen	Least-squares support vector regression for identification of periodic systems	2010	<i>Proceedings IFAC Workshop on Periodic Control Systems (PSYCO 2010)</i>	
Kian, Q., A. Di Lieto, J. Corander, P. Auvinen, and D. Greco	Reanalysis of bipolar disorder and schizophrenia gene expression complements the Kraepelian dichotomy	2011	<i>Proceedings of the 11th International Conference on Systems Biology, ICSB 2010</i>	
Lastusilta, T. and T. Westerlund	A comparative study of solving Quadratic Assignment Problems using some standard MINLP solvers	2011	<i>SIMULTECH 2011 1st International Conference on Simulation and Modeling Methodologies, Technologies and Applications</i>	
Lundell, A. and T. Westerlund	Finding an optimized set of transformations for convexifying non-convex MINLP problems	2012	<i>Proceedings of the 11th International Symposium on Process Systems Engineering 15-19 July 2012, Singapore. Ed. by I.A. Karimi and R. Srinivasan.</i>	
Mølgaard, B., T. Hussein, J. Corander, and K. Hämmäri	Bayesian forecast model for size-fractionated urban particle number concentrations	2011	<i>Proceedings of the Finnish Center of Excellence and Graduate School in 'Physics, Chemistry, Biology and Meteorology of Atmospheric Composition and Climate Change'. Ed. by M. Kulmala, J. Bäck, H. K. Lappalainen, and T. Nieminen. Report series in aerosol science 126. 2011, pp. 270–271.</i>	

Nissfolk, O., R. Pörn, T. Westerlund, and F. Jansson	A Mixed Integer Quadratic Reformulation of the Quadratic Assignment Problem with Rank-1 Matrix	2012	<i>Proceedings of the 11th International Symposium on Process Systems Engineering</i> 15-19 July 2012, Singapore. Ed. by I.A. Karimi and R. Srinivasan.	
Nyberg, M. and K.-M. Björk	Solving a complex production planning problem with mathematical programming techniques	2010	<i>Proceedings of ICAOR 10</i>	
Pörn, R., O. Nissfolk, F. Jansson, and T. Westerlund	The Coulomb glass – modeling and computational experience with a large scale 0-1 QP problem	2011	<i>21st European Symposium on Computer Aided Process Engineering</i> . Ed. by E.N. Pistikopoulos, M.C. Georgiadis, and A.C. Kokossis. Vol. 29. Computer Aided Chemical Engineering. Elsevier, pp. 658–662. doi: 10.1016/B978-0-444-53711-9.50132-2.	
Skjäl, A., A. Lundell, and T. Westerlund	Global optimization of C^2 constraints by convex reformulations	2011	<i>Chemical Engineering Transactions</i> 24 (2011), pp. 373–378. doi: 10.3303/CET1124063.	
Skjäl, A., R. Misener, and C. A. Floudas	A generalization of classical aBB underestimation to include bilinear terms	2012	<i>22nd European Symposium on Computer Aided Process Engineering</i> . Ed. by D. Bogle and M. Fairweather. Vol. 30. Computer Aided Chemical Engineering. Submitted. Elsevier, pp. 1202–1206.	

Staffans, O.	The Reciprocal Symmetry in State/Signal Systems in Continuous Time	2010	<i>Proceedings of the 19th International Symposium on Mathematical Theory of Networks and Systems – MTNS 2010</i>	
Toivonen, H. T. and L. Hietarinta	Floquet-type representations of linear periodic systems – general case	2011	<i>Proceedings of the 18th IFAC World Congress, 2011</i> . Ed. by S. Bittanti, A. Cenedese, and S. Zampieri. Vol. 18. doi: 10.3182/20110828-6-IT-1002.02385.	
Westerlund, T., A. Lundell, and J. Westerlund	On convex relaxations in nonconvex optimization	2011	<i>Chemical Engineering Transactions</i> 24, pp. 331–336. doi: 10.3303/CET1124056.	

4.1.3 Scientific monographs

Authors	Title	Publication year	Publisher etc.	Additional comments
* Högnäs, G. and A. Mukherjea	<i>Probability Measures on Semigroups. Convolution Products, Random Walks and Random Matrices</i>	2011	Springer Science+Business Media, New York	Second edition

4.1.4 Book chapters and other scientific publications

Authors	Title	Publication year	Journal/Publisher etc.	Additional comments
Arov, D., M. Kurula, and O. Staffans	Boundary control state/signal systems and boundary triplets	2012	Cambridge University Press	Chapter in: <i>Boundary Relations, Extension Theory, and Applications</i>
Arov, D., M. Kurula, and O. Staffans	Passive state/signal system and conservative boundary relations	2012	Cambridge University Press	Chapter in: <i>Boundary Relations, Extension Theory, and Applications</i>
Budzbán, G. and G. Högnäs	Convolution products of probability measures on a compact semi-group with applications to random measures	2010	American Mathematical Society	Chapter in: <i>Algebraic Methods in Statistics and Probability II</i>
Eronen, V.-P., M. Mäkelä, and T. Westerlund	Nonsmooth Extended Cutting-Plane Method for Generally Convex MINLP Problems	2012	TUCS Technical Report	
Eronen, V.-P., M. Mäkelä, and T. Westerlund.	On convex nonsmooth MINLP	2010	TUCS Technical Report	
Högnäs, G.	Special issue on Stochastic Difference Equations, Journal of Difference Equations and Applications 18, No. 8	2012	Taylor & Francis	Guest editor
Lundell, A. and T. Westerlund	Global optimization of mixed-integer signomial programming problems	2012	Springer, New York	Chapter in: <i>Mixed Integer Non-linear Programming. Vol. 154. The IMA Volumes in Mathematics and its Applications.</i>

4.2 Academic theses completed in CoE funded subjects

4.2.1 Doctoral theses

Name	Title of thesis	Publication year	Impact
Kurula, Mikael	Towards input/output-free modelling of linear infinite-dimensional systems in continuous time	2010	5
Lastusilta, Toni	GAMS MINLP solver comparisons and some improvements to the AlphaECP algorithm	2011	5
Lindberg, Peter	The knapsack problem approach in solving partial hedging problems of options	2012	4

4.2.2 Licentiate theses

No licentiate theses has been published during the reported period.

4.2.3 Master's theses

Name	Title of thesis	Publication year	Impact
Anckar, Andreas	Hausdorffdimensionen av mått	2011	4
Neovius, Jesper	A two-tier optimization model for a production-distribution network, a case study	2010	4
Niemelä, Max	En komparativ studie om binärklassificerare med tillämpning inom kriminologi	2012	3
Nissfolk, Otto	Minimering av totalenergin för elektronkonfigurationen i oordnade material	2010	5
Nyberg, Axel	En modification av extended cutting plane algoritmen för lösandet av kvadratiska tilldelningsproblem	2010	5
Pensar, Johan	Labeled Directed Acyclic Graphs	2011	5

Rosama, Noora	Design of radial flow reactors	2011	4
Sundqvist, Dennis	Mapping of heat streams and heat pumps in heat recovery systems for paper machines	2010	4
Storbjörk, Jan	Modelling and Simulation of Power Plant Generators	2012	4

5 OTHER SCIENTIFIC ACTIVITIES

5.1 Cooperation

5.1.1 Visits to other research groups by the CoE members

CoE member	Research group	University	Duration of stay
Corander, Jukka	Prof. William P Hanage	Harvard School of Public Health	29.10-5.11.2010 3.7-1.8.2012
Corander, Jukka	Prof. Timo Koski	Royal Institute of Technology	10.6.-31.7.2011.
Corander, Jukka	Prof. Ziheng Yang	University College London/Chinese Academy of Sciences, Beijing	6-14.11.2010
Glader, Christer	Institut fuer Angewandte Analysis	TU Bergakademie, Freiberg, Germany	14-23.6.2010
Glader, Christer	Institut fuer Angewandte Analysis	TU Bergakademie, Freiberg, Germany	18-27.6.2012
Kurula, Mikael	Hybrid Systems, Department of Applied Mathematics	University of Twente, The Netherlands	1.2.2011-30.6.2012
Nyberg, Axel	Grossmann Research Group	Carnegie Mellon University, USA	13.1-13.6.2012
Skjäl, Anders	Computer-Aided Systems Laboratory (CASL)	Princeton University, USA	28.3-8.4.2011
Staffans, Olof	Mathematisches Forschungsinstitut Oberwolfach	Germany	11-18.12.2010

Staffans, Olof	Prof. George Weiss	Tel Aviv University, Israel	14-24.10.2010
Staffans, Olof	Mathematical Control Theory group, Prof. Mark Opmeer	University of Bath, UK	3-12.5.2010

5.1.2 Visits to the CoE by scientists from other research groups

Name	Academic degree/Position	University	Duration of stay
Arov, Damir	Professor	South Ukrainian State Pedagogical University, Odessa	3 months 2010, 2 months 2011, 2 months 2012
Borodin, Andrei	Professor	St Petersburg State University, Russia	24.4-4.5.2010
Bussieck, Michael	Ph.D. / Senior Research Analyst	GAMS Development Corporation	30.11-2.12.2011
Floudas, Christodoulos	Professor of Chemical and Biological Engineering	Princeton University, USA	2-4.11.2010
Grossmann, Ignacio	Professor of Chemical Engineering	Carnegie Mellon University	7-9.12.2011
Kravanja, Zdravko	Professor of Chemical Engineering	University of Maribor	1-2.12.2011
Yor, Marc	Professor	Université Paris VI	8-15.8.2010

5.1.3 Seminars or other larger meetings arranged by the CoE

Event	Target group	Location	Date
Annual OSE seminar 2011	OSE group members, public	Åbo Akademi University, Turku, Finland	8.12.2011
Annual OSE seminar 2010	OSE group members, public	Åbo Akademi University, Turku, Finland	3.11.2010

5.2 Expert assignments by CoE members and achieved awards

5.2.1 Scientific positions of trust

CoE member	Title	Journal/organization etc.	Term
Corander, Jukka	Associate editor	Scandinavian Journal of Statistics	2009-
Högnäs, Göran	Associate editor	Journal of Theoretical Probability	1988-
Högnäs, Göran	Associate editor	Journal of Difference Equations and Applications	2008-
Högnäs, Göran	Board member	Estonian Doctoral School of Mathematics and Statistics	2009-
Högnäs, Göran	Board member	Finnish National Graduate School in Mathematics and its Applications	2012-
Högnäs, Göran	Chairman	Collegial Body of Åbo Akademi University	2010-2012
Salminen, Paavo	Chairman of the Board, Director	Finnish Doctoral Programme in Stochastic and Statistics	2010-
Salminen, Paavo	Editor in chief	Scandinavian Journal of Statistics	2010-2012
Staffans, Olof	Board member	Finnish National Graduate School in Mathematics and its Applications	2010-2011
Westerlund, Tapio	Board member	MINLP World	2001-
Westerlund, Tapio	Editorial board member	International Journal of Chemical Engineering	2008-
Westerlund, Tapio	Associate editor, board member	Journal of Global Optimization	2012-
Westerlund, Tapio	Board member	National Graduate School in Energy Science and Technology	2002-2010

Members of the group have also been peer reviewers for a number of journals in the fields, *e.g.* Automatica, Computers and Chemical Engineering, Control Engineering Practice, European Journal of Operational Research, IEEE Transactions on Control Systems Technology, IET Control Theory & Applications, INFORMS Journal on Computing, International Journal of Control, Journal of Global Optimization, Journal of Mechanical Engineering Science, Journal of Process Control, Mechanical Systems and Signal Processing, SIAM Journal on Control and Optimization, Systems and Control Letters.

5.2.2 External PhD examinations

CoE member	Doctoral candidate	University	Dissertation date
Corander, Jukka	Nettelblad, Carl	Uppsala University	2012
Högnäs, Göran	Service, Robert	University of Helsinki, Finland	2012
Toivonen, Hannu	Komperød, Magnus	Telemark University College	2012
Toivonen, Hannu	Mendelson, Alexander	Aalto University	2010
Westerlund, Tapio	Sanchez, Luis Perez	University of Pisa, Italy	2010
Westerlund, Tapio	Laukkanen, Timo	Aalto University School of Science and Technology, Helsinki, Finland	2012

5.2.3 Membership in review panels for academic or other positions

CoE member	Position reviewed	University or other organization	Year
Högnäs, Göran	Assistant professor	University of Texas Pan American, Edinburg, USA	2010
Högnäs, Göran	Assistant professor	University of Linköping, Sweden	2011
Högnäs, Göran	Adjunct professor	University of Jyväskylä, Finland	2011
Salminen, Paavo	Associate professor	KTH Royal Institute of Technology, Stockholm, Sweden	2012
Westerlund, Tapio	Adjunct professor in Applied Mathematics	University of Turku, Finland	2011
Westerlund, Tapio	Professor in Process Plant Design	Aalto University School of Science and Technology, Helsinki, Finland	2010

5.2.4 Membership in ad hoc evaluation panels

CoE member	Evaluation	University or other organization	Year
Corander, Jukka	Center of Excellence application in mathematics	Research Council of Norway	2011
Corander, Jukka	Grant review panel for mathematics and computational sciences	Research Council of Norway	2012
Westerlund, Tapio	Application: "Nonlinear Global Optimization"	National Research Foundation, South Africa	2010

5.2.5 Awards and honors

CoE member	Award or honor	Awarding University or other organization	Year
Högnäs, Göran	Knight 1 st class of the Order of the White Rose of Finland	President of the Republic of Finland	2011
Kurula, Mikael	Postdoctoral fellowship 25.000 €	Ruth and Nils-Erik Stenbäck Foundation	2012
Staffans, Olof	Professor E.J. Nyström prize 20.000 €	The Finnish Society of Sciences and Letters	2010

5.3 Other possible scientific activities

Professor Jukka Corander is a representative in the European Science Foundation for physical sciences and engineering in the management committee of ESF Forward Look in personalized medicine (www.esf.org/iPM). Furthermore, senior members of the OSE group have during the current project-period been active in several program committees, scientific advisory boards etc. for international conferences, such as: Complex Systems Design and Management (CSDM) 2010, Paris; 10th International Conference on Chemical and Process Engineering (ICheap-10), Florence Italy, 2011; 21st European Conference on Process Systems Engineering (ESCAPE21) Greece 2011; 22nd European Conference on Process Systems Engineering (ESCAPE22) in UK London 2012; Global Optimization Workshop (GOW'12) in Natal, Brazil 2012; Foundations of Computer-Aided Process Operations (FOCAP 2012) in Georgia USA 2012; the 3rd Nordic-Baltic Biometric Conference 2011, NBBC11, Turku, Finland, and the 23rd European Conference on Process Systems Engineering (ESCAPE23) in Finland 2013 (384 submissions from 46 countries).

6 POPULAR SCIENCE AND PUBLICITY

Event	Additional description	Date
Volvo Ocean Race Game	PhD student Mikael Nyberg in the OSE group recently won an lap in the world's largest virtual regatta (Volvo Ocean Race Game), with over 178.000 registered boats. He took help of one hundred year old weather maps as well as his knowledge in optimization. The prize was a trip to New Zealand. Featured also in MfÅA (Meddelanden från Åbo Akademi , the internal magazine of ÅAU) 4/2012.	2.3.2012
Annual OSE seminar 2011	The annual OSE seminar 2011 featured in MfÅA 2/2012.	3.2.2012
Interview with Prof. Ralf Östermark	Professor Ralf Östermark is interviewed about fast parallel computations in MfÅA 17/2011.	25.11.2011
Interview with Dr. Andreas Lundell	Post doc researcher Andreas Lundell presents his research within OSE in MfÅA 17/2011.	25.11.2011
"25 hottest articles" in Discrete Optimization	The article: "Global Solution of Optimization Problems with Signomial Parts, Discrete Optimization, 5, Issue 1, February 2008, 108-120" from the OSE group on place five among the "Top 25 Hottest Articles" in Discrete Optimization during January to March 2011.	7.6.2011
Broadcasting company YLE reporting on on-going research about the spreading of polio	Finland's national public service broadcasting company YLE is reporting about the research in the spreading of polio published in The New England Journal of Medicine by professor Corander et al.	4.3.2011
Article about the OSE research	OSE Research presented in Åbo Akademi News Bulletin 1/2011.	1.2.2011
Article about advances in QAP	Advances in solving Quadratic Assignment Problems within OSE is presented in MfÅA 1/2011.	21.1.2011
"Top 10 cited articles" in Computers and Chemical Engineering	Adjoint professor Iiro Harjunoski, is co-author of the article "State-of-the-art review of optimization methods for short-term scheduling of batch processes" listed as the most cited article published in CACE in the last five years. The other authors are C.A. Méndez, J. Cerdá, J. M. Fahl och I.E. Grossmann.	12.1.2011
Article about the annual OSE seminar 2010	Featured in MfÅA 16/20102	15.11.2010

7 ADDITIONAL RESEARCH FUNDING

Granting organization	CoE Project leader	Project title	Amount of funding	Time span
Academy of Finland	Corander, Jukka	BioBayes: advanced statistical tools for molecular biology	318.000 €	1.1.2008-31.12.2011
Academy of Finland	Salminen, Paavo	Applied Markov processes -- fluid queues, optimal stopping and population dynamics	375.000 €	1.1.2009-31.12.2011
Academy of Finland	Salminen, Paavo	The Finnish Graduate School in Stochastics and Statistics	146.200 €	1.1.2010-31.12.2013
Academy of Finland	Salminen, Paavo	Doctoral programme in stochastics and statistics	87.030 €	1.1.2012-31.12.2015
Academy of Finland	Staffans, Olof	Passive and conservative state/signal systems in discrete and continuous time	126.860 €	1.8.2009-31.12.2010
Academy of Finland	Toivonen, Hannu	Controller design of linear time-periodic systems with specific applications in flexural and torsional vibration control of electrical machines	110.000 €	1.1.2009-31.12.2012
Academy of Finland	Westerlund, Tapio	Large Scale Mixed-Integer Global Optimization	330.000 €	1.1.2009-31.12.2012
EU ISEC programme	Corander, Jukka	Statistical linking of amphetamine samples using raw massspectrometer data	45.000 €	1.7.2012-31.1.2013
Outotec Research Oy	Corander, Jukka	Statistical prediction of thermodynamical properties of molecules in computational chemistry	425.700 €	1.1.2009-31.12.2013

Appendix

The following ten publications (nine journal papers and one monograph) have been selected as a sample of the research performed in the CoE in Optimization and Systems Engineering. The publications are attached to the report as PDF-files.

Arov, D., M. Kurula, and O. Staffans	Canonical state/signal shift realizations of passive continuous time behaviors	2011	<i>Complex Analysis and Operator Theory</i> 5 (2), pp. 331–402. issn: 1661-8254. doi: 10.1007/s11785-010-0128-8.
Christensen, S., P. Salminen, and B. Q. Ta	Optimal stopping of strong Markov processes	2012	<i>Stochastic Processes and their Application</i> . Accepted for publication.
Corander, J., Y. Cui, T. Koski, and J. Sirén	Have I seen you before? Principles of Bayesian predictive classification revisited	2012	<i>Statistics and Computing</i> issn: 0960-3174. doi: 10.1007/s11222-011-9291-7.
Eronen, V.-P., M. M. Mäkelä, and T. Westerlund	On the generalization of ECP and OA methods to nonsmooth convex MINLP problems	2012	<i>Optimization</i> doi: 10.1080/02331934.2012.712118.
Lundell, A., A. Skjäl, and T. Westerlund	A reformulation framework for global optimization	2012	<i>Journal of Global Optimization</i> , pp. 1–27. issn: 0925-5001. doi: 10.1007/s10898-012-9877-4.
Nyberg, A. and T. Westerlund	A new exact discrete linear reformulation of the quadratic assignment problem	2012	<i>European Journal of Operational Research</i> 220.2, pp. 314–319. issn: 0377-2217. doi: 10.1016/j.ejor.2012.02.010.
Skjäl, A., T. Westerlund, R. Misener, and C. A. Floudas	A generalization of the classical aBB convex underestimation via diagonal and non-diagonal quadratic terms	2012	<i>Journal of Optimization Theory and Applications</i> 154 (2). doi: 10.1007/s10957-012-0033-6, pp. 462–490. issn: 0022-3239.
Staffans, O. J. and G. Weiss	A physically motivated class of scattering passive linear systems	2012	<i>SIAM Journal on Control and Optimization</i> 50.5 , pp. 3083–3112. doi: 10.1137/110846403.
Toivonen, H. T., A. Meinander, T. Asaoka, M. Westerlund, F. Pettersson, A. Mikhailov, J. E. Eriksson, and H. Saxén	Modeling reveals that dynamic regulation of c-FLIP levels determines cell-to-cell distribution of CD95-mediated apoptosis	2011	<i>Journal of Biological Chemistry</i> 286.21, pp. 18375–18382. doi: 10.1074/jbc.M110.177097.
Högnäs, G. and A. Mukherjea	<i>Probability Measures on Semigroups. Convolution Products, Random Walks and Random Matrices</i>	2011	Springer Science+Business Media, New York

Canonical State/Signal Shift Realizations of Passive Continuous Time Behaviors

Damir Z. Arov · Mikael Kurula · Olof J. Staffans

Received: 30 August 2010 / Accepted: 22 December 2010 / Published online: 25 January 2011
© Springer Basel AG 2011

Abstract This work is devoted to the construction of canonical passive and conservative state/signal shift realizations of arbitrary passive continuous time behaviors. By definition, a passive future continuous time behavior is a maximal nonnegative right-shift invariant subspace of the Kreĭn space $L^2([0, \infty); \mathcal{W})$, where $\mathcal{W}$ is a Kreĭn space, and the inner product in $L^2([0, \infty); \mathcal{W})$ is the one inherited from $\mathcal{W}$. A state/signal system $\Sigma = (V; \mathcal{X}, \mathcal{W})$, with a Hilbert state space $\mathcal{X}$ and a Kreĭn signal space $\mathcal{W}$, is a dynamical system whose classical trajectories (x, w) on $[0, \infty)$ satisfy $x \in C^1([0, \infty); \mathcal{X})$, $w \in C([0, \infty); \mathcal{W})$, and

$$(\dot{x}(t), x(t), w(t)) \in V, \quad t \in [0, \infty),$$

where the generating subspace V is a given subspace of the node space $\mathfrak{K} := \mathcal{X} \times \mathcal{X} \times \mathcal{W}$. Passivity of this systems means that V is maximal nonnegative with respect to a certain Kreĭn space inner product on $\mathfrak{K}$, and that $(z, 0, 0) \in V$ implies $z = 0$. We

Communicated by Bernd Kirstein.

D. Z. Arov thanks Åbo Akademi for its hospitality and the Academy of Finland and the Magnus Ehrnrooth Foundation for their financial support during his visits to Åbo in 2003–2010.

D. Z. Arov
Division of Applied Mathematics and Informatics, Institute of Physics and Mathematics,
South-Ukrainian Pedagogical University, 65020 Odessa, Ukraine

M. Kurula · O. J. Staffans (✉)
Department of Mathematics, Åbo Akademi University, 20500 Åbo, Finland
e-mail: olof.staffans@abo.fi
URL: <http://users.abo.fi/staffans/>

M. Kurula
URL: <http://users.abo.fi/mkurula/>

present three canonical passive shift models: (a) an observable and co-energy preserving model, (b) a controllable and energy preserving model, and (c) a simple conservative model. In order to construct these models we first introduce the notions of the input map, the output map, and the past/future map of a passive state/signal system. Our canonical passive state/signal shift realizations are analogous to the corresponding de Branges–Rovnyak type input/state/output realizations of a given Schur function.

Keywords Passive · Conservative · Behavior · State/signal system · de Branges–Rovnyak model · Input/state/output system · Input map · Output map · Past/future map · Transfer function · Krein space

List of Symbols

$\mathbb{R}, \mathbb{R}^+, \mathbb{R}^-$	$\mathbb{R} := (-\infty, \infty), \mathbb{R}^+ := [0, \infty), \mathbb{R}^- = (-\infty, 0]$
$\mathbb{Z}, \mathbb{Z}^+, \mathbb{Z}^-$	$\mathbb{Z} = \{0, \pm 1, \pm 2, \dots\}, \mathbb{Z}^+ = \{0, 1, 2, \dots\},$ $\mathbb{Z}^- = \{-1, -2, \dots\}$
$\overline{\Omega}$	The closure of Ω
$\mathcal{B}(\mathcal{U}; \mathcal{Y})$	The space of bounded linear operators from $\mathcal{U}$ to $\mathcal{Y}$
$\text{dom}(A), \text{im}(A), \text{ker}(A)$	The domain, range, and kernel of the operator A
$A _{\mathcal{Z}}$	The restriction of the operator A to $\mathcal{Z}$
$1_{\mathcal{X}}$	The identity operator on $\mathcal{X}$
$(\cdot, \cdot)_{\mathcal{X}}$	The inner product in the Hilbert space $\mathcal{X}$
$[\cdot, \cdot]_{\mathcal{W}}$	The inner product in the Kreĭn space $\mathcal{W}$
$-\mathcal{K}$	The anti-space of the Kreĭn space $\mathcal{K}$. This is the same topological vector space as $\mathcal{K}$, but it has a different inner product $[\cdot, \cdot]_{-\mathcal{K}} := -[\cdot, \cdot]_{\mathcal{K}}$
τ^t	$(\tau^t w)(s) = w(s + t), s, t \in \mathbb{R}$ (this is a left shift if $t > 0$)
τ_+^t	$(\tau_+^t w)(s) = w(s + t), s, t \in \mathbb{R}^+$ (this is a left shift if $t > 0$)
τ_-^t	$(\tau_-^t w)(s) = w(s + t)$ if $s + t \leq 0$, $(\tau_-^t w)(s) = 0$ if $s + t > 0$. Here $s \in \mathbb{R}^-, t \in \mathbb{R}^+$
τ^{*t}	$(\tau^{*t} w)(s) = (\tau^{-1} w)(s) = w(s - t), s, t \in \mathbb{R}$ (this is a right shift if $t > 0$)
τ_+^{*t}	$(\tau_+^{*t} w)(s) = w(s - t)$ if $s - t \geq 0$ and $(\tau_+^{*t} w)(s) = 0$ if $s - t < 0$. Here $s, t \in \mathbb{R}^+$
τ_-^{*t}	$(\tau_-^{*t} w)(s) = w(s - t)$ for all $s \in \mathbb{R}^-, t \in \mathbb{R}^+$
π_I, π_+, π_-	$(\pi_I w)(s) = w(s)$ if $s \in I$, $(\pi_I w)(s) = 0$ if $s \notin I$. We abbreviate $\pi_- = \pi_{\mathbb{R}^-}$ and $\pi_+ = \pi_{\mathbb{R}^+}$
$C(I; \mathcal{X}), BUC(I; \mathcal{X}), C^1(I; \mathcal{X})$	The spaces of continuous, bounded uniformly continuous, or continuously differentiable functions, respectively, on I with values in $\mathcal{X}$, with the standard norms

$L_{\text{loc}}^2(I; \mathcal{W})$	The space of functions from I to $\mathcal{W}$ which belong locally to L^2
$K^2(I; \mathcal{X}), K^2(\mathcal{W}), K_+^2(\mathcal{W}), K_-^2(\mathcal{W})$	See (1.10) and (1.11)
$\mathcal{H}(\mathcal{Z}), \mathcal{H}^0(\mathcal{Z})$	See Sect. 2.2
$\mathfrak{W}, \mathfrak{W}_+, \mathfrak{W}_-$	A passive two-sided, future, or past behavior, respectively, on the Kreĭn signal space $\mathcal{W}$
$\mathcal{H}_+, \mathcal{H}(\mathfrak{W}_+)$	$\mathcal{H}_+ := \mathcal{H}(\mathfrak{W}_+)$ is defined in Theorem 5.1
$\mathcal{Q}_+$	$\mathcal{Q}_+ : w \mapsto w + \mathfrak{W}_+$ is the quotient map $K_+^2(\mathcal{W}) \mapsto K_+^2(\mathcal{W})/\mathfrak{W}_+$
$\mathcal{H}_+^0, \mathcal{H}^0(\mathfrak{W}_+)$	$\mathcal{H}_+^0 := \mathcal{H}^0(\mathfrak{W}_+) := \mathcal{Q}_+ \mathfrak{W}_+^{[\perp]}$
$\mathcal{K}(\mathfrak{W}_+)$	$\mathcal{K}(\mathfrak{W}_+) := \mathcal{Q}_+^{-1} \mathcal{H}(\mathfrak{W}_+)$
$\mathcal{H}_-, \mathcal{H}(\mathfrak{W}_-^{[\perp]})$	$\mathcal{H}_- := \mathcal{H}(\mathfrak{W}_-^{[\perp]})$ is defined in Theorem 5.4
$\mathcal{Q}_-$	$\mathcal{Q}_- : w \mapsto w + \mathfrak{W}_-^{[\perp]}$ is the quotient map $K_-^2(\mathcal{W}) \mapsto K_-^2(\mathcal{W})/\mathfrak{W}_-^{[\perp]}$
$\mathcal{H}_-^0, \mathcal{H}^0(\mathfrak{W}_-)$	$\mathcal{H}_-^0 := \mathcal{H}^0(\mathfrak{W}_-) := \mathcal{Q}_- \mathfrak{W}_-^{[\perp]}$
$\mathcal{K}(\mathfrak{W}_-)$	$\mathcal{K}(\mathfrak{W}_-) := \mathcal{Q}_-^{-1} \mathcal{H}(\mathfrak{W}_-^{[\perp]})$
$\Gamma_{\mathfrak{W}}$	The past/future map of $\mathfrak{W}$. See Definition 5.8 and Sect. 7
$\mathcal{D}(\mathfrak{W})$	$\mathcal{D}(\mathfrak{W})$ is defined before Lemma 5.9
$\mathcal{Q}$	The quotient map $K^2(\mathcal{W}) \mapsto K^2(\mathcal{W})/(\mathfrak{W}_+ + \mathfrak{W}_-^{[\perp]})$
$\mathcal{D}^0(\mathfrak{W})$	$\mathcal{D}^0(\mathfrak{W}) = \mathcal{Q}(\mathfrak{W} + \mathfrak{W}_-^{[\perp]})$
$\mathcal{L}(\mathfrak{W})$	$\mathcal{L}(\mathfrak{W}) := \mathcal{Q}^{-1} \mathcal{D}(\mathfrak{W})$
P_+	The projection of $K^2(\mathcal{W})/(\mathfrak{W}_+ + \mathfrak{W}_-^{[\perp]})$ onto $K_+^2(\mathcal{W})/\mathfrak{W}_+$
P_-	The projection of $K^2(\mathcal{W})/(\mathfrak{W}_+ + \mathfrak{W}_-^{[\perp]})$ onto $K_-^2(\mathcal{W})/\mathfrak{W}_-^{[\perp]}$
$\mathcal{L}_{\pm}$	See Lemma 5.9
$\Pi_{\pm}$	The orthogonal projection of $\mathcal{H}_- \oplus \mathcal{H}_+$ onto $\mathcal{H}_{\pm}$
$\mathfrak{B}_{\Sigma}, \mathfrak{C}_{\Sigma}$	The input and output maps are defined in Sect. 6
$\mathfrak{G}_+^{\Sigma}$	See (6.3)

Contents

1 Introduction	334
2 Kreĭn Spaces	341
2.1 Some Kreĭn Space Results	341
2.2 The Hilbert Space $\mathcal{H}(\mathcal{Z})$	346

3	Passive and Conservative S/S Systems	347
3.1	Basic Properties of Trajectories of Passive S/S Systems	347
3.2	Classical Trajectories and the Generating Subspace	353
3.3	More on Externally Generated Stable Trajectories	356
3.4	Passive Past, Two-sided, and Future Behaviors	358
3.5	Intertwined Systems	365
4	Anti-Passive Adjoint State/Signal Systems	366
4.1	Anti-Passive State/Signal Systems	366
4.2	Anti-Passive Behaviors	370
5	The Hilbert Spaces $\mathcal{H}(\mathfrak{M}_+)$, $\mathcal{H}(\mathfrak{M}_-^{[1]})$, and $\mathcal{D}(\mathfrak{M})$	372
5.1	The Hilbert Space $\mathcal{H}_+ := \mathcal{H}(\mathfrak{M}_+)$	372
5.2	The Hilbert Space $\mathcal{H}_- := \mathcal{H}(\mathfrak{M}_-^{[1]})$	374
5.3	The Past/Future Map $\Gamma_{\mathfrak{M}}$ and the Hilbert Space $\mathcal{D}(\mathfrak{M})$	375
6	The Output and Input Maps	380
6.1	The Output Map $\mathfrak{C}_{\Sigma}$	380
6.2	The Input Map $\mathfrak{B}_{\Sigma}$	382
6.3	The Adjoints of $\mathfrak{C}_{\Sigma}$ and $\mathfrak{B}_{\Sigma}$	385
6.4	The Backward Reachable and Unobservable Subspaces	386
7	The Past/Future Map of a Passive System	386
8	The Canonical Co-Energy Preserving Model	389
9	The Canonical Energy Preserving Model	393
10	The Canonical Simple Conservative Model	396

1 Introduction

A linear continuous time invariant s/s (state/signal) system $\Sigma = (V; \mathcal{X}, \mathcal{W})$ has a Hilbert (state) space, a Kreĭn (signal) space, and a closed (generating) subspace V of the (node) space $\mathfrak{K} = \begin{bmatrix} \mathcal{X} \\ \mathcal{W} \end{bmatrix}$ that satisfies some additional conditions, among them the condition

$$\begin{bmatrix} z \\ 0 \\ 0 \end{bmatrix} \in V \Rightarrow z = 0. \quad (1.1)$$

Condition (1.1) means that V is the graph of some linear operator $G: \begin{bmatrix} \mathcal{X} \\ \mathcal{W} \end{bmatrix} \rightarrow \mathcal{X}$ with domain $\text{dom}(G) \subset \begin{bmatrix} \mathcal{X} \\ \mathcal{W} \end{bmatrix}$. Since V is assumed to be closed, the operator G is closed. The reason for taking $\mathcal{W}$ to be a Kreĭn space instead of a Hilbert space will be explained later when we come to the notion of a *passive* s/s system.

Let $I \subset \mathbb{R}$ be a time interval with positive length. By a *classical trajectory* of Σ on I we mean a pair of functions $\begin{bmatrix} x \\ w \end{bmatrix} \in \begin{bmatrix} C^1(I; \mathcal{X}) \\ C(I; \mathcal{W}) \end{bmatrix}$ satisfying

$$\Sigma : \begin{bmatrix} \dot{x}(t) \\ x(t) \\ w(t) \end{bmatrix} \in V, \quad t \in I, \quad (1.2)$$

or equivalently,

$$\Sigma : \begin{bmatrix} x(t) \\ w(t) \end{bmatrix} \in \text{dom}(G) \text{ and } \dot{x}(t) = G \begin{bmatrix} x(t) \\ w(t) \end{bmatrix}, \quad t \in I. \quad (1.3)$$

By a *generalized trajectory* of Σ on I we mean a pair of functions $\begin{bmatrix} x \\ w \end{bmatrix} \in \begin{bmatrix} C(I; \mathcal{X}) \\ L^2_{\text{loc}}(I; \mathcal{W}) \end{bmatrix}$ which is the limit in this space of a sequence $\begin{bmatrix} x_n \\ w_n \end{bmatrix}$ of classical trajectories of Σ on I .

We call Σ *passive* if its generating subspace V satisfies two additional conditions. The first condition is related to the fact that its classical trajectories should satisfy the (power) inequality

$$\frac{d}{dt} \|x(t)\|_{\mathcal{X}}^2 \leq [w(t), w(t)]_{\mathcal{W}}, \quad t \in I, \quad (1.4)$$

or equivalently,

$$-(\dot{x}(t), x(t))_{\mathcal{X}} - (\dot{x}(t), x(t))_{\mathcal{X}} + [w(t), w(t)]_{\mathcal{W}} \geq 0, \quad t \in I. \quad (1.5)$$

Above $\|x(t)\|_{\mathcal{X}}^2$ stands for the *internal energy* stored in the state x at time t and $[w(t), w(t)]_{\mathcal{W}}$ represents the power flow (energy flow per time unit) into the system at time t through the signal $w(t)$. Incidentally, this explains why we need to allow the inner product in $\mathcal{W}$ to be indefinite: If the inner product in $\mathcal{W}$ is positive, then no energy can leave the system through the signal, and if the inner product in $\mathcal{W}$ is negative, then no energy can enter the system. The inequality (1.4) says that the system has no internal energy sources.

By (1.2), a sufficient condition for (1.4) and (1.5) to hold is that

$$-(z, x)_{\mathcal{X}} - (z, x)_{\mathcal{X}} + [w, w]_{\mathcal{W}} \geq 0, \quad \begin{bmatrix} z \\ x \\ w \end{bmatrix} \in V. \quad (1.6)$$

This makes it natural to introduce the following (strictly indefinite) Kreĭn space inner product in the node space $\mathfrak{K}$:

$$\left[\begin{bmatrix} z_1 \\ x_1 \\ w_1 \end{bmatrix}, \begin{bmatrix} z_2 \\ x_2 \\ w_2 \end{bmatrix} \right]_{\mathfrak{K}} = -(z_1, x_2) - (x_1, z_2) + [w_1, w_2]_{\mathcal{W}}. \quad (1.7)$$

Then (1.6) says that V is a nonnegative subspace of $\mathfrak{K}$ with respect to the inner product (1.7), and (1.5) can be rewritten in the form

$$\left[\begin{bmatrix} \dot{x}(t) \\ x(t) \\ w(t) \end{bmatrix}, \begin{bmatrix} \dot{x}(t) \\ x(t) \\ w(t) \end{bmatrix} \right]_{\mathfrak{K}} = -\frac{d}{dt} \|x(t)\|_{\mathcal{X}}^2 + [w(t), w(t)]_{\mathcal{W}} \geq 0, \quad t \in I. \quad (1.8)$$

The first condition that we require of the system $\Sigma = (V; \mathcal{W}, \mathcal{W})$, in addition to (1.1), in order to call it passive is that V is a *nonnegative subspace of the node space* $\mathfrak{K}$ with respect to the inner product in (1.7).

The second condition that we require of Σ in order to be passive is a maximality condition. It is not enough to require V to be nonnegative in $\mathfrak{K}$, but it should be even *maximal nonnegative*, i.e., it should not be properly contained in any other nonnegative subspace of $\mathfrak{K}$. This condition is analogous to the condition that one needs to impose on an operator A in order for A to generate a C_0 contraction semigroup on a Hilbert space $\mathcal{X}$: It is not enough that A is dissipative, but it must, in fact, be *maximal dissipative*.

Thus, summarizing the preceding discussion, $\Sigma = (V; \mathcal{X}, \mathcal{W})$ is a passive s/s system if V is a maximal nonnegative subspace of $\mathfrak{K}$ with respect to the inner product in (1.7) and (1.1) holds. Note, in particular, that the maximality of V implies that V is closed.

One often encounters passive s/s systems where (1.4)–(1.6) hold as equalities instead of inequalities. Such systems are called *energy preserving*. Thus, a passive energy preserving system is characterized by the fact that V is maximal nonnegative and neutral, i.e., $V \subset V^{[\perp]}$, where $V^{[\perp]}$ is the orthogonal companion to V in $\mathfrak{K}$. If instead V is maximal nonnegative and co-neutral, i.e., $V^{[\perp]} \subset V$, then Σ is called *co-energy preserving*. Finally, if V is Lagrangian, i.e., if $V = V^{[\perp]}$, then V is called *conservative*.

By integrating (1.4) over the interval $[s, t]$ one can rewrite (1.4) in the equivalent form

$$\|x(t)\|_{\mathcal{X}}^2 - \|x(s)\|_{\mathcal{X}}^2 \leq \int_s^t [w(v), w(v)]_{\mathcal{W}} dv, \quad s, t \in I, s \leq t. \quad (1.9)$$

By continuity of the integral, (1.9) remains valid for all generalized trajectories of Σ on I . The right-hand side of (1.9) can be interpreted as the inner product of the function w restricted to the interval $[s, t]$ with itself in a certain Kreĭn space. For each interval I with positive length we define the Kreĭn space $K^2(I; \mathcal{W})$ to be the space which coincides with $L^2(I; \mathcal{W})$ as a topological vector space, equipped with the inner product

$$[w_1, w_2]_{K^2(I; \mathcal{W})} := \int_I [w_1(s), w_2(s)]_{\mathcal{W}} ds. \quad (1.10)$$

The quadratic form $[w, w]_{K^2(I; \mathcal{W})}$ measures the amount of energy that enters the s/s system $\Sigma = (V; \mathcal{X}, \mathcal{W})$ through the signal w during the time interval I when its evolution is described by a generalized trajectory $\begin{bmatrix} x \\ w \end{bmatrix}$ with $w \in L^2(I; \mathcal{W})$. We shall be especially interested in the cases where $I = \mathbb{R}^+$, $I = \mathbb{R}$, or $I = \mathbb{R}^-$, and denote

$$K^2(\mathcal{W}) := K^2(\mathbb{R}; \mathcal{W}), \quad K_{\pm}^2(\mathcal{W}) := K^2(\mathbb{R}^{\pm}; \mathcal{W}). \quad (1.11)$$

In view of (1.9), the family $\mathfrak{K}_{0,t} := \begin{bmatrix} \mathcal{X} \\ K^2([0,t]; \mathcal{W}) \end{bmatrix}$, $t \in \mathbb{R}^+$, of Kreĭn spaces which the indefinite inner products

$$\left[\begin{bmatrix} z_1 \\ x_1 \\ w_1 \end{bmatrix}, \begin{bmatrix} z_2 \\ x_2 \\ w_2 \end{bmatrix} \right]_{\mathfrak{K}_{0,t}} = -(z_1, z_2) + (x_1, x_2) + [w_1, w_2]_{K^2([0,t]; \mathcal{W})} \quad (1.12)$$

enters naturally into the theory of passive s/s systems. Indeed, inequality (1.9) says that if we denote

$$\mathcal{T}_{0,t} := \left\{ \begin{bmatrix} x(t) \\ x(0) \\ \pi_{[0,t]} w \end{bmatrix} \middle| \begin{bmatrix} x \\ w \end{bmatrix} \text{ is a generalized trajectory of } \Sigma \text{ on } [0, t] \right\},$$

then $\mathcal{T}_{0,t}$ is a nonnegative subspace of $\mathcal{K}_{0,t}$ for all $t \in \mathbb{R}^+$. As will be shown in Theorem 3.5 below, it is even maximal nonnegative, and this fact is an important ingredient in the complete characterization given in that theorem of a passive s/s system in terms of its trajectories $\begin{bmatrix} x \\ w \end{bmatrix}$ on $\mathbb{R}^+$ satisfying $w \in K_+^2(\mathcal{W})$. Moreover, the generating subspace V can be recovered from the family $\mathcal{T}_{0,t}$ in a way that is analogous to the definition of the generator of a C_0 semigroup.

Intuitively, the state space $\mathcal{X}$ of Σ plays the role of an internal memory, and at each time t the state vector $x(t)$ contains the part of the past history of the system which may have some influence on the future dynamics. All the exchange of information with the environment takes place via the signal part w of a trajectory $\begin{bmatrix} x \\ w \end{bmatrix}$. Two s/s systems Σ_1 and Σ_2 with the same signal space $\mathcal{W}$ are *externally equivalent* if they cannot be distinguished from each other by observing only the signal parts w of the trajectories $\begin{bmatrix} x \\ w \end{bmatrix}$ of the two systems on the time interval $I = \mathbb{R}^+ := [0, \infty)$ whose initial states are zero (i.e., the two systems start “from rest” at time $t = 0$). Trajectories of this type are called *externally generated* on $\mathbb{R}^+$. We call an externally generated trajectory $\begin{bmatrix} x \\ w \end{bmatrix}$ on $\mathbb{R}^+$ *stable* if $w \in L^2(\mathbb{R}^+; \mathcal{W})$, and by the (stable) *future behavior* $\mathfrak{W}_+^\Sigma$ of Σ we mean the set of all the signal parts w of the externally generated stable trajectories $\begin{bmatrix} x \\ w \end{bmatrix}$ of Σ on $\mathbb{R}^+$.

The future behavior $\mathfrak{W}_+^\Sigma$ of a passive s/s system $\Sigma = (V; \mathcal{X}, \mathcal{W})$ has two characteristic properties. One of them is fairly obvious, namely that $\mathfrak{W}_+^\Sigma$ is *right-shift invariant* in the Kreĭn space $K_+^2(\mathcal{W})$. This follows from the fact that if $\begin{bmatrix} x \\ w \end{bmatrix}$ is an externally generated trajectory of Σ on $\mathbb{R}^+$, and if we shift this trajectory to the right by the amount t and define $\begin{bmatrix} x(s) \\ w(s) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ for $0 \leq s < t$, then we obtain another externally generated trajectory on $\mathbb{R}^+$. This means that if we denote the right-shift semigroup in $K_+^2(\mathcal{W})$ by τ_+^* (i.e., $(\tau_+^{*t} w)(s) = w(s - t)$ if $s \geq t$ and $(\tau_+^{*t} w)(s) = 0$ otherwise), then $\tau_+^{*t} \mathfrak{W}_+^\Sigma \subset \mathfrak{W}_+^\Sigma$ for all $t \in \mathbb{R}^+$.

According to the above discussion, $\mathfrak{W}_+^\Sigma$ is also a nonnegative subspace of $K_+^2(\mathcal{W})$. As we shall show in Sect. 3, $\mathfrak{W}_+^\Sigma$ is even *maximal nonnegative*. This is the second characteristic property of $\mathfrak{W}_+^\Sigma$ that we mentioned above. In particular, $\mathfrak{W}_+^\Sigma$ is closed in $K_+^2(\mathcal{W})$.

The above facts lead us to the following definition. By a *passive future behavior* on the Kreĭn signal space $\mathcal{W}$ we mean a maximal nonnegative, right-shift invariant subspace of $K_+^2(\mathcal{W})$. According to the above discussion, the future behavior of a passive s/s system with signal space $\mathcal{W}$ is a *passive future behavior* on $\mathcal{W}$.

This paper ends with a study of the inverse problem, in which we construct three different *canonical shift realizations of an arbitrary passive future behavior* $\mathfrak{W}_+$. By this we mean the following. A passive s/s system $\Sigma = (V; \mathcal{X}, \mathcal{W})$ is a *realization* of $\mathfrak{W}_+$ if the future behavior $\mathfrak{W}_+^\Sigma$ of Σ is equal to $\mathfrak{W}_+$. By a *canonical realization* we mean a realization that is completely determined by the given data $\mathfrak{W}_+$. By a *shift realization* we mean a realization whose dynamics can be interpreted (in a generalized sense) as a compressed shift in some space of functions with values in the given signal space $\mathcal{W}$.

Our three canonical realizations have the following characteristic properties: (a) the first realization is observable and co-energy preserving, (b) the second is controllable and energy preserving, and (c) the third is simple and conservative (see Sect. 3 for the

precise definitions). It will be shown that every other passive s/s realization that has one of the properties (a), (b), or (c) is unitarily similar to the corresponding canonical realization that we have constructed, and for this reason we also call them *canonical models* of passive s/s systems with one of the properties (a), (b), or (c).

The canonical models mentioned above are analogous to the three canonical de Branges–Rovnyak shift models of types (a), (b), and (c) with a given scattering matrix that belongs to the Schur class over the right half-plane, i.e., it is an analytic function whose values are contractive operators from one Hilbert space $\mathcal{U}$ to another Hilbert space $\mathcal{V}$. They were originally presented in [10, 11] (in discrete time), and they can also be found in, e.g., [1] and in [16, 17]. Indeed, there is a two-sided connection between our canonical models and the de Branges–Rovnyak models analogous to the one described in [6] in the discrete time case. Since the continuous time result looks more or less the same as the discrete time results (with the unit disc replaced by the right half-plane), and since this article is already quite long, we have chosen to here give only a very short outline of this connection. A future passive behavior $\mathfrak{M}_+$ can be mapped into the frequency domain by use of the Laplace transform, and we denote the image by $\widehat{\mathfrak{M}}_+$. This is a maximal nonnegative shift-invariant subspace of the Kreĭn–Hardy space $H_+^2(\mathcal{W})$ over the right-half plane with values in $\mathcal{W}$ (the inner product in $H_+^2(\mathcal{W})$ is the one inherited from $\mathcal{W}$). Each fundamental decomposition $\mathcal{W} = -\mathcal{W}_- \boxplus \mathcal{W}_+$ gives rise to a fundamental decomposition $H_+^2(\mathcal{W}) = -H_+^2(\mathcal{W}_-) \boxplus H_+^2(\mathcal{W}_+)$, and with respect to this decomposition $\widehat{\mathfrak{M}}_+$ has a graph representation

$$\widehat{\mathfrak{M}}_+ = \left\{ \left[\begin{array}{c} \widehat{S}\widehat{w}_+ \\ \widehat{w}_+ \end{array} \right] \mid \widehat{w}_+ \in H_+^2(\mathcal{W}_+) \right\},$$

where $\widehat{S}$ is a multiplication operator whose symbol φ is a Schur function mapping $\mathcal{W}_+$ into $\mathcal{W}_-$. This symbol φ is called a scattering matrix. It is uniquely determined by $\widehat{\mathfrak{M}}_+$ and the decomposition $\mathcal{W} = -\mathcal{W}_- \boxplus \mathcal{W}_+$, but of course, different fundamental decompositions of $\mathcal{W}$ lead to different scattering matrices. From our canonical s/s models of a passive s/s system with a given future behavior $\mathfrak{M}_+$ we can derive the corresponding de Branges–Rovnyak canonical models with the given scattering matrix φ by first passing to the i/s/o representation of the given s/s system corresponding to the fundamental decomposition $\mathcal{W} = -\mathcal{W}_- \boxplus \mathcal{W}_+$, and then applying certain unitary maps. It is also in principle possible to proceed in the opposite direction, i.e., to start with one of the de Branges–Rovnyak models and to define a state signal system in terms of an i/s/o scattering representation, but this does not lead to a canonical s/s model, since the result depends on some arbitrarily chosen fundamental decomposition of the signal space. See [6] for details. By replacing the fundamental decompositions of $\mathcal{W}$ used above by Lagrangian decompositions or orthogonal decompositions (these are defined in Sect. 2 below) one can also derive canonical models whose transfer functions belong to certain subclasses of Nevanlinna functions or Potapov functions. We shall return to this elsewhere.

We end this introduction with a short overview of the remaining sections. In Sect. 2 we review the notion of a Kreĭn space and present some Kreĭn space results that will be needed later. Some background on passive s/s systems is presented in Sect. 3. In addition to the future behaviors $\mathfrak{M}_+^\Sigma$ of a passive s/s system Σ we also introduce the

past behavior $\mathfrak{W}_-^\Sigma$ and the two-sided behavior $\mathfrak{W}^\Sigma$ of Σ . They are in principle constructed in the same way as $\Sigma_+^\mathfrak{W}$, but with $I = \mathbb{R}^+$ replaced by $I = \mathbb{R}^- = (-\infty, 0]$ or $I = \mathbb{R}$, respectively. We also introduce the general notions of a *passive past behaviors* and a *passive two-sided behaviors* on a given signal space $\mathcal{W}$ (without reference to any s/s system). It will always be true that if $\mathfrak{W}$ is a passive two-sided behavior on $\mathcal{W}$, then

$$\mathfrak{W}_- := \{w \in K_-^2(\mathcal{W}) \mid w \text{ is the restriction to } \mathbb{R}^- \text{ of a function in } \mathfrak{W}\}$$

is a passive past behavior on $\mathcal{W}$, and

$$\mathfrak{W}_+ := \{w \in \mathfrak{W} \mid w(t) = 0 \text{ for } t < 0\}.$$

is a passive future behavior on $\mathcal{W}$. As shown in Proposition 3.15 below, either one of $\mathfrak{W}_-$ and $\mathfrak{W}_+$ determines $\mathfrak{W}$ uniquely. It is also true that the past and two-sided behaviors of a passive s/s system Σ are *passive* past and two-sided behaviors.

Many of the results in Sect. 3 are either taken from [13, 15] or are straightforward extensions of results in [5]. However, it also contains some significant new results, the most important of which is Theorem 3.5 which gives necessary and sufficient conditions on a subspace $\mathcal{T}_+$ of $\left[\begin{smallmatrix} BUC(\mathbb{R}^+; \mathcal{X}) \\ K_+^2(\mathcal{W}) \end{smallmatrix} \right]$ to be the family all stable future generalized trajectories of some passive s/s system. Adjoint systems and behaviors, as well as anti-passive time reflected s/s systems are studied in Sect. 4.

In Sect. 5.2 we present two Hilbert spaces $\mathcal{H}(\mathfrak{W}_+)$ and $\mathcal{H}(\mathfrak{W}_-^{[\perp]})$ that play fundamental roles in the rest of this article. Here $\mathcal{H}(\mathfrak{W}_+)$ is the subspace of the quotient $K_+^2(\mathcal{W})/\mathfrak{W}_+$ consisting of all those equivalence classes whose $\mathcal{H}(\mathfrak{W}_+)$ -norm, defined in (5.8) below, is finite. The Hilbert space $\mathcal{H}(\mathfrak{W}_-^{[\perp]})$ is constructed in a similar way, with $\mathfrak{W}_+$ replaced by the orthogonal companion to a passive past behavior $\mathfrak{W}_-$, interpreted as a maximal nonnegative subspace of $-K_-^2(\mathcal{W})$. Both of these spaces are special cases of the spaces $\mathcal{H}(\mathcal{Z})$ introduced and studied in [4], where $\mathcal{Z}$ is a maximal nonnegative subspace of a Kreĭn space $\mathcal{X}$. A short review of the spaces $\mathcal{H}(\mathcal{Z})$ is included in Sect. 2.

In Sect. 5.3 we introduce the past/future map $\Gamma_{\mathfrak{W}}$ of a passive two-sided behavior $\mathfrak{W}$. This map plays a decisive role in our study of the construction of our three canonical realizations. It is a contraction from $\mathcal{H}(\mathfrak{W}_-^{[\perp]})$ to $\mathcal{H}(\mathfrak{W}_+)$, and it is uniquely determined by the property that if $w \in \mathfrak{W}$ and if w_- and w_+ are the restrictions of w to $\mathbb{R}^-$ and $\mathbb{R}^+$, respectively, then the image under $\Gamma_{\mathfrak{W}}$ of the equivalence class in $K_-^2(\mathcal{W})/\mathfrak{W}_-^{[\perp]}$ containing w_- is the equivalence class in $K_+^2(\mathcal{W})/\mathfrak{W}_+$ containing w_+ . This map is used to construct a third Hilbert space $\mathcal{D}(\mathfrak{W})$, which is continuously contained in $\left[\begin{smallmatrix} \mathcal{H}(\mathfrak{W}_-^{[\perp]}) \\ \mathcal{H}(\mathfrak{W}_+) \end{smallmatrix} \right]$. The space $\mathcal{D}(\mathfrak{W})$ has the property that the operators

$T_- := \left[\begin{smallmatrix} \Gamma_{\mathfrak{W}} \\ 1_{\mathcal{H}(\mathfrak{W}_-^{[\perp]})} \end{smallmatrix} \right]$ and $T_+ := \left[\begin{smallmatrix} 1_{\mathcal{H}(\mathfrak{W}_+)} \\ \Gamma_{\mathfrak{W}}^* \end{smallmatrix} \right]$ map $\mathcal{H}(\mathfrak{W}_-^{[\perp]})$ and $\mathcal{H}(\mathfrak{W}_+)$ isometrically onto subspaces $\mathcal{L}_-$ and $\mathcal{L}_+$ of $\mathcal{D}(\mathfrak{W})$, respectively, where $\mathcal{L}_- + \mathcal{L}_+$ is dense in $\mathcal{D}(\mathfrak{W})$ and the angle operator $P_{\mathcal{L}_+|\mathcal{L}_-}$ (the orthogonal projection onto $\mathcal{L}_+$ restricted to $\mathcal{L}_-$) is given by $P_{\mathcal{L}_+|\mathcal{L}_-} = T_+ \Gamma_{\mathfrak{W}} T_-^{-1}$. Thus, the angle operator $P_{\mathcal{L}_+|\mathcal{L}_-}$ between the subspaces $\mathcal{L}_-$ and $\mathcal{L}_+$ in $\mathcal{D}(\mathfrak{W})$ is a unitary image of $\Gamma_{\mathfrak{W}}$.

In Sect. 6 we develop the passive s/s systems theory further and introduce the input map $\mathfrak{B}_\Sigma$ and the output map $\mathfrak{C}_\Sigma$ of a passive s/s system Σ . Here $\mathfrak{B}_\Sigma$ is a contraction from $\mathcal{H}(\mathfrak{W}_-^{[\perp]})$ to $\mathcal{X}$, and it is the unique extension to $\mathcal{H}(\mathfrak{W}_-^{[\perp]})$ of the map from the equivalence class in $K_-^2(\mathcal{W})/\mathfrak{W}_-^{[\perp]}$ containing the signal part w of an externally generated stable trajectory $\begin{bmatrix} x \\ w \end{bmatrix}$ on $\mathbb{R}^-$ to $x(0)$. The operator $\mathfrak{C}_\Sigma$ is a contraction from $\mathcal{X}$ to $\mathcal{H}(\mathfrak{W}_+)$, and it is equal to the map from the initial state $x(0)$ of a stable trajectory $\begin{bmatrix} x \\ w \end{bmatrix}$ on $\mathbb{Z}^+$ to its signal part w factored over the future behavior $\mathfrak{W}_+$. As shown in Sect. 7, $\Gamma_{\mathfrak{W}} = \mathfrak{C}_\Sigma \mathfrak{B}_\Sigma$ whenever Σ is a passive s/s system with two-sided behavior $\mathfrak{W}$.

Finally, in Sects. 8, 9 and 10 we present our canonical observable co-energy preserving s/s shift model, controllable energy preserving s/s shift model, and simple conservative s/s shift model, respectively, whose future, past, and two-sided behaviors coincide with the given triple of passive behaviors $\mathfrak{W}_+$, $\mathfrak{W}_-$, and $\mathfrak{W}$, respectively. These models are canonical in the sense that they are uniquely determined by the given data $\mathfrak{W}_+$, $\mathfrak{W}_-$, or $\mathfrak{W}$ (any one of these three behaviors determines the other two uniquely). The state spaces in the co-energy preserving canonical model, the energy preserving canonical model, and the simple conservative canonical model are $\mathcal{H}(\mathfrak{W}_+)$, $\mathcal{H}(\mathfrak{W}_-^{[\perp]})$, and $\mathcal{D}(\mathfrak{W})$, respectively. In all cases the dynamics of the models are described by means of a generalized compression of a shift acting in the state space.

This article may be regarded as an blend of [4–6] on one hand and of [13, 15] on the other hand. In the first three of these canonical models of passive s/s systems were obtained in a discrete time setting, and in the last two a s/s theory is developed for the continuous time setting, including the passive case. Some preliminary steps towards the development of a s/s theory in continuous time were taken already in [9] by Ball and Staffans. See, in particular, [9] for a discussion of the connection with the theory of passive and conservative behaviors presented in the papers [20, 21, 23, 24] and the monograph [18]. As explained in [3], part of the motivation comes from classical passive time-invariant circuit theory found in, e.g., [7, 22].

An (inner) direct sum decomposition of a Hilbert or Kreĭn space $\mathcal{W}$ into two closed subspaces $\mathcal{Y}$ and $\mathcal{U}$ will be denoted by $\mathcal{W} = \mathcal{Y} \dot{+} \mathcal{U}$, and the corresponding complementary projections onto $\mathcal{Y}$ and $\mathcal{U}$ will be denoted by $P_{\mathcal{Y}}^{\mathcal{U}}$ and $P_{\mathcal{U}}^{\mathcal{Y}}$. If, in addition, $\mathcal{Y}$ and $\mathcal{U}$ are orthogonal to each other, then we write $\mathcal{W} = \mathcal{Y} \oplus \mathcal{U}$ in the case of a Hilbert space and $\mathcal{W} = \mathcal{Y} \boxplus \mathcal{U}$ in the case of a Kreĭn space. In the orthogonal case the subspaces $\mathcal{Y}$ and $\mathcal{U}$ become Hilbert or Kreĭn spaces when we let them inherit the inner product from $\mathcal{W}$, and we denote the (orthogonal) projections of $\mathcal{W}$ onto $\mathcal{Y}$ and $\mathcal{U}$ by $P_{\mathcal{Y}}$ and $P_{\mathcal{U}}$, respectively.

We denote the (external) direct sum of two Hilbert or Kreĭn spaces $\mathcal{Y}$ and $\mathcal{U}$ by $\begin{bmatrix} \mathcal{Y} \\ \mathcal{U} \end{bmatrix}$. By this we mean the Cartesian product of $\mathcal{Y}$ and $\mathcal{U}$ equipped with the standard algebraic operations and standard product topology. We sometimes equip $\begin{bmatrix} \mathcal{Y} \\ \mathcal{U} \end{bmatrix}$ with the induced Kreĭn space inner product (in the Kreĭn space notation)

$$\left[\begin{bmatrix} y_1 \\ u_1 \end{bmatrix}, \begin{bmatrix} y_2 \\ u_2 \end{bmatrix} \right]_{\mathcal{Y} \boxplus \mathcal{U}} = [y_1, y_2]_{\mathcal{Y}} + [u_1, u_2]_{\mathcal{U}}. \quad (1.13)$$

After identifying $\begin{bmatrix} \mathcal{Y} \\ 0 \end{bmatrix}$ with $\mathcal{Y}$ and $\begin{bmatrix} 0 \\ \mathcal{U} \end{bmatrix}$ with $\mathcal{U}$ we can in this case identify $\begin{bmatrix} \mathcal{Y} \\ \mathcal{U} \end{bmatrix}$ with $\mathcal{Y} \boxplus \mathcal{U}$. However, we shall often instead use a different Kreĭn space inner product in $\begin{bmatrix} \mathcal{Y} \\ \mathcal{U} \end{bmatrix}$ of the type

$$\left[\begin{bmatrix} y_1 \\ u_1 \end{bmatrix}, \begin{bmatrix} y_2 \\ u_2 \end{bmatrix} \right]_{\begin{bmatrix} \mathcal{Y} \\ \mathcal{U} \end{bmatrix}} = \left(\begin{bmatrix} y \\ u \end{bmatrix}, J \begin{bmatrix} y \\ u \end{bmatrix} \right)_{\mathcal{Y} \boxplus \mathcal{U}},$$

where J is a given signature operator in $\mathcal{Y} \boxplus \mathcal{U}$. With respect to this inner product $\mathcal{Y}$ and $\mathcal{U}$ may or may not be orthogonal. Analogous notations are used for direct sums with three or more components.

2 Kreĭn Spaces

2.1 Some Kreĭn Space Results

Throughout this work both the signal space $\mathcal{W}$ and the node space $\mathfrak{K}$ will be Kreĭn spaces. We therefore begin with a review of some Kreĭn space notions and results that will be needed here.

A Kreĭn space $\mathcal{W}$ is a vector space with an inner product $[\cdot, \cdot]_{\mathcal{W}}$ that satisfies all the standard properties required by an inner product, except for the condition $[w, w]_{\mathcal{W}} > 0$ for nonzero w , with the additional property that $\mathcal{W}$ can be decomposed into a direct sum $\mathcal{W} = -\mathcal{Y} \dot{+} \mathcal{U}$ in such a way that the following conditions are satisfied:

- (i) $\mathcal{U}$ and $-\mathcal{Y}$ are orthogonal to each other with respect to the inner product $[\cdot, \cdot]_{\mathcal{W}}$, i.e., $[y, u]_{\mathcal{W}} = 0$ for all $u \in \mathcal{U}$ and all $y \in -\mathcal{Y}$.
- (ii) $\mathcal{U}$ is a Hilbert space with the inner product $(u, u')_{\mathcal{U}} := [u, u']_{\mathcal{W}}$, $u, u' \in \mathcal{U}$, inherited from $\mathcal{W}$.
- (iii) $-\mathcal{Y}$ is an anti-Hilbert space with the inner product $[y, y']_{-\mathcal{Y}} := [y, y']_{\mathcal{W}}$, $y, y' \in -\mathcal{Y}$, inherited from $\mathcal{W}$.

Here and later we shall use the notation $-\mathcal{Y}$ for the *anti-space* of a vector space $\mathcal{Y}$ equipped with a (possibly indefinite) inner product. This is the same topological vector space as $\mathcal{Y}$, but the inner product $[\cdot, \cdot]_{\mathcal{Y}}$ in $\mathcal{Y}$ has been replaced by the inner product $[y, y']_{-\mathcal{Y}} := -[y, y']_{\mathcal{Y}}$, $y, y' \in -\mathcal{Y}$. The condition that $-\mathcal{Y}$ is an *anti-Hilbert space* with the inner product inherited from $\mathcal{W}$ is equivalent to saying that $\mathcal{Y}$ is a Hilbert space with the inner product $(y, y')_{\mathcal{Y}} := -[y, y']_{\mathcal{W}}$, $y, y' \in -\mathcal{Y}$, inherited from $-\mathcal{W}$. Since $\mathcal{Y}$ and $\mathcal{U}$ are orthogonal to each other we shall denote the direct sum by $\mathcal{W} = -\mathcal{Y} \boxplus \mathcal{U}$.

Any decomposition $\mathcal{W} = -\mathcal{Y} \boxplus \mathcal{U}$ with the properties listed above is called a *fundamental decomposition* of $\mathcal{W}$. If the space $\mathcal{W}$ itself is neither a Hilbert space nor an anti-Hilbert space, then it has infinite many fundamental decompositions. If $\mathcal{W} = -\mathcal{Y} \boxplus \mathcal{U}$ is a fundamental decomposition of $\mathcal{W}$, then

$$[w, w]_{\mathcal{W}} = -\|y\|_{\mathcal{Y}}^2 + \|u\|_{\mathcal{U}}^2, \quad w = u + y, \quad u \in \mathcal{U}, \quad y \in \mathcal{Y}. \quad (2.1)$$

The dimensions of the positive space $\mathcal{U}$ and the negative space $-\mathcal{Y}$ do not depend on the particular fundamental decomposition. These dimensions are called the positive and negative indices of $\mathcal{W}$, respectively, and they are denoted by $\text{ind}_+ \mathcal{W}$ and $\text{ind}_- \mathcal{W}$.

An arbitrary choice of fundamental decomposition $\mathcal{W} = -\mathcal{Y} \boxplus \mathcal{U}$ determines a Hilbert space norm on $\mathcal{W}$ by

$$\|w\|_{\mathcal{Y} \boxplus \mathcal{U}}^2 = \|y\|_{\mathcal{Y}}^2 + \|u\|_{\mathcal{U}}^2, \quad w = u + y, \quad u \in \mathcal{U}, \quad y \in \mathcal{Y}. \quad (2.2)$$

While the norm $\|\cdot\|_{\mathcal{Y} \boxplus \mathcal{U}}$ itself depends on the choice of fundamental decomposition $\mathcal{W} = -\mathcal{Y} \boxplus \mathcal{U}$, all these norms are equivalent and the resulting strong and weak topologies are each independent of the choice of the fundamental decomposition. Thus, we can define topological notions, such as convergence, or closedness, with respect to any one of these norms. Any norm on $\mathcal{W}$ arising in this way from some choice of fundamental decomposition $\mathcal{W} = -\mathcal{Y} \boxplus \mathcal{U}$ for $\mathcal{W}$ we shall call an *admissible norm* on $\mathcal{W}$, and we shall refer to the corresponding positive inner product on $\mathcal{Y} \oplus \mathcal{U}$ as an *admissible Hilbert space inner product* on $\mathcal{W}$.

A subspace $\mathcal{Z}$ of $\mathcal{W}$ is *nonnegative* if every vector $w \in \mathcal{Z}$ is nonnegative ($[w, w]_{\mathcal{W}} \geq 0$), it is *neutral* if every vector $w \in \mathcal{Z}$ is neutral ($[w, w]_{\mathcal{W}} = 0$), and *nonpositive* if every vector $w \in \mathcal{Z}$ is nonpositive ($[w, w]_{\mathcal{W}} \leq 0$). A nonnegative subspace which is not strictly contained in any other nonnegative subspace is called *maximal nonnegative*, and the notion of a *maximal nonpositive subspace* is defined in an analogous way. Every nonnegative subspace is contained in some maximal nonnegative subspace, and every nonpositive subspace is contained in some maximal nonpositive subspace. Maximal nonnegative or nonpositive subspaces are always closed.

The *orthogonal companion* $\mathcal{Z}^{[\perp]}$ of an arbitrary subset $\mathcal{Z} \subset \mathcal{W}$ with respect to the Kreĭn space inner product $[\cdot, \cdot]_{\mathcal{W}}$ consists of all vectors in $\mathcal{W}$ that are orthogonal to all vectors in $\mathcal{Z}$, i.e.,

$$\mathcal{Z}^{[\perp]} = \{w' \in \mathcal{W} \mid [w', w]_{\mathcal{W}} = 0 \text{ for all } w \in \mathcal{Z}\}.$$

This is always a closed subspace of $\mathcal{W}$, and $\mathcal{Z} = (\mathcal{Z}^{[\perp]})^{[\perp]}$ if and only if $\mathcal{Z}$ is a closed subspace. If $\mathcal{W}$ is a Hilbert space, then we write $\mathcal{Z}^\perp$ instead of $\mathcal{Z}^{[\perp]}$. A subspace $\mathcal{Z}$ is neutral if and only if $\mathcal{Z} \subset \mathcal{Z}^{[\perp]}$. If instead $\mathcal{Z}^{[\perp]} \subset \mathcal{Z}$ (i.e., $\mathcal{Z}^{[\perp]}$ is neutral), then we call $\mathcal{Z}$ *co-neutral*. A subspace $\mathcal{Z} \subset \mathcal{W}$ is called *Lagrangian* if $\mathcal{Z} = \mathcal{Z}^{[\perp]}$.

A direct sum decomposition $\mathcal{W} = \mathcal{F} \dot{+} \mathcal{E}$ of $\mathcal{W}$ where both $\mathcal{F}$ and $\mathcal{E}$ are neutral is called a *Lagrangian decomposition* of $\mathcal{W}$. The subspaces $\mathcal{F}$ and $\mathcal{E}$ are automatically Lagrangian in this case. Such a decomposition exists if and only if $\text{ind}_+ \mathcal{W} = \text{ind}_- \mathcal{W}$ (this index may be finite or infinite).

If we fix a fundamental decomposition $\mathcal{W} = -\mathcal{Y} \boxplus \mathcal{U}$, then we may view elements of $\mathcal{W}$ as consisting of column vectors

$$w = \begin{bmatrix} y \\ u \end{bmatrix} \in \begin{bmatrix} -\mathcal{Y} \\ \mathcal{U} \end{bmatrix},$$

where we view $\mathcal{Y}$ and $\mathcal{U}$ as Hilbert spaces, and the Kreĭn space inner product on $\mathcal{W}$ is given by

$$\begin{aligned} \begin{bmatrix} y \\ u \end{bmatrix}, \begin{bmatrix} y' \\ u' \end{bmatrix} \Big|_{\mathcal{W}} &= \left(\begin{bmatrix} y \\ u \end{bmatrix}, \begin{bmatrix} -1_{\mathcal{Y}} & 0 \\ 0 & 1_{\mathcal{U}} \end{bmatrix} \begin{bmatrix} y' \\ u' \end{bmatrix} \right)_{\mathcal{Y} \oplus \mathcal{U}} \\ &= -(y, y')_{\mathcal{Y}} + (u, u')_{\mathcal{U}}. \end{aligned} \quad (2.3)$$

In this representation, nonnegative, neutral, nonpositive, and Lagrangian subspaces are characterized as follows.

Proposition 2.1 *Let $\mathcal{W}$ be a Kreĭn space represented in the form $\mathcal{W} = \begin{bmatrix} -\mathcal{Y} \\ \mathcal{U} \end{bmatrix}$ with Kreĭn space inner product equal to the quadratic form $[\cdot, \cdot]_J$ induced by the operator $J = \begin{bmatrix} -1_{\mathcal{Y}} & 0 \\ 0 & 1_{\mathcal{U}} \end{bmatrix}$ in the Hilbert space inner product of $\begin{bmatrix} \mathcal{Y} \\ \mathcal{U} \end{bmatrix}$ as explained above, and let $\mathcal{Z}$ be a subspace of $\mathcal{W}$. Then the following claims are true:*

- (i) $\mathcal{Z}$ is nonnegative if and only if there is a linear Hilbert space contraction $A_+ : \mathcal{Z}_+ \mapsto \mathcal{Y}$ from some domain $\mathcal{Z}_+ \subset \mathcal{U}$ into $\mathcal{Y}$ such that

$$\mathcal{Z} = \begin{bmatrix} A_+ \\ 1_{\mathcal{U}} \end{bmatrix} \mathcal{Z}_+ = \left\{ \begin{bmatrix} A_+ z_+ \\ z_+ \end{bmatrix} \mid z_+ \in \mathcal{Z}_+ \right\}. \quad (2.4)$$

$\mathcal{Z}$ is maximal nonnegative if and only if, in addition, $\mathcal{Z}_+ = \mathcal{U}$.

- (ii) $\mathcal{Z}$ is nonpositive if and only if there is a linear contraction $A_- : \mathcal{Z}_- \mapsto \mathcal{U}$ from some domain $\mathcal{Z}_- \subset \mathcal{Y}$ into $\mathcal{U}$ such that

$$\mathcal{Z} = \begin{bmatrix} 1_{\mathcal{Y}} \\ A_- \end{bmatrix} \mathcal{Z}_- = \left\{ \begin{bmatrix} z_- \\ A_- z_- \end{bmatrix} \mid z_- \in \mathcal{Z}_- \right\}. \quad (2.5)$$

$\mathcal{Z}$ is maximal nonpositive if and only if, in addition, $\mathcal{Z}_- = \mathcal{Y}$.

- (iii) $\mathcal{Z}$ is neutral if and only if there is an isometry A_+ mapping a subspace $\mathcal{Z}_+$ of $\mathcal{U}$ onto a subspace $\mathcal{Z}_-$ of $\mathcal{Y}$, or equivalently, an isometry A_- mapping $\mathcal{Z}_- \subset \mathcal{Y}$ isometrically onto $\mathcal{Z}_+ \subset \mathcal{U}$, such that

$$\mathcal{Z} = \begin{bmatrix} A_+ \\ 1_{\mathcal{U}} \end{bmatrix} \mathcal{Z}_+ = \begin{bmatrix} 1_{\mathcal{Y}} \\ A_- \end{bmatrix} \mathcal{Z}_-. \quad (2.6)$$

$\mathcal{Z}$ is Lagrangian if and only if, in addition, $\mathcal{Z}_+ = \mathcal{U}$ and $\mathcal{Z}_- = \mathcal{Y}$.

- (iv) $\mathcal{Z}$ is maximal nonnegative if and only if $\mathcal{Z}$ is closed and $\mathcal{Z}^{[\perp]}$ is maximal nonpositive. More precisely, if $\mathcal{Z}$ has the representation (2.4) with $\mathcal{Z}_+ = \mathcal{U}$, then $\mathcal{Z}^{[\perp]}$ has the representation

$$\mathcal{Z}^{[\perp]} = \begin{bmatrix} 1_{\mathcal{Y}} \\ A_+^* \end{bmatrix} \mathcal{Y}, \quad (2.7)$$

where A_+^* is computed with respect to the Hilbert space inner product in $\mathcal{Y}$ (instead of the anti-Hilbert space inner product in $-\mathcal{Y}$ inherited from $\mathcal{W}$).

- (v) $\mathcal{Z}$ is maximal nonnegative if and only if $\mathcal{Z}$ is closed and nonnegative and $\mathcal{Z}^{[\perp]}$ is nonpositive. In particular, $\mathcal{Z}$ is Lagrangian if and only if $\mathcal{Z}$ is both maximal nonnegative and maximal nonpositive.

Proof See [2, Section 1.8, pp. 48–64] or the following theorems in [8]: Theorem 11.7 on p. 54, Theorems 4.2 and 4.4 on pp. 105–106, and Lemma 4.5 on p. 106. $\square$

The fundamental decompositions that we have considered above are a special case of *orthogonal decompositions* $\mathcal{W} = -\mathcal{Y} \boxplus \mathcal{U}$ of $\mathcal{W}$, where $\mathcal{Y}$ and $\mathcal{U}$ are orthogonal with respect to $[\cdot, \cdot]_{\mathcal{W}}$, and both $\mathcal{Y}$ and $\mathcal{U}$ are Kreĭn spaces with the inner products inherited from $-\mathcal{W}$ and $\mathcal{W}$, respectively. Thus, if $w = y + u$ with $y \in \mathcal{Y}$ and $u \in \mathcal{U}$, then

$$[w, w]_{\mathcal{W}} = [y, y]_{\mathcal{W}} + [u, u]_{\mathcal{W}} = -[y, y]_{\mathcal{Y}} + [u, u]_{\mathcal{U}}. \quad (2.8)$$

This orthogonal decomposition is fundamental if and only if $\mathcal{Y}$ and $\mathcal{U}$ are Hilbert spaces, i.e., if they are both nonnegative Kreĭn spaces.

The next lemma, proved in [6], will be used later to find out if certain subspaces of a Kreĭn space with a special orthogonal decomposition are maximal nonnegative, or maximal nonpositive, or Lagrangian.

Lemma 2.2 ([6, Lemma 2.2]) *Let $\mathcal{X}$ and $\mathcal{Y}$ be two Hilbert spaces and $\mathcal{K}$ a Kreĭn space, and let $\mathfrak{K}$ be the Kreĭn space $\mathfrak{K} = \begin{bmatrix} -\mathcal{Y} \\ \mathcal{X} \\ \mathcal{K} \end{bmatrix} = -\mathcal{Y} \boxplus \mathcal{X} \boxplus \mathcal{K}$.*

- (i) *A nonnegative subspace $\mathcal{Z}$ of $\mathfrak{K}$ is maximal nonnegative if and only if conditions (a) and (b) below hold:*
 - (a) *For each $x \in \mathcal{X}$ there exists some $y \in \mathcal{Y}$ and $w \in \mathcal{K}$ such that $\begin{bmatrix} y \\ x \\ w \end{bmatrix} \in \mathcal{Z}$;*
 - (b) *The set of all $w \in \mathcal{K}$ for which there exists some $y \in \mathcal{Y}$ such that $\begin{bmatrix} y \\ 0 \\ w \end{bmatrix} \in \mathcal{Z}$ is maximal nonnegative in $\mathcal{K}$.*
- (ii) *A nonpositive subspace $\mathcal{Z}$ of $\mathfrak{K}$ is maximal nonpositive if and only if conditions (c) and (d) below hold:*
 - (c) *For each $y \in \mathcal{Y}$ there exists some $x \in \mathcal{X}$ and $w \in \mathcal{K}$ such that $\begin{bmatrix} y \\ x \\ w \end{bmatrix} \in \mathcal{Z}$;*
 - (d) *The set of all $w \in \mathcal{K}$ for which there exists some $x \in \mathcal{X}$ such that $\begin{bmatrix} 0 \\ x \\ w \end{bmatrix} \in \mathcal{Z}$ is maximal nonpositive in $\mathcal{K}$.*
- (iii) *A neutral subspace $\mathcal{Z}$ of $\mathfrak{K}$ is Lagrangian if and only if conditions (a)–(d) above hold.*

Lemma 2.3 *Let $\mathfrak{K}$ be a Kreĭn space with the orthogonal decomposition $\mathfrak{K} = \mathfrak{K}_1 \boxplus \mathfrak{K}_2$, and let $\mathcal{Z}$ be a subspace of $\mathfrak{K}$. Then*

$$(P_{\mathfrak{K}_2} \mathcal{Z})^{[\perp]} = \mathcal{Z}^{[\perp]} \cap \mathfrak{K}_2 \quad \text{and} \quad (\mathcal{Z}^{[\perp]} \cap \mathfrak{K}_2)^{[\perp]} = \overline{P_{\mathfrak{K}_2} \mathcal{Z}}, \quad (2.9)$$

where the orthogonal companions on the left-hand sides are computed with respect to $\mathfrak{K}_2$. In particular, if $P_{\mathfrak{K}_2} \mathcal{Z}$ is closed, then $P_{\mathfrak{K}_2} \mathcal{Z} = (\mathcal{Z}^{[\perp]} \cap \mathfrak{K}_2)^{[\perp]}$.

Proof That $(P_{\mathfrak{K}_2}\mathcal{Z})^{[\perp]} = \mathcal{Z}^{[\perp]} \cap \mathfrak{K}_2$ follows from the following chain of equivalences:

$$\begin{aligned} z^\dagger &\in (P_{\mathfrak{K}_2}\mathcal{Z})^{[\perp]} \\ \Leftrightarrow z^\dagger &\in \mathfrak{K}_2 \quad \text{and} \quad [z^\dagger, P_{\mathfrak{K}_2}z]_{\mathfrak{K}} = 0 \quad \text{for all } z \in \mathcal{Z} \\ \Leftrightarrow z^\dagger &\in \mathfrak{K}_2 \quad \text{and} \quad [z^\dagger, z]_{\mathfrak{K}} = 0 \quad \text{for all } z \in \mathcal{Z} \\ \Leftrightarrow z^\dagger &\in \mathcal{Z}^{[\perp]} \cap \mathfrak{K}_2. \end{aligned}$$

This implies that $\overline{P_{\mathfrak{K}_2}\mathcal{Z}} = ((P_{\mathfrak{K}_2}\mathcal{Z})^{[\perp]})^{[\perp]} = (\mathcal{Z}^{[\perp]} \cap \mathfrak{K}_2)^{[\perp]}$. $\square$

Lemma 2.4 *Let $\mathfrak{K}$ be a Kreĭn space with the orthogonal decomposition $\mathfrak{K} = \mathfrak{K}_1 \boxplus \mathfrak{K}_2$, and let $\mathcal{Z}$ be a maximal nonnegative subspace of $\mathfrak{K}$. Then the following conditions are equivalent:*

- (i) $P_{\mathfrak{K}_2}\mathcal{Z}$ is a nonnegative subspace of $\mathfrak{K}_2$;
- (ii) $P_{\mathfrak{K}_2}\mathcal{Z}$ is a maximal nonnegative subspace of $\mathfrak{K}_2$;
- (iii) $\mathcal{Z} \cap \mathfrak{K}_1$ is a maximal nonnegative subspace of $\mathfrak{K}_1$;
- (iv) $P_{\mathfrak{K}_1}|_{\mathcal{Z}}$ is a contraction $\mathcal{Z} \rightarrow \mathfrak{K}_1$, i.e.,

$$[z, z]_{\mathfrak{K}} \leq [P_{\mathfrak{K}_1}z, P_{\mathfrak{K}_1}z]_{\mathfrak{K}_1} \quad \text{for all } z \in \mathcal{Z}.$$

- (v) $P_{\mathfrak{K}_1}\mathcal{Z}^{[\perp]}$ is a nonpositive subspace of $\mathfrak{K}_1$;
- (vi) $P_{\mathfrak{K}_1}\mathcal{Z}^{[\perp]}$ is a maximal nonpositive subspace of $\mathfrak{K}_1$;
- (vii) $\mathcal{Z}^{[\perp]} \cap \mathfrak{K}_2$ is a maximal nonpositive subspace of $\mathfrak{K}_2$;
- (viii) $P_{\mathfrak{K}_2}|_{\mathcal{Z}^{[\perp]}}$ is an expansion $\mathcal{Z}^{[\perp]} \rightarrow \mathfrak{K}_2$, i.e.,

$$[z^\dagger, z^\dagger]_{\mathfrak{K}} \geq [P_{\mathfrak{K}_2}z^\dagger, P_{\mathfrak{K}_2}z^\dagger]_{\mathfrak{K}_2} \quad \text{for all } z^\dagger \in \mathcal{Z}^{[\perp]}.$$

When these equivalent conditions holds we have $(\mathcal{Z} \cap \mathfrak{K}_1)^{[\perp]} = P_{\mathfrak{K}_1}\mathcal{Z}^{[\perp]}$ and $P_{\mathfrak{K}_2}\mathcal{Z} = (\mathcal{Z}^{[\perp]} \cap \mathfrak{K}_2)^{[\perp]}$, where the orthogonal companions on the right-hand sides are computed in $\mathfrak{K}_1$ and $\mathfrak{K}_2$, respectively.

Proof We first show that (i), (ii), (iv), and (vii) are equivalent to each other, and that (analogously) (v), (vi), (viii), and (iii) are equivalent to each other, and then complete the proof of the equivalence of the conditions (i)–(viii) by showing that (iii) $\Rightarrow$ (i) and (vii) $\Rightarrow$ (v). The final claim follows from Lemma 2.3.

(i) $\Leftrightarrow$ (ii): Trivially (ii) $\Rightarrow$ (i). If $P_{\mathfrak{K}_2}\mathcal{Z}$ is not maximal nonnegative in $\mathfrak{K}_2$, then $P_{\mathfrak{K}_2}\mathcal{Z}$ is properly contained in some nonnegative subspace $\mathcal{Z}_2$ of $\mathfrak{K}_2$. This implies that $\mathcal{Z}$ is properly contained in the nonnegative subspace $\mathcal{Z} \vee \begin{bmatrix} \{0\} \\ \mathcal{Z}_2 \end{bmatrix}$ of $\mathfrak{K}$, and hence $\mathcal{Z}$ cannot be maximal. Thus (i) $\Rightarrow$ (ii).

(i) $\Leftrightarrow$ (iv): Since $\mathfrak{K} = \mathfrak{K}_1 \boxplus \mathfrak{K}_2$ we have

$$[z, z]_{\mathfrak{K}} = [P_{\mathfrak{K}_1}z, P_{\mathfrak{K}_1}z]_{\mathfrak{K}_1} + [P_{\mathfrak{K}_2}z, P_{\mathfrak{K}_2}z]_{\mathfrak{K}_2} \quad \text{for all } z \in \mathcal{Z}.$$

Thus, (i) $\Leftrightarrow$ (iv).

(ii) $\Leftrightarrow$ (vii): This follows from Proposition 2.1(iv) and Lemma 2.3.

(v) $\Leftrightarrow$ (vi): This follows from the equivalence (i) $\Leftrightarrow$ (ii) if we replace $\mathfrak{K}$ by its anti-space $-\mathfrak{K}$, interchange $\mathfrak{K}_1$ and $\mathfrak{K}_2$, and also interchange $\mathcal{Z}$ and $\mathcal{Z}^{[\perp]}$.

(v) $\Leftrightarrow$ (viii): This follows from the equivalence (i) $\Leftrightarrow$ (iv) if we replace $\mathfrak{K}$ by its anti-space $-\mathfrak{K}$, interchange $\mathfrak{K}_1$ and $\mathfrak{K}_2$, and also interchange $\mathcal{Z}$ and $\mathcal{Z}^{[\perp]}$.

(vi) $\Leftrightarrow$ (iii): This follows from Proposition 2.1(iv) and Lemma 2.3.

(iii) $\Rightarrow$ (i): Suppose that (i) does not hold. Then there exists a vector $z_0 \in \mathcal{Z}$ such that $[P_{\mathfrak{K}_2} z_0, P_{\mathfrak{K}_2} z_0]_{\mathfrak{K}_2} < 0$. In particular, since $\mathcal{Z}$ is nonnegative, this implies that $P_{\mathfrak{K}_2} z_0 \notin \mathcal{Z}$, and consequently, $P_{\mathfrak{K}_1} z_0 = z_0 - P_{\mathfrak{K}_2} z_0 \notin \mathcal{Z}$. Thus, $\mathcal{Z} \cap \mathfrak{K}_1$ is a proper subset of $P_{\mathfrak{K}_1} z_0 \vee (\mathcal{Z} \cap \mathfrak{K}_1)$. We claim that this subspace is nonnegative. This is true because for all $z \in \mathcal{Z}$ and all $\lambda \in \mathbb{C}$, we have $\lambda z_0 + z \in \mathcal{Z}$, and hence

$$[\lambda P_{\mathfrak{K}_1} z_0 + z, \lambda P_{\mathfrak{K}_1} z_0 + z]_{\mathfrak{K}} = [\lambda z_0 + z, \lambda z_0 + z]_{\mathfrak{K}} - |\lambda|^2 [P_{\mathfrak{K}_2} z_0, P_{\mathfrak{K}_2} z_0]_{\mathfrak{K}_2} \geq 0.$$

Thus, if (i) is false, then so is (iii).

(vii) $\Rightarrow$ (v): This follows from the implication (iii) $\Rightarrow$ (i) if we replace $\mathfrak{K}$ by its anti-space $-\mathfrak{K}$, interchange $\mathfrak{K}_1$ and $\mathfrak{K}_2$, and also interchange $\mathcal{Z}$ and $\mathcal{Z}^{[\perp]}$. $\square$

2.2 The Hilbert Space $\mathcal{H}(\mathcal{Z})$

In [4] a Hilbert space $\mathcal{H}(\mathcal{Z})$ was constructed, starting from an arbitrary maximal nonnegative subspace $\mathcal{Z}$ of a Kreĭn space. Below we give a short review of this construction. It will be used later in the construction of the state spaces of our canonical s/s signal realizations.

Let $\mathcal{Z}$ be a maximal nonnegative subspace of the Kreĭn space $\mathcal{K}$, and let $\mathcal{K}/\mathcal{Z}$ be the quotient of $\mathcal{K}$ modulo $\mathcal{Z}$. We define $\mathcal{H}(\mathcal{Z})$ by

$$\mathcal{H}(\mathcal{Z}) = \{h \in \mathcal{K}/\mathcal{Z} \mid \sup\{-[x, x]_{\mathcal{K}} \mid x \in h\} < \infty\}. \quad (2.10)$$

It turns out that $\sup\{-[x, x]_{\mathcal{K}} \mid x \in h\} \geq 0$ for all $h \in \mathcal{H}(\mathcal{Z})$, that $\mathcal{H}(\mathcal{Z})$ is a subspace of $\mathcal{K}/\mathcal{Z}$, that $\mathcal{H}(\mathcal{Z})$ is a Hilbert space with the norm

$$\|h\|_{\mathcal{H}(\mathcal{Z})} = (\sup\{-[x, x]_{\mathcal{K}} \mid x \in h\})^{1/2}, \quad h \in \mathcal{H}(\mathcal{Z}), \quad (2.11)$$

and that $\mathcal{H}(\mathcal{Z})$ is continuously contained in $\mathcal{K}/\mathcal{Z}$ (where we use the standard quotient topology in $\mathcal{K}/\mathcal{Z}$, induced by some arbitrarily chosen admissible Hilbert space norm in $\mathcal{K}$). We denote the equivalence class $h \in \mathcal{K}/\mathcal{Z}$ that contains a particular vector $x \in \mathcal{K}$ by $h = x + \mathcal{Z}$. Thus, with this notation, (2.10) and (2.11) can be rewritten in the form

$$\mathcal{H}(\mathcal{Z}) = \{x + \mathcal{Z} \in \mathcal{K}/\mathcal{Z} \mid \|x + \mathcal{Z}\|_{\mathcal{H}(\mathcal{Z})}^2 < \infty\}, \quad (2.12)$$

$$\|x + \mathcal{Z}\|_{\mathcal{H}(\mathcal{Z})}^2 = \sup\{-[x + z, x + z]_{\mathcal{K}} \mid z \in \mathcal{Z}\}, \quad x \in \mathcal{K}. \quad (2.13)$$

A very important (and easily proved fact) is that if we define

$$\mathcal{H}^0(\mathcal{Z}) := \{z^\dagger + \mathcal{Z} \mid z^\dagger \in \mathcal{Z}^{[\perp]}\}, \quad (2.14)$$

then $\mathcal{H}^0(\mathcal{Z})$ is a subspace of $\mathcal{H}(\mathcal{Z})$. However, even more is true: $\mathcal{H}^0(\mathcal{Z})$ is a *dense subspace* of $\mathcal{H}(\mathcal{Z})$, and

$$[x + \mathcal{Z}, z^\dagger + \mathcal{Z}]_{\mathcal{H}(\mathcal{Z})} = -[x, z^\dagger]_{\mathcal{K}}, \quad x + \mathcal{Z} \in \mathcal{H}(\mathcal{Z}), \quad z^\dagger \in \mathcal{Z}^{[\perp]}, \quad (2.15)$$

$$\|z^\dagger + \mathcal{Z}\|_{\mathcal{H}(\mathcal{Z})}^2 = -[z^\dagger, z^\dagger]_{\mathcal{K}}, \quad z^\dagger \in \mathcal{Z}^{[\perp]}. \quad (2.16)$$

Thus, $\mathcal{H}(\mathcal{Z})$ may be interpreted as a completion of $\mathcal{H}^0(\mathcal{Z})$. See [4] for more details.

3 Passive and Conservative State/Signal Systems

3.1 Basic Properties of Trajectories of Passive S/S Systems

We already gave a short introduction to passive s/s (state/signal) systems, and now describe this notion in more detail.

In the following definition and throughout the remainder of this paper, the interval I is assumed to be *closed* and *nontrivial*, i.e., it should have a nonempty interior. Thus, it is either a finite interval $I = [t_0, t_1]$, or a semi-finite interval $I = (-\infty, t_1]$ or $I = [t_0, \infty)$, or the full real line $I = \mathbb{R} = (-\infty, \infty)$.

Definition 3.1 Let $\mathcal{X}$ be a Hilbert space and $\mathcal{W}$ a Kreĭn space.

- (i) By a *passive s/s node* in continuous time we mean a triple $\Sigma = (V; \mathcal{X}, \mathcal{W})$ where V is a maximal nonnegative subspace of the Kreĭn *node space* $\mathfrak{K} := \begin{bmatrix} \mathcal{X} \\ \mathcal{W} \end{bmatrix}$ equipped with the inner product (1.7) satisfying (1.1).
- (ii) A *classical trajectory* generated by a subspace V of $\mathfrak{K}$ on an interval I is a pair of functions $\begin{bmatrix} x \\ w \end{bmatrix} \in \begin{bmatrix} C^1(I; \mathcal{X}) \\ C(I; \mathcal{W}) \end{bmatrix}$ satisfying (1.2).
- (iii) A (*generalized*) *trajectory* generated by a subspace V of $\mathfrak{K}$ on an interval I is a pair of functions $\begin{bmatrix} x \\ w \end{bmatrix} \in \begin{bmatrix} C(I; \mathcal{X}) \\ L_{\text{loc}}^2(I; \mathcal{W}) \end{bmatrix}$ which can be approximated by a sequence of classical trajectories $\begin{bmatrix} x_n \\ w_n \end{bmatrix}$ in such a way that $x_n \rightarrow x$ in $\mathcal{X}$ locally uniformly on I , and $w_n \rightarrow w$ in $L_{\text{loc}}^2(I; \mathcal{W})$.
- (iv) The passive s/s node Σ together with its families of classical and generalized trajectories is called a *passive s/s system*, and it is denoted by the same symbols as the node.
- (v) By a *past*, *two-sided*, or *future* trajectory of Σ we mean a trajectory of Σ on $\mathbb{R}^-$, $\mathbb{R}$, or $\mathbb{R}^+$, respectively.¹
- (vi) A (generalized) trajectory $\begin{bmatrix} x \\ w \end{bmatrix}$ of a passive s/s system $\Sigma = (V; \mathcal{X}, \mathcal{W})$ on an interval I is *externally generated* if the following condition holds: If I has a finite left end-point t_0 , then we require that $x(t_0) = 0$, and if the left end-point of I is $-\infty$, then we require that $\lim_{t \rightarrow -\infty} x(t) = 0$ and that $w \in L^2((-\infty, T]; \mathcal{W})$ for every finite $T \in I$.
- (vii) A (generalized) trajectory $\begin{bmatrix} x \\ w \end{bmatrix}$ of a passive s/s system $\Sigma = (V; \mathcal{X}, \mathcal{W})$ is *stable* if x is bounded on I and $w \in L^2(I; \mathcal{W})$.

¹ Instead of “two-sided trajectory” the name “full trajectory” as used in the discrete time settings of [5] and [6].

As the following lemma shows, the boundedness condition on x in Definition 3.1 (vii) is often redundant.

Lemma 3.2 *Let $\Sigma = (V; \mathcal{X}, \mathcal{W})$ be a passive s/s system, let I be an interval, and let $\begin{bmatrix} x \\ w \end{bmatrix}$ be a (generalized) trajectory of Σ on the closed interval I . Assume further that at least one of the conditions (i) or (ii) below holds:*

- (i) *The interval I is bounded to the left.*
- (ii) *$\begin{bmatrix} x \\ w \end{bmatrix}$ is externally generated.*

Then $\begin{bmatrix} x \\ w \end{bmatrix}$ is stable if and only if $w \in L^2(I; \mathcal{W})$, or equivalently, if and only if $P_{\mathcal{U}}w \in L^2(I; \mathcal{U})$ for some fundamental decomposition $\mathcal{W} = -\mathcal{Y} \boxplus \mathcal{U}$ of $\mathcal{W}$. In the case where $\begin{bmatrix} x \\ w \end{bmatrix}$ is externally generated we have, in addition,

$$\begin{aligned} \|x(t)\|_{\mathcal{X}}^2 &\leq \int_{s \in I; s \leq t} [w(s), w(s)]_{\mathcal{W}} \, ds \\ &= \int_{s \in I; s \leq t} (\|P_{\mathcal{U}}w(s)\|_{\mathcal{U}}^2 - \|P_{\mathcal{Y}}w(s)\|_{\mathcal{Y}}^2) \, ds, \quad t \in I. \end{aligned} \quad (3.1)$$

In particular, if $I = \mathbb{R}^-$ and $\begin{bmatrix} x \\ w \end{bmatrix}$ is an externally generated stable past trajectory, then

$$\|x(0)\|_{\mathcal{X}}^2 \leq [w, w]_{K_-^2(\mathcal{W})} = \|P_{\mathcal{U}}w\|_{L_-^2(\mathcal{U})}^2 - \|P_{\mathcal{Y}}w\|_{L_-^2(\mathcal{Y})}^2. \quad (3.2)$$

Proof In Sect. 1 we outlined a proof of (1.9), and the results mentioned above follow from (1.9). $\square$

In the following lemma we list some elementary properties of the set of all trajectories of a passive s/s system $\Sigma = (V; \mathcal{X}, \mathcal{W})$. The notations τ^t and τ_+^t were explained at the end of Sect. 1.

Lemma 3.3 *The stable trajectories of a passive s/s system $\Sigma = (V; \mathcal{X}, \mathcal{W})$ have the following properties:*

- (i) *If $\begin{bmatrix} x \\ w \end{bmatrix}$ is a classical or generalized stable trajectory on some interval I and $t \in \mathbb{R}$, then $\begin{bmatrix} \tau^t x \\ \tau^t w \end{bmatrix}$ is a classical or generalized stable trajectory, respectively, of Σ on the interval $I - t := \{s \in \mathbb{R} \mid s + t \in I\}$, and $\begin{bmatrix} x \\ w \end{bmatrix}$ is externally generated on I if and only if $\begin{bmatrix} \tau^t x \\ \tau^t w \end{bmatrix}$ is externally generated on $I - t$.*
- (ii) *The restriction of a classical or generalized stable trajectory on some interval I' to a subinterval $I \subset I'$ is a classical or generalized stable trajectory of Σ on I , respectively, and if I and I' have the same left end-point, then the restricted trajectory is externally generated if and only if the original trajectory is externally generated.*
- (iii) *If $\begin{bmatrix} x \\ w \end{bmatrix}$ is a classical or generalized stable trajectory on $\mathbb{R}^+$, then $\begin{bmatrix} \tau_+^t x \\ \tau_+^t w \end{bmatrix}$ is a classical or generalized stable trajectory on $\mathbb{R}^+$ for all $t \in \mathbb{R}^+$.*

- (iv) *The set of all stable (generalized) trajectories and the set of all externally generated stable (generalized) trajectories of Σ on some interval I (finite or infinite) are closed subspaces of $\begin{bmatrix} BUC(I; \mathcal{X}) \\ L^2(I; \mathcal{W}) \end{bmatrix}$.*

Proof Claims (i)–(iii) follow immediately from Definition 3.1.

In order to prove (iv) we let $\begin{bmatrix} x_n \\ w_n \end{bmatrix}$ be a sequence of stable trajectories of Σ on I converging to $\begin{bmatrix} x \\ w \end{bmatrix}$ in $\begin{bmatrix} BUC(I; \mathcal{X}) \\ L^2(I; \mathcal{W}) \end{bmatrix}$. Both $C(I; \mathcal{X})$ and $L^2_{\text{loc}}(I; \mathcal{W})$ are Fréchet spaces, and by the definition of a generalized trajectory of Σ on I , we can find a sequence of classical trajectories $\begin{bmatrix} \tilde{x}_n \\ \tilde{w}_n \end{bmatrix}$ of Σ on I such that the distance from $\begin{bmatrix} \tilde{x}_n \\ \tilde{w}_n \end{bmatrix}$ to $\begin{bmatrix} x_n \\ w_n \end{bmatrix}$ in $\begin{bmatrix} C(I; \mathcal{X}) \\ L^2_{\text{loc}}(I; \mathcal{W}) \end{bmatrix}$ tends to zero as $n \rightarrow \infty$. Then $\begin{bmatrix} \tilde{x}_n \\ \tilde{w}_n \end{bmatrix}$ also tends to $\begin{bmatrix} x \\ w \end{bmatrix}$ in $\begin{bmatrix} C(I; \mathcal{X}) \\ L^2_{\text{loc}}(I; \mathcal{W}) \end{bmatrix}$ as $n \rightarrow \infty$. Thus, $\begin{bmatrix} x \\ w \end{bmatrix}$ is a trajectory of Σ on I . By assumption, $\begin{bmatrix} x \\ w \end{bmatrix} \in \begin{bmatrix} BUC(I; \mathcal{X}) \\ L^2(I; \mathcal{W}) \end{bmatrix}$, and hence $\begin{bmatrix} x \\ w \end{bmatrix}$ is stable. If all x_n tend to zero at the left end-point of I , then so does x (because of the uniform convergence), and hence $\begin{bmatrix} x \\ w \end{bmatrix}$ is externally generated if all the trajectories $\begin{bmatrix} x_n \\ w_n \end{bmatrix}$ are externally generated. $\square$

In the proof of Lemma 3.3 we made only marginal use of the passivity of Σ (i.e., we did not use any other properties of V than the closedness). However, in the proof of the following lemma we shall make significant use of the passivity of Σ (or more precisely, of the fact that every passive s/s system is well-posed in the sense of [13]).

Lemma 3.4 *The set of (generalized) trajectories of a passive s/s system $\Sigma = (V; \mathcal{X}, \mathcal{W})$ has the following properties:*

- (i) *Let $\mathcal{W} = -\mathcal{Y} \boxplus \mathcal{U}$ be a fundamental decomposition of $\mathcal{W}$. Then, for each $x_0 \in \mathcal{X}$, each closed interval I with a finite left end-point t_0 , and each $u \in L^2(I; \mathcal{U})$, there exists a unique stable trajectory $\begin{bmatrix} x \\ w \end{bmatrix}$ of Σ on I satisfying $x(t_0) = x_0$ and $P_{\mathcal{U}}w = u$.*
- (ii) *Let $\begin{bmatrix} x_1 \\ w_1 \end{bmatrix}$ be a stable trajectory of Σ on the finite interval $I_1 = [t_0, t_1]$, and let $\begin{bmatrix} x_2 \\ w_2 \end{bmatrix}$ be a stable trajectory of Σ on a closed interval I_2 with left end-point t_1 . Then the concatenation $\begin{bmatrix} x \\ w \end{bmatrix}$ defined by*

$$\begin{bmatrix} x(t) \\ w(t) \end{bmatrix} := \begin{cases} \begin{bmatrix} x_1(t) \\ w_1(t) \end{bmatrix}, & t \in I_1, \\ \begin{bmatrix} x_2(t) \\ w_2(t) \end{bmatrix}, & t \in I_2 \setminus \{t_1\}, \end{cases}, \quad (3.3)$$

is a stable trajectory of Σ on $I := I_1 \cup I_2$ if and only if $x_1(t_1) = x_2(t_1)$.

- (iii) *Every stable trajectory on some finite interval $I = [t_0, t_1]$ can be extended to a stable trajectory of Σ on $[t_0, \infty)$. This extension can be chosen so that $\pi_{[t_1, \infty)} P_{\mathcal{U}}w = u$ for an arbitrary $u \in L^2([t_1, \infty); \mathcal{U})$, and it is uniquely determined by u .*
- (iv) *A pair of functions $\begin{bmatrix} x \\ w \end{bmatrix}$ on an interval $[t_0, \infty)$ is a stable trajectory of Σ on $[t_0, \infty)$ if and only if the restriction of $\begin{bmatrix} x \\ w \end{bmatrix}$ to every finite subinterval $[t_0, t_1]$ of $[t_0, \infty)$ is a trajectory of Σ on $[t_0, t_1]$ and, in addition, $P_{\mathcal{U}}w \in L^2([t_0, \infty), \mathcal{U})$ for some fundamental decomposition $\mathcal{W} = -\mathcal{Y} \boxplus \mathcal{U}$ of $\mathcal{W}$.*

Proof (i) By Lemma 3.3(i), it suffices to prove the case where $t_0 = 0$. By [15, Proposition 5.8], the i/o pair $(\mathcal{U}, \mathcal{Y})$ is L^2 -well-posed for Σ . Theorem 6.6 of [13] then yields that for every $x_0 \in \mathcal{X}$ and $u \in L^2_{\text{loc}}(\mathbb{R}^+; \mathcal{U})$ the system Σ has a unique future trajectory $\begin{bmatrix} x \\ w \end{bmatrix}$, such that $x(0) = x_0$ and $P_{\mathcal{U}}w = u$. According to Lemma 3.2, a sufficient condition for the stability of $\begin{bmatrix} x \\ w \end{bmatrix}$ is that $w \in K^2(\mathbb{R}^+; \mathcal{W})$. Indeed, this condition is satisfied due to the fact that, by (1.9),

$$\|P_{\mathcal{Y}}w\|_{L^2(\mathbb{R}^+; \mathcal{Y})}^2 \leq \|x(0)\|_{\mathcal{X}}^2 + \|u\|_{L^2(\mathbb{R}^+; \mathcal{U})}^2.$$

(ii) Claim (ii) is proved in [13, Proposition 3.8]

(iii) Claim (iii) follows from (i) and (ii).

(iv) According to [13, Proposition 3.9], $\begin{bmatrix} x \\ w \end{bmatrix}$ is a trajectory of Σ on $[t_0, \infty)$ if and only if the restriction of $\begin{bmatrix} x \\ w \end{bmatrix}$ to every finite subinterval $[t_0, t_1]$ is a trajectory of Σ on $[t_0, t_1]$. By Lemma 3.2, this trajectory is stable if and only if $P_{\mathcal{U}} \in L^2([t_0, \infty); \mathcal{U})$. $\square$

The following theorem plays a key role in our extension of many of the discrete time results developed in [5] and [6] to a continuous time setting (see Remark 3.17 below). In this theorem we need the family $\mathfrak{K}_{0,t} := \begin{bmatrix} \mathcal{X} \\ K^2([0,t]; \mathcal{W}) \end{bmatrix}$, $t \in \mathbb{R}^+$, of Kreĭn spaces with the indefinite inner products (1.12) as well as the Kreĭn space $\mathfrak{L}_{0,\infty} := \begin{bmatrix} \mathcal{X} \\ K^2(\mathbb{R}^+; \mathcal{W}) \end{bmatrix}$ with the natural inner product (1.13). It follows immediately that $\mathfrak{K}_{0,t}$ and $\mathfrak{L}_{0,\infty}$ are Kreĭn spaces with fundamental decompositions $\mathfrak{K}_{0,t} = -\mathfrak{K}_{0,t,-} \boxplus \mathfrak{K}_{0,t,+}$ and $\mathfrak{L}_{0,\infty} = -\mathfrak{L}_{0,\infty,-} \boxplus \mathfrak{L}_{0,\infty,+}$, where

$$\begin{aligned} \mathfrak{K}_{0,t,+} &:= \begin{bmatrix} \{0\} \\ \mathcal{X} \\ L^2([0,t]; \mathcal{U}) \end{bmatrix}, & \mathfrak{K}_{0,t,-} &:= \begin{bmatrix} \mathcal{X} \\ \{0\} \\ L^2([0,t]; \mathcal{Y}) \end{bmatrix}, \quad t \geq 0, \\ \mathfrak{L}_{0,\infty,+} &:= \begin{bmatrix} \mathcal{X} \\ L^2(\mathbb{R}^+; \mathcal{U}) \end{bmatrix}, & \mathfrak{L}_{0,\infty,-} &:= \begin{bmatrix} \{0\} \\ L^2(\mathbb{R}^+; \mathcal{Y}) \end{bmatrix}. \end{aligned} \quad (3.4)$$

and $\mathcal{W} = -\mathcal{Y} \boxplus \mathcal{U}$ is an arbitrary fundamental decomposition of $\mathcal{W}$.

Theorem 3.5 *Let $\mathcal{X}$ be a Hilbert space, let $\mathcal{W}$ be a Kreĭn space, and let $\mathcal{T}_+$ be a subspace of $\begin{bmatrix} BUC(\mathbb{R}^+; \mathcal{X}) \\ K^2(\mathbb{R}^+; \mathcal{W}) \end{bmatrix}$. Define*

$$\begin{aligned} \mathcal{T}_{0,t} &:= \left\{ \left[\begin{array}{c} x(t) \\ x(0) \\ \pi_{[0,t]}w \end{array} \right] \mid \begin{bmatrix} x \\ w \end{bmatrix} \in \mathcal{T}_+ \right\}, \quad t \in \mathbb{R}^+, \\ \mathcal{S}_{0,\infty} &:= \left\{ \left[\begin{array}{c} x(0) \\ w \end{array} \right] \mid \begin{bmatrix} x \\ w \end{bmatrix} \in \mathcal{T}_+ \right\}. \end{aligned} \quad (3.5)$$

Then the subspace $\mathcal{T}_+$ is the family of all stable future trajectories of some passive s/s system $\Sigma = (V; \mathcal{X}, \mathcal{W})$ if and only if the following three conditions hold:

(i) $\mathcal{T}_+$ is left-shift invariant, i.e.,

$$\tau'_+ \mathcal{T}_+ := \{ \tau'_+ \begin{bmatrix} x \\ w \end{bmatrix} \mid \begin{bmatrix} x \\ w \end{bmatrix} \in \mathcal{T}_+ \} \subset \mathcal{T}_+, \quad t \in \mathbb{R}^+; \quad (3.6)$$

- (ii) For all $t \in \mathbb{R}^+$, $\mathcal{T}_{0,t}$ is a maximal nonnegative subspace of $\mathfrak{K}_{0,t}$.
- (iii) $\mathcal{S}_{0,\infty}$ is a maximal nonnegative subspace of $\mathfrak{L}_{0,\infty}$.

Proof This proof is based on [13, Lemma 4.7], and we refer the reader to [13] for the precise definitions of some of the s/s and i/s/o notions that we use in this proof. The monograph [19] can be used as an alternative source for basic results on well-posed i/s/o systems.

Throughout this proof we let $\mathcal{W} = -\mathcal{U} \boxplus \mathcal{Y}$ be a fundamental decomposition of $\mathcal{W}$, and let $\mathfrak{K}_{0,t} = -\mathfrak{K}_{0,t,-} \boxplus \mathfrak{K}_{0,t,+}$ and $\mathfrak{L}_{0,\infty} = -\mathfrak{L}_{0,\infty,-} \boxplus \mathfrak{L}_{0,\infty,+}$ be the corresponding fundamental decompositions defined in (3.4).

Step 1: Necessity of properties (i)–(iii). Let $\mathcal{T}_+$ be the space of stable future trajectories of the passive s/s system Σ . Then the left-shift invariance of $\mathcal{T}_+$ follows from Lemma 3.3(iii). The nonnegativity of $\mathcal{T}_{0,t}$ in $\mathfrak{K}_{0,t}$ and the nonnegativity of $\mathcal{S}_{0,\infty}$ in $\mathfrak{L}_{0,\infty}$ follow from (1.9) with $t_1 = 0$ and $t_2 = t$. To see that $\mathcal{T}_{0,t}$ is maximal nonnegative we argue as follows. It follows from Lemma 3.4(i) that the projection of $\mathcal{T}_{0,t}$ onto $\mathfrak{K}_{0,t,+}$ is all of $\mathfrak{K}_{0,t,+}$ and the projection of $\mathcal{S}_{0,\infty}$ onto $\mathfrak{L}_{0,\infty,+}$ is all of $\mathfrak{L}_{0,\infty,+}$, and consequently, by Proposition 2.1(i), $\mathcal{T}_{0,t}$ and $\mathcal{S}_{0,\infty}$ are maximal nonnegative.

Step 2: Characterization of the closure of $\mathcal{T}_+$ in $\begin{bmatrix} C(\mathbb{R}^+; \mathcal{X}) \\ L_{\text{loc}}^2(\mathbb{R}^+; \mathcal{W}) \end{bmatrix}$. For the proof of the sufficiency of (i)–(iii) we let $\mathcal{T}_+$ be a subspace of $\begin{bmatrix} BUC(\mathbb{R}^+; \mathcal{X}) \\ K^2(\mathbb{R}^+; \mathcal{W}) \end{bmatrix}$ with properties (i)–(iii). Since this part of our proof is based on Lemma [13, Lemma 4.7] we must show that the closure $\overline{\mathcal{T}_+}$ of $\mathcal{T}_+$ in $\begin{bmatrix} BUC(\mathbb{R}^+; \mathcal{X}) \\ L_{\text{loc}}^2(\mathbb{R}^+; \mathcal{W}) \end{bmatrix}$ has the following three properties:

- (i') $\overline{\mathcal{T}_+}$ is left-shift invariant;
- (ii') For all $\begin{bmatrix} x \\ w \end{bmatrix} \in \overline{\mathcal{T}_+}$ and all $t \in \mathbb{R}^+$ we have

$$\|x(t)\|_{\mathcal{X}}^2 + \int_0^t \|P_{\mathcal{Y}}w(s)\|_{\mathcal{Y}}^2 ds \leq \|x(0)\|_{\mathcal{X}}^2 + \int_0^t \|P_{\mathcal{U}}w(s)\|_{\mathcal{U}}^2 ds; \quad (3.7)$$

- (iii') For all $x_0 \in \mathcal{X}$ and $u \in L_{\text{loc}}^2(\mathbb{R}^+; \mathcal{U})$ there exist a unique $\begin{bmatrix} x \\ w \end{bmatrix} \in \overline{\mathcal{T}_+}$ satisfying $x(0) = x_0$ and $P_{\mathcal{U}}w = u$.

Clearly (i') follows from the left-shift invariance of $\mathcal{T}_+$. Moreover, the uniqueness in (iii') follows from (ii'), so it suffices to prove existence in (iii') in addition to (ii').

Fix $x_0 \in \mathcal{X}$ and $u \in L_{\text{loc}}^2(\mathbb{R}^+; \mathcal{W})$, and let $n \in \mathbb{Z}^+$. By the maximal nonnegativity of $\mathcal{T}_{[0,n]}$ in $\mathfrak{K}_{[0,n]}$, the definition of $\mathcal{T}_{[0,n]}$, and Proposition 2.1(i), there exists some $\begin{bmatrix} x_n \\ w_n \end{bmatrix} \in \mathcal{T}_+$ such that $x(0) = x_0$ and $\pi_{[0,n]}P_{\mathcal{U}}w_n = \pi_{[0,n]}u$. By the nonnegativity of $\mathcal{T}_{0,t}$ for all $t \in \mathbb{R}^+$, (3.7) holds with $\begin{bmatrix} x \\ w \end{bmatrix}$ replaced by $\begin{bmatrix} x_n \\ w_n \end{bmatrix}$. We claim that $\begin{bmatrix} x_n \\ w_n \end{bmatrix}$ tends to a limit in $\begin{bmatrix} C(\mathbb{R}^+; \mathcal{X}) \\ L_{\text{loc}}^2(\mathbb{R}^+; \mathcal{W}) \end{bmatrix}$ as $n \rightarrow \infty$. Indeed, for all $T > 0$ and all $m, n \geq T$, we get from (3.7) applied to $\begin{bmatrix} x_n - x_m \\ w_n - w_m \end{bmatrix}$, combined with the conditions $x_n(0) = x_m(0) = x_0$ and $\pi_{[0,T]}P_{\mathcal{U}}w_n = \pi_{[0,T]}u = \pi_{[0,T]}P_{\mathcal{U}}w_m$, that

$$\begin{aligned}
\sup_{0 \leq t \leq T} \|x_n(t) - x_m(t)\|_{\mathcal{X}}^2 &\leq \|x_n(0) - x_m(0)\|_{\mathcal{X}}^2 \\
&\quad + \int_0^T \|P_{\mathcal{U}}(w_n(s) - w_m(s))\|_{\mathcal{W}}^2 ds = 0, \\
\int_0^T \|P_{\mathcal{Y}}(w_n(s) - w_m(s))\|_{\mathcal{V}}^2 ds &\leq \int_0^T \|P_{\mathcal{U}}(w_n(s) - w_m(s))\|_{\mathcal{W}}^2 ds = 0.
\end{aligned}$$

Thus, if we define $\begin{bmatrix} x \\ w \end{bmatrix}$ by $\begin{bmatrix} x(s) \\ w(s) \end{bmatrix} = \begin{bmatrix} x_n(s) \\ w_n(s) \end{bmatrix}$ for $s \in [n-1, n)$, $n \in \mathbb{Z}^+$, then $\pi_{[0,n)} \begin{bmatrix} x_n \\ w_n \end{bmatrix} = \pi_{[0,n)} \begin{bmatrix} x \\ w \end{bmatrix}$, $n \in \mathbb{Z}^+$, and hence $\begin{bmatrix} x_n \\ w_n \end{bmatrix} \rightarrow \begin{bmatrix} x \\ w \end{bmatrix}$ in $\begin{bmatrix} C(\mathbb{R}^+; \mathcal{W}) \\ L_{\text{loc}}^2(\mathbb{R}^+; \mathcal{W}) \end{bmatrix}$ as $n \rightarrow \infty$. Since each $\begin{bmatrix} x_n \\ w_n \end{bmatrix} \in \mathcal{T}_+$, the limit $\begin{bmatrix} x \\ w \end{bmatrix}$ belongs to $\overline{\mathcal{T}_+}$, and (3.7) holds since we know that (3.7) holds with $\begin{bmatrix} x \\ w \end{bmatrix}$ replaced by $\begin{bmatrix} x_n \\ w_n \end{bmatrix}$ for all n . This proves (ii') and (iii').

Step 3: Existence of L^2 -well-posed scattering passive i/s/o representation. By Step 2 and [13, Lemma 4.7], there exists a L^2 -well-posed i/s/o system $\begin{bmatrix} \mathfrak{A} & \mathfrak{B} \\ \mathfrak{C} & \mathfrak{D} \end{bmatrix}$, such that

$$\begin{aligned}
\overline{\mathcal{T}_+} &= \left\{ \begin{bmatrix} x \\ w \end{bmatrix} \in \begin{bmatrix} C(\mathbb{R}^+; \mathcal{X}) \\ K_{\text{loc}}^2(\mathbb{R}^+; \mathcal{W}) \end{bmatrix} \mid \forall t \geq 0 : \right. \\
&\quad \left. \begin{bmatrix} x(t) \\ P_{\mathcal{Y}} w \end{bmatrix} = \begin{bmatrix} \mathfrak{A}^t & \mathfrak{B} \tau^t \\ \mathfrak{C} & \mathfrak{D} \end{bmatrix} \begin{bmatrix} x(0) \\ P_{\mathcal{U}} w \end{bmatrix} \right\}.
\end{aligned} \tag{3.8}$$

That this system is scattering passive follows from (3.7) and [19, Lemma 11.1.4].

Step 4: Existence of passive s/s system. By Step 2 and [13, Theorem 6.6], there exists a L^2 -well-posed s/s system $\Sigma = (V; \mathcal{X}, \mathcal{W})$ such that $\overline{\mathcal{T}_+}$ is the set of all future (generalized) trajectories of Σ . The same theorem says that the decomposition $-\mathcal{Y} \boxplus \mathcal{U}$ is admissible, and it follows from (3.8) that the system $\begin{bmatrix} \mathfrak{A} & \mathfrak{B} \\ \mathfrak{C} & \mathfrak{D} \end{bmatrix}$ is an i/s/o representation of Σ . Every L^2 -well-posed s/s system has a unique (maximal) generating subspace V in the sense of [13, Theorem 6.4], and by that theorem, this subspace V is given by

$$V = \left[\begin{array}{c} A \& B \\ [1 \ 0] \\ C \& D + [0 \ 1] \end{array} \right] \text{dom} \left(\begin{bmatrix} A \& B \\ C \& D \end{bmatrix} \right), \tag{3.9}$$

where $\begin{bmatrix} A \& B \\ C \& D \end{bmatrix} : \begin{bmatrix} \mathcal{X} \\ \mathcal{U} \end{bmatrix} \supset \text{dom} \left(\begin{bmatrix} A \& B \\ C \& D \end{bmatrix} \right) \rightarrow \begin{bmatrix} \mathcal{X} \\ \mathcal{Y} \end{bmatrix}$ is the system node of $\begin{bmatrix} \mathfrak{A} & \mathfrak{B} \\ \mathfrak{C} & \mathfrak{D} \end{bmatrix}$ (see, e.g., [19, Section 4.6] or [13, Sect. 5] for the definition of the system node of a well-posed linear i/s/o system). By [19, Theorem 11.1.5] and [15, Proposition 5.6], V is maximal nonnegative in the node space $\mathfrak{K}$ (note that condition (iii) in [19, Theorem 11.1.5] is identical to condition (ii) in [15, Proposition 5.6]). Consequently, Σ is passive.

Step 5: $\mathcal{T}_+$ is the space of stable future trajectories of Σ . By construction $\mathcal{T}_+ \subset \overline{\mathcal{T}_+} \cap \begin{bmatrix} BUC(\mathbb{R}^+; \mathcal{X}) \\ K^2(\mathbb{R}^+; \mathcal{W}) \end{bmatrix}$, and every $\begin{bmatrix} w \\ w \end{bmatrix} \in \overline{\mathcal{T}_+}$ is a future trajectory of Σ . Consequently, every $\begin{bmatrix} x \\ w \end{bmatrix} \in \mathcal{T}_+$ is a stable future trajectory of Σ . Conversely, let $\begin{bmatrix} x \\ w \end{bmatrix}$ be a stable future trajectory of Σ . Since $\mathcal{S}_{0,\infty}$ is maximal nonnegative in $\mathfrak{L}_{0,\infty}$ and $\mathfrak{L}_{0,\infty} = -\mathfrak{L}_{0,\infty,-} \boxplus \mathfrak{L}_{0,\infty,+}$

is a fundamental decomposition of $\mathfrak{K}_+$, there exists some $\begin{bmatrix} \tilde{x} \\ \tilde{w} \end{bmatrix} \in \mathcal{T}_+$ with $\tilde{x}(0) = x(0)$ and $P_{\mathcal{U}}\tilde{w} = P_{\mathcal{U}}w$. Since $\begin{bmatrix} x \\ w \end{bmatrix} - \begin{bmatrix} \tilde{x} \\ \tilde{w} \end{bmatrix} \in \overline{\mathcal{T}_+}$, we can apply (3.7) to $\begin{bmatrix} x-\tilde{x} \\ w-\tilde{w} \end{bmatrix}$, and conclude that $x = \tilde{x}$ and $P_{\mathcal{U}}w = P_{\mathcal{U}}\tilde{w}$. Thus, $\begin{bmatrix} x \\ w \end{bmatrix} = \begin{bmatrix} \tilde{x} \\ \tilde{w} \end{bmatrix} \in \mathcal{T}_+$, and we have proved that every stable future trajectory of Σ belongs to $\mathcal{T}_+$. $\square$

3.2 Classical Trajectories and the Generating Subspace

We originally defined the notion of a trajectory of a passive s/s system $\Sigma = (V; \mathcal{X}, \mathcal{W})$ by means of the generating subspace V . Below we shall study the converse problem: how to recreate the generating subspace from the family of all classical trajectories. (We have already encountered one result of this type in the proof of Theorem 3.5.) For simplicity we primarily restrict ourselves to future trajectories, i.e., trajectories defined on $\mathbb{R}^+$.

We begin with a preliminary lemma which gives a universal method to construct a sequence of classical approximations of an arbitrary future trajectory.

Lemma 3.6 *Let $\begin{bmatrix} x \\ w \end{bmatrix}$ be a future trajectory of the passive s/s system $\Sigma = (V; \mathcal{X}, \mathcal{W})$. For each $n \in \mathbb{Z}^+$, define $\begin{bmatrix} x_n \\ w_n \end{bmatrix}$ by*

$$\begin{bmatrix} x_n(t) \\ w_n(t) \end{bmatrix} := n \int_t^{t+1/n} \begin{bmatrix} x(s) \\ w(s) \end{bmatrix} ds, \quad t \in \mathbb{R}^+. \quad (3.10)$$

Then $\begin{bmatrix} x_n \\ w_n \end{bmatrix}$ is a classical future trajectory of Σ , and $\begin{bmatrix} x_n \\ w_n \end{bmatrix} \rightarrow \begin{bmatrix} x \\ w \end{bmatrix}$ in $\begin{bmatrix} C(\mathbb{R}^+; \mathcal{X}) \\ L^2_{\text{loc}}(\mathbb{R}^+; \mathcal{W}) \end{bmatrix}$ as $n \rightarrow \infty$. If $\begin{bmatrix} x \\ w \end{bmatrix}$ is stable, then so is $\begin{bmatrix} x_n \\ w_n \end{bmatrix}$, and $\begin{bmatrix} x_n \\ w_n \end{bmatrix} \rightarrow \begin{bmatrix} x \\ w \end{bmatrix}$ in $\begin{bmatrix} BUC(\mathbb{R}^+; \mathcal{X}) \\ L^2(\mathbb{R}^+; \mathcal{W}) \end{bmatrix}$ as $n \rightarrow \infty$.

This result is essentially contained in [14, Corollary 2.4]. For the convenience of the reader we have included a proof.

Proof of Lemma 3.6. Clearly, each $\begin{bmatrix} x_n \\ w_n \end{bmatrix} \in \begin{bmatrix} C^1(\mathbb{R}^+; \mathcal{X}) \\ C(\mathbb{R}^+; \mathcal{W}) \end{bmatrix}$. Let $\begin{bmatrix} x^k \\ w^k \end{bmatrix}$ be a sequence of classical future trajectories of Σ converging to $\begin{bmatrix} x \\ w \end{bmatrix} \in \begin{bmatrix} C(\mathbb{R}^+; \mathcal{X}) \\ L^2_{\text{loc}}(\mathbb{R}^+; \mathcal{W}) \end{bmatrix}$. Since the trajectories $\begin{bmatrix} x^k \\ w^k \end{bmatrix}$ are classical, they satisfy $\begin{bmatrix} \dot{x}^k(s) \\ x^k(s) \\ w^k(s) \end{bmatrix} \in V$ for all $s \in \mathbb{R}^+$, and since V is closed, also

$$n \int_t^{t+1/n} \begin{bmatrix} \dot{x}^k(s) \\ x^k(s) \\ w^k(s) \end{bmatrix} ds \in V.$$

As $k \rightarrow \infty$,

$$\begin{bmatrix} x_n^k(t) \\ w_n^k(t) \end{bmatrix} = n \int_t^{t+1/n} \begin{bmatrix} x^k(s) \\ w^k(s) \end{bmatrix} ds \rightarrow n \int_t^{t+1/n} \begin{bmatrix} x(s) \\ w(s) \end{bmatrix} ds = \begin{bmatrix} x_n(t) \\ w_n(t) \end{bmatrix}$$

and

$$n \int_t^{t+1/n} \dot{x}^k(s) ds = n[x^k(t+1/n) - x^k(t)] \rightarrow n[x(t+1/n) - x(t)] = \dot{x}_n(t).$$

Consequently, $n \int_t^{t+1/n} \begin{bmatrix} \dot{x}^k(s) \\ x^k(s) \\ w^k(s) \end{bmatrix} ds \rightarrow \begin{bmatrix} \dot{x}_n(t) \\ x_n(t) \\ w_n(t) \end{bmatrix}$ as $k \rightarrow \infty$, and since V is closed, we find that $\begin{bmatrix} \dot{x}_n(t) \\ x_n(t) \\ w_n(t) \end{bmatrix} \in V$ for all $t \in \mathbb{R}^+$. Thus, $\begin{bmatrix} x_n \\ w_n \end{bmatrix}$ is a classical future trajectory of Σ .

All the additional claims in Lemma 3.6 follow from standard properties of approximate identities (the scalar versions of these results are found in many places, such as [12, p. 67], and the vector-valued versions can be proved in the same way). $\square$

Proposition 3.7 *Let $\Sigma = (V; \mathcal{X}, \mathcal{W})$ be a passive s/s system. Then the following claims are true:*

- (i) *If $\begin{bmatrix} x \\ w \end{bmatrix}$ is a (generalized) trajectory of Σ on some interval $[t_0, t_0 + h]$, and if both $z_0 := \lim_{t \rightarrow t_0+} \frac{1}{t}(x(t) - x(t_0))$ and $w_0 := \lim_{t \rightarrow t_0+} \frac{1}{t} \int_{t_0}^t w(s) ds$ exist, then $\begin{bmatrix} z_0 \\ x(t_0) \\ w_0 \end{bmatrix} \in V$.*
- (ii) *For each $\begin{bmatrix} z_0 \\ x_0 \\ w_0 \end{bmatrix} \in V$ there exists a stable future classical trajectory $\begin{bmatrix} x \\ w \end{bmatrix}$ satisfying $\begin{bmatrix} \dot{x}(0) \\ x(0) \\ w(0) \end{bmatrix} = \begin{bmatrix} z_0 \\ x_0 \\ w_0 \end{bmatrix}$ with the additional property that w is locally absolutely continuous and $\begin{bmatrix} \dot{x} \\ x \\ w \end{bmatrix}$ is a stable future trajectory of Σ . In particular,*

$$V = \left\{ \begin{bmatrix} \dot{x}(0) \\ x(0) \\ w(0) \end{bmatrix} \middle| \begin{bmatrix} x \\ w \end{bmatrix} \text{ is a future classical trajectory of } \Sigma \right\}. \quad (3.11)$$

- (iii) *A (generalized) trajectory $\begin{bmatrix} x \\ w \end{bmatrix}$ of Σ on some interval I is classical if and only if $\begin{bmatrix} x \\ w \end{bmatrix} \in \begin{bmatrix} C^1(I; \mathcal{X}) \\ C(I; \mathcal{W}) \end{bmatrix}$.*
- (iv) *There is a one-to-one correspondence between the passive s/s system $\Sigma = (V; \mathcal{X}, \mathcal{W})$ and the set of all classical future trajectories of Σ , and also between Σ and the set of all generalized future trajectories of Σ .*

Parts of this proposition are also found in [14, Theorem 3.1 and Corollary 3.2].

Proof of Proposition 3.7. (i) By the shift-invariance expressed in Lemma 3.3(i) it suffices to treat the case $t_0 = 0$. If $h < \infty$, then we first extend $\begin{bmatrix} x \\ w \end{bmatrix}$ to a trajectory

defined on all of $\mathbb{R}^+$ in an arbitrary way; cf. Lemma 3.4(iii). Let $\begin{bmatrix} x_n \\ w_n \end{bmatrix}$ be the family of classical approximations of $\begin{bmatrix} x \\ w \end{bmatrix}$ defined in Lemma 3.6. Then, for all $n \in \mathbb{Z}^+$,

$$\begin{bmatrix} \dot{x}_n(0) \\ x_n(0) \\ w_n(0) \end{bmatrix} = \frac{1}{n} \begin{bmatrix} x(1/n) - x(0) \\ \int_0^{1/n} x(s) \, ds \\ \int_0^{1/n} w(s) \, ds \end{bmatrix} \in V.$$

This tends to $\begin{bmatrix} z_0 \\ x(0) \\ w_0 \end{bmatrix}$ as $n \rightarrow \infty$, and since V is closed, it follows that $\begin{bmatrix} z_0 \\ x(0) \\ w_0 \end{bmatrix} \in V$.

(ii) Let $\mathcal{W} = -\mathcal{Y} \boxplus \mathcal{U}$ be a fundamental decomposition of $\mathcal{W}$, and let $\begin{bmatrix} \mathfrak{A} & \mathfrak{B} \\ \mathfrak{C} & \mathfrak{D} \end{bmatrix}$ be the i/s/o representation of Σ constructed in the proof of Theorem 3.5. Since $\begin{bmatrix} z_0 \\ x_0 \\ w_0 \end{bmatrix} \in V$, it follows from (3.9) that $\begin{bmatrix} x_0 \\ w_0 \end{bmatrix} \in \text{dom} \left(\begin{bmatrix} A \& B \\ C \& D \end{bmatrix} \right)$, and that

$$\begin{bmatrix} z_0 \\ w_0 \end{bmatrix} = \begin{bmatrix} A \& B \\ C \& D \end{bmatrix} \begin{bmatrix} x_0 \\ u_0 \end{bmatrix} + \begin{bmatrix} 0 \\ u_0 \end{bmatrix}.$$

Let u be an arbitrary function in $C^\infty(\mathbb{R}^+; \mathcal{U})$ with compact support and with $u(0) = u_0$, define

$$\begin{bmatrix} x(t) \\ y \end{bmatrix} := \begin{bmatrix} \mathfrak{A}^t & \mathfrak{B} \tau^t \\ \mathfrak{C} & \mathfrak{D} \end{bmatrix} \begin{bmatrix} x(0) \\ u \end{bmatrix}, \quad t \in \mathbb{R}^+,$$

and take $w = u + y$. By Theorem 3.5 and its proof, $\begin{bmatrix} x \\ w \end{bmatrix}$ is a future trajectory of Σ , and it follows from (3.7) that this trajectory is stable. Moreover, $\begin{bmatrix} x(0) \\ w(0) \end{bmatrix} = \begin{bmatrix} x_0 \\ w_0 \end{bmatrix}$.

By [19, Theorem 4.6.11], $\begin{bmatrix} x \\ w \end{bmatrix} \in \begin{bmatrix} C^1(\mathbb{R}^+; \mathcal{X}) \\ C(\mathbb{R}^+; \mathcal{W}) \end{bmatrix}$, $\dot{x}(0) = z_0$, and y is locally absolutely continuous with a distribution derivative $\dot{y} \in L^2_{\text{loc}}(\mathbb{R}^+; \mathcal{Y})$. In particular, by part (i), $\begin{bmatrix} x \\ w \end{bmatrix}$ is a classical trajectory of Σ , and w is locally absolutely continuous with $\dot{w} \in L^2_{\text{loc}}(\mathbb{R}^+; \mathcal{W})$.

For each $n \in \mathbb{Z}^+$, we define the sequence $\begin{bmatrix} x_n \\ w_n \end{bmatrix}$ by

$$\begin{bmatrix} x_n(t) \\ w_n(t) \end{bmatrix} = n \int_t^{t+1/n} \begin{bmatrix} \dot{x}(s) \\ \dot{w}(s) \end{bmatrix} \, ds = \begin{bmatrix} n(x(t+1/n) - x(t)) \\ n(w(t+1/n) - w(t)) \end{bmatrix}, \quad n \in \mathbb{Z}^+.$$

The set of all classical future trajectory of Σ is a left-shift invariant subspace, and consequently each $\begin{bmatrix} x_n \\ w_n \end{bmatrix}$ is a classical future trajectory of Σ . In the same way as in the proof of Lemma 3.6 (with $\begin{bmatrix} x \\ w \end{bmatrix}$ replaced by $\begin{bmatrix} \dot{x} \\ \dot{w} \end{bmatrix}$) we find that $x_n \rightarrow \dot{x}$ in $C(\mathbb{R}^+; \mathcal{X})$ and $w_n \rightarrow \dot{w}$ in $L^2_{\text{loc}}(\mathbb{R}^+; \mathcal{W})$ as $n \rightarrow \infty$. It then follows from Lemma 3.3(iv) that the restriction of $\begin{bmatrix} \dot{x} \\ \dot{w} \end{bmatrix}$ to each finite interval $[0, t_2]$ is trajectory of Σ on $[0, t_2]$, and from Lemma 3.4(iv) that $\begin{bmatrix} \dot{x} \\ \dot{w} \end{bmatrix}$ is a stable future trajectory of Σ .

(iii) In the case where the right end-point of I is $+\infty$ claim (iii) follows from (i) (combined with the obvious fact that every classical trajectory is also a generalized trajectory). If I has a finite right end-point t_1 , then $\begin{bmatrix} \dot{x}(t_1-1/n) \\ x(t_1-1/n) \\ w(t_1-1/n) \end{bmatrix} \in V$ for all sufficiently

large n and $\begin{bmatrix} \dot{x}(t_1-1/n) \\ x(t_1-1/n) \\ w(t_1-1/n) \end{bmatrix} \rightarrow \begin{bmatrix} \dot{x}(t_1) \\ x(t_1) \\ w(t_1) \end{bmatrix}$ as $n \rightarrow \infty$. Since V is closed, this implies that $\begin{bmatrix} \dot{x}(t_1) \\ x(t_1) \\ w(t_1) \end{bmatrix} \in V$. Thus, also in this case $\begin{bmatrix} x \\ w \end{bmatrix}$ is a classical trajectory on the full interval I .

(iv) Clearly, the generating subspace V of Σ determines the sets of all future smooth and generalized trajectories of Σ uniquely. Conversely, formula (3.11) defines V uniquely in terms of the set of all classical future trajectories of Σ , and (iii) defines the set of all classical future trajectories of Σ uniquely in terms of the set of all generalized future trajectories of Σ . $\square$

Proposition 3.8 *Let $\Sigma = (V; \mathcal{X}, \mathcal{W})$ be a passive s/s system, and let $\begin{bmatrix} x \\ w \end{bmatrix}$ be a future trajectory of Σ for which w is locally absolutely continuous and $\dot{w} \in L^2_{\text{loc}}(\mathbb{R}^+; \mathcal{W})$. Then $\begin{bmatrix} x \\ w \end{bmatrix}$ is a classical trajectory if and only if $\dot{x}(0) := \lim_{t \rightarrow 0+} \frac{1}{t}(x(t) - x(0))$ exists.*

Proof The existence of $\dot{x}(0)$ is necessary for $\begin{bmatrix} x \\ w \end{bmatrix}$ to be a classical solution. Conversely, if $\dot{x}(0)$ exists, then it follows from Proposition 3.7(i) that $\begin{bmatrix} \dot{x}(0) \\ x(0) \\ w(0) \end{bmatrix} \in V$. That $x \in C^1(\mathbb{R}^+; \mathcal{X})$ then follows from [19, Theorem 4.6.11] in the same way as in the preceding proof, and by Proposition 3.7, this implies that $\begin{bmatrix} x \\ w \end{bmatrix}$ is a classical solution. $\square$

3.3 More on Externally Generated Stable Trajectories

Here we continue our study of externally generated trajectories begun in Sect. 3.1. In particular, we now allow the left end-point of the interval I on which the trajectories are defined to be $-\infty$.

Lemma 3.9 *Let $\Sigma = (V; \mathcal{X}, \mathcal{W})$ be a passive s/s system.*

- (i) *Let $\begin{bmatrix} x \\ w \end{bmatrix}$ be an externally generated stable trajectory $\begin{bmatrix} x \\ w \end{bmatrix}$ of Σ on $[t_0, \infty)$. Then there exists a sequence of stable classical trajectories $\begin{bmatrix} x_n \\ w_n \end{bmatrix}$ of Σ on $[t_0, \infty)$ which satisfies $\begin{bmatrix} x_n(t_0) \\ w_n(t_0) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ for all n and tends to $\begin{bmatrix} x \\ w \end{bmatrix}$ in $\begin{bmatrix} BUC([t_0, \infty); \mathcal{X}) \\ L^2([t_0, \infty); \mathcal{W}) \end{bmatrix}$ as $n \rightarrow \infty$.*
- (ii) *If $\begin{bmatrix} x \\ w \end{bmatrix}$ is an externally generated stable trajectory of Σ on the interval $[t_0, \infty)$, and if we define $x(t) = 0$ and $w(t) = 0$ for $t < t_0$, then this extended pair of functions is an externally generated stable two-sided trajectory of Σ .*
- (iii) *Let $\mathcal{W} = -\mathcal{Y} \boxplus \mathcal{U}$ be a fundamental decomposition of $\mathcal{W}$, and let I be a nontrivial closed interval. Then, for each $u \in L^2(I; \mathcal{U})$ there exists a unique externally generated stable two-sided trajectory $\begin{bmatrix} x \\ w \end{bmatrix}$ of Σ on I satisfying $P_{\mathcal{U}}w = u$.*
- (iv) *Let $\begin{bmatrix} x_1 \\ w_1 \end{bmatrix}$ be a stable externally generated trajectory of Σ on the interval $I_1 = (-\infty, t_1]$, and let $\begin{bmatrix} x_2 \\ w_2 \end{bmatrix}$ be a stable trajectory of Σ on an interval I_2 with left end-point t_1 . Then the concatenation $\begin{bmatrix} x \\ w \end{bmatrix}$ defined by (3.3) is a stable trajectory of Σ on $I := I_1 \cup I_2$ if and only if $x_1(t_1) = x_2(t_1)$.*
- (v) *Every stable trajectory on the interval $I = (-\infty, t_1]$ can be extended to a stable two-sided trajectory of Σ . This extension can be chosen so that*

$\pi_{[t_1, \infty)} P_{\mathcal{U}} w = u$ for an arbitrary $u \in L^2([t_1, \infty); \mathcal{U})$, and it is uniquely determined by u .

Proof (i) By Lemma 3.3(i), it suffices to prove the case where $t_0 = 0$. Let $\mathcal{W} = -\mathcal{Y} \boxplus \mathcal{U}$ be a fundamental decomposition of $\mathcal{W}$, and let $\{u_n\}$ be a sequence of $\mathcal{U}$ -valued C^∞ functions with compact support such that $u_n(0) = 0$ for each n and $u_n \rightarrow P_{\mathcal{U}} w$ in $L^2(\mathbb{R}^+, \mathcal{U})$ as $n \rightarrow \infty$. Let $\begin{bmatrix} x_n \\ w_n \end{bmatrix}$ be the stable future trajectory of Σ with $x_n(0) = 0$ given by Lemma 3.4(i). By [19, Theorem 4.6.11], $\begin{bmatrix} x \\ w \end{bmatrix} \in \begin{bmatrix} C^1(\mathbb{R}^+; \mathcal{X}) \\ C(\mathbb{R}^+; \mathcal{W}) \end{bmatrix}$ and $w_n(0) = 0$. By Proposition 3.7(iii), each $\begin{bmatrix} x_n \\ w_n \end{bmatrix}$ is a classical future trajectory of Σ . It follows from Lemma 3.2 that $\begin{bmatrix} x_n \\ w_n \end{bmatrix}$ is stable, and that $\begin{bmatrix} x_n \\ w_n \end{bmatrix} \rightarrow \begin{bmatrix} x \\ w \end{bmatrix}$ in $\begin{bmatrix} BUC(\mathbb{R}^+; \mathcal{X}) \\ L^2(\mathbb{R}^+; \mathcal{W}) \end{bmatrix}$ as $n \rightarrow \infty$.

(ii) Let $\begin{bmatrix} x_n \\ w_n \end{bmatrix}$ be a sequence of classical stable future trajectories of Σ with the properties listed in (i). If we define $\begin{bmatrix} x_n(t) \\ w_n(t) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ for $t < 0$, then each $\begin{bmatrix} x_n \\ w_n \end{bmatrix}$ is a classical stable two-sided trajectory of Σ (note that $\dot{x}_n(0) = 0$ since $\begin{bmatrix} x_n(t) \\ w_n(t) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} \dot{x}(0) \\ x(0) \\ w(0) \end{bmatrix} \in V$.) This extended sequence converges to the extended version of $\begin{bmatrix} x \\ w \end{bmatrix}$ in $\begin{bmatrix} BUC(\mathbb{R}; \mathcal{X}) \\ L^2(\mathbb{R}; \mathcal{W}) \end{bmatrix}$ as $n \rightarrow \infty$, and by Lemma 3.3(iv), the limit is a stable externally generated two-sided trajectory of Σ .

(iii) In the case where I has a finite left end-point the claim (iii) follows from Lemma 3.4(i). Thus, we may here assume that the left end-point of I is $-\infty$. If the right end-point of I is finite, then we start by extending u to all of $\mathbb{R}$ by defining $u(t) = 0$ for $t \notin I$.

For each $n \in I$, define $u_n = P_{[-n, \infty)} u$. Then $u_n \rightarrow u$ in $L^2(\mathbb{R}; \mathcal{U})$ as $n \rightarrow \infty$. Let $\begin{bmatrix} x_n \\ w_n \end{bmatrix}$ be the externally generated trajectory of Σ on $[-n, \infty)$ satisfying $P_{\mathcal{U}} w_n = u_n$ given by Lemma 3.4(i), and use (ii) to extend this trajectory to a two-sided externally generated trajectory, which we still denote by $\begin{bmatrix} x_n \\ w_n \end{bmatrix}$. It follows from Lemma 3.2 that $\begin{bmatrix} x_n \\ w_n \end{bmatrix}$ is a Cauchy sequence in $\begin{bmatrix} BUC(\mathbb{R}; \mathcal{X}) \\ L^2(\mathbb{R}; \mathcal{W}) \end{bmatrix}$ as $n \rightarrow \infty$, and hence it converges to a limit $\begin{bmatrix} x \\ w \end{bmatrix}$ in this space. By Lemma 3.3, this limit is an externally generated two-sided trajectory of Σ . Clearly $P_{\mathcal{U}} w = u$. The uniqueness of this trajectory follows from Lemma 3.2.

(iv) The necessity of the condition $x_1(t_1) = x_2(t)$ for $\begin{bmatrix} x \\ w \end{bmatrix}$ to be a trajectory is obvious, since x is required to be continuous at t_1 .

Let $\mathcal{W} = -\mathcal{Y} \boxplus \mathcal{U}$ be a fundamental decomposition of $\mathcal{W}$, and let $\begin{bmatrix} x' \\ w' \end{bmatrix}$ be the unique externally generated two-sided trajectory of Σ given by (iii) which satisfies $P_{\mathcal{U}} w' = P_{\mathcal{U}} w$. Then the restriction of $\begin{bmatrix} x' \\ w' \end{bmatrix}$ to I_1 is an externally generated trajectory of Σ on I_1 , and by (iii), this restriction is equal to $\begin{bmatrix} x_1 \\ w_1 \end{bmatrix}$. On the other hand, the restriction of $\begin{bmatrix} x' \\ w' \end{bmatrix}$ to I_2 is a trajectory of Σ on I_2 , and by Lemma 3.4(i), this restriction is equal to $\begin{bmatrix} x_2 \\ w_2 \end{bmatrix}$. Thus, $\begin{bmatrix} x \\ w \end{bmatrix} = \begin{bmatrix} x' \\ w' \end{bmatrix}$, and so $\begin{bmatrix} x \\ w \end{bmatrix}$ is an externally generated trajectory of Σ on I .

(v) That (v) is true follows from (iv) and Lemma 3.4(i). $\square$

3.4 Passive Past, Two-sided, and Future Behaviors

We recall the following definition from Sect. 1.

Definition 3.10 By the (stable) *behavior* $\mathfrak{W}^\Sigma(I)$ of the passive s/s system Σ on the closed and nontrivial interval I we mean the set of all the signal parts w of all externally generated stable trajectories $\begin{bmatrix} x \\ w \end{bmatrix}$ of Σ on I .

In the special cases $I = \mathbb{R}^-$, $I = \mathbb{R}$, and $I = \mathbb{R}^+$ we denote these behaviors by $\mathfrak{W}_-^\Sigma$, $\mathfrak{W}^\Sigma$, and $\mathfrak{W}_+^\Sigma$, and refer to them as the *past, two-sided, and future behaviors* of Σ , respectively.²

Lemma 3.11 *To each $w \in \mathfrak{W}_+^\Sigma$ there exists a unique $x \in C(\mathbb{R}^+; \mathcal{X})$ such that $\begin{bmatrix} x \\ w \end{bmatrix}$ is an externally generated stable trajectory of Σ on $\mathbb{R}^+$, and this function satisfies $x \in BUC(\mathbb{R}^+; \mathcal{X})$. The same statement remains true if we replace $\mathfrak{W}_+^\Sigma$ by $\mathfrak{W}^\Sigma$ or by $\mathfrak{W}_-^\Sigma$ and at the same time replace $\mathbb{R}^+$ by $\mathbb{R}$ or $\mathbb{R}^-$, respectively.*

Proof This follows from the definitions of $\mathfrak{W}_+^\Sigma$, $\mathfrak{W}^\Sigma$, and $\mathfrak{W}_-^\Sigma$ and Lemmas 3.4(i) and 3.9(iii). $\square$

The shift (semi)groups τ^t , τ_-^t , and τ_+^t used in the following lemma were defined at the end of Sect. 1.

Lemma 3.12 *The past, two-sided, and future behaviors $\mathfrak{W}_-^\Sigma$, $\mathfrak{W}^\Sigma$, and $\mathfrak{W}_+^\Sigma$ of a passive s/s system $\Sigma = (V; \mathcal{X}, \mathcal{W})$ have the following properties:*

- (i) $\mathfrak{W}_\pm$ are right-shift invariant and $\mathfrak{W}^\Sigma$ is bilaterally shift-invariant, i.e.,

$$\begin{aligned} \tau_\pm^{*t} \mathfrak{W}_\pm^\Sigma &\subset \mathfrak{W}_\pm^\Sigma, & t \in \mathbb{R}^+, \\ \tau^t \mathfrak{W}^\Sigma &= \mathfrak{W}^\Sigma, & t \in \mathbb{R}. \end{aligned} \quad (3.12)$$

- (ii) $\mathfrak{W}_\pm^\Sigma$ can be recovered from $\mathfrak{W}^\Sigma$ by the formulas

$$\begin{aligned} \mathfrak{W}_-^\Sigma &= \pi_- \mathfrak{W}^\Sigma := \{w_- \in K_-^2(\mathcal{W}) \mid w_- = \pi_- w \text{ for some } w \in \mathfrak{W}^\Sigma\}, \\ \mathfrak{W}_+^\Sigma &= \mathfrak{W}^\Sigma \cap K_+^2(\mathcal{W}) := \{w \in \mathfrak{W}^\Sigma \mid w(t) = 0 \text{ for } t < 0\}. \end{aligned} \quad (3.13)$$

- (iii) $\mathfrak{W}_\pm^\Sigma$ is a maximal nonnegative subspace of $K_\pm^2(\mathcal{W})$ and $\mathfrak{W}^\Sigma$ is a maximal nonnegative subspace of $K^2(\mathcal{W})$.

Proof (i) By Lemma 3.3(i), $\tau^t \mathfrak{W}^\Sigma = \mathfrak{W}^\Sigma$ for all $t \in \mathbb{R}$. By Lemma 3.3(i)–(ii), $\tau_-^{*t} \mathfrak{W}_-^\Sigma \subset \mathfrak{W}_-^\Sigma$ (and actually even $\tau_-^{*t} \mathfrak{W}_-^\Sigma = \mathfrak{W}_-^\Sigma$) for all $t \in \mathbb{R}^+$. That $\tau_+^{*t} \mathfrak{W}_+^\Sigma \subset \mathfrak{W}_+^\Sigma$ for all $t \in \mathbb{R}^+$ follows from Lemmas 3.3(i)–(ii) and 3.9(ii).

(ii) If $w \in \mathfrak{W}_-^\Sigma$, then by Lemma 3.3(ii), $w_- := \pi_- w \in \mathfrak{W}_-^\Sigma$. Conversely, according to Lemma 3.9, every $w_- \in \mathfrak{W}_-^\Sigma$ can be extended to a function $w \in \mathfrak{W}^\Sigma$. Analogously, by Lemma 3.3(ii), if $w \in \mathfrak{W}^\Sigma \cap K_+^2(\mathcal{W})$ then $w \in \mathfrak{W}_+^\Sigma$, and if $w \in \mathfrak{W}_+^\Sigma$ and we

² Instead of “two-sided behavior” the name “full behavior” was used in the discrete time settings of [5] and [6].

extend w to $K^2(\mathcal{W})$ by defining $w(t) = 0$ for $t < 0$, then by Lemma 3.9(ii), the extended function belongs to $\mathfrak{W}^\Sigma$.

(iii) That $\mathfrak{W}^\Sigma$, $\mathfrak{W}^\Sigma$, and $\mathfrak{W}_+^\Sigma$ are nonnegative follows from Lemma 3.2. To see that they are *maximal* nonnegative it suffices to take an arbitrary fundamental decomposition $\mathcal{W} = -\mathcal{Y} \boxplus \mathcal{U}$ of $\mathcal{W}$ and use Lemmas 3.4(i) and 3.9(iii) and Proposition 2.1(i). $\square$

At this point we must warn the reader that if $\mathfrak{W}$ is an arbitrary maximal nonnegative bilaterally shift-invariant subspace of $K^2(\mathcal{W})$ and if we define $\mathfrak{W}_- = \pi_- \mathfrak{W}$ and $\mathfrak{W}_+ = \mathfrak{W} \cap K_+^2(\mathcal{W})$ (as in (3.12)) then it need *not be true* that $\mathfrak{W}_-$ is maximal nonnegative in $K_-^2(\mathcal{W})$ or that $\mathfrak{W}_+$ is maximal nonnegative in $K_+^2(\mathcal{W})$. A discrete time counter example is given in [5, Examples 2.7 and 2.14], and the same example can easily be modified to become a continuous time counter example. The following lemma clarifies the situation.

Lemma 3.13 *Let $\mathfrak{W}$ be a maximal nonnegative subspace $\mathfrak{W}$ of $K^2(\mathcal{W})$, and define $\mathfrak{W}_-$ and $\mathfrak{W}_+$ by*

$$\mathfrak{W}_- := \pi_- \mathfrak{W}, \quad \mathfrak{W}_+ := \mathfrak{W} \cap K_+^2(\mathcal{W}), \quad (3.14)$$

Then the following conditions are equivalent:

- (i) $\mathfrak{W}_-$ is a maximal nonnegative subspace of $K_-^2(\mathcal{W})$.
- (ii) $\mathfrak{W}_+$ is a maximal nonnegative subspace of $K_+^2(\mathcal{W})$.
- (iii) For some fundamental decomposition $\mathcal{W} = -\mathcal{Y} \boxplus \mathcal{U}$ the following implication is valid: If $w \in \mathfrak{W}$ and $\pi_- \pi_{\mathcal{U}} w = 0$, then $\pi_- \pi_{\mathcal{Y}} w = 0$.
- (iv) For every fundamental decomposition $\mathcal{W} = -\mathcal{Y} \boxplus \mathcal{U}$ the following implication is valid: If $w \in \mathfrak{W}$ and $\pi_- \pi_{\mathcal{U}} w = 0$, then $\pi_- \pi_{\mathcal{Y}} w = 0$.

Proof This follows from Lemma 2.4 with the substitutions $\mathfrak{K} \rightarrow K^2(\mathcal{W})$, $\mathfrak{K}_1 \rightarrow K_+^2(\mathcal{W})$, $\mathfrak{K}_2 \rightarrow K_-^2(\mathcal{W})$, $\mathcal{Z} \rightarrow \mathfrak{W}$, $\mathcal{Y}_2 \rightarrow L_-^2(\mathcal{Y})$, and $\mathcal{U}_2 \rightarrow L_-^2(\mathcal{U})$. Note that conditions (i) and (ii) in Lemma 2.4 do not depend on the particular fundamental decomposition used in part (iii) of that lemma, so if (iii) holds for one fundamental decomposition, then it holds for every fundamental decomposition. $\square$

Motivated by Lemmas 3.12 and 3.13 we make the following definition:

Definition 3.14 Let $\mathcal{W}$ be a Kreĭn space.

- (i) A maximal nonnegative right-shift invariant subspace of $K_-^2(\mathcal{W})$ is called a *passive past behavior* on the (signal) space $\mathcal{W}$.
- (ii) A maximal nonnegative right-shift invariant subspace $\mathfrak{W}_+$ of $K_+^2(\mathcal{W})$ is called a *passive future behavior* on the (signal) space $\mathcal{W}$.
- (iii) A maximal nonnegative bilaterally shift invariant subspace $\mathfrak{W}$ of $K^2(\mathcal{W})$ which satisfies the equivalent conditions (i)–(iv) listed in Lemma 3.13 is called a *passive two-sided behavior* on the Kreĭn (signal) space $\mathcal{W}$.

Proposition 3.15 *Let $\Sigma = (V; \mathcal{X}, \mathcal{W})$ be a passive s/s system. Then the past, two-sided, and future behaviors of Σ are passive past, two-sided, and future behaviors, respectively, on $\mathcal{W}$ in the sense of Definition 3.14.*

Proof This follows from Lemma 3.12 and Definition 3.14. $\square$

As we shall see in Lemma 3.18 below, each one of these behaviors determine the two others uniquely.

Passive past and future behaviors actually have slightly stronger shift-invariance properties than what is explicitly required in Definition 3.14.

Lemma 3.16 *Let $\mathcal{W}$ be a Kreĭn space.*

- (i) *Every passive past behavior $\mathfrak{W}_-$ on $\mathcal{W}$ satisfies $\tau_-^{*t}\mathfrak{W}_- = \mathfrak{W}_-$ for all $t \in \mathbb{R}^+$.*
- (ii) *Every passive future behavior $\mathfrak{W}_+$ on $\mathcal{W}$ satisfies*

$$\tau_+^{*t}\mathfrak{W}_+ = \{w \in \mathfrak{W}_+ \mid w(s) = 0 \text{ for almost all } s \in [0, t]\}$$

for all $t \in \mathbb{R}^+$.

Before proving this lemma we make the following remark.

Remark 3.17 Many of our subsequent results (as well as Lemmas 3.13 and 3.18 above) can be regarded as continuous time versions of the corresponding discrete time results given in [5,6]. In many cases the proofs given in [5,6] can be adapted to the present setting by performing some simple substitutions. As a general rule, all those notions defined [5,6] have a natural counterpart presented here, they should be replaced by that counterpart. The discrete time right shifts S_- , S , and S_+ are replaced by τ_-^{*t} , τ^{-t} , and τ_+^{*t} , $t \in \mathbb{R}^+$, and the discrete time left shifts S_-^* , S^{-1} , and S_+^* are replaced by τ_-^t , τ^t , and τ_+^t , $t \in \mathbb{R}^+$. The discrete time trajectories in [5,6] are throughout replaced by *generalized continuous time trajectories* (i.e., no *classical* trajectories enter in these translations). The main difference between the discrete time and the continuous time cases is that in the proofs one should *not* replace the discrete time generating subspace V by the continuous time generating subspace V . Instead, in computations involving *future trajectories* one should through *replace V by the subspaces $\mathcal{T}_{0,t}$ defined in (3.5)*, and the discrete time node space $\mathfrak{K}$ *should be replaced by the Kreĭn space $\mathfrak{K}_{0,t}$ with the inner product (1.12)*, and we throughout *use Theorem 3.5 to characterize the passivity of a continuous time s/s system*, and not the original Definition 3.1. In connection with *past trajectories* we replace the discrete time generating subspace V by a left-shifted versions $\mathcal{T}_{[-t,0]}$ and $\mathfrak{K}_{[-t,0]}$ of $\mathcal{T}_{0,t}$ and $\mathfrak{K}_{0,t}$. This has the consequence that whenever a discrete time formula contains the term $[w(0), w(0)]_{\mathcal{W}}$ it should be replaced by $\int_0^t [w(s), w(s)]_{\mathcal{W}} ds$, and analogously $[w(-1), w(-1)]_{\mathcal{W}}$ should be replaced by $\int_{-t}^0 [w(s), w(s)]_{\mathcal{W}} ds$. As a consequence of these changes, the continuous time proofs are often slightly shorter than the discrete time proofs, since there is no need to build trajectories on a finite time interval $[0, T]$ from scratch, as sometimes happens in the discrete time setting.

If the proof of some particular result given below can be obtained from the corresponding result in [5] or [6] by performing the substitutions listed above, then we sometimes omit the proof, and refer the reader to [5] or [6]. We do this, in particular, if the conversion is straightforward and the proof is of significant length. If the proof is short, or if the conversion is less straightforward, or if the proof is important for the general understanding of the theory we write it out in full detail.

As an example on how to convert discrete time results to continuous time results, let us look at the graph representation of passive behaviors used in the proof of [5, Theorem 2.11]. (These graph representations are needed, among others, for the proof of Lemma 3.16.) Let $\mathcal{W} = -\mathcal{Y} \boxplus \mathcal{U}$ be a fundamental decomposition of $\mathcal{W}$. Then $K^2(\mathcal{W}) = -L^2(\mathcal{Y}) \boxplus L^2(\mathcal{U})$ and $K_{\pm}^2(\mathcal{W}) = -L_{\pm}^2(\mathcal{Y}) \boxplus L_{\pm}^2(\mathcal{U})$ are fundamental decompositions of the Kreĭn spaces $K^2(\mathcal{W})$ and $K_{\pm}^2(\mathcal{W})$, respectively. By assertion (i) and (iv) of Proposition 2.1, every passive past, two-sided, and future behavior $\mathfrak{W}_-$, $\mathfrak{W}$, and $\mathfrak{W}_+$ on $\mathcal{W}$ and their orthogonal companions have graph representations with respect to the above fundamental decompositions of the type

$$\begin{aligned} \mathfrak{W}_{\pm} &= \left\{ \begin{bmatrix} \mathfrak{D}_{\pm} u \\ u \end{bmatrix} \mid u \in L_{\pm}^2(\mathcal{U}) \right\}, & \mathfrak{W} &= \left\{ \begin{bmatrix} \mathfrak{D} u \\ u \end{bmatrix} \mid u \in L^2(\mathcal{U}) \right\}, \\ \mathfrak{W}_{\pm}^{[\perp]} &= \left\{ \begin{bmatrix} y \\ \mathfrak{D}_{\pm}^* y \end{bmatrix} \mid y \in L_{\pm}^2(\mathcal{Y}) \right\}, & \mathfrak{W}^{[\perp]} &= \left\{ \begin{bmatrix} y \\ \mathfrak{D}^* y \end{bmatrix} \mid y \in L^2(\mathcal{Y}) \right\}, \end{aligned} \quad (3.15)$$

where $\mathfrak{D}_{\pm}$ and $\mathfrak{D}$ are linear contractions between the respective L^2 -spaces. It follows from Lemma 3.13 and Definition 3.14 that $\pi_- \mathfrak{D}|_{L_{+}^2(\mathcal{U})} = 0$ and $\pi_+ \mathfrak{D}^*|_{L_{-}^2(\mathcal{U})} = 0$, i.e., $\mathfrak{D}$ and $\mathfrak{D}^*$ are *causal* and *anti-causal*, respectively. Since $\tau_{\pm}^{*t} \mathfrak{W}_{\pm} \subset \mathfrak{W}_{\pm}$ for all $t \in \mathbb{R}^+$ and $\tau^t \mathfrak{W} = \mathfrak{W}$ for all $t \in \mathbb{R}$, it follows from (3.15) that $\mathfrak{D}_{\pm}$ are right-shift invariant and $\mathfrak{D}$ is bilaterally shift invariant, i.e.,

$$\tau_{\pm}^{*t} \mathfrak{D}_{\pm} = \mathfrak{D}_{\pm} \tau_{\pm}^{*t} \text{ for all } t \in \mathbb{R}^+ \text{ and } \tau^t \mathfrak{D} = \mathfrak{D} \tau^t \text{ for all } t \in \mathbb{R}. \quad (3.16)$$

Furthermore, if the three behaviors $\mathfrak{W}_{\pm}$ and $\mathfrak{W}$ are related to each other by the relations (3.14)–(3.19), then $\mathfrak{D}_{\pm}$ and $\mathfrak{D}$ are related to each other by

$$\begin{aligned} \mathfrak{D}_+ &= \mathfrak{D}|_{L_{+}^2(\mathcal{U})}, & \mathfrak{D}_- &= \pi_- \mathfrak{D}|_{L_{-}^2(\mathcal{U})}, \\ \mathfrak{D}_+^* &= \pi_+ \mathfrak{D}^*|_{L_{+}^2(\mathcal{U})}, & \mathfrak{D}_-^* &= \mathfrak{D}^*|_{L_{-}^2(\mathcal{U})}. \end{aligned} \quad (3.17)$$

Proof of Lemma 3.16. Let $\mathcal{W} = -\mathcal{Y} \boxplus \mathcal{U}$ be a fundamental decomposition of $\mathcal{W}$.

(i) By (3.15), the right-shift invariance of $\mathfrak{D}_-$, and the fact that $\tau_-^{*t} L_{-}^2(\mathcal{U}) = L_{-}^2(\mathcal{U})$, we have

$$\begin{aligned} \tau_-^{*t} \mathfrak{W}_- &= \left\{ \begin{bmatrix} \tau_-^{*t} \mathfrak{D}_- u \\ \tau_-^{*t} u \end{bmatrix} \mid u \in L_{-}^2(\mathcal{U}) \right\} = \left\{ \begin{bmatrix} \mathfrak{D}_- \tau_-^{*t} u \\ \tau_-^{*t} u \end{bmatrix} \mid u \in L_{-}^2(\mathcal{U}) \right\} \\ &= \left\{ \begin{bmatrix} \mathfrak{D}_- u \\ u \end{bmatrix} \mid u \in L_{-}^2(\mathcal{U}) \right\} = \mathfrak{W}_-. \end{aligned}$$

(ii) If $u \in \tau_+^{*t} \mathfrak{W}_+$, then by the shift-invariance of $\mathfrak{W}_+$, $\tau_+^{*t} w \in \mathfrak{W}_+$, and of course, w vanishes on $[0, t]$. Conversely, let $w \in \mathfrak{W}_+$ vanish on $[0, t]$. Define $u_1 = \tau_+^t P_{\mathcal{U}} w$, and $w_1 = \begin{bmatrix} \mathfrak{D}_+ u_1 \\ u_1 \end{bmatrix}$. Then

$$\tau_+^{*t} w_1 = \begin{bmatrix} \tau_+^{*t} \mathfrak{D}_+ u_1 \\ \tau_+^{*t} u_1 \end{bmatrix} = \begin{bmatrix} \mathfrak{D}_+ \tau_+^{*t} u_1 \\ \tau_+^{*t} u_1 \end{bmatrix} = \begin{bmatrix} \mathfrak{D}_+ P_{\mathcal{U}} w \\ P_{\mathcal{U}} w \end{bmatrix} = w.$$

Consequently, $w \in \tau_+^{*t} \mathfrak{W}_+$. □

The following lemma complements Lemmas 3.12 and 3.13.

Lemma 3.18 *Let $\mathcal{W}$ be a Kreĭn space.*

(i) *If $\mathfrak{W}_-$ is a passive past behavior on $\mathcal{W}$, and if we define $\mathfrak{W}$ by*

$$\mathfrak{W} = \bigcap_{t \in \mathbb{R}^+} \{w \in K^2(\mathcal{W}) \mid \pi_- \tau^t w \in \mathfrak{W}_-\}, \quad (3.18)$$

then $\mathfrak{W}$ is a passive two-sided behavior on $\mathcal{W}$ and $\mathfrak{W}_- = \pi_- \mathfrak{W}$.

(ii) *If $\mathfrak{W}_+$ is a passive future behavior on $\mathcal{W}$, and if we define $\mathfrak{W}$ by*

$$\mathfrak{W} = \bigvee_{t \in \mathbb{R}^+} \tau^t \mathfrak{W}_+, \quad (3.19)$$

then $\mathfrak{W}$ is a passive two-sided behavior on $\mathcal{W}$, and $\mathfrak{W}_+ = \mathfrak{W} \cap K_+^2(\mathcal{W})$.

(iii) *Let $\mathfrak{W}$ be a passive two-sided behavior on the Kreĭn signal space $\mathcal{W}$, and define $\mathfrak{W}_-$ and $\mathfrak{W}_+$ by (3.14). Then $\mathfrak{W}_-$ is a passive past behavior on $\mathcal{W}$, $\mathfrak{W}_+$ is a passive future behavior on $\mathcal{W}$, and $\mathfrak{W}$ can be recovered from $\mathfrak{W}_+$ and from $\mathfrak{W}_-$ by means of formulas (3.18) and (3.19).*

Proof (i) Let $\mathfrak{W}_-$ be a passive past behavior on $\mathcal{W}$, and define $\mathfrak{W}$ by (3.19). Denote

$$\mathfrak{W}_-^t = \{w \in K^2(\mathcal{W}) \mid \pi_- \tau^t w \in \mathfrak{W}_-\}, \quad t \in \mathbb{R},$$

so that $\mathfrak{W} = \bigcap_{t \in \mathbb{R}^+} \mathfrak{W}_-^t$. Then $\mathfrak{W}_-^t = \tau^{-t} \mathfrak{W}_-^0$ since

$$w \in \mathfrak{W}_-^t \Leftrightarrow \pi_- \tau^t w \in \mathfrak{W}_- \Leftrightarrow \tau^t w \in \mathfrak{W}_-^0 \Leftrightarrow w \in \tau^{-t} \mathfrak{W}_-^0.$$

Thus $\tau^s \mathfrak{W}_-^t = \tau^{s-t} \mathfrak{W}_-^0 = \mathfrak{W}_-^{t-s}$ for all $s, t \in \mathbb{R}$. If $w \in \mathfrak{W}_-^t$, or equivalently, $\pi_- \tau^t w \in \mathfrak{W}_-$, then the right-shift invariance of $\mathfrak{W}_-$ implies that for all $s \in \mathbb{R}^+$,

$$\pi_- \tau^{t-s} w = \tau_-^{*s} \pi_- \tau^t w \in \mathfrak{W}_-.$$

and consequently $\mathfrak{W}_-^t \subset \mathfrak{W}_-^{t-s}$ for all $t \in \mathbb{R}$ and $s \in \mathbb{R}^+$. In particular, this implies that for all $t \in \mathbb{R}$,

$$\tau^t \mathfrak{W} = \tau^t \bigcap_{s \in \mathbb{R}^+} \mathfrak{W}_-^s = \bigcap_{s \in \mathbb{R}^+} \tau^t \mathfrak{W}_-^s = \bigcap_{s \in \mathbb{R}^+} \mathfrak{W}_-^{s-t} = \bigcap_{r \geq -t} \mathfrak{W}_-^r = \bigcap_{r \in \mathbb{R}^+} \mathfrak{W}_-^r = \mathfrak{W}.$$

Thus, $\mathfrak{W}$ is bilaterally shift-invariant.

We next show that $\mathfrak{W}$ is nonnegative in $K^2(\mathcal{W})$. If $w \in \mathfrak{W}_-^t$ for some $t \in \mathbb{R}$, then it follows from the definition of $\mathfrak{W}_-^t$ and the maximal nonnegativity of $\mathfrak{W}_-$ that

$$0 \leq \int_{\mathbb{R}^-} [(\tau^t w)(s), (\tau^t w)(s)]_{\mathcal{W}} ds = \int_{-\infty}^t [w(s), w(s)]_{\mathcal{W}} ds.$$

If $w \in \mathfrak{W} = \bigcap_{t \in \mathbb{R}^+} \mathfrak{W}_-^t$, then we can let $t \rightarrow \infty$ to get $[w, w]_{K^2(\mathcal{W})} \geq 0$. Thus, $\mathfrak{W}$ is a nonnegative subspace of $K^2(\mathcal{W})$.

To prove that $\mathfrak{W}$ is maximal nonnegative in $K^2(\mathcal{W})$ we let $K^2(\mathcal{W}) = -L^2(\mathcal{Y}) \boxplus L^2(\mathcal{U})$ be a fundamental decomposition of $K^2(\mathcal{W})$, where $\mathcal{W} = -\mathcal{Y} \boxplus \mathcal{U}$ is a fundamental decomposition of $\mathcal{W}$. Let u be an arbitrary function in $L^2(\mathcal{U})$. By the definition of $\mathfrak{W}_-^t$, the maximal nonnegativity of $\mathfrak{W}_-$, and Proposition 2.1(i), for each $n \in \mathbb{Z}^+$ there exists some $w_n \in \mathfrak{W}_-^n$ such that $\pi_- P_{\mathcal{U}} \tau^n w_n = \pi_- u$, or equivalently, $\pi_{(-\infty, n]} P_{\mathcal{U}} w_n = \pi_{(-\infty, n]} u$. Moreover, $\pi_{(-\infty, n]} P_{\mathcal{Y}} w_n$ is uniquely determined by $\pi_{(-\infty, n]} u$. Since $\mathfrak{W}_-^m \subset \mathfrak{W}_-^n$ for all $m, n \in \mathbb{Z}^+, n \geq m$, this implies that $\pi_{(-\infty, m]} w_n = \pi_{(-\infty, m]} w_m$ for all $n \geq m$. If we use the Hilbert space norm in $\mathcal{W}$ induced by the decomposition $\mathcal{W} = -\mathcal{Y} \boxplus \mathcal{U}$, then

$$\|\pi_{(-\infty, n]} w_n\|_{L^2(-\infty, n]; \mathcal{W})} \leq 2\|\pi_{(-\infty, n]} u\|_{L^2(-\infty, n]; \mathcal{U})} \leq 2\|u\|_{L^2(\mathcal{U})}.$$

Define $w(t) = w_0(t)$ for $t \leq 0$, and $w(t) = w_n(t)$ for $t \in (n-1, n], n \geq 1$. Then $\pi_{(-\infty, n]} \tau^n w = \pi_{(-\infty, n]} \tau^n w_n \in \mathfrak{W}_-$, and consequently $w \in \bigcap_{n \in \mathbb{Z}^+} \mathfrak{W}_-^n = \mathfrak{W}$. By Proposition 2.1(i), $\mathfrak{W}$ is maximal nonnegative.

Trivially, $\pi_- \mathfrak{W} \subset \pi_- \mathfrak{W}_-^0 = \mathfrak{W}_-$. Conversely, take some arbitrary $w_- \in \mathfrak{W}_-$. Let $\mathcal{W} = -\mathcal{Y} \boxplus \mathcal{U}$ be a fundamental decomposition of $\mathcal{W}$, and define $u(t) = P_{\mathcal{U}} w_-(t)$ for $t \in \mathbb{R}^-$ and $u(t) = 0$ for $t > 0$. Let w be the corresponding function in $\mathcal{W}$ constructed in the preceding paragraph. Then $\pi_- w \in \mathfrak{W}_-$ and $P_{\mathcal{U}} \pi_- w = P_{\mathcal{U}} w_-$. Consequently, since every function in $\mathfrak{W}_-$ is uniquely determined by its $\mathcal{U}$ -component, we have $\pi_- w = w_-$. Thus, $\mathfrak{W}_- \subset \pi_- \mathfrak{W}$. Together with the inclusion $\pi_- \mathfrak{W} \subset \mathfrak{W}_-$ this gives $\pi_- \mathfrak{W} = \mathfrak{W}_-$. By Definition 3.14, $\mathfrak{W}$ is a passive two-sided behavior.

(ii) Let $\mathfrak{W}_+$ be a passive future behavior on $\mathcal{W}$, and define $\mathfrak{W}$ by (3.19). Denote

$$\mathfrak{W}_+^t = \tau^t \mathfrak{W}_+, \quad t \in \mathbb{R},$$

so that $\mathfrak{W} = \bigvee_{t \in \mathbb{R}^+} \mathfrak{W}_+^t$. Trivially $\tau^s \mathfrak{W}_+^t = \tau^{s+t} \mathfrak{W}_+ = \mathfrak{W}_+^{s+t}$ for all $s, t \in \mathbb{R}$. The right-shift invariance of $\mathfrak{W}_+$ implies that $\mathfrak{W}_+^s \subset \mathfrak{W}_+^t$ for all $s \leq t$. In particular, for all $t \in \mathbb{R}$,

$$\tau^t \mathfrak{W} = \tau^t \bigvee_{s \in \mathbb{R}^+} \mathfrak{W}_+^s = \bigvee_{s \in \mathbb{R}^+} \tau^t \mathfrak{W}_+^s = \bigvee_{s \in \mathbb{R}^+} \mathfrak{W}_+^{s+t} = \bigvee_{r \geq t} \mathfrak{W}_+^r = \bigvee_{r \in \mathbb{R}^+} \mathfrak{W}_+^r = \mathfrak{W}.$$

Thus, $\mathfrak{W}$ is bilaterally shift-invariant.

We next show that $\mathfrak{W}$ is nonnegative in $K^2(\mathcal{W})$. If $w \in \mathfrak{W}_+^t$ for some $t \in \mathbb{R}$, then it follows from the definition of $\mathfrak{W}_+^t$ and the maximal nonnegativity of $\mathfrak{W}_+$ that

$$0 \leq \int_{\mathbb{R}^+} [(\tau^{-t} w)(s), (\tau^{-t} w)(s)]_{\mathcal{W}} ds = \int_{\mathbb{R}} [w(s), w(s)]_{\mathcal{W}} ds.$$

Thus, each of the subspaces $\mathfrak{W}_+^t$ is nonnegative, and hence so is the closed linear hull $\mathfrak{W} = \bigvee_{t \in \mathbb{R}^+} \mathfrak{W}_+^t$.

To prove that $\mathfrak{W}$ is maximal nonnegative in $K^2(\mathcal{W})$ we let $K^2(\mathcal{W}) = -L^2(\mathcal{Y}) \boxplus L^2(\mathcal{U})$ be a fundamental decomposition of $K^2(\mathcal{W})$, where $\mathcal{W} = -\mathcal{Y} \boxplus \mathcal{U}$ is a fundamental decomposition of $\mathcal{W}$. Let u be an arbitrary function in $L^2(\mathcal{U})$. By the definition of $\mathfrak{W}_+$, the maximal nonnegativity of $\mathfrak{W}_+$, and Proposition 2.1(i), for each $n \in \mathbb{Z}^+$ there exists some $w_n \in \mathfrak{W}_+$ such that $P_{\mathcal{U}}w_n = \pi_{[-n, \infty)}u$. If we use the Hilbert space norm in $\mathcal{W}$ induced by the decomposition $\mathcal{W} = -\mathcal{Y} \boxplus \mathcal{U}$, then for all $m, n \in \mathbb{Z}^+$, $m \geq n$,

$$\|w_m - w_n\|_{L^2(\mathbb{R}; \mathcal{W})} \leq 2\|u\|_{L^2([-m, -n]; \mathcal{U})}.$$

Thus, w_n is a Cauchy sequence in $L^2(\mathcal{W})$ which converges to a limit w in $L^2(\mathcal{W})$. Since each $w_n \in \mathfrak{W}_+$, we have $w \in \bigvee_{t \in \mathbb{R}^+} \mathfrak{W}_+^t = \mathfrak{W}$. Thus, $P_{\mathcal{U}}\mathfrak{W} = L^2(\mathcal{U})$, and by Proposition 2.1(i), $\mathfrak{W}$ is maximal nonnegative.

By Lemma 3.16, for each $t \in \mathbb{R}^+$ we have $\mathfrak{W}_+ = \mathfrak{W}_+^t \cap K_+^2(\mathcal{W})$. Thus $\mathfrak{W}_+ \subset (\bigvee_{t \in \mathbb{R}^+} \mathfrak{W}_+^t) \cap K_+^2(\mathcal{W}) = \mathfrak{W} \cap K_+^2(\mathcal{W})$. On the other hand, $\mathfrak{W} \cap K_+^2(\mathcal{W})$ is a nonnegative subspace of $K_+^2(\mathcal{W})$ whereas $\mathfrak{W}_+$ is a *maximal* nonnegative subspace of $K_+^2(\mathcal{W})$ contained in $K_+^2(\mathcal{W}) \cap \mathfrak{W}$. Thus, $\mathfrak{W}_+ = K_+^2(\mathcal{W}) \cap \mathfrak{W}$. By Definition 3.14, $\mathfrak{W}$ is a passive two-sided behavior.

(iii) Let $\mathfrak{W}$ be a passive two-sided behavior on $\mathcal{W}$, and define $\mathfrak{W}_-$ and $\mathfrak{W}_+$ by (3.14). It follows from Definition 3.14 that $\mathfrak{W}_+$ is a maximal nonnegative subspace of $K_+^2(\mathcal{W})$ and that $\mathfrak{W}_-$ is a maximal nonnegative subspace of $K_-^2(\mathcal{W})$. The right-shift invariance of $\mathfrak{W}_{\pm}$ follows from the bilateral shift invariance of $\mathfrak{W}$ and (3.14). Thus, $\mathfrak{W}_+$ and $\mathfrak{W}_-$ are passive future and past behaviors, respectively. This proves the first two claims in (i).

We continue with the proof of (3.18). Denote the right-hand side of (3.18) by $\tilde{\mathfrak{W}}$. By (i), $\tilde{\mathfrak{W}}$ is a (maximal) nonnegative subspace of $K^2(\mathcal{W})$, and it follows from Definition 3.14 that $\mathfrak{W} \subset \tilde{\mathfrak{W}}$. Since $\mathfrak{W}$ is maximal nonnegative, we must have $\mathfrak{W} = \tilde{\mathfrak{W}}$, and consequently (3.18) holds.

We finally prove (3.19). Denote the right-hand side of (3.19) by $\tilde{\mathfrak{W}}$. By (ii), $\tilde{\mathfrak{W}}$ is a maximal nonnegative subspace of $K^2(\mathcal{W})$, and it follows from Definition 3.14 that $\tilde{\mathfrak{W}} \subset \mathfrak{W}$. Since $\mathfrak{W}$ is nonnegative, we must have $\mathfrak{W} = \tilde{\mathfrak{W}}$, and consequently (3.19) holds. $\square$

Lemma 3.19 *Let $\mathfrak{W}_-$ be a passive past behavior on a Kreĭn space $\mathcal{W}$. Then the set of all $w \in \mathfrak{W}_-$ with compact support is a dense subspace of $\mathfrak{W}_-$.*

Proof By Lemma 3.18(iii),

$$\mathfrak{W}_- = \pi_- \mathfrak{W} = \pi_- \bigvee_{t \in \mathbb{R}^+} \tau^t \mathfrak{W}_+ = \bigvee_{t \in \mathbb{R}^+} \pi_- \tau^t \mathfrak{W}_+,$$

where each function in $\pi_- \tau^t \mathfrak{W}_+$ has compact support. $\square$

Lemma 3.20 *Let $\mathfrak{W}_+$ be a passive future behavior on $\mathcal{W}$, and define the $[0, t]$ -sections $\mathfrak{W}_{[0, t]}$ of $\mathfrak{W}_+$ by*

$$\mathfrak{W}_{[0,t]} := \pi_{[0,t]} \mathfrak{W}_+, \quad t \in \mathbb{R}^+. \quad (3.20)$$

Then each $\mathfrak{W}_{[0,t]}$ is a maximal nonnegative subspace of $K^2([0, t]; \mathcal{W})$.

Proof By Lemma 3.16, $\mathfrak{W}_+ \cap K^2([t, \infty); \mathcal{W}) = \tau_+^{*t} \mathfrak{W}_+$, and therefore $\mathfrak{W}_+ \cap K^2([t, \infty); \mathcal{W})$ is maximal nonnegative in $K^2([t, \infty); \mathcal{W})$. This fact, combined with Lemma 2.4 with the substitutions $\mathfrak{K} \rightarrow K_+^2(\mathcal{W})$, $\mathfrak{K}_1 \rightarrow K^2([t, \infty); \mathcal{W})$, $\mathfrak{K}_2 \rightarrow K^2([0, t]; \mathcal{W})$, and $\mathcal{Z} \rightarrow \mathfrak{W}_+$, implies that $\pi_{[0,t]} \mathfrak{W}_+$ is maximal nonnegative in $K^2([0, t]; \mathcal{W})$. $\square$

3.5 Intertwined Systems

Definition 3.21 Let $\Sigma_1 = (V_1; \mathcal{X}_1; \mathcal{W})$ and $\Sigma_2 = (V_2; \mathcal{X}_2; \mathcal{W})$ be two passive s/s systems (with the same signal space $\mathcal{W}$).

- (i) A bounded linear operator $E: \mathcal{X}_1 \rightarrow \mathcal{X}_2$ *intertwines* the two passive s/s systems Σ_1 and Σ_2 if the formula

$$(x_1, w) \mapsto (Ex_1, w) \quad (3.21)$$

defines a map from the set of all stable future trajectories $\begin{bmatrix} x_1 \\ w \end{bmatrix}$ of Σ_1 onto the set of all stable future trajectories $\begin{bmatrix} x_2 \\ w \end{bmatrix}$ of Σ_2 satisfying $x_2(0) \in \text{im}(E)$.

- (ii) Σ_1 and Σ_2 are *boundedly intertwined* if there exists an operator $E \in \mathcal{B}(\mathcal{X}_1; \mathcal{X}_2)$ which intertwines Σ_1 and Σ_2 . The operator E is called an intertwining operator between Σ_1 and Σ_2 .
- (iii) Σ_1 and Σ_2 are *contractively intertwined* if there exists a contraction $E \in \mathcal{B}(\mathcal{X}_1; \mathcal{X}_2)$ which intertwines Σ_1 and Σ_2 .
- (iv) Σ_1 and Σ_2 are *similar* if there exists a boundedly invertible operator $E \in \mathcal{B}(\mathcal{X}_1; \mathcal{X}_2)$ which intertwines Σ_1 and Σ_2 . The operator E is called a similarity operator between Σ_1 and Σ_2 .
- (v) Σ_1 and Σ_2 are *unitarily similar* if there exists a unitary operator $E \in \mathcal{B}(\mathcal{X}_1; \mathcal{X}_2)$ which intertwines Σ_1 and Σ_2 .

Note, in particular, that if Σ_1 and Σ_2 are boundedly intertwined, then they have the same future behavior.

- Definition 3.22**
- (i) The s/s system $\tilde{\Sigma} = (\tilde{V}; \tilde{\mathcal{X}}; \mathcal{W})$ is called an *orthogonal outgoing dilation* of the s/s system $\Sigma = (V; \mathcal{X}; \mathcal{W})$ and Σ is called an *orthogonal outgoing compression onto $\mathcal{X}$* of $\tilde{\Sigma}$, if $\mathcal{X} \subset \tilde{\mathcal{X}}$ and the orthogonal projection of $\tilde{\mathcal{X}}$ onto $\mathcal{X}$ intertwines $\tilde{\Sigma}$ and Σ .
 - (ii) The s/s system $\tilde{\Sigma}$ is called an *incoming dilation* of Σ and Σ is called an *incoming compression* of $\tilde{\Sigma}$ if $\mathcal{X} \subset \tilde{\mathcal{X}}$ and the embedding operator $\mathcal{X} \hookrightarrow \tilde{\mathcal{X}}$ intertwines Σ and $\tilde{\Sigma}$.

4 The Anti-Passive Adjoint State/Signal Systems

4.1 Anti-Passive State/Signal Systems

According to Definition 3.1, the generating subspace V of a passive s/s system $\Sigma = (V; \mathcal{X}, \mathcal{W})$ is required to be maximal nonnegative. Consequently, by Proposition 2.1(iv), its orthogonal companion $V^{[\perp]}$ is maximal nonpositive.

Definition 4.1 Let $\mathcal{X}$ be a Hilbert space and $\mathcal{W}$ a Kreĭn space.

- (i) By an *anti-passive s/s node* in continuous time we mean a triple $\Sigma^\dagger = (V^\dagger; \mathcal{X}, \mathcal{W})$ where $V^\dagger$ is a maximal nonpositive subspace of the Kreĭn node space $\mathfrak{K} := \begin{bmatrix} \mathcal{X} \\ \mathcal{X} \\ \mathcal{W} \end{bmatrix}$ equipped with the inner product (1.7), with the additional property that if $\begin{bmatrix} z \\ 0 \\ 0 \end{bmatrix} \in V^\dagger$, then $z = 0$.
- (ii) *Classical* and *generalized trajectories* of an anti-passive s/s system $\Sigma^\dagger$ are defined in the same way as in the case of a passive system [see Definition 3.1(i)–(ii)].
- (iii) The anti-passive s/s node $\Sigma^\dagger$ together with its families of classical and generalized trajectories is called an *anti-passive s/s system*, and it denoted by the same symbols as the node.
- (iv) By a *past*, *two-sided*, or *future* trajectory of an anti-passive system $\Sigma^\dagger$ we mean a trajectory of Σ on $\mathbb{R}^-$, $\mathbb{R}$, or $\mathbb{R}^+$, respectively.
- (v) A (generalized) trajectory $\begin{bmatrix} x^\dagger \\ w^\dagger \end{bmatrix}$ of an anti-passive s/s system $\Sigma^\dagger = (V^\dagger; \mathcal{X}, \mathcal{W})$ on an interval I is *backward externally generated* if the following condition holds: If I has a finite right end-point t_1 , then we require that $x^\dagger(t_1) = 0$, and if the right end-point of I is ∞ , then we require that $\lim_{t \rightarrow \infty} x^\dagger(t) = 0$ and that $w^\dagger \in L^2([T, \infty); \mathcal{W})$ for every finite $T \in I$.
- (vi) A (generalized) trajectory $\begin{bmatrix} x^\dagger \\ w^\dagger \end{bmatrix}$ of an anti-passive s/s system $\Sigma^\dagger = (V^\dagger; \mathcal{X}, \mathcal{W})$ is *stable* if x is bounded on I and $w \in L^2(I; \mathcal{W})$.

To distinguish between trajectories of a passive s/s system and an anti-passive system we often denote the trajectories of an anti-passive system by $\begin{bmatrix} x^\dagger \\ w^\dagger \end{bmatrix}$.

Remark 4.2 Since the generating subspace $V^\dagger$ of an anti-passive system is maximal nonpositive in the node space $\mathfrak{K}$, it is maximal nonnegative in the anti-space $-\mathfrak{K}$. The inner product in $-\mathfrak{K}$ is given by

$$\left[\begin{bmatrix} z_1 \\ x_1 \\ w_1 \end{bmatrix}, \begin{bmatrix} z_2 \\ x_2 \\ w_2 \end{bmatrix} \right]_{-\mathfrak{K}} = -(-z_1, x_2) - (x_1, -z_2) - [w_1, w_2]_{\mathcal{W}}. \quad (4.1)$$

Recall that the z -component represents the time derivative $\dot{x}^\dagger(t)$ of a classical trajectory $\begin{bmatrix} x^\dagger \\ w^\dagger \end{bmatrix}$, the x -component represents the state $x^\dagger(t)$ itself, and the w -component represents the signal $w^\dagger(t)$. The change of sign in the z -component in (4.1) compared to (1.7) can be interpreted as a reflection of the time direction, since $\frac{d}{dt}x^\dagger(-t) = -\dot{x}^\dagger(-t)$,

and the change of sign in the w -component amounts to the replacement of the signal space $\mathcal{W}$ by its anti-space. This means that the theory of anti-passive s/s systems is identical to the theory of passive s/s systems, apart from a reflection of the time axis, and a change of sign in the signal component. Because of the reflection, externally generated stable trajectories of a passive s/s system correspond to backward externally generated trajectories of the corresponding reflected system. All the results listed in Sects. 3.1–3.3 have anti-passive counterparts, where the past and future have changed places. Note, in particular, that the basic inequalities (1.6), (1.8), and (1.9) are reversed. We shall not give a complete list here, but only formulate those results that we actually use. See also Remark 4.12 below.

Lemma 4.3 $\Sigma = (V; \mathcal{X}, \mathcal{W})$ is a passive s/s system if and only if $\Sigma^{[\perp]} = (V^{[\perp]}; \mathcal{X}, \mathcal{W})$ is an anti-passive s/s system.

Proof Suppose that $\Sigma := (V; \mathcal{X}, \mathcal{W})$ is a passive s/s system. By Proposition 2.1(iv), $V^{[\perp]}$ is maximal nonpositive subspace of the node space $\mathfrak{K}$ since V is maximal non-negative in $\mathfrak{K}$. That $V^{[\perp]}$ also satisfies the additional condition that if $\begin{bmatrix} z^\dagger \\ 0 \\ 0 \end{bmatrix} \in V^\dagger$, then $z^\dagger = 0$ follows from [15, Corollary 4.8]. That also the converse claim is true follows from Remark 4.2. $\square$

Definition 4.4 The anti-passive dual of a passive s/s system $\Sigma = (V; \mathcal{X}, \mathcal{W})$ is the anti-passive s/s system $\Sigma^{[\perp]} := (V^{[\perp]}; \mathcal{X}, \mathcal{W})$.

Above we have defined the anti-passive dual of a s/s system by means of its generating subspace. It can alternatively be characterized by means of the orthogonality between the trajectories of the original system and its dual, as described in the following theorem.

Theorem 4.5 Let $\Sigma = (V; \mathcal{X}, \mathcal{W})$ be a passive s/s system, and let $\Sigma^{[\perp]} = (V^{[\perp]}; \mathcal{X}, \mathcal{W})$ be its anti-passive dual.

- (i) For each interval I with finite right end-point t_2 , the pair of functions $\begin{bmatrix} x^\dagger \\ w^\dagger \end{bmatrix} \in \begin{bmatrix} C(I; \mathcal{X}) \\ L^2(I; \mathcal{W}) \end{bmatrix}$ is a stable trajectory of $\Sigma^{[\perp]}$ in I if and only if

$$(x^\dagger(t_2), x(t_2))_{\mathcal{X}} = (x^\dagger(t_1), x(t_1))_{\mathcal{X}} + \int_{t_1}^{t_2} [w^\dagger(s), w(s)]_{\mathcal{W}} ds \quad (4.2)$$

for all $t_1 \in I$ and all stable trajectories $\begin{bmatrix} x \\ w \end{bmatrix}$ of Σ on $[t_1, t_2]$.

- (ii) For each interval I with right end-point ∞ , the pair of functions $\begin{bmatrix} x^\dagger \\ w^\dagger \end{bmatrix} \in \begin{bmatrix} C(I; \mathcal{X}) \\ L^2(I; \mathcal{W}) \end{bmatrix}$ is a backward externally generated stable trajectory of $\Sigma^{[\perp]}$ in I if and only if $x^\dagger(t) \rightarrow 0$ as $t \rightarrow \infty$ and

$$0 = (x^\dagger(t_1), x(t_1))_{\mathcal{X}} + \int_{t_1}^{\infty} [w^\dagger(s), w(s)]_{\mathcal{W}} ds \quad (4.3)$$

for all $t_1 \in I$ and all stable trajectories $\begin{bmatrix} x \\ w \end{bmatrix}$ of Σ on $[t_1, \infty)$.

Proof The proofs of claims (i) and (ii) are almost identical to each other, so we only prove (i), and leave the proof of (ii) to the reader.

The necessity of (4.2) is quite obvious: if both of the trajectories are classical, then one gets (4.2) by integrating the equation

$$0 = \left[\begin{bmatrix} \dot{x}^\dagger(s) \\ x^\dagger(s) \\ w^\dagger(s) \end{bmatrix}, \begin{bmatrix} \dot{x}(s) \\ x(s) \\ w(s) \end{bmatrix} \right]_{\mathfrak{K}} = - \frac{d}{ds} (x^\dagger(s), x(s))_{\mathcal{X}} + [w^\dagger(s), w(s)]_{\mathcal{W}}$$

over the interval $[t_1, t_2]$. In the case of generalized trajectories we first approximate $\begin{bmatrix} x^\dagger \\ w^\dagger \end{bmatrix}$ and $\begin{bmatrix} x \\ w \end{bmatrix}$ by sequences of classical trajectories, and then pass to the limit to get (4.2) for generalized trajectories $\begin{bmatrix} x^\dagger \\ w^\dagger \end{bmatrix}$ and $\begin{bmatrix} x \\ w \end{bmatrix}$.

Conversely, let $\begin{bmatrix} x^\dagger \\ w^\dagger \end{bmatrix} \in \begin{bmatrix} C(I; \mathcal{X}) \\ L^2(I; \mathcal{W}) \end{bmatrix}$ satisfy (4.2) for all $t_1 \leq t_2$, $t_1 \in I$ and all trajectories $\begin{bmatrix} x \\ w \end{bmatrix}$ of Σ on $[t_1, t_2]$. Fix $t_0 \in I$. By Lemma 3.4(i) and Remark 4.2, there exists a stable trajectory $\begin{bmatrix} x_1^\dagger \\ w_1^\dagger \end{bmatrix}$ of $\Sigma^{[\perp]}$ on $[t_0, t_2]$ with $x_1^\dagger(t_2) = x^\dagger(t_2)$ and $P_{\mathcal{Y}}\pi_{[t_0, t_2]}w_1^\dagger = P_{\mathcal{Y}}\pi_{[t_0, t_2]}w^\dagger$. By the first part of the proof and by our assumption on $\begin{bmatrix} x^\dagger \\ w^\dagger \end{bmatrix}$, for all $t_1 \in [t_0, t_2]$,

$$\begin{aligned} 0 &= (x^\dagger(t_1) - x_1^\dagger(t_1), x(t_1))_{\mathcal{X}} + \int_{t_1}^{t_2} [w^\dagger(s) - w_1^\dagger(s), w(s)]_{\mathcal{W}} ds \\ &= (x^\dagger(t_1) - x_1^\dagger(t_1), x(t_1))_{\mathcal{X}} + \int_{t_1}^{t_2} [P_{\mathcal{U}}(w^\dagger(s) - w_1^\dagger(s)), P_{\mathcal{U}}w(s)]_{\mathcal{W}} ds. \end{aligned}$$

By Lemma 3.4(i), the pair $\begin{bmatrix} x(t_1) \\ P_{\mathcal{U}}P_{[t_1, t_2]}w \end{bmatrix}$ can be an arbitrary vector in $\mathcal{X} \times L^2([t_1, t_2]; \mathcal{U})$, and consequently $x^\dagger(t_1) = x_1^\dagger(t_1)$ and $\pi_{[t_1, t_2]}w^\dagger = \pi_{[t_1, t_2]}w_1^\dagger$. Thus, the restriction of $\begin{bmatrix} x^\dagger \\ w^\dagger \end{bmatrix}$ to any finite interval $[t_0, t_2]$ of I is a trajectory of $\Sigma^{[\perp]}$ on $[t_0, t_2]$, and by Lemma 3.4(iv) and Remark 4.2, $\begin{bmatrix} x^\dagger \\ w^\dagger \end{bmatrix}$ is a stable trajectory of $\Sigma^{[\perp]}$ on I . $\square$

Corollary 4.6 *Let $\Sigma = (V; \mathcal{X}, \mathcal{W})$ be a passive s/s system, and let $\Sigma^{[\perp]} = (V^{[\perp]}; \mathcal{X}, \mathcal{W})$ be the anti-passive dual of Σ .*

- (i) If $\begin{bmatrix} x \\ w \end{bmatrix}$ is a stable past trajectory of Σ and $\begin{bmatrix} x^\dagger \\ w^\dagger \end{bmatrix}$ a stable past trajectory of $\Sigma^{[\perp]}$, then $\lim_{t \rightarrow -\infty} (x(t), x^\dagger(t))_{\mathcal{X}}$ exists, and

$$(x^\dagger(0), x(0))_{\mathcal{X}} = \lim_{t \rightarrow -\infty} (x^\dagger(t), x(t))_{\mathcal{X}} + [w^\dagger, w]_{K_-^2(\mathcal{W})}. \quad (4.4)$$

- (ii) If $\begin{bmatrix} x \\ w \end{bmatrix}$ is a stable future trajectory of Σ and $\begin{bmatrix} x^\dagger \\ w^\dagger \end{bmatrix}$ a stable future trajectory of $\Sigma^{[\perp]}$, then $\lim_{t \rightarrow \infty} (x(t), x^\dagger(t))_{\mathcal{X}}$ exists, and

$$\lim_{t \rightarrow \infty} (x^\dagger(t), x(t))_{\mathcal{X}} = (x^\dagger(0), x(0))_{\mathcal{X}} + [w^\dagger, w]_{K^2(\mathbb{R}^+; \mathcal{W})}. \quad (4.5)$$

- (iii) If $\begin{bmatrix} x \\ w \end{bmatrix}$ is a stable two-sided trajectory of Σ and $\begin{bmatrix} x^\dagger \\ w^\dagger \end{bmatrix}$ is a stable two-sided trajectory of $\Sigma^\dagger$, then $\lim_{t \rightarrow -\infty} (x(t), x^\dagger(t))_{\mathcal{X}}$ and $\lim_{t \rightarrow \infty} (x(t), x^\dagger(t))_{\mathcal{X}}$ exist, and

$$\lim_{t \rightarrow \infty} (x^\dagger(t), x(t))_{\mathcal{X}} = \lim_{t \rightarrow -\infty} (x^\dagger(t), x(t))_{\mathcal{X}} + [w^\dagger, w]_{K^2(\mathbb{R}; \mathcal{W})}. \quad (4.6)$$

Proof This follows immediately from Theorem 4.5. $\square$

Definition 4.7 A passive s/s system $\Sigma = (V; \mathcal{X}, \mathcal{W})$ is *energy preserving* if $V \subset V^{[\perp]}$, it is *co-energy preserving* if $V^{[\perp]} \subset V$, and it is *conservative* if $V = V^{[\perp]}$. Analogously, an anti-passive s/s system $\Sigma^\dagger = (V^\dagger; \mathcal{X}, \mathcal{W})$ is energy preserving, co-energy preserving, or conservative if $V^\dagger \subset (V^\dagger)^{[\perp]}$, $(V^\dagger)^{[\perp]} \subset V^\dagger$, or $V^\dagger = (V^\dagger)^\perp$, respectively.

Thus, in particular, a conservative s/s system is at the same time both passive and anti-passive.

Lemma 4.8 Let $\Sigma = (V; \mathcal{X}, \mathcal{W})$ be a passive s/s system, and let $\Sigma^{[\perp]} = (V^{[\perp]}; \mathcal{X}, \mathcal{W})$ be its anti-passive dual.

- (i) Σ is energy preserving if and only if every trajectory of Σ on every nontrivial interval I is also a trajectory of $\Sigma^{[\perp]}$ on I .
- (ii) Σ is co-energy preserving if and only if every trajectory of $\Sigma^{[\perp]}$ on every nontrivial interval I is also a trajectory of Σ on I .
- (iii) Σ is conservative if and only if Σ and $\Sigma^{[\perp]}$ have the same set of trajectories on every nontrivial interval I .

The same claims remain true if we restrict I to belong to the family of all nontrivial finite subintervals of $\mathbb{R}^+$.

Proof This follows from Definitions 3.1 and 4.7 combined with Lemma 3.3 and Proposition 3.7, which imply that the generating subspace is uniquely determined by the set of all trajectories on some arbitrarily small interval. $\square$

Lemma 4.9 Let $\Sigma = (V; \mathcal{X}, \mathcal{W})$ be a passive s/s node with the space $\mathcal{T}_+$ of stable future trajectories, and define $\mathcal{T}_{0,t}$ and $\mathfrak{K}_{0,t}$ by (3.5) and (1.12), respectively. Then Σ is energy preserving, co-energy preserving or conservative if and only if $\mathcal{T}_{0,t}$ is a neutral, co-neutral or Lagrangian subspace, respectively of $\mathfrak{K}_{0,t}$ for all $t \in \mathbb{R}^+$.

Proof If Σ is energy preserving, then the argument leading up to (1.9) shows that $\mathcal{T}_{0,t}$ is a neutral subspace of $\mathfrak{K}_{0,t}$ for all $t \in \mathbb{R}^+$. Conversely, suppose that $\mathcal{T}_{0,t}$ is a neutral subspace of $\mathfrak{K}_{0,t}$ for all $t \in \mathbb{R}^+$. Then it follows from Theorem 4.5 that every trajectory $\begin{bmatrix} x \\ w \end{bmatrix}$ of Σ on some interval $[0, t]$ is also a trajectory of $\Sigma^{[\perp]}$ on $[0, t]$, and Lemma 4.8 then shows that Σ is neutral.

Recall that Σ is co-energy preserving if and only if its anti-passive dual $\Sigma^{[\perp]}$ is energy preserving. Let $\mathcal{T}_-^\dagger$ be the family of all stable past trajectories of $\Sigma^{[\perp]}$, and denote

$$\mathcal{T}_{t,0}^\dagger := \left\{ \left[\begin{array}{c} x^\dagger(t) \\ x^\dagger(0) \\ \pi_{[t,0]} w^\dagger \end{array} \right] \middle| \begin{bmatrix} x^\dagger \\ w^\dagger \end{bmatrix} \in \mathcal{T}_-^\dagger \right\}, \quad t \in \mathbb{R}^-. \quad (4.7)$$

By Remark 4.2 and the part of Lemma 4.9 which we have already established, $\Sigma^{[\perp]}$ is energy-preserving if and only if $\mathcal{T}_{t,0}^\dagger$ is a neutral subspace of $\mathfrak{K}_{t,0}$ for all $t \in \mathbb{R}^-$, or equivalently, if and only if $\tau^{-t} \mathcal{T}_{-t,0}^\dagger$ is a $\mathfrak{K}_{0,t}$ for all $t \in \mathbb{R}^+$.

We claim that $\tau^{-t} \mathcal{T}_{-t,0}^\dagger = \mathcal{T}_{0,t}^{[\perp]}$. It follows from Theorem 4.5 that $\tau^{-t} \mathcal{T}_{-t,0}^\dagger \subset \mathcal{T}_{0,t}^{[\perp]}$. On the other hand, by Theorem 3.5 and Remark 4.2, $\tau^{-t} \mathcal{T}_{-t,0}^\dagger$ is maximal nonpositive, whereas by Theorem 3.5 and Proposition 2.1(iv), $\mathcal{T}_{0,t}^{[\perp]}$ is (maximal) nonpositive. Thus, $\tau^{-t} \mathcal{T}_{-t,0}^\dagger = \mathcal{T}_{0,t}^{[\perp]}$, as claimed.

Since $\tau^{-t} \mathcal{T}_{-t,0}^\dagger = \mathcal{T}_{0,t}^{[\perp]}$, we find that Σ is energy-preserving if and only if $\mathcal{T}_{0,t}$ is a neutral subspace of $\mathfrak{K}_{0,t}$ for all $t \in \mathbb{R}^+$.

Finally, Σ is conservative if and only if Σ is at the same time both energy preserving and co-energy preserving, and by the above argument, this is equivalent to the condition that $\mathcal{T}_{0,t}$ is both a neutral and a co-neutral subspace of $\mathfrak{K}_{0,t}$. $\square$

4.2 Anti-Passive Behaviors

Definition 4.10 Let $\mathcal{W}$ be a Kreĭn space.

- (i) A maximal nonpositive left-shift invariant subspace of $K_-^2(\mathcal{W})$ is called an *anti-passive past behavior* on the (signal) space $\mathcal{W}$.
- (ii) A maximal nonpositive left-shift invariant subspace of $K_+^2(\mathcal{W})$ is called an *anti-passive future behavior* on the Kreĭn (signal) space $\mathcal{W}$.
- (iii) A maximal nonpositive bilaterally shift-invariant subspace $\mathfrak{W}^\dagger$ of $K^2(\mathcal{W})$ is called an *anti-passive two-sided behavior* on the Kreĭn (signal) space $\mathcal{W}$ if $\mathfrak{W}_+^\dagger := \pi_+ \mathfrak{W}^\dagger$ is a maximal nonpositive subspace of $K_+^2(\mathcal{W})$, or equivalently, if $\mathfrak{W}_-^\dagger := \mathfrak{W}^\dagger \cap K_-^2(\mathcal{W})$ is a maximal nonpositive subspace of $K_-^2(\mathcal{W})$.

Indeed, by Lemma 3.13 and Remark 4.2, the two conditions given in part (iii) of the above definition are equivalent.

- Lemma 4.11** (i) A closed subspace $\mathfrak{W}_+$ of $K_+^2(\mathcal{W})$ is a passive future behavior on $\mathcal{W}$ if and only if $\mathfrak{W}_+^{[\perp]}$ is an anti-passive future behavior on $\mathcal{W}$.
- (ii) A closed subspace $\mathfrak{W}_-$ of $K_-^2(\mathcal{W})$ is a passive past behavior on $\mathcal{W}$ if and only if $\mathfrak{W}_-^{[\perp]}$ is an anti-passive past behavior on $\mathcal{W}$.
- (iii) A closed subspace $\mathfrak{W}$ of $K^2(\mathcal{W})$ is a passive two-sided behavior on $\mathcal{W}$ if and only if $\mathfrak{W}^{[\perp]}$ is an anti-passive two-sided behavior on $\mathcal{W}$.

Proof (i) By definition, $\mathfrak{W}_+$ is a passive future behavior if and only if $\mathfrak{W}_+$ is maximal nonnegative in $K_+^2(\mathcal{W})$ and right-shift invariant. Since $\mathfrak{W}_+$ is assumed to be closed, according to Proposition 2.1(iv), $\mathfrak{W}_+$ is maximal nonnegative if and only if $\mathfrak{W}_+^{[\perp]}$ is maximal nonpositive. It is also easy to see that $\mathfrak{W}_+$ is right-shift invariant if and only if $\mathfrak{W}_+^{[\perp]}$ is left-shift invariant. Thus, $\mathfrak{W}_+$ is a passive future behavior if and only if $\mathfrak{W}_+^{[\perp]}$ is an anti-passive future behavior.

(ii) The proof of the claim about the past behaviors is analogous.

(iii) In the case of two-sided behaviors, by arguing in the same way as above we find that $\mathfrak{W}$ is maximal nonnegative and bilaterally shift-invariant if and only if $\mathfrak{W}^{[\perp]}$ is maximal nonpositive and bilaterally shift-invariant. By the continuous time version of [5, Lemma 3.5] (cf. Remark 3.17),

$$\mathfrak{W}^{[\perp]} \cap K_-^2(\mathfrak{W}) = (\pi_- \mathfrak{W})^{[\perp]}.$$

Thus, by Lemma 3.13, Remark 4.2, and Definitions 3.14 and 4.10, $\mathfrak{W}$ is a passive two-sided behavior if and only if $\mathfrak{W}^{[\perp]}$ is an anti-passive two-sided behavior. $\square$

Remark 4.12 It is easy to see that $\mathfrak{W}_+$, $\mathfrak{W}$, and $\mathfrak{W}_-$ are passive future, two-sided, or past behaviors on the signal space $\mathcal{W}$ if and only if the time-reflected versions of these behaviors are anti-passive past, two-sided, or future behaviors, respectively, on the signal space $-\mathcal{W}$. This implies that all the results in Sect. 3.4 have anti-passive counterparts, where the past and the future have been interchanged with each other. We shall not give a complete list here, but only formulate those results that we actually use. See also Remark 4.2 above.

Lemma 4.13 Let $\mathfrak{W}_+$ be a passive future behavior on a Kreĭn space $\mathcal{W}$. Then the set of all $w^\dagger \in \mathfrak{W}_+^{[\perp]}$ with compact support is a dense subspace of $\mathfrak{W}_+^{[\perp]}$.

Proof This follows from Lemmas 4.11 and 3.19 and Remark 4.12. $\square$

Definition 4.14 By the (stable) *backward behavior* induced by the anti-passive s/s system $\Sigma^\dagger$ on the closed and nontrivial interval I we mean the set of all the signal parts $w^\dagger$ of all backward externally generated stable trajectories $\begin{bmatrix} x^\dagger \\ w^\dagger \end{bmatrix}$ of $\Sigma^\dagger$.

In the special cases $I = \mathbb{R}^-$, $I = \mathbb{R}$, and $I = \mathbb{R}^+$ we denote these behaviors by ${}^\dagger\mathfrak{W}_-^{\Sigma^\dagger}$, ${}^\dagger\mathfrak{W}^{\Sigma^\dagger}$, and ${}^\dagger\mathfrak{W}_+^{\Sigma^\dagger}$, and refer to them as the *past*, *two-sided*, and *future backward behaviors* of $\Sigma^\dagger$, respectively.

Proposition 4.15 The past, two-sided, and future backward behaviors of an anti-passive s/s system are anti-passive past, two-sided, and future behaviors, respectively, in the sense of Definition 4.10.

Proof This follows from Proposition 3.15 and Remark 4.2. $\square$

Proposition 4.16 *The past, two-sided, and future backward behaviors ${}^\dagger\mathfrak{W}_-^{\Sigma^{[\perp]}}$, ${}^\dagger\mathfrak{W}^{\Sigma^{[\perp]}}$, and ${}^\dagger\mathfrak{W}_+^{\Sigma^{[\perp]}}$ of the anti-causal dual $\Sigma^{[\perp]} = (V^{[\perp]}; \mathcal{X}, \mathcal{W})$ of the passive s/s systems $\Sigma = (V; \mathcal{X}, \mathcal{W})$ are given by*

$${}^\dagger\mathfrak{W}_-^{\Sigma^{[\perp]}} = (\mathfrak{W}_-^\Sigma)^{[\perp]}, \quad {}^\dagger\mathfrak{W}^{\Sigma^{[\perp]}} = (\mathfrak{W}^\Sigma)^{[\perp]}, \quad \text{and} \quad {}^\dagger\mathfrak{W}_+^{\Sigma^{[\perp]}} = (\mathfrak{W}_+^\Sigma)^{[\perp]}, \quad (4.8)$$

where $\mathfrak{W}_-^\Sigma$, $\mathfrak{W}^\Sigma$, and $\mathfrak{W}_+^\Sigma$ are the past, two-sided, and future behaviors of Σ .

Proof These three identities are in principle proved in the same way, so we only prove one of them. If $\begin{bmatrix} x \\ w \end{bmatrix}$ and $\begin{bmatrix} x^\dagger \\ w^\dagger \end{bmatrix}$ are stable externally and backward externally generated trajectories of Σ and $\Sigma^{[\perp]}$, respectively, then by Corollary 4.6, $[w, w^\dagger]_{k^2(\mathcal{W})} = 0$. This implies that ${}^\dagger\mathfrak{W}^{\Sigma^{[\perp]}} \subset (\mathfrak{W}^\Sigma)^{[\perp]}$. Since ${}^\dagger\mathfrak{W}^{\Sigma^{[\perp]}}$ is maximal nonpositive and $(\mathfrak{W}^\Sigma)^{[\perp]}$ is nonpositive, this implies that ${}^\dagger\mathfrak{W}^{\Sigma^{[\perp]}} = (\mathfrak{W}^\Sigma)^{[\perp]}$. $\square$

5 The Hilbert Spaces $\mathcal{H}(\mathfrak{W}_+)$, $\mathcal{H}(\mathfrak{W}_-^{[\perp]})$, and $\mathcal{D}(\mathfrak{W})$

In this subsection we shall present three special Hilbert spaces that play a central role throughout the rest of this article. These Hilbert spaces will be used as the state spaces of three of our canonical passive s/s realizations of a given passive behavior. These first two of them are special cases of the Hilbert space $\mathcal{H}(\mathcal{Z})$ constructed in [4] and described in Sect. 2.2, where $\mathcal{Z}$ is a maximal nonnegative subspace of a Kreĭn space $\mathcal{K}$, and the third is constructed from the first two and an angle operator, called the past/future map.

We begin by adapting the spaces $\mathcal{H}(\mathcal{Z})$ from Sect. 2.2 to the case where $\mathcal{Z}$ is either a passive future or an anti-passive past behavior.

5.1 The Hilbert Space $\mathcal{H}_+ := \mathcal{H}(\mathfrak{W}_+)$

Let $\mathfrak{W}_+$ be a given passive future behavior on a Kreĭn signal space $\mathcal{W}$, i.e., $\mathfrak{W}_+$ is a maximal nonnegative right-shift invariant subspace of $K_+^2(\mathcal{W})$. We take $\mathcal{K} = K_+^2(\mathcal{W})$ and $\mathcal{Z} = \mathfrak{W}_+$ in the discussion in Sect. 2.2. Adapting our earlier formulas to this case we get the following result.

Theorem 5.1 *Let $\mathfrak{W}_+$ be a passive future behavior on the Kreĭn space $K_+^2(\mathcal{W})$. Denote the quotient map $K_+^2(\mathcal{W}) \mapsto K_+^2(\mathcal{W})/\mathfrak{W}_+$ by Q_+ , and define $\mathcal{H}_+ := \mathcal{H}(\mathfrak{W}_+)$ and $\|\cdot\|_{\mathcal{H}(\mathfrak{W}_+)}$ by*

$$\|h_+\|_{\mathcal{H}(\mathfrak{W}_+)}^2 = \sup\{-[w_+, w_+]_{K_+^2(\mathcal{W})} \mid w_+ \in h_+\}, \quad h_+ \in K_+^2(\mathcal{W})/\mathfrak{W}_+, \quad (5.1)$$

$$\mathcal{H}(\mathfrak{W}_+) = \{h_+ \in K_+^2(\mathcal{W})/\mathfrak{W}_+ \mid \|h_+\|_{\mathcal{H}(\mathfrak{W}_+)}^2 < \infty\}. \quad (5.2)$$

- (i) $\mathcal{H}(\mathfrak{W}_+)$ equipped with the norm $\|\cdot\|_{\mathcal{H}(\mathfrak{W}_+)}$ is a Hilbert space that is continuously contained in $K_+^2(\mathcal{W})/\mathfrak{W}_+$.

(ii) *The image*

$$\mathcal{H}_+^0 := \mathcal{H}^0(\mathfrak{W}_+) := Q_+ \mathfrak{W}_+^{[\perp]} \quad (5.3)$$

of $\mathfrak{W}_+^{[\perp]}$ under Q_+ is a dense subspace of $\mathcal{H}(\mathfrak{W}_+)$, and

$$\|Q_+ w_+^\dagger\|_{\mathcal{H}(\mathfrak{W}_+)}^2 = -[w_+^\dagger, w_+^\dagger]_{K_+^2(\mathcal{W})}, \quad w_+^\dagger \in \mathfrak{W}_+^{[\perp]}. \quad (5.4)$$

(iii) *Denote the inverse image of $\mathcal{H}(\mathfrak{W}_+)$ under Q_+ by*

$$\mathcal{K}(\mathfrak{W}_+) := Q_+^{-1} \mathcal{H}(\mathfrak{W}_+). \quad (5.5)$$

Then

$$\begin{aligned} (Q_+ w_+^\dagger, Q_+ w_+)_{\mathcal{H}(\mathfrak{W}_+)} &= -[w_+^\dagger, w_+]_{K_+^2(\mathcal{W})}, \\ \text{if } w_+^\dagger &\in \mathfrak{W}_+^{[\perp]} \text{ and } w_+ \in \mathcal{K}(\mathfrak{W}_+). \end{aligned} \quad (5.6)$$

(iv) *The restriction $Q_+|_{\mathcal{K}(\mathfrak{W}_+)}$ is closed and surjective as an operator $K_+^2(\mathcal{W}) \rightarrow \mathcal{H}(\mathfrak{W}_+)$, and it has a bounded right-inverse.*

(v) *If $w_+^k \in \mathcal{K}(\mathfrak{W}_+)$ and $Q_+ w_+^k \rightarrow Q_+ w_+$ in $\mathcal{H}(\mathfrak{W}_+)$ for some $w_+ \in \mathcal{K}(\mathfrak{W}_+)$, then there exists a sequence $z_+^k \in \mathfrak{W}_+$ such that $w_+^k + z_+^k \rightarrow w_+$ in $K_+^2(\mathcal{W})$.*

Proof Claims (i)–(iii) follow from the discussion in Sect. 2.2. Claims (iv) and (v) follow from the more detailed discussion of $\mathcal{H}(\mathcal{Z})$ given in [4, p. 2597]. $\square$

Lemma 5.2 *Let $\mathfrak{W}_+$ be a passive future behavior on the Kreĭn space $\mathcal{W}$. Then the set*

$$\mathcal{H}_0^0(\mathfrak{W}_+) := \{Q_+ w_+^\dagger \mid w_+^\dagger \in \mathfrak{W}_+^{[\perp]} \text{ has compact support}\}$$

(which is contained in $\mathcal{H}^0(\mathfrak{W}_+)$) is a dense subspace of $\mathcal{H}(\mathfrak{W}_+)$.

Proof Let $w_+^\dagger \in \mathfrak{W}_+^{[\perp]}$. Then by Lemma 4.13, there exists a sequence $w_+^k \in \mathfrak{W}_+^{[\perp]}$, where each w_+^k has compact support, such that $w_+^k \rightarrow w_+^\dagger$ in $K_+^2(\mathcal{W})$ as $k \rightarrow \infty$. This implies that $[w_+^k - w_+^\dagger, w_+^k - w_+^\dagger]_{K_+^2(\mathcal{W})} \rightarrow 0$ as $n \rightarrow \infty$, and according to (5.4), this means that $w_+^k + \mathfrak{W}_+ \rightarrow w_+^\dagger + \mathfrak{W}_+$ in $\mathcal{H}(\mathfrak{W}_+)$ as $k \rightarrow \infty$. Since $\mathcal{H}_0^0(\mathfrak{W}_+)$ is dense in $\mathcal{H}(\mathfrak{W}_+)$, this proves the lemma. $\square$

Lemma 5.3 *If $w_+ \in \mathcal{K}(\mathfrak{W}_+)$, where $\mathfrak{W}_+$ is a passive future behavior on the Kreĭn space $\mathcal{W}$, then $\tau_+^t w_+ \in \mathcal{K}(\mathfrak{W}_+)$ for all $t \in \mathbb{R}^+$, and*

$$\|Q_+ \tau_+^t w_+\|_{\mathcal{H}(\mathfrak{W}_+)}^2 \leq \|Q_+ w_+\|_{\mathcal{H}(\mathfrak{W}_+)}^2 + \int_0^t [w_+(s), w_+(s)]_{\mathcal{W}} ds. \quad (5.7)$$

If $w_+ \in \mathfrak{W}_+^{[\perp]}$, then $w_+ \in \mathcal{K}(\mathfrak{W}_+)$ and (5.7) holds with equality.

Proof We have for all $w_+ \in \mathcal{K}(\mathfrak{W}_+)$, all $z \in \mathfrak{W}_+$, and all $t \in \mathbb{R}^+$ (recall that $\tau_+^{*t}\mathfrak{W}_+ \subset \mathfrak{W}_+$)

$$\begin{aligned} -[\tau_+^t w_+ + z, \tau_+^t w_+ + z]_{K_+^2(\mathcal{W})} &= -[\tau_+^t(w_+ + \tau_+^{*t}z), \tau_+^t(w_+ + \tau_+^{*t}z)]_{K_+^2(\mathcal{W})} \\ &= -[w_+ + \tau_+^{*t}z, w_+ + \tau_+^{*t}z]_{K_+^2(\mathcal{W})} + \int_0^t [w_+(s), w_+(s)]_{\mathcal{W}} ds \\ &\leq \|Q_+ w_+\|_{\mathcal{H}(\mathfrak{W}_+)}^2 + \int_0^t [w_+(s), w_+(s)]_{\mathcal{W}} ds. \end{aligned}$$

From here we get (5.7) by taking the supremum over all $z \in \mathfrak{W}_+$. If $w_+ \in \mathfrak{W}_+^{[\perp]}$, then $Q_+ w_+ \in \mathcal{H}^0(\mathfrak{W}_+) \subset \mathcal{H}(\mathfrak{W}_+)$, and by (5.1),

$$\begin{aligned} \|Q_+ \tau_+^t w_+\|_{\mathcal{H}(\mathfrak{W}_+)}^2 - \|Q_+ w_+\|_{\mathcal{H}(\mathfrak{W}_+)}^2 \\ = -[\tau_+^t w_+, \tau_+^t w_+]_{K_+^2(\mathcal{W})} + [w_+, w_+]_{K_+^2(\mathcal{W})} = \int_0^t [w_+(s), w_+(s)]_{\mathcal{W}} ds. \quad \square \end{aligned}$$

5.2 The Hilbert Space $\mathcal{H}_- := \mathcal{H}(\mathfrak{W}_-^{[\perp]})$

Let $\mathfrak{W}_-$ be a given passive past behavior on a Kreĭn signal space $\mathcal{W}$, i.e., $\mathfrak{W}_-$ is a maximal nonnegative right-shift invariant subspace of $K_-^2(\mathcal{W})$. Then $\mathfrak{W}_-^{[\perp]}$ is a maximal nonpositive left-shift invariant subspace of $K_-^2(\mathcal{W})$, and hence it can be interpreted as a maximal nonnegative left-shift invariant subspace of the anti-space $-K_-^2(\mathcal{W})$ of $K_-^2(\mathcal{W})$. This time we take $\mathcal{K} = -K_-^2(\mathcal{W})$ and $\mathcal{Z} = \mathfrak{W}_-^{[\perp]}$ in the definition of $\mathcal{H}(\mathcal{Z})$. Adapting our earlier formulas to this case we get the following result.

Theorem 5.4 *Let $\mathfrak{W}_-$ be a passive past behavior on the Kreĭn space $K_-^2(\mathcal{W})$, and interpret $\mathfrak{W}_-^{[\perp]}$ as a maximal nonnegative left-shift invariant subspace of the anti-space $-K_-^2(\mathcal{W})$. Denote the quotient map $-K_-^2(\mathcal{W}) \mapsto -K_-^2(\mathcal{W})/\mathfrak{W}_-^{[\perp]}$ by Q_- , and define $\mathcal{H}_- := \mathcal{H}(\mathfrak{W}_-^{[\perp]})$ and $\|\cdot\|_{\mathcal{H}(\mathfrak{W}_-^{[\perp]})}$ by*

$$\|h_-\|_{\mathcal{H}(\mathfrak{W}_-^{[\perp]})}^2 = \sup\{[w_-, w_-]_{K_-^2(\mathcal{W})} \mid w_- \in h_-\}, \quad h_- \in -K_-^2(\mathcal{W})/\mathfrak{W}_-^{[\perp]}, \quad (5.8)$$

$$\mathcal{H}(\mathfrak{W}_-^{[\perp]}) = \{h_- \in -K_-^2(\mathcal{W})/\mathfrak{W}_-^{[\perp]} \mid \|h_-\|_{\mathcal{H}(\mathfrak{W}_-^{[\perp]})}^2 < \infty\}. \quad (5.9)$$

- (i) $\mathcal{H}(\mathfrak{W}_-^{[\perp]})$ equipped with the norm $\|\cdot\|_{\mathcal{H}(\mathfrak{W}_-^{[\perp]})}$ is a Hilbert space that is continuously contained in $-K_-^2(\mathcal{W})/\mathfrak{W}_-^{[\perp]}$.
- (ii) The image

$$\mathcal{H}_-^0 := \mathcal{H}^0(\mathfrak{W}_-^{[\perp]}) = Q_- \mathfrak{W}_- \quad (5.10)$$

of $\mathfrak{W}_-$ under Q_- is a dense subspace of $\mathcal{H}(\mathfrak{W}_-^{[\perp]})$, and

$$\|Q_-w_-\|_{\mathcal{H}(\mathfrak{W}_-^{[\perp]})}^2 = [w_-, w_-]_{K_-^2(\mathcal{W})}, \quad w_- \in \mathfrak{W}_-. \quad (5.11)$$

(iii) Denote the inverse image of $\mathcal{H}(\mathfrak{W}_-^{[\perp]})$ under Q_- by

$$\mathcal{K}(\mathfrak{W}_-^{[\perp]}) := Q_-^{-1}\mathcal{H}(\mathfrak{W}_-^{[\perp]}). \quad (5.12)$$

Then

$$\begin{aligned} (Q_-w_-, Q_-v_-)_{\mathcal{H}(\mathfrak{W}_-^{[\perp]})} &= [w_-, v_-]_{K_-^2(\mathcal{W})}, \\ \text{if } w_- \in \mathfrak{W}_- \text{ and } v_- \in \mathcal{K}(\mathfrak{W}_-^{[\perp]}). \end{aligned} \quad (5.13)$$

- (iv) The restriction $Q_-|_{\mathcal{K}(\mathfrak{W}_-^{[\perp]})}$ is closed and surjective as an operator $K_-^2(\mathcal{W}) \rightarrow \mathcal{H}(\mathfrak{W}_-^{[\perp]})$, and it has a bounded right-inverse.
- (v) If $w_-^k \in \mathcal{K}(\mathfrak{W}_-^{[\perp]})$ and $Q_-w_-^k \rightarrow Q_-w_-$ in $\mathcal{H}(\mathfrak{W}_-^{[\perp]})$ for some $w_- \in \mathcal{K}(\mathfrak{W}_-^{[\perp]})$, then there exists a sequence $z_-^k \in \mathfrak{W}_-^{[\perp]}$ such that $w_-^k + z_-^k \rightarrow w_-$ in $K_-^2(\mathcal{W})$.

Proof The proof is analogous to the proof of Theorem 5.1. $\square$

Lemma 5.5 Let $\mathfrak{W}_-$ be a passive past behavior on the Kreĭn space $\mathcal{W}$. Then the set

$$\mathcal{H}_0^0(\mathfrak{W}_-^{[\perp]}) := \{Q_-w_- \mid w_- \in \mathfrak{W}_- \text{ has compact support}\}$$

(which is contained in $\mathcal{H}^0(\mathfrak{W}_-^{[\perp]})$) is a dense subspace of $\mathcal{H}(\mathfrak{W}_-^{[\perp]})$.

Proof This follows from Lemma 5.2 and Remark 4.12. $\square$

Lemma 5.6 If $w_- \in \mathcal{K}(\mathfrak{W}_-^{[\perp]})$, then $\tau_-^{*t}w_- \in \mathcal{K}(\mathfrak{W}_-^{[\perp]})$ and

$$\|Q_- \tau_-^{*t}w_-\|_{\mathcal{H}(\mathfrak{W}_-^{[\perp]})}^2 \leq \|Q_-w_-\|_{\mathcal{H}(\mathfrak{W}_-^{[\perp]})}^2 - \int_{-t}^0 [w_-(s), w_-(s)]_{\mathcal{W}} ds. \quad (5.14)$$

If $w_- \in \mathfrak{W}_-$, then $w_- \in \mathcal{K}(\mathfrak{W}_-^{[\perp]})$ and (5.14) holds with equality.

Proof This follows from Lemma 5.3 and Remark 4.12. $\square$

5.3 The Past/Future Map $\Gamma_{\mathfrak{W}}$ and the Hilbert Space $\mathcal{D}(\mathfrak{W})$

In Sect. 10 we shall also need the quotient space $K^2(\mathcal{W})/(\mathfrak{W}_+ \dot{+} \mathfrak{W}_-^{[\perp]})$. Here $\mathfrak{W}_+ \dot{+} \mathfrak{W}_-^{[\perp]}$ is a closed subspace of $K^2(\mathcal{W})$ since the sum $\mathfrak{W}_+ \dot{+} \mathfrak{W}_-^{[\perp]}$ is direct in $K^2(\mathcal{W})$.

We denote the quotient map $K^2(\mathcal{W}) \mapsto K^2(\mathcal{W})/(\mathfrak{W}_+ \dot{+} \mathfrak{W}_-^{[\perp]})$ by Q . Thus,

$$\begin{aligned} Q_- w_- &:= w_- + \mathfrak{W}_-^{[\perp]}, \quad w_- \in K_-^2(\mathcal{W}), \\ Q_+ w_+ &:= w_+ + \mathfrak{W}_+, \quad w_+ \in K_+^2(\mathcal{W}), \\ Qw &:= w + (\mathfrak{W}_+ \dot{+} \mathfrak{W}_-^{[\perp]}), \quad w \in K^2(\mathcal{W}). \end{aligned} \quad (5.15)$$

To shorten the notations we shall in the sequel throughout use the abbreviations

$$\begin{aligned} \mathcal{H}_- &:= \mathcal{H}(\mathfrak{W}_-^{[\perp]}), \quad \mathcal{H}_-^0 := \mathcal{H}^0(\mathfrak{W}_-^{[\perp]}), \\ \mathcal{H}_+ &:= \mathcal{H}(\mathfrak{W}_+), \quad \mathcal{H}_+^0 := \mathcal{H}^0(\mathfrak{W}_+). \end{aligned} \quad (5.16)$$

Each vector in the quotient space $K^2(\mathcal{W})/(\mathfrak{W}_+ \dot{+} \mathfrak{W}_-^{[\perp]})$ is an equivalence class of the type $x := w + (\mathfrak{W}_+ \dot{+} \mathfrak{W}_-^{[\perp]})$ for some $w \in K^2(\mathcal{W})$. Since $K^2(\mathcal{W}) = K_+^2(\mathcal{W}) \boxplus K_-^2(\mathcal{W})$, and since $\mathfrak{W}_+$ is a closed subspace of $K_+^2(\mathcal{W})$ and $\mathfrak{W}_-^{[\perp]}$ is a closed subspace of $K_-^2(\mathcal{W})$, it follows that we can identify $K^2(\mathcal{W})/(\mathfrak{W}_+ \dot{+} \mathfrak{W}_-^{[\perp]})$ with the product space $\begin{bmatrix} K_+^2(\mathcal{W})/\mathfrak{W}_+ \\ K_-^2(\mathcal{W})/\mathfrak{W}_-^{[\perp]} \end{bmatrix}$.

We denote the projections of $K^2(\mathcal{W})/(\mathfrak{W}_+ \dot{+} \mathfrak{W}_-^{[\perp]})$ onto $K_+^2(\mathcal{W})/\mathfrak{W}_+$ and $K_-^2(\mathcal{W})/\mathfrak{W}_-^{[\perp]}$ by P_+ and P_- , respectively. Thus, $P_\pm$ is the operator which for each $w \in K^2(\mathcal{W})$ maps $x = Qw$ into $Q_\pm \pi_\pm w$. Since $\mathcal{H}_+$ is continuously contained in $K_+^2(\mathcal{W})/\mathfrak{W}_+$ and $\mathcal{H}_-$ is continuously contained in $K_-^2(\mathcal{W})/\mathfrak{W}_-^{[\perp]}$, this means that $\begin{bmatrix} \mathcal{H}_+ \\ \mathcal{H}_- \end{bmatrix}$ can be interpreted as a continuously contained subspace of $K^2(\mathcal{W})/(\mathfrak{W}_+ \dot{+} \mathfrak{W}_-^{[\perp]})$.

Lemma 5.7 *Let $\mathfrak{W}$ be a passive two-sided behavior on $\mathcal{W}$ with the corresponding passive past behavior $\mathfrak{W}_- = \pi_- \mathfrak{W}$ and passive future behavior $\mathfrak{W}_+ = \mathfrak{W} \cap K_+^2(\mathcal{W})$. Then there exists a unique contraction $\Gamma_{\mathfrak{W}}: \mathcal{H}_- \rightarrow \mathcal{H}_+$ satisfying*

$$\Gamma_{\mathfrak{W}} Q_- \pi_- w = Q_+ \pi_+ w, \quad w \in \mathfrak{W}, \quad (5.17)$$

where Q_- is the quotient map $K_-^2 \mapsto K_-^2/\mathfrak{W}_-^{[\perp]}$ and Q_+ is the quotient map $K_+^2 \mapsto K_+^2/\mathfrak{W}_+$.

Proof The proof is essentially the same as the proof of [5, Lemma 6.1] (see Remark 3.17). $\square$

Definition 5.8 The contraction $\Gamma_{\mathfrak{W}}: \mathcal{H}_- \rightarrow \mathcal{H}_+$ in Lemma 5.7 is called the *past/future map* of the two-sided behavior $\mathfrak{W}$.

Throughout the rest of this section we let $\mathfrak{W}$ be a passive two-sided behavior on $\mathcal{W}$, and define the corresponding passive past and future behaviors $\mathfrak{W}_-$ and $\mathfrak{W}_+$ by (3.14).

Let

$$A_{\mathfrak{W}} := \begin{bmatrix} 1_{\mathcal{H}_+} & \Gamma_{\mathfrak{W}} \\ \Gamma_{\mathfrak{W}}^* & 1_{\mathcal{H}_-} \end{bmatrix}. \quad (5.18)$$

This is a bounded linear operator on $\mathcal{H}_+ \oplus \mathcal{H}_-$. It is nonnegative since $\Gamma_{\mathfrak{W}}$ is a contraction $\mathcal{H}_- \rightarrow \mathcal{H}_+$, and by the Schwarz inequality, for all $\begin{bmatrix} x_+ \\ x_- \end{bmatrix} \in \mathcal{H}_+ \oplus \mathcal{H}_-$,

$$\begin{aligned} \left(\begin{bmatrix} x_+ \\ x_- \end{bmatrix}, A_{\mathfrak{W}} \begin{bmatrix} x_+ \\ x_- \end{bmatrix} \right)_{\mathcal{H}_+ \oplus \mathcal{H}_-} &= \|x_+\|_{\mathcal{H}_+}^2 + 2\Re(x_+, \Gamma_{\mathfrak{W}}x_-)_{\mathcal{H}_+} + \|x_-\|_{\mathcal{H}_-}^2 \\ &\geq \|x_+\|_{\mathcal{H}_+}^2 - 2\|x_+\|_{\mathcal{H}_+}\|x_-\|_{\mathcal{H}_-} + \|x_-\|_{\mathcal{H}_-}^2 \geq 0. \end{aligned}$$

We define $\mathcal{D}(\mathfrak{W})$ to be the range of $A_{\mathfrak{W}}^{1/2}$, with the range norm, i.e.,

$$\left\| \begin{bmatrix} x_+ \\ x_- \end{bmatrix} \right\|_{\mathcal{D}(\mathfrak{W})} = \left\| (A_{\mathfrak{W}}^{1/2})^{[-1]} \begin{bmatrix} x_+ \\ x_- \end{bmatrix} \right\|_{\mathcal{H}_+ \oplus \mathcal{H}_-},$$

where $(A_{\mathfrak{W}}^{1/2})^{[-1]}$ is the pseudo-inverse of $A_{\mathfrak{W}}^{1/2}$, i.e., $\begin{bmatrix} x'_+ \\ x'_- \end{bmatrix} := (A_{\mathfrak{W}}^{1/2})^{[-1]} \begin{bmatrix} x_+ \\ x_- \end{bmatrix}$ is the unique vector in $\overline{\text{im}(A_{\mathfrak{W}})} = \overline{\text{im}(A_{\mathfrak{W}}^{1/2})}$ which satisfies $\begin{bmatrix} x_+ \\ x_- \end{bmatrix} = A_{\mathfrak{W}}^{1/2} \begin{bmatrix} x'_+ \\ x'_- \end{bmatrix}$. With respect to this inner product in the range space the operator $A_{\mathfrak{W}}^{1/2}|_{\overline{\text{im}(A_{\mathfrak{W}})}}$ is a unitary operator mapping $\overline{\text{im}(A_{\mathfrak{W}})}$ onto $\mathcal{D}(\mathfrak{W})$. In particular, $\mathcal{D}(\mathfrak{W})$ is a Hilbert space.

Lemma 5.9 Define $A_{\mathfrak{W}}$ by (5.18).

- (i) $\text{im}(A_{\mathfrak{W}})$ is a dense subset of the Hilbert space $\mathcal{D}(\mathfrak{W})$, $\mathcal{D}(\mathfrak{W})$ is a dense subspace of $\overline{\text{im}(A_{\mathfrak{W}})}$, and $\mathcal{D}(\mathfrak{W})$ is continuously contained in $\mathcal{H}_+ \oplus \mathcal{H}_-$.
- (ii) $A_{\mathfrak{W}}$ is bounded as an operator $\mathcal{H}_+ \oplus \mathcal{H}_- \rightarrow \mathcal{D}(\mathfrak{W})$.
- (iii) If $x \in \mathcal{D}(\mathfrak{W})$ and $y = A_{\mathfrak{W}}x'$, then $y \in \mathcal{D}(\mathfrak{W})$, and $(x, y)_{\mathcal{D}(\mathfrak{W})} = (x, x')_{\mathcal{H}_+ \oplus \mathcal{H}_-}$.
- (iv) $T_- := A_{\mathfrak{W}}|_{\mathcal{H}_-} = \begin{bmatrix} \Gamma_{\mathfrak{W}} \\ 1_{\mathcal{H}_-} \end{bmatrix}$ is an isometry $\mathcal{H}_- \rightarrow \mathcal{D}(\mathfrak{W})$.
- (v) $T_+ := A_{\mathfrak{W}}|_{\mathcal{H}_+} = \begin{bmatrix} 1_{\mathcal{H}_+} \\ \Gamma_{\mathfrak{W}}^* \end{bmatrix}$ is an isometry $\mathcal{H}_+ \rightarrow \mathcal{D}(\mathfrak{W})$.

Proof The proof is essentially the same as the proof of [6, Lemma 3.1] (see Remark 3.17). $\square$

Lemma 5.10 Denote $\mathcal{L}_{\pm} := \text{im}(T_{\pm})$, where $T_{\pm}$ are the operators defined in Lemma 5.9. Then $\mathcal{L}_+ + \mathcal{L}_-$ is dense in $\mathcal{D}(\mathfrak{W})$, and

$$P_{\mathcal{L}_+}|_{\mathcal{L}_-} = T_+ \Gamma_{\mathfrak{W}} T_-^*|_{\mathcal{L}_-}. \quad (5.19)$$

Proof That $\mathcal{L}_+ + \mathcal{L}_- = \text{im}(A_{\mathfrak{W}})$ is dense in $\mathcal{D}(\mathfrak{W})$ is contained in part (i) of Lemma 5.9.

Let $x \in \mathcal{D}(\mathfrak{W})$, $x_- \in \mathcal{L}_-$, and define $h_- := T_-^* x_-$. Then $x_- = T_- h_-$ (since $T_- : \mathcal{H}_- \rightarrow \mathcal{L}_-$ is unitary). Define $x_+ := P_{\mathcal{L}_+} x$ and $h_+ = T_+^* x_+$, so that $x_+ = T_+ h_+$. Then

$$\begin{aligned}
 (x, P_{\mathcal{L}_+} x_-)_{\mathcal{D}(\mathfrak{W})} &= (P_{\mathcal{L}_+} x, x_-)_{\mathcal{D}(\mathfrak{W})} = (P_{\mathcal{L}_+} x, A_{\mathfrak{W}} \begin{bmatrix} 0 \\ h_- \end{bmatrix})_{\mathcal{D}(\mathfrak{W})} \\
 &= (P_{\mathcal{L}_+} x, \begin{bmatrix} 0 \\ h_- \end{bmatrix})_{\mathcal{H}_+ \oplus \mathcal{H}_-} = \left(\begin{bmatrix} h_+ \\ \Gamma_{\mathfrak{W}}^* h_+ \end{bmatrix}, \begin{bmatrix} 0 \\ h_- \end{bmatrix} \right)_{\mathcal{H}_+ \oplus \mathcal{H}_-} \\
 &= (\Gamma_{\mathfrak{W}}^* h_+, h_-)_{\mathcal{H}_-} = (h_+, \Gamma_{\mathfrak{W}} h_-)_{\mathcal{H}_+} \\
 &= (T_+^* P_{\mathcal{L}_+} x, \Gamma_{\mathfrak{W}} T_-^* x_-)_{\mathcal{H}_+} = (P_{\mathcal{L}_+} x, T_+ \Gamma_{\mathfrak{W}} T_-^* x_-)_{\mathcal{H}_+} \\
 &= (x, T_+ \Gamma_{\mathfrak{W}} T_-^* x_-)_{\mathcal{H}_+}
 \end{aligned}$$

This proves (5.19). $\square$

Remark 5.11 Lemma 5.10 may be reformulated as follows. The space $\mathcal{D}(\mathfrak{W})$ is equal to the closed linear span of its subspaces $\mathcal{L}_+ \vee \mathcal{L}_-$, where $\mathcal{L}_{\pm}$ are the unitary images in $\mathcal{D}(\mathfrak{W})$ of $\mathcal{H}_{\pm}$ under $T_{\pm}$, and the angle operator $K := P_{\mathcal{L}_+}|_{\mathcal{L}_-}$ between $\mathcal{L}_-$ and $\mathcal{L}_+$ is given by (5.19) in terms of $T_{\pm}$ and the past/future map $\Gamma_{\mathfrak{W}}$. In particular,

$$\begin{aligned}
 1_{\mathcal{L}_-} - K^* K &= 1_{\mathcal{L}_-} - T_- \Gamma_{\mathfrak{W}}^* \Gamma_{\mathfrak{W}} T_-^*|_{\mathcal{L}_-} = T_- (1_{\mathcal{L}_-} - \Gamma_{\mathfrak{W}}^* \Gamma_{\mathfrak{W}}) T_-^*|_{\mathcal{L}_-}, \\
 1_{\mathcal{L}_+} - K K^* &= 1_{\mathcal{L}_+} - T_+ \Gamma_{\mathfrak{W}} \Gamma_{\mathfrak{W}}^* T_+^*|_{\mathcal{L}_+} = T_+ (1_{\mathcal{L}_+} - \Gamma_{\mathfrak{W}} \Gamma_{\mathfrak{W}}^*) T_+^*|_{\mathcal{L}_+}.
 \end{aligned}$$

This leads to the following conclusions:

- (i) The following conditions are equivalent:
 - (a) K is an isometry;
 - (b) $\mathcal{D}(\mathfrak{W}) = \mathcal{L}_+$;
 - (c) $\Gamma_{\mathfrak{W}}$ is an isometry $\mathcal{H}_- \rightarrow \mathcal{H}_+$.
- (ii) The following conditions are equivalent:
 - (a) K is a co-isometry;
 - (b) $\mathcal{D}(\mathfrak{W}) = \mathcal{L}_-$;
 - (c) $\Gamma_{\mathfrak{W}}$ is a co-isometry $\mathcal{H}_- \rightarrow \mathcal{H}_+$.

In the sequel we shall throughout interpret $A_{\mathfrak{W}}$ as a bounded linear operator $\mathcal{H}_+ \oplus \mathcal{H}_- \rightarrow \mathcal{D}(\mathfrak{W})$, instead of interpreting $A_{\mathfrak{W}}$ as a self-adjoint operator in $\mathcal{H}_+ \oplus \mathcal{H}_-$. In particular, in this setting the operator $A_{\mathfrak{W}}$ is not self-adjoint unless $\mathcal{D}(\mathfrak{W}) = \mathcal{H}_+ \oplus \mathcal{H}_-$, i.e., unless $\Gamma_{\mathfrak{W}} = 0$. When the duality in the range space is taken with respect to the inner product in $\mathcal{D}(\mathfrak{W})$ instead of the inner product in $\mathcal{H}_+ \oplus \mathcal{H}_-$ then the operator $A_{\mathfrak{W}}^*$ becomes a bounded linear operator $\mathcal{D}(\mathfrak{W}) \rightarrow \mathcal{H}_+ \oplus \mathcal{H}_-$.

Recall that we denoted the projections of $K^2(\mathcal{W})/(\mathfrak{W}_+ \dot{+} \mathfrak{W}_-^{[\perp]})$ onto $K_+^2(\mathcal{W})/\mathfrak{W}_+$ and $K_-^2(\mathcal{W})/\mathfrak{W}_-^{[\perp]}$ by P_+ and P_- , respectively. We denote the restrictions of $P_{\pm}$ to $\begin{bmatrix} \mathcal{H}_- \\ \mathcal{H}_+ \end{bmatrix}$ by $\Pi_{\pm}$, so that $\Pi_{\pm} \begin{bmatrix} x_+ \\ x_- \end{bmatrix} = x_{\pm}$ for all $\begin{bmatrix} x_+ \\ x_- \end{bmatrix} \in \begin{bmatrix} \mathcal{H}_+ \\ \mathcal{H}_- \end{bmatrix}$.

Lemma 5.12 Let $A_{\mathfrak{W}}$ be the operator defined in (5.18), interpreted as bounded linear operator $\mathcal{H}_+ \oplus \mathcal{H}_- \rightarrow \mathcal{D}(\mathfrak{W})$, whose adjoint $A_{\mathfrak{W}}^*$ is a bounded linear operator $\mathcal{D}(\mathfrak{W}) \rightarrow \mathcal{H}_+ \oplus \mathcal{H}_-$.

- (i) $A_{\mathfrak{W}}^*$ is equal to the embedding operator $\mathcal{D}(\mathfrak{W}) \hookrightarrow \begin{bmatrix} \mathcal{H}_+ \\ \mathcal{H}_- \end{bmatrix}$.
- (ii) $(A_{\mathfrak{W}}|_{\mathcal{H}_+})^* = \Pi_+|_{\mathcal{D}(\mathfrak{W})}$ and $(A_{\mathfrak{W}}|_{\mathcal{H}_-})^* = \Pi_-|_{\mathcal{D}(\mathfrak{W})}$. (In the computation of these adjoints we interpret $A_{\mathfrak{W}}|_{\mathcal{H}_{\pm}}$ as operators $\mathcal{H}_{\pm} \rightarrow \mathcal{D}(\mathfrak{W})$.)

Proof By Lemma 5.9(iii), for all $x \in \mathcal{D}(\mathfrak{W})$ and all $y' \in \mathcal{H}_+ \oplus \mathcal{H}_-$,

$$(x, A_{\mathfrak{W}}y')_{\mathcal{D}(\mathfrak{W})} = (x, y')_{\mathcal{H}_+ \oplus \mathcal{H}_-}.$$

This proves claim (i). If we in the same computation replace $y' \in \mathcal{H}_+ \oplus \mathcal{H}_-$ by either $y' \in \mathcal{H}_+$ or $y' \in \mathcal{H}_-$ we get claim (ii). $\square$

We let $\mathcal{L}(\mathfrak{W})$ be the inverse image of $\mathcal{D}(\mathfrak{W})$ under Q , i.e.,

$$\mathcal{L}(\mathfrak{W}) := Q^{-1}\mathcal{D}(\mathfrak{W}) := \{w \in K^2(\mathcal{W}) \mid Qw \in \mathcal{D}(\mathfrak{W})\}, \quad (5.20)$$

and denote

$$\mathcal{D}^0(\mathfrak{W}) := Q(\mathfrak{W} + \mathfrak{W}^{\perp}) := \{Q(z + z^{\dagger}) \mid z \in \mathfrak{W}, z^{\dagger} \in \mathfrak{W}^{\perp}\}. \quad (5.21)$$

As we shall see in the following lemma, $\mathcal{D}^0(\mathfrak{W})$ is a dense subspace of $\mathcal{D}(\mathfrak{W})$.

- Lemma 5.13** (i) If $z \in \mathfrak{W}$ and $z^{\dagger} \in \mathfrak{W}^{\perp}$, then $Q(z + z^{\dagger}) = A_{\mathfrak{W}} \begin{bmatrix} Q_+ \pi_+ z^{\dagger} \\ Q_- \pi_- z \end{bmatrix}$. In particular, $\mathcal{D}^0(\mathfrak{W}) \subset \text{im}(A_{\mathfrak{W}})$ and $\mathfrak{W} + \mathfrak{W}^{\perp} \subset \mathcal{L}(\mathfrak{W})$.
- (ii) $\mathcal{D}^0(\mathfrak{W})$ is a dense subspace of $\mathcal{D}(\mathfrak{W})$.
- (iii) If $w \in \mathcal{L}(\mathfrak{W})$, $z \in \mathfrak{W}$, and $z^{\dagger} \in \mathfrak{W}^{\perp}$, then

$$(Qw, Qz)_{\mathcal{D}(\mathfrak{W})} = (Q_- \pi_- w, Q_- \pi_- z)_{\mathcal{H}_-} = [\pi_- w, \pi_- z]_{K_-^2(\mathcal{W})} \quad (5.22)$$

$$(Qw, Qz^{\dagger})_{\mathcal{D}(\mathfrak{W})} = (Q_+ \pi_+ w, Q_+ \pi_+ z^{\dagger})_{\mathcal{H}_+} = -[\pi_+ w, \pi_+ z^{\dagger}]_{K_+^2(\mathcal{W})}. \quad (5.23)$$

In particular,

$$\|Qz\|_{\mathcal{D}(\mathfrak{W})}^2 = \|Q_- \pi_- z\|_{\mathcal{H}_-}^2 = [\pi_- z, \pi_- z]_{K_-^2(\mathcal{W})}, \quad z \in \mathfrak{W}, \quad (5.24)$$

$$\|Qz^{\dagger}\|_{\mathcal{D}(\mathfrak{W})}^2 = \|Q_+ \pi_+ z^{\dagger}\|_{\mathcal{H}_+}^2 = -[\pi_+ z^{\dagger}, \pi_+ z^{\dagger}]_{K_+^2(\mathcal{W})}, \quad z^{\dagger} \in \mathfrak{W}^{\perp}. \quad (5.25)$$

Proof The proof is essentially the same as the proof of [6, Lemma 3.3] (see Remark 3.17). $\square$

- Lemma 5.14** (i) If $w \in \mathcal{L}(\mathfrak{W})$, then $\tau^t w \in \mathcal{L}(\mathfrak{W})$ for all $t \in \mathbb{R}$, and

$$\|Q\tau^t w\|_{\mathcal{D}(\mathfrak{W})}^2 = \int_0^t [w(s), w(s)]_{\mathcal{W}} ds + \|Qw\|_{\mathcal{D}(\mathfrak{W})}^2, \quad t \in \mathbb{R}. \quad (5.26)$$

(ii) If $w_1, w_2 \in \mathcal{L}(\mathfrak{W})$, then for all $t \in \mathbb{R}$,

$$(Qw_1, Q\tau^t w_2)_{\mathcal{D}(\mathfrak{W})} = \int_0^t [w_1(s-t), w_2(s)]_{\mathcal{W}} ds + (Q\tau^{-t} w_1, Qw_2)_{\mathcal{D}(\mathfrak{W})}. \quad (5.27)$$

Proof The proof is essentially the same as the proof of [6, Lemma 3.4] (see Remark 3.17). $\square$

6 The Output and Input Maps

6.1 The Output Map $\mathfrak{C}_\Sigma$

We begin by presenting the output map $\mathfrak{C}_\Sigma$ of a passive s/s system $\Sigma = (V; \mathcal{X}, \mathcal{W})$ with future behavior $\mathfrak{W}_+$. This is an operator from the state space $\mathcal{X}$ into the Hilbert space $\mathcal{H}_+$ which was defined in (5.2). As in Theorem 5.1 we denote the quotient map $K_+^2(\mathcal{W}) \mapsto K_+^2(\mathcal{W})/\mathfrak{W}_+$ by Q_+ , and the inverse image of $\mathcal{H}_+$ under Q_+ by $\mathcal{K}(\mathfrak{W}_+)$.

Lemma 6.1 *Let $\Sigma = (V; \mathcal{X}; \mathcal{W})$ be a passive s/s system with future behavior $\mathfrak{W}_+$. If $\begin{bmatrix} x \\ w \end{bmatrix}$ is a stable future trajectory of Σ , then*

$$w \in \mathcal{K}(\mathfrak{W}_+) \quad \text{and} \quad \|Q_+ w\|_{\mathcal{H}_+} \leq \|x(0)\|_{\mathcal{X}}. \quad (6.1)$$

Proof Let $\begin{bmatrix} x \\ w \end{bmatrix}$ be a stable future trajectory of Σ , let $z \in \mathfrak{W}_+$, and let $\begin{bmatrix} x_1 \\ z \end{bmatrix}$ be the corresponding externally generated stable future trajectory of Σ . Then $\begin{bmatrix} x+x_1 \\ w+z \end{bmatrix}$ is a stable future trajectory of Σ , and by (1.9),

$$-[w+z, w+z]_{K_+^2(\mathcal{W})} \leq \|x(0) + x_1(0)\|_{\mathcal{X}}^2 = \|x(0)\|_{\mathcal{X}}^2.$$

Taking the supremum over all $z \in \mathfrak{W}_+$ we find that (6.1) holds. $\square$

Lemma 6.2 *Let $\Sigma = (V; \mathcal{X}; \mathcal{W})$ be a passive s/s system with future behavior $\mathfrak{W}_+$. Then the formula*

$$\mathfrak{C}_\Sigma x_0 = \left\{ Q_+ w \mid \begin{array}{l} w \text{ is the signal part of some stable future} \\ \text{trajectory } \begin{bmatrix} x \\ w \end{bmatrix} \text{ of } \Sigma \text{ with } x(0) = x_0 \end{array} \right\} \quad (6.2)$$

defines a linear contraction $\mathfrak{C}_\Sigma: \mathcal{X} \rightarrow \mathcal{H}_+$.

Proof Let $\begin{bmatrix} x \\ w \end{bmatrix}$ be a stable future trajectory of Σ . If $\begin{bmatrix} x_1 \\ w_1 \end{bmatrix}$ is another stable future trajectory of Σ with the same initial state $x_1(0) = x(0)$, then $w_1 - w \in \mathfrak{W}_+$, and conversely, if $w_1 - w \in \mathfrak{W}_+$, then there exist a stable future trajectory $\begin{bmatrix} x_1 \\ w_1 \end{bmatrix}$ with $x_1(0) = x(0)$. Thus, the set of all signal parts w of the stable future trajectories $\begin{bmatrix} x \\ w \end{bmatrix}$ of Σ with fixed initial state $x(0) = x_0$ is an equivalence class in $K_+^2(\mathcal{W})/\mathfrak{W}_+$. By

(6.1), the map $\mathfrak{C}_\Sigma$ from x_0 to this equivalence class is a contraction $\mathcal{X} \rightarrow \mathcal{H}_+$. It is easy to see that this map is linear, and by Lemma 3.4(i), the domain of $\mathfrak{C}_\Sigma$ is all of $\mathcal{X}$. $\square$

Definition 6.3 The contraction $\mathfrak{C}_\Sigma$ in Lemma 6.2 above is called the *output map* of Σ .

In our next lemma we need the inverse image of $\text{im}(\mathfrak{C}_\Sigma)$ under Q_+ which we denote by

$$\mathfrak{S}_+^\Sigma = Q_+^{-1} \cdot \text{im}(\mathfrak{C}_\Sigma). \quad (6.3)$$

Thus, $\mathfrak{S}_+^\Sigma$ consists of the signal parts w of all stable future trajectories $\begin{bmatrix} x \\ w \end{bmatrix}$ of Σ . By Lemma 6.1, $\mathfrak{S}_+^\Sigma \subset \mathcal{K}(\mathfrak{W}_+)$, where $\mathcal{K}(\mathfrak{W}_+)$ is the space defined in (5.5).

Lemma 6.4 Let $\Sigma = (V; \mathcal{X}; \mathcal{W})$ be a passive s/s system with future behavior $\mathfrak{W}_+$ and output map $\mathfrak{C}_\Sigma$, and define $\mathfrak{S}_+^\Sigma$ by (6.3). Then every stable future trajectory $\begin{bmatrix} x \\ w \end{bmatrix}$ of Σ satisfies

$$w \in \mathfrak{S}_+^\Sigma, \quad \tau_+^t w \in \mathfrak{S}_+^\Sigma, \quad \text{and} \quad \mathfrak{C}_\Sigma x(t) = Q_+ \tau_+^t w, \quad t \in \mathbb{R}^+. \quad (6.4)$$

Proof That $w \in \mathfrak{S}_+^\Sigma$ follows immediately from (6.3). To get (6.4) we simply shift the trajectory $\begin{bmatrix} x \\ w \end{bmatrix}$ to the left by the amount t and apply (6.2) with x_0 replaced by $x(t)$. $\square$

Definition 6.5 By an *unobservable future trajectory* of a passive s/s system Σ we mean a future trajectory of Σ of the type $\begin{bmatrix} x \\ 0 \end{bmatrix}$ (i.e., the signal part is identically zero). The *unobservable subspace* $\mathfrak{U}_\Sigma$ of Σ consists of all the initial states $x(0)$ of all unobservable future trajectories of Σ . The system Σ is *observable* if $\mathfrak{U}_\Sigma = \{0\}$.

By Lemma 3.2, every unobservable future trajectory of a passive s/s system is stable.

Lemma 6.6 The unobservable subspace $\mathfrak{U}_\Sigma$ of a passive s/s system $\Sigma = (V; \mathcal{X}; \mathcal{W})$ is equal to the null space of its output map $\mathfrak{C}_\Sigma$.

Proof It follows directly from Definition 6.5 and Lemma 6.4 that if $x_0 \in \mathfrak{U}_\Sigma$, then $\mathfrak{C}_\Sigma x_0 = \mathfrak{W}_+$, and hence $\mathfrak{C}_\Sigma x_0$ is the zero element in $\mathcal{H}_+$. Conversely, suppose that $x_0 \in \ker(\mathfrak{C}_\Sigma)$, i.e., $\mathfrak{C}_\Sigma x_0 = \mathfrak{W}_+$. By Lemma 3.4(i), there exists a stable future trajectory $\begin{bmatrix} x_1 \\ w_1 \end{bmatrix}$ of Σ with $x_1(0) = x_0$, and by Lemma 6.4, $w_1 \in \mathfrak{C}_\Sigma x_0 = \mathfrak{W}_+$. Let $\begin{bmatrix} x_2 \\ w_1 \end{bmatrix}$ be the externally generated future trajectory of Σ whose signal part is w_1 (cf. Lemma 3.11), and define $x = x_1 - x_2$. Then $\begin{bmatrix} x \\ 0 \end{bmatrix}$ is an unobservable future trajectory of Σ with $x(0) = x_0$, and hence $x_0 \in \mathfrak{U}_\Sigma$. $\square$

Lemma 6.7 If the passive s/s system $\Sigma = (V; \mathcal{X}; \mathcal{W})$ is observable, then $\begin{bmatrix} x \\ w \end{bmatrix}$ is a stable future trajectory of Σ if and only if (6.4) holds.

Proof The necessity of (6.4) follows from Lemma 6.4 and (6.3). Conversely, suppose that (6.4) holds. According to (6.3) there exists at least one stable future trajectory

$\begin{bmatrix} x_1 \\ w \end{bmatrix}$ of Σ , and by Lemma 6.4, (6.4) holds with x replaced by x_1 . By Lemma 6.6 and the observability assumption on Σ , $\mathfrak{C}_\Sigma$ is injective, and hence (6.4) implies that $x(t) = x_1(t)$ for all $t \in \mathbb{R}^+$. This implies that $\begin{bmatrix} x \\ w \end{bmatrix}$ is a stable future trajectory of Σ . $\square$

Lemma 6.8 *Let $\Sigma = (V; \mathcal{X}, \mathcal{W})$ be a passive s/s system with output map $\mathfrak{C}_\Sigma$. Then $\begin{bmatrix} x \\ w \end{bmatrix}$ is a stable future trajectory of Σ if and only if $x = x_1 + x_2$, where $\begin{bmatrix} x_1 \\ 0 \end{bmatrix}$ is an unobservable future trajectory of Σ and $\begin{bmatrix} x_2 \\ w \end{bmatrix}$ is a stable future trajectory of Σ with $x_2(0) \in (\ker(\mathfrak{C}_\Sigma))^\perp$. This decomposition is unique, and (6.4) also holds with x replaced by x_2 .*

Proof Trivially, if x has a decomposition of the type described in the lemma, then $\begin{bmatrix} x \\ w \end{bmatrix}$ is a stable future trajectory of Σ .

Conversely, let $\begin{bmatrix} x \\ w \end{bmatrix}$ be a stable future trajectory of Σ . Define $x_1(0) = P_{\mathfrak{U}_\Sigma} x(0)$ and $x_2(0) = P_{\mathfrak{U}_\Sigma^\perp} x(0)$. Then $x(0) = x_1(0) + x_2(0)$ and $x_1(0) \in \mathfrak{U}_\Sigma$. The latter condition implies that $x_1(0)$ is the initial state of some unobservable trajectory $\begin{bmatrix} x_1 \\ 0 \end{bmatrix}$ of Σ . Define $x_2 = x - x_1$. Then $\begin{bmatrix} x_2 \\ w \end{bmatrix}$ is a stable future trajectory of Σ and $x = x_1 + x_2$. That (6.4) also holds with x replaced by x_2 follows from the fact that $\begin{bmatrix} x_2 \\ w \end{bmatrix}$ is a stable future trajectory of Σ . $\square$

6.2 The Input Map $\mathfrak{B}_\Sigma$

We now proceed to the construction of the input map $\mathfrak{B}_\Sigma$ of a passive s/s system Σ . This is an operator from the Hilbert space $\mathcal{H}_-$ defined in (5.9) into the state space $\mathcal{X}$. As in Theorem 5.4 we denote the quotient map $K_-^2(\mathcal{W}) \mapsto K_-^2(\mathcal{W})/\mathfrak{W}_-^{[\perp]}$ by Q_- , the image of $\mathfrak{W}_-$ under Q_- by $\mathcal{H}_-^0$, and the inverse image of $\mathcal{H}_-$ under Q_- by $\mathcal{K}(\mathfrak{W}_-)$.

Before giving the formal definition of the input map of a passive s/s system, let us explain the underlying idea. Let $\Sigma = (V; \mathcal{X}; \mathcal{W})$ be a passive s/s system with past behavior $\mathfrak{W}_-$. By Lemma 3.11, to each $w \in \mathfrak{W}_-$ there exists a unique stable externally generated past trajectory $\begin{bmatrix} x \\ w \end{bmatrix}$ of Σ . It is easy to see that the map $\mathfrak{B}_\Sigma$ from $w \in \mathfrak{W}_-$ to $x(0)$ is a linear operator. By Lemma 3.2, this operator is a contraction with respect to the semi-norm in $\mathfrak{W}_-$ inherited from $K_-^2(\mathcal{W})$. In particular, if $w \in \mathfrak{W}_-$ is neutral, i.e., if $[w, w]_{K_-^2(\mathcal{W})} = 0$, then $\mathfrak{B}_\Sigma w = 0$. After factoring out the maximal neutral subspace $\mathfrak{W}_0 := \mathfrak{W}_- \cap \mathfrak{W}_-^{[\perp]}$ from $\mathfrak{W}_-$, the space $\mathfrak{W}_-/\mathfrak{W}_0$ becomes a unitary space (the noncomplete version of a Hilbert space), and the operator $\mathfrak{B}_\Sigma$ becomes a contraction $\mathfrak{W}_-/\mathfrak{W}_0 \rightarrow \mathcal{X}$. It follows from Theorem 5.4(ii) that the space $\mathcal{H}_-$ has a natural interpretation as a completion of the unitary space $\mathfrak{W}_-/\mathfrak{W}_0$. Therefore, the contraction $\mathfrak{B}_\Sigma$ from $\mathfrak{W}_-/\mathfrak{W}_0$ to $\mathcal{X}$ has a natural extension to a contraction $\mathcal{H}_- \rightarrow \mathcal{X}$. In the following lemma this extension of $\mathfrak{B}_\Sigma$ has been denoted by $\tilde{\mathfrak{B}}_\Sigma$.

Lemma 6.9 *Let $\Sigma = (V; \mathcal{X}; \mathcal{W})$ be a passive s/s system with past behavior $\mathfrak{W}_-$. Then there exist a unique linear contraction $\mathfrak{B}_\Sigma: \mathcal{H}_- \rightarrow \mathcal{X}$ whose restriction to $\mathcal{H}_-^0$ is given by*

$$\mathfrak{B}_\Sigma Q_- w = x(0), \quad w \in \mathfrak{W}_-, \quad (6.5)$$

where $\begin{bmatrix} x \\ w \end{bmatrix}$ is the unique stable externally generated past trajectory of Σ whose signal part is w (cf. Lemma 3.11).

Proof Let $w \in \mathfrak{W}_-$, and let $\begin{bmatrix} x \\ w \end{bmatrix}$ be the externally generated stable past trajectory of Σ with signal part w . Then by (3.2) and (5.11)

$$\|x(0)\|_{\mathcal{X}}^2 \leq [w, w]_{K_-^2(\mathcal{W})} = \|Q_-w\|_{\mathcal{H}_-}^2.$$

This implies that the mapping $Q_-w \rightarrow x(0)$ is a linear contraction $\mathcal{H}_-^0 \rightarrow \mathcal{X}$. Since $\mathcal{H}_-^0$ is dense in $\mathcal{H}_-$, this mapping has a unique extension to a linear contraction $\mathfrak{B}_\Sigma: \mathcal{H}_- \rightarrow \mathcal{X}$. $\square$

Definition 6.10 The contraction $\mathfrak{B}_\Sigma$ in Lemma 6.9 is called the *input map* of Σ .

In the construction of our three canonical realizations in Sects. 8–10 we shall make crucial use of the following lemma.

Lemma 6.11 Let $\Sigma = (V; \mathcal{X}; \mathcal{W})$ be a passive s/s system with past behavior $\mathfrak{W}_-$, future behavior $\mathfrak{W}_+$, two-sided behavior $\mathfrak{W}$, input map $\mathfrak{B}_\Sigma$, and output map $\mathfrak{C}_\Sigma$.

- (i) A pair of functions $\begin{bmatrix} x \\ w \end{bmatrix}$ is an externally generated stable past trajectory of Σ if and only if

$$w \in \mathfrak{W}_- \quad \text{and} \quad x(t) = \mathfrak{B}_\Sigma Q_- \pi_- \tau^t w, \quad t \in \mathbb{R}^-. \quad (6.6)$$

- (ii) A pair of functions $\begin{bmatrix} x \\ w \end{bmatrix}$ is an externally generated stable two-sided trajectory of Σ if and only if

$$w \in \mathfrak{W} \quad \text{and} \quad x(t) = \mathfrak{B}_\Sigma Q_- \pi_- \tau^t w, \quad t \in \mathbb{R}. \quad (6.7)$$

In this case

$$\mathfrak{C}_\Sigma x(t) = Q_+ \pi_+ \tau^t w, \quad t \in \mathbb{R}. \quad (6.8)$$

- (iii) A pair of functions $\begin{bmatrix} x \\ w \end{bmatrix}$ is an externally generated stable future trajectory of Σ if and only if

$$w \in \mathfrak{W}_+ \quad \text{and} \quad x(t) = \mathfrak{B}_\Sigma Q_- \pi_- \tau^t w, \quad t \in \mathbb{R}^+. \quad (6.9)$$

In this case

$$\mathfrak{C}_\Sigma x(t) = Q_+ \pi_+ \tau^t w, \quad t \in \mathbb{R}^+. \quad (6.10)$$

Proof The proof of (i) is an easy modification of the proof of the first half of (ii), and (iii) is a special case of (ii), so let us only give the proof of (ii).

Let $\begin{bmatrix} x \\ w \end{bmatrix}$ be an externally generated stable two-sided trajectory of Σ . Then $w \in \mathfrak{W}$, and (6.5) implies that (6.7) holds with $t = 0$. By shifting the trajectory to the left or

right by the amount $|t|$ and applying (6.5) to the shifted trajectory we get (6.6) for all values of $t \in \mathbb{R}$.

Conversely, let $w \in \mathfrak{W}$. By Lemma 3.11, there exists a unique stable externally generated two-sided trajectory $\begin{bmatrix} x \\ w \end{bmatrix}$ of Σ , and by the first part of the proof, the function x is given by (6.7).

That also (6.8) holds follows from Lemma 6.4 and the fact that the restriction to $\mathbb{R}^+$ of any left- or right-shifted externally generated stable two-sided trajectory of Σ is a stable future trajectory of Σ . $\square$

Definition 6.12 Let $\Sigma = (V; \mathcal{X}, \mathcal{W})$ be a passive s/s system.

- (i) The *finite time reachable subspace* of Σ is the set

$$\left\{ x_0 \in \mathcal{X} \mid \begin{array}{l} x_0 = x(0) \text{ for some (stable) past} \\ \text{trajectory of } \Sigma \text{ with compact support} \end{array} \right\}.$$

- (ii) The *infinite time exactly reachable subspace* of Σ is the set

$$\left\{ x_0 \in \mathcal{X} \mid \begin{array}{l} x_0 = x(0) \text{ for some stable externally} \\ \text{generated past trajectory of } \Sigma \end{array} \right\}.$$

- (iii) The $\mathcal{H}_-$ -*exactly reachable subspace* of Σ is the range of the input map $\mathfrak{B}_\Sigma$ of Σ .
- (iv) Σ is *exactly reachable in one of the above senses* if the corresponding reachable subspace is all of $\mathcal{X}$.
- (v) The closure of the subspace in (i) is called the (*approximately*) *reachable subspace*.
- (vi) The system Σ is *controllable* if the approximately reachable subspace is all of $\mathcal{X}$.

Lemma 6.13 All the different types of exactly reachable subspaces in Definition 6.12 have the same closure, equal to the approximately reachable subspace.

Proof The three different types of exactly reachable subspaces defined in Definition 6.12 are (in the order that they appear) the range of the restriction of $\mathfrak{B}_\Sigma$ to the space $\mathcal{H}_0^0(\mathfrak{W}_-^{[1]})$ defined in Lemma 5.2, the range of the restriction of $\mathfrak{B}_\Sigma$ to the space $\mathcal{H}_-^0$, and the full range of $\mathfrak{B}_\Sigma$. That these three subspaces have the same closure follows from the fact that when one restricts the bounded linear operator $\mathfrak{B}_\Sigma$ to a dense subset of its domain, then the closure of its range remains the same. $\square$

Lemma 6.14 All the different types of exactly reachable subspaces in Definition 6.12 and also the approximately reachable subspace are strongly invariant in the sense that if $\begin{bmatrix} x \\ w \end{bmatrix}$ is a future trajectory of Σ whose initial state $x(0)$ belongs to one of these subspaces, then $x(t)$ stays in this subspace for all $t \in \mathbb{R}^+$.

Proof In the case of the finite time reachable subspace and the infinite time exactly reachable subspace the strong invariance follows from the general properties of trajectories (see Lemmas 3.3 and 3.4). In the case of the $\mathcal{H}_-$ -exactly reachable subspace the

strong invariance follows from Lemma 7.8 below. Finally, in the case of the approximately reachable subspace the strong invariance follows from the density and the strong invariance of the finitely reachable subspace. $\square$

Lemma 6.15 *If Σ is a passive energy preserving s/s system, then the input map $\mathfrak{B}_\Sigma$ of Σ is an isometry. If, in addition, Σ is controllable, then $\mathfrak{B}_\Sigma$ is unitary.*

Proof That $\mathfrak{B}_\Sigma$ is an isometry follows from the fact that we have equality in (3.2) whenever Σ is energy preserving (because then (1.9) holds as an equality). In particular, $\text{im}(\mathfrak{B}_\Sigma)$ is closed. If, in addition, Σ is controllable, then $\text{im}(\mathfrak{B}_\Sigma)$ is dense in $\mathcal{X}$, and hence equal to $\mathcal{X}$. $\square$

A partial converse to Lemma 6.15 is given in Corollary 9.9 below.

6.3 The Adjoints of $\mathfrak{C}_\Sigma$ and $\mathfrak{B}_\Sigma$

The rest of this section is devoted to the study of the adjoints of the input and output maps of a passive s/s system.

Lemma 6.16 *Let $\Sigma^\dagger = (V^\dagger; \mathcal{X}, \mathcal{W})$ be an anti-passive s/s system with past and future behaviors $\mathfrak{W}_-^\dagger$ and $\mathfrak{W}_+^\dagger$, respectively. Let $\mathfrak{W}_- := (\mathfrak{W}_-^\dagger)^{[\perp]}$ and $\mathfrak{W}_+ := (\mathfrak{W}_+^\dagger)^{[\perp]}$ be the corresponding passive past and future behaviors (see Lemma 4.11), and let Q_- and Q_+ be the quotient maps $K_-^2(\mathcal{W}) \rightarrow K_-^2(\mathcal{W})/\mathfrak{W}_-^{[\perp]}$ and $K_+^2(\mathcal{W}) \rightarrow K_+^2(\mathcal{W})/\mathfrak{W}_+^{[\perp]}$, respectively.*

- (i) *There exists a unique contraction $\mathfrak{B}_{\Sigma^\dagger}^\dagger: \mathcal{H}_+ \rightarrow \mathcal{X}$ such that $\begin{bmatrix} x^\dagger \\ w^\dagger \end{bmatrix}$ is an externally generated stable future trajectory of $\Sigma^\dagger$ if and only if*

$$w^\dagger \in \mathfrak{W}_+^\dagger \quad \text{and} \quad x^\dagger(t) = \mathfrak{B}_{\Sigma^\dagger}^\dagger Q_+ \tau^t w^\dagger, \quad t \in \mathbb{R}^+. \quad (6.11)$$

- (ii) *There exists a unique contraction $\mathfrak{C}_{\Sigma^\dagger}^\dagger: \mathcal{X} \rightarrow \mathcal{H}_-$ satisfying*

$$\mathfrak{C}_{\Sigma^\dagger}^\dagger x^\dagger(-t) = Q_- \tau_-^{*t} w^\dagger, \quad t \in \mathbb{R}^+, \quad (6.12)$$

for every stable past trajectory $\begin{bmatrix} x^\dagger \\ w^\dagger \end{bmatrix}$ of $\Sigma^\dagger$.

Proof This follows from Lemmas 6.11 and Lemma 6.4 and Remark 4.2. $\square$

Definition 6.17 The contractions $\mathfrak{B}_{\Sigma^\dagger}^\dagger$ and $\mathfrak{C}_{\Sigma^\dagger}^\dagger$ are called the backward input and backward output maps of $\Sigma^\dagger$, respectively.

Lemma 6.18 *Let $\Sigma = (V; \mathcal{X}, \mathcal{W})$ be a passive s/s system with input map $\mathfrak{B}_\Sigma$ and output map $\mathfrak{C}_\Sigma$, and let $\Sigma^{[\perp]}$ be the anti-passive dual of Σ , with the backward input map $\mathfrak{B}_{\Sigma^{[\perp]}}^\dagger$ and backward output map $\mathfrak{C}_{\Sigma^{[\perp]}}^\dagger$. Then $\mathfrak{B}_{\Sigma^{[\perp]}}^\dagger = \mathfrak{C}_\Sigma^*$ and $\mathfrak{C}_{\Sigma^{[\perp]}}^\dagger = \mathfrak{B}_\Sigma^*$.*

Proof The proof of this lemma is essentially the same as the proof of [5, Lemma 5.19] (see Remark 3.17). $\square$

Lemma 6.19 *If Σ is a co-energy preserving passive s/s system, then the output map $\mathfrak{C}_\Sigma$ of Σ is a co-isometry. If, in addition, Σ is observable, then $\mathfrak{C}_\Sigma$ is unitary.*

Proof The first claim follows from the fact that if Σ is co-energy preserving, then the anti-passive dual $\Sigma^{[\perp]}$ is energy preserving (in the backward time direction), and hence its input map $\mathfrak{B}_{\Sigma^{[\perp]}} = \mathfrak{C}_\Sigma^*$ is an isometry (see Lemma 6.15 and Remark 4.2). The second claim follows from the first claim since $\mathfrak{C}_\Sigma$ is injective iff Σ is observable. $\square$

A partial converse to Lemma 6.19 is given in Corollary 8.8 below.

6.4 The Backward Reachable and Unobservable Subspaces

Our definitions of the reachable and unobservable subspaces $\mathfrak{R}_\Sigma$ and $\mathfrak{U}_\Sigma$ have a built-in direction of time. If $\Sigma^\dagger = (V^\dagger; \mathcal{X}, \mathcal{W})$ is an anti-passive system, then we denote the backward counterparts of $\mathfrak{R}_\Sigma$ and $\mathfrak{U}_\Sigma$ by $\mathfrak{R}_{\Sigma^\dagger}^\dagger$ and $\mathfrak{U}_{\Sigma^\dagger}^\dagger$, respectively. Thus, $\mathfrak{R}_{\Sigma^\dagger}^\dagger$ is the closure in $\mathcal{X}$ of all states $x(t)$ that appear in backward externally generated past trajectories $\begin{bmatrix} x^\dagger \\ w^\dagger \end{bmatrix}$ of $\Sigma^\dagger$, and $\mathfrak{U}_{\Sigma^\dagger}^\dagger$ consists of all $x_0^\dagger \in \mathcal{X}$ with the property that there exists some future trajectory $\begin{bmatrix} x^\dagger \\ w^\dagger \end{bmatrix}$ of $\Sigma^\dagger$ which with $x^\dagger(0) = x_0^\dagger$ for which $w^\dagger$ vanishes identically. We call $\mathfrak{R}_{\Sigma^\dagger}^\dagger$ the (approximately) *backward reachable subspace* and $\mathfrak{U}_{\Sigma^\dagger}^\dagger$ the *backward unobservable subspace*.³ By an *backward unobservable trajectory* we mean a past trajectory $\begin{bmatrix} x^\dagger \\ w^\dagger \end{bmatrix}$ of $\Sigma^\dagger$ for which $w^\dagger$ vanishes identically.

If Σ is conservative, then it is both passive and anti-passive, and hence both the forward reachable and unobservable subspaces $\mathfrak{R}_\Sigma$ and $\mathfrak{U}_\Sigma$ as well as the backward reachable and unobservable subspaces $\mathfrak{R}_\Sigma^\dagger$ and $\mathfrak{U}_\Sigma^\dagger$ are well-defined. A two-sided trajectory $\begin{bmatrix} x \\ w \end{bmatrix}$ of a conservative system Σ whose signal part w vanishes identically will be called a *bilaterally unobservable trajectory*. The restriction of such a trajectory to $\mathbb{R}^+$ is unobservable, and the restriction to $\mathbb{R}^-$ is backward unobservable.

Definition 6.20 A conservative system is *simple* if it does not have any nontrivial bilaterally unobservable trajectories.

It follows from Lemmas 6.6 and 6.13 and their anti-passive counterparts combined with Lemma 6.18 that $\mathfrak{R}_{\Sigma^{[\perp]}}^\dagger = \mathfrak{U}_\Sigma^{[\perp]}$ and $\mathfrak{U}_{\Sigma^{[\perp]}}^\dagger = \mathfrak{R}_\Sigma^{[\perp]}$. Moreover, a conservative system $\Sigma = (V; \mathcal{X}, \mathcal{W})$ is simple if and only if $\mathfrak{U} \cap \mathfrak{U}^\dagger = \{0\}$, or equivalently, if and only if $\mathfrak{R} \vee \mathfrak{R}^\dagger = \mathcal{X}$.

7 The Past/Future Map of a Passive System

We here take a closer look at the past/future map $\Gamma_{\mathfrak{W}}$ introduced in Definition 5.8 in the case where $\mathfrak{W}$ is the two-sided behavior of a passive s/s system Σ .

³ In stochastic realization theory $\mathfrak{R}^\dagger$ is called the controllable subspace and $\mathfrak{U}^\dagger$ the *unconstructable subspace*.

Definition 7.1 In the case where $\mathfrak{W}$ is the two-sided behavior of a passive s/s system Σ , then we call the past/future map $\Gamma_{\mathfrak{W}}$ introduced in Definition 5.8 the *past/future map* of Σ , and denote it alternatively by Γ_{Σ} .

Lemma 7.2 The past/future map Γ_{Σ} of a passive s/s system $\Sigma = (V; \mathcal{X}, \mathcal{W})$ factors into the product

$$\Gamma_{\Sigma} = \mathfrak{C}_{\Sigma} \mathfrak{B}_{\Sigma} \quad (7.1)$$

of the input map $\mathfrak{B}_{\Sigma}$ and the output map $\mathfrak{C}_{\Sigma}$ of Σ . In particular, if $\Sigma_i, i = 1, 2$, are two passive s/s systems which have the same behaviors, and if we denote the input and output maps of Σ_i by $\mathfrak{B}_{\Sigma_i}$ and $\mathfrak{C}_{\Sigma_i}$, respectively, then $\mathfrak{C}_{\Sigma_1} \mathfrak{B}_{\Sigma_1} = \mathfrak{C}_{\Sigma_2} \mathfrak{B}_{\Sigma_2}$.

Proof Let $\begin{bmatrix} x \\ w \end{bmatrix}$ be an externally generated stable two-sided trajectory of Σ . Then the restriction of $\begin{bmatrix} x \\ w \end{bmatrix}$ to $\mathbb{R}^-$ is an externally generated stable past trajectory and the restriction of $\begin{bmatrix} x \\ w \end{bmatrix}$ to $\mathbb{R}^+$ is a stable future trajectory of Σ . Thus, by (6.6), $x(0) = \mathfrak{B}_{\Sigma} \pi_- w$ and by (6.4), $\mathfrak{C}_{\Sigma} x(0) = \pi_+ w + \mathfrak{W}_+$. Thus, the two operators Γ_{Σ} and $\mathfrak{C}_{\Sigma} \mathfrak{B}_{\Sigma}$ coincide on the dense subspace $\mathcal{H}_-^0$ of $\mathcal{H}_-$, and hence on all of $\mathcal{H}_-$. If the systems $\Sigma_i, i = 1, 2$ have the same two-sided behavior $\mathfrak{W}$, then they also have the same past/future map $\Gamma_{\mathfrak{W}}$. Thus $\mathfrak{C}_{\Sigma_1} \mathfrak{B}_{\Sigma_1} = \Gamma_{\mathfrak{W}} = \mathfrak{C}_{\Sigma_2} \mathfrak{B}_{\Sigma_2}$. $\square$

We next turn to the corresponding operator induced by an anti-passive two-sided behavior $\mathfrak{W}^{\dagger}$.

Lemma 7.3 Let $\mathfrak{W}^{\dagger}$ be an anti-passive two-sided behavior with the corresponding anti-passive past behavior $\mathfrak{W}_-^{\dagger} = \mathfrak{W}^{\dagger} \cap K_-^2(\mathcal{W})$ and future behavior $\mathfrak{W}_+^{\dagger} = \pi_+ \mathfrak{W}^{\dagger}$. Let $\mathfrak{W}_- := (\mathfrak{W}_-^{\dagger})^{[\perp]}$ and $\mathfrak{W}_+ := (\mathfrak{W}_+^{\dagger})^{[\perp]}$ be the passive past and future behaviors, respectively, induced by the passive two-sided behavior $\mathfrak{W} = (\mathfrak{W}^{\dagger})^{[\perp]}$. Then $\mathfrak{W}_- = \pi_- \mathfrak{W}$ and $\mathfrak{W}_+ = \mathfrak{W} \cap K_+^2(\mathcal{W})$.

Let Q_- and Q_+ be the quotient maps $K_-^2(\mathcal{W}) \rightarrow K_-^2(\mathcal{W})/\mathfrak{W}_-^{[\perp]}$ and $K_+^2(\mathcal{W}) \rightarrow K_+^2(\mathcal{W})/\mathfrak{W}_+^{[\perp]}$, respectively. Then there is a unique contraction $\Gamma_{\mathfrak{W}^{\dagger}}: \mathcal{H}_+ \rightarrow \mathcal{H}_-$ satisfying

$$\Gamma_{\mathfrak{W}^{\dagger}} Q_+ \pi_+ w^{\dagger} = Q_- \pi_- w^{\dagger}, \quad w^{\dagger} \in \mathfrak{W}^{\dagger}. \quad (7.2)$$

Proof The first claim follows from Lemmas 2.3, 3.18(iii), and 4.11, where we consider the orthogonal decomposition $K^2(\mathcal{W}) = K_-^2(\mathcal{W}) \boxplus K_+^2(\mathcal{W})$ in Lemma 2.3. The second claim follows from Remark 4.12 and Lemma 5.7. $\square$

Definition 7.4 The contraction $\Gamma_{\mathfrak{W}^{\dagger}}: \mathcal{H}_+ \rightarrow \mathcal{H}_-$ in Lemma 7.3 is called the *future/past map* of the anti-passive two-sided behavior $\mathfrak{W}^{\dagger}$. If $\mathfrak{W}^{\dagger}$ is the two-sided behavior of a passive anti-causal s/s system $\Sigma^{\dagger}$, then we also call $\Gamma_{\mathfrak{W}^{\dagger}}$ the future/past map of $\Sigma^{\dagger}$ and denote it by $\Gamma_{\Sigma^{\dagger}}$.

Lemma 7.5 The future/past map $\Gamma_{\Sigma^{\dagger}}$ of the anti-passive two-sided behavior $\mathfrak{W}^{\dagger}$ induced by an anti-passive reflected s/s system $\Sigma^{\dagger}$ factors into the product

$$\Gamma_{\Sigma^{\dagger}} = \mathfrak{C}_{\Sigma^{\dagger}}^{\dagger} \mathfrak{B}_{\Sigma^{\dagger}}^{\dagger} \quad (7.3)$$

of the backward input map $\mathfrak{B}_{\Sigma^{\dagger}}$ of $\Sigma^{\dagger}$ and the backward output map $\mathfrak{C}_{\Sigma^{\dagger}}^{\dagger}$ of $\Sigma^{\dagger}$.

Proof This follows from Lemma 7.2 and Remarks 4.2 and 4.12. $\square$

Lemma 7.6 *The adjoint of the past/future map $\Gamma_{\mathfrak{W}}$ of the two-sided behavior $\mathfrak{W}$ is the future/past map $\Gamma_{\mathfrak{W}^{[\perp]}}$ of the dual behavior $\mathfrak{W}^{[\perp]}$.*

Proof This follows from Lemmas 7.2, 6.18, and 7.5. $\square$

Lemma 7.7 *Let $\mathfrak{W}$ be a passive two-sided behavior with the corresponding passive past behavior $\mathfrak{W}_- = \pi_- \mathfrak{W}$ and passive future behavior $\mathfrak{W}_+ = \mathfrak{W} \cap K_+^2(\mathcal{W})$. Let $w \in K^2(\mathcal{W})$, and suppose that $\pi_- w \in \mathcal{K}(\mathfrak{W}_-^{[\perp]})$, $\pi_+ w \in \mathcal{K}(\mathfrak{W}_+)$, and that*

$$Q_+ \pi_+ w = \Gamma_{\mathfrak{W}} Q_- \pi_- w, \quad (7.4)$$

where $\Gamma_{\mathfrak{W}}$ is the past/future map of $\mathfrak{W}$.

(i) *For all $t \in \mathbb{R}^+$, $\pi_- \tau^t w \in \mathcal{K}(\mathfrak{W}_-^{[\perp]})$, $\pi_+ \tau^t w \in \mathcal{K}(\mathfrak{W}_+)$,*

$$Q_+ \pi_+ \tau^t w = \Gamma_{\mathfrak{W}} Q_- \pi_- \tau^t w, \quad t \in \mathbb{R}^+, \quad (7.5)$$

$$\begin{aligned} \|Q_- \pi_- \tau^t w\|_{\mathcal{H}_-}^2 &= \|Q_- \pi_- w\|_{\mathcal{H}_-}^2 + \int_0^t [w(s), w(s)]_{\mathcal{W}} ds, \\ t &\in \mathbb{R}^+. \end{aligned} \quad (7.6)$$

(ii) *There exists a sequence $w^k \in \mathfrak{W}$ such that*

$$Q_+ \pi_+ \tau^t w^k \rightarrow Q_+ \pi_+ \tau^t w \quad \text{in } \mathcal{H}_+, \quad t \in \mathbb{R}^+, \quad (7.7)$$

$$Q_- \pi_- \tau^t w^k \rightarrow Q_- \pi_- \tau^t w \quad \text{in } \mathcal{H}_-, \quad t \in \mathbb{R}^+, \quad (7.8)$$

$$\pi_+ w^k \rightarrow w_+ \quad \text{in } K_+^2(\mathcal{W}), \quad (7.9)$$

as $k \rightarrow \infty$, where the convergence in (7.7) and (7.8) is uniform in t .

Proof The proof is essentially the same as the proof of [5, Lemma 6.8] (see Remark 3.17). $\square$

Lemma 7.8 *Let $\Sigma = (V; \mathcal{X}, \mathcal{W})$ be a passive s/s system with input map $\mathfrak{B}_\Sigma$, past behavior $\mathfrak{W}_-$, future behavior $\mathfrak{W}_+$, and past/future map Γ_Σ . Then the following two conditions are equivalent:*

- (i) $\begin{bmatrix} x \\ w_+ \end{bmatrix}$ is a stable future trajectory of Σ satisfying $x(0) \in \text{im}(\mathfrak{B}_\Sigma)$;
- (ii) $w_+ \in \mathcal{K}(\mathfrak{W}_+)$, and there exists some $w_- \in \mathcal{K}(\mathfrak{W}_-^{[\perp]})$ such that

$$\begin{aligned} Q_+ w_+ &= \Gamma_\Sigma Q_- w_- \quad \text{and} \\ x(t) &= \mathfrak{B}_\Sigma Q_- \pi_- \tau^t (w_- + w_+), \quad t \in \mathbb{R}^+. \end{aligned} \quad (7.10)$$

The function w_- in (7.10) can be chosen to be any $w_- \in \mathcal{K}(\mathfrak{W}_-^{[\perp]})$ satisfying $x(0) = \mathfrak{B}_\Sigma Q_- w_-$.

Proof The proof of the implication (i) $\Rightarrow$ (ii) and the proof of the final claim are essentially the same as the first part of the proof of [5, Lemma 6.9] (see Remark 3.17).

Conversely, suppose that (ii) holds. Then, in particular, $x(0) = \mathfrak{B}_\Sigma Q_- w_- \in \text{im}(\mathfrak{B}_\Sigma)$ and $Q_+ w_+ = \Gamma_\Sigma Q_- w_-$. By Lemma 7.2, $Q_+ w_+ = \mathfrak{C}_\Sigma x_0 \in \text{im}(\mathfrak{C}_\Sigma)$. By (6.2), there exists a stable future trajectory $\begin{bmatrix} x_1 \\ w_+ \end{bmatrix}$ of Σ . By Lemma 6.4, $\mathfrak{C}_\Sigma x_1(t) = Q_+ \tau_+^t w_+$ for all $t \in \mathbb{R}^+$. On the other hand, by assumption, $x(t) = \mathfrak{B}_\Sigma Q_- \pi_- \tau^t(w_- + w_+)$ for all $t \in \mathbb{R}^+$, and hence by Lemma 7.7,

$$\mathfrak{C}_\Sigma x(t) = \mathfrak{C}_\Sigma \mathfrak{B}_\Sigma Q_- \pi_- \tau^t(w_- + w_+) = \Gamma_\Sigma Q_- \pi_- \tau^t(w_- + w_+) = Q_+ \pi_+ \tau^t w_+.$$

Thus, $\mathfrak{C}_\Sigma(x(0) - x_1(0)) = 0$. By Lemma 6.6, there exists an unobservable future trajectory $\begin{bmatrix} x_2 \\ 0 \end{bmatrix}$ of Σ with $x_2(0) = x(0) - x_1(0)$. Define $x_3 = x_1 + x_2$. Then $\begin{bmatrix} x_3 \\ w_+ \end{bmatrix}$ is a stable future trajectory of Σ , $x_3(0) = x(0) = \mathfrak{B}_\Sigma Q_- w_-$, and by the implication (i) $\Rightarrow$ (ii),

$$x_3(t) = \mathfrak{B}_\Sigma Q_- \pi_- \tau^t(w_- + w_+) = x(t), \quad t \in \mathbb{R}^+.$$

Thus, $\begin{bmatrix} x \\ w_+ \end{bmatrix}$ is a stable future trajectory of Σ . □

8 The Canonical Observable Co-Energy Preserving Model

In this section we shall construct a canonical model $\Sigma_{\text{oce}}^{\mathfrak{M}_+} = (V_{\text{oce}}^{\mathfrak{M}_+}; \mathcal{H}_+, \mathcal{W})$ of a passive observable co-energy preserving s/s system with a given passive future behavior $\mathfrak{M}_+$.

Theorem 8.1 *Let $\mathfrak{M}_+$ be a passive future behavior on the Kreĭn space $\mathcal{W}$ with the corresponding two-sided behavior $\mathfrak{M}$ defined by (3.19) and past behavior $\mathfrak{M}_- := \pi_- \mathfrak{M}$. Define $\mathcal{H}_+$ as in Theorem 5.4 and $\mathcal{K}(\mathfrak{M}_+)$ as in (5.12).*

(i) *Define*

$$\mathcal{T}_+ := \left\{ \begin{bmatrix} x \\ w_+ \end{bmatrix} \in \begin{bmatrix} C(\mathbb{R}^+; \mathcal{H}_+) \\ \mathcal{K}(\mathfrak{M}_+) \end{bmatrix} \mid x(t) = Q_+ \tau_+^t w_+, \quad t \in \mathbb{R}^+ \right\}. \quad (8.1)$$

Then $\mathcal{T}_+$ is the set of all stable future trajectories of a passive observable co-energy preserving s/s system $\Sigma_{\text{oce}}^{\mathfrak{M}_+}$ with state space $\mathcal{H}_+$ whose future behavior is equal to $\mathfrak{M}_+$ and two-sided behavior is equal to $\mathfrak{M}$.

- (ii) *The input map of $\Sigma_{\text{oce}}^{\mathfrak{M}_+}$ is the past/future map $\Gamma_{\mathfrak{M}}$ of $\mathfrak{M}$, and the output map of $\Sigma_{\text{oce}}^{\mathfrak{M}_+}$ is the identity on $\mathcal{H}_+$.*
- (iii) *A pair of functions $\begin{bmatrix} x \\ w_- \end{bmatrix}$ is an externally generated stable past trajectory of $\Sigma_{\text{oce}}^{\mathfrak{M}_+}$ if and only if*

$$w_- \in \mathfrak{M}_- \quad \text{and} \quad x(-t) = \Gamma_{\mathfrak{M}} Q_- \tau_-^{*t} w_-, \quad t \in \mathbb{R}^+. \quad (8.2)$$

Proof We define the Kreĭn spaces $\mathfrak{K}_{0,t}$ and $\mathfrak{L}_{0,\infty}$ as in the paragraph before Theorem 3.5 with $\mathcal{X}$ replaced by $\mathcal{H}_+$, and the subspaces $\mathcal{T}_{0,t}$ and $\mathcal{S}_{0,\infty}$ by (3.5) with $\mathcal{T}_+$ defined by (8.1).

Step 1: $\mathcal{T}_{0,t}$ is a maximal nonnegative subspace of $\mathfrak{K}_{0,t}$. That $\mathcal{T}_{0,t}$ is a nonnegative subspace of $\mathfrak{K}_{0,t}$ follows from Lemma 5.3. The maximality of $\mathcal{T}_{0,t}$ follows from Lemma 3.20 and Lemma 2.2(i) with $\mathcal{Y} = \mathcal{X} = \mathcal{H}_+$.

Step 2: $(\mathcal{T}_{0,t})^{[\perp]} \subset \mathcal{T}_{0,t}$. Fix $t \in \mathbb{R}^+$ and define $\mathring{\mathcal{T}}_{[0,t]}$ by

$$\mathring{\mathcal{T}}_{[0,t]} := \left\{ \left[\begin{array}{c} Q_+ \tau_+^t z^\dagger \\ Q_+ z^\dagger \\ \pi_{[0,t]} z^\dagger \end{array} \right] \middle| z^\dagger \in \mathfrak{W}_+^{[\perp]} \right\}. \quad (8.3)$$

Then $\mathring{\mathcal{T}}_{[0,t]} \subset \mathcal{T}_{0,t}$ since $\mathcal{H}_+^0 \subset \mathcal{H}_+$. We claim that $(\mathring{\mathcal{T}}_{[0,t]})^{[\perp]} = \mathcal{T}_{0,t}$. Clearly, this implies that $(\mathcal{T}_{0,t})^{[\perp]} \subset \mathcal{T}_{0,t}$ since $(\mathcal{T}_{0,t})^{[\perp]} = ((\mathring{\mathcal{T}}_{[0,t]})^{[\perp]})^{[\perp]}$ is the closure of $\mathring{\mathcal{T}}_{[0,t]}$.

A vector $k = \begin{bmatrix} x_1 \\ x_0 \\ w \end{bmatrix}$ belongs to $\mathring{\mathcal{T}}_{[0,t]}^{[\perp]}$ if and only if $x_1, x_0 \in \mathcal{H}_+$, $w \in K^2([0, t]; \mathcal{W})$, and

$$-(x_1, Q_+ \tau_+^t z^\dagger)_{\mathcal{H}_+} + (x_0, Q_+ z^\dagger)_{\mathcal{H}_+} + [w, z^\dagger]_{K^2([0,t]; \mathcal{W})} = 0, \quad z^\dagger \in \mathfrak{W}_+^{[\perp]}. \quad (8.4)$$

Since $\mathfrak{W}_+$ is τ_+^* -invariant, its orthogonal companion $\mathfrak{W}_+^{[\perp]}$ is τ_+ -invariant, i.e., $\tau_+^t z^\dagger \in \mathfrak{W}_+^{[\perp]}$ whenever $z^\dagger \in \mathfrak{W}_+^{[\perp]}$ and $t \in \mathbb{R}^+$. By (5.6), for every $v_1 \in x_1$ and $v_0 \in x_0$, (8.4) can therefore be rewritten in the form

$$[v_1, \tau_+^t z^\dagger]_{K_+^2(\mathcal{W})} - [v_0, z^\dagger]_{K_+^2(\mathcal{W})} + [w, z^\dagger]_{K^2([0,t]; \mathcal{W})} = 0, \quad z^\dagger \in \mathfrak{W}_+^{[\perp]}. \quad (8.5)$$

Extend w to a function in $K_+^2(\mathcal{W})$ by defining $w(s) = 0$ for $s > t$. Then (8.5) can be rewritten as

$$[\tau_+^{*t} v_1 - v_0 + w, z^\dagger]_{K_+^2(\mathcal{W})} = 0, \quad z^\dagger \in \mathfrak{W}_+^{[\perp]}.$$

Since $(\mathfrak{W}_+^{[\perp]})^{[\perp]} = \mathfrak{W}_+$, this is equivalent to

$$\tau_+^{*t} v_1 - v_0 + w = z$$

for some $z \in \mathfrak{W}_+$. Define $v = v_0 + z$. Then $v \in x_0$, and

$$\tau_+^{*t} v_1 - v + w = 0.$$

This is equivalent to the pair of equations

$$w = \pi_{[0,t]} v \quad \text{and} \quad v_1 = \tau_+^t v.$$

Thus, $\begin{bmatrix} x_1 \\ x_0 \\ w \end{bmatrix} \in (\mathring{\mathcal{T}}_{[0,t]})^{[\perp]}$ if and only if $x_0 = Q_+ v$, $x_1 = Q_+ \tau_+^t v$, and $w = \pi_{[0,t]} v$ for some $v \in \mathcal{K}(\mathfrak{W}_+)$, or equivalently, if and only if $k \in \mathcal{T}_{0,t}$.

Step 3: $\mathcal{T}_+$ is the set of stable future trajectories of a passive co-energy preserving s/s system $\Sigma_{\text{oce}}^{\mathfrak{W}_+}$. By (8.1) and Lemma 5.3, $\mathcal{T}_+$ is left-shift invariant, and by Steps 1 and 2, $\mathcal{T}_{0,t}$ is a maximal nonnegative co-neutral subspace of $\mathfrak{K}_{0,t}$ for all $t \in \mathbb{R}^+$. By (5.1), $\mathcal{S}_{0,\infty}$ is nonnegative in $\mathfrak{L}_{0,\infty}$, and by the definition of $\mathcal{K}(\mathfrak{W}_+)$, the maximal nonnegativity of $\mathfrak{W}_+$, and Lemma 2.2(i) with $\mathcal{Y} = \{0\}$ and $\mathcal{X} = \mathcal{H}_+$, $\mathcal{S}_{0,\infty}$ is maximal nonnegative in $\mathfrak{L}_{0,\infty}$. By Theorem 3.5, $\mathcal{T}_+$ is the set of stable trajectories of a passive s/s system, and by Lemma 4.9, this system is co-energy preserving.

Step 4: The future behavior of $\Sigma_{\text{oce}}^{\mathfrak{W}_+}$ is equal to $\mathfrak{W}_+$. This follows from the definition (8.1) of $\mathcal{T}_+$.

Step 5: $\mathfrak{B}_{\Sigma_{\text{oce}}^{\mathfrak{W}_+}} = \Gamma \mathfrak{W}$ and $\mathfrak{C}_{\Sigma_{\text{oce}}^{\mathfrak{W}_+}} = 1_{\mathcal{H}_+}$. By (8.1), if $\begin{bmatrix} x \\ w \end{bmatrix}$ is a stable future trajectory of $\Sigma_{\text{oce}}^{\mathfrak{W}_+}$, then $Q_+ w = x(0)$. On the other hand, by Lemma 6.4, $Q_+ w = \mathfrak{C}_{\Sigma_{\text{oce}}^{\mathfrak{W}_+}} x(0)$. Thus, $\mathfrak{C}_{\Sigma_{\text{oce}}^{\mathfrak{W}_+}} = 1_{\mathcal{H}_+}$. By Lemma 7.2, $\Gamma \mathfrak{W} = \mathfrak{C}_{\Sigma_{\text{oce}}^{\mathfrak{W}_+}} \mathfrak{B}_{\Sigma_{\text{oce}}^{\mathfrak{W}_+}} = \mathfrak{B}_{\Sigma_{\text{oce}}^{\mathfrak{W}_+}}$.

Step 6: $\Sigma_{\text{oce}}^{\mathfrak{W}_+}$ is observable. This follows from Step 5 and Lemma 6.6.

Step 7: (8.2) holds. This follows from Step 5 and Lemma 6.11. $\square$

Theorem 8.2 The generating subspace of the s/s system $\Sigma_{\text{oce}}^{\mathfrak{W}_+}$ in Theorem 8.1 is given by

$$V_{\text{oce}}^{\mathfrak{W}_+} = \left\{ \begin{bmatrix} Q_+ \dot{w}_+ \\ Q_+ w_+ \\ w_+(0) \end{bmatrix} \in \begin{bmatrix} \mathcal{H}_+ \\ \mathcal{H}_+ \\ \mathcal{W} \end{bmatrix} \left| \begin{array}{l} w_+ \in \mathcal{K}(\mathfrak{W}_+) \text{ is locally absolutely} \\ \text{continuous with } \dot{w} \in K_+^2(\mathcal{W}), \text{ and} \\ \lim_{t \rightarrow 0+} \frac{1}{t} Q_+(\tau_+^t w_+ - w_+) \text{ exists in } \mathcal{H}_+. \end{array} \right. \right\} \quad (8.6)$$

Proof Before proving (8.6) we first show that (8.6) is equivalent to the formula

$$V_{\text{oce}}^{\mathfrak{W}_+} = \left\{ \begin{bmatrix} \lim_{t \rightarrow 0+} \frac{1}{t} Q_+(\tau_+^t w_+ - w_+) \\ Q_+ w_+ \\ w_+(0) \end{bmatrix} \in \begin{bmatrix} \mathcal{H}_+ \\ \mathcal{H}_+ \\ \mathcal{W} \end{bmatrix} \left| \begin{array}{l} w_+ \in \mathcal{K}(\mathfrak{W}_+) \text{ is locally} \\ \text{absolutely continuous with} \\ \dot{w} \in K_+^2(\mathcal{W}), \text{ and} \\ \lim_{t \rightarrow 0+} \frac{1}{t} Q_+(\tau_+^t w_+ - w_+) \\ \text{exists in } \mathcal{H}_+. \end{array} \right. \right\} \quad (8.7)$$

Indeed, as $t \rightarrow 0+$, the function $\frac{1}{t}(\tau_+^t w_+ - w_+)$ tends to $\dot{w}_+$ in $K_+^2(\mathcal{W})$, and since $Q_+|_{\mathcal{K}(\mathfrak{W}_+)}$ is closed as an operator $K_+^2(\mathcal{W}) \rightarrow \mathcal{H}_+$, this implies that $\lim_{t \rightarrow 0+} \frac{1}{t} Q_+(\tau_+^t w_+ - w_+) = \lim_{t \rightarrow 0+} Q_+ \frac{1}{t}(\tau_+^t w_+ - w_+) = Q_+ \dot{w}_+$. Thus, (8.6) and (8.7) are equivalent.

Let $w_+ \in \mathcal{K}(\mathfrak{W}_+) \cap W^{1,2}(\mathbb{R}^+; \mathcal{W})$, and suppose that $\lim_{t \rightarrow 0+} \frac{1}{t} Q_+(\tau_+^t w_+ - w_+)$ exists in $\mathcal{H}_+$. Define $x(t) := Q_+ \tau_+^t w_+$ for $t \in \mathbb{R}^+$. Then $\begin{bmatrix} x \\ w_+ \end{bmatrix}$ is a generalized stable future trajectory of $\Sigma_{\text{oce}}^{\mathfrak{W}_+}$. By Proposition 3.8, this trajectory is even classical.

In particular, $\begin{bmatrix} \dot{x}(0) \\ x(0) \\ w_+(0) \end{bmatrix} \in V_{\text{oce}}^{\mathfrak{M}_+}$, where $\dot{x}(0) = \lim_{t \rightarrow 0+} \frac{1}{t} Q_+(\tau_+^t w_+ - w_+)$. Thus the right-hand side of (8.7) is contained in $V_{\text{oce}}^{\mathfrak{M}_+}$. The opposite inclusion follows from Proposition 3.7(ii). $\square$

Corollary 8.3 *The system $\Sigma_{\text{oce}}^{\mathfrak{M}_+}$ is approximately null-controllable in the sense that the set of all the initial states $x(0)$ of all future trajectories of $\Sigma_{\text{oce}}^{\mathfrak{M}_+}$ with compact support is dense in $\mathcal{X}_{\text{oce}} = \mathcal{H}_+$.*

Proof This follows from Theorem 8.1 and Lemma 5.2, since $\tau_+^t w_+ = 0$ for all sufficiently large t whenever $w_+ \in \mathcal{H}_0^0(\mathfrak{M}_+)$. $\square$

Theorem 8.4 *Let $\Sigma = (V; \mathcal{X}, \mathcal{W})$ be a passive s/s system with output map $\mathfrak{C}_\Sigma$ and future behavior $\mathfrak{M}_+$. Then Σ and $\Sigma_{\text{oce}}^{\mathfrak{M}_+} = (V_{\text{oce}}^{\mathfrak{M}_+}; \mathcal{H}_+, \mathcal{W})$ are contractively intertwined by $\mathfrak{C}_\Sigma$.*

Proof Let $\begin{bmatrix} x \\ w \end{bmatrix}$ be a stable future trajectory of Σ . By Lemmas 6.1 and 6.4, $w \in \mathcal{K}(\mathfrak{M}_+)$ and $\mathfrak{C}_\Sigma x(t) = Q_+ \tau_+^t w$, $t \in \mathbb{R}^+$. Define $x_o(t) = Q_+ \tau_+^t w$, $t \in \mathbb{R}^+$. Then $\begin{bmatrix} x_o \\ w \end{bmatrix}$ is a stable future trajectory of $\Sigma_{\text{oce}}^{\mathfrak{M}_+}$, and $x_o(t) = \mathfrak{C}_\Sigma x(t)$, $t \in \mathbb{R}^+$.

Conversely, let $\begin{bmatrix} x_o \\ w \end{bmatrix}$ be a stable future trajectory of $\Sigma_{\text{oce}}^{\mathfrak{M}_+}$ satisfying $x_o(0) \in \text{im}(\mathfrak{C}_\Sigma)$. Then $Q_+ w = x_o(0)$, and hence $w \in Q_+^{-1} \text{im}(\mathfrak{C}_\Sigma) = \mathfrak{S}_+^\Sigma$. By (6.2), there exists a stable future trajectory $\begin{bmatrix} x \\ w \end{bmatrix}$ of Σ (whose signal part is equal to the given signal w). By Lemma 6.4, $\mathfrak{C}_\Sigma x(t) = Q_+ \tau_+^t w = x_o(t)$, $t \in \mathbb{R}^+$. $\square$

Theorem 8.5 *Every observable and co-energy preserving passive s/s system Σ with future behavior $\mathfrak{M}_+$ is unitarily similar to the system $\Sigma_{\text{oce}}^{\mathfrak{M}_+}$. The unitary similarity transformation is the output map $\mathfrak{C}_\Sigma$ of Σ .*

Proof By Lemma 6.19, the output map $\mathfrak{C}_\Sigma$ is unitary, and by Theorem 8.4, $\mathfrak{C}_\Sigma$ intertwines Σ and $\Sigma_{\text{oce}}^{\mathfrak{M}_+}$. $\square$

Definition 8.6 We call the system $\Sigma_{\text{oce}}^{\mathfrak{M}_+}$ the *canonical model* of an observable passive co-energy preserving s/s system with future behavior $\mathfrak{M}_+$.

Corollary 8.7 *Any two observable and co-energy preserving realizations of a given passive future behavior $\mathfrak{M}_+$ are unitarily similar to each other.*

Proof This follows from Theorem 8.5. $\square$

Corollary 8.8 *A passive s/s system Σ is observable and co-energy preserving if and only if its output map $\mathfrak{C}_\Sigma$ is unitary.*

Proof This follows from Lemma 6.19 and Theorem 8.4. $\square$

9 The Canonical Controllable Energy Preserving Model

In this section we shall construct a canonical model $\Sigma_{\text{cep}}^{\mathfrak{M}_-} = (V_{\text{cep}}^{\mathfrak{M}_-}; \mathcal{X}_{\text{cep}}^{\mathfrak{M}_-}, \mathcal{W})$ of a passive controllable energy preserving s/s system with a given passive past behavior $\mathfrak{M}_-$. The results for this model are analogous to the results on the model $\Sigma_{\text{oce}}^{\mathfrak{M}_+}$ obtained in the preceding section. The state space of $\Sigma_{\text{cep}}^{\mathfrak{M}_-}$ is the Hilbert space $\mathcal{H}(\mathfrak{M}_-^{[\perp]})$ presented in Theorem 5.4.

Theorem 9.1 *Let $\mathfrak{M}_-$ be a passive past behavior on the Kreĭn space $\mathcal{W}$ with the corresponding two-sided behavior $\mathfrak{M}$ defined by (3.18) and future behavior $\mathfrak{M}_+ := \mathfrak{M} \cap K_+^2(\mathcal{W})$. Define $\mathcal{H}_-$ as in Theorem 5.4 and define $\mathcal{K}(\mathfrak{M}_+)$ and $\mathcal{K}(\mathfrak{M}_-^{[\perp]})$ as in (5.5) and (5.12), respectively.*

(i) *Define*

$$\mathcal{T}_+ := \left\{ \begin{bmatrix} x \\ w_+ \end{bmatrix} \in \begin{bmatrix} C(\mathbb{R}^+; \mathcal{H}_-) \\ \mathcal{K}(\mathfrak{M}_+) \end{bmatrix} \left| \begin{array}{l} Q_+ w_+ = \Gamma_{\mathfrak{M}} Q_- w_- \text{ and} \\ x(t) = Q_- \pi_- \tau^t(w_- + w_+) \\ \text{for some } w_- \in \mathcal{K}(\mathfrak{M}_-^{[\perp]}) \text{ satisfying} \\ x(0) = Q_- w_-, \end{array} \right. \right\} \quad (9.1)$$

Then $\mathcal{T}_+$ is the set of all stable future trajectories of a passive controllable energy preserving s/s system $\Sigma_{\text{cep}}^{\mathfrak{M}_-}$ with state space $\mathcal{H}_-$ whose past behavior is equal to $\mathfrak{M}_-$.

- (ii) *The input map of $\Sigma_{\text{cep}}^{\mathfrak{M}_-}$ is the identity on $\mathcal{H}_-$ and the output map of $\Sigma_{\text{cep}}^{\mathfrak{M}_-}$ is the past/future map $\Gamma_{\mathfrak{M}}$ of $\mathfrak{M}$.*
 (iii) *A pair of functions $\begin{bmatrix} x \\ w \end{bmatrix}$ is an externally generated stable past trajectory of $\Sigma_{\text{cep}}^{\mathfrak{M}_-}$ if and only if*

$$w \in \mathfrak{M}_- \quad \text{and} \quad x(-t) = Q_- \tau_-^{*t} w, \quad t \geq 0. \quad (9.2)$$

Proof We define the Kreĭn spaces $\mathfrak{K}_{0,t}$ and $\mathfrak{L}_{0,\infty}$ as in the paragraph before Theorem 3.5 with $\mathcal{X}$ replaced by $\mathcal{H}_-$, and the subspaces $\mathcal{T}_{0,t}$ and $\mathcal{S}_{0,\infty}$ by (3.5) with $\mathcal{T}_+$ defined by (9.1).

Step 1: $\mathcal{T}_{0,t}$ is a maximal nonnegative neutral subspace of $\mathfrak{K}_{0,t}$. That $\mathcal{T}_{0,t}$ is a neutral subspace of $\mathfrak{K}_{0,t}$ follows from (7.6). It follows from (9.1) that to every $x_0 \in \mathcal{H}_-$ there exists some $\begin{bmatrix} x \\ w \end{bmatrix} \in \mathcal{T}_+$ such that $x(0) = x_0$. Moreover, if $\begin{bmatrix} x \\ w \end{bmatrix} \in \mathcal{T}_+$ with $x(0) = 0$, then $w \in \mathfrak{M}_+$. These two facts together with Lemma 2.2 with $\mathcal{Y} = \mathcal{X} = \mathcal{H}_-$ imply that $\mathcal{T}_{0,t}$ is maximal nonnegative.

Step 2: $\mathcal{T}_+$ is the set of stable future trajectories of a passive energy preserving s/s system $\Sigma_{\text{cep}}^{\mathfrak{M}_-}$. By (9.1) and Lemma 7.7, $\mathcal{T}_+$ is left-shift invariant, and by Step 1, each $\mathcal{T}_{0,t}$ is a maximal nonnegative and neutral subspace of $\mathfrak{K}_{0,t}$. By letting $t \rightarrow \infty$ and using the nonnegativity of $\mathcal{T}_{0,t}$ we find that $\mathcal{S}_{0,\infty}$ is nonnegative, and the maximality of $\mathcal{S}_{0,\infty}$ is proved in the same way as the maximality of $\mathcal{T}_{0,t}$. By Theorem 3.5, $\mathcal{T}_+$ is the set of stable trajectories of a passive s/s system, and by Lemma 4.9, this system is energy preserving.

Step 3: The past, two-sided, and future behaviors of $\Sigma_{\text{cep}}^{\mathfrak{M}_-}$ are equal to $\mathfrak{M}_-$, $\mathfrak{M}$, and $\mathfrak{M}_+$, respectively. That the future behavior of $\Sigma_{\text{cep}}^{\mathfrak{M}_-}$ is equal to $\mathfrak{M}_+$ follows from the definition (9.1) of $\mathcal{T}_+$, and the remaining claims then follow from Proposition 3.15.

Step 4: $\mathfrak{B}_{\Sigma_{\text{cep}}^{\mathfrak{M}_-}} = 1_{\mathcal{H}_-}$ and $\mathfrak{C}_{\Sigma_{\text{cep}}^{\mathfrak{M}_-}} = \Gamma_{\mathfrak{M}}$. Take $w_- = 0$ and $w_+ \in \mathfrak{M}_+$ in (9.1), define $w = w_- + w_+$, and define $x(t) = Q_- \pi_- \tau^t w$ for $t \in \mathbb{R}$. Then by Step 2 and Lemma 3.9(ii), $\begin{bmatrix} x \\ w \end{bmatrix}$ is a two-sided trajectory of $\Sigma_{\text{cep}}^{\mathfrak{M}_-}$ supported on $\mathbb{R}^+$. If we left-translate this trajectory by the amount $s > 0$, then we get another stable two-sided trajectory $\begin{bmatrix} x_s \\ w_s \end{bmatrix} := \begin{bmatrix} \tau^s x \\ \tau^s w \end{bmatrix}$ of $\Sigma_{\text{cep}}^{\mathfrak{M}_-}$ supported on $[-s, \infty)$. This trajectory satisfies $(x_s)(0) = Q_- \pi_- w_s$. On the other hand, by Lemma 6.9, $x_s(0) = \mathfrak{B}_{\Sigma_{\text{cep}}^{\mathfrak{M}_-}} Q_- \pi_- w_s$. Thus, $\mathfrak{B}_{\Sigma_{\text{cep}}^{\mathfrak{M}_-}} Q_- \pi_- w_s = Q_- \pi_- w_s$. By varying w_+ and s we can in this way generate all the stable two-sided trajectories of $\Sigma_{\text{cep}}^{\mathfrak{M}_-}$ whose support are bounded to the left, and consequently, the restriction of $\mathfrak{B}_{\Sigma_{\text{cep}}^{\mathfrak{M}_-}}$ to the space $\mathcal{H}_0^0(\mathfrak{M}_-^{[\perp]})$ defined in Lemma 5.5 is the identity. By Lemma 5.5, $\mathcal{H}_0^0(\mathfrak{M}_-^{[\perp]})$ is dense in $\mathcal{H}_-$, and thus $\mathfrak{B}_{\Sigma_{\text{cep}}^{\mathfrak{M}_-}} = 1_{\mathcal{H}_-}$. By Lemma 7.2, $\Gamma_{\mathfrak{M}} = \mathfrak{C}_{\Sigma_{\text{cep}}^{\mathfrak{M}_-}} \mathfrak{B}_{\Sigma_{\text{cep}}^{\mathfrak{M}_-}} = \mathfrak{C}_{\Sigma_{\text{cep}}^{\mathfrak{M}_-}}$.

Step 5: $\Sigma_{\text{cep}}^{\mathfrak{M}_-}$ is controllable. This follows from Step 4 and Lemma 6.13.

Step 6: A pair of functions $\begin{bmatrix} x \\ w \end{bmatrix}$ is an externally generated stable past trajectory of $\Sigma_{\text{cep}}^{\mathfrak{M}_-}$ if and only if (9.2) holds. This follows from Step 4 and Lemma 6.11(i). $\square$

Corollary 9.2 *Every stable past trajectory $\begin{bmatrix} x \\ w \end{bmatrix}$ of $\Sigma_{\text{cep}}^{\mathfrak{M}_-}$ (not necessarily externally generated) satisfies*

$$w \in \mathcal{K}(\mathfrak{M}_-^{[\perp]}) \quad \text{and} \quad x(-t) = Q_- \tau_-^{*t} w, \quad t \geq 0. \quad (9.3)$$

Proof Since $\Sigma_{\text{cep}}^{\mathfrak{M}_-}$ is energy preserving, it follows from Lemma 4.8 that every stable past trajectory $\begin{bmatrix} x \\ w \end{bmatrix}$ of $\Sigma_{\text{cep}}^{\mathfrak{M}_-}$ is also a stable past trajectory of the anti-passive dual $\Sigma^{\dagger}$ of $\Sigma_{\text{cep}}^{\mathfrak{M}_-}$. By applying the reflected version of Theorem 8.1 to the system $\Sigma^{\dagger}$ we find that (9.3) holds. $\square$

Corollary 9.3 *The system $\Sigma_{\text{cep}}^{\mathfrak{M}_-}$ is both $\mathcal{H}_-$ -exactly controllable and observable in backward time in the sense that if the signal part w of a stable past trajectory $\begin{bmatrix} x \\ w \end{bmatrix}$ of $\Sigma_{\text{cep}}^{\mathfrak{M}_-}$ is zero, then also the state part x is zero.*

Proof The first claim follows from the fact that the input map is the identity, and the second claim follows from (9.3). $\square$

Lemma 9.4 *Let $\mathfrak{M}_-$ be a passive past behavior on the Kreĭn space $\mathcal{W}$. Then the generating subspace $V_{\text{cep}}^{\mathfrak{M}_-}$ of the controllable and energy preserving s/s system $\Sigma_{\text{cep}}^{\mathfrak{M}_-}$ in Theorem 9.1 is a closed subspace of the subspace*

$$(V_{\text{cep}}^{\mathfrak{M}_-})^{\perp} = \left\{ \begin{bmatrix} Q_- \dot{w}_- \\ Q_- w_- \\ w_-(0) \end{bmatrix} \in \begin{bmatrix} \mathcal{H}_- \\ \mathcal{H}_- \\ \mathcal{W} \end{bmatrix} \left| \begin{array}{l} w_- \in K(\mathfrak{M}_-^{\perp}) \text{ is locally absolutely} \\ \text{continuous with } \dot{w} \in K_-^2(\mathcal{W}) \text{ and} \\ \lim_{t \rightarrow 0+} \frac{1}{t} Q_- (\tau_-^{*t} w - w) \text{ exists in } \mathcal{H}_-. \end{array} \right. \right\}. \quad (9.4)$$

Proof The above subspace is the generating subspace of the anti-causal dual of $\Sigma_{\text{cep}}^{\mathfrak{M}_-}$. That system is a co-energy preserving anti-passive s/s realization of the anti-passive past behavior $\mathfrak{M}_-^{\perp}$, and its generating subspace is obtained from (8.6) through a time reflection and the replacement of $\mathfrak{M}_+$ by $\mathfrak{M}_-^{\perp}$. Since $\Sigma_{\text{cep}}^{\mathfrak{M}_-}$ is energy-preserving, $V_{\text{cep}}^{\mathfrak{M}_-}$ is a closed subspace of its orthogonal companion. $\square$

The exact description of $V_{\text{cep}}^{\mathfrak{M}_-}$ will be given in Theorem 10.9 below.

Theorem 9.5 *Let $\Sigma = (V; \mathcal{X}, \mathcal{W})$ be a passive s/s system with input map $\mathfrak{B}_\Sigma$ and past behavior $\mathfrak{M}_-$. Then $\Sigma_{\text{cep}}^{\mathfrak{M}_-}$ and Σ are intertwined by $\mathfrak{B}_\Sigma$.*

Proof This follows from Lemma 7.8 and Theorem 9.1. $\square$

Theorem 9.6 *Every controllable and energy preserving passive s/s system Σ with past behavior $\mathfrak{M}_-$ is unitarily similar to the system $\Sigma_{\text{cep}}^{\mathfrak{M}_-}$. The unitary similarity transformation is the inverse of the input map $\mathfrak{B}_\Sigma$ of Σ .*

Proof By Lemma 6.15, the input map $\mathfrak{B}_\Sigma$ is unitary, and by Theorem 9.5, $\mathfrak{B}_\Sigma^{-1}$ intertwines Σ and $\Sigma_{\text{cep}}^{\mathfrak{M}_-}$. $\square$

Definition 9.7 We call the system $\Sigma_{\text{cep}}^{\mathfrak{M}_-}$ the *canonical model* of a passive controllable energy preserving s/s system with past behavior $\mathfrak{M}_-$.

Corollary 9.8 *Any two controllable and energy preserving realizations of a given passive past behavior $\mathfrak{M}_-$ are unitarily similar to each other.*

Proof This follows from Theorem 9.6. $\square$

Corollary 9.9 *A passive s/s system Σ is controllable and energy preserving if and only if its input map $\mathfrak{B}_\Sigma$ is unitary.*

Proof This follows from Lemma 6.15 and Theorem 9.5. $\square$

Theorem 9.10 *The operator $\Gamma_{\mathfrak{M}}$ intertwines the two s/s systems $\Sigma_{\text{cep}}^{\mathfrak{M}_-}$ and $\Sigma_{\text{oce}}^{\mathfrak{M}_+}$.*

Proof This follows from Theorem 8.4 and also from Theorem 9.5. $\square$

10 The Canonical Simple Conservative Model

We finally develop a canonical model for a conservative simple state/signal system with a given passive two-sided behavior $\mathfrak{W}$ (see Definition 6.20 for the notion of a simplicity of a conservative system).

Theorem 10.1 *Let $\mathfrak{W}$ be a passive two-sided behavior on the Kreĭn space $\mathcal{W}$, and let $\mathfrak{W}_- = \pi_- \mathfrak{W}$ and $\mathfrak{W}_+ = \mathfrak{W} \cap K_+^2(\mathcal{W})$ be the corresponding passive past and future behaviors. Let $\mathcal{D}(\mathfrak{W})$ be the range space of the operator $A_{\mathfrak{W}}^{1/2}$, where $A_{\mathfrak{W}}$ is the nonnegative self-adjoint operator on $\mathcal{H}_+ \oplus \mathcal{H}_-$ defined by (5.18), and define $\mathcal{L}(\mathfrak{W})$ by (5.20).*

(i) *Define*

$$\mathcal{T} := \left\{ \begin{bmatrix} x \\ w \end{bmatrix} \in \begin{bmatrix} BUC(\mathbb{R}; \mathcal{X}) \\ \mathcal{L}(\mathfrak{W}) \end{bmatrix} \mid x(t) = Q\tau^t w, t \in \mathbb{R} \right\}. \quad (10.1)$$

Then $\mathcal{T}$ is the set of all stable two-sided trajectories of a simple conservative s/s system $\Sigma_{sc}^{\mathfrak{W}}$ with state space $\mathcal{D}(\mathfrak{W})$ whose two-sided behavior is equal to $\mathfrak{W}$.

(ii) *The input map of $\Sigma_{sc}^{\mathfrak{W}}$ is $\mathfrak{B}_{\Sigma_{sc}^{\mathfrak{W}}} = \begin{bmatrix} \Gamma_{\mathfrak{W}} \\ 1_{\mathcal{H}_-} \end{bmatrix}$ with $(\mathfrak{B}_{\Sigma_{sc}^{\mathfrak{W}}})^* = \Pi_-|_{\mathcal{D}(\mathfrak{W})}$, the output map of $\Sigma_{sc}^{\mathfrak{W}}$ is $\mathfrak{C}_{\Sigma_{sc}^{\mathfrak{W}}} = \Pi_+|_{\mathcal{D}(\mathfrak{W})}$ with $\mathfrak{C}_{\Sigma_{sc}^{\mathfrak{W}}}^* = \begin{bmatrix} 1_{\mathcal{H}_+} \\ \Gamma_{\mathfrak{W}}^* \end{bmatrix}$.*

Proof We define the Kreĭn spaces $\mathfrak{K}_{0,t}$ and $\mathfrak{L}_{0,\infty}$ as in the paragraph before Theorem 3.5 with $\mathcal{X}$ replaced by $\mathcal{H}_+$, and the subspaces $\mathcal{T}_{0,t}$ and $\mathcal{S}_{0,\infty}$ by (3.5), with $\mathcal{T}_+ := \pi_+ \mathcal{T}$, with $\mathcal{T}$ defined by (10.1).

Step 1: $\mathcal{T}_{0,t}$ is a Lagrangian subspace of $\mathfrak{K}_{0,t}$. By (5.26), $\mathcal{T}_{0,t}$ is a neutral subspace of $\mathfrak{K}_{0,t}$. To prove that $\mathcal{T}_{0,t}$ is a Lagrangian we shall use Lemma 2.2 with $\mathcal{Y} = \mathcal{X} = \mathcal{D}(\mathfrak{W})$. Clearly condition (a) in that lemma holds because of the definition of $\mathcal{L}(\mathfrak{W})$, and (c) holds because of Lemma 5.14. The set described in condition (b) is equal to the section $\mathfrak{W}_{[0,t]} = \pi_{[0,t]} \mathfrak{W}_+$, which according to Lemma 3.20 is maximal nonnegative, and the set described in condition (d) is equal to $\tau_+^{*t} \pi_{[-t,0]} \mathfrak{W}_-$, which according to Lemma 3.20 and Remark 4.2 is maximal nonpositive.

Step 2: $\mathcal{S}_{0,\infty}$ is a maximal nonnegative subspace of $\mathfrak{L}_{0,\infty}$. By dropping the term $\|Q\tau^t w\|_{\mathcal{D}(\mathfrak{W})}^2$ in (5.26) and letting $t \rightarrow \infty$ we find that $\mathcal{S}_{0,\infty}$ is a nonnegative subspace of $\mathfrak{L}_{0,\infty}$. The proof of the maximality of $\mathcal{S}_{0,\infty}$ is analogous to (but simpler than) the proof of the maximality of $\mathcal{T}_{0,t}$ given in Step 1.

Step 3: $\mathcal{T}_+ := \pi_+ \mathcal{T}$ is the set of stable future trajectories of a conservative s/s system $\Sigma_{sc}^{\mathfrak{W}}$. By (10.1) and Lemma 5.14, $\mathcal{T}$ is shift invariant in $\begin{bmatrix} BUC(\mathbb{R}; \mathcal{X}) \\ K^2(\mathcal{W}) \end{bmatrix}$. In particular, $\mathcal{T}_+ := \pi_+ \mathcal{T}$ is then left-shift invariant. By Steps 1 and 2, $\mathcal{T}_{0,t}$ is a Lagrangian subspace of $\mathfrak{K}_{0,t}$ for all $t \in \mathbb{R}^+$ and $\mathcal{S}_{0,\infty}$ is a maximal nonnegative subspace of $\mathfrak{L}_{0,\infty}$. By Theorem 3.5, $\mathcal{T}_+$ is the set of stable trajectories of a passive s/s system, and by Lemma 4.9, this system is conservative.

Step 4: $\mathcal{T}$ is the set of stable two-sided trajectories of $\Sigma_{sc}^{\mathfrak{W}}$. This follows from Step 3 and Remark 4.2.

Step 5: The behavior of $\Sigma_{sc}^{\mathfrak{W}}$ is equal to $\mathfrak{W}$. If $w \in \mathfrak{W}_+$, then $Qw \in \mathcal{D}(\mathfrak{W})$, and it follows from (10.1) that $\begin{bmatrix} x \\ w \end{bmatrix}$, where $x(t) = Q\tau^t w, t \in \mathbb{R}^+$, is an externally generated

stable future trajectory of $\Sigma_{\text{sc}}^{\mathfrak{W}}$. This implies that $\mathfrak{W}_+ \subset \mathfrak{W}_+^{\Sigma_{\text{sc}}^{\mathfrak{W}}}$. Since $\mathfrak{W}_+$ is maximal nonnegative and $\mathfrak{W}_+^{\Sigma_{\text{sc}}^{\mathfrak{W}}}$ is nonnegative, this implies that $\mathfrak{W}_+ = \mathfrak{W}_+^{\Sigma_{\text{sc}}^{\mathfrak{W}}}$. From this follows that also $\mathfrak{W}_- = \mathfrak{W}_-^{\Sigma_{\text{sc}}^{\mathfrak{W}}}$ and $\mathfrak{W} = \mathfrak{W}^{\Sigma_{\text{sc}}^{\mathfrak{W}}}$.

Step 6: The input map of $\Sigma_{\text{sc}}^{\mathfrak{W}}$ is $\begin{bmatrix} \Gamma_{\mathfrak{W}} \\ 1_{\mathcal{H}_-} \end{bmatrix}$. According to Lemma 6.9, the operator $\mathfrak{B}_{\Sigma_{\text{sc}}^{\mathfrak{W}}}$ is the unique operator $\mathcal{H}_- \rightarrow \mathcal{D}(\mathfrak{W})$ which satisfies $\mathfrak{B}_{\Sigma_{\text{sc}}^{\mathfrak{W}}} Q_- \pi_- w = x(0)$ for every $w \in \mathfrak{W}$, where x is the state component of the unique externally generated trajectory $\begin{bmatrix} x \\ w \end{bmatrix}$ whose signal part is w . Let $w \in \mathfrak{W}$. By (10.1),

$$x(0) = Qw = \begin{bmatrix} Q_+ \pi_+ w \\ Q_- \pi_- w \end{bmatrix} = \begin{bmatrix} \Gamma_{\mathfrak{W}} \\ 1_{\mathcal{H}_-} \end{bmatrix} Q_- \pi_- w.$$

Thus, $\mathfrak{B}_{\Sigma_{\text{sc}}^{\mathfrak{W}}} = \begin{bmatrix} \Gamma_{\mathfrak{W}} \\ 1_{\mathcal{H}_-} \end{bmatrix}$.

Step 7: The output map of Σ is $\Pi_+|_{\mathcal{D}(\mathfrak{W})}$. According to Lemma 6.2, $\mathfrak{C}_{\Sigma_{\text{sc}}^{\mathfrak{W}}}$ is the operator which maps $x_0 \in \mathcal{D}(\mathfrak{W})$ into the equivalence class consisting of all the signal parts w of all stable future trajectories $\begin{bmatrix} x \\ w \end{bmatrix}$ of $\Sigma_{\text{sc}}^{\mathfrak{W}}$ satisfying $x(0) = x_0$. Let $x_0 \in \mathcal{D}(\mathfrak{W})$, and choose some $w_0 \in \mathcal{L}(\mathfrak{W})$ such that $Qw_0 = x_0$. It follows from (10.1) that $\begin{bmatrix} x \\ w_0 \end{bmatrix}$, where $x(t) = Q\tau^t w_0$, $t \in \mathbb{R}^+$, is a stable future trajectory of $\Sigma_{\text{sc}}^{\mathfrak{W}}$ satisfying $x(0) = x_0$. If $\begin{bmatrix} x_1 \\ w_1 \end{bmatrix}$ is another stable future trajectory of $\Sigma_{\text{sc}}^{\mathfrak{W}}$ satisfying $x_1(0) = x(0) = x_0$, then $\begin{bmatrix} x-x_1 \\ w_0-w_1 \end{bmatrix}$ is an externally generated stable future trajectory of $\Sigma_{\text{sc}}^{\mathfrak{W}}$, and hence $w_1 - w_0 \in \mathfrak{W}_+$. Thus, the equivalence class of all the signal parts w of all stable future trajectories $\begin{bmatrix} x \\ w \end{bmatrix}$ of $\Sigma_{\text{sc}}^{\mathfrak{W}}$ satisfying $x(0) = x_0$ is equal to $Q_+ \pi_+ w_0 = \Pi_+ x_0$. Consequently, $\mathfrak{C}_{\Sigma_{\text{sc}}^{\mathfrak{W}}} = \Pi_+|_{\mathcal{D}(\mathfrak{W})}$.

Step 8: $\Sigma_{\text{sc}}^{\mathfrak{W}}$ is simple. According to Lemma 5.10, the linear span of the ranges of $\mathfrak{B}_{\Sigma_{\text{sc}}^{\mathfrak{W}}} = \begin{bmatrix} \Gamma_{\mathfrak{W}} \\ 1_{\mathcal{H}_-} \end{bmatrix}$ and $\mathfrak{C}_{\Sigma_{\text{sc}}^{\mathfrak{W}}}^* = \begin{bmatrix} 1_{\mathcal{W}_+} \\ \Gamma_{\mathfrak{W}}^* \end{bmatrix}$ is dense in the state space $\mathcal{D}(\mathfrak{W})$, and hence $\Sigma_{\text{sc}}^{\mathfrak{W}}$ is simple. $\square$

Theorem 10.2 *The generating subspace of the s/s system $\Sigma_{\text{sc}}^{\mathfrak{W}}$ in Theorem 10.1 is given by*

$$V_{\text{sc}}^{\mathfrak{W}} = \left\{ \begin{bmatrix} Q\dot{w} \\ Qw \\ w(0) \end{bmatrix} \in \begin{bmatrix} \mathcal{D}(\mathfrak{W}) \\ \mathcal{D}(\mathfrak{W}) \\ \mathcal{W} \end{bmatrix} \left| \begin{array}{l} w \in \mathcal{L}(\mathfrak{W}) \text{ is locally absolutely} \\ \text{continuous with } \dot{w} \in K^2(\mathcal{W}), \text{ and} \\ \lim_{t \rightarrow 0} \frac{1}{t} Q(\tau^t w - w) \text{ exists in } \mathcal{D}(\mathfrak{W}). \end{array} \right. \right\} \quad (10.2)$$

Proof The proof is essentially the same as the proof of Theorem 8.2 with $\mathbb{R}^+$ replaced by $\mathbb{R}$, Q_+ replaced by Q , and $\mathcal{K}(\mathfrak{W}_+)$ replaced by $\mathcal{L}(\mathfrak{W})$. For the converse direction one needs the fact that for a conservative system part (ii) of Proposition 3.7 holds in the following modified form:

- (ii') For each $\begin{bmatrix} z_0 \\ x_0 \\ w_0 \end{bmatrix} \in V$ there exists a stable two-sided classical trajectory $\begin{bmatrix} x \\ w \end{bmatrix}$ satisfying $\begin{bmatrix} \dot{x}(0) \\ x(0) \\ w(0) \end{bmatrix} = \begin{bmatrix} z_0 \\ x_0 \\ w_0 \end{bmatrix}$ with the additional property that w is locally abso-

lutely continuous and $\begin{bmatrix} \dot{x} \\ \dot{w} \end{bmatrix}$ is a stable two-sided trajectory of Σ . In particular,

$$V = \left\{ \begin{bmatrix} \dot{x}(0) \\ x(0) \\ w(0) \end{bmatrix} \middle| \begin{bmatrix} x \\ w \end{bmatrix} \text{ is a two-sided classical trajectory of } \Sigma \right\}. \quad (10.3)$$

That (ii') holds for conservative systems follows from Proposition 3.7 and Remark 4.2. $\square$

Let $\mathfrak{R}$ be the reachable subspace, $\mathfrak{U}$ the unobservable subspace, $\mathfrak{R}^\dagger$ the backward reachable subspace, and $\mathfrak{U}^\dagger$ the backward unobservable subspace of $\Sigma_{sc}^{\mathfrak{W}}$. As we noticed earlier, $\mathfrak{R}^\dagger = \mathfrak{U}^\perp$ and $\mathfrak{U}^\dagger = \mathfrak{R}^\perp$. By Lemma 5.12 and Theorem 10.1,

$$\begin{aligned} \mathfrak{R} &= \text{im}(\mathfrak{B}_{\Sigma_{sc}^{\mathfrak{W}}}) = \text{im}\left(\begin{bmatrix} \Gamma_{\mathfrak{W}} \\ 1_{\mathcal{H}_-} \end{bmatrix}\right) = \{[\Gamma_{\mathfrak{W}} x_-] \mid x_- \in \mathcal{H}_-\}, \\ \mathfrak{U}^\dagger &= \ker(\mathfrak{B}_{\Sigma_{sc}^{\mathfrak{W}}}) = \ker(\Pi_-|_{\mathcal{D}(\mathfrak{W})}) = \{Qw \mid w \in \mathcal{L}(\mathfrak{W}) \cap K_+^2(\mathcal{W})\}, \\ \mathfrak{R}^\dagger &= \text{im}(\mathfrak{C}_{\Sigma_{sc}^{\mathfrak{W}}}^*) = \text{im}\left(\begin{bmatrix} 1_{\mathcal{H}_+} \\ \Gamma_{\mathfrak{W}}^* \end{bmatrix}\right) = \left\{\begin{bmatrix} x_+ \\ \Gamma_{\mathfrak{W}}^* x_+ \end{bmatrix} \mid x_+ \in \mathcal{H}_+\right\}, \\ \mathfrak{U} &= \ker(\mathfrak{C}_{\Sigma_{sc}^{\mathfrak{W}}}) = \ker(\Pi_+|_{\mathcal{D}(\mathfrak{W})}) = \{Qw \mid w \in \mathcal{L}(\mathfrak{W}) \cap K_-^2(\mathcal{W})\}. \end{aligned} \quad (10.4)$$

The orthogonal projections onto these subspaces are given by

$$\begin{aligned} P_{\mathfrak{R}} &= \mathfrak{B}_{\Sigma_{sc}^{\mathfrak{W}}} \mathfrak{B}_{\Sigma_{sc}^{\mathfrak{W}}}^* = \begin{bmatrix} \Gamma_{\mathfrak{W}} \Pi_-|_{\mathcal{D}(\mathfrak{W})} \\ \Pi_-|_{\mathcal{D}(\mathfrak{W})} \end{bmatrix}, \\ P_{\mathfrak{U}^\dagger} &= 1_{\mathcal{D}(\mathfrak{W})} - P_{\mathfrak{R}} = \Pi_+|_{\mathcal{D}(\mathfrak{W})} - \Gamma_{\mathfrak{W}} \Pi_-|_{\mathcal{D}(\mathfrak{W})}, \\ P_{\mathfrak{R}^\dagger} &= \mathfrak{C}_{\Sigma_{sc}^{\mathfrak{W}}}^* \mathfrak{C}_{\Sigma_{sc}^{\mathfrak{W}}} = \begin{bmatrix} \Pi_+|_{\mathcal{D}(\mathfrak{W})} \\ \Gamma_{\mathfrak{W}}^* \Pi_+|_{\mathcal{D}(\mathfrak{W})} \end{bmatrix}, \\ P_{\mathfrak{U}} &= 1_{\mathcal{D}(\mathfrak{W})} - P_{\mathfrak{R}^\dagger} = \Pi_-|_{\mathcal{D}(\mathfrak{W})} - \Gamma_{\mathfrak{W}}^* \Pi_+|_{\mathcal{D}(\mathfrak{W})}, \end{aligned} \quad (10.5)$$

Theorem 10.3 *Let $\Sigma = (V; \mathcal{X}, \mathcal{W})$ be a conservative s/s system with behavior $\mathfrak{W}$, input map $\mathfrak{B}_\Sigma$, output map $\mathfrak{C}_\Sigma$, reachable subspace $\mathfrak{R} = \text{im}(\mathfrak{B}_\Sigma)$, unobservable subspace $\mathfrak{U}_\Sigma = \ker(\mathfrak{C}_\Sigma)$, backward reachable subspace $\mathfrak{R}_\Sigma^\dagger = \text{im}(\mathfrak{C}_\Sigma^*)$ and backward unobservable subspace $\mathfrak{U}_\Sigma^\dagger = \ker(\mathfrak{B}_\Sigma^*)$.*

(i) *The operator*

$$\mathfrak{C}_\Sigma^{\text{bil}} := \begin{bmatrix} \mathfrak{C}_\Sigma \\ \mathfrak{B}_\Sigma^* \end{bmatrix} \quad (10.6)$$

is a co-isometry from $\mathcal{X}$ onto $\mathcal{D}(\mathfrak{W})$, with kernel $\mathcal{X}_0 := \ker(\mathfrak{C}_\Sigma^{\text{bil}}) = \mathfrak{U} \cap \mathfrak{U}^\dagger$. Thus, Σ is simple if and only if $\mathfrak{C}_\Sigma^{\text{bil}}$ is injective.

(ii) *Define $\mathfrak{B}_\Sigma^{\text{bil}} := (\mathfrak{C}_\Sigma^{\text{bil}})^*$. Then $\mathfrak{B}_\Sigma^{\text{bil}}$ is an isometry $\mathcal{D}(\mathfrak{W}) \rightarrow \mathcal{X}$ with range $\mathcal{X}_0^\perp = \overline{\mathfrak{R} + \mathfrak{R}^\dagger}$, which is uniquely determined by the fact that*

$$[\mathfrak{C}_\Sigma^* \ \mathfrak{B}_\Sigma] = \mathfrak{B}_\Sigma^{\text{bil}} A_{\mathfrak{W}}, \quad (10.7)$$

where $A_{\mathfrak{W}}$ is the operator defined in (5.18). In particular, $\mathfrak{B}_\Sigma^{\text{bil}}$ is surjective if and only if Σ is simple.

- (iii) A two-sided trajectory $\begin{bmatrix} x \\ w \end{bmatrix}$ of Σ is stable if and only if $w \in K^2(\mathcal{W})$.
- (iv) If $\begin{bmatrix} x \\ w \end{bmatrix}$ is a stable two-sided trajectory of Σ , then $w \in \mathcal{L}(\mathfrak{W})$, $Q\tau^t w = \mathfrak{C}_\Sigma^{\text{bil}} x(t)$, and $P_{\mathcal{X}_0^\perp} x(t) = \mathfrak{B}_\Sigma^{\text{bil}} Q\tau^t w$ for all $t \in \mathbb{R}$.
- (v) Conversely, let $w \in \mathcal{L}(\mathfrak{W})$, and define $x(t) = \mathfrak{B}_\Sigma^{\text{bil}} Q\tau^t w$, $t \in \mathbb{R}$. Then $\begin{bmatrix} x \\ w \end{bmatrix}$ is a stable two-sided trajectory of Σ .
- (vi) The state component x of a stable two-sided trajectory $\begin{bmatrix} x \\ w \end{bmatrix}$ of Σ is determined uniquely by the signal component w if and only if Σ is simple.

Proof The proof of this theorem is essentially the same as the proof of parts (1)–(6) of [6, Theorem 4.1] (see Remark 3.17). $\square$

Corollary 10.4 Let $\mathfrak{W}$ be a two-sided behavior on the Kreĭn space $\mathcal{W}$. Then the pair of functions $\begin{bmatrix} x \\ w \end{bmatrix}$ is a stable two-sided trajectory of $\Sigma_{\text{sc}}^{\mathfrak{W}}$ if and only if

$$w \in \mathcal{L}(\mathfrak{W}) \quad \text{and} \quad x(t) = Q\tau^t w, \quad t \in \mathbb{R}.$$

Proof This follows from Theorem 10.3. $\square$

Theorem 10.5 Every simple conservative s/s system Σ with two-sided behavior $\mathfrak{W}$ is unitarily similar to the system $\Sigma_{\text{sc}}^{\mathfrak{W}}$. The unitary similarity transformation is the map $\mathfrak{C}_\Sigma^{\text{bil}}$ defined in (10.6).

Proof Let $\begin{bmatrix} x_+ \\ w_+ \end{bmatrix}$ be a stable future trajectory of Σ . By Lemma 3.4 and Remark 4.2, this trajectory can be extended to a stable two-sided trajectory $\begin{bmatrix} x \\ w \end{bmatrix}$ of Σ . For each $t \in \mathbb{R}$, define $x_1(t) = Q\tau^t w$. Then $\begin{bmatrix} x_1 \\ w \end{bmatrix}$ is a stable two-sided trajectory of $\Sigma_{\text{sc}}^{\mathfrak{W}}$, and by Theorem 10.3(iv), $x_1(t) = \mathfrak{C}_\Sigma^{\text{bil}} x(t)$ for all $t \in \mathbb{R}$, and hence, in particular, for all $t \in \mathbb{R}^+$.

Conversely, let $\begin{bmatrix} x_1 \\ w_+ \end{bmatrix}$ be a stable future trajectory of $\Sigma_{\text{sc}}^{\mathfrak{W}}$. This trajectory can be extended to a stable two-sided trajectory $\begin{bmatrix} x_1 \\ w \end{bmatrix}$, after which it satisfies $x_1(t) = Q\tau^t w$ for all $t \in \mathbb{R}$. For each $t \in \mathbb{R}$ we can define $x(t) = \mathfrak{B}_\Sigma^{\text{bil}} \tau^t Qw$. Then by Theorem 10.3(v), $\begin{bmatrix} x \\ w \end{bmatrix}$ is a stable two-sided trajectory of Σ and $x(t) = \mathfrak{B}_\Sigma^{\text{bil}} Q\tau^t w$ for all $t \in \mathbb{R}$, and hence, in particular, for all $t \in \mathbb{R}^+$.

Since $\mathfrak{C}_\Sigma^{\text{bil}}$ is unitary we conclude that Σ is unitarily similar to $\Sigma_{\text{sc}}^{\mathfrak{W}}$ with similarity operator $\mathfrak{C}_\Sigma^{\text{bil}}$. $\square$

Definition 10.6 We call the system $\Sigma_{\text{sc}}^{\mathfrak{W}}$ the *canonical model* of a simple conservative s/s system with two-sided behavior $\mathfrak{W}$.

Corollary 10.7 Any two simple conservative realizations of a given passive two-sided behavior $\mathfrak{W}$ are unitarily similar to each other.

Proof This follows from Theorem 10.5. $\square$

Definition 10.8 We call the operators $\mathfrak{C}_\Sigma^{\text{bil}}$ and $\mathfrak{B}_\Sigma^{\text{bil}}$ defined in Theorem 10.3 the two-sided (or bilateral) *output and input maps*, respectively, of the conservative s/s system Σ .

As we shall show below, the two models in Sects. 8 and 9 can be obtained from $\Sigma_{sc}^{\mathfrak{M}}$ by first performing an orthogonal compression, and then applying a unitary similarity transform.

By Theorem 8.4, the output map $\mathfrak{C}_{\Sigma_{sc}^{\mathfrak{M}}}$ intertwines $\Sigma_{sc}^{\mathfrak{M}}$ and the co-energy preserving system $\Sigma_{oce}^{\mathfrak{M}+}$. Since $\Sigma_{sc}^{\mathfrak{M}}$ is conservative, it follows from Lemma 6.19 that $\mathfrak{C}_{\Sigma_{sc}^{\mathfrak{M}}} = \Pi_+|_{\mathcal{D}(\mathfrak{M})}$ is a co-isometry, and the restriction of $\mathfrak{C}_{\Sigma_{sc}^{\mathfrak{M}}}$ to $\mathfrak{R}^\dagger$ is a unitary map of $\mathfrak{R}^\dagger$ onto $\mathcal{H}_+$. Clearly $\Sigma_{oce}^{\mathfrak{M}+}$ is unitarily similar to the system $\Sigma_{sc}^o := (V_{sc}^o; \mathfrak{R}^\dagger, \mathcal{W})$ that we get by applying $\mathfrak{C}_{\Sigma_{sc}^{\mathfrak{M}}}^* = \begin{bmatrix} 1_{\mathcal{H}_+} \\ \Gamma_{\mathfrak{M}}^* \end{bmatrix}$ to the state of the state of $\Sigma_{oce}^{\mathfrak{M}+}$. The set of all future stable trajectories of Σ_{sc}^o is given by

$$\left\{ \begin{bmatrix} x \\ w_+ \end{bmatrix} \in \begin{bmatrix} C(\mathbb{R}^+; \mathfrak{R}^\dagger) \\ \mathcal{K}(\mathfrak{M}_+) \end{bmatrix} \mid x(t) = \begin{bmatrix} 1_{\mathcal{H}_+} \\ \Gamma_{\mathfrak{M}}^* \end{bmatrix} Q_+ \tau_+^t w_+, \quad t \in \mathbb{R}^+ \right\}. \quad (10.8)$$

Since $w_+ \in \mathcal{K}(\mathfrak{M}_+)$ if and only if it can be written in the form $w_+ = \pi_+ w$ for some $w \in \mathcal{L}(\mathfrak{M})$, we can replace the parameter w_+ by $\pi_+ w$ with $w \in \mathcal{L}(\mathfrak{M})$, after which (10.8) can be written in the equivalent form

$$\left\{ \begin{bmatrix} x \\ w_+ \end{bmatrix} \in \begin{bmatrix} C(\mathbb{R}^+; \mathfrak{R}^\dagger) \\ \mathcal{K}(\mathfrak{M}_+) \end{bmatrix} \mid x(t) = P_{\mathfrak{R}^\dagger} Q \tau^t w, \quad w_+ = \pi_+ w, \quad w \in \mathcal{L}(\mathfrak{M}), \quad t \in \mathbb{R}^+ \right\}. \quad (10.9)$$

Comparing this to (10.1) we find that Σ_{sc}^o is the orthogonal outgoing compression of $\Sigma_{sc}^{\mathfrak{M}}$ onto the backward reachable subspace $\mathfrak{R}^\dagger$ of $\Sigma_{sc}^{\mathfrak{M}}$.

By Theorem 9.5, the input map $\mathfrak{B}_{\Sigma_{sc}^{\mathfrak{M}}}$ intertwines the energy preserving system $\Sigma_{cep}^{\mathfrak{M}}$ with $\Sigma_{sc}^{\mathfrak{M}}$. Since $\Sigma_{sc}^{\mathfrak{M}}$ is conservative, it follows from Lemma 6.15 that $\mathfrak{B}_{\Sigma_{sc}^{\mathfrak{M}}} = \begin{bmatrix} \Gamma_{\mathfrak{M}} \\ 1_{\mathcal{H}_-} \end{bmatrix}$ is an isometry whose range is the reachable subspace $\mathfrak{R}$ of $\Sigma_{sc}^{\mathfrak{M}}$. Clearly $\Sigma_{cep}^{\mathfrak{M}}$ is unitarily similar to the system $\Sigma_{sc}^c := (V_{sc}^c; \mathfrak{R}, \mathcal{W})$ that we get by applying $\mathfrak{B}_{\Sigma_{sc}^{\mathfrak{M}}}$ to the state of $\Sigma_{cep}^{\mathfrak{M}}$. The set of all future stable trajectories of Σ_{sc}^c is given by

$$\left\{ \begin{bmatrix} x \\ w_+ \end{bmatrix} \in \begin{bmatrix} C(\mathbb{R}^+; \mathfrak{R}) \\ \mathcal{K}(\mathfrak{M}_+) \end{bmatrix} \mid x = Q \tau^t w, \quad w_+ = \pi_+ w, \quad w \in Q^{-1}\mathfrak{R} \right\}. \quad (10.10)$$

Comparing this to (10.1) we find that Σ_{sc}^c is an orthogonal incoming compression of $\Sigma_{sc}^{\mathfrak{M}}$ onto the reachable subspace $\mathfrak{R}$ of $\Sigma_{sc}^{\mathfrak{M}}$.

Theorem 10.9 *Let $\mathfrak{M}_-$ be a passive past behavior on the Kreĭn space $\mathcal{W}$. Then the generating subspace $V_{cep}^{\mathfrak{M}_-}$ of the canonical model $\Sigma_{cep}^{\mathfrak{M}_-}$ in Theorem 9.1 is given by*

$$V_{cep}^{\mathfrak{M}_-} = \left\{ \begin{bmatrix} Q_- \pi_- \dot{w} \\ Q_- \pi_- w \\ w(0) \end{bmatrix} \in \begin{bmatrix} \mathcal{H}_- \\ \mathcal{H}_- \\ \mathcal{W} \end{bmatrix} \mid \begin{array}{l} w \in \text{im} \left(\begin{bmatrix} \Gamma_{\mathfrak{M}_-} \\ 1_{\mathcal{H}_-} \end{bmatrix} \right) \text{ is locally absolutely} \\ \text{continuous with } \dot{w} \in K^2(\mathcal{W}), \text{ and} \\ \lim_{t \rightarrow 0+} \frac{1}{t} Q_- \pi_- (\tau^t w - w) \text{ exists in } \mathcal{H}_-. \end{array} \right\} \quad (10.11)$$

Proof As we established above, $\Sigma_{\text{cep}}^{\mathfrak{M}_-}$ is unitarily similar to the orthogonal incoming compression Σ_{sc}^c of $\Sigma_{\text{sc}}^{\mathfrak{M}}$ onto its reachable subspace $\mathfrak{R}$. That subspace is strongly invariant in the sense that if $\begin{bmatrix} x \\ w \end{bmatrix}$ is a future trajectory of $\Sigma_{\text{sc}}^{\mathfrak{M}}$ satisfying $x(0) \in \mathfrak{R}$, then $x(t) \in \mathfrak{R}$ for all $t \geq 0$ (see Lemma 6.14). In particular, $\begin{bmatrix} x \\ w \end{bmatrix}$ is a smooth future trajectory of Σ_{sc}^c if and only if $\begin{bmatrix} x \\ w \end{bmatrix}$ is a smooth future trajectory of $\Sigma_{\text{sc}}^{\mathfrak{M}}$ and $x(0) \in \mathfrak{R}$. This, combined with Proposition 3.7 and Theorem 10.2 implies that the generating subspace V_{sc}^c of Σ_{sc}^c is given by

$$V_{\text{sc}}^c = V_{\text{sc}}^{\mathfrak{M}} \cap \left[\begin{array}{c} \mathcal{D}(\mathfrak{M}) \\ \mathfrak{R} \\ \mathcal{W} \end{array} \right] = \left\{ \left[\begin{array}{c} Q\dot{w} \\ Qw \\ w(0) \end{array} \right] \in \left[\begin{array}{c} \mathfrak{R} \\ \mathfrak{R} \\ \mathcal{W} \end{array} \right] \left| \begin{array}{l} w \in Q^{-1}\mathfrak{R} \text{ is locally absolutely} \\ \text{continuous with } \dot{w} \in K^2(\mathcal{W}), \text{ and} \\ \lim_{t \rightarrow 0+} \frac{1}{t} Q(\tau^t w - w) \text{ exists in } \mathfrak{R}. \end{array} \right. \right\} \quad (10.12)$$

From here we get the generating subspace $V_{\text{cep}}^{\mathfrak{M}_-}$ of $\Sigma_{\text{cep}}^{\mathfrak{M}_-}$ by applying the unitary operator $(\mathfrak{B}_{\text{sc}}^{\mathfrak{M}})^* = \Pi_-$ to the two state components. This leads to formula (10.11). $\square$

References

1. Alpay, D., Dijksma, A., Rovnyak, J., de Snoo, H.: Schur functions, operator colligations, and reproducing kernel Hilbert spaces. In: Operator Theory: Advances and Applications, vol. 96. Birkhäuser-Verlag, Basel (1997)
2. Azizov, T.Y., Iokhvidov, I.S.: Linear Operators in Spaces with an Indefinite Metric. Wiley, New York (1989)
3. Arov, D.Z., Staffans, O.J.: State/signal linear time-invariant systems theory. Part I: discrete time systems. In: The State Space Method, Generalizations and Applications. Operator Theory: Advances and Applications, vol. 161, pp. 115–177. Birkhäuser-Verlag, Basel (2005)
4. Arov, D.Z., Staffans, O.J.: A Kreĭn space coordinate free version of the de Branges complementary space. J. Funct. Anal. **256**, 3892–3915 (2009)
5. Arov, D.Z., Staffans, O.J.: Two canonical passive state/signal shift realizations of passive discrete time behaviors. J. Funct. Anal. **257**, 2573–2634 (2009)
6. Arov, D.Z., Staffans, O.J.: Canonical conservative state/signal shift realizations of passive discrete time behaviors. J. Funct. Anal. **259**, 3265–3327 (2010)
7. Belevitch, V.: Classical Network Theory. Holden-Day, San Francisco (1968)
8. Bognár, J.: Indefinite inner product spaces. Ergebnisse der Mathematik und ihrer Grenzgebiete, vol. 78. Springer-Verlag, Berlin (1974)
9. Ball, J.A., Staffans, O.J.: Conservative state-space realizations of dissipative system behaviors. Integral Equ. Oper. Theory **54**, 151–213 (2006)
10. de Branges, L., Rovnyak, J.: Canonical models in quantum scattering theory. In: Perturbation Theory and its Applications in Quantum Mechanics (Proc. Adv. Sem. Math. Res. Center, U.S. Army, Theoret. Chem. Inst., Univ. of Wisconsin, Madison, Wis., 1965), pp. 295–392. Wiley, New York (1966)
11. de Branges, L., Rovnyak, J.: Square Summable Power Series. Holt, Rinehart and Winston, New York (1966)
12. Gripenberg, G., Londen, S.-O., Staffans, O.: Volterra integral and functional equations. In: Encyclopedia of Mathematics and its Applications, vol. 34. Cambridge University Press, Cambridge (1990)
13. Kurula, M., Staffans, O.J.: Well-posed state/signal systems in continuous time. Complex Anal. Oper. Theory **4**, 319–390 (2009)
14. Kurula, M., Staffans, O.J.: Connections between classical and generalized trajectories of a state/signal system. Complex Anal. Oper. Theory (2011, accepted). doi:[10.1007/s11785-010-0129-7](https://doi.org/10.1007/s11785-010-0129-7)

15. Kurula, M.: On passive and conservative state/signal systems in continuous time. *Integral Equ. Oper. Theory* **67**, 377–424, 449 (2010)
16. Nikolskiĭ, N.K., Vasyunin, V.I.: A unified approach to function models, and the transcription problem. In: *The Gohberg Anniversary Collection*, vol. II (Calgary, AB, 1988). *Operator Theory Advances and Applications*, vol. 41, pp. 405–434. Birkhäuser, Basel (1989)
17. Nikolskiĭ, N.K., Vasyunin, V.I.: Elements of spectral theory in terms of the free function model. I. Basic constructions. In: *Holomorphic Spaces* (Berkeley, CA, 1995). *Math. Sci. Res. Inst. Publ.*, vol. 33, pp. 211–302. Cambridge University Press, Cambridge (1998)
18. Polderman, J.W., Willems, J.C.: *Introduction to Mathematical Systems Theory: A Behavioral Approach*. Springer-Verlag, New York (1998)
19. Staffans, O.J.: *Well-Posed Linear Systems*. Cambridge University Press, Cambridge (2005)
20. Willems, J.C.: Dissipative dynamical systems. Part I: general theory. *Arch. Ration. Mech. Anal.* **45**, 321–351 (1972)
21. Willems, J.C.: Dissipative dynamical systems. Part II: linear systems with quadratic supply rates. *Arch. Ration. Mech. Anal.* **45**, 352–393 (1972)
22. Wohlers, M.R.: *Lumped and Distributed Passive Networks*. Academic Press, New York (1969)
23. Willems, J.C., Trentelman, H.L.: On quadratic differential forms. *SIAM J. Control Optim.* **36**(5), 1703–1749 (1998)
24. Willems, J.C., Trentelman, H.L.: Synthesis of dissipative systems using quadratic differential forms: Parts I–II. *IEEE Trans. Autom. Control* **47**, 53–86 (2002)

Optimal stopping of strong Markov processes

Sören Christensen*, Paavo Salminen†, and Bao Quoc Ta †

Abstract

We characterize the value function and the optimal stopping time for a large class of optimal stopping problems where the underlying process to be stopped is a fairly general Markov process. The main result is inspired by recent findings for Lévy processes obtained essentially via the Wiener-Hopf factorization. The main ingredient in our approach is the representation of the β -excessive functions as expected suprema. A variety of examples is given.

Keywords: optimal stopping problem, Markov processes, Hunt processes, Lévy processes, supremum representation for excessive functions

1 Introduction

Consider a real-valued strong Markov process $X = (X_t)_{t \geq 0}$. Let $\mathbb{P}_x$ and $\mathbb{E}_x$ stand for the probability and the expectation, respectively, associated with X when initiated from x . The natural filtration generated by X is denoted by $\mathcal{F} = (\mathcal{F}_t)_{t \geq 0}$ and $\mathcal{M}$ is the set of all stopping times with respect to $\mathcal{F}$.

We are interested in studying the following optimal stopping problem: Find a function V (value function) and a stopping time τ^* (optimal stopping time) such that

$$V(x) := \sup_{\tau \in \mathcal{M}} \mathbb{E}_x \left(e^{-\beta \tau} G(X_\tau) \right) = \mathbb{E}_x \left(e^{-\beta \tau^*} G(X_{\tau^*}) \right), \quad (*)$$

*Christian-Albrechts-Universität, Mathematisches Seminar, Ludewig-Meyn-Str.4, D-24098 Kiel, Germany, email: christensen@math.uni-kiel.de.

†Åbo Akademi, Department of Natural Sciences, Mathematics and Statistics, FIN-20500 Åbo, Finland, email: phsalmin@abo.fi, tbao@abo.fi

where the function G (reward function) is assumed to have some regularity properties to be specified later. On $\{\tau = \infty\}$ we define

$$e^{-\beta\tau}G(X_\tau) = \limsup_{t \rightarrow \infty} e^{-\beta t}G(X_t).$$

Recall that a non-negative, measurable function u is called β -excessive for X if the following two conditions hold:

$$\mathbb{E}_x\left(e^{-\beta t}u(X_t)\right) \leq u(x) \quad \forall t \geq 0, x \in S,$$

$$\lim_{t \rightarrow 0} \mathbb{E}_x\left(e^{-\beta t}u(X_t)\right) = u(x) \quad \forall x \in S.$$

A basic result, see Theorem 1 p. 124 in Shiriyayev [25], is that if the reward function G is lower semicontinuous then the value function V exists and is characterized as the smallest β -excessive majorant of G .

During the last decade the theory of optimal stopping for Lévy processes has been developed strongly. This research has been focused on analyzing the validity of the smooth pasting condition, see Alili and Kyprianou [1], Christensen and Irle [5], on the one hand, and solving particular problems, see Boyarchenko and Levendorskii [4], Mordecki [17], Novikov and Shiriyayev [19], [20], Kyprianou and Surya [15], Surya [26], Mordecki and Salminen [18], Deligiannidis, Le and Utev [9], on the other hand. The methodology to derive explicit solutions is in many cases based on the Wiener-Hopf factorization.

For random walks in discrete time optimal stopping problems with the reward functions $G(x) = x^+$ and $G(x) = (e^x - 1)^+$ were solved already by Darling et al. in [8]. In Mordecki [17] the results in [8] were lifted to the framework of Lévy processes with the focus to the reward functions $G(x) = (K - e^x)^+$ (put option) and $G(x) = (e^x - K)^+$ (call option). Furthermore, for the value function $G(x) = (x^+)^{\nu}$, $\nu > 0$, the problem was solved by Novikov and Shiriyayev in [19] and [20] for random walks and Lévy processes using Appell polynomials and Appell functions, see also Kyprianou and Surya [15] and Alili and Kyprianou [1]. In addition to the Wiener-Hopf factorization, the fluctuation identities constitute an essential tool in these solutions.

This technique was generalized under certain conditions to be applicable for more general reward functions independently in Surya [26] and Deligiannidis et al. [9]. In particular, in [26] it is assumed that there exists a function Q such that the reward function has the representation

$$G(x) = \mathbb{E}_0\left(Q(x + M_T)\right), \tag{1}$$

where T is an exponentially with parameter $\beta > 0$ distributed random variable independent of X and $M_t := \sup_{0 \leq s \leq t} X_s$ is the running maximum of

the Lévy process X . The function Q should have a unique positive root x^* such that $Q(x) < 0$ for $x < x^*$ and $Q(x) > 0$ and non-decreasing for $x > x^*$. Then the stopping problem (*) is, in fact, one-sided and the optimal stopping time is

$$\tau^* := \inf\{t : X_t \geq x^*\}.$$

Moreover, the value function is expressed in [26] as

$$V(x) = \mathbb{E}_0 \left(Q(x + M_T) 1_{\{x + M_T \geq x^*\}} \right).$$

In [9] the starting point is the assumption that there exists a function g such that

$$G(x) = \mathbb{E}_0 \left(g(x + X_T) \right). \quad (2)$$

To see the connection between Q in (1) and g in (2) recall the Wiener-Hopf factorization

$$X_T \stackrel{(d)}{=} M_T + I_T^\circ, \quad (3)$$

where $I_T^\circ \stackrel{(d)}{=} I_T := \inf_{0 \leq t \leq T} X_t$ and independent of X . Consequently,

$$Q(x) = \mathbb{E}_0 \left(g(x + I_T) \right). \quad (4)$$

Since in [9] the stopping problem is analyzed via the function

$$x \rightarrow \mathbb{E}_0(g(x + I_T))$$

the relationship (4) implies that the results in [26] and [9] are very close to each other.

Our main result characterizes the solution of the optimal stopping problem (*) similarly as in [19], [26] and [9] but is valid for very general strong Markov processes, e.g., for diffusions, Lévy processes and continuous time Markov chains. The main ingredient in our approach is the representation of β -excessive functions in the form

$$x \mapsto \mathbb{E}_x \left(\sup_{0 \leq t \leq T} f(X_t) \right)$$

for a function $f \geq 0$. This kind of representations have been studied in Föllmer and Knispel [11] and El Karoui and Meziou [10].

The structure of the paper is the following: In the next section we state and prove the main theorems which lead to explicit solutions to the optimal stopping problem (*) for general Hunt processes X and reward functions G with some representation properties. Depending on the form of the representing function $\hat{f}$ the optimal stopping time is one-sided or two-sided. We furthermore present in section 2.2 a way for finding $\hat{f}$ explicitly. In section 2.3 the connection to the representing measure approach is discussed. In Section 3 the results are illustrated with a variety of examples.

2 Main results

2.1 Verification theorems

We use the notation from the introduction and let $X = (X_t)_{t \geq 0}$ be a Hunt process taking values in a subset S of $\mathbb{R}$, e.g., an interval or a finite set. We refer to [18] for references and further discussions on Hunt processes. For a full account see Blumenthal and Gettoor [2], Chung and Walsh [7] and Sharpe [24]. In particular, X has the strong Markov property and is quasi left continuous with right continuous sample paths with left limits. For simplicity, we suppose that the lifetime of X is a.s. infinite.

Remark 2.1. *In some problems it is of interest to study X , e.g., up to the passage time of a level. In such cases X is viewed as a Hunt process with lifetime equal to the passage time. For instance, a Brownian motion X killed at the first passage time*

$$\zeta_0 := \inf\{t \geq 0 : X_t < 0\}$$

is a Hunt process with lifetime ζ_0 . It is not difficult to see how to refine our approach to cover these and more general lifetimes, and we leave this task to the interested reader.

Throughout the paper, the notation T is used for an exponentially distributed random variable assumed to be independent of X and having the mean $1/\beta$ where $\beta > 0$ is the discounting rate introduced in (*). We remark that a function u is β -excessive if it is 0-excessive, i.e, excessive, for X killed at T .

The following lemma is the key to our approach for the optimal stopping problem (*). It is proved similarly as Proposition 2.1 in [11]. We give a short proof for the sake of completeness.

Lemma 2.2. *Let $f : S \mapsto \mathbb{R}_+$ be an upper semicontinuous function and define $u(x) := \mathbb{E}_x\left(\sup_{0 \leq t \leq T} f(X_t)\right)$. Then the function $u : S \mapsto \mathbb{R} \cup \{+\infty\}$ is β -excessive.*

Proof. By the upper semicontinuity of f and the right continuity of the sample paths of X the function $t \mapsto \sup_{0 \leq s \leq t} f(X_s)$ is right continuous and,

hence, measurable. It follows that u is well-defined. Consider for $t > 0$

$$\begin{aligned}
u(x) &= \mathbb{E}_x \left(\sup_{0 \leq s \leq T} f(X_s) \right) \\
&\geq \mathbb{E}_x \left(\sup_{t \leq s \leq T} f(X_s); t \leq T \right) = \mathbb{E}_x \left(\mathbb{E}_x \left(\sup_{t \leq s \leq T} f(X_s) | \mathcal{F}_t \right); t \leq T \right) \\
&= \mathbb{E}_x \left(\mathbb{E}_x \left(\sup_{t \leq s \leq T} f(X_s) | X_t \right); t \leq T \right) = \mathbb{E}_x \left(u(X_t); t \leq T \right) \\
&= \mathbb{E}_x \left(e^{-\beta t} u(X_t) \right),
\end{aligned}$$

and, therefore,

$$\mathbb{E}_x \left(u(X_t); t \leq T \right) \leq u(x).$$

Next,

$$\begin{aligned}
\lim_{t \rightarrow 0} \mathbb{E}_x \left(e^{-\beta t} u(X_t) \right) &= \lim_{t \rightarrow 0} \mathbb{E}_x \left(\sup_{t \leq s \leq T} f(X_s); t \leq T \right) \\
&= \mathbb{E}_x \left(\sup_{0 \leq s \leq T} f(X_s) \right) \\
&= u(x),
\end{aligned}$$

where the second equality holds by the monotone convergence. $\square$

Example 2.3. Consider a regular linear diffusion X on $\mathbb{R}$. Let $x \rightarrow u_\beta(x, y)$ be the β -resolvent kernel associated with X . As is well-known, the function $x \rightarrow u_\beta(x, y)$ is β -excessive for any fixed y . We find the representation $u_\beta(\cdot, y)$ as expected supremum. Recall that

$$\mathbb{E}_x \left(e^{-\beta H_y} \right) = \frac{u_\beta(x, y)}{u_\beta(y, y)} = \begin{cases} \frac{\psi_\beta(x)}{\psi_\beta(y)}, & \text{if } x \leq y, \\ \frac{\varphi_\beta(x)}{\varphi_\beta(y)}, & \text{if } x \geq y, \end{cases} \quad (5)$$

where $H_y := \inf\{t : X_t = y\}$ and ψ_β and φ_β are the fundamental increasing and decreasing, respectively, solutions for the generalized differential operator associated with X , see [3]. Introducing

$$x \rightarrow f_y(x) = \begin{cases} u_\beta(y, y), & \text{if } x = y, \\ 0, & \text{otherwise,} \end{cases}$$

we have by (5)

$$\begin{aligned}
\mathbb{E}_x \left(\sup_{t \leq T} f_y(X_t) \right) &= u_\beta(y, y) \mathbb{P}_x(H_y < T) \\
&= u_\beta(y, y) \mathbb{E}_x(e^{-\beta H_y}) \\
&= u_\beta(x, y).
\end{aligned}$$

For another example, consider the β -excessive function $\psi_\beta \wedge 1$. We assume that there exists a point y such that $\psi_\beta(y) = 1$. Notice that $\psi_\beta(x) \geq 1$ for all $x \geq y$ since ψ_β is non-decreasing. Define

$$x \rightarrow f_y^\uparrow(x) = \mathbf{1}_{[y, +\infty)}(x).$$

Then, calculating as above, we have

$$\mathbb{E}_x \left(\sup_{t \leq T} f_y^\uparrow(X_t) \right) = \begin{cases} 1, & \text{if } x \geq y, \\ \psi_\beta(x), & \text{if } x \leq y. \end{cases}$$

We consider now the optimal stopping problem (*) and assume throughout the paper that the reward function G is non-negative, lower semicontinuous and satisfies

$$\mathbb{E}_x \left(\sup_{t \geq 0} e^{-\beta t} G(X_t) \right) < \infty \quad (6)$$

Our main result which characterizes solutions of (*) is as follows:

Theorem 2.4. (*One-sided case*). Assume that there exists an upper semicontinuous function $\hat{f} : S \mapsto \mathbb{R}$ and a point $x^* \in S$ such that

- (a) (i) $\hat{f}(x) \leq 0$ for $x \leq x^*$,
- (ii) $\hat{f}(x)$ is positive and non-decreasing for $x > x^*$.
- (b) (i) $\mathbb{E}_x \left(\sup_{0 \leq t \leq T} \hat{f}(X_t) \right) = G(x)$ for $x \geq x^*$
- (ii) $\mathbb{E}_x \left(\sup_{0 \leq t \leq T} \hat{f}(X_t) \right) \geq G(x)$ for $x \leq x^*$.

Then the value function of the optimal stopping problem (*) is given by

$$V(x) = \mathbb{E}_x \left(\sup_{0 \leq t \leq T} \hat{f}(X_t) \mathbf{1}_{\{X_t \geq x^*\}} \right) = \mathbb{E}_x \left(\hat{f}(M_T) \mathbf{1}_{\{M_T \geq x^*\}} \right) \quad (7)$$

and

$$\tau^* := \inf \{ t \geq 0 : X_t > x^* \}$$

is an optimal stopping time.

Proof. Since the function $x \mapsto \hat{f}(x) \mathbf{1}_{\{x \geq x^*\}}$ is non-decreasing, the second equality in (7) holds. Let $\hat{V}$ denote the function given via these expressions, i.e.,

$$\hat{V}(x) := \mathbb{E}_x \left(\sup_{0 \leq t \leq T} \hat{f}(X_t) \mathbf{1}_{\{X_t \geq x^*\}} \right).$$

By Lemma 2.2, we know that $\hat{V}$ is β -excessive, so that $(e^{-\beta t}\hat{V}(X_t))_{t \geq 0}$ is a non-negative càdlàg supermartingale (see [7] p. 101). Hence by optional sampling we have for all x and all stopping times τ

$$\hat{V}(x) \geq \mathbb{E}_x \left(e^{-\beta \tau} \hat{V}(X_\tau) \right). \quad (8)$$

By condition (a) we furthermore have $\hat{f}(x)1_{\{x \geq x^*\}} \geq \hat{f}(x)$ for all x , therefore using (b) we obtain from (8)

$$\hat{V}(x) \geq \mathbb{E}_x \left(e^{-\beta \tau} G(X_\tau) \right).$$

To conclude the proof we show that

$$\hat{V}(x) \leq \mathbb{E}_x \left(e^{-\beta \tau^*} G(X_{\tau^*}) \right). \quad (9)$$

The value function V being the smallest β -excessive majorant of G it then follows that $\hat{V} = V$ and τ^* is an optimal stopping time, as claimed. To prove (9) consider

$$\begin{aligned} \hat{V}(x) &= \mathbb{E}_x \left(\sup_{0 \leq t \leq T} \hat{f}(X_t) 1_{\{X_t \geq x^*\}} \right) \\ &= \mathbb{E}_x \left(\sup_{0 \leq t \leq T} \hat{f}(X_t) 1_{\{X_t \geq x^*\}}; \tau^* < T \right) \\ &= \mathbb{E}_x \left(\hat{f}(M_T) 1_{\{M_T \geq x^*\}}; \tau^* < T \right) \\ &= \mathbb{E}_x \left(\hat{f}(M_T); \tau^* < T \right) \\ &= \mathbb{E}_x \left(\hat{f} \left(\sup_{\tau^* \leq t \leq T} X_t \right); \tau^* < T \right) \\ &= \mathbb{E}_x \left(e^{-\beta \tau^*} \mathbb{E}_{X_{\tau^*}} \left(\hat{f} \left(\sup_{0 \leq t \leq T} X_t \right) \right) \right) \end{aligned}$$

by the strong Markov property. For the last equality note that by the memoryless property of the exponential distribution

$$\begin{aligned} &\mathbb{E}_x \left(\hat{f} \left(\sup_{\tau^* \leq t \leq T} X_t \right); \tau^* < T \right) \\ &= \int_0^\infty \mathbb{E}_x \left(\mathbb{E}_x \left(\hat{f} \left(\sup_{\tau^* \leq t \leq r} X_t \right) \middle| \mathcal{F}_{\tau^*} \right) 1_{\{\tau^* \leq r\}} \right) \mathbb{P}(T \in dr) \\ &= \mathbb{E}_x \left(\mathbb{E}_{X_{\tau^*}} \left(\int_0^\infty \hat{f} \left(\sup_{0 \leq t \leq r} X_t \right) \mathbb{P}(T \in dr) \right) e^{-\beta \tau^*} \right) \\ &= \mathbb{E}_x \left(e^{-\beta \tau^*} \mathbb{E}_{X_{\tau^*}} \left(\hat{f} \left(\sup_{0 \leq t \leq T} X_t \right) \right) \right). \end{aligned}$$

Since $\hat{f}(\sup_{0 \leq t \leq T} X_t) \leq \sup_{0 \leq t \leq T} \hat{f}(X_t)$ we obtain

$$\begin{aligned} \hat{V}(x) &\leq \mathbb{E}_x \left(e^{-\beta \tau^*} \mathbb{E}_{X_{\tau^*}} \left(\sup_{0 \leq t \leq T} \hat{f}(X_t) \right) \right) \\ &= \mathbb{E}_x \left(e^{-\beta \tau^*} G(X_{\tau^*}) \right), \end{aligned}$$

where we used (b)(ii) in the last step. This shows (9) and the claim is proved. $\square$

Remark 2.5. *By applying Theorem 2.4 to the process $-X$, the reward function $x \mapsto G(-x)$ and to the function $x \mapsto \hat{f}(-x)$ one obtains the analogous result for a threshold to the left applicable, e.g., to put-type options.*

Although one-sided stopping rules as in the previous theorem often arise naturally in many examples (especially for monotonic reward functions and discounting, see also [6] for a general discussion), for other problems a two-sided stopping rule is needed. The following theorem gives the corresponding result for this case.

Theorem 2.6. *(Two-sided case) Assume that there exists a lower semicontinuous function $\hat{f} : S \mapsto \mathbb{R}$ and points $x_*, x^* \in S$ such that*

- (a) (i) $\hat{f}(x) \leq 0$ for $x_* < x < x^*$,
- (ii) $\hat{f}(x)$ is non-increasing for $x < x_*$ and non-decreasing for $x > x^*$,
- (b) (i) $\mathbb{E}_x \left(\sup_{0 \leq t \leq T} \hat{f}(X_t) \right) = G(x)$ for $x \notin [x_*, x^*]$,
- (ii) $\mathbb{E}_x \left(\sup_{0 \leq t \leq T} \hat{f}(X_t) \right) \geq G(x)$ for $x \in [x_*, x^*]$.

Then the value function V of the optimal stopping problem () satisfies*

$$\begin{aligned} V(x) &= \mathbb{E}_x \left(\sup_{0 \leq t \leq T} \hat{f}(X_t) 1_{\{X_t \notin [x_*, x^*]\}} \right) \\ &= \mathbb{E}_x \left(\hat{f}(I_T) 1_{\{I_T \leq x_*\}} \vee \hat{f}(M_T) 1_{\{M_T \geq x^*\}} \right) \end{aligned} \tag{10}$$

and

$$\tau^* := \inf \{ t \geq 0 : X_t \notin [x_*, x^*] \}$$

is an optimal stopping time.

Proof. This theorem is proved similarly as Theorem 2.4. The second equality in (10) follows from the monotonic properties of $\hat{f}$. Letting

$$\hat{V}(x) := \mathbb{E}_x \left(\sup_{0 \leq t \leq T} \hat{f}(X_t) 1_{\{X_t \notin [x_*, x^*]\}} \right)$$

it is seen from Lemma 2.2 that $\hat{V}$ is β -excessive and, by conditions (a) and (b),

$$\hat{V}(x) \geq \mathbb{E}_x \left(e^{-\beta\tau} G(X_\tau) \right)$$

for all stopping times $\tau \in \mathcal{M}$. It remains to show that

$$\hat{V}(x) \leq \mathbb{E}_x \left(e^{-\beta\tau^*} G(X_{\tau^*}) \right). \quad (11)$$

For this, consider

$$\begin{aligned} \hat{V}(x) &= \mathbb{E}_x \left(\sup_{0 \leq t \leq T} \hat{f}(X_t) 1_{\{X_t \geq x^*\}} \vee \sup_{0 \leq t \leq T} \hat{f}(X_t) 1_{\{X_t \leq x_*\}}; \tau^* < T \right) \\ &= \mathbb{E}_x \left(\hat{f}(M_T) 1_{\{M_T \geq x^*\}} \vee \hat{f}(I_T) 1_{\{I_T \leq x_*\}}; \tau^* < T \right) \\ &= \mathbb{E}_x \left(\hat{f} \left(\sup_{\tau^* \leq t \leq T} X_t \right) \vee \hat{f} \left(\inf_{\tau^* \leq t \leq T} X_t \right); \tau^* < T \right) \\ &= \mathbb{E}_x \left(e^{-\beta\tau^*} \mathbb{E}_{X_{\tau^*}} \left(\hat{f} \left(\sup_{0 \leq t \leq T} X_t \right) \right) \vee \left(\hat{f} \left(\inf_{0 \leq t \leq T} X_t \right) \right) \right), \end{aligned}$$

where the last equality is obtained by the strong Markov property. Since

$$\hat{f} \left(\sup_{0 \leq t \leq T} X_t \right) \vee \hat{f} \left(\inf_{0 \leq t \leq T} X_t \right) \leq \sup_{0 \leq t \leq T} \hat{f}(X_t),$$

we have

$$\begin{aligned} \hat{V}(x) &\leq \mathbb{E}_x \left(e^{-\beta\tau^*} \mathbb{E}_{X_{\tau^*}} \left(\sup_{0 \leq t \leq T} \hat{f}(X_t) \right) \right) \\ &= \mathbb{E}_x \left(e^{-\beta\tau^*} G(X_{\tau^*}) \right). \end{aligned}$$

Thus the proof is complete. $\square$

Remark 2.7. *The methodologies in [9] and [26] seem to be applicable only for one-sided problems (for Lévy processes). It is remarkable that the present approach provides a clean characterization also in two-sided cases (for general Hunt processes).*

2.2 Computing $\hat{f}$

To apply Theorem 2.4 and 2.6 for a given reward function G the question arises how to find a function $\hat{f}$ fulfilling the conditions stated therein. For the related problem for underlying Lévy processes, Surya [26] gave a general way for sufficiently regular bounded functions G using the Fourier transform, but this seems to be specific for Lévy processes and is based on a certain assumption on the zeros of the Wiener-Hopf factors of X . We now discuss a more general way.

Let G be a given reward function and assume that for all x

$$\mathbb{E}_x\left(e^{-\beta t}G(X_t)\right) \rightarrow 0 \quad \text{as } t \rightarrow \infty. \quad (12)$$

Note that, by dominated convergence, (12) is fulfilled if in addition to (6) it also holds

$$\lim_{t \rightarrow \infty} e^{-\beta t}G(X_t) = 0 \quad \mathbb{P}_x\text{-a.s.}$$

Furthermore, assume that the function G is in the domain of the extended infinitesimal generator of X killed at T (cf. [21, VII.1]). Then, by definition, there exists a function $\tilde{f}$ such that

$$\int_0^t |\tilde{f}(X_s)| ds < \infty \quad \mathbb{P}_x\text{-a.s.}$$

and the process

$$\left(e^{-\beta t}G(X_t) - G(X_0) + \int_0^t e^{-\beta s}\tilde{f}(X_s)ds\right)_{t \geq 0}$$

is a $(\mathcal{F}_t, \mathbb{P}_x)$ -martingale. Consequently,

$$\mathbb{E}_x\left(e^{-\beta t}G(X_t)\right) - G(x) + \mathbb{E}_x\left(\int_0^t e^{-\beta s}\tilde{f}(X_s)ds\right) = 0,$$

and as $t \rightarrow \infty$ we obtain using (12)

$$G(x) = \lim_{t \rightarrow \infty} \mathbb{E}_x\left(\int_0^t e^{-\beta s}\tilde{f}(X_s)ds\right).$$

Hence, if, e.g.,

$$\mathbb{E}_x\left(\int_0^\infty e^{-\beta s}|\tilde{f}(X_s)|ds\right) < \infty,$$

the Lebesgue dominated convergence theorem yields

$$G(x) = \frac{1}{\beta} \mathbb{E}_x\left(\tilde{f}(X_T)\right). \quad (13)$$

Notice that for G in the domain of the infinitesimal generator A of X we have $\tilde{f} = -(A - \beta)G$. In this case equation (13) reflects one of the fundamental properties of the resolvent operator.

Remark 2.8. *For optimal stopping problems for Lévy processes it was assumed in [9] that a representation of type (13) (see also (2)) holds for G . Hence, the above considerations are useful when applying the methods in [9].*

Imitating the approach for optimal stopping of Lévy processes, we next express G as the expectation of a function of M_T (cf. (1) and (2)). Obviously, it holds

$$\mathbb{E}_x(\tilde{f}(X_T)) = \mathbb{E}_x(\mathbb{E}_x(\tilde{f}(X_T)|M_T)).$$

We assume that the regular conditional probability distribution

$$\mathbb{P}_x(X_T \in dy | M_T = z) := \mathbb{P}_x(X_T \in dy, M_T \in dz) / \mathbb{P}_x(M_T \in dz) \quad (14)$$

exists and has the needed regularity properties so that it holds

$$\frac{1}{\beta} \mathbb{E}_x(\tilde{f}(X_T)) = \mathbb{E}_x(Q(M_T; x)), \quad (15)$$

where

$$Q(z; x) := \frac{1}{\beta} \int_{-\infty}^z \tilde{f}(y) \mathbb{P}_x(X_T \in dy | M_T = z). \quad (16)$$

These observations are now applied to characterize solutions for optimal stopping problems for Lévy processes and linear diffusions via Theorem 2.4. Lévy processes are discussed first. The following lemma is well known. We present a short proof for the sake of completeness.

Lemma 2.9. *For a Lévy process X (which is not a subordinator or a compound Poisson process) it holds that*

$$\mathbb{P}_x(X_T \in dy | M_T = z) = \mathbb{P}_z(I_T \in dy). \quad (17)$$

In particular, the conditional distribution does not depend on x .

Proof. The claim follows from the Wiener-Hopf factorization (3) and the spatial homogeneity of X . Indeed,

$$\begin{aligned} \mathbb{P}_x(X_T \in dy, M_T \in dz) &= \mathbb{P}_0(x + X_T \in dy, x + M_T \in dz) \\ &= \mathbb{P}_0(x + M_T + I_T^\circ \in dy, x + M_T \in dz) \\ &= \mathbb{P}_0(z + I_T^\circ \in dy, x + M_T \in dz) \\ &= \mathbb{P}_z(I_T \in dy) \mathbb{P}_x(M_T \in dz). \end{aligned}$$

□

Proposition 2.10. *Consider the optimal stopping problem (*) in case of a Lévy process X (as introduced in Lemma 2.9). Assume that there exists a function $\tilde{f}$ such that (13) holds and define*

$$Q(z) := \frac{1}{\beta} \int_{-\infty}^0 \tilde{f}(z+y) \mathbb{P}_0(I_T \in dy). \quad (18)$$

If Q fulfils (a) in Theorem 2.4 then the value function of the optimal stopping problem () is given as in (7) with $\hat{f} = Q \vee 0$.*

Proof. Since Q is assumed to fulfil (a) in Theorem 2.4 also $\hat{f}$ fulfils this condition. We need to verify that $\hat{f}$ satisfies (b) in Theorem 2.4. Definition (18) of Q , Lemma 2.9, and formula (15) yield

$$G(x) = \frac{1}{\beta} \mathbb{E}_x(\tilde{f}(X_T)) = \mathbb{E}_x(Q(M_T)).$$

Consequently, since $Q \leq \hat{f}$ we have for all x

$$G(x) = \mathbb{E}_x(Q(M_T)) \leq \mathbb{E}_x(\hat{f}(M_T)) \leq \mathbb{E}_x\left(\sup_{0 \leq t \leq T} \hat{f}(X_t)\right). \quad (19)$$

In case $x > x^*$ - since Q satisfies (a) in Theorem 2.4 - (19) holds with equalities instead of inequalities, and this completes the proof. $\square$

Next we consider the optimal stopping problem (*) for a regular linear diffusion. For simplicity, it is assumed that X is conservative and takes values on the whole $\mathbb{R}$. The following result is extracted from [3, II.19]. We use the notation from Example 2.3.

Lemma 2.11. *For a conservative regular linear diffusion X on $\mathbb{R}$ it holds that*

$$\mathbb{P}_x(X_T \in dy | M_T = z) = \frac{\psi_\beta(y) m(dy)}{\int_{-\infty}^z \psi_\beta(u) m(du)}, \quad (20)$$

where m is the speed measure. In particular, the conditional distribution does not depend on x .

The next result can now be proved analogously as Proposition 2.10.

Proposition 2.12. *Consider the optimal stopping problem (*) for a diffusion X as introduced above. Assume that there exists a function $\tilde{f}$ such that (13) holds and define*

$$Q(z) := \frac{1}{\beta} \frac{\int_{-\infty}^z \tilde{f}(y) \psi_\beta(y) m(dy)}{\int_{-\infty}^z \psi_\beta(y) m(dy)}. \quad (21)$$

If Q fulfils (a) in Theorem 2.4 then the value function of the optimal stopping problem () is given as in (7) with $\hat{f} = Q \vee 0$.*

To apply Proposition 2.12 it is useful to have condition on hand to guarantee that Q fulfills (a) in Theorem 2.4. The following Lemma gives such a condition in terms of the function $\tilde{f}$ which is known explicitly in many situations of interest.

Lemma 2.13. *In the setting of Proposition 2.12 assume that there exists $\tilde{x}$ such that $\tilde{f}(x) \leq 0$ for $x \leq \tilde{x}$ and $\tilde{f}(x)$ is non-decreasing for $x \geq \tilde{x}$. Then $Q \leq 0$ or there exists x^* such that Q fulfils (a) in Theorem 2.4.*

Proof. For simplicity we assume that the measure m is absolutely continuous, i.e. $m(dy) = m(y)dy$, and let

$$x^* = \inf\{z : \int_{-\infty}^z \tilde{f}(y)\psi_{\beta}(y) m(y)dy > 0\}.$$

If $x^* = \infty$, then the claim trivially holds. Therefore, we assume $x^* < \infty$. It remains to prove that Q is non-decreasing for $x > x^*$. Obviously $x^* > \tilde{x}$ and therefore by the assumption on $\tilde{f}$ we get

$$\int_{-\infty}^z (\tilde{f}(z) - \tilde{f}(y))\psi_{\beta}(y) m(y)dy \geq 0.$$

Using this fact, by differentiating the function Q we immediately see that $\frac{d}{dx}Q(x) \geq 0$. $\square$

Remark 2.14. *It is critical that the function Q defined in (18) for Lévy processes and in (21) for diffusions does not depend on x . This clearly follows from the fact that the conditional distribution of X_T given M_T does not depend on the initial state of the process. Path decomposition theorems for strong Markov processes imply such results in general, see Millar [16]. This observation yields a method for finding $\hat{f}$ for general underlying processes.*

2.3 Connection to the representing measure approach

Another approach to the solution of optimal stopping problems is the observation that finding the solution of such problems is equivalent to finding the Choquet-type integral representation of the value function in terms of the resolvent kernel $u_{\beta}(x, y)$. The Radon measure σ that appears in this representation characterizes the β -excessive value function V , and furthermore, the stopping region can be described as the support of σ .

This approach was first discussed in [22] for optimal stopping problems of diffusions. In [18] a verification theorem based on this approach applicable for general Hunt processes is presented. Moreover, in [18] the appearance of a function of the maximum in the solution of optimal stopping problems for

Lévy processes is explained via this approach and the Wiener-Hopf factorization.

The critical point to utilize the approach is to find a candidate solution for the measure σ . In [23], this was carried out explicitly for spectrally positive Lévy processes in the Novikov-Shiryaev problem in terms of the Appell polynomials of X_T using the Laplace transforms. In Proposition 2.15 we extend these results by pointing out a connection between this approach and our approach in this paper for a general Hunt process X which does not have negative jumps, that is,

$$\mathbb{P}_x\left(\inf_{t \geq 0}(X_t - X_{t-}) \geq 0\right) = 1 \quad \forall x \in S. \quad (22)$$

We work in the setting of Hunt processes satisfying the set of assumptions as given in [18, Section 2]. In particular, it is assumed that there exists a dual resolvent with respect to a duality measure m and that Hypothesis (B) from [13] holds. Moreover, we assume that the Laplace transform of the first hitting time $H_x := \inf\{t : X_t = x\}$ admits for all $x_0 < x$ the representation.

$$\mathbb{E}_x(e^{-\beta H_{x_0}}) = \frac{u_\beta(x, x_0)}{u_\beta(x_0, x_0)} \quad \forall x_0 < x \in S, \quad (23)$$

where $T_x = \inf\{t > 0 : X_t = x\}$. This representation is valid for a large class of processes. A sufficient condition is that (I) and (II) below hold true, see [2, Chapter V.3 and Exercise VI.4.18].

- (I) The mapping $x \mapsto \int_S u_\beta(x, y)f(y)m(dy)$ is continuous, where f is a bounded and measurable function with compact support.
- (II) Every point $x \in S$ is regular for $\{x\}$, that is

$$\mathbb{P}_x(T_x = 0) > 0, \quad (24)$$

where $T_x = \inf\{t > 0 : X_t = x\}$.

Assumption (I) is (4.1) in [2, p. 284], and assumption (II) implies that X has a local time at every point of the state space. Condition (23) furthermore holds true whenever X is a spectrally positive Lévy process, see [14, Theorem 3.12, Corollary 8.9].

Note that, in case X have positive jumps, overshoot occurs in the problem. Nonetheless it is possible to find an explicit representation of the measure σ .

Proposition 2.15. *Let X be a Hunt process as described above with no negative jumps that fulfills condition (23). Assume that for the optimal stopping*

problem (*) there exists a function $\tilde{f}$ such that (13) holds. Moreover, assume that the function

$$Q(z) := \frac{1}{\beta} \int_{-\infty}^z \tilde{f}(y) \mathbb{P}(X_T \in dy | M_T = z)$$

fulfills (a) in Theorem 2.4. Then the value function V of (*) has the representation

$$V(x) = \int_{[x^*, \infty) \cap S} u_\beta(x, y) \sigma(dy) \quad \forall x \in S,$$

where the measure σ is given by

$$\sigma(dy) = \begin{cases} \tilde{f}(y)m(dy), & \text{for } y > x^*, \\ 0, & \text{for } y < x^*, \end{cases} \quad (25)$$

and

$$\sigma(\{x^*\}) = \frac{\mathbb{E}_{x^*}(\tilde{f}(X_T)1_{\{X_T \leq x^*\}})}{\beta u_\beta(x^*, x^*)}. \quad (26)$$

In the particular case that X is a spectrally positive Lévy process, then $\sigma(\{x^*\}) = 0$, and

$$V(x) = E_x(\tilde{f}(X_T)1_{\{X_T > x^*\}}).$$

Proof. We remark first that similarly as in the proof of Proposition 2.10 it can be verified that the value function V of the problem satisfies the equality

$$V(x) = \mathbb{E}_x(Q(M_T)1_{\{M_T \geq x^*\}}),$$

where x^* is the point associated with Q via (a) in Theorem 2.4. Now we have

$$\begin{aligned} V(x) &= \mathbb{E}_x(Q(M_T)1_{\{M_T \geq x^*\}}) = \frac{1}{\beta} \mathbb{E}_x\left(\mathbb{E}_x(\tilde{f}(X_T)|M_T)1_{\{M_T \geq x^*\}}\right) \\ &= \frac{1}{\beta} \mathbb{E}_x\left(\mathbb{E}_x(\tilde{f}(X_T)1_{\{M_T \geq x^*\}}|M_T)\right) = \frac{1}{\beta} \mathbb{E}_x(\tilde{f}(X_T)1_{\{M_T \geq x^*\}}) \\ &= \frac{1}{\beta} \mathbb{E}_x(\tilde{f}(X_T)1_{\{X_T > x^*\}}) + \frac{1}{\beta} \mathbb{E}_x(\tilde{f}(X_T)1_{\{M_T \geq x^*, X_T \leq x^*\}}). \end{aligned}$$

Since it is assumed that X moves continuously down we rewrite the second summand

$$\mathbb{E}_x(\tilde{f}(X_T)1_{\{M_T \geq x^*, X_T \leq x^*\}}) = \mathbb{E}_x(\tilde{f}(X_T)1_{\{T_{x^*} < T, X_T \leq x^*\}}),$$

and, by the strong Markov property,

$$\begin{aligned}\mathbb{E}_x \left(\tilde{f}(X_T) 1_{\{M_T \geq x^*, X_T \leq x^*\}} \right) &= \mathbb{E}_x \left(\mathbb{E}_x \left(\tilde{f}(X_T) 1_{\{X_T \leq x^*\}} | \mathcal{F}_{T_{x^*}} \right) 1_{\{T_{x^*} < T\}} \right) \\ &= \mathbb{E}_x \left(e^{-\beta T_{x^*}} \mathbb{E}_{X_{T_{x^*}}} \left(\tilde{f}(X_T) 1_{\{X_T \leq x^*\}} \right) \right) \\ &= \mathbb{E}_{x^*} \left(\tilde{f}(X_T) 1_{\{X_T \leq x^*\}} \right) \mathbb{E}_x \left(e^{-\beta T_{x^*}} \right).\end{aligned}$$

Letting $c = \mathbb{E}_{x^*} \left(\tilde{f}(X_T) 1_{\{X_T \leq x^*\}} \right)$ and using assumption (23) we obtain

$$\begin{aligned}V(x) &= \frac{1}{\beta} \mathbb{E}_x \left(\tilde{f}(X_T) 1_{\{X_T > x^*\}} \right) + \frac{1}{\beta} \mathbb{E}_x \left(\tilde{f}(X_T) 1_{\{M_T \geq x^*, X_T \leq x^*\}} \right) \\ &= \int_{(x^*, \infty) \cap S} \tilde{f}(y) u_\beta(x, y) m(dy) + \frac{c}{\beta u_\beta(x^*, x^*)} u_\beta(x, x^*) \\ &= \int_{[x^*, \infty) \cap S} u_\beta(x, y) \sigma(dy)\end{aligned}$$

with σ as in (25) and in (26). It remains to prove that σ is indeed non-negative. To this end, we use the fact that each β -excessive function u locally integrable with respect to the duality measure m can be decomposed uniquely in the form

$$u(x) = \int u_\beta(x, y) \sigma_u(dy) + h_u(x), \quad (27)$$

where σ_u is a non-negative Radon measure and h_u is a β -harmonic function, see [13, Theorem 2 p. 505 and Proposition 13.1 p. 523]. Since the value function V is β -excessive we may apply (27) to obtain

$$\int u_\beta(x, y) \sigma_V(dy) + h_V(x) = V(x) = \int_{[x^*, \infty) \cap S} u_\beta(x, y) \sigma(dy),$$

where σ_V is the non-negative Radon measure associated with V . By the Hahn decomposition we can write $\sigma - \sigma_V = \sigma^+ - \sigma^-$ for some non-negative measures σ^+ and σ^- . Therefore,

$$h_V(x) + \int u_\beta(x, y) \sigma^-(dy) = \int u_\beta(x, y) \sigma^+(dy).$$

By the uniqueness of the decomposition in (27) we obtain $h_V = 0$ and $\sigma^+ = \sigma^-$. Hence, $\sigma = \sigma_V$ is indeed a non-negative measure. This proves the first claim.

For spectrally positive Lévy processes it is well known that $-I_T$ is $Exp(\lambda)$ -distributed, where λ is the unique positive root of the equation

$$\mathbb{E}(e^{\lambda X_1}) = e^\beta.$$

Therefore, by the definition of Q and Proposition 17 we see that Q is continuous. In particular, $Q(x^*) = 0$. Using the Wiener-Hopf factorization (3) we obtain

$$\begin{aligned}\mathbb{E}_{x^*}(\tilde{f}(X_T)1_{\{X_T \leq x^*\}}) &= \mathbb{E}_0(\tilde{f}(x^* + M_T + I_T^\circ)1_{\{x^* + M_T + I_T^\circ \leq x^*\}}) \\ &= \int_0^\infty \mathbb{E}_0(\tilde{f}(x^* + t + I_T^\circ)1_{\{t + I_T^\circ \leq 0\}})\mathbb{P}(M_T \in dt).\end{aligned}$$

Therefore, $\mathbb{E}_{x^*}(\tilde{f}(X_T)1_{\{X_T < x^*\}}) = 0$, since

$$\begin{aligned}\mathbb{E}_0(\tilde{f}(x^* + t + I_T^\circ)1_{\{t + I_T^\circ \leq 0\}}) &= \int_{-\infty}^{-t} \tilde{f}(x^* + t + s)\lambda e^{\lambda s} ds \\ &= e^{-\lambda t} \int_{-\infty}^0 \tilde{f}(x^* + y)\lambda e^{\lambda y} dy \\ &= e^{-\lambda t} Q(x^*) = 0.\end{aligned}$$

Consequently,

$$\sigma(\{x^*\}) = \frac{\mathbb{E}_{x^*}(\tilde{f}(X_T)1_{\{X_T \leq x^*\}})}{\beta u_\beta(x^*, x^*)} = 0.$$

□

Remark 2.16. *By inspecting the previous proof one can see that assumption (23) can be weakened by assuming that (23) holds for $x_0 = x^*$ only.*

3 Examples

In this section we demonstrate the applicability of the above theory by solving some optimal stopping problems for diffusions, Lévy processes, and continuous time Markov chains.

3.1 Ornstein-Uhlenbeck process

Let X be an Ornstein-Uhlenbeck diffusion with parameter $\gamma > 0$, i.e. the differential operator associated to X is

$$Af = \frac{1}{2} \frac{d^2}{dx^2} f - \gamma x \frac{d}{dx} f.$$

We consider the optimal stopping problem (*) for X with the reward function

$$G_n(x) = (x^+)^n, \quad n \in \mathbb{N}.$$

For general Lévy-driven Ornstein-Uhlenbeck processes it was proved in [6] that the optimal stopping rule is one-sided, but no explicit way for finding the boundary explicitly in general was given there. Here we want characterize the optimal boundary for the Ornstein-Uhlenbeck diffusion using our approach. We have

$$(\beta - A)x^n = x^{n-2} \left((\beta + \gamma n)x^2 - \frac{n(n-1)}{2} \right) =: \tilde{f}_n(x).$$

Using Itô's lemma it is seen that $\tilde{f}_n$ indeed has the mean value property for X_T , i.e.,

$$x^n = \mathbb{E}_x \left(\tilde{f}_n(X_T) \right) \quad \forall x$$

Recall that for Ornstein-Uhlenbeck processes the speed measure m is given by

$$m(dx) = 2e^{-\gamma x^2} dx$$

and

$$\psi_\beta(x) = e^{\gamma x^2/2} \mathcal{D}_{-\beta/\gamma}(-x\sqrt{2\gamma}),$$

where $\mathcal{D}_\nu$ denotes the parabolic cylinder function, see [3, A1.24,II.7]. Defining $Q_n = Q$ as in (21) we obtain

$$x^n = \mathbb{E}_x \left(Q_n(M_T) \right).$$

Using arguments similar to those in Lemma 2.13 it is easy to see that Q has a unique positive root x^* , which is given by the positive solution of the equation

$$\int_{-\infty}^x \tilde{f}(y) e^{-\gamma y^2/2} \mathcal{D}_{-\beta/\gamma}(-y\sqrt{2\gamma}) dy = 0. \quad (28)$$

An approximative value of x^* can be found using numerical methods. Notice that Q_n can also attain positive values on the negative half-line. However, this problem can be dealt with by considering the function

$$\hat{f}(x) = \begin{cases} 0, & x \leq 0, \\ Q(x), & x > 0, \end{cases}$$

which obviously satisfies the requirements of Theorem 2.4. Therefore, we obtain that the stopping time

$$\tau^* = \inf\{t : X_t \geq x^*\}$$

is optimal.

3.2 American put in a Lévy market

In this subsection we review the well known example of a perpetual American put option in a Lévy driven market in the light of our approach (cf. the solutions in [17] and [1], see also [5] for further discussions). Although the solution is well-known we want to demonstrate how to find the representing function $\hat{f}$ systematically with the approach discussed in Subsection 2.2. Since the reward function $G(x) = K - e^x$, $K > 0$, is non-increasing we use the modified version of Theorem 2.4, see Remark 2.5. The American call option can be handled similarly.

The generator A of X is given by

$$Af(x) = \frac{c^2}{2} \frac{d^2}{dx^2} f(x) + b \frac{d}{dx} f(x) + \int_{\mathbb{R}} \left(f(x+y) - f(x) - y \frac{d}{dx} f(x) 1_{\{|y|<1\}} \right) \pi(dy), \quad (29)$$

where (b, c, π) is the Lévy triple of X and f is assumed to be in the domain of A . We wish to apply $\beta - A$ to $G(x) = K - e^x$. For this ansatz to work we assume that

$$\mathbb{E}_0(e^{X_1}) < e^\beta. \quad (30)$$

At the end we will see that the formula obtained is a solution for general X . We have

$$\begin{aligned} (\beta - A)G(x) &= \beta(K - e^x) + e^x \left(\frac{c^2}{2} + b + \int_{\mathbb{R}} (e^y - 1 - y 1_{|y|<1}) \pi(dy) \right) \\ &= \beta K - (\beta - \psi(1))e^x =: \tilde{f}(x), \end{aligned}$$

where $\psi(1) = \log \mathbb{E}_0(e^{X_1}) < \beta$. To find $\hat{f}$ we calculate, cf. Proposition 2.10,

$$Q(x) = \frac{1}{\beta} \mathbb{E}_0(\tilde{f}(x + M_T)) = K - \frac{\beta - \psi(1)}{\beta} \mathbb{E}_0(e^{M_T}) e^x.$$

By noting that

$$\frac{\beta - \psi(1)}{\beta} = \frac{1}{\mathbb{E}_0(e^{X_T})} \quad (31)$$

and applying the Wiener-Hopf factorization (3) we obtain

$$Q(x) = K - \mathbb{E}_0(e^{I_T})^{-1} e^x.$$

Now we have a candidate for $\hat{f}$ by our ansatz, i.e., $\hat{f} = Q \vee 0$, and formula (31) also makes sense for general Lévy processes without the integrability assumption (30). Indeed, one immediately sees that for general X we have

$$K - e^x = \mathbb{E}_x(Q(I_T)).$$

We can apply Theorem 2.4 (together with Remark 2.5) to $\hat{f}$ and obtain that $x^* = \log(K\mathbb{E}_0(e^{I_T}))$ is the optimal stopping boundary and the value function is given by

$$V(x) = \mathbb{E}_x \left(\hat{f}(I_T) 1_{\{I_T < x^*\}} \right) = \mathbb{E}_x \left(\left(K - \frac{e^{I_T}}{\mathbb{E}_0(e^{I_T})} \right)^+ \right).$$

3.3 Novikov-Shiryaev optimal stopping problem

In [19] Novikov and Shiryaev solved the optimal stopping problem (*) with $G_n(x) = (x^+)^n$, $n = 1, 2, \dots$, for random walks and in [15] Kyprianou and Surya found the solution for Lévy processes. Here we discuss shortly how the solution can be found with the methodology presented in this paper. For simplicity, we assume that $\beta > 0$.

Let X be a Lévy process with the generator A as given in (29). We fix the exponent $n \in \{1, 2, \dots\}$ for the reward function G_n and assume that the Lévy measure π satisfies

$$\int_{(-\infty, -1) \cup (1, +\infty)} |y|^n \pi(dy) < \infty. \quad (32)$$

This condition implies that $\mathbb{E}_x(|X_t|^k) < \infty$ for all x , $t \geq 0$, and $k = 1, 2, \dots, n$. Using the independence of increments, it is easily seen that $\mathbb{E}_x(X_t^n)$ is a polynomial in t . Consequently, also the moments of X_T up to n exist.

Remark 3.1. *In [15] the solution is found under the weaker condition that (only) the integral over $(1, +\infty)$ in (32) is finite. Under this weaker condition the moments of M_T up to n exist but the moments of I_T may not exist. Working under the stronger condition reveals the rôle of the Appell polynomials of X_T which we find interesting and important. In fact, since the function Q_n , see (35), that is, the Appell polynomial associated with M_T , found in this way satisfies conditions (a) and (b) in Theorem 2.4 it is clear that our solution is indeed valid under the weaker condition and, hence, coincides with the solution in by Novikov and Shiryaev.*

We follow the receipt in Section 2.2 and operate with $\beta - A$ on x^n . This yields

$$\begin{aligned} \tilde{f}_n(x) &:= (\beta - A)(x^n) \\ &= \beta x^n - \left(\frac{c^2}{2} n(n-1) x^{n-2} + b n x^{n-1} \right) \\ &\quad - \int_{\mathbb{R}} \left((x+y)^n - x^n - y n x^{n-1} 1_{\{|y| \leq 1\}} \right) \pi(dy), \end{aligned}$$

where the integral term is well defined due to (32). Clearly, $\tilde{f}_n$ is an n th order polynomial and satisfies the DE

$$\frac{d}{dx}\tilde{f}_n(x) = n\tilde{f}_{n-1}(x) \quad (33)$$

with $f_0 = 1$. We claim that

$$\mathbb{E}_x\left(\tilde{f}_n(X_T)\right) = \mathbb{E}_0\left(\tilde{f}_n(X_T + x)\right) = \beta x^n, \quad (34)$$

which, then, together with (33) shows that $x \mapsto \tilde{f}_n(x)/\beta$ is the n th order Appell polynomial associated with X_T , (for properties of the Appell polynomials, see [19] and [23]). To prove (33) we use standard computations, legitimate by (32), with the Itô formula for Lévy processes, see, e.g., [12, p. 612], to obtain

$$\begin{aligned} e^{-\beta t}\mathbb{E}_x(X_t^n) - x^n &= \mathbb{E}_x\left(\int_0^t e^{-\beta s}\left[\frac{c^2}{2}n(n-1)X_s^{n-2} + bnX_s^{n-1} - \beta X_s^n\right] ds\right) \\ &\quad + \mathbb{E}_x\left(\int_0^t e^{-\beta s}\int_{\mathbb{R}}[(X_s + y)^n - X_s^n - ynX_s^{n-1}1_{\{|y|\leq 1\}}]\pi(dy) ds\right). \end{aligned}$$

Multiplying both sides with β , letting $t \rightarrow \infty$ and using that the moments of X_t are polynomials in t we obtain

$$\begin{aligned} -\beta x^n &= \mathbb{E}_x\left(\int_0^\infty \beta e^{-\beta s}\left[\frac{c^2}{2}n(n-1)X_s^{n-2} + bnX_s^{n-1} - \beta X_s^n\right] ds\right) \\ &\quad + \mathbb{E}_x\left(\int_0^\infty \beta e^{-\beta s}\int_{\mathbb{R}}[(X_s + y)^n - X_s^n - ynX_s^{n-1}1_{\{|y|\leq 1\}}]\pi(dy) ds\right), \end{aligned}$$

which proves (34). Next we apply Proposition 2.10 to find the function Q_n and then $\hat{f}$. We write formula (18) therein as

$$Q_n(x) = \int_{-\infty}^0 Q_n^{(X_T)}(x+y)\mathbb{P}_0(I_T \in dy), \quad (35)$$

where $Q_n^{(X_T)}(x) := \tilde{f}_n(x)/\beta$. Recall from [23] the identity

$$Q_n^{(X_T)}(x+y) = \sum_{k=0}^n \binom{n}{k} Q_k^{(M_T)}(x) Q_{n-k}^{(I_T)}(y) \quad (36)$$

where $Q_i^{(M_T)}$ and $Q_i^{(I_T)}$ denote the i th Appell polynomials of M_T and I_T , respectively. Identity (36) follows easily from the Wiener-Hopf factorization

and the definition of the Appell polynomials. Substituting (36) into (35) and using the normalizations

$$Q_0^{(I_T)} \equiv 1 \quad \text{and} \quad \mathbb{E}_0\left(Q_i^{(I_T)}(I_T)\right) = 0 \text{ for all } i = 1, 2, \dots, n$$

yields $Q_n = Q_n^{(M_T)}$. Hence, our approach leads to the same characterization of the optimal stopping time in terms of $Q_n^{(M_T)}$ as in [19]. The crucial property to be able to apply Theorem 2.4, i.e., that $Q_n^{(M_T)}$ has a unique positive zero x^* and is non-decreasing for $x > x^*$, is proved in [19] Lemma 5. Consequently, it now easily follows that

$$\hat{f}(x) := \begin{cases} Q_n^{(M_T)}(x), & x > 0, \\ 0, & x \leq 0, \end{cases}$$

satisfies conditions (a) and (b) in Theorem 2.4 and the solution of the optimal stopping problem results.

Note that Proposition 2.15 together with the fact that $x \mapsto \tilde{f}_n(x)/\beta$ is the n th order Appell polynomial associated with X_T reproduces the formula of the representing measure for spectrally positive Lévy processes obtained in [23, Corollary 8].

3.4 Absorbing Brownian motion

Let X be a Brownian motion on $\mathbb{R}_+$ absorbed at 0, i.e., $X_t = 0$ for $t \geq H_0$. We solve the optimal stopping problem (*) with $G_n(x) = x^n$, $n \geq 1$. The first step is to find $\tilde{f}_n$ such that (13) holds. Guided by Subsection 2.2 we set

$$\tilde{f}_n(x) = \left(\beta - \frac{1}{2} \frac{d^2}{dx^2}\right) G_n(x) = x^{n-2} \left(\beta x^2 - \frac{n(n-1)}{2}\right) =: \beta P_n(x). \quad (37)$$

Instead of checking (13) via a direct calculation, we recall that P_n is the n th order Appell polynomial associated with a standard Brownian motion $(W_t)_{t \geq 0}$ started from 0 and evaluated at T , see [23, Example 3.9]. Hence, for all x

$$x^n = \mathbb{E}_0\left(P_n(x + W_T)\right) = \mathbb{E}_x\left(P_n(W_T)\right) = \frac{1}{\beta} \mathbb{E}_x\left(\tilde{f}_n(W_T)\right),$$

and, by the strong Markov property,

$$\begin{aligned} \mathbb{E}_x\left(\tilde{f}_n(W_T)\right) &= \mathbb{E}_x\left(\tilde{f}_n(W_T)1_{\{T < H_0\}}\right) + \mathbb{E}_x\left(\tilde{f}_n(W_T)1_{\{T > H_0\}}\right) \\ &= \mathbb{E}_x\left(\tilde{f}_n(X_T)1_{\{X_T > 0\}}\right) + \mathbb{E}_x\left(\mathbb{E}_x(\tilde{f}_n(W_T) \mid W_{H_0})1_{\{T > H_0\}}\right) \\ &= \mathbb{E}_x\left(\tilde{f}_n(X_T)1_{\{X_T > 0\}}\right). \end{aligned}$$

Consequently, we have

$$G_n(x) = \frac{1}{\beta} \mathbb{E}_x \left(\tilde{f}_n(X_T) 1_{\{X_T > 0\}} \right),$$

which is (13) if $n \neq 2$. To cover also the case $n = 2$ we abandon (13) and instead try to find a function Q_n such that

$$\mathbb{E}_x \left(\tilde{f}_n(X_T) 1_{\{X_T > 0\}} \right) = \mathbb{E}_x \left(Q_n(M_T) \right).$$

For this, the conditional distribution of X_T given M_T is needed, see (16). Since X is not conservative, this distribution has an atom at 0 and, hence, the formula in Lemma 2.11 is not valid. However, from [3, II.19] we may deduce

$$\mathbb{P}_x(M_T \in dz)/dz = \frac{\partial}{\partial z} \left(\frac{\psi_\beta(x)}{\psi_\beta(z)} \right)$$

and

$$\mathbb{P}_x(M_T \in dz, X_T \in dy) = 2 \frac{\psi_\beta(x) \psi_\beta(y)}{\psi_\beta^2(z)} dz dy.$$

where $\psi_\beta(x) = \text{sh}(x\sqrt{2\beta})$, $x \geq 0$. see *ibid.* p. 121. Consequently, the conditional density of X_T given M_T is given by

$$f_{X_T}(y|M_T = z) = \frac{2\psi_\beta(y)}{\psi'_\beta(z)}.$$

and, therefore,

$$\begin{aligned} Q_n(z) &:= \frac{1}{\beta} \int_0^z \tilde{f}_n(y) \frac{2\psi_\beta(y)}{\psi'_\beta(z)} dy \\ &= \frac{2}{\beta \sqrt{2\beta} \text{ch}(z\sqrt{2\beta})} \int_0^z y^{n-2} \left(\beta y^2 - \frac{n(n-1)}{2} \right) \text{sh}(y\sqrt{2\beta}) dy. \end{aligned}$$

Since the integrand changes its sign from $-$ to $+$ one can see (after some easy calculations) that Q_n satisfies (a) of Theorem 2.4. Although Proposition 2.12 is not directly applicable it can be proved analogously as therein that (b) of Theorem 2.4 holds for $\hat{f}_n := Q_n \vee 0$. Therefore, the optimal stopping boundary x^* is given as the unique positive solution to the equation

$$\int_0^z y^{n-2} \left(\beta y^2 - \frac{n(n-1)}{2} \right) \text{sh}(y\sqrt{2\beta}) dy = 0. \quad (38)$$

Integrating by parts, (38) is seen to be equivalent with

$$z \text{ch}(z\sqrt{2\beta}) - \frac{n}{\sqrt{2\beta}} \text{sh}(z\sqrt{2\beta}) = 0. \quad (39)$$

Notice that for $n = 1$ this equation does not have a positive solution and the optimal stopping rule is to stop immediately – in other words, reward $G(x) = x$ is β -excessive.

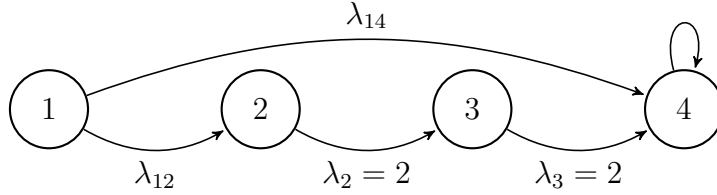
It is also interesting to observe that (39) is of the form

$$G_n(z) \frac{d\psi(z)}{dz} - \psi(z) \frac{dG_n(z)}{dz} = 0,$$

and, hence, our approach yields in this case the same characterization of the stopping point as given in Remark (ii) p. 97 in [22].

3.5 Markov chain

To see the general applicability of the theory we treat in this subsection an easy example with an underlying finite-state Markov chain. Let $(X_t)_{t \geq 0}$ be a continuous-time Markov chain with state space $E = \{1, \dots, 4\}$ and transition rates as given in the following figure.



To obtain explicit results let us assume that $\beta = 1$ and $G(1) = 3$, $G(2) = 1 = G(3)$ and $G(4) = 2$. We could use the method described in Subsection 2.2 to find $\hat{f}$ such that $G(x) = \mathbb{E}_x(\sup_{t \leq T} \hat{f}(X_t))$ for all $x \in E$, but it may be instructive to find it directly: Since 4 is an absorbing state, we take $\hat{f}(4) = G(4) = 2$. To find $\hat{f}(3)$ we make the ansatz that $\hat{f}(3) \leq \hat{f}(4)$ and obtain

$$1 = G(3) = \mathbb{E}_3 \left(\sup_{t \leq T} \hat{f}(X_t) \right) = \hat{f}(4) \mathbb{P}_3(S_3 \leq T) + \hat{f}(3) \mathbb{P}_3(S_3 > T),$$

where the sojourn time S_3 at state 3 is exponentially distributed with parameter 2 and independent of T . Solving for $\hat{f}(3)$ yields

$$\hat{f}(3) = \frac{\lambda_3 + \beta}{\beta} \left(1 - 2 \frac{\lambda_3}{\lambda_3 + \beta} \right) = -1.$$

By a similar calculation we obtain that

$$\hat{f}(2) = \frac{(\beta + \lambda_2)^2}{\beta^2 + 2\beta\lambda_2} \left(1 - \hat{f}(4) \frac{\lambda_2^2}{(\beta + \lambda_2)^2} \right) = \frac{1}{5}.$$

Since 1 is a global maximum point of G taking

$$\hat{f}(1) = G(1) = 3$$

yields the right representation. We see that the assumptions of Theorem 2.6 are fulfilled (with $x_* = 2$, $x^* = 4$). Therefore

$$\tau^* = \inf\{t \geq 0 : X_t \neq 3\}$$

is an optimal stopping time.

Acknowledgements. This paper was partly written during the first author's stay at Åbo Akademi in the project *Applied Markov processes – fluid queues, optimal stopping and population dynamics, Project number: 127719 (Academy of Finland)*. He would like to express his gratitude for the hospitality and support.

References

- [1] L. Alili and A. E. Kyprianou. Some remarks on first passage of Lévy processes, the American put and pasting principles. *Ann. Appl. Probab.*, 15(3):2062–2080, 2005.
- [2] R. M. Blumenthal and R. K. Gettoor. *Markov processes and potential theory*. Pure and Applied Mathematics, Vol. 29. Academic Press, New York, 1968.
- [3] A. N. Borodin and P. Salminen. *Handbook of Brownian motion—facts and formulae*. Probability and its Applications. Birkhäuser Verlag, Basel, second edition, 2002.
- [4] S. Boyarchenko and S. Levendorskiĭ. *Irreversible decisions under uncertainty*, volume 27 of *Studies in Economic Theory*. Springer, Berlin, 2007. Optimal stopping made easy.
- [5] S. Christensen and A. Irle. A note on pasting conditions for the American perpetual optimal stopping problem. *Statist. Probab. Lett.*, 79(3):349–353, 2009.
- [6] S. Christensen, A. Irle, and A. Novikov. An elementary approach to optimal stopping problems for AR(1) sequences. *Sequential Analysis*, 30(1):79–93, 2011.
- [7] K. L. Chung and J. B. Walsh. *Markov processes, Brownian motion, and time symmetry*, volume 249 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, New York, second edition, 2005.
- [8] D. A. Darling, T. Liggett, and H. M. Taylor. Optimal stopping for partial sums. *Ann. Math. Statist.*, 43:1363–1368, 1972.
- [9] G. Deligiannidis, H. Le, and S. Utev. Optimal stopping for processes with independent increments, and applications. *J. Appl. Probab.*, 46(4):1130–1145, 2009.
- [10] N. El Karoui and A. Meziou. Max-Plus decomposition of supermartingales and convex order. Application to American options and portfolio insurance. *Ann. Probab.*, 36(2):647–697, 2008.
- [11] H. Föllmer and T. Knispel. A representation of excessive functions as expected suprema. *Probab. Math. Statist.*, 26(2):379–394, 2006.

- [12] M. Jeanblanc, M. Yor, and M. Chesney. *Mathematical methods for financial markets*. Springer Finance. Springer-Verlag London Ltd., London, 2009.
- [13] H. Kunita and T. Watanabe. Markov processes and Martin boundaries. I. *Illinois J. Math.*, 9:485–526, 1965.
- [14] A. E. Kyprianou. *Introductory lectures on fluctuations of Lévy processes with applications*. Universitext. Springer-Verlag, Berlin, 2006.
- [15] A. E. Kyprianou and B. A. Surya. On the Novikov-Shiryaev optimal stopping problems in continuous time. *Electron. Comm. Probab.*, 10:146–154 (electronic), 2005.
- [16] P.W. Millar. A path decomposition for markov processes. *Ann. Probab.*, 6(2):345–348, 1978.
- [17] E. Mordecki. Optimal stopping and perpetual options for Lévy processes. *Finance Stoch.*, 6(4):473–493, 2002.
- [18] E. Mordecki and P. Salminen. Optimal stopping of Hunt and Lévy processes. *Stochastics*, 79(3-4):233–251, 2007.
- [19] A. Novikov and A. Shiryaev. On an effective case of the solution of the optimal stopping problem for random walks. *Teor. Veroyatn. Primen.*, 49(2):373–382, 2004.
- [20] A. Novikov and A. Shiryaev. On solution of the optimal stopping problem for processes with independent increments. *Stochastics*, 79(3-4):393–406, 2007.
- [21] D. Revuz and M. Yor. *Continuous martingales and Brownian motion*, volume 293 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, third edition, 1999.
- [22] P. Salminen. Optimal stopping of one-dimensional diffusions. *Math. Nachr.*, 124:85–101, 1985.
- [23] P. Salminen. Optimal stopping, Appell polynomials, and Wiener-Hopf factorization. *Stochastics*, 83(4-6):611–622, 2011.
- [24] M. Sharpe. *General theory of Markov processes*, volume 133 of *Pure and Applied Mathematics*. Academic Press Inc., Boston, MA, 1988.

- [25] A. N. Shiryaev. *Optimal stopping rules*, volume 8 of *Stochastic Modelling and Applied Probability*. Springer-Verlag, Berlin, 2008. Translated from the 1976 Russian second edition by A. B. Aries, Reprint of the 1978 translation.
- [26] B. A. Surya. An approach for solving perpetual optimal stopping problems driven by Lévy processes. *Stochastics*, 79(3-4):337–361, 2007.

Have I seen you before? Principles of Bayesian predictive classification revisited

Jukka Corander · Yaqiong Cui · Timo Koski ·
Jukka Sirén

Received: 26 March 2011 / Accepted: 20 September 2011
© Springer Science+Business Media, LLC 2011

Abstract A general inductive Bayesian classification framework is considered using a simultaneous predictive distribution for test items. We introduce a principle of generative supervised and semi-supervised classification based on marginalizing the joint posterior distribution of labels for all test items. The simultaneous and marginalized classifiers arise under different loss functions, while both acknowledge jointly all uncertainty about the labels of test items and the generating probability measures of the classes. We illustrate for data from multiple finite alphabets that such classifiers achieve higher correct classification rates than a standard marginal predictive classifier which labels all test items independently, when training data are sparse. In the supervised case for multiple finite alphabets the simultaneous and the marginal classifiers are proven to become equal under generalized exchangeability when the amount of training data increases. Hence, the marginal classifier can be interpreted as an asymptotic approximation to the simultaneous classifier for finite sets of training data. It is also shown that such convergence is not guaranteed in the semi-supervised setting, where the marginal classifier does not provide a consistent approximation.

Keywords Classification · Exchangeability · Inductive learning · Predictive inference

1 Introduction

We consider the generic problem of classification, *i.e.*, assigning some items into a discrete set of classes using observed characteristics (features) of the items and assessing the uncertainty related to the assignments conditional on all relevant information available. We consider explicitly the following two distinct settings for classification: the *supervised* classification, where all the eligible classes are *a priori* given, and the *semi-supervised* classification, where a certain set of eligible classes is *a priori* determined, while the items are not forced to solely be allocated to such classes, but can also form previously unknown groups during the classification task. It should be noticed that the concept of semi-supervised classification is used in various distinct meanings, and apart from the definition used here, it may refer to a situation where partial (for example pairwise) constraints are imposed on an otherwise unsupervised classification (Basu 2005). Semi-supervised learning subsumes in general a wide range of approaches to solving learning tasks (Chapelle et al. 2006). For the development of supervised classification over nearly half a century, see Bailey (1965), Chow and Liu (1968), Devroye et al. (1996), Ripley (1996), Duda et al. (2000), Hand and Yu (2001) and Bishop (2007).

We focus on developing inductive principles for learning from specific assumptions of probabilistic invariance concerning the joint predictive model for training and future (test) data sets. Such principles will in general provide guidelines for assessing the uncertainty related to a question such as the one posed in the title of this work, “Have I seen

J. Corander (✉) · Y. Cui · J. Sirén
Department of Mathematics and Statistics, University of Helsinki,
00014 Helsinki, Finland
e-mail: jukka.corander@helsinki.fi

T. Koski
Department of Mathematics, Royal Institute of Technology,
100 44 Stockholm, Sweden

J. Corander
Department of Mathematics, Åbo Akademi University,
20500 Åbo, Finland

you before?”, given some noisy or highly variable measurements for training and test items. That is, here one needs to judge in probabilistic terms whether a test item resembles well enough those present in a particular class of training data to make the claim that we have indeed seen something like it before. For a philosophical discussion about probabilistic induction see Howson (2000) and for an account of the theory of inductive inference see Solomonoff (1964, 2008). An operationalization of the inductive learning approach is obtained by combining: (i) a model of *simultaneous classification* or *data labeling* defined by generating random urns and assigning data items to them, (ii) a predictive model for future data given training data, and (iii) an algorithm that can compute the sought predictive probabilities.

In supervised classification it is typical to classify test items or samples individually, independently of the classification decisions made for any other item in the test set. This follows from the generating *i.i.d.* assumption, where the features for any test item are conditionally independent of the features for any other item given the fixed generative probability measure. However, inductive learning theory based on predictive modeling implies that the test items *should be* classified *simultaneously*, because the item data can be considered conditionally independent only given their *joint labeling* with respect to the *a priori* specified classes. Marginal dependence exists between the test items in the predictive probability distribution, since the underlying generating probability measure is not exactly known. We will show that this dependence vanishes at the limit when the amount of training data tends to infinity, and consequently, the two learning approaches become equivalent.

The simultaneous predictive supervised classification principle was pioneered by Geisser (1964, 1966, 1993) with a primary focus on a Gaussian modeling context, but apart from applications in speech recognition and image analysis (Nádas 1985; Huo and Lee 2000; Ripley 1988, 1996), it seems to have received very limited attention in the statistical or machine learning literature. Here we operationalize this approach by combining a generative model for classification structures with the predictive machinery to derive inductive classifiers. The classification principles themselves are universally applicable to any data types, but here we restrict ourselves to multivariate discrete data in the derivation of explicit predictive probabilities.

We also expand the inductive learning in two respects. Firstly, we develop the corresponding simultaneous classifier for a semi-supervised situation, where some or all test items are allowed to be generated from *a priori* unknown models lacking training data, while at the same time training data exist for a finite set of putative sources of the test items. Secondly, we introduce a predictive classification principle which uses marginal posterior distributions of item labels

derived from the simultaneous classifier. Such a classifier is referred to as a *marginalized*, in contrast to the standard *marginal* classifier which treats all test items independently of each other. The marginalized classifier arises under the predictive learning framework (supervised or semi-supervised) for any particular item when the labels of remaining items are considered as nuisance parameters. To apply the predictive classifiers considered in this work we consider two alternative algorithms, one based on deterministic greedy optimization and the other on a stochastic optimization algorithm implemented with a Gibbs sampler (Robert and Casella 2005).

The outline of the paper is as follows. In Sect. 2 we derive the predictive supervised and semi-supervised classification framework using random urn models, followed by the explicit formulations of predictive probability distributions for the considered classifiers. Asymptotic relationships between simultaneous and marginal classifiers are investigated in Sect. 3 and in Sect. 4 we discuss algorithms for the implementation of the classifiers. In Sect. 5 we compare the finite sample performance of the classifiers in numerical experiments involving both real and synthetic data sets. The final section provides some remarks and a discussion, while technical details related to proofs are given in an Appendix at the end of the paper.

2 Predictive classification using random urns and classification structures

Let M be a set of m training items $i \in M$ for which we observe finite vectors $\mathbf{z}_i$ of d features, such that each element z_{ij} in $\mathbf{z}_i$ belongs to a finite alphabet, $z_{ij} \in \mathcal{X}_j = \{1, \dots, r_j\}$, $r_j \geq 1$, $j = 1, \dots, d$. The joint feature space is represented by the Cartesian product

$$\mathbf{z}_i \in \mathcal{X} = \times_{j=1}^d \mathcal{X}_j. \quad (1)$$

Let T denote the joint classification, *i.e.*, labels of the training data items in M , such that each element $T_i \in \{1, \dots, k\}$ assigns an item $i \in M$ unambiguously to a class indexed by the integer c , $1 \leq c \leq k$. The k classes and the labels in T are assumed to have been established *a priori* using some suitable auxiliary information. The total collection of training feature data is denoted by $\mathbf{z}^{(M)}$ and the subset of training feature data for the class c by $\mathbf{z}^{(M_c)}$, $M_c \subset M$, with the size $m_c = |M_c|$. Note that $\mathbf{z}^{(M)}$ does not include the labels specified by the training data, as the labels are instead defined by T . Let now N be a set of n test items $i \in N$ for which we observe finite feature vectors $\mathbf{x}_i \in \mathcal{X}$ defined in the same space as $\mathbf{z}_i$, such that the total collection of test feature data is denoted as $\mathbf{x}^{(N)}$. For any subset $A \subset N$ we let $\mathbf{x}^{(A)}$ be the corresponding subset of the test feature data. In particular,

$N_c \subset N$ will be used for the set of test items assigned to a particular class and $n_c = |N_c|$ is the size of this set.

By a classification structure S we refer to an unambiguous assignment of all the test items in N to a finite set of classes, which may either be: (1) the k classes present in T , or (2) contain some classes lacking training data, depending on whether one considers a supervised or a semi-supervised classification scenario, respectively. We let $\mathcal{S}$ denote the family of all such structures in both of the mentioned cases. Notice that the cardinality of $\mathcal{S}$ differs between the two situations. For supervised classification $\mathcal{S}$ equals the Cartesian product $\{1, \dots, k\}^n$, whereas for the semi-supervised case $\mathcal{S}$ has a substantially larger cardinality and its definition will be explicitly considered in Sect. 2.3. In the supervised classification each element in S is defined analogously to T , such that $S_i \in \{1, \dots, k\}$.

In the semi-supervised case the actual label values assigned to the classes residing outside those represented by training data lack a meaningful interpretation because they are permutable. To make this explicit in the notation, we let here the number of previously known classes in S equal k_1 (i.e. $\arg \max_{i \in M} T_i = k_1$) and the number of previously unknown classes equal k_2 . Let r , $0 \leq r \leq n$, be the number of test items allocated to the k_1 first mentioned classes, then the remaining $n - r$ items are in the latter k_2 classes. The statistically relevant information in S about the latter $n - r$ items is then comprised in terms of an unordered partition $(s_1, \dots, s_{k_2}), s_{c'} \subset N$, with permutable class indices $c' = 1, \dots, k_2$. A classification structure S can still be unambiguously defined by a vector of integers $S_i \in \{1, \dots, k_1 + k_2\}$, where one can use any one-to-one mapping of the k_2 index values of $(s_1, \dots, s_{k_2})$ into the integers $k_1 + 1, \dots, k_1 + k_2$.

2.1 Predictive classifiers

We consider first the joint conditional distribution of S , the training data $\mathbf{z}^{(M)}$ and the test data $\mathbf{x}^{(N)}$, given the fixed labeling T . Let $p(S|T) > 0$, $S \in \mathcal{S}$, be a prior probability for a classification structure S conditional on T . A class of such priors is derived later using random urn models in Sect. 2.3, however, for the derivation of the predictive classifier we let the prior remain implicitly defined. To obtain the posterior probability for S , we define first the probability of all feature data $(\mathbf{x}^{(N)}, \mathbf{z}^{(M)})$ conditional on T as

$$p(\mathbf{x}^{(N)}, \mathbf{z}^{(M)}|T) = \sum_{S \in \mathcal{S}} p(\mathbf{x}^{(N)}|\mathbf{z}^{(M)}, S, T) p(\mathbf{z}^{(M)}|T) p(S|T). \quad (2)$$

Here $p(\mathbf{x}^{(N)}|\mathbf{z}^{(M)}, S, T)$ is the (posterior) predictive probability of test data given the training data (see, e.g. Bernardo and Smith 1994). Formally, in the case of a parametric family for the features defined by $\theta \in \Theta$, the predictive probability will equal an infinite mixture of data distributions over

a set of random parameters θ , such that

$$p(\mathbf{x}^{(N)}|\mathbf{z}^{(M)}, S, T) = \int_{\Theta} p(\mathbf{x}^{(N)}|\theta, S, T) dF(\theta|\mathbf{z}^{(M)}, T), \quad (3)$$

where $dF(\theta|\mathbf{z}^{(M)}, T)$ is a conditional probability measure for the model parameters given the training data. This is predictive in the sense that it assigns a probability on future observations by learning about the parameters from the training data (Geisser 1993).

We note that the classification principles considered in this section hold more universally for arbitrary data types, but explicit expressions for the predictive probabilities will be given in Sect. 2.2 for the finite alphabet case. It is natural to assume that the distribution of the training feature data does not depend on S , whereby the posterior probability of S equals

$$p(S|\mathbf{x}^{(N)}, \mathbf{z}^{(M)}, T) = \frac{p(\mathbf{x}^{(N)}|\mathbf{z}^{(M)}, S, T) p(S|T)}{\sum_{S \in \mathcal{S}} p(\mathbf{x}^{(N)}|\mathbf{z}^{(M)}, S, T) p(S|T)}, \quad (4)$$

since the term $p(\mathbf{z}^{(M)}|T)$ cancels out. Also, in the above expression it is implicitly assumed that the training data labeling T is fixed, such that no prior distribution $p(T)$ needs to be introduced.

If the n test items are considered independently of each other, we obtain the standard *marginal* predictive classifier (Ripley 1996, Chap. 2.4), which is subsequently denoted as $PPRED_1$. This classifier yields the following posterior distribution for the label of item $i \in N$ in the supervised case:

$$p(S_i = c|\mathbf{z}^{(M)}, \mathbf{x}_i, T) = \frac{p(\mathbf{x}_i|\mathbf{z}^{(M_c)}, S_i = c, T) p(S_i = c)}{\sum_{c=1}^k p(\mathbf{x}_i|\mathbf{z}^{(M_c)}, S_i = c, T) p(S_i = c)}, \quad c = 1, \dots, k \quad (5)$$

where $p(\mathbf{x}_i|\mathbf{z}^{(M_c)}, S_i = c, T)$ is the predictive probability of $\mathbf{x}_i$ in class c and $p(S_i = c)$ is the prior probability of label c . A predictive classifier that assigns the label S_i by maximizing the posterior probability can be shown to minimize the averaged risk of misclassification (Ripley 1996, Chap. 2.4). A proof of optimality of the predictive classifier has also been given by Nádas (1985) by considering the expected payoff for the case where training and test items arrive sequentially, motivated by an application to speech recognition. For more recent work on the use of Bayesian predictive classification in solving such problems, see also Huo and Lee (2000).

For the semi-supervised case, the corresponding posterior distribution of labels for $i \in N$ equals

$$p(S_i = c | \mathbf{z}^{(M)}, \mathbf{x}_i, T) \propto \begin{cases} p(\mathbf{x}_i | \mathbf{z}^{(M_c)}, S_i = c, T) p(S_i = c), & c = 1, \dots, k_1 \\ p(\mathbf{x}_i | S_i = k_1 + 1) p(S_i = k_1 + 1), \end{cases} \quad (6)$$

where $p(\mathbf{x}_i | S_i = k_1 + 1)$ is the *prior predictive probability* (Bernardo and Smith 1994) of $\mathbf{x}_i$ in class lacking training data, which is defined as $p(\mathbf{x}^{(N)} | S_i = k_1 + 1) = \int_{\Theta} p(\mathbf{x}^{(N)} | \theta, S_i = k_1 + 1) dF(\theta)$. The constant of proportionality in the above distribution equals

$$p(\mathbf{x}_i | S_i = k_1 + 1) p(S_i = k_1 + 1) + \sum_{c=1}^{k_1} p(\mathbf{x}_i | \mathbf{z}^{(M_c)}, S_i = c, T) p(S_i = c). \quad (7)$$

A simultaneous predictive classifier, denoted as p_{PRED_2} , assigns jointly the posterior probability on the collection of labels for all items according to (4). As a principle, this was first suggested by Geisser (1966) in the context of Gaussian discriminant analysis. A simultaneous predictive classifier which assign the labels S by maximizing the posterior (4), is optimal under the (zero-one) loss function rewarding only decisions where all predicted labels are correct (Ripley 1996, Chap. 2.4). Ripley (1996) notes that the marginal classifier does not have as high a prediction accuracy as possible when $n > 1$, since one can benefit from further learning about the model parameters by using the other test items when θ is not known.

To make the distinction between the supervised and the semi-supervised simultaneous classifiers explicit, the posterior probability under the latter can be written as

$$p(S | \mathbf{z}^{(M)}, \mathbf{x}^{(N)}, T) = \frac{p(\mathbf{x}^{(N^{(1)})} | \mathbf{z}^{(M)}, S^{(1)}, T) p(\mathbf{x}^{(N^{(2)})} | S^{(2)}) p(S^{(2)} | T)}{\sum_{S \in \mathcal{S}} p(\mathbf{x}^{(N^{(1)})} | \mathbf{z}^{(M)}, S^{(1)}, T) p(\mathbf{x}^{(N^{(2)})} | S^{(2)}) p(S^{(2)} | T)}, \quad (8)$$

where $S^{(1)}, S^{(2)}$ are the subvectors of the joint labeling S corresponding to the k_1 and k_2 classes, respectively, $N^{(1)}, N^{(2)}$ are the corresponding subsets of N , and $p(\mathbf{x}^{(N^{(2)})} | S^{(2)})$ is the joint prior predictive probability of the data in all previously unknown classes.

A *marginalized* predictive classifier, denoted as p_{PRED_3} , assigns in the supervised case the following posterior probability distribution for the label of item $i \in N$:

$$p(S_i = c | \mathbf{z}^{(M)}, \mathbf{x}^{(N)}, T) = \sum_{\{S \in \mathcal{S}: S_i = c\}} p(S | \mathbf{z}^{(M)}, \mathbf{x}^{(N)}, T), \quad c = 1, \dots, k, \quad (9)$$

which equals the marginal posterior distribution of the label for the item. In image analysis it has been shown that a simultaneous labeling of pixels, such that the marginal pos-

terior for each label is maximized, is an optimal classification rule under the loss function that sanctions the number of incorrect labels (Ripley 1988, p. 114). This optimality result, albeit not derived in the predictive classification context, applies also to the classifier p_{PRED_3} , such that when (9) is maximized for all $i \in N$, the classifier minimizes the same expected loss. Also, Ripley (1996, Chap. 2.4) considers the related idea of using the remaining test data $\mathbf{x}_1, \dots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \dots, \mathbf{x}_n$ to improve learning about model parameters when assigning the label S_i under diagnostic (non-generative) classification. This is distinct from the predictive paradigm considered here and the idea of using marginal posteriors of labels seems not to have appeared earlier in the predictive classification context, except when test items are ordered in time as in speech recognition (Nádas 1985; Huo and Lee 2000).

In the semi-supervised marginalized case we obtain the same probability distribution as in (9) for all items $i \in N^{(1)}$, whereas for any item $i \in N^{(2)}$ we have

$$p(S_i > k_1 | \mathbf{z}^{(M)}, \mathbf{x}^{(N)}, T) = 1 - \sum_{c=1}^{k_1} \sum_{\{S \in \mathcal{S}: S_i = c\}} p(S | \mathbf{z}^{(M)}, \mathbf{x}^{(N)}, T). \quad (10)$$

Notice that (10) is the marginal posterior probability that an item represents a previously unknown source.

In summary, the differences between the three alternative forms of predictive classifiers for both the supervised and semi-supervised setting can be characterized as follows. Firstly, the marginal classifier is the most basic predictive classifier, which uses only the existing training data to define the posterior distribution of the label of each test item, separately for each test item. In the supervised case this amounts to comparing the posterior predictive probabilities of the observed feature data for each putative class out of the k classes in total. In the semi-supervised case, a single additional source lacking training information is allowed for, and thus, the posterior predictive probabilities arising under the k known classes are compared with the prior predictive probability of the feature data resulting from the assignment to an *a priori* empty class. Intuitively, the latter prediction can attain high values when the test item feature observations are very unlikely to occur under all classes for which training data exist.

Secondly, a simultaneous predictive classifier uses the information about the distribution of features shared by all items that are assigned to the same class, *i.e.* data from both the training and all test items simultaneously. When the training data are sparse, the simultaneous predictive probability (3) may differ considerably from the product of n marginal predictive probabilities calculated separately according to (5) for each test item. The primary reason for this difference is that the simultaneous classifier considers the

joint probability of observing any specific combination of feature values over the set of test items that are assigned to the same class, instead of the item-wise marginal probabilities used by the marginal classifier. In the semi-supervised case the test items assigned to an *a priori* unknown class can meaningfully share the same label, which is not feasible under the marginal semi-supervised classifier. In particular, the simultaneous semi-supervised classifier can then benefit from the fact that data from multiple test items are jointly available to make inferences about the underlying class distribution.

Finally, the marginalized classifiers take the predictive inferences one step further from the simultaneous classifier, by considering the labeling decision separately for each test item, such that the labels of the remaining test items are treated as nuisance parameters. Thus, the marginalized classifiers still acknowledge the predictive modeling of the test data in a simultaneous fashion, but derive from this a decision rule for each individual item by using the marginal posterior distribution of the label. This marginal posterior is in general different from the one resulting in the marginal classifier because the latter ignores the feature data from the remaining items. As discussed above, the optimality of this marginalization strategy follows from the application of an intuitively appealing risk function which aims at minimizing the total number of incorrect labels.

2.2 Predictive distributions under generalized exchangeability

In the previous section we introduced the predictive classifiers without reference to explicit expressions for the actual probabilities as functions of the observed data. Here we derive such expressions for data from multiple finite alphabets. Corander et al. (2007, 2009) used a random urn model in conjunction with certain generalizations of exchangeability and the de Finetti representation theorem to arrive at a predictive probabilistic representation of classification uncertainty in an unsupervised (clustering) situation. This representation is based on particular assumptions of general probabilistic invariance principles, which are investigated in detail in Bernardo and Smith (1994) and Kallenberg (2005). We will here apply the invariances to obtain explicit predictive probabilities for the supervised and semi-supervised classifiers.

We assume that the observed sequences of features $\mathbf{x}^{(N_c)}, \mathbf{z}^{(M_c)}$ assigned to the same class are *unrestrictedly infinitely exchangeable* (see Definitions 4.2, 4.3, and 4.13, and Propositions 4.2 and 4.18 in Bernardo and Smith 1994). This assumption implies that, if we combine any permutation of the item values $x_{1j}, \dots, x_{n_{cj}}, z_{1j}, \dots, z_{m_{cj}}$, for a fixed $j = 1, \dots, d$, with arbitrary corresponding permutations over the remaining features, the same predictive probability mass function for the data is obtained. Furthermore, we

also obtain a characterizations of the test and training data sets in terms of the sufficient statistics $n_{cjl}, m_{cjl}, l = 1, \dots, r_j$, where n_{cjl} and m_{cjl} represent the number of copies of value l for feature j observed among the test and training items in the class c , respectively.

The unique probabilistic characterization of the data under unrestricted infinite exchangeability assumed to hold over the classes of S then equals by (3)

$$p(\mathbf{x}^{(N)} | \mathbf{z}^{(M)}, S, T) = \int_{\Theta} \prod_{c=1}^k \prod_{j=1}^d \prod_{l=1}^{r_j} \theta_{cjl}^{n_{cjl}} dF(\theta | \mathbf{z}^{(M)}, T), \quad (11)$$

i.e., the test data have a product multinomial likelihood

$$p(\mathbf{x}^{(N)} | \theta, S, T) = \prod_{c=1}^k \prod_{j=1}^d \prod_{l=1}^{r_j} \theta_{cjl}^{n_{cjl}}, \quad (12)$$

where θ denotes jointly the multinomial probabilities $(\theta_{c11}, \dots, \theta_{c1r_1}, \dots, \theta_{cd1}, \dots, \theta_{cdr_d})$, $c = 1, \dots, k$ and $F(\theta | \mathbf{z}^{(M)}, T)$ is a probability measure over the space Θ . This measure represents our posterior beliefs about the limits of the relative frequencies of observing certain feature values in the different classes, given the observed training data. The extension of the unrestricted exchangeability assumption over the classes implies that the observed values of features within an arbitrary class provide no information about the values observed in any other class. Thus, the assumed exchangeability corresponds to assuming conditional independence of all the features over the considered classes, and leads to a product multinomial likelihood for the data conditional on any θ .

If a class c is void of training data, the corresponding element in the distribution (11) is a prior predictive model for the observed items in the class, which is often also termed as the *marginal likelihood* or the *evidence*. By using the so-called *sufficientness* postulate (Zabell 1982), the prior beliefs about the relative frequencies for a feature can be explicitly expressed in terms of the Dirichlet $(\{\lambda_{cjl}\}_{l=1}^{r_j})$ distribution, which is conjugate for the multinomial likelihood. When this is combined with the product form of (11) over features and classes arising from the generalized exchangeability, we can write the predictive probability as

$$p(\mathbf{x}^{(N)} | \mathbf{z}^{(M)}, S, T) = \prod_{c=1}^k \prod_{j=1}^d \frac{\Gamma(m_c + 1)}{\Gamma(n_c + m_c + 1)} \times \prod_{l=1}^{r_j} \frac{\Gamma(n_{cjl} + m_{cjl} + \lambda_{cjl})}{\Gamma(m_{cjl} + \lambda_{cjl})}, \quad (13)$$

where it is assumed for simplicity that the Dirichlet hyperparameters are constants defined as in the reference prior with

$\lambda_{cjl} = 1/r_j$ (for details on reference priors, see Bernardo and Smith 1994). We note that under the marginal predictive supervised classifier p_{PRED_1} the predictive probability is a special case of (13) with $n = 1$, $\mathbf{x}^{(N)}$ consisting thus of a single data vector only. Finally, the corresponding predictive probability for the semi-supervised classifier is an extension of (13):

$$p(\mathbf{x}^{(N)} | \mathbf{z}^{(M)}, S, T) = \prod_{c=1}^{k_1} \left[\prod_{j=1}^d \frac{\Gamma(m_c + 1)}{\Gamma(n_c + m_c + 1)} \prod_{l=1}^{r_j} \frac{\Gamma(n_{cjl} + m_{cjl} + \lambda_{cjl})}{\Gamma(m_{cjl} + \lambda_{cjl})} \right] \times \prod_{c=k_1+1}^{k_1+k_2} \left[\prod_{j=1}^d \frac{\Gamma(1)}{\Gamma(n_c + 1)} \prod_{l=1}^{r_j} \frac{\Gamma(n_{cjl} + \lambda_{cjl})}{\Gamma(\lambda_{cjl})} \right],$$

which is obtained by combining the posterior predictive probability over the k_1 classes having training data with the prior predictive probability over the k_2 classes lacking training data.

2.3 Priors for classification structures

To perform inductive inference (Bernardo and Smith 1994) about the classification structures it is necessary to specify a prior probability distribution $p(S)$ for S , which will here be derived by generalizing the stochastic urn model in Corander et al. (2007, 2009). Let $u \in \mathbb{Z}^+$ be in general a stochastic number of urns into which the items in N are randomly allocated. The classification structure S is then taken to be the contents of the k nonempty urns which leads to a distribution over the space $\mathcal{S}$. For the supervised classification the number of urns u is fixed to k ($p(u = k | T) = 1$) and the urns are identified by T , such that the prior probability of S can be expressed as

$$p(S|T) = \sum_{u=1}^{\infty} \pi(S|u, T) p(u|T) = \pi(S|k, T) = k^{-n}, \quad (14)$$

where $\pi(S|u, T)$ is the conditional probability of the classification structure S , which equals u^{-n} for every S (by counting).

In semi-supervised classification a structure S must allow for situations where *all* or *none* of the test items carry labels among the k_1 classes with respect to T . The family $\mathcal{S}$ of such classification structures and its cardinality can be obtained by considering the known and the new classes separately. The r items in the k_1 classes can be chosen from N in $\binom{n}{r}$ ways and allocated to the k_1 classes in k_1^r ways. The remaining $n - r$ items are then randomly allocated to a

stochastic number of urns and form a partition of the corresponding item set into k_2 new classes. The number of classification structures is consequently obtained by summing over the possible values of r as

$$|\mathcal{S}| = \sum_{r=0}^n \binom{n}{r} k_1^r B_{n-r} = \sum_{r=0}^n \binom{n}{r} k_1^r e^{-1} \sum_{k_2=1}^{\infty} \frac{k_2^{n-r}}{k_2!}, \quad (15)$$

where B_{n-r} is the Bell number for $n - r$ items. Any particular structure $S \in \mathcal{S}$ will in the semi-supervised case be representable by an $n \times 1$ vector with r entries taking values in the set $\{1, \dots, k_1\}$ and $n - r$ entries taking values in the set $\{k_1 + 1, \dots, k_1 + k_2\}$. However, it should be noticed that the latter k_2 labels are permutable, since these classes are entirely determined in terms of the items assigned to them and not by any auxiliary information.

Given the number of urns u , each item of N is randomly allocated to an urn corresponding to a class in T with probability k_1/u . In general, the conditional probability of a random partition S with k classes, given the number of urns u , equals $u^{(k)} u^{-n}$, for $u \geq k$, and zero for $u < k$, where $u^{(0)} = 1$ and $u^{(k)} = u(u-1) \cdots (u-k+1)$. Since there can be some urns which do not receive any items in the random allocation, k will be the number of resulting non-empty urns out of u in total. For a derivation of this urn based model for generating random partitions, see Stam (1983).

Finally, the conditional probability $\pi(S|u, T)$ of a classification structure S with r items in the k_1 known classes and k_2 new classes can be computed by considering the r and $n - r$ items separately and it has the expression

$$\begin{aligned} \pi(S|u, T) &= \left(\left(\frac{k_1}{u} \right)^r \left(\frac{1}{k_1} \right)^r \right) \left(\left(\frac{u - k_1}{u} \right)^{n-r} \frac{(u - k_1)^{(k_2)}}{(u - k_1)^{n-r}} \right) \\ &= \frac{(u - k_1)^{(k_2)}}{u^n}, \end{aligned} \quad (16)$$

for $u \geq k_1$. Note that the number of urns u can be much larger than k .

A uniform distribution for S is obtained as follows, by extending the result of Stam (1983). Let the distribution for the number of urns u be

$$p(u|T) = |\mathcal{S}|^{-1} e^{-1} \frac{u^n}{(u - k_1)!}, \quad (17)$$

where the cardinality $|\mathcal{S}|$ is defined by (15). Now the conditional distribution of S given T equals

$$\begin{aligned} p(S|T) &= \sum_{u=1}^{\infty} \pi(S|u, T) p(u|T) \\ &= \sum_{u=k_1+k_2}^{\infty} \pi(S|u, T) p(u|T) \end{aligned}$$

$$\begin{aligned}
&= \sum_{u=k_1+k_2}^{\infty} |\mathcal{S}|^{-1} e^{-1} \frac{u^n}{(u-k_1)!} \frac{(u-k_1)^{(k_2)}}{u^n} \\
&= |\mathcal{S}|^{-1} e^{-1} \sum_{u=k_1+k_2}^{\infty} \frac{1}{(u-k_1-k_2)!} = |\mathcal{S}|^{-1}.
\end{aligned}$$

3 Asymptotic relationships between predictive classifiers

Here we characterize the behavior of the simultaneous predictive classifiers in relation to the marginal classifiers when the amount of training data increases. Recall that the logarithm of the data predictive probability (13) for the simultaneous classifier $PPRED_2$ equals under the previous derivations:

$$\begin{aligned}
&\log p(\mathbf{x}^{(N)} | \mathbf{z}^{(M)}, S, T) \\
&= \sum_{c=1}^k \sum_{j=1}^d \log \frac{\Gamma(m_c + 1)}{\Gamma(n_c + m_c + 1)} \\
&\quad + \sum_{c=1}^k \sum_{j=1}^d \sum_{l=1}^{r_j} \log \frac{\Gamma(n_{cjl} + m_{cjl} + \lambda_{cjl})}{\Gamma(m_{cjl} + \lambda_{cjl})}. \quad (18)
\end{aligned}$$

We can derive a comparable expression for the logarithm of the data predictive probability (13) of the marginal classifier $PPRED_1$ by considering:

$$\begin{aligned}
&\log p(\mathbf{x}_1, \dots, \mathbf{x}_n | \mathbf{z}^{(M)}, S, T) \\
&= \sum_{c=1}^k \sum_{i: S_i=c} \log p(\mathbf{x}_i | \mathbf{z}^{(M)}, T, S_i = c), \quad (19)
\end{aligned}$$

where S now refers to the classification structure resulting from the assignment of the n items into the classes one by one. Using Stirling's approximation to the gamma function terms in (18) and (19), it follows (for details see Appendix) that for any particular classification structure S , the difference in the logarithms of the data predictive probability between the simultaneous and marginal classifiers tends to zero as the amount of training data increases, i.e.

$$\begin{aligned}
&\log p(\mathbf{x}^{(N)} | \mathbf{z}^{(M)}, S, T) \\
&\quad - \log p(\mathbf{x}_1, \dots, \mathbf{x}_n | \mathbf{z}^{(M)}, S, T) \xrightarrow{m \rightarrow \infty} 0,
\end{aligned}$$

when the sampling process of training data from each class is infinitely unrestrictedly exchangeable. The limits of the relative frequencies of feature values exist and we assume that they are strictly positive (see Appendix). Consequently, the classifiers $PPRED_1$ and $PPRED_2$ will become equivalent with increasing training data set size, as the optimal classification according to one will coincide with that of the other, which is formally stated in the theorem below.

Theorem 1 *Asymptotic equivalence of supervised predictive classifiers. Assume that all limits of the relative frequencies of feature values are strictly positive under an infinitely exchangeable sampling process:*

$$\frac{m_{cjl}}{m_c} \xrightarrow{m \rightarrow \infty} \beta_{cjl} > 0,$$

$$c = 1, \dots, k, j = 1, \dots, d, l = 1, \dots, r_j.$$

Then, the optimal classifications under the simultaneous predictive supervised classifier $PPRED_2$ and the marginal predictive supervised classifier $PPRED_1$ are asymptotically equal as the amount of training data m tends to infinity.

Proof The proof is given in Appendix.

The limiting form of the predictive classifier coincides with the classifier obtained assuming *i.i.d.* sampling under a fixed probability measure. Intuitively, the above theorem tells us that no additional inductive learning can take place in the optimal simultaneous classifier when the amount of training information increases sufficiently. This has a practical consequence in classification applications, as numerical implementation of simultaneous classifier rules is considerably more computation intensive than that of marginal classifiers. However, the expected accuracies of the two classifiers may differ when the amount of training data is small, as illustrated in the next section.

In contrast to the supervised case, the simultaneous and marginal semi-supervised classifications need not coincide, even in the presence of infinite amounts of training data. In a sense, this can be seen as a simple consequence of the fact that the marginal semi-supervised classifier cannot make inductive statements about whether any two test items that are both labeled with an *a priori* unknown class should be considered to represent the same or different unknown classes. The simultaneous classifier, on the other hand, makes such statements in the predictive distribution over the joint labels. However, the discrepancy between the two classifiers goes deeper than this, such that even if we consider an identical labeling S under both of them, the resulting predictive probabilities can still disagree when the amount of training data increases. This result is formally stated in the second theorem below, which shows that the disagreement arises among the unknown classes, as expected. $\square$

Theorem 2 *Asymptotic discrepancy of semi-supervised predictive classifiers. Under the same assumptions as in Theorem 1, the difference in the logarithms of the data predictive probability between the simultaneous and marginal semi-supervised classifiers tends to:*

$$\sum_{c=k_1+1}^{k_1+k_2} \sum_{j=1}^d \sum_{l=1}^{r_j} \log \left(\frac{n_{cjl} + \lambda_{cjl}}{1 + \lambda_{cjl}} \right)^{n_{cjl}}$$

$$- \sum_{c=k_1+1}^{k_1+k_2} \sum_{j=1}^d n_c \log \left(\frac{n_c + 1}{2} \right),$$

when the amount of training data m tends to infinity.

Proof The proof is similar to that of Theorem 1 and is also given in [Appendix](#). $\square$

4 Learning algorithms for predictive classifiers

Here we describe a set possible algorithms for performing the learning tasks. In all algorithms described below the indicated probabilities are proportional to an unknown normalizing constant which equals the sum of predictive test data probabilities over the space of classification structures. Since the constant does not affect the maximization or sampling operators, it is tacitly ignored in the definitions.

The marginal supervised classifier is computationally straightforward under maximum *a posteriori* rule, as the classifier can be implemented according to

$$\hat{p}_{PRED_1} : \arg \max_{c=1, \dots, k} p(S_i = c | \mathbf{z}^{(M)}, \mathbf{x}_i, T),$$

independently for all $i \in N$. (20)

Equivalently, the same independent maximization algorithm applies to the semi-supervised classifier with the slight modification that also the class $c = k_1 + 1$ lacking training data is considered as an eligible label.

Let $S_c^{(i)}$ denote a classification structure obtained from S by relocating item i to class c and keeping the remaining labels unchanged. If $S_i = c$ in S , then $S_c^{(i)}$ is identical to S . A straightforward extension of the above deterministic algorithms to optimizing the simultaneous predictive classifiers can be defined as

$$\hat{p}_{PRED_2} : \arg \max_{S \in \mathcal{S}} p(S | \mathbf{x}^{(N)}, \mathbf{z}^{(M)}, T),$$
(21)

with the following operators:

- A1. Initialize S by S_0 obtained with $\hat{p}_{PRED_1}$.
- A2. Until no change occurs in S , do repeatedly for all $i \in N$:

$$\arg \max_{S \in \{S_1^{(i)}, \dots, S_k^{(i)}\}} p(S | \mathbf{x}^{(N)}, \mathbf{z}^{(M)}, T). \quad (22)$$

Thus, $\hat{p}_{PRED_2}$ is obtained otherwise analogously to $\hat{p}_{PRED_1}$, except that the simultaneous posterior assignment probabilities of all test items are compared when assessing the k possible labels of any $i \in N$ at an arbitrary iteration in A2. Similarly to the above comment, also this algorithm can be generalized to the semi-supervised classification by extending the space $\mathcal{S}$ to include even classes lacking

training data. By extending the definition of $S_c^{(i)}$ such that it also holds for $c > k$, where $k = k_1 + k_2$ is the number of classes present in S , relocation of items into previously empty classes is allowed. The algorithm for the simultaneous semi-supervised classifier is then based on the following operators:

- B1. Initialize S by S_0 obtained with the semi-supervised version of $\hat{p}_{PRED_1}$, such that each item assigned to an unknown source is in a singleton class.
- B2. Until no change occurs in S , do repeatedly for all $i \in N$:

$$\arg \max_{S \in \{S_1^{(i)}, \dots, S_{k_1+k_2+1}^{(i)}\}} p(S | \mathbf{x}^{(N)}, \mathbf{z}^{(M)}, T). \quad (23)$$

Here the optimization step compares the simultaneous posterior assignment probabilities of all test items when considering the $k_1 + k_2 + 1$ possible sources for the test item i , such that the last source corresponds to a singleton class consisting of i alone. By a simple modification of the proof in Gyllenberg et al. (1997a) it can be established that the above deterministic greedy algorithms will converge to local maxima over the space of classification structure in a finite number of steps in A2 and B2.

Marginalized classifiers p_{PRED_3} cannot be sensibly optimized using the greedy operators defined above, as they require summing over the space of classification structures to obtain the marginal posterior probabilities of labels. However, Gibbs sampler (Robert and Casella 2005) can be consistently used to optimize both the simultaneous and the marginalized supervised classifiers using the following sampling procedure:

- C1. Let $S_{(t)}$ denote the instance of S at iteration $t = 0, 1, \dots$ (parentheses are used in the subindex to distinguish this for an element of $\mathcal{S}$). Initialize the algorithm by S_0 obtained with $\hat{p}_{PRED_1}$.
- C2. For $t = 1, 2, \dots$, sample sequentially for all $i \in N$, a classification structure $S_{(t)}$ from the discrete distribution defined by the k probabilities:

$$\tilde{p}(S | \mathbf{x}^{(N)}, \mathbf{z}^{(M)}, T), \quad S \in \mathcal{S} : S_i = c, c = 1, \dots, k, \quad (24)$$

where each probability satisfies $\tilde{p}(S | \mathbf{x}^{(N)}, \mathbf{z}^{(M)}, T) \propto p(\mathbf{x}^{(N)} | \mathbf{z}^{(M)}, S, T)$ (as defined in (13)) and for each of the k structures the labels of the other test items in $N \setminus \{i\}$ are unchanged.

- C3. As an output of the algorithm store both the sequence $S_{(0)}, S_{(1)}, \dots$ of generated classification structures and the corresponding probabilities $p(\mathbf{x}^{(N)} | \mathbf{z}^{(M)}, S_{(0)}, T), p(\mathbf{x}^{(N)} | \mathbf{z}^{(M)}, S_{(1)}, T), \dots$

Table 1 Item-wise correct classification rates for different predictive supervised classifiers. The estimates are based on 100 synthetic data sets generated as described in the text. The Gibbs sampler algorithm was run for 5,000 iterations

	$\hat{p}_{PRED_1}$	$\hat{p}_{PRED_2}$	$\hat{p}_{PRED_2}^{Gibbs}$	$\hat{p}_{PRED_3}$
$k = 5, d = 10, m = 10, n = 25 :$	0.6088	0.6388	0.6452	0.6680
$k = 5, d = 10, m = 10, n = 50 :$	0.6220	0.6672	0.6708	0.6988
$k = 10, d = 10, m = 20, n = 50 :$	0.4392	0.4414	0.4360	0.4666

Using the output from the Gibbs sampler (C3), one can define the optimal simultaneous and marginalized supervised classifiers:

$$\hat{p}_{PRED_2}^{Gibbs} : \arg \max_{S \in \{S_{(0)}, S_{(1)}, \dots\}} p(\mathbf{x}^{(N)} | \mathbf{z}^{(M)}, S, T) \quad (25)$$

$$\hat{p}_{PRED_3} : \arg \max_{c=1, \dots, k} \sum_{S \in \{S_{(0)}, S_{(1)}, \dots\}} I(S_i = c), \quad \text{for all } i \in N, \quad (26)$$

where $I(S_i = c)$ is the indicator function of test item i being labeled with c in any classification structure S generated by the Gibbs sampler. Ripley (1988, p. 114) described an analogous use of the Gibbs sampler to jointly obtain marginal posterior model labels for pixels in image analysis. Standard properties of the Gibbs sampler algorithm (Robert and Casella 2005) establish the following result.

Theorem 3 *The classifiers $\hat{p}_{PRED_2}^{Gibbs}$ and $\hat{p}_{PRED_3}$ learned by the Gibbs sampler are consistent estimators of the optimal predictive supervised classifiers, such that $\hat{p}_{PRED_2}^{Gibbs} \rightarrow \arg \max_{S \in \mathcal{S}} p(S | \mathbf{z}^{(M)}, \mathbf{x}^{(N)}, T)$ and $\hat{p}_{PRED_3} \rightarrow \arg \max_{c=1, \dots, k} \sum_{\{S \in \mathcal{S} : S_i = c\}} p(S | \mathbf{z}^{(M)}, \mathbf{x}^{(N)}, T)$, for all $i \in N$, almost surely as $t \rightarrow \infty$.*

The Gibbs sampler algorithm in C1–C3 extends in a straightforward manner to the semi-supervised classifiers by extending the operator C2 such that it samples the label of item i from the normalized test data predictive probabilities corresponding to the k_1 classes with training data and the k_2 classes lacking training data in the classification structure S at the given iteration, as well as an additional singleton class containing only item i . A comparable consistency result can therefore be obtained for such an implementation, however, we abstain here from the details.

5 Numerical illustrations of predictive classifier performance

We conducted a set of experiments with both synthetic and real data to demonstrate finite sample performance of dif-

ferent predictive classifiers considered in the previous sections. Results for supervised and semi-supervised classification tasks are presented in Sect. 5.1 and Sect. 5.2, respectively.

5.1 Experiments for supervised classification

For each combination of number of features d and classes k , both training and test samples were generated as follows to obtain a single replicate data set used in the learning. Firstly, generating parameters were independently sampled for feature and class from the Beta(1,1) distribution (that is the Uniform(0,1) distribution). Conditionally on these values, $m_c + n_c$ binary data vectors were generated for each class. The training and test sets thus contain $m = m_c k$ and $n = n_c k$ items, respectively. The item-wise rates of accurate classification over 100 replicate data sets (each with different values for the generating model parameters) are shown in Table 1.

The results in Table 1 illustrate that the simultaneous classifier achieves higher rates of correct classification than the marginal classifier and that the difference can increase as a function of increasing size of the test data set in relation to the amount of training data available. Furthermore, the novel inductive learning principle based on the marginalized posterior probabilities performs consistently best. This is not particularly surprising, as the marginalization averages the evidence supporting a particular source of a test item, which has a stabilizing effect on the inferences in the presence of sparse training data. We observed also a reduced difference between the alternative classifiers observed for the case with largest number of classes. This is not due to a diminishing advantage of the inductive learning approach itself, but a manifestation of the failure of the adopted algorithms to identify the true global optimum over the space of classification structures.

In particular, in the more demanding experiment the simultaneous classifier implemented using Gibbs sampler fails to improve upon the marginal classifier. Therefore, we examined in detail the convergence behavior for multiple independent runs of the algorithm for some individual data sets and concluded that there was a lack of convergence even after 50,000 iterations. In contrast, the greedy optimization of the simultaneous classifier converged in general

Fig. 1 Image of the bacterial data summarized in the Table 2. Each column represents a binary phenotypic characteristic of an isolate coded with *black/white* colors and each row is an observed sample. The *black horizontal lines* indicate class boundaries

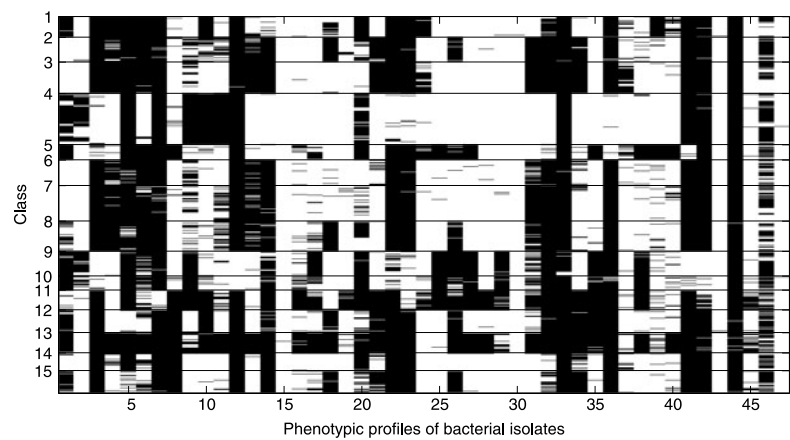


Table 2 Summary of the *Enterobacteriaceae* data used in the real data experiments. Column marked 'Size' specifies the number of isolates in each class. More details on data can be found in Gyllenberg et al. (1997b)

Class	Size	Species present
1	50	<i>Escherichia vulneris</i>
2	60	<i>Escherichia coli</i>
3	76	<i>E. coli</i>
4	125	<i>Klebsiella pneumoniae</i> , <i>K. planticola</i> , <i>K. oxytoca</i> , <i>K. terrigena</i>
5	37	<i>Cedecea neteri</i> , <i>C. davisae</i>
6	62	<i>E. coli</i>
7	86	<i>E. coli</i>
8	73	<i>E. coli</i>
9	60	<i>Serratia marcescens</i>
10	34	<i>S. marcescens</i>
11	48	<i>Providencia stuartii</i>
12	55	<i>Citrobacter freundii</i>
13	49	<i>Shigella prov</i> , <i>S. boydii</i>
14	43	<i>Citrobacter amalonaticus</i>
15	55	<i>C. freundii</i>

to a local mode very rapidly, so that the time to convergence was nearly always only twice the total time required by the marginal classifier.

We now compare performance of the marginal and simultaneous classifiers for a real data set. Due to the larger size of the data and convergence problems of the Gibbs sampler algorithm experienced for the case with $k = 10$ in the synthetic experiments, we only focus on comparing $\hat{p}_{PRED_1}$ and $\hat{p}_{PRED_2}$ to assess the relative effect of simultaneous inductive learning. The real data consist of 912 bacterial isolates representing $k = 15$ distinct classes which are summarized in Table 2. The data contain binary profiles over $d = 47$ features and have earlier been considered in an unsupervised classification setting in Gyllenberg et al. (1997b) and Corander et al. (2009). In particular, in Corander et al. (2009) approximately 130 classes were obtained from a database of 5313 bacterial isolates. Among these we picked at random a subset of 15 classes with sizes varying between 34 and 124 samples. An image of the class-wise data is shown in Fig. 1.

To compare the marginal and simultaneous classifiers in identifying the origin of a bacterial sample in the presence of varying amount of training and test data, we sampled without replacement j bacterial isolates, for $j = 1, \dots, 33$, from each of the 15 classes and comprised the test data set N from these $15j$ samples, while the remainder of the data were assigned to M given the fixed classification from the earlier unsupervised analysis. Thus, the smallest and largest test data sets contained 15 and 495 bacterial isolates, respectively. A total of 100 random replicates of the test data set were generated for each value of j and these were classified using the greedy optimization algorithms implementing $\hat{p}_{PRED_1}$ and $\hat{p}_{PRED_2}$. Figure 2 shows the item-wise correct classification rates for the two classifiers as a function of j . The asymptotic equivalence of the classifiers is well reflected by the nearly identical results for small values of j , while the difference of the two methods increases steadily as a function of the size of the test data sets. This demonstrates in a realistic supervised setting that gains in classifi-

cation accuracy can be obtained by using the simultaneous classifier instead of a marginal approach, even if it is only implemented using the deterministic greedy optimization.

5.2 Experiments for semi-supervised classification

We show two types of experiments with synthetic data in the semi-supervised setting. Firstly, otherwise equivalent simulation setup as in the supervised experiment was used, except that no training data was generated for the last class out of k classes in total. Also, use of Gibbs sampler was excluded due to convergence problems observed in initial experiments. Thus, the results reported in Table 3 only compare the relative performance of the two classifiers $\hat{p}_{\text{PRED}_1}$ and $\hat{p}_{\text{PRED}_2}$. However, here we report separately the correct classification rate for all data and for items from the class lacking training data. This distinction is made to illustrate a particular characteristic of the marginal classifier.

Table 3 shows that the simultaneous semi-supervised classifier achieves higher overall correct classification rates than the marginal classifier for the whole data even when im-

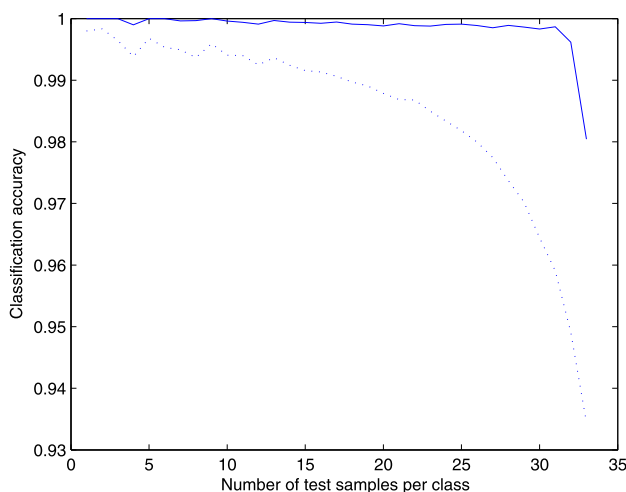


Fig. 2 The item-wise correct classification rates as a function of test data set size. The curves correspond to the two classifiers $\hat{p}_{\text{PRED}_1}$ (lower, dashed curve) and $\hat{p}_{\text{PRED}_2}$ (upper, continuous curve). Results are based on random splits of the bacterial data described in the text and Table 2

Table 3 Item-wise correct classification rates for different predictive semi-supervised classifiers. The estimates are based on 100 synthetic data sets generated as described in the text. The column label (a) refers

	$\hat{p}_{\text{PRED}_1}(\text{a})$	$\hat{p}_{\text{PRED}_2}(\text{a})$	$\hat{p}_{\text{PRED}_1}(\text{b})$	$\hat{p}_{\text{PRED}_2}(\text{b})$
$k = 5, d = 10, m = 10, n = 25 :$	0.5390	0.5679	0.4750	0.4418
$k = 5, d = 10, m = 10, n = 50 :$	0.5436	0.5904	0.4904	0.4862
$k = 5, d = 10, m = 25, n = 25 :$	0.6280	0.6498	0.4826	0.4886
$k = 5, d = 10, m = 25, n = 50 :$	0.6252	0.6603	0.4794	0.5238

plemented using the deterministic greedy algorithm, which is in general converging only to local optima in our experiments. However, by examining the two last columns of Table 3 referring to the rate for items from a class lacking training data, it is seen that the marginal classifier achieves higher rate for the two cases with smaller training sets for the other classes. This is an artefact caused in this case by a strong tendency of the marginal classifier to assign a large fraction of the items in general to an unknown source in the presence of very sparse training data for the *a priori* existing classes. This appears to be due to stochastic patterns of deviation between the training and test items, even when they are generated by the same model, causing the predictive data distribution to favor an unknown source.

In order to also compare the relative performance of all three inductive learning principles discussed here, we examined a situation where all classification structures can be investigated enumeratively, such that the effect of algorithms does not confound the comparison of the learning principles. The generating model is here exactly the same as in the first semi-supervised experiment, however, smaller number of training and test items are considered, such that there are in total 674 distinct simultaneous classification structures. Results reported in Table 4 demonstrate that both the simultaneous and marginalized classifiers perform better than the marginal classifier, and that in accordance with the earlier results the marginalized classifier achieves the highest correct classification rate.

6 Discussion

The predictive model considered here is based on an assumption of predictive invariance corresponding to a generalization of de Finetti's representation theorem for data from finite alphabets (Bernardo and Smith 1994; Kallenberg 2005). Although this representation is related to the standard assumption of *i.i.d.* data, there exists an important conceptual difference between modeling data generating mechanism through an *i.i.d.* assumption and predictive invariance. Namely, the former is by necessity defined conditional on a fixed probability measure, which in reality is nearly always

to rate calculated over all items and the label (b) refers to rate calculated only for the items from the class lacking training data

Table 4 Item-wise correct classification rates for three different semi-supervised classification principles based on complete enumeration over classification structures given 1,000 replicate data sets from the generating model. The superscript ‘exact’ for $\hat{p}_{PRED_1}$ and $\hat{p}_{PRED_2}$ indicates that the predictive distribution over classification structures was calculated enumeratively. The abbreviations used in the table are as follows: m_c number of training samples from class c , n_c number

of test samples from class c , k_1 number of classes with training data, k_2 number of classes lacking training data (in the generating model). For these two experimental settings there are 674 distinct simultaneous classification structures, ranging from the S where all the test data are assigned into a single *a priori* existing class to the case where all test samples are assigned into separate singleton classes lacking training data

	$\hat{p}_{PRED_1}$	$\hat{p}_{PRED_2}^{exact}$	$\hat{p}_{PRED_3}^{exact}$
$k_1 = 2, k_2 = 1, d = 10, m_1 = m_2 = 2, n_1 = n_2 = 2, n_3 = 1 :$	0.5782	0.5996	0.6018
$k_1 = 2, k_2 = 1, d = 10, m_1 = m_2 = 2, n_1 = 2, n_2 = 1, n_3 = 2 :$	0.4592	0.5952	0.6174

unknown and can only be learned from finite data to some degree of accuracy. In contrast, the predictive approach inductively defines distribution for any future data set by updating knowledge under axioms of probability using the information from available data, while coherently acknowledging the remaining uncertainty about underlying generating probability measure. Such uncertainty can be considerable in the presence of sparse training data and it should intuitively affect any learning decisions made. When all parameters and hidden variables are integrated out in the predictive approach, the test items will not be independent. The two modeling approaches are therefore united in the predictive sense only at the limit when the amount of training data tends to infinity and the range of eligible labels is fixed *a priori*. It would be interesting to make a more detailed analysis of the rate at which the simultaneous classifier converges to the marginal one under some generic models, such as the exponential family. However, since such an investigation is beyond the scope of this paper, we leave this issue to future work.

In our computational experiments with both synthetic and real data, greedy optimization of the simultaneous classifier performs perhaps surprisingly well and it adds generally only moderate amount of computation time compared to the marginal classifier. Gibbs sampler performs satisfactorily for a range of situations, however, its convergence times may become prohibitive in more challenging classification scenarios, as it is typically expected that the convergence rate of the algorithm grows exponentially as a function of the number of test items. An attractive solution to these problems is to use more advanced stochastic computation algorithms that use more efficient search operators in the space of classification structures. Standard Markov chain Monte Carlo (MCMC) algorithms used for Dirichlet process models provide one attractive alternative (Jain and Neal 2007). Another alternative is provided by the class of non-reversible MCMC algorithms introduced by Corander et al. (2006), which were applied to the learning problem for predictive unsupervised classification (clustering). This class of MCMC algorithms enjoys the attractive property of enabling large degree of freedom in the design of efficient

search operators, while still providing consistent estimates of the sought probabilities. Such property is particularly desirable in applications where the results need to be provided within a relatively short time span. An adaptive extension of the family of non-reversible MCMC algorithms with time-varying proposals would be a particularly interesting target for further work on efficient methods of computation for joint classifiers.

Acknowledgements Work of JC, YC, JS was funded by the European Research Council grant no. 239784 and work of JS was also funded by the FICS graduate school. The authors would like to thank Henrik Nyman and two anonymous reviewers for useful comments and constructive suggestions.

Appendix

Here we provide the technical details related to the asymptotic characteristics of the marginal and simultaneous classifiers stated in Theorems 1 and 2. We recall first the logarithm of the predictive probability (13) for p_{PRED_2} which equals under the assumptions stated earlier:

$$\begin{aligned} \log p(\mathbf{x}^{(N)} | \mathbf{z}^{(M)}, S, T) \\ = \sum_{c=1}^k \sum_{j=1}^d \log \frac{\Gamma(m_c + 1)}{\Gamma(n_c + m_c + 1)} \\ + \sum_{c=1}^k \sum_{j=1}^d \sum_{l=1}^{r_j} \log \frac{\Gamma(n_{cjl} + m_{cjl} + \lambda_{cjl})}{\Gamma(m_{cjl} + \lambda_{cjl})}. \end{aligned} \quad (27)$$

Using Stirling’s approximation on the gamma function terms in (27) it follows that,

$$\begin{aligned} \log p(\mathbf{x}^{(N)} | \mathbf{z}^{(M)}, S, T) \\ = - \sum_{c=1}^k \sum_{j=1}^d \left(m_c + \frac{1}{2} \right) \log \left(1 + \frac{n_c}{m_c + 1} \right) \\ + \sum_{c=1}^k \sum_{j=1}^d \sum_{l=1}^{r_j} \left(m_{cjl} + \lambda_{cjl} - \frac{1}{2} \right) \end{aligned}$$

$$\begin{aligned}
& \times \log \left(1 + \frac{n_{cjl}}{m_{cjl} + \lambda_{cjl}} \right) \\
& - \sum_{c=1}^k \sum_{j=1}^d n_c \log(n_c + m_c + 1) \\
& + \sum_{c=1}^k \sum_{j=1}^d \sum_{l=1}^{r_j} n_{cjl} \log(n_{cjl} + m_{cjl} + \lambda_{cjl}) \\
& = - \sum_{c=1}^k \sum_{j=1}^d \left(m_c + \frac{1}{2} \right) \left[\frac{n_c}{m_c + 1} - \frac{1}{2} \left(\frac{n_c}{m_c + 1} \right)^2 + \dots \right] \\
& + \sum_{c=1}^k \sum_{j=1}^d \sum_{l=1}^{r_j} \left(m_{cjl} + \lambda_{cjl} - \frac{1}{2} \right) \left[\frac{n_{cjl}}{m_{cjl} + \lambda_{cjl}} \right. \\
& \quad \left. - \frac{1}{2} \left(\frac{n_{cjl}}{m_{cjl} + \lambda_{cjl}} \right)^2 + \dots \right] \\
& - \sum_{c=1}^k \sum_{j=1}^d n_c \log(n_c + m_c + 1) \\
& + \sum_{c=1}^k \sum_{j=1}^d \sum_{l=1}^{r_j} n_{cjl} \log(n_{cjl} + m_{cjl} + \lambda_{cjl}),
\end{aligned}$$

where a series expansion of logarithm was used after the second equality sign, and thus we get

$$\begin{aligned}
& \log p(\mathbf{x}^{(N)} | \mathbf{z}^{(M)}, S, T) \\
& \approx - \sum_{c=1}^k \sum_{j=1}^d n_c \log(n_c + m_c + 1) \\
& + \sum_{c=1}^k \sum_{j=1}^d \sum_{l=1}^{r_j} n_{cjl} \log(n_{cjl} + m_{cjl} + \lambda_{cjl}), \quad (28)
\end{aligned}$$

where it is assumed that m_c is sufficiently large for all classes and that all limits of the relative frequencies of feature values are strictly positive under an infinitely exchangeable sampling process of the data, i.e.

$$\begin{aligned}
& \frac{m_{cjl}}{m_c} \xrightarrow{m \rightarrow \infty} \beta_{cjl} > 0, \\
& c = 1, \dots, k, j = 1, \dots, d, l = 1, \dots, r_j. \quad (29)
\end{aligned}$$

For the marginal predictive classifier p_{PRED_1} we can derive the comparable expression by considering

$$\sum_{c=1}^k \sum_{i: S_i=c} \log p(\mathbf{x}_i | \mathbf{z}^{(M)}, T, S_i = c), \quad (30)$$

which can be expanded equivalently to (27) and we get

$$\begin{aligned}
& \sum_{c=1}^k \sum_{i: S_i=c} \log p(\mathbf{x}_i | \mathbf{z}^{(M)}, T, S_i = c) \\
& \approx - \sum_{c=1}^k \sum_{j=1}^d n_c \log(m_c + 2) \\
& + \sum_{c=1}^k \sum_{i: S_i=c} \sum_{j=1}^d \sum_{l=1}^{r_j} n_{i;cjl} \log(n_{i;cjl} + m_{cjl} + \lambda_{cjl}), \quad (31)
\end{aligned}$$

where the sufficient statistic $n_{i;cjl}$ equals 1 if $\mathbf{x}_i$ carries the value l for feature j and $S_i = c$, and $n_{i;cjl} = 0$ otherwise.

Consider now the relationship between data predictive probabilities for the two classifiers under the same classification structure, by investigating the difference between (28) and (31), which can be expressed as

$$\begin{aligned}
& - \sum_{c=1}^k \sum_{j=1}^d n_c \log(n_c + m_c + 1) + \sum_{c=1}^k \sum_{j=1}^d n_c \log(m_c + 2) \\
& + \sum_{c=1}^k \sum_{j=1}^d \sum_{l=1}^{r_j} n_{cjl} \log(n_{cjl} + m_{cjl} + \lambda_{cjl}) \\
& - \sum_{c=1}^k \sum_{i: S_i=c} \sum_{j=1}^d \sum_{l=1}^{r_j} n_{i;cjl} \log(n_{i;cjl} + m_{cjl} + \lambda_{cjl}) \\
& = - \sum_{c=1}^k \sum_{j=1}^d n_c \log \left(\frac{n_c + m_c + 1}{m_c + 2} \right) \\
& - \sum_{c=1}^k \sum_{j=1}^d \sum_{l=1}^{r_j} \log \left(\frac{n_{cjl} + m_{cjl} + \lambda_{cjl}}{1 + m_{cjl} + \lambda_{cjl}} \right)^{n_{cjl}} \xrightarrow{m \rightarrow \infty} 0,
\end{aligned}$$

where the last step necessitates the use of the condition stated in (6). Thus, for any particular classification structure according to the joint rule, the log predictive probability of the data is asymptotically the same as the log predictive probability according to the corresponding marginal classifier, which implies that the two classifiers p_{PRED_1} and p_{PRED_2} will lead to equivalent classifications.

The result stated in Theorem 2 follows by similar arguments. The predictive probability for simultaneous semi-supervised classification $p(\mathbf{x}^{(N)} | \mathbf{z}^{(M)}, S, T)$ equals

$$\begin{aligned}
& \prod_{c=1}^{k_1} \prod_{j=1}^d \frac{\Gamma(m_c + 1)}{\Gamma(n_c + m_c + 1)} \prod_{l=1}^{r_j} \frac{\Gamma(n_{cjl} + m_{cjl} + \lambda_{cjl})}{\Gamma(m_{cjl} + \lambda_{cjl})} \\
& \times \prod_{c=k_1+1}^{k_1+k_2} \prod_{j=1}^d \frac{\Gamma(1)}{\Gamma(n_c + 1)} \prod_{l=1}^{r_j} \frac{\Gamma(n_{cjl} + \lambda_{cjl})}{\Gamma(\lambda_{cjl})}, \quad (32)
\end{aligned}$$

and the corresponding probability for the marginal classifier is

$$p(\mathbf{x}_1, \dots, \mathbf{x}_n | \mathbf{z}^{(M)}, S, T) = \prod_{c=1}^{k_1+k_2} \sum_{i:S_i=c} p(\mathbf{x}_i | \mathbf{z}^{(M)}, S_i = c, T). \quad (33)$$

Note that the predictive probability for semi-supervised classification can also be written in the form of the supervised classification as follows

$$p(\mathbf{x}^{(N)} | \mathbf{z}^{(M)}, S, T) = \prod_{c=1}^{k_1+k_2} \prod_{j=1}^d \frac{\Gamma(m_c + 1)}{\Gamma(n_c + m_c + 1)} \times \prod_{l=1}^{r_j} \frac{\Gamma(n_{cjl} + m_{cjl} + \lambda_{cjl})}{\Gamma(m_{cjl} + \lambda_{cjl})}, \quad (34)$$

while $m_c = 0$ and $m_{cjl} = 0$ for all $c = k_1 + 1, \dots, k_1 + k_2$. Then, according to the same derivation in supervised case, the logarithm of (32) can be expressed as

$$\log p(\mathbf{x}^{(N)} | \mathbf{z}^{(M)}, S, T) \approx - \sum_{c=1}^{k_1+k_2} \sum_{j=1}^d n_c \log(n_c + m_c + 1) + \sum_{c=1}^{k_1+k_2} \sum_{j=1}^d \sum_{l=1}^{r_j} n_{cjl} \log(n_{cjl} + m_{cjl} + \lambda_{cjl}), \quad (35)$$

where $m_c = 0$ and $m_{cjl} = 0$ for all $c = k_1 + 1, \dots, k_1 + k_2$.

Similarly, we obtain for the marginal case

$$\log p(\mathbf{x}_1, \dots, \mathbf{x}_n | \mathbf{z}^{(M)}, S, T) \approx - \sum_{c=1}^{k_1+k_2} \sum_{j=1}^d n_c \log(m_c + 2) + \sum_{c=1}^{k_1+k_2} \sum_{i \in s_c} \sum_{j=1}^d \sum_{l=1}^{r_j} n_{i;cjl} \log(n_{i;cjl} + m_{cjl} + \lambda_{cjl}). \quad (36)$$

The difference between the two log probabilities is thus

$$= - \sum_{c=1}^{k_1+k_2} \sum_{j=1}^d n_c \log\left(\frac{n_c + m_c + 1}{m_c + 2}\right) + \sum_{c=1}^{k_1+k_2} \sum_{j=1}^d \sum_{l=1}^{r_j} \log\left(\frac{n_{cjl} + m_{cjl} + \lambda_{cjl}}{1 + m_{cjl} + \lambda_{cjl}}\right)^{n_{cjl}}, \quad (37)$$

which tends to

$$\sum_{c=k_1+1}^{k_1+k_2} \sum_{j=1}^d \sum_{l=1}^{r_j} \log\left(\frac{n_{cjl} + \lambda_{cjl}}{1 + \lambda_{cjl}}\right)^{n_{cjl}}$$

$$- \sum_{c=k_1+1}^{k_1+k_2} \sum_{j=1}^d n_c \log\left(\frac{n_c + 1}{2}\right), \quad (38)$$

when $m \rightarrow \infty$.

References

- Basu, S.: Semi-supervised clustering: probabilistic models, algorithms and experiments. Ph.D. thesis, Department of Computer Sciences, UT at Austin (2005)
- Bailey, N.T.J.: Probability methods of diagnosis based on small samples. In: Mathematics and Computer Science in Biology and Medicine. H.M. Stationery Office, London (1965)
- Bernardo, J.M., Smith, A.F.M.: Bayesian Theory. Wiley, Chichester (1994)
- Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York (2007)
- Chapelle, O., Schölkopf, B., Zien, A.: Introduction to semi-supervised learning. In: Chapelle, O., et al. (eds.) Semi-Supervised Learning, pp. 1–12. MIT Press, Cambridge (2006)
- Chow, C.K., Liu, C.N.: Approximating discrete probability distributions with dependence trees. IEEE Trans. Inf. Theory **14**, 462–467 (1968)
- Corander, J., Gyllenberg, M., Koski, T.: Bayesian model learning based on a parallel MCMC strategy. Stat. Comput. **16**, 355–362 (2006)
- Corander, J., Gyllenberg, M., Koski, T.: Random partition models and exchangeability for Bayesian identification of population structure. Bull. Math. Biol. **69**, 797–815 (2007)
- Corander, J., Gyllenberg, M., Koski, T.: Bayesian unsupervised classification framework based on stochastic partitions of data and a parallel search strategy. Adv. Data Anal. Classif. **3**, 3–24 (2009)
- Devroye, L., Györfi, L., Lugosi, G.: A Probabilistic Theory of Pattern Recognition. Springer, New York (1996)
- Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley, New York (2000)
- Geisser, S.: Posterior odds for multivariate normal classifications. J. R. Stat. Soc. B **26**, 69–76 (1964)
- Geisser, S.: Predictive discrimination. In: Krishnajah, P.R. (ed.) Multivariate Analysis, pp. 149–163. Academic Press, New York (1966)
- Geisser, S.: Predictive Inference: An Introduction. Chapman & Hall, London (1993)
- Gyllenberg, M., Koski, T., Verlaan, M.: Classification of binary vectors by stochastic complexity. J. Multivar. Anal. **63**, 47–72 (1997a)
- Gyllenberg, H.G., Gyllenberg, M., Koski, T., Lund, T., Schindler, J., Verlaan, J.: Classification of *Enterobacteriaceae* by minimization of stochastic complexity. Microbiology **143**, 721–732 (1997b)
- Hand, D.J., Yu, K.: Idiot's Bayes: not so stupid after all. Int. Stat. Rev. **69**, 385–398 (2001)
- Howson, C.: Hume's Problem: Induction and the Justification of Belief. Oxford University Press, Oxford (2000)
- Huo, Q., Lee, C.-H.: A Bayesian predictive classification approach to robust speech recognition. IEEE Trans. Speech Audio Process. **8**, 200–204 (2000)
- Jain, S., Neal, R.M.: Splitting and merging components of a nonconjugate Dirichlet process mixture model. Bayesian Anal. **3**, 445–472 (2007)
- Kallenberg, O.: Probabilistic Symmetries and Invariance Principles. Springer, New York (2005)
- Nádas, A.: Optimal solution of a training problem in speech recognition. IEEE Trans. Acoust. Speech Signal Process. **33**, 326–329 (1985)

- Ripley, B.D.: Statistical Inference for Spatial Processes. Cambridge University Press, Cambridge (1988)
- Ripley, B.D.: Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge (1996)
- Robert, C.P., Casella, G.: Monte Carlo Statistical Methods, 2nd edn. Springer, New York (2005)
- Solomonoff, R.J.: A formal theory of inductive inference. Inf. Control **7**, 1–22 (1964)
- Solomonoff, R.J.: Three kinds of probabilistic induction: universal distributions and convergence theorems. Comput. J. **51**, 566–570 (2008)
- Stam, A.J.: Generation of a random partition of a finite set by an urn model. J. Comb. Theory, Ser. A **35**, 231–240 (1983)
- Zabell, S.L.: W.E. Johnson's 'sufficientness' postulate. Ann. Stat. **10**, 1091–1099 (1982)

On the generalization of ECP and OA methods to nonsmooth convex MINLP problems

Ville-Pekka Eronen^{a*}, Marko M. Mäkelä^a and Tapio Westerlund^b

^a*Department of Mathematics and Statistics, University of Turku, FI-20014 Turku, Finland;* ^b*Process Design and Systems Engineering Laboratory, Åbo Akademi University, Biskopsgatan 8, FI-20500 Åbo, Finland*

(Received 2 December 2010; final version received 10 July 2012)

In this article, generalization of some mixed-integer nonlinear programming algorithms to cover convex nonsmooth problems is studied. In the extended cutting plane method, gradients are replaced by the subgradients of the convex function and the resulting algorithm shall be proved to converge to a global optimum. It is shown through a counterexample that this type of generalization is insufficient with certain versions of the outer approximation algorithm. However, with some modifications to the outer approximation method a special type of nonsmooth functions for which the subdifferential at any point is a convex combination of a finite number of subgradients at the point can be considered. Numerical results with extended cutting plane method are also reported.

Keywords: nonsmooth optimization; extended cutting plane algorithm; outer approximation algorithm; subgradient

AMS Subject Classifications: 90C11; 90C25; 90C56

1. Introduction

Several methods for mixed-integer nonlinear programming (MINLP) problems have been developed during the past few decades and they can be divided into three main classes, namely cutting plane (CP) (see, e.g. [21,22]), branch-and-bound (BB) (see, e.g. [6,16]) and outer approximation (OA) (see, e.g. [4,5,24]) type of methods.

At the same time, nonsmooth optimization (NSO) has also been widely studied. Different kinds of subgradient (see, e.g. [1,2,20]) and bundle methods (see e.g. [7,12–15,17,19]) have been introduced to solve continuous, possibly NSO problems. However, the combination of these two optimization areas is amazingly rare, although, both bundle methods for NSO and CP methods for MINLP have their origin in the same classical CP method of Kelley [10]. In addition, the OA method [5] was considered with exact penalty function and some variants that are nonsmooth.

In this article, generalizations of some MINLP algorithms to cover convex nonsmooth MINLP problems are studied. In the extended CP (ECP) [22] method,

*Corresponding author. Email: vpoero@utu.fi

gradients are replaced by the subgradients of convex function and the resulting algorithm shall be proved to converge to global optimum. It is shown through a counterexample that this type of generalization is insufficient with certain versions of OA algorithm. However, with some modifications to the OA method, a special type of nonsmooth functions for which subdifferentials at any point is a convex combination of finite number of subgradients at the point can be considered.

The BB method is not discussed extensively here, since the generalization is quite evident. The BB method solves the nonsmooth MINLP if continuous NSO problems can be solved to global optimum in each node. This can be done, for example, by the bundle method presented in [12], if the objective and constraint functions are convex.

This article is organized as follows. Section 2 introduces the general problem discussed. In Section 3 the generalization of the ECP method is shown to converge to a global optimum. In Section 4 it is proved that the simple generalization discussed above is insufficient for the OA method. The algorithm is then modified a bit to cover special types of nonsmooth functions. Section 5 illustrates the generalizations through an example and in Section 6 the ECP method is applied to some small nonsmooth MINLP problems. Section 7 summarizes the results.

2. The optimization problem

In this article, the following MINLP problem is considered:

$$\begin{aligned} \min & f(x, y) \\ \text{s.t. } & g_j(x, y) \leq 0, \quad j = 1, \dots, J, \\ & x \in X, y \in Y, \end{aligned} \quad (\hat{P})$$

where functions $f: \mathbb{R}^n \times \mathbb{Z}^m \rightarrow \mathbb{R}$ and $g_j: \mathbb{R}^n \times \mathbb{Z}^m \rightarrow \mathbb{R}$, $j = 1, \dots, J$ are convex, possibly nonsmooth functions, X is a compact, convex polyhedral set

$$X = \{x \mid Ax \leq b\},$$

where A is $(p \times n)$ matrix and b is $(p \times 1)$ matrix and Y is a finite set in $\mathbb{Z}^m$. It is assumed that the functions g_j are nonlinear. If there are linear constraints with both continuous and integer variables these would be treated in sequel quite similarly as the constraints in the set X .

The aim of this article is to generalize the well-known ECP [22] and OA [5] methods for MINLP problems to also cover nonsmooth problems of form $(\hat{P})$.

Nonsmoothness implies that gradients may not exist at every point in the set $X \times Y$. To deal with this problem subgradients for convex functions are used instead.

Definition 2.1 Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be a convex function. The *subgradient* of the function f at the point x^0 is any vector $\xi(x^0)$ that satisfies the condition

$$f(x^0) + \xi(x^0)^T(x - x^0) \leq f(x), \quad \text{for all } x \in \mathbb{R}^n. \quad (1)$$

The set of all vectors $\xi(x^0)$ satisfying condition (1) is called *subdifferential* at the point x^0 and is denoted by $\partial f(x^0)$.

It can be proved that the subdifferential of a convex function is a nonempty, convex and compact set. Also, a gradient of a differentiable convex function is always a subgradient [15].

In our attempt to generalize the ECP and OA methods, gradients are replaced by subgradients. In the following, we shall prove that generalization will be successful with the ECP method presented in [22] but not with the OA method presented in [5]. However, after some modifications to the OA method it will also cover problems of form $(\hat{P})$ with some further assumptions on objective and constraint functions. Next, some useful definitions and lemmas needed in the convergence proofs are presented.

Definition 2.2 Function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is *locally Lipschitz continuous* at point x^0 , if there exist $\epsilon > 0$ and K such that the inequality

$$|f(y) - f(z)| \leq K \|y - z\| \quad (2)$$

holds for all $y, z \in B(x^0, \epsilon)$. Here K is called a local Lipschitz constant of function f at point x^0 and $B(x^0, \epsilon)$ is an open ball with a centre x^0 and radius ϵ .

Note that convex functions are locally Lipschitz continuous [15].

LEMMA 2.3 Let K be a local Lipschitz constant of a convex function f at point x^0 . Then, the inequality

$$\|\xi(x^0)\| \leq K \quad (3)$$

holds for all $\xi(x^0) \in \partial f(x^0)$.

Proof See [15, p. 14]. ■

A convex combination of sets $A_1, \dots, A_I$ is denoted by

$$\text{conv}\{A_1, \dots, A_I\} := \left\{ \lambda_1 a_1 + \dots + \lambda_I a_I \mid a_i \in A_i, \sum_{i=1}^I \lambda_i = 1, \lambda_i \geq 0, i = 1, \dots, I \right\}.$$

LEMMA 2.4 If functions $f_i, i = 1, \dots, I$ are convex then the function

$$f(x) := \max_{i=1, \dots, I} \{f_i(x)\}$$

is also convex and

$$\partial f(x) = \text{conv}\{\partial f_i(x) \mid i \in \mathcal{I}(x)\},$$

where $\mathcal{I}(x) = \{i \mid f_i(x) = f(x)\}$.

Proof See [15, p. 47]. ■

3. Generalization of ECP method

In the ECP method [22], the objective function of problem $(\hat{P})$ is turned into a constraint by introducing a constraint

$$g_{J+1}(x, y, \mu) := f(x, y) - \mu \leq 0$$

and minimizing variable μ . Function g_{J+1} is convex since it is a sum of two convex functions $f(x, y)$ and $-\mu$. Also, the scalars $\mu_{\min}$ and $\mu_{\max}$ such that the inequalities

$$\mu_{\min} \leq f(x, y) \leq \mu_{\max} \quad (4)$$

hold for all $x \in X$ and $y \in Y$ for which $g_j(x, y) \leq 0$, $j = 1, \dots, J$, should be known *a priori*. Then, we may introduce linear constraints $\mu_{\min} \leq \mu$ and $\mu \leq \mu_{\max}$ to the problem. This is necessary in order to keep the decision space compact and to guarantee the convergence to a global optimum.

Introducing variable $z = (x, y, \mu) \in \mathbb{R}^n \times \mathbb{Z}^m \times \mathbb{R}$ will simplify notations. Now, the constraint functions g_j are treated as functions of z . The use of the function

$$\tilde{g}(z) := \max_{j=1, \dots, J+1} \{g_j(z)\}$$

will facilitate the proof. According to Lemma 2.4 $\tilde{g}$ is convex and

$$\partial \tilde{g}(z) = \text{conv}\{\partial g_i(z) \mid g_i(z) = \tilde{g}(z)\}.$$

If we denote $L := X \times Y \times [\mu_{\min}, \mu_{\max}]$ and $N := \{z \mid \tilde{g}(z) \leq 0\}$ then the problem ($\hat{P}$) can be stated as follows:

$$\begin{aligned} \min \quad & z_{m+n+1} \\ \text{s.t.} \quad & z \in N \cap L. \end{aligned} \quad (P)$$

THEOREM 3.1 *Point $(\hat{x}, \hat{y}, \hat{z}_{m+n+1})$ is a global optimum of the problem (P) iff the point $(\hat{x}, \hat{y})$ is a global optimum of problem ($\hat{P}$).*

Proof Due to box constraints (4) and constraints of problems ($\hat{P}$) and (P) a point (x, y) is feasible in the problem ($\hat{P}$) iff $(x, y, f(x, y))$ is feasible in the problem (P). If $(\tilde{x}, \tilde{y}, \tilde{z}_{m+n+1})$ is a global optimum of the problem (P) then, $\tilde{z}_{m+n+1} = f(\tilde{x}, \tilde{y})$ due to constraint $f - z_{m+n+1} \leq 0$. Now, assume that $(\hat{x}, \hat{y})$ is a global optimum of the problem ($\hat{P}$) and $(\tilde{x}, \tilde{y}, f(\tilde{x}, \tilde{y}))$ is a global optimum of the problem (P). Since $(\hat{x}, \hat{y}, f(\hat{x}, \hat{y}))$ is feasible in (P) the inequality $f(\tilde{x}, \tilde{y}) \leq f(\hat{x}, \hat{y})$ holds. Since the point $(\tilde{x}, \tilde{y})$ is feasible in the problem ($\hat{P}$) also $f(\hat{x}, \hat{y}) \leq f(\tilde{x}, \tilde{y})$. Hence, $f(\hat{x}, \hat{y}) = f(\tilde{x}, \tilde{y})$ and the theorem is proved. ■

In the ECP method, the MILP problems where nonlinear constraint $\tilde{g}$ is linearized at previous solution points are solved in each step. In the beginning the nonlinear constraints are left out and the problem

$$\begin{aligned} \min \quad & z_{m+n+1} \\ & z \in L \end{aligned} \quad (P^0)$$

is solved. Then, at iteration $i > 0$ we solve the problem

$$\begin{aligned} \min \quad & z_{m+n+1} \\ \text{s.t.} \quad & \tilde{g}(z^k) + \tilde{\xi}(z^k)^T(z - z^k) \leq 0, \quad k = 0, \dots, i-1, \\ & z \in L, \end{aligned} \quad (P^i)$$

where z^k is the solution of problem (P^k). The algorithm is as follows:

ECP Algorithm 3.2

Step 0 Set $i=0$, give some small feasibility tolerance $\epsilon_g > 0$ and formulate problem (P^0).

Step 1 Solve MILP problem (P^i) . Let the solution point be z^i .

Step 2 Check whether z^i satisfies the nonlinear constraints that is $\tilde{g}(z^i) \leq \epsilon_g$. If the nonlinear constraints are satisfied stop, point z^i is a global optimum of the problem (P) .

Step 3 Create a new problem (P^{i+1}) by adding the constraint $l^i(z) = \tilde{g}(z^i) + \tilde{\xi}(z^i)^T(z - z^i) \leq 0$, where $\tilde{\xi}(z^i) \in \partial\tilde{g}(z^i)$ is arbitrary.

Step 4 Set $i = i + 1$ and go to step 1.

Note, that in step 1 the optimization problem can be solved by using any suitable MILP solver.

Next, the convergence proof of the algorithm to a global optimum is given. The proof necessitates that $\epsilon_g = 0$. First, we show that CPs introduced in the algorithm do not cut off any part of the feasible set of the problem (P) . Let z^i be a solution point of the problem (P^i) . Due to convexity of the function $\tilde{g}$ the inequality

$$l^i(z) = \tilde{g}(z^i) + \tilde{\xi}(z^i)^T(z - z^i) \leq \tilde{g}(z) \leq 0 \quad (5)$$

holds for all $z \in N \cap L$. Thus, any feasible point of (P) remains feasible in MILP problems. If we denote the feasible set of problem (P^i) by

$$\Omega^i = \{z \mid z \in \Omega^{i-1}, l^{i-1}(z) \leq 0\}$$

and $\Omega^0 = L$ we get a sequence

$$N \cap L \subseteq \dots \subseteq \Omega^i \subseteq \dots \subseteq \Omega^0. \quad (6)$$

If the algorithm ends with a finite number of iterations say at i , the last solution point is feasible in the problem (P) . This point is also a global optimum of the problem (P) , since the point is optimum at the feasible set of problem (P^i) and this set includes the feasible set of the problem (P) according to (6).

Next, we assume that the algorithm does not stop after a finite number of iterations. Since $l^i(z^i) = \tilde{g}(z^i) > 0$ the constraint (5) makes the point z^i infeasible in subsequent MILP problems. Hence, the algorithm generates a sequence of different points. The solution sequence belongs to the compact set L . Since L is bounded, the classical Bolzano–Weierstrass theorem implies that the sequence has an accumulation point. It remains to prove that the accumulation point is feasible in (P) .

LEMMA 3.3 *An accumulation point generated by the ECP algorithm in the problem (P) is a feasible point.*

Proof Let the accumulation point be $\hat{z}$. Since the sequence (z^j) belongs to the closed set L , also $\hat{z} \in L$. To prove the feasibility of $\hat{z}$ it suffices to prove that $\hat{z} \in N$. Let $\epsilon > 0$ be arbitrary and $\delta > 0$ be so small that local Lipschitz condition (2) with Lipschitz constant K holds for $\tilde{g}$ in $B(\hat{z}, \delta)$. Denote $\hat{\epsilon} = \min\{\frac{\epsilon}{3K}, \delta\}$. Since $\hat{z}$ is the accumulation point, there exist indices $k > j$ such that for MILP solution points z^j and z^k the relation $z^j, z^k \in B(\hat{z}, \hat{\epsilon})$ holds. Since the algorithm does not terminate after a finite number of iterations, points in the solution sequence (z^j) are in the closed set $\{z \mid \tilde{g}(z) \geq 0\}$. The accumulation point $\hat{z}$ also belongs to this set, because the set is closed. Since $k > j$, the inequality $l^j(z^k) \leq 0$ holds, and thus the inequality

$$\tilde{g}(\hat{z}) \leq |\tilde{g}(\hat{z}) - l^j(z^k)| = |\tilde{g}(\hat{z}) - \tilde{g}(z^j) - \tilde{\xi}(z^j)^T(z^k - z^j)|$$

is also true. From the triangle and the Cauchy–Schwarz inequalities it follows that

$$\left| \tilde{g}(\hat{z}) - \tilde{g}(z^j) - \tilde{\xi}(z^j)^T(z^k - z^j) \right| \leq \left| \tilde{g}(\hat{z}) - \tilde{g}(z^j) \right| + \left\| \tilde{\xi}(z^j) \right\| \|z^k - z^j\|.$$

Then the local Lipschitz continuity and Lemma 2.3 imply

$$\begin{aligned} \left| \tilde{g}(\hat{z}) - \tilde{g}(z^j) \right| + \left\| \tilde{\xi}(z^j) \right\| \|z^k - z^j - \hat{z} + \hat{z}\| &\leq K \|\hat{z} - z^j\| + K \|z^k - \hat{z}\| + K \|\hat{z} - z^j\| \\ &< K \frac{\epsilon}{3K} + K \frac{\epsilon}{3K} + K \frac{\epsilon}{3K} = \epsilon. \end{aligned}$$

Thus, $\tilde{g}(\hat{z}) = 0$ proving the lemma. ■

We sum up the convergence result in the following theorem.

THEOREM 3.4 *ECP Algorithm 3.2 with $\epsilon_g = 0$ converges to a global optimum when solving the problem (P).*

Proof As stated earlier, the added CPs do not cut off any part of the feasible set of the problem (P). Thus, the feasible set of the problem (P) is a subset of the feasible set of any occurring MILP problem. If the solution of the MILP problem is feasible in (P) it is also a global optimum. If the algorithm stops after a finite number of iterations a feasible point will be obtained. If the algorithm generates an infinite sequence it will have an accumulation point that satisfies the constraints according to Lemma 3.3. Hence, the algorithm converges to a global optimum. ■

It is important to note that there can be many accumulation points, but all the accumulation points are, possibly different, points defining the global optimum.

As a consequence of the above results, the ECP method presented in [22] can be applied also to nonsmooth problems without any modifications. In [8] the authors mentioned a version of ECP method in which every violated constraint was linearized at each step. Also this version can be applied to nonsmooth problems. The proof of the convergence is quite similar to that given above and it is presented in the appendix.

4. About the generalization of OA method

Next, the classical OA method presented in [5] is considered. It is assumed here that the objective function and the constraint functions of the problem ($\hat{P}$) are continuously differentiable. A short presentation of the algorithm follows.

The OA method proceeds by alternating between a continuous nonlinear programming problem and an MILP problem. The nonlinear problem corresponds to the problem ($\hat{P}$) with fixed integer value $y = y^i$, that is, the problem

$$\begin{aligned} \min & f(x, y^i) \\ \text{s.t. } & g_j(x, y^i) \leq 0, \quad j = 1, \dots, J, \\ & x \in X \end{aligned} \tag{NLP}_{y^i}$$

is solved at iteration i . If (NLP_{y^i}) is feasible, an upper bound UB^i is obtained for the objective function. The best upper bound is stored for the MILP problem as

$$UBD^i = \min \{ UB^j \mid j \leq i \text{ and } (NLP_{y^j}) \text{ is feasible} \}.$$

If (NLP_{y^i}) is infeasible a feasibility problem (F_{y^i}) is solved instead. A special case of feasibility problems presented in [5] is

$$\begin{aligned} \min \quad & \mu \\ \text{s.t.} \quad & g_j(x, y^i) \leq \mu, \quad j = 1, \dots, J, \\ & x \in X. \end{aligned} \quad (F_{y^i})$$

Indices for which (NLP_{y^i}) is feasible is stored in set T^i :

$$T^i = \{j \mid j \leq i, (NLP_{y^j}) \text{ was feasible}\}.$$

Similarly, S^i stores indices where (NLP_{y^i}) was infeasible:

$$S^i = \{j \mid j \leq i, (NLP_{y^j}) \text{ was infeasible}\}.$$

Let x^i be the solution of (NLP_{y^i}) or (F_{y^i}) . The MILP problem is then formulated as follows:

$$\begin{aligned} \min \quad & \mu \\ \text{s.t.} \quad & \mu < UBD^i, \\ & f(x^i, y^i) + \nabla f(x^i, y^i)^T \begin{pmatrix} x - x^i \\ y - y^i \end{pmatrix} \leq \mu \quad \forall i \in T^i, \\ & g_j(x^i, y^i) + \nabla g_j(x^i, y^i)^T \begin{pmatrix} x - x^i \\ y - y^i \end{pmatrix} \leq 0 \quad \forall i \in T^i \cup S^i, j = 1, \dots, J, \\ & x \in X, y \in Y. \end{aligned} \quad (M^i)$$

Let the solution of (M^i) be (x^{i+1}, y^{i+1}) . Next, the problem $(NLP_{y^{i+1}})$ or, if infeasible, the problem $(F_{y^{i+1}})$ is solved. The OA algorithm proceeds as follows:

Algorithm 4.1 [5]

Step 0 Let $y^0 \in \mathbb{Z}^m$ be given. Set $i=0$, $T^{-1} = \emptyset$, $S^{-1} = \emptyset$ and $UBD^{-1} = \infty$.

Step 1 Solve the problem (NLP_{y^i}) , or the feasibility problem (F_{y^i}) , if (NLP_{y^i}) is infeasible, and let the solution be x^i .

Step 2 Linearize the objective and constraint functions about (x^i, y^i) . Set $T^i = T^{i-1} \cup \{i\}$ or $S^i = S^{i-1} \cup \{i\}$ as appropriate.

Step 3 If (NLP_{y^i}) is feasible and $f(x^i, y^i) < UBD^{i-1}$ then update current best point by setting $x^* = x^i$, $y^* = y^i$ and $UBD^i = f(x^i, y^i)$.

Step 4 Solve the problem (M^i) , giving a new integer assignment y^{i+1} to be tested in the algorithm. If (M^i) is infeasible stop: the solution is (x^*, y^*) , else set $i = i + 1$ and go to step 1.

In addition to the assumptions of the problem $(\hat{P})$, the constraints are supposed to satisfy the Slater constraint qualification condition in problems (NLP_{y^i}) and (F_{y^i}) for every y^i . This means that for every y^i there should exist a point $x \in X$ such that $g_j(x, y^i) < 0$ for all $j=1, \dots, J$. This requirement enables us to apply the KKT-conditions.

The proof that OA method converges is based on the following three steps.

LEMMA 4.2 *If (NLP_{y^i}) is infeasible and (x^i, y^i) is the solution to the (F_{y^i}) problem, then $y = y^i$ is infeasible in the constraints*

$$g_j(x^i, y^i) + \nabla g_j(x^i, y^i)^T \begin{pmatrix} x - x^i \\ y - y^i \end{pmatrix} \leq 0, \quad j = 1, \dots, J$$

for all $x \in X$.

Proof See [5, Lemma 1]. ■

LEMMA 4.3 *If (NLP_{y^i}) is feasible and (x^i, y^i) is the solution point, then $y = y^i$ is infeasible in the constraints*

$$\begin{aligned} \mu &< UBD^i, \\ f(x^i, y^i) + \nabla f(x^i, y^i)^T \begin{pmatrix} x - x^i \\ y - y^i \end{pmatrix} &\leq \mu, \\ g_j(x^i, y^i) + \nabla g_j(x^i, y^i)^T \begin{pmatrix} x - x^i \\ y - y^i \end{pmatrix} &\leq 0, \quad j = 1, \dots, J \end{aligned}$$

for all $x \in X$.

Proof This proof can be found in [5] as a part of the proof of Theorem 2. ■

The two previous lemmas show that the integer part of the solution of the feasible (M^i) problem is different from the previous solutions. Since Y is finite the algorithm ends up with an infeasible (M^i) problem. Now, it remains to prove that when (M^i) is infeasible the upper bound UBD^i is a global optimum. This follows from the fact that linearizations in (M^i) do not cut off any feasible point of problem (P) if the objective function f is smaller than UBD^i at the feasible point. This in turn follows from the convexity of the objective and constraint functions, especially, from inequality (1) for gradients. The proof can be found in [5, Theorem 2].

4.1. Counterexample

Next, we shall show that substitution of the gradient by an arbitrary subgradient in OA is insufficient. Let

$$g(x, y) = \max \left\{ -\frac{3}{2} - x + y, -\frac{7}{2} + y + x \right\}$$

be the convex, nonsmooth constraint function and consider the following problem:

$$\begin{aligned} \min \quad & 2x - y \\ \text{s.t.} \quad & g(x, y) \leq 0, \\ & y - 4x - 1 \leq 0, \\ & 0 \leq x \leq 2, y \in Y, \end{aligned} \tag{E}$$

where $Y = \{0, 1, 2, 3, 4, 5\}$. Let the initial point be $y^0 = 3$.

There are no feasible points in the problem (NLP_{y^0}) and thus the problem (F_{y^0})

$$\begin{aligned} \min \quad & \mu \\ \text{s.t.} \quad & \max \left\{ \frac{3}{2} - x, -\frac{1}{2} + x \right\} \leq \mu, \\ & 2 - 4x \leq 0, \\ & 0 \leq x \leq 2 \end{aligned}$$

will be solved. The solution is $x^0 = 1$ with $\mu = \frac{1}{2}$. Then g will be linearized at point $(x^0, y^0) = (1, 3)$ for the next MILP problem. Both the functions $-\frac{3}{2} - x + y$ and $-\frac{7}{2} + y + x$ have the same value $\frac{1}{2}$ at the point (x^0, y^0) and thus according to Lemma 2.4 the subdifferential is

$$\partial g(1, 3) = \{(\alpha, 1)^T \mid \alpha \in [-1, 1]\}. \quad (7)$$

Since we may select an arbitrary subgradient we can choose for example $\xi(z^0) = (1, 1)^T$. Thus, the new linear constraint is

$$\frac{1}{2} + (1, 1)(x - 1, y - 3)^T \leq 0 \Rightarrow x + y - \frac{7}{2} \leq 0.$$

The resulting problem (M^0) is

$$\begin{aligned} \min \quad & 2x - y \\ \text{s.t.} \quad & x + y - \frac{7}{2} \leq 0, \\ & y - 4x - 1 \leq 0, \\ & 0 \leq x \leq 2, y \in Y. \end{aligned}$$

The solution point of problem (M^0) is $(\frac{1}{2}, 3)$. Here $y^1 = y^0$ and $F_{y^1} \equiv F_{y^0}$, thus OA may generate an infinite loop between points $(1, 3)$ and $(\frac{1}{2}, 3)$. Both of them are infeasible, but the problem (E) has a feasible point $(0, 1)$ for example. Thus, OA with subgradients could not find an optimum.

In [5], the authors propose an alternative objective function to the feasibility problem (F_{y^i}) utilizing the l_1 -norm of the violated constraint functions. In the previous counterexample there was only one nonlinear constraint function and, thus, l_1 -norm and used l_∞ -norm coincide. Hence, the counterexample applies for both versions of the feasibility problem.

4.2. Modified OA method

If the integer variables of the solutions obtained in problems (M^i) are explicitly cut off with integer cuts from the following problems (M^i) then the generalization would be sufficient. This follows from the inequality (1) which ensures that feasible solutions for which f is smaller than UBD^i are not cut off in problem (M^i) . However, as pointed out in [5], integer cuts are practical only with binary variables.

When tracking down why properties proved for continuously differentiable functions in [5] are not valid here, the reason was noticed to be the KKT-conditions.

We know that for differentiable functions the KKT-conditions for the (F_{y^i}) problem are the equations

$$\nabla\mu + \sum_{j=1}^J \lambda_j \nabla(g_j(x, y^i) - \mu) + \eta = 0, \quad (8)$$

$$\begin{aligned} \lambda_j(g_j(x, y^i) - \mu) &= 0, \quad j = 1, \dots, J, \\ \lambda_j &\geq 0, \quad j = 1, \dots, J, \end{aligned} \quad (9)$$

where $\eta \in N_X(x)$ and $N_X(x)$ is the normal cone of the set X at the point x . Since X is a convex set the inequality

$$\eta^T(x^0 - x) \leq 0 \quad (10)$$

holds for all $x^0 \in X$ and $\eta \in N_X(x)$. For nonsmooth functions the KKT-conditions reads

$$0 \in \partial\mu + \sum_{j=1}^J \lambda_j \partial(g_j(x, y^i) - \mu) + N_X(x), \quad (11)$$

$$\begin{aligned} \lambda_j(g_j(x, y^i) - \mu) &= 0, \quad j = 1, \dots, J, \\ \lambda_j &\geq 0, \quad j = 1, \dots, J. \end{aligned} \quad (12)$$

Observe that in the nonsmooth case, Equation (11) does not hold for all subgradients but only for certain ones contrary to the smooth case where (8) is an equality. If we could always find such subgradients that the equality holds for (11) also the proof in [5] would hold and OA would converge also in the nonsmooth case.

A special class of nonsmooth functions can be dealt with in the following way. Consider a nonsmooth function f for which at any point the subdifferential can be generated by a convex combination of finite number of subgradients. That is, for a point z there exist subgradients $\xi^1(z), \dots, \xi^K(z)$ such that

$$\partial f(z) = \left\{ \alpha_k \xi^k(z) \mid \sum_{k=1}^K \alpha_k = 1, \alpha_k \geq 0, k = 1, \dots, K \right\}. \quad (13)$$

Note that we must know these subgradients explicitly at every point where we make linearizations. An example of such a function is the maximum of a finite number of convex differentiable functions [15]. It should be noted that this type of functions can be dealt with also the exact penalty function formulation used in [5]. Also, the maximum of a finite number of differentiable functions can be replaced by several differentiable functions (see, e.g. [9, p. 366]) and those can be dealt with the OA method for differentiable functions. In the generalization of the OA method to work with these kinds of functions possibly many linearizations are made instead of one in step 2 of Algorithm 4.1. For example, consider the function f that should be linearized at point (x^i, y^i) for the next MILP problem (M^i) and its subdifferential,

which is of the form (13). Then, new linearizations for (M^i) would be

$$\begin{aligned} f(x^i, y^i) + \xi^1(x^i, y^i)^T \begin{pmatrix} x - x^i \\ y - y^i \end{pmatrix} &\leq 0, \\ &\vdots \\ f(x^i, y^i) + \xi^K(x^i, y^i)^T \begin{pmatrix} x - x^i \\ y - y^i \end{pmatrix} &\leq 0. \end{aligned}$$

For a certain linearized function subgradients in different linearizations must be such that the subdifferential can be constructed with a convex combination of those subgradients. This is different from the requirements for the ECP method where it was sufficient to know only one arbitrary subgradient.

Next, we shall prove that with this generalization Lemma 4.2 holds.

LEMMA 4.4 *If (NLP_{y^i}) is infeasible and (x^i, y^i) is the solution to the problem (F_{y^i}) , then $y = y^i$ is infeasible in the constraints*

$$g_j(x^i, y^i) + \xi_j^k(x^i, y^i)^T \begin{pmatrix} x - x^i \\ y - y^i \end{pmatrix} \leq 0, \quad j = 1, \dots, J, k = 1, \dots, K^{ij}, \quad (14)$$

where $\xi_j^1, \dots, \xi_j^{K^{ij}}$ are such subgradients that any vector from the subdifferential $\partial g_j(x^i, y^i)$ is a convex combination of those subgradients.

Proof Suppose by contradiction, y^i is feasible in constraints (14). Since x^i is optimal in (F_{y^i}) , the KKT-conditions hold and

$$0 \in \partial \mu + \sum_{j=1}^J \lambda_j \partial (g_j(x^i, y^i) - \mu) + N_X(x^i).$$

From the component corresponding to μ we see that

$$\sum_{j=1}^J \lambda_j = 1. \quad (15)$$

From the other components we see that there exist appropriate subgradients $\xi_j(x^i, y^i)$ and a normal vector η such that the equation

$$0 = \sum_{j=1}^J \lambda_j \xi_j(x^i, y^i) + \eta \quad (16)$$

holds. Due to (13) for all j we can choose constants α_{jk} such that

$$\begin{aligned} \xi_j(x^i, y^i) &= \sum_{k=1}^{K^{ij}} \alpha_{jk} \xi_j^k(x^i, y^i), \\ \sum_{k=1}^{K^{ij}} \alpha_{jk} &= 1, \quad j = 1, \dots, J, \\ \alpha_{jk} &\geq 0, \quad j = 1, \dots, J, k = 1, \dots, K^{ij}. \end{aligned}$$

Multiplying all the inequalities in (14) with corresponding constants α_{jk} , inserting $y = y^i$ and summing over k we obtain

$$\begin{aligned} \sum_{k=1}^{K^j} \alpha_{jk} g_j(x^i, y^i) + \sum_{k=1}^{K^j} \alpha_{jk} \xi_j^k(x^i, y^i)^T \begin{pmatrix} x - x^i \\ y^i - y^i \end{pmatrix} &\leq 0, \quad j = 1, \dots, J \\ \Rightarrow g_j(x^i, y^i) + \xi_j(x^i, y^i)^T \begin{pmatrix} x - x^i \\ 0 \end{pmatrix} &\leq 0, \quad j = 1, \dots, J. \end{aligned}$$

Multiplying these equations with the KKT-multipliers λ_j and summing them up results in the equation,

$$\sum_{j=1}^J \lambda_j g_j(x^i, y^i) + \sum_{j=1}^J \lambda_j \xi_j(x^i, y^i)^T \begin{pmatrix} x - x^i \\ 0 \end{pmatrix} \leq 0.$$

From the KKT-condition (12) and Equation (15) we see that $\sum_j \lambda_j g_j(x^i, y^i) = \mu$. Also, since $\eta \in N_X(x^i)$ equation $\eta^T(x - x^i) \leq 0$ holds. Summing this equation and making substitution $\sum_j \lambda_j g_j(x^i, y^i) = \mu$ results in the equation,

$$\mu + \sum_j \lambda_j \xi_j(x^i, y^i)^T \begin{pmatrix} x - x^i \\ 0 \end{pmatrix} + \eta^T(x - x^i) \leq 0. \quad (17)$$

From Equation (16) we know that

$$\left(\sum_j \lambda_j \xi_j(x^i, y^i) + \eta \right)^T (x - x^i) = 0,$$

thus inequality (17) results in $\mu \leq 0$ implying there is a feasible solution for the problem (NLP_{y^i}) . This is impossible since the problem (F_{y^i}) was solved. Hence, the lemma is proved. $\blacksquare$

Similarly, corresponding Lemma 4.3 holds. The proof of this lemma is very similar to the previous one. Also, if a point is feasible and the objective function f is smaller than UBD^i at the point, then the point is feasible in problem (M^i) . This follows from the definition of a subgradient and convexity of the objective and constraint functions. Thus, the modified OA method converges. Also, if functions are continuously differentiable this procedure reduces to the OA method in [5].

5. An example problem

Next, the example problem (E) is solved with the modified OA and ECP methods. First, we continue the solution process of OA from point $(1, 3)$ obtained in previous section. Now, for constraint g_1 two CPs are generated, one with gradient $(1, 1)^T$ and the other with $(-1, 1)^T$, resulting in the constraints

$$\begin{aligned} x + y - \frac{7}{2} &\leq 0 \quad \text{and} \\ -x + y - \frac{3}{2} &\leq 0. \end{aligned}$$

The resulting MILP problem is

$$\begin{aligned} \min \quad & 2x - y \\ \text{s.t.} \quad & x + y - \frac{7}{2} \leq 0, \\ & -x + y - \frac{3}{2} \leq 0, \\ & y - 4x - 1 \leq 0, \\ & 0 \leq x \leq 2, y \in Y, \end{aligned}$$

and the solution is $(0, 1)^T$. The problem (NLP_{y^1}) has the same optimum and (M^1) is infeasible. Thus, the modified OA algorithm ends up with the solution $(0, 1)^T$ which is a global optimum.

For the ECP method the first MILP problem is

$$\begin{aligned} \min \quad & 2x - y \\ \text{s.t.} \quad & y - 4x - 1 \leq 0, \\ & 0 \leq x \leq 2, y \in Y. \end{aligned}$$

The solution point $(1, 5)^T$ is not feasible in (E) since $g(1, 5) = \frac{5}{2} \geq 0$. Both the linear functions of g have the same value and the subdifferential is the same as in (7). According to Algorithm 3.2 we may choose any subgradient from the subdifferential. We will see how the algorithm proceeds if we would choose subgradients $(1, 1)$, $(0, 1)$ or $(-1, 1)$, respectively. Results are summarized in Table 1.

As can be seen from Table 1, the ECP method converged to optimal point in this example with all chosen subgradients. However, different choices of subgradients resulted in different number of iterations. The global optimum was found fastest with the subgradient $(-1, 1)^T$.

6. Numerical examples

In this section, the ECP method is applied to four two-dimensional problems. The problems are modified from known convex nonsmooth test problems from the literature. Two-dimensional cases of the original problems are of the form

$$\min f(x_1, x_2), \tag{GNP}$$

Table 1. Quantities in the ECP algorithm with different choices of subgradient.

ξ^i	i	x^i	y^i	$g(x^i, y^i)$	$2x - y$
$(1, 1)$	0	1	5	$\frac{5}{2}$	-3
	1	$\frac{1}{2}$	3	1	$-\frac{3}{2}$
	2	0	1	$-\frac{1}{2}$	-1
$(0, 1)$	1	$\frac{1}{4}$	2	1	$-\frac{7}{4}$
	2	0	1	$-\frac{1}{2}$	-1
$(-1, 1)$	1	0	1	$-\frac{1}{2}$	-1

Table 2. Numerical results.

Problem	Minimum	μ^*	$g(x, y, \mu^*)$	Iterations
<i>MAXQ</i>	0	-4.07×10^{-5}	2.65×10^{-4}	14
<i>MXHILB</i>	0	0	0	4
<i>Chained LQ</i>	-1.25	-1.250	2.52×10^{-4}	15
<i>Chained CB3 I</i>	2	2.000	6.08×10^{-6}	24

where $x_1, x_2 \in \mathbb{R}$. The objective functions that are considered are

$$\max\{x_1^2, x_2^2\} \quad (MAXQ)$$

$$\max\left\{\left|x_1 + \frac{1}{2}x_2\right|, \left|\frac{1}{2}x_1 + \frac{1}{3}x_2\right|\right\}, \quad (MXHILB)$$

$$\max\{-x_1 - x_2, -x_1 - x_2 + (x_1^2 + x_2^2 - 1)\}, \quad (Chained LQ)$$

$$\max\{x_1^4 + x_2^2, (2 - x_1)^2 + (2 - x_2)^2, 2e^{-x_1+x_2}\}. \quad (Chained CB3 I)$$

The objective functions are taken from [3,11,18,23].

In the modifications the second variable is restricted to discrete values, that is, $x_2 = y$ and $y \in \mathbb{Z}$. From now on we denote $x = x_1$. Also, the objective function is transformed to a constraint function by introducing constraint function $g(x, y, \mu) = : f(x, y) - \mu$ and constraint $g \leq 0$, where $\mu \in \mathbb{R}$ is the new objective function. Furthermore, appropriate box constraints for the variables are introduced in order to make the feasible set compact without cutting off the global minimum. By doing all the modifications to the problem (*GNP*) we end up with the problem

$$\begin{aligned} \min \mu \\ \text{subject to } f(x, y) - \mu &\leq 0, \\ -5 &\leq x \leq 5, \\ -5 &\leq y \leq 5, \\ -5 &\leq \mu \leq 5. \end{aligned}$$

Since the functions are nonsmooth and convex the considered problems are nonsmooth convex MINLP problems. Table 2 shows the results for different functions. The minimum obtained by the algorithm is denoted μ^* .

From Table 2, we see that the correct objective function value was found in every problem with low number of iterations. In these test problems, there were as many evaluations of nonlinear function as there were iterations. The constant value $\epsilon_g = 10^{-3}$ was used as a feasibility tolerance. The gradients were calculated from the analytical formula. Subgradient was chosen to be the gradient of an active constraint function, which is valid since this gradient belongs to the subdifferential according to Lemma 2.4.

Note that the global optimum in problem (*Chained LQ*) is not the same as in [23]. This is due to discretization of variable x_2 . The global optimum of the original problem is attained in $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ which is not feasible when x_2 is restricted to $\mathbb{Z}$.

7. Conclusions

The ECP and OA methods with nonsmooth MINLP problems were studied. The ECP method could solve the problem if subgradients are used in linearizations when the gradient does not exist. An example showed that different choices of subgradients may affect how fast the global optimum is found. The OA method presented in [5] could not solve the problem generally, but with slight modifications to the algorithm a special class of nonsmooth functions could be dealt with. The ECP method was applied to some small nonsmooth MINLP problems which the algorithm solved efficiently and correctly.

Acknowledgements

This research was supported by the grant no. 127992 of the Academy of Finland.

References

- [1] A. Beck and M. Teboulle, *Mirror descent and nonlinear projected subgradient methods for convex optimization*, Oper. Res. Lett. 31 (2003), pp. 167–175.
- [2] A. Ben-Tal and A. Nemirovski, *Non-Euclidean restricted memory level method for large-scale convex optimization*, Math. Program. 102 (2005), pp. 407–456.
- [3] C. Charalambous and A.R. Conn, *An efficient method to solve the minimax problem directly*, SIAM J. Numer. Anal. 15 (1978), pp. 162–187.
- [4] M.A. Duran and I.E. Grossmann, *An outer-approximation algorithm for a class of mixed-integer nonlinear programs*, Math. Program. 36 (1986), pp. 307–339.
- [5] R. Fletcher and S. Leyffer, *Solving mixed integer nonlinear programs by outer approximation*, Math. program. 66 (1994), pp. 327–349.
- [6] R. Fletcher and S. Leyffer, *Numerical experience with lower bounds for MIQP branch-and-bound*, Numerical Analysis Rep. NA/151, University of Dundee, 1995.
- [7] M. Gaudioso and M.F. Monaco, *Variants to the cutting plane approach for convex nondifferentiable optimization*, Optimization 25 (1992), pp. 65–75.
- [8] I. Grossmann, *Review of nonlinear mixed-integer and disjunctive programming techniques*, Optim. Eng. 3 (2002), pp. 227–252.
- [9] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms I*, Springer-Verlag, Berlin, 1993.
- [10] J.E. Kelley, *The cutting-plane method for solving convex programs*, J. SIAM 8 (1960), pp. 703–712.
- [11] K.C. Kiwiel, *An ellipsoid trust region bundle method for nonsmooth convex optimization*, SIAM J. Control Optim. 27 (1989), pp. 737–757.
- [12] K.C. Kiwiel, *Proximity control in bundle methods for convex nondifferentiable minimization*, Math. Program. 46 (1990), pp. 105–122.
- [13] L. Lukšan and J. Vlček, *Globally convergent variable metric method for convex nonsmooth unconstrained minimization*, J. Optim. Theory Appl. 102 (1999), pp. 593–613.
- [14] M.M. Mäkelä, *Survey of bundle methods for nonsmooth optimization*, Optim. Methods Softw. 17 (2002), pp. 1–29.
- [15] M.M. Mäkelä and P. Neittaanmäki, *Nonsmooth Optimization: Analysis and Algorithm with Applications to Optimal Control*, World Scientific, Singapore, 1992.
- [16] I. Quesada and I.E. Grossmann, *An LP/NLP based branch-and-bound algorithm for convex MINLP optimization problems*, Comput. Chem. Eng. 16 (1992), pp. 937–947.

- [17] C. Sagastizábal and M. Solodov, *An infeasible bundle method for nonsmooth convex constrained optimization without a penalty function or a filter*, SIAM J. Optim. 16 (2005), pp. 146–169.
- [18] H. Schramm, *Eine Kombination von Bundle- und Trust-Region-Verfahren zur Lösung nichtdifferenzierbarer Optimierungsprobleme*, Ph.D. thesis, Bayreuther Mathematische Schriften, No. 30, Universität Bayreuth.
- [19] H. Schramm and J. Zowe, *A version of the bundle idea for minimizing a nonsmooth function: Conceptual idea, convergence analysis, numerical results*, SIAM J. Optim. 2 (1992), pp. 121–152.
- [20] N.Z. Shor, *Minimization Methods for Non-Differentiable Functions*, Springer-Verlag, Berlin, 1985.
- [21] C. Still and T. Westerlund, *Extended cutting plane algorithm*, in *Encyclopedia of Optimization*, C.A. Floudas and P.M. Pardalos, eds, Kluwer Academic, Dordrecht, 2001, pp. 53–61.
- [22] T. Westerlund and F. Pettersson, *An extended cutting plane method for solving convex MINLP problems*, Comput. Chem. Eng. 19 (1995), pp. 131–136.
- [23] R.S. Womersley, *Numerical methods for structured problems in nonsmooth optimization*, Ph.D. thesis, Department of Mathematics, University of Dundee, 1981.
- [24] X. Yuan, L. Pibouleau, and S. Domenech, *Experiments in process synthesis via mixed-integer programming*, Chem. Eng. Process. 25 (1989), pp. 99–116.

Appendix: the proof of convergence of the other version of the ECP method

A version of ECP method in which every violated constraint will be linearized at each iteration is mentioned in [8]. This version will here be referred to as the other version of the ECP method. Next we shall prove that also the other version of the ECP method converges in the MINLP problem ($\hat{P}$). The proof is quite similar to the proof presented in Section 3 for the first version. Again, we assume $\epsilon_g = 0$. Denote

$$J^i = \{j \mid g_j(z^i) > 0\},$$

where z^i is the solution point at iteration i . Thus, J^i is the set of indices of the constraint functions that are violated at iteration i . Also, denote the linearization of the j th constraint function at iteration i

$$l_j^i(z) := g_j(z^i) + \xi_j(z^i)^T(z - z^i),$$

where $\xi_j(z^i) \in \partial g_j(z^i)$.

First we show that the CPs will not cut off any of the feasible points. Consider the i th iteration. Let $j \in J^i$ be arbitrary and $z_0 \in N \cap L$ be a feasible point. Then, by the convexity of the constraint functions, we have

$$l_j^i(z_0) = g_j(z^i) + \xi_j(z^i)^T(z_0 - z^i) \leq g_j(z_0) \leq 0$$

and, thus, the constraint $l_j^i(z) \leq 0$ does not cut off the point z_0 .

Next we show that if the solution point z^i at iteration i is not feasible then z^i will be cut off by the CPs introduced at the i th iteration. Let $j \in J^i$ be arbitrary. Then

$$l_j^i(z^i) = g_j(z^i) + \xi_j(z^i)^T(z^i - z^i) = g_j(z^i) > 0$$

and, thus, z^i cannot be a solution at the sequent iterations.

By the arguments used in the convergence proof for ECP method in Section 3 we either get a global optimum after a finite number of iterations or the sequence of solutions (z^i) contains an accumulation point $\hat{z}$. The following theorem proves that the accumulation point will be a feasible point. This implies that $\hat{z}$ is a global optimum as deduced in Section 3.

THEOREM A.1 *An accumulation point generated by the other version of the ECP method is a feasible point.*

Proof Assume that (z^i) has only one accumulation point $\hat{z}$. If there would be many accumulation points we could consider, instead of (z^i) , a subsequence of (z^i) that has only one accumulation point.

Denote

$$\tilde{J} = \{j \mid g_j \text{ is linearized infinitely many times.}\}.$$

By the definition of $\tilde{J}$ there exists $I \in \mathbb{N}$ such that for all $r > I$ and $z^r \in (z^i)$ we have $g_j(z^r) \leq 0$ if $j \in \{1, \dots, J\} \setminus \tilde{J}$. That is, after some number of iterations the only constraints that are violated are those that are violated infinitely many times. Let K_j be the Lipschitz constant of the constraint function g_j in the ball $B(\hat{z}, \delta_j)$, where $\delta_j > 0$. Denote $K = \max_{j \in \tilde{J}} \{K_j\}$ and $\delta = \min_{j \in \tilde{J}} \{\delta_j\}$. Then, for any $j \in \tilde{J}$ the inequality (2) holds for g_j with Lipschitz constant K in the ball $B(\hat{z}, \delta)$. Let $\epsilon > 0$ and $j \in \tilde{J}$ be arbitrary. Denote $\hat{\epsilon} = \min\{\frac{\epsilon}{3K}, \delta\}$. Let i and k be indices such that $k > i$, $i > I$, $z^i, z^k \in B(\hat{z}, \hat{\epsilon})$ and $g_j(z^i) > 0$. Since constraint $g_j(z^i) \leq 0$ is violated at iteration i it will be linearized then. By the deductions used in the Lemma 3.3 we have

$$\begin{aligned} g_j(\hat{z}) &\leq |g_j(\hat{z}) - l_j^i(z^k)| \\ &= |g_j(\hat{z}) - g_j(z^i) - \xi_j(z^i)^T(z^k - z^i)| \\ &\leq |g_j(\hat{z}) - g_j(z^i)| + |\xi_j(z^i)^T(z^k - z^i)| \\ &\leq K\|\hat{z} - z^i\| + K\|z^k - z^i\| \\ &< K\frac{\epsilon}{3K} + K\frac{2\epsilon}{3K} = \epsilon. \end{aligned}$$

Since ϵ can be arbitrarily small, $g_j(\hat{z}) \leq 0$ for all $j \in \tilde{J}$. ■

A reformulation framework for global optimization

Andreas Lundell · Anders Skjäl · Tapio Westerlund

Received: 30 September 2011 / Accepted: 11 February 2012
© Springer Science+Business Media, LLC. 2012

Abstract In this paper, we present a global optimization method for solving nonconvex mixed integer nonlinear programming (MINLP) problems. A convex overestimation of the feasible region is obtained by replacing the nonconvex constraint functions with convex underestimators. For signomial functions single-variable power and exponential transformations are used to obtain the convex underestimators. For more general nonconvex functions two versions of the so-called α BB-underestimator, valid for twice-differentiable functions, are integrated in the actual reformulation framework. However, in contrast to what is done in branch-and-bound type algorithms, no direct branching is performed in the actual algorithm. Instead a piecewise convex reformulation is used to convexify the entire problem in an extended variable-space, and the reformulated problem is then solved by a convex MINLP solver. As the piecewise linear approximations are made finer, the solution to the convexified and overestimated problem will form a converging sequence towards a global optimal solution. The result is an easily-implementable algorithm for solving a very general class of optimization problems.

Keywords Global optimization · Reformulation technique · Convex underestimators · Mixed integer nonlinear programming · Twice-differentiable functions · Signomial functions · Piecewise linear functions · α BB-underestimator · SGO-algorithm

1 Introduction

The scope of this article is to show how α BB-type underestimators [11,32] can easily be included in the signomial global optimization (SGO) algorithm [22,25,29], utilizing similar convex reformulation techniques as already present in the SGO-algorithm. Thus, the class of problems that can be solved using the algorithm is extended to include all twice-differentiable functions.

A. Lundell (✉) · A. Skjäl · T. Westerlund
Center of Excellence in Optimization and Systems Engineering, Åbo Akademi University,
Biskopsgatan 8, Turku, 20500, Finland
e-mail: andreas.lundell@abo.fi

In Sect. 2, the original SGO-algorithm and the transformation techniques for nonconvex signomial terms are introduced. The original α BB-underestimator is described in Sect. 3 and the implementation of the α BB-underestimator in the reformulation framework together with an illustrative example is given in Sect. 4. In Sect. 5, a new method for excluding infeasible regions of the domain is presented. The extension of the SGO-algorithm for handling twice-differentiable functions is discussed in Sect. 6. To exemplify the iterative nature of the algorithm, it is applied to a mixed integer nonlinear programming (MINLP) problem containing nonconvex trigonometric and signomial functions in Sect. 7. Finally, some conclusions and ideas for future research are provided in Sect. 8.

The problem type for which global optimal solutions can be obtained with the presented method is

$$\begin{aligned} & \text{minimize} && f(\mathbf{x}), \\ & \text{subject to} && \mathbf{Ax} = \mathbf{a}, \quad \mathbf{Bx} \leq \mathbf{b} \\ & && \underbrace{\mathbf{g}(\mathbf{x}) + \boldsymbol{\sigma}(\mathbf{x}) + \mathbf{q}(\mathbf{x})}_{\mathbf{h}(\mathbf{x})} \leq 0, \\ & && \mathbf{x} = [x_1, x_2, \dots, x_N]^T, \quad \underline{\mathbf{x}} \leq \mathbf{x} \leq \bar{\mathbf{x}}, \end{aligned} \quad (1)$$

where f is a convex objective function, $\mathbf{A}$ and $\mathbf{B}$ are matrices of the size $N \times N$ and $\mathbf{a}$ and $\mathbf{b}$ are vectors of the size $N \times 1$. Furthermore, the nonlinear inequality constraints $\mathbf{h}(\mathbf{x}) \leq 0$ are composed of twice-differentiable nonconvex functions $\mathbf{g}$, signomial functions $\boldsymbol{\sigma}$ and convex functions $\mathbf{q}$. The variable-vector $\mathbf{x}$ may contain both integer- and real-valued variables, and the variables are bounded by appropriate lower and upper bounds $\underline{\mathbf{x}}$ and $\bar{\mathbf{x}}$. The objective function f is also allowed to be nonlinear and nonconvex by replacing the objective function with a new variable μ and including an objective function constraint of the type $f(\mathbf{x}) - \mu \leq 0$. Equality constraints can be handled by relaxing them to corresponding positive and negative inequality constraints.

A signomial function is defined as a sum of signomial terms, where each term consists of a product of power functions. Thus, a signomial function of N variables and J signomial terms can be expressed mathematically as

$$\sigma(\mathbf{x}) = \sum_{j=1}^J c_j \prod_{i=1}^N x_i^{pji}, \quad (2)$$

where the coefficients c_j and the powers pji are real-valued. If a certain variable x_i does not exist in the j -th term, then $pji = 0$. The signomial class of functions contain several often occurring functions in mathematical programming problems such as polynomials, including bi- and trilinear terms.

In the SGO-algorithm, power and exponential transformations are used to convexify and underestimate the signomial functions $\boldsymbol{\sigma}$. These have been explained thoroughly before [22, 24] and are also briefly introduced in Sect. 2.2. In this paper, the algorithm is extended to also allow for the solution of MINLP problems containing general twice-differentiable functions in the constraints, i.e., problem (1). Since all signomials are twice-differentiable (except for possible singularities at zero), they can alternatively be regarded to be part of the functions $\mathbf{q}$ if convex or $\mathbf{g}$ if nonconvex. Therefore nonconvex signomial functions can be handled using either the convexification techniques in the original SGO-algorithm or using those presented in this paper for general twice-differentiable functions. However, as shown in the example in Sect. 7, using the signomial transformation strategies can be beneficial in some cases, for

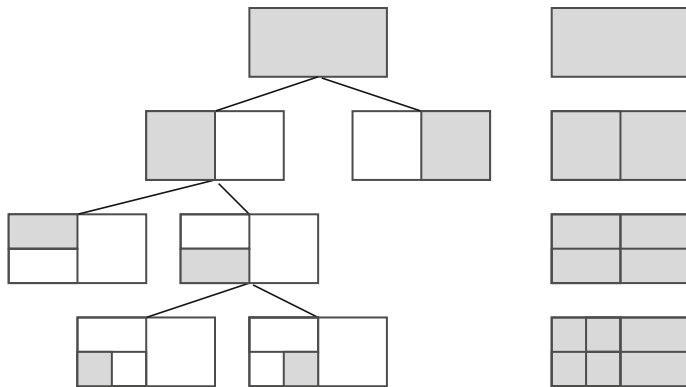


Fig. 1 *Left* in a branching algorithm, the feasible region is partitioned and different subproblems are considered and solved in each node. *Right* in a reformulation algorithm, one convex overestimated MINLP problem is solved (where the whole domain is considered) in each iteration and no direct branching strategy is needed

example, the combinatorial complexity of the reformulated problem may be smaller, or the tightness of the convex underestimators may be increased.

In Fig. 1, the difference between using a branching and a reformulation framework is illustrated. In a branching scheme, only part of the feasible region is investigated in each node, whereas in our reformulation framework, the whole feasible region is considered in each iteration.

The reformulation technique also allows us to obtain a convex relaxation of desired accuracy of the nonconvex problem beforehand, giving a convex reformulation of any nonconvex problem of type (1). The reformulated problem can then be solved directly using a convex MINLP solver. Unfortunately, improved tightness results in an increased size of the convex reformulation. Increasing the tightness in a sequential manner will on the other hand typically result in a final convex relaxation of smaller size. In the proposed algorithm, the iterative procedure can also be started from a sufficiently tight convex relaxation, which is not possible in a branching strategy.

Some other methods for solving nonconvex MINLP problems to global optimality have been given in [14, 15, 33, 34].

2 The signomial global optimization algorithm

The original SGO-algorithm uses single-variable exponential or power transformations applied to the nonconvex signomial terms to convexify them. (The convexity of signomial terms can be ascertained using the rules put forward by Maranas and Floudas [30].) As the relationship between the transformation and original variables are approximated using PLFs, an overestimation of the nonconvex feasible region is obtained. The PLFs are updated iteratively, and the solution to the overestimated problem will converge to the global optimal solution of the original nonconvex problem.

2.1 Earlier work

The SGO-algorithm has been under development for over 10 years. However, many of the convexification techniques used in the algorithm have been used much longer. For instance,

the exponential transformation has been used in geometric programming (GP) since the 1960's. GP is a special case of signomial programming (SP) where only positive signomial terms are allowed. An overview of the early years of GP was given by Peterson [35]. Additionally, algorithms for solving continuous SP problems have been given earlier, for example, by Dembo [10] and Rijckaert and Martens [38]. A newer result is that of Maranas and Floudas, who presented a branch-and-bound algorithm for solving SP problems in [31].

Pörn et al. [37] took the first steps towards the SGO-algorithm when they combined the exponential transformation and PLF-approximations for reformulating nonconvex signomial problems. This work was extended by Pörn et al. [36] and Björk [8], by also including certain power transformations for convexification of nonconvex signomial terms. Westerlund and Westerlund [45] described the predecessor to the SGO-algorithm: the generalized geometric programming extended cutting plane (GGPECP) algorithm. The algorithm was tightly integrated with the α ECP solver used for solving the reformulated convex MINLP subproblems. The GGPECP-algorithm was described in more detail by Westerlund [43]. Finally, in [22] and [23], Lundell et al. combined the GGPECP-algorithm with an optimization framework for the transformations used to convexify the signomial terms into the SGO-algorithm. Additionally, the SGO-algorithm included the option to use other MINLP solvers except α ECP. The SGO-algorithm was further extended by Lundell and Westerlund [24–28]. The most complete and up-to-date description of the algorithm was given by Lundell and Westerlund [29].

Similar reformulation techniques as those in the SGO-algorithm have also been presented by Li et al. [17], Lin and Tsai [19] as well as Tsai and Lin [41, 42].

2.2 The transformation technique

The transformation technique for nonconvex signomial functions has the following effect on a generalized signomial constraint in problems of type (1)

$$q_m(\mathbf{x}) + \sigma_m(\mathbf{x}) \leq 0 \xrightarrow{(i)} q_m(\mathbf{x}) + \sigma_m^C(\mathbf{x}, \mathbf{X}) \leq 0 \xrightarrow{(ii)} q_m(\mathbf{x}) + \sigma_m^C(\mathbf{x}, \hat{\mathbf{X}}) \leq 0.$$

In step (i), the nonconvex signomial terms are transformed to convex form using single-variable transformations T_{ji} of the type $x_i = T_{ji}(X_{ji})$, where X_{ji} is a new transformation variable. However, since the relationship between the original and transformation variables must be included in the problem to retain the information, the problem is not convexified in this step even though the signomial constraints are. Instead, the nonconvexities have only been moved to the nonlinear equality constraints $x_i = T_{ji}(X_{ji})$. Therefore, in step (ii), the equality constraints are approximated using PLFs. Since the transformations T are chosen such that they will always result in an underestimation of the original nonconvex term when the inverse transformations are at the same time approximated by PLFs, cf. [21, 22, 43], the feasible region of the constraint will be larger than that of the original constraint as well as convex, i.e., a convex relaxation has been obtained.

Two different types of single-variables transformations, power and exponential, are used in the algorithm. Through convexity analysis of the signomial terms [21, 30, 36] in combination with study of the underestimation properties of the transformations, different types of combined transformation schemes for each term can be deduced. For positive signomial terms, four transformation schemes are applicable: the positive power transformation (PPT), the negative power transformation (NPT), the exponential transformation (ET) and the mixed power and exponential transformation (MPET); for negative terms only the power transformation (PT) scheme is applicable.

The NPT transformation scheme is based on the fact that a positive signomial term with only negative powers is convex [21,30].

Definition 1 (*Negative power transformation—NPT*) The NPT convex underestimator for positive signomial terms is obtained by applying the transformation

$$x_i = X_i^{Q_i}, \quad Q_i < 0,$$

to all variables x_i with positive powers ($p_i > 0$) in the term, as long as the inverse transformation $X_i = x_i^{1/Q_i}$ is approximated by a PLF $\widehat{X}_i$.

The PPT transformation scheme is based on the fact that a positive signomial term, where all powers are negative except for one and where the sum of the powers is greater than or equal to one, is convex [21,30].

Definition 2 (*Positive power transformation—PPT*) The PPT convex underestimator for positive signomial terms is obtained by applying the transformation

$$x_i = X_i^{Q_i}$$

to all variables with positive powers in the term, where the transformation powers $Q_i < 0$ for all indices i , except for one ($i = k$), for which $Q_k \geq 1$. Furthermore, the condition

$$\sum_{i:p_i>0} p_i Q_i + \sum_{i:p_i<0} p_i \geq 1$$

must be fulfilled and the inverse transformation $X_i = x_i^{1/Q_i}$ approximated by a PLF $\widehat{X}_i$.

The ET transformation scheme is based on the exponential transformation, which is used in GP to reformulate nonconvex posynomial problems into convex form.

Definition 3 (*Exponential transformation—ET*) The ET convex underestimator for positive signomial terms is obtained by applying the transformation

$$x_i = e^{X_i}$$

to the individual variables with positive powers in the term, as long as the inverse transformation $X_i = \ln x_i$ is approximated by a PLF $\widehat{X}_i$.

Finally, a combination of the ET and NPT transformation schemes can also be used to convexify a positive term. In fact, the ET and NPT schemes can be regarded as special cases of this transformation type, which is called the mixed power and exponential transformation (MPET). The MPET scheme is valid since a term of the form $e^{p_1 x_1} \dots e^{p_m x_m} x_{m+1}^{p_{m+1}} \dots x_N^{p_N}$, where $p_1, \dots, p_m > 0$ and $p_{m+1}, \dots, p_N < 0$, is convex [21,36] and the transformed terms are underestimated when the transformations are replaced with linear approximations.

Definition 4 (*Mixed power and exponential transformation—MPET*) The MPET convex underestimator for positive signomial terms is obtained by applying either of the transformations

$$x_i = X_i^{Q_i}, \quad Q_i < 0, \quad \text{or} \quad x_i = e^{X_i}$$

to all variables x_i with positive powers in the term, as long as the inverse transformations $X_i = x_i^{1/Q_i}$ and $X_i = \ln x_i$ respectively are approximated by PLFs $\widehat{X}_i$.

For negative terms, the PT transformation scheme is based on the fact that a negative signomial term is convex if all powers are positive and the sum of the powers is less than or equal to one [21,30].

Definition 5 (*Power transformation for negative terms—PT*) The PT convex underestimator for a negative signomial term is obtained by applying the transformation

$$x_i = X_i^{Q_i},$$

where the transformation powers used are $0 < Q_i \leq 1$ for all variables with positive powers and $Q_i < 0$ for all variables with negative powers, to the individual variables in the term. Furthermore, the condition

$$0 < \sum_i p_i Q_i \leq 1,$$

must be fulfilled and the inverse transformations $X_i = x_i^{1/Q_i}$ approximated by a PLF $\widehat{X}_i$.

In the case of the power transformations, the powers in the single-variable transformations are themselves also parameters that have a large impact on the numerical properties of the transformed constraint as shown by Lundell and Westerlund [25]. Therefore, a technique for determining an optimized set of transformations for convexifying signomial terms in a MINLP problem was introduced in [23]. By formulating and solving an auxiliary mixed integer linear programming (MILP) problem, the set of transformations can be optimized with respect to, e.g., the total number of transformations or number of original variables transformed. Additional enhancements to the MILP method, e.g., including the exponential transformation, have been the topic of several additional papers by Lundell et al. [22,24,28].

2.3 Termination criterion

The global optimal solution to a nonconvex signomial problem is found when the nonconvex constraints are fulfilled to an ϵ -tolerance in the optimal solution point of an overestimated convex MINLP problem. That is, for the nonconvex constraints $m = 1, \dots, M$ in the problem, the following is true

$$\max_m (\sigma_m(\mathbf{x}^*) + q_m(\mathbf{x}^*)) \leq \epsilon, \quad (3)$$

where $\mathbf{x}^*$ is the solution point to the convexified MINLP problem in the current iteration. If the solution point is not optimal, i.e., condition (3) is not fulfilled, new breakpoints are added to the PLFs, and a new tighter convex overestimated MINLP problem is solved. Since a nonconvex objective function is rewritten as an additional constraint, the objective function does not need to be handled separately, although a different ϵ -tolerance could be used.

2.4 Strategies for iteratively adding breakpoints to the PLFs

Several different types of strategies for adding breakpoints to the PLFs can be constructed and they have been previously described by Lundell [21] and Lundell and Westerlund [29]. The simplest strategy is to add the midpoint of the current interval where the solution is found as the new breakpoints. Westerlund showed in [43] that this strategy guarantees finding a global optimal solution within an ϵ -tolerance.

Another possibility for adding the breakpoints is to add the solution value for the variable in question to the corresponding PLFs. A third option is to add the point where the approximation error, caused by replacing the equality relation between the original and transformation

variable with a PLF-approximation, is the largest. The location of this point depends on the type of transformation. According to the results given by Lundell and Westerlund [29], the most deviating point in the interval $[\underline{x}, \bar{x}]$ is for a power transformation (with transformation power Q)

$$\tilde{x} = \left(Q \cdot \frac{\bar{x}^{1/Q} - \underline{x}^{1/Q}}{\bar{x} - \underline{x}} \right)^{\frac{Q}{1-Q}}$$

and for an exponential transformation

$$\tilde{x} = \frac{\bar{x} - \underline{x}}{\ln(\bar{x}/\underline{x})}.$$

A combination of any of these breakpoint strategies can also be used. Naturally, more than one breakpoint can be added per transformation variable in each iteration, but this leads to the combinatorial complexity of the reformulated problem growing faster, which may result in longer solution times for the subproblems in each iteration.

3 The α BB-underestimator

In the original α BB-method, a convex underestimator is obtained by adding a quadratic function to the original nonconvex function; the quadratic function is chosen in such a way that it “overpowers” any nonconvexities of the original function. The original α BB-underestimator gives an underestimation of the nonconvex twice-differentiable function g in the domain $[\underline{x}_1, \bar{x}_1] \times [\underline{x}_2, \bar{x}_2] \times \cdots \times [\underline{x}_N, \bar{x}_N]$ according to:

$$g(\mathbf{x}) + \sum_{i=1}^N \underbrace{\alpha_i (x_i - \underline{x}_i)(x_i - \bar{x}_i)}_{=A_i} \leq g(\mathbf{x}). \quad (4)$$

If the function in the left-hand-side (LHS) of expression (4) is to be convex, the Hessian matrix of the function must be positive semidefinite in the domain. This can be guaranteed as long as the positive parameters α_i are selected large enough, cf. [11]. Also, the underestimation property is clear, since all the added functions A_i are always less than or equal to zero in the domain $x \in [\underline{x}_1, \bar{x}_1] \times [\underline{x}_2, \bar{x}_2] \times \cdots \times [\underline{x}_N, \bar{x}_N]$.

An α BB-like convexification was introduced by Maranas and Floudas [30] and Liu and Floudas [20], where it was used for expressing sufficiently smooth functions as a difference of two convex functions. The α BB methodology was used in a general framework for global optimization in Androulakis et al. [7]. A thorough account of theory and implementation was given in Adjiman et al. [2,3]. Additionally, the textbook by Floudas [11] contains further theory and describes how α BB has been used in applications.

Some variants of α BB have been developed for specialized problems or as alternatives to the original method. Two algorithms were suggested by Adjiman et al. [1,4] for MINLP problems, SMIN- α BB for problems where the participation of the binary variables is restricted and GMIN- α BB for more general problems. An alternative convexification method, γ BB, using another form of underestimator, but with α BB-like properties, was described by Akrotirianakis and Floudas [5,6]. Floudas also proved in [12,13] that α BB and γ BB give the only possible functional forms if both shift-invariance and sign-invariance is required. Meyer and Floudas [32] described spline α BB, a method where α -values calculated on “subsubdomains”

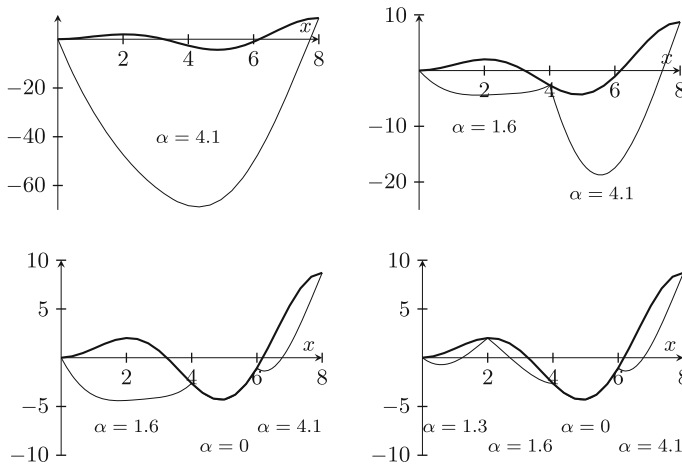


Fig. 2 The optimal α BB-underestimator for the function in Example 1, i.e., $h(x) = x \sin x + x/10$, illustrated as the number of partitions increase. The α -values used in the different partitions of the domain are indicated in the plots

are used to construct splines. These splines are often tighter than the original α BB-underestimator on the subdomain.

To obtain a convex underestimator that is as tight as possible, ideally the smallest possible values of the α_i 's should be chosen. For a function of only one variable, the smallest α in the interval $[\underline{x}, \bar{x}]$ can be found by taking the second derivative of $g(x) + \alpha(x - \underline{x})(x - \bar{x})$, i.e.,

$$\frac{d^2}{dx^2} (g(x) + \alpha(x - \underline{x})(x - \bar{x})) = g''(x) + 2\alpha, \quad (5)$$

and then finding the minimum value for $\alpha \geq 0$ such that

$$\alpha \geq -\frac{1}{2}g''(x) \quad \forall x \in [\underline{x}, \bar{x}] \quad (6)$$

is fulfilled. For functions of more than one variable, however, other methods are generally needed and several techniques for determining the parameters α are described more thoroughly by Adjiman et al. [2]. In the examples in this paper the scaled Gerschgorin method is used, if not stated otherwise.

Example 1 α BB-underestimators for the function $h(x) = x \sin x + x/10$, $0 \leq x \leq 8$, is illustrated in Fig. 2, as the variable-space is partitioned. The partitioning can be introduced in the form of a branch-and-bound type algorithm, where subproblems corresponding to the individual intervals are solved. Optimal α -values are determined by finding the minimal values such that expression (6) is fulfilled, so the underestimators shown in the figure are actually the tightest underestimators of this type in each interval. For example if the variable space is partitioned into the intervals $[0, 4]$ and $[4, 8]$ the optimal α BB-underestimator will be

$$\begin{cases} h(x) + 1.6x(x - 4) & \text{if } 0 \leq x < 4 \\ h(x) + 4.10293(x - 4)(x - 8) & \text{if } 4 \leq x \leq 8. \end{cases} \quad (7)$$

If this type of underestimator is used in a branch-and-bound type framework, different values for the parameters α can be used in different nodes, corresponding to different variable

intervals, of the branching tree. The underestimators, as illustrated in Fig. 2, are of course not convex in the entire domain, but rather piecewise convex when more than one interval is used. Therefore, it is not possible to directly use the underestimator to replace the nonconvex function in our reformulation framework.

To implement the α BB-underestimator in a global optimization algorithm, normally some type of branch-and-bound method is used. The variable-space is partitioned, and an over-estimated convex subproblem is solved in each node of the branching tree using a convex MINLP solver. What we want to accomplish is instead a reformulation of the problem which is convex in an extended variable-space consisting of the transformation variables and variables needed for the PLFs. The feasible region of the reformulated problem contains all solutions to the original nonconvex problem, and is then iteratively reduced (in the domain of the original variables) until the global optimal solution is found.

Since signomial functions are twice-differentiable, it would of course be possible to use the α BB-strategy for all nonconvex functions. However, as was shown by Lundell [21] and Lundell and Westerlund [25], the exponential and power transformations may in many cases give a tighter convex underestimator than the α BB-underestimator for signomial functions. Therefore, it may be beneficial to also include the power and exponential transformations for signomials in a global optimization algorithm.

4 Implementation of the α BB-underestimator in a piecewise linear reformulation framework

Our technique for including the α BB-underestimator in the reformulation framework is described in this section. For the case of simplicity, first the univariate case is considered in Sect. 4.1 and a generalization to problems with N variables is given in Sect. 4.2. A brief introduction to the reformulation technique presented in this section was previously given by Skjäl et al. [40].

4.1 Formulation of the underestimator in the one-dimensional case

In a piecewise convex reformulation framework as that of the SGO-algorithm, additional binary and continuous variables (or continuous special ordered sets of type 2 (SOS2) variables) are used to partition the domain of the variable x (using consecutive breakpoints ω_k , $k = 1, \dots, K$). The goal is to add a function convex enough to overpower the nonconvexities of the original constraint in the whole domain, and then compensate the underestimation error by subtracting it, or an estimate of it, defined by an additional variable. Since the additional variable does not affect the convexity of the expression, it is still convex. Thus, when applying the original α BB-underestimator in the reformulated form, the convexified constraint becomes

$$g(x) + \alpha(x - \underline{x})(x - \bar{x}) - W \leq 0, \quad (8)$$

where α is large enough so that the sum $g(x) + \alpha(x - \underline{x})(x - \bar{x}) - W$ is convex. This can be simplified by noting that the inclusion of the variable bounds $\underline{x}$ and $\bar{x}$ is not necessary for the convexity of expression (8), so the reformulated constraint can in fact be rewritten as

$$g(x) + \alpha x^2 - W \leq 0. \quad (9)$$

For convexification, the LHS of (9) should be convex, and to obtain an underestimation, the compensation-factor W should satisfy

$$W \geq \alpha x^2. \quad (10)$$

From expression (9), we observe that the tightest underestimation is obtained when

$$W = \alpha x^2, \quad (11)$$

but including an equality constraint (11) will only move the nonconvexities from the original inequality constraint $g(x) \leq 0$ to the added equality constraint. Observe, however, that the tightest linear approximation overestimating W can be obtained by approximating (11) with a PLF $\widehat{W}$ with breakpoints on $(x, \alpha x^2)$. Due to the negative sign, the LHS of (9) will be underestimated in the whole domain when replacing W with $\widehat{W}$. Now, both the convexity requirement is met and the underestimation condition (10) is fulfilled.

Thus, by replacing W with $\widehat{W}$, we have

$$g(x) + \alpha x^2 - \widehat{W} \leq g(x) + \alpha x^2 - W = g(x), \quad (12)$$

where $g(x) + \alpha x^2 - \widehat{W}$ is a reformulated convex underestimator of the nonconvex function. This is an example of a lifting reformulation, since the reformulated problem is convex in an extended variable space [18].

Although the expressions for the original α BB-underestimator $g(x) + \alpha(x - \underline{x})(x - \bar{x})$ and the reformulated α BB-underestimator (12) are different, the next theorem shows that they both render the same underestimation in the subdomains where the same α -values are used.

Theorem 1 *The convex underestimator $g(x) + \alpha x^2 - \widehat{W}$, where $\widehat{W}$ is a PLF approximating $W = \alpha x^2$, gives the same value for the underestimator in all breakpoint intervals $[\underline{x}, \bar{x}]$ of the PLF as the original α BB-underestimator $g(x) + \alpha(x - \underline{x})(x - \bar{x})$ in the corresponding interval, under the condition that the same α -value is used in both formulations of the underestimator.*

Proof Considering an arbitrary interval $[\underline{x}, \bar{x}]$ of the PLF, the linear approximation is a straight line through the points $\alpha \underline{x}^2$ and $\alpha \bar{x}^2$, i.e.,

$$\widehat{W} = \alpha \underline{x}^2 + \frac{\alpha \bar{x}^2 - \alpha \underline{x}^2}{\bar{x} - \underline{x}} (x - \underline{x}).$$

Therefore, the underestimator can be expressed as

$$\begin{aligned} g(x) + \alpha x^2 - \widehat{W} &= g(x) + \alpha x^2 - \left(\alpha \underline{x}^2 + \frac{\alpha \bar{x}^2 - \alpha \underline{x}^2}{\bar{x} - \underline{x}} (x - \underline{x}) \right) \\ &= \dots = g(x) + \alpha(x - \underline{x})(x - \bar{x}). \end{aligned}$$

Thus, the value of the underestimator is the same in all breakpoint intervals of the PLFs where the same α -values are used. $\square$

A disadvantage when using the reformulated convex underestimator (9) is, however, that only one α -value can be used, while the original α BB-underestimator can utilize different α -values in different subdomains. In order to satisfy the convexification requirement over the whole domain, the maximal α -value must therefore be used in the reformulated underestimator (9).

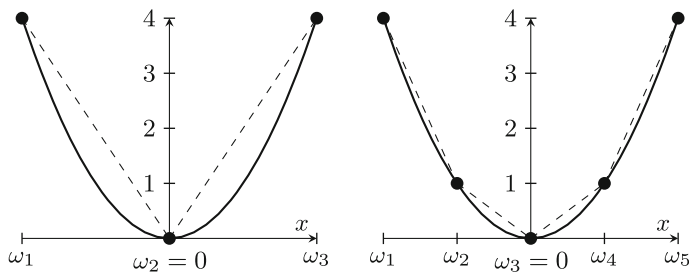


Fig. 3 The correction function $W(x)$ (solid) and its overestimations $\widehat{W}(x)$ (dashed) given by PLFs with three (left) and five (right) breakpoints respectively

As more breakpoints are used in the PLFs, the approximation $\widehat{W}$ becomes tighter with respect to W , as illustrated in Fig. 3. However, additional breakpoints also correspond to a transformed problem of larger combinatorial complexity, as more binary variables are needed for the PLFs. Therefore, in the global optimization algorithm described later, instead of adding many breakpoints to the initial approximation, additional breakpoints are added iteratively until the global solution is obtained.

4.2 General formulation of the underestimator

The formulation for the underestimator in the one-dimensional case can easily be extended to a general multidimensional one of N variables by replacing the nonconvex constraint

$$g(\mathbf{x}) \leq 0, \quad \mathbf{x} = (x_1, x_2, \dots, x_N), \quad (13)$$

with the following one:

$$g(\mathbf{x}) + \sum_{i=1}^N \alpha_i x_i^2 - \widehat{W} \leq 0. \quad (14)$$

If the parameters α are large enough, the expression (14) will be convex.

The feasible region (in the variables $\mathbf{x}$) of this constraint will be the same as that of the original nonconvex constraint if $\widehat{W} = W$, where

$$W = \sum_{i=1}^N \alpha_i x_i^2. \quad (15)$$

However, as stated previously, this constraint is nonconvex and is therefore replaced with a linearized variant. If the i -th variable has K_i consecutive breakpoints belonging to the ordered set Ω_i , where $\Omega_i = \{\omega_{i,1}, \dots, \omega_{i,K_i}\}$, $\omega_{i,1} = \underline{x}_i$ and $\omega_{i,K_i} = \bar{x}_i$, then the equality constraint can be approximated as

$$\widehat{W} = \sum_{i=1}^N \sum_{k=1}^{K_i-1} A_{i,k} b_{i,k} + (A_{i,k+1} - A_{i,k}) s_{i,k}, \quad (16)$$

where $A_{i,k_i} = \alpha_i \omega_{i,k_i}^2$, and

$$x_i = \sum_{k=1}^{K_i-1} \omega_{i,k} b_{i,k} + (\omega_{i,k+1} - \omega_{i,k}) s_{i,k}, \quad (17)$$

$$0 \leq s_{i,k} \leq b_{i,k}, \quad k = 1, \dots, K_i - 1, \text{ and} \quad (18)$$

$$\sum_{k=1}^{K_i-1} b_{i,k} = 1. \quad (19)$$

Here, the binaries $b_{i,k}$ indicate whether the variable x_i is in the interval $[\omega_{i,k}, \omega_{i,k+1}]$ ($b_{i,k} = 1$), or not ($b_{i,k} = 0$). This formulation for the PLFs is sharp, i.e., the linear programming relaxation of the PLF is not strictly dominated by the convex relaxation of the concave function, cf. [16]. Observe that the PLFs can also be expressed by using other formulations, e.g., employing SOS2-sets.

To summarize, all that is needed for a convex reformulation of any nonconvex MINLP problem of type (1) is to replace all nonconvex inequality constraints $g(\mathbf{x}) \leq 0$ with expression (14) and add the corresponding constraints (16–19) to the problem. If some of the other functions σ and q appear in the constraint function h in problem (1), then just replace g with the LHS of (14) and add the corresponding expressions (16–19) to the problem. In the latter case, the signomial functions σ must be handled using the techniques in Sect. 2.2 or included in g . Hence, no branching method need to be implemented but instead the problem can be solved using any MINLP solver available for the specified convex problem.

4.3 Comparisons to the α BB-underestimator

In a branch-and-bound framework, different α -values, and hence also different convex α BB-underestimators, can be used in each node. In this section, we show how the reformulated α BB-underestimators compare to the original one.

4.3.1 Comparison of the original α BB-underestimator to the reformulation

As was indicated in Sect. 4.1, we note that it is not possible to directly utilize all the properties of the original α BB-underestimator in the reformulated form. In this case, the α -value to be used in the reformulated version must be the largest of all the α -values in each breakpoint interval of each variable, i.e.,

$$\alpha_i = \max_k \alpha_{i,k} \quad k = 1, \dots, K_i. \quad (20)$$

If we only have one variable x_i in a twice-differentiable function g with nonzero α -values in a certain constraint in problem (1), the underestimator will be equal to the original α BB-underestimator in the interval(s) with the largest $\alpha_{i,k}$ value, and less tight in the rest of the intervals. If we have more than one variable with nonzero α -values in a constraint, then the reformulated α BB-underestimator will only be exact in such an interval where all of these variables happen to assume their maximal α -values.

In the intervals where the original α BB uses a smaller α -value α_k , the difference between the original underestimator and the underestimator $g(x) + \alpha x^2 - \bar{W}$ in the point x is

$$E_k = (\alpha - \alpha_k) (x - \underline{x}) (\bar{x} - x). \quad (21)$$

The maximal error between the two underestimators occurs in the midpoint of the interval, i.e., $x = (\underline{x} + \bar{x})/2$ and is thus

$$E_{\max} = \frac{1}{4} (\alpha - \alpha_k) (\bar{x} - \underline{x})^2. \quad (22)$$

The fact that the underestimator will be less tight in intervals where $\alpha_k < \alpha$ is a disadvantage when using the original α BB-underestimator in the reformulation framework. Furthermore, in this case, the underestimator is not equal to the original function g in intervals where the function is convex, i.e., $\alpha_k = 0$, as with the original α BB-underestimator. This will later on be illustrated in Example 2. If g is the objective function of a nonconvex MINLP problem, the minimum of the convex underestimator is the solution to the convexified problem and then the underestimation error has a direct impact on the quality of the solution. As a comment, if the nonconvex functions are in the constraints of the problem, then a large underestimation error might not have a very large direct impact on the feasible region corresponding to the constraint since we are only interested in whether the constraint is fulfilled or not, cf. the results given by Westerlund et al. [44].

However, since there is no requirement that the α -values decrease—only that they do not increase—as the domain is partitioned, the fact that we have to select the largest α -value does not affect the convergence of the method. Thus, the proof given by Floudas [11] for the convergence of the lower bound is valid here as well since the α -values never grow as new breakpoints are added. Note that we do not need to consider the upper bound in the reformulation framework, since we know that as soon as all (nonconvex) constraints are fulfilled we have obtained an ϵ -global optimal solution due to the fact that the whole domain is considered in each iteration.

4.3.2 Comparison of the spline α BB-underestimator to the reformulation

The drawbacks connected to using the largest α -value in the reformulated form when using the original α BB-underestimator can, however, be partly avoided by replacing the original α BB-underestimator with the so-called spline α BB-underestimator introduced by Meyer and Floudas [32].

When using the spline α BB-underestimator in the reformulation framework, it is possible to take advantage of different α -values in different parts of the domain to be convexified over. This gives a tighter underestimator (in fact as tight as the original α BB-underestimator if we have a completely separable function g), since we do not need to choose the largest α -value for each variable over its whole domain.

The spline α BB-underestimator is a smooth convex piecewise polynomial expression

$$S_i(x_i) = \begin{cases} \alpha_{i,1}x_i^2 + \beta_{i,1}x_i + \gamma_{i,1} & \text{if } x_i \in [\omega_{i,1}, \omega_{i,2}] \\ \alpha_{i,2}x_i^2 + \beta_{i,2}x_i + \gamma_{i,2} & \text{if } x_i \in [\omega_{i,2}, \omega_{i,3}] \\ \vdots & \vdots \\ \alpha_{i,K_i-1}x_i^2 + \beta_{i,K_i-1}x_i + \gamma_{i,K_i-1} & \text{if } x_i \in [\omega_{i,K_i-1}, \omega_{i,K_i}], \end{cases} \quad (23)$$

where $\alpha_{i,k}$, $\beta_{i,k}$ and $\gamma_{i,k}$ are parameters related to the k -th breakpoint interval of the variable x_i , i.e., $[\omega_{i,k}, \omega_{i,k+1}]$. The convexity requirement is for each interval guaranteed by the sufficiently large $\alpha_{i,k}$ -value while the parameters $\beta_{i,k}$ and $\gamma_{i,k}$ ensure the continuity and smoothness of the spline function. If the spline α BB-underestimator is used, then the PLF used to fulfill the underestimation condition is applied to $S(x)$, i.e., the constraint is rewritten as

$$g(\mathbf{x}) + \sum_{i=1}^N S_i(x_i) - \widehat{W} \leq 0, \quad (24)$$

where

$$\widehat{W} = \sum_{i=1}^N \hat{S}_i, \quad (25)$$

i.e., a sum of the PLFs of the spline functions S_i for variables with an α -value greater than zero in any breakpoint interval.

Naturally, the grid points of the PLF then coincide with those of the spline function, and the different α -values can directly be utilized in the reformulated form. It is, of course, possible to determine the parameters α , β and γ in smaller intervals than those specified by the breakpoints in the PLFs. Since S is a convex function no matter how the parameters are chosen, the linearization $\widehat{W}$, which is piecewise with respect to the breakpoints in the PLFs, will always overestimate the spline function.

As the next theorem shows, the reformulated spline α BB-underestimator will render exactly the same numerical underestimation (with respect to the original variables) as the original α BB-underestimator used in a branching framework under the condition that the nonconvex function g is completely separable.

Theorem 2 *The reformulated spline α BB-underestimator $g(\mathbf{x}) + \sum_{i=1}^N S_i(x_i) - \widehat{W}$, where S_i is given by Eq. (23) and $\widehat{W}$ is a sum of PLFs approximating the functions S_i , gives exactly the same underestimator (in the original variables) as the original α BB-underestimator (as used in a piecewise convex manner) in all breakpoint intervals if $g(\mathbf{x})$ is completely separable with respect to addition.*

Proof For simplicity, we consider the case with one variable only which has no affect on the validity of the proof since the nonconvex function g is separable (and can hence be grouped into functions $g_i(x_i)$ considered separately). Thus, we consider the arbitrary breakpoint interval $[\underline{x}, \bar{x}]$, and assume that the corresponding spline α BB-underestimator in this interval is $S(x) = \alpha x^2 + \beta x + \gamma$. Thus, the linear approximation $\widehat{W}$ is a line through the points $(\underline{x}, S(\underline{x}))$ and $(\bar{x}, S(\bar{x}))$, i.e.,

$$\widehat{W} = S(\underline{x}) + \frac{S(\bar{x}) - S(\underline{x})}{\bar{x} - \underline{x}} (x - \underline{x}).$$

Then the convex underestimator will be

$$\begin{aligned} g(x) + S(x) - \widehat{W} &= g(x) + S(x) - \left(S(\underline{x}) + \frac{S(\bar{x}) - S(\underline{x})}{\bar{x} - \underline{x}} (x - \underline{x}) \right) \\ &= \dots = g(x) + \alpha(x - \underline{x})(x - \bar{x}). \end{aligned}$$

□

If parts of the nonconvex function g is not separable, we must use the same $\alpha_{i,k}$ -value for each nonseparable variable x_j in an entire “slice”,

$$[\underline{x}_{j,k}, \bar{x}_{j,k}] \times \prod_{\substack{i \neq j \\ x_i \text{ not sep.}}} [\underline{x}_i, \bar{x}_i], \quad (26)$$

of the domain, in comparison to the original α BB-underestimator, which may use a different α -values in each subregion of the multidimensional domain.

To illustrate the reformulation technique, we now conclude this section with an example.

Example 2 In this example, the same function as considered in Example 1, i.e., the function

$$h(x) = \underbrace{x \sin x}_{g(x)} + \underbrace{x/10}_{q(x)}, \quad (27)$$

is underestimated using the reformulation technique with both the original and spline α BB-underestimators. The underestimators are also compared to the α BB-underestimator as used in a branch-and-bound framework. The function consists of a nonconvex and a convex term, and only the former needs to be convexified and underestimated.

If the breakpoints for the variable x are, e.g., $\omega_1 = 0$, $\omega_2 = 4$ and $\omega_3 = 8$, then we will have two different intervals where the α -values should be determined. In this case the optimal α -values are $\alpha_1 = 1.6$ and $\alpha_2 = 4.10293$ respectively.

Thus, in the original α BB-underestimator, the convex underestimators

$$g(x) + 1.6x(x - 4) + q(x) \quad \text{and} \quad g(x) + 4.10293(x - 4)(x - 8) + q(x)$$

are used in the intervals $x \in [0, 4]$ and $x \in [4, 8]$ respectively.

The expressions for the piecewise reformulated α BB-underestimator are on the other hand

$$g(x) \boxed{+\alpha x^2 - \widehat{W}} + q(x),$$

where $\alpha = 4.10293$ according to Eq. (20) and the following constraints are used to model the PLFs

$$\begin{aligned} \widehat{W} &= \alpha \cdot 0^2 b_1 + (\alpha \cdot 4^2 - \alpha \cdot 0^2) s_1 + \alpha \cdot 4^2 b_2 + (\alpha \cdot 8^2 - \alpha \cdot 4^2) s_2 \\ &= 65.647 s_1 + 65.647 b_2 + 196.941 s_2, \\ x &= 0 b_1 + (4 - 0) s_1 + 4 b_2 + (8 - 4) s_2 = 4 s_1 + 4 b_2 + 4 s_2, \\ 0 &\leq s_1 \leq b_1, \quad 0 \leq s_2 \leq b_2, \quad \text{and} \quad b_1 + b_2 = 1. \end{aligned}$$

The expression for the piecewise reformulated spline α BB-underestimator is

$$g(x) \boxed{+S(x) - \widehat{W}} + q(x),$$

where S is the piecewise smooth convex spline-function defined by

$$S(x) = \begin{cases} \alpha_1 x^2 + \beta_1 x + \gamma_1 & \text{if } 0 \leq x \leq 4 \\ \alpha_2 x^2 + \beta_2 x + \gamma_2 & \text{if } 4 \leq x \leq 8. \end{cases} \quad (28)$$

The α 's are the same as for the original α BB-underestimator. The additional parameters β and γ are determined using the continuity and smoothness requirements of the polynomials, as well as, the additional condition that S is defined to be zero at the endpoints of the entire interval, i.e., $[0, 8]$. In this particular case the parameters are $\alpha_1 = 1.6$, $\beta_1 = -17.80586$, $\gamma_1 = 0$ and $\alpha_2 = 4.10293$, $\beta_2 = -37.82930$, $\gamma_2 = 40.04688$. The expressions for the PLFs (28) are the same, except that $\widehat{W}$ changes according to

$$\begin{aligned} \widehat{W} &= S(0) b_1 + [S(4) - S(0)] s_1 + S(4) b_2 + [S(8) - S(4)] s_2 \\ &= -47.1503 s_1 - 47.1503 b_2 + 47.1503 s_2. \end{aligned} \quad (29)$$

The underestimators are illustrated in Fig. 4, for some different sets of breakpoints. As additional breakpoints are added, the underestimation error decreases. Also, from the figure it can be seen that the reformulated original α BB-underestimator coincides with the original

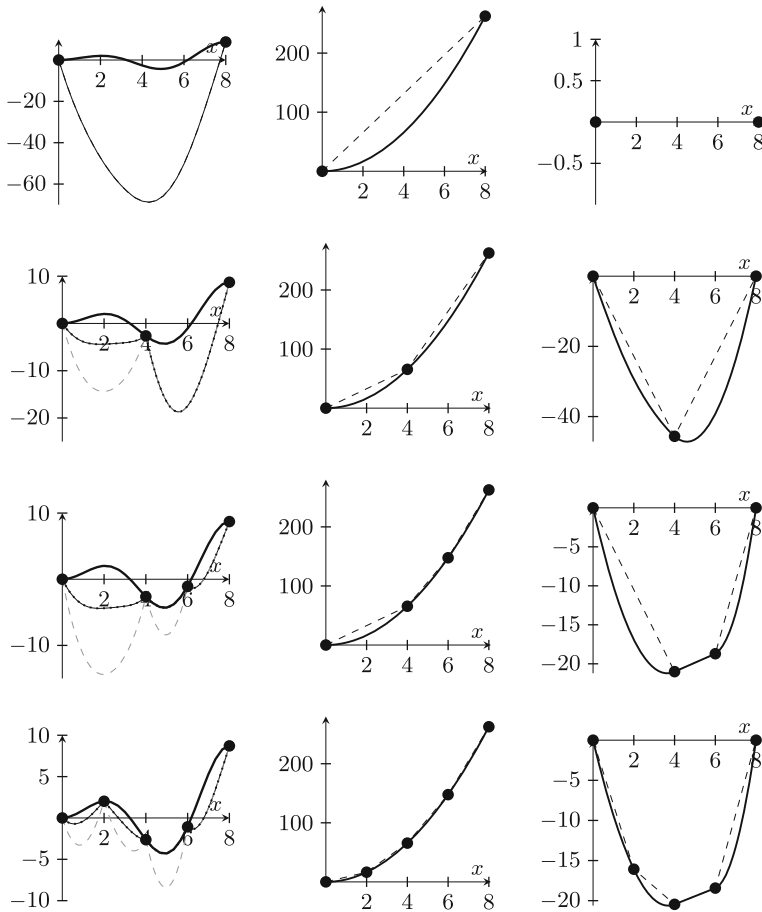


Fig. 4 Left the original α BB-underestimator (thin, black) for the function $h(x) = x \sin x + x/10$ (thick, black) compared to the reformulated original α BB-underestimator (dashed) and the reformulated spline α BB-underestimator. The spline version coincides with the original α BB-underestimator in all points. Middle the piecewise approximation $\tilde{W}$ of the function $W = \alpha x^2$, used in the reformulated original α BB-underestimator. Right the piecewise approximation of the function $S(x)$ used in the reformulated spline α BB-underestimator

α BB-underestimator only in the interval(s) with the largest α -value. From the figure, it is also found that the reformulated spline α BB-underestimator now coincides with the the original α BB-underestimator in all intervals, since g is a function of one variable, i.e., it is separable, cf. Theorem 2.

5 Exclusion of infeasible intervals

In a branch-and-bound type algorithm, nodes that are infeasible are discarded and thus the complexity of the problem is reduced. Unfortunately, in a reformulation method as that of the SGO-algorithm this cannot be done directly. Using bound-tightening strategies, loose variable bounds can, however, be tightened, but since the whole domain is considered in each iteration, only the end points of the variable bounds are altered in such a process.

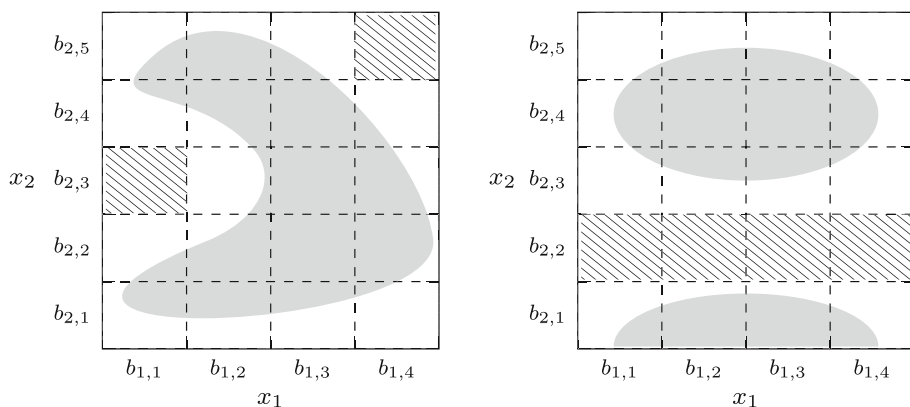


Fig. 5 Left the regions $R_1 = \{(1, 1), (2, 3)\}$ and $R_2 = \{(1, 4), (2, 5)\}$ excluded using the expressions $b_{1,1} + b_{2,3} \leq 1$ and $b_{1,4} + b_{2,5} \leq 1$. Right the region $R_3 = \{(2, 2)\}$ excluded using the expression $b_{2,2} \leq 0$. The grey region indicates the feasible region of the original nonconvex problem

To remedy this, a method for removing infeasible regions from the variable-space is introduced in this section. It is based on adding additional linear constraints to the transformed problem which exclude the infeasible regions of the original nonconvex problem (not the convex reformulated one). This may remove valid solutions in the overestimated problem, which are not valid in the nonconvex problem. An excluded region is defined as a set of variables (not necessarily containing all variables in the problem) and indices that correspond to a specific breakpoint interval of each variable. The regions that can be excluded are determined by the breakpoints of the PLFs, and all or only a selection of the regions are then checked for feasibility.

To exemplify the region exclusion technique, assume that we have a problem where the variables x_1 and x_2 are involved in either a power or exponential transformation or the α BB-underestimator and that they have the sets of consecutive breakpoints $\Omega_1 = \{\omega_{1,1}, \dots, \omega_{1,K_1}\}$ and $\Omega_2 = \{\omega_{2,1}, \dots, \omega_{2,K_2}\}$ respectively. If expressions of the type (17–19) are used for the PLFs, we have a binary variable $b_{i,k}$ corresponding to each breakpoint interval for the i -th variable, and these binaries can be used to exclude infeasible regions. For example, assume that the regions corresponding to the first breakpoint interval of x_1 and third breakpoint interval of x_2 as well as the fourth breakpoint interval of x_1 and fifth breakpoint interval of x_2 , do not fulfill some of the constraints in the problem. These regions are described by the sets

$$R_1 = \{(1, 1), (2, 3)\} \quad \text{and} \quad R_2 = \{(1, 4), (2, 5)\},$$

where the index pair in each element of the sets corresponds to the variable index and breakpoint index respectively. These regions can then be excluded by introducing the linear constraints

$$b_{1,1} + b_{2,3} \leq 1 \quad \text{and} \quad b_{1,4} + b_{2,5} \leq 1,$$

since this prevents the two binaries in each constraint from being equal to one at the same time. The excluded regions are illustrated in Fig. 5.

If there is no pair corresponding to a variable transformed in the problem, all the breakpoint intervals for the transformed variable missing are also excluded. For example, if we have the variables x_1 and x_2 involved in transformations and the second breakpoint interval of x_2 excluded by the constraint $b_{2,2} \leq 0$, i.e., $R_3 = \{(2, 2)\}$, actually all the regions

$$\{(1, 1), (2, 2)\}, \{(1, 2), (2, 2)\}, \dots, \{(1, K_1 - 1), (2, 2)\},$$

(where K_1 is the number of breakpoint intervals for x_1) will be excluded from the feasible region of the problem. The excluded regions R_1 , R_2 and R_3 are also shown in Fig. 5.

In general, using this type of notation for the sets, a region R_l corresponds to the exclusion constraint

$$\sum_{(r_1, r_2) \in R_l} b_{r_1, r_2} \leq \text{card } R_l - 1.$$

Now, how is it possible to obtain the infeasible regions? One method is to solve a subproblem consisting of some or all of the expressions in the original problem, but where the variable bounds have been replaced with those of the region in question. However, since this is often almost as difficult a problem to solve as the original one, other techniques are needed. The one described here, is to use interval arithmetic to obtain lower and upper bounds of the constraints, provided the variable bounds are the one of the region being investigated. Interval arithmetic evaluations are computationally quite cheap so this method is more suitable than solving an optimization problem in each region.

The test for feasibility is done on one constraint at a time (either all, or a subset of the linear and nonlinear constraints in the original nonconvex problem can be tested), for a variable-space corresponding to all the variables in the constraint. If there are variables that are not modified using either the α BB-reformulation method or signomial transformation strategies, these variables are assumed to take values between the upper and lower bounds of the variables. If the constraint is of the type “less than or equal to”, it is checked whether the lower bound of the LHS (obtained through interval arithmetic) is less than or equal to the right-hand-side (RHS) and if the constraint is of the type “greater than or equal to”, it is tested whether the upper bound of the LHS is greater than the RHS. Finally, for an equality constraint, it is tested if the RHS-value is contained in the interval the LHS assumes. If the statement is false in the region, i.e., the constraint is infeasible, it is added to the set of excluded regions R . In this case the region does not need to be tested further for infeasibility and we can continue with the next constraint.

The region exclusion technique is especially suited for equality constraints, since often a large part of the region specified by only the bounds of the variables is, in fact, infeasible.

6 Extending the SGO-algorithm to the α GO- and α SGO-algorithms

By combining the underestimation technique presented earlier in Sect. 4 with an algorithm of the SGO-algorithm type as explained in Sect. 2 or more thoroughly by Lundell et al. [21, 22, 25, 29], a new global optimization method for solving MINLP problems containing nonconvex signomial and general twice-differentiable functions can be obtained. Since signomials are twice-differentiable, there are actually two different paths that can be followed: Either the α BB-underestimator (in the original or spline form) is used for all nonconvex functions (including the signomials) or the single-variable transformation techniques from Sect. 2 is used on the signomials and the α BB-underestimator on the rest of the nonconvex functions if such are present. These two cases are from here on named the α GO-algorithm and the α SGO-algorithm respectively. Of course, it is also possible in the α SGO-algorithm to have a mixture of these strategies, where some of the signomial terms are underestimated using the α BB-underestimator.

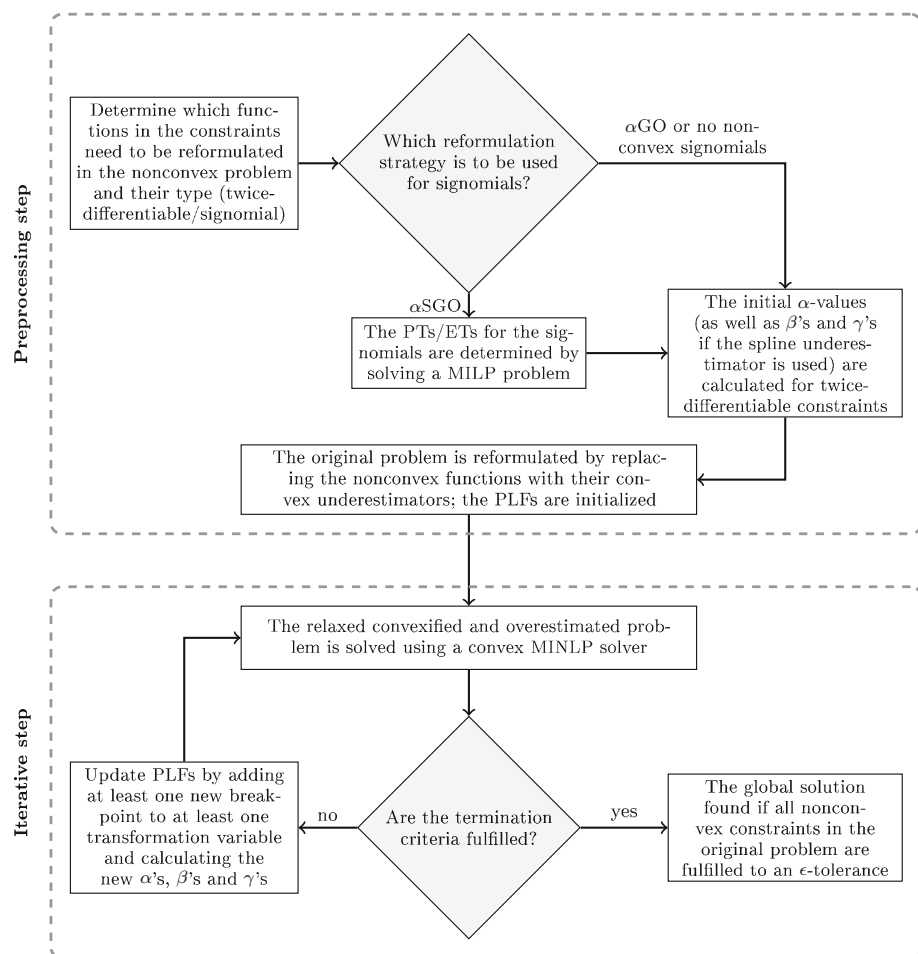


Fig. 6 A simplified flowchart of the α GO- and α SGO-algorithm

Only minor modifications are needed in the SGO-algorithm to include the α BB-underestimator in the reformulated form and to allow for general twice-differentiable functions in the nonconvex MINLP problem: The α -values (and the β - and γ -values if the spline-version is used) need to be calculated (and updated in each iteration as new breakpoints are added), and the formulations for the underestimators from Sects. 4.2 and 4.3 need to be included in the reformulated problem. Also, the termination criterion (3) needs to be updated to include the general nonconvex functions g , i.e., it will become:

$$\max_m (g_m(\mathbf{x}^*) + \sigma_m(\mathbf{x}^*) + q_m(\mathbf{x}^*)) \leq \epsilon. \quad (30)$$

In Fig. 6, a simplified flowchart of the α GO- and α SGO-algorithms is shown.

6.1 The solver implementation

The reformulation framework presented in this paper has been implemented as the α GO solver, a global optimization solver for nonconvex twice-differentiable problems of the

type (1). The solver implements both the α GO- and α SGO-algorithms and is based on the SIGOPT solver [21, 26], which implements the SGO-algorithm for solving nonconvex signomial problems.

Since SIGOPT already utilized Wolfram Mathematica for parsing and symbolic manipulation, it was natural to also calculate the α -values using the interval arithmetic functions in Mathematica in the α GO solver. However, it would be straight-forward to replace this implementation with other available interval arithmetic packages, e.g., Boost [9].

The α GO solver interfaces with the General algebraic modeling system (GAMS) [39] for solving the auxiliary MILP problem for obtaining the optimized set of signomial transformations, as well as the convex MINLP problems. In the first iteration, a GAMS problem file is created containing the reformulated MINLP problem. The breakpoints for the PLFs as well as the α -, β - and γ -values are specified separately in a parameter file, and this file is the only one modified in subsequent iterations.

By interfacing the α GO solver with GAMS, it is possible to use any of the suitable MINLP solvers available in GAMS, providing increased flexibility since certain solvers may be more efficient for certain types of problems. Different MINLP solvers could even be used in different iterations. In order to be able to define the spline α BB-underestimator as a spline polynomial function in the GAMS model, collaboration with the GAMS Development Corporation led to the introduction of the extrinsic function `pwpcclib` in GAMS version 23.7.

7 The global optimization algorithms applied to a nonconvex MINLP problem

Here, the α GO- and α SGO-algorithms are applied to a discrete trigonometric optimization problem. The problem has a convex objective function f , and only one nonconvex constraint $g + \sigma + q \leq 0$. The variable x_1 is real-valued and x_2 is integer-valued.

$$\begin{aligned} \text{minimize} \quad & f(x_1, x_2) = (2x_1 - 4)^2 + (x_2 - 13/2)^2 \\ \text{subject to} \quad & h(x_1, x_2) = \underbrace{x_1 \cos^2 x_2 + x_2 \sin^2 x_1}_{g(x_1, x_2)} \underbrace{- 3/x_2}_{\sigma(x_2)} + \underbrace{x_1/2 - 5/2}_{q(x_1)} \leq 0, \\ & 2 \leq x_1 \leq 4, \quad 2 \leq x_2 \leq 8, \quad x_1 \in \mathbb{R}, \quad x_2 \in \mathbb{Z}. \end{aligned} \quad (31)$$

The nonconvex function $g + \sigma$ can be regarded either as a twice-differentiable function or as a sum of a twice-differentiable nonconvex trigonometric function g and a nonconvex signomial function σ . Thus either the α SGO- or the α GO-algorithm can be used to solve the problem. In the first case, the α BB-underestimator is used on the entire function $g + \sigma$; in the second case, a PT is used to transform and underestimate the negative signomial term, while the α BB-underestimator is used to convexify and underestimate the trigonometric part.

If the α GO-algorithm is used, either the reformulated original α BB- or the spline α BB-underestimator can be used for handling the nonconvex constraint. When the reformulated α BB-underestimator is used, then the constraint is replaced with the following reformulation

$$x_1 \cos^2 x_2 + x_2 \sin^2 x_1 - 3/x_2 \left[+\alpha_1 x_1^2 + \alpha_2 x_2^2 - \widehat{W} \right] + x_1/2 - 5/2 \leq 0, \quad (32)$$

where the following expressions define linearized correction factor $\widehat{W}$:

$$\begin{aligned}\widehat{W} = & \sum_{k=1}^{K_1-1} A_{1,k} b_{1,k} + (A_{1,k+1} - A_{1,k}) s_{1,k} \\ & + \sum_{k=1}^{K_2-1} A_{2,k} b_{2,k} + (A_{2,k+1} - A_{2,k}) s_{2,k},\end{aligned}\quad (33)$$

Here, $A_{i,k} = \alpha_i \omega_{i,k}^2$ and $\omega_{i,k}$ is the k -th breakpoint for the variable x_i . Also, the following linear constraints defining the relationship between the original variables and the additional variables b and s in the PLFs, need to be specified as

$$x_1 = \sum_{k=1}^{K_1-1} \omega_{1,k} b_{1,k} + (\omega_{1,k+1} - \omega_{1,k}) s_{1,k} \quad \text{and} \quad (34)$$

$$x_2 = \sum_{k=1}^{K_2-1} \omega_{2,k} b_{2,k} + (\omega_{2,k+1} - \omega_{2,k}) s_{2,k}. \quad (35)$$

The real variables s should only be nonzero if the corresponding binary variable b is equal to one, i.e.,

$$0 \leq s_{i,k_i} \leq b_{i,k_i}, \quad \forall i \in \{1, 2\}, \quad k_i \in \{1, \dots, K_i - 1\}. \quad (36)$$

Finally, only one binary variable can be active at once for each variable, so the expressions

$$\sum_{k=1}^{K_1-1} b_{1,k} = 1 \quad \text{and} \quad \sum_{k=1}^{K_2-1} b_{2,k} = 1 \quad (37)$$

must also be included.

If the reformulated spline α BB-underestimator is used, the nonconvex constraint will be replaced with

$$x_1 \cos^2 x_2 + x_2 \sin^2 x_1 - 3/x_2 \left[+S_1(x_1) + S_2(x_2) - \widehat{W} \right] + x_1/2 - 5/2 \leq 0, \quad (38)$$

where S_1 and S_2 are the convex piecewise spline underestimators and $\widehat{W}$ is defined according to

$$\begin{aligned}\widehat{W} = & \sum_{k=1}^{K_1-1} S_1(\omega_{1,k}) b_{1,k} + (S_1(\omega_{1,k+1}) - S_1(\omega_{1,k})) s_{1,k} \\ & + \sum_{k=1}^{K_2-1} S_2(\omega_{2,k}) b_{2,k} + (S_2(\omega_{2,k+1}) - S_2(\omega_{2,k})) s_{2,k},\end{aligned}\quad (39)$$

The expressions (34–37) are the same as in the case of the reformulated original α BB-underestimator. The piecewise convex underestimator obtained from the reformulated spline α BB-underestimator, as well as the spline function and its linearization are shown in Fig. 7.

If instead the α SGO-algorithm is used, i.e., a power transformation is applied to the negative nonconvex signomial term, the convexified and underestimated constraint will be of the form

$$x_1 \cos^2 x_2 + x_2 \sin^2 x_1 \left[+\alpha_1 x_1^2 + \alpha_2 x_2^2 - \widehat{W} - 3\widehat{X}_2^{-Q} \right] + x_1/2 - 5/2 \leq 0, \quad (40)$$

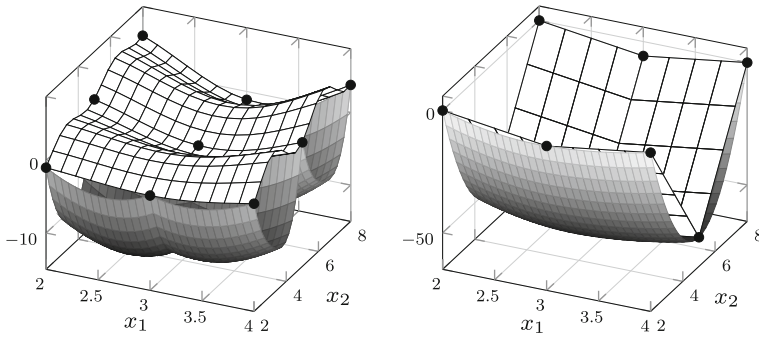


Fig. 7 The reformulated spline α BB-underestimator when the breakpoints $x_1 = 3$ and $x_2 = 5$ have been added. *Left* the original nonconvex function $g + \sigma + q$ and the piecewise convex underestimator. *Right* the spline function $S_1 + S_2$ and its linear approximation $\widehat{W}$. Note that the underestimation and approximation are exact in the illustrated points

if the reformulated original α BB-underestimator is used and

$$x_1 \cos^2 x_2 + x_2 \sin^2 x_1 \left[+S_1(x_1) + S_2(x_2) - \widehat{W} - 3\widehat{X}_2^{-Q} \right] + x_1/2 - 5/2 \leq 0, \quad (41)$$

if the reformulated spline α BB-underestimator is used. In both cases, the parameter $Q \in [-1, 0)$. One additional constraint to Eqs. (33–37) is needed to express the PLF-approximation:

$$\widehat{X}_2 = \sum_{k=1}^{K_2-1} \omega_{2,k}^{1/Q} b_{2,k} + \left(\omega_{2,k+1}^{1/Q} - \omega_{2,k}^{1/Q} \right) s_{2,k}. \quad (42)$$

Note that it is possible to use the same variables $b_{2,k}$ and $s_{2,k}$ in all transformations where the variable x_2 is involved (even a mixture of α BB-reformulations and signomial transformations). Also note that the numerical values for the parameters α_1 and α_2 in constraints (32) and (40) are not necessarily the same. The same is true about the functions S_1 and S_2 in the constraints (38) and (41).

The results obtained by solving the problem are given in Table 1 when region exclusion is disabled and Table 2 when region exclusion is enabled. The value for the error-tolerance ϵ in the constraints was 0.01. The midpoint breakpoint strategy is used, i.e., new breakpoints are added to all transformation variables in the midpoints of the breakpoint intervals of the solution point in the previous iteration. Since x_2 is an integer variable, only integer-valued breakpoints are added. The solver SBB in GAMS was used for solving the reformulated MINLP problems. As can be seen from the results, both the implementations of the α SGO-algorithm and the α GO-algorithm gives the same solution to an ϵ -tolerance using both the spline and original reformulated α BB-underestimators. However, the number of iterations required were different in some of the cases.

Illustrations of the overestimated feasible regions in each iteration in the case of the α GO-algorithm using the reformulated spline underestimator is given in Fig. 8. The results indicate the benefits of excluding infeasible regions. In this case, the reason for the smaller number of iterations required when using the exclusion strategy is since it saves us from adding unnecessary breakpoints in regions that are not part of the feasible region of the nonconvex problem. This can be seen by comparing the fourth iteration in Fig. 8a to the fourth one in Fig. 8b.

Table 1 The solution process of the example in Sect. 7 when no region exclusion of infeasible regions is performed

Iterations	Regions	$f(x_1, x_2)$	x_1	x_2	$h(x_1, x_2)$
(a) The α GO-algorithm with the reformulated normal α BB-underestimator					
1	1	0.2500	2.0	7	4.9959
2	4	0.2500	2.0	6	4.8048
3	9	0.2500	2.0	7	4.9959
4	16	3.3280	2.51914	5	0.0619
5	20	3.3523	2.52495	5	0.0378
6	24	3.3715	2.52951	5	0.0190
7	28	3.3843	2.53252	5	0.0066
(b) The α GO-algorithm with the reformulated spline α BB-underestimator					
1	1	0.2500	2.0	7	4.9959
2	4	0.2500	2.0	6	4.8048
3	9	0.2500	2.0	7	4.9959
4	16	3.3630	2.52749	5	0.0273
5	20	3.3767	2.53074	5	0.0139
6	24	3.3848	2.53263	5	0.0061
(c) The α SGO-algorithm with the reformulated normal α BB-underestimator					
1	1	0.5002	2.25008	7	3.7125
2	4	1.5175	2.56292	7	1.9034
3	9	3.3251	2.51843	5	0.0648
4	12	3.3499	2.52438	5	0.0402
5	15	3.3700	2.52916	5	0.0204
6	18	3.3837	2.53239	5	0.0071
(d) The α SGO-algorithm with the reformulated spline α BB-underestimator					
1	1	0.2500	2.0	7	4.8048
2	4	0.2500	2.0	6	4.8048
3	9	0.2500	2.0	7	4.9959
4	16	3.3630	2.52749	5	0.0273
5	20	3.3767	2.53074	5	0.0139
6	24	3.3848	2.53263	5	0.0061

8 Conclusions and further directions

In this paper, a combination of two techniques for global optimization is combined into the α GO- and α SGO-algorithms that can be used to solve any nonconvex MINLP problem containing signomial and twice-differentiable functions to global optimality. The algorithms utilize the SGO-framework, where PLFs are used to reformulate the nonconvex signomial constraint to a convex relaxed form in a higher dimensional space. By iteratively refining the PLFs, the overestimation of the feasible region in the original domain is made tighter, and the global solution can be found within an ϵ -tolerance. Since the SGO-algorithm itself can only handle nonconvex signomial functions, an extension of the method has been included for general twice-differentiable functions. The α BB-underestimator is a suitable technique,

Table 2 The solution process of the example in Sect. 7 when region exclusion of infeasible regions is performed

Iterations	Regions	Regions excl.	Domain excl. (%)	$f(x_1, x_2)$	x_1	x_2	$h(x_1, x_2)$
(a) The α GO-algorithm with the reformulated normal α BB-underestimator							
1	1	0	0	0.2500	2.0	7	4.9959
2	4	0	0	0.2500	2.0	6	4.8048
3	9	1	8.3	1.2500	2.5	7	2.2495
4	16	5	20.8	3.3516	2.52479	5	0.0384
5	20	7	21.9	3.3711	2.52941	5	0.0194
6	24	9	21.9	3.3842	2.53249	5	0.0067
(b) The α GO-algorithm with the reformulated spline α BB-underestimator							
1	1	0	0	0.2500	2.0	7	4.9959
2	4	0	0	0.2500	2.0	6	4.8048
3	9	1	8.3	1.2500	2.5	7	2.2495
4	16	5	20.8	3.3768	2.53074	5	0.0139
5	20	7	21.9	3.3848	2.53263	5	0.0061
(c) The α SGO-algorithm with the reformulated normal α BB-underestimator							
1	1	0	0	0.5001	2.25008	7	3.7126
2	4	0	0	1.5175	2.56292	7	1.9034
3	9	1	8.3	3.3251	2.51843	5	0.0648
4	12	1	8.3	3.3499	2.52438	5	0.0402
5	15	1	8.3	3.3700	2.52916	5	0.0204
6	18	2	9.4	3.3837	2.53239	5	0.0071
(d) The α SGO-algorithm with the reformulated spline α BB-underestimator							
1	1	0	0	0.2500	2.0	7	4.9959
2	4	0	0	0.2500	2.0	6	4.8048
3	9	1	8.3	1.2500	2.5	7	2.2495
4	16	5	20.8	3.3767	2.53074	5	0.0138
5	20	7	21.9	3.3848	2.53263	5	0.0061

and it was shown here, how it can be integrated, in its original or spline form, into the SGO-framework using the same type of PLF-reformulation technique. This results in an easily implemented algorithm for general twice-differentiable MINLP problems.

Since there are certain degrees of freedom regarding the way nonconvex functions are transformed to convex form, e.g., choosing between power and exponential transformations versus using the reformulated α BB-underestimator for nonconvex signomial terms, the optimization method for the signomial transformations can be extended to also include the α BB-underestimators. Thus, e.g., the set of transformations resulting in the minimal number of transformation variables required can easily be obtained, reducing the variable-space of the reformulated problem.

Since the α GO-algorithm considers the whole domain in each iteration in contrast to a branching algorithm, certain techniques can be implemented to reduce the domain by removing infeasible regions in the nonconvex problem which are feasible in the reformulated convex overestimated problem. This is achieved by including certain restrictions on the binary variables used in the PLFs. This exclusion method could, however, probably be

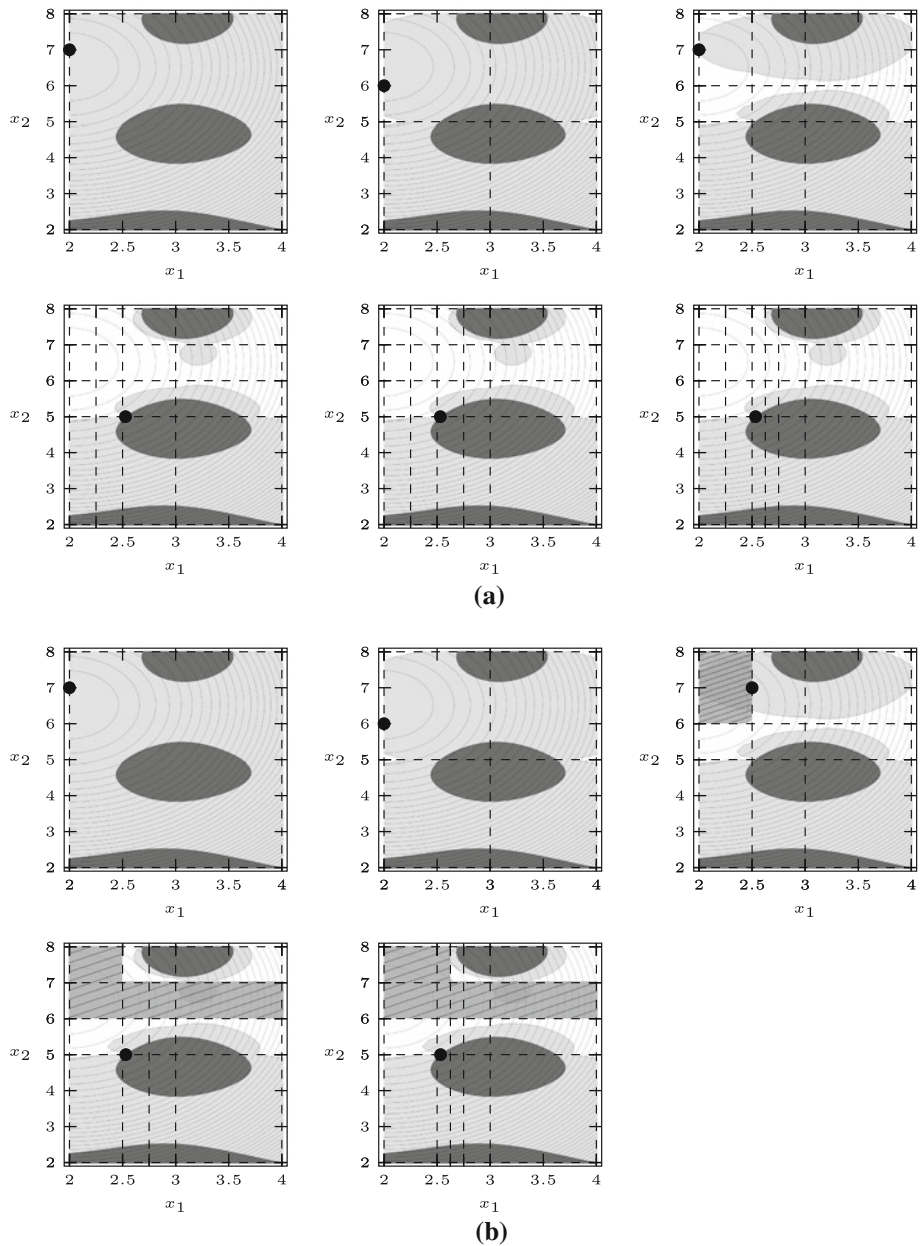


Fig. 8 Illustrations of the subproblems in each iteration when solving the problem in Sect. 7 with or without exclusion of the infeasible regions when using the α GO-algorithm. Each plot correspond to one iteration of the algorithm. The level curves of the convex objective function are shown in the figure. The vertical and horizontal dashed lines indicate the breakpoints. The dark gray areas indicate the integer-relaxed feasible region of the nonconvex problem and the light gray areas the overestimated integer-relaxed feasible region in the different iterations. The dark slanted rectangular regions are the excluded regions. The point represents the solution in each iteration. **a** Without exclusion of infeasible regions, **b** with exclusion of infeasible regions

further enhanced. For instance, it is possible to increase the variable bounds excluded, to mimic bound reduction techniques used in branch-and-bound based algorithms.

One other interesting approach would be to develop a hybrid reformulation and branching algorithm, so that the reformulation strategies can be used in certain nodes within a branching tree. In such a hybrid method, an entire subdomain in the node can be considered without additional branching (and for example fathomed) by solving only one reformulated problem.

Acknowledgments Financial support from the Foundation of Åbo Akademi University, as part of the grant for the Center of Excellence in Optimization and Systems Engineering, is gratefully acknowledged. We also acknowledge Michael Bussieck and Lutz Westermann from GAMS Development Corporation for their help with implementing the spline polynomial as an extrinsic function in GAMS.

References

- Adjiman, C.S., Androulakis, I.P., Floudas, C.A.: Global optimization of mixed-integer nonlinear problems. *AIChE J.* **46**(9), 1769–1797 (2000)
- Adjiman, C.S., Dallwig, S., Floudas, C.A., Neumaier, A.: A global optimization method, α BB, for general twice-differentiable constrained NLPs—I. Theoretical advances. *Comput. Chem. Eng.* **22**(9), 1137–1158 (1998)
- Adjiman, C.S., Dallwig, S., Floudas, C.A., Neumaier, A.: A global optimization method, α BB, for general twice differentiable NLPs—II. Implementation and computational results. *Comput. Chem. Eng.* **22**, 1159–1179 (1998)
- Adjiman, C.S., Androulakis, I.P., Floudas, C.A.: Global optimization of MINLP problems in process synthesis and design. *Comput. Chem. Eng.* **21**, 445–450 (1997)
- Akrotirianakis, I.G., Floudas, C.A.: Computational experience with a new class of convex underestimators: box-constrained NLP problems. *J. Glob. Optim.* **29**(3), 249–264 (2004)
- Akrotirianakis, I.G., Floudas, C.A.: A new class of improved convex underestimators for twice continuously differentiable constrained NLPs. *J. Glob. Optim.* **30**(4), 367–390 (2004)
- Androulakis, I.P., Maranas, C.D., Floudas, C.A.: α BB: a global optimization method for general constrained nonconvex problems. *J. Glob. Optim.* **7**, 337–363 (1995)
- Björk, K.-M.: A global optimization method with some heat exchanger network applications. Ph.D. thesis, Åbo Akademi University (2002)
- Brönnimann, H., Melquiond, G., Pion, S.: The design of the Boost interval arithmetic library. *Theor. Comput. Sci.* **351**, 111–118 (2006)
- Dembo, R.S.: Current state of the art of algorithms and computer software for geometric programming. *J. Optim. Theory Appl.* **26**(2), 149–183 (1978)
- Floudas, C.A.: Deterministic global optimization. Theory, methods and applications. Number 37 in *Nonconvex Optimization and Its Applications*. Kluwer Academic Publishers, Dordrecht (2000)
- Floudas, C.A., Kreinovich, V.: On the functional form of convex underestimators for twice continuously differentiable functions. *Optim. Lett.* **1**, 187–192 (2007)
- Floudas, C.A., Kreinovich, V.: Towards optimal techniques for solving global optimization problems: symmetry-based approach. In: Törn, A., Zilinskas, J. (eds.) *Models and Algorithms for Global Optimization*, pp. 21–42. Springer US, Boston, MA (2007)
- Horst, R., Pardalos, P.M., Romeijn, H.E.: Handbook of global optimization. Number 2 in *Nonconvex Optimization and Its Applications*. Kluwer Academic Publishers, Dordrecht (2002)
- Horst, R., Pardalos, P.M., Thoai, N.V.: Introduction to Global Optimization. *Nonconvex Optimization and its Applications*. Kluwer Academic Publishers, Dordrecht (2000)
- Jeroslow, R.G., Lowe, J.K.: Modelling with integer variables. In: *Mathematical Programming at Oberwolfach II*, vol. 22 of *Mathematical Programming Studies*, pp. 167–184. Springer, Berlin (1984)
- Li, H.-L., Tsai, J.F., Floudas, C.A.: Convex underestimation for posynomial functions of positive variables. *Optim. Lett.* **2**(3), 333–340 (2008)
- Liberti, L., Cafieri, S., Tarissan, F.: Reformulations in mathematical programming: a computational approach. In: Abraham, A., Hassanien, A.-E., Siarry, P., Engelbrecht, A. (eds.) *Foundations of Computational Intelligence Volume 3*, vol. 203 of *Studies in Computational Intelligence*, pp. 153–234. Springer, Berlin (2009)
- Lin, M.-H., Tsai, J.-F.: Range reduction techniques for improving computational efficiency in global optimization of signomial geometric programming problems. *Eur. J. Oper. Res.* **216**(1), 17–25 (2012)

20. Liu, W.B., Floudas, C.A.: A remark on the GOP algorithm for global optimization. *J. Glob. Optim.* **3**, 519–521 (1993)
21. Lundell, A.: Transformation techniques for signomial functions in global optimization. Ph.D. thesis, Åbo Akademi University (2009)
22. Lundell, A., Westerlund, J., Westerlund, T.: Some transformation techniques with applications in global optimization. *J. Glob. Optim.* **43**(2), 391–405 (2009)
23. Lundell, A., Westerlund, T.: Optimization of power transformations in global optimization. *Chem. Eng. Trans.* **11**, 95–100 (2007)
24. Lundell, A., Westerlund, T.: Exponential and power transformations for convexifying signomial terms in MINLP problems. In: Bruzzone, L. (ed.) *Proceedings of the 27th IASTED International Conference on Modelling, Identification and Control*, pp. 154–159. ACTA Press, Anaheim
25. Lundell, A., Westerlund, T.: Convex underestimation strategies for signomial functions. *Optim. Methods Softw.* **24**, 505–522 (2009)
26. Lundell, A., Westerlund, T.: Implementation of a convexification technique for signomial functions. In: Jezowski, J., Thullie, J. (eds.) *Proceedings of the 19th European Symposium on Computer Aided Process Engineering*, vol. 26 of *Computer Aided Chemical Engineering*, pp. 579–583. Elsevier, Amsterdam (2009)
27. Lundell, A., Westerlund, T.: On the relationship between power and exponential transformations for positive signomial functions. *Chem. Eng. Trans.* **17**, 1287–1292 (2009)
28. Lundell, A., Westerlund, T.: Optimization of transformations for convex relaxations of MINLP problems containing signomial functions. In: de Brito Alves, R.M., do Nascimento, C.A.O., Biscaia, E.C. (eds.) *Proceedings of the 10th International Symposium on Process Systems Engineering: Part A*, vol. 27 of *Computer Aided Chemical Engineering*, pp. 231–236. Elsevier, Amsterdam (2009)
29. Lundell, A., Westerlund, T.: Global optimization of mixed-integer signomial programming problems. In: Lee, J., Leyffer, S. (eds.) *Mixed Integer Nonlinear Programming*, vol. 154 of *The IMA Volumes in Mathematics and its Applications*, pp. 349–369. Springer, New York, NY (2012)
30. Maranas, C.D., Floudas, C.A.: Finding all solutions of nonlinearly constrained systems of equations. *J. Glob. Optim.* **7**, 143–182 (1995)
31. Maranas, C.D., Floudas, C.A.: Global optimization in generalized geometric programming. *Comput. Chem. Eng.* **21**(4), 351–369 (1997)
32. Meyer, C.A., Floudas, C.A.: Convex underestimation of twice continuously differentiable functions by piecewise quadratic perturbation: Spline α BB underestimators. *J. Glob. Optim.* **32**(2), 221–258 (2005)
33. Pardalos, P.M., Romeijn, H.E., Tuy, H.: Recent developments and trends in global optimization. *J. Comput. Appl. Math.* **124**(1–2), 209–228 (2000)
34. Pardalos, P.M., Rosen, J.B.: *Constrained Global Optimization: Algorithms and Applications*, vol. 268 of *Lecture notes in Computer Science*. Springer, Berlin (1987)
35. Peterson, E.L.: The origins of geometric programming. *Ann. Oper. Res.* **105**, 15–19 (2001)
36. Pörn, R., Björk, K.-M., Westerlund, T.: Global solution of optimization problems with signomial parts. *Discret. Optim.* **5**, 108–120 (2008)
37. Pörn, R., Harjunkski, I., Westerlund, T.: Convexification of different classes of non-convex MINLP problems. *Comput. Chem. Eng.* **23**, 439–448 (1999)
38. Rijckaert, M.J., Martens, X.M.: Comparison of generalized geometric programming algorithms. *J. Optim. Theory Appl.* **26**(2), 205–242 (1978)
39. Rosenthal, R.E.: *GAMS—A User's Guide*. GAMS Development Corporation, Washington, DC (2008)
40. Skjäl, A., Lundell, A., Westerlund, T.: Global optimization with C^2 constraints by convex reformulations. *Chem. Eng. Trans.* **24**, 373–378 (2011)
41. Tsai, J.F., Lin, M.-H.: Global optimization of signomial mixed-integer nonlinear programming problems with free variables. *J. Glob. Optim.* **42**(1), 39–49 (2008)
42. Tsai, J.F., Lin, M.-H.: An efficient global approach for posynomial geometric programming problems. *INFORMS J. Comput.* **23**, 483–492 (2011)
43. Westerlund, T.: Some transformation techniques in global optimization. In: Liberti, L., Maculan, N. (eds.) *Global Optimization: From Theory to Implementation* vol. 84 of *Nonconvex Optimization and its Applications*, pp. 47–74. Springer, Berlin (2005)
44. Westerlund, T., Lundell, A., Westerlund, J.: On convex relaxations in nonconvex optimization. *Chem. Eng. Trans.* **24**, 331–336 (2011)
45. Westerlund, T., Westerlund, J.: GGPECP—An algorithm for solving non-convex MINLP problems by cutting plane and transformation techniques. *Chem. Eng. Trans.* **3**, 1045–1050 (2003)



Discrete Optimization

A new exact discrete linear reformulation of the quadratic assignment problem

Axel Nyberg*, Tapio Westerlund

Process Design and Systems Engineering, Åbo Akademi University, FIN-20500 Åbo, Finland

ARTICLE INFO

Article history:

Received 11 March 2011

Accepted 7 February 2012

Available online 16 February 2012

Keywords:

Combinatorial optimization
 Quadratic assignment problem
 Discrete linear reformulation
 Mixed integer programming
 Global Optimization

ABSTRACT

The quadratic assignment problem (QAP) is a challenging combinatorial problem. The problem is NP-hard and in addition, it is considered practically intractable to solve large QAP instances, to proven optimality, within reasonable time limits. In this paper we present an attractive mixed integer linear programming (MILP) formulation of the QAP. We first introduce a useful non-linear formulation of the problem and then a method of how to reformulate it to a new exact, compact discrete linear model. This reformulation is efficient for QAP instances with few unique elements in the flow or distance matrices. Finally, we present optimal results, obtained with the discrete linear reformulation, for some previously unsolved instances (with the size $n = 32$ and 64), from the quadratic assignment problem library, QAPLIB.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

The quadratic assignment problem was introduced by Koopmans and Beckmann (1957) in the basic form:

$$\min \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \sum_{l=1}^n a_{ij} b_{kl} \cdot x_{ik} x_{jl} \quad (1)$$

subject to

$$\sum_{i=1}^n x_{ij} = 1, \quad j = 1, \dots, n; \quad (2)$$

$$\sum_{j=1}^n x_{ij} = 1, \quad i = 1, \dots, n; \quad (3)$$

$$x_{ij} \in \{0, 1\}, \quad i, j = 1, \dots, n; \quad (4)$$

a_{ij} are given distances between locations and b_{kl} flows between facilities defined in the matrices, **A** and **B**. The QAP is an NP-hard problem and many different formulations and methods for this problem have been suggested. These include heuristics (Taillard, 1991), linear reformulation techniques (Adams et al., 2007) and convex quadratic programming (Anstreicher and Brixius, 2001) to name a few. Computing lower bounds for QAPs using SDP relaxations has recently also been widely studied (Peng et al., 2010). In fact over one hundred papers have been published in the subject since 1999 according to Loiola et al. (2007). However, to this date only a few QAPs of size $n \geq 30$ from QAPLIB (Burkard et al., 1997; Hahn and Anjos, 2002)

have been solved to proven optimality including the famous nug30 and the kra instances (Anstreicher et al., 2002). Most of these instances have been solved using computers connected in parallel and the CPU times required to solve the models, calculated for a single computer, are immense. It should however be mentioned that in addition to the instances in the QAPLIB Drezner et al. (2005) reported solutions of some special classes of QAPs as big as $n = 75$.¹

This paper presents a compact exact discrete linear reformulation (DLR) of the QAP that can be solved using general MILP solvers. This formulation gives promising results on instances where one of the matrices has few unique elements per row. Various different MILP formulations have been presented through the years. The model with the smallest amount of variables and constraints, is to our knowledge, the model presented by Kaufman and Broeckx (1978). Their model contains n^2 binary variables, n^2 real variables and $2n + n^2$ constraints. The formulation gives, however, very poor lower bounds and the tightening of the formulation has recently been studied by Zhang et al. (2010). The MILP model in this paper contains $\alpha \cdot n^2$ binary variables, $\beta \cdot n^2$ real variables and $\gamma \cdot n + \delta \cdot n^2$ constraints, where α , β , γ and δ are problem specific parameters. In the worst case, our formulation has n^2 binary variables, n^3 real variables and $2n + n^3$ constraints. However, this is almost never the case. Values for these parameters for some instances in the QAPLIB are found in Sections 3–4. With respect to the notations, vectors and matrices are written in bold face throughout the paper.

* Corresponding author. Tel.: +358 407783015.

E-mail addresses: axnyberg@abo.fi (A. Nyberg), twesterl@abo.fi (T. Westerlund).¹ In addition, after the submission of this paper, Fischetti, Monaci and Salvagin reported the solution of the instance esc128 in QAPLIB with $n = 128$.

2. QAP-reformulations

Consider two permutation vectors $\mathbf{p}$ and $\tilde{\mathbf{p}}$ where $p_i = k$ if facility i is at location k and $\tilde{p}_i = k$ if facility k is at location i . The objective function in Eq. (1), can then be written as:

$$\sum_{i=1}^n \sum_{j=1}^n a_{p_i p_j} b_{ij} = \sum_{i=1}^n \sum_{j=1}^n a_{ij} b_{\tilde{p}_i \tilde{p}_j}. \quad (5)$$

In matrix form Eqs. (2) and (3) can be written as $\mathbf{X}\mathbf{e} = \mathbf{X}^T \mathbf{e} = \mathbf{e}$ where $\mathbf{e}$ is a vector with all elements equal to 1 and $\mathbf{X}$ is the so-called permutation matrix containing the binary variables, x_{ij} . The permutation vectors $\mathbf{p}$ and $\tilde{\mathbf{p}}$ are given by $\mathbf{p} = \mathbf{X}\mathbf{q}$ and $\tilde{\mathbf{p}} = \mathbf{X}^T \mathbf{q}$ where $\mathbf{q}^T = (1, 2, \dots, n)$. Using the properties of the permutation vectors, Eq. (5) can be rewritten in the form:

$$\sum_{i=1}^n \sum_{j=1}^n a_{p_i p_j} b_{ij} = \sum_{i=1}^n \sum_{j=1}^n a'_{ij} b'_{ij}, \quad (6)$$

where a'_{ij} and b'_{ij} are given by:

$$a'_{ij} = \sum_{k=1}^n a_{kj} x_{ik} \quad \forall i, j, \quad (7)$$

$$b'_{ij} = \sum_{k=1}^n b_{ik} x_{kj} \quad \forall i, j. \quad (8)$$

The row sum and the column sum of the binary variables x are equal to 1 according to Eqs. (2) and (3).

Using more compact matrix notation, Eqs. (6)–(8) can also be written as $\mathbf{X}\mathbf{A}\bullet\mathbf{B}\mathbf{X}$ where $\bullet$ stands for the scalar product of the matrices, defined as the sum of the product of the corresponding elements. This representation is considered later on in Section 3. From Eq. (6) we find that the objective function is a sum of bilinear terms, $a'_{ij} b'_{ij}$, of discrete variables, where each variable can obtain, at most, n different numerical values, defined by the elements of corresponding columns and rows of the matrices $\mathbf{A}$ and $\mathbf{B}$ respectively. We can formulate a linear relaxation, w_{ij} , of the bilinear term $a'_{ij} b'_{ij}$ by considering the discrete nature of the variable b'_{ij} as follows:

$$w_{ij} \geq \sum_{m=1}^{M_i} B_i^m z_{ij}^m \quad (9)$$

$$\sum_{m=1}^{M_i} z_{ij}^m = a'_{ij} \quad (10)$$

$$z_{ij}^m \leq \bar{A}_j \sum_{k \in K_i^m} x_{kj} \quad \forall m = 1, \dots, M_i \quad (11)$$

B_i^m are the constant unique values of the elements b_{ij} in row i of the matrix $\mathbf{B}$ and M_i the number of unique elements in the row, where $M_i \leq n$. z_{ij}^m are nonnegative variables of which one will be equal to a'_{ij} while the others will be zero. $\bar{A}_j = \max_i a_{ij}$ (i.e. the largest element in column j of the matrix $\mathbf{A}$). The index sets K_i^m are connected to the unique values of the elements (with corresponding binary variables) of row i in the $\mathbf{B}$ matrix and are defined as follows:

$$K_i^m = \{j | b_{ij} = B_i^m\} \quad \forall j \quad \wedge m = 1, \dots, M_i. \quad (12)$$

Since both a'_{ij} and b'_{ij} are nonnegative variables and the objective is to minimize a sum of such bilinear terms we will get an exact discrete linear relaxed reformulation of the QAP by including the constraints in Eqs. (10) and (11) and by replacing the bilinear terms $a'_{ij} b'_{ij}$ in the objective function with the corresponding variables w_{ij} . The complete formulation is given in the next section.

2.1. A simple example

To conclude this section we will illustrate the exact relaxation of one bilinear term as given in Eqs. (9)–(12). We consider the following QAP with $n = 5$ where:

$$\mathbf{A} = \begin{bmatrix} 0 & 3 & 5 & 9 & 6 \\ 3 & 0 & 2 & 6 & 9 \\ 5 & 2 & 0 & 8 & 10 \\ 9 & 6 & 8 & 0 & 2 \\ 6 & 9 & 10 & 2 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 0 & 4 & 3 & 7 & 7 \\ 4 & 0 & 4 & 10 & 4 \\ 3 & 4 & 0 & 2 & 3 \\ 7 & 10 & 2 & 0 & 4 \\ 7 & 4 & 3 & 4 & 0 \end{bmatrix}$$

Now, as an illustrative example, consider the relaxation of the bilinear term $a'_{ij} b'_{ij}$, with $i = 2$ and $j = 3$, i.e. $a'_{23} b'_{23}$. According to Eqs. (7) and (8) a'_{23} and b'_{23} are given by:

$$a'_{23} = 5x_{21} + 2x_{22} + 0x_{23} + 8x_{24} + 10x_{25},$$

$$b'_{23} = 4x_{13} + 0x_{23} + 4x_{33} + 10x_{43} + 4x_{53},$$

where according to Eqs. (2) and (3):

$$x_{13} + x_{23} + x_{33} + x_{43} + x_{53} = 1,$$

$$x_{21} + x_{22} + x_{23} + x_{24} + x_{25} = 1.$$

Since row two ($i = 2$) in the matrix $\mathbf{B}$ has three unique values, namely 0, 4 and 10 and Eq. (2) holds, (i.e. $x_{13} + x_{23} + x_{33} + x_{43} + x_{53} = 1$), we get $b'_{23} \in \{0, 4, 10\} = \{B_2^1, B_2^2, B_2^3\}$ and $M_2 = 3$. Thus according to Eq. (9):

$$w_{23} \geq 0z_{23}^1 + 4z_{23}^2 + 10z_{23}^3.$$

Furthermore, according to Eq. (10):

$$z_{23}^1 + z_{23}^2 + z_{23}^3 = 5x_{21} + 2x_{22} + 0x_{23} + 8x_{24} + 10x_{25},$$

where the RHS is equal to a'_{23} . The index sets K_2^m (connected to the unique values of the elements of row two in the $\mathbf{B}$ matrix) are, according to Eq. (12), in this case $K_2^1 = \{2\}$; $K_2^2 = \{1, 3, 5\}$ and $K_2^3 = \{4\}$. Furthermore, $\bar{A}_3 = 10$ (i.e. the largest value in column three of the $\mathbf{A}$ matrix). From Eq. (11) we, thus, obtain:

$$z_{23}^1 \leq 10x_{23},$$

$$z_{23}^2 \leq 10(x_{13} + x_{33} + x_{53}),$$

$$z_{23}^3 \leq 10x_{43}.$$

Since only one of the binary variables above will be equal to 1 and the others equal to 0 only one of the z_{23}^m variables can be nonzero and, thus, according to Eq. (10) equal to a'_{23} . Furthermore, observe that the binary variables above are connected to the parameters B_2^m , i.e. the unique values of b'_{23} .

As we minimize w_{23} , the inequality as in Eq. (9) will be active and we obtain:

$$w_{23} = 0z_{23}^1 + 4z_{23}^2 + 10z_{23}^3,$$

which result in an exact discrete linear relaxation of the bilinear term. In a similar way bilinear terms for all i, j in the QAP can be handled. Since, in a QAP, both a'_{ij} and b'_{ij} are discrete variables, we can do the discretization in either a'_{ij} or b'_{ij} . In Fig. 1 both discretizations are illustrated. The left figure correspond to the example above. From the figures we observe that the representation is a relaxation of the bilinear term for continuous values of the variables a'_{ij} and b'_{ij} but an exact reformulation of it if one of the variables is a discrete variable. A similar type of relaxation for continuous variables in pooling problems, using piecewise linear expressions, is given in Gounaris et al. (2009).

3. The discrete linear QAP reformulation

Now we reformulate every bilinear term in the QAP using the method described in Section 2. Thus we get an exact discrete linear reformulation (DLR) of the QAP. The discretization can be done in either a'_{ij} or b'_{ij} , as mentioned before. In the below reformulation, the discretization is done in the variables b'_{ij} .

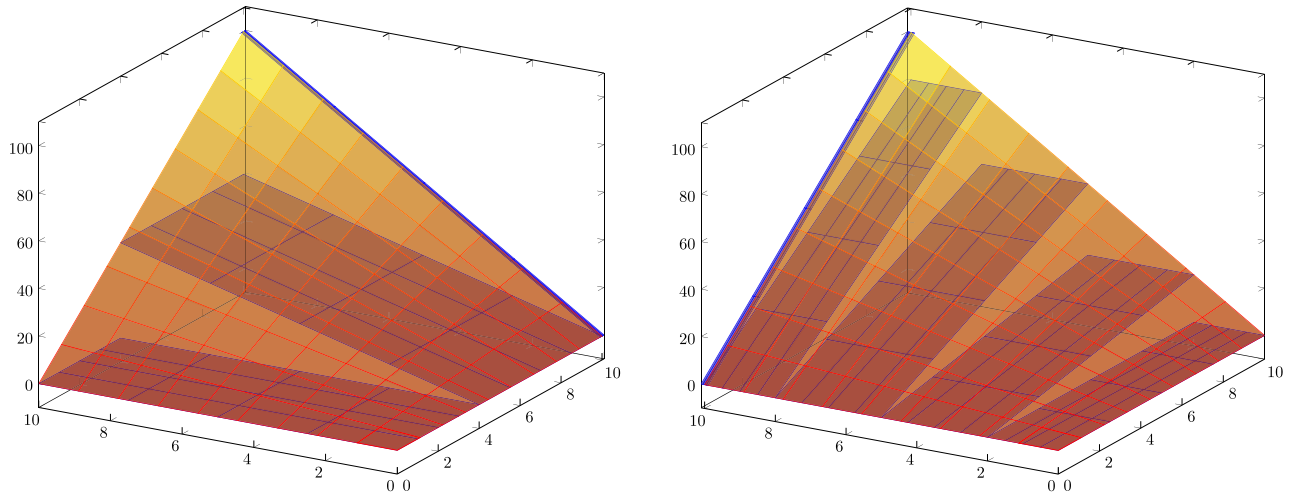


Fig. 1. Bilinear term $a'_{23}b'_{23}$ discretized in b'_{23} (to the left) and in a'_{23} (to the right).

$$\min \sum_{i=1}^n \sum_{j=1}^n \sum_{m=1}^{M_i} B_i^m z_{ij}^m \quad (13)$$

subject to

$$\sum_{i=1}^n x_{ij} = 1 \quad \forall j, \quad (14)$$

$$\sum_{j=1}^n x_{ij} = 1 \quad \forall i, \quad (15)$$

$$\left. \begin{aligned} z_{ij}^m &\leq \bar{A}_j \sum_{k \in K_i^m} x_{kj} \quad m = 1, \dots, M_i, \\ \sum_{m=1}^{M_i} z_{ij}^m &= \sum_{k=1}^n a_{kj} x_{ik} \end{aligned} \right\} \quad \forall i, j, \quad (16)$$

$$x_{ij} \in \{0, 1\} \quad z_{ij}^m \in [0, \bar{A}_j] \quad \forall i, j \wedge m = 1, \dots, M_i, \quad (17)$$

where

$$\bar{A}_j = \max_i a_{ij} \quad \forall j, \quad (18)$$

$$K_i^m = \{j | b_{ij} = B_i^m\} \quad \forall i, j \wedge m = 1, \dots, M_i, \quad (19)$$

$$b_{ij} \in \{B_i^1, B_i^2, \dots, B_i^{M_i}\} \quad \forall i, j. \quad (20)$$

The values of a_{ij} and b_{ij} are defined in the matrices **A** and **B**. The amount of variables and constraints in the model is dependent on the number of unique values in the rows or columns of the matrices (depending on which one is discretized). The number of binary variables is $\alpha \cdot n^2$ and the number of real variables $\beta \cdot n^2$ while the number of constraints is $\gamma \cdot n + \delta \cdot n^2$. In Table 1 the values of α, β, γ and δ are given for certain instances from QAPLIB when reformulated in this manner. In Table 1 the solution times when solving the instances to optimality in the different discretized forms are also given. As can be noted, fewer constraints and variables do not always result in shorter CPU times. Furthermore, it can be observed that the two discretization directions will result in highly different solution times. When the problem size increases, the choice of the direction will have a crucial importance on whether the problem can be solved or not. This can already be seen at the root node, since the value of the lower bound differs greatly depending on which way the model has been discretized.

Remark. The two optional discretization directions can algorithmically be implemented in different ways. In case the implementation is simply done by switching the matrices in the objective

Table 1

Parameters α, β, γ and δ as well as solution times in seconds, when discretizing in A or B

Instance	in A					in B				
	α	β	γ	δ	time (s)	α	β	γ	δ	time (s)
chr12a	1	2.83	2	3.83	10	1	11.33	2	12.33	2
esc16a	1	3.12	2	2.75	33	1	2.5	2	3.12	11
nug12	1	6.17	2	5.17	59	1	5.33	2	6.33	1186

function, $\mathbf{XA} \bullet \mathbf{BX}$, it should, however, be noted that the permutation matrix **X** must then be considered as its transpose as well (and the permutation vector p would be $\bar{p}$) since $\mathbf{XA} \bullet \mathbf{BX} = \mathbf{BX} \bullet \mathbf{XA} = \mathbf{X}^T \mathbf{BX} \bullet \mathbf{A} = \mathbf{X}^T \mathbf{B} \bullet \mathbf{AX}^T$.

From the formulation in Eqs. (13)–(20) it may, further, be observed that if $\bar{A}_j = 0$ then the corresponding variables, $z_{ij}^m \forall i, m$ will be equal to zero. This reduces the number of variables, constraints in Eq. (16) and corresponding terms in the objective function Eq. (13). On the other hand if $B_i^m = 0$ it only reduces corresponding terms in the objective function, but does not affect the number of variables and constraints.

4. Special structures of the QAP

Many of the instances in the QAPLIB (e.g. esc32a–esc32h) have matrices where all elements in a row and the corresponding column are equal to zero. It is important to consider such special structures of QAP since they result in multiple equal solutions. We therefore, introduce a way to exclude such solutions in the DLR model. By first considering the **B** matrix and partitioning the matrices in the objective function (Eq. (6), written in matrix form) we obtain,

$$\mathbf{XA} \bullet \mathbf{BX} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 \\ \mathbf{X}_3 & \mathbf{X}_4 \end{bmatrix} \begin{bmatrix} \mathbf{A}_{B1} & \mathbf{A}_{B2} \\ \mathbf{A}_{B3} & \mathbf{A}_{B4} \end{bmatrix} \bullet \begin{bmatrix} \mathbf{B}_1 & \mathbf{B}_2 \\ \mathbf{B}_3 & \mathbf{B}_4 \end{bmatrix} \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 \\ \mathbf{X}_3 & \mathbf{X}_4 \end{bmatrix} =$$

$$\begin{bmatrix} \mathbf{X}_1 \mathbf{A}_{B1} + \mathbf{X}_2 \mathbf{A}_{B3} & \mathbf{X}_1 \mathbf{A}_{B2} + \mathbf{X}_2 \mathbf{A}_{B4} \\ \mathbf{X}_3 \mathbf{A}_{B1} + \mathbf{X}_4 \mathbf{A}_{B3} & \mathbf{X}_3 \mathbf{A}_{B2} + \mathbf{X}_4 \mathbf{A}_{B4} \end{bmatrix} \bullet \begin{bmatrix} \mathbf{B}_1 \mathbf{X}_1 + \mathbf{B}_2 \mathbf{X}_3 & \mathbf{B}_1 \mathbf{X}_2 + \mathbf{B}_2 \mathbf{X}_4 \\ \mathbf{B}_3 \mathbf{X}_1 + \mathbf{B}_4 \mathbf{X}_3 & \mathbf{B}_3 \mathbf{X}_2 + \mathbf{B}_4 \mathbf{X}_4 \end{bmatrix}$$

When $\mathbf{B}_2, \mathbf{B}_3, \mathbf{B}_4 = \mathbf{0}$ we obtain:

$$\mathbf{XA} \bullet \mathbf{BX} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \mathbf{A}_{B1} & \mathbf{A}_{B2} \\ \mathbf{A}_{B3} & \mathbf{A}_{B4} \end{bmatrix} \bullet \mathbf{B}_1 \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{bmatrix} \mathbf{A} \bullet \mathbf{B}_1 \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{bmatrix}$$

$$= \mathbf{X}_B \mathbf{A} \bullet \mathbf{B}_1 \mathbf{X}_B$$

If the number of remaining rows and columns in the B matrix is n_B then $n - n_B$ rows of the permutation matrix have been removed. The original QAP contains $n!$ permutations, but since $(n - n_B)$ rows of the permutation matrix have been eliminated, $(n - n_B)!$ permutations have as well. The number of permutations in the reduced problem is thus $n!/(n - n_B)!$. In Table 2 and 3 examples of the influence of the compression is illustrated by some instances from QAPLIB.

As we apply this method to certain instances in QAPLIB we can decrease the number of binary variables and constraints. The difference in the formulation after the removal of the unnecessary binary variables is that the index i should be summed up to only n_B and Eq. (14) should now be written as an inequality constraint,

$$\sum_{i=1}^{n_B} x_{ij} \leq 1, \quad j = 1, \dots, n; \quad (21)$$

However, in Eq. (21) all binary variables may be equal to zero and forcing all other variables in Eq. (16) to be equal to zero as well. Then the equality constraint in Eq. (16) would not hold. Therefore, an additional variable, z_{ij}^0 , with a corresponding relaxation constraint must be added. This constraint is given by:

$$z_{ij}^0 \leq \bar{A}_j \left(1 - \sum_{k=1}^{n_B} x_{kj} \right) \quad \forall i, j. \quad (22)$$

The variable, z_{ij}^0 , is then added to the equality constraint in Eq. (16) by starting the summation in the LHS from $m = 0$. Also, in a similar way, rows and corresponding columns with all elements zero in **A** can be utilized to reduce the number of columns in the permutation matrix. In the latter case the index j should be summed up to only n_A and Eq. (15) must be written as an inequality constraint:

$$\sum_{j=1}^{n_A} x_{ij} \leq 1, \quad i = 1, \dots, n; \quad (23)$$

where n_A is the number of remaining rows and columns in the A matrix. No additional relaxation constraints connected to the inequality constraint Eq. (23) need be added in this case, as long as we have discretized in B. On the other hand, if we discretize in A an additional variable and a corresponding relaxation constraint should be connected to Eq. (23) as well.

Additional remarks. If both the matrices **A** and **B** have rows (and columns) with all elements equal to zero we obtain:

$$\mathbf{X}\mathbf{A} \bullet \mathbf{B}\mathbf{X} = \mathbf{X}_B \mathbf{A} \bullet \mathbf{B}_1 \mathbf{X}_B = \mathbf{X}_A \mathbf{A}_1 \bullet \mathbf{B}\mathbf{X}_A = \tilde{\mathbf{X}} \mathbf{A}_1 \bullet \mathbf{B}_1 \tilde{\mathbf{X}},$$

where

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_B \\ - \end{bmatrix} = \begin{bmatrix} \mathbf{X}_A & | & - \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{X}} & | & - \\ - & & - \end{bmatrix} \quad 8$$

Table 2

Variables, $(\alpha + \beta)n^2$, and constraints, $\gamma n + \delta n^2$, in the model discretized in A with and without compression.

Instance	No compression				Compressed				$(n - n_B)!$
	α	β	γ	δ	α	β	γ	δ	
esc32a	1	2.94	2	3.72	0.78	2.72	1.8	3.49	7!
esc32b	1	2.50	2	3.25	0.75	2.25	1.8	2.99	8!
esc32c	1	2.72	2	3.31	0.59	2.31	1.6	2.89	13!
esc32d	1	2.06	2	2.62	0.56	1.62	1.6	2.17	14!
esc32e	1	1.34	2	1.62	0.28	0.62	1.3	0.88	23!
esc32g	1	1.41	2	1.62	0.22	0.62	1.2	0.82	25!
esc32h	1	2.62	2	3.22	0.59	2.22	1.6	2.80	13!
esc64a	1	1.39	2	1.73	0.34	0.73	1.3	1.07	42!
esc128	1	1.26	2	1.50	0.24	0.50	1.2	0.74	97!
tai64c	1	1.20	2	1.41	0.20	0.41	1.2	0.60	51!
tai256c	1	1.36	2	1.72	0.36	0.72	1.4	1.08	164!

Table 3

Variables, $(\alpha + \beta)n^2$, and constraints, $\gamma n + \delta n^2$, in the model discretized in B with and without compression.

Instance	No compression				Compressed				$(n - n_B)!$
	α	β	γ	δ	α	β	γ	δ	
esc32a	1	3.91	2	4.69	0.78	3.91	1.8	4.68	7!
esc32b	1	3.75	2	4.50	0.75	3.75	1.8	4.49	8!
esc32c	1	2.97	2	3.56	0.59	2.97	1.6	3.55	13!
esc32d	1	2.81	2	3.38	0.56	2.81	1.6	3.36	14!
esc32e	1	1.41	2	1.69	0.28	1.41	1.3	1.67	23!
esc32g	1	1.09	2	1.31	0.22	1.09	1.2	1.29	25!
esc32h	1	2.97	2	3.56	0.59	2.97	1.6	3.55	13!
esc64a	1	2.06	2	2.41	0.34	2.06	1.3	2.40	42!
esc128	1	1.70	2	1.94	0.24	1.70	1.2	1.93	97!
tai64c	1	3.05	2	3.25	0.20	3.05	1.2	3.24	51!
tai256c	1	15.09	2	15.45	0.36	15.09	1.4	15.45	164!

and

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & | & \mathbf{0} \\ \hline \mathbf{0} & | & \mathbf{0} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{B}_1 & | & \mathbf{0} \\ \hline \mathbf{0} & | & \mathbf{0} \end{bmatrix}$$

$\tilde{\mathbf{X}}$, $\mathbf{A}_1$ and $\mathbf{B}_1$ are $n_B \times n_A$, $n_A \times n_A$ and $n_B \times n_B$ matrices respectively. If both **A** and **B** are compressed an additional constraint to Eqs. (21)–(23) is needed to ensure the remaining number of permutations. This constraint is written as follows:

$$\sum_{i=1}^{n_B} \sum_{j=1}^{n_A} x_{ij} \geq n_B + n_A - n. \quad (24)$$

The permutations eliminated after compressing in both **A** and **B** are:

$$\min \left\{ \frac{n!(n - n_B)!}{n_A!}, \frac{n!(n - n_A)!}{n_B!} \right\} \quad (25)$$

while the number of permutations left is:

$$\max \left\{ \frac{n_A!}{(n - n_B)!}, \frac{n_B!}{(n - n_A)!} \right\}. \quad (26)$$

From Eq. (26) we find that only one permutation remains when $n_B + n_A - n = 0$. In this case the optimal value of the QAP is 0 and all $x_{ij} = 0$ in the compressed model. $n_B + n_A - n = 0$ when $n_B = n_A = \frac{n}{2}$.

Symmetries In addition to compressing the matrices, optional symmetry strategies can be utilized. Considering the objective function written as $\mathbf{X}\mathbf{A}\mathbf{B}\mathbf{X}$ we observe that when both **A** and **B** are

Table 4

Solution times (in seconds) for solving some instances to optimality utilizing symmetry (S) and compression (C) (when possible). DLR indicate the solution time when solving without S and C.

Instance	Disc in	DLR	S	C	C + S
nug12	A	59	20		
scr12	B	9.6	3.6		
chr12a	B	1.6	0.8		
tai12a	A	246	157		
rou12	B	1187	216		
esc16a	B	11.4	10.0	9.9	7.1
esc16b	B	158	>1200	71	>1200
esc16c	B	286	264	168	250

Table 5

Solution results when solving the instances esc32a, esc32c, esc32d, esc64a and tai64c from the QAPLIB

Instance	BKS	old LB	DLR	Nodes	Time (s)
esc32a	130	103	130	110365472	1618580
esc32c	642	616	642	1473284	24365
esc32d	200	191	200	2922791	36256
esc64a	116	98	116	62365	16370
tai64c	1855928	1855928	1855928	385103	182983

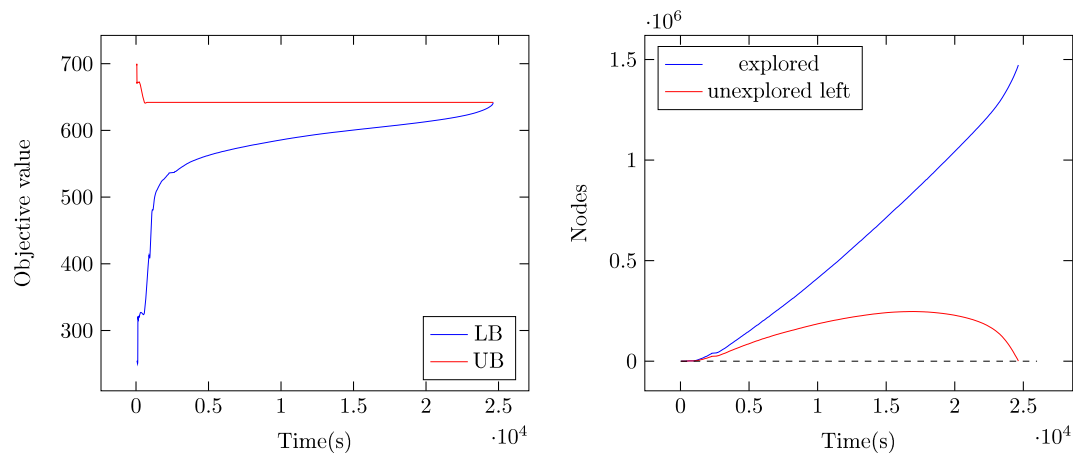


Fig. 2. Solution progress for the instance esc32c.

Table 6

Optimal permutation vectors obtained with the DLR model for the instances esc32a, esc32b, esc32c, esc32d, esc64a and tai64c in the QAPLIB.

Instance	
esc32a	11, 3, 7, 23, 19, 27, *, 14, 20, 17, 28, 9, 12, 4, 8, 2, 26, 24, 32, *, 22, 25, 6, 18, 29, 10, 30, *, *, *, *, *, * $S \in \{1, 5, 13, 15, 16, 21, 31\}$
esc32c	15, 12, *, 13, 22, 8, 24, 23, 20, 19, 4, 2, 1, 7, 6, 3, 5, 18, 17, 21, *, *, *, *, *, *, *, *, *, * $S \in \{9, 10, 11, 14, 16, 25, 26, 27, 28, 29, 30, 31, 32\}$
esc32d	18, 29, 10, 2, 25, 32, 22, 20, 24, 17, 30, 9, 1, 26, 31, 21, 19, 23, *, *, *, *, *, *, *, *, *, * $S \in \{3, 4, 5, 6, 7, 8, 11, 12, 13, 14, 15, 16, 27, 28\}$
esc64a	18, 29, 10, 2, 25, 32, 22, 20, 24, 17, 30, 9, 1, 26, 31, 21, 19, 23, *, *, *, *, *, *, *, *, *, * $S \in \{3, 4, 5, 6, 7, 8, 11, 12, 13, 14, 15, 16, 27, 28\}$
tai64c	1, 15, 32, 35, 61, 41, 29, 18, 12, 45, 63, 59, 47, *, ..., * $S \in \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14, 16, 17, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 30, 31, 33, 34, 36, 37, 38, 39, 40, 42, 43, 44, 46, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 60, 62, 64\}$

symmetric we can write either $\mathbf{XA} \bullet \mathbf{BX} = \mathbf{A} \bullet \mathbf{X}^T \mathbf{BX} = \tilde{\mathbf{A}} \bullet \mathbf{X}^T \mathbf{BX}$ or $\mathbf{XA} \bullet \mathbf{BX} = \mathbf{B} \bullet \mathbf{XAX}^T = \mathbf{B} \bullet \mathbf{XAX}^T$ where $\tilde{\mathbf{A}} = 2 \cdot \mathbf{U}_A - \mathbf{D}_A$ and $\mathbf{B} = 2 \cdot \mathbf{U}_B - \mathbf{D}_B$. $\mathbf{U}_B$ and $\mathbf{U}_A$ are the upper (or lower) triangular parts and $\mathbf{D}_A$ and $\mathbf{D}_B$ the diagonal parts of the corresponding matrices. This reduces the number of variables and constraints. In Table 4 solution times for solving some smaller instances from the QAPLIB to optimality utilizing symmetry as well as compression (when possible) are given. When the symmetry option has been used it has been applied to the matrix not being discretized. Only the $\mathbf{A}$ matrix in the three last instances can be compressed. Table 4 shows that utilizing symmetry and compression reduces the solution time in many cases.

Some of the QAPLIB instances have persymmetric (symmetric w.r.t. the counter diagonal), centrosymmetric (symmetric w.r.t. the matrix center) or bisymmetric (symmetric and centrosymmetric) matrices. These properties can be utilized as well. For example, a centrosymmetric matrix $\mathbf{A}$ has the property $\mathbf{MA} = \mathbf{AM}$, where $\mathbf{M}$ is a counter-diagonal matrix, with elements equal to one in the counter diagonal and all entries off the counter-diagonal equal to zero. In this case two permutation vectors, $\mathbf{p}_1 = \mathbf{Xq}$ and $\mathbf{p}_2 = \mathbf{XMq}$ result in the same value of the objective function. In order to break this symmetry the first row of the permutation matrix can be partitioned to a left and a right half, $(\mathbf{x}_L | \mathbf{x}_R)$. Letting $\mathbf{x}_L = \mathbf{0}$ the symmetry will be broken and only one of the solutions will remain valid. In a similar way if the $\mathbf{B}$ matrix is centrosymmetric, a permutation vector $\mathbf{p}_3 = \mathbf{MXq}$ result in the same value of the objective function as with $\mathbf{p}_1$. This symmetry can be broken, for example, by including zero elements in the upper half of the first column in the permutation matrix. Breaking the centrosymmetry eliminates multiple solutions. There are several other symmetry options as well and several of them have been introduced in the literature (Anstreicher, 2003; de Klerk and Sotirov, 2010). However, we conclude this section only by stating that symmetries in the matrices have not been considered in greater detail in this paper, but we will introduce some alternative options for special structures of the $\mathbf{A}$ and $\mathbf{B}$ matrices in a forthcoming paper.

5. Results

By using the model formulation presented in this paper we have solved some previously unsolved instances from the QAPLIB to proven optimality. In Table 5 the results obtained with the DLR model are given. The first two columns show the best known solutions and the best lower bounds for the instances, reported in the QAPLIB.² The verified optimal solution obtained with the discrete linear reformulated MIP model, the total number of nodes visited and the solution time in seconds, as reported by Gurobi, are given in the last three columns. All instances in Table 5 have been discretized in $\mathbf{B}$ and compression was utilized in the $n = 64$ instances. All instances have been solved on a single PC with an Intel i7 4-core 2.8 GigaHertz processor and 6 GigaByte RAM, except for the esc64a instance where a PC with an Intel i7 6-core 3.2 GigaHertz processor was used. As a MILP solver Gurobi (4.0.1) with default parameter settings was used. The node files were stored on the hard drive, thus reducing the amount of RAM needed to solve the models. In Fig. 2, to the left, the best integer and best node solutions versus the solution time are illustrated when solving the instance esc32c. To the right, in the same figure, the number of nodes explored and nodes remaining versus the solution time are shown. The permutation vectors for the problems solved to optimality are given in Table 6, in the Appendix. Surprisingly the esc64a instance was solved in about 4.5 hours and tai64c in about 51 hours. Even though the problems with $n = 64$ are huge we were still able to solve them to proven optimality. This is both because of the efficiency of the discrete reformulation as well as the reduction of the permutation matrix by using compression. When solving the esc32a, esc32c and esc32d instances compression did, however, not result in a shorter solution time.

² The instance tai64c is a so-called grey pattern problem having a simplified QAP structure. The problem has previously been solved to proven optimality by Dreznier (2006), using a special purpose algorithm utilizing the simpler structure. The problem is included in Table 5, since we have solved it in the QAP form and it was reported as unsolved in the QAPLIB.

Table 5 indicates that the solution times are quite moderate when comparing them to previously solved instances of $n \geq 30$ (Anstreicher et al., 2002; Adams et al., 2007). A good branching strategy is crucial for any large MILP problem as indicated in Achterberg et al. (2005). In all the test runs we, however, used default settings for Gurobi. A drawback with some recent methods is the amount of RAM required as indicated in (Adams et al., 2007; Hahn et al., 2008). With the DLR-model presented in this paper the amount of RAM is, on the other hand, not a crucial issue. Even on the largest problem from the QAPLIB, tai256c, the DLR can be written as a MILP model with $0.36n^2$ binary and $0.72n^2$ continuous variables as well as $1.4n + 1.08n^2$ constraints.

Having the opportunity to test the efficiency of new solution approaches on instances from the QAPLIB is a great resource. Results obtained with several different solvers on the same instances can easily be found in the QAPLIB from well documented references, thanks to the maintainers Hahn and Anjos (2002). The unsolved and recently solved problems in the library are however very challenging. This restricts the possibilities of doing comprehensive numerical comparisons, because of the CPU-time required to solve these problems. Already the CPU-time for a single instance might end up lasting weeks, even when solving it using parallel computation (Anstreicher et al., 2002). This practical limitation has also, in our case, limited the amount of comparisons done in this paper.

6. Conclusions

The exact discrete linear reformulation, DLR, presented in this paper is shown to be compact and from numerical experiments we have found that it gives very good results, especially on sparse instances with few unique elements per row. We have presented solutions to never before solved instances in the QAPLIB. The run times for problems solved with the DLR-model are reasonably short (taking into account the combinatorial complexity of the instances) and the RAM memory required is not a crucial issue with this model. However, we do not think that any QAP can be solved more efficiently with the given reformulation than with some other approach but we are of course happy that some of the instances in the QAPLIB that had never been solved to optimality before could be solved with this formulation. Default settings in the MILP solver Gurobi were used when conducting the experiments, whilst the node files were stored on a hard drive. A good branching strategy could probably still speed up the solution times. Symmetry options were not considered in greater detail in this paper, but will be considered in a forthcoming paper.

Acknowledgements

The financial support from the Academy of Finland project 127992 is gratefully acknowledged. We also want to acknowledge professor Peter Hahn at the University of Pennsylvania as well as Miguel Anjos at École Polytechnique Montreal for their kind support in connection with the QAPLIB.

Appendix A. Optimal permutations

Below are the optimal permutation vectors for the instances esc32a, esc32b, esc32c, esc32d, esc64a and tai64c obtained with the DLR QAP formulation. For each instance, the elements indicated by * can be replaced with any permutation of the elements in the corresponding set S.

References

- Achterberg, T., Koch, T., Martin, A., 2005. Branching rules revisited. *Operations Research Letters* 33 (1), 42–54.
- Adams, W.P., Guignard, M., Hahn, P.M., Hightower, W.L., 2007. A level-2 reformulation-linearization technique bound for the quadratic assignment problem. *European Journal of Operational Research* 180 (3), 983–996.
- Anstreicher, K., Brixius, N., Goux, J.-P., Linderoth, J., 2002. Solving large quadratic assignment problems on computational grids. *Mathematical Programming* 91, 563–588.
- Anstreicher, K.M., 2003. Recent advances in the solution of quadratic assignment problems. *Mathematical Programming, Ser. B* 97, 27–42.
- Anstreicher, K.M., Brixius, N.W., 2001. A new bound for the quadratic assignment problem based on convex quadratic programming. *Mathematical Programming* 89, 341–357. doi:10.1007/PL00011402 <http://dx.doi.org/10.1007/PL00011402>.
- Burkard, R., Çela, D.E., Karisch, S., Rendl, F., 1997. Qaplib – a quadratic assignment problem library. *Journal of Global Optimization* 10, 391–403.
- de Klerk, E., Sotirov, R., 2010. Exploiting group symmetry in semidefinite programming relaxations of the quadratic assignment problem. *Mathematical Programming* 122, 225–246. doi:10.1007/s10107-008-0246-5 <http://dx.doi.org/10.1007/s10107-008-0246-5>.
- Drezner, Z., 2006. Finding a cluster of points and the grey pattern quadratic assignment problem. *OR Spectrum* 28, 417–436. doi:10.1007/s00291-005-0010-7 <http://dx.doi.org/10.1007/s00291-005-0010-7>.
- Drezner, Z., Hahn, P., Taillard, E., 2005. Recent advances for the quadratic assignment problem with special emphasis on instances that are difficult for meta-heuristic methods. *Annals of Operations Research* 139, 65–94. doi:10.1007/s10479-005-3444-z <http://dx.doi.org/10.1007/s10479-005-3444-z>.
- Gounaris, C.E., Misener, R., Floudas, C.A., 2009. Computational comparison of piecewise linear relaxations for pooling problems. *Industrial & Engineering Chemistry Research* 48 (12), 5742–5766 <http://pubs.acs.org/doi/abs/10.1021/ie8016048>.
- Hahn, P., Anjos, M., 2002. Qaplib – a quadratic assignment problem library online. University of Pennsylvania, School of Engineering and Applied Science, <http://www.seas.upenn.edu/qaplib/>.
- Hahn, P.M., Zhu, Y.-R., Guignard, M., Hightower, W.L., 2008. A level-3 reformulation-linearization technique based bound for the quadratic assignment problem. *Optimization-Online* 09. URL <http://www.optimization-online.org/DB_HTML/2008/09/2094.html>.
- Kaufman, L., Broeckx, F., 1978. An algorithm for the quadratic assignment problem using bender's decomposition. *European Journal of Operational Research* 2 (3), 207–211.
- Koopmans, T., Beckmann, M., 1957. Assignment problems and location of economic activities. *Econometrica* 25, 53–76.
- Loiola, E., Abreu, N., Boaventura-Netto, P., Hahn, P., Querido, T., 2007. A survey for the quadratic assignment problem. *European Journal of Operational Research* 176 (2), 657–690.
- Peng, J., Mittelman, H., Li, X., 2010. A new relaxation framework for quadratic assignment problems based on matrix splitting. *Mathematical Programming Computation* 2 (1), 59–77 <http://dx.doi.org/10.1007/s12532-010-0012-6>.
- Taillard, E., 1991. Robust taboo search for the quadratic assignment problem. *Parallel Computing* 17 (4–5), 443–455.
- Zhang, H., Beltran-Royo, C., Ma, L., 2010. Solving the quadratic assignment problem by means of general purpose mixed integer linear programming solvers. *Statistics and Operations Research, Rey Juan Carlos University (URJC), C/Tulipán s/n, 28933, M=stoles (Madrid), Spain 04, 1.*

A Generalization of the Classical α BB Convex Underestimation via Diagonal and Nondiagonal Quadratic Terms

A. Skjäl · T. Westerlund · R. Misener ·
C.A. Floudas

Received: 17 January 2012 / Accepted: 14 March 2012 / Published online: 30 March 2012
© Springer Science+Business Media, LLC 2012

Abstract The classical α BB method determines univariate quadratic perturbations that convexify twice continuously differentiable functions. This paper generalizes α BB to additionally consider nondiagonal elements in the perturbation Hessian matrix. These correspond to bilinear terms in the underestimators, where previously all nonlinear terms were separable quadratic terms. An interval extension of Gerschgorin's circle theorem guarantees convexity of the underestimator. It is shown that underestimation parameters which are optimal, in the sense that the maximal underestimation error is minimized, can be obtained by solving a linear optimization model.

Theoretical results are presented regarding the instantiation of the nondiagonal underestimator that minimizes the maximum error. Two special cases are analyzed to convey an intuitive understanding of that optimally-selected convexifier. Illustrative examples that convey the practical advantage of these new α BB underestimators are presented.

Keywords Global optimization · α BB · Convex relaxations · Nonconvex optimization

A. Skjäl · T. Westerlund
Center of Excellence in Optimization and Systems Engineering, Åbo Akademi University, Turku,
Finland

R. Misener · C.A. Floudas (✉)
Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ
08544-5263, USA
e-mail: floudas@titan.princeton.edu

1 Introduction

Deterministic global optimization algorithms aim at finding a minimum in a nonconvex optimization model to a given precision (ε -global) within finite time. The three major components of a typical deterministic global minimization algorithm are:

- Generating and solving convex relaxations of the nonconvex nonlinear program (NLP) that rigorously guarantee lower bounds on the global solution.
- Finding feasible solutions via local optimization or heuristics to bound the global solution from above. Termination of the algorithm is based on comparisons between the best feasible solution and the lower bounds generated through convex relaxations.
- Dividing and conquering the feasible set (*e.g.*, through branch-and-bound or branch-and-cut) to generate a sequence of convex relaxations converging to the global optimum.

Although all three components of deterministic global optimization methods are challenging in their own right and have been addressed in a wide array of excellent research articles, review papers, and textbooks [1–10], the most challenging of the triptych is the generation of tight convex underestimators [5].

Expression-specific underestimators have been designed for an array of functional forms, but there still exists a need for underestimators that effectively address generic, multivariate terms [5]. Complementary techniques for addressing generic functional forms include: (i) the α BB methodology that generates convexifying quadratic or exponential relaxations on expression aggregates via an interval Hessian matrix [1, 2, 11–15], (ii) directed acyclic graph representations that partition expressions into their component parts by introducing auxiliary variables [3, 10, 16–19] and (iii) systematic subgradient propagation that recursively relaxes expressions [20–22]. Because the methodological trade-offs between α BB and tree-based representations produce complementary convergence behavior [23], the most generic global optimization tools are hybrid algorithms that opportunistically exploit the tightest relaxation at each node of a global optimization search tree [24].

Motivated to improve the relative tightness of the classical α BB convexification method, we generalize the classical α BB from exclusively considering univariate quadratic perturbations to additionally incorporating bilinear terms. An interval extension of Gerschgorin's circle theorem is employed to both guarantee convexity and design a linear optimization model that selects the nondiagonal underestimator with the smallest possible maximum underestimation error.

This paper begins in Sect. 2 by situating the present work within the context of previous research, and continues in Sect. 3 with a brief discussion of interval arithmetic. Section 4 presents novel theoretical contributions by discussing both the classical and nondiagonal α BB underestimators. Section 5 proposes minimizing the maximum underestimation error as a metric for optimally selecting underestimator parameters and proves a number of theorems, which explicate when and how the new underestimators improve the classical choice. The present work and a previous investigation into nondiagonal extensions of the α BB convexification methodology [13] are compared in Sect. 6. Section 7 presents some detailed examples and a table of test problems. Conclusions are drawn and discussed in Sect. 8.

2 Literature Review

We consider nonlinear optimization problems (NLPs) of the type:

$$\begin{aligned} \min \quad & f_0(\mathbf{x}), \\ \text{s.t.} \quad & f_m(\mathbf{x}) \leq 0, \quad m \in \{1, 2, \dots, M\}, \\ & x_i^L \leq x_i \leq x_i^U, \quad i \in \{1, 2, \dots, n\}, \end{aligned}$$

where the functions $f_0, f_1, \dots, f_M$ are twice continuously differentiable. Equality constraints are handled as follows:

$$g(\mathbf{x}) = 0 \quad \Leftrightarrow \quad g(\mathbf{x}) \leq 0 \wedge -g(\mathbf{x}) \leq 0.$$

Finite variable bounds (x_i^L, x_i^U) must be directly supplied by the user or indirectly inferred from the constraints. It is worth noting that prior work immediately extends this NLP formulation to nonconvex mixed-integer nonlinear programs (nonconvex MINLP) [25, 26].

The α BB convexification methodology originated with Maranas and Floudas [27], Liu and Floudas [14] defining and characterizing a parameter α_i that, when multiplied by a univariate quadratic, permitted any nonconvex function with a Lipschitz continuous gradient to be expressed as the difference of two convex functions [28]. Further characterizations of the α BB underestimators with respect to their convergence behavior and the necessary and sufficient conditions for their existence have been investigated by Zlobec [29–31] and Bompadre and Mitsos [23].

A global optimization framework exploiting the α BB convexification methodology was first described in Androulakis et al. [32]. This framework was expanded to include a method to calculate α based on stability characteristics of polynomial families [33]. A comprehensive exposition of theory, rigorous approaches to calculate the α parameters, and implementation of the α BB-based global optimization algorithm, was presented by Adjiman et al. [1, 2]. The important extension to nonuniform shift matrices was also described in these papers. The application of α BB to mixed-integer nonlinear optimization problems was studied by Adjiman et al. [25, 26]. Two variants were described, SMIN- α BB for problems with special structure, and GMIN- α BB for more general mixed-integer problems [25, 26]. An algorithm using the same quadratic underestimation properties as α BB has been more recently introduced by Chang et al. [34].

The reader is referred to the textbook of Floudas [5] for the most complete description of the classical α BB and extensive discussion on its applications. It is important to mention that the classical α BB underestimators have been used in a variety of applications including: system identification, parameter estimation, optimal control, reactor network synthesis, semiinfinite programming, bilevel programming, molecular structure prediction, peptide and protein folding, design under uncertainty, solving constrained systems of equations, phase equilibrium calculations, dynamic systems, and solving subproblems within an augmented Lagrangian framework [15, 24, 35–59].

Univariate exponential perturbations were introduced as an alternative to quadratic perturbations by Akrotirianakis and Floudas [12]. The quadratic functions $-\alpha_i(x_i^U - x_i)(x_i - x_i^L)$ in the classical α BB were replaced by functions of the form:

$$-(1 - e^{\gamma_i(x_i - x_i^L)})(1 - e^{\gamma_i(x_i^U - x_i)}),$$

where the parameters $\gamma_1, \gamma_2, \dots, \gamma_n$ are determined through the α values via an iterative scheme. This exponential alternative to the classical α BB is denoted as γ BB. Akrotirianakis and Floudas [11] discussed their computational experience for the γ BB underestimators on box-constrained problems. Floudas and Kreinovich [60, 61] proved that the classical and the exponential α BB underestimators are the only two functional forms possessing both shift-invariance and sign-invariance, and are therefore the two natural choices for any α BB-like convexification methodology.

Classical α BB underestimators have also served as building blocks for more advanced underestimators. Piecewise α BB, a variant where different α values are calculated for individual subdomains and the underestimators splined together, was proposed by Meyer and Floudas [62]. Gounaris and Floudas [63] presented an alternative piecewise underestimation scheme in which the univariate case converges to the convex envelope as the partitioning is refined. Gounaris and Floudas [64] further extended the methodology from the univariate to the multivariate case using two-dimensional projections. Finally, the possibility of nondiagonal α BB underestimators was first explored by Akrotirianakis et al. [13]. We detail the similarities between our underestimators and those of Akrotirianakis et al. [13] in Sect. 6.

3 Interval Arithmetic Definitions

Determining convexifying perturbations via any variant of the α BB methodology requires bounds on the Hessian over the domain of interest. Interval arithmetic places bounds on expressions by recursively overestimating each of the participating functions and operators. The reader is directed to Neumaier [65] and Wolfe [66] for an excellent treatment of interval arithmetic and an extensive survey on the application of interval arithmetic to equation solving and optimization, respectively.

In this section, the basic interval arithmetic notation used to derive the results is introduced [65, Chap. 1]. An interval h' is defined as

$$h' \equiv [\underline{h}, \bar{h}] := \{h \in \mathbb{R} \mid \underline{h} \leq h \leq \bar{h}\},$$

where the lower and upper bound of the interval are $\underline{h}$ and $\bar{h}$, respectively. The midpoint and radius of an interval are defined [65]:

$$\text{mid}(h') := \frac{\underline{h} + \bar{h}}{2}, \quad \text{rad}(h') := \frac{\bar{h} - \underline{h}}{2},$$

and the absolute value function is generalized to intervals as [65]:

$$|h'| = \max\{|\underline{h}|, |\bar{h}|\}.$$

Interval arithmetic is computationally inexpensive, but the overestimate may be large if the same variable appears multiple times in an expression. This overrelaxation of the bounds arises from independent treatment of multiply-occurring variables. For instance, $x \cdot x$, $x \in [-2, 3]$ is overestimated as $[-6, 9]$, but $[0, 9]$ would be the optimal estimate.

An alternative, optimization-based methodology for determining these interval estimates is to minimize and maximize the expression under the bound constraints and possibly additional linear and convex constraints from the problem. These subproblems provide tighter bounds than interval arithmetic. A decreased number of nodes in the branch-and-bound tree may justify the increased computational effort needed to find better estimates. Note that interval propagation and constraint satisfaction methods also offer tighter bounds than natural interval extensions, but are less computationally demanding than optimality-based bounding [67, 68].

4 Theoretical Justification of the Nondiagonal Underestimators

Let $H(\mathbf{x})$ be the Hessian of $f(\mathbf{x})$ at $\mathbf{x} = [x_1, \dots, x_n]$:

$$H(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{\partial^2 f(\mathbf{x})}{\partial x_{n-1} \partial x_n} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_{n-1}} & \frac{\partial^2 f(\mathbf{x})}{\partial x_n^2} \end{bmatrix} \\ = \begin{bmatrix} h_{1,1}(\mathbf{x}) & h_{1,2}(\mathbf{x}) & \cdots & h_{1,n}(\mathbf{x}) \\ h_{2,1}(\mathbf{x}) & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & h_{n-1,n}(\mathbf{x}) \\ h_{n,1}(\mathbf{x}) & \cdots & h_{n,n-1}(\mathbf{x}) & h_{n,n}(\mathbf{x}) \end{bmatrix},$$

and let $[H]$ be an interval Hessian with elements:

$$h'_{ij} \equiv [\underline{h}_{ij}, \overline{h}_{ij}] := \{h_{ij} \in \mathbb{R} \mid \underline{h}_{ij} \leq h_{ij} \leq \overline{h}_{ij}\}.$$

The elements $h'_{ij} \in [H]$ rigorously bound all Hessian values of f on the hyperrectangle

$$D = [x_1^L, x_1^U] \times \cdots \times [x_n^L, x_n^U]$$

and can be viewed as a superset of all possible values for $h_{ij}(\mathbf{x})$.

The perturbation Hessian H^P , which consists of diagonal and nondiagonal elements, is represented as

$$H^P = \begin{bmatrix} 2\alpha_1 & \beta_{1,2} & \cdots & \beta_{1,n} \\ \beta_{2,1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \beta_{n-1,n} \\ \beta_{n,1} & \cdots & \beta_{n,n-1} & 2\alpha_n \end{bmatrix}.$$

Two conditions are necessary for the underestimator; (i) it must be convex and (ii) it must underestimate f on D . The classical α BB method satisfies these conditions by adding the perturbation:

$$-\sum_i \alpha_i (x_i^U - x_i)(x_i - x_i^L),$$

where the nonnegative parameters $\alpha_1, \alpha_2, \dots, \alpha_n$ are sufficiently large to guarantee that any instance of $[H] + \text{diag}(2\alpha)$ is positive semidefinite.

Adjiman et al. [2] presented several methods to determine the α values. Our proposed generalization, which builds upon the best-performing approach of Adjiman et al. [2], is based on interval arithmetic and Gerschgorin's circle theorem [69]. No scaling is used (*i.e.*, $d = 1$ in [2, Theorem 3.13]).

To guarantee convexity of the nondiagonal underestimator, Theorem 4.1 proves that it suffices to demand that the selected perturbation satisfies:

$$\underline{h}_{ii} + 2\alpha_i - \sum_{j \neq i} |h'_{ij} + \beta_{ij}| \geq 0, \quad \forall i, \quad (1)$$

where $\underline{h}_{ii}$ is the lower bound of the diagonal interval Hessian element, h'_{ij} are the nondiagonal elements of the interval Hessian, and α_i and β_{ij} are the diagonal and nondiagonal elements of the perturbation Hessian H^P , respectively. We write summations informally (*e.g.*, $\sum_i$, $\sum_{j>i}$, $\sum_{j \neq i}$, where i and j range over $\{1, 2, \dots, n\}$). Note that, by the definition of the absolute value function introduced in Sect. 3:

$$|h'_{ij} + \beta_{ij}| = \max\{|\underline{h}_{ij} + \beta_{ij}|, |\overline{h}_{ij} + \beta_{ij}|\}.$$

Theorem 4.1 *If*

$$\underline{h}_{ii} + 2\alpha_i - \sum_{j \neq i} |h'_{ij} + \beta_{ij}| \geq 0$$

for all i , then any instance of $[H] + H^P$ is positive semidefinite.

Proof Assume that $\underline{h}_{ii} + 2\alpha_i - \sum_{j \neq i} |h'_{ij} + \beta_{ij}| \geq 0$ for all i . By Gerschgorin's circle theorem, the least eigenvalue of $[H] + H^P$ is bounded from below by

$$\min_i \left\{ \underline{h}_{ii} + 2\alpha_i - \sum_{j \neq i} |h'_{ij} + \beta_{ij}| \right\}.$$

This shows that any instance of $[H] + H^P$ is positive semidefinite. $\square$

Prior to describing a measure to optimize, we need to choose how the perturbation Hessian H^P will be realized as a function. An important property of the classical α BB underestimator is that it matches the original function at the vertices of the hyperrectangle D . A joint diagonal/nondiagonal α BB underestimator is proposed such that this property is preserved. After starting with

$$-\sum_i \alpha_i (x_i^U - x_i)(x_i - x_i^L) + \sum_i \sum_{j>i} \beta_{ij} x_i x_j,$$

a piecewise-linear function (PLF) is added to make the perturbation equal to zero at all vertices of D . To achieve this, we study an individual term $\beta_{ij} x_i x_j$. There are two different triangulations of $[x_i^L, x_i^U] \times [x_j^L, x_j^U]$ (i.e., add one diagonal or the other). On each of these triangulations there is a unique PLF that matches the vertices of D to zero. These functions are related to the concave and convex envelopes of bilinear terms [20]. On one triangulation, the unique PLF is minus the concave envelope of $\beta_{ij} x_i x_j$, on the other it is minus the convex envelope. The applicability of the PLF can be verified via the observation that bilinear terms on box domains have vertex polyhedral envelopes [70, 71]. The convex envelope will not be used here to avoid the binary variables required to model disjunctions.

Finally, the following perturbation function $P(\mathbf{x}, \mathbf{z})$ is introduced:

$$P(\mathbf{x}, \mathbf{z}) = -\sum_i \alpha_i (x_i^U - x_i)(x_i - x_i^L) + \sum_i \sum_{j>i} (\beta_{ij} x_i x_j + |\beta_{ij}| z_{ij}), \quad (2)$$

where z_{ij} is an additional auxiliary variable representing the negative of the concave envelope of $\text{sgn}(\beta_{ij}) x_i x_j$:

$$\begin{aligned} z_{ij} &= \begin{cases} -\min\left(\frac{x_i x_j^L + x_i^U x_j - x_i^U x_j^L}{x_i x_j^U + x_i^L x_j - x_i^L x_j^U}\right), & \text{if } \beta_{ij} > 0, \\ \max\left(\frac{x_i x_j^L + x_i^L x_j - x_i^L x_j^L}{x_i x_j^U + x_i^U x_j - x_i^U x_j^U}\right), & \text{if } \beta_{ij} < 0, \end{cases} \\ &= \begin{cases} \max\left(\frac{-x_i x_j^L - x_i^U x_j + x_i^U x_j^L}{-x_i x_j^U - x_i^L x_j + x_i^L x_j^U}\right), & \text{if } \beta_{ij} > 0, \\ \max\left(\frac{x_i x_j^L + x_i^L x_j - x_i^L x_j^L}{x_i x_j^U + x_i^U x_j - x_i^U x_j^U}\right), & \text{if } \beta_{ij} < 0, \end{cases} \end{aligned} \quad (3)$$

The contribution from the off-diagonal β_{ij} is shown in Fig. 1. Observe in Eq. (2) that the perturbation element $H_{ij}^P = \beta_{ij}$ is derived from the contribution $\beta_{ij} x_i x_j$, and that the affine correction $|\beta_{ij}| z_{ij}$ matches the perturbation function $P(\mathbf{x}, \mathbf{z})$ to zero at the corners of D without affecting the convexity.

Theorem 4.2 Let $P(\mathbf{x}, \mathbf{z})$ and z_{ij} be given by (2) and (3), respectively. Then $f(\mathbf{x}) + P(\mathbf{x}, \mathbf{z})$ underestimates $f(\mathbf{x})$ on the whole box domain D .

Proof We will show that $P(\mathbf{x}, \mathbf{z})$ is composed of nonpositive terms. The univariate terms

$$-\alpha_i (x_i^U - x_i)(x_i - x_i^L)$$

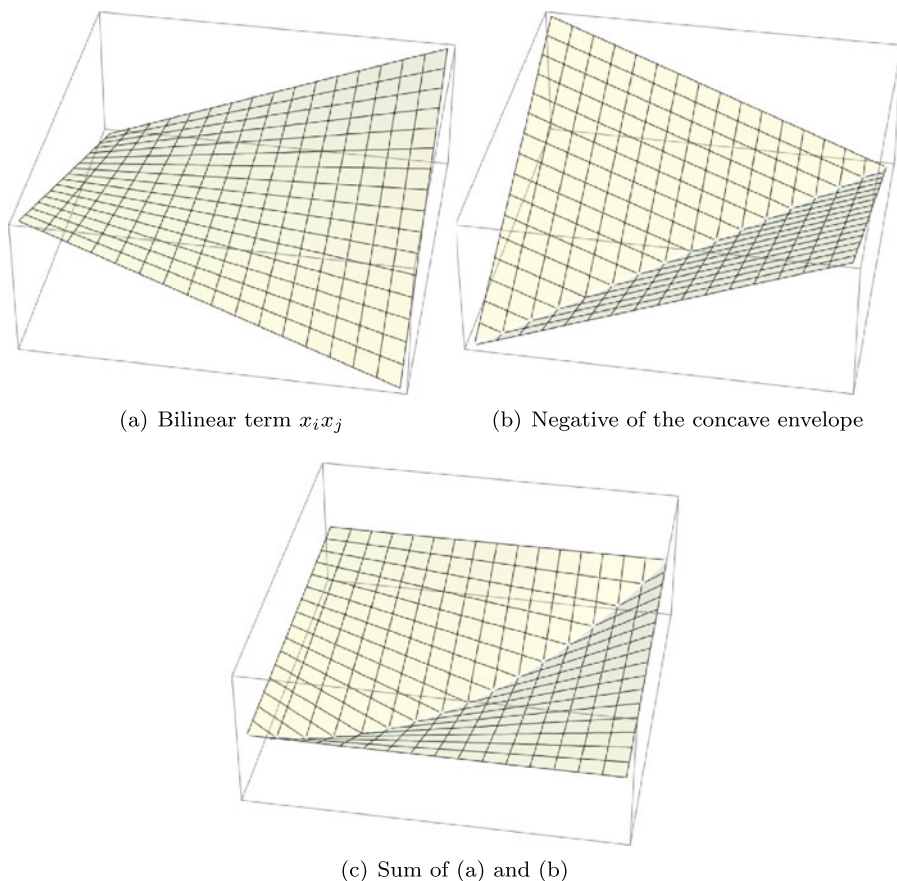


Fig. 1 A bilinear term (a), the negative of its concave envelope (b), and their sum (c). The contribution is 0 along the edges and most negative in the midpoint

are clearly nonpositive on D since $\alpha_i \geq 0$, $x_i^U - x_i \geq 0$, and $x_i - x_i^L \geq 0$. Assuming that Eq. (3) is satisfied, the expression:

$$\underbrace{\beta_{ij} x_i x_j}_{\varphi} + \underbrace{|\beta_{ij}| z_{ij}}_{-\hat{\varphi}}$$

is composed of a function minus its concave envelope. Since $\varphi \leq \hat{\varphi}$ by definition, $\varphi - \hat{\varphi} \leq 0$, and thus also the nondiagonal contributions are nonpositive. $\square$

In general, piecewise-linear functions need to be modeled using disjunctions with binary or special-ordered set (SOS) variables. However, the right-hand side of Eq. (3) is a convex expression and the equality can be relaxed to an inequality $\geq$. This inequality can in turn be modeled by

$$\begin{aligned} z_{ij} &\geq (-x_i x_j^L - x_i^U x_j + x_i^U x_j^L) \\ z_{ij} &\geq (-x_i x_j^U - x_i^L x_j + x_i^L x_j^U) \end{aligned} \quad (4)$$

if $\beta_{ij} > 0$, and:

$$\begin{aligned} z_{ij} &\geq (x_i x_j^L + x_i^L x_j - x_i^L x_j^L) \\ z_{ij} &\geq (x_i x_j^U + x_i^U x_j - x_i^U x_j^U) \end{aligned} \quad (5)$$

if $\beta_{ij} < 0$.

Using the perturbation function $P(\mathbf{x}, \mathbf{z})$ from Eq. (2), together with the constraints (4) and (5), will change neither (i) the minimum objective value of the underestimator $f_0(\mathbf{x}) + P_0(\mathbf{x}, \mathbf{z})$ nor (ii) the feasible set of the convexified constraint $f_m(\mathbf{x}) + P_m(\mathbf{x}, \mathbf{z})$ projected onto D , as compared to using (2) and (3) together. To state this equivalence as a theorem, we introduce the notation:

$$\phi_{ij}^m(\mathbf{x}) := \begin{cases} \max(-x_i x_j^L - x_i^U x_j + x_i^U x_j^L, -x_i x_j^U - x_i^L x_j + x_i^L x_j^U), & \text{if } \beta_{ij}^m > 0 \\ \max(x_i x_j^L + x_i^L x_j - x_i^L x_j^L, x_i x_j^U + x_i^U x_j - x_i^U x_j^U), & \text{if } \beta_{ij}^m < 0. \end{cases}$$

The index m ranges over the constraints, $m = 1, 2, \dots, M$, and the objective, $m = 0$.

Theorem 4.3 *The two problems:*

$$\begin{aligned} \min \quad & F_{P1} = f_0(\mathbf{x}) - \sum_i \alpha_i^0 (x_i^U - x_i) (x_i - x_i^L) + \sum_i \sum_{j>i} (\beta_{ij}^0 x_i x_j + |\beta_{ij}^0| z_{ij}^0), \\ \text{s.t.} \quad & f_m(\mathbf{x}) - \sum_i \alpha_i^m (x_i^U - x_i) (x_i - x_i^L) \\ & + \sum_i \sum_{j>i} (\beta_{ij}^m x_i x_j + |\beta_{ij}^m| z_{ij}^m) \leq 0, \quad \forall m : m \geq 1, \\ & z_{ij}^m = \phi_{ij}^m(\mathbf{x}), \quad \forall m, i, j : j > i, \\ & \mathbf{x} \in X \end{aligned} \quad (6)$$

and

$$\begin{aligned} \min \quad & F_{P2} = f_0(\mathbf{x}) - \sum_i \alpha_i^0 (x_i^U - x_i) (x_i - x_i^L) + \sum_i \sum_{j>i} (\beta_{ij}^0 x_i x_j + |\beta_{ij}^0| z_{ij}^0), \\ \text{s.t.} \quad & f_m(\mathbf{x}) - \sum_i \alpha_i^m (x_i^U - x_i) (x_i - x_i^L) \\ & + \sum_i \sum_{j>i} (\beta_{ij}^m x_i x_j + |\beta_{ij}^m| z_{ij}^m) \leq 0, \quad \forall m : m \geq 1, \\ & z_{ij}^m \geq \phi_{ij}^m(\mathbf{x}), \quad \forall m, i, j : j > i, \\ & \mathbf{x} \in X \end{aligned} \quad (7)$$

have the same optimal objective value. Furthermore, projections of the feasible sets onto X are the same.

Proof Let $\mathbf{z}$ be a vector consisting of all the variables z_{ij}^0, z_{ij}^m and let ϕ be a vector function composed of all the functions ϕ_{ij}^0, ϕ_{ij}^m .

We start with the feasible set projections. Define:

$$S_1 = \{\mathbf{x} \in X : (\mathbf{x}, \mathbf{z}) \text{ is P1-feasible for some vector } \mathbf{z}\}$$

$$S_2 = \{\mathbf{x} \in X : (\mathbf{x}, \mathbf{z}) \text{ is P2-feasible for some vector } \mathbf{z}\}.$$

Clearly, $S_1 \subseteq S_2$. We will show that the converse also holds. Assume that $\tilde{\mathbf{x}} \in S_2$. Then there exists a vector $\tilde{\mathbf{z}} \geq \phi(\tilde{\mathbf{x}})$ such that $(\tilde{\mathbf{x}}, \tilde{\mathbf{z}})$ is P2-feasible. The point $(\tilde{\mathbf{x}}, \phi(\tilde{\mathbf{x}}))$ is P1-feasible since, for all constraints, $m = 1, 2, \dots, M$,

$$\begin{aligned} f_m(\tilde{\mathbf{x}}) - \sum_i \alpha_i^m (x_i^U - \tilde{x}_i)(\tilde{x}_i - x_i^L) + \sum_i \sum_{j>i} (\beta_{ij}^m \tilde{x}_i \tilde{x}_j + |\beta_{ij}^m| \phi_{ij}^m(\tilde{\mathbf{x}})) \\ \leq f_m(\tilde{\mathbf{x}}) - \sum_i \alpha_i^m (x_i^U - \tilde{x}_i)(\tilde{x}_i - x_i^L) + \sum_i \sum_{j>i} (\beta_{ij}^m \tilde{x}_i \tilde{x}_j + |\beta_{ij}^m| \tilde{z}_{ij}^m) \leq 0. \end{aligned}$$

The first inequality holds because of the property $\tilde{\mathbf{z}} \geq \phi(\tilde{\mathbf{x}})$, and the second inequality follows from the assumption that $(\tilde{\mathbf{x}}, \tilde{\mathbf{z}})$ is P2-feasible. Thus, $\tilde{\mathbf{x}} \in S_1$, $S_2 \subseteq S_1$, and we infer that $S_1 = S_2$.

The sequel proves that the optimal objective values, denoted F_{P1}^* and F_{P2}^* , are the same. Since 7 is a relaxation of 6, it follows that $F_{P1}^* \geq F_{P2}^*$. To show that the converse holds, let $(\mathbf{x}^*, \mathbf{z}^*)$ be an optimal point for 7. The point $(\mathbf{x}^*, \phi(\mathbf{x}^*))$ is 6-feasible (as shown in the first part of the proof) and it has the objective value:

$$\begin{aligned} F_{P1}(\mathbf{x}^*, \phi(\mathbf{x}^*)) \\ = f_0(\mathbf{x}^*) - \sum_i \alpha_i^0 (x_i^U - x_i^*)(x_i^* - x_i^L) + \sum_i \sum_{j>i} (\beta_{ij}^0 x_i^* x_j^* + |\beta_{ij}^0| \phi_{ij}^0(\mathbf{x}^*)) \\ \leq f_0(\mathbf{x}^*) - \sum_i \alpha_i^0 (x_i^U - x_i^*)(x_i^* - x_i^L) + \sum_i \sum_{j>i} (\beta_{ij}^0 x_i^* x_j^* + |\beta_{ij}^0| z_{ij}^{*0}) \\ = F_{P2}^*. \end{aligned}$$

The inequality follows from $\mathbf{z}^* \geq \phi(\mathbf{z}^*)$ (which is part of 7-feasibility). Now $F_{P1}^* \leq F_{P1}(\mathbf{x}^*, \phi(\mathbf{x}^*)) \leq F_{P2}^*$ and we conclude that $F_{P1}^* = F_{P2}^*$. $\square$

It would appear that applying the nondiagonal underestimators to the objective function and M constraints would demand up to $(M+1)\frac{n(n-1)}{2}$ additional variables z_{ij}^m and $(M+1)n(n-1)$ additional linear constraints. However, the upper bound on the number of additional variables and constraints beyond those required by the classical α BB methodology is independent of the number of underestimated constraints.

Based on

$$\operatorname{sgn}(\beta_{ij}^m)x_i x_j = \begin{cases} x_i x_j, & \text{if } \beta_{ij}^m > 0, \\ 0, & \text{if } \beta_{ij}^m = 0, \\ -x_i x_j, & \text{if } \beta_{ij}^m < 0, \end{cases}$$

observe that, for fixed (i, j) , there are at most two different corrections needed, one for positive β_{ij}^m and one for negative β_{ij}^m . Introducing the additional variables z_{ij}^+ and z_{ij}^- , the perturbation for one constraint (or the objective function) is

$$P_m(\mathbf{x}, \mathbf{z}) = - \sum_i \alpha_i^m (x_i^U - x_i)(x_i - x_i^L) \\ + \sum_i \sum_{j>i} \left(\beta_{ij}^m x_i x_j + |\beta_{ij}^m| \begin{cases} z_{ij}^+, & \text{if } \beta_{ij}^m > 0 \\ 0, & \text{if } \beta_{ij}^m = 0 \\ z_{ij}^-, & \text{if } \beta_{ij}^m < 0 \end{cases} \right).$$

The constraints

$$\begin{cases} z_{ij}^+ \geq (-x_i x_j^L - x_i^U x_j + x_i^U x_j^L) \\ z_{ij}^+ \geq (-x_i x_j^U - x_i^L x_j + x_i^L x_j^U) \end{cases} \quad \begin{cases} z_{ij}^- \geq (x_i x_j^L + x_i^L x_j - x_i^L x_j^L) \\ z_{ij}^- \geq (x_i x_j^U + x_i^U x_j - x_i^U x_j^U) \end{cases}$$

are added as needed. There will be at most $n(n-1)$ new z_{ij}^+ , z_{ij}^- variables and $2n(n-1)$ new linear constraints.

5 Selection of Nondiagonal Convex Underestimators

In this section, an expression for the maximum underestimation error is derived, and its minimization is outlined.

Theorem 5.1 *An underestimator $P_m(\mathbf{x}, \mathbf{z})$, realized according to (2) and (3), with $\alpha \geq \mathbf{0}$, will have the largest perturbation error in the centerpoint of D , $\mathbf{x}^M := (\mathbf{x}^L + \mathbf{x}^U)/2$.*

Proof The following shows that the perturbation is largest at the centerpoint for every individual contribution in (2). It is clearly true for the diagonal contributions $-\alpha_i(x_i^U - x_i)(x_i - x_i^L)$ [1]. For the nondiagonal contributions, the proof is through case analysis. The concave envelope triangulates the rectangle $[x_i^L, x_i^U] \times [x_j^L, x_j^U]$ along one diagonal or the other depending on the sign of β_{ij} . On both triangles, the contribution will be most negative in $\mathbf{x}^M$. We present the case with positive β_{ij} and the triangle with vertices in (x_i^L, x_j^L) , (x_i^U, x_j^U) , (x_i^L, x_j^U) . The other cases are analogous.

The relevant expression in this triangle is

$$g(x_i, x_j) = \beta_{ij}(x_i x_j - x_i x_j^U - x_i^L x_j + x_i^L x_j^U).$$

The only critical point of g is at $(x_i, x_j) = (x_i^L, x_j^U)$, and thus the minimal value is not found in the interior of the triangle. Along the edges $x_i = x_i^L$ and $x_j = x_j^U$, the expression g is identically 0. Along the diagonal, it will take the values:

$$\begin{aligned} g^{\text{diag}}(\lambda) &= g(\lambda x_i^L + (1-\lambda)x_i^U, \lambda x_j^L + (1-\lambda)x_j^U) \\ &= \lambda(\lambda-1)\beta_{ij}(x_i^U - x_i^L)(x_j^U - x_j^L) \\ &= \left[\left(\lambda - \frac{1}{2} \right)^2 - \frac{1}{4} \right] \beta_{ij}(x_i^U - x_i^L)(x_j^U - x_j^L). \end{aligned}$$

The minimum of g^{diag} is found at $\lambda = 1/2$ (i.e., the centerpoint). □

Corollary 5.1 *The maximum underestimation error (positive sign) when realizing a perturbation with $\alpha \geq 0$, according to (2) and (3), is*

$$\sum_i \alpha_i \left(\frac{x_i^U - x_i^L}{2} \right)^2 + \sum_i \sum_{j>i} |\beta_{ij}| \frac{(x_i^U - x_i^L)(x_j^U - x_j^L)}{4}. \quad (8)$$

Proof By Theorem 5.1, the maximum underestimation error is realized at the centerpoint. To see that Eq. (8) represents the absolute value of the perturbation in the centerpoint, note that, for univariate α_i contributions, it was shown that the expression is $\alpha_i (\frac{x_i^U - x_i^L}{2})^2$ [1]. The centerpoint value for a β_{ij} contribution turns out to be

$$-|\beta_{ij}| \frac{(x_i^U - x_i^L)(x_j^U - x_j^L)}{4},$$

whether $\beta_{ij} > 0$ or $\beta_{ij} < 0$. Using $g^{\text{diag}}(\lambda)$ as in the previous proof, we show the case when β_{ij} is positive. (The perturbation is continuous, so it is sufficient to verify the expression for one of the triangles.)

$$\begin{aligned} g^{\text{diag}}\left(\frac{1}{2}\right) &= \left[\left(\frac{1}{2} - \frac{1}{2} \right)^2 - \frac{1}{4} \right] \beta_{ij}(x_i^U - x_i^L)(x_j^U - x_j^L) \\ &= -\frac{1}{4} \beta_{ij}(x_i^U - x_i^L)(x_j^U - x_j^L) = -|\beta_{ij}| \frac{(x_i^U - x_i^L)(x_j^U - x_j^L)}{4}. \quad \square \end{aligned}$$

To optimize the choice of underestimator, the objective (8) is minimized under constraint (1), positivity of the α_i 's, and symmetry of the Hessian. This is denoted as (9):

$$\begin{aligned}
\min_{\alpha, \beta} \quad & F_{NLP} = \sum_i \frac{\alpha_i}{4} (x_i^U - x_i^L)^2 + \sum_i \sum_{j>i} \frac{|\beta_{ij}|}{4} (x_i^U - x_i^L)(x_j^U - x_j^L), \\
\text{s.t.} \quad & \underline{h}_{ii} + 2\alpha_i - \sum_{j \neq i} |h'_{ij} + \beta_{ij}| \geq 0, \quad \forall i, \\
& \alpha_i \geq 0, \quad \forall i, \\
& \beta_{ij} = \beta_{ji}, \quad \forall i, j : j > i.
\end{aligned} \tag{9}$$

Because the diagonal solution (*i.e.*, $\beta = \mathbf{0}$) is a feasible point of (9), the maximum perturbation error of the nondiagonal α BB underestimator will be at least as good as the classical α BB underestimator.

The question arises whether $\alpha_i \geq 0$ is needed or whether a negative value on some α_i could give a better underestimator. Negative α values are associated with some difficulties. First, $-\alpha_i(x_i^U - x_i^L)(x_i^U - x_i^L) \geq 0$, $\mathbf{x} \in D$ for $\alpha_i < 0$. This positive contribution cannot be totally countered by the other perturbations. A negative constant needs to be added so as to meet the underestimation property, but adding a negative constant violates the corner-matching property. Secondly (assuming that a negative constant is added), for negative α values, we have no simple explicit expression for the maximum underestimation error. Replacing α_i by $|\alpha_i|$ in (8) provides an upper bound on the underestimation error, but the optimal solution for this new objective would not contain negative α values. Based on these remarks, this paper maintains the restriction $\alpha_i \geq 0$.

The absolute values and the generalized absolute values ($|h'_{ij} + \beta_{ij}|$) are piecewise-linear, convex functions. Auxiliary variables could be introduced to model (9) as a piecewise-linear problem with the same optimal (α, β) . However, by first proving some results about (9), a simple linear optimization problem without auxiliary variables emerges.

The optimization problem (9) provides insight as to how the nondiagonal quadratic terms β_{ij} improve the classical α BB underestimation. Using the Gerschgorin method, the classical underestimator parameters are [2]:

$$\alpha_i = \frac{1}{2} \max \left(0, -\underline{h}_{ii} + \sum_{j \neq i} |h'_{ij}| \right), \tag{10}$$

while the solution to (9) will satisfy:

$$2\alpha_i = \max \left(0, -\underline{h}_{ii} + \sum_{j \neq i} |h'_{ij} + \beta_{ij}| \right). \tag{11}$$

Two observations are worth noting. First, choosing a smaller α_i than in the classical α BB method is possible only if $|h'_{ij} + \beta_{ij}| < |h'_{ij}|$ for some i (*i.e.*, β_{ij} makes $h'_{ij} + \beta_{ij}$ more centered around zero than h'_{ij} by itself). For instance, adding 0.6 to the interval $[-1.5, 0.3]$ will center the result around zero, $[-0.9, 0.9]$, making the absolute value of the interval smaller.

Secondly, there is no possible gain from choosing β_{ij} such that $|h'_{ij} + \beta_{ij}| > |h'_{ij}|$ (*i.e.*, making the interval $h'_{ij} + \beta_{ij}$ less centered than h'_{ij}). This will only set harder

restrictions on α_i and α_j and give a higher objective value than $\beta_{ij} = 0$. Theorems 5.2 and 5.3 present results for two special cases.

Theorem 5.2 *If all off-diagonal elements of the interval Hessian $[H]$ are centered around 0 ($\underline{h}_{ij} = -\overline{h}_{ij}$ or equivalently $\text{mid}(h'_{ij}) = 0$), then the classical α BB underestimator is the unique optimal solution to (9).*

Proof We will first demonstrate that, for any nondiagonal solution (i.e., $\beta \neq 0$), there exists a diagonal solution with $\beta = 0$ that has a better objective value. This proves that all optimal solutions are diagonal. The second half of this proof shows that the optimal solution is unique and found by the classical α BB method.

Assume that $\text{mid}(h'_{ij}) = 0$ for all i and j , $i \neq j$ and let (α^0, β^0) be a nondiagonal feasible solution to (9) with $\beta^0 \neq 0$ for some (i, j) . The absolute value $|h'_{ij} + \beta_{ij}|$ can be rewritten:

$$\begin{aligned} |h'_{ij} + \beta_{ij}| &= |\text{mid}(h'_{ij} + \beta_{ij})| + \text{rad}(h'_{ij} + \beta_{ij}) \\ &= |\text{mid}(h'_{ij}) + \text{mid}(\beta_{ij})| + \text{rad}(h'_{ij}) + \text{rad}(\beta_{ij}) = |\beta_{ij}| + \text{rad}(h'_{ij}). \end{aligned}$$

The first equality follows from the identity $|x'| = |\text{mid}(x')| + \text{rad}(x')$, [65, p. 22]. The point $(\alpha^0, 0)$ is a feasible solution since

$$\begin{aligned} \underline{h}_{ii} + 2\alpha_i^0 - \sum_{j \neq i} |h'_{ij} + 0| &= \underline{h}_{ii} + 2\alpha_i^0 - \sum_{j \neq i} \text{rad}(h'_{ij}) \\ &\geq \underline{h}_{ii} + 2\alpha_i^0 - \sum_{j \neq i} (|\beta_{ij}^0| + \text{rad}(h'_{ij})) \\ &= \underline{h}_{ii} + 2\alpha_i^0 - \sum_{j \neq i} |h'_{ij} + \beta_{ij}^0| \geq 0. \end{aligned}$$

The objective value has improved by dropping the nondiagonal parameters:

$$F_{NLP}(\alpha^0, 0) - F_{NLP}(\alpha^0, \beta^0) = - \sum_i \sum_{j > i} \frac{|\beta_{ij}^0|}{4} (x_i^U - x_i^L)(x_j^U - x_j^L) < 0,$$

with strict inequality because of the assumption $\beta^0 \neq 0$.

Restricting the search to diagonal solutions ($\beta = 0$), (9) simplifies to

$$\begin{aligned} \min_{\alpha} \quad & \sum_i \frac{\alpha_i}{4} (x_i^U - x_i^L)^2, \\ \text{s.t.} \quad & \underline{h}_{ii} + 2\alpha_i - \sum_{j \neq i} |h'_{ij}| \geq 0, \quad \forall i, \\ & \alpha_i \geq 0, \quad \forall i. \end{aligned}$$

This problem is separable and has the classical α BB underestimator as the unique solution:

$$\alpha_i = \frac{1}{2} \max \left(0, -\underline{h}_{ii} + \sum_{j \neq i} |h'_{ij}| \right). \quad \square$$

Theorem 5.3 proves that, under a certain condition, an optimal choice of β is to center all off-diagonal elements $h'_{ij} + \beta_{ij}$ around 0.

Theorem 5.3 *If*

$$\underline{h}_{ii} - \sum_{j \neq i} \text{rad}(h'_{ij}) \leq 0, \quad \forall i \quad (12)$$

then

$$\begin{cases} \alpha_i = -\frac{1}{2}\underline{h}_{ii} + \frac{1}{2} \sum_{j \neq i} \text{rad}(h'_{ij}), & \forall i \\ \beta_{ij} = -\text{mid}(h'_{ij}), & \forall i, j : j \neq i. \end{cases} \quad (13)$$

is an optimal solution to the model (9).

Proof Assume that $\underline{h}_{ii} - \sum_{j \neq i} \text{rad}(h'_{ij}) \leq 0$ for all i . The constraint $\alpha_i \geq 0$ in (9) is redundant because the first constraint gives

$$\begin{aligned} 2\alpha_i &\geq -\underline{h}_{ii} + \sum_{j \neq i} |h'_{ij} + \beta_{ij}| \\ &= -\underline{h}_{ii} + \sum_{j \neq i} \underbrace{\text{rad}(h'_{ij} + \beta_{ij})}_{=\text{rad}(h'_{ij})} + \sum_{j \neq i} \underbrace{|\text{mid}(h'_{ij} + \beta_{ij})|}_{\geq 0} \\ &\geq -\underline{h}_{ii} + \sum_{j \neq i} \text{rad}(h'_{ij}) \geq 0. \end{aligned}$$

The equality follows from the identity $|x| = |\text{mid}(x)| + \text{rad}(x)$ [65, p. 22]. The last inequality holds by the assumption in Eq. (12). The inequality constraint:

$$\underline{h}_{ii} + 2\alpha_i - \sum_{j \neq i} |h'_{ij}| \geq 0 \quad (14)$$

will be active for all i , since it is the only nonredundant restriction on α_i and the objective function improves when α_i is decreased. Because constraint (14) is active, the inequality can be replaced with equality.

To show the optimality of suggested solution (13), a variable transformation is introduced. The suggested optimal point is translated to the origin in the new coordinates,

$$\begin{cases} A_i = \alpha_i + \frac{1}{2}\underline{h}_{ii} - \frac{1}{2} \sum_{j \neq i} \text{rad}(h'_{ij}), & \forall i \\ B_{ij} = \beta_{ij} + \text{mid}(h'_{ij}), & \forall i, j : j \neq i. \end{cases}$$

After simplification via basic interval arithmetic, (9) is expressed in the new coordinates as:

$$\begin{aligned} \min_{A, B} \quad F_{AB} &= \sum_i \frac{A_i}{4} (x_i^U - x_i^L)^2 \\ &\quad + \sum_i \sum_{j>i} \frac{|B_{ij} - \text{mid}(h'_{ij})|}{4} (x_i^U - x_i^L)(x_j^U - x_j^L) + C, \\ \text{s.t.} \quad A_i &= \frac{1}{2} \sum_{j \neq i} |B_{ij}|, \quad \forall i, \\ B_{ij} &= B_{ji}, \quad \forall i, j : j > i, \end{aligned}$$

where the constant C is:

$$C = \sum_i \frac{1}{8} \left(-h_{ii} + \sum_{j \neq i} \text{rad}(h'_{ij}) \right) (x_i^U - x_i^L)^2.$$

Clearly, $(A, B) = \mathbf{0}$ is a feasible solution. The sequel shows that this objective is at least as good as an arbitrary feasible point. Assume that $(A^0, B^0) \neq \mathbf{0}$ is a feasible solution. The constraints guarantee that $B_{kl}^0 = B_{lk}^0 \neq 0$ for some (k, l) . Define:

$$\begin{cases} A_i^1 = A_i^0, & \forall i : i \notin \{k, l\} \\ A_k^1 = A_k^0 - \frac{1}{2} |B_{kl}^0| \\ A_l^1 = A_l^0 - \frac{1}{2} |B_{lk}^0| \\ B_{ij}^1 = B_{ij}^0, & \forall i, j : i \neq j, (i, j) \notin \{(k, l), (l, k)\} \\ B_{kl}^1 = B_{lk}^1 = 0. \end{cases}$$

The point (A^1, B^1) is also feasible. Denote $\Delta B = |B_{kl}^0|$. The change in objective value is

$$\begin{aligned} \Delta F_{AB} &= F_{AB}(A^1, B^1) - F_{AB}(A^0, B^0) \\ &= -\frac{\Delta B}{8} (x_k^U - x_k^L)^2 - \frac{\Delta B}{8} (x_l^U - x_l^L)^2 \\ &\quad + \underbrace{(|\text{mid}(h'_{kl})| - |B_{kl}^0 - \text{mid}(h'_{kl})|)}_{\leq |\text{mid}(h'_{kl}) + B_{kl}^0 - \text{mid}(h'_{kl})| = |B_{kl}^0|} \frac{1}{4} (x_k^U - x_k^L)(x_l^U - x_l^L) \\ &\leq -\frac{\Delta B}{8} (x_k^U - x_k^L)^2 - \frac{\Delta B}{8} (x_l^U - x_l^L)^2 + \frac{\Delta B}{4} (x_k^U - x_k^L)(x_l^U - x_l^L) \\ &= -\frac{\Delta B}{8} ((x_k^U - x_k^L) - (x_l^U - x_l^L))^2 \leq 0. \end{aligned}$$

The first inequality is found by applying the reverse triangle inequality. Since $\Delta F_{AB} \leq 0$, the objective value in (A^1, B^1) is at least as good as in the initial

point (A^0, B^0) . Setting $(A^0, B^0) \leftarrow (A^1, B^1)$ and continuing iteratively, the objective monotonically improves until the process terminates at $(A, B) = \mathbf{0}$. Since the initialization point was arbitrary, $(A, B) = \mathbf{0}$ is an optimal solution. $\square$

A stronger assumption implying condition (12) in Theorem 5.3 is $\frac{h_{ii}}{2} < 0, \forall i$. The conditions in Theorems 5.2 and 5.3 can be checked with fewer than $\frac{1}{2}n^2$ and $2n^2$ floating point operations, respectively. Theorem 5.4 forms the basis for the derivation, in Corollary 5.2, of a linear optimization problem (16) to find the optimal nondiagonal underestimator.

Theorem 5.4 Any optimal solution to (9) satisfies

$$\min\{0, -\text{mid}(h'_{ij})\} \leq \beta_{ij} \leq \max\{0, -\text{mid}(h'_{ij})\}, \quad \forall i, j : j \neq i. \quad (15)$$

Proof To prove that a feasible solution violating Eq. (15) cannot be optimal, assume that (α^0, β^0) is a feasible solution that does not satisfy the constraints (15). Let β^0_{kl} be one of the violating variables. The proof is by case analysis, $\text{mid}(h'_{kl}) \geq 0$ or < 0 , and β^0_{kl} is below or above the interval in Eq. (15). The following explicitly addresses the case where $\text{mid}(h'_{kl}) \geq 0$ and $\beta^0_{kl} < -\text{mid}(h'_{kl})$; the three other cases are analogous.

Define

$$\begin{cases} \alpha^1_i = \alpha^0_i, & \forall i \\ \beta^1_{ij} = \beta^0_{ij}, & \forall i, j : i \neq j, (i, j) \notin \{(k, l), (l, k)\} \\ \beta^1_{kl} = \beta^1_{lk} = -\text{mid}(h'_{kl}). \end{cases}$$

The point (α^1, β^1) is also feasible since

$$\begin{aligned} & \underline{h}_{kk} + 2\alpha^1_k - \sum_{j \neq k} |h'_{kj} + \beta^1_{kj}| \\ &= \underline{h}_{kk} + 2\alpha^0_k - \sum_{j \notin \{k, l\}} |h'_{kj} + \beta^0_{kj}| - \underbrace{|h'_{kl} - \text{mid}(h'_{kl})|}_{=\min_{\beta_{kl}} (|h'_{kl} + \beta_{kl}|)} \\ &\geq \underline{h}_{kk} + 2\alpha^0_k - \sum_{j \neq k} |h'_{kj} + \beta^0_{kj}| \geq 0, \end{aligned}$$

and analogously for $i = l$. The objective value is better (lower) in the new point,

$$\begin{aligned} F_{NLP}(\alpha^1, \beta^1) - F_{NLP}(\alpha^0, \beta^0) &= (|\beta^1_{kl}| - |\beta^0_{kl}|)(x^U_k - x^L_k)(x^U_l - x^L_l) \\ &= (\text{mid}(h'_{kl}) + \beta^0_{kl})(x^U_k - x^L_k)(x^U_l - x^L_l) \\ &< (\text{mid}(h'_{kl}) - \text{mid}(h'_{kl}))(x^U_k - x^L_k)(x^U_l - x^L_l) = 0. \end{aligned}$$

Setting $(A^0, B^0) \leftarrow (A^1, B^1)$ and continuing iteratively, all violating variables β^0_{ij} are exchanged for the upper or the lower limit until the process terminates at a point satisfying the constraints (15). The objective value strictly improves in every step

and, therefore, (α^0, β^0) was not optimal. The point (α^0, β^0) was arbitrary, and thus the theorem is proven. $\square$

Corollary 5.2 *Define:*

$$J_i^+ := \{j : j \neq i, \text{mid}(h'_{ij}) \geq 0\}, \quad J_i^- := \{j : j \neq i, \text{mid}(h'_{ij}) < 0\}.$$

The problem formulation (9) has the same optimal solutions as the following linear optimization problem (16):

$$\begin{aligned} \min_{\alpha, \beta} \quad F_{LP} = & \sum_i \frac{\alpha_i}{4} (x_i^U - x_i^L)^2 - \sum_i \sum_{\substack{j > i \\ j \in J_i^+}} \frac{\beta_{ij}}{4} (x_i^U - x_i^L)(x_j^U - x_j^L) \\ & + \sum_i \sum_{\substack{j > i \\ j \in J_i^-}} \frac{\beta_{ij}}{4} (x_i^U - x_i^L)(x_j^U - x_j^L), \\ \text{s.t.} \quad & \underline{h}_{ii} + 2\alpha_i - \sum_{j \in J_i^+} (\overline{h}_{ij} + \beta_{ij}) + \sum_{j \in J_i^-} (\underline{h}_{ij} + \beta_{ij}) \geq 0, \quad \forall i, \\ & \alpha_i \geq 0, \quad \forall i, \\ & \beta_{ij} = \beta_{ji}, \quad \forall i, j : j > i, \\ & \min\{0, -\text{mid}(h'_{ij})\} \leq \beta_{ij} \leq \max\{0, -\text{mid}(h'_{ij})\}, \quad \forall i, j. \end{aligned} \quad (16)$$

Proof The expression $|h'_{ij} + \beta_{ij}|$ is linear on the interval $(\beta_{ij} \in \beta'_{ij})$:

$$\beta'_{ij} \equiv [\min\{0, -\text{mid}(h'_{ij})\}, \max\{0, -\text{mid}(h'_{ij})\}].$$

On the interval β'_{ij} :

$$|h'_{ij} + \beta_{ij}| = \begin{cases} \overline{h}_{ij} + \beta_{ij}, & \text{mid}(h'_{ij}) \geq 0 \\ -\underline{h}_{ij} - \beta_{ij}, & \text{mid}(h'_{ij}) < 0. \end{cases}$$

Since all optimal solutions, according to Theorem 5.4, are found in interval β'_{ij} , we can disaggregate β by sign (i.e., through sets J_i^+ and J_i^-) and substitute linear expressions for the absolute values in (9). Substituting for all i and j , $j \neq i$ yields linear model (16). $\square$

The last theorem of this section relates the improvements on the classical α BB underestimators to the difference in bound interval lengths. An interpretation of the result is that no improvement of the maximum underestimation error is possible if all variable bound intervals are of the same length. In such a situation, the classical α BB underestimator is one of the optimal choices. There may exist a nondiagonal underestimator dominating the classical choice at all points except the centerpoint and the corners, but the maximum error objective cannot discern this.

Theorem 5.5 *If $x_k^U - x_k^L = x_l^U - x_l^L$ for some pair (k, l) , $k \neq l$, then there exists an optimal solution to (9) with $\beta_{kl} = 0$.*

calculate $[H]$ of the nonconvex function f	#by interval arithmetic or other methods
$b_1 \leftarrow \text{True}, b_2 \leftarrow \text{True}$	#flags to recognize special cases
for each pair of integers $(i, j), 1 \leq i < j \leq n$	#check assumption in Theorem 5.2
if $\text{mid}(h'_{ij}) \neq 0$	
$b_1 \leftarrow \text{False}$	
break	
end if	
end for	
for each integer $i, 1 \leq i \leq n$	#check assumption in Theorem 5.3
if $h_{ii} - \sum_{j \neq i} \text{rad}(h'_{ij}) \leq 0$	
$b_2 \leftarrow \text{False}$	
break	
end if	
end for	
if $b_1 = \text{True}$	#using Theorem 5.2
calculate α according to Eq. (10), $\beta \leftarrow 0$	
else if $b_2 = \text{True}$	#using Theorem 5.3
calculate α and β according to Eq. (13)	
else	#the general case, Corollary 5.2
calculate α and β by solving the LP model (16)	
end if	

Fig. 2 Pseudocode demonstrating how Theorems 5.2, 5.3 and Corollary 5.2 calculate the parameters α and β of the generalized α BB underestimator

Proof Assume $x_k^U - x_k^L = x_l^U - x_l^L$ and let (α^0, β^0) be an optimal solution to (9). If $\beta_{kl}^0 = 0$, we are done. Assume $\beta_{kl}^0 \neq 0$ and denote $\Delta\beta = |\beta_{kl}^0|$. Define (α^1, β^1) as

$$\begin{cases} \alpha_i^1 = \alpha_i^0, & \forall i \notin \{k, l\} \\ \alpha_k^1 = \alpha_k^0 + \frac{1}{2}\Delta\beta \\ \alpha_l^1 = \alpha_l^0 + \frac{1}{2}\Delta\beta \\ \beta_{ij}^1 = \beta_{ij}^0, & \forall i, j : i \neq j, (i, j) \notin \{(k, l), (l, k)\} \\ \beta_{kl}^1 = \beta_{lk}^1 = 0. \end{cases}$$

The new point satisfies the constraints in (9); the modifications will not decrease any of the left-hand sides in (1). The difference in objective values is

$$\begin{aligned} F_{NLP}(\alpha^1, \beta^1) - F_{NLP}(\alpha^0, \beta^0) &= \frac{\Delta\beta}{8}(x_k^U - x_k^L)^2 + \frac{\Delta\beta}{8}(x_l^U - x_l^L)^2 - \frac{\Delta\beta}{4}(x_k^U - x_k^L)(x_l^U - x_l^L) \\ &= \frac{\Delta\beta}{8}((x_k^U - x_k^L) - (x_l^U - x_l^L))^2 = 0, \end{aligned}$$

and thus (α^1, β^1) is also optimal. $\square$

The procedure for generating the nondiagonal α BB underestimator is summarized as pseudocode in Fig. 2. The assumptions of Theorems 5.2 and 5.3 are efficiently

checked with computational complexity $\mathcal{O}(n^2)$ and applied, if appropriate. Otherwise, the solution is found by solving the LP model in Corollary 5.2. Therefore, in the general case, the computational cost of the procedure is equivalent to solving the auxiliary LP problem (16) [72].

6 Comparison to Previous Work

Although the derivation is different, the underestimator generated via the methodology discussed in this paper is identical to one introduced by Akrotirianakis et al. [13]. This section outlines the contributions of Akrotirianakis et al. [13], but uses slightly different notation.

Recall the definition in Corollary 5.2:

$$J_i^+ := \{j : j \neq i, \text{mid}(h'_{ij}) \geq 0\}, \quad J_i^- := \{j : j \neq i, \text{mid}(h'_{ij}) < 0\}.$$

The sets B^i and Γ^i introduced by Akrotirianakis et al. [13] are identical to J_i^+ and J_i^- , respectively. The latter are used here to facilitate comparison. Akrotirianakis et al. [13] presented the following function $P_M(\mathbf{x})$ with nonnegative parameters α_i^M , γ_{ij}^- , and γ_{ij}^+ :

$$\begin{aligned} P_M(\mathbf{x}) = & - \sum_i \alpha_i^M (x_i^U - x_i)(x_i - x_i^L) + \sum_i \sum_{\substack{j>i \\ j \in J_i^-}} \gamma_{ij}^- (x_i + x_j)^2 \\ & + \sum_i \sum_{\substack{j>i \\ j \in J_i^+}} \gamma_{ij}^+ (x_i - x_j)^2, \end{aligned} \quad (17)$$

where

$$\begin{aligned} \alpha_i^M = \max & \left[0, \frac{1}{2} \left(-h_{ii} - \sum_{j \in J_i^-} 2\gamma_{ij}^- - \sum_{j \in J_i^+} 2\gamma_{ij}^+ + \sum_{j \in J_i^-} |h'_{ij} + 2\gamma_{ij}^-| \right. \right. \\ & \left. \left. + \sum_{j \in J_i^+} |h'_{ij} - 2\gamma_{ij}^+| \right) \right]. \end{aligned}$$

After a piecewise-linear shift that matches function $P_M(\mathbf{x})$ to zero at the corner points, $P_M(\mathbf{x})$ is a convexifying underestimator to the original nonconvex function $f(\mathbf{x})$ [13].

Observe that Eq. (17) and Eq. (18) have the same nonlinear terms:

$$\begin{aligned} P_{M'}(\mathbf{x}) = & - \sum_i \alpha_i^{M'} (x_i^U - x_i)(x_i - x_i^L) + \sum_i \sum_{\substack{j>i \\ j \in J_i^-}} 2\gamma_{ij}^- x_i x_j \\ & - \sum_i \sum_{\substack{j>i \\ j \in J_i^+}} 2\gamma_{ij}^+ x_i x_j, \end{aligned} \quad (18)$$

when

$$\alpha_i^{M'} = \max \left[0, \frac{1}{2} \left(-\underline{h}_{ii} + \sum_{j \in J_i^-} |h'_{ij} + 2\gamma_{ij}^-| + \sum_{j \in J_i^+} |h'_{ij} - 2\gamma_{ij}^+| \right) \right].$$

Equation (18) is (up to the affine shift $\sum_i \alpha_i^M (x_i^U x_i + x_i^L x_i - x_i^L x_i^U)$) associated with the classical α BB) identical to the $P_m(\mathbf{x}, \mathbf{z})$ definition developed in Sect. 4 under parameter transformation $\beta_{ij} = 2\gamma_{ij}^- - 2\gamma_{ij}^+$.

Because the linear and constant terms in the nondiagonal generalizations are used to match the underestimating and original nonconvex function at the corner points, equivalence of the nonlinear perturbation terms immediately implies equivalence of the underestimators. Analogously to our analysis in Sect. 4, Akrotirianakis et al. [13] observed that the maximum separation distance of the nondiagonal underestimator occurs at the center point, and that minimizing the maximum separation distance is therefore a natural objective for optimizing the choice of nondiagonal underestimator. Akrotirianakis et al. [13] also proposed the following linear optimization model for the calculation of the convexification parameters:

$$\begin{aligned} \min_{\gamma^+, \gamma^-} \quad & \sum_i \left[-\underline{h}_{ii} + \sum_{j \in J_i^+} (\overline{h}_{ij} - 2\gamma_{ij}^+) - \sum_{j \in J_i^-} (\underline{h}_{ij} + 2\gamma_{ij}^-) \right] \frac{(x_i^U - x_i^L)^2}{8} \\ & + \sum_i \sum_{\substack{j > i \\ j \in J_i^+}} \frac{\gamma_{ij}^+}{2} (x_i^U - x_i^L)(x_j^U - x_j^L) \\ & + \sum_i \sum_{\substack{j > i \\ j \in J_i^-}} \frac{\gamma_{ij}^-}{2} (x_i^U - x_i^L)(x_j^U - x_j^L), \\ \text{s.t.} \quad & -\underline{h}_{ii} + \sum_{j \in J_i^+} (\overline{h}_{ij} - 2\gamma_{ij}^+) - \sum_{j \in J_i^-} (\underline{h}_{ij} + 2\gamma_{ij}^-) \geq 0, \quad \forall i, \\ & 0 \leq \gamma_{ij}^+ \leq -\frac{1}{2} \text{mid}(h'_{ij}), \quad \forall i, j : j \in J_i^+, \\ & 0 \leq \gamma_{ij}^- \leq \frac{1}{2} \text{mid}(h'_{ij}), \quad \forall i, j : j \in J_i^-. \end{aligned}$$

By the preceding analysis, the convex underestimators generated by the methodology discussed in this paper are equivalent to those developed by Akrotirianakis et al. [13]. The entirely novel contributions in this paper stem from (i) an alternative development of the underestimators that requires only a single set of parameters β_{ij} , (ii) intuitive, in-depth analysis of the nondiagonal underestimators for interesting special cases, and (iii) comparative studies between the classical and nondiagonal underestimators.

7 Computational Studies of the Classical and Nondiagonal α BB Underestimators

The six-hump camel back function:

$$f(x_1, x_2) = 4x_1^2 - 2.1x_1^4 + \frac{1}{3}x_1^6 + x_1x_2 - 4x_2^2 + 4x_2^4 \quad (19)$$

$$x_1 \in [-1.9, 1.9], \quad x_2 \in [-1.1, 1.1],$$

motivates the behavior of the proposed non-diagonal underestimator. In the α BB methodology [1, 2], terms for which the convex envelope is known (e.g., x_1x_2 , $-4x_2^2$) would be underestimated individually, but for the purposes of demonstration, we underestimate the entire aggregate function.

The Hessian of Eq. (19),

$$H(x_1, x_2) = \begin{bmatrix} 8 - 25.2x_1^2 + 10x_1^4 & 1 \\ 1 & -8 + 48x_2^2 \end{bmatrix},$$

yields alternatively an optimally calculated interval Hessian:

$$[H(x_1, x_2)]_{\text{OPT}} = \begin{bmatrix} [-7.88, 47.35] & 1 \\ 1 & [-8.00, 50.08] \end{bmatrix}$$

and an interval Hessian calculated via interval arithmetic:

$$[H(x_1, x_2)]_{\text{INT}} = \begin{bmatrix} [-82.97, 138.32] & 1 \\ 1 & [-8.00, 50.08] \end{bmatrix}.$$

The optimal calculation of the interval Hessians is the ideal choice, and careful bounding of the interval Hessian is an important component of designing good α BB underestimators [73]. Using the Gerschgorin method for interval matrices to calculate the classical α values yields [2, 69]:

$$H_{\text{OPT}}^P = \begin{bmatrix} 2\alpha_1 & \\ & 2\alpha_2 \end{bmatrix} = \begin{bmatrix} 8.88 & \\ & 9.00 \end{bmatrix}$$

$$H_{\text{INT}}^P = \begin{bmatrix} 2\alpha_1 & \\ & 2\alpha_2 \end{bmatrix} = \begin{bmatrix} 83.97 & \\ & 9.00 \end{bmatrix}.$$

Considering the nondiagonal Hessian, recall that linear model (16), which is solved to obtain the best realization of the nondiagonal perturbation Hessian, minimizes Eq. (8), the perturbation at the midpoint. The effect of this minimization is to push the centers of the off-diagonal interval Hessian elements to zero. In this simple example, the off-diagonal elements of the Hessian are identically equal to one, and hence the solution of linear model (16) yields:

$$H_{\text{OPT}}^P = \begin{bmatrix} 2\alpha_1 & \beta_{12} \\ \beta_{12} & 2\alpha_2 \end{bmatrix} = \begin{bmatrix} 7.88 & -1.00 \\ -1.00 & 8.00 \end{bmatrix}$$

$$H_{\text{INT}}^P = \begin{bmatrix} 2\alpha_1 & \beta_{12} \\ \beta_{12} & 2\alpha_2 \end{bmatrix} = \begin{bmatrix} 82.97 & -1.00 \\ -1.00 & 8.00 \end{bmatrix}.$$

Consider the globally optimal values of the original nonlinear program and the root node relaxations of both the classical and nondiagonal α BB underestimators for both the interval Hessians calculated via global optimization and natural interval extensions, respectively:

	NLP G.O.	Classical α BB		Nondiagonal α BB	
		$[H]_{\text{OPT}}$	$[H]_{\text{INT}}$	$[H]_{\text{OPT}}$	$[H]_{\text{INT}}$
f^*	-1.032	-21.466	-157.01	-21.146	-156.69
x_1^*	-0.090	0.000	0.000	0.000	0.000
x_2^*	0.713	0.000	0.000	0.000	0.000

The small size of the improvements over the classical underestimators is a result of the diagonal elements of the Hessian matrix shifting by a relatively small increment. Considering a factor of 5 multiplying the bilinear term x_1x_2 in Eq. (19), the optimally calculated classical and nondiagonal perturbation Hessians are

$$H_{\text{OPT}}^P = \begin{bmatrix} 2\alpha_1 & \\ & 2\alpha_2 \end{bmatrix} = \begin{bmatrix} 12.88 & \\ & 13.00 \end{bmatrix}$$

$$H_{\text{OPT}}^P = \begin{bmatrix} 2\alpha_1 & \beta_{12} \\ \beta_{12} & 2\alpha_2 \end{bmatrix} = \begin{bmatrix} 7.88 & -5.00 \\ -5.00 & 8.00 \end{bmatrix}.$$

The classical and nondiagonal perturbation Hessians calculated via natural interval extensions are

$$H_{\text{INT}}^P = \begin{bmatrix} 2\alpha_1 & \\ & 2\alpha_2 \end{bmatrix} = \begin{bmatrix} 87.97 & \\ & 13.00 \end{bmatrix}$$

$$H_{\text{INT}}^P = \begin{bmatrix} 2\alpha_1 & \beta_{12} \\ \beta_{12} & 2\alpha_2 \end{bmatrix} = \begin{bmatrix} 82.97 & -5.00 \\ -5.00 & 8.00 \end{bmatrix}.$$

Because considering off-diagonal elements in the perturbation Hessian have a more profound effect when the off-diagonal elements of the interval Hessian are farther from zero, the improvement in the underestimator is more significant. Note that the effect on the lower bound is more significant for the optimally-calculated interval Hessian because the diagonal elements in H_{OPT}^P are on the same order of magnitude as the off-diagonal elements:

	NLP G.O.	Classical α BB		Nondiagonal α BB	
		$[H]_{\text{OPT}}$	$[H]_{\text{INT}}$	$[H]_{\text{OPT}}$	$[H]_{\text{INT}}$
f^*	-6.81	-31.11	-166.65	-29.51	-165.05
x_1^*	-1.86	0.000	0.000	0.000	0.000
x_2^*	1.03	0.000	0.000	0.000	0.000

Figure 3 extends this analysis. Using the six-hump camelback function as a basis, consider a number of possible factors multiplying the bilinear term x_1x_2 . As expected

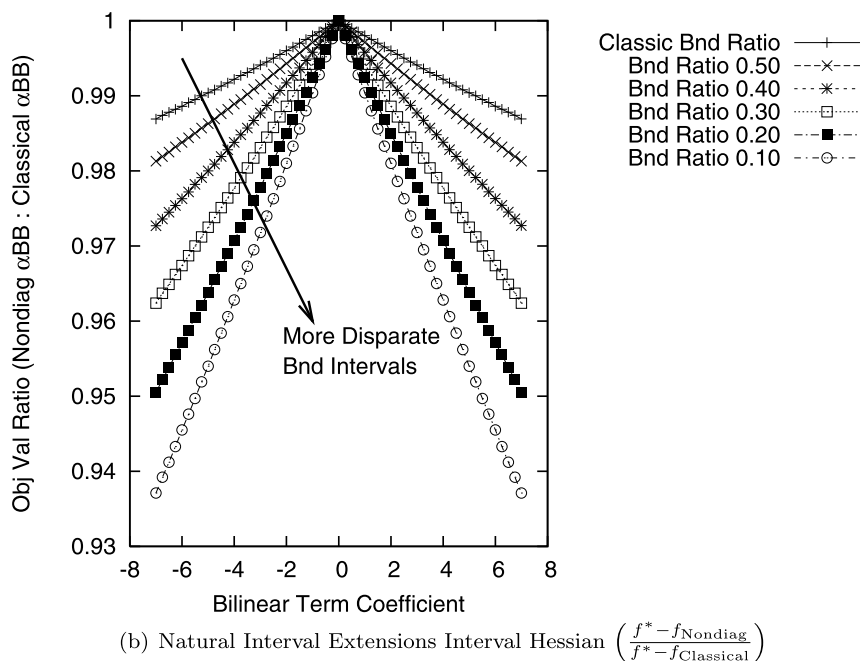
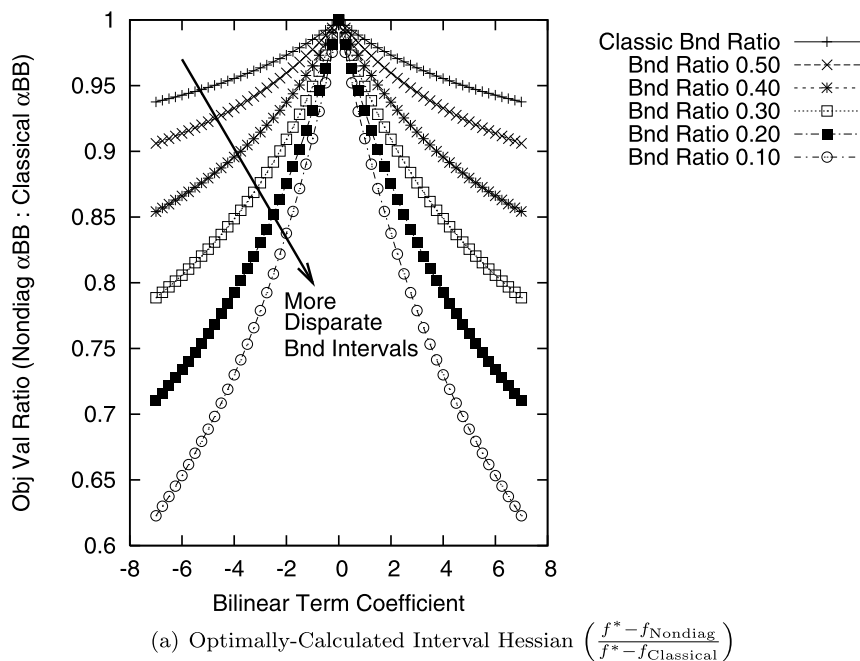


Fig. 3 Comparison of the nondiagonal to classical α BB underestimators for a range of off-diagonal Hessian values and bound ratios. Observe that nondiagonal α BB has a more significant advantage over classical α BB for larger off-diagonal shifts and more disparate bound intervals

Table 1 The stationary point of the convex relaxations generated by the classical and nondiagonal α BB underestimators. The test problems are taken from Floudas et al. [74], Gounaris and Floudas [64]. Case studies highlighted in bold are ones where the nondiagonal α BB underestimator offers at least a ten percent improvement over the classical choice

Function	Domain	INT/OPT	G.O.	Classical LB	Nondiagonal LB
1 $\cos(x_1) \sin(x_2) - \frac{x_1}{x_2^2+1}$	$[-1, 2] \times [-1, 1]$	INT	-2.02	-15.88	-15.85
	$[-1, 2] \times [-1, 1]$	OPT	-2.02	-5.17	-5.13
2 $x_1x_2 + x_1x_2x_3$	$[-1, 1]^3$	INT,OPT	-2	-4	-4
	$[0.5, 1] \times [-1, 1]^2$	INT,OPT	-2	-2.72	-2.46
3 $\begin{cases} (x_1 + 10x_2)^2 + 5(x_3 - x_4)^2 \\ + (x_2 - 2x_3)^4 \\ + 10(x_1 - x_4)^4 \end{cases}$	$[0, 1]^4$	INT,OPT	0	-43.22	-43.08
	$[0, \frac{1}{5}] \times [0, 1] \times [0, 1]^2$	INT,OPT 0	0	-26.68	-21.84
	$[0, 1] \times [0, \frac{1}{5}] \times [0, 1]^2$	INT,OPT 0	0	-43.22	-36.41
4 $0.4x_1^{2/3}x_3^{-2/3} + 0.4x_2^{2/3}x_4^{2/3}$	$[0.1, 10]^4$	INT,OPT	0.04	-778.67	-778.53
	$[9, 10] \times [0.1, 10]^3$	INT,OPT	0.39	-283.02	-263.37
5 $e^{-x_1x_2-x_1x_3-x_2x_3}$	$[0, 3]^3$	INT	2 E-12	-243.00	-243.00
	$[0, 3]^3$	OPT	2 E-12	-54.00	-54.00
	$[0, 1] \times [0, 3]^2$	INT	3 E-7	-95.98	-84.48
	$[0, 1] \times [0, 3]^2$	OPT	3 E-7	-20.80	-16.80
6 $\begin{cases} (2x_1 + x_2 - 3)^2 \\ + (x_1x_2 - 1)^2 \end{cases}$	$[0, 4]^2$	INT,OPT	0	-231.05	-230.90
	$[0, 1] \times [0, 4]$	INT,OPT 0	0	-33.25	-22.00
7 $(1 + x_1 - e^{x_2})^2$	$[0, 1] \times [0, 2]$	INT	0	-14.46	-13.18
	$[0, 1] \times [0, 2]$	OPT	0	-8.24	-6.94

from Theorem 5.2, the classical and nondiagonal α BB match each other when the off-diagonal elements of the interval Hessian are centered at 0, and the nondiagonal α BB underestimator improves with respect to the classical α BB underestimator for larger coefficients. Furthermore, for more disparate bound intervals (*i.e.*, smaller values of $\frac{x_2^U - x_2^L}{x_1^U - x_1^L}$), the nondiagonal α BB underestimators have an increased improvement versus the classical α BB. Theorem 5.5 in Sect. 5 does not prove that the nondiagonal α BB underestimator becomes a better choice as the bound ratio decreases. However, it does imply that the maximum underestimation error cannot be improved significantly for similarly-spaced intervals.

Figure 3 also illustrates that, for this function, accurate calculation of the interval Hessian (*i.e.*, optimization versus natural interval extensions) gives particular advantage to the nondiagonal underestimator because the nondiagonal elements are of the same order of magnitude as the off-diagonal ones.

Test problems from the work of Floudas et al. [74] and Gounaris and Floudas [64] are presented in Table 1. Observe that each of the case studies is analyzed with a variety of domain bounds, because as in the preceding analysis of the six-hump camelback function, the advantage of the nondiagonal α BB underestimators is related to the relative bound interval lengths. Each of the examples in Table 1 are considered with an interval Hessian calculated by both natural interval extensions (*INT*) and optimization (*OPT*).

The penultimate and ultimate columns in Table 1 compare the stationary point of the convex relaxations generated by the classical and nondiagonal α BB underestimators, respectively. These root node relaxations are additionally contrasted with the global optimum (G.O.). Because the maximum perturbation error of the nondiagonal α BB underestimator is at least as small as the classical one, the nondiagonal α BB underestimator dominates in all the test cases. Observe that the case studies where the two methods coincide (*i.e.*, the bounds $[-1, 1]^3$ on Example 3 and the bounds $[0, 3]^3$ on Example 5) are problem instantiations with equal bound intervals. As described by Theorem 5.5, an optimal solution to (9) is the classical α BB underestimator, so our optimization routine merely generated the classical choice. However, there may be many solutions to (9), some of which dominate the classical underestimator at all points except the center and the corners. Those underestimators have been found serendipitously in the instantiations of Examples 3, 4, and 6 with identical bound intervals. Therefore, an interesting extension of this work would design nondiagonal α BB underestimators using a multiobjective, hierarchical approach distinguishing between underestimators with the same maximum underestimation error.

Furthermore, in the six instantiations where the nondiagonal α BB underestimator improves the classical α BB underestimator by more than 10 % (highlighted in bold in Table 1), the bound intervals are particularly disparate.

Considering alternative methodologies for determining bounds on the interval Hessian, see in Table 1 that underestimators determined by optimally-calculated the interval Hessians dominate those determined by natural interval extensions.

Additionally, for the disparate bound interval instantiations of Examples 5 and 7, observe that the nondiagonal α BB underestimator is disproportionately better than the classical one when the interval Hessian is calculated optimally. For example, the (*INT*) case of Example 7 improves the classical α BB underestimator by 9.7 %, but the (*OPT*) case generates an improvement of 32.0 %. This relative advantage of optimally-calculated bounds on the interval Hessian is not universal, but as we observed for the six-hump camelback function, it is related in these examples to the relative magnitudes of the on- and off-diagonal elements of the interval Hessian. Recall that bounds for the interval Hessian are tighter than natural interval extensions, but less expensive than optimally-calculated ones that could be calculated via interval propagation and constraint satisfaction [67, 68].

8 Conclusions

This paper extends the classical α BB convexification methodology from solely considering univariate quadratic terms $\alpha_i x_i^2$ to incorporating nondiagonal bilinear terms $\beta_{ij} x_i x_j$. Theorems are presented relating both the interval Hessian and the variable bounds to the improvements that one can expect from this nondiagonal α BB underestimator. A number of examples illustrating the advantages of the improved underestimators are provided.

Generating these improved underestimators requires only the solution of a linear optimization problem (16). The low computational cost of solving the linear optimization problem can be further reduced to a number of floating point operations

scaling as n^2 under the special cases addressed by Theorems 5.2 and 5.3. Using the nondiagonal underestimators in a constrained nonlinear problem requires at most $n(n-1)$ additional variables and $2n(n-1)$ additional linear constraints compared to the classical α BB convexifiers.

Acknowledgements A.S. gratefully acknowledges financial support from the Center of Excellence in Optimization and Systems Engineering at Åbo Akademi University. C.A.F and R.M. are thankful for support from the National Science Foundation (CBET – 0827907). This material is based upon work supported by the National Science Foundation Graduate Research Fellowship to R.M. under Grant No. DGE-0646086.

References

1. Adjiman, C.S., Androulakis, I.P., Floudas, C.A.: A global optimization method, α BB, for general twice differentiable NLPs-II. Implementation and computational results. *Comput. Chem. Eng.* **22**, 1159–1179 (1998)
2. Adjiman, C.S., Dallwig, S., Floudas, C.A., Neumaier, A.: A global optimization method, α BB, for general twice differentiable NLPs-I. Theoretical advances. *Comput. Chem. Eng.* **22**, 1137–1158 (1998)
3. Belotti, P., Lee, J., Liberti, L., Margot, F., Wächter, A.: Branching and bounds tightening techniques for non-convex MINLP. *Optim. Methods Softw.* **24**, 597–634 (2009)
4. Floudas, C.A.: *Nonlinear and Mixed-Integer Optimization: Fundamentals and Applications*. Oxford University Press, New York (1995)
5. Floudas, C.A.: *Deterministic Global Optimization: Theory, Methods and Applications*. Nonconvex Optimization and Its Applications. Kluwer Academic Publishers, Dordrecht (2000)
6. Floudas, C.A., Gounaris, C.E.: A review of recent advances in global optimization. *J. Glob. Optim.* **45**, 3–38 (2009)
7. Floudas, C.A., Pardalos, P.M.: State-of-the-art in global optimization—computational methods and applications—preface. *J. Glob. Optim.* **7**, 113 (1995)
8. Floudas, C.A., Akrotirianakis, I.G., Caratzoulas, S., Meyer, C.A., Kallrath, J.: Global optimization in the 21st century: advances and challenges. *Comput. Chem. Eng.* **29**, 1185–1202 (2005)
9. Sherali, H.D., Adams, W.P.: *A Reformulation-Linearization Technique for Solving Discrete and Continuous Nonconvex Problems*. Nonconvex Optimization and Its Applications. Kluwer Academic Publishers, Dordrecht (1999)
10. Tawarmalani, M., Sahinidis, N.V.: *Convexification and Global Optimization in Continuous and Mixed-Integer Nonlinear Programming: Theory, Algorithms, Software, and Applications*. Nonconvex Optimization and Its Applications. Kluwer Academic Publishers, Norwell (2002)
11. Akrotirianakis, I.G., Floudas, C.A.: Computational experience with a new class of convex underestimators: Box-constrained NLP problems. *J. Glob. Optim.* **29**, 249–264 (2004)
12. Akrotirianakis, I.G., Floudas, C.A.: A new class of improved convex underestimators for twice continuously differentiable constrained NLPs. *J. Glob. Optim.* **30**, 367–390 (2004)
13. Akrotirianakis, I.G., Meyer, C.A., Floudas, C.A.: The role of the off-diagonal elements of the Hessian matrix in the construction of tight convex underestimators for nonconvex functions. In: *Discovery Through Product and Process Design*. Foundations of Computer-Aided Process Design, pp. 501–504 (2004)
14. Liu, W.B., Floudas, C.A.: A remark on the GOP algorithm for global optimization. *J. Glob. Optim.* **3**, 519–521 (1993)
15. Maranas, C.D., Floudas, C.A.: Finding all solutions of nonlinearly constrained systems of equations. *J. Glob. Optim.* **7**, 143–182 (1995)
16. Liberti, L.: Reformulations in mathematical programming: automatic symmetry detection and exploitation. *Math. Program.* (2010). doi:[10.1007/s10107-010-0351-0](https://doi.org/10.1007/s10107-010-0351-0)
17. Liberti, L., Cafieri, S., Tarissan, F.: Reformulations in mathematical programming: A computational approach. In: Abraham, A., Hassanien, A.E., Siarry, P. (eds.) *Foundations of Computational Intelligence*, vol. 3, pp. 153–234. Springer, New York (2009)
18. Smith, E.M.B., Pantelides, C.C.: A symbolic reformulation/spatial branch-and-bound algorithm for the global optimisation of nonconvex MINLPs. *Comput. Chem. Eng.* **23**(4–5), 457–478 (1999)

19. Tawarmalani, M., Sahinidis, N.V.: A polyhedral branch-and-cut approach to global optimization. *Math. Program.* **103**, 225–249 (2005)
20. McCormick, G.P.: Computability of global solutions to factorable nonconvex programs: Part 1—convex underestimating problems. *Math. Program.* **10**, 147–175 (1976)
21. Mitsos, A., Chachuat, B., Barton, P.I.: McCormick-based relaxations of algorithms. *SIAM J. Optim.* **20**, 573–601 (2009)
22. Scott, J.K., Stuber, M.D., Barton, P.I.: Generalized McCormick relaxations. *J. Glob. Optim.* **51**, 569–606 (2011)
23. Bompadre, A., Mitsos, A.: Convergence rate of McCormick relaxations. *J. Glob. Optim.* **52**, 1–28 (2011)
24. Gatzke, E.P., Tolsma, J.E., Barton, P.I.: Construction of convex relaxations using automated code generation techniques. *Optim. Eng.* **3**, 305–326 (2002)
25. Adjiman, C.S., Androulakis, I.P., Floudas, C.A.: Global optimization of MINLP problems in process synthesis and design. *Comput. Chem. Eng.* **21**(Suppl. S), S445–S450 (1997)
26. Adjiman, C.S., Androulakis, I.P., Floudas, C.A.: Global optimization of mixed-integer nonlinear problems. *AIChE J.* **46**, 1769–1797 (2000)
27. Maranas, C.D., Floudas, C.A.: A global optimization approach for Lennard-Jones microclusters. *J. Chem. Phys.* **97**, 7667–7678 (1992)
28. Zlobec, S.: On the Liu-Floudas convexification of smooth programs. *J. Glob. Optim.* **32**, 401–407 (2005)
29. Zlobec, S.: Characterization of convexifiable functions. *Optimization* **55**, 251–261 (2006)
30. Zlobec, S.: Saddle-point optimality: A look beyond convexity. *J. Glob. Optim.* **29**, 97–112 (2004)
31. Zlobec, S.: On two simple decompositions of Lipschitz functions. *Optimization* **57**, 249–261 (2008)
32. Androulakis, I.P., Maranas, C.D., Floudas, C.A.: α BB: A global optimization method for general constrained nonconvex problems. *J. Glob. Optim.* **7**, 337–363 (1995)
33. Adjiman, C.S., Floudas, C.A.: Rigorous convex underestimators for general twice-differentiable problems. *J. Glob. Optim.* **9**, 23–40 (1996)
34. Chang, M.H., Park, Y.C., Lee, T.Y.: A new global optimization method for univariate constrained twice-differentiable NLP problems. *J. Glob. Optim.* **39**, 79–100 (2007)
35. Birgin, E., Floudas, C.A., Martínez, J.: Global minimization using an augmented Lagrangian method with variable lower-level constraints. *Math. Program.* **125**, 139–162 (2010)
36. Bravo, J.M., Alamo, A.T., Redondo, M.J., Camacho, E.F.: An algorithm for bounded-error identification of nonlinear systems based on DC functions. *Automatica* **44**, 437–444 (2008)
37. Esposito, W.R., Floudas, C.A.: Global optimization in parameter estimation of nonlinear algebraic models via the error-in-variables approach. *Ind. Eng. Chem. Res.* **37**, 1841–1858 (1998)
38. Esposito, W.R., Floudas, C.A.: Deterministic global optimization in nonlinear optimal control problems. *J. Glob. Optim.* **17**, 97–126 (2000)
39. Esposito, W.R., Floudas, C.A.: Global optimization for the parameter estimation of differential-algebraic systems. *Ind. Eng. Chem. Res.* **39**, 1291–1310 (2000)
40. Esposito, W.R., Floudas, C.A.: Deterministic global optimization in isothermal reactor network synthesis. *J. Glob. Optim.* **22**, 59–95 (2002)
41. Floudas, C.A., Stein, O.: The adaptive convexification algorithm: a feasible point method for semi-infinite programming. *SIAM J. Optim.* **18**, 1187–1208 (2007)
42. Floudas, C.A., Gümus, Z.H., Ierapetritou, M.G.: Global optimization in design under uncertainty: feasibility test and flexibility index problems. *Ind. Eng. Chem. Res.* **40**, 4267–4282 (2001)
43. Gümus, Z.H., Floudas, C.A.: Global optimization of nonlinear bilevel programming problems. *J. Glob. Optim.* **20**, 1–31 (2001)
44. Klepeis, J.L., Floudas, C.A.: A comparative study of global minimum energy conformations of hydrated peptides. *J. Comput. Chem.* **20**, 636–654 (1999)
45. Klepeis, J.L., Floudas, C.A.: Free energy calculations for peptides via deterministic global optimization. *J. Chem. Phys.* **110**, 7491–7512 (1999)
46. Klepeis, J.L., Floudas, C.A.: Deterministic global optimization and torsion angle dynamics for molecular structure prediction. *Comput. Chem. Eng.* **24**, 1761–1766 (2000)
47. Klepeis, J.L., Floudas, C.A.: Ab initio tertiary structure prediction of proteins. *J. Glob. Optim.* **25**, 113–140 (2003)
48. Klepeis, J.L., Floudas, C.A.: ASTRO-FOLD: a combinatorial and global optimization framework for ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. *Biophys. J.* **85**, 2119–2146 (2003)

49. Klepeis, J.L., Floudas, C.A., Morikis, D., Lambris, J.D.: Predicting peptide structures using NMR data and deterministic global optimization. *J. Comput. Chem.* **20**, 1354–1370 (1999)
50. Klepeis, J.L., Pieja, M., Floudas, C.A.: A new class of hybrid global optimization algorithms for peptide structure prediction: integrated hybrids. *Comput. Phys. Commun.* **151**, 121–140 (2003)
51. Maranas, C.D., Floudas, C.A.: A deterministic global optimization approach for molecular-structure determination. *J. Chem. Phys.* **100**, 1247–1261 (1994)
52. McDonald, C.M., Floudas, C.A.: Decomposition based and branch and bound global optimization approaches for the phase equilibrium problem. *J. Glob. Optim.* **5**, 205–251 (1994)
53. McDonald, C.M., Floudas, C.A.: Global optimization for the phase and chemical equilibrium problem: application to the NRTL equation. *Comput. Chem. Eng.* **19**, 1111–1141 (1995)
54. McDonald, C.M., Floudas, C.A.: Global optimization for the phase stability problem. *AIChE J.* **41**, 1798–1814 (1995)
55. McDonald, C.M., Floudas, C.A.: Global optimization and analysis for the Gibbs free energy function for the UNIFAC, Wilson, and ASOG equations. *Ind. Eng. Chem. Res.* **34**, 1674–1687 (1995)
56. Mitsos, A., Lemonidis, P., Lee, C.K., Barton, P.I.: Relaxation-based bounds for semi-infinite programs. *SIAM J. Optim.* **19**, 77–113 (2008)
57. Papamichail, I., Adjiman, C.S.: A rigorous global optimization algorithm for problems with ordinary differential equations. *J. Glob. Optim.* **24**, 1–33 (2002)
58. Papamichail, I., Adjiman, C.S.: Proof of convergence for a global optimization algorithm for problems with ordinary differential equations. *J. Glob. Optim.* **33**, 83–107 (2005)
59. Zhu, Y.S., Inoue, K.: Calculation of chemical and phase equilibrium based on stability analysis by QBB algorithm: application to NRTL equation. *Chem. Eng. Sci.* **56**, 6915–6931 (2001)
60. Floudas, C.A., Kreinovich, V.: On the functional form of convex underestimators for twice continuously differentiable functions. *Optim. Lett.* **1**, 187–192 (2007)
61. Floudas, C.A., Kreinovich, V.: Towards optimal techniques for solving global optimization problems: Symmetry-based approach. In: Törn, A., Zilinskas, J. (eds.) *Models and Algorithms for Global Optimization*, pp. 21–42. Springer, New York (2007)
62. Meyer, C.A., Floudas, C.A.: Convex underestimation of twice continuously differentiable functions by piecewise quadratic perturbation: spline α BB underestimators. *J. Glob. Optim.* **32**, 221–258 (2005)
63. Gounaris, C.E., Floudas, C.A.: Tight convex underestimators for C^2 -continuous problems: I. Univariate functions. *J. Glob. Optim.* **42**, 51–67 (2008)
64. Gounaris, C.E., Floudas, C.A.: Tight convex underestimators for C^2 -continuous problems: II. Multivariate functions. *J. Glob. Optim.* **42**, 69–89 (2008)
65. Neumaier, A.: *Interval Methods for Systems of Equations*. Encyclopedia of Mathematics and Its Applications. Cambridge University Press, Cambridge (1990)
66. Wolfe, M.A.: Interval mathematics, algebraic equations and optimization. *J. Comput. Appl. Math.* **124**, 263–280 (2000)
67. Domes, F., Neumaier, A.: Constraint propagation on quadratic constraints. *Constraints* **15**, 404–429 (2010)
68. Granvilliers, L., Benhamou, F.: Algorithm 852: RealPaver: an interval solver using constraint satisfaction techniques. *ACM Trans. Math. Softw.* **32**, 138–156 (2006)
69. Gerschgorin, S.: Über die abgrenzung der eigenwerte einer matrix. *Izv. Akad. Nauk SSSR, Ser. Fiz. Mat.* **6**, 749–754 (1931)
70. Tardella, F.: On a class of functions attaining their maximum at the vertices of a polyhedron. *Discrete Appl. Math.* **22**, 191–195 (1988/89)
71. Tardella, F.: Existence and sum decomposition of vertex polyhedral convex envelopes. *Optim. Lett.* **2**, 363–375 (2008)
72. Vanderbei, R.J.: *Linear Programming: Foundations and Extensions*. International Series in Operations Research & Management Science. Springer, Berlin (2008). ISBN 9780387743875
73. Hertz, D., Adjiman, C.S., Floudas, C.A.: Two results on bounding the roots of interval polynomials. *Comput. Chem. Eng.* **23**, 1333–1339 (1999)
74. Floudas, C.A., Pardalos, P.M., Adjiman, C.S., Esposito, W.R., Günius, Z.H., Harding, S.T., Klepeis, J.L., Meyer, C.A., Schweiger, C.A.: *Handbook of Test Problems in Local and Global Optimization*. Kluwer Academic Publishers, Dordrecht (1999)

A PHYSICALLY MOTIVATED CLASS OF SCATTERING PASSIVE LINEAR SYSTEMS*

OLOF J. STAFFANS[†] AND GEORGE WEISS[‡]

Abstract. We introduce a class of scattering passive linear systems motivated by examples from mathematical physics. The state space of the system is $X = H \oplus E$, where H and E are Hilbert spaces. We also have a Hilbert space E_0 which is dense in E , with continuous embedding, and E'_0 is the dual of E_0 with respect to the pivot space E . The input space is the same as the output space, and it is denoted by U . The semigroup generator has the structure $A = \begin{bmatrix} 0 & -L \\ L^* & G - \frac{1}{2}K^*K \end{bmatrix}$, where $L \in \mathcal{L}(E_0, H)$ and $K \in \mathcal{L}(E_0, U)$ are such that $\begin{bmatrix} L \\ K \end{bmatrix}$, with domain E_0 , is closed as an unbounded operator from E to $H \oplus U$. The operator $G \in \mathcal{L}(E_0, E'_0)$ is such that $\operatorname{Re} \langle Gw_0, w_0 \rangle \leq 0$ for all $w_0 \in E_0$. The observation operator is $C = \begin{bmatrix} 0 & -K \end{bmatrix}$, the control operator is $B = -C^*$, and the output equation is $y = Cx + u = -Kw + u$, where u is the input function, $x = \begin{bmatrix} z \\ w \end{bmatrix}$ is the state trajectory, and y is the corresponding output function. We show that this system is scattering passive (hence, well-posed) and that classical solutions of the system equation $\dot{x} = Ax + Bu$ satisfy $\frac{d}{dt} \|x(t)\|^2 = \|u(t)\|^2 - \|y(t)\|^2 + 2\operatorname{Re} \langle Gw, w \rangle$. Moreover, the dual system satisfies a similar power balance equation. Hence, this system is scattering conservative if and only if $\operatorname{Re} \langle Gw_0, w_0 \rangle = 0$ for all $w_0 \in E_0$. We give two examples involving the beam equation and one with Maxwell's equations.

Key words. scattering passive system, scattering conservative system, system node, Cayley transform, beam equation, Maxwell's equations

AMS subject classifications. 93C20, 93A30, 47A40, 35Q61

DOI. 10.1137/110846403

1. The main results. Given four operators A, B, C, D on appropriate Hilbert spaces, a natural question is whether they determine a scattering passive or conservative (in particular, well-posed) linear system via the equations $\dot{x} = Ax + Bu$, $y = Cx + Du$. This was studied for the first time in Arov and Nudelman [4], using earlier results about discrete-time scattering passive systems and translating those results using the internal Cayley transform. More results about scattering passive systems were derived in Staffans and Weiss [35] (where they were called dissipative systems), and relatively simple necessary and sufficient conditions for a system node to be scattering conservative were provided in Malinen, Staffans, and Weiss [24]. A good overview of these results can be found in the book by Staffans [34], and the connection with impedance passive and conservative systems is studied in Staffans [31, 32, 33].

It is of interest to identify large classes of systems where the operators A, B, C, D have a special structure observed in models of mathematical physics, which implies that the system is scattering passive or conservative. Indeed, if we then find a system with this special structure, we do not have to take the trouble of checking the conditions for scattering passivity or conservativity given in the papers listed earlier. (This kind of checking need not be straightforward.) Such a special class of conservative

*Received by the editors August 31, 2011; accepted for publication (in revised form) August 16, 2012; published electronically October 10, 2012. This research was supported by Israel Science Foundation grant 701/10 and by Academy of Finland grant 130358/09.

<http://www.siam.org/journals/sicon/50-5/84640.html>

[†]Department of Mathematics, Åbo Akademi University, FIN-20500 Åbo, Finland (Olof.Staffans@abo.fi).

[‡]School of Electrical Engineering, Tel Aviv University, Ramat Aviv 69978, Israel (gweiss@eng.tau.ac.il).

systems (“from thin air”) has been introduced in Weiss and Tucsnak [46] and further studied in Tucsnak and Weiss [38] and in Staffans [33]. In this paper we give a larger special class, which includes the systems introduced in [46] and also others. We were led to introduce this class by our failure to fit Maxwell’s equations into the framework of [46]. We illustrate the new theory with two short examples based on the beam equation (only one of which falls within the class of systems treated in [46] and [38]), and we also outline an application to Maxwell’s equations. To keep the length of this paper within reasonable limits we have postponed a more complete treatment of Maxwell’s equations to the follow-up paper [44].

In this paper we consider a linear system Σ whose state space X can be decomposed as $X = H \oplus E$, where H and E are Hilbert spaces. The Hilbert space U is both the input space and the output space of Σ . We identify H , E , and U with their duals H' , E' , and U' . The Hilbert space E_0 is a dense subspace of E and the embedding $E_0 \hookrightarrow E$ is continuous. We denote by E'_0 the dual of E_0 with respect to the pivot space E so that

$$E_0 \subset E \subset E'_0$$

densely and with continuous embeddings. Such triples of Hilbert spaces are often encountered in the abstract treatment of partial differential equations. We denote $X_0 = H \oplus E_0$ so that $X'_0 = H \oplus E'_0$. We decompose the state of Σ as follows:

$$x_0 = \begin{bmatrix} z_0 \\ w_0 \end{bmatrix}, \quad z_0 \in H, \quad w_0 \in E.$$

We assume that we have three bounded operators

$$(1.1) \quad L \in \mathcal{L}(E_0, H), \quad K \in \mathcal{L}(E_0, U), \quad G \in \mathcal{L}(E_0, E'_0)$$

such that

$$(1.2) \quad \operatorname{Re} \langle Gw_0, w_0 \rangle_{E'_0, E_0} \leq 0 \quad \forall w_0 \in E_0,$$

and we define $\bar{A} \in \mathcal{L}(X_0, X'_0)$, $B \in \mathcal{L}(U, X'_0)$, and $\bar{C} \in \mathcal{L}(X_0, U)$ by

$$(1.3) \quad \bar{A} = \begin{bmatrix} 0 & -L \\ L^* & G - \frac{1}{2}K^*K \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ K^* \end{bmatrix}, \quad \bar{C} = [0 \quad -K].$$

The equations of the system are

$$(1.4) \quad \dot{x}(t) = \bar{A}x(t) + Bu(t), \quad y(t) = \bar{C}x(t) + u(t),$$

where x is the state trajectory, u is the input function, and y is the output function. Note that the differential equation above is an equation in X'_0 .

We define the domain $\mathcal{D}(A)$ by

$$(1.5) \quad \mathcal{D}(A) = \{x_0 \in X_0 \mid \bar{A}x_0 \in X\}$$

and we denote by A and C the restrictions of $\bar{A}$ and $\bar{C}$ to $\mathcal{D}(A)$. More explicitly,

$$(1.6) \quad \mathcal{D}(A) = \left\{ \begin{bmatrix} z_0 \\ w_0 \end{bmatrix} \in X_0 \mid L^*z_0 + (G - \frac{1}{2}K^*K)w_0 \in E \right\}.$$

Under the assumptions made so far, A is not necessarily closed. But for (1.4) to define a scattering passive system, we need A to be the generator of a strongly continuous semigroup of operators on X . One way to overcome this problem would be to assume that L is closed. This would indeed work, but it would be too restrictive: it would eliminate Maxwell's equations, which we would like to fit into this abstract framework. A better alternative is to assume the following weaker condition:

$$(1.7) \quad \begin{bmatrix} L \\ K \end{bmatrix} \quad (\text{with domain } E_0) \text{ is closed as an unbounded operator } E \rightarrow H \oplus U.$$

As we shall see later (in Theorem 1.4), this assumption implies that A is m-dissipative (i.e., A is dissipative and $I + A$ has a bounded inverse) and hence it generates a semigroup of contractions.

Informal statement of the main result. *Under the assumptions (1.1)–(1.3) and (1.7), the equations of (1.4) determine a scattering passive system with state space X . This system is scattering conservative if and only if we have equality in (1.2).*

Scattering passive systems and scattering conservative systems will be formally defined in section 3. The above statement sounds simple and it contains the essence of what we are proving in this paper. However, a precise and formal statement is much more complicated. One reason for this is that we have to make it clear what we mean by the claim that certain equations determine a system. Another reason is that we want to give a precise description of the system in terms of its system node, semigroup generator, control and observation operators, and transfer function.

Before stating a more precise version of the above informal result, let us explain how systems of the above type arise in the modeling of physical systems. To do this we introduce $K_0 = \frac{1}{\sqrt{2}}K$ and remark that (1.7) holds if and only if it holds with K_0 in place of K . The formal dynamical system

$$(1.8) \quad \begin{aligned} \begin{bmatrix} \dot{z}(t) \\ \dot{w}(t) \end{bmatrix} &= \begin{bmatrix} 0 & -L \\ L^* & G \end{bmatrix} \begin{bmatrix} z(t) \\ w(t) \end{bmatrix} + \begin{bmatrix} 0 \\ K_0^* \end{bmatrix} e(t), & t \geq 0, \\ f(t) &= \begin{bmatrix} 0 & K_0 \end{bmatrix} \begin{bmatrix} z(t) \\ w(t) \end{bmatrix}, & t \geq 0, \end{aligned}$$

often arises in physical modeling in the so-called impedance setting, where e stands for an effort variable and f stands for a flow variable. In Proposition 6.2 below we prove that if (1.1), (1.2), and (1.7) hold, then the above system is *impedance passive* in a weak sense, i.e., the operator

$$(1.9) \quad T = \begin{bmatrix} 0 & -L & 0 \\ L^* & G & K_0^* \\ 0 & -K_0 & 0 \end{bmatrix}$$

(with the appropriate domain) is m-dissipative. This implies that

$$(1.10) \quad \frac{d}{dt} \|x(t)\|^2 \leq 2\operatorname{Re} \langle e(t), f(t) \rangle$$

for all classical solutions $x(\cdot) = \begin{bmatrix} z(\cdot) \\ w(\cdot) \end{bmatrix}$ of (1.8). The physical interpretation of this inequality is that $\frac{1}{2}\|x(t)\|^2$ is the energy in the system, while $\operatorname{Re} \langle e(t), f(t) \rangle$ is the power entering the system. For an *impedance conservative* system the inequality in (1.10) holds as an equality. For closed $\begin{bmatrix} L \\ K_0 \end{bmatrix}$ the system will be impedance conservative

if and only if we have equality in (1.2). We refer to [2, 3, 5, 23, 31, 32, 33] for details on impedance passive and conservative systems.

Unfortunately, the system (1.8) is usually not well-posed. Well-posedness means for some (hence, for every) $t > 0$, the state $x(t)$ and the flow f (restricted to $[0, t]$) depend continuously on the initial state $x(0)$ and the effort e (restricted to $[0, t]$), and this does not follow from the above assumptions. In general, we do not even have a strongly continuous semigroup describing the evolution of the state $x(t)$ if $e = 0$. However, it is possible to reformulate the above problem into a *scattering* setting by choosing suitable combinations of e and f as input and output signals. If we define

$$(1.11) \quad u = \frac{1}{\sqrt{2}}(e + f), \quad y = \frac{1}{\sqrt{2}}(e - f),$$

then (1.8) becomes

$$\begin{aligned} \begin{bmatrix} \dot{z}(t) \\ \dot{w}(t) \end{bmatrix} &= \begin{bmatrix} 0 & -L \\ L^* & G - K_0^* K_0 \end{bmatrix} \begin{bmatrix} z(t) \\ w(t) \end{bmatrix} + \begin{bmatrix} 0 \\ \sqrt{2} K_0^* \end{bmatrix} u(t), & t \geq 0, \\ y(t) &= \begin{bmatrix} 0 & \sqrt{2} K_0 \end{bmatrix} \begin{bmatrix} z(t) \\ w(t) \end{bmatrix} + u(t), & t \geq 0, \end{aligned}$$

and instead of (1.10) we get the inequality that is characteristic of scattering passive systems:

$$(1.12) \quad \frac{d}{dt} \|x(t)\|^2 \leq \|u(t)\|^2 - \|y(t)\|^2.$$

After substituting $\sqrt{2} K_0 = K$ we arrive at the system (1.3)–(1.4).

The transformation (1.11) is called the *external Cayley transform*. It can be interpreted as a negative output feedback combined with a feed-forward term and a rescaling, as will be explained in more detail in section 5 (see Figure 1). As we prove in Theorem 5.2 and Remark 5.3, the system (1.3)–(1.4) is *scattering passive if and only if the operator T in (1.9) is m -dissipative*. Scattering passivity implies well-posedness. The system (1.3)–(1.4) is scattering conservative if the inequality in (1.2) holds as an equality. (Scattering conservative systems will be defined in section 3, and for them equality holds in (1.12).)

After this digression we now return to a more precise statement of our main results. The following three theorems use terminology that will be recalled in section 3.

THEOREM 1.1. *Let H, E, U, E_0 , and X_0 be as at the beginning of this section and let the operators L, K , and G be as (1.1), (1.2), and (1.7). Define the operator S_{sca} by*

$$S_{\text{sca}} = \begin{bmatrix} [A \& B]_{\text{sca}} \\ [C \& D]_{\text{sca}} \end{bmatrix},$$

where

$$(1.13) \quad [A \& B]_{\text{sca}} = \begin{bmatrix} 0 & -L & 0 \\ L^* & G - \frac{1}{2} K^* K & K^* \end{bmatrix}, \quad [C \& D]_{\text{sca}} = \begin{bmatrix} 0 & -K & I \end{bmatrix},$$

both have the domain

$$(1.14) \quad \mathcal{D}(S_{\text{sca}}) = \left\{ \begin{bmatrix} z_0 \\ w_0 \\ u_0 \end{bmatrix} \in X_0 \times U \mid L^* z_0 + (G - \frac{1}{2} K^* K) w_0 + K^* u_0 \in E \right\}.$$

Then S_{sca} is a scattering passive (hence, well-posed) system node with input space U , state space $X = H \oplus E$ and output space U .

The following proposition is an easy consequence of the above theorem, using known general properties of well-posed system nodes. It tells us that the equation associated with the system node S_{sca} (see (1.17) below) has plenty of classical solutions. These coincide with the classical solutions of (1.4).

We denote by $\mathcal{H}_{\text{loc}}^1((0, \infty); U)$ the space of those functions on $(0, \infty)$ whose restriction to $(0, n)$ is in $\mathcal{H}^1((0, n); U)$ for every $n \in \mathbb{N}$.

PROPOSITION 1.2. *We use the notation and the assumptions of Theorem 1.1. The space $\mathcal{D}(S_{\text{sca}})$ is dense in $H \oplus E \oplus U$. It is a Hilbert space with the norm*

$$\left\| \begin{bmatrix} z_0 \\ w_0 \\ u_0 \end{bmatrix} \right\|_{\mathcal{D}(S_{\text{sca}})}^2 = \|z_0\|^2 + \|w_0\|^2 + \|u_0\|^2 + \|Lw_0\|^2 + \|L^*z_0 + (G - \tfrac{1}{2}K^*K)w_0 + K^*u_0\|^2.$$

(On the right-hand side, we have used the norms of H, E , and U .)

If the input function u and the initial state $\begin{bmatrix} z(0) \\ w(0) \end{bmatrix}$ of S_{sca} satisfy

$$(1.15) \quad u \in \mathcal{H}_{\text{loc}}^1((0, \infty); U), \quad \begin{bmatrix} z(0) \\ w(0) \\ u(0) \end{bmatrix} \in \mathcal{D}(S_{\text{sca}}),$$

then the corresponding state trajectory $\begin{bmatrix} z \\ w \end{bmatrix}$ and output function y of S_{sca} satisfy

$$(1.16) \quad \begin{bmatrix} z \\ w \end{bmatrix} \in C^1([0, \infty); X), \quad \begin{bmatrix} z \\ w \\ u \end{bmatrix} \in C([0, \infty); \mathcal{D}(S_{\text{sca}})), \quad y \in \mathcal{H}_{\text{loc}}^1((0, \infty); Y),$$

and

$$(1.17) \quad \begin{bmatrix} \dot{z}(t) \\ \dot{w}(t) \\ y(t) \end{bmatrix} = S_{\text{sca}} \begin{bmatrix} z(t) \\ w(t) \\ u(t) \end{bmatrix} \quad \forall t \geq 0.$$

THEOREM 1.3. *We use the notation and the assumptions of Theorem 1.1. If the functions $u, x = \begin{bmatrix} z \\ w \end{bmatrix}$, and y are as in (1.15)–(1.17), then they satisfy the following power balance equation for every $t \geq 0$:*

$$(1.18) \quad \frac{d}{dt} \|x(t)\|^2 = \|u(t)\|^2 - \|y(t)\|^2 + 2\operatorname{Re} \langle Gw(t), w(t) \rangle.$$

The dual system node S_{sca}^* has the same structure but with L, K , and G replaced with $-L, -K$, and G^* . Hence, its classical solutions satisfy the same power balance equation (1.18). Therefore, S_{sca} is scattering conservative if and only if

$$(1.19) \quad \operatorname{Re} \langle Gw_0, w_0 \rangle = 0 \quad \forall w_0 \in E_0.$$

THEOREM 1.4. *We use the notation and the assumptions of Theorem 1.1. The semigroup generator A of S_{sca} is $\bar{A}$ from (1.3) restricted to $\mathcal{D}(A)$ from (1.6).*

Let $X_1 = \mathcal{D}(A)$ with the norm $\|x\|_1 = \|(I - A)x\|$, and let X_{-1} be the completion of X with respect to the norm $\|x\|_{-1} = \|(I - A)^{-1}x\|$. Then

$$(1.20) \quad X_1 \subset X_0 \subset X \subset X'_0 \subset X_{-1},$$

where all the embeddings are continuous and dense. A has a unique extension to an operator $A_{-1} \in \mathcal{L}(X, X_{-1})$, whose restriction to X_0 is $\overline{A}$ from (1.3).

The control operator B of S_{sca} is as in (1.3) and the observation operator C of S_{sca} is $\overline{C}$ from (1.3) restricted to $\mathcal{D}(A)$. The transfer function of S_{sca} is

$$\mathbf{G}(s) = I - K \left[sI + \frac{1}{2}K^*K - G + \frac{1}{s}L^*L \right]^{-1} K^*$$

for all s in the open right half-plane.

We mention that it follows immediately from the above theorem, combined with Theorem 1.3, that

$$A^* = \begin{bmatrix} 0 & L \\ -L^* & G^* - \frac{1}{2}K^*K \end{bmatrix},$$

$$\mathcal{D}(A^*) = \left\{ \begin{bmatrix} z_0 \\ w_0 \end{bmatrix} \in X_0 \mid -L^*z_0 + (G^* - \frac{1}{2}K^*K)w_0 \in E \right\}.$$

In section 2 we give several comments on the above results and we explain some connections with other classes of systems and with second order differential equations on Hilbert spaces. We give two examples based on the Euler–Bernoulli beam equation and outline an application to Maxwell’s equations.

To understand the theorems stated earlier, the reader should know the concepts of system node, well-posed system node, scattering passive system node, scattering conservative system node, classical solution, and well-posed system. The definitions of these concepts and some of their properties will be recalled in section 3. The concepts of semigroup generator, control operator, observation operator, and transfer function of a system node will also be recalled in section 3.

The internal and external Cayley transformations are discussed in sections 4 and 5. The proof of Theorem 1.1 is given in section 6. The proofs of Proposition 1.2 and Theorem 1.3 are also in section 6.

In section 7 we give a more explicit formula for the resolvent $(sI - \overline{A})^{-1}$ (as an operator matrix) and we prove Theorem 1.4.

A preliminary version of our results was presented in the conference paper [43].

2. Comments on the main results and examples.

Remark 2.1. If S_{sca} is a system node of the above structure, and moreover L is surjective, then for any $\begin{bmatrix} z \\ w \end{bmatrix}, u, y$ that satisfy (1.16) and (1.17) it is possible to identify a *hidden variable* q such that

$$z(t) = -Lq(t), \quad w(t) = \dot{q}(t).$$

Indeed, first we have to find an initial value for q such that $z(0) = -Lq(0)$ (this $q(0)$ may be nonunique). Then, for every $t \geq 0$ put $q(t) = q(0) + \int_0^t w(\sigma) d\sigma$. The hidden variable q satisfies the differential equation

$$\ddot{q}(t) + (\frac{1}{2}K^*K - G)\dot{q}(t) + L^*Lq(t) = K^*u(t),$$

where all the terms are in E'_0 , $q \in C^1([0, \infty); E_0)$, and $\dot{q} \in C^2([0, \infty); E)$. The difference between two different hidden variables associated to the same solution of (1.17) is a constant in $\text{Ker } L$. An alternative state for this system is $(q, \dot{q})$, taking values in the new state space $E_0 \oplus E$ (this is often used in the literature). Note that

since L is not assumed to be closed, in general it is not possible to restrict L^*L to a positive (in particular, self-adjoint) operator from $\mathcal{D}(L^*L) \subset E$ to E . If L is closed, then this becomes possible; see Theorem 13.13 in Rudin [27].

Systems described by second order differential equations like the one above, with similar assumptions but with L closed, have been studied in Jacob and Morris [15, 16]. They have examined various properties of the associated transfer functions. For more material on second order linear differential equations with operator coefficients we refer to section VI.3 in Engel and Nagel [12] and to Fattorini [13].

As already mentioned, the expression $\frac{1}{2}\|x(t)\|^2 = \frac{1}{2}(\|z(t)\|^2 + \|w(t)\|^2)$ is interpreted as the energy of the system at time t . After introducing q , the second term $\frac{1}{2}\|w(t)\|^2 = \frac{1}{2}\|\dot{q}(t)\|^2$ can be interpreted as the *kinetic energy*, and the first term $\frac{1}{2}\|z(t)\|^2 = \frac{1}{2}\|Lq(t)\|^2$ can be interpreted as the *potential energy* stored in Σ .

Remark 2.2. An interesting generalization of the class of systems discussed here arises if instead of the quadratic expression for the potential energy $V(z(t)) = \frac{1}{2}\|z(t)\|^2$, we allow it to be a more general nonlinear C^1 function of $z(t)$, denoted by V . Here, we assume that all the Hilbert spaces are real, which is more realistic for nonlinear systems. The energy in the system is defined by $\mathcal{E}(t) = V(z(t)) + \frac{1}{2}\|w(t)\|^2$. The equations of the system are postulated to be

$$(2.1) \quad \begin{aligned} \frac{d}{dt} \begin{bmatrix} z(t) \\ w(t) \end{bmatrix} &= \begin{bmatrix} 0 & -L \\ L^* & G - \frac{1}{2}K^*K \end{bmatrix} \begin{bmatrix} (\nabla V)(z(t)) \\ w(t) \end{bmatrix} + \begin{bmatrix} 0 \\ K^* \end{bmatrix} u(t) \\ y(t) &= -Kw(t) + u(t), \end{aligned}$$

with the same assumptions on L, K, G as before. This is an instance of a scattering version of a port-Hamiltonian system, studied by van der Schaft and collaborators [7, 28], Villegas [40], and others. For every solution of the above equations we have the power balance equation

$$2\dot{\mathcal{E}}(t) = \|u(t)\|^2 - \|y(t)\|^2 + 2\langle Gw(t), w(t) \rangle \quad \forall t \geq 0.$$

For most systems in this class, it is challenging to prove the existence and uniqueness of suitable solutions of (2.1) for a dense set of initial states and input functions. We refer to Yao and Weiss [47] for a nonlinear scattering passive Rayleigh type beam equation that can be put into this framework.

Remark 2.3. The conservative systems “from thin air” considered in [38] and [46] are a subclass of the systems discussed here. Indeed, if we assume that L has an inverse $L^{-1} \in \mathcal{L}(H, E_0)$, $G = 0$ and if we denote $A_0 = L^*L$, $C_0 = K$, then we obtain the conservative systems from [38, 46]. From $L^{-1} \in \mathcal{L}(H, E_0)$ we see that L is closed (from E to H) so that the assumption (1.7) is satisfied. The operator $A_0 \in \mathcal{L}(E_0, E'_0)$ can be restricted to a positive operator on E with domain $\mathcal{D}(A_0)$ (see also Remark 2.1) and then $E_0 = \mathcal{D}(A_0^{\frac{1}{2}})$. The state vector used in [38, 46] is $\begin{bmatrix} q \\ \dot{q} \end{bmatrix} = \begin{bmatrix} q \\ w \end{bmatrix}$, where q is the hidden variable introduced in Remark 2.1 ($q = -L^{-1}z$), and the corresponding state space is $E_0 \oplus E$. Due to this choice of state vector, there is no need to ever mention the space H or the operator L , the theory of systems “from thin air” uses only E , U , A_0 , and C_0 . In this case, our spaces E and E_0 correspond to what is denoted in [46] by H and $H_{\frac{1}{2}}$. The transformation from conservative systems “from thin air” to systems of the type discussed here can be found also in Schnaubelt and Weiss [29] (the proof of Theorem 5.1).

Example 2.4 (clamped Euler–Bernoulli beam). Consider $E = H = L^2[0, l]$, $E_0 = \mathcal{H}_0^2(0, l)$, and let $\alpha \in L^\infty[0, l]$ be positive and bounded from below, i.e., $\alpha(\xi) \geq \varepsilon > 0$

for almost every $\xi \in [0, l]$. Let $\kappa \in \mathcal{H}^{-2}(0, l)$ so that κ may contain several “Dirac pulses” and their derivatives placed in the interval $(0, l)$. We define $L \in \mathcal{L}(E_0, H)$ and $K \in \mathcal{L}(E_0, \mathbb{C})$ by

$$Lw = \alpha w'', \quad Kw = \langle w, \kappa \rangle_{\mathcal{H}_0^2, \mathcal{H}^{-2}}.$$

We consider $G = 0$. The operator L is the composition of a bounded and invertible operator (multiplication by α) with a closed operator (second derivative). Therefore, L is closed as an operator from E to H , so that (1.7) holds. Then according to Theorems 1.1 and 1.3, S_{sca} as in (1.13)–(1.14) is a scattering conservative system node with state space $L^2[0, l] \oplus L^2[0, l]$. Since L is invertible, this system fits into the class of conservative systems “from thin air” that were described in Remark 2.3. For every solution of (1.16)–(1.17) we can introduce the hidden variable $q = -L^{-1}z$, as explained in Remark 2.1, and we obtain that $q \in C^1([0, \infty); \mathcal{H}_0^2(0, l))$, $q \in C^2([0, \infty); L^2[0, l])$ and q satisfies the following Euler–Bernoulli beam equation with damping:

$$\ddot{q}(t) + \frac{\kappa}{2} \langle \dot{q}(t), \kappa \rangle_{\mathcal{H}_0^2, \mathcal{H}^{-2}} + (\alpha^2 q(t))'' = \kappa u(t)$$

with the (clamped) boundary conditions $q(0) = q(l) = 0$, $q'(0) = q'(l) = 0$. Here, the second derivative of $\alpha^2 q''$ is considered in the sense of distributions on $(0, l)$. The corresponding output signal is of course given by $y(t) = -\langle \dot{q}(t), \kappa \rangle + u(t)$.

Example 2.5 (free Euler–Bernoulli beam). This is a variation of the previous example, with noninvertible (but still closed) L . Consider again $E = H = L^2[0, l]$ but we take away the boundary conditions from E_0 : $E_0 = \mathcal{H}^2(0, l)$. Let α be as in Example 2.4. We denote by $\mathcal{H}^2(0, l)'$ the dual of $\mathcal{H}^2(0, l)$ with respect to the pivot space E . Let $\tilde{\kappa} \in \mathcal{H}^2(0, l)'$; then $\tilde{\kappa}$ has a unique decomposition of the form

$$(2.2) \quad \langle \varphi, \tilde{\kappa} \rangle_{\mathcal{H}^2, \mathcal{H}^{2'}} = a_0 \varphi(0) + b_0 \varphi'(0) + a_1 \varphi(l) + b_1 \varphi'(l) + \langle \varphi - \varphi_0, \kappa \rangle_{\mathcal{H}_0^2, \mathcal{H}^{-2}}$$

for all $\varphi \in \mathcal{H}^2(0, l)$, where $\kappa \in \mathcal{H}^{-2}(0, l)$. Here φ_0 is the polynomial of order at most three that satisfies the same (four) boundary conditions as φ so that $\varphi - \varphi_0 \in \mathcal{H}_0^2(0, l)$. We define the operators $L \in \mathcal{L}(E_0, H)$ and $K \in \mathcal{L}(E_0, \mathbb{C})$ by

$$Lw = \alpha w'', \quad Kw = \langle w, \tilde{\kappa} \rangle_{\mathcal{H}^2, \mathcal{H}^{2'}}.$$

We consider $G = 0$. L is closed, for the same reason as in the previous example. According to Theorems 1.1 and 1.3, S_{sca} as in (1.13)–(1.14) is a scattering conservative system node with state space $L^2[0, l] \oplus L^2[0, l]$. Since L is not invertible, this system does not fit into the class of conservative systems “from thin air.” Nevertheless, L is surjective; hence for every classical solution of (1.17) we can introduce the hidden variable q , as explained in Remark 2.1, and we obtain that $q \in C^1([0, \infty); \mathcal{H}^2(0, l))$, $q \in C^2([0, \infty); L^2[0, l])$ and q satisfies the following equation:

$$(2.3) \quad \ddot{q}(t) + \frac{\tilde{\kappa}}{2} \langle \dot{q}(t), \tilde{\kappa} \rangle_{\mathcal{H}^2, \mathcal{H}^{2'}} + L^*(\alpha q(t))'' = \tilde{\kappa} u(t).$$

The output signal is $y(t) = -\langle \dot{q}(t), \tilde{\kappa} \rangle + u(t)$.

In general, we cannot say more about the abstract equation (2.3). However, we can get a better understanding of it by assuming a special structure for $\tilde{\kappa}$. In what follows, we assume that κ from (2.2) is in $L^2[0, l]$. Equivalently, $\tilde{\kappa}$ is given by

$$(2.4) \quad \langle \varphi, \tilde{\kappa} \rangle_{\mathcal{H}^2, \mathcal{H}^{2'}} = c_0 \varphi(0) + d_0 \varphi'(0) + c_1 \varphi(l) + d_1 \varphi'(l) + \langle \varphi, \kappa \rangle.$$

It is not a difficult computation to express the numbers c_0, d_0, c_1, d_1 in terms of the numbers a_0, b_0, a_1, b_1 and the function κ , but this is not needed here: we may take (2.4) as the definition of $\tilde{\kappa}$ with $\kappa \in L^2[0, l]$. To understand the meaning of (2.3) with $\tilde{\kappa}$ given by (2.4), we rewrite (2.3) in weak form,

$$\langle \varphi, \ddot{q}(t) \rangle + \langle \varphi, \tilde{\kappa} \rangle_{\mathcal{H}^2, \mathcal{H}^{2'}} \left[\frac{1}{2} \langle \dot{q}(t), \tilde{\kappa} \rangle_{\mathcal{H}^2, \mathcal{H}^{2'}} - u(t) \right] + \langle \alpha \varphi'', \alpha q(t)'' \rangle = 0,$$

for every $\varphi \in \mathcal{H}^2(0, l)$. Using (2.4) this becomes

$$(2.5) \quad \left[c_0 \varphi(0) + d_0 \varphi'(0) + c_1 \varphi(l) + d_1 \varphi'(l) + \langle \varphi, \kappa \rangle \right] \cdot \left[\frac{1}{2} \langle \dot{q}(t), \tilde{\kappa} \rangle_{\mathcal{H}^2, \mathcal{H}^{2'}} - u(t) \right] + \langle \varphi, \ddot{q}(t) \rangle + \langle \varphi'', \alpha^2 q(t)'' \rangle = 0$$

for every $t \geq 0$ and every $\varphi \in \mathcal{H}^2(0, l)$. In particular, for $\varphi \in \mathcal{H}_0^2(0, l)$ we obtain

$$\langle \varphi, \kappa \rangle \cdot \left[\frac{1}{2} \langle \dot{q}(t), \tilde{\kappa} \rangle_{\mathcal{H}^2, \mathcal{H}^{2'}} - u(t) \right] + \langle \varphi, \ddot{q}(t) \rangle + \langle \varphi'', \alpha^2 q(t)'' \rangle = 0.$$

This shows that (for each $t \geq 0$) the last term (which depends on φ'') can be extended continuously to all $\varphi \in L^2[0, l]$. Hence, $\alpha^2 q(t)''$ belongs to the domain of the adjoint of the second derivative operator with domain $\mathcal{H}_0^2(0, l)$, which is $\mathcal{H}^2(0, l)$:

$$\alpha^2 q(t)'' \in \mathcal{H}^2(0, l) \quad \forall t \geq 0.$$

Armed with this knowledge, we return to (2.5) and perform integration by parts twice. We obtain that for all $t \geq 0$ and for all $\varphi \in \mathcal{H}^2(0, l)$,

$$(2.6) \quad \left[c_0 \varphi(0) + d_0 \varphi'(0) + c_1 \varphi(l) + d_1 \varphi'(l) + \langle \varphi, \kappa \rangle \right] \cdot \left[\frac{1}{2} \langle \dot{q}(t), \tilde{\kappa} \rangle_{\mathcal{H}^2, \mathcal{H}^{2'}} - u(t) \right] + \langle \varphi, \ddot{q}(t) \rangle + \varphi' \alpha^2 \overline{q(t)''} \Big|_0^l - \varphi \left(\alpha^2 \overline{q(t)''} \right)' \Big|_0^l + \langle \varphi, (\alpha^2 q(t)'')'' \rangle = 0.$$

In particular, for $\varphi \in \mathcal{H}_0^2(0, l)$ we obtain

$$\langle \varphi, \kappa \rangle \cdot \left[\frac{1}{2} \langle \dot{q}(t), \tilde{\kappa} \rangle_{\mathcal{H}^2, \mathcal{H}^{2'}} - u(t) \right] + \langle \varphi, \ddot{q}(t) \rangle + \langle \varphi, (\alpha^2 q(t)'')'' \rangle = 0,$$

which is equivalent to

$$\ddot{q}(t) + \frac{\kappa}{2} \langle \dot{q}(t), \tilde{\kappa} \rangle_{\mathcal{H}^2, \mathcal{H}^{2'}} + (\alpha^2 q(t)'')'' = \kappa u(t).$$

The last equation is the (second order in time) partial differential equation that describes the behavior of our system in the open interval $(0, l)$. We recognize this as an Euler–Bernoulli beam equation with a nonlocal distributed damping term (which may depend also on the boundary values of $\dot{q}$). The physical interpretation of $q(t)$ is the vertical displacement and the input $u(t)$ is acting on the beam as a distributed vertical force with intensity proportional to κ .

We want to understand the boundary conditions that q satisfies. For this we have to go back to (2.6), where we take φ such that $\varphi(0) = 1$, $\varphi(l) = 0$, $\varphi'(0) = 0$, and $\varphi'(l) = 0$ and $\|\varphi\|_{L^2}$ is very small (it is an easy exercise to see that this is possible). Then, in the limit as $\|\varphi\|_{L^2} \rightarrow 0$, we obtain

$$c_0 \left[\frac{1}{2} \langle \dot{q}(t), \tilde{\kappa} \rangle_{\mathcal{H}^2, \mathcal{H}^{2'}} - u(t) \right] + \left(\alpha^2 \overline{q(t)}'' \right)'(0) = 0.$$

If it happens that $c_0 = 0$, then this equation means that there is no force acting on the beam at the left end (where $\xi = 0$). The system corresponding to a $c_0 \neq 0$ can be interpreted as being obtained from the system with $c_0 = 0$ by closing a boundary force feedback loop and adding a boundary force input for the beam, because the last term in the above equation is the force acting at $\xi = 0$.

Similarly, in (2.6) we now take φ such that $\varphi(0) = 0$, $\varphi(l) = 0$, $\varphi'(0) = 1$, and $\varphi'(l) = 0$ and $\|\varphi\|_{L^2}$ is very small. Then, in the limit as $\|\varphi\|_{L^2} \rightarrow 0$, we obtain

$$d_0 \left[\frac{1}{2} \langle \dot{q}(t), \tilde{\kappa} \rangle_{\mathcal{H}^2, \mathcal{H}^{2'}} - u(t) \right] - \left(\alpha^2 \overline{q(t)}'' \right)'(0) = 0.$$

For $d_0 = 0$ this equation means that there is no torque acting on the beam at $\xi = 0$. The system corresponding to a $d_0 \neq 0$ can be interpreted as being obtained from the system with $d_0 = 0$ by closing a boundary torque feedback loop and adding a boundary torque input for the beam, because the last term in the above equation is the torque acting at $\xi = 0$.

A similar analysis can be done for the other end of the beam.

We mention that if we assume that $\tilde{\kappa} = \delta_x$, where $x \in (0, l)$ (a Dirac mass at x), then we obtain a free beam with a local force feedback at x . If we assume that $\tilde{\kappa} = \delta'_x$, then we obtain a free beam with a local torque feedback at x .

Example 2.6 (Maxwell's equations). In this example we consider Maxwell's equations on a bounded domain $\Omega \subset \mathbb{R}^3$ with Lipschitz boundary Γ . We denote the electric and magnetic fields by $\mathbf{E}$ and $\mathbf{H}$, respectively. We consider the system described by the Maxwell equations on Ω , assuming that the materials in Ω are linear, homogeneous, and isotropic, they have no conductivity (hence, there are no currents), and there are no charges and no external sources of electric field:

$$(2.7) \quad \mu \frac{\partial \mathbf{H}}{\partial t} = -\operatorname{rot} \mathbf{E}, \quad \varepsilon \frac{\partial \mathbf{E}}{\partial t} = \operatorname{rot} \mathbf{H},$$

$$(2.8) \quad \operatorname{div}(\mu \mathbf{H}) = 0, \quad \operatorname{div}(\varepsilon \mathbf{E}) = 0.$$

Here ε (the electric permittivity) and μ (the magnetic permeability) are positive numbers. We denote by ν the unit normal outward vector field on Γ (this is defined almost everywhere on Γ). We denote by $\gamma_0 \mathbf{E}$ and $\gamma_0 \mathbf{H}$ the traces of $\mathbf{E}$ and $\mathbf{H}$ on Γ and denote the tangential component of $\gamma_0 \mathbf{E}$ by $\pi_\tau \mathbf{E}$ so that

$$\pi_\tau \mathbf{E} = (\nu \times \gamma_0 \mathbf{E}) \times \nu.$$

Note that $\nu \times \gamma_0 \mathbf{E}$ is the same as $\pi_\tau \mathbf{E}$ rotated 90° around the normal direction to Γ . We define the input function u and the output function y by (cf. [28])

$$(2.9) \quad \begin{aligned} u &= \frac{1}{\sqrt{2}} \left(\nu \times \gamma_0 \mathbf{H} + \pi_\tau \mathbf{E} \right), \\ y &= \frac{1}{\sqrt{2}} \left(\nu \times \gamma_0 \mathbf{H} - \pi_\tau \mathbf{E} \right). \end{aligned}$$

In this brief example we consider only the case when $\mu = 1$ and $\varepsilon = 1$, because this case fits directly into the framework of Theorem 1.1. (The generalization to other values of μ and ε is easy, and for nonconstant μ and ε it becomes more tricky; a much more general and detailed treatment of Maxwell's equations is in [44].)

All the spaces that we use to analyze the Maxwell equations are real Hilbert spaces, consisting of real-valued functions. The input and output space consists of tangential vector fields on Γ :

$$(2.10) \quad U = \{u \in L^2(\Gamma, \mathbb{R}^3) \mid u \cdot \nu = 0\}.$$

The state space is $X = E \oplus E$, where $E = L^2(\Omega; \mathbb{R}^3)$. Thus, for a state $x = \begin{bmatrix} \mathbf{H} \\ \mathbf{E} \end{bmatrix}$, the expression $\|x\|^2 = \|\mathbf{H}\|^2 + \|\mathbf{E}\|^2$ is twice the physical energy. Systems described by Maxwell's equations were considered by many authors (see, for instance, [11, 20, 26, 28]), but we are not aware of works that define the input and output variables in a similar way to (2.9).

For any $\mathbf{E} \in E$, $\text{rot } \mathbf{E}$ is defined in the sense of distributions on Ω . We know from Theorem 2 on p. 204 of Dautray and Lions [8] that if $\mathbf{E} \in E$ is such that $\text{rot } \mathbf{E} \in E$, then its tangential trace $\mathbf{E}_\tau$ is well defined as an element of $\mathcal{H}^{-\frac{1}{2}}(\Gamma; \mathbb{R}^3)$. Hence, it makes sense to define a dense subspace of E as

$$E_0 = \{\mathbf{E} \in E \mid \text{rot } \mathbf{E} \in E, \pi_\tau \mathbf{E} \in U\},$$

which is a Hilbert space with the norm

$$\|\mathbf{E}\|_{E_0}^2 = \|\mathbf{E}\|_E^2 + \|\text{rot } \mathbf{E}\|_E^2 + \|\pi_\tau \mathbf{E}\|_U^2.$$

We define $L \in \mathcal{L}(E_0, E)$ and $K \in \mathcal{L}(E_0, U)$ by

$$L\mathbf{E} = \text{rot } \mathbf{E}, \quad K\mathbf{E} = \sqrt{2}\pi_\tau \mathbf{E}.$$

From the way the space E_0 and the operators L and K are defined, it is easy to see that $\begin{bmatrix} L \\ K \end{bmatrix}$ is a closed operator, as required in (1.7). It is also not difficult to see that L itself is not closed. We take $G = 0$. According to Theorems 1.1 and 1.3 (with $H = E$), these operators determine via (1.13) a scattering conservative system node S_{sca} . It can be shown (see [44]) that the equations of (1.17) that are satisfied by sufficiently smooth trajectories of S_{sca} are equivalent to (2.7) and (2.9). (The proof is not straightforward.) The equations (2.8) do not follow from (1.17), but it does follow that $\text{div } \mathbf{H}$ and $\text{div } \mathbf{E}$ are constant in time. In particular, if they are zero at the initial time $t = 0$, then they remain zero for all $t \geq 0$.

For much more detail on this example we refer the reader to our article [44], which can be regarded as a continuation of the present article. That article treats a more general case, where a part of the boundary is superconductive and the remaining part plays the same role as the full boundary Γ does above. In addition, in [44] the material in Ω is allowed to have a nonzero conductivity and nonconstant coefficients ε and μ , and the formulas (2.9) contain an additional coefficient function r .

3. System nodes, well-posed systems, scattering passive systems, and conservative systems. First we recall some simple facts about strongly continuous semigroups. Let $\mathbb{T} = (\mathbb{T}_t)_{t \geq 0}$ be a strongly continuous semigroup of bounded linear operators on the Hilbert space X with generator A . We define on X a new norm by

$$(3.1) \quad \|x\|_{-1} = \|(\beta I - A)^{-1}x\|,$$

where $\beta \in \rho(A)$ is fixed. The choice of β is not important, because different choices lead to equivalent norms. The generator A determines two additional Hilbert spaces as follows: X_1 is $\mathcal{D}(A)$ with the norm $\|x\|_1 = \|(\beta I - A)x\|$ (this norm is equivalent to the graph norm), while X_{-1} is the completion of X with respect to the norm $\|\cdot\|_{-1}$. It is possible to extend A to an operator $A_{-1} \in \mathcal{L}(X, X_{-1})$ which generates a strongly continuous semigroup $\mathbb{T}_{-1}$ on X_{-1} . For every $t \geq 0$, $\mathbb{T}_{-1,t}$ is an extension of $\mathbb{T}_t$. The extended semigroup is isomorphic to $\mathbb{T}$ via the unitary operator $\beta I - A_{-1} \in \mathcal{L}(X, X_{-1})$. Often we denote these extensions of A and of $\mathbb{T}_t$ by the same symbols A and $\mathbb{T}_t$. (We refer to our books [34, 39] for more details.)

The first aim of this section is to recall some facts about system nodes following [34, 18, 17]. The idea of system nodes goes back to Smuljan in 1986, using a different terminology. The concept was formalized in Staffans [30, 34].

DEFINITION 3.1. *Let U , X , and Y be Hilbert spaces. An operator*

$$S : \mathcal{D}(S) \rightarrow X \oplus Y \quad \text{with } \mathcal{D}(S) \subset X \oplus U$$

is called a system node on (U, X, Y) if it has the following properties:

1. *S is closed (as an operator from $X \oplus U$ to $X \oplus Y$).*
2. *We partition $S = \begin{bmatrix} A \& B \\ C \& D \end{bmatrix}$. The operator $A : \mathcal{D}(A) \rightarrow X$ defined by*

$$(3.2) \quad Ax = A \& B \begin{bmatrix} x \\ 0 \end{bmatrix}, \quad \mathcal{D}(A) = \{x \in X \mid \begin{bmatrix} x \\ 0 \end{bmatrix} \in \mathcal{D}(S)\}$$

is the generator of a strongly continuous semigroup on X .

3. *The operator $A \& B$ (with $\mathcal{D}(A \& B) = \mathcal{D}(S)$) can be extended to an operator $\begin{bmatrix} A_{-1} & B \end{bmatrix} \in \mathcal{L}(X \oplus U, X_{-1})$, where X_{-1} is defined as above.*
4. *$\mathcal{D}(S) = \{\begin{bmatrix} x \\ u \end{bmatrix} \in X \oplus U \mid A_{-1}x + Bu \in X\}$.*

It is easy to see that if S is a system node on (U, X, Y) , then $\mathcal{D}(S)$ is dense in $X \oplus U$ and $A \& B$ is closed (with domain $\mathcal{D}(S)$). Hence, the graph norm on $\mathcal{D}(S)$ is equivalent to the graph norm of the operator $A \& B$ on the same domain, defined by

$$(3.3) \quad \|\begin{bmatrix} x \\ u \end{bmatrix}\|_{\mathcal{D}(S)}^2 = \|x\|^2 + \|u\|^2 + \|A_{-1}x + Bu\|^2.$$

The operator A is called the *semigroup generator* of S and B is called the *control operator* of S . The operator $C \in \mathcal{L}(X_1, Y)$ defined by

$$(3.4) \quad Cx = C \& D \begin{bmatrix} x \\ 0 \end{bmatrix} \quad \forall x \in \mathcal{D}(A)$$

is called the *observation operator* of S . In this paper we usually write A instead of A_{-1} . The *transfer function* of S is the $\mathcal{L}(U, Y)$ -valued analytic function defined by

$$(3.5) \quad \mathbf{G}(s) = C \& D \begin{bmatrix} (sI - A)^{-1}B \\ I \end{bmatrix} \quad \forall s \in \rho(A).$$

It is easy to see that for all $s, \beta \in \rho(A)$ we have

$$(3.6) \quad \mathbf{G}(s) - \mathbf{G}(\beta) = C \left[(sI - A)^{-1} - (\beta I - A)^{-1} \right] B.$$

Combining (3.4) with (3.5) we easily get the useful formula

$$(3.7) \quad C \& D \begin{bmatrix} x \\ u \end{bmatrix} = C \left[x - (sI - A)^{-1}Bu \right] + \mathbf{G}(s)u,$$

valid for all $\begin{bmatrix} x \\ u \end{bmatrix} \in \mathcal{D}(S)$ and all $s \in \rho(A)$. This, together with $A \& B \begin{bmatrix} x \\ u \end{bmatrix} = Ax + Bu$, shows that S is completely determined by A, B, C , and $\mathbf{G}(s)$ (for a single s).

Define the space

$$(3.8) \quad Z = D(A) + (\beta I - A)^{-1}BU,$$

which is a Hilbert space with the norm

$$(3.9) \quad \|z\|_Z^2 = \inf \{ \|x\|_1^2 + \|v\|^2 \mid x \in X_1, v \in U, z = x + (\beta I - A)^{-1}Bv \}.$$

Note that if $\begin{bmatrix} x \\ v \end{bmatrix} \in \mathcal{D}(S)$, then $x \in Z$ and $\|x\|_Z \leq m \|\begin{bmatrix} x \\ v \end{bmatrix}\|_{\mathcal{D}(S)}$ for some $m > 0$ independent of x and v . The system node is called *compatible* if C has a continuous extension to an operator $\overline{C} \in \mathcal{L}(Z, Y)$. In this case, we may define the operator $D \in \mathcal{L}(U, Y)$ by $D = \mathbf{G}(\beta) - \overline{C}(\beta I - A)^{-1}B$ and it follows from (3.6) that D is independent of $\beta \in \rho(A)$. Then $C \& D$ and S can be split to take their form, which is familiar from finite-dimensional systems theory,

$$C \& D \begin{bmatrix} x \\ v \end{bmatrix} = \overline{C}x + Dv, \quad S = \begin{bmatrix} A & B \\ \overline{C} & D \end{bmatrix},$$

and we have

$$\mathbf{G}(s) = \overline{C}(sI - A)^{-1}B + D \quad \forall s \in \rho(A).$$

A system node S is usually associated with the equation

$$(3.10) \quad \begin{bmatrix} \dot{x}(t) \\ y(t) \end{bmatrix} = S \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} \quad \forall t \geq 0,$$

or equivalently, using the notation A, B , and $C \& D$ from Definition 3.1,

$$(3.11) \quad \dot{x}(t) = Ax(t) + Bu(t), \quad y(t) = C \& D \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} \quad \forall t \geq 0.$$

DEFINITION 3.2. Let S be a closed linear operator from $X \oplus U$ to $X \oplus Y$ with domain $\mathcal{D}(S)$ (but S need not be a system node).

A triple (x, u, y) is called a *classical solution* of (3.10) on $[0, \infty)$ if

- (a) $x \in C^1([0, \infty); X)$,
- (b) $u \in C([0, \infty); U)$, $y \in C([0, \infty); Y)$,
- (c) $\begin{bmatrix} x(t) \\ u(t) \end{bmatrix} \in \mathcal{D}(S)$ for all $t \geq 0$,
- (d) equation (3.10) holds.

A triple (x, u, y) is called a *generalized solution* of (3.10) on $[0, \infty)$ if

- (e) $x \in C([0, \infty); X)$,
- (f) $u \in L_{\text{loc}}^2([0, \infty); U)$, $y \in L_{\text{loc}}^2([0, \infty); Y)$,
- (g) there exists a sequence (x_k, u_k, y_k) of classical solutions of (3.10) such that $x_n \rightarrow x$ in $C([0, \infty); X)$, $u_k \rightarrow u$ in $L_{\text{loc}}^2([0, \infty); U)$, $y_k \rightarrow y$ in $L_{\text{loc}}^2([0, \infty); Y)$.

We remark that it follows easily from conditions (a)–(d) above that every classical solution of (3.10) on $[0, \infty)$ also satisfies

- (h) $\begin{bmatrix} x \\ u \end{bmatrix} \in C([0, \infty); \mathcal{D}(S))$,

where the continuity is with respect to the graph norm of S on $\mathcal{D}(S)$.

The following proposition guarantees that for a system node, we have plenty of classical solutions of the system equation (3.10).

PROPOSITION 3.3. Let S be a system node on (U, X, Y) . If $u \in C^2([0, \infty); U)$ and $\begin{bmatrix} x_0 \\ u(0) \end{bmatrix} \in \mathcal{D}(S)$, then (3.10) has a unique classical solution (x, u, y) satisfying $x(0) = x_0$. Moreover, this classical solution satisfies

$$x \in C^2([0, \infty); X_{-1}).$$

For the proof we refer to Lemma 4.7.8 in [34] or Proposition 4.2.11 in [39]. (Various versions of (parts of) this proposition can be found in the literature.)

For any $\tau \geq 0$ and $u \in L^2([0, \infty); U)$, let us denote by $\mathbf{P}_\tau u$ the restriction of u to $[0, \tau]$. Let us denote by $\mathcal{D}$ the space of all the pairs $(x_0, u) \in X \oplus L^2([0, \infty); U)$ which satisfy the assumptions of Proposition 3.3. Notice that $\mathcal{D}$ is dense in $X \oplus L^2([0, \infty); U)$. Hence, the corresponding space $\mathcal{D}_\tau$ of pairs $(x_0, \mathbf{P}_\tau u)$ is dense in $X \oplus L^2([0, \tau]; U)$. The last proposition allows us to define the operators Σ_τ from $\mathcal{D}_\tau$ to $X \oplus L^2([0, \tau]; Y)$ such that for any solution of (3.10) and for any $\tau \geq 0$,

$$(3.12) \quad \begin{bmatrix} x(\tau) \\ \mathbf{P}_\tau y \end{bmatrix} = \Sigma_\tau \begin{bmatrix} x_0 \\ \mathbf{P}_\tau u \end{bmatrix}.$$

DEFINITION 3.4. *The system node S is called well-posed if for some (hence, for every) $\tau > 0$, the operator Σ_τ from (3.12) has a continuous extension*

$$\Sigma_\tau \in \mathcal{L}(X \oplus L^2([0, \tau], U), X \oplus L^2([0, \tau], Y)).$$

In this case, the family $(\Sigma_\tau)_{\tau \geq 0}$ is called a well-posed linear system.

For such systems we refer to [34] and the references therein. Every well-posed system node is compatible; see Theorem 3.4 in [35]. For well-posed system nodes Proposition 3.3 can be modified to obtain a stronger statement, as follows.

PROPOSITION 3.5. *Let S be a well-posed system node on (U, X, Y) . Assume that $u \in \mathcal{H}_{\text{loc}}^1((0, \infty); U)$ and $\begin{bmatrix} x_0 \\ u(0) \end{bmatrix} \in \mathcal{D}(S)$. Then (3.10) has a unique classical solution (x, u, y) satisfying $x(0) = x_0$. Moreover, we have*

$$y \in \mathcal{H}_{\text{loc}}^1((0, \infty); Y).$$

For the proof see Theorem 4.6.11 in [34] or Theorem 3.1 in [35].

DEFINITION 3.6. *The system node S is called scattering passive if all the classical solutions of (3.10) satisfy*

$$\frac{d}{dt} \|x(t)\|^2 \leq \|u(t)\|^2 - \|y(t)\|^2 \quad \forall t \geq 0.$$

An equivalent condition is that all the generalized solutions of (3.10) satisfy

$$(3.13) \quad \|x(\tau)\|^2 + \int_0^\tau \|y(t)\|^2 dt \leq \|x(0)\|^2 + \int_0^\tau \|u(t)\|^2 dt \quad \forall t \geq 0.$$

A third equivalent condition is that the operators Σ_τ from (3.12) are *contractions*. In this case, the well-posed linear system $(\Sigma_\tau)_{\tau \geq 0}$ is called a *scattering passive linear system*. Such systems have been studied in [4, 24, 31, 32, 34, 35] and other references. (In [24] and [35] such systems were called dissipative.)

The system node S is called *scattering energy preserving* if the power balance equation

$$\frac{d}{dt} \|x(t)\|^2 = \|u(t)\|^2 - \|y(t)\|^2 \quad \forall t \geq 0$$

holds for all classical solutions of (3.10). Clearly, such systems nodes are scattering passive. The corresponding scattering passive system $(\Sigma_\tau)_{\tau \geq 0}$ is then *scattering energy preserving*, which means that all its trajectories satisfy (3.13) with equality. In other words, the operators Σ_τ are isometric.

The *dual* of a system node S on (U, X, Y) is simply its adjoint S^* . It can be verified that S^* is a system node on (Y, X, U) . The semigroup generator A^d , the control operators B^d , the observation operator C^d , and transfer functions $\mathbf{G}^d$ of the dual system node S^* are related to the corresponding operators for S as follows:

$$(3.14) \quad A^d = A^*, \quad B^d = C^*, \quad C^d = B^*, \quad \mathbf{G}^d(s) = \mathbf{G}(\bar{s})^*.$$

The system node S^* is well-posed (or scattering passive) if and only if S is well-posed (or scattering passive). This follows from Theorems 3.4 and 3.5 in [36].

The system node S is called *scattering conservative* if both S and S^* are scattering energy preserving. The corresponding scattering-passive system $(\Sigma_\tau)_{\tau \geq 0}$ is then also called *scattering conservative*. For such a system, the operators Σ_τ are unitary. For scattering conservative systems we refer to [4, 21, 24, 28, 31, 36, 38, 45, 46] and the references therein. In particular, relatively simple necessary and sufficient conditions for a system node to be scattering conservative have been established in [24].

4. The internal Cayley transformation. We denote by $\mathbb{C}_+$ the open right half-plane in $\mathbb{C}$.

DEFINITION 4.1. The internal Cayley transform of a system node S with parameter $\alpha \in \rho(A) \cap \mathbb{C}_+$ is the following operator from $X \oplus U$ to $X \oplus Y$:

$$(4.1) \quad \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} = \begin{bmatrix} (\bar{\alpha}I + A)(\alpha I - A)^{-1} & \sqrt{2\operatorname{Re} \alpha}(\alpha I - A)^{-1}B \\ \sqrt{2\operatorname{Re} \alpha}C(\alpha I - A)^{-1} & \mathbf{G}(\alpha) \end{bmatrix},$$

where A, B, C are as in Definition 3.1 and $\mathbf{G}$ is the transfer function of S .

The internal Cayley transformation has been used in [4, 24, 25, 31, 32, 33, 34, 35] and other references. It should not be confused with the usual *Cayley transform* of the operator S with parameter $\alpha \in \rho(S) \cap \mathbb{C}_+$, which is $S^C = (\bar{\alpha}I + S)(\alpha I - S)^{-1}$.

The operator defined in (4.1) can be interpreted as a discrete-time system node on (U, X, Y) , which determines a discrete-time system with input space U , state space X , and output space Y via the equations

$$(4.2) \quad x_{k+1} = \mathbf{A}x_k + \mathbf{B}u_k, \quad y_k = \mathbf{C}x_k + \mathbf{D}u_k.$$

It is easy to check that if $\alpha > 0$, then we have

$$\begin{bmatrix} \frac{x_{k+1} - x_k}{h} \\ \frac{y_k}{\sqrt{h}} \end{bmatrix} = S \begin{bmatrix} \frac{x_{k+1} + x_k}{2} \\ \frac{u_k}{\sqrt{h}} \end{bmatrix}, \quad \text{where } h = \frac{2}{\alpha}.$$

The transformation that leads from the system S to the discrete-time system described in (4.2) (with $\alpha > 0$) is called *Tustin discretization* with time step h in the engineering literature. The transfer function of this discrete-time system is

$$\mathbf{C}(zI - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D} = \mathbf{G}\left(\frac{\alpha z - \bar{\alpha}}{z + 1}\right) \quad \text{for } |z| > 1.$$

PROPOSITION 4.2. Let $S = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$ be a system node with semigroup generator A and let $\mathbf{S} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}$ be its internal Cayley transform with parameter $\alpha \in \rho(A) \cap \mathbb{C}_+$.

Then the operator $\mathbf{A}$ does not have -1 as an eigenvalue, and $\mathbf{S}$ can alternatively be computed from S in the following way: The operator

$$E = \begin{bmatrix} \frac{I}{\sqrt{2\operatorname{Re} \alpha}} & 0 \\ 0 & I \end{bmatrix} \left(\begin{bmatrix} \alpha I & 0 \\ 0 & I \end{bmatrix} - \begin{bmatrix} A & B \\ 0 & 0 \end{bmatrix} \right)$$

maps $\mathcal{D}(S)$ one-to-one onto $X \oplus U$ and

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} = \begin{bmatrix} -I & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \sqrt{2\operatorname{Re}\alpha}I & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ C\&D \end{bmatrix} E^{-1}.$$

The inverse of E , denoted by $\mathbf{E} = E^{-1}$, is given by

$$(4.3) \quad \mathbf{E} = \begin{bmatrix} \frac{I}{\sqrt{2\operatorname{Re}\alpha}} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} I + \mathbf{A} & \mathbf{B} \\ 0 & I \end{bmatrix}$$

and S can be recovered from $\mathbf{S}$ via the formula

$$(4.4) \quad \begin{bmatrix} A\&B \\ C\&D \end{bmatrix} = \begin{bmatrix} \alpha I & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \sqrt{2\operatorname{Re}\alpha}I & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} -I & 0 \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \mathbf{E}^{-1}.$$

Conversely, suppose that $\mathbf{S} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \in \mathcal{L}(X \oplus U; X \oplus Y)$, that $\mathbf{A}$ does not have -1 as an eigenvalue, and that for some $\alpha \in \mathbb{C}_+$ the operator

$$(4.5) \quad A = (\alpha\mathbf{A} - \bar{\alpha}I)(\mathbf{A} + I)^{-1}, \quad \mathcal{D}(A) = \operatorname{Ran}(\mathbf{A} + I)$$

generates a strongly continuous semigroup on X . Then $\alpha \in \rho(A)$ and $\mathbf{E}$ from (4.3) is injective and has dense range. Denote the range of $\mathbf{E}$ by $\mathcal{D}(S)$, and on this domain define the operator $S = \begin{bmatrix} A\&B \\ C\&D \end{bmatrix}$ by (4.4).

Then S is a system node whose semigroup generator is A from (4.5). The internal Cayley transform of S with parameter α is $\mathbf{S}$.

For the proof we refer to Propositions 5.1 and 5.2 in [4] or Lemma 7.1 in [32]. Note that according to the above proposition, the formulas (4.3)–(4.4) define the inverse internal Cayley transformation.

PROPOSITION 4.3. *Let S be a scattering passive system node. Then for every $\alpha \in \mathbb{C}_+$, the internal Cayley transform of S is a contraction.*

Conversely, suppose that $\mathbf{S} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}$ is a contraction and that -1 is not an eigenvalue of $\mathbf{A}$. Then for every $\alpha \in \mathbb{C}_+$, the operator S defined by (4.3)–(4.4), with $\mathcal{D}(S)$ being the range of $\mathbf{E}$, is a scattering passive system node.

For the proof we refer to Theorem 7.1(ii) in [32]. (Various parts of this can also be found in [4], [24], and [34].)

PROPOSITION 4.4. *If S is a scattering passive system node, then S is scattering conservative if and only if its internal Cayley transform $\mathbf{S}$ is unitary.*

For the proof we refer again to [4], [24], [32], or [34].

5. The external Cayley transformation. A concept that is closely related to the concept of a scattering passive (or conservative) system node is the concept of an *impedance passive* (or *conservative*) system node. Still following the terminology of [31, 32], these are system nodes which have equal input and output spaces and for which the trajectories (the solutions of (3.10)) satisfy

$$\|x(\tau)\|^2 - \|x(0)\|^2 \leq 2 \int_0^\tau \operatorname{Re} \langle e(t), f(t) \rangle dt \quad \forall \tau \in [0, T)$$

(or the corresponding equality). Here, we have denoted the input signal by e (sometimes called the effort) and the output signal by f (sometimes called the flow), and of course (3.10) should be written with these signals in place of u and y .

It is always possible to transform an impedance passive or conservative system node S_{imp} into a scattering passive or conservative system node S_{sca} by the *external*

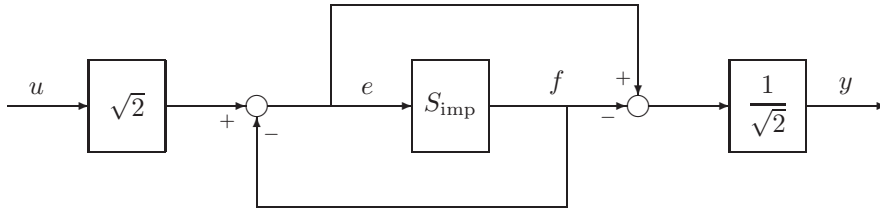


FIG. 1. The system node S_{sca} with input u and output y , as obtained from the system node S_{imp} (with input e and output f) via (1.11).

Cayley transformation (sometimes called the diagonal transformation) which redefines the input and the output as in (1.11). The inverse transformation is given by the same formulas, only with the places of u, y and e, f reversed, as is easy to see. The external Cayley transformation has been employed in many works; see, for example, Macchelli et al. [22], Staffans [31, 32, 33], and Weiss [42]. In the literature, often an extra parameter $\beta \in \mathbb{C}_+$ is included in the definition, but here we have taken $\beta = 1$. The external Cayley transformation can be understood also as an output feedback transformation (combined with a feed-forward term and a rescaling), as Figure 1 (approximately reproduced from [42]) shows. It is easy to see from this figure that the relationship between the transfer functions of S_{imp} and S_{sca} is

$$\mathbf{G}_{\text{sca}} = (I - \mathbf{G}_{\text{imp}})(I + \mathbf{G}_{\text{imp}})^{-1}.$$

In the papers [31, 32] the precise relationship between the system nodes S_{imp} and S_{sca} has been determined, and this is conveyed in the following proposition.

PROPOSITION 5.1. *Suppose that $S_{\text{imp}} = \begin{bmatrix} [A \& B]_{\text{imp}} \\ [C \& D]_{\text{imp}} \end{bmatrix}$ is an impedance passive system node. Then the operator*

$$(5.1) \quad E_{\text{imp}} := \begin{bmatrix} I & 0 \\ 0 & \frac{I}{\sqrt{2}} \end{bmatrix} \left(\begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ [C \& D]_{\text{imp}} \end{bmatrix} \right)$$

with domain $\mathcal{D}(S_{\text{imp}})$ is injective. Denote the range of this operator by $\mathcal{D}(S_{\text{sca}})$, and define the operator S_{sca} (with domain $\mathcal{D}(S_{\text{sca}})$) by

$$(5.2) \quad S_{\text{sca}} = \begin{bmatrix} [A \& B]_{\text{sca}} \\ [C \& D]_{\text{sca}} \end{bmatrix} := \begin{bmatrix} 0 & 0 \\ 0 & -I \end{bmatrix} + \left(\begin{bmatrix} 0 & 0 \\ 0 & \sqrt{2}I \end{bmatrix} + \begin{bmatrix} [A \& B]_{\text{imp}} \\ 0 & 0 \end{bmatrix} \right) E_{\text{imp}}^{-1}.$$

Then S_{sca} is a scattering passive system node. The inverse of E_{imp} is $E_{\text{imp}}^{-1} = E_{\text{sca}}$, where

$$(5.3) \quad E_{\text{sca}} := \begin{bmatrix} I & 0 \\ 0 & \frac{I}{\sqrt{2}} \end{bmatrix} \left(\begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ [C \& D]_{\text{sca}} \end{bmatrix} \right),$$

and S_{imp} can be recovered from S_{sca} via

$$(5.4) \quad S_{\text{imp}} = \begin{bmatrix} 0 & 0 \\ 0 & -I \end{bmatrix} + \left(\begin{bmatrix} 0 & 0 \\ 0 & \sqrt{2}I \end{bmatrix} + \begin{bmatrix} [A \& B]_{\text{sca}} \\ 0 & 0 \end{bmatrix} \right) E_{\text{sca}}^{-1}.$$

Notice that S_{sca} is obtained from S_{imp} by the same formulas by which S_{imp} is obtained from S_{sca} . However, there is a hidden asymmetry here: the external Cayley

transformation will not yield every possible scattering passive system node. The range of the external Cayley transformation (when applied to impedance conservative system nodes) will be characterized in Remark 5.4. This range is contained in those scattering passive system nodes for which the operator $\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}$ from (4.1) does not have -1 as an eigenvalue. When the backward transformation (5.3)–(5.4) is applied to a scattering passive system node that does not satisfy the above eigenvalue condition, then it results in a “multivalued operator.” (It follows from (5.7) below that E_{sca} is injective if and only if $\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}$ does not have -1 as an eigenvalue.) Both the single-valued case and the multivalued case can be analyzed with the methods described in Arov, Kurula, and Staffans [1, 2], Behrndt, Hassi, and de Snoo [6], Derkach et al. [9], Hassi, Malamud, and Mogilevskii [14], Kurula [17], Kurula and Staffans [18], Kurula et al. [19], etc. Here we shall restrict ourselves to the single-valued case.

To prove Theorem 1.1, we need the following generalization of Proposition 5.1, which can be applied to an operator S_{imp} that is not necessarily a system node. The proof of this theorem is based on the fact that if one applies to a system node S both the internal Cayley transform with parameter $\alpha = 1$ and the external Cayley transform (in any order), then the resulting operator is $\begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix} (I+S)(I-S)^{-1}$. (Here $(I+S)(I-S)^{-1}$ is the usual Cayley transform of S with parameter $\alpha = 1$.)

THEOREM 5.2. *Let $S_{\text{imp}} = \begin{bmatrix} [A\&B]_{\text{imp}} \\ [C\&D]_{\text{imp}} \end{bmatrix}$ be an operator in $X \oplus U$ with domain $\mathcal{D}(S_{\text{imp}})$ such that $T := \begin{bmatrix} [A\&B]_{\text{imp}} \\ -[C\&D]_{\text{imp}} \end{bmatrix}$ (with the same domain) is m -dissipative. Then the operator E_{imp} from (5.1) is injective on $\mathcal{D}(S_{\text{imp}})$. We denote its range by $\mathcal{D}(S_{\text{sca}})$ and we define S_{sca} (with domain $\mathcal{D}(S_{\text{sca}})$) by (5.2). Then S_{sca} is a scattering passive system node and $E_{\text{imp}}^{-1} = E_{\text{sca}}$ from (5.3).*

We denote by A_{sca} , B_{sca} , and C_{sca} the semigroup generator, the control operator, and the observation operator of S_{sca} , and we denote by $\mathbf{G}_{\text{sca}}$ its transfer function. Then, for all $s \in \mathbb{C}_+$,

$$(5.5) \quad \begin{bmatrix} (sI - A_{\text{sca}})^{-1} & \frac{1}{\sqrt{2}}(sI - A_{\text{sca}})^{-1}B_{\text{sca}} \\ \frac{1}{\sqrt{2}}C_{\text{sca}}(sI - A_{\text{sca}})^{-1} & \frac{1}{2}(I + \mathbf{G}_{\text{sca}}(s)) \end{bmatrix} = \left(\begin{bmatrix} sI & 0 \\ 0 & I \end{bmatrix} - \begin{bmatrix} [A\&B]_{\text{imp}} \\ -[C\&D]_{\text{imp}} \end{bmatrix} \right)^{-1}.$$

The operator S_{imp} can be recovered from S_{sca} via the formulas (5.3)–(5.4).

The system node S_{sca} is scattering conservative if and only if T is skew-adjoint.

Proof. Define $\mathbf{S}$ to be the Cayley transform of T with parameter 1, meaning that $\mathbf{S} = (I + T)(I - T)^{-1}$, so that

$$(5.6) \quad \frac{1}{2}(\mathbf{S} + I) = \frac{1}{2} \left(\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} + \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} \right) = \left(\begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} - \begin{bmatrix} [A\&B]_{\text{imp}} \\ -[C\&D]_{\text{imp}} \end{bmatrix} \right)^{-1}.$$

Then according to classical results about the Cayley transformation (see Theorem 3.4.9 in [34] or p. 167 in [37]), $\mathbf{S}$ is a contraction, and $\mathbf{S}$ is unitary if and only if T is skew-adjoint. Moreover, -1 is not an eigenvalue of $\mathbf{S}$ and hence also not of $\mathbf{A}$. Indeed, if there would exist a nonzero $x \in X$ such that $\mathbf{A}x = -x$, then from $\|\mathbf{S}\| \leq 1$ we see that $\mathbf{C}x = 0$ and hence $\mathbf{S} \begin{bmatrix} x \\ 0 \end{bmatrix} = - \begin{bmatrix} x \\ 0 \end{bmatrix}$, a contradiction.

Since $\mathbf{S}$ is contractive and $\mathbf{A}$ does not have -1 as an eigenvalue, it follows from the converse part of Proposition 4.3 that $\mathbf{S}$ has an inverse internal Cayley transform with parameter $\alpha = 1$, denoted S_{sca} , and this is a scattering passive system node. According to the last part of Proposition 4.2, the semigroup generator of S_{sca} is

$$A_{\text{sca}} = (\mathbf{A} - I)(\mathbf{A} + I)^{-1}, \quad \mathcal{D}(A_{\text{sca}}) = \text{Ran}(\mathbf{A} + I)$$

and the internal Cayley transform of S_{sca} is $\mathbf{S}$.

Denote the transfer function of S_{sca} by $\mathbf{G}_{\text{sca}}$. If in formula (4.1) we take $\alpha = 1$, add the identity on $X \oplus U$ to both sides, and divide by 2, then we get

$$\frac{1}{2} \left(\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} + \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} \right) = \begin{bmatrix} (I - A_{\text{sca}})^{-1} & \frac{1}{\sqrt{2}}(I - A_{\text{sca}})^{-1}B_{\text{sca}} \\ \frac{1}{\sqrt{2}}C_{\text{sca}}(I - A_{\text{sca}})^{-1} & \frac{1}{2}(I + \mathbf{G}_{\text{sca}}(1)) \end{bmatrix}.$$

This combined with (5.6) gives (5.5) with $s = 1$.

Now we show that S_{sca} (which in the proof we have defined differently from in the theorem) is indeed given by (5.2). We factor the right-hand side above, obtaining

$$(5.7) \quad \frac{1}{2} \left(\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} + \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} \right) = E_{\text{sca}} \begin{bmatrix} (I - A_{\text{sca}})^{-1} & (I - A_{\text{sca}})^{-1}B_{\text{sca}} \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & \frac{I}{\sqrt{2}} \end{bmatrix},$$

where E_{sca} is defined as in (5.3). This factorization is checked by direct computation, using (3.4) and (3.5). The last factor on the right-hand side maps $X \oplus U$ one-to-one onto itself, and the second factor maps $X \oplus U$ one-to-one onto $\mathcal{D}(S_{\text{sca}})$, whereas, according to (5.6), the left-hand side maps $X \oplus U$ one-to-one onto $\mathcal{D}(S_{\text{imp}})$. Consequently, E_{sca} maps $\mathcal{D}(S_{\text{sca}})$ one-to-one onto $\mathcal{D}(S_{\text{imp}})$. Inverting both sides of (5.5) with $s = 1$ we get the following identity, valid on $\mathcal{D}(S_{\text{imp}})$:

$$(5.8) \quad \begin{bmatrix} I & 0 \\ 0 & \sqrt{2}I \end{bmatrix} \left(\begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} - \begin{bmatrix} [A \& B]_{\text{sca}} \\ 0 & 0 \end{bmatrix} \right) E_{\text{sca}}^{-1} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} - \begin{bmatrix} [A \& B]_{\text{imp}} \\ -[C \& D]_{\text{imp}} \end{bmatrix}.$$

The bottom row of this identity implies that the bottom row of E_{sca}^{-1} is equal to $\frac{1}{\sqrt{2}}([0 \ I] + [C \& D]_{\text{imp}})$. Trivially, $[I \ 0] E_{\text{sca}} = [I \ 0]$, and hence the top row of E_{sca}^{-1} is $[I \ 0]$. Thus, $E_{\text{sca}}^{-1} = E_{\text{imp}}$, as defined in (5.1). In particular, E_{imp} is injective and maps $\mathcal{D}(S_{\text{imp}})$ one-to-one onto $\mathcal{D}(S_{\text{sca}})$. If we multiply (5.8) by E_{imp}^{-1} to the right and discard the bottom row, then we get the top row of (5.2) (since $[I \ 0] E_{\text{imp}}^{-1} = [I \ 0] E_{\text{sca}} = [I \ 0]$). That the bottom row of (5.2) also holds follows from the fact that $E_{\text{imp}}^{-1} = E_{\text{sca}}$. Thus, S_{sca} is indeed given by (5.2).

Now we show that S_{imp} can be recovered from S_{sca} via (5.3)–(5.4). It is easy to see that the top row of (5.4) follows from the top row of (5.2) (since $E_{\text{sca}} = E_{\text{imp}}^{-1}$). The bottom row of (5.4) follows from $E_{\text{sca}}^{-1} = E_{\text{imp}}$ and (5.1).

Recall (from this proof) that T is skew-adjoint if and only if $\mathbf{S}$ is unitary. According to Proposition 4.4, $\mathbf{S}$ is unitary if and only if S_{sca} is scattering conservative.

It remains to prove (5.5) for general $s \in \mathbb{C}_+$. Let us denote by $\mathcal{R}(s)$ the left-hand side of (5.5). Using (3.4), (3.5), and (5.3) we factor

$$\begin{aligned} \mathcal{R}(s) &= E_{\text{sca}} \begin{bmatrix} (sI - A_{\text{sca}})^{-1} & (sI - A_{\text{sca}})^{-1}B_{\text{sca}} \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & \frac{I}{\sqrt{2}} \end{bmatrix} \\ &= E_{\text{sca}} \left(\begin{bmatrix} sI & 0 \\ 0 & I \end{bmatrix} - \begin{bmatrix} [A \& B]_{\text{sca}} \\ 0 & 0 \end{bmatrix} \right)^{-1} \begin{bmatrix} I & 0 \\ 0 & \frac{I}{\sqrt{2}} \end{bmatrix}. \end{aligned}$$

Using (5.4) we obtain

$$\begin{aligned} \mathcal{R}(s) &= \left(\begin{bmatrix} I & 0 \\ 0 & \sqrt{2}I \end{bmatrix} \left(\begin{bmatrix} sI & 0 \\ 0 & I \end{bmatrix} - \begin{bmatrix} [A \& B]_{\text{sca}} \\ 0 & 0 \end{bmatrix} \right) E_{\text{sca}}^{-1} \right)^{-1} \\ &= \left(\begin{bmatrix} sI & 0 \\ 0 & I \end{bmatrix} - \begin{bmatrix} [A \& B]_{\text{imp}} \\ -[C \& D]_{\text{imp}} \end{bmatrix} \right)^{-1}. \quad \square \end{aligned}$$

Remark 5.3. By examining the proof of Theorem 5.2 we find that the converse of this theorem is also true in the following form: If S_{sca} is a scattering passive system node, and if its internal Cayley transform $\mathbf{S}$ with parameter $\alpha = 1$ does not have -1 as an eigenvalue, then its external Cayley transform S_{imp} given by (5.3)–(5.4) is well defined, and it satisfies the assumption of Theorem 5.2.

Remark 5.4. It follows from Theorem 5.2 and Proposition 4.11 in Ball and Staffans [5] that a scattering conservative system node S_{sca} with input and output space U and state space X is the image of an impedance conservative system node under the external Cayley transformation if and only if the operator E_{sca} from (5.3) is injective and the projection of the range of E_{sca} onto U is all of U .

PROPOSITION 5.5. *We use the notation and the assumption of Theorem 5.2. The triple (x, u, y) is a classical (or generalized) solution of $\begin{bmatrix} \dot{x}(t) \\ y(t) \end{bmatrix} = S_{\text{sca}} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix}$ on $[0, \infty)$ if and only if the triple (x, e, f) , where e, f satisfy (1.11), is a classical (or generalized) solution of $\begin{bmatrix} \dot{x}(t) \\ f(t) \end{bmatrix} = S_{\text{imp}} \begin{bmatrix} x(t) \\ e(t) \end{bmatrix}$ on $[0, \infty)$.*

Proof. First we prove the claim about the classical solutions. Clearly the transformation (1.11) preserves continuity, so that conditions (a) and (b) in Definition 3.2 are equivalent for the two systems. Thus, it suffices to prove that for any fixed $t \geq 0$, $\begin{bmatrix} x(t) \\ u(t) \end{bmatrix} \in \mathcal{D}(S_{\text{sca}})$ if and only if $\begin{bmatrix} x(t) \\ e(t) \end{bmatrix} \in \mathcal{D}(S_{\text{imp}})$ and that $\begin{bmatrix} \dot{x}(t) \\ y(t) \end{bmatrix} = S_{\text{sca}} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix}$ if and only if $\begin{bmatrix} \dot{x}(t) \\ f(t) \end{bmatrix} = S_{\text{imp}} \begin{bmatrix} x(t) \\ e(t) \end{bmatrix}$.

Suppose that the vectors $x(t), \dot{x}(t) \in X$ and $u(t), y(t) \in U$ are such that $\begin{bmatrix} x(t) \\ u(t) \end{bmatrix} \in \mathcal{D}(S_{\text{sca}})$ and $\begin{bmatrix} \dot{x}(t) \\ y(t) \end{bmatrix} = S_{\text{sca}} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix}$. Then by (1.11),

$$\begin{bmatrix} x(t) \\ e(t) \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & \frac{I}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} x(t) \\ u(t) + y(t) \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & \frac{I}{\sqrt{2}} \end{bmatrix} \left(\begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ [C \& D]_{\text{sca}} \end{bmatrix} \right) \begin{bmatrix} x(t) \\ u(t) \end{bmatrix}.$$

We have obtained that $\begin{bmatrix} x(t) \\ e(t) \end{bmatrix} = E_{\text{sca}} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix}$, and hence $\begin{bmatrix} x(t) \\ e(t) \end{bmatrix} \in \mathcal{D}(S_{\text{imp}})$. Combining $\begin{bmatrix} x(t) \\ u(t) \end{bmatrix} = E_{\text{sca}}^{-1} \begin{bmatrix} x(t) \\ e(t) \end{bmatrix}$ with (5.4) we obtain

$$S_{\text{imp}} \begin{bmatrix} x(t) \\ e(t) \end{bmatrix} = \begin{bmatrix} 0 \\ -e(t) \end{bmatrix} + \begin{bmatrix} \dot{x}(t) \\ \sqrt{2}u(t) \end{bmatrix} = \begin{bmatrix} \dot{x}(t) \\ f(t) \end{bmatrix},$$

as claimed in the proposition. The converse direction is proved similarly.

The claim about the generalized solutions of the two equations follows from the corresponding relationship between the classical solutions of these equations. $\square$

The above proposition shows that the transformation described by (5.1) and (5.2) is indeed the external Cayley transformation (defined earlier via (1.11)).

PROPOSITION 5.6. *We use the notation and the assumption of Theorem 5.2. The adjoint operator S_{imp}^* has an external Cayley transform, denoted S_{sca}^d , and this is a scattering passive system node. Moreover,*

$$(5.9) \quad S_{\text{sca}}^d = \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix} S_{\text{sca}}^* \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix}.$$

Thus, S_{sca}^d is scattering conservative if and only if T is skew-adjoint.

Proof. The operator

$$(5.10) \quad T^d = \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix} S_{\text{imp}}^* = \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix} T^* \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix}$$

is m -dissipative since T is m -dissipative. Thus, by Theorem 5.2, the external Cayley transform of S_{imp}^* , denoted by S_{sca}^d , is a scattering passive system node. To establish the connection (5.9) between S_{sca}^d and S_{sca}^* we use (5.5). We denote by A_{sca}^d , B_{sca}^d , and C_{sca}^d the semigroup generator, the control operator and the observation operator of S_{sca}^d , and we denote by $\mathbf{G}_{\text{sca}}^d$ its transfer function. Then according to (5.5) we have

$$\begin{bmatrix} (I - A_{\text{sca}}^d)^{-1} & \frac{1}{\sqrt{2}}(I - A_{\text{sca}}^d)^{-1}B_{\text{sca}}^d \\ \frac{1}{\sqrt{2}}C_{\text{sca}}^d(I - A_{\text{sca}}^d)^{-1} & \frac{1}{2}(I + \mathbf{G}_{\text{sca}}^d(1)) \end{bmatrix} = (I - T^d)^{-1}.$$

Using here (5.10) and again (5.5), we obtain

$$\begin{aligned} & \begin{bmatrix} (I - A_{\text{sca}}^d)^{-1} \frac{1}{\sqrt{2}}(I - A_{\text{sca}}^d)^{-1}B_{\text{sca}}^d & \\ \frac{1}{\sqrt{2}}C_{\text{sca}}^d(I - A_{\text{sca}}^d)^{-1} & \frac{1}{2}(I + \mathbf{G}_{\text{sca}}^d(1)) \end{bmatrix} \\ &= \begin{bmatrix} (I - A_{\text{sca}}^*)^{-1} & -\frac{1}{\sqrt{2}}(I - A_{\text{sca}}^*)^{-1}C_{\text{sca}}^* \\ -\frac{1}{\sqrt{2}}B_{\text{sca}}^*(I - A_{\text{sca}}^*)^{-1} & \frac{1}{2}(I + \mathbf{G}_{\text{sca}}^*(1)) \end{bmatrix}. \end{aligned}$$

This shows that

$$(5.11) \quad A_{\text{sca}}^d = A_{\text{sca}}^*, \quad B_{\text{sca}}^d = -C_{\text{sca}}^*, \quad C_{\text{sca}}^d = -B_{\text{sca}}^*, \quad \mathbf{G}_{\text{sca}}^d(1) = \mathbf{G}_{\text{sca}}^*(1).$$

We introduce the system node $\tilde{S}$ by $\mathcal{D}(\tilde{S}) = \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix} \mathcal{D}(S_{\text{sca}})$,

$$\tilde{S} = \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix} S_{\text{sca}} \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix}.$$

It is easy to verify that the semigroup generator of $\tilde{S}$ is $\tilde{A} = A_{\text{sca}}$, its control operator is $\tilde{B} = -B_{\text{sca}}$, its observation operator is $\tilde{C} = -C_{\text{sca}}$, and its transfer function is $\tilde{\mathbf{G}}(s) = \mathbf{G}_{\text{sca}}(s)$. As mentioned after (3.7), S_{sca}^d is completely determined by A_{sca}^d , B_{sca}^d , C_{sca}^d , and $\mathbf{G}_{\text{sca}}^d(1)$. Comparing (5.11) with (3.14), we see that S_{sca}^d must be the dual of $\tilde{S}$, i.e., $S_{\text{sca}}^d = \tilde{S}^*$. From here we get that (5.9) holds.

Finally, we know from Theorem 5.2 that S_{sca}^d is scattering conservative if and only if T^d is skew-adjoint. We see from (5.10) that T^d is skew-adjoint if and only if T is. $\square$

6. The proof of Theorem 1.1. We start with a lemma about computing the adjoint of a certain type of matrix of operators, which then leads to a result about matrices of operators that are m -dissipative. Much useful material about matrices of operators can be found in the unpublished book of Engel [10] (but the specific results that we need here do not seem to follow from those in [10]). For the concept of the dual space with respect to a pivot space and related concepts we refer, for example, to section 2.9 in [39]. As in [39], the pairing between a Hilbert space V and its dual V' with respect to the pivot space H is considered to be linear in the first component and antilinear in the second, so that if both elements in the pairing belong to H , then the pairing coincides with their inner product in H .

LEMMA 6.1. *Let H_1 and H_2 be Hilbert spaces that are identified with their dual spaces. Let $\mathcal{D}(\Lambda)$ be a dense subspace of H_2 and let $\Lambda : \mathcal{D}(\Lambda) \rightarrow H_1$ be closed. We regard $\mathcal{D}(\Lambda)$ as a Hilbert space with the graph norm of Λ , and let $\mathcal{D}(\Lambda)'$ be the dual of $\mathcal{D}(\Lambda)$ with respect to the pivot space H_2 . Then $\Lambda \in \mathcal{L}(\mathcal{D}(\Lambda), H_1)$ and its adjoint (in the sense of bounded operators) is $\Lambda^* \in \mathcal{L}(H_1, \mathcal{D}(\Lambda)')$. Let $G \in \mathcal{L}(\mathcal{D}(\Lambda), \mathcal{D}(\Lambda)')$. We define*

$$T = \begin{bmatrix} 0 & -\Lambda \\ \Lambda^* & G \end{bmatrix}$$

with domain

$$\mathcal{D}(T) = \left\{ \begin{bmatrix} z \\ w \end{bmatrix} \in H_1 \times \mathcal{D}(\Lambda) \mid \Lambda^* z + Gw \in H_2 \right\}.$$

Then $\mathcal{D}(T)$ is dense in $H_1 \oplus H_2$ and the adjoint of T (as an unbounded operator on $H_1 \oplus H_2$) is

$$(6.1) \quad T^* = \begin{bmatrix} 0 & \Lambda \\ -\Lambda^* & G^* \end{bmatrix}$$

with domain

$$(6.2) \quad \mathcal{D}(T^*) = \left\{ \begin{bmatrix} z_0 \\ w_0 \end{bmatrix} \in H_1 \times \mathcal{D}(\Lambda) \mid -\Lambda^* z_0 + G^* w_0 \in H_2 \right\}.$$

Proof. It is clear that $\Lambda \in \mathcal{L}(\mathcal{D}(\Lambda), H_1)$ and its adjoint Λ^* (in the sense of bounded operators) belongs to $\mathcal{L}(H_1, \mathcal{D}(\Lambda)')$. When we regard Λ as an unbounded operator from H_2 to H_1 , then it has another adjoint (in the sense of unbounded operators) which maps $\mathcal{D}(\Lambda^*) \subset H_1$ to H_2 . The adjoint in the sense of bounded operators is an extension of the adjoint in the sense of unbounded operators, as is easy to check. For this reason, we use the same notation Λ^* for both of them. Since Λ is closed, we have that $\mathcal{D}(\Lambda^*)$ is dense in H_1 .

Now we show that $\mathcal{D}(T)$ is dense in $H_1 \oplus H_2$. Let $\begin{bmatrix} z_0 \\ w_0 \end{bmatrix} \in \mathcal{D}(T)^\perp$. Since $\mathcal{D}(\Lambda^*) \times \{0\} \subset \mathcal{D}(T)$, we must have $\langle z, z_0 \rangle = 0$ for all $z \in \mathcal{D}(\Lambda^*)$, and since $\mathcal{D}(\Lambda^*)$ is dense in H_1 we get that $z_0 = 0$. By the construction of the space $\mathcal{D}(\Lambda)'$, the operator $I + \Lambda^* \Lambda$ is a continuous bijection of $\mathcal{D}(\Lambda)$ onto $\mathcal{D}(\Lambda)'$, and consequently

$$\begin{bmatrix} I & \Lambda^* \end{bmatrix} \begin{bmatrix} I \\ \Lambda \end{bmatrix} (I + \Lambda^* \Lambda)^{-1} = I.$$

Thus, $\begin{bmatrix} I & \Lambda^* \end{bmatrix}$ maps $\mathcal{D}(\Lambda) \times H_1$ onto $\mathcal{D}(\Lambda)'$, and so for every $w \in \mathcal{D}(\Lambda)$ there exists some $w_2 \in \mathcal{D}(\Lambda)$ and $z \in H_1$ such that $Gw = w_2 - \Lambda^* z$. Hence, for every $w \in \mathcal{D}(\Lambda)$ there is some $z \in H_1$ such that $\Lambda^* z + Gw \in H_2$, i.e., $\begin{bmatrix} z \\ w \end{bmatrix} \in \mathcal{D}(T)$. Thus, the condition that $\begin{bmatrix} 0 \\ w_0 \end{bmatrix}$ is orthogonal to every $\begin{bmatrix} z \\ w \end{bmatrix} \in \mathcal{D}(T)$ implies that w_0 is orthogonal to $\mathcal{D}(\Lambda)$, and hence $w_0 = 0$. This proves that $\mathcal{D}(T)$ is dense in $H_1 \oplus H_2$.

By the definition of the adjoint of an unbounded operator, an element $x_0 = \begin{bmatrix} z_0 \\ w_0 \end{bmatrix}$ from $H_1 \oplus H_2$ belongs to $\mathcal{D}(T^*)$ if and only if the functional $F(x) = \langle Tx, x_0 \rangle$, which is defined for $x = \begin{bmatrix} z \\ w \end{bmatrix} \in \mathcal{D}(T)$, has a continuous extension to all $H_1 \oplus H_2$. We have

$$F \left(\begin{bmatrix} z \\ w \end{bmatrix} \right) = -\langle \Lambda w, z_0 \rangle_{H_1} + \langle \Lambda^* z + Gw, w_0 \rangle_{H_2}.$$

Suppose that $\begin{bmatrix} z_0 \\ w_0 \end{bmatrix} \in \mathcal{D}(T^*)$. Since for $z \in \mathcal{D}(\Lambda^*)$ and $w = 0$ we have

$$F \left(\begin{bmatrix} z \\ 0 \end{bmatrix} \right) = \langle \Lambda^* z, w_0 \rangle$$

and this is a continuous function of z , it follows that $w_0 \in \mathcal{D}(\Lambda^{**}) = \mathcal{D}(\Lambda)$. It follows that for any $\begin{bmatrix} z \\ w \end{bmatrix} \in \mathcal{D}(T)$,

$$(6.3) \quad \begin{aligned} F \left(\begin{bmatrix} z \\ w \end{bmatrix} \right) &= -\langle \Lambda w, z_0 \rangle_{H_1} + \langle \Lambda^* z + Gw, w_0 \rangle_{\mathcal{D}(\Lambda)', \mathcal{D}(\Lambda)} \\ &= \langle w, -\Lambda^* z_0 + G^* w_0 \rangle_{\mathcal{D}(\Lambda), \mathcal{D}(\Lambda)'} + \langle z, \Lambda w_0 \rangle_{H_1}. \end{aligned}$$

Since F has a continuous extension to $H_1 \oplus H_2$, it follows that $-\Lambda^* z_0 + G^* w_0 \in H_2$. Thus we have shown that

$$\mathcal{D}(T^*) \subset \left\{ \begin{bmatrix} z_0 \\ w_0 \end{bmatrix} \in H_1 \times \mathcal{D}(\Lambda) \mid -\Lambda^* z_0 + G^* w_0 \in H_2 \right\}.$$

Conversely, suppose that $x_0 = \begin{bmatrix} z_0 \\ w_0 \end{bmatrix} \in H_1 \times \mathcal{D}(\Lambda)$ is such that $-\Lambda^* z_0 + G^* w_0 \in H_2$. Then the expression in (6.3) depends continuously on $\begin{bmatrix} z \\ w \end{bmatrix} \in H_1 \oplus H_2$. Doing the computations leading to (6.3) backward, we obtain that $F(x)$ (defined for $x \in \mathcal{D}(T)$) has a continuous extension to $H_1 \oplus H_2$, so that $x_0 \in \mathcal{D}(T^*)$. Thus, we have proved (6.2). Notice that the expression in (6.3) is in fact $\langle x, T^* x_0 \rangle$ with T^* as in the lemma. This confirms the formula for T^* given in the lemma. $\square$

PROPOSITION 6.2. *With the assumptions of Lemma 6.1, assume additionally that*

$$(6.4) \quad \operatorname{Re} \langle G w_0, w_0 \rangle \leq 0 \quad \forall w_0 \in \mathcal{D}(\Lambda).$$

Then T is m -dissipative (as an unbounded operator on $H_1 \oplus H_2$).

Moreover, T is skew-adjoint if and only if we always have equality in (6.4).

Proof. According to Lemma 6.1, T^* is an operator with a similar structure as T , so that we can apply the lemma to T^* , obtaining that $\mathcal{D}(T^*)$ is dense and $T^{**} = T$. This implies that T is closed. (So far we have not used (6.4).) It is easy to check that both T and T^* are dissipative (as unbounded operators on $H_1 \oplus H_2$). By, e.g., Proposition 3.1.11 in [39], T is m -dissipative.

To prove the last part of the proposition, first we note that if T is skew-adjoint, then so is G , so that $\operatorname{Re} \langle G w_0, w_0 \rangle = 0$ for all $w_0 \in \mathcal{D}(\Lambda)$. Conversely, if G is such that we always have equality in (6.4), then we have $\langle G w_1, w_2 \rangle = -\langle w_1, G w_2 \rangle$ for all $w_1, w_2 \in \mathcal{D}(\Lambda)$. (See the proof of Proposition 3.2.2 in [39] for the easy argument.) Since G is a bounded operator from $\mathcal{D}(\Lambda)$ to $\mathcal{D}(\Lambda)'$, it follows that $G^* = -G$. Now we see from (6.1) that $T^* = -T$. $\square$

Proof of Theorem 1.1. We use the notation and the assumptions in the theorem. Introduce

$$K_0 = \frac{1}{\sqrt{2}} K, \quad \Lambda = \begin{bmatrix} L \\ K_0 \end{bmatrix};$$

then Λ is a closed operator from E to $H \oplus U$ with domain $\mathcal{D}(\Lambda) = E_0$. Notice that Λ and G satisfy the assumptions of Proposition 6.2 with $H \oplus U$ and E in place of H_1 and H_2 . According to Proposition 6.2, the operator

$$\tilde{T} = \begin{bmatrix} 0 & -\Lambda \\ \Lambda^* & G \end{bmatrix} = \begin{bmatrix} 0 & 0 & -L \\ 0 & 0 & -K_0 \\ L^* & K_0^* & G \end{bmatrix}$$

with domain

$$\mathcal{D}(\tilde{T}) = \left\{ \begin{bmatrix} z_0 \\ u_0 \\ w_0 \end{bmatrix} \in H \times U \times E_0 \mid L^* z_0 + K_0^* u_0 + G w_0 \in E \right\}$$

is m -dissipative on $H \oplus U \oplus E$. Moreover, $\tilde{T}$ is skew-adjoint if and only if G is such that we always have equality in (1.2). We interchange the places of U and E in the space $H \oplus U \oplus E$; then the entries of $\tilde{T}$ get permuted accordingly (we interchange

the last two rows and then the last two columns). This gives that the operator

$$T = \begin{bmatrix} 0 & -L & 0 \\ L^* & G & K_0^* \\ 0 & -K_0 & 0 \end{bmatrix}$$

is m-dissipative on $H \oplus E \oplus U$ with domain

$$\mathcal{D}(T) = \left\{ \begin{bmatrix} z_0 \\ w_0 \\ u_0 \end{bmatrix} \in H \times E_0 \times U \mid L^* z_0 + G w_0 + K_0^* u_0 \in E \right\}.$$

T is skew-adjoint if and only if G is such that we always have equality in (1.2).

We want to use the above T as the m-dissipative operator appearing in Theorem 5.2. For this, we partition T horizontally into the operators

$$(6.5) \quad [A \& B]_{\text{imp}} = \begin{bmatrix} 0 & -L & 0 \\ L^* & G & K_0^* \end{bmatrix}, \quad -[C \& D]_{\text{imp}} = \begin{bmatrix} 0 & -K_0 & 0 \end{bmatrix},$$

both with domain $\mathcal{D}(T)$, and we define

$$(6.6) \quad S_{\text{imp}} = \begin{bmatrix} [A \& B]_{\text{imp}} \\ [C \& D]_{\text{imp}} \end{bmatrix}, \quad \mathcal{D}(S_{\text{imp}}) = \mathcal{D}(T).$$

Then according to Theorem 5.2, the operator S_{sca} defined in (5.2) is a scattering passive system node. We partition S_{sca} as in (5.2),

$$S_{\text{sca}} = \begin{bmatrix} [A \& B]_{\text{sca}} \\ [C \& D]_{\text{sca}} \end{bmatrix}$$

and we want to express these in terms of L, K , and G . For this, according to (5.2), we have to compute the inverse (on its range) of E_{imp} from (5.1):

$$\begin{aligned} E_{\text{imp}} &= \begin{bmatrix} I & 0 \\ 0 & \frac{I}{\sqrt{2}} \end{bmatrix} \left(\begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ [C \& D]_{\text{imp}} \end{bmatrix} \right) \\ &= \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & \frac{I}{\sqrt{2}} \end{bmatrix} \left(\begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & K_0 & 0 \end{bmatrix} \right) = \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & \frac{K_0}{\sqrt{2}} & \frac{I}{\sqrt{2}} \end{bmatrix}, \end{aligned}$$

which is defined on $\mathcal{D}(T)$. It is easy to see that E_{imp} is indeed injective (as stated in Theorem 5.2) and its range $\mathcal{D}(S_{\text{sca}}) = E_{\text{imp}} \mathcal{D}(T)$ is given by (1.14). It is also easy to check that

$$E_{\text{imp}}^{-1} = \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & -K_0 & \sqrt{2}I \end{bmatrix}.$$

Substituting this and (6.5) into (5.2), we obtain that on $\mathcal{D}(S_{\text{sca}})$,

$$\begin{aligned} \begin{bmatrix} [A \& B]_{\text{sca}} \\ [C \& D]_{\text{sca}} \end{bmatrix} &= \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -I \end{bmatrix} + \begin{bmatrix} 0 & -L & 0 \\ L^* & G & K_0^* \\ 0 & 0 & \sqrt{2}I \end{bmatrix} \cdot \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & -K_0 & \sqrt{2}I \end{bmatrix} \\ &= \begin{bmatrix} 0 & -L & 0 \\ L^* & G - K_0^* K_0 & \sqrt{2}K_0^* \\ 0 & -\sqrt{2}K_0 & 2I - I \end{bmatrix} = \begin{bmatrix} 0 & -L & 0 \\ L^* & G - \frac{1}{2}K^* K & K^* \\ 0 & -K & I \end{bmatrix}. \end{aligned}$$

According to Theorem 5.2, S_{sca} is a scattering passive system node with state space

$H \oplus E$ and with input and output space U . We also know from the same theorem that S_{sca} is scattering conservative if and only if G is skew-adjoint. $\square$

Note that in the above proof, we have shown a bit more than what was required. Indeed, we have also shown (without any extra effort) that S_{sca} is scattering conservative if and only if G satisfies (1.2). This also follows from Theorem 1.3.

Remark 6.3. Our assumption (1.7) that the operator $\begin{bmatrix} L \\ K \end{bmatrix}$ is closed is not far from being a necessary condition for the conclusion of Theorem 1.1 to be valid, in the following sense. Being a system node, the operator S_{sca} is closed. As can easily be seen, this implies that also its external Cayley transform S_{imp} is closed, which implies in particular that the second column of S_{imp} is closed. This second column is $\begin{bmatrix} -L \\ G \\ K_0 \end{bmatrix}$; see (6.5) and (6.6). If, in addition, we assume that $G \in \mathcal{L}(E)$, then the domain of this column is E_0 and it follows that $\begin{bmatrix} L \\ K \end{bmatrix}$ has to be closed (as a densely defined operator from E to $H \oplus U$), which is (1.7).

Proof of Proposition 1.2. All the claims in this proposition follow from standard properties of a well-posed system node. See, in particular, (3.3), claim (h) after Definition 3.2, and Proposition 3.5. $\square$

Proof of Theorem 1.3. We shall use the operator $S_{\text{imp}} = \begin{bmatrix} [A \& B]_{\text{imp}} \\ [C \& D]_{\text{imp}} \end{bmatrix}$, whose components $[A \& B]_{\text{imp}}$ and $[C \& D]_{\text{imp}}$ are as introduced in the proof of Theorem 1.1 (see (6.5)). According to the same proof, S_{sca} is obtained from S_{imp} via the external Cayley transformation (5.1)–(5.2). If $(\begin{bmatrix} z \\ w \end{bmatrix}, u, y)$ is a classical solution of $\begin{bmatrix} \dot{z}(t) \\ \dot{w}(t) \\ y(t) \end{bmatrix} = S_{\text{sca}} \begin{bmatrix} z(t) \\ w(t) \\ u(t) \end{bmatrix}$ and if we define the signals e, f by (1.11), then by Proposition 5.5, $(\begin{bmatrix} z \\ w \end{bmatrix}, e, f)$ is a classical solution of $\begin{bmatrix} \dot{z}(t) \\ \dot{w}(t) \\ f(t) \end{bmatrix} = S_{\text{imp}} \begin{bmatrix} z(t) \\ w(t) \\ e(t) \end{bmatrix}$. Thus,

$$\begin{bmatrix} \dot{z}(t) \\ \dot{w}(t) \\ -f(t) \end{bmatrix} = \begin{bmatrix} 0 & -L & 0 \\ L^* & G & K_0^* \\ 0 & -K_0 & 0 \end{bmatrix} \begin{bmatrix} z(t) \\ w(t) \\ e(t) \end{bmatrix} \quad \forall t \geq 0.$$

By taking the real part of the inner product of both sides in this equation with the vector $[z(t) \ w(t) \ f(t)]^T$ and recalling that $w(t) \in E_0$, we get that for all $t \geq 0$,

$$\frac{d}{dt} \|z(t)\|^2 + \frac{d}{dt} \|w(t)\|^2 - 2\operatorname{Re} \langle e(t), f(t) \rangle = 2\operatorname{Re} \langle Gw(t), w(t) \rangle.$$

Here $2\operatorname{Re} \langle e(t), f(t) \rangle = \|u(t)\|^2 - \|y(t)\|^2$, and (1.18) follows.

The formula for the adjoint system follows from Proposition 5.6. The operator S_{imp}^* is of the same form as S_{imp} , apart from the fact that L and G are replaced by $-L$ and G^* . When S_{sca}^d is multiplied by $\begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix}$ to the left and right this has the effect of replacing K with $-K$. $\square$

7. The resolvent and the transfer function. In this section we continue to investigate the special class of systems described in section 1, and we prove Theorem 1.4. We begin with two preliminary propositions.

PROPOSITION 7.1. *Consider the spaces $H, E, E_0, E'_0, X_0, X'_0$ as introduced at the beginning of section 1. If the operators L, K , and G are as in (1.1), (1.2), and (1.7), then for every $s \in \mathbb{C}_+$ the operator $P(s) \in \mathcal{L}(E_0, E'_0)$ defined by*

$$(7.1) \quad P(s) = sI + \frac{1}{2}K^*K - G + \frac{1}{s}L^*L$$

has an inverse $V(s) \in \mathcal{L}(E'_0, E_0)$.

Proof. Because of condition (1.7), the space E_0 is complete with the graph norm of $\begin{bmatrix} L \\ K \end{bmatrix}$, which is

$$\|w\|_0^2 = \|w\|_E^2 + \|Lw\|_H^2 + \|Kw\|_U^2 \quad \forall w \in E_0.$$

Using the closed graph theorem, it follows that the original norm $\|\cdot\|_{E_0}$ on E_0 is equivalent to the above graph norm. In this proof, we may therefore replace $\|\cdot\|_{E_0}$ with $\|\cdot\|_0$, but we still keep the old notation $\|\cdot\|_{E_0}$ for the norm on E_0 and the notation $\|\cdot\|_{E'_0}$ for the corresponding dual norm on E'_0 .

For every $w \in E_0$ and $s \in \mathbb{C}_+$ we have

$$\begin{aligned} \|P(s)w\|_{E'_0} \cdot \|w\|_{E_0} &\geq \operatorname{Re} \langle P(s)w, w \rangle_{E'_0, E_0} \\ &= (\operatorname{Re} s) \|w\|_E^2 + \frac{1}{2} \|Kw\|_U^2 + \left(\operatorname{Re} \frac{1}{s} \right) \|Lw\|_H^2 - \langle Gw, w \rangle_{E'_0, E_0} \\ &\geq (\operatorname{Re} s) \|w\|_E^2 + \frac{1}{2} \|Kw\|_U^2 + \left(\operatorname{Re} \frac{1}{s} \right) \|Lw\|_H^2 \geq c \|w\|_{E_0}^2, \end{aligned}$$

where $c = \min\{\operatorname{Re} s, \frac{1}{2}, \operatorname{Re} \frac{1}{s}\}$. This implies that $P(s)$ is injective and has closed range. The operator

$$P(s)^* = \overline{s}I + \frac{1}{2}K^*K - G^* + \frac{1}{\overline{s}}L^*L$$

has the same structure as $P(s)$, so that by the same argument as in the first step, it is also injective and has closed range. Since $[\operatorname{Ran} P(s)]^\perp = \operatorname{Ker} P(s)^*$ (the orthogonal complement is with respect to the duality pairing; see, for instance, Theorem 4.12 in Rudin [27]), this implies that $\operatorname{Ran} P(s)$ is dense in E'_0 . Hence, $\operatorname{Ran} P(s) = E'_0$, and thus $P(s)$ has a bounded inverse, which we denote by $V(s)$. $\square$

PROPOSITION 7.2. *We use the notation and the assumptions of Proposition 7.1 and S_{sca} is the system node from Theorem 1.1. Let A be the semigroup generator of S_{sca} , let B and C be the control and observation operators of S_{sca} , and let $\mathbf{G}$ be its transfer function. Then for every $s \in \mathbb{C}_+$, we have*

$$\begin{aligned} (sI - A)^{-1} &= \begin{bmatrix} \frac{1}{s}I - \frac{1}{s^2}LV(s)L^* & -\frac{1}{s}LV(s) \\ \frac{1}{s}V(s)L^* & V(s) \end{bmatrix} \\ (7.2) \quad (sI - A)^{-1}B &= \begin{bmatrix} -\frac{1}{s}LV(s) \\ V(s) \end{bmatrix} K^*, \\ C(sI - A)^{-1} &= -K \begin{bmatrix} -\frac{1}{s}V(s)L^* & V(s) \end{bmatrix} \\ \mathbf{G}(s) &= I - KV(s)K^*, \end{aligned}$$

where $V(s) = P(s)^{-1}$ is the operator defined in Proposition 7.1.

Proof. Define S_{imp} as in (6.5) and (6.6), so that S_{imp} satisfies the assumption in Theorem 5.2. It is easy to see that on $\mathcal{D}(S_{\text{imp}})$ we have

$$\begin{bmatrix} sI & 0 \\ 0 & I \end{bmatrix} - \begin{bmatrix} [A \& B]_{\text{imp}} \\ -[C \& D]_{\text{imp}} \end{bmatrix} = \left[\begin{array}{cc|c} sI & L & 0 \\ -L^* & sI - G & -K_0^* \\ \hline 0 & K_0 & I \end{array} \right],$$

where $K_0 = \frac{1}{\sqrt{2}}K$. The operator on the right-hand side above has a natural extension to a bounded linear operator from $X_0 \oplus U$ to $X'_0 \oplus U$. This extended operator can be

factored as

$$\begin{aligned} & \left[\begin{array}{cc|c} sI & L & 0 \\ -L^* & sI - G & -K_0^* \\ \hline 0 & K_0 & I \end{array} \right] \\ &= \left[\begin{array}{cc|c} I & 0 & 0 \\ -\frac{1}{s}L^* & I & -K_0^* \\ \hline 0 & 0 & I \end{array} \right] \left[\begin{array}{cc|c} sI & 0 & 0 \\ 0 & P(s) & 0 \\ \hline 0 & 0 & I \end{array} \right] \left[\begin{array}{cc|c} I & \frac{1}{s}L & 0 \\ 0 & I & 0 \\ \hline 0 & K_0 & I \end{array} \right], \end{aligned}$$

where the first factor is a boundedly invertible operator on $X'_0 \oplus U$, the second factor is a boundedly invertible operator from $X_0 \oplus U$ to $X'_0 \oplus U$, and the last factor is a boundedly invertible operator on $X_0 \oplus U$ (we have used Proposition 7.1). By inverting the sides of the last formula, we get that

$$\begin{aligned} & \left(\begin{bmatrix} sI & 0 \\ 0 & I \end{bmatrix} - \begin{bmatrix} [A \& B]_{\text{imp}} \\ -[C \& D]_{\text{imp}} \end{bmatrix} \right)^{-1} \\ &= \left[\begin{array}{cc|c} I & -\frac{1}{s}L & 0 \\ 0 & I & 0 \\ \hline 0 & -K_0 & I \end{array} \right] \left[\begin{array}{cc|c} \frac{1}{s}I & 0 & 0 \\ 0 & V(s) & 0 \\ \hline 0 & 0 & I \end{array} \right] \left[\begin{array}{cc|c} I & 0 & 0 \\ \frac{1}{s}L^* & I & K_0^* \\ \hline 0 & 0 & I \end{array} \right]. \end{aligned}$$

We have seen in the proof of Theorem 1.1 that S_{sca} is obtained from S_{imp} via the external Cayley transformation (5.1) and (5.2). Using (5.5) (where now $A_{\text{sca}} = A$, $B_{\text{sca}} = B$, $C_{\text{sca}} = C$, and $\mathbf{G}_{\text{sca}} = \mathbf{G}$) we get (7.2). $\square$

Remark 7.3. It is well known that if A is the generator of a semigroup of contractions, then $\|(sI - A)^{-1}\| \leq \frac{1}{\operatorname{Re} s}$ for all $s \in \mathbb{C}_+$. Combining this fact with (7.2), we obtain that

$$\|V(s)\|_{\mathcal{L}(E)} \leq \frac{1}{\operatorname{Re} s} \quad \forall s \in \mathbb{C}_+.$$

Remark 7.4. With the assumptions and the notation of Proposition 7.2, the space Z from (3.8), (3.9) is a subspace of X_0 with continuous embedding. Moreover, the operator $\tilde{C}$ from (1.3), when restricted to Z , is in $\mathcal{L}(Z, U)$. All this is easy to prove using Proposition 7.1 and the expression of $(I - A)^{-1}B$ from (7.2).

Proof of Theorem 1.4. The given formulas for the operators A and C follow directly from (1.13), (3.2), and (3.4).

Now we prove (1.20). The continuity and density of the embeddings $X_0 \subset X \subset X'_0$ follow directly from the continuity and density of the embeddings $E_0 \subset X \subset E'_0$.

By the definition of $\mathcal{D}(A)$ we have that $X_1 \subset X_0$. We claim that X_1 is dense in X_0 . To prove this assume that $\begin{bmatrix} z' \\ w' \end{bmatrix} \in X'_0$ and $\langle \begin{bmatrix} z' \\ w' \end{bmatrix}, (I - A)^{-1} \begin{bmatrix} z \\ w \end{bmatrix} \rangle_{X'_0, X_0} = 0$ for all $\begin{bmatrix} z \\ w \end{bmatrix} \in X$. We have to show that this implies $\begin{bmatrix} z' \\ w' \end{bmatrix} = 0$. By (7.2),

$$\begin{aligned} & \left\langle \begin{bmatrix} z' \\ w' \end{bmatrix}, (I - A)^{-1} \begin{bmatrix} z \\ w \end{bmatrix} \right\rangle_{X'_0, X_0} \\ &= \left\langle \begin{bmatrix} z' \\ w' \end{bmatrix}, \begin{bmatrix} I - LV(1)L^* & -LV(1) \\ V(1)L^* & V(1) \end{bmatrix} \begin{bmatrix} z \\ w \end{bmatrix} \right\rangle_{X'_0, X_0} \\ &= \langle z', z - LV(1)L^*z - LV(1)w \rangle_H + \langle w', V(1)L^*z + V(1)w \rangle_{E'_0, E_0} \\ &= \langle z' - LV(1)^*L^*z' + LV(1)^*w', z \rangle_H + \langle -V(1)^*L^*z' + V(1)^*w', w \rangle_E. \end{aligned}$$

Thus $z' - LV(1)^*L^*z' + LV(1)^*w' = 0$ and $-V(1)^*L^*z' + V(1)^*w' = 0$. The second of these equations gives $L^*z' = w'$, which substituted in the first equation gives $z' = 0$. Thus, $\begin{bmatrix} z' \\ w' \end{bmatrix} = 0$ and this proves that X_1 is dense in X_0 .

Since both X_1 and X_0 are continuously embedded in X , it follows that the embedding from X_1 to X_0 is a closed operator from X_1 to X_0 . By the closed graph theorem, the embedding $X_1 \hookrightarrow X_0$ is continuous.

Recall that A^* has the same structure as A with L , K , and G replaced with $-L$, $-K$, and G^* . Therefore, by the same argument as above, if we replace A by A^* and denote the domain of A^* by X_1^d , then X_1^d is continuously and densely embedded in X_0 . Recall from Proposition 2.10.2 in [39] that X_{-1} is the dual of X_1^d with respect to the pivot space X . Therefore, by duality X_0' is densely and continuously embedded in X_{-1} . This completes the proof of (1.20).

Let $A_{-1} \in \mathcal{L}(X, X_{-1})$ be the usual extension of the generator A , as explained at the beginning of section 3. Since X_0 is continuously embedded in X , the restriction of A_{-1} to X_0 is in $\mathcal{L}(X_0, X_{-1})$. Since X_0' is continuously embedded in X_{-1} , the operator $\overline{A}$ may also be regarded as an operator in $\mathcal{L}(X_0, X_{-1})$. Since the restrictions of these two operators to X_1 are equal and since X_1 is dense in X_0 , it follows that these two operators are in fact equal, as stated in the theorem.

The two remaining claims about B and $\mathbf{G}(s)$ follow from Proposition 7.2. $\square$

REFERENCES

- [1] D.Z. AROV, M. KURULA, AND O.J. STAFFANS, *Canonical state/signal shift realizations of passive continuous time behaviors*, Complex Anal. Oper. Theory, 5 (2011), pp. 331–402.
- [2] D.Z. AROV, M. KURULA, AND O.J. STAFFANS, *Boundary control state/signal systems and boundary triplets*, in Operator Methods for Boundary Value Problems, S. Hassi, H.S.V. de Snoo, and F.H. Szafraniec, eds., Cambridge University Press, Cambridge, UK, 2012.
- [3] D.Z. AROV, M. KURULA, AND O.J. STAFFANS, *Passive state/signal systems and conservative boundary relations*, in Operator Methods for Boundary Value Problems, S. Hassi, H.S.V. de Snoo, and F.H. Szafraniec, eds., Cambridge University Press, Cambridge, UK, 2012.
- [4] D.Z. AROV AND M.A. NUDELMAN, *Passive linear stationary dynamical scattering systems with continuous time*, Integral Equations Operator Theory, 24 (1996), pp. 1–45.
- [5] J.A. BALL AND O.J. STAFFANS, *Conservative state-space realizations of dissipative system behaviors*, Integral Equations Operator Theory, 54 (2006), pp. 151–213.
- [6] J. BEHRNDT, S. HASSI, AND H. DE SNOO, *Boundary relations, unitary colligations, and functional models*, Complex Anal. Oper. Theory, 3 (2009), pp. 57–98.
- [7] M. DALSMO AND A.J. VAN DER SCHAFT, *On representations and integrability of mathematical structures in energy-conserving physical systems*, SIAM J. Control Optim., 37 (1999), pp. 54–91.
- [8] R. DAUTRAY AND J.-L. LIONS, *Mathematical Analysis and Numerical Methods for Science and Technology. Vol. 3: Spectral Theory and Applications*, Springer-Verlag, Berlin, 1990.
- [9] V. DERKACH, S. HASSI, M. MALAMUD, AND H. DE SNOO, *Boundary relations and their Weyl families*, Trans. Amer. Math. Soc., 358 (2006), pp. 5351–5400.
- [10] K.J. ENGEL, *Operator Matrices and Systems of Evolution Equations*, manuscript, L'Aquila, Italy, 1999.
- [11] M. ELLER, J.E. LAGNESE, AND S. NICAISE, *Stabilization of heterogeneous Maxwell's equations by linear or nonlinear boundary feedbacks*, Electron. J. Differential Equations, 2002 (2002), pp. 1–26.
- [12] K.J. ENGEL AND R. NAGEL, *One-Parameter Semigroups for Linear Evolution Equations*, Grad. Texts in Math. 194, Springer-Verlag, New York, 2000.
- [13] H.O. FATTORINI, *Second Order Linear Differential Equations in Banach Spaces*, North-Holland, Amsterdam, 1985.
- [14] S. HASSI, M. MALAMUD, AND V. MOGILEVSKII, *Generalized resolvents and boundary triplets for dual pairs of linear relations*, Methods Funct. Anal. Topology, 11 (2005), pp. 170–187.

- [15] B. JACOB, K.A. MORRIS, AND C. TRUNK, *Minimum-phase behaviour of damped second-order systems with position measurements*, IEEE Trans. Automat. Control, 52 (2007), pp. 1654–1665.
- [16] B. JACOB AND K.A. MORRIS, *Second-order systems with acceleration measurements*, IEEE Trans. Automat. Control, 57 (2012), pp. 690–700.
- [17] M. KURULA, *On passive and conservative state/signal systems in continuous time*, Integral Equations Oper. Theory, 67 (2010), pp. 377–424, with erratum on p. 449.
- [18] M. KURULA AND O.J. STAFFANS, *Well-posed state/signal systems in continuous time*, Complex Anal. Oper. Theory, 4 (2009), pp. 319–390.
- [19] M. KURULA, H. ZWART, A. VAN DER SCHAFT, AND J. BEHRNDT, *Dirac structures and their composition on Hilbert spaces*, J. Math. Anal. Appl., 372 (2010), pp. 402–422.
- [20] J.E. LAGNESE, *A singular perturbation problem in exact controllability of the Maxwell system*, ESAIM Control Optim. Calc. Var., 6 (2001), pp. 275–289.
- [21] M.S. LIVSHITS, *On linear physical systems connected with the external world by coupling channels*, Izv. Akad. Nauk. SSSR Ser. Matem. 27 (1963), pp. 993–1030 (in Russian).
- [22] A. MACCHELLI, S. STRAMIGIOLI, A. VAN DER SCHAFT, AND C. MELCHIORRI, *Scattering for infinite dimensional port Hamiltonian systems*, in Proceedings of the CDC02, Las Vegas, 2002.
- [23] J. MALINEN AND O.J. STAFFANS, *Impedance passive and conservative boundary control systems*, Complex Anal. Oper. Theory, 1 (2007), pp. 279–300.
- [24] J. MALINEN, O.J. STAFFANS, AND G. WEISS, *When is a linear system conservative?*, Quart. Appl. Math., 64 (2006), pp. 61–91.
- [25] R. OBER AND S. MONTGOMERY-SMITH, *Bilinear transformation of infinite-dimensional state-space systems and balanced realisations of nonrational transfer functions*, SIAM J. Control Optim., 28 (1990), pp. 438–465.
- [26] K.D. PHUNG, *Contrôle et stabilisation d’ondes électromagnétiques*, ESAIM Control, Optim. Calc. Var., 5 (2000), pp. 87–137.
- [27] W. RUDIN, *Functional Analysis*, McGraw-Hill, New York, 1973.
- [28] A.J. VAN DER SCHAFT AND B.M. MASCHKE, *Hamiltonian formulation of distributed-parameter systems with boundary energy flow*, J. Geometry Phys., 42 (2002), pp. 166–194.
- [29] R. SCHNAUBELT AND G. WEISS, *Two classes of passive time-varying well-posed linear systems*, Math. Control, Signals Systems, 21 (2010), pp. 265–301.
- [30] O.J. STAFFANS, *J-energy preserving well-posed linear systems*, Appl. Math. Comput. Sci., 11 (2001), pp. 1361–1378.
- [31] O.J. STAFFANS, *Passive and conservative continuous-time impedance and scattering systems. Part I: Well-posed systems*, Math. Control Signals Systems, 15 (2002), pp. 291–315.
- [32] O.J. STAFFANS, *Passive and conservative infinite-dimensional impedance and scattering systems (from a personal point of view)*, in Mathematical Systems Theory in Biology, Communication, Computation, and Finance, IMA Vol. Math. Appl. 134, Springer-Verlag, New York, 2002, pp. 375–414.
- [33] O.J. STAFFANS, *Stabilization by collocated feedback*, in Directions in Mathematical Systems Theory and Optimization, Lecture Notes in Control and Inform. Sci. 286, A. Rantzer and C.I. Byrnes, eds., Springer-Verlag, Berlin, 2002 pp. 261–278.
- [34] O.J. STAFFANS, *Well-Posed Linear Systems*, Cambridge University Press, Cambridge, UK, 2004.
- [35] O.J. STAFFANS AND G. WEISS, *Transfer functions of regular linear systems. Part II: The system operator and the Lax-Phillips semigroup*, Trans. American Math. Soc., 354 (2002), pp. 3229–3262.
- [36] O.J. STAFFANS AND G. WEISS, *Transfer functions of regular linear systems. Part III: Inversions and duality*, Integral Equations Oper. Theory, 49 (2004), pp. 517–558.
- [37] B. SZ.-NAGY AND C. FOIAŞ, *Harmonic Analysis of Operators on Hilbert Space*, North-Holland, Amsterdam, 1970.
- [38] M. TUCSNAK AND G. WEISS, *How to get a conservative well-posed linear system out of thin air, Part II: Controllability and stability*, SIAM J. Control Optim., 42 (2003), pp. 907–935.
- [39] M. TUCSNAK AND G. WEISS, *Observation and Control for Operator Semigroups*, Birkhäuser-Verlag, Basel, 2009.
- [40] J. VILLEGAS, *A Port-Hamiltonian Approach to Distributed Parameter Systems*, Ph.D. thesis, University of Twente, Netherlands, 2007.
- [41] G. WEISS, *Transfer functions of regular linear systems. Part I: Characterizations of regularity*, Trans. Amer. Math. Soc., 342 (1994), pp. 827–854.
- [42] G. WEISS, *Optimal control of systems with a unitary semigroup and with colocated control and observation*, Systems Control Lett., 48 (2003), pp. 329–340.

- [43] G. WEISS, *A special class of conservative linear systems comprising the Maxwell equations*, in Proceedings of the MTNS Symposium, Leuven, Belgium, 2004.
- [44] G. WEISS AND O.J. STAFFANS, *Maxwell's Equations as a Scattering Passive Linear System*, <http://users.abo.fi/staffans> (2012).
- [45] G. WEISS, O.J. STAFFANS, AND M. TUCSNAK, *Well-posed linear systems -a survey with emphasis on conservative systems*, Appl. Math. Comput. Sci., 11 (2001), pp. 101–127.
- [46] G. WEISS AND M. TUCSNAK, *How to get a conservative well-posed linear system out of thin air, Part I: Well-posedness and energy balance*, ESAIM Control Optim. Calc. Var., 9 (2003), pp. 247–274.
- [47] P.F. YAO AND G. WEISS, *Global smooth solutions and exponential stability for a nonlinear beam*, SIAM J. Control Optim., 45 (2007), pp. 1931–1964.

Modeling Reveals That Dynamic Regulation of c-FLIP Levels Determines Cell-to-Cell Distribution of CD95-mediated Apoptosis*

Received for publication, August 19, 2010, and in revised form, February 11, 2011. Published, JBC Papers in Press, February 15, 2011, DOI 10.1074/jbc.M110.177097

Hannu T. Toivonen[‡], Annika Meinander[§], Tomoko Asaoka^{§||1}, Mia Westerlund[‡], Frank Pettersson[¶],
Andrey Mikhailov[§], John E. Eriksson^{§||2}, and Henrik Saxén^{¶13}

From the Departments of [‡]Information Technologies, [§]Biosciences and [¶]Chemical Engineering, Åbo Akademi University, FI-20500 Turku, Finland and ^{||}Turku Centre for Biotechnology, Åbo Akademi University and University of Turku, FI-20520 Turku, Finland

The expression levels of caspase-8 inhibitory c-FLIP proteins play an important role in regulating death receptor-mediated apoptosis, as their concentration at the moment when the death-inducing signaling complex (DISC) is formed determines the outcome of the DISC signal. Experimental studies have shown that c-FLIP proteins are subject to dynamic turnover and that their stability and expression levels can be rapidly altered. Even though the influence of c-FLIP on the apoptotic behavior of a single cell has been captured in mathematical simulation studies, the effect of c-FLIP turnover and stability has not been investigated. In this study, a mathematical model of apoptosis was developed to analyze how the dynamic turnover and stability of the c-FLIP isoforms regulate apoptotic signaling for both individual cells and cell populations. Intercellular parameter and concentration distributions were used to describe the behavior of cell populations. Monte-Carlo simulations of cell populations showed that c-FLIP turnover is a key determinant of death receptor responses. The fact that the developed model simulates the state of whole cell populations makes it possible to validate it by comparison with empirical data. The proposed modeling approach can be used to further determine limiting factors in the DISC signaling process.

Apoptosis, or programmed cell death, is a mechanism of active controlled cellular elimination, which, together with cell growth and division, regulates key processes such as embryonic development, immune responses, and tissue homeostasis. Upon stimulation, death receptors such as Fas/CD95/Apo-1 are activated and recruit cytoplasmic proteins, forming a death-inducing signaling complex (DISC).⁴ In the DISC, the initiator caspase-8 can be activated by cleavage of procaspase-8, resulting in downstream activation of effector caspases and propaga-

tion of the apoptotic signal. Caspase-8 activation is regulated by the c-FLIP (cellular FLICE inhibitory protein) inhibitory proteins, which exist in three splice variants, a long form (c-FLIP_L) and two short forms (c-FLIP_S and c-FLIP_R). The DISC-interacting proteins procaspase-8 and c-FLIP_S all contain two death effector domains, but c-FLIP_S and c-FLIP_R lack the longer protease-like domain of procaspase-8 and c-FLIP_L (1). The c-FLIP isoforms have been shown to primarily delay or inhibit death receptor-mediated apoptosis (2), although the long c-FLIP variant is known to have a dual function and may also act as a caspase activator (3–6). Recently, Fricker *et al.* (7) proposed an underlying mechanism for the dual role of c-FLIP_L and developed a mathematical model, which was verified experimentally.

Experiments on human K562 erythroleukemia cells and WM35 melanoma cells (8) demonstrated that c-FLIP_S is more apt to degradative ubiquitylation and consequent degradation than c-FLIP_L. Furthermore, the anti-apoptotic activity of c-FLIP_S and its ubiquitin-mediated degradation are functionally and structurally independent of each other (8). The destabilizing sequence in c-FLIP_S implies distinct degradation rates of the c-FLIP isoforms, allowing differential regulation of their concentration levels.

It has also been found that heat shock sensitizes cells to CD95-mediated apoptosis in Jurkat and HeLa cells, independently of the expression of Hsp70 (9, 10). The study demonstrated that a 30-min heat shock treatment is sufficient to decrease the levels of c-FLIP_L and c-FLIP_S and to sensitize the cells to apoptosis mediated by death receptors. Hyperthermia also sensitizes activated primary human T lymphocytes to CD95-induced apoptosis, primarily due to induced down-regulation of c-FLIP_S (10). It was also demonstrated that the hyperthermia does not affect the synthesis of c-FLIP but rather rapidly enhances the proteasomal degradation of c-FLIP_L and c-FLIP_S.

Mathematical modeling and *in silico* analysis of biological systems have become important tools for simulation-based hypothesis testing (11–14). The apoptotic signaling pathway has a complex dynamic behavior involving several reactions and feedback loops, and it may be difficult to infer its overall response by examining individual parts of the system separately. Mathematical modeling and simulation can therefore provide valuable insights into the behavior of apoptotic signaling pathways. A number of mathematical models of apoptosis

* This work was in part supported by the Finnish Funding Agency for Technology and Innovation and Åbo Akademi University, the Academy of Finland, the Sigrid Jusélius Foundation, Finnish Cancer Foundations, and the Research Institute of the Åbo Akademi University Foundation.

¹ Supported by the Turku Graduate School for Biomedical Sciences.

² To whom correspondence may be addressed: Dept. of Biosciences, Åbo Akademi University, Tykistökatu 6B, FI-20500 Turku, Finland. Fax: 358-2-333-8000; E-mail: john.eriksson@abo.fi.

³ To whom correspondence may be addressed: Dept. of Chemical Engineering, Åbo Akademi University, Piispankatu 8, FI-20500 Turku, Finland. Fax: 358-2-215-4792; E-mail: henrik.saxen@abo.fi.

⁴ The abbreviation used is: DISC, death-inducing signaling complex.

have been proposed, including ordinary differential equation models (15–22), stochastic models (18, 23), and Petri net models (24, 25). One of the most detailed differential equation approaches was presented by Bentele *et al.* (18). The model predicted CD95 signaling to be characterized by a threshold behavior, which depends on c-FLIP concentration in the DISC and the initial concentration of the death ligand CD95L (6).

Experimental studies (8–10) indicate that c-FLIP proteins are subject to active ubiquitylation, thereby yielding high turnover rates. Consequently, their expression levels can change rapidly when the ubiquitylation and stability of c-FLIP are affected. In this study, mathematical modeling was used to mimic experimental studies of apoptosis in which c-FLIP levels are dynamically regulated. Synthesis and degradation of the isoforms c-FLIP_S and c-FLIP_L were introduced within the framework of the apoptosis model of Bentele *et al.* (18), and it is demonstrated that the dynamic regulation of c-FLIP plays an important role for apoptosis. A problem in verifying these findings is that experimental studies of apoptosis usually report the proportion of cell deaths during a given time interval. To overcome this limitation of single-cell models, we extended the model to take into account the fact that cells are not identical copies of each other (26) but may show differences in size, intercellular structure, and protein concentrations. We studied cell populations in which the reaction rate constants and initial protein concentrations in the various cells are statistically distributed. The statistical treatment and Monte Carlo simulations made it possible to compare the outcome of the simulations with results from empirical studies, in which the proportion of apoptotic cells is reported. This approach was also employed to study the effect of down- and up-regulation of c-FLIP under conditions similar to those used in the previously reported heat shock experiments (9, 10). To mimic experimental studies of apoptosis in which c-FLIP levels are dynamically regulated and in which this regulation is a key factor for the outcome (6, 8, 10), the model was here modified to encompass dynamic c-FLIP turnover.

The results corroborate the role of c-FLIP-targeted protein degradation as a fundamental determinant of death receptor responses. A further asset of the developed model is the compatibility with cell population data, enabling convenient communication between empirical and modeled data. This feature was employed to validate the modeling results using empirical cell population data. The model can be used to study, apart from ubiquitylation-mediated degradation, other rate-limiting steps in DISC signaling such as the recently defined isoform-specific phosphorylation processes of c-FLIP (27, 28) that partly regulate ubiquitylation and partly regulate other aspects.

MATERIALS AND METHODS

Mathematical Model—The mathematical modeling of apoptosis originated from the approach of Bentele *et al.* (18), in which the apoptotic signaling process is described using ordinary differential equations for 41 chemical components in 32 reactions. The reaction rates are modeled by mass action and Michaelis-Menten kinetics. The model exhibits a threshold mechanism in apoptotic behavior, which depends on the stationary c-FLIP concentration in the DISC and the initial con-

centration of ligand. The model parameters were determined using experiments performed on the human B lymphoblastoid cell line SKW 6.4. To fit experimental data, the model incorporates both the long and short c-FLIP isoforms to account for their different reaction mechanisms. The two short c-FLIP isoforms, c-FLIP_S and c-FLIP_R, have been shown to affect caspase-8 activation and apoptosis sensitivity in a similar way (8, 10, 27, 33), and they have similar half-lives. As the c-FLIP property that is important for this study is its half-life, and as the half-lives of the short forms of c-FLIP, *i.e.* c-FLIP_S and c-FLIP_R, are the same in all the model cell lines (8, 10, 27, 33), we can assume that c-FLIP_S and c-FLIP_R will behave in the same way. Therefore, no distinction has been made between the two short isoforms in this study. Instead, both short forms are included as one factor, c-FLIP_{S/R}.

Besides its anti-apoptotic role, the long c-FLIP isoform can also act as an activator of caspase-8 (4, 5). Experimental evidence supported by mathematical modeling shows that c-FLIP_L has a pro-apoptotic role only at high ligand concentration or upon overexpression of one of the short c-FLIP isoforms (7). The pro-apoptotic roles of c-FLIP_L have been shown only under conditions of ectopic overexpression. These conditions are, however, not relevant for this study, in which the focus is to investigate apoptosis in cell populations in which the endogenous c-FLIP levels are down-regulated due to increased turnover and/or proteasomal degradation. We have therefore used the mechanism presented in Ref. 18, in which the pro-apoptotic role of c-FLIP_L is excluded.

Dynamic Regulation of c-FLIP Levels—In the mathematical model presented by Bentele *et al.* (18), the simplifying assumption was made that, upon stimulation, only the c-FLIP initially present in the cell takes part in the reactions, whereas the effect of synthesis and degradation of c-FLIP is not taken into account. However, the reported high turnover rate of c-FLIP (8, 10) implies that c-FLIP bound in the DISC is rapidly replenished with newly synthesized c-FLIP, tending to counteract changes in c-FLIP levels. Moreover, a number of experimental reports (8, 10, 27) indicate that regulation of the stability and consequent concentration levels of c-FLIP plays an important role in controlling apoptotic sensitivity.

To describe the experimentally found high turnover rate of c-FLIP and the fact that the levels of c-FLIP_{S/R} and c-FLIP_L are dynamically regulated by affecting degradation rates, the rate of c-FLIP degradation was modeled as a first-order reaction (Equation 1),

$$r_{\text{deg}}^A(t) = k_{\text{deg}}^A[\text{c-FLIP}_A](t) \quad (A = \text{S/R, L}) \quad (\text{Eq. 1})$$

where [c-FLIP](*t*) denotes the concentration of c-FLIP at time *t*, and $k_{\text{deg}}^{\text{S/R}}$ and $k_{\text{deg}}^{\text{L}}$ are the degradation rate constants of the two c-FLIP isoforms considered in the model. The degradation rates can be estimated from experimentally determined half-lives ($t_{1/2}^A$) as $k_{\text{deg}}^A = \ln 2/t_{1/2}^A$. Poukkula *et al.* (8) determined the half-lives of c-FLIP_{S/R} and c-FLIP_L in human K562 erythroleukemia cells by knocking out *de novo* synthesis and found them to be $t_{1/2}^{\text{S/R}} = 40$ min and $t_{1/2}^{\text{L}} = 2$ h, respectively, corresponding to the degradation rate constants $k_{\text{deg}}^{\text{S/R}} = 0.0173 \text{ min}^{-1}$ and $k_{\text{deg}}^{\text{L}} = 0.0058 \text{ min}^{-1}$.

At steady state, the synthesis rates must equal the degradation rates, so it follows that the synthesis rates $r_{\text{syn}}^{\text{S/R}}$ and $r_{\text{syn}}^{\text{L}}$ are related to the steady-state concentrations $[\text{c-FLIP}_{\text{S/R}}]_{\text{ss}}$ and $[\text{c-FLIP}_{\text{L}}]_{\text{ss}}$ as in Equation 2.

$$r_{\text{syn}}^A = k_{\text{deg}}^A [\text{c-FLIP}_A]_{\text{ss}} \quad (A = \text{S/R, L}) \quad (\text{Eq. 2})$$

Equation 2 gives a quantitative form for the results presented by Poukkula *et al.* (8), *i.e.* if the synthesis rate is constant, the expression level of c-FLIP can be efficiently regulated by affecting the stability of c-FLIP, defined by the degradation rate constant.

Time to Apoptosis—In the study by Bentele *et al.* (18), poly(ADP-ribose) polymerase cleavage indicated the occurrence of apoptosis, and the time of poly(ADP-ribose) polymerase cleavage was used to determine relative times to apoptosis. It turns out, however, that this characterization of the time to apoptosis gives consistently too short times compared with experimental results (6, 8, 10) in which different approaches using flow cytometry had been used to determine the occurrence of apoptosis. Harrington *et al.* (22) used the maximum value of activated caspase-3 as an indicator of apoptosis. To introduce a definition that would better reflect the time at which apoptosis is detected, the time to apoptosis (t_{ap}) was taken as the time elapsed from exposure to ligand to the instant when the concentration of activated caspase-3 had decreased to half of its maximum value. In addition, the original model (18) was modified by assuming that apoptosis does not affect ligands in the cell exterior, receptors on the cell membrane, or ligand-receptor complexes.

Modeling of Cell Populations—For a given cell type, individual cells may show large variations in apoptotic sensitivity. Experimental studies of apoptosis therefore often report the proportions of cells that undergo apoptosis during given time intervals. In contrast, mathematical models of apoptosis based on differential equations describe the behavior of a single cell with specified kinetic parameters and molecular concentrations. Therefore, simulation results from differential equation models are not directly compatible with experimental data, making both model validation and use for prediction purposes difficult.

There are two major sources of random variations in cellular reactions. The first one is intrinsic noise (29) caused by the inherent stochastic nature of chemical reactions. The intrinsic noise causes large fluctuations when the numbers of the relevant molecules are small, but its effect decreases rapidly with increasing concentrations. Stochastic models, which take into account the intrinsic noise, describe the reactions at a molecular level (30). The second important source of randomness in cellular processes stems from variations between individual cells, which may differ both in the concentrations of various molecules and in the chemical reaction rates. Such variations can be described by modeling a cell population in which the parameters in the cells are randomly distributed. The behavior of the cell population can be studied using Monte Carlo simulations, in which the concentrations and kinetic parameters are selected randomly from the associated distributions. This approach has been applied to investigate robustness properties

of biochemical networks (31, 32), including apoptosis models (23).

To account for differences between the behavior of different cells in a population, we introduced distributions on initial concentrations, $[X_i](0)$, and chemical reaction rate constants, k_i . Randomly distributed parameters that take only positive values can be modeled using the log-normal distribution (23), in which the logarithm of a parameter is normally distributed. Also measurements have indicated log-normal distributions of concentrations in cells (26). We thus assumed the initial concentrations and the rate constants in the cells to be independent stochastic variables given by Equations 3 and 4,

$$[X_i](0) = [X_i]^0 e^{\epsilon_i} \quad (\epsilon_i \sim N(0, \sigma_{\epsilon_i})) \quad (\text{Eq. 3})$$

$$k_i = k_i^0 e^{\eta_i} \quad (\eta_i \sim N(0, \sigma_{\eta_i})) \quad (\text{Eq. 4})$$

where $N(0, \sigma)$ denotes a normally distributed stochastic variable with zero mean and variance σ , and $[X_i]^0$ and k_i^0 are the nominal initial concentrations and rate factors, respectively, which were here taken as those given in the study by Bentele *et al.* (18). The model was programmed in MATLAB (The MathWorks, Inc., Natick, MA).

RESULTS

We used the modified apoptosis model to study how c-FLIP turnover and intercellular variability affect death receptor-mediated apoptosis. First, the effect of the synthesis and degradation rates of c-FLIP on the apoptotic process was analyzed. Then, intercellular variability was introduced to describe the apoptotic behavior of cell populations. Finally, the cell population model with dynamic regulation of c-FLIP levels was used to reproduce reported experimental results (10), studying recovery of a cell population that has been sensitized to apoptosis by a heat shock, followed by exposure to ligands after various time periods of recovery.

Effect of c-FLIP Turnover and Degradation Rate—The effect of c-FLIP turnover in the nominal (single-cell) model is shown in Fig. 1A, which depicts the concentrations of some key components in the apoptotic process for different turnover rates. The initial (steady-state) concentration of c-FLIP was taken to be the same as the initial concentration in the previously reported model (18), *i.e.* $[\text{c-FLIP}_{\text{S/R}}](0) = [\text{c-FLIP}_{\text{L}}](0) = 0.2937$. Ligand exposure occurred at time $t = 0$, with an initial ligand concentration of 200 ng/ml, corresponding to a ligand/receptor ratio of 0.2. The *upper panels* of Fig. 1A illustrate the response obtained with the original model (18), in which neither synthesis nor degradation of c-FLIP is taken into account. The levels of free c-FLIP decrease as c-FLIP is bound to the DISC, and no new molecules are synthesized. The simulation indicates cell death execution at $t_{\text{ap}} = 141$ min. The corresponding simulation using the c-FLIP turnover rates reported by Poukkula *et al.* (8) and Meinander *et al.* (10) is shown in the *middle panels* of Fig. 1A. Although the c-FLIP levels are now better maintained by synthesis of new molecules, c-FLIP turnover has only a minor effect on the apoptotic behavior or on the time to apoptosis ($t_{\text{ap}} = 144$ min). To confirm that the apoptotic signaling depends mainly on the initial (steady-state) concentration of c-FLIP and not on the replacement of consumed

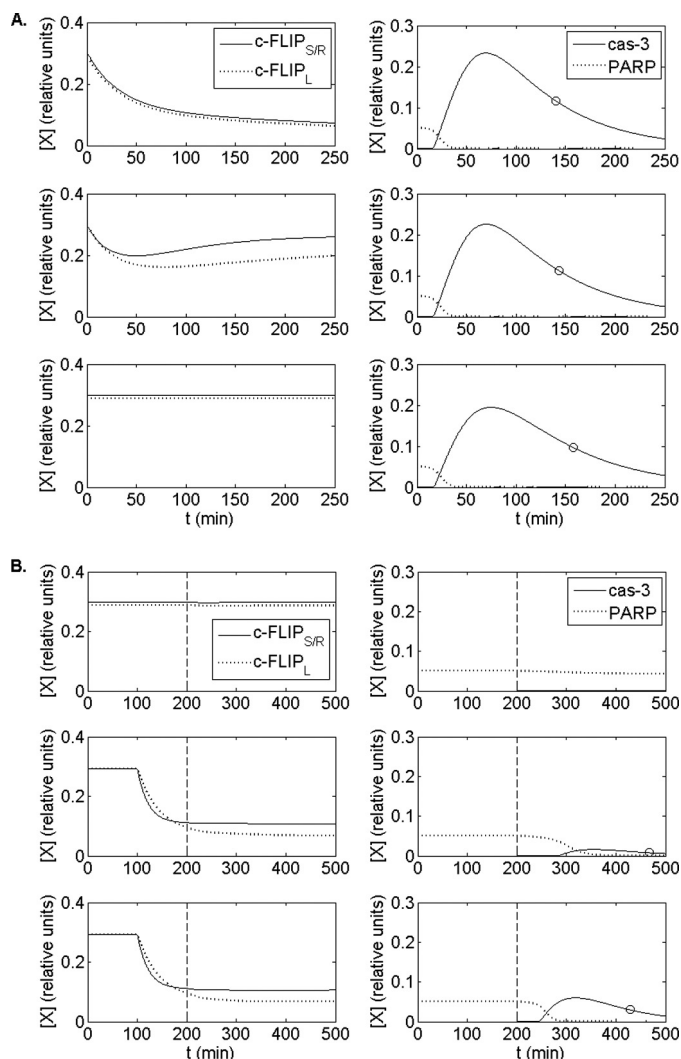


FIGURE 1. Single-cell simulations with c-FLIP turnover and regulation. Concentrations of c-FLIP_{S/R} and c-FLIP_L (left panels) and activated caspase-3 (cas-3) and poly(ADP-ribose) polymerase (PARP) (right panels) are depicted as functions of time when using the nominal initial concentrations and reaction rate parameters of the model in the study by Bentele *et al.* (18) with [L](0) = 200 ng/ml (ligand/receptor = 0.2). t_{ap} is indicated by circles. **A**, simulations at various c-FLIP turnover rates. *Upper panels*, no synthesis or degradation of c-FLIP. The concentrations of both c-FLIP isoforms are identical (indicated by slightly shifted curves for the purpose of illustration), as the reaction rate constants that describe their binding to the DISC are the same in the nominal model (18). t_{ap} = 141 min. *Middle panels*, synthesis and degradation of c-FLIP. The degradation rates of c-FLIP_{S/R} and c-FLIP_L correspond to half-lives of 40 min and 2 h, respectively. The initial steady-state c-FLIP concentrations are the same as reported by Bentele *et al.* (18). t_{ap} = 144 min. *Lower panels*, infinite turnover rate of c-FLIP, with the same steady-state c-FLIP concentrations as reported previously (18). t_{ap} = 158 min. **B**, simulation considering synthesis and degradation of c-FLIP. The receptors were exposed to ligands at $t = 200$ min (indicated by dashed vertical lines), with [L](200) = 5 ng/ml. *Upper panels*, degradation rates of c-FLIP_{S/R} and c-FLIP_L corresponding to half-lives of 40 min and 2 h, respectively (no apoptosis). *Middle panels*, down-regulation of c-FLIP. Degradation rates of free c-FLIP_{S/R} and c-FLIP_L correspond to half-lives of 40 min and 2 h, respectively, at $t < 100$ min and are changed to correspond to half-lives of 15 and 30 min at $t = 100$ min. t_{ap} = 268 min after exposure to ligands. *Lower panels*, same as the middle panels but with DISC-bound c-FLIP being degraded 100 times faster than free c-FLIP. t_{ap} = 228 min after exposure to ligands.

c-FLIP by synthesis, the responses for infinite turnover rates ($r_{syn}^{S/R}$ and $r_{syn}^L \rightarrow \infty$) were simulated and are illustrated in the lower panels of Fig. 1A. These results show that, although the dynamic responses of the c-FLIP concentrations depend

strongly on the turnover rates, they have little effect on the concentrations of caspases associated with the DISC or on the apoptotic process. The results are similar for other initial ligand and c-FLIP concentrations (data not shown). Hence, the apoptotic behavior of the model can be concluded to depend primarily on the concentration of c-FLIP during DISC formation, whereas the turnover rate has a much smaller impact.

For a constant synthesis rate, the steady-state concentration of c-FLIP (and hence, apoptotic sensitivity) is affected by the degradation rate of c-FLIP (*cf.* Equation 2). This is illustrated in Fig. 1B, which depicts the results of simulations in which the death receptors were activated by death ligands at 5 ng/ml (ligand/receptor = 0.005) at $t = 200$ min. In the upper panels of Fig. 1B, the c-FLIP turnover rates and levels are the ones reported by Meinander *et al.* (10) (Fig. 1A, middle panels), and no apoptosis is triggered. The middle panels of Fig. 1B show the effect of increased degradation rates of c-FLIP, *e.g.* caused by heat shock (10). Following the experiments reported by Meinander *et al.* (10), the half-lives of c-FLIP_S and c-FLIP_L were reduced at $t = 100$ min from $t_{1/2}^{S/R} = 40$ min to 15 min and from $t_{1/2}^L = 2$ h to 30 min, respectively. This corresponds to an increase in the degradation rate constants (Equation 1) from $k_{deg}^{S/R} = 0.0173 \text{ min}^{-1}$ to 0.0462 min^{-1} and from $k_{deg}^L = 0.0058 \text{ min}^{-1}$ to 0.0231 min^{-1} , resulting in down-regulation of c-FLIP levels and consequent apoptosis 268 min after exposure to ligands.

The results described above refer to a situation in which both DISC-bound c-FLIP and free c-FLIP are degraded. In addition, we studied a scenario in which the degradation of the DISC-bound portion of c-FLIP could be separated from degradation of free c-FLIP, a situation freeing the c-FLIP-bound binding sites of the DISC, leaving them open for further reactions. The lower panels of Fig. 1B show the response obtained when DISC-bound c-FLIP was degraded 100 times faster than free c-FLIP. As expected, an increase in apoptotic sensitivity was obtained, and apoptosis now occurred 228 min after exposure to ligand. This and other simulation studies (data not shown) indicate that the fraction (DISC-bound versus free) of c-FLIP that is degraded does not have a major effect on apoptotic sensitivity but that the overall degradation is the determinant.

Thus, Fig. 1 indicates that the apoptotic behavior is determined primarily by the c-FLIP concentration at ligand exposure and that this concentration can be regulated by the degradation rate of c-FLIP. This is supported by a study (10) that concluded that the apoptotic sensitivity of the cells depends mainly on the c-FLIP levels at the time of receptor ligation. This conclusion was based on experiments in which no sensitization to apoptosis was observed when cells were exposed to CD95 ligand prior to heat shock. In this situation, the c-FLIP concentrations are at normal levels at the time of ligand stimulation, in contrast to the situation in which cells have been subjected to heat shock prior to ligand stimulation, when c-FLIP levels have been down-regulated due to proteasomal degradation (10). These findings (10) could be confirmed by simulations indicating that, when the heat shock is initiated after ligand stimulation, its effect on the apoptotic process is negligible.

In the model used here, the long and short c-FLIP forms affect the apoptotic process in similar ways, as the nominal

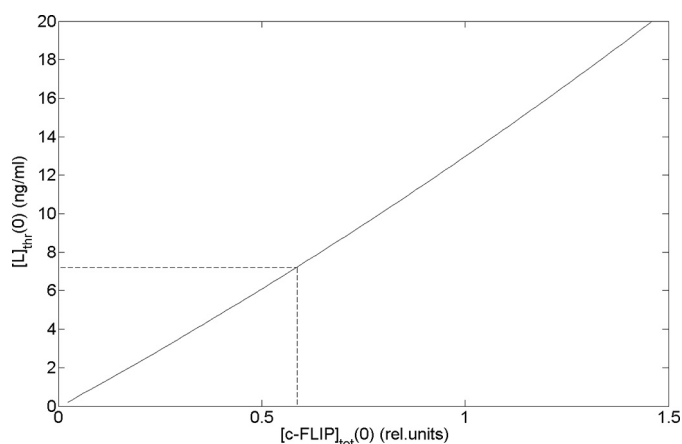


FIGURE 2. Effect of total initial c-FLIP concentration on the threshold ligand concentration required for signaling. Shown is $[L]_{thr}(0)$ for apoptosis as a function of total initial c-FLIP concentration ($[c-FLIP]_{tot}(0) = [c-FLIP_{S/R}](0) + [c-FLIP_L](0)$). Below the solid curve no apoptosis occurs, whereas concentrations above the curve result in apoptosis. For the initial c-FLIP concentrations used by Bentele *et al.* (18), the threshold ligand concentration is 7.2 ng/ml (dashed line). *rel. units*, relative units.

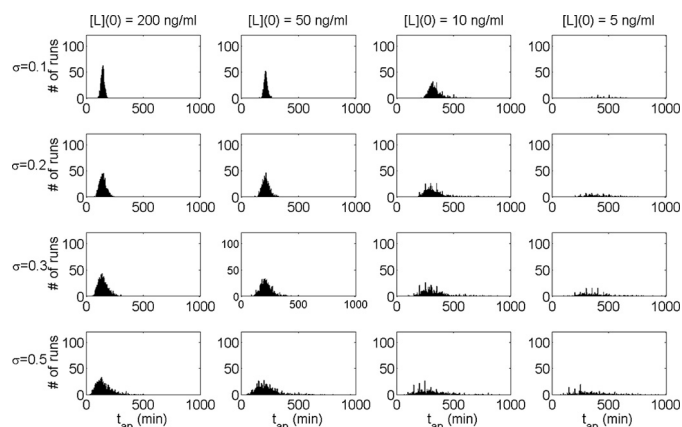


FIGURE 3. Simulated distributions of time to apoptosis. Results based on Monte-Carlo simulations with 1000 cells are shown as histograms with a bar width of 4 min. The same standard deviation (σ) was used for all initial concentrations and kinetic parameters. The panels from top to bottom illustrate results for $\sigma = 0.1, 0.2, 0.3$, and 0.5 , with $[L](0)$ (left to right) = 200, 50, 10, and 5 ng/ml. The corresponding proportions of non-apoptotic runs are as follows: for $\sigma = 0.1$: 0.0, 0.0, 10.4, and 93.1%; for $\sigma = 0.2$: 0.0, 0.1, 26.6, and 77.9%; for $\sigma = 0.3$: 0.0, 0.6, 4.2, and 68.7%; and for $\sigma = 0.5$: 0.6, 6.4, 40.7, and 61.5%.

reaction rates in the DISC are the same for both isoforms. Fig. 2 summarizes how the apoptotic sensitivity in the nominal model depends on the total initial concentration of c-FLIP, showing that the threshold initial ligand concentration, $[L]_{thr}(0)$, for triggering apoptosis increases almost linearly with the total initial c-FLIP concentration.

Apoptotic Response of Cell Populations—Monte Carlo simulations were performed to study the apoptotic response of a cell population with intercellular variations. The number of cells used in the simulations was 1000, and all initial concentrations and kinetic parameters in the simulation model were assumed to have the same standard deviation (σ) in the log-normal distribution (Equations 3 and 4).

Fig. 3 shows the simulated distributions of the time to apoptosis after exposure to ligands obtained for different initial ligand concentrations and various σ values. The simulation time was limited to 1000 min, and simulations in which apopto-

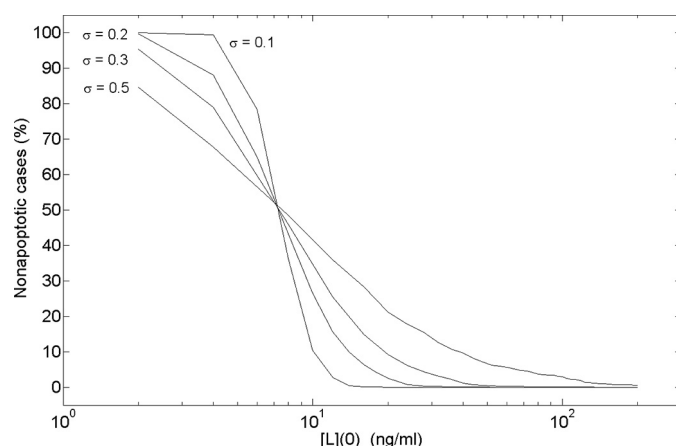


FIGURE 4. Proportion of non-apoptotic runs. Shown is the proportion of non-apoptotic runs as a function of initial ligand concentration for different σ values in the log-normal distribution of the model parameters. A threshold occurred at $[L](0) \approx 7.2$ ng/ml.

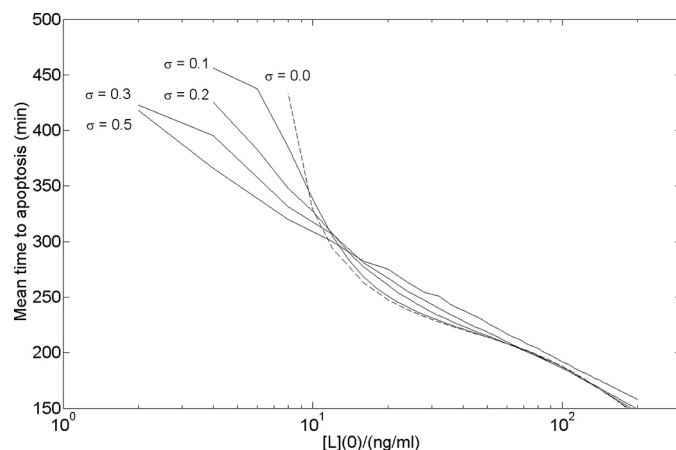


FIGURE 5. Mean time to apoptosis. Shown is the mean time to apoptosis for cell populations at different initial ligand concentrations for different σ values in the log-normal distribution of the model parameters. Populations with few ($<1\%$) apoptotic cases have been excluded.

sis had not occurred within this time frame were considered non-apoptotic. This approximation was motivated by the fact that, in extensive tests with a longer time frame, there were extremely few simulated apoptotic events with $t_{ap} > 1000$ min. Fig. 3 shows that, for moderate intercellular parameter variations (*upper panels*), the time to apoptosis is quite well defined but becomes smeared out when the parameter variations are large. It is also interesting to note that, despite the statistical parameter distributions, a quite clear threshold behavior for apoptosis is observed with respect to initial ligand concentrations. Fig. 4 illustrates the proportion of simulated non-apoptotic events as a function of the initial ligand concentration in the simulation runs. An apoptotic threshold at $[L](0) \approx 7.2$ ng/ml is observed for small σ values. This is consistent with the nominal model (Fig. 2). Note that, at the nominal threshold ligand concentration for apoptosis, only the direction of parameter variation, and not its magnitude, determines whether the cell will undergo apoptosis or not, and therefore, the proportion of simulated non-apoptotic events is independent of the standard deviation of the parameter distributions at this point (where the curves intersect in Fig. 4).

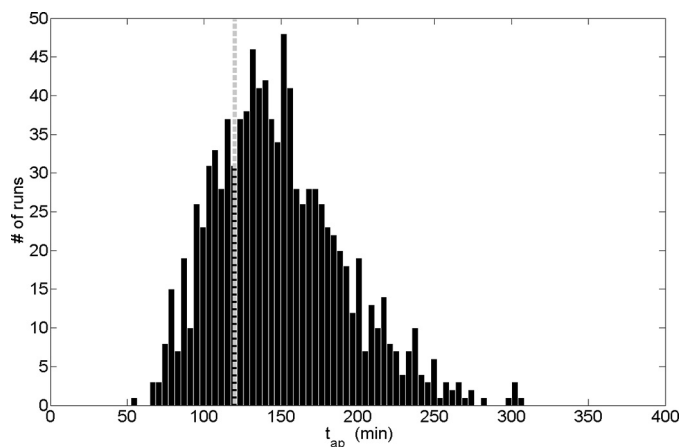


FIGURE 6. **Proportion of apoptotic runs after a given time period.** Shown is the distribution of time to apoptosis for $\sigma = 0.3$ and $[L](0) = 200$ ng/ml. Apoptosis occurred within 2 h (i.e. left of vertical dashed line) in 26.6% of the simulated cases.

Fig. 5 shows the mean time to apoptosis among the simulated apoptotic events, disregarding cases with very few ($<1\%$) apoptotic cells. For sufficiently large standard deviations ($\sigma > 0.3$), the mean time to apoptosis is a nearly linear function of the logarithm of the ligand concentration, roughly given by $t_{ap} \approx 440 - 130 \log([L](0))$ when using the previous definition of t_{ap} (in minutes) and expressing ligand concentration in nanograms/ml. Observe that this relationship could not be predicted using a single-cell model with nominal parameters ($\sigma = 0$), which results in a completely different relationship between ligand concentration and mean time to apoptosis (Fig. 5).

In experimental studies of apoptosis, the results of standard measurement methods such as different types of flow cytometry analysis can be summarized in the number of apoptotic cells in a cell population, and the apoptotic process is then often illustrated using bar graphs showing the proportion of apoptotic cells obtained at given times after exposure to ligands, for example (6, 10). The results of the stochastic simulations can be represented in a form compliant with those of the experimental studies by depicting the proportion of apoptotic events as a function of time. Fig. 6 shows a magnification of the panel of Fig. 3 corresponding to $\sigma = 0.3$ and $[L](0) = 200$ ng/ml, where the simulated apoptotic events within 2 h appear to the left of the vertical dashed line. The proportion of apoptotic runs within 2 h (using the present definition of the time to apoptosis) obtained for $\sigma = 0.1, 0.2, 0.3$, and 0.5 and $[L](0) = 200$ ng/ml (cf. Fig. 3) are 4.2, 19.2, 26.6, and 34.7%, respectively. Thus, the degree of stochasticity in the initial protein concentrations and reaction rate parameters has a substantial effect on the proportion of cells that are predicted to undergo apoptosis within a given time interval.

Fig. 7 depicts the proportions of apoptotic cells within 2, 3, 4, 5, and 16 h after exposure to ligands of different concentrations for simulations with $\sigma = 0.3$. These simulation results are in qualitative agreement with experimental findings presented in the literature (6, 18). For example, when comparing simulations with the experimental data presented in Fig. 1A in the study by Lavrik *et al.* (6), the simulations and experimental data both indicate that the proportion of cell deaths occurring during

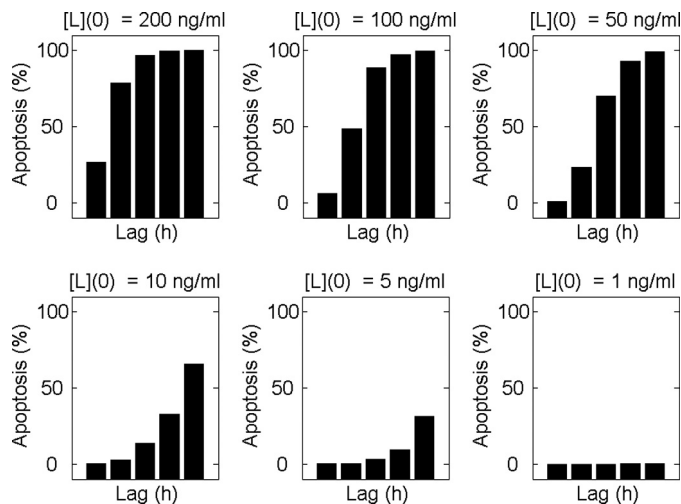


FIGURE 7. **Proportion of cell deaths as a function of time.** Shown is the occurrence of apoptosis within 2, 3, 4, 5, h and 16 h at different initial ligand concentrations for $\sigma = 0.3$ in the log-normal distribution of the parameters.

various time periods after ligand exposure depends on ligand concentration. At higher concentrations, most of the cell deaths occur within 5 h. At lower ligand concentrations, the total number of deaths is smaller, and a larger fraction of the cell deaths occur at later time points.

Example of Modeling Recovery after Heat Shock—To more specifically compare modeling results with experimental data, we investigated a specific case of sensitization to apoptosis caused by c-FLIP destabilization. Meinander *et al.* (10) studied the effect of heat shock treatment on human T lymphocytes and found that the degradation rate of c-FLIP increased considerably, whereas synthesis rates were unaffected. The half-lives of the two isoforms decreased from 40 and 120 min to 15 and 30 min, respectively. The investigators also studied how the down-regulation of c-FLIP during a 30-min heat shock treatment affected the sensitivity to apoptosis when the cells were exposed to ligands after various periods of recovery from the heat shock.

The single-cell simulation in Fig. 1B indicated that down-regulation of c-FLIP levels made the cell more sensitive to apoptosis when subjected to ligands. To mimic the experimental conditions reported by Meinander *et al.* (10), the following program was simulated. A 30-min period of steady-state conditions, in which the synthesis of both c-FLIP_{S/R} and c-FLIP_L counterbalances their net degradations, was followed by a 30-min period in which the degradation rates were kept high (using the same parameters as in the simulation shown in the middle panels of Fig. 1B). Next, the original degradation rates were restored, resulting in a recovery of the c-FLIP concentrations to their original levels. This dynamic behavior is illustrated in the upper panel of Fig. 8, which depicts the responses of the c-FLIP concentrations of the original single-cell model (18) complemented by synthesis and degradation of c-FLIP. The curves clearly show the more rapid response of c-FLIP_{S/R} and also illustrate that the c-FLIP levels did not reach their new steady-state values during the 30-min heat shock period. The implications of these changes on sensitivity to apoptosis for the nominal single-cell model are illustrated in the middle and lower panels of Fig. 8, which show how the threshold ligand

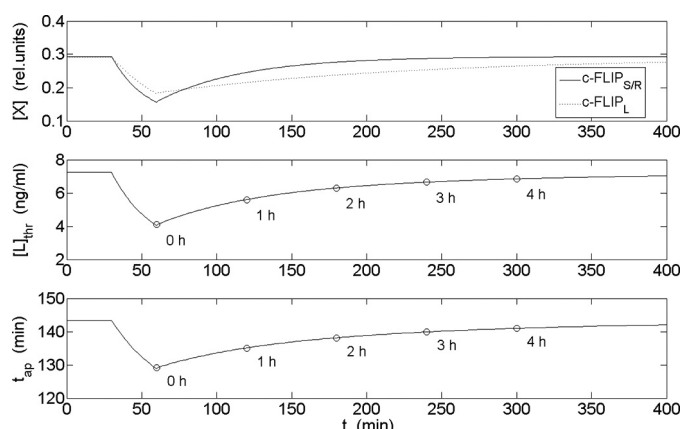


FIGURE 8. Dynamic c-FLIP regulation in heat shock experiments. Shown is a single-cell simulation of down- and up-regulation of c-FLIP in heat shock experiments reported by Meinander *et al.* (10). Degradation rates were increased at $t = 30$ min and were restored to their initial values at $t = 60$ min. *Upper panel*, concentrations of c-FLIP isoforms during heat shock and subsequent recovery period when there was no ligand exposure. *Middle panel*, threshold ligand concentration for apoptosis when the cells were exposed to ligands at various times during heat shock experiment. For 0, 1, 2, 3, and 4 h and infinite recovery periods after heat shock, $[L]_{thr}$ (indicated by circles) = 4.0, 5.6, 6.3, 6.7, 6.9, and 7.2 ng/ml, respectively. *Lower panel*, t_{ap} after ligand stimulation when cells were exposed to ligands (200 ng/ml) at various times during heat shock experiment. For 0, 1, 2, 3, and 4 h and infinite recovery periods after heat shock, t_{ap} = 129, 135, 138, 140, 141, and 143 min, respectively.

concentration for triggering apoptosis and the time to apoptosis varied with time during the heat shock and the subsequent recovery period.

It is now interesting to use the cell population model to mimic the experiments in Ref. 10, studying the effect of introducing death ligands after different periods of recovery following a heat shock treatment. Fig. 9A illustrates the simulated proportions of cells that underwent apoptosis within 2 h after exposure to ligands ($[L] = 200$ ng/ml) for populations with different σ values. Ligands were introduced at 0, 1, 2, 3, and 4 h into the recovery period. The simulation results for $\sigma = 0.2$ and 0.3 qualitatively agree with the experimental findings reported in Fig. 3b in Ref. 10, which are reanalyzed here in Fig. 9B. The simulations give a somewhat faster recovery after heat shock and a slightly higher proportion of apoptotic cells than experimentally observed. These discrepancies are actually what might be expected, considering the fact that the dynamics of the stabilization and destabilization of c-FLIP have been ignored, and the real down- and up-regulation rates of c-FLIP would therefore be somewhat slower than those shown in Fig. 8. It is also worthwhile to observe that, although the previously compiled model (18), which was further developed in this study, was developed for a human B lymphoblastoid cell line (SKW 6.4), its simulations fitted well with data obtained from other cells, indicating that the model has universal applicability.

Note that the results shown in Fig. 9 cannot be directly deduced from the single-cell simulations in Fig. 8 alone. However, the time to apoptosis in the deterministic model (Fig. 8, lower panel) is slightly above 2 h throughout the heat shock experiment. Thus, the single-cell model predicts that, for a cell population, the proportion of apoptotic cells obtained 2 h after

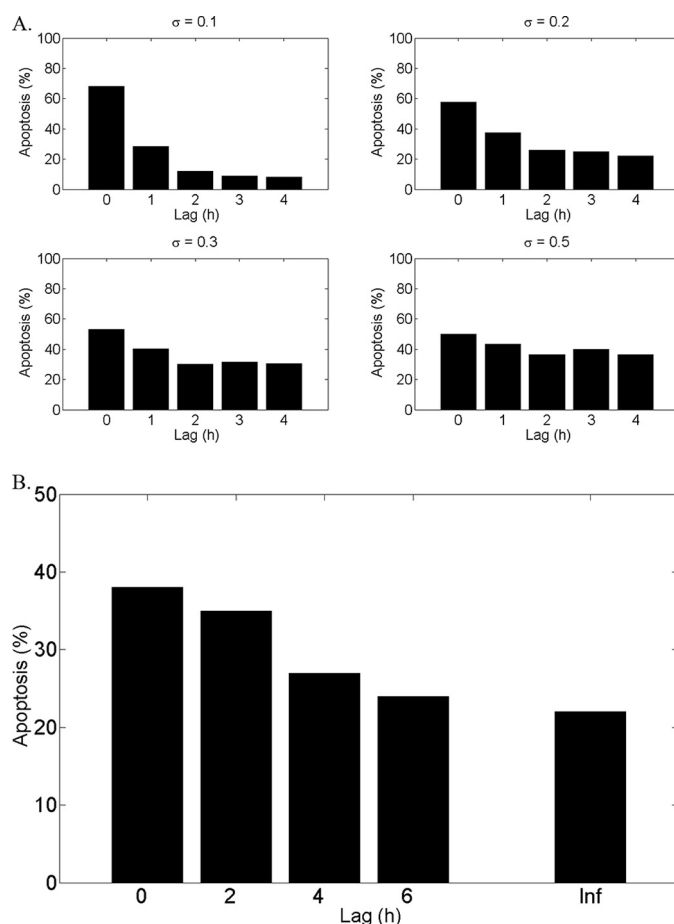


FIGURE 9. Proportions of apoptotic cells in heat shock experiments. A, simulated results of a 30-min down-regulation of c-FLIP_{S/R} and c-FLIP_L followed by a recovery period of 0, 1, 2, 3, or 4 h before exposing the receptors to ligand. The bars show the occurrence of apoptosis within 2 h after ligand stimulation (200 ng/ml) in populations with $\sigma = 0.1, 0.2, 0.3$, and 0.5 using log-normal distribution of the parameters. B, experimental results from Ref. 10 showing proportions of apoptotic cells obtained within 2 h after exposure to ligands (200 ng/ml) 0, 2, 4, and 6 h into the recovery period following a 30-min heat shock treatment.

ligand stimulation is strongly correlated with the apoptotic sensitivity during the experiment.

DISCUSSION

We used a novel approach based on mathematical modeling and simulation of cell populations to study the role of dynamic regulation of the expression levels of c-FLIP proteins in controlling death receptor-mediated apoptosis. A completely novel feature of our model, compared with previous approaches (7, 18), is inclusions of equations to account for synthesis and degradation of c-FLIP, with rate parameters derived from experimental observations. Simulations with the model confirmed that the concentrations of c-FLIP at the moment when the cell receptors are exposed to ligands largely determine whether the apoptotic machinery is triggered, whereas the high turnover rates of c-FLIP allow for efficient regulation of c-FLIP concentrations by affecting its stability.

Experimental studies of apoptosis often determine the proportions of cells in a cell population that undergo apoptosis after a given time, obtained as an integral measure of results determined for a large bulk of cells (e.g. in flow cytometry). To

be able to use such averaged quantities for model validation and comparison with simulation results, we modeled cell populations with intercellular variations by introducing log-normally distributed protein concentrations and reaction parameters. In this way, the proportion of apoptotic cells within a given time frame could be estimated using Monte Carlo simulations. As an example, the approach was applied to simulate heat shock experiments, in which down-regulation of c-FLIP levels was followed by recovery periods of various lengths before exposure to ligands. The proportions of apoptotic cells predicted by the simulation model were in qualitative agreement with experimental findings. These results demonstrate that a model developed for one cell line (B cell line SKW 6.4) could be applied to qualitatively predict experimental results obtained from other cells.

A recent model proposed significant non-genetic variability in cell populations subjected to TRAIL-induced apoptosis (26). Interestingly, our model predicts that variations of this kind occur among any given cell population, largely as a result of intercell variability in the fine-tuning of c-FLIP levels.

The results of this study and the developed model provide insight into how small system perturbations may have pronounced effects on cell sensitivity. In this way, the system characteristics and especially limiting factors among the key molecules governing death receptor responses may be determined under different conditions. For example, the different roles of c-FLIP proteins (7, 8, 34), reflected in their different chemical reactions and rate parameters as well as concentration levels in individual cell lines, may now be considered as a function of the factors affecting the stability of c-FLIP proteins. This could even be an outset for modeling different therapies targeted at c-FLIP levels and/or DISC signaling. Parameter estimation in population models based on experimental findings will also yield possibilities to estimate rate-limiting steps among the processes that determine the regulation of c-FLIP isoforms and thereby the processing of DISC signals. For example, as mentioned in the Introduction, c-FLIP isoforms are subjected to active isoform-specific phosphorylation (27, 28) and ubiquitylation and degradation (9, 10). All the post-translational modifications seem to have distinct roles and features not only in regulating the stability of the proteins but also in determining their functions. Further applications of the developed model will enable prediction of both the molecular functions and the relative importance of these individual processes as determinants of DISC signaling.

REFERENCES

1. Ueffing, N., Keil, E., Freund, C., Kühne, R., Schulze-Osthoff, K., and Schmitz, I. (2008) *Cell Death Differ.* **15**, 773–782
2. Krueger, A., Baumann, S., Krammer, P. H., and Kirchhoff, S. (2001) *Mol. Cell. Biol.* **21**, 8247–8254
3. Boatright, K. M., Deis, C., Denault, J. B., Sutherlin, D. P., and Salvesen, G. S.

- (2004) *Biochem. J.* **382**, 651–657
4. Chang, D. W., Xing, Z., Pan, Y., Algeciras-Schimnich, A., Barnhart, B. C., Yaish-Ohad, S., Peter, M. E., and Yang, X. (2002) *EMBO J.* **21**, 3704–3714
5. Micheau, O., Thome, M., Schneider, P., Holler, N., Tschopp, J., Nicholson, D. W., Briand, C., and Grütter, M. G. (2002) *J. Biol. Chem.* **277**, 45162–45171
6. Lavrik, I. N., Golks, A., Riess, D., Bentele, M., Eils, R., and Krammer, P. H. (2007) *J. Biol. Chem.* **282**, 13664–13671
7. Fricker, N., Beaudouin, J., Richter, P., Eils, R., Krammer, P. H., and Lavrik, I. N. (2010) *J. Cell Biol.* **190**, 377–389
8. Poukkula, M., Kaunisto, A., Hietakangas, V., Denessiouk, K., Katajamäki, T., Johnson, M. S., Sistonen, L., and Eriksson, J. E. (2005) *J. Biol. Chem.* **280**, 27345–27355
9. Tran, S. E., Meinander, A., Holmström, T. H., Rivero-Müller, A., Heiskanen, K. M., Linnau, E. K., Courtney, M. J., Mosser, D. D., Sistonen, L., and Eriksson, J. E. (2003) *Cell Death Differ.* **10**, 1137–1147
10. Meinander, A., Söderström, T. S., Kaunisto, A., Poukkula, M., Sistonen, L., and Eriksson, J. E. (2007) *J. Immunol.* **178**, 3944–3953
11. Bhalla, U. S., and Iyengar, R. (1999) *Science* **283**, 381–387
12. Endy, D., and Brent, R. (2001) *Nature* **409**, 391–395
13. Kitano, H. (2002) *Science* **295**, 1662–1664
14. Alberghina, L., and Westerhoff, H. V. (eds) (2005) *Systems Biology: Definitions and Perspectives*, Springer-Verlag, Berlin
15. Bagci, E. Z., Vodovotz, Y., Billiar, T. R., Ermentrout, G. B., and Bahar, I. (2006) *Biophys. J.* **90**, 1546–1559
16. Legewie, S., Blüthgen, N., and Herzog, H. (2006) *PLoS Comput. Biol.* **2**, 1061–1072
17. Stucki, J. W., and Simon, H. U. (2005) *J. Theor. Biol.* **234**, 123–131
18. Bentele, M., Lavrik, I., Ulrich, M., Stösser, S., Heermann, D. W., Kalthoff, H., Krammer, P. H., and Eils, R. (2004) *J. Cell Biol.* **166**, 839–851
19. Eissing, T., Conzelmann, H., Gilles, E. D., Allgöwer, F., Bullinger, E., and Scheurich, P. (2004) *J. Biol. Chem.* **279**, 36892–36897
20. Fussenegger, M., Bailey, J. E., and Varner, J. (2000) *Nat. Biotechnol.* **18**, 768–774
21. Albeck, J. G., Burke, J. M., Spencer, S. L., Lauffenburger, D. A., and Sorger, P. K. (2008) *PLoS Biol.* **6**, 2831–2852
22. Harrington, H. A., Ho, K. L., Ghosh, S., and Tung, K. C. (2008) *Theor. Biol. Med. Model.* **5**, 26
23. Eissing, T., Allgöwer, F., and Bullinger, E. (2005) *IEEE Proc. Syst. Biol.* **152**, 221–228
24. Li, C., Ge, Q. W., Nakata, M., Matsuno, H., and Miyano, S. (2007) *J. Biosci.* **32**, 113–127
25. Heiner, M., Koch, I., and Will, J. (2004) *BioSystems* **75**, 15–28
26. Spencer, S. L., Gaudet, S., Albeck, J. G., Burke, J. M., and Sorger, P. K. (2009) *Nature* **459**, 428–432
27. Kaunisto, A., Kochin, V., Asaoka, T., Mikhailov, A., Poukkula, M., Meinander, A., and Eriksson, J. E. (2009) *Cell Death Differ.* **16**, 1215–1226
28. Shi, B., Tran, T., Sobkoviak, R., and Pope, R. M. (2009) *J. Biol. Chem.* **284**, 14513–14523
29. Elowitz, M. B., Levine, A. J., Siggia, E. D., and Swain, P. S. (2002) *Science* **297**, 1183–1186
30. Gillespie, D. T. (1976) *J. Comp. Phys.* **22**, 403–434
31. Barkai, N., and Leibler, S. (1997) *Nature* **387**, 913–917
32. Blüthgen, N., and Herzog, H. (2003) *J. Theor. Biol.* **225**, 293–300
33. Golks, A., Brenner, D., Fritsch, C., Krammer, P. H., and Lavrik, I. N. (2005) *J. Biol. Chem.* **280**, 14507–14513
34. Neumann, L., Pforr, C., Beaudouin, J., Pappa, A., Fricker, N., Krammer, P. H., Lavrik, I. N., and Eils, R. (2010) *Mol. Syst. Biol.* **6**, 352

Probability and Its Applications

Göran Högnäs
Arunava Mukherjea

**Probability
Measures
on Semigroups**

Convolution Products,
Random Walks
and Random Matrices

Second Edition

 Springer

Probability and Its Applications

Published in association with the Applied Probability Trust

Editors: J. Gani, C.C. Heyde, P. Jagers, T.G. Kurtz

Uncorrected Proof



For other titles published in this series, go to
www.springer.com/series/1560

About the Authors:

Göran Högnäs is Professor of Applied Mathematics at his alma mater Åbo Akademi University, from which he received his Ph.D. in Mathematics in 1974. Professor Högnäs is the co-author of a text on applied mathematics as well as co-editor of the Proceedings of the Third Finnish-Soviet Symposium on Probability Theory and Mathematical Statistics. He serves  the editorial board of the Journal of Theoretical Probability. Professor Hogna was the director of the Finnish Graduate School in Stochastics and Statistics (1998–2009).

Arunava Mukherjea is Professor of Mathematics at the University of South Florida in Tampa. He received his Ph.D. in Mathematics from Wayne State University and is co-author of Real and Functional Analysis (second edition) and Measures on Topological Semigroups. As visiting professor he spent sojourns at various academic institutions including Tata Institute of Fundamental Research in Mumbai, India and Université Paul Sabatier in Toulouse, France. Professor Mukherjea was a Fulbright Scholar and is the editor-in-chief of the Journal of Theoretical Probability. Arunava Mukherjea is Professor of Mathematics at the University of Texas – Pan American since 2007. He is an associate editor of the Journal of Theoretical Probability (2005–present) and was its editor-in-chief before 2005.

Göran Högnäs · Arunava Mukherjea

Probability Measures on Semigroups

Convolution Products, Random Walks,
and Random Matrices



Springer

Göran Högnäs
Åbo Akademi University
Fänriksgatan 3B
20500 Åbo
Finland
ghognas@abo.fi

Arunava Mukherjea
University of Texas Pan America
Department of Mathematics
West University Drive 12010
Edinburg Texas, 78539
Arunava.mukherjea@gmail.com

Series Editors:

Søren Asmussen
Department of Mathematical Sciences
Aarhus University
Ny Munkegade
8000 Aarhus C
Denmark
asmus@imf.au.dk

Peter Jagers
Mathematical Statistics
Chalmers University of Technology
and Göteborg (Gothenburg) University
412 96 Göteborg
Sweden
jagers@chalmers.se

Joe Gani
Centre for Mathematics
and its Applications
Mathematical Sciences Institute
Australian National University
Canberra, ACT 0200
Australia
gani@maths.anu.edu.au

Thomas G. Kurtz
Department of Mathematics
University of Wisconsin - Madison
480 Lincoln Drive
Madison, WI 53706-1388
USA
kurtz@math.wisc.edu

ISSN 1431-7028

ISBN 978-0-387-77547-0

e-ISBN 978-0-387-77548-7

DOI 10.1007/978-0-387-77548-7

Library of Congress Control Number: xxxxxxxxx

© Springer Science+Business Media, LLC 2011

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Cover design: WMXDesign

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface to the Second Edition

1

In this new edition, we have corrected all errors (mathematical and otherwise) that we could find or were brought to our attention from the book's first edition by our colleagues and students. We have also updated the references, all notes and comments at the end of each chapter, and more importantly, added exercises at the end of each section in each chapter. We have also added new results often in the main text and in the appendices.

The book, we feel, can be a useful reference for courses in the area of probability on algebraic structures. It can also serve as a text for graduate or senior undergraduate students for an one semester course on "Probability measures on semigroups" designed around the following four core topics:

- (i) Completely simple semigroups: their structure and the Rees-Suschkewitsch theorem (see Chapter 1)
- (ii) Convolution products of probability measures on locally compact semigroups, their weak convergence, their weak limits, while idempotent probability measures, with supports always closed completely simple subsemigroups, and structure theorem for idempotent probability measures (see Chapter 2)
- (iii) Recurrent random walks on semigroups: characterization of the set of recurrent states as closed and completely simple subsemigroups (see Chapter 3)
- (iv) Tightness and convergence in distribution of products of d by d i.i.d. random nonnegative matrices: The Kesten-Spitzer theory and various generalizations to random real matrices (see Chapter 4)

The first chapter of the book can also serve as a text for an one semester course on Semigroups for senior undergraduate or graduate students. Sufficiently many exercises of varying levels of difficulty have been included to help the instructor of such a course.

Semigroups are very general structures and we often come across them in various contexts in science and engineering. (See the preface to the first edition.) The results that we have presented here on weak convergence or random walks or random matrices using semigroup ideas are for the most part complete and best possible. Still

there are other results that remain to be completed. These are all mentioned in the 31
notes and comments at the end of each chapter, and this, we hope, will keep the 32
readership of this book enthusiastic and interested for some time to come. 33

Abo Akademi University, Abo
AQ1 Finland and University of Texas, Edinburg, Texas

Goran Hognas 34
Arunava Mukherjea 35

Uncorrected Proof

AUTHOR QUERY

AQ1. Please provide date and year

Uncorrected Proof

Contents

1

1 Semigroups 1 2

1.1 Introduction 1 3

1.2 Homomorphisms, Quotients, and Products 6 4

1.3 Semigroups with Zero 10 5

1.4 The Rees–Suschkewitsch Representation Theorem 12 6

1.5 Topological Semigroups 22 7

1.6 Semigroups of Matrices 33 8

1.7 Semigroups of Infinite Dimensional Matrices 52 9

1.8 Notes and Comments 60 10

2 Probability Measures on Topological Semigroups 63 11

2.1 Introduction 63 12

2.2 Invariant and Idempotent Probability Measures 64 13

2.3 Weak Convergence of Convolution Products of Probability Measures 83 14

2.4 Weak Convergence of Convolution Products
of Nonidentical Probability Measures 137 16

2.5 Notes and Comments 168 17

3 Random Walks on Semigroups 171 18

3.1 Introduction 171 19

3.2 Discrete Semigroups 180 20

3.3 Locally Compact Groups 198 21

3.4 Compact Semigroups 219 22

3.5 Completely Simple Semigroups 241 23

3.6 Notes and Comments 247 24

4 Random Matrices 253 25

4.1 Introduction 253 26

4.2 Recurrent Random Walks in Nonnegative Matrices 254 27

4.3 Tightness of Products of I.I.D. Random Matrices: Weak Convergence 284 28

4.4 Invariant Measures for Random Walks in Nonnegative
Matrices: Laws of Large Numbers 333 30

4.5 Notes and Comments 355 31

**A Products of I.I.D. Random Stochastic Matrices: Their
Skeletons and Convergence in Distribution 359 32 33**

B An Example Due to Chamayou and Letac 365 34

**C Asymptotic Behavior of $\|X_n X_{n-1} \dots X_0 u\|$ for I.I.D.
Random Nonnegative Matrices 369 35 36**

**D Continuous Singularity of the Limit Distribution of
Products of I.I.D. Stochastic Matrices 383 37 38**

**E Rate of Decay of Concentration Functions on Discrete
Groups, by T. M. Retzlaff 397 39 40**

E.1 Preliminaries 397 41

E.2 Irreducible Measures 404 42

E.3 Adapted Measures 408 43

References 413 44

Index 423 45

Chapter 1

Semigroups

1

2

1.1 Introduction

3

Chapter 1 contains the basics of semigroups: definitions, elementary concepts, 4
and fundamental examples. We assume some familiarity with standard notions of 5
point-set topology (see [124, 167]); the algebraic portions of Chap. 1 are, however, 6
completely self-contained. Without going into any details whatsoever, it is perhaps 7
prudent to remark at this point that our main interest centers around asymptotics, 8
invariance questions, etc. Our treatment is a reflection of this. We concentrate on 9
algebraic concepts corresponding to such phenomena as absorption, stability, and 10
invariance: zeros, simple semigroups, minimal ideals, maximal subgroups, and so 11
on. We strive to keep the digressions at a minimum. Clifford and Preston [43] of- 12
fers a wealth of information on all aspects of algebraic semigroups, and this text is 13
recommended to any reader interested in a much more elaborate treatment of this 14
fascinating subject. 15

Sections 1.1–1.5 contain basic material necessary for the development of all 16
subsequent chapters, while Sects. 1.6 and 1.7, which deal with more specific ap- 17
plications, can be skipped at first reading. 18

Arguably, the most important notion in mathematics is that of a *mapping* (or 19
function or *transformation*). The ultimate goal of the research presented in this 20
book is to describe the long-term behavior of random transformations of some set. 21
Transformations of a set form a semigroup in a natural way. Indeed, we see that any 22
semigroup is (algebraically) a transformation semigroup in a canonical way. *Linear* 23
transformations of a vector space form another family of fundamental examples. 24
We devote considerable effort to those semigroups, which incidentally may just as 25
well be viewed as semigroups of *matrices*. 26

Let S be a set. If S is endowed with an associative binary operation [which 27
we call *multiplication* and denote by a dot, $(\cdot)$ or simply by juxtaposition], then the 28
 S is called a *semigroup*. Strictly speaking, the semigroup is the pair $(S, \cdot)$, but the 29
intended operation is usually quite clear from the context. When we are dealing with 30
a specific application, we of course use the established notation. 31

If s is an element of a semigroup S and A and B are subsets of S , then we de- 32
note by sA the set $\{sa \in S | a \in A\}$ and by AB the set $\{ab \in S | a \in A, b \in B\}$. 33

(As is, of course, defined analogously.) Note that products with more than two factors, such as abc and $aBcS$, are well defined due to the associativity of the multiplication. $aa, aaa, \dots$ are usually written $a^2, a^3, \dots$

A nonempty subset T of S is called a *subsemigroup* if it is stable under multiplication, i.e., if $TT \subset T$. If T is also a group, we call it a *subgroup* of S . As we see later, it is important in many applications to identify the subgroups of a given semigroup S .

A subsemigroup L of S is called a *left ideal*, if $SL \subset L$; *right ideals* are similarly defined. A nonempty subsemigroup I is a *bilateral* or *two-sided ideal* or just *ideal* of S if it is both a right and a left ideal: $SI \subset I, IS \subset I$. S is said to be *left (right) simple* if it contains no proper left (right) ideal. Similarly, S is *simple* if the only ideal of S is S itself. A left (right) ideal is a *principal* left (right) ideal if it is of the form $\{a\} \cup Sa$ ($\{a\} \cup aS$) for some $a \in S$.

Note that S is left simple if and only if for any given $a, b \in S$, the equation $xa = b$ is soluble. (Sa is a left ideal of S for all $a \in S$. On the other hand, any left ideal L contains a subset of the form Sa .)

An element $e \in S$ is called a *left (right) identity element* of S if $es = s$ ($se = s$) for every $s \in S$. e is a *two-sided identity element* of S , or simply *identity* of S , if it is both a right and a left identity of S . It is easy to see that the identity is unique if it exists.

An element z of S is called a *left (right) zero element* of S if $zs = z$ ($sz = z$) for all $s \in S$. If z is both a left zero and a right zero of S , we simply call it a *zero* of S . A semigroup has at most one zero.

The semigroup S is said to be *left (right) cancellative* if for any $a \in S$, the equation $ax = ay$ ($xa = ya$) in S implies $x = y$.

An element $a \in S$ is *idempotent* if $a^2 = a$. Zeros and identities are idempotent. An idempotent is in a trivial fashion the identity element of a subgroup of S .

Elements a and b of S are said to *commute* if $ab = ba$. If all elements of S commute with each other, S is called a *commutative* or *abelian* semigroup. In abelian semigroups, the operation is often called *addition* and denoted by $+$, the identity element by 0 and the inverse of a by $-a$.

To put the preceding concepts into perspective, we now investigate some of their relationships to a group structure. Most textbooks define a *group* as a nonempty set G with an associative binary operation with identity element e and inverses; i.e., for all $a \in G$, there is a $b \in G$ such that $ab = ba = e$, (see, for example, [62] or [222]). There is, however, a multitude of alternative, seemingly weaker, but in fact equivalent definitions, see [43], Chap. 1.

In our context, Proposition 1.1 is a convenient characterization of a group.

Proposition 1.1. *A semigroup is a group if and only if it is both left and right simple.*

Proof. A group is clearly both left and right simple. Conversely, let S be a semigroup which is both left and right simple. For any $a \in S$, the equation $ae = a$ has a solution. On the other hand, Sa is all of S ; hence, e is a right identity of S . In the same way, we can produce a left identity f which turns out to be equal to e . (We have $f = fe = e$.)

It is just as straightforward to obtain the two-sided inverse for an element $a \in S$.
Let b and c be the solutions of the equations $ab = e$ and $ca = e$, respectively. Then
 $b = eb = (ca)b = c(ab) = ce = c$. Thus, S is a group. $\square$

If A is a subsemigroup of S we denote by $\langle A \rangle$ all the elements in S that can
be expressed as finite products of the elements in A . In the case $\langle A \rangle = S$, we call
 A the *generator* of S . If a semigroup is generated by a single element, we call it
monothetic or *cyclic*.

Remark 1.1. In a finite semigroup S , $\langle x \rangle$ contains a subgroup for every $x \in S$, cf.
Exercise 1.6.

Example 1.1 (The semigroup of transformations of a set).

Let X be any finite or infinite nonempty set. If f and g are mappings from X
to itself, we define as usual the *composition* $f \circ g$ of f and g by $(f \circ g)(x) =$
 $f(g(x))$, $x \in X$. (The domain of the mappings is always understood to be all of X .)
The composition of mappings is an associative operation; hence, the set of all map-
pings from X into X forms a semigroup with composition as multiplication. This
semigroup is called the *full transformation semigroup on X* and denoted by $\mathcal{T}_X$.
A more complete treatment of $\mathcal{T}_X$ is given in [43], Chap. 2.2.

$\mathcal{T}_X$ has an identity element, the identity mapping ι : $\iota(x) = x$, $x \in X$. The
constant mappings are left zeros, since $c \circ f = c$ for all f if $c(x) = x_0$, $x \in X$
(where x_0 is a particular element of X). On the other hand, $\mathcal{T}_X$ does not admit any
right zeros unless of course X is a singleton.

Define the *range* $R(f)$ of an $f \in \mathcal{T}_X$ to be the set $f(X) \equiv \{f(x) \mid x \in X\}$
(where $\equiv$ means equal by definition). Clearly, $R(f \circ g) \subset R(f)$. If R_0 is a subset
of X , then mappings with range inside R_0 , $\{f \in \mathcal{T}_X \mid R(f) \subset R_0\}$, form a right
ideal of $\mathcal{T}_X$.

The *partition* $\pi(f)$ of X generated by an element $f \in \mathcal{T}_X$ is the equivalence
relation on X defined by $x\pi(f)y \iff f(x) = f(y)$, $x, y \in X$. In other words,
elements are $\pi(f)$ -equivalent if and only if they have the same image under f . The
 $\pi(g)$ -equivalence implies $\pi(f \circ g)$ -equivalence, $\pi(f \circ g) \supset \pi(g)$. Consequently,
if π_0 is a given equivalence relation on X , then those $f \in \mathcal{T}_X$ with $\pi(f) \supset \pi_0$
form a left ideal of $\mathcal{T}_X$.

Define the *rank* of f to be the cardinality of $R(f)$, which we denote by $|R(f)|$.
Note that the cardinality of $R(f)$ is the same as that of the quotient space $X/\pi(f)$
[the number of $\pi(f)$ -equivalence classes]. Mappings with rank no larger than a
given cardinality r , $\{f \in \mathcal{T}_X \mid |R(f)| \leq r\}$, form a two-sided ideal of the semi-
group $\mathcal{T}_X$.

The constant mappings, i.e., the mappings of rank 1, form the minimal two-sided
ideal of $\mathcal{T}_X$. This subsemigroup is a *left zero semigroup*, since all its elements are
left zeros.

Let us now identify the subgroups of $\mathcal{T}_X$. Since any subgroup has an idempotent
as its identity element, our first task is to determine idempotents in $\mathcal{T}_X$. Let e be
idempotent with range R and partition π , whence $e(x) = x$, $x \in R$. As before
 $|R| = |X/\pi|$. This is possible if and only if R is a *complete set of representatives* of

the π -equivalence classes, i.e., every equivalence class contains exactly one element of R . In the terminology of Clifford and Preston [43] R is a *cross section* of π .

Suppose f belongs to a subgroup G of $\mathcal{T}_X$ and the identity element of G is the idempotent e previously discussed. We can immediately conclude from the relation $e \circ f = f \circ e = f$ that the range of f is R and the partition corresponding to f is π . Furthermore, f is a one-to-one mapping from R to R precisely because R is a cross section of π . We can then construct a g belonging to $\mathcal{T}_X$ with the following properties: g has range R and partition π , and the restriction of g to R is the inverse of the restriction of f to R . It is then clear that $g \circ f = f \circ g = e$, in other words, the inverse of f in the subgroup G is g . Our result is thus the following, see [43], Theorem 2.10: *f belongs to a subgroup of $\mathcal{T}_X$ if and only if $R(f)$ is a complete set of representatives of $\pi(f)$. The sets $\{f \mid R(f) = R, \pi(f) = \pi\}$ where $|R| = |X/\pi|$ are groups if and only if R is a complete set of representatives of π .*

If there is an f whose range is *not* a cross section of $\pi(f)$, then $f \circ f$ has a smaller range than f : $R(f \circ f)$ is a proper subset of $R(f)$. Clearly, such an f cannot belong to a group.

The preceding discussion holds almost verbatim for any transformation semigroup on X ; i.e., any subsemigroup S of $\mathcal{T}_X$. The conditions are necessary only in the general case. For example, $f \in S$ can belong to a subgroup of S only if $R(f)$ is a cross section of $\pi(f)$. For a transformation semigroup S on a *finite* set X we have the converse: If elements of a subsemigroup of S have common range R and partition π , where R is a cross section of π , then it is a group. (Each element f is of finite order, i.e., some power of f equals the identity mapping e on R with the given partition π .)

Let X be countably infinite and G a subgroup of infinite rank of $\mathcal{T}_X$. If e is the identity element of G , then we can construct a mapping α with the properties that α restricted to the range of e is a bijection onto all of X (practically the definition of an infinite subset of X) and α has the same partition as e . If we denote the inverse of α restricted to the range of e by β , then β is injective, and has the same range as e . Hence, $\alpha \circ \beta = \iota$ (the identity mapping on X) and $\beta \circ \alpha = e$. For any $g \in G$, $\alpha \circ g \circ \beta$ is a bijection on X . Conversely, the elements of G can be written $\beta \circ h \circ \alpha$, where the h 's are bijections on X .

The semigroup $\mathcal{B}_X$ of relations on the set X consists of all subsets of the Cartesian product $X \times X$ equipped with the composition operation $\circ$: For any two σ and ν in $X \times X$ define $(x, y) \in \sigma \circ \nu$ if and only if there is a $z \in X$ such that $(x, z) \in \nu$ and $(z, y) \in \sigma$.

Exercise 1.1. Show that a semigroup with more than one element contains a subsemigroup $S' \neq S$.

Exercise 1.2. Prove that a finite semigroup that is cancellative is in fact a group.

Exercise 1.3. Show that if a semigroup S contains a left identity element e , such that there for any $x \in S$ exists a $y \in S$ such that $yx = e$, then S is a group.

Exercise 1.4. Prove that the following are groups: 163

- (i) A left simple semigroup that contains a left identity. 164
- (ii) A cancellative and simple semigroup containing an idempotent 165

Exercise 1.5. Let S be a semigroup so that for every $x \in S$ there exists an $a \in S$ such that $x = xax$. Show that the following are equivalent: 166

- (a) S has exactly one idempotent 168
- (b) S is cancellative 169
- (c) S is a group 170

Exercise 1.6. Let S be a finite semigroup generated by a single element x . Prove that there exists $m, r \in \mathbb{N}$, such that $x^{m+r} = x^r$ and 171

$$\langle x \rangle = \{x, x^2, \dots, x^{m+r-1}\}.$$

Also show that the set $\{x^r, x^{r+1}, \dots, x^{m+r-1}\}$ is a subgroup of S . 173

Exercise 1.7. Let S be a finite commutative semigroup. Show that S can be partitioned into maximal subsemigroups $\{P(e(i)) | i = 1, 2, \dots, r\}$, where $\{e(1), e(2), \dots, e(r)\}$ is the set of all idempotents of S , and x is in $P(e(i))$ iff $x^k = e(i)$ for some positive integer k . Further, show that there are smallest positive integers K and D such that for all x in S , $x^K = x^{K+D}$. 174

Exercise 1.8. Verify that the definition of the operation $\circ$ defined above on the semigroup $\mathcal{B}_X$ of relations on the set X coincides with the composition on $\mathcal{T}_X$ defined in Example 1.1 in case the relations are transformations of X . ($f \in \mathcal{T}_X$ can be viewed as a relation $f \in \mathcal{B}_X$ when we identify f with its graph, a subset of $X \times X$: $(x, y) \in f$ if and only if $y = f(x)$.) 175

Show that if a relation σ is reflexive and transitive, then it is idempotent, i.e., $\sigma \circ \sigma = \sigma$. 176

Exercise 1.9. A $d \times d$ real matrix is called *circulant* and denoted by $\text{circ}(x(0), x(1), \dots, x(d-1))$ if its first row is $x(0), x(1), \dots, x(d-1)$, its second row $x(d-1), x(0), x(1), \dots, x(d-2)$, the third row $x(d-2), x(d-1), x(0), \dots, x(d-3)$ and so on: 177

$$\begin{pmatrix} x(0) & x(1) & x(2) & \dots & x(d-1) \\ x(d-1) & x(0) & x(1) & \dots & x(d-2) \\ \dots & \dots & \dots & \dots & \dots \\ x(1) & x(2) & x(3) & \dots & x(0) \end{pmatrix}.$$

Prove that the circulant matrices form a commutative semigroup with respect to matrix multiplication. 178

Exercise 1.10. Let x be a $d \times d$ real matrix. Show that the following are equivalent: 192

- (i) x is circulant; 193
- (ii) x^T , the transpose of x , is circulant; 194
- (iii) $xP = Px$, where P is the permutation matrix $\text{circ}(0, 1, 0, \dots, 0)$; 195
- (iv) $x = f(P)$, where f is a polynomial of order less than d . 196

1.2 Homomorphisms, Quotients, and Products

197

A mapping ϕ between two semigroups $(S, \cdot)$ and $(T, *)$ is called a *semigroup* 198
homomorphism (antihomomorphism) if 199

$$\phi(a \cdot b) = \phi(a) * \phi(b) \quad (\phi(a \cdot b) = \phi(b) * \phi(a)), \quad a, b \in S.$$

The ϕ is said to be a *semigroup isomorphism (anti-isomorphism)* if it is bijective 200
(i.e., onto and one-to-one) as well. We however usually suppress the explicit ref- 201
erence to semigroups when it is clear from the context that we are dealing with a 202
semigroup structure. The S and T are *isomorphic (as semigroups)* if there exists a 203
semigroup isomorphism between them. Using the antiisomorphism $a \cdot b \mapsto b * a$ 204
we can always convert an antihomomorphism to a homomorphism, if need be. 205

As an example, consider the set of *left translations* on a given semigroup S : 206
 $\lambda_a, a \in S$, where λ_a is defined by $\lambda_a(x) = ax, x \in S$. Clearly, $\lambda_{ab} = \lambda_a \circ \lambda_b$. The 207
 λ is thus a homomorphism from S to the full transformation semigroup on S , $\mathcal{T}_S$. 208
Right translations define in a similar way a semigroup antihomomorphism from S 209
to $\mathcal{T}_S$. If in addition $ax = bx$ for all $x \in S$ implies $a = b$ (the left translations *act* 210
effectively on S), then λ is injective and S is *isomorphic to a subsemigroup of* $\mathcal{T}_S$. 211
In particular this is the case when S has a right identity element. 212

If we extend the idea of left translations slightly, we obtain the useful result that 213
any semigroup S is *isomorphic to some transformation semigroup*. We need only 214
take $X = S \cup 1$ and define $a1 = a, a \in S$. Left translations $\lambda_a, a \in S$ on X define 215
an injective homomorphism from S into $\mathcal{T}_X$. 216

An equivalence relation ρ on a semigroup S compatible with the multiplication 217
is called a *congruence* on S . More formally, ρ is a congruence if for all $a, b, s \in S$, 218
 apb implies $aspsb$ and $sapsb$. If ρ is a congruence on S , then the multiplication of 219
equivalence classes in the natural way will be a well-defined operation on S/ρ . (For 220
 $a, b \in S$, we take $[a][b] = [ab]$, where $[a]$ is the equivalence class containing a). 221
The semigroup thus obtained is called the *quotient or factor semigroup* of $S \bmod \rho$. 222

The discussion in the preceding paragraphs shows that if ρ is the congruence on 223
 S defined by 224

$$apb \iff ax = bx \text{ for all } x \in S,$$

then S/ρ is isomorphic to a subsemigroup of $\mathcal{T}_S$. 225

Definition 1.1 presents another useful congruence in semigroup theory. 226

Definition 1.1. Let I be an ideal of S . If we define a relation ρ by

227

$$a\rho b \iff a = b \text{ or else both } a \text{ and } b \in I,$$

then ρ is a congruence on S . The corresponding factor semigroup is usually written S/I , and it is called the *Rees factor* (or *quotient*) *semigroup* of $S \bmod I$. The intuitive idea behind the Rees factor semigroup is to lump all elements of I together into a single zero.

228
229
230
231

Take a semigroup S and view it as transformation semigroup on X . For $f \in S$, define a *matrix* $B_f(x, y)$ indexed by X according to the following prescription: $B_f(x, y) = \delta_{f(y)x}$, i.e., $B_f(x, y) = 1$ if $x = f(y)$ and 0 otherwise, $x, y \in X$. Multiplying the matrices according to the usual rules of matrix multiplication, we see that indeed $B_{f \circ g} = B_f B_g$, [see [53] where the matrices A_f have the anti-homomorphism property instead]. Hence, any semigroup can be described as a semigroup of 0–1 matrices, *transformation matrices*, with exactly one 1 in each column. If there is exactly one 1 in each row as well, we obtain the familiar *permutation* matrices corresponding to bijections on the set X .

232
233
234
235
236
237
238
239
240

For semigroups $(S, \cdot)$ and $(T, *)$, we obtain a new structure on their cartesian product by the rule

241
242

$$(s, t) \star (s', t') = (s \cdot s', t * t').$$

The resulting semigroup $(S \times T, \star)$ is called the *direct product* of $(S, \cdot)$ and $(T, *)$. Direct products with several factors are defined analogously.

243
244

Let G be a group and E be a right zero semigroup (where $ee' = e'$, $e, e' \in E$). Consider the direct product of G and E . Multiplication in $G \times E$ is given by $(g, e)(g', e') = (gg', e')$.

245
246
247

The usefulness of this structure is due to the fact that any right group has this representation.

248
249

Definition 1.2. A *right group* (*left group*) is a semigroup that is right simple (left simple) and left cancellative (right cancellative).

250
251

Alternative characterizations are given in Proposition 1.2

252

Proposition 1.2. For a semigroup S , the following statements are equivalent:

253

- (i) S is a right group;
- (ii) For any $a, b \in S$, the equation $ax = b$ has one and only one solution;
- (iii) S is right simple and contains an idempotent;
- (iv) S is isomorphic to the direct product of a group G and a right zero semigroup E .

254
255
256
257
258

Proof. The equivalence of (i) and (ii) follows immediately from the definitions and the characterization of right simplicity in Sect. 1.1.

259
260

Let S satisfy (ii). Then for any $a \in S$ the equation $ax = a$ has a solution e , say. We have $ae = ae$; by left cancellativity, $ee = e$. Thus (ii) implies (iii).

261
262

Assume (iii) and let $ax = ay$ for some $a, x, y \in S$. There is an idempotent e in S , e is a left identity of $S = eS$. There is a $b \in S$ such that $ab = e$. $(ba)(ba) = b(ab)a = bea = ba$, so ba is an idempotent too and consequently a left identity of S . We finally obtain $ax = ay \implies bax = bay \implies x = y$; i.e., (i) holds. The equation $(g, e)(x, y) = (g', e')$ in the direct product $G \times E$ has the unique solution $(x, y) = (g^{-1}g', e')$. This shows that (iv) implies (ii).

Suppose now that S satisfies (iii), and, equivalently, (i) and (ii). Let E be the set of idempotents in S . We saw above that any idempotent is a left identity, so $ee' = e', e, e' \in E$. In other words, E is a right zero (sub)semigroup (of S). Take an $e \in E$. We show next that Se is a subgroup of S . Se is right simple: $seSe = Se$. Se is also left simple. To see this, take an element $se \in Se$ and let $t \in S$ be such that $set = e$. Then $se(te) = (set)e = ee = e$ and $se(tese) = (sete)se = ese = see$ (since e is a left identity of S), so that $tese = e$ by left cancellability. For an arbitrary $ue \in Se$, the equation $(xe)(se) = ue$ can be solved, namely by $x = uet \in S$. By Proposition 1.1, Se is a group. Take a particular idempotent e_0 and let $Se_0 = G$.

Consider the map $(g, e) \mapsto ge, g \in G, e \in E$. Call it ϕ . We prove that ϕ is the desired isomorphism.

The map ϕ is a homomorphism: For $g, h \in G$ and $e, f \in E$, $geh f = ghf$ because each idempotent is a left identity.

ϕ is injective. To see this, let $ge = hf$. Then $g = ge_0 = g(ee_0) = (ge)e_0 = (hf)e_0 = h(fe_0) = he_0 = h$. Left cancellativity then yields $e = f$.

ϕ is surjective, since any $a \in S$ can be written in the form ae for some idempotent e [see the proof of (ii) $\implies$ (iii) above]. e_0 is a left identity of S , so $a = a(e_0e) = (ae_0)e$. $\square$

To understand right groups concretely, let us look at the full transformation semigroup $\mathcal{T}_X$ and its subsemigroups. The possible idempotents and subgroups were characterized in Sect. 1.1. Note that the subgroups consist of bijections (permutations) of the set R . It is not difficult to see that subgroups are *isomorphic to a permutation group on R* (i.e., a subgroup of the symmetric group $\mathcal{G}_R$ on R consisting of all the bijections from R to R).

It is evident that a right group $S \subset \mathcal{T}_X$ consists of mappings with a common range R_0 ; otherwise, it could not possibly be right simple. It follows from the assumption that S is closed under multiplication that the common range R_0 is a cross section of all partitions generated by elements of S .

Indeed, we obtain Proposition 1.3 in the case of a finite X :

Proposition 1.3. *A subsemigroup S of $\mathcal{T}_X$ is a right group if and only if the elements of S have common range.*

Proof. The only if statement is immediate, so let us concentrate on the sufficiency, the if statement. Any $f \in S$ is bijective on the common range, so left cancellativity follows. Since X is finite, the powers $f^n, n = 1, 2, 3, \dots$ form a group whose identity is a left identity of S . The equation $f \circ g = h$ thus has a solution $g = f^{r-1} \circ h$ (where r is the order of the group generated by f). In other words, S is right simple and hence a right group. $\square$

To see how the concepts work for countably infinite X , let us return to the last paragraph of Example 1.1. If H is a subgroup of the symmetric group $\mathcal{G}_X$, then $\beta \circ H \circ \alpha$ is an infinite-rank subgroup of $\mathcal{T}_X$. Conversely, any such subgroup is isomorphic to $\beta \circ H \circ \alpha$ for some H, α , and β . Right groups are then obtained by varying the partition in the construction of α : Any right group of infinite rank is isomorphic to $\beta \circ H \circ E$, where E is some set of mappings constructed exactly as α but with the partition varying; of course, the fundamental property of the range as a cross section of the partition must be maintained. In particular for an $\alpha' \in E$, $\alpha' \circ \beta = \iota$ and $\beta \circ \alpha'$ is the identity of some group of mappings with the same range as e .

Definition 1.3. We conclude this section by introducing another important product structure, the *Rees product*. Let E be a left zero semigroup, F a right zero semigroup, G a group, and ϕ a function from $F \times E$ to G . Define a multiplication on $E \times G \times F$ by

$$(e, g, f)(e', g', f') = (e, g\phi(f, e')g', f').$$

Note that this product is direct if the *sandwich function* ϕ maps everything onto the identity element of the group G . Such a ϕ is termed *trivial*.

We emphasize at this point that ϕ is a completely arbitrary function from $F \times E$ to G . Different choices of ϕ may produce isomorphic semigroups. If, for example, ϕ maps everything onto a constant $c \in G$, then the resulting Rees product is isomorphic to the direct product of E , G , and F . We will return briefly to this question in Sect. 1.4 (Proposition 1.10).

Remark 1.2. The cylinder subsets of the form $\{e\} \times G \times \{f\}$, called *cells*, are all groups isomorphic to G . The identity of such a group is the element $(e, (\phi(f, e))^{-1}, f)$. Any subsemigroup of the form $\{e\} \times G \times B$ (where $B \subset F$) of the Rees product is a right group. This fact is an immediate consequence of Definition 1.2.

Exercise 1.11. Let $\phi : S \rightarrow T$ be a homomorphism, where S and T are semigroups. Show that ϕ maps subsemigroups of S to subsemigroups of T , and also idempotents of S to idempotents of T . Show further that if ϕ is onto, then right ideals of S are mapped to right ideals of T , and the identity of S to the identity of T .

Exercise 1.12. Let S be a semigroup and T a subsemigroup of S , such that $xT = Tx$ for every $x \in S$. Prove that ρ , defined by $a\rho b \Leftrightarrow aT = bT$, is a congruence.

Exercise 1.13. Prove that every equivalence relation on a semigroup of left zeros is congruence.

Exercise 1.14. Let S be a semigroup and let ρ be a congruence on S . Prove that if e is an idempotent, $e \in S$, then its equivalence class e/ρ is a subsemigroup of S , and is an idempotent in the quotient S/ρ . Also, prove that if S is finite and x/ρ is an idempotent of S/ρ , then x/ρ contains an idempotent.

Exercise 1.15. Prove that the following are equivalent: 343

- (i) S is a right group 344
- (ii) There exists a right identity element $e \in S$ such that $e \in aS$ for every $a \in S$. 345
- (iii) For each element $a \in S$, Sa contains a left identity element of S . 346

Exercise 1.16. Show that if $X = \{1, 2, 3, 4\}$ and $R = \{1, 2\}$, then the semigroup S 347
of all maps from X into X with range R is a right group, but not a group. Find all 348
the eight elements of S and the right group decomposition as the direct product of a 349
group and a right zero semigroup. 350

1.3 Semigroups with Zero 351

Recall that an element z of a semigroup S is a *zero* if $sz = zs = z$ for all $s \in S$. 352
A zero is unique if it exists. We henceforth adhere to the common convention and 353
denote a zero by 0 . 354

The notions of (right, left) simplicity are trivial in the presence of a zero 0 . For 355
instance, S is simple if and only if $S = \{0\}$. For semigroups with zero, it is therefore 356
useful for many purposes to restrict some of the definitions to *nonzero* elements of S . 357
It is often practical to have a special notation for these elements; A^* is the generic 358
notation for the nonzero elements of the set A 359

$$A^* \equiv A \setminus \{0\} \equiv \{a \in A \mid a \neq 0\}.$$

An ideal $I \neq \{0\}$ of a semigroup S is said to be *0-minimal* if $\{0\}$ is the only ideal 360
of S properly contained in I . A 0-minimal left (right) ideal is defined analogously. 361

A semigroup S is called *right (left) 0-simple* if $S^2 \neq \{0\}$ and its only right (left) 362
ideals are 0 and S itself. S is called 0-simple if $S^2 \neq \{0\}$ and 0 is the only proper 363
two-sided ideal of S . 364

A semigroup S with the property that all products are 0 , $S^2 = \{0\}$, is called a 365
null semigroup. Such a semigroup obviously satisfies the second condition of the 366
preceding definitions. 367

Elements (nonzero elements) a and b of a semigroup with 0 are called *divisors* 368
of zero (*proper divisors of zero*) if $ab = 0$. 369

Propositions 1.1 and 1.2 have their counterparts for semigroups with 0 , which 370
follow immediately from the original Propositions and Lemma 1.1. 371

Proposition 1.4. A semigroup with 0 is a group with 0 if and only if it is both left 372
and right 0-simple. 373

Proposition 1.5. For a semigroup S with 0 , the following statements are equivalent: 374

- (i) S^* is a right group; 375
- (ii) For any $a \in S^*$, $b \in S$ the equation $ax = b$ has one and only one solution; 376
- (iii) S is right 0-simple and contains a nonzero idempotent; 377

- (iv) S is isomorphic to the Rees quotient of the direct product of a group with zero G^0 and a right zero semigroup E modulo $E \times \{0\}$.

Lemma 1.1. A right (left) 0-simple semigroup has no proper divisors of zero.

Proof. Let S be a right 0-simple semigroup and assume that a and b are proper divisors of zero. For a given a , solutions of the equation $ax = 0$ form a right ideal of S . Consequently by our assumptions on S , the existence of one nonzero solution b implies that all elements of S are solutions to $ax = 0$; i.e., $aS = \{0\}$. Those as with this property also form a right ideal of S . Again that right ideal has to be all of S , implying $SS = \{0\}$, which contradicts the assumption of right 0-simplicity. $\square$

Definition 1.4. A product structure analogous to the Rees product in Sect. 1.2 is the Rees product over a group with zero which is defined as follows: Let G^0 be a group with a zero 0 adjoined. Let E and F be as in the definition in Sect. 1.2. ϕ denotes a map from $F \times E$ to G^0 . Form the product $E \times G^0 \times F$ with the same multiplication rule as before

$$(e, g, f)(e', g', f') = (e, g\phi(f, e')g', f'), e, e' \in E, g, g' \in G^0, f, f' \in F.$$

This defines a semigroup. The set $I \equiv E \times \{0\} \times F$ consisting of triples with zero middle factor is a two-sided ideal of the semigroup. The Rees product over G^0 is then obtained by collapsing all of I into a zero. More precisely, the Rees product over the group with zero is the Rees quotient

$$E \times G^0 \times F \text{ mod } I.$$

The cells $\{e\} \times G \times \{f\}$ are again groups with identity $(e, (\phi(f, e))^{-1}, f)$ provided the middle term exists, i.e., $\phi(f, e) \neq 0$. These cells are called *group cells*. Null cells are characterized by $\phi(f, e) = 0$. The null cells (with 0 adjoined) are null subsemigroups of the Rees product over G^0 .

The sandwich function ϕ is said to be *regular* if the mappings $\phi(f, \cdot)$ and $\phi(\cdot, e)$ are not identically 0 for any $f \in F$ or $e \in E$. In other words, for each $f \in F$ there is an $e \in E$ (and for each $e \in E$ there is an $f \in F$) such that $\phi(f, e) \neq 0$. If ϕ is pictured as a matrix indexed by $F \times E$, then it is regular if and only if there are no zero rows or zero columns.

Remark 1.3. The condition that ϕ be regular is necessary and sufficient for the Rees product over G^0 to be 0-simple: Take elements $(e, g, f), (e', g', f') \in E \times G \times F$. The equation

$$(x, y, z)(e, g, f)(x', y', z') = (e', g', f')$$

can be solved if and only if $\phi(z, e)$ and $\phi(f, x')$ can be chosen to be nonzero. In that case, a solution is provided by $x = e', y = (\phi(z, e)g\phi(f, x'))^{-1}g, z' = f'$, and y' the identity of G .

In contrast to the case without 0, sets of the form $\{e\} \times G^0 \times B$ (where $B \subset F$) are not necessarily right 0-simple unless B is all of F .

In general semigroup theory, a semigroup S is *regular* if $a \in aSa$ for all $a \in S$. This notion and the regularity of the sandwich function just discussed are consistent in the following sense: The Rees product over a group with 0 is regular as a semigroup if and only if its sandwich function is regular. To see this, consider the equation $(e, g, f) = (e, g, f)(x, y, z)(e, g, f)$, which can be solved provided the functions $\phi(f, \cdot)$ and $\phi(\cdot, e)$ are not identically 0.

Exercise 1.17. Let S be a semigroup with zero 0, and let $S \neq 0$. Prove that S is 0-simple if and only if $SxS = S$ for every $x \neq 0$ of S .

Exercise 1.18. Let S be a 0-simple semigroup, and let for every $x \neq 0$ in S , $x^n \neq 0$ where $n = 1, 2, \dots$. Show that S contains no proper divisors of zero.

Exercise 1.19. Let M be a 0-minimal ideal of a semigroup S with zero. Prove that either $M = 0$ or M is a 0-simple subsemigroup of S .

Exercise 1.20. Define J_k as the set of elements of rank at most k for a transformation semigroup on a finite set X . Let S be a subsemigroup of $\mathcal{T}_n$. Identify the divisors of zero in S/J_k . Prove that J_n/J_{n-1} does not contain any divisors of zero. If $S = \mathcal{T}_n$, show that J_k/J_{k-1} has divisors of zero for $2 \leq k < n$.

1.4 The Rees–Suschkewitsch Representation Theorem

The Rees–Suschkewitsch representation theorem comes in two versions or even four if we take into account the topological considerations to be presented in Sect. 1.5. In Sect. 1.4, we discuss the algebraic case only for semigroups with and without 0. Needless to say, our strategy is dictated by applications where both versions appear naturally.

We will begin by studying the case without 0 in some detail. As seen in Chap. 4, this case is more prevalent in the applications. We only outline proofs of results in the latter part of Sect. 1.4, where the semigroups with 0 are treated. However, we point out the main pitfalls in going from one theory to the other.

Definition 1.5. An idempotent in S is said to be a *primitive idempotent* if it is minimal with respect to the partial order $\leq$ on the set $E(S)$ of idempotents of S defined by

$$e \leq f \iff ef = fe = e \quad (e, f \in E(S)).$$

In $\mathcal{T}_X$ constant mappings are primitive idempotents. If S is a subsemigroup of $\mathcal{T}_X$, X finite, then primitive idempotents must have minimal rank. All idempotents in a Rees product $E \times G \times F$ are primitive. (If e and f are idempotents such that $ef = fe = e$ then they belong to the same group, which of course has only one idempotent element, its identity.)

Definition 1.6. If S is a simple semigroup, i.e., a semigroup without proper two-sided ideals, then it is termed *completely simple* if it contains a primitive idempotent.

As the following example shows, this supplementary condition is by no means vacuous. Let S be the subsemigroup

$$\left\{ \begin{pmatrix} a & 0 \\ b & 1 \end{pmatrix} \mid a, b > 0 \right\}$$

of the semigroup of 2×2 real matrices (with ordinary matrix multiplication as operation). Clearly, S has no idempotent element, since the identity matrix does not belong to S . Consider the equation

$$\begin{pmatrix} x & 0 \\ y & 1 \end{pmatrix} \begin{pmatrix} a & 0 \\ b & 1 \end{pmatrix} \begin{pmatrix} u & 0 \\ v & 1 \end{pmatrix} = \begin{pmatrix} c & 0 \\ d & 1 \end{pmatrix},$$

where a, b, c, d are given positive real numbers. We can solve this equation in S . One solution is given by

$$\begin{pmatrix} x & 0 \\ y & 1 \end{pmatrix} = \begin{pmatrix} 3bc/ad & 0 \\ b/a & 1 \end{pmatrix}, \quad \begin{pmatrix} u & 0 \\ v & 1 \end{pmatrix} = \begin{pmatrix} d/3b & 0 \\ d/3 & 1 \end{pmatrix}.$$

Hence S is simple, but not completely simple.

Before we proceed to the formulation of the structure theorem for completely simple semigroups, we present some preparatory results, useful in their own right.

Proposition 1.6. If S is a simple semigroup, then

- (i) If the idempotent $e \in S$ is primitive, then Se and eS are minimal left and right ideals, respectively; the set $eSe = Se \cap eS$ is a group.
- (ii) S is completely simple if and only if it contains at least one minimal left ideal and at least one minimal right ideal.

By a minimal (left, right, two-sided) ideal of a semigroup S we mean of course a (right, left, two-sided) ideal that does not properly contain a (right, left, two-sided) ideal of S .

Proof. (i) Our first task is to show that $xS = eS$ for any x of the form es . $SxS = S$, hence $txu = e$, for some $t, u \in S$. In fact, we can choose t from eSe and u from Se , since e is idempotent and x belongs to eS .

Look at xut . We have $(xut) \cdot (xut) = xuet = xut$. Hence, xut is idempotent. Furthermore $e(xut) = xut = (xut)e$. On the other hand, e was assumed to be primitive. Thus, $xut = e$, i.e., $xS = eS$. Hence, eS is a minimal right ideal. The minimality of Se can be established in the same way.

Clearly, $eSe \subset Se \cap eS$. Then e is both right and left identity (thus the unique identity) for $Se \cap eS$, which implies that $Se \cap eS \subset eSe$. The subsemigroup eSe is right and left simple by the preceding minimality result. Proposition 1.1 tells us that eSe is a group.

(ii) The only if-statement follows from (i).

Let $R(L)$ be a minimal right (left) ideal of the simple semigroup S . Then for $r, r' \in R$, $rr'S = rR = R$; a similar statement holds for L . Thus, $RLRL = RL$ and RL is both right and left simple and hence a group by Proposition 1.1.

Let e be the identity of the group RL . The e is a primitive idempotent: If $fe = ef = f$, then f is an idempotent belonging to RL , so it has to coincide with the identity e of the group. Thus S is completely simple. $\square$

We can now formulate the important structure theorem due to Rees and Suschkewitsch.

Theorem 1.1. *A semigroup S is completely simple if and only if it is isomorphic to a Rees product $E \times G \times F$ with some sandwich function ϕ .*

Proof. It is quite straightforward and easy to show that a Rees product is completely simple.

To prove the converse, let S be a completely simple semigroup. Let e_0 be a primitive idempotent. Then Se_0 and e_0S are minimal left and right ideals, respectively, by Proposition 1.6, part (i). Their intersection is a group, call it G and e_0 is the identity of G . Denote the set of idempotents of Se_0 (e_0S) by E (F). E is a left zero semigroup, since for e , an idempotent of Se_0 , we have $Se_0 = Se$ and e is a right identity of Se . Similarly, F is a right zero semigroup: $f_1f_2 = f_2$ for $f_1, f_2 \in F$.

Define a Rees product $E \times G \times F$ with multiplication rule $(e, g, f)(e', g', f') = (e, gfe'g', f')$. This product is well defined, since $FE \subset e_0Se_0 = G$. The sandwich function simply maps (f, e') to fe' .

We now claim that the mapping $\Psi : E \times G \times F \mapsto S$ defined by $\Psi(e, g, f) = egf$ is a semigroup isomorphism. Ψ is clearly a homomorphism. To see that it is injective, notice that

$$\begin{aligned} (egf)(e_0egfe_0)^{-1} &= (egf) \cdot e_0 \cdot ((e_0e) \cdot g \cdot (fe_0))^{-1} \\ &= (eg) \cdot (fe_0)(fe_0)^{-1} \cdot g^{-1} \cdot (e_0e)^{-1} \\ &= e \cdot ge_0g^{-1} \cdot e_0 = ee_0 = e. \end{aligned}$$

Also $(e_0egfe_0)^{-1}(egf) = f$ and

$$e_0egfe_0 = e_0ge_0 = g.$$

(All inverses are taken in the group $G \equiv e_0Se_0$.) Hence, Ψ is injective with inverse

$$s \mapsto (s(e_0se_0)^{-1}, e_0se_0, (e_0se_0)^{-1}s)$$

on its range.

To complete the proof, we have to show that Ψ is surjective; i.e., its range is all of S . To this end, let $s \in S$ and $e \in E$. Then $se \in Se_0$, which is a left group, by Propositions 1.2 and 1.6 (i); and hence for some idempotent $e_1 \in Se_0$, $se = e_1se = (e_1e_0)se = e_1(e_0se)$. Since $e_0se \in e_0Se_0 \equiv G$ so we can infer that se belongs to

the set EG and thus $s(egf) \in EGF$ for any $s \in S$. Similarly, $(egf)s \in EGF$. Hence EGF , the range of Ψ , is a two-sided ideal of S and consequently equal to S , since S is simple by assumption. $\square$

Definition 1.7. The description of a completely simple semigroup S as a Rees product $E \times G \times F$ is called its *Rees–Suschkewitsch representation*, and E , G , and F are called the *left*, *group* (or *middle*), and *right factor* of S , respectively.

Corollary 1.1. A completely simple semigroup S is the union of all its minimal right ideals R_α as well as the union of all its minimal left ideals L_β , where α and β range over some index sets. Furthermore, S is the union of the isomorphic groups $R_\alpha L_\beta$.

Proof. From Theorem 1.1, it is immediate that S is the union of the minimal right ideals $\{e\} \times G \times F$ and the union of the minimal left ideals $E \times G \times \{f\}$. (Index sets for α and β can thus be identified with E and F , respectively). S can also be written as the union of the isomorphic groups $\{e\} \times G \times \{f\}$. $\square$

Definition 1.8. A completely simple minimal two-sided ideal of a semigroup is called its *kernel*. A finite semigroup always has a kernel. This insight goes back to pioneering work by Suschkewitsch, [219].

Proposition 1.7. A semigroup S with at least one minimal left ideal and at least one minimal right ideal contains a completely simple minimal two-sided ideal K . In particular any finite semigroup S has a minimal two-sided ideal K ; K is a completely simple subsemigroup of S .

Proof. Let K consist of those elements of S that generate a minimal left ideal. Thus, K is the union of all the minimal left ideals of S . The set $Ssx \subset Sx$ is a minimal left ideal of S if $x \in K$ and $s \in S$. The left ideal Sxs is minimal as well: Any $txs \in Sxs$ generates a (minimal) left ideal that contains xs on account of the minimality of Sx . Thus, K is a two-sided ideal. It is a minimal two-sided ideal, since for any $k, x \in K$, we have $x \in Skx$ and hence $SkS = K$.

Let R be a minimal right ideal of S . Then R is a minimal right ideal of K as well. Note that the minimal left ideals Sx are minimal left ideals Kx of K too. Thus, we can conclude that the simple semigroup K is completely simple by Proposition 1.6.

If S is finite, then we note that x generates a minimal left ideal of S if and only if $|Sx|$, the cardinality of Sx , is minimal. Hence, S always admits a minimal left ideal and a minimal right ideal. $\square$

AQ1 In the important case of a finite transformation semigroup, the kernel has a simple interpretation (Table 1.1).

Proposition 1.8. Let S be a subsemigroup of $\mathcal{T}_X$, where X is a finite set. Then the completely simple minimal ideal K of S consists precisely of those elements that have minimal rank.

tl.1 **Table 1.1** An egg box picture of the structure
of a completely simple semigroup as the union
of minimal right ideals R_α as well as the union
of minimal left ideals L_β . $R_\alpha L_\beta = R_\alpha \cap L_\beta \equiv$
 $G_{\alpha\beta}$ are isomorphic groups; $L_0 = Se_0$, $R_0 =$
 e_0S . The idempotents of the first column (row)
make up the sets E (F)

tl.1		L_0		L_β	L_γ
tl.2	R_0	G_{00}		$G_{0\beta}$	$G_{0\gamma}$
tl.3					
tl.4	R_α	$G_{\alpha 0}$		$G_{\alpha\beta}$	$G_{\alpha\gamma}$
tl.5	R_δ	$G_{\delta 0}$		$G_{\delta\beta}$	$G_{\delta\gamma}$

Proof. The reasoning is based on the discussion of $\mathcal{T}_X$ (X finite) in Sects. 1.1 549
and 1.2. 550

The set of mappings K of minimal rank r constitutes a two-sided ideal of S . The 551
ranges of *all* the elements of K are cross-sections of *all* the possible partitions gen- 552
erated by the elements of K , otherwise a product of two elements could be formed 553
with lower rank than r . 554

Any $k \in K$ generates a group of mappings with the same range and the same 555
partition as k . The subset of K consisting of mappings with common range R_0 , 556
say, is a right ideal. By Proposition 1.3, it is a right group and hence a minimal 557
right ideal. Similarly, the set of mappings in K with partition π_0 constitutes a min- 558
imal left ideal. By Proposition 1.6(ii) or Corollary 1.1, K is the completely simple 559
minimal ideal of S . $\square$ 560

The treatment of finite dimensional linear transformations is not very different 561
from the case of finite transformation semigroups. These important semigroups will 562
be discussed more fully in Sect. 1.6. 563

Proposition 1.9. Let S be a semigroup of $d \times d$ real matrices under matrix mul- 564
tiplication. Suppose that S admits a completely simple minimal ideal K . Then K 565
consists precisely of the elements of S of minimal rank. 566

This semigroup is no longer necessarily finite. Thus, the existence of a com- 567
pletely simple minimal ideal is not assured in general. As usual, the rank of a matrix 568
(interpreted as a linear transformation on $\mathbb{R}^d$) is the dimension ($\leq d$) of its range. 569

Proof. The set of matrices of minimal rank is an ideal of S hence it contains the 570
minimal ideal K . Let x be a matrix of minimal rank. We have to show that $x \in K$. If 571
 e is a primitive idempotent of K , then $K = SeS$. In particular, $xex \in K$. Since both 572
 x and xex are matrices of minimal rank, they have the same range and null space. 573
Since $xex \in K$, there is an idempotent $e' \in K$ with $(xex)e' = e'(xex) = xex$. The 574
idempotent matrix e' is the projection onto the range of xex along the null space of 575
 xex . Then we also have $xe' = e'x = x$, $x \in K$. $\square$ 576

Remark 1.4. If S is known to be a simple semigroup of matrices and to contain 577
an idempotent e , then this idempotent is necessarily primitive. Hence such an S is 578

completely simple. To see this, note first that the elements of S have minimal rank. 579
If f is idempotent and $ef = fe = f$, then e and f have common null space and 580
range, so e must equal f . 581

It was pointed out in connection with the definition of a Rees product that differ- 582
ent sandwich functions may produce isomorphic semigroups. A constant sandwich 583
function, for example, gives rise to a semigroup isomorphic to a direct product of 584
 E , G , and F . 585

Proposition 1.10. *Suppose S and S' are completely simple semigroups with Rees– 586
Suschkewitsch representation $E \times G \times F$ with sandwich functions ϕ and ϕ' , 587
respectively. Then S and S' are isomorphic if there exist functions u and v from 588
 F to G and E to G , respectively, satisfying 589*

$$\phi(f, e) = u(f)\phi'(f, e)v(e), \quad (f, e) \in F \times E.$$

Proof. It is readily verified that Ψ defined by $\Psi(e, g, f) = (e, v(e)gu(f), f)$ is a 590
surjective and injective homomorphism from S to S' . $\square$ 591

Actually, the relationship between S and S' in Proposition 1.10 is essentially, 592
disregarding some obvious technical points, the necessary and sufficient condition 593
for two semigroups to be isomorphic. For a complete account of the homomorphism 594
theory for Rees products, see [43]. 595

As promised in the beginning of Sect. 1.4, we also treat the case of semigroups 596
with 0. The philosophy in many books, notably [43], is to consider these semigroups 597
first. The theory for semigroups without 0 then follows rather automatically and ef- 598
fortlessly. We have chosen to proceed differently, however, because 0 does make 599
many proofs and arguments much more tedious and involved, although not neces- 600
sarily any more difficult, obscuring the main ideas relevant in applications in later 601
chapters. 602

The discussion given here provides the principal line of thought; we feel it is not 603
too difficult for the reader to work out the details from our outline and proofs of 604
corresponding results without 0. 605

Some basic definitions in the theory of semigroups with 0 were given in Sect. 1.3. 606

Definition 1.9. Let S be a semigroup with 0. A *primitive idempotent* of S is a 607
nonzero idempotent minimal with respect to the partial order $\leq$, defined 608

$$e \leq f \iff ef = fe = e, \quad e, f \in E(S^*),$$

where $E(S^*)$ denotes the set of idempotents of $S \setminus \{0\}$. 609

Definition 1.10. A *completely 0-simple* semigroup is a 0-simple semigroup con- 610
taining a primitive idempotent. 611

A basic result is Proposition 1.11 612

Proposition 1.11. *If S is a 0-simple semigroup, then:*

- (i) *If the idempotent $e \in S$ is primitive, then Se and eS are 0-minimal left and right ideals, respectively; the set $eSe = Se \cap eS$ is a group with 0.*
- (ii) *S is completely 0-simple if and only if it contains at least one 0-minimal left ideal and at least one 0-minimal right ideal, and moreover*
- (iii) *A completely 0-simple semigroup S is the union of all its 0-minimal left (right) ideals.*

Proof (Sketch of proof). First part of (i): We note that the assumption of 0-simplicity implies that $SxS = S$ for all nonzero elements of S . If $x \in (eS)^* \equiv eS \setminus \{0\}$, then we prove that $xS = eS$ to establish the 0-minimality of eS .

As in the proof of Proposition 1.6, part (i), we look at an element xut , $t \in (eSe)^*$, $u \in (Se)^*$. This element is idempotent with the property that $e(xut) = (xut)e = xut$. In the present situation, we have to exclude the possibility $xut = 0$ before drawing the conclusion $xut = e$ and $xS = eS$. Indeed, $t(xut) = et = t$, so $xut \neq 0$. Thus, eS and Se have the desired minimality properties. $\square$

We will collect the most important features of a completely 0-simple semigroup in Proposition 1.12, whose detailed proof can be found in [43], Chap. 2.

Proposition 1.12. *Let S be a completely 0-simple semigroup. Then*

- (i) *S is the union of sets of the form xSx , $x \in S$, which either coincide or intersect only at 0;*
- (ii) *Each xSx equals eSf for some nonzero idempotents e and f ;*
- (iii) *The set eSf is either a group with 0 (a group cell), in which case it contains an idempotent e' such that $eSf = e'Se'$, or a null semigroup (a null cell) with $(eSf)^2 = 0$;*
- (iv) *For a fixed idempotent e_0 , all cells can be written xSy , where $x \in Se_0$ and $y \in e_0S$. In particular if eSf is a group cell, then we take $e \in E(Se_0)^*$ and $f \in E(e_0S)^*$*
- (v) *For $x \in S$, $x^2 \neq 0$, there exists an idempotent e in xSx , which is a group with zero, such that $ex = x \in xSx$.*

Theorem 1.2. *A semigroup S with 0 is completely 0-simple if and only if it is isomorphic to a regular Rees product $X \times G^0 \times Y$ over a group with 0.*

Proof (Sketch). In Remark 1.3, we saw that regularity of a Rees product (or equivalently its sandwich function) is necessary and sufficient for it to be 0-simple. Given a regular Rees product over G^0 with sandwich function ϕ , we can immediately exhibit a primitive idempotent: the identity $(x, (\phi(y, x))^{-1}, y)$ of a group cell.

Let S be a completely 0-simple semigroup. Construction of the Rees product is analogous to that in the proof of Theorem 1.1. with one important exception: X is defined, using the Axiom of Choice if necessary, as a set of nonzero representatives of equivalence classes with respect to the relation

$$x\mathcal{R}y \iff xSe_0 = ySe_0$$

on Se_0 [see Proposition 1.12(iv)] with the proviso that if such a class contains
an idempotent $\neq 0$ (i.e., is a group cell), then this idempotent is chosen as repre-
sentative. Of course, Y is determined in the same manner. The function Ψ maps
 $X \times G^0 \times Y$ homomorphically into S . We then show that the map is in fact surjec-
tive. It is also injective on the set of elements *not* belonging to the set $X \times \{0\} \times Y$;
the set $X \times \{0\} \times Y$ is mapped onto 0.

The inverse Ψ_0 is given by the explicit formula, as in Theorem 1.1, but only on
group cells. On null cells we have to appeal to the particular choice of
representatives. $\square$

Remark 1.5. If there are no proper divisors of zero in S , then X is just $E(Se_0)^*$ and
 $Y = E(e_0S)^*$ as in Theorem 1.1. There are in other words only group cells.

AQ2 Propositions 1.13 and 1.14 illustrate some particular cases (Table 1.2).

Proposition 1.13. Any finite 0-simple semigroup is completely 0-simple.

Proof. We must establish that there is a nonzero idempotent because then there will
be a minimal one too.

Every element of S generates a subsemigroup with an idempotent. If 0 is the only
idempotent, we can conclude that all elements of S are nilpotent: There is a k such
that $a^k = 0$ for all $a \in S$. This implies that the finite semigroup S itself is nilpotent,
since $a \notin aS$ for all $a \in S^*$: $a = as$ implies $a = as^k$ for $k = 1, 2, \dots$ and hence
 $a = 0$. Thus, all nonzero elements of the sequence of products $a_1, a_1a_2, a_1a_2a_3, \dots$
($a_i \in S, i = 1, 2, \dots$) must be different so $S^k = \{0\}$ for all k sufficiently large.
(In fact, any k larger than the number of elements of S will do.) However by our
assumption, $S^2 = S$ and hence $S^n = S$ for all $n = 1, 2, 3, \dots$ $\square$

Recall that for a transformation semigroup on a finite set X , J_k is defined as the
set of elements of rank at most k . We then have the result in Proposition 1.14.

t2.1 **Table 1.2** An egg box picture of the structure of a
completely 0-simple semigroup S . The S is the union
of sets R_α as well as the union of sets L_β . $R_\alpha L_\beta =$
 $R_\alpha \cap L_\beta \equiv G_{\alpha\beta}^0$ are either group cells (in which case,
they are isomorphic groups with 0) or null cells. Note
that the cells are not disjoint: 0 belongs to every cell.
 $Se_0 = L_0, e_0S = R_0$. The set X (Y) is made up of
representatives of cells in the first column (row). In the
case of a group cell, the identity of the group is chosen
as representative

t2.2		L_0		L_β	L_γ
t2.3	R_0	G_{00}^0		$G_{0\beta}^0$	$G_{0\gamma}^0$
t2.4					
t2.5	R_α	$G_{\alpha 0}^0$		$G_{\alpha\beta}^0$	$G_{\alpha\gamma}^0$
t2.6	R_δ	$G_{\delta 0}^0$		$G_{\delta\beta}^0$	$G_{\delta\gamma}^0$

Proposition 1.14. *Consider the full transformation semigroup on the finite set X . Then the Rees factor semigroup J_k/J_{k-1} is completely 0-simple.*

Proof. We saw in Sect. 1.1 that transformations with a given partition mapping a particular k -element subset of X onto itself form a group; in this case, the common range is automatically a cross section of the given partition. Right groups and left groups of transformations on X were studied in Sect. 1.2.

Take e to be the identity of such a subgroup of rank k , i.e., e is the identity mapping on a k -element subset $e(X)$ that is a cross section of the partition $\pi(e)$. We claim that $SeS = S$, where S is the Rees quotient J_k/J_{k-1} .

Let a be a transformation on X with range $a(X)$ and partition $\pi(a)$. Let the range have cardinality k ; it is however not assumed that $a(X)$ is a cross section of $\pi(a)$. We exhibit transformations (of rank k) s, t such that $a = set$.

Let the mapping u be defined as follows: u is a bijection from $a(X)$ to $e(X)$ (both of which have cardinality k); u maps all elements outside of $a(X)$ onto a single point in $e(X)$. Let $t = ua$.

Let s be the inverse of u in the following sense: $su(= seu)$ is the identity mapping on $a(X)$; outside of $e(X)$, s maps, say, all elements onto a single point in $a(X)$. Then $a = seua = set$.

Also, $SaS = S$. To see this, note that it is possible for any cross section C of $\pi(a)$ to find an element g of rank k that maps $a(X)$ onto C [and the elements outside of $a(X)$ onto arbitrary elements of C]. The mapping ga then has partition $\pi(a)$ and range C , which means that some power of ga is the identity element e of the group of mappings with this particular partition and range. We saw that $SeS = S$ and of course, $SaS \supset SgaS \supset SeS$.

Thus, the finite semigroup S is 0-simple and hence completely 0-simple. $\square$

As yet another illustration of the close kinship between finite transformation semigroups and semigroups of finite dimensional linear transformations, we offer the following result, whose proof is given in Sect. 1.6.

Consider the semigroup M_d of all $d \times d$ real or complex matrices. If the ideal consisting of matrices of rank no larger than k is denoted by T_k , then the Rees factor semigroup T_k/T_{k-1} is completely 0-simple.

Actually, Propositions 1.13 and 1.14 can be viewed as special cases of the following very convenient characterization (Theorem 1.3).

Theorem 1.3. *A (0-)simple semigroup S is completely (0-)simple if and only if some power of each element of S lies in a subgroup of S .*

Before proving Theorem 1.3, let us look briefly at simple semigroups with only nonprimitive idempotents.

Definition 1.11. Let $\mathcal{C} = \mathcal{C}(p, q)$ consist of the empty word e and all the words generated by the two-element alphabet $\{p, q\}$ subject to the single generating rule $pq = e$. Under the operation of juxtaposition of words $\mathcal{C}(p, q)$ is then a semigroup called the *bicyclic semigroup* generated by p and q . We can prove that elements of $\mathcal{C}(p, q)$ are uniquely representable in the form $q^m p^n$, $m, n \geq 0$, if we agree to write e as $q^0 p^0$. The multiplication rule is

$$(q^m p^n)(q^{m'} p^{n'}) = \begin{cases} q^{m+m'-n} p^{n'}, & \text{if } m' \geq n; \\ q^m p^{n'+n-m'}, & \text{if } m' < n. \end{cases}$$

The idempotents are of the form $q^n p^n$, $n \geq 0$. It is easily seen that there are no primitive idempotents

$$(q^{n+1} p^{n+1})(q^n p^n) = q^{n+1} p^{n+1}; \quad (q^n p^n)(q^{n+1} p^{n+1}) = q^{n+1} p^{n+1}$$

$\mathcal{C}(p, q)$ is simple, however. To see this, note that $e \in \mathcal{C}q^n p^n \mathcal{C}$, since $(q^0 p^n)(q^n p^n)(q^n p^0) = e$. Hence, $\mathcal{C}$ is simple but not completely simple.

A concrete example of a bicyclic semigroup is the subsemigroup of $\mathcal{T}_N$ (where N is the set of nonnegative integers) generated by the identity map $\iota = \iota_N$ and the two maps α and β

$$\alpha(n) = \begin{cases} 0, & \text{if } n = 0, \\ n - 1, & \text{if } n > 0. \end{cases}$$

$$\beta(n) = n + 1, n \in N.$$

Then $\alpha\beta = \iota$ and $\beta\alpha \neq \iota$. Transformations of N generated by α and β can be written $\beta^m \alpha^n$, $m, n \geq 0$ in a unique fashion, where ι is of course, $\beta^0 \alpha^0$.

The bicyclic semigroup is in a sense *the* example of a semigroup that is simple without being completely simple.

Proposition 1.15. *If e is any nonzero idempotent of a (0-)simple semigroup S that is not completely (0-)simple, then S contains a bicyclic subsemigroup with identity e .*

Proof (for the case with 0). The e is by definition of a completely 0-simple a non-primitive idempotent of S . Then there exists another nonzero idempotent $f \in S$ such that $ef = fe = f$. S is 0-simple, so $e = xfy$ for some $x, y \in S^* \equiv S \setminus \{0\}$. Set $p = exf$ and $q = fye$, then $pq = eee = e$. Direct calculations show that qp is also idempotent and nonzero, but $fqp = qpf = qp$ so $qp \leq f < e$ in the ordering of idempotents. In particular, $qp \neq e$. The preceding relations $pq = e$, $qp \neq e$, and the fact that e is a two-sided inverse of the subsemigroup generated by p and q , guarantee that this subsemigroup is isomorphic to $\mathcal{C}$, [see [43], Sect. 1.12.] $\square$

Proof (of Theorem 1.3, case with 0). The square of each element of a completely 0-simple semigroup lies in a group (Definition 1.4).

Conversely assume that some power of each element of S is contained in a subgroup of S . Let $a \in S^*$. Then $a \in S^* a S^*$, since S is 0-simple. Let $a = xay$. By iterating this equality, we obtain $a = x^n a y^n$. For some n , x^n belongs to a subgroup of S , indeed of S^* . Let $e \neq 0$ be the identity element of this subgroup.

Suppose S is not completely 0-simple. Then e is not a primitive idempotent, and by Proposition 1.15, e is the identity element of a bicyclic subsemigroup of S with generators p and q , say $(pq = e, qp \neq e)$. For some n , p^n belongs to a group with identity f . Let r be the inverse of p^n in that group: $rp^n = p^n r = f$. We have $fe = fp^n q^n = p^n q^n = e$ and $fe = rp^n e = rp^n = f$, so $f = e$. But then

$r = p^n$ and $p^n q^n = q^n p^n = e$, which is not possible in the bicyclic semigroup generated by p and q . Hence, S must be completely 0-simple. $\square$

Remark 1.6. Semigroups S with some power of each element in a subgroup of S are termed *pseudoinvertible*. Thus, a simple semigroup is completely simple if and only it is pseudoinvertible.

Exercise 1.21. Prove that a Rees product is completely simple.

Exercise 1.22. Show that if $\bigcap_{x \in S} SxS$ is nonempty, then it is isomorphic to a Rees product $E \times G \times F$ with some sandwich function ϕ .

Exercise 1.23. Prove that a Rees product with a sandwich function that maps everything onto a constant is isomorphic to a direct product of E , G , and F .

Exercise 1.24. Prove that a subsemigroup U of $\mathcal{T}_n$ is completely simple if and only if $\text{rank}(\alpha) = \text{rank}(\beta)$ for all $\alpha, \beta \in U$.

Exercise 1.25. Let $\mathcal{C} = \mathcal{C}(p, q)$ be a bicyclic semigroup (as in Definition 1.11). Consider the subsets:

$$\begin{aligned} S_1 &= \{q^{4i+5}p^{4j+5} : i, j \geq 0\} \\ S_2 &= \{q^{4i+7}p^{4j+7} : i, j \geq 0\} \end{aligned}$$

Prove that both S_1 and S_2 are subsemigroups and that they are isomorphic to $\mathcal{C}$. Prove that their union $S = S_1 \cup S_2$ is also a subsemigroup.

Exercise 1.26. Show by using the properties of minimal ideals that the bicyclic semigroup contains no minimal left or minimal right ideal.

1.5 Topological Semigroups

Let S be a semigroup. Suppose S is also endowed with a Hausdorff topology. If the map $S \times S \ni (s, t) \mapsto st \in S$ is continuous S is called a *topological semigroup*. (The topology on $S \times S$ is the product topology.) In other words, a semigroup whose multiplication is jointly continuous in the factors is a topological semigroup.

If the multiplication is only *separately continuous* in the factors, we call S a *semitopological semigroup*. Equivalently S is semitopological if the right translations ρ_a and the left translations λ_a are continuous mappings from S to S for all $a \in S$. However, we always assume that our semigroups are topological unless explicitly stated otherwise. Similarly, the topology is always assumed to be Hausdorff.

Subsemigroups of a topological semigroup are automatically topological semi- groups with respect to the induced topology. By an *open (closed, compact, locally compact) subsemigroup (ideal) of S* we mean, respectively, an open (closed, compact, locally compact) subset of S that is algebraically a subsemigroup (ideal) of S .

We collect some elementary facts about subsets of a topological semigroup in Proposition 1.16.

Proposition 1.16. *Let S be a topological semigroup. Let F be a closed subset, O an open subset, A an arbitrary subset and C, C_1, C_2 compact subsets of S .*

- (i) *For every element $s \in S$, sC, Cs , and C_1C_2 are compact, $s^{-1}F \equiv \lambda_s^{-1}(F) \equiv \{t \in S | st \in F\}$, $Fs^{-1} \equiv \rho_s^{-1}(F) \equiv \{t \in S | ts \in F\}$ are closed, and $s^{-1}O, Os^{-1}$,*

$$A^{-1}O \equiv \bigcup_{s \in A} s^{-1}O, OA^{-1} \equiv \bigcup_{s \in A} Os^{-1}$$

are open subsets of S . If T is a subsemigroup of S , then so is $\bar{T}$, the closure of T .

- (ii) *If in addition S is algebraically a group, then the translations are homeomorphisms from S to S . In particular for any $s \in S$, sF, Fs, CF are closed, and sO, Os, AO, OA are open subsets of S . In addition, C^{-1} is closed. Furthermore, sU (Us) is a neighborhood of s if and only if U is a neighborhood of the identity e . If the net $x_\alpha \rightarrow x$ and $x_\alpha^{-1} \rightarrow y$, then $y = x^{-1}$.*
- (iii) *The set of idempotent elements of S , $E(S)$, is closed. If e is an idempotent of a closed subsemigroup T , then eT, Te, eTe are closed.*

Note that (ii), $\lambda_s^{-1} = \lambda_{s^{-1}}$, so our notation in (i) for $s^{-1}F$ is consistent with left translation of the set F by the element s^{-1} . (C^{-1} is of course a notation for the set $\{s^{-1} | s \in C\}$.)

Proof. (i) C_1C_2 is the image under multiplication of the compact set $C_1 \times C_2$ in $S \times S$ and hence compact. Likewise, sC, Cs are continuous images of a compact set.

The sets $s^{-1}F, Fs^{-1}$ are preimages of closed sets, hence closed, and $s^{-1}O, Os^{-1}$ are preimages of open sets, hence open. $A^{-1}O, OA^{-1}$ are in turn unions of open sets, hence also open. If t_α, u_α are nets in T converging to t and u , respectively, then $t_\alpha u_\alpha \rightarrow tu$, so $tu \in \bar{T}$.

(ii) Since $\lambda_s (\rho_s)$ has the continuous inverse $\lambda_{s^{-1}} (\rho_{s^{-1}})$, most of the assertions follow directly from the results in (i).

To see that C^{-1} is closed and to prove the last assertion of (ii), consider a net c_α^{-1} converging to an element $g \in S$. The net c_α has a subnet, say, c_β , converging to some c in the compact set C . Then $e = c_\beta c_\beta^{-1}$. By the joint continuity of the multiplication, the limit of the right-hand side equals cg , so $g = c^{-1}$ and hence $g \in C^{-1}$.

(iii) The statements are immediate consequences of, respectively, the joint continuity and the continuity of the multiplication. $\square$

One of our basic examples is the semigroup M_d of $d \times d$ real matrices. If we endow M_d with the usual topology of $\mathbb{R}^{d^2}$, then matrix multiplication is jointly continuous, so M_d is a topological semigroup with respect to this topology. It is locally compact, since $\mathbb{R}^{d^2}$ is. The subgroup $GL(d)$ of invertible matrices is an open subset of $M_d - g \in GL(d) \iff \det(g) \neq 0$ – and thus a locally compact space. Hence by Ellis' theorem (Theorem 1.4) $GL(d)$ is a topological group. [Actually, we could just as well have used the explicit formula for the matrix inverse to see that inversion is continuous in $GL(d)$.]

More generally, the continuous mappings of $\mathbb{R}^d$ into itself form a semigroup with an ordinary composition of mappings as binary operation. Under the topology of pointwise convergence, this semigroup is only semitopological: If $f_n \rightarrow f$ and $g_n \rightarrow g$ pointwise, then $f_n \circ g \rightarrow f \circ g$ and $f \circ g_n \rightarrow f \circ g$ pointwise. The operation is not jointly continuous; take for example $g_n(x) = \frac{1}{n}$ and $f_n(x) = n|x|e^{-n|x|}$ for $x \in \mathbb{R}^1$. We have $f_n(x), g_n(x) \rightarrow 0$ but $f_n(g_n(x)) = e^{-1}$.

On the other hand, if we restrict ourselves to some smaller class of mappings, then the composition of mappings may very well be jointly continuous. Such smaller classes are for example, the contraction mappings from $\mathbb{R}^d$ into itself and the affine mappings of $\mathbb{R}^d$ into itself.

Recall that a *topological group* is a topological space that is algebraically a group with jointly continuous multiplication and continuous inversion. If G is (algebraically) a subgroup of the topological semigroup S , then multiplication is automatically jointly continuous on G , but the inversion $g \mapsto g^{-1}$ need not be continuous in general.

Fortunately, we have at our disposal a powerful theorem by Ellis [65, 66] that settles the matter in most cases of interest to us.

Theorem 1.4 (Ellis's Theorem). *Let G be a locally compact space that is algebraically a group. Suppose further that multiplication in G is separately continuous. Then G is a topological group.*

We do not prove Theorem 1.4 in its general form. Instead, we prove a version directly adapted to the case of subgroups of a locally compact topological semigroup:

Theorem 1.5. *Let S be a locally compact topological semigroup. If S is algebraically a group, then it is a topological group.*

By Proposition 1.1, we immediately have Corollary 1.2.

Corollary 1.2. *Let S be a locally compact topological semigroup. Then S is a topological group if and only if it is both left and right simple.*

Before we prove Theorem 1.5 we need Lemmas 1.2 and 1.3.

Lemma 1.2. *Assume S as in Theorem 1.5. Let E be a countable subset of S and let x be a limit point of E . Then x^{-1} is a limit point of E^{-1} .*

Proof. Let x_α be a net in E converging to x . We need to show that x_α^{-1} converges to some y , because then y is in fact x^{-1} .

Set $B = E \cup \{x\}$, then

$$D \equiv \bigcup_{n=-\infty}^{\infty} B^n$$

is the group generated by B . Let $A = \bar{D}$, the closure of D ; then A is a subsemigroup of S , since D is one by Proposition 1.16(i). $A \subset DV^{-1}$, where V is a compact neighborhood of the identity. Thus, $A = \bigcup_{d \in D} (dV^{-1} \cap A) = \bigcup_{d \in D} (d(V^{-1} \cap A))$, because D is a group and A is its closure. $d(V^{-1} \cap A)$ is a closed set [Proposition 1.16(ii)] for all $d \in D$. A is a locally compact space (it is a closed subset of the locally compact space S) expressible as a countable union of closed sets. By the Baire category theorem, one of the sets $d(V^{-1} \cap A)$ has nonempty interior relative to A . For this d , there is an open set N with $N \cap A$ nonempty and $N \cap A \subset d(V^{-1} \cap A)$. Consequently, $N \cap D$ is also nonempty.

Take $c \in N \cap D$. Then $xc^{-1}(N \cap A) = xc^{-1}N \cap A$ is an open set (relative to A) containing x . There is a neighborhood U of x such that $x \in U \subset xc^{-1}(N \cap A) \subset xc^{-1}dV^{-1}$. Hence, $U^{-1} \subset Vd^{-1}cx^{-1}$, a compact set. Consequently, a subnet x_β^{-1} converges to some point in S , which by Proposition 1.16(ii) must then be x^{-1} . $\square$

Lemma 1.3. For C a compact subset of S (with S as in Theorem 1.5), C^{-1} is compact.

Proof. Let V be a compact neighborhood of the identity. We show that C^{-1} can be covered by a finite number of translates of V . If this claim is false, then there exists a sequence of elements $x_n^{-1} \in C^{-1}$ such that $x_n^{-1} \notin \bigcup_{k=1}^{n-1} x_k^{-1}V$. Select E_n such that $\{x_k | k \geq n\}$.

Since C is compact $\bigcap_{n=1}^{\infty} \bar{E}_n$ is nonempty. Take a point x in this intersection. Let U be a neighborhood of the identity, with $U^2 \subset V$. Since $x \in \bar{E}_1$, $x_m \in Ux$ for some m . Then $x^{-1} \in x_m^{-1}U$, and $x \in \text{cl}(E_{m+1})$ implies (Lemma 1.2) that $x^{-1} \in \text{cl}(E_{m+1}^{-1})$. Hence for some $n > m$, $x_n^{-1} \in x^{-1}U \subset x_m^{-1}U^2 \subset x_m^{-1}V$. This is a contradiction because x_n^{-1} was chosen outside of $x_m^{-1}V$ for all $m < n$. $\square$

Proof (of Theorem 1.5). It suffices to prove the continuity at e . Let U be an open neighborhood of e . We must find a compact neighborhood V of e , such that $V^{-1} \subset U$. Suppose that this is impossible; i.e., $V^{-1} \setminus U$ is not empty for $V \in \mathcal{C}$, the family of compact neighborhoods of the identity. Since the sets $V^{-1} \setminus U$ are compact we infer that the intersection $\bigcap_{V \in \mathcal{C}} V^{-1} \setminus U$ is also nonempty. But this is impossible, since $e \in \bigcap_{V \in \mathcal{C}} V^{-1}$ and $e \in U$. $\square$

Remark 1.7. Most of our applications require the topology to be second countable; i.e., there is a countable base for the topology. For this case, Theorem 1.5 follows directly from Lemmas 1.3 and 1.2. The set $\text{cl}(\{x_n\})$ is compact if $x_n \rightarrow x$, so $\text{closure}(\{x_n^{-1}\})$ is compact too. Thus, all subsequences of $\{x_n^{-1}\}$ have convergent subsequences, all converging by Lemma 1.16(ii) to x^{-1} .

Our next objective is to study the ideal structure and the Rees–Suschkewitsch representation theory in this topological setting.

Proposition 1.17. *Let S be a compact semigroup. Then each left (right) ideal of S contains at least one minimal left (right) ideal and each minimal left (right) ideal is closed and principal, i.e., of the form Sa (aS) for some $a \in S$.*

Proof. Given a left ideal I , define the family of sets $\mathcal{F} \equiv \{J \mid J \text{ a closed left ideal of } S, J \subset I\}$. For any $a \in I$, Sa is closed [Proposition 1.16(i)] and a left ideal, hence a member of $\mathcal{F}$. The usual inclusion relation is a partial order on $\mathcal{F}$. Furthermore, any linearly ordered subfamily of $\mathcal{F}$ has a minimal element, since sets in $\mathcal{F}$ are compact. By Zorn's lemma, there exists at least one minimal element (with respect to inclusion) in $\mathcal{F}$. Call this minimal closed left ideal I_0 . It remains to prove that I_0 is a minimal left ideal of S . Suppose that $I_1 \subset I_0$, I_1 ideal of S . Then for $a \in I_1$, Sa is a closed ideal of S included in I . Hence, $Sa = I_0$ and $I_0 = Sa \subset I_1$ so $I_1 = I_0$. The last argument also shows that I_0 must be equal to Sx for any $x \in I_0$. $\square$

Theorem 1.6. *Let S be a compact semigroup. Then S has a kernel K , i.e., a minimal two-sided ideal. K is a closed, completely simple semigroup.*

Proof. Let K be the set of all those $x \in S$ that generate a minimal left ideal. By Proposition 1.17 K can also be described as the union of minimal left ideals. K is a left ideal since all the elements $Sx, x \in K$, also generate minimal left ideals. K is a right ideal because xt generates a minimal left ideal for each $x \in K$ and each $t \in S$. Thus, K is a bilateral ideal.

Let $x, k \in K$ be arbitrary. Then $x \in Skx$. Hence $SkS = K$ for each $k \in K$, so K is minimal. It is also a principal ideal and thus closed.

The minimality properties of the elements of K imply that $Sx = Kx, xS = xK, SxS = KxK$ for $x \in K$ so we can conclude, by Proposition 1.6, that K is completely simple. $\square$

Remark 1.8. Note the similarity between this proof and that of Proposition 1.7, the finite semigroup case.

Corollary 1.3. *If S is a compact abelian semigroup, then its kernel K is a compact abelian topological group.*

Proof. In the abelian case, there is only one minimal left (and right) ideal because Sx and Sy can never be disjoint. K is then left and right simple. It is a group by Proposition 1.1 and a topological group by Ellis's Theorem. $\square$

Remark 1.9. (i) Let $a \in S$, where S is second countable and compact but not necessarily abelian. $A \equiv \text{cl}(\{a^n \mid n \geq 1\})$ is called the *monothetic* subsemigroup generated by a . The kernel K of A is a compact abelian topological group. K is generated by au , where u is the identity of K ; K is simply $Au = Au^2 = uAu$. (ii) Any compact subsemigroup S of a group G is in fact a topological group: The group identity e belongs to S , and it is its only idempotent; moreover, $eS = Se = S$.

The set of $d \times d$ stochastic matrices form a compact semigroup S . Recall that $P = (p_{ij})$ is stochastic if $0 \leq p_{ij} \leq 1$ and $\sum_{j=1}^d p_{ij} = 1$. (If the row sum is ≤ 1 , then the matrix is said to be substochastic; these matrices also form a compact semigroup.) A stochastic matrix of rank 1 has identical rows, and it is easy to see that such a matrix m by itself constitutes a minimal left ideal $\{m\} = Sm$. On the other hand, the set mS is precisely the set of all stochastic matrices of rank 1. So those matrices will constitute the kernel K of S . We treat matrix semigroups in greater detail in Sect. 1.6.

We now turn to locally compact semigroups. We note in passing that compact semigroups with 0 quite naturally fit in here because removal of 0 makes the rest of the space locally compact. In case the semigroup S has no proper divisors of zero, the treatment is straightforward (since $S^* \equiv S \setminus \{0\}$ is then a subsemigroup of S). In the general case however, much more elaborate methods are called for.

A *topological isomorphism* between two semigroups S and S' means a mapping $\Psi : S \mapsto S'$ that is both a semigroup isomorphism and a homeomorphism.

Proposition 1.18. *Let S be a locally compact semigroup that is a right group. Then S is topologically isomorphic to the direct product $G \times Y$, where G is a locally compact topological group and Y a locally compact right zero semigroup.*

Proof. As in the proof of Proposition 1.2, we take $e_0 \in E(S)$, the idempotents of S , $G = Se_0$ and $Y = E(S)$. Then, by Proposition 1.16(iii), G and Y are closed subsets of S , hence locally compact spaces. G is a group (algebraically) and consequently a topological group by Ellis's Theorem.

$\Phi : G \times Y \mapsto S$ defined by $\Phi(g, y) = gy$ is an isomorphism (Proposition 1.2). It is also continuous, and has the continuous inverse

$$s \mapsto (se_0, (se_0)^{-1}s), s \in S.$$

We use here the continuous inversion in G and the fact that any idempotent of S is a left identity: $(se_0)^{-1}s \cdot (se_0)^{-1}s = (se_0)^{-1}[se_0(se_0)^{-1}s] = (se_0)^{-1}s$. $\square$

Theorem 1.7. *Let S be a locally compact semigroup with a completely simple kernel K . Let $e_0 \in E(K)$, $G = e_0Ke_0$, $X = E(Ke_0)$, $Y = E(e_0K)$. Define a Rees product $X \times G \times Y$ with multiplication*

$$(x, g, y)(x', g', y') = (x, gyx'g', y').$$

Then K is closed and topologically isomorphic to the Rees product $X \times G \times Y$ just defined.

Proof. We showed the algebraic portions of the result in Theorem 1.1. There it was proved that Ψ , defined by

$$\Psi(x, g, y) = xgy, \quad x \in X, g \in G, y \in Y$$

is the required isomorphism, and if

971

$$\Psi'_0(s) = (s(e_0se_0)^{-1}, e_0se_0, (e_0se_0)^{-1}s), s \in S$$

then Ψ_0 , the restriction of Ψ'_0 to K , is its inverse.

972

Let us first note that X and Y are closed subsets of S and G is algebraically a group. The set $e_0Se_0 = e_0Ke_0$ is closed, hence locally compact and a topological group by Ellis' Theorem. This means that both Ψ and Ψ_0 are continuous. To see that K is closed, let $k_\alpha \rightarrow s \in S$. Then $k_\alpha = \Psi(\Psi'_0(k_\alpha)) \rightarrow \Psi(\Psi'_0(s)) \in K$ since $\Psi \circ \Psi'_0$ is a continuous function on S and the image of Ψ is K . $\square$

977

Corollary 1.4. *Let S be a completely 0-simple compact semigroup without proper divisors of zero. Then $S^* \equiv S \setminus \{0\}$ is topologically isomorphic to a Rees product $X \times G \times Y$, where the factors are locally compact.*

978

979

980

Proof. S^* is a locally compact subsemigroup of S containing primitive idempotents by assumption. S^* is simple since $SxS = S$ for all $x \in S^*$ hence completely simple. $\square$

983

Compactness of S does not assure the existence of any minimal ideals in S^* even in the absence of zero divisors. One example is $[0, 1]$ under ordinary multiplication. If, however, the compact second-countable semigroup S is known to be 0-simple, then a structure theorem can be established.

984

985

986

987

Proposition 1.19. *Let S be a compact second-countable 0-simple semigroup. Then S is completely 0-simple. 0 is an isolated point of S .*

988

989

We first prove Lemma 1.4.

990

Lemma 1.4. *Let a be any element of a topological semigroup S . Let C_1 and C_2 be compact subsets of S . If U is any open set containing the compact set C_1aC_2 , then there are open sets $O_1 \supset C_1, V$ containing $a, O_2 \supset C_2$, such that $O_1VO_2 \subset U$.*

991

992

993

Proof. By the joint continuity of the multiplication, there are open sets $O(x), V_{x,y}(a), O(y)$ containing, respectively, $x \in C_1, a$ and $y \in C_2$ such that $O(x)V_{x,y}(a)O(y) \subset U$. Since the sets $O(x) \times O(y), (x, y) \in C_1 \times C_2$, form an open covering of the compact set $C_1 \times C_2$, we can find a finite subcovering

994

995

996

997

$$\bigcup_{i=1}^n O(x_i) \times O(y_i) \supset C_1 \times C_2$$

such that $O(x_i)V_{x_i,y_i}(a)O(y_i) \subset U, i = 1, 2, \dots, n$. Then

998

$$\bigcup_{i=1}^n O(x_i) \bigcap_{i=1}^n V_{x_i,y_i}(a) \bigcup_{i=1}^n O(y_i) \subset U.$$

$\square$ 999

Proof (of Proposition 1.19). S is compact and 0-simple. Then $SVS \subset U \neq S$ implies that $V = \{0\}$, so we can conclude, from Lemma 1.4, that $\{0\}$ is open and hence isolated.

S contains a nonzero idempotent: Take $a \in S^*$. Then $a = xay$ for some $x, y \in S^*$. Furthermore, $a = x^na y^n$ so $x^n \neq 0$, and the monothetic semigroup generated by x is contained in S^* , a compact set. Its identity $e \neq 0$ is idempotent.

Since S^* is compact, the partial order $\leq$ on $E(S^*)$ must have a minimal element, a primitive idempotent. Hence, S is completely 0-simple. $\square$

Remark 1.10. The fact that $\{0\}$ is open is important from a probabilistic point of view. As seen in subsequent chapters, the natural assumption is that our probability measures will in a sense charge all open subsets of the semigroup; i.e., those sets will have positive probability.

Theorem 1.8. Let S be a compact second-countable 0-simple semigroup. Then S is topologically isomorphic to a Rees product $X \times G^0 \times Y$, where X and Y are compact left and right zero semigroups, respectively, and G^0 is a compact topological group with 0.

For a proof of Theorem 1.8, see [185], Chap. 2.

Remark 1.11. As we saw in the proof of Theorem 1.2 the choice of the side factors X and Y is somewhat arbitrary in the presence of zero divisors. The proof of Theorem 1.8 shows that it is possible to make that choice in a *continuous* fashion.

Before we go into the Rees–Suschkewitsch structure theory for locally compact semigroups with 0, let us look at a specific example in some detail.

Example 1.2. Let S be the set of real 2×2 matrices of rank ≤ 1 . S is a locally compact second-countable semigroup with 0 under ordinary matrix multiplication. The zero is not an isolated point of S .

There are plenty of idempotent elements in S . The idempotents are the 1-dimensional projection matrices and they have the following general form

$$\begin{pmatrix} 0 & c \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ b & 1 \end{pmatrix} \text{ or } \begin{pmatrix} (1+cb)^{-1} & (1+cb)^{-1}c \\ b(1+cb)^{-1} & b(1+cb)^{-1}c \end{pmatrix},$$

where a, b, c, d are real numbers, with $1+cb \neq 0$ in the third case. Direct calculations show that the idempotents are primitive: If $e \neq 0$ is an idempotent and $ef = fe = f$ for another idempotent f , then either $f = e$ or $f = 0$. It can also be shown that S is 0-simple, hence completely 0-simple. S contains many divisors of zero, so S^* is not a subsemigroup.

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

is 0 as well as the product of

1032

$$\begin{pmatrix} (1+cb)^{-1} & (1+cb)^{-1}c \\ b(1+cb)^{-1} & b(1+cb)^{-1}c \end{pmatrix} \text{ and } \begin{pmatrix} (1+c'b')^{-1} & (1+c'b')^{-1}c' \\ b'(1+c'b')^{-1} & b'(1+c'b')^{-1}c' \end{pmatrix}$$

when $1+cb' = 0$.

1033

When we move to more general cases, we encounter two main difficulties: The completely 0-simple semigroups may not be topological, owing to a discontinuity at 0, and we may be unable to choose the side factors X and Y in a continuous way. In such cases, the Rees–Suschkewitsch structure theorem may be valid only locally in a neighborhood of some nonzero idempotent.

Theorem 1.9. *Let S be a completely 0-simple semigroup with jointly continuous multiplication at all $(x, y) \in S \times S$ such that $xy \neq 0$. Suppose further that S^* is a locally compact space. Let e_0 be a nonzero idempotent. Then $G_0 \equiv (e_0 S e_0)^*$ is a locally compact topological group. If $A_0 \equiv E(S e_0)^*$ and $B_0 \equiv E(e_0 S)^*$, then $A_0 G_0 B_0$ is an open subset of S and the map*

$$\Psi(a, g, b) = agb$$

is a homeomorphism from $A_0 \times G_0 \times B_0$ to $A_0 G_0 B_0$.

1044

Remark 1.12. Suppose S has no divisors of zero. Then $A_0 G_0 B_0 = S^*$. This was the content of Corollary 1.4.

1045

1046

Proof (of Theorem 1.9). S^* is a locally compact space with jointly continuous multiplication whenever 0 is not involved. Ellis's theorem guarantees then that all closed or open (algebraic) subgroups of S^* are topological groups.

1047

1048

1049

Since S is completely 0-simple, there is a primitive idempotent $e_0 \in S^*$. $(e_0 S e_0)^*$ is a closed subgroup of S^* by Theorem 1.2, hence a topological group. A_0 and B_0 are also locally compact spaces. To see this, consider an idempotent $e \in A_0$. There is a compact neighborhood $O(e)$ of e such that $O(e)^2$ does not contain 0. Multiplication is then jointly continuous in $O(e)$. Hence, $O(e) \cap A_0$ is a compact neighborhood of e relative to A_0 [see Proposition 1.16(iii)].

1050

1051

1052

1053

1054

1055

By the same argument used in the proof of Theorem 1.1, we see that Ψ is a homeomorphism from $A_0 \times G_0 \times B_0$ to $A_0 G_0 B_0$. Note that $0 \notin A_0 G_0 B_0$ because $e_0(A_0 G_0 B_0)e_0 \subset G_0$, since $a \in A_0$ ($b \in B_0$) has the property that $e_0 a = e_0 a e_0 \neq 0$ ($b e_0 = e_0 b e_0 \neq 0$). In general, $A_0 G_0 B_0$ is not a subsemigroup of S because $B_0 A_0$ may contain 0.

1056

1057

1058

1059

1060

Consider a net $s_\alpha \rightarrow agb$, $a \in A_0$, $g \in G_0$, $b \in B_0$. Again, $e_0(agb)e_0 \neq 0$ so $e_0 s_\alpha e_0 \neq 0$ eventually. For those α , consider the unique representation of S described in Theorem 1.2: $s_\alpha = x_\alpha g_\alpha y_\alpha$, $x_\alpha \in S e_0$, $y_\alpha \in e_0 S$, $g_\alpha \in G_0$.

1061

1062

1063

Recall that x_α is nonidempotent if and only if $x_\alpha S e_0$ is a null cell: $(x_\alpha S e_0)^2 = 0$. This is the case if and only if $e_0 x_\alpha = 0$.

1064

1065

We see that $e_0 x_\alpha \neq 0$ so x_α is an idempotent of $S e_0$; hence, it belongs to A_0 . Similarly $y_\alpha \in B_0$.

1066

1067

t3.1 **Table 1.3** An egg box picture of the completely simple semigroup S . $R_\alpha L_\beta \equiv G_{\alpha\beta}^0$ are either group cells or null cells. $Se_0 = L_0, e_0S = R_0$. The set $X(Y)$ is made up of representatives of cells in the first column (row). In the case of a group cell, the identity of the group is chosen as representative. This is denoted by a (g) , the opposite case by a (n) in Table 1.3. The boldface cells (with 0 removed) constitute the set $A_0G_0B_0$

t3.2		L_0		L_β	L_γ	L_ϕ
t3.3	R_0	$\mathbf{G}_{00}^0(g)$		$G_{0\beta}^0(n)$	$\mathbf{G}_{0\gamma}^0(g)$	$\mathbf{G}_{0\phi}^0(g)$
t3.4						
t3.5	R_α	$\mathbf{G}_{\alpha 0}^0$		$G_{\alpha\beta}^0$	$\mathbf{G}_{\alpha\gamma}^0$	$\mathbf{G}_{\alpha\phi}^0$
t3.6	R_δ	$G_{\delta 0}^0$		$G_{\delta\beta}^0$	$G_{\delta\gamma}^0$	$G_{\delta\phi}^0$
t3.7	R_ε	$G_{\varepsilon 0}^0$		$G_{\varepsilon\beta}^0$	$G_{\varepsilon\gamma}^0$	$G_{\varepsilon\phi}^0$

Since $x_\alpha g_\alpha y_\alpha \in A_0G_0B_0$ eventually, the set $A_0G_0B_0$ is open in S .

□ 1068

Corollary 1.5. Let $N \equiv \{x \in S | x^2 = 0\}$. Then $G_0B_0 = e_0S \setminus N$ is a right group, topologically isomorphic to $G_0 \times B_0$.

1069

1070

Proof. The set B_0 consists of idempotents of the group cells of $R_0 \equiv e_0S$. Hence, $b \in B_0$ if and only if $e_0Sb = G_0b \cup \{0\}$ is not a null cell. Therefore, $G_0B_0 = e_0S \setminus N$. B_0 and G_0 are locally compact spaces, a right zero semigroup (a consequence of $e_0be_0 = e_0$) and a topological group, respectively. $e_0S \setminus N$ is right simple, hence a right group. The result then follows from Proposition 1.18.

□ 1075

Example 1.3. Consider again the semigroup S of 2×2 matrices of rank ≤ 1 (see

1076

Example 1.2). Let e_0 be the idempotent $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$. Then

1077

$$G = e_0Se_0 = \left\{ \begin{pmatrix} a & 0 \\ 0 & 0 \end{pmatrix} \mid a \in \mathfrak{R} \right\}$$

is a group with 0. The left ideal Se_0 is $\left\{ \begin{pmatrix} a & 0 \\ b & 0 \end{pmatrix} \mid a, b \in \mathfrak{R} \right\}$. The idempotents of Se_0

1078

are of the form $\begin{pmatrix} 1 & 0 \\ b & 0 \end{pmatrix}$, $b \in \mathfrak{R}$. The group cell with identity $\begin{pmatrix} 1 & 0 \\ b & 0 \end{pmatrix}$ is $\left\{ \begin{pmatrix} a & 0 \\ ab & 0 \end{pmatrix} \mid a \in \mathfrak{R} \right\}$.

1080

There is only one null cell in Se_0 ; it is $\left\{ \begin{pmatrix} 0 & 0 \\ a & 0 \end{pmatrix} \mid a \in \mathfrak{R} \right\}$.

1081

The right ideal e_0S is $\left\{ \begin{pmatrix} a & c \\ 0 & 0 \end{pmatrix} \mid a, b \in \mathfrak{R} \right\}$, and the idempotents of e_0S are $\begin{pmatrix} 1 & c \\ 0 & 0 \end{pmatrix}$, $c \in \mathfrak{R}$.

1082

1083

The cell with the left representative $\begin{pmatrix} 1 & 0 \\ b & 0 \end{pmatrix}$, $b \neq 0$, and right representative $\begin{pmatrix} 1 & c \\ 0 & 0 \end{pmatrix}$, $c \neq 0$ (somewhere in the middle of the egg box) is a group cell if and only if

$$\begin{pmatrix} 1 & c \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ b & 0 \end{pmatrix} = \begin{pmatrix} 1+bc & 0 \\ 0 & 0 \end{pmatrix} \neq \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

The group cells are isomorphic to the multiplicative group $\mathbb{R}^*$ of nonzero reals.

Exercise 1.27. Show that a compact simple semigroup is completely simple.

Exercise 1.28. Let S be a compact cancellative semigroup. Prove that S is a group.

Exercise 1.29. Let S be a simple compact semigroup, and E the set of idempotents in S . Show that if $e, f \in E$, then eSf is a group.

Exercise 1.30. Let S be a compact semigroup and A a left simple subsemigroup. Show that $\overline{A}$ is also a left simple subsemigroup.

Exercise 1.31. Prove that a compact semitopological group is a topological group.

Exercise 1.32. Let S be a compact semigroup and let x be any element of S . Prove that $\{x, x^2, \dots\}$ contains a closed abelian group G , such that

$$G = \bigcap_{k=1}^{\infty} \overline{\{x^k, x^{k+1}, \dots\}}.$$

Exercise 1.33. Let $a, b \in S$, $a \neq b$, where S is a topological semigroup, and let $A = \{a, a^2, \dots\}$ and $B = \{b, b^2, \dots\}$. Prove that

- (i) If $S = A$ is compact and has an identity, then S is a group.
- (ii) If $S = A$ is compact and if $\{a\}$ is not isolated, then S is a topological group.
- (iii) If $S = A = B$ is compact, then it is a compact group.

Exercise 1.34. Let S be a multiplicative semigroup of real numbers, x , $x \in (0, 1)$ (with usual topology). Show that the kernel of S is empty.

Exercise 1.35. Let $S = \{(x, y) | 0 \leq x \leq 1, 0 \leq y \leq 1\}$ and define multiplication in S by $(x_1, y_1)(x_2, y_2) = (0, y_2)$. Find the kernel of S .

Exercise 1.36. An element a of a topological group G is *compact* if $\overline{\{a, a^2, \dots\}}$ is compact. Consider the general linear group $GL(2, \mathbb{C})$ (the set of nonsingular complex 2×2 matrices). Let $z_n = e^{\frac{2\pi i}{n}}$ for $n = 2, 3, \dots$. Show that $\begin{pmatrix} z_n & 1 \\ 0 & 1 \end{pmatrix}$ generates a finite subgroup of $GL(2, \mathbb{C})$. Thus, it is a compact element. Also show that the set of compact elements of $GL(2, \mathbb{C})$ is not closed.

Exercise 1.37. Prove that if S is a compact subsemigroup of a topological group G , then S is a subgroup. 1111
1112

Exercise 1.38. Prove that every closed proper subgroup of the circle group $\{x \in \mathbf{C} \mid |x| = 1\}$ is finite and cyclic. 1113
1114

Exercise 1.39. A compact topological semigroup satisfies the following compactness condition: For any two compact subsets A and B , the sets $AB^{-1} \equiv \{y \in S \mid yx \in A \text{ for some } x \in B\}$ and $A^{-1}B \equiv \{y \in S \mid xy \in B \text{ for some } x \in A\}$ are both compact. 1115
1116
1117
1118

Prove that a completely simple topological semigroup S with a compact group factor in its usual product decomposition satisfies the above compactness condition iff S is compact. 1119
1120
1121

1.6 Semigroups of Matrices

1122

In Sect. 1.6, we study multiplicative semigroups of finite dimensional matrices in some detail. We begin with the structure theory for semigroups of $d \times d$ real matrices, viewed as linear transformations on $\mathfrak{R}^d$. We do not treat the complex case, but most of our results carry over, as the reader will see without difficulty. We give as complete a picture as possible of completely simple (and completely 0-simple) semigroups of matrices. Needless to say, there may be many different descriptions, each useful for dealing with particular problems. The second part of Sect. 1.6 is devoted to the theory of nonnegative matrices. In this special case, more concrete results are obtained. The term nonnegative matrix, vector, etc., always means that all the entries of the matrix (vector) are ≥ 0 . 1123
1124
1125
1126
1127
1128
1129
1130
1131
1132

Let M_d be the set of $d \times d$ real matrices. Under ordinary matrix multiplication, this set is a semigroup. If we use the euclidean distance on $\mathfrak{R}^{d^2}$ to define a norm on M_d (this is called the *Frobenius norm* in matrix theory), then M_d becomes a locally compact second-countable space with jointly continuous multiplication. The usual operator (L_p) norms on M_d (where elements of M_d are considered as linear operators on $\mathfrak{R}^d$) are equivalent to the Frobenius norm. Hence, they all define the same locally compact topology. 1133
1134
1135
1136
1137
1138
1139

When the coordinate system is fixed, we interpret $v \in \mathfrak{R}^d$ as a d -dimensional column vector. It is transformed by an element $x \in M_d$ to $xv \in \mathfrak{R}^d$ by matrix multiplication. 1140
1141
1142

The *range* $R(x)$ is the subspace $x\mathfrak{R}^d \equiv \{xv \mid v \in \mathfrak{R}^d\}$. Its dimension is called the *rank* of x , $\text{rank}(x) = \dim(R(x))$. The *null-space* $N(x)$ is the kernel of the linear homomorphism x ; in other words $N(x) \equiv \{v \in \mathfrak{R}^d \mid xv = 0\}$. For any $x \in M_d$: $d = \text{rank}(x) + \dim(N(x))$. 1143
1144
1145
1146

The rank is a decreasing function in the following sense: $\text{rank}(xy)$, and $\text{rank}(yx) \leq \text{rank}(x)$, for any $x, y \in M_d$. Elementary considerations also show that $\text{rank}(xy) = \text{rank}(x)$ if and only if $R(x) \cap N(y) = \{0\}$. $yxv = 0$ if and only if either $xv \neq 0$ and $xv \in N(y)$ or $v \in N(x)$. We have thus proved Proposition 1.20. 1147
1148
1149
1150

- Proposition 1.20.** (i) *The sets $T_k \equiv \{x \in M_d \mid \text{rank}(x) \leq k\}$ ($k = 0, 1, \dots, d$) are two-sided ideals of M_d .* 1151
1152
(ii) *$\text{rank}(yx) = \text{rank}(x)$ if and only if $N(y) \cap R(x) = \{0\}$.* 1153

The idempotents of M_d are by definition projection matrices $e^2 = e$. Geometrically, e is a projection onto $R(e)$ along $N(e)$. Of course, $R(e) \cap N(e) = \{0\}$ by Proposition 1.20(ii). The idempotent e is an orthogonal projection if and only if it is symmetric (see, for example [18]). Conversely, any projection operator on $\Re^d$ can be described by an idempotent matrix. 1154
1155
1156
1157
1158

Note that for nonzero e , $\|e\| \geq 1$ for any operator norm $\|\cdot\|$ on M_d because $\|e^2\| \leq \|e\|^2$. 1159
1160

- Proposition 1.21.** (i) *x and y generate the same left (right) ideal if and only if they have the same null-space (range),* 1161
1162
(ii) *Any idempotent e is the identity element of the group of transformations with the same range $R(e)$ and the same null-space $N(e)$ as e . This group is isomorphic to the general linear group $GL(k)$ of $k \times k$ nonsingular matrices, where $k = \text{rank}(e)$.* 1163
1164
1165
1166

For a proof, we refer to [43], p. 57. 1167

Theorem 1.10. *The quotient semigroup T_k/T_{k-1} is completely 0-simple ($k = 1, 2, \dots, d$).* 1168
1169

Proof. Recall that $S = T_k/T_{k-1}$ is the set of matrices of rank $\leq k$ such that all elements of rank $\leq k-1$ are identified with 0. There are idempotents in $T_k \setminus T_{k-1}$ (the projections onto k -dimensional subspaces). They are easily seen to be primitive in T_k/T_{k-1} , since lower rank idempotents are considered as 0. We have to show that $SaS = S$ for all nonzero $a \in S$, i.e., for all matrices a, b of rank k , the equation $xay = b$ is solvable in T_k . 1170
1171
1172
1173
1174
1175

First take a k -dimensional subset R with $N(a) \cap R = \{0\}$ and an element g of rank k mapping $R(a)$ onto R (and its orthogonal complement to 0, say). Then $N(ga) = N(a)$ and $R(ga) = R$. By Proposition 1.21, there is an element $h \in T_k$ such that $hga = gah = e$, where e is the idempotent with range R and null-space $N(a)$. $e \in SaS$. 1176
1177
1178
1179
1180

We next show that $b \in SeS$, which proves $b \in SaS$. Let u map $R(b)$ onto R and the orthogonal complement $R(b)^\perp$ to 0. Consider ub . $R(ub) = R$ and $N(ub) = N(b)$. Let s restricted to R be the inverse of u restricted to $R(b)$ and let s map $R^\perp$ to 0. Then $sub = seub = b$. $\square$ 1181
1182
1183
1184

Corollary 1.6. *Let e_0 be an idempotent of rank k . Then there is a left (right) zero locally compact topological semigroup A_0 (B_0) and a topological group G_0 such that the product $A_0 \times G_0 \times B_0$ is homeomorphic with an open set containing e_0 . The correspondence is a homomorphism on pairs that are not zero divisors. The topological group G_0 is isomorphic to $GL(k)$, A_0 consists of all idempotents with null space $N(e_0)$ and B_0 of all idempotents with range $R(e_0)$.* 1185
1186
1187
1188
1189
1190

Proof. M_d is locally compact, so A_0, B_0, G_0 (defined in Theorem 1.9) are locally compact spaces. $S = T_k/T_{k-1}$ is a locally compact space with jointly continuous multiplication whenever 0 is not involved.

The correspondence $\Psi : A_0 \times G_0 \times B_0 \mapsto A_0 G_0 B_0$, defined by $\Psi(a, g, b) = agb$, is a homeomorphism. It is algebraically a partial homomorphism on $A_0 \times G_0 \times B_0$ (it preserves products as long as 0 is not involved). $\square$

Remark 1.13. In the Rees–Suschkewitsch representation, cells aG_0b are group cells or null cells according to whether the product ba of the idempotents is of rank k or strictly less than k , respectively. This in turn is equivalent to $N(b) \cap R(a) = \{0\}$ or $\neq \{0\}$, respectively. (Notice that ab is always in G_0 , thus of rank k .)

Remark 1.14. (i) Let S be a subsemigroup of M_d . Suppose S is simple. As we remarked before, it is easy to see that all idempotents in S are of minimal rank, hence primitive. So in applications, the main problem that we encounter in this respect is to prove that the semigroup is simple (or 0-simple as the case may be) and find one idempotent $\neq 0$ in S , (see Remark 1.4).

(ii) In the completely simple situation, we note the following: If the idempotents a and b have idempotent products (ab and ba are idempotent), then the Rees product $X \times G \times Y$ is *direct*, i.e., the sandwich function maps $Y \times X$ to the identity of the group G . Naturally, such a situation facilitates the analysis considerably, since the components evolve independently from each other.

Example 1.4. An important semigroup S studied by Bougerol in [16] consists of matrices having the following block matrix form

$$\begin{pmatrix} a & c & e \\ 0 & k & f \\ 0 & 0 & b \end{pmatrix},$$

where the diagonal block have dimensions $d_1 \times d_1, d_2 \times d_2, d_3 \times d_3$, $d_1 + d_2 + d_3 = d$. The central block k is an orthogonal matrix; the other blocks are arbitrary (and may be 0).

Clearly the matrices of the form

$$\begin{pmatrix} 0 & c & e \\ 0 & k & f \\ 0 & 0 & 0 \end{pmatrix}$$

form an ideal of S . The idempotents of lowest rank are

$$\begin{pmatrix} 0 & c & cf \\ 0 & I & f \\ 0 & 0 & 0 \end{pmatrix}$$

with I the $d_2 \times d_2$ identity matrix and c, f arbitrary (subject to the dimension requirements of course). 1218
1219

$$\begin{pmatrix} 0 & c & cf \\ 0 & I & f \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & c' & c'f' \\ 0 & I & f' \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & c & cf' \\ 0 & I & f' \\ 0 & 0 & 0 \end{pmatrix}$$

so the product is direct. One can also verify that indeed $R \cap N' = \{0\}$ for the range 1220
1221 R and null space N' of any two idempotents of lowest rank ($= d_2$).

Example 1.5. Let us return to the 2×2 matrices of rank 1. In Example 1.2, we saw 1222
1223 that

$$\begin{pmatrix} (1+cb)^{-1} & (1+cb)^{-1}c \\ b(1+cb)^{-1} & b(1+cb)^{-1}c \end{pmatrix}$$

(where $(1+cb) \neq 0$) is idempotent. But the product 1224

$$\begin{pmatrix} (1+cb)^{-1} & (1+cb)^{-1}c \\ b(1+cb)^{-1} & b(1+cb)^{-1}c \end{pmatrix} \begin{pmatrix} (1+c'b')^{-1} & (1+c'b')^{-1}c' \\ b'(1+c'b')^{-1} & b'(1+c'b')^{-1}c' \end{pmatrix}$$

is usually not an idempotent unless $b = b'$ or $c = c'$. We note that an idempotent 1225
1226 matrix of rank 1 can also be characterized as follows: rows are proportional and the trace is 1. If rows are proportional and the trace is 1, then the matrix is of the form 1227
1228 xy^T , where x and y are two column vectors with scalar product $x^T y = y^T x = 1$.
1229 (y^T denotes the transpose of y .)
1230

Let

$$e = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}, \quad f = \begin{pmatrix} 1 & -2 \\ 0 & 0 \end{pmatrix}.$$

Then 1231

$$ef = \begin{pmatrix} 1 & -2 \\ 1 & -2 \end{pmatrix}, \quad fe = \begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix},$$

$$efe = \begin{pmatrix} -1 & 0 \\ -1 & 0 \end{pmatrix} = -e, \quad fef = \begin{pmatrix} -1 & 2 \\ 0 & 0 \end{pmatrix} = -f,$$

e and f generate a finite semigroup of rank 1 matrices 1232

$$\{e, f, ef, fe, -e, -f, -ef, -fe\}.$$

Example 1.6. The following matrix semigroups are abelian:

1233

$$(i) \ S = \left\{ \begin{pmatrix} 1 & a & b \\ 0 & 1 & a \\ 0 & 0 & 1 \end{pmatrix} \mid a \geq 0, b \in \mathfrak{R} \right\},$$

1234

$$(ii) \ S = \left\{ \begin{pmatrix} a & b \\ b & a \end{pmatrix} \mid a, b \in \mathfrak{R} \right\},$$

1235

$$(iii) \ S = \left\{ \begin{pmatrix} a & b & c \\ c & a & b \\ b & c & a \end{pmatrix} \mid a, b, c \in \mathfrak{R} \right\}.$$

1236

(i) does not admit a completely simple minimal ideal: S is not even simple, since the entry a at (1,2) and (2,3) is increasing; if $a > 0$, the corresponding element generates an ideal properly contained in S . The subsemigroup with $a = 0$ is a group isomorphic with $(\mathfrak{R}, +)$.

1237

1238

1239

1240

The minimal ideal in (ii) is $\{0\}$. S has divisors of zero, for example,

1241

$$\begin{pmatrix} a & -a \\ -a & a \end{pmatrix} \begin{pmatrix} a & a \\ a & a \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Idempotents of rank 1 are

1242

$$\begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}$$

and

1243

$$\begin{pmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{pmatrix}.$$

Subsemigroups of matrices of rank ≤ 1 is not 0-simple, since

1244

$$\begin{pmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{pmatrix} \notin S \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}.$$

(iii) S has of course a zero. The rank 1 matrices are

1245

$$a \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{pmatrix}, \ a \neq 0$$

so they form a group isomorphic with $(\mathfrak{R}^*, \cdot)$. (Complex matrices would have more idempotents of rank 1, i.e., let the three cube roots of 1 be a, b, c .)

1246

1247

Idempotents are projections onto some subspace of $\mathfrak{R}^d$ along another, complementary subspace; we give another characterization later. To see the origin of the ideas, let us first look at the corresponding result for nonnegative matrices.

1248

1249

1250

Theorem 1.11. Let P be a $d \times d$ nonnegative idempotent matrix of rank $k > 0$. Then there is a unique partition of $\{1, 2, 3, \dots, d\}$, called the basis of P , into classes $\{T, C_1, C_2, \dots, C_k\}$ such that the following hold:

- (i) $T = T_r \cup T_c$, where $T_r \equiv \{i \mid \text{the } i\text{th row of } P \text{ is a zero row}\}$ and $T_c \equiv \{i \mid \text{the } i\text{th column of } P \text{ is a zero column}\}$;
- (ii) $P_{ij} = 0$ whenever $i \in C_s$ and $j \in C_t$, ($t \neq s$);
- (iii) In any $C_s \times C_s$ -block ($1 \leq s \leq k$) of P , the rows are all proportional, each entry is strictly positive, and the sum of the diagonal entries is 1;
- (iv) If $i \in T_c$ then for $j, h \in C_r$ ($1 \leq r \leq k$)

$$\frac{P_{ij}}{P_{jj}} = \frac{P_{ih}}{P_{jh}};$$

- (v) If $j \in T_r$ then for $i, h \in C_r$ ($1 \leq r \leq k$)

$$\frac{P_{ih}}{P_{ii}} = \frac{P_{hj}}{P_{hi}}.$$

Conversely any such matrix is idempotent of rank k provided the $T_c \times T_r$ block satisfies the idempotency condition, cf. Exercise 1.43, and all other blocks are zero.

The proof will be given (for the infinite dimensional nonnegative case) in Sect. 1.7. Real idempotent matrices have a similar but weaker and nonunique structure.

Theorem 1.12. Let P be a $d \times d$ idempotent matrix of rank k . Then there is a partition $\{T, B, C_1, C_2, \dots, C_k\}$ of $D \equiv \{1, 2, \dots, d\}$ such that

- (i) $T = T_r \cup T_c$, where $T_r \equiv \{i \mid \text{the } i\text{th row of } P \text{ is a zero row}\}$ and $T_c \equiv \{i \mid \text{the } i\text{th column of } P \text{ is a zero column}\}$;
- (ii) The rank of the restriction of P to $C_s \times D$ is 1, $1 \leq s \leq k$;
- (iii) If $u_s \in C_s$ for $1 \leq s \leq k$ and $P_{hj} = \gamma(h, u_s)P_{u_s j}$ for $h \in C_s$ then for $i \in B$ there exist real numbers $a_t(i)$, $1 \leq t \leq k$, such that the following conditions hold

$$P_{ij} = \sum_{t=1}^k a_t(i)P_{u_t j}, \quad (1.1)$$

$$\sum_{i \in C_s} \gamma(i, u_s)P_{u_s i} + \sum_{i \in B} a_s(i)P_{u_s i} = 1, \quad (1.2)$$

$$\sum_{i \in C_t} \gamma(i, u_t)P_{u_s i} + \sum_{i \in B} a_t(i)P_{u_s i} = 0, \text{ for } s \neq t, \quad (1.3)$$

$$a_t(h) = \sum_{i \in C_t} \gamma(i, u_t)P_{hi} + \sum_{i \in B} a_t(i)P_{hi}, h \in B. \quad (1.4)$$

- (iv) For $i \in T_c$ and $1 \leq j \leq d$,

$$P_{ij} = \sum_{t=1}^k \left[\sum_{h \in C_t} P_{ih} \gamma(h, u_t) + \sum_{h \in B} a_t(h) P_{ih} \right] P_{u_t j}.$$

Conversely any real matrix P with a partition $\{T, B, C_1, C_2, \dots, C_k\}$ such that (i), (ii), (iii), and (iv) hold is idempotent of rank k . 1274 1275

Proof. We observe first that P restricted to the complement of T is idempotent. 1276 Also the rank of $P|_{T^c}$ is k . We assume then with no loss of generality that T is 1277 empty. 1278

Let $R_{m_1}, \dots, R_{m_k}$ be k linearly independent rows of P . Let $C_i \equiv \{j \mid \text{the } j\text{th row of } P \text{ is proportional to the row } R_{m_i}\}$. Then the C_i s are disjoint and we denote 1279 1280 by B the remaining part of D , $B \equiv (\bigcup_{i=1}^k C_i)^c$. 1281

Let u_t be some fixed but arbitrary element of C_t . Then 1282

$$P_{hj} = \gamma(h, u_t) P_{u_t j}, \quad h \in C_t, \quad 1 \leq j \leq d$$

for some real numbers $\gamma(h, u_t) \neq 0$. Clearly for $i \in B$, we have (1.1): 1283

AQ4

$$P_{ij} = \sum_{t=1}^k a_t(i) P_{u_t j}, \quad (1.1)$$

because the remaining rows are linear combinations of the rows $u_1, u_2, \dots, u_k$. 1284

Let us write down the idempotency requirement 1285

$$P_{ij} = \sum_{t=1}^k \left(\sum_{h \in C_t} P_{ih} P_{hj} + \sum_{h \in B} P_{ih} P_{hj} \right).$$

Then 1286

$$P_{usj} = \sum_{t=1}^k P_{u_t j} \left[\sum_{h \in C_t} \gamma(h, u_t) P_{u_s h} \right] + \sum_{h \in B} P_{u_s h} P_{hj}, \quad 1 \leq j \leq d,$$

where the last term is 1287

$$\sum_{h \in B} \sum_{t=1}^k P_{u_s h} a_t(h) P_{u_t j} = \sum_{t=1}^k P_{u_t j} \left[\sum_{h \in B} a_t(h) P_{u_s h} \right].$$

By the linear independence of the “representative rows” $u_1, u_2, \dots, u_k$ we have (1.2) 1288 and (1.3) 1289

$$1 = \sum_{h \in C_s} \gamma(h, u_s) P_{u_s h} + \sum_{h \in B} a_s(h) P_{u_s h} \quad (1.2)$$

and

$$0 = \sum_{h \in C_t} \gamma(h, u_t) P_{u_s h} + \sum_{h \in B} a_t(h) P_{u_s h}, \text{ for } s \neq t. \quad (1.3)$$

Also if $i \in B$, then (1.4) can be written as 1291

$$a_t(h) = \sum_{i \in C_t} \gamma(i, u_t) P_{hi} + \sum_{i \in B} a_t(i) P_{hi}, h \in B. \quad (1.4)$$

Conversely, conditions (1.1)–(1.4) imply that P is idempotent. 1292

Now if T is not empty, we must make sure that the corresponding rows are linear combinations of rows $u_1, \dots, u_k$. Furthermore, the idempotency requirement has to be satisfied for these rows. This is the meaning of (iv). 1293
1294
1295

Remark 1.15. The trace of P is equal to its rank k . This can be seen from (1.2) and (1.3). If B is empty we infer immediately from (1.2) that the $C_s \times C_s$ block is not identically 0; hence, it has rank 1. Also, the partition into classes is then *unique*; we call it the *basis* for P . 1296
1297
1298
1299

If B is not empty, then the diagonal $C_s \times C_s$ blocks may be 0, since the rectangular matrix indexed by $C_s \times D$ being by construction nonzero. Take, for example 1300
1301

$$P = \begin{pmatrix} 0 & -1 & 1 \\ a & 1+a & -a \\ a & a & 1-a \end{pmatrix}, a \neq 0$$

where $C_1 = \{1\}$, $C_2 = \{2\}$, $B = \{3\}$ and $P_{C_1 \times C_1} = 0$. For $a = 1$, we cannot even correct the situation by choosing C_1 and B differently. 1302
1303

Another way to see that the trace of P is equal to its rank is that the trace is invariant under change of coordinates. Hence, the trace of a k -dimensional projection equals that of a $k \times k$ identity matrix. 1304
1305
1306

Remark 1.16. (i) Let 1307

$$P = \begin{pmatrix} 1/2 & 1+d & e-1 & a & b+1 \\ 1/2 & 1 & -1 & a & b \\ 1/4 & 1/2 & -1/2 & a/2 & b/2 \\ 0 & d & e & 0 & 1 \\ 0 & d & e & 0 & 1 \end{pmatrix},$$

where $a + b = -1/2$, $d + e/2 = 0$. The partition of P into classes is not unique. We can take $B = \{1\}$, $C_1 = \{2, 3\}$, $C_2 = \{4, 5\}$. We can, however, also have $C_1 = \{1\}$, $C_2 = \{4, 5\}$, $B = \{2, 3\}$. 1308
1309
1310

(ii) Let

1311

$$P = \begin{pmatrix} 0 & b & b & 3b/4 \\ 0 & 1 & 1 & 3/4 \\ 1 & -b & 1/3 & -1/2 \\ -4/3 & 4b/3 & -4/9 & 2/3 \end{pmatrix}, b \neq 0.$$

P is idempotent of rank 2. $C_1 = \{1, 2\}$, $C_2 = \{3, 4\}$. B is empty. Then the diagonal elements of the $C_1 \times C_1$ and $C_2 \times C_2$ blocks sum to 1. If $b = 0$, then P is also idempotent (by continuity for instance). Then $T = \{1\}$, $C_1 = \{2\}$, $C_2 = \{3, 4\}$.

(iii) For full-rank P , $C_s = \{s\}$ and $\gamma(s, s) = 1$ ($s = 1, 2 \dots d$). B is empty. Equation (1.2) then says that the diagonal elements are all 1, whereas the off-diagonal elements are all 0 by (1.3).

The idempotent P is the identity element of a group of matrices, with the same range and null space as P . It is natural to ask whether the basis structure carries over in some way to the group elements. The answer is yes.

With the same notation as in Theorem 1.12, we have Theorem 1.13.

Theorem 1.13. Let G be a group of $d \times d$ matrices of rank k with identity P . Let $\{T, B, C_1, \dots, C_k\}$ be the partition for P . Then for any $M \in G$:

- (i) The rank of M is k and the $C_s \times D$ block of M has rank 1;
- (ii) M is completely determined by its values on its u_s -rows, $1 \leq s \leq k$, and the group identity P . Also there is a $k \times k$ matrix m of full rank (k) such that

$$M_{u_s j} = \sum_{t=1}^k m_{ts} P_{u_t j}, \quad 1 \leq j \leq d, \quad 1 \leq s \leq k,$$

and

$$m_{ts} = \sum_{j \in C_t} \gamma(j, u_t) M_{u_s j} + \sum_{j \in B} a_t(j) M_{u_s j}.$$

(iii) The map $M \leftrightarrow m$ defines an isomorphism between G and a subgroup of the general linear group $GL(k)$.

Proof (Sketch). The details can be found in [152]:

Since $MM^{-1} = M^{-1}M = P$, the zero rows and the zero columns of M are the same as those of P .

Next consider, for example, the equality

$$P_{ij} = \sum_{s=1}^k a_s(i) P_{u_s j}, \quad i \in B, \quad 1 \leq j \leq d. \quad (1.1)$$

Postmultiplication by M shows that the same equality remains true when P is replaced by M . Similarly, computations of the proof of Theorem 1.12 can be worked out for M instead of P . The matrix m is defined by

$$m_{st} = \sum_{j \in C_s} \gamma(j, u_s) M_{ujj} + \sum_{j \in B} a_s(j) M_{ujj} \quad (1.5)$$

with inverse

$$m'_{st} = \sum_{j \in C_s} \gamma(j, u_s) (M^{-1})_{ujj} + \sum_{j \in B} a_s(j) (M^{-1})_{ujj}. \quad (1.6)$$

The isomorphism properties follow by direct calculations. $\square$

Corollary 1.7. *If G is a locally compact subgroup of M_d , then G is topologically isomorphic to a locally compact topological subgroup of $GL(k)$.*

Proof. Corollary 1.7 is an immediate consequence of Corollary 1.6 but it can also be seen directly by noting that m and m' are continuous functions of M . This is because the partition does not vary with $M \in G$. Also, G is a topological group by Ellis' theorem, so the inversion is continuous. $\square$

Example 1.7. If S is a completely simple semigroup of matrices, the partition very definitely is not invariant. Consider the 3×3 matrices

$$e = \begin{pmatrix} 0 & b & 0 \\ 0 & 1 & 0 \\ -2 & 2b & 1 \end{pmatrix}, \quad f = \begin{pmatrix} 1 & 1/2 & -1/2 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

If $b \neq 0$, $S = \{e, f\}$ is a completely simple semigroup, in fact a right zero semigroup of matrices of rank 2. For e , $C_1 = \{1, 2\}$, $C_2 = \{3\}$ whereas f has the partition $C_1 = \{1\}$, $C_2 = \{2, 3\}$.

The structure theory for groups and completely simple semigroups of real matrices is complicated and generally speaking difficult to use. It is also hampered by the nonuniqueness of the corresponding partitions.

In Theorem 1.11, we saw that the structure of a nonnegative idempotent matrix is much simpler, as is to be expected. Moreover, it turns out that the same basis can be used in for completely simple semigroups.

Let us first study a group G of nonnegative matrices. We also assume that G is locally compact. Let P be the identity of G . Assume that P has rank k . Then by Theorem 1.11, there is a unique basis $\{T, C_1, C_2, \dots, C_k\}$ for P . Let $M \in G$. Then we saw in the proof of Theorem 1.13 that M has precisely the same zero rows and zero columns as P . Of course, M also has rank k . As in the proof of Theorem 1.12, for $1 \leq t \leq k$, we have

$$P_{hj} = \gamma(h, u_t) P_{ujj}, \quad h \in C_t, \quad 1 \leq j \leq d \quad (1.7)$$

with $\gamma(h, u_t) > 0$. Since $M = PM$,

1363

$$M_{hj} = \gamma(h, u_t)M_{u_t j}, \quad h \in C_t, \quad 1 \leq j \leq d. \quad (1.8)$$

Hence, the rank of M restricted to $C_t \times D$ is 1.

1364

Suppose that $M_{ij} > 0$. Let $i \in C_s$. We claim that there is precisely one C_t with $M_{ij} > 0$ for $i \in C_s, j \in C_t$. Furthermore, this correspondence is one-to-one: $M_{ij} > 0$ and $M_{i'j} > 0$ for j in some C_t implies that i and i' belong to the same C_r ($1 \leq r \leq k$).

1365

1366

1367

1368

First of all, let us point out that since all elements of G have the same zero rows and zero columns,

1369

1370

$$\sum_{h \notin T} P_{ih} N_{hj} = N_{ij} = \sum_{h \notin T} N_{ih} P_{hj}, \quad i, j = 1, 2, \dots, d \quad (1.9)$$

and

1371

$$(NM)_{ij} = \sum_{h \notin T} N_{ih} M_{hj}, \quad i, j = 1, 2, \dots, d, \quad (1.10)$$

where M and N are arbitrary matrices in G . Let us now prove our claim. For P the correspondence $s \leftrightarrow t$ is simply the identity mapping by Theorem 1.11: $P_{ij} > 0$, $i \in C_s, j \in C_t$ if and only if $s = t$.

1372

1373

1374

In particular, $M_{ij} > 0$ for all $i \in C_s$ means that the i th row is not a zero row and that j can be chosen from the complement of T ; i.e., $j \in C_t$ for some t . Similarly, $(M^{-1})_{il} > 0, l \in C_r$ for some r

1375

1376

1377

$$0 < P_{ii} = \sum_{h \notin T} M_{ih} (M^{-1})_{hi}.$$

Hence, $(M^{-1})_{hi} > 0$ for some $h \notin T$. Suppose $(M^{-1})_{hi} > 0$ for an $h \notin C_t$. Then

1378

$$0 < (M^{-1})_{hi} M_{ij} \leq P_{hj}$$

with h and j in different C classes, contradicting Theorem 1.11.

1379

$M_{il} > 0$ for $l \notin C_t$ is impossible too:

1380

$$(M^{-1})_{hi} M_{il} \leq P_{hl} = 0.$$

Equation (1.8) shows that $M_{ij} > 0$ for one $i \in C_s$ implies $M_{i'j} > 0$ for all $i' \in C_s$. On the other hand, if $M_{ij'} = 0$ for some $j' \in C_t$ then $M_{i'j'} = 0$ for all $i' \in C_s$ and hence

1381

1382

1383

$$0 = \sum_{h \in C_s} (M^{-1})_{jh} M_{hj'} = \sum_{h \notin T} (M^{-1})_{jh} M_{hj'} = P_{jj'},$$

a contradiction, since j and j' both belong to C_t . Hence, all entries in $M|_{C_s \times C_t}$ are positive. This proves the claim.

1384

1385

Another way of phrasing the claim is that there is a permutation π on D such that 1386

$$M_{ij} > 0, i \in C_s, j \in C_t \iff t = \pi(s). \quad 1387$$

Furthermore, the preceding proof shows that 1388

$$(M^{-1})_{ij} > 0, i \in C_s, j \in C_t \iff t = \pi^{-1}(s),$$

where π^{-1} is the inverse permutation. 1389

Let us write 1390

$$\pi_M(s) = t \iff M|_{C_s \times C_t} > 0 \quad (1.11)$$

(in the sense that all the entries in that block are strictly positive). π is an antihomomorphism of G into the group of permutations of $D \equiv \{1, 2, \dots, d\}$ 1391

$$\pi_M \circ \pi_N = \pi_{NM}. \quad (1.12)$$

This follows at once from the observation that the expression 1393

$$(NM)_{ij} = \sum_{h \in T} N_{ih} M_{hj}, i \in C_s, j \in C_t$$

is positive if and only if $h \in C_{\pi_N(s)}$ and $j \in C_{\pi_M(\pi_N(s))}$. 1394

In general, π is far from injective, since the group of permutations is a finite group. The matrices M and N have the same image under π if and only if 1395

$$N|_{C_s \times C_{\pi(s)}} = \mu_s M|_{C_s \times C_{\pi(s)}}, 1 \leq s \leq k \quad (1.13)$$

for some strictly positive numbers $\mu_1, \mu_2, \dots, \mu_k$. (Here, $\pi_M = \pi_N$ is denoted by just π .) Clearly, the condition (1.13) is sufficient, since, by (1.9), the values of both matrices are uniquely determined by their values on $T^c \times T^c$ (and by P). 1398

To prove the converse, let $i \in C_s, j \in C_t, t = \pi(s)$. Equation (1.7) says 1400

$$P_{hj} = \gamma(h, u_t) P_{u_t j} \text{ for } h \in C_t.$$

Premultiplication by M gives 1401

$$M_{ij} = \sum_{h \in C_t} M_{ih} \gamma(h, u_t) P_{u_t j} = P_{u_t j} \sum_{h \in C_t} M_{ih} \gamma(h, u_t)$$

$$N_{ij} = P_{u_t j} \sum_{h \in C_t} N_{ih} \gamma(h, u_t)$$

Equation (1.8) shows that $M_{ih} = \gamma(i, u_s) M_{u_s h}, N_{ih} = \gamma(i, u_s) N_{u_s h}$, so 1403

$$N_{ih} = M_{ih} \frac{N_{u_s h}}{M_{u_s h}} \equiv \mu_s(h) M_{ih}$$

independently of $i \in C_s$. Hence,

1404

$$\frac{N_{ij}}{M_{ij}} = \frac{\sum_h \mu_s(h) M_{ih} \gamma(h, u_t)}{\sum_h M_{ih} \gamma(h, u_t)},$$

which is independent of $j \in C_t$.

1405

Let us now collect the results of the preceding discussion of nonnegative matrices in Theorem 1.14.

1406

1407

Theorem 1.14. *Let G be a group of nonnegative matrices with identity P of rank k . Let the basis of P be $\{T, C_1, C_2, \dots, C_k\}$. Then*

1408

1409

- (i) *Each $M \in G$ has rank k , the same zero rows and zero columns as P ; the rank of $M|_{C_s \times D}$ is 1 for all s , $1 \leq s \leq k$;*
- (ii) *Each $M \in G$ is uniquely determined by P and $M|_{T^c \times T^c}$;*
- (iii) *There is an antihomomorphism π from G into a subgroup H of the group of permutations on $\{1, 2, \dots, k\}$ such that for $M \in G$ and $1 \leq s \leq k$*

1410

1411

1412

1413

1414

$$M|_{C_s \times C_t}$$

is a zero block if $t \neq \pi(s)$ and a strictly positive block of proportional rows if $t = \pi(s)$;

1415

1416

- (iv) *The constants of proportionality in the strictly positive blocks depend on s only ($1 \leq s \leq k$); they are the same for all group elements.*

1417

1418

Remark 1.17. For any given P , π_M , and the positive numbers $v_M(s) = M_{u_s \pi(s)}$, $1 \leq s \leq k$, determine M uniquely. v_M can be any strictly positive vector in $\mathbb{R}^k$ and π_M any permutation on $\{1, 2, \dots, k\}$.

1419

1420

1421

Proposition 1.22. *Let G be as in Theorem 1.14. The subset $G_1 \equiv \{M \in G \mid \pi_M(s) = s, 1 \leq s \leq k\}$ is a subgroup of G . If G is a locally compact group, then G_1 is a locally compact topological group. Moreover, G_1 is topologically isomorphic to a subgroup of the additive group $\mathbb{R}^k$.*

1422

1423

1424

1425

Proof. Clearly, the set G_1 is open and closed in G . Let

1426

$$\mu_M(s) = \frac{M_{u_s u_s}}{P_{u_s u_s}}, 1 \leq s \leq k.$$

For $M, N \in G_1$,

1427

$$(MN)_{u_s u_s} = \sum_{h \in C_s} M_{u_s h} N_{h u_s} = \sum_{h \in C_s} \mu_M(s) P_{u_s h} \mu_N(s) P_{h u_s}$$

so $\mu_{MN}(s) = \mu_M(s) \mu_N(s)$, also $\mu_P(s) = 1$. Thus, μ maps G_1 to the multiplicative group $(\mathbb{R}_+)^k$ of k -dimensional strictly positive vectors (under pointwise

1428

1429

multiplication). This correspondence is injective and continuous. The inverse is also continuous from $(\mathfrak{N}_+)^k$ to G_1 . $(\mathfrak{N}_+)^k$ is in turn topologically isomorphic to the additive group $\mathfrak{N}^k$. Hence, G_1 is topologically isomorphic to a subgroup of $\mathfrak{N}^k$. $\square$

As a corollary, we obtain the beautiful and perhaps surprising result in Corollary 1.8.

Corollary 1.8. *Any compact group of nonnegative matrices is finite. If the group contains a strictly positive matrix (i.e., one with all entries > 0), then it is a singleton.*

Proof. Let $M, N \in G$, a compact group of nonnegative matrices. Let P be of rank k . Then using notations from Theorem 1.14 and its proof, $\pi_{MN^{-1}}$ is the identity permutation. Hence, MN^{-1} generates a compact subgroup G_2 included in the subgroup G_1 of Proposition 1.22. The only compact subgroup of $\mathfrak{N}^k$ is $\{0\}$. By Proposition 1.22, G_2 is then the inverse image of $\{0\}$ under the correspondence $\mathfrak{N}^k \leftrightarrow G_1$; i.e., $G_2 = \{P\}$ [the vector $\mu_M(s) = 1$, $1 \leq s \leq k$, for all $M \in G_2$.] This shows that given a permutation π on $\{1, 2, \dots, k\}$, there is at most one element M of G with $\pi_M = \pi$. Hence, G can have at most finitely many $(k!)$ elements. If G contains a strictly positive matrix then $k = 1$, so G is a singleton. $\square$

The next question we address is whether the nice structure of a single group of nonnegative matrices is preserved when we look at many groups at once. To be more precise, we ask whether the completely simple semigroups of those matrices admit the same kind of structure.

Let S be a completely simple semigroup of nonnegative matrices. Assume all elements have the same rank k . There are idempotent elements in S , identities of the isomorphic groups that constitute S . In the uncomplicated case when no zero rows or zero columns are present, the structure is easily determined.

Proposition 1.23. *Suppose S is a completely simple semigroup of nonnegative matrices, none of which has any zero rows or zero columns. Then all idempotent elements of S have the same basis. In particular, all $C_s \times C_s$ blocks are strictly positive.*

Proof. Let P be an idempotent with basis $\{C_1, C_2, \dots, C_k\}$. Let $M \in S$ and let P' be the identity of the group MSM and call its basis $\{C'_1, C'_2, \dots, C'_k\}$. PMP belongs to the group PSP , since it has the same basis as P .

Let $M_{ij} > 0$. Then $(PMP)_{ij} \geq P_{ii}M_{ij}P_{jj} > 0$. This implies that the C' class containing j is contained in a C class. By varying j , we infer that each C' class is contained in some C class. But the bases have both k elements and both exhaust $\{1, 2, \dots, d\}$. Hence, they must be equal. $\square$

Remark 1.18. Proposition 1.23 does not imply that S is a group. We can construct many different idempotents by varying the entries in each $C_s \times C_s$ block ($1 \leq s \leq k$), subject to the requirements that the trace be 1, the rank 1 and the entries all positive.

Example 1.8. Let S be the completely simple semigroup $\{e, f\}$, where

1469

$$e = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 1/2 \\ 0 & 1/2 & 1/2 \end{pmatrix}, f = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

e and f have different bases. However, f is the limit of idempotents with the same basis as e . Thus, we can go slightly beyond the construction of Remark 1.18 and still obtain completely simple semigroups.

1470

1471

1472

To obtain an analog to Proposition 1.23 in the presence of zero rows and zero columns, we must introduce a condition on these. Recall that for a given idempotent P with basis $\{T, C_1, \dots, C_k\}$ the set $T = T_c \cup T_r$ consists of those indices which correspond to zero rows (T_r) and those which correspond to zero columns (T_c).

1473

1474

1475

1476

$$\forall u \in T_r, \exists i \in C_s, j \in C_t, s \neq t, P_{ui}, P_{uj} > 0 \quad (1.14)$$

$$\forall u \in T_c, \exists i \in C_s, j \in C_t, s \neq t, P_{iu}, P_{ju} > 0 \quad (1.15)$$

Note that (1.15) is not satisfied in Example 1.8.

1477

Theorem 1.15. Let S be a completely simple semigroup of nonnegative matrices. Suppose that there is an idempotent $P \in S$ that satisfies (1.14) and (1.15). Then an arbitrary element M of S shares the same basis, in the following sense: For each C_s there is $C_{\pi(s)}$ (where $\pi = \pi_{PM}$) such that

1478

1479

1480

1481

$$M_{|C_s \times C_t} = 0, \text{ if } t \neq \pi(s),$$

$$M_{|C_s \times C_{\pi(s)}} \neq 0$$

(meaning that the entries are not all zeros)

1482

$$M_{|C_s \times T_c} = 0, s = 1, 2, \dots, k,$$

$$M_{|T_r \times T_c} = 0,$$

$$M_{|T_r \times C_s} = 0, s = 1, 2, \dots, k.$$

Moreover if all the idempotents satisfy (1.14) and (1.15), then all elements in S have the same zero rows and zero columns. In particular, the $C_s \times C_s$ block of the idempotents are strictly positive.

1483

1484

1485

Proof (Sketch). As in the proof of Proposition 1.23, we look at a positive M_{ij} , $i \in C_s, j \in C_t$

1486

1487

$$M_{ij} > 0 \implies P_{ii} M_{ij} P_{jj} > 0,$$

which implies that $j \in C_{\pi(s)}$, where $\pi = \pi_{PM}$. Hence, $M_{|C_s \times C_t} = 0$ if $t \neq \pi(s)$ while $M_{|C_s \times D} \neq 0$ by the requirement that the rank of M be k .

1488

1489

Equations (1.14) and (1.15) can in a similar way be extended to all of C_s and C_t

1490

$$P_{ii} > 0 \implies P_{ii'} \geq P_{ii} P_{ii'} > 0 \quad \forall i' \in C_s.$$

This is used to obtain, for example

1491

$$M_{|C_s \times T_c} = 0$$

as follows. For $i \in C_s, j \in C_t, t \neq s$,

1492

$$\begin{aligned} 0 = (PMP)_{ij} &= \sum_{h \in C_s, l \in C_t} P_{ih} M_{hl} P_{lj} \\ &+ \sum_{\substack{h \in C_s, \\ l \in T_c}} P_{ih} M_{hl} P_{lj} + \sum_{\substack{h \in T_r, \\ l \in T_c \cup C_t}} P_{ih} M_{hl} P_{lj}. \end{aligned}$$

Then we can show that $M_{hl} = 0$ since the second term is 0.

1493

For the details of the proof, we refer to [152].

□ 1494

We now turn to *stochastic* matrices, i.e., $d \times d$ matrices with nonnegative entries M_{ij} with row sums $\sum_{j=1}^d M_{ij} = 1$. If all column sums are also 1, the matrix is said to be *doubly stochastic* or *bistochastic*. If row sums are ≤ 1 , then the matrix is called *substochastic*.

1495

1496

1497

1498

Our interest in these matrices is of course motivated by their extensive use in probability theory as transition probability matrices for Markov chains. To analyze the evolution of the Markov chain, we must study products of transition probability matrices. The asymptotic behavior of the chain can often be described by idempotent stochastic matrices appearing as limits of products of transition probability matrices.

1499

1500

1501

1502

1503

Remark 1.19. Any bistochastic matrix is a convex combination of permutation matrices (i.e., 0–1 matrices with exactly one nonzero element in each row and each column) by a famous theorem of Birkhoff. Any stochastic matrix is a convex combination of stochastic 0–1 matrices, which could then be called transformation matrices. The representation is not unique, but the idempotents of the generated semigroup are (see discussion in [101]).

1504

1505

1506

1507

1508

1509

Topological semigroups of (sub)stochastic matrices are *compact*, since entries are bounded by 1. Then we can use the preceding theory to obtain powerful results. Using notations from Theorems 1.11 and 1.12, we state Theorem 1.16.

1510

1511

1512

Theorem 1.16. Let P be a $d \times d$ stochastic idempotent matrix of rank k . Then there is a basis $\{T_c, C_1, \dots, C_k\}$ for P such that $P_{|C_s \times C_s}$ has identical positive rows of sum 1 and $P_{|C_s \times C_t} = 0, s \neq t$.

1513

1514

1515

If $i \in T_c$,

1516

$$\frac{P_{ij}}{P_{jj}} = \frac{P_{ih}}{P_{jh}}, \quad j, h \in C_s.$$

Conversely, any stochastic matrix with these properties is idempotent.

1517

Proof. Since P is stochastic, T_r is empty, therefore P is 0 on $C_s \times (C_s)^c$. The rows of $P_{|C_s \times C_s}$ are proportional with row sum 1, hence equal. The rest is clear. □

1518

1519

Corollary 1.9. Let P be a $d \times d$ bistochastic idempotent matrix of rank k . Then there is a basis $\{C_1, \dots, C_k\}$ for P such that $P_{ij} = \frac{1}{|C_s|}$ for $i, j \in C_s$ and $P_{ij} = 0$ for $i \in C_s, j \in C_t, t \neq s$ ($1 \leq s, t \leq k$). 1520
1521
1522

Proof. This time rows and columns of $P|_{C_s \times C_s}$ are also equal and sum to 1. Hence, the block consist of identical entries. 1523
□ 1524

Locally compact groups of stochastic matrices are finite by Corollary 1.8. Completely simple semigroups of stochastic matrices need not share the same basis, as Example 1.8 shows. Another example is $S = \{e, f\}$, where 1525
1526
1527

$$e = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, f = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

If T is empty for all idempotents of S , then Proposition 1.23 says that they all have a common basis. In particular, we have Corollary 1.10. 1528
1529

Corollary 1.10. If S is a completely simple semigroup of bistochastic matrices, then S is a finite group. 1530
1531

Proof. By Corollary 1.9, there is only one possible idempotent for each basis. Then S consists of just one group. 1532
□ 1533

Remark 1.20. If (1.15) is satisfied – probabilistically speaking this means that the Markov chain governed by P can jump from a transient state to at least two different recurrent classes – Theorem 1.15 is applicable. Note that T_r is empty. 1534
1535
1536

Exercise 1.40. Let G be a compact group of rank one $d \times d$ real matrices. Show that G is either a singleton or a two-point group $\{e, -e\}$. 1537
1538

Exercise 1.41. Verify that the minimum number of elements for a nontrivial Rees product is 8. Also show that the set $\{a, b, c, d, e, f, g, h\}$ constitutes such a Rees product, when 1539
1540
1541

$$a = \begin{pmatrix} O & I \\ O & I \end{pmatrix}, b = \begin{pmatrix} O & I \\ O & M \end{pmatrix}, c = \begin{pmatrix} O & M \\ O & M \end{pmatrix}, d = \begin{pmatrix} O & M \\ O & I \end{pmatrix},$$

$$e = \begin{pmatrix} I & O \\ I & O \end{pmatrix}, f = \begin{pmatrix} I & O \\ M & O \end{pmatrix}, g = \begin{pmatrix} M & O \\ I & O \end{pmatrix}, h = \begin{pmatrix} M & O \\ M & O \end{pmatrix},$$

and I, M , and O are the 2×2 blocks 1542

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \text{ and } \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Exercise 1.42. Find a projection Q , such that $PQP = -P$, when $R(P) = \{(x, y) | y = x\}$ and $N(P) = \{(x, y) | y = -\frac{1}{3}x\}$. Also show that such a Q does 1543
1544

not exist if $N(P) = \{(x, y) | y = -x\}$ instead. Verify, e.g., by inspection of the multiplication tables, that the semigroup generated by P and Q is isomorphic to the one in Exercise 1.41.

Exercise 1.43. Let e be a nonnegative idempotent of rank k with basis $\{T, C_1, C_2, \dots, C_k\}$, where $T = T_c \cup T_r$ as in Theorem 1.11. Show that there are column vectors (of appropriate dimensions) $\lambda_i > 0, \alpha_i > 0, \delta_i \geq 0, \gamma_i \geq 0, i = 1, 2, \dots, k$, so that the $C_i \times C_i$ blocks are of the form $\lambda_i \alpha_i^T$, the $C_i \times T_r$ blocks of the form $\lambda_i \gamma_i^T$, the $T_c \times C_i$ blocks of the form $\delta_i \alpha_i^T$, $T_c \times T_r$ block equals $\sum_{i=1}^k \delta_i \gamma_i^T$, and all other blocks are all-zero ones.

Exercise 1.44. Let S be the multiplicative semigroup generated by the stochastic matrices

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, B = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ a & 1-a & 0 \end{pmatrix},$$

where $0 \leq a \leq 1$. Determine the kernel K of S and its product representation when $a = 0$ and when $a > 0$.

Exercise 1.45. Let S be the closed multiplicative semigroup generated by a subset A of $\{\text{circ}(x, x-1, x-1) | x \in \mathbb{R}\}$. (For the definition of circulant matrices, see Exercise 1.9.) Suppose that A contains an idempotent other than the identity matrix. Find the kernel of S .

Exercise 1.46. Let G be a compact abelian group of $d \times d$ real matrices of rank r , $0 < r \leq d$. Show that G is topologically isomorphic to a compact abelian group of $r \times r$ invertible matrices of determinant 1 or -1 .

Exercise 1.47. Find the kernel of the multiplicative semigroup S given by

$$S = \{\text{circ}(a, b, c) | a, b, c \in \mathbb{R}, a^2 + b^2 + c^2 - ab - bc - ca = 1\}.$$

Exercise 1.48. Let S be a multiplicative semigroup of $d \times d$ nonnegative matrices, which have no zero rows or zero columns. Let S have the usual Euclidean topology of $\mathbb{R}^{d^2}$. Then S may not be compact, but S satisfies the compactness condition of Exercise 1.39. Show this!

[Actually, more is true: Let S be a multiplicative semigroup of $d \times d$ nonnegative matrices with the usual topology. Then if A and B are two compact subsets of S , then the set $AB^{-1} (A^{-1}B)$ in Exercise 1.39 is compact whenever none of the matrices in B (A) has any zero row or zero column.]

Exercise 1.49. Let S be the set of 2×2 stochastic matrices of the form

$$\begin{pmatrix} p & 1-p \\ q & 1-q \end{pmatrix},$$

where $0 \leq p, q \leq 1$. There is a natural one-to-one mapping from S to the unit square, namely the one which maps each matrix to the point whose coordinates are the entries of the first column (p, q) . The matrix multiplication in S induces a corresponding operation ("multiplication") in the unit square.

Show that multiplication from the left with (p, q) always moves a point (x, y) in the same direction, the step length being a multiple of $x - y$.

Take a matrix M in S and its corresponding image in the unit square. Describe the image of $\langle M \rangle$ geometrically.

Show that the kernel of S is isomorphic to the main diagonal in the unit square.

Exercise 1.50. Let M and N be two matrices in S of Exercise 1.49. For simplicity, assume that both correspond to points in the lower right hand part of the unit square, where the first coordinate is greater than the second.

Describe the image of $\langle M, N \rangle$ geometrically.

Exercise 1.51. For each of the following Exercises 1.51–1.54, assume that A is an $n \times n$ adjacency matrix for a 2-out, strongly connected directed graph $G = (V, E)$. r (or b) will always be a function on the set $V = \{1, 2, \dots, n\}$ and R will denote the 0–1-matrix representing r , i.e., $R_{ji} = 1$ iff $r(i) = j$. If r and b are such that $A = R + B$, then we call $\{r, b\}$ an *edge-coloring* of G , since we can think of the edge (i, j) being colored r ("red") if $r(i) = j$. Let $S = \langle r, b \rangle$ be the semigroup generated by r and b (for "blue") under composition – we call S a coloring semigroup – with kernel K .

- (i) Show that $\text{rank}(K) = 1$ implies that the graph G is aperiodic.
- (ii) Let w be any element of S . Show that $\text{range}(Kw) = \bigcup_{k \in K} \text{range}(kw) = V$.
- (iii) Let $e = e^2$ be any idempotent in K and $H = eKe$. Show that H is transitive on its range, i.e., show that for any $i \in V$, $\{k(i) | k \in H\} = \text{range}(H)$.
- (iv) Suppose $(ef)^2 = ef$ where e and f are any two idempotents in K . Show that K is a left group.

Exercise 1.52. The following problems refer to a congruence on S introduced by Jarkko Kari in [121]. For $i, j \in V$ write $i \equiv j$ if for any $w_1 \in S$, there exists a $w_2 \in S$ such that $w_2 w_1(i) = w_2 w_1(j)$.

- (i) Show that $\equiv$ induces an equivalence relation on V such that if $r(i) = j$ for $i, j \in V$, then $r([i]) \subset [j]$, whence $[r([i])] = [j]$, where $[i]$ is the equivalence class under $\equiv$ containing i . Thus, $\equiv$ is a congruence on the graph G and if $\equiv$ is nontrivial, then $G/\equiv$ is a quotient graph with fewer vertices than G .
- (ii) Let $\Pi = \{P_1, P_2, \dots, P_m\}$ be the set of partitions of K . Let $\equiv_{P_t}$ be the equivalence relation induced by the partition P_t . Show that $i \equiv j$ if and only if $i \equiv_{P_t} j$ for all $P_t, t = 1, 2, \dots, m$.
- (iii) Suppose K is a left group. Show that $G/\equiv$ is both 2-in and 2-out. In fact, show that both r and b become permutations when restricted to the vertices of $G/\equiv$.
- (iv) Suppose $k_1, k_2 \in K$ and let $R_1 = \text{range}(k_1)$ and $R_2 = \text{range}(k_2)$. Suppose that $R_1 \setminus (R_1 \cap R_2) = \{i\}$ and $R_2 \setminus (R_1 \cap R_2) = \{j\}, i \neq j$. Show that $i \equiv j$.

- (v) Let $A = R + B = R' + B'$ (representing two different colorings). Let $\Delta = R - B$ and $\Delta' = R' - B'$. Show that $\text{rank}(\Delta) = \text{rank}(\Delta')$. (Here rank refers to the standard rank for matrices.)

Exercise 1.53. Define the *trace* of a transformation $f \in \mathcal{T}_n$ to be the number of fixed points of the transformation: $\text{trace}(f) = |\{j : f(j) = j\}|$.

- (i) Show that a transformation e is idempotent iff $\text{trace}(e) = \text{rank}(e)$.
 (ii) Show that for any transformations f, g , $\text{trace}(fg) = \text{trace}(gf)$.

Exercise 1.54. For the following problems, we will assume that $S = \langle r, b \rangle$ is solvable. A transformation semigroup is called solvable if for any finite set of transformations $s_1, s_2, \dots, s_m \in S$ and any permutation ϕ on m symbols, $\text{trace}(s_1 s_2 \cdots s_m) = \text{trace}(s_{\phi(1)} s_{\phi(2)} \cdots s_{\phi(m)})$.

- (i) Show that if S is solvable with kernel K then K is a left group. (Hint: Show that if $e, f \in K$ are idempotent transformations, then $(ef)^2 = ef$, and use 1.51(iv).)
 (ii) Given transformations r, b let p, q be such that r^p and b^q are idempotents. Let $S' = \langle r^p, b^q \rangle$ with kernel K' . Suppose that $\text{rank}(K') = m$. Show that $\text{trace}(r^p b^q) = m$. (Hint: Recall that if $w \in K \cong X \times H \times Y$, then $w^{|H|}$ is idempotent.) Deduce that for some ℓ , $(r^p b^q)^\ell \in K$.
 (iii) Use (ii) to show that for some m , $(rb)^m \in K$.

Exercise 1.55. Let P be a real $d \times d$ idempotent matrix with rank k , $1 \leq k \leq d$. Show that there is a unique $d \times k$ matrix L and a unique $k \times d$ matrix M such that $P = LM$ and $ML = I_k$, the $k \times k$ identity matrix. (Such matrices are treated by Chakraborty and Mukherjea in [32].)

Exercise 1.56. Let S be a multiplicative semigroup of real $d \times d$ matrices such that all matrices have the same rank k , $1 \leq k \leq d$. Assume also that there is at least one idempotent e in S . Show that

- (i) S is completely simple if and only if eSe is a group.
 (ii) S can be embedded in a completely simple semigroup.

1.7 Semigroups of Infinite Dimensional Matrices

Section 1.7 discusses the theory of semigroups of countably infinite matrices, usually indexed by the positive integers. We begin by developing some general theory but rather quickly the discussion focuses on matrices with nonnegative entries. Here, we can present good, in a certain sense optimal, analogues of the results from Sect. 1.6. In the process, we prove Theorem 1.11, which was used but not proved in Sect. 1.6. Consider a multiplicative semigroup S of real matrices indexed by the set of positive integers E (which we call the *state space* of S) such that for each $x \in S$,

$$\sup_{i \in E} \sum_{j \in E} |x_{ij}| < \infty. \quad (1.16)$$

Each $x \in S$ can be regarded as a linear operator (also denoted by x) on the infinite dimensional vector space

$$l_\infty \equiv \{v = (v_1, v_2, \dots)\}$$

of real bounded sequences

$$[x(v)]_i \equiv \sum_{j \in E} x_{ij} v_j. \quad (1.17)$$

Remark 1.21. The index set E is chosen to be the positive integers only for convenience. Any countable set works just as well. The condition (1.17) is needed to assure that the operation in (1.17) is welldefined.

Let $R(x)$ denote the range of the operator x defined in (1.17). Then we define the set $m(S)$ by

$$m(S) \equiv \{x \in S \mid R(xy) = R(x) \text{ for any } y \in S\},$$

where $m(S)$ is the analog of the set of matrices of minimal rank in Proposition 1.9.

Proposition 1.24. Suppose that S has a completely simple kernel K . Then $K = m(S)$.

Proof. To show that $K \subset m(S)$, let x be any element of K and y any element of S . Then $xy \in K$ and there is a $w \in K$ such that $x = xyw$ (since xK is a minimal right ideal). In particular, $R(x) = R(xyw) \subset R(xy)$. But $R(x) \supset R(xy)$ trivially, so $R(x) = R(xy)$ and $x \in m(S)$.

$m(S)$ is nonempty, since it contains K . Let $z \in m(S)$. If x is any element of K , then $zx \in K$. zx belongs to a subgroup of K ; let e be the identity of this subgroup and y the inverse of zx : $y(zx) = e = (zx)y$. Then $R(e) = R(zxy) = R(z)$, since $z \in m(S)$. But e acts as the identity mapping on $R(e) = R(z)$ so $ez = z$. Then $z \in K$ since e is. $\square$

Example 1.9. (i) Let V be a countably infinite set. Let S be the set of all those injective maps from V into itself that are not surjective. Then S is a semigroup under the usual composition of maps. Each map $x \in S$ can be regarded as an infinite dimensional matrix

$$(M_x)_{ij} = \begin{cases} 1, & \text{if } x(i) = j, \\ 0, & \text{otherwise.} \end{cases}$$

Then $(M_x v)_i = v_{x(i)}$, $i \in V$, and $M_x M_y v)_i = (M_x(v_{y(\cdot)}))_i = v_{y(x(i))} = (M_{yx} v)_i$. The ranges of M_x and x are related as follows: $j \in V$ belongs to the range of x if and only if $\delta_j \in R(M_x)$, where δ_j is the vector in l_∞ with all zero entries except for a 1 as the j th coordinate.

We write $x, R(x)$ for the matrices and their ranges, and the semigroup of matrices is called S as well. $m(S)$ is empty for this semigroup: For any $x \in S$, let

- $v \in R(x)$ so that $x(w) = v$ for some $w \in l_\infty$. But for some $y \in S$, $w \notin R(y)$.
 $R(xy)$ does not contain v , so $R(xy) \neq R(x)$.
(ii) Suppose S is a subsemigroup of the set V^V of all maps from V into itself. Suppose S contains an element x with $R(x)$ finite. (We now regard the semigroup elements as maps, not matrices.) Then the set

$$A \equiv \{y \in S \mid |R(y)| \leq |R(z)| \text{ for any } z \in S\}$$

(where $|C|$ stands for the cardinality of the set C) is the completely simple kernel $m(S)$ of S .

The proof is similar to the proofs of Propositions 1.8 and 1.3. The ranges involved are all finite, and the partitions also consist of a finite number of classes.

We defined a semigroup S to be pseudoinvertible if some power of each element lies in a subgroup of S (Remark 1.6).

Theorem 1.17. *Let S be a pseudoinvertible semigroup of infinite dimensional matrices satisfying (1.16). If $m(S)$ is nonempty, then it is the completely simple minimal ideal of S .*

Proof. By Theorem 1.3, we need only prove that $m(S)$ is simple; i.e., $m(S)x = m(S)$ for all $x \in m(S)$. In fact if $x \in m(S)$ and $y, z \in S$ then $R(xyz) = R(x) = R(xy)$ and $R(yxz) = R(yx)$, since $R(xz) = R(x)$. Hence, $xy, yx \in m(S)$ so $m(S)$ is an ideal of S .

For I , another ideal of S , we note that $I \cap m(S)$ is an ideal as well. We can then take I to be a subset of $m(S)$. Let $w \in m(S)$, $y \in I$, $x = wy$. Then $x \in I$. Since S is pseudoinvertible $x^k \in G$, where G is a subgroup of S . If e is the identity of G , then e , having the form $x^k z$, belongs to I . Let $x' = x^{k-1}$ be the inverse of x in G . Then $R(e) = R(xx') = R(wyx') = R(w)$, since $w \in m(S)$. But $ew = w$, since e acts as the identity on $R(e) = R(w)$. Then $w \in I$, since $e \in I$. Thus $m(S)$ is the minimal ideal and it is clearly simple. $\square$

Remark 1.22. Let $x \in m(S)$. Then $l_\infty = R(x) \cup N(x)$ and $R(x) \cap N(x) = \{0\}$ or equivalently

$$l_\infty = R(x) \oplus N(x),$$

where $N(x) \equiv \{v \in l_\infty \mid x(v) = 0\}$.

Proof. Let e be the identity of the subgroup G containing x . Then $R(e) = R(ex) = R(x)$. $e(v) = 0 \implies 0 = xe(v) = x(v)$ and $x(v) = 0 \implies 0 = x^{-1}x(v) = e(v)$, where x^{-1} is the inverse of x in G . Now take $v \in R(x) \cap N(x) = R(e) \cap N(e)$. Then $v = e(v) = 0$. For $v \in l_\infty$, write $v = e(v) + v - e(v)$. Plainly $e(v) \in R(e) = R(x)$ and $v - e(v) \in N(e) = N(x)$. $\square$

Remark 1.23. For $x \in m(S)$, the subgroup G of $m(S)$ containing x is a subgroup of all bijections from $R(x)$ to itself.

As pointed out at the beginning of Sect. 1.7, more can be said if we specialize in nonnegative matrices satisfying (1.16). Theorem 1.18 was already used extensively in the study of finite dimensional nonnegative matrices.

Theorem 1.18. *Let e be an idempotent nonnegative matrix. Then there is a unique partition of the state space E , called the basis of e*

$$E = T \cup C_1 \cup C_2 \cup \dots$$

such that

- (i) If $i \in T$, then the i th row or the i th column (or both) of e consists of zeros only;
- (ii) For $s \geq 1$, $i \in C_s$, $j \in C_s$, there exist $\gamma_s(i, j) > 0$, $\bar{\gamma}_s(i, j) > 0$ such that

$$e_{ih} = \gamma_s(i, j)e_{jh}, e_{hi} = \bar{\gamma}_s(i, j)e_{hj}, h \in E; \quad (1.18)$$

- (iii) For $s, t \geq 1$, $s \neq t$, $i \in C_s$, $j \in C_t$,

$$e_{ii} > 0 \text{ and } e_{ij} = 0. \quad (1.19)$$

Proof. Define T as in (i). $e_{ij} = \sum_k e_{ik}e_{kj}$, so the products can be taken over the complement T^c of T . In other words, the restriction of e to $T^c \times T^c$ is also idempotent. If $i \in T^c$ and $e_{ij} = 0$ for all $j \in T^c$, then the whole row is 0, contradicting the choice of i : $e_{ik} = \sum_{j \in T^c} e_{ij}e_{jk} = 0$. Similarly, the j th column of $e|_{T^c \times T^c}$ is also nonzero.

For the moment, we assume that T is empty. We first establish

$$e_{ii} > 0 \text{ for all } i \in E. \quad (1.20)$$

Assume the contrary: $e_{ii} = 0$ for some $i \in E$, $i = 1$, say. Let A be the set $\{j \in E \mid e_{1j} > 0\}$. A is not empty, since we assumed that e had no zero rows or zero columns. Then

$$0 = e_{1j} = \sum_{k \in A} e_{1k}e_{kj} \text{ for } j \notin A$$

so $e_{kj} = 0$ for $j \notin A$, $k \in A$; i.e., e restricted to $A \times A^c$ is a zero block.

Call the restrictions of e to $A^c \times A^c$, $A^c \times A$, and $A \times A$, P_1 , P_2 , and P_3 , respectively. Then e^2 has the blocks P_1^2 , $P_1P_2 + P_2P_3$ and P_3^2 , respectively. By idempotency, we obtain equations

$$\begin{aligned} P_1 &= P_1^2, \\ P_2 &= P_1P_2 + P_2P_3, \\ P_3 &= P_3^2. \end{aligned}$$

Algebraic manipulations show that $P_1 P_2 P_3 = 0$. ($P_1 P_2 = P_1^2 P_2 + P_1 P_2 P_3 =$ 1742
 $P_1 P_2 + P_1 P_2 P_3$). Also we obtain a contradiction 1743

$$(P_1 P_2 P_3)_{kl} \geq (P_1)_{kj} (P_2 P_3)_{js} > 0$$

unless $P_2 P_3 = 0$. [$(P_1)_{kj} > 0$ for some $k \in A^c$ otherwise there would be a zero 1744
column in e .] 1745

Finally $P_2 = 0$. This is because P_3 cannot have a zero row, so $(P_3)_{kj} > 0$ for 1746
all k and some j (depending on k). Then $(P_2 P_3)_{ij} \geq (P_2)_{ik} (P_3)_{kj}$, which may be 1747
 > 0 unless $P_2 = 0$. 1748

We now have a zero row in e : Row 1 is the first row of P_2 ($1 \in A^c$), and by 1749
definition of A , $(P_1)_{1j} = 0$ for all $j \in A^c$. But e has no zero row; hence (1.20) 1750
follows. 1751

Our next claim is 1752

$$e_{ij} = 0 \iff e_{ji} = 0. \quad (1.21)$$

The proof of this claim rests on a similar decomposition of e into blocks. Suppose 1753
 $e_{1j} = 0$ for some j . Define the set $B \equiv \{k \mid e_{1k} > 0\}$. B is nonempty; by (1.20) 1754
 $1 \in B$ and $j \notin B$. 1755

For $k \in B^c$, $0 = e_{1k} = \sum_{l \in B} e_{1l} e_{lk}$. Hence, e restricted to $B \times B^c$ is 0. Call 1756
the other $B \times B$, $B^c \times B$, $B^c \times B^c$ blocks, respectively, P_4 , P_5 , and P_6 . Using the 1757
relation $e^2 = e$, we get 1758

$$P_4 = P_4^2,$$

$$P_5 = P_5 P_4 + P_6 P_5,$$

$$P_6 = P_6^2.$$

By a simple argument as in the proof of (1.20), we end up with $P_5 = 0$; hence, 1761
 $e_{j1} = 0$. 1762

If $C_i = \{j \mid e_{ij} > 0\}$, then $e|_{C_i \times C_i}$ is a strictly positive idempotent matrix. (1.22)

This follows immediately, since $(e^2)_{ii} = e_{ii} = \sum_k e_{ik} e_{ki} = \sum_{k \in C_i} e_{ik} e_{ki}$, by (1.21). 1763
Likewise $(e^2)_{ij} = e_{ij} = \sum_{k \in C_i} e_{ik} e_{kj}$. 1764

Finally, 1765

A strictly positive idempotent matrix D has rank 1. (1.23)

Write $D_{ir} = a(i, r, k) D_{kr}$. $(D^2)_{kr} = D_{kr} \geq D_{ki} D_{ir}$. 1766

$$a(i, r, k) = \frac{D_{ir}}{D_{kr}} \leq \frac{1}{D_{ki}}.$$

Then 1767

$$\sup_r a(i, r, k) = \beta(i, k) \leq \frac{1}{D_{ki}} < \infty.$$

1768

1.7 Semigroups of Infinite Dimensional Matrices

57

Also for any t, k and r

1769

$$\frac{D_{tr}}{D_{kr}} \geq D_{tk} > 0.$$

We now have

1770

$$\sum_t [\beta(i, k) - a(i, t, k)] D_{kt} D_{tr} = \beta(i, k) D_{kr} - D_{ir} = [\beta(i, k) - a(i, r, k)] D_{kr}.$$

Therefore for any t ,

1771

$$0 \leq [\beta(i, k) - a(i, t, k)] D_{kt} \frac{D_{tr}}{D_{kr}} \leq \beta(i, k) - a(i, r, k),$$

and

1772

$$0 \leq [\beta(i, k) - a(i, t, k)] D_{kt} D_{tk} \leq \inf_r (\beta(i, k) - a(i, r, k)) = 0.$$

Hence $\beta(i, k) = a(i, t, k)$ for each t , so $D_{ir} = \beta(i, k) D_{kr}$, whence $\text{rank}(D) = 1$.

1773

It is clear from (1.20)–(1.23) that we can now partition E into disjoint classes $\{C_1, C_2, \dots\}$, where i and j belong to the same class if and only if $e_{ij} > 0$. Also $e_{|C_s \times C_s}$ is a rank 1 idempotent matrix for all $s = 1, 2, \dots$

1774

1775

1776

Returning to the original matrix e (with nonempty T), it is easy to verify that

1777

- (i) If the i th column of e is a zero column then for j and k in the same C -class of T^c

1778

1779

$$\frac{e_{ij}}{e_{jj}} = \frac{e_{ik}}{e_{jk}}.$$

Consider

1780

$$(e^2)_{ij} = e_{ij} = \sum_{h \in T^c} e_{ih} e_{hj} = \sum_{h \in C_s} e_{ih} \beta(h, j) e_{jj}$$

and

1781

$$e_{ik} = \sum_{h \in T^c} e_{ih} e_{hk} = \sum_{h \in C_s} e_{ih} \beta(h, j) e_{jk};$$

- (ii) If the i th row of e is a zero row, then for j and k in the same C -class of T^c

1782

$$\frac{e_{ji}}{e_{jj}} = \frac{e_{ki}}{e_{kj}}.$$

From (i) and (ii) and (1.23), we conclude that (1.18) is satisfied.

1783

This finishes the proof of the Theorem.

□ 1784

Let us now turn to other elements x of the group G with identity e . We find that

1785

- (i) The zero rows and the zero columns are the same because of the relations $ex = xe = x$ and $xx^{-1} = x^{-1}x = e$;
- (ii) Since $x = xe = ex$ we can replace e by x in (1.18);
- (iii) For a given C -class C_s on the basis of e , there exist unique s' and s'' such that the block, $C_k \times C_s$ of x is strictly positive or an all zero block, depending on

1786

1787

1788

1789

1790

whether or not $k = s'$; the block $C_s \times C_k$ of x is a strictly positive block or an all zero one depending on whether or not $k = s''$.

To prove (iii), let $i \in C_s$. Since $e = xx^{-1}$, we have

$$0 \leq e_{ii} = \sum_{u \in T^c} x_{iu}(x^{-1})_{ui} \quad (1.24)$$

therefore, there is a s' such that for some $u \in C_{s''}$, $x_{iu} > 0$, $(x^{-1})_{ui} > 0$. This means that $x_{iu'} > 0$ for all $u' \in C_{s''}$: $x_{iu'} \geq x_{iu}e_{uu'} > 0$. Conversely, $x_{i'u'} > 0$ for all $i' \in C_{s'}$.

On the other hand,

$$x_{ij} > 0 \Rightarrow e_{uj} \geq (x^{-1})_{ui}x_{ij} > 0 \Rightarrow j \in C_{s'}.$$

Similar arguments show that $x_{ij} > 0$ for $i \in T^c$, $j \in C_s$ if and only if $i \in C_{s'}$ for some unique s' .

Observation (iii) and (1.24) show that there exists a permutation π_x on the set of C -classes E_0 (E_0 is the quotient space $T^c / \equiv$ where $i \equiv j$ if and only if $e_{ij} > 0$) defined as follows:

For $s, t \in E_0$, $\pi_x(s) = t$ if and only if the $C_s \times C_t$ block of e is all positive.

Equation (1.24) shows that $\pi_{x^{-1}}(t) = s \iff \pi_x(s) = t$. More generally $\pi_{xy} = \pi_y \circ \pi_x$, since

$$\pi_{xy}(s) = t \iff \text{for } i \in C_s : (xy)_{ij} > 0 \text{ for } j \in C_t$$

and

$$(xy)_{ij} = \sum_k x_{ik}y_{kj},$$

where the summation gives a positive result if and only if $k \in C_{\pi_x(s)}$ and $j \in C_{\pi_y(\pi_x(s))}$.

Hence, π is an antihomomorphism from G to the group H of permutations on E_0 . As in Sect. 1.6, see especially the proof of (1.13), we see that $\pi_x = \pi_e$ (the identity permutation) if and only if

$$x|_{C_s \times C_s} = \mu(s)e|_{C_s \times C_s},$$

for some positive constants $\mu(s)$, $s \in E_0$. (The finite sums of Sect. 1.6 are replaced by possibly infinite convergent sums.)

Let $x, y \in G_1 \equiv \{z \in G \mid \pi_z = \pi_e\}$, the kernel of the homomorphism π . Then G_1 is a group, and the constants of proportionality preserve multiplication: $\mu_x(s)\mu_y(s) = \mu_{xy}(s)$, see the proof of Proposition 1.22.

Suppose that the group G is compact (with respect to a topology where convergence implies pointwise convergence). If $x \in G$ and $\pi_x = \pi_e$, then $\mu_x(s) = 1$ for all $s \in E_0$, if not, either $x|_{C_s \times C_s}^n$ or $x|_{C_s \times C_s}^{-n}$ goes to infinity for some $s \in E_0$. This comment implies that for a compact G , π is one-to-one. G can (algebraically) be

mapped isomorphically onto a subgroup of the permutations of E_0 . If E_0 is finite, then G is finite, see Corollary 1.8.

Theorem 1.19. *Let G be a multiplicative group of infinite dimensional nonnegative matrices satisfying (1.16). Let e be the identity of G with basis $\{T, C_1, C_2, \dots\}$. Then there is an antihomomorphism π from G into the group of permutations on the set E_0 of C -classes, with the following properties:*

- (i) *For each $x \in G$, the block $C_s \times C_t$ is a strictly positive block or an all-zero block depending on whether or not $t = \pi_x(s)$;*
- (ii) *If G is compact with respect to a topology as least as strong as the topology of pointwise convergence and if E_0 is finite, then G is finite; indeed if G consists of only strictly positive matrices then G is a singleton;*
- (iii) *If G is a group of infinite dimensional stochastic matrices, then G is finite if E_0 is finite.*

Proof. It remains to prove only (iii). (Notice that the infinite dimensional stochastic matrices do not in general form a relatively compact set.) (iii) follows since x^n and x^{-n} are bounded on $C_s \times C_s$; hence $\mu_x(s) = 1$ for all s . Then π is injective, so G has to be finite if E_0 is. $\square$

The proof of Proposition 1.23 can be modified to yield the infinite dimensional analog, Proposition 1.25.

Proposition 1.25. *If S is a completely simple semigroup of infinite bistochastic matrices, then S is a group.*

Proof. Let the common partition be $\{C_1, C_2, \dots\}$. Let e be an idempotent of S . On $C_s \times C_s$ both the row sum and the column sum are 1. Hence, C_s is a finite set and $e_{ij} = \frac{1}{|C_s|}$ for all $i, j \in C_s$ ($s = 1, 2, \dots$). Since there is only one idempotent in S , S is a group. $\square$

Exercise 1.57. Find a matrix semigroup that is isomorphic to the bicyclic semigroup. How can you represent $\alpha, \beta, \beta^m, \alpha^n, \beta\alpha$ (see p. 21) as matrices?

Exercise 1.58. Suppose that S is a completely simple semigroup of infinite dimensional nonnegative matrices satisfying (1.16) and with no zero rows or columns. Show that there is a unique partition $\{C_1, C_2, \dots\}$ of E , the positive integers, such that the following hold:

- (i) If e is an idempotent in S , then the above partition is the basis of e . (ii) For each $x \in S$, the rank of x restricted to the $C_k \times C_k$ block is 1, for each C_k . A left simple semigroup that contains an idempotent.
- (ii) For each $x \in S$, the rank of x restricted to the $C_k \times C_k$ block is 1, for each C_k .
- (iii) There is an antihomomorphism p from S into the group of permutations on the set $E_0 = \{1, 2, \dots, m\}$, where $m \leq \infty$ is the number of C -classes in the partition above, such that for $x \in S$, $i \in C_k$, $x_{ij} > 0$ precisely when $j \in C_{p(x)(k)}$.

Exercise 1.59. Let e be an idempotent infinite bistochastic matrix. Show that the partition E_0 is infinite.

1.8 Notes and Comments

1862

At the previous turn of the century there was a lively discussion on the axioms of group theory. In that connection, some basic semigroup properties were found, but the activity did not give rise, yet, to a theory of semigroup proper. Through the work of Suschkewitsch and Rees, semigroup theory began to take on a life of its own. Suschkewitsch [219] and Rees [196] are considered to be the breakthrough papers.

The comprehensive works on the algebraic theory of semigroups, viz., [42, 43, 136] contain a large bibliography and many historical notes, albeit scattered throughout the text.

The theory of topological semigroups seems to have been first presented in book form by Paalman-de Miranda [185], whose first edition appeared in 1964. Later monographs are, for example, [9, 10, 29, 93].

Historical notes are to be found in all of these works.

The probabilistically oriented monographs by Rosenblatt [204] and Mukherjea and Tserpes [176] provide some necessary background on topological semigroups as well.

The exercises are drawn from various sources, most frequently from Paalman-de Miranda [185] and Clifford and Preston [43]. Also [10, 89] have been sources for the exercises.

Section 1.1

1881

The basic definitions are given in [43], Chap. 1.4.

The semigroup $\mathcal{B}_X$ of relations on the set X plays an important role in algebraic coding theory, see [11]. Simon [213] is another textbook in theoretical computer science using semigroup theory in an essential way, to develop the theory of automata.

Section 1.5

1887

The general references for this section are [9, 115, 185]. The fundamental Rees–Suschkewitsch theorem was proved by Wallace [234]. The structure theory of locally compact completely 0-simple semigroups (Theorem 1.9) is due to Owen [183, 184], see also [111]. Example 1.2 was studied in [184] and [104].

Section 1.6

1892

As pointed out in the text, Mukherjea [152] studies the completely simple semigroups of matrices in more detail. Clark [41] treats the kernel of a matrix semigroup.

Högnäs [108] studies a semigroup with a closely related structure: the semigroup of analytic functions with a common fixed point (with composition of functions as semigroup operation). In many respects, the essential features of the whole semigroup are expressed by some finite dimensional subsemigroups, which turn out to be semigroups of matrices.

A matrix application of a different kind is furnished by Meyer [143] in a study of a finite irreducible Markov transition probability matrix P and the associated invariant probability measure π . P belongs to a group of singular matrices. The inverse of P in that group may be used to describe how π varies as a function of P .

The striking result on compact groups of nonnegative matrices, Corollary 1.8 was found by Brown [21] and Flor [68]. The proof here, however, follows the one given in [155].

The facts about bistochastic matrices (Theorem 1.16 and its Corollaries) are taken from [137].

Simple examples of nondirect Rees products (such as Exercise 1.41) are given in [113].

The 2×2 stochastic matrices in Exercise 1.49 are treated, e.g., in [176, 204].

We gratefully acknowledge the contribution by Greg Budzban and Phil Feinsilver who compiled the Exercises 1.51–1.54 for our use. They are taken from [24, 25].

The semigroup operation in $G/\equiv$ in Exercise 1.52 is the following: The product of $[i]$ and $[j]$ is defined to be the equivalence class containing ij , cf. [43], p. 16.

The result in Exercise 1.52(iv) was used by Trahtman in his proof of the Road Coloring Problem, though it was stated using different terminology, see [226].

The definition of solvable transformation used in Exercise 1.54 is a consequence of a result by Radjavi and Rosenthal [193].

The decomposition of the idempotent P in Exercise 1.55 is due to Chakraborty and Mukherjea in [32]. The Exercise 1.56 is also taken from there.

There is nowadays a large body of work dedicated to a linear algebra based on the semiring operations $(\max, +)$ on $\mathbb{R}_+ \cup \{-\infty\}$, i.e., the “addition” is taking the maximum and the “multiplication” is ordinary addition. This so-called Max-Plus algebra and some similar constructs have been invented and reinvented many times for different applications, such as queuing theory, analysis of networks, graph theory, Markov decision processes, and discrete event systems. The minimal common core of these different theories is presented in a survey by Gaubert and Max Plus (a working group) in [74].

Section 1.7

The material in this section is mostly from [156]. Theorem 1.18 appears in [75]. Its finite-dimensional version was first discovered by Denny Brown. For related results, see [53].

AUTHOR QUERIES

- AQ1. Please check if the Table 1.1 inserted here is appropriate.
- AQ2. Please check if the Table 1.2 inserted here is appropriate.
- AQ3. Please check equation 1.1 is repeating three times and 1.2–1.4 are repeating two times.
- AQ4. Here onward equation number is repeat again please check.

Uncorrected Proof

Chapter 2

Probability Measures on Topological Semigroups

1

2

3

2.1 Introduction

4

This chapter forms the core of this book. It introduces and develops the key concepts, methods, and results (needed for the rest of this book) that involve convolution products of probability measures and their weak convergence. Most of the results presented here are in their final form. However, there are a number of problems, as the reader will readily discover, which are natural, important, and still waiting to be solved. For example, the asymptotic behavior of the sequence of convolution powers of a probability measure μ , though reasonably completely known in a compact or discrete semigroup (see Theorems 2.7 and 2.13) is not completely clear in the noncompact nondiscrete situation. Also, the problem of weak convergence of convolution products of a sequence of nonidentical probability measures, even though reasonably resolved in a compact abelian semigroup as well as in a discrete group or a discrete abelian semigroup (see Theorem 2.22 and 2.27, Corollary 2.6) is far from being resolved even in the case of compact groups.

The central result (that is needed for much of the work covered in this book) is the characterization of idempotent probability measures on locally compact or complete metric semigroups. It is very interesting and pleasantly surprising that such measures are supported by a class of semigroups, called completely simple semigroups, for which there already exists a rich theory. (Such semigroups are studied in detail in Chap. 1.) It is this information that is needed for the identification of all such measures (see Theorem 2.2). The final form of this result evolved in stages over many years through the works of many authors (see [88, 90, 123, 125, 138, 172, 191, 235]). This result was first obtained in compact groups in [123], in compact semigroups in [191] and in [88] independently, and in locally compact (or complete metric) semigroups in [172]. Idempotent probability measures have invariance properties (see Proposition 2.4) and, as such, we have also looked into invariant measures in some detail. Thus, Sect. 2.2 deals with idempotent and invariant measures.

In Sect. 2.3, we utilize the results of Sect. 2.2 to describe weak convergence of convolution products of probability measures. The main results in this section are Theorems 2.7, 2.9, 2.10, and 2.11–2.13 and Proposition 2.6.

Results from Sect. 2.2 are again used in Sect. 2.4 to describe the weak convergence of convolution products of nonidentical probability measures. The main results in this section are Theorems 2.22, 2.26, 2.27, and also Corollary 2.6.

In Chap. 2, S always (unless otherwise mentioned) denotes a locally compact Hausdorff second countable topological semigroup. (Often assertions in various results in Chap. 2 are valid in more general topological structures; however, this is not pointed out explicitly, since the serious reader is able to see this by carefully analyzing the proofs.)

$\mathcal{B}$ denotes the class of Borel subsets of S (generated by the open subsets). Let $P(S)$ be the set of all regular probability measures μ on $\mathcal{B}$. Recall (see [186]) that a measure μ in $P(S)$ is called *regular* if for any $\epsilon > 0$, there exists a compact subset K_{ep} in $\mathcal{B}$ such that $\mu(S - K_{ep}) < \epsilon$. It follows that for $B \in \mathcal{B}$, $\mu(B) = \inf\{\mu(V) : V \text{ open, } V \supset B\} = \sup\{\mu(K) : K \text{ compact, } K \subset B\}$. An infinite nonnegative measure μ is called *regular* when this approximation property still holds and for K compact, $\mu(K) < \infty$. In a complete separable metric space, all measures in $P(S)$ are regular. (See [186, p. 29].)

In what follows, we write:

$$Bx^{-1} \equiv \{y \in S \mid yx \in B\};$$

$$x^{-1}B \equiv \{y \in S \mid xy \in B\}, \quad B \in \mathcal{B};$$

$$C(S) \equiv \text{the space of all bounded real continuous functions on } S;$$

$$C_c(S) \equiv \text{the space of functions in } C(S) \text{ with compact support.}$$

2.2 Invariant and Idempotent Probability Measures

For $\mu \in P(S)$, the measures μ_x , defined by $\mu_x(B) = \mu(Bx^{-1})$, and ${}_x\mu$, defined by ${}_x\mu(B) = \mu(x^{-1}B)$, are both regular. The reason is the following: Given $\epsilon > 0$, if K is a compact subset of Bx^{-1} such that $\mu(Bx^{-1} - K) < \epsilon$, then $Kx \equiv \{yx \mid y \in K\}$ is a compact subset of B , and moreover

$$\mu((B - Kx)x^{-1}) \leq \mu(Bx^{-1} - K) < \epsilon,$$

since $K \subset (Kx)x^{-1}$. Thus, μ_x is regular relative to compact sets from inside (and, therefore, relative to open sets from outside). The same is the case with ${}_x\mu$.

Proposition 2.1. *Let $\mu \in P(S)$, $B \in \mathcal{B}$, and V be an open set in $\mathcal{B}$. Then $\mu(Vx^{-1})$ is a lower semicontinuous function of x and $\mu(Bx^{-1})$ is $\mathcal{B}$ -measurable. The same is true for $\mu(x^{-1}V)$ and $\mu(x^{-1}B)$.*

Proof. Notice that for $f \in C(S)$, $\int f(sx)\mu(ds)$ is a continuous function of x (since μ is regular), and moreover

$$\mu(Vx^{-1}) = \sup \left\{ \int f(sx)\mu(ds) : f \in C(S), 0 \leq f \leq 1, f = 0 \text{ on } S - V \right\}.$$

This implies the first assertion in Proposition 2.1. This also means that

63

$$\mathcal{F} = \{B \in \mathcal{B} \mid \mu(Bx^{-1}) \text{ is a } \mathcal{B}\text{-measurable function of } x\}$$

contains all open sets and all closed sets. Furthermore, $\mathcal{F}$ is a monotone class containing the class

64

65

$$\mathcal{F} = \{V \cap W : V \text{ is an open set in } \mathcal{B}, \text{ and } W \text{ is a closed set in } \mathcal{B}\}.$$

Notice that $\mathcal{F}_0$ is closed under finite intersections and the complement of any set in $\mathcal{F}_0$ is a finite disjoint union of sets in $\mathcal{F}_0$; thus it belongs to $\mathcal{F}$. It follows easily that $\mathcal{B} \subset \mathcal{F}$, since $\mathcal{F}$ contains the algebra generated by $\mathcal{F}_0$. $\square$

66

67

68

Definition 2.1. Let $\mu, \nu \in P(S)$ and $f \in C(S)$. Then the iterated integral

69

$$I(f) = \iint f(xy) \mu(dx) \nu(dy)$$

is welldefined. By the Riesz representation theorem (see [167, Theorem 5.10]), there is a unique regular probability measure λ on $\mathcal{B}$ such that for any $f \in C_c(S)$,

70

71

$$\int f d\lambda = I(f).$$

The measure λ is called $\mu * \nu$, the convolution of μ and ν . Notice that it follows using Fubini's theorem that

72

73

$$\begin{aligned} \int f d(\mu * \nu) &= \iint f(xy) \mu(dx) \nu(dy) \\ &= \iint f(xy) \nu(dy) \mu(dx), \end{aligned} \tag{2.1}$$

for each f in $C_c(S)$.

74

Actually, (2.1) leads to a more convenient formula for the convolution.

75

Proposition 2.2. Let $\mu, \nu \in P(S)$ and $B \in \mathcal{B}$. Then

76

$$\mu * \nu(B) = \int \mu(Bx^{-1}) \nu(dx) = \int \nu(x^{-1}B) \mu(dx). \tag{2.2}$$

Proof. For $B \in \mathcal{B}$, define

77

$$\lambda(B) = \int \mu(Bx^{-1}) \nu(dx). \tag{2.3}$$

Then λ is a probability measure on $\mathcal{B}$. We claim that λ is regular. To prove this claim, notice that for any open set $G \subset S$,

$$\begin{aligned} \lambda(G) &= \int \mu(Gx^{-1}) \nu(dx) = \int \left[\sup_{\substack{f \in C_c(S) \\ 0 \leq f \leq 1 \\ f=0 \text{ on } G^c}} \int f(sx) \mu(ds) \right] \nu(dx) \\ &\geq \sup_{\substack{f \in C_c(S) \\ 0 \leq f \leq 1 \\ f=0 \text{ on } G^c}} \iint f(sx) \mu(ds) \mu(dx) \\ &= \sup \int f d(\mu * \nu) \text{ (sup over the same class as before)} \\ &= \mu * \nu(G), \end{aligned}$$

since $\mu * \nu$ is regular.

Since S is a metric space, by [186, p. 27], λ is regular with respect to open sets from outside and closed sets from inside. It follows that $\lambda = \mu * \nu$, since

$$\begin{aligned} \lambda(B) &= \inf\{\lambda(G) : G \text{ open}, B \subset G\} \\ &\geq \inf\{\mu * \nu(G) : G \text{ open}, B \subset G\} \\ &= \mu * \nu(B) \text{ (}\forall \text{ Borel set } B\text{)}. \end{aligned} \quad \square$$

We remark that if $X_1, X_2, \dots, X_n$ are independent random variables with values in S such that the distribution of X_i is $\mu_i \in P(S)$, then the product $X_1, X_2, \dots, X_n$ has distribution $\mu_1 * \mu_2 * \dots * \mu_n$. (The convolution operation is associative.)

There is an important interplay between the convolution operation in $P(S)$ and a topology, called the weak topology, in $P(S)$. We present it next.

Definition 2.2. A sequence μ_n in $P(S)$ is said to converge weakly to μ in $P(S)$ if for every f in $C(S)$,

$$\lim_{n \rightarrow \infty} \int f d\mu_n = \int f d\mu.$$

To denote weak convergence, we often write:

$$\mu_n \xrightarrow{w} \mu \text{ or } \mu_n \Rightarrow \mu.$$

Let us denote by $B(S)$ the set of all (nonnegative) regular measures μ on $\mathcal{B}$ such that $\mu(S) \leq 1$.

Definition 2.3. A sequence μ_n in $B(S)$ is said to converge vaguely (or in the weak* topology) to μ in $B(S)$ if for every continuous function f on S with compact support

$$\lim_{n \rightarrow \infty} \int f d\mu_n = \int f d\mu.$$

To denote weak* convergence, we write:

97

$$\mu_n \xrightarrow{w^*} \mu.$$

Let us remark that $B(S)$ as a closed subset of the unit ball in the dual (adjoint) space of $C_c(S)$ is compact in the weak* topology (by the Banach–Alaoglu theorem of functional analysis, see [168, p. 63]). For a sequence μ_n in $P(S)$, if $\mu_n \xrightarrow{w^*} \mu$ and μ is in $P(S)$, then $\mu_n \xrightarrow{w} \mu$. $P(S)$ is compact in the weak topology. In fact, $P(S)$ is a compact metric space in its weak topology if and only if S is compact (note that S is already a metric space, see [186, p. 45]). In any case, the weak topology of $P(S)$ is metrizable (in our context).

We remark that $\mu_n \xrightarrow{w} \mu$ in $P(S)$ if and only if

$$\forall \text{ closed set } F \subset S, \mu(F) \geq \lim_{n \rightarrow \infty} \sup \mu_n(F)$$

if and only if

$$\forall \text{ open set } G \subset S, \mu(G) \leq \lim_{n \rightarrow \infty} \inf \mu_n(G).$$

(This can be verified easily, see [186, p. 40])

Proposition 2.3. *Convolution is jointly continuous in $P(S)$ with respect to the weak topology.*

Proof. Let $\mu_n \xrightarrow{w} \mu$ and $\sigma_n \xrightarrow{w} \sigma$. Let $f \in C(S)$ and $\epsilon > 0$. There exists a compact set K such that for $n \geq 1$,

$$\begin{aligned} \mu_n(K) &> 1 - \epsilon, & \mu(K) &> 1 - \epsilon \\ \sigma_n(K) &> 1 - \epsilon, & \sigma(K) &> 1 - \epsilon. \end{aligned} \quad (2.4)$$

Write: $g_n(t) \equiv \int_K f(st) \mu_n(ds)$. Then the sequence $\{g_n : n \geq 1\}$ is a uniformly bounded equicontinuous sequence; moreover, for each t in S ,

$$\lim_{n \rightarrow \infty} \sup \left| \int_K f(st) \mu_n(ds) - \int_K f(st) \mu(ds) \right| < 2\epsilon \|f\|. \quad (2.5)$$

Calling the second integral in (2.5) $g(t)$, we notice that if

$$\left| \int_K g_n(t) \sigma_n(dt) - \int_K g(t) \sigma_n(dt) \right| > 3\epsilon \|f\|$$

for infinitely many n , then there is a subsequence (n_i) such that

$$g_{n_i} \rightarrow g_0 \text{ uniformly on } K \quad (2.6)$$

and

116

$$\left| \int_K g_0 d\sigma_n - \int_K g d\sigma_n \right| > 3\epsilon \|f\|$$

for infinitely many n . But then (2.5) and (2.6) contradict the last inequality. It follows that eventually

117
118

$$\left| \int_K \int_K f(st) \mu_n(ds) \sigma_n(dt) - \int_K \int_K f(st) \mu(ds) \sigma(dt) \right| < 3\epsilon \|f\|.$$

Proposition 2.2 now follows from (2.4).

□ 119

Notice that if μ^n denotes the n th convolution power of μ , $\mu \in P(S)$ and if $\mu^n \xrightarrow{w} \nu$, $\nu \in P(S)$, then since convolution is jointly continuous (by Proposition 2.3),

120
121

$$\mu * \nu = \nu * \mu = \nu. \quad (2.7)$$

The next result characterizes all such weak limits when S is abelian.

122

First we need a definition. For $\mu \in P(S)$, we define $S(\mu)$, the *support* of μ , by

123

$$S(\mu) = \{x \in S : \mu(V) > 0 \text{ whenever } V \text{ is open and } x \in V\}.$$

Note that $S(\mu)$ is the smallest closed subset of S such that $\mu(S(\mu)) = 1$. Note that for μ, ν in $P(S)$,

124
125

$$S(\mu * \nu) = \overline{S(\mu) * S(\nu)}. \quad (2.8)$$

This is a basic formula, and it will be used often. It follows that if μ is an idempotent (i.e., $\mu = \mu * \mu$) probability measure in $P(S)$, then $S(\mu)$ is a closed subsemigroup of S .

126
127
128

Theorem 2.1. *Let S be abelian and μ, ν in $P(S)$ such that*

129

$$\mu * \nu = \nu. \quad (2.9)$$

Then $\nu \in P(S)$ satisfies (2.9) if and only if for ever $B \in \mathcal{B}$, $\nu(B) = \nu(By^{-1})$ for each $y \in S(\mu)$.

130
131

Proof. Let $B \in \mathcal{B}$ and let $f(x) = \nu(Bx^{-1})$. Then (2.9) implies that

132

$$f(x) = \int f(xy) \mu(dy). \quad (2.10)$$

Let (X_n) be a sequence of independent S -valued random variables, each with distribution μ . Let $x \in S$. Write: $S_n = xX_1X_2 \cdots X_n$.

133
134

If $\mathcal{F}_n$ is the σ -algebra generated by $X_1, X_2, \dots, X_n$, then $f(S_n)$ is a martingale with respect to $(\mathcal{F}_n)$. The reason is the following: For $A \in \mathcal{F}_n$,

$$\begin{aligned} E[I_A \cdot f(S_{n+1})] &= \int E[I_A f(S_n X_{n+1}) | X_{n+1} = y] \mu(dy) \\ &= \int \left[\int I_A \cdot f(S_n y) dP \right] \mu(dy) \\ &= \int \left[\int f(S_n y) \mu(dy) \right] \cdot I_A \cdot dP \\ &= \int f(S_n) I_{AdP} \text{ [using (2.10)],} \\ &= E[I_A \cdot f(S_n)], \end{aligned}$$

which means that $E[f(S_{n+1}) | \mathcal{F}_n] = f(S_n)$. Thus, $(f(S_n))$ is a bounded non-negative martingale, and therefore, $\lim_{n \rightarrow \infty} f(S_n) = Z(x)$ exists almost surely, and for $n \geq 1$,

$$E[Z(x) | \mathcal{F}_n] = f(S_n) \text{ almost surely.}$$

This means that

$$E[f(S_n)] = E[Z(x)] \equiv a, \text{ say.}$$

By the Hewitt–Savage zero-one law (see [38, p. 255]), for $n \geq 1$,

$$f(S_n) = E[Z(x) | \mathcal{F}_n] = a, \text{ almost surely.}$$

Then for $n = 1$, we have almost surely

$$f(xX_1) = a = E[f(xX_1)] = \int f(xy) \mu(dy) = f(x),$$

so that

$$f(x) = f(xy),$$

for μ -almost all y . This means that for $x \in S$, $y \in S(\mu)$, and any $B \in \mathcal{B}$,

$$\nu(Bx^{-1}) = \nu(By^{-1}x^{-1}),$$

which implies

$$\nu(B) = \int \nu(Bx^{-1}) \mu(dx) = \int \nu(By^{-1}x^{-1}) \mu(dx) = \nu(By^{-1}).$$

□ 146

Our next result considers the same problems in Theorem 2.1 when S is *not* assumed to be abelian. The method is necessarily different here, and also the result, which is the best possible in a sense, but not so complete as in Theorem 2.1.

Proposition 2.4. Let $\mu, \nu \in P(S)$ such that

150

$$\nu = \mu * \nu = \nu * \mu. \quad (2.11)$$

Then for $B \in \mathcal{B}$, $x \in S(\nu)$, $y \in S(\mu)$, we have

151

$$\begin{aligned} \nu(Bx^{-1}y^{-1}) &= \nu(Bx^{-1}) \\ \nu(y^{-1}x^{-1}B) &= \nu(x^{-1}B). \end{aligned} \quad (2.12)$$

Proof. Proposition 2.2 and hypothesis (2.11) imply that for any bounded Borel measurable function f on S , we have

152

153

$$\int f \, d\nu = \iint f(st)\nu(ds)\mu(dt) = \iint f(st)\mu(ds)\nu(dt). \quad (2.13)$$

Let $x \in S(\nu)$, and let K be any compact subset of S . Let $\nu(Kx^{-1}) = a$ and $\epsilon > 0$. Since ν_x is regular, there exist open sets G and W and a closed set C such that $K \subset W \subset C \subset G$ such that $\nu(Gx^{-1}) < a + \epsilon$. Notice that the set

154

155

156

$$A = \{s \in S : \nu(Cs^{-1}) < a + \epsilon\}$$

is open by Proposition 2.1, and contains x so that $\nu(A) > 0$. Define the function g on S by

157

158

$$g(s) = \max \{ \nu(Cs^{-1}) - a - \epsilon, 0 \}.$$

Then it follows from (2.13) that

159

$$\int \left[\int g(st)\mu(ds) - g(t) \right] \nu(dt) = 0. \quad (2.14)$$

Using (2.11), we have for $t \in S$,

160

$$\nu(Ct^{-1}) - a - \epsilon = \int [\nu(Ct^{-1}s^{-1}) - a - \epsilon] \mu(ds) \leq \int g(st)\mu(ds) \geq 0$$

so that for $t \in S$,

161

$$g(t) \leq \int g(st)\mu(ds). \quad (2.15)$$

We have from (2.14) and (2.15) that

162

$$g(t) = \int g(st)\mu(ds), \quad (\nu) \text{ almost everywhere.}$$

For $t \in A$, $g(t) = 0$ therefore $g(st) = 0$ for almost all s (with respect to μ), for almost all t (with respect to ν) in A . Since

163

164

$$g(st) = 0 \Rightarrow \nu(Wt^{-1}s^{-1}) \leq a + \epsilon,$$

and, since W (as well as Wt^{-1} , for each t) is an open set, functions $s \rightarrow \nu$ 165
 $(Wt^{-1}s^{-1})$ and $t \rightarrow \nu_s(Wt^{-1})$ are both lower semicontinuous, it follows by 166
 Proposition 2.1 that for all t in A and all s in $S(\mu)$, we must have 167

$$\nu(Wt^{-1}s^{-1}) \leq a + \epsilon.$$

Since $x \in A$ and $K \subset W$,

168

$$\nu(Kx^{-1}s^{-1}) \leq \nu(Kx^{-1})$$

for all $s \in S(\mu)$. Since by (2.11),

169

$$\int [\nu(Kx^{-1}) - \nu(Kx^{-1}s^{-1})] \mu(ds) = 0,$$

it follows that for almost all s (with respect to μ),

170

$$\nu(Kx^{-1}) = \nu(Kx^{-1}s^{-1}).$$

From the upper semicontinuity of the function $s \rightarrow \nu(Kx^{-1}s^{-1})$, we then have 171

$$\nu(Kx^{-1}) = \nu(Kx^{-1}s^{-1})$$

whenever $x \in S(\nu)$ and $s \in S(\mu)$. Since ν is regular, $\nu(Bx^{-1}) = \nu(Bx^{-1}s^{-1})$ for 172
 $x \in S(\nu)$, $s \in S(\mu)$ and $B \in \mathcal{B}$. The second result in (2.12) follows similarly. $\square$ 173

Let us remark that the equalities in (2.12) are not sufficient to ensure (2.11) al- 174
 ways. If S is a special semigroup such that 175

- (i) S contains a zero 0; 176
- (ii) $xy = 0$ for all x, y in S , then, for all x, y in S , 177

$$By^{-1}x^{-1} = By^{-1} = \emptyset, \text{ if } 0 \notin B$$

and

178

$$By^{-1}x^{-1} = By^{-1} = S, \text{ if } 0 \in B.$$

The same is true for $x^{-1}y^{-1}B$ and $y^{-1}B$. Thus, (2.12) holds, whereas

179

$$\nu * \mu = \delta_0 \text{ (the unit mass at 0)} \quad \forall \nu, \mu \in P(S).$$

In Sect. 2.3, we show that the result of Theorem 2.1 continues to hold when S is a 180
 group (not necessarily abelian) in the following sense: For μ, ν in $P(S)$, $\nu = \mu * \nu$ 181
 if and only if $\nu = {}_y\nu$ for each y in the closed subgroup generated by $S(\mu)$. 182

Let us also remark that the result $\nu(By^{-1}) = \nu(B) = \nu(y^{-1}B)$ for $y \in S(\mu)$, 183
 instead of (2.12), cannot hold in general in Proposition 2.5, unless S is special. 184
 This is clear from our results on idempotent probability measures (i.e., those μ in 185

$P(S)$ for which $\mu = \mu * \mu$), which we present next. Note that when the weak limit of the convolution iterates (μ^n) of a probability measure μ in $P(S)$ exists, it is idempotent. There is an important connection between invariant and idempotent probability measures that are exploited in our characterization of idempotent probability measures.

Definition 2.4. A (nonnegative) regular measure μ , not necessarily finite, on the Borel sets of S is called r^* -invariant (ℓ^* -invariant) if for $B \in \mathcal{B}$ and $x \in S$, $\mu(Bx^{-1}) = \mu(B)$ [$\mu(x^{-1}B) = \mu(B)$].

Notice that for a (nonnegative) regular measure μ , $\mu(S - S(\mu)) = 0$ and $S(\mu)$ is a closed subset of S .

If μ is r^* -invariant on S , then for any $x \in S$, $\mu_x = \mu$ so that $S(\mu) = S(\mu_x) = S(\mu) \cdot x$. Similarly, if μ is ℓ^* -invariant, then for any $x \in S$, $S(\mu) = S(\mu)_x = xS(\mu)$.

Lemma 2.1. If μ is an r^* -invariant (ℓ^* -invariant) measure on S , then for $x \in S(\mu)$, $xS(\mu)(S(\mu) \cdot x)$ is left-cancellative (right-cancellative).

Proof. Let μ be a r^* -invariant measure on S . Then for any $s \in S(\mu)$, $\overline{S(\mu)s} = S(\mu)$. Let x, y, z , and w be in $S(\mu)$ such that $(xy)(xz) = (xy)(xw)$.

Now $\overline{S(\mu) \cdot (xyx)} = S(\mu)$ so there exist (t_n) in $S(\mu)$ such that

$$t_n \cdot (xyx) \rightarrow x, \quad n \rightarrow \infty$$

This means that

$$\begin{aligned} t_n(xy x)z &\rightarrow xz \\ \parallel \\ t_n(xy x)w &\rightarrow xw \end{aligned}$$

so $xz = xw$. Thus, $xS(\mu)$ is left-cancellative. Similarly, when μ is ℓ^* -invariant, $S(\mu)x$ is right-cancellative. $\square$

Proposition 2.5. Let μ be a measure on S such that it is both r^* - and ℓ^* -invariant. Then μ is a unimodular Haar measure on its support, which is a locally compact topological group.

Proof. By Lemma 2.1, for $x, y \in S(\mu)$, $x \cdot S(\mu) \cdot y$ is bicancellative since $xS(\mu)y \subset [xS(\mu)] \cap [S(\mu)y]$. We prove Proposition 2.5 by showing that $xS(\mu)y$ contains an idempotent element e , because then for any x, y in $S(\mu)$, by the remark preceding Lemma 2.1,

$$S(\mu) = \overline{S(\mu)e} = S(\mu)e \subset S(\mu)y \subset S(\mu),$$

and

$$S(\mu) = \overline{eS(\mu)} = eS(\mu) \subset xS(\mu) \subset \overline{xS(\mu)} = S(\mu).$$

Thus, $S(\mu)$ is both left and right simple, and, therefore, a group. It follows that it is a topological group (see Theorem 1.4, Chap. 1), since it is locally compact.

Let us now prove that for $x, y \in S(\mu)$, $xS(\mu)y$ contains an idempotent. Let K 217
be a compact subset of $S(\mu)$ such that $\mu(K) > 0$. Let L be a compact subset of 218
 $S(\mu)$ so that 219

$$\mu(x^{-1}(xKy) \cap L) > 0.$$

Note that $x^{-1}(xKy)y^{-1} \supset K$ and thus, $\mu(x^{-1}(xKy)) \geq \mu(K) > 0$. 220

Write $A = LC \cup L$, $C \equiv xKy$. Define the finite regular measure λ by 221

$$\lambda(B) = \mu(x^{-1}B \cap A), \quad B \subset S(\mu).$$

Define the maps θ and β by 222

$$\begin{aligned} \theta(s, t) &= (s, ts) \\ \beta(s, t) &= (t, s), \quad s, t \in S(\mu). \end{aligned}$$

Using Fubini's theorem, we have 223

$$\begin{aligned} \lambda \times \lambda(\theta(C \times C)) &= \int_C \lambda(Cs) \lambda(ds) \\ &\geq \int_C \mu(x^{-1}(xKys) \cap Ls) \lambda(ds) \\ &\geq \int_C \mu([x^{-1}(xKys) \cap Ls]s^{-1}) \lambda(ds) \\ &\geq \int_C \mu(x^{-1}(xKy) \cap L) \lambda(ds) > 0. \end{aligned}$$

Now, let E be any compact subset of the set 224

$$F \equiv D \times D - \theta(D \times D), \quad D \equiv xS(\mu)y.$$

Then for any $s \in D$, 225

$$E_s = \{t : (s, t) \in E\} \subset D - Ds$$

so that $E_s s^{-1} \cap D = \emptyset$. Therefore, $\lambda(E_s) \leq \mu(x^{-1}E_s) = \mu(x^{-1}E_s s^{-1} y^{-1}) = 0$, 226
since $S(\mu) \cap [x^{-1}E_s s^{-1} y^{-1}] = \emptyset$. Thus $\lambda \times \lambda(E) = 0$ whenever E is a compact 227
subset of F . Since $\lambda \times \lambda(\beta\theta(C \times C)) > 0$ (β being measure preserving) and since 228
 $\beta\theta(C \times C)$ is contained in $D \times D$, it follows that 229

$$\beta\theta(C \times C) \cap \theta(D \times D) \neq \emptyset.$$

This means that there exist u, v, w, z in D such that 230

$$(z, wz) = (vu, u)$$

so that $z = vu = vwz$ or $vwz = (vw)^2 z$. Since D is cancellative, $vw = (vw)^2$ is an 231
idempotent element in D . $\square$ 232

Our next result is a basic one needed for much of the limit theory in this book. This result characterizes the structure of all idempotent probability measures on S , and its proof here depends on Propositions 2.4 and 2.5.

Theorem 2.2. Assume $\mu \in P(S)$ and $\mu = \mu * \mu$. Then $S(\mu)$ is a closed completely simple subsemigroup of S . Furthermore, the structure of μ can be described as follows:

Let $e = e^2 \in S(\mu)$ and let $E(S(\mu)e) \times eS(\mu)e \times E(eS(\mu))$ be the usual product representation of $S(\mu)$ (see Theorem 1.1, Chap. 1). There exists $\mu_1 \in P(X)$, $\mu_3 \in P(Z)$, where $X = E(S(\mu)e)$, $Z = E(eS(\mu))$ such that $\mu = \mu_1 * \mu_2 * \mu_3$, where μ_2 is the Haar probability measure on $eS(\mu)e$, which is a compact group. The measure μ can also be regarded as the product measure $\mu_1 \times \mu_2 \times \mu_3$. Conversely, if μ_1 and μ_3 are in $P(S)$ and μ_2 is the Haar probability measure on some compact subgroup $G \subset S$ such that $S(\mu_3)S(\mu_1) \subset G$, then $\mu_1 * \mu_2 * \mu_3$ is an idempotent probability measure on S . Here, as usual, $E(A)$ is the set of all idempotent elements in the set A .

Proof. It follows from (2.8) and Proposition 2.4 that

$$\left. \begin{aligned} \overline{S(\mu)S(\mu)} &= S(\mu); \\ \mu(Bx^{-1}y^{-1}) &= \mu(Bx^{-1}), \\ \mu(y^{-1}x^{-1}B) &= \mu(x^{-1}B), \end{aligned} \right\} \quad (2.16) \quad (2.17)$$

for $x, y \in S(\mu)$, $B \in \mathcal{B}$.

Consider the probability measure μ_x for $x \in S(\mu)$, given by $\mu_x(B) = \mu(Bx^{-1})$. Notice that for any $z \in S(\mu)$,

$$\mu_x(B(zx)^{-1}) = \mu(Bx^{-1}z^{-1}x^{-1}) = \mu(Bx^{-1}(xz)^{-1}) = \mu(Bx^{-1}) = \mu_x(B)$$

by (2.17) whenever $B \in \mathcal{B}$. Also if $z_n \in S(\mu)$ and $z_n x \rightarrow z$ as $n \rightarrow \infty$, then for any compact subset $K \subset B$, using Proposition 2.1, we have:

$$\mu_x(Kz^{-1}) \geq \lim_{n \rightarrow \infty} \sup \mu_x(K(z_n x)^{-1}) = \mu_x(K)$$

and for any open subset $V \supset B$,

$$\mu_x(Vz^{-1}) \leq \lim_{n \rightarrow \infty} \int \mu_x(V(z_n x)^{-1}) = \mu_x(V).$$

Thus, it follows easily (using the regularity of μ) that for $B \in \mathcal{B}$,

$$\mu_x(Bz^{-1}) = \mu_x(B) \quad (2.18)$$

for all z in $S(\mu_x) = \overline{S(\mu) \cdot x}$.

Similarly, for any $x \in S(\mu)$,

257

$${}_x\mu(z^{-1}B) = {}_x\mu(B) \quad (2.19)$$

whenever $z \in S({}_x\mu) = \overline{{}_xS(\mu)}$ and $B \in \mathcal{B}$.

258

Let us now consider for $x \in S(\mu)$, the regular probability measure λ given by

259

$$\lambda(B) = \mu(x^{-1}Bx^{-1}), \quad x \in S(\mu).$$

Its support is $\overline{{}_xS(\mu)x}$. It is clear from the preceding considerations that μ is both ℓ^* and r^* -invariant on its support. It follows from Proposition 2.7 that $\overline{{}_xS(\mu)x}$ is a topological group; therefore, it must be compact, since it supports that Haar probability measure λ . (See [168, Problem 6.7.2].)

260

261

262

263

Now we claim that for any $x \in S(\mu)$,

264

$$\begin{aligned} S(\mu)x & \text{ is a left group,} \\ xS(\mu) & \text{ is a right group.} \end{aligned} \quad (2.20)$$

To prove (2.20), write $H = \overline{{}_xS(\mu)x}$; then by (2.18), for $z \in H$,

265

$$\overline{Hz} = S(\mu_{xz}) = S(\mu_x) = H. \quad (2.21)$$

Now for $z \in H$, $H \cap \overline{{}_xS(\mu)x}$ is a nonempty left ideal of the group $\overline{{}_xS(\mu)x}$, which implies that $\overline{{}_xS(\mu)x} \subset Hz$. This means that the identity e of $\overline{{}_xS(\mu)x}$ belongs to Hz , so that by (2.21)

266

267

268

$$H \supset Hz \supset H \cdot e = \overline{H} \cdot e = H. \quad (2.22)$$

It follows from (2.21) and (2.22) that H is a left group, establishing that $S(\mu)x$ is a left group in (2.20). The other part in (2.20) is verified similarly.

269

270

Because of (2.20), it follows (by Proposition 1.6 and Theorem 1.7 in Chap. 1) that because $S(\mu)$ has a completely simple kernel that is the union of all minimal left ideals; that is, $\cup\{S(\mu)x : x \in S(\mu)\}$, which is dense in $S(\mu)$ by (2.16). Since the kernel is known to be closed (see Theorem 1.7), it follows that $S(\mu)$ is a completely simple subsemigroup of S .

271

272

273

274

275

For the second part of Theorem 2.2, let $e = e^2 \in S(\mu)$. Then, $\delta_e * \mu * \delta_e$ is the Haar probability measure of the compact group $eS(\mu)e$, by remarks preceding (2.20). Moreover, notice that by (2.17), we have:

276

277

278

$$\begin{aligned} \mu * \delta_e * \mu(B) &= \int \mu * \delta_e(Bx^{-1}) \mu(dx) = \int \mu(Bx^{-1}e^{-1}) \mu(dx) \\ &= \int \mu(Bx^{-1}) \mu(dx) = \mu^2(B) = \mu(B), \end{aligned}$$

so that

$$\mu = \mu * \delta_e * \mu. \quad (2.23)$$

Also notice that by (2.23),

$$\mu = (\mu * \delta_e) * (\delta_e * \mu).$$

By (2.18) and (2.19), $\mu_e = \mu * \delta_e$ is r^* -invariant on its support $S(\mu) \cdot e$, which is a left group by (2.20), and ${}_e\mu = \delta_e * \mu$ is ℓ^* -invariant on its support $e \cdot S(\mu)$, which is a right group by (2.20).

Since $S(\mu)e$ is a left group, $S(\mu) \cdot e$ is topologically isomorphic to $E(S(\mu)e) \times eS(\mu)e$ via the mapping $\Phi : (x, y) \rightarrow x \cdot y$. We define the probability measure μ_1 on $E(S(\mu)e)$ by

$$\mu_1(A) = \mu_e(\Phi(A \times eS_1(\mu)e))$$

and the probability measure μ_2 on $eS(\mu)e$ by

$$\mu_2(B) = \mu_e(\Phi(E(S(\mu)e) \times B)).$$

It is easily verified that μ_2 is r^* invariant on $eS(\mu)e$, since μ_e is r^* -invariant; therefore, μ_2 is the Haar measure (and both r^* - and ℓ^* -invariant) on the compact group $eS(\mu)e$. Also, it is verified easily that

$$\mu_1 * \mu_2(\Phi(A \times B)) = \mu_1(A)\mu_2(B), \quad (2.24)$$

where $A \subset E(S(\mu)e)$, $B \subset eS(\mu)e$. Now notice that by (2.23),

$$\mu_e = \mu_e * \mu_e.$$

Therefore, if A and B are as in (2.24),

$$\begin{aligned} \mu_e(\Phi(A \times B)) &= \mu_e * \mu_e(\Phi(A \times B)) = \int \mu_e(x^{-1}\Phi(A \times B)) \mu_e(dx) \\ &= \int_{\Phi(A \times eS(\mu)e)} \mu_e(\Phi(E(S(\mu)e) \times g^{-1}B)) \mu_e(d\Phi(f, g)) \\ &= \mu_1(A)\mu_2(B). \end{aligned} \quad (2.25)$$

Notice that (2.24) and (2.25) show that both μ_e and $\mu_1 * \mu_2$ induce, via the map Φ , the same probability measure $\mu_1 \times \mu_2$ on $E(S(\mu)e) \times eS(\mu)e$. It follows that

$$\mu_e = \mu_1 * \mu_2. \quad (2.26)$$

Similarly,

$${}_e\mu = \mu_2 * \mu_3, \quad (2.27)$$

where μ_3 is the probability measure on $E(eS(\mu))$ defined by

297

$$\mu_3(C) = {}_e\mu(\Phi(eS(\mu)e \times C)).$$

Theorem 2.2 now follows from (2.26) and (2.27), since

298

$$\mu = \mu_e * {}_e\mu = (\mu_1 * \mu_2) * (\mu_2 * \mu_3) = \mu_1 * \mu_2 * \mu_3.$$

□ 299

Theorem 2.3. Suppose that $\mu \in P(S)$ and μ is r^* -invariant. Then $S(\mu)$ is a left 300
group and consequently topologically isomorphic to $E(S(\mu)) \times eS(\mu)$, under the 301
map $\Phi(x, y) = xy$, where e is a fixed idempotent in $S(\mu)$ and $eS(\mu)$ is a compact 302
group. Moreover if μ_1 is the probability measure on $E(S(\mu))$ defined by 303

$$\mu_1(A) = \mu(\Phi(A \times eS(\mu))),$$

then $\mu = \mu_1 \times \mu_2$, where μ_2 is the Haar probability measure on $eS(\mu)$.

304

Proof. Notice that any r^* -invariant probability measure on S is idempotent. Since 305
for $x \in S(\mu)$, by (2.20), $S(\mu) \cdot x$ is a left group, then for an idempotent e in $S(\mu)x$, 306
we have, by comments preceding Lemma 2.1, 307

$$S(\mu) = \overline{S(\mu)e} = S(\mu)e \subset S(\mu)x \subset \overline{S(\mu)x} = S(\mu)$$

So that $S(\mu)$ is a left group. The rest of Theorem 2.3 follows from Theorem 2.2. □ 308

Note that whenever μ is r^* -invariant and consequently, $S(\mu)$ is a left group by 309
Theorem 2.3, then since $S(\mu)$ is right-cancellative, for $B \subset S(\mu)$ and $x \in S(\mu)$, 310

$$Bxx^{-1} \cap S(\mu) = B$$

so that whenever B is a compact subset of $S(\mu)$,

311

$$\mu(Bx) = \mu(Bxx^{-1}) = \mu(Bxx^{-1} \cap S(\mu)) = \mu(B).$$

This property of μ is called right-invariance. Thus, every r^* -invariant probability 312
measure on S is right-invariant when restricted to its support. Conversely, if μ is a 313
right-invariant probability measure in $P(S)$, then μ is r^* -invariant. The reason is 314
the following: for $B \in \mathcal{B}$, $x \in S$, any compact set $A \subset Bx^{-1}$, we have $\mu(A) =$ 315
 $\mu(Ax) \leq \mu(B)$, since $Ax \subset (Bx^{-1})x \subset B$. This means that $\mu(Bx^{-1}) \leq \mu(B)$ 316
for any $B \in \mathcal{B}$, and therefore $1 - \mu(Bx^{-1}) = \mu(B^c x^{-1}) \leq \mu(B^c) = 1 - \mu(B)$, 317
so that μ is r^* -invariant. This right-invariance property helps us characterize all 318
 r^* -invariant (possibly infinite) measures on S . 319

Theorem 2.4. Let μ be a regular r^* -invariant measure (not necessarily finite) on 320
 S . Then $S(\mu)$ is a left group if and only if μ is right-invariant on its support. Fur- 321
thermore, if $S(\mu)$ is a left group and $S(\mu)$ has the usual left group representation 322

$E \times G$, where $E \equiv E(S(\mu))$ and $G \equiv eS(\mu)$, G a group, then μ is of the form $\mu_1 \times \mu_2$ where $\mu_1 \in P(E)$ and μ_2 is a right Haar measure on G . Conversely, any measure of the form $\mu_1 \times \mu_2$ (as here) on a left group $E \times G$ is both r^* - and right-invariant.

Proof. Suppose μ is an r^* -invariant, right-invariant regular measure on S . Note that since μ is r^* -invariant, $\overline{S(\mu)x} = S(\mu)$ for $x \in S$. Consider the map π from $S(\mu)$ to L , the space of bounded linear operators on $L_2(S(\mu), \mu)$ of norm 1, defined by

$$\pi(x)[f](s) = f(sx).$$

Note that $\|f\|_2 = \|\pi(x)[f]\|_2$ since μ is r^* -invariant. Thus, $\pi(x)$ is an isometric linear operator on L_2 . We show that it is unitary. It suffices to show that $\pi(x)$ is surjective. (See [168, Proposition 7.17, p. 153].) To this end, let $f \in L_2$. Then there exist simple functions f_n , where each f_n is a linear combination of indicator functions of compact subsets of S such that

$$\|f_n - f\|_2 < \frac{1}{n}.$$

If $f_n = \sum c_i I_{K_i}$, then let g_n be defined by $g_n = \sum c_i I_{K_i x}$. It follows from the right and r^* -invariance of μ that

$$\|I_K(s) - I_{Kx}(sx)\|_2 = \mu(Kxx^{-1} - K) = 0,$$

since $K \subset Kxx^{-1}$ and $\mu(K) = \mu(Kx) = \mu(Kxx^{-1})$. Thus, the sequence $(g_n(sx))$ is Cauchy in L_2 . By the r^* -invariance of μ ,

$$\|g_n(sx) - g_m(sx)\|_2 = \|g_n(s) - g_m(s)\|_2$$

so that (g_n) is Cauchy. Then there exists g in L_2 such that

$$\lim_{n \rightarrow \infty} \|g_n - g\|_2 = 0.$$

Since $\|g_n(sx) - f_n(s)\|_2 = 0$, it follows that $g(sx) = f(s)$ in L_2 ; then $\pi(x)[g] = f$. This proves that $\pi(x)$ is surjective.

Next we claim that for each x in $S(\mu)$, $xS(\mu)$ is bicancellative. This is established once we show that π is an anti-isomorphism from $xS(\mu)$ into the group of unitary operators on L_2 . All that is left to be shown to this end is that π is 1-1 on $xS(\mu)$. If not, then there exists $s_1, s_2 \in S(\mu)$ such that $\pi(xs_1) = \pi(xs_2)$; but $xs_1 \neq xs_2$. This means that there exists at least one $y \in S(\mu)$ such that $yx s_1 \neq yx s_2$, since $\overline{S(\mu)x} = S(\mu)$, where μ is r^* -invariant. But then there is a continuous bounded function f on $S(\mu)$ such that

$$\{s \in S(\mu) \mid \|f(sx s_1) - f(sx s_2)\| > 0\}$$

is an open set with positive μ measure. Then $\pi(xs_1)[f] \neq \pi(xs_2)[f]$, which is a contradiction.

Thus, for each $x \in S(\mu)$, since $xS(\mu)$ is anti-isomorphic to a subsemigroup of a group, it is bicancellative. Therefore, as in the proof of Proposition 2.5, we can also show that for any $x, y \in S(\mu)$, $xS(\mu)y$ contains an idempotent. (In this part of the proof of Proposition 2.5, we did not use the ℓ^* -invariance of μ .) Thus, if $e = e^2 \in xS(\mu)y \subset S(\mu)y$,

$$S(\mu) = \overline{S(\mu) \cdot e} = S(\mu)e \subset S(\mu)y \subset \overline{S(\mu)y} = S(\mu),$$

so that for any $y \in S(\mu)$,

$$S(\mu)y = S(\mu).$$

This means that $S(\mu)$ is a left group.

For the last part of Theorem 2.4, let us assume that $S(\mu)$ is a left group with its usual product representation $E \times G$, so that

$$\Phi : E \times G \rightarrow S(\mu),$$

where $\Phi(x, y) = xy$ is a topological isomorphism and E, G inherit the topology from $S(\mu)$. Here, $G = e_0S(\mu)$, a group, and $E \equiv E(S(\mu))$, where e_0 is a fixed idempotent in $S(\mu)$. Let h be a continuous function on E with compact support such that $h \geq 0$. For f , a continuous function with compact support on G , we define

$$I(f) = \int h(e)f(g)\mu_0(d(e, g)),$$

where $\mu_0(A) = \mu(\Phi(A))$, $A \subset E \times G$. Notice that if $f_y(g) = f(gy)$ for g and y in G , then

$$I(f_y) = \int h(e)f_y(g)\mu_0(d(e, g)) = I(f),$$

since μ_0 is r^* -invariant and $h(e)f_y(g)$ is the right translate of the function $(e, g) \rightarrow h(e)f(g)$ by the element (e, y) , $e \in E$. Since I is a right-invariant positive linear functional on $C_c(G)$, there exists a right Haar measure μ_h on G such that for $f \in C_c(G)$,

$$\int f(g)\mu_h(dg) = \int h(e)f(g)\mu(d(e, g)).$$

Thus, given a right Haar measure μ_2 on G , there exists a positive number $\alpha(h)$ such that

$$\alpha(h) \int f(g)\mu_2(dg) = \int h(e)f(g)\mu(d(e, g)).$$

Clearly, α (with $\alpha(h) = \alpha(h^+) - \alpha(h^-)$) is a positive linear functional on $C_c(E)$ so that there is a nonnegative regular measure μ_1 on E such that for $h \in C_c(E)$,

$$\alpha(h) = \int h(e)\mu_1(de)$$

so that for $h \in C_c(E)$, $f \in C_c(G)$,

374

$$\int h(e) f(g) \mu(d(e, g)) = \iint h(e) f(g) \mu_1 \times \mu_2(d(e, g)).$$

The rest of Theorem 2.4 follows easily by an application of the Stone–Weierstrass theorem (see [167, p. 78]). $\square$

375

376

Before we close Sect. 2.2, we discuss two related questions, one on (infinite) regular measures μ satisfying

377

378

$$\mu(B) = \int \mu(Bx^{-1}) \mu(dx) \quad (2.28)$$

for all Borel sets B (for which the integral on the right makes sense), and the other on the existence of right-invariant (i.e., $\mu(Kx) = \mu(K)$ for compact $K \subset S$, $x \in S$) measures on S .

379

380

381

Notice that (2.28) implies that

382

$$\mu(S) = \mu(S) \cdot \mu(S) \quad (2.29)$$

since for $x \in S$, $Sx^{-1} = S$, so that a measure satisfying (2.28) must be, if not a zero measure, either an idempotent probability measure or an infinite measure. When S is a group, it has been proven in [148] that there cannot exist an infinite regular measure μ satisfying (2.28). The proof is quite involved (and works in contexts much more general than groups), and thus is not presented here. However, we point out that the proof is based on the following observation: If μ is a solution of (2.28), then for every x in $S(\mu)$, there are continuous homomorphisms β_x and γ_x from $S(\mu)$ into the positive reals such that

383

384

385

386

387

388

389

390

$$\mu(Bx^{-1}y^{-1}) = \beta_x(y) \mu(Bx^{-1})$$

and

391

$$\mu(y^{-1}x^{-1}B) = \gamma_x(y) \mu(x^{-1}B),$$

for every Borel set $B \in \mathcal{B}$ and x, y in $S(\mu)$. Furthermore, there cannot exist an infinite regular measure μ on S when S is a group such that

392

393

$$\mu(By^{-1}) = \beta(y) \mu(B),$$

for $B \in \mathcal{B}$ and $y \in S(\mu)$, where β is a continuous homomorphism from S into the positive reals such that $\int \beta d\mu = 1$.

394

395

Finally, we consider the question of the existence of right-invariant measures on S . Recall that μ is right-invariant if μ is regular, and for compact $K \subset S$ and $x \in S$, $\mu(Kx) = \mu(K)$. If μ is finite and right-invariant, then μ is r^* -invariant (as was noted earlier); therefore, finite right-invariant measures are completely known and fully characterized earlier. The problem is then with those that are infinite. How do they arise?

396

397

398

399

400

401

Suppose that right translations are open in S ; that is, the maps $x \rightarrow x \cdot y$ are open maps for each y in S . Then if μ is a right-invariant measure on S and there are x, y, z in $S(\mu)$ such that

$$xy = zy, \quad x \neq z,$$

then there are disjoint open sets $V(x), V(z)$ containing x and z , respectively, such that

$$\begin{aligned} \mu(V(x) \cup V(z)) &= \mu(V(x)y \cup V(z)y) \\ &= \mu(V(x)y) + \mu(V(z)y) - \mu(V(x)y \cap V(z)y) \\ &< \mu(V(x)) + \mu(V(z)) \end{aligned}$$

since μ is right-invariant (so that $S(\mu)$ is a subsemigroup of S), $V(x)y \cap V(z)y$ is nonempty open and it intersects $S(\mu)$. This means that right-invariance and open right translations imply that $S(\mu)$ is right-cancellative. Theorem 2.5 is given in [174].

Theorem 2.5. *Let S be right reversible (that is, $Sx \cap Sy \neq \emptyset$ for every $x, y \in S$). (This property holds for example when S is abelian.) Suppose S has translations open. Then S can support a left- and right-invariant measure μ if and only if S can be topologically embedded (via a topological isomorphism $\alpha : S \rightarrow \alpha(S) \subset G$) as an open subspace in a locally compact Hausdorff topological group G and μ on S is the restriction of a unimodular Haar measure on G .*

Note that in Theorem 2.5, to embed S topologically as an open subspace in a topological group, it is necessary for translation maps in S to be open.

It is clear that Theorem 2.5 gives only a partial solution to the problem of characterizing all right-invariant measures on S . The reader can consult [3] for other (partial) results in this context. The problem in the case when S is abelian has a more satisfactory solution (see Theorem 2.6) given in [187].

Theorem 2.6. *Let S be abelian. If μ is right-invariant on S , then there is a locally compact abelian group G and a continuous homomorphism $q : S \rightarrow G$ such that $q(S)$ has a nonempty interior and that there is a Haar measure m on G such that for compact $K \subset S$, $\mu(K) = m(q(K))$. The map q is injective if and only if S is cancellative; it is a homeomorphism onto $q(S)$ if and only if S has the following property:*

$$s_n \rightarrow s, \quad s_n t_n \rightarrow st \Rightarrow t_n \rightarrow t.$$

Section 2.2 Exercises

Below, S is always a locally compact Hausdorff second countable topological semigroup (unless otherwise mentioned).

Exercise 2.1. Let $\mu \in P(S)$ and $\mathcal{F} = \{B \subset S : B \text{ is a Borel set such that}$ 432
given $\epsilon > 0$, there is an open set u and a closed set V such that $V \subset B \subset U$ and 433
 $\mu(U - V) < \epsilon\}$. Show that $\mathcal{F}$ is the class of all Borel subsets of S . (Hint: $\mathcal{F}$ is a 434
 σ -algebra and it contains all closed subsets, since in a metric space, a closed subset 435
is the countable intersection of decreasing open subsets.) 436

Exercise 2.2. Show that for $\mu \in P(S)$, given any Borel subset $B \subset S$ and any 437
 $\epsilon > 0$, there exist an open set U and a compact subset K such that $K \subset B \subset U$ and 438
 $\mu(U - K) < \epsilon$. 439

Exercise 2.3. Show that for $\mu \in P(S)$, the set $S(\mu)$, defined by 440

$$S(\mu) = \{x \in S : \mu(V) > 0 \text{ for any open set } V \text{ containing } x\},$$

is a nonempty closed subset of S and $\mu(S(\mu)) = 1$. This set $S(\mu)$ is the support 441
of μ . 442

Exercise 2.4. Let μ_1 and μ_2 be in $P(S)$ such that $\mu_1(B) \geq \mu_2(B)$ for any Borel 443
set $B \subset S$. Show that $\mu_1 = \mu_2$. 444

Exercise 2.5. Show that in $P(S)$, μ_n converges to μ weakly iff for any open subset 445
 $G \subset S$, $\mu(G) \leq \liminf_{n \rightarrow \infty} \mu_n(G)$. 446

Exercise 2.6. Let $\mu_n, n \geq 1$, be in $P(S)$ and $\mu \in B(S)$ such that μ_n converges to 447
 μ in the weak*-topology. Show that μ_n converges to μ weakly iff $\mu \in P(S)$. 448

Exercise 2.7. Explain how the inequalities in (2.4) follow under the assumptions 449
made there. 450

Exercise 2.8. Prove the formula (2.8). 451

Exercise 2.9. Prove (2.13) assuming (2.11). 452

Exercise 2.10. Let μ be an infinite nonnegative regular measure (definition given at 453
the end of Sect. 2.1) on the Borel subsets of S . Show that $\mu(S - S(\mu)) = 0$, when 454
 $S(\mu)$ is as defined in Exercise 2.3. 455

Exercise 2.11. Let $\mu \in P(S)$ or μ be as given in Exercise 2.10. For $x \in S$, define 456
 $\mu_x(B) = \mu(Bx^{-1})$. Show that μ_x is a regular measure and $S(\mu_x) = \overline{S(\mu) \cdot x}$. 457

Exercise 2.12. Let $\mu \in P(S)$ and V be an open subset of S . Show that if $x_n \in S$ 458
and $x_n \rightarrow x$, then 459

$$\mu(Vx^{-1}) \leq \liminf \mu(Vx_n^{-1}).$$

Give the corresponding inequality when V is a closed subset of S . 460

Exercise 2.13. Let $X \times G \times Y, X \cup G \cup Y \subset S$, be the usual product representation 461
of a completely simple semigroup S . Here, G is a compact group and $Y \cdot X \subset G$. 462
Let $\mu_1 \in P(X)$, $\mu_3 \in P(Y)$ and μ_2 be the Haar probability measure on G . Show 463
that $\mu_1 * \mu_2 * \mu_3$ is an idempotent probability on S . 464

Exercise 2.14. Let μ be an r^* -invariant probability measure on the Borel subsets of S . Show that μ is right invariant (that is, for compact $K \subset S(\mu)$ and $x \in S(\mu)$, $\mu(Kx) = \mu(K)$).

Exercise 2.15. Let $\mu \in P(S)$ be such that for any compact $K \subset S$ and $x \in S$, $\mu(Kx) \geq \mu(K)$. Show that $S(\mu)$ is a left group. (Hint: Show that μ is r^* -invariant.)

2.3 Weak Convergence of Convolution Products of Probability Measures

As before, S is a locally compact Hausdorff second countable topological semi-group and $C_c(S)$ is the vector space of real continuous functions on S with compact support. It is then wellknown (from Banach–Alaoglu’s theorem in functional analysis) that the unit ball in the dual of $C_c(S)$ is weak* compact. Thus, the set

$$B(S) \equiv \{\mu \mid \mu \text{ is a nonnegative regular Borel measure on } S \text{ with } \mu(S) \leq 1\}$$

is compact in the weak* topology. Recall: A net (μ_α) in $B(S)$, w^* converges to μ in $B(S)$ if and only if for every f in $C_c(S)$, $\int f d\mu_\alpha \rightarrow \int f d\mu$. However, $P(S) \equiv \{\mu \in B(S) : \mu(S) = 1\}$ need not be weak* compact, unless S is compact. Note that in $P(S)$, weak* compactness is equivalent to weak compactness, and thus $P(S)$ is weak* compact if and only if S is compact. (See [186, Theorem 6.4, p. 45].) For a subset $\Gamma \subset P(S)$, the weak* closure of Γ in $P(S)$ is weak* compact, if Γ is tight; that is, given $\epsilon > 0$, there is a compact subset $K_\epsilon \subset S$ such that

$$\mu \in \Gamma \Rightarrow \mu(K_\epsilon) > 1 - \epsilon. \quad (2.30)$$

The reason for this is obvious since $\mu \in w^*$ -closure of Γ and Γ is tight only if $\mu \in P(S)$, and since $B(S)$ is w^* -compact.

In Sect. 2.3, we present various results on the weak (and weak*) convergence of convolution products of probability measures of the form

$$\mu_{k,n} \equiv \mu_{k+1} * \mu_{k+2} * \cdots * \mu_n \quad (n > k).$$

Let us consider a few simple examples.

Example 2.1. Suppose there is an ideal $I \in \mathcal{B}$ of S such that for a sequence μ_n in $P(S)$, we have:

$$\sum_{i=1}^{\infty} \mu_{k_i, k_i + r_i}(I) = \infty, \quad (2.31)$$

for sequences (k_i) and (r_i) of positive integers with the property that for $i \geq 1$,

$$k_i + r_i < k_{i+1}, \quad 1 \leq r_i \leq r_{i+1}.$$

Here, we show that for $s \geq 1$, $\lim_{t \rightarrow \infty} \mu_{s,t}(I) = 1$.

To this end, if (X_i) is a sequence of independent S -valued random variables such that the distribution of X_i is μ_i , it follows from (2.31) that

$$\sum_{i=1}^{\infty} \Pr(X_{k_i+1} X_{k_i+2} \cdots X_{k_i+r_i} \in I) = \infty.$$

By the Borel–Cantelli lemma, we then have

$$\Pr\left(\bigcap_{m=1}^{\infty} \bigcup_{i=m}^{\infty} (X_{k_i+1} X_{k_i+2} \cdots X_{k_i+r_i} \in I)\right) = 1.$$

Hence, given $\epsilon > 0$, for every $m \geq 1$, there exists $n(m)$ such that

$$\Pr\left(\bigcup_{i=m}^{n(m)} (X_{k_i+1} X_{k_i+2} \cdots X_{k_i+r_i} \in I)\right) > 1 - \epsilon.$$

But notice that the set inside the parentheses is contained in the set

$$\{X_s X_{s+1} \cdots X_t \in I\},$$

whenever $s \leq k_m + 1$ and $t \geq k_i + r_i$ for $i = n(m)$. This is true because I is an ideal. This means that for $s \geq 1$, there exists $t(s)$, depending on ϵ , such that

$$t \geq t(s) \Rightarrow \mu_{s,t}(I) > 1 - \epsilon.$$

It is clear that if I is a singleton say, $I = \{0\}$, then for each $s \geq 1$, the sequence $\mu_{s,n}$ converges weakly to the unit mass at 0.

Example 2.2. Let S be the (multiplicative) topological semigroup of nonnegative reals with usual topology. Let μ be the normalized Lebesgue measure on the interval $[0, e]$, where $e = \exp(1)$. We show that the sequence (μ^n) converges vaguely (that is, in the weak* sense) to the measure λ given by $\lambda(\{0\}) = 1/2$ and $\lambda(0, \infty) = 0$. To show this, we use the Central Limit Theorem.

Let β be the probability measure on the reals $(-\infty, \infty)$, under addition, induced by μ and the map $\pi : x \rightarrow e^x$, so that

$$\beta(B) = \mu(\pi(B)).$$

Let (X_n) be an independent and identically distributed (i.i.d.) sequence of real random variables with distribution β . Then it follows that

$$m = E(X_i) = \frac{1}{e} \int_0^e \ln x \, dx = 0$$

and

$$\sigma^2 \equiv \text{Var}(X_i) = E(X_i^2) - m^2 > 0.$$

Notice that for $0 < k < \infty$,

$$\mu^n([0, k]) = \beta^n((-\infty, \ln k]).$$

Let $Y_n = \sum_{i=1}^n X_i$. Then, for any $k > 0$, $\beta^n((-\infty, \ln k]) = \Pr(Y_n \in (-\infty, \ln k]) =$

$\Pr\left([Y_n - nm]/s\sqrt{n} \in \left(-\infty, \frac{\ln k}{s\sqrt{n}}\right]\right)$ and this converges by the Central Limit Theorem to

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^0 e^{-(1/2)x^2} dx = \frac{1}{2}.$$

Example 2.3. In Example 2.2, if we had considered the interval $[0, a]$, $a > 0$, (instead of the interval $[0, e]$), but defined the probability measure μ in the same manner, then by using the Central Limit Theorem, we would easily obtain a weak convergence result for (μ^n) that is strikingly different. If $a > e$, the sequence converges vaguely to the zero measure, which would mean that for any $\alpha > 0$,

$$\lim_{n \rightarrow \infty} \mu^n([0, \alpha]) = 0.$$

If $a < e$, the behavior is just the opposite; in this case, for any $\alpha > 0$,

$$\lim_{n \rightarrow \infty} \mu^n([0, \alpha]) = 1,$$

that is, the sequence μ^n converges weakly to the unit mass at 0.

Example 2.4. Let $S = [0, 1]$, the real numbers x , $0 \leq x \leq 1$, with the usual topology and multiplication. Then S is a compact topological semigroup. Consider any sequence of probability measures μ_n in $P(S)$. Let (X_n) be a sequence of independent S -valued random variables such that X_n has distribution μ_n . Write

$$X_{k,n} \equiv X_{k+1} X_{k+2} \cdots X_n, \quad n > k.$$

Notice that $Y_k = \lim_{n \rightarrow \infty} X_{k,n}$ exists pointwise. Let $A_k \equiv \{Y_k > 0\}$. Then $A_k \subset A_{k+1}$. Notice that on the set

$$E \equiv \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k,$$

$\lim_{m \rightarrow \infty} Y_m = 1$. Since E is a tail event ($A_k \subset A_{k+1}$), it follows by the Kolmogorov zero-one law (see [38, p. 254]) that $\Pr(E) = 0$ or 1. Also, on $S - E$, $\lim_{m \rightarrow \infty} Y_m = 0$. This means that the limit

$$Y = \lim_{m \rightarrow \infty} Y_m$$

always exists, and either $Y = 0$ almost surely or $Y = 1$ almost surely.

It is now clear from the preceding discussion that for each $k \geq 1$, the sequence $\mu_{k,n}$ ($\equiv \mu_{k+1} * \cdots * \mu_n$) as $n \rightarrow \infty$ converges weakly to $\nu_k \in P(S)$, where ν_k is the distribution of Y_k , and that as $k \rightarrow \infty$, ν_k converges weakly to either the unit mass at 0 or the unit mass at 1. Notice that for any bounded continuous function f on S ,

$$\int f d\mu_{k,n} = \int f(X_{k,n}) dP \rightarrow \int f(Y_k) dP = \int f d\nu_k,$$

by the standard bounded convergence theorem of integration theory.

Example 2.5. Let S be the circle group $\{z \text{ is a complex number such that } |z| = 1\}$ with usual topology and multiplication. Let $\mu_n \in P(S)$ be defined by

$$\mu_n \left(\exp \left(\frac{2\pi i}{\sqrt{n}} \right) \right) = \mu_n \left(\exp \left(-\frac{2\pi i}{\sqrt{n}} \right) \right) = \frac{1}{2}.$$

Then notice that the r th Fourier coefficient of μ_n is given by

$$\text{fr}(\mu_n) = \int t^r \mu_n(dt) = \cos \left(\frac{2\pi r}{\sqrt{n}} \right).$$

Let r be any integer. Then we have

(i) There exists m such that

$$n \geq m \Rightarrow \left| \sin \left(\frac{\pi r}{\sqrt{n}} \right) \right| > \frac{1}{2} \cdot \left| \frac{\pi r}{\sqrt{n}} \right|.$$

(ii) For each $r \neq 0$,

$$\prod_{n=m}^{\infty} \cos \left(\frac{2\pi r}{\sqrt{n}} \right) = 0,$$

since

$$\sum_{n=m}^{\infty} \left[1 - \cos \left(\frac{2\pi r}{\sqrt{n}} \right) \right] \geq \frac{1}{2} \pi^2 r^2 \sum_{n=m}^{\infty} \frac{1}{n} = \infty,$$

where m is as in (i). This means that for each nonzero integer r ,

$$\lim_{n \rightarrow \infty} \text{fr}(\mu_{k,n}) = 0.$$

Notice that if λ is the Haar probability measure on the circle group, then for any integer r , $\text{fr}(\lambda) = 0$, so that

$$\lim_{n \rightarrow \infty} \text{fr}(\mu_{k,n}) = \text{fr}(\lambda).$$

Then it follows (see [186, Theorem 3.3, p. 76]) that $\mu_{k,n}$ converges weakly to the Haar probability measure on S as $n \rightarrow \infty$.

Theorem 2.7 (below) is a general result on the weak convergence of the sequence of averages of convolution powers of a probability measure. It is not essential in Theorem 2.7 for S to be locally compact; for example, Theorem 2.7 also holds when S is a complete metric separable topological semigroup.

Theorem 2.7. *Let $\mu \in P(S)$ and suppose that the sequence (μ^n) is tight; that is, $\{\mu^n : n \geq 1\}$ satisfies (2.30). Suppose also that*

$$S = \text{cl} \left[\bigcup_{n=1}^{\infty} S(\mu)^n \right].$$

Let $\mathcal{K} = \{\lambda \in P(S) : \lambda \text{ is a weak limit point of the sequence } (\mu^n)\}$. Also, let us define:

$$S_0 = \bigcup \{S(\lambda) : \lambda \in \mathcal{K}\}$$

and

$$S_1 = \overline{S_0}.$$

Then the following assertions hold:

- (i) *The sequence $1/n \sum_{k=1}^n \mu^k$ converges weakly to a probability measure ν such that $\nu = \nu * \nu = \mu * \nu = \nu * \mu$, and $S(\nu)$ is the closed completely simple kernel of S with a compact group factor.*
- (ii) *The set $\mathcal{K}$ is a group with respect to convolution as multiplication. The support $S(\nu)$ of ν coincides with the set S_1 . If η is the identity of $\mathcal{K}$, then $S(\eta) \subset S_1$, and they have the following product representations:*

$$S_1 = X \times H \times Y, \quad S(\eta) = X \times H_1 \times Y$$

(the usual product representation for completely simple semigroups), where H_1 is a closed normal subgroup of the compact group H , and X, Y have their usual meanings. Also if $\lambda \in \mathcal{K}$ and $\lambda \neq \eta$, then

$$S(\lambda) = X \times gH_1 \times Y, \quad g \in H - H_1,$$

if the product representation of the idempotent probability measure η (see Theorem 2.2) is $\eta_1 * \eta_2 * \eta_3$, then the product representation of λ is $\eta_1 * (\eta_2 * \delta_g) * \eta_3$, where $\eta_1 \in P(X)$, $\eta_3 \in P(Y)$, and η_2 is the Haar probability measure on H_1 . (Here, equality for sets is used in the sense of a topological isomorphism.) Furthermore, the set S_0 is also a completely simple subsemigroup of S , and $\mathcal{K} = \{\eta\}$ if and only if $\mu * \eta = \eta$.

- (iii) *The sequence μ^n converges weakly if and only if there does not exist a closed normal subgroup H_0 of H such that*

$$S(\mu) \cdot (X \times H_0 \times Y) \subset X \times gH_0 \times Y$$

- for some $g \in H - H_0$ and $YX \subset H_0$ if and only if there does not exist a closed
normal subgroup H_0 of H such that $YX \subset H_0$ and $eS(\mu)e \subset gH_0$ for some
 $g \in H - H_0$, where e is the identity of the group H .
(iv) The sequence μ^n converges weakly if and only if $\liminf_n (S(\mu)^n)$ is nonempty,
where $\liminf_n (S(\mu)^n)$ is defined as $\{x \in S : \text{for every open set } V(x) \text{ contain-}$
ing x , there exists a positive integer N such that $n \geq N$ implies that $V(x) \cap$
 $S(\mu)^n$ is nonempty $\}$.

Proof. (i) Write:

$$\mu_n \equiv 1/n \sum_{k=1}^n \mu^k.$$

Then, for $k \geq 1$, $\mu^k * \mu_n = \mu_n * \mu^k$ and

$$\lim_{n \rightarrow \infty} \|\mu_n - \mu^k * \mu_n\| = 0. \quad (2.32)$$

Since the sequence μ^n is tight, the same is true for the sequence μ_n so that
 $\{\mu_n : n \geq 1\}$ is weakly relatively compact. Let ν_1 and ν_2 be two weak limit points
of (μ_n) . By (2.32), for every $k \geq 1$,

$$\begin{aligned} \mu^k * \nu_1 &= \nu_1 * \mu^k = \nu_1, \\ \mu^k * \nu_2 &= \nu_2 * \mu^k = \nu_2. \end{aligned} \quad (2.33)$$

It follows that for every $n \geq 1$,

$$\begin{aligned} \mu_n * \nu_1 &= \nu_1 * \mu_n = \nu_1 \\ \mu_n * \nu_2 &= \nu_2 * \mu_n = \nu_2, \end{aligned} \quad (2.34)$$

which then implies that $\nu_1 = \nu_2$ ($\equiv \nu$, say), and $\mu * \nu = \nu * \mu = \nu = \nu * \nu$. It follows
easily that $S(\nu)$ is the completely simple semigroup with a compact group factor
(since it is the support of an idempotent probability measure; see Theorem 2.2),
which is also the kernel of S (since $\text{cl}[S(\mu)S(\nu)] = \text{cl}[S(\nu)S(\mu)] = S(\nu)$).

(ii) Notice that if T is a compact Hausdorff (second countable) topological semi-
group, then for any element $x \in T$, the subsemigroup

$$A(x) = \text{cl}\{x^n : n \geq 1\}$$

is a compact abelian subsemigroup. Therefore, it has a closed kernel (see Corol-
lary 1.3) K . If B denotes the set $\{y \mid y \text{ is a limit point of the sequence } x^n\}$, then it
is easy to verify that

(i) $K \subset B$;

(ii) $Bz = zB = B$ for every $z \in B$.

This means that B is a group containing the ideal K . For $k \in K$, $k^{-1} \in B$, so that the identity element of B belongs to K , which means $B = K$.

From the preceding arguments, it follows that the set $\mathcal{H}$ is a group with respect to convolution. Let us now prove that the set S_1 given in the theorem is the closed, completely simple kernel of S . To this end, it suffices to show that $S_1 = S(\nu)$, where ν is as in Assertion (i).

Let U be an open set containing $S(\nu)$. Then we claim that

$$\lim_{n \rightarrow \infty} \mu^n(U) = 1. \quad (2.35)$$

Let $\epsilon > 0$. Since (μ^n) is tight, there exists a compact subset $K_1 \subset S$ such that

$$\mu^n(K_1) > 1 - \epsilon, \quad n \geq 1. \quad (2.36)$$

Let K_2 be a compact subset of $S(\nu)$ such that

$$\nu(K_2) > 1 - \epsilon. \quad (2.37)$$

Since $S(\nu)$ is an ideal of S , the set $K_1 K_2 \subset S(\nu) \subset U$, and therefore, there exists an open set $V \supset K_2$ such that

$$K_1 V \subset U. \quad (2.38)$$

Now $\mu_n \rightarrow \nu$ weakly; therefore,

$$1 - \epsilon < \nu(V) \leq \liminf_{n \rightarrow \infty} \mu_n(V),$$

which means that there exists a positive integer m such that

$$\mu^m(V) > 1 - \epsilon. \quad (2.39)$$

By (2.36), (2.38) and (2.39), for all $n \geq 1$,

$$\mu^{n+m}(U) \geq \mu^n(K_1) \mu^m(V) > (1 - \epsilon)^2.$$

This proves (2.35).

Let us now write

$$S(\nu) = \bigcap_{n=1}^{\infty} U_n, \quad U_n \text{ open}, \quad U_n \supset \overline{U_{n+1}}.$$

Let $\lambda \in \mathcal{H}$ and $x \in S(\lambda)$. Suppose that $x \notin \overline{U_k}$ for some $k > 1$. Then there is an open set $V(x)$ containing x such that $V(x) \cap U_k = \emptyset$. Since $\lambda(V(x)) > 0$, there exists infinitely many n such that for some $\delta > 0$,

$$\mu^n(V(x)) > \delta,$$

which means that

$$\lim_{n \rightarrow \infty} \inf \mu^n(U_k) \leq 1 - \delta.$$

This contradicts (2.35). This means that for each $\lambda \in \mathcal{K}$,

$$S(\lambda) \subset \bigcap_{k=1}^{\infty} \overline{U_k} = S(v),$$

which implies that

$$S_1 \subset S(v).$$

Since for $\lambda \in \mathcal{K}$, $\mu * \lambda$ as well as $\lambda * \mu$ is in $\mathcal{K}$, it is clear that S_1 is an ideal of S . Since $S(v)$ is simple, it follows that $S_1 = S(v)$.

Let $\eta = \eta * \eta$ be the identity of the group $\mathcal{K}$. Then $S(\eta) (\subset S_1)$ as well as S_1 is a closed, completely simple subsemigroup of S with a compact group factor. Let $e = e^2$ be an idempotent element in $S(\eta)$, and let us define

$$\begin{aligned} X &\equiv E(S_1 e), & X_1 &\equiv E(S(\eta) e), \\ Y &\equiv E(e S_1), & Y_1 &\equiv E(e S(\eta)), \\ H &\equiv e S_1 e, & H_1 &\equiv e S(\eta) e. \end{aligned}$$

(Here, $E(D)$ denotes the set of idempotents in the set D .) Then it follows from Theorem 1.7 that

$$S_1 = X \times H \times Y, \quad S(\eta) = X_1 \times H_1 \times Y_1,$$

where equality means topologically isomorphic. Let us now show that $X = X_1$ and $Y = Y_1$. Let $(x, h, y) \in S_1$. It follows from the definition of S_1 that there exist $\lambda_n \in \mathcal{K}$, elements (x_n, h_n, y_n) in $S(\lambda_n)$ such that $(x_n, h_n, y_n) \rightarrow (x, h, y)$ as $n \rightarrow \infty$. By the group property of $\mathcal{K}$, there exist $\beta_n \in \mathcal{K}$ such that $\beta_n * \lambda_n = \lambda_n * \beta_n = \eta$. This means that $x_n \in X_1$ and $y_n \in Y_1$. It is easily verified that X_1 and Y_1 are both closed in S so that $x \in X_1$ and $y \in Y_1$. Thus, $X = X_1$ and $Y = Y_1$. Consequently, $YX \subset H_1$.

Now we prove the last assertions (including the normality of H_1) in Assertion (ii). Let $\lambda \in \mathcal{K}$. Since $\eta * \lambda = \lambda$, it is clear that if $(x, g, y) \in S(\lambda)$, then

$$X \times H_1 g \times \{y\} \equiv (X \times H_1 \times Y)(x, g, y) \subset S(\lambda).$$

Since $\lambda * \eta = \lambda$ and $YX \subset H_1$, then

$$X \times H_1 g \times Y \equiv (X \times H_1 g \times \{y\}) \cdot (\{x\} \times \{(yx)^{-1}\} \times Y) \subset S(\lambda).$$

Similarly, $X \times g H_1 \times Y \subset S(\lambda)$. Thus, we must have

$$S(\lambda) \equiv X \times U_\lambda \times Y,$$

where the set U_λ is a union of right cosets of H_1 , and similarly, it is also a union of
left cosets of H_1 . Since there exists $\beta \in \mathcal{K}$ such that $\lambda * \beta = \eta$, it is not difficult to
verify that U_λ consists only of a single right coset of H_1 . It follows similarly that

$$S(\lambda) (\equiv X \times H_1 g \times Y) = X \times g H_1 \times Y.$$

Thus, whenever $(x, g, y) \in S(\lambda)$ for some $\lambda \in \mathcal{K}$, we have $H_1 g = g H_1$. If $g \in H$,
then there exist $\lambda_n \in \mathcal{K}$ and $g_n \in H$ such that

$$S(\lambda_n) = X \times H_1 g_n \times Y,$$

$H_1 g_n = g_n H_1$, and $g_n \rightarrow g$. If $h \in H_1$, then there exist $h_n \in H_1$ such that

$$h g_n = g_n h_n.$$

Since H_1 is compact, it follows that $h g \in g H_1$ proving H_1 is a normal subgroup
of H .

Now we must show that if $S(\lambda) = X \times g H_1 \times Y$, then $g \notin H_1$ whenever $\lambda \neq \eta$.
To prove this, notice that if $g \in H_1$, then $S(\lambda) = S(\eta)$. Since $\lambda * \eta = \eta * \lambda = \lambda$, it
follows from Proposition 2.5 that for any Borel subset $B \subset S(\lambda)$ and u, v in $S(\lambda)$,

$$\lambda(Bu^{-1}v^{-1}) = \lambda(Bu^{-1}). \quad (2.40)$$

It follows from (2.40) that for $u \in S(\lambda)$,

$$\lambda^2(Bu^{-1}) = \int \lambda(Bu^{-1}v^{-1}) \lambda(dv) = \lambda(Bu^{-1}).$$

This means that since $\lambda^2 * \eta = \lambda^2$ and $\lambda * \eta = \lambda$ and $S(\eta) = S(\lambda)$,

$$\lambda^2(B) = \int \lambda^2(Bu^{-1}) \eta(du) = \int \lambda(Bu^{-1}) \eta(du) = \lambda(B),$$

for all Borel subsets $B \subset S(\lambda)$. This implies that $\lambda = \lambda^2$, so that $\lambda = \eta$, which is a
contradiction.

Now about the product representation of λ : First notice that $\lambda = \eta * \lambda * \eta$.
Following Theorem 2.2, consider the product representation $\eta = \eta_1 * \eta_2 * \eta_3$,
where $\eta_1 \in P(X)$, $\eta_3 \in P(Y)$, and η_2 is the Haar probability measure on H_1 . Then
 $\lambda = \eta_1 * [\eta_2 * \{(\eta_3 * \lambda * \eta_1) * \eta_2 * \delta_{g^{-1}}\} * \delta_g * \eta_3] = \eta_1 * (\eta_2 * \delta_g) * \eta_3$, since
the support of $\eta_3 * \lambda * \eta_1$ is $g H_1$, so that $\eta_3 * \lambda * \eta_1 * \eta_2 * \delta_{g^{-1}} \in P(H_1)$ and
consequently the product inside $\{ \}$ is simply η_2 .

The proof of Assertion (ii) will be complete if we show that the set S_0 is also
a completely simple semigroup. Notice that if J is an ideal of S_0 , then for $u \in J$,
 $u S_0 \cap S(\lambda) \neq \emptyset$ for some $\lambda \in \mathcal{K}$. Therefore, since $\mathcal{K}$ is a group, $u S_0 \cap S(\eta) \neq \emptyset$.
This means that

$$S(\eta) u S_0 \cap S(\eta) \neq \emptyset.$$

Since $S(\eta)uS_0 \cap S(\eta)$ is an ideal of $S(\eta)$, and $S(\eta)$ is simple,

666

$$S(\eta) \subset S(\eta)uS_0 \subset J.$$

For $\lambda \in \mathcal{K}$, $\lambda * \eta = \lambda$ and therefore, $S(\lambda) = S(\lambda)S(\eta)$. Consequently, $S(\lambda) \subset J$.
Thus, $J = S_0$, and S_0 is simple. Since $S_0 \subset S_1$ and S_1 is completely simple, it
follows that S_0 is a completely simple subsemigroup of S .

667

668

669

(iii) The if part in the first if-and-only-if statement follows from Assertion (ii)
since if μ^n does not converge weakly, then $\mu * \eta \neq \eta$, and the support of $\mu * \eta$ must
be of the form

670

671

672

$$X \times gH_1 \times Y, \quad g \in H - H_1,$$

where H_1 is a normal subgroup of H . This means that

673

$$S(\mu)(X \times H_1 \times Y) \subset S(\mu * \nu) = X \times gH_1 \times Y.$$

Now suppose there exists a closed normal subgroup H_0 of H such that $YX \subset$
 H_0 and

674

675

$$S(\mu)(X \times H_0 \times Y) \subset X \times gH_0 \times Y, \quad g \in H - H_0. \quad (2.41)$$

Then this implies that for each positive integer n ,

676

$$S(\mu)^n(X \times H_0 \times Y) \subset X \times g^n H_0 \times Y. \quad (2.42)$$

Since

677

$$S = \text{cl} \left[\bigcup_{n=1}^{\infty} S(\mu)^n \right]$$

and $SS_1 = S_1$, it is clear that

678

$$H = \text{cl} \left[\bigcup_{n=1}^{\infty} g^n H_0 \right].$$

Suppose there is a (smallest) positive integer n such that $g^n \in H_0$. In that case, H_0
and gH_0 are both open subsets of H . If β is a probability measure with its support
contained in $X \times H_0 \times Y$, then

679

680

681

$$S(\mu^n * \beta) \subset X \times H_0 \times Y \quad (2.43)$$

for infinitely many n and also

682

$$S(\mu^n * \beta) \subset X \times gH_0 \times Y \quad (2.44)$$

for infinitely many other values of n . Since the sets on the right of (2.43) and (2.44)
are both closed and open in $X \times H \times Y$, it is then clear that the sequence $\mu^n * \beta$ (and
therefore the sequence μ^n) cannot converge weakly. Now let us consider the other

683

684

685

possibility when $g^n \notin H_0$ for all positive integers n . In this case, we consider a limit point h of the infinite sequence g^n . Notice that $hH_0 \cap hgH_0 = \emptyset$, and therefore there are open subsets V_1 and V_2 with disjoint closures such that $hH_0 \subset V_1$ and $ghH_0 \subset V_2$.

Since $g^n H_0 \subset V_1$ for infinitely many n and again $g^n H_0 \subset V_2$ for infinitely many other values of n , we see that

$$S(\mu^n * \beta) \subset X \times V_1 \times Y \quad (2.45)$$

for infinitely many n and also

$$S(\mu^n * \beta) \subset X \times V_2 \times Y \quad (2.46)$$

for infinitely many other values of n , where $\beta \in P(S)$ and $S(\beta) \subset X \times H_0 \times Y$. It follows from (2.45) and (2.46) that the sequence $\mu^n * \beta$ (and therefore, the sequence μ^n) cannot converge weakly. This completes the proof of the first if-and-only-if statement in Assertion (iii).

To complete the proof of Assertion (iii), it is enough to show that (2.41) is equivalent to $eS(\mu)e \subset gH_0$, where e is the identity of H . To this end, suppose that

$$eS(\mu)e \subset gH_0. \quad (2.47)$$

Let $x \in S(\mu)$ and $y = (e_1, h, f_1)$, where $h \in H_0$, $e_1 \in X$ and $f_1 \in Y$. Notice that $ex = e(ex) \in eS_1$ so we can write

$$ex = (e, h', f), \quad h' \in H, \quad f \in Y.$$

Also, $exe = (e, h', f)(e, e, e) = (e, h', e)$, since $fe = e$, where Y is a right zero semigroup. By (2.47) $h' = exe \in gH_0$. Now $exy = (e, h', f)(e_1, h, f_1) = (e, h'fe_1h, f_1)$ so that $exy \in X \times gH_0 \times Y$, since $h' \in gH_0$ and $YX \subset H_0$. Let us now write

$$xy = (e_2, h_2, f_2), \quad h_2 \in H.$$

Then we have

$$exy = (e, e, e)(e_2, h_2, f_2) = (e, h_2, f_2)$$

since $ee_2 = e$, and this means that $h_2 \in gH_0$; thus (2.41) holds. Conversely, (2.41) implies that

$$S(\mu) \cdot (XH_0Y) \subset X \cdot gH_0 \cdot Y$$

so that

$$S(\mu)e \subset X \cdot gH_0 \cdot Y,$$

and therefore,

$$eS(\mu)e \subset (eX)(gH_0)(Ye) = gH_0,$$

which implies (2.47).

(iv) If $\mu^n \xrightarrow{w} \nu$, then for any x in $S(\nu)$ and any open set V containing x ,

$$0 < \nu(V) \leq \liminf_{n \rightarrow \infty} \mu^n(V),$$

which means

$$S(v) \subset \lim_{n \rightarrow \infty} \inf (S(\mu)^n).$$

To prove the converse, let us assume that

$$\lim_{n \rightarrow \infty} \inf (S(\mu)^n) \neq \emptyset.$$

Then it is easily verified that

$$\lim_{n \rightarrow \infty} \inf (S(\mu)^n)$$

is an ideal of S . Suppose that the sequence μ^n does not converge weakly. Then, by Assertion (iii) it follows that there exists a closed normal subgroup H_0 of H such that for each positive integer n and for some $g \in H - H_0$,

$$eS(\mu)^n e \subset g^n H_0, \quad e = e^2 \in H. \quad (2.48)$$

This follows using the same argument used preceding (iv) above. Let $x \in H$. Since H is a topological group, given any open set $U(x)$ containing x , there exist open sets $U(y)$ containing $y \in H$ and $U(z)$ containing $z \in H$ such that

$$U(x) \cap H \supset [eU(y)e][eU(z)e]^{-1}. \quad (2.49)$$

Now S_1 , as the kernel of S , is a subset of

$$\lim_{n \rightarrow \infty} \inf (S(\mu)^n),$$

and this means there exists a positive integer N such that

$$U(y) \cap S(\mu)^N \neq \emptyset, \quad U(z) \cap S(\mu)^N \neq \emptyset.$$

It follows from (2.48) that

$$eU(y)e \cap g^N H_0 \neq \emptyset, \quad eU(z)e \cap g^N H_0 \neq \emptyset.$$

It follows from (2.49) that

$$U(x) \cap H \cap H_0 \neq \emptyset.$$

This means that $x \in H_0$. Thus, $H_0 = H$, and this is a contradiction. $\square$

Noting that a completely simple subsemigroup (with a compact group factor) in a group is a compact subgroup, one can now easily obtain the following corollary.

Corollary 2.1. *Let S be a compact group and $\mu \in P(S)$. Let $\eta = \eta^2$ be the identity of the group*

$$\mathcal{K} = \{v \in P(S) : v \text{ is a weak limit point of the sequence } \mu^n\},$$

where convolution is the multiplication. Assume that

730

$$S = \text{cl} \left(\bigcup_{n=1}^{\infty} S(\mu)^n \right).$$

(Recall that a compact cancellative semigroup is a compact topological group so that the smallest closed subsemigroup containing $S(\mu)$ in a compact group is also the smallest closed subgroup containing $S(\mu)$). Then the following assertions are equivalent:

- (i) The sequence (μ^n) is weakly convergent; 735
- (ii) The set $\liminf_{n \rightarrow \infty} (S(\mu)^n)$ is nonempty; 736
- (iii) The sets 737

$$\liminf_{n \rightarrow \infty} (S(\mu)^n) \text{ and } \limsup_{n \rightarrow \infty} (S(\mu)^n)$$

are equal; if $v \in \mathcal{K}$, then

738

$$S(v) \subset \limsup_{n \rightarrow \infty} (S(\mu)^n).$$

Note that $\limsup_{n \rightarrow \infty} A_n \equiv \{x \in S \mid \text{if } V \text{ is open and } x \in V, \text{ then } V \cap A_n \neq \emptyset \text{ for infinitely many } n\}.$

739

740

(iv) S is the smallest closed subgroup containing the set

741

$$\cup \{S(\mu)^n \cdot S(\mu)^{-n} : n \geq 1\};$$

- (v) $S(\mu)$ is not contained in a proper coset of any closed normal subgroup of S ; 742
- (vi) $S(\mu)$ is not contained in any proper coset of $S(\eta)$; 743
- (vii) $\eta(Bx^{-1}) = \eta(x^{-1}B) = \eta(B)$ for $x \in S$, $B \subset S$. 744

Proof. The proof is left to the reader.

□ 745

In Sect. 2.2, when S is abelian, we determined in Theorem 2.1 all the v -invariant ($v \in P(S)$) probability measures μ . Corollary 2.2 and 2.3 address the same problem when S is a group.

746

747

748

Corollary 2.2. Suppose that S is a group and μ, v are in $P(S)$. Then $\mu = \mu * v$ if and only if $\mu = \mu_x$ for each x in the smallest closed subgroup of S containing $S(v)$.

749

750

751

Proof. The if part is trivial. For the only-if part, let $\mu = \mu * v$. Then we have for $n \geq 1$,

752

753

$$\mu = \mu * v^n. \quad (2.50)$$

754

Given $\epsilon > 0$, let K be a compact subset such that $\mu(K) > 1 - \epsilon$. Then

755

$$\begin{aligned} 1 - \epsilon < \mu(K) &= \int v^n(x^{-1}K) \mu(dx) \\ &\leq \int_K v^n(x^{-1}K) \mu(dx) + \epsilon \\ &\leq v^n(K^{-1}K) + \epsilon, \end{aligned}$$

where $K^{-1}K = \cup \{x^{-1}K : x \in K\}$. Since the set $K^{-1}K$ is a subset of S (which is a topological group), it is compact and, consequently, the sequence v^n is tight. We can now use Theorem 2.7 so that

757
758

$$\frac{1}{n} \sum_{k=1}^n v^k$$

converges weakly to some idempotent probability measure β . Also we have from (2.50), $\mu = \mu * 1/n \sum_{k=1}^n v^k$. Since, by Proposition 2.3, convolution is continuous with respect to weak topology, we have

759

760

761

$$\mu = \mu * \beta, \quad (2.51)$$

where $\beta = \beta * \beta$ and, consequently, by Theorem 2.2, $S(\beta)$ is a compact subgroup of S , $S(\beta) \supset S(v)$, since $v * \beta = \beta * v = \beta$, (and β is the Haar probability measure on $S(\beta)$).

762

763

764

Now let K be any compact subset of S . Since the map $x \rightarrow \mu(Kx^{-1})$ is upper-semicontinuous (by Proposition 2.1), there exists $z \in S(\beta)$, which is compact, such that

765

766

767

$$\mu(Kz^{-1}) = \sup \{ \mu(Kx^{-1}) : x \in S(\beta) \}.$$

It follows from (2.51) that

768

$$\mu(Kz^{-1}) = \int \mu(Kz^{-1}x^{-1}) \beta(dx)$$

so that

769

$$\mu(Kz^{-1}) = \mu(Kz^{-1}x^{-1}) \quad (2.52)$$

for β -almost all x in $S(\beta)$. Suppose there exist x_0 in $S(\beta)$ such that $\mu(Kz^{-1}x_0^{-1}) < \mu(Kz^{-1})$. Then since $x \rightarrow \mu(Kz^{-1}x^{-1})$ is upper-semicontinuous,

770

771

$$\{x \in S \mid \mu(Kz^{-1}x^{-1}) < \mu(Kz^{-1})\}$$

is an open set containing x_0 in $S(\beta)$ (and, therefore, has positive β -measure) and this contradicts that (2.52) holds almost everywhere (β). Thus, we have:

$$\mu(Kz^{-1}) = \mu(Kz^{-1}x^{-1})$$

for all x in $S(\beta)$. Since $S(\beta)$ is a group, it is clear that $\mu(K) = \mu(Kx^{-1})$ for all x in $S(\beta)$. $\square$

As far as we know, it is still an unsolved problem to determine all the ν -invariant probability measures μ , where μ and ν are in $P(S)$ and S is nonabelian or not a group. However, when S is compact, or more generally, when the sequence ν^n is tight, we have Corollary 2.3.

Corollary 2.3. *Let $\nu \in P(S)$ such that the sequence (ν^n) is tight and*

$$S = \text{cl} \left(\bigcup_{n=1}^{\infty} S(\nu)^n \right).$$

Let $\mu \in P(S)$ such that

$$\mu * \nu = \mu. \quad (2.53)$$

Then the following assertions are valid:

- (i) *S has a completely simple kernel $K = S(\nu_0)$, where ν_0 is the weak limit of*

$$\frac{1}{n} \sum_{k=1}^n \nu^k$$

and $\nu * \nu_0 = \nu_0 * \nu = \nu_0$.

- (ii) *$S(\mu) \subset K$ and $\mu = \mu * \mu$.*

- (iii) *If $e = e^2$ is an idempotent element in $S(\mu)$, which is a completely simple subsemigroup of K , then the product representations of $S(\mu)$ and K take the form*

$$\begin{aligned} K &= X \times G \times Y \\ S(\mu) &= A \times G \times Y, \quad A \subset X, \end{aligned} \quad (2.54)$$

where $X \equiv E(Ke)$, $Y \equiv E(eK)$, $G \equiv eKe$; moreover, μ and ν_0 have the following representation:

$$\begin{aligned} \nu_0 &= \nu_1 * \nu_2 * \nu_3, \\ \mu &= \mu_1 * \nu_2 * \nu_3, \end{aligned} \quad (2.55)$$

where $\nu_1 \in P(X)$, $\mu_1 \in P(A)$, $\nu_3 \in P(Y)$, and ν_2 is the Haar probability measure on G . Conversely, if ν_0 is the weak limit of

$$\frac{1}{n} \sum_{k=1}^n \nu^k$$

and takes the form of (2.55), then any $\mu \in P(S)$ written as in (2.55) satisfies (2.53).

Proof. Assertion (i) follows from Theorem 2.7. Suppose now that (2.53) holds for some $\mu \in P(S)$. Then

$$\mu * \left(\frac{1}{n} \sum_{k=1}^n v^k \right) = \mu, \quad n \geq 1;$$

therefore, $\mu * v_0 = \mu$ and $S(\mu) = \overline{S(\mu)S(v_0)} \subset S(v_0) = K$. Now let $x \in S(\mu)$ and B be any Borel set. Then $\mu * v_0 = \mu$ implies that

$$\mu(x^{-1}B) = \int v_0(y^{-1}x^{-1}B) \mu(dy) = v_0(x^{-1}B), \quad (2.56)$$

by Proposition 2.5, since $v_0 = v_0 * v_0$. This means that

$$\begin{aligned} \mu * \mu(B) &= \int \mu(x^{-1}B) \mu(dx) \\ &= \int v_0(x^{-1}B) \mu(dx) \\ &= \mu * v_0(B) = \mu(B) \end{aligned}$$

so that $\mu = \mu * \mu$. By Theorem 2.2, $S(\mu)$ is a completely simple subsemigroup of $K = S(v_0)$. Since $S(v_0) = \text{cl}[S(\mu)S(v_0)]$, the representations in (2.54) follow immediately. By Theorem 2.2, v_0 takes the form of (2.55), where $v_2 * v_3 = \delta_e * v_0$ (see (2.27)). By (2.56), $\delta_e * \mu = \delta_e * v_0$; therefore, since $\mu = \mu * \mu$, μ must take the form given in (2.55).

For the converse part, if μ and v_0 take the form of (2.55) and $v_0 * v = v_0$, then since $S(v_3 * v_1) \subset YX \subset G$,

$$v_2 * (v_3 * v_1) = v_2 = v_2 * v_2.$$

Therefore, it follows that $\mu * v_0 = \mu$ and, consequently,

$$\begin{aligned} \mu * v &= (\mu * v_0) * v \\ &= \mu * (v_0 * v) \\ &= \mu * v_0 = \mu. \end{aligned}$$

□

Notice that in Corollary 2.3, if (2.53) is replaced by the stronger condition

$$\mu * v = v * \mu = \mu$$

then μ has to coincide with v_0 there. However, what happens when this equation holds, but the sequence v^n is not tight? Look at Example 2.2. There, with v as the normalized Lebesgue measure on $[0, e]$, where $e = \exp(1)$,

$$\frac{1}{n} \sum_{k=1}^n v^k,$$

converges vaguely to the measure $1/2 \cdot \delta_{\{0\}}$, and with $\mu = \delta_{\{0\}}$, $\mu * \nu = \nu * \mu = \mu$. 813
Corollary 2.4 shows that this type of behavior cannot occur in the presence of a mild 814
compactness condition. 815

Corollary 2.4. *Suppose that all the assumptions in Corollary 2.3 (except the one on 816
the tightness of ν^n) hold. Suppose also that S satisfies the following compactness 817
condition: 818*

$$K \text{ compact, } x \in S \Rightarrow x^{-1}K \text{ is compact.} \quad (2.57)$$

Then (2.53) implies that $\frac{1}{n} \sum_{k=1}^n \nu^k$ converges weakly to ν_0 in $P(S)$ and, conse- 819
quently, all the results in Corollary 2.3 remain valid. 820

Proof. Let λ be a weak* limit point of the sequence 821

$$\nu_n \equiv \frac{1}{n} \sum_{k=1}^n \nu^k.$$

If all such weak* limit points are probability measures, then it follows from 822
Theorem 2.7 that the sequence $\frac{1}{n} \sum_{k=1}^n \nu^k$ converges weakly to some ν_0 in $P(S)$, and 823
the rest of Corollary 2.4 then follows exactly as in Corollary 2.3. Thus it suffices to 824
show that $\lambda \in P(S)$. 825

To this end, let $f \in C_c(S)$ and $x \in S$. Define $f_x(y) \equiv f(xy)$. Then (2.57) im- 826
plies that $f_x \in C_c(S)$. Let (n_k) be the subsequence such that ν_{n_k} weak* converges 827
to λ . Then let us define the functions g_k and g by 828

$$g_k(x) = \int f(xy) \nu_{n_k}(dy),$$

$$g(x) = \int f(xy) \lambda(dy).$$

It is clear that for each $x \in S$, $g_k(x)$ converges, as $k \rightarrow \infty$, to $g(x)$; therefore, by 829
the bounded convergence theorem, for $f \in C_c(S)$, we have: 830

$$\begin{aligned} \int f(x) \mu(dx) &= \int f(x) \mu * \nu_{n_k}(dx) \\ &= \iint f(xy) \mu(dx) \nu_{n_k}(dy) \\ &= \int g_k(x) \mu(dx) \\ &\rightarrow \int g(x) \mu(dx) \\ &= \iint f(xy) \mu(dx) \lambda(dy) \\ &= \int f(x) \mu * \lambda(dx). \end{aligned}$$

This means that $\mu = \mu * \lambda$ and, consequently, $\lambda \in P(S)$.

□ 831

In Examples 2.2 and 2.3, we found that even though $\{0\}$ was an ideal of the multiplicative semigroup $[0, \infty)$ (with the usual real-line topology),

$$\lim_{n \rightarrow \infty} \sup \mu^n([0, a]) < 1 \quad (2.58)$$

for $a > 0$, when μ is the normalized Lebesgue measure on $[0, b]$, $b \geq e$, where $e = \exp(1)$. However, if we consider the Lebesgue measure β on the compact semigroup $[0, 1]$ with usual topology, then β^n converges weakly to $\delta_{\{0\}}$, the unit mass at 0. Proposition 2.6 explains this result.

Proposition 2.6. Suppose S is compact, $\mu \in P(S)$, and

$$S = \text{cl} \left(\bigcup_{n=1}^{\infty} S(\mu)^n \right).$$

Then for any open set $G \supset K$, where K is the kernel of S ,

$$\lim_{n \rightarrow \infty} \mu^n(G) = 1.$$

In fact, there exist $0 < \delta < 1$ and a positive integer m such that for $n \geq m$,

$$\mu^n(S \setminus G) \leq \delta^{[n/m]},$$

where $[n/m]$ is the greatest integer $\leq \frac{n}{m}$. In other words, the convergence is geometrically fast.

Proof. Let $K \subset G$, G open. Notice that S is compact and $KS \cup SK \cup SKS \subset K \subset G$. Thus, there exist open sets V and W such that $K \subset V \subset W \subset G$ such that

$$VS \cup SV \cup SVS \subset W, \quad WS \subset G. \quad (2.59)$$

There exists a positive integer m such that $\mu^m(V) = 1 - \delta > 0$. Let (X_n) be an i.i.d. sequence of S -valued random variables, each with distribution μ^m . Then, for $1 \leq k < m$, using (2.59), we have:

$$\begin{aligned} \mu^{nm+k}(G) &\geq \mu^{nm}(W) \mu^k(S) \\ &= \mu^{nm}(W) \\ &= P(X_1 X_2 \dots X_n \in W) \\ &\geq P\left(\bigcup_{i=1}^n \{X_i \in V\}\right) \\ &= 1 - \delta^{nm}. \end{aligned}$$

The proof is now clear.

□ 848

We obtain a similar result with an assumption different from compactness (see Proposition 2.7). 849 850

Proposition 2.7. *Let I be a Borel set that is an ideal of S . Suppose that for some positive integer m , $\mu^m(I) > 0$ for some $\mu \in P(S)$. Then the sequence $\mu^n(I)$ monotonically increases to 1.* 851 852 853

Proof. Notice that $I \supset I \cdot S$ so that $\mu^{n+1}(I) \geq \mu^n(I)\mu(S) = \mu^n(I)$, for all positive integers n . Also, notice that 854 855

$$\begin{aligned}\mu^{n+1}(I^c) &= \int \mu^n(I^c x^{-1}) \mu(dx) \\ &\leq \mu^n(I^c) \mu(I^c) \leq [\mu(I^c)]^{n+1}.\end{aligned}$$

Thus, $(\mu^m)^{n+1}(I^c) \leq [\mu^m(I^c)]^{n+1}$. Now the proof follows the same lines as in Proposition 2.6 since $SIS \subset I$. 856 857 $\square$

An important question in the context of weak* convergence is what makes the sequence μ^n for $\mu \in P(S)$ a tight sequence when μ^n does not converge vaguely (that is, in the weak* sense) to the zero measure. Note that in Example 2.2, the sequence μ^n is not tight nor does it converge vaguely to the zero measure. Although this problem is not completely solved (to the best of our knowledge), we present a number of satisfactory results in this context. First, we prove Lemma 2.2. 858 859 860 861 862 863

Lemma 2.2. *Let μ_n be a sequence in $P(S)$ such that the subsequence μ_{0,n_t} , where $\mu_{k,n} = \mu_{k+1} * \cdots * \mu_n$, has at least one weak* limit point in $P(S)$. Suppose that S has the property such that convolution as a map from $P(S) \times B(S) \rightarrow B(S)$ is continuous in the weak* sense. Then there is a subsequence $(p_t) \subset (n_t)$ such that for each positive integer k ,* 864 865 866 867 868

$$\begin{aligned}\mu_{k,p_t} &\xrightarrow{w} \lambda_k \in P(S), \\ \lambda_{p_t} &\xrightarrow{w} \lambda = \lambda * \lambda \in P(S), \\ \lambda_k * \lambda &= \lambda_k.\end{aligned}$$

Proof. Suppose $\mu_{0,n_t} \xrightarrow{w} \lambda_0 \in P(S)$. Note that w^* -convergence is weak convergence when the limit is in $P(S)$. Now for each positive integer t , 869 870

$$y_{n_t} \equiv (\mu_{0,n_t}, \mu_{1,n_t}, \dots, \mu_{n_t-1}, 0, 0, 0, \dots)$$

are elements in the product space 871

$$Y = \prod_{i=1}^{\infty} X_i, \quad X_i \equiv B(S), \text{ for each } i,$$

where each X_i has weak* topology, Y has the product topology so that Y is compact, since $B(S)$ is w^* -compact. Since Y is compact (and also first countable), there 872 873

is a subsequence $(m_t) \subset (n_t)$ such that $y_{m_t} \rightarrow y \in Y$, in the topology of Y . This means that for each $k \geq 0$, there exists $\lambda_k \in B(S)$ such that

$$\mu_{k,m_t} \rightarrow \lambda_k.$$

Since convolution is continuous as a map from $P(S) \times B(S) \rightarrow B(S)$, it follows that for each $k \geq 1$,

$$\mu_{0,m_t} = \mu_{0,k} * \mu_{k,m_t} \rightarrow \mu_{0,k} * \lambda_k$$

in the w^* -sense, and this means that

$$\mu_{0,k} * \lambda_k = \lambda_0, \quad k \geq 1.$$

However, since $\lambda_0 \in P(S)$, this implies that $\lambda_k \in P(S)$ for each $k \geq 1$. Let $(p_t) \subset (m_t)$ be a subsequence such that $\lambda_{p_t} \rightarrow \lambda \in B(S)$ in the weak* sense. Now for fixed integer s and $t > s$ such that $p_s > k$, we have

$$\mu_{k,p_s} * \mu_{p_s,p_t} = \mu_{k,p_t}.$$

Again by the continuity of convolution, it follows that given $k \geq 0$, for each s such that $p_s > k$,

$$\mu_{k,p_s} * \lambda_{p_s} = \lambda_k$$

which in turn implies that

$$\lambda_k * \lambda = \lambda_k, \quad k \geq 1.$$

Since $\lambda_k \in P(S)$, $\lambda \in P(S)$. The last equation implies that $\lambda * \lambda = \lambda$. $\square$

Lemma 2.3 shows when the convolution operation as a map from $P(S) \times B(S)$ into $B(S)$ is continuous. The compactness conditions in the next two lemmas hold when S is either compact or a group.

Lemma 2.3. Suppose S satisfies the following compactness condition:

$$K \text{ compact and } x \in S \Rightarrow x^{-1}K \text{ is compact.}$$

This condition holds when S is either compact or a group.

If $\mu_n \rightarrow \mu$ weakly in $P(S)$ and $v_n \rightarrow v \in B(S)$ in the weak* sense with each $v_n \in P(S)$, then $\mu_n * v_n \rightarrow \mu * v$ in the weak* topology.

Proof. Let $f \in C_c(S)$. Then for each $s \in S$, $t \rightarrow f(st)$ is in $C_c(S)$. Hence, if

$$g_n(s) \equiv \int f(st)v_n(dt), \quad g(s) \equiv \int f(st)v(dt)$$

then $\lim_{n \rightarrow \infty} g_n(s) = g(s)$, $s \in S$. Since ν is a regular measure, it is easily seen that g is a bounded continuous function on S . Also, by Egoroff's theorem in analysis, given $\epsilon > 0$, there exists a compact set K such that $\mu(K) < \epsilon$ and on $S - K$, $g_n \rightarrow g$ uniformly. Since $\mu_n \rightarrow \mu$ weakly,

$$\lim_{n \rightarrow \infty} \sup \mu_n(K) \leq \mu(K) < \epsilon.$$

Then we have

$$\begin{aligned} \left| \int g_n d\mu_n - \int g d\mu \right| &\leq \left| \int_K g_n d\mu_n - \int_K g d\mu_n \right| + \int_{K^c} |g_n - g| d\mu_n \\ &\quad + \left| \int g d\mu_n - \int g d\mu \right|, \end{aligned}$$

which shows that

$$\lim_{n \rightarrow \infty} \int g_n d\mu_n = \int g d\mu.$$

This means that

$$\begin{aligned} \int f d\mu_n * \nu_n &= \iint f(st) \mu_n(ds) \nu_n(dt) \\ &= \int \left[\int f(st) \nu_n(dt) \right] \mu_n ds \\ &= \int g_n(s) \mu_n(ds) \rightarrow \int g d\mu = \int f d\mu * \nu. \end{aligned}$$

□ 901

Lemma 2.4. Let S satisfy the following compactness condition:

$A^{-1}B$ is compact whenever A and B are compact.

Let $\mu_n \in P(S)$ and $\mu_{0,n} \equiv \mu_1 * \mu_2 * \cdots * \mu_n$. Then

$$\sup \left\{ \lim_{n \rightarrow \infty} \sup_{x \in S} \mu_{0,n}(Kx^{-1}) : K \text{ compact} \right\} \equiv b$$

is either 0 or 1.

Proof. First note that

$$\mu_{0,n+1}(Kx^{-1}) = \int \mu_{0,n}(K(yx)^{-1}) \mu_{n+1}(dy) \leq \sup_{x \in S} \mu_{0,n}(Kx^{-1}),$$

which shows that

$$\lim_{n \rightarrow \infty} \sup_{x \in S} \mu_{0,n}(Kx^{-1})$$

exists.

Suppose that $0 < b < 1$. Choose

908

$$c = b + \frac{b(1-b)}{4}.$$

Then $b < c < 1$ and $c(1+c) < 2b$. Let K be any compact set. Then we have for some $k \geq 1$ and $q > 0$,

909
910

$$\sup_{x \in S} \mu_{0,k}(Kx^{-1}) < c - q. \quad (2.60)$$

Notice that there exists a compact set K_1 (depending only on k) such that

911

$$\mu_{0,k}(K_1) > \frac{b+c}{2c}.$$

Then given any compact set A , for n sufficiently large, we must have:

912

$$\mu_{k,n}(Ax^{-1}) < c, \quad x \in S. \quad (2.61)$$

Otherwise, since $(K_1A)x^{-1} \supset K_1(Ax^{-1})$,

913

$$\mu_{0,n}((K_1A)x^{-1}) \geq \mu_{0,k}(K_1)\mu_{k,n}(Ax^{-1}) \geq \frac{b+c}{2c} \cdot c = \frac{b+c}{2} > b$$

for infinitely many n , contradicting the definition of b .

914

Now let D be a compact set such that

915

$$\mu_{0,k}(S-D) < \frac{c}{2} - q. \quad (2.62)$$

Let $E = D^{-1}K$. Then E is compact by our assumption. For n sufficiently large and any $x \in S$, using (2.60)–(2.62) we have

916
917

$$\begin{aligned} \mu_{0,n}(Kx^{-1}) &= \int \mu_{0,k}(Kx^{-1}y^{-1}) \mu_{k,n}(dy) \\ &= \int_{Ex^{-1}} + \int_{S-Ex^{-1}} \\ &\leq (c-q)\mu_{k,n}(Ex^{-1}) + \left(\frac{c}{2}-q\right)[1-\mu_{k,n}(Ex^{-1})] \\ &= \left(\frac{c}{2}-q\right) + \frac{c}{2}\mu_{k,n}(Ex^{-1}) \\ &< \frac{c+c^2}{2} \\ &< b, \end{aligned}$$

which contradicts the definition of b . This proves Lemma 2.4.

□ 918

We remark that the compactness condition in Lemma 2.4 holds when S is compact or S is a group. This condition holds in other cases also; for example, when S is a multiplicative semigroup of $d \times d$ nonnegative matrices (with usual topology from the Euclidean d^2 -space), and A is a compact subset of matrices in S that have no zero rows or zero columns, then $A^{-1}B$ is a compact subset of S whenever B is compact (and the set AB^{-1} in this case is compact if A and B are compact and B consists of only those matrices in S that have no zero rows or zero columns).

We present a useful result in Theorem 2.8.

Theorem 2.8. *Suppose that S satisfies the following compactness condition:*

- *If A and B are compact subsets of S , then the sets $A^{-1}B$ and AB^{-1} , where $A^{-1}B \equiv \cup \{x^{-1}B : x \in A\}$ and $AB^{-1} \equiv \cup \{Ax^{-1} : x \in B\}$ are compact.*

Let $\mu_n \in P(S)$, $n \geq 1$. Then one of the following two assertions is always valid:

- $\mu_{0,n}$ converges vaguely (that is, in the weak*-sense) to the zero measure as $n \rightarrow \infty$.*
- There exist $x_n \in S$, $n \geq 1$, such that for each $k \geq 1$,*

$$\mu_{k,n} * \delta_{x_n} \rightarrow \nu_k \in P(S) \text{ as } n \rightarrow \infty.$$

Thus, when S is compact, Assertion (ii) holds; and when S is a group either Assertion (i) or (ii) holds.

Proof. Suppose that Assertion (i) does not hold. Then there exists $\epsilon > 0$, a compact set K_1 and a subsequence (n_i) of positive integers such that

$$\mu_{0,n_i}(K_1) > \epsilon, \quad i \geq 1. \quad (2.63)$$

By Lemma 2.4, for $0 < \delta < \epsilon$, there exist elements $x_{n_i} \in S$ and a compact set K_2 such that for $i \geq 1$,

$$\mu_{0,n_i}(K_2 x_{n_i}^{-1}) > 1 - \delta. \quad (2.64)$$

It follows from (2.63) and (2.64) that

$$K_1 \cap K_2 x_{n_i}^{-1} \neq \emptyset, \quad i \geq 1,$$

which means that

$$x_{n_i} \in K_1^{-1} K_2, \quad i \geq 1.$$

Then from (2.64) we have for the compact set $A_\delta \equiv K_2 (K_1^{-1} K_2)^{-1}$,

$$\mu_{0,n_i}(A_\delta) > 1 - \delta. \quad (2.65)$$

Thus, the subsequence $[\mu_{0,n_i}]$ is tight.

By Lemmas 2.2 and 2.3, there exists a further subsequence $(p_i) \subset (n_i)$ such that 944
for each $k \geq 1$, 945

$$\begin{aligned}\mu_{k,p_i} &\xrightarrow{w} \nu_k \in P(S) \\ \nu_{p_i} &\xrightarrow{w} \nu = \nu * \nu^2 \\ \nu_k * \nu &= \nu_k.\end{aligned}\tag{2.66}$$

It follows from Theorem 2.2 that $S(\nu)$ is a completely simple semigroup with a 946
compact group factor. The compactness condition in this theorem then implies that 947
 $S(\nu)$ is also compact. Let us now write: 948

$$S(\nu) = \bigcap_{n=1}^{\infty} G_n,$$

where G_n is a sequence of open sets with compact closure. Now there exists a sub- 949
sequence $(s_i) \subset (p_i)$ such that 950

$$\mu_{s_i, s_{i+1}} \xrightarrow{w} \nu$$

and for each $i \geq 1, j \geq 1$, 951

$$\mu_{s_i, s_{i+j}}(G_i) > 1 - \frac{1}{i}.$$

Then for any positive integer m such that $s_i < m \leq s_{i+1}$, there exists z_m in S such 952
that $\mu_{s_i, m}(G_i z_m^{-1}) > 1 - \frac{1}{i}$, since 953

$$\int \mu_{s_i, m}(G_i y^{-1}) \mu_{m, s_{i+2}}(dy) = \mu_{s_i, s_{i+2}}(G_i) > 1 - \frac{1}{i}.$$

Let $z \in S(\nu)$. We claim that for all nonnegative integers k , the sequence $\mu_{k,n} * \delta_{z_n z}$ 954
converges weakly to $\nu_k * \delta_z$. 955

To prove this claim, let λ_k be a weak* limit point of $(\mu_{k,n} * \delta_{z_n z})$. Then given 956
any sequence (t_j) , there is a subsequence (m_j) such that for some subsequence 957
 $(s_{i(j)})$ of (s_i) such that 958

$$s_{i(j)} < m_j \leq s_{i(j)+1}$$

and 959

$$\mu_{k, m_j} * \delta_{z_{m_j} z} \rightarrow \lambda_k \text{ (vaguely).}\tag{2.67}$$

The sequence $\mu_{s_{i(j)}, m_j} \delta_{z_{m_j}}$ is tight (because of the way we constructed the se- 960
quence z_m), and it follows that any weak limit point β of this sequence must satisfy 961
 $S(\beta) \subset S(\nu)$. Also note that for any Borel set B , 962

$$\begin{aligned}\nu * \beta * \delta_z(B) &= \int \nu(B z^{-1} y^{-1}) \beta(dy) \\ &= \nu(B z^{-1}) \\ &= \nu * \delta_z(B)\end{aligned}$$

by Proposition 2.4 (since $\nu = \nu^2$). Since we have

963

$$\mu_{k,m_j} * \delta_{z_{m_j} z} = \mu_{k,s_{i(j)}} * \mu_{s_{i(j)},m_j} * \delta_{z_{m_j} z},$$

it is clear that

964

$$\begin{aligned} \lambda_k &= \nu_k * \beta * \delta_z \\ &= (\nu_k * \nu) * \beta * \delta_z \\ &= \nu_k * (\nu * \beta * \delta_z) \\ &= \nu_k * (\nu * \delta_z) \\ &= \nu_k * \delta_z. \end{aligned}$$

This means that for each $k \geq 1$,

965

$$\mu_{k,n} * \delta_{z_n z} \xrightarrow{w} \nu_k * \delta_z.$$

□ 966

Let us remark that with a little extra work, Assertion (i) of Theorem 2.8 can be strengthened when S is a group to the assertion that

967

968

$$\lim_{n \rightarrow \infty} \sup_{x \in S} \mu_{0,n}(Kx^{-1}) = 0$$

for any compact subset $K \subset S$. Theorem 2.8 leads to the following important result.

969

970

Theorem 2.9. *Let S be a group and $\mu \in P(S)$. Suppose S is noncompact and S has no proper closed subgroup containing $S(\mu)$. Then the sequence μ^n converges vaguely to the zero measure.*

971

972

973

Proof. Suppose

974

$$S_0 = \text{cl} \left(\bigcup_{n=1}^{\infty} S(\mu)^n \right)$$

and that μ^n does not converge vaguely to the zero measure; then there exists a compact subset $K \subset S_0$ such that for some $\delta > 0$,

975

976

$$\mu^n(K) > \delta \quad (2.68)$$

for infinitely many n . This means that there exists an element x in K such that for any open set $N(x)$ containing x

977

978

$$\lim_{n \rightarrow \infty} \sup \mu^n(N(x)) > 0. \quad (2.69)$$

In other words, if (X_n) is a sequence of independent S -valued random variables, each with distribution μ , and if Z_n is defined by $Z_n = X_n X_{n-1} \cdots X_1$, then Z_n has distribution μ^n and (2.69) implies that

979

980

981

$$\Pr(Z_n \in N(x) \text{ infinitely often}) > 0.$$

We prove in Chap. 3 that this infinitely often probability is either 0 or 1 when S is a group, and therefore,

$$\Pr(Z_n \in N(x) \text{ infinitely often}) = 1 \quad (2.70)$$

for any open $N(x)$ containing x . Now let e be the identity of S and $y \in S_0$. Let N be a given open set such that $e \in N$. There exist open sets N_1 and N_2 containing e such that

$$N_1 N_1^{-1} \subset N, \quad N_2 x y^{-1} \subset x y^{-1} N_1.$$

Now there exists a positive integer k such that $\mu^k(N_1 y) > 0$, since $y \in S_0$. Then we have

$$\begin{aligned} & \{Z_{k+n} \in N_2 x \text{ infinitely often}\} \\ & \subset \{Z_k \notin N_1 y\} \cup \{Z_{k+n} Z_k^{-1} \in x y^{-1} N \text{ infinitely often}\}, \end{aligned}$$

which means that

$$\Pr(Z_n \in x y^{-1} N \text{ infinitely often}) > 0$$

so that $x y^{-1} \in S_0$, and it is a recurrent state for the random walk (Z_n) whenever $y \in S_0$. Thus, $e \in S_0$ and it is recurrent. Then for any $y \in S_0$, repeating the previous argument, $e \cdot y^{-1} \in S_0$. It is also clear that S_0 is a closed semigroup, so that $S_0 = S$.

Now by Theorem 2.8, since μ^n does not converge vaguely to the zero measure, there exist elements x_n in S such that for each $k \geq 0$,

$$\mu^{n-k} * \delta_{x_n} \rightarrow \nu_k \in P(S).$$

Then there exists n_0 such that

$$n \geq n_0 \Rightarrow \mu^n(K_1 x_n^{-1}) > 1 - \delta,$$

for some compact set K_1 ; it follows from (2.68) that

$$K_1 x_n^{-1} \cap K \neq \emptyset,$$

for infinitely many n . This means that $x_n \in K^{-1} K_1$, which is compact, for infinitely many n . Let z be a limit point of the sequence x_n , so that $\nu_0 * \delta_{z^{-1}}$ is a weak limit point of (μ^n) . By Lemma 2.2, there is a subsequence n_t such that for $k \geq 0$,

$$\begin{aligned} \mu^{n_t-k} & \xrightarrow{w} \lambda_k \in P(S), \\ \lambda_{n_t} & \xrightarrow{w} \lambda = \lambda * \lambda \in P(S). \end{aligned} \quad (2.71)$$

It follows from (2.71) (since S is second countable, the weak topology in $P(S)$ is metrizable) that there is a subsequence $(m_t) \subset (n_t)$ such that

$$\mu^{m_{t+1}-m_t} \xrightarrow{w} \lambda.$$

For any weak* limit point ν_k of $\mu^{m_{t+1}-m_t-k}$, we have by Lemma 2.3,

$$\mu^k * \nu_k = \lambda, \quad k \geq 1. \quad (2.72)$$

Since $S(\lambda) \equiv H$ is a compact subgroup by Theorem 2.2, there exists (by (2.72)) $x_k \in S(\nu_k)$ such that

$$S(\mu)^k \subset Hx_k^{-1};$$

also, for $k \geq 1$, there exists h_k in H such that

$$x^k = h_k x_k^{-1}, \quad S(\mu)^k \subset Hx^k, \quad x \in S(\mu). \quad (2.73)$$

By [89, p. 85], either the set $\{x^k : k \geq 1\}$ has compact closure or the sequence $x^k \rightarrow \infty$ as $k \rightarrow \infty$. It follows from (2.73) that since S_0 is noncompact, $x^k \rightarrow \infty$. But then given any compact set A ,

$$A \cap Hx^k = \emptyset$$

for all but finitely many k , since otherwise $x^k \in H^{-1}A$ for infinitely many k . Thus, $\mu^k(A) = 0$ for all but finitely many k . $\square$

Theorem 2.10 further generalizes Theorem 2.9.

Theorem 2.10. *Let S be a completely simple semigroup with (usual) product representation $X \times G \times Y$. Suppose $\mu \in P(S)$ and*

$$S = \text{cl} \left(\bigcup_{n=1}^{\infty} S(\mu)^n \right).$$

Then μ^n converges to the zero measure vaguely if and only if G is a noncompact group.

Proof. The proof consists of several steps consisting of noting that observations made earlier are still valid in this more general situation.

Step 1. Given $\epsilon > 0$, there still exist compact subsets $K_1 \subset X$, $K_2 \subset Y$ such that

$$\mu(K_1 \times G \times K_2) > \sqrt{1 - \epsilon}. \quad (2.74)$$

Notice that for $n \geq 1$,

1020

$$\begin{aligned} & (K_1 \times G \times K_2) \cdot \underbrace{S \cdot S \cdots S}_{n \text{ times}} \cdot (K_1 \times G \times K_2) \\ &= K_1 \times G \times K_2, \end{aligned}$$

which means that for $n \geq 1$,

1021

$$\mu^{n+2}(K_1 \times G \times K_2) > 1 - \epsilon. \quad (2.75)$$

Thus, if G is compact, the sequence μ^n is tight. This proves the only if part of Theorem 2.10. 1022
1023

Step 2. Let K be a compact subset of X . Then it easily verified that 1024

$$\begin{aligned} & \left[(A_1 \times A_2 \times A_3)^{-1} (B_1 \times B_2 \times B_3) \right] \cap [K \times G \times Y] \\ & \subset K \times \left[(A_3 K)^{-1} A_2^{-1} B_2 \right] \times B_3, \end{aligned} \quad (2.76)$$

which means that if A, B are compact subsets of S , then 1025

$$A^{-1}B \cap K \times G \times Y$$

is closed in S and therefore compact (from (2.76)). 1026

Step 3. Using Step 2 and the same proof as in Lemma 2.4 along with the regularity property in (2.75), it is easily verified that 1027
1028

$$\sup \left\{ \lim_{n \rightarrow \infty} \sup_{x \in S} \mu^n(Kx^{-1}) : K \text{ compact} \right\}$$

is either 0 or 1. 1029

Step 4. Suppose from now on that G is noncompact and (if possible) for some subsequence n_i and some compact subset $A \subset S$, 1030
1031

$$\mu^{n_i}(A) > \epsilon > 0, \quad i \geq 1. \quad (2.77)$$

In Step 4, we show that (μ^{n_i}) is tight. Notice that 1032

$$\mu^{n_i}(A) = \int \mu^{n_i-1}(Ax^{-1}) \mu(dx) \leq \sup_{x \in S} \mu^{n_i-1}(Ax^{-1})$$

and therefore, from Step 3, given $0 < 2\delta < \epsilon$, there is a compact set B and elements $x_{n_i} \in S$ such that 1033
1034

$$\mu^{n_i+1}(Bx_{n_i}^{-1}) > 1 - \delta, \quad i \geq 1. \quad (2.78)$$

1035

Choose the compact set $K_1 \subset X$ such that

1036

$$\mu(K_1 \times G \times Y) > 1 - \delta.$$

Then we have

1037

$$\mu^{n_i+1}(Bx_{n_i}^{-1}) = \int \mu^{n_i}(Bx_{n_i}^{-1}y^{-1})\mu(dy) \leq \int_{K_1 \times G \times Y} + \delta,$$

so that there exists $y \in K_1 \times G \times Y$ such that for $z_{n_i} = yx_{n_i} (\in K_1 \times G \times Y)$,

1038

$$\mu^{n_i}(Bz_{n_i}^{-1}) > 1 - 2\delta. \quad (2.79)$$

It follows from (2.77) and (2.79) that

1039

$$A \cap Bz_{n_i}^{-1} \neq \emptyset$$

so that

1040

$$z_{n_i} \in A^{-1}B \cap K_1 \times G \times Y \equiv C,$$

say, for $i \geq 1$. By Step 2, C is compact. Then it follows from (2.79) and (2.75),

1041

with ϵ replaced by δ , that for $i \geq 1$,

1042

$$\mu^{n_i}(BC^{-1} \cap (K_1 \times G \times K_2)) > 1 - 3\delta,$$

where $K_1 \subset X$, $K_2 \subset Y$ are compact subsets. It is easily verified that the set $BC^{-1} \cap [K_1 \times G \times K_2]$ is compact. Thus, the subsequence (μ^{n_i}) is tight.

1043

1044

Step 5. Using the same argument as in the proof of Lemma 2.2, there is a subsequence $(m_i) \subset (n_i)$ such that for each $k \geq 0$,

1045

1046

$$\mu^{m_i-k} \rightarrow \lambda_k \quad (2.80)$$

vaguely as $i \rightarrow \infty$. Let us show that $\lambda_k \in P(S)$, $k \geq 0$.

1047

Let k be a fixed positive integer. Since (μ^{m_i}) is tight, using Step 4 and (2.75), given $\epsilon > 0$, we can find compact subsets $K_1 \subset X$, $L \subset G$, $K_2 \subset Y$ such that

1048

1049

$$\mu^{m_i}(K_1 \times L \times K_2) > 1 - \epsilon, \quad i \geq 1,$$

$$\mu^k(K_1 \times L \times K_2) > 1 - \epsilon,$$

(2.81)

$$\mu^{n+2}(K_1 \times G \times K_2) > 1 - \epsilon, \quad n \geq 1.$$

Thus,

1050

$$1 - \epsilon < \int \mu^{m_i-k}((K_1 \times L \times K_2)y^{-1})\mu^k(dy)$$

or

1051

$$1 - 2\epsilon \leq \mu^{m_i-k}((K_1 \times L \times K_2)(K_1 \times L \times K_2)^{-1}).$$

It follows from (2.81) that for $i \geq 1$,

1052

$$\mu^{m_i-k} \left((K_1 \times L \times K_2) (K_1 \times L \times K_2)^{-1} \cap [K_1 \times G \times K_2] \right) > 1 - 3\epsilon.$$

The set inside the parentheses is easily verified to be compact. This proves that (μ^{m_i-k}) is tight, so that $\lambda_k \in P(S)$. 1053
1054

Notice that it follows from (2.81) that given $\epsilon > 0$, we can obtain the same compact set 1055
1056

$$A_\epsilon \equiv (K_1 \times L \times K_2) (K_1 \times L \times K_2)^{-1} \cap (K_1 \times G \times K_2)$$

such that for $i > j \geq 1$,

1057

$$\mu^{m_i-m_j} (A_\epsilon) > 1 - 3\epsilon.$$

This means that the sequence λ_{m_j} is tight. Let $(s_i) \subset (m_i)$ be such that 1058

$$\lambda_{s_i} \xrightarrow{w} \lambda \in P(S). \quad (2.82)$$

For $k < j$, (2.80) gives

1059

$$\mu^{j-k} * \mu^{m_i-j} \xrightarrow{w} \lambda_k.$$

Then

1060

$$\lambda_j * \mu^{j-k} = \mu^{j-k} * \lambda_j = \lambda_k, \quad k < j. \quad (2.83)$$

Taking $j = s_i$ in (2.83), we have:

1061

$$\lambda * \lambda_k = \lambda_k * \lambda = \lambda_k, \quad k \geq 1. \quad (2.84)$$

Therefore,

1062

$$\lambda * \lambda = \lambda. \quad (2.85)$$

Let us now show that for each $k \geq 1$, (λ_{s_i+k}) is tight. To this end, as in (2.81), given $\epsilon > 0$, we can find compact sets K_1, L', K_2 such that 1063
1064

$$\mu^{s_i+k} (K_1 \times L' \times K_2) > 1 - \epsilon, \quad i \geq 1,$$

and any given $j \geq 1$,

1065

$$1 - \epsilon < \int \mu^{s_i-(s_j+k)} ((K_1 \times L' \times K_2) y^{-1}) \mu^{s_j+k}(dy)$$

so that

1066

$$1 - 2\epsilon < \mu^{s_i-(s_j+k)} \left((K_1 \times L' \times K_2) (K_1 \times L' \times K_2)^{-1} \right).$$

If we define

1067

$$A'_\epsilon = (K_1 \times L' \times K_2) (K_1 \times L' \times K_2)^{-1} \cap (K_1 \times G \times K_2),$$

then for $i, j \geq 1$ with $s_i > s_j + k$, we have

1068

$$\mu^{s_i - (s_j + k)} (A'_\epsilon) > 1 - 3\epsilon. \quad (2.86)$$

It follows from (2.86) and (2.80) that the subsequence $(\lambda_{s_j + k})$ is tight for each $k \geq 1$. From (2.83),

1069
1070

$$\lambda_{r+k} * \mu^r = \mu^r * \lambda_{r+k} = \lambda_k \quad k \geq 1, \quad r \geq 1,$$

so that

1071

$$\lambda_{r+s_i} * \mu^r = \mu^r * \lambda_{r+s_i} = \lambda_{s_i}, \quad i \geq 1.$$

If ν_r is a weak* limit point of (λ_{r+s_i}) , then $\nu_r \in P(S)$ and

1072

$$\nu_r * \mu^r = \mu^r * \nu_r = \lambda \quad r \geq 1. \quad (2.87)$$

From (2.84),

1073

$$\lambda * \nu_r = \nu_r * \lambda = \nu_r, \quad r \geq 1. \quad (2.88)$$

Since

1074

$$S = \text{cl} \left[\bigcup_{k=1}^{\infty} S(\mu)^k \right],$$

it follows from (2.87) that $S(\lambda)$, which is a closed, completely simple subsemigroup with a compact group factor by Theorem 2.2 (λ is idempotent), must take the form

1075
1076

$$S(\lambda) = X \times H \times Y, \quad (2.89)$$

where H is a compact subgroup of G and $YX \subset H$.

1077

We claim that

1078

$$S(\mu) \subset X \times gH \times Y, \quad g \in G - H, \quad (2.90)$$

and H is a normal subgroup of G .

1079

To prove (2.90), let (x_1, g_1, y_1) be an element of $S(\mu)$ and let $(x, g, y) \in S(\nu_1)$. Then it follows from (2.86) and (2.87) that

1080
1081

$$(x_1, g_1, y_1) (x, g, y) \in X \times H \times Y,$$

which, since $YX \subset H$, implies that

1082

$$g \in Hg_1^{-1}H. \quad (2.91)$$

Using (2.91), we have

1083

$$S(\mu) \subset X \times H g_1 H \times Y \quad (2.92)$$

$$S(v_1) \subset X \times H g_1^{-1} H \times Y. \quad (2.93)$$

From (2.88), we have

1084

$$(X \times H \times Y) \cdot S(v_1) \subset S(v_1). \quad (2.94)$$

Let $(x, h_1 g_1^{-1} h_2, y) \in S(v_1)$, where h_1 and h_2 are in H . Then by (2.94),

1085

$$(X \times H \times Y) (x, h_1 g_1^{-1} h_2, y) = X \times H g_1^{-1} h_2 \times \{y\} \subset S(v_1). \quad (2.95)$$

By (2.92), a typical element in $S(\mu)$ is $(x_1, h_3 g_1 h_4, y_1)$, where h_3 and h_4 are in H .

1086

Then from (2.87) and (2.95),

1087

$$(x_1, h_3 g_1 h_4, y_1) \cdot (X \times H g_1^{-1} h_2 \times \{y\}) \subset X \times H \times Y,$$

which means that

1088

$$g_1 H g_1^{-1} \subset H \quad \text{or} \quad g_1 H \subset H g_1.$$

From (2.87), $v_1 * \mu = \lambda$, we obtain similarly

1089

$$g_1^{-1} H g_1 \subset H \quad \text{or} \quad H g_1 \subset g_1 H.$$

It follows from (2.92) that

1090

$$S(\mu) \subset X \times g_1 H \times Y, \quad (2.96)$$

Therefore for $k \geq 1$,

1091

$$S(\mu)^k \subset X \times g_1^k H \times Y. \quad (2.97)$$

Since

1092

$$S = X \times G \times Y = \text{cl} \left[\bigcup_{k=1}^{\infty} S(\mu)^k \right],$$

it follows that

1093

$$G = \text{cl} \left(\bigcup_{k=1}^{\infty} g_1^k H \right). \quad (2.98)$$

Since $g_1 H = H g_1$, (2.98) implies that H is a normal (and compact) subgroup of G .

1094

Since G is noncompact, $\{g_1^k : k \geq 1\}$ cannot have compact closure. Therefore, g_1^k

1095

must converge to infinity as $k \rightarrow \infty$, see [89, p. 85]. This means that given any

1096

compact subset $M \subset G$,

1097

$$M \cap g_1^k H \neq \emptyset, \quad (2.99)$$

for at most finitely many k , since (2.99) implies that $g_1^k \in MH^{-1}$, a compact subset of G . This contradicts (2.77) and the proof of Theorem 2.10 is complete. $\square$

Proposition 2.8. *Suppose S satisfies the following compactness condition: Kx^{-1} and $x^{-1}K$ are compact for compact K and $x \in S$. Let $\mu \in P(S)$ and*

$$S = \text{cl} \left[\bigcup_{n=1}^{\infty} S(\mu)^n \right].$$

Then the sequence

$$\mu_n \equiv \frac{1}{n} \sum_{k=1}^n \mu^k$$

converges to the zero measure vaguely if and only if S is noncompact.

(Note that in Proposition 2.8, the compactness conditions mentioned are necessary. The completely simple semigroup $X \times G \times Y$ has the compactness condition “ Kx^{-1} is compact for compact K and $x \in S$ ” when Y is compact, and it has the other compactness condition when X is compact. Yet, this semigroup can support an idempotent probability measure when G is a compact group even when X and Y are not compact (as long as X and Y can support probability measures). See Theorem 2.2.)

Proof. Because of Theorem 2.7, we need to prove only the if part. Notice that if K is a compact subset of S , then for S , Kx^{-1} and $x^{-1}K$ are both compact, and

$$\begin{aligned} x \cdot (x^{-1}K)^c &\subset K^c, \\ (Kx^{-1})^c \cdot x &\subset K^c. \end{aligned} \tag{2.100}$$

Now consider the one-point compactification $S^* = S \cup \{\infty\}$ of S . For $x \in S$, let us define,

$$x \cdot \infty = \infty \cdot x = \infty, \quad \infty \cdot \infty = \infty.$$

It is then clear that S^* is a compact separately continuous (that is, the maps $x \rightarrow x \cdot y$ and $y \rightarrow x \cdot y$ are continuous) semigroup. It is easily verified that $P(S)$ is also a separately continuous semigroup with respect to convolution and weak topology. We now consider μ as a probability measure on S^* , so it is clear that S^* is the closed semigroup generated by $S(\mu)$. Then using the same procedure as in Theorem 2.7, we observe that there is an idempotent probability measure λ such that $\mu_n \rightarrow \lambda$ weakly as $n \rightarrow \infty$. This means

$$\mu * \lambda = \lambda * \mu = \lambda, \tag{2.101}$$

which implies that $S(\lambda)$ is an ideal of S^* , thus $\infty \in S(\lambda)$. Also, it can be proven (following the proof of Theorem 2.2) that even though S^* is only separately

continuous, $S(\lambda)$ contains a kernel of S^* , which is dense in $S(\lambda)$. This means that $S(\lambda) = \{\infty\}$. Thus, for compact $K \subset S$,

$$\lim_{n \rightarrow \infty} \sup \mu_n(K) \leq \lambda(K) = 0.$$

□ 1126

A stronger version of the preceding result holds under a stronger compactness assumption.

Theorem 2.11. *Let $\mu \in P(S)$ and*

$$S = \text{cl} \left[\bigcup_{n=1}^{\infty} S(\mu)^n \right].$$

Suppose the following compactness condition holds in S : For compact subsets A and B , $A^{-1}B$ and AB^{-1} are compact. Then μ^n converges vaguely to the zero measure if and only if S is noncompact.

(Note that when S is the multiplicative semigroup of $d \times d$, d finite, nonnegative matrices (with no zero rows nor zero columns) with usual topology, then S has the compactness conditions assumed in Theorem 2.11.)

Proof. Suppose there is a compact set $K \subset S$, a subsequence (n_k) , and $\epsilon > 0$ such that for $k \geq 1$,

$$\mu^{n_k}(K) > \epsilon. \quad (2.102)$$

Let us define the set $H \subset S$ by

$$H = \left\{ x \in S : \sum_{n=1}^{\infty} \mu^n(N(x)) = \infty \text{ for any open set } N(x) \text{ containing } x \right\}.$$

Then (2.102) implies that H is nonempty. It is clear that H is an ideal of S , since if $x \in H$ and $y \in S$, then given any open set N containing xy , there are open sets $N(x)$, $N(y)$ such that $x \in N(x)$, $y \in N(y)$, and $\text{cl}[N(x)N(y)] \subset N$; therefore, if $\mu^k(N(y)) > 0$ (such a positive integer k always exists), then

$$\begin{aligned} \sum_{n=1}^{\infty} \mu^{n+k}(N) &= \int \sum_{n=1}^{\infty} \mu^n(Nz^{-1}) \mu^k(dz) \\ &\geq \int_{N(y)} \sum_{n=1}^{\infty} \mu^n(Nz^{-1}) \mu^k(dz) \\ &\geq \int_{N(y)} \sum_{n=1}^{\infty} \mu^n(N(x)) \mu^k(dz) = \infty. \end{aligned}$$

Thus, $HS \subset H$; similarly, $SH \subset H$. We claim that H is a closed completely simple kernel of S .

To prove this claim, we use an argument from Chap.3. Let $X_1, X_2, \dots$ be a sequence of independent S -valued random variables, each with distribution μ , and let $Z_n = X_1 X_2 \cdots X_n$. Then given any positive integer k , we can find a positive integer m with $1 \leq m \leq k$ such that

$$\sum_{i=0}^{\infty} \Pr(Z_{m+ik} \in N(x)) = \infty, \quad (2.103)$$

where $N(x)$ is any open set containing x , an element in H . Also, we have

$$\begin{aligned} 1 &\geq \Pr(Z_n \in N(x) \text{ finitely often}) \\ &\geq \sum_{i=0}^{\infty} \Pr(Z_{m+ik} \in N(x), Z_n \notin N(x) \text{ for } n \geq m + (i+1)k) \\ &\geq \sum_{i=0}^{\infty} \Pr(Z_{m+ik} \in N(x), X_{m+ik+1} \cdots X_n \notin N(x)^{-1}N(x) \\ &\quad \text{for all } n \geq m + (i+1)k) \\ &= \Pr(Z_n \notin N(x)^{-1}N(x) \text{ for all } n \geq k) \cdot \sum_{i=0}^{\infty} \Pr(Z_{m+ik} \in N(x)). \end{aligned}$$

It follows from (2.103) that for each positive integer k ,

$$\Pr(Z_n \notin N(x)^{-1}N(x) \text{ for all } n \geq k) = 0. \quad (2.104)$$

This means that for any open set $N(x)$ (with compact closure), the set $N(x)^{-1}N(x) \neq \emptyset$. Thus, there exist

$$z_n \in N(x)^{-1}N(x) \subset \text{cl}(N(x)^{-1}N(x)),$$

which is compact. This means that there is a sequence x_n such that

$$x_n \rightarrow x, \quad x_n z_n \rightarrow x.$$

It follows that $x \in xS$. Let $y \in xS$. Then we claim that $x \in yS$ and this will prove that xS is a minimal right ideal of S . If $x \notin yS$, then by our compactness assumptions in Theorem 2.11, there must exist open sets $N(x)$ containing x and $N(y)$ containing y such that $N(y)^{-1}N(x)$ is empty. Write $y = xz$. Let $W(x)$ and $W(z)$ be open subsets containing x and z , respectively, such that

$$W(x) \subset N(x), \quad W(x)W(z) \subset N(y)$$

so that

$$W(z)^{-1}[W(x)^{-1}W(x)] = \emptyset.$$

There is a positive integer k such that

1160

$$\Pr(Z_k \in W(z)) > 0.$$

But using (2.104), we have

1161

$$\begin{aligned} \Pr(Z_k \in W(z)) &= \Pr(Z_k \in W(z), Z_n \in W(x)^{-1}W(x) \text{ infinitely often}) \\ &\leq \Pr(Z_k \in W(z), X_{k+1} \cdots X_n \in W_z^{-1}[W(x)^{-1}W(x)] \\ &\quad \text{infinitely often}) \\ &= 0 \end{aligned}$$

which is a contradiction. Thus for every $y \in xS$, $x \in xS$ and $x \in yS$ whenever $x \in H$. This means that xS is a minimal right ideal. Similarly, for $x \in H$, Sx is a minimal left ideal. By Proposition 1.6 and Theorem 1.7, S has a closed completely simple kernel I , which is the union of all minimal right ideals. It is clear that $I = H$.

To complete the proof of Theorem 2.11, we show that the sequence (μ^{n_k}) in (2.102) is tight. By Lemma 2.4, given $0 < \delta < \epsilon$, there exists a compact subset A , elements a_n in S , and a positive integer n_0 such that for $n > n_0$,

$$\mu^n(Aa_n^{-1}) > 1 - \delta. \quad (2.105)$$

By (2.102) and (2.105), we have

1169

$$Aa_{n_k}^{-1} \cap K \neq \emptyset, \quad n_k > n_0.$$

This means that $a_{n_k} \in K^{-1}A$ for $n_k > n_0$. Thus, if $B \equiv A(K^{-1}A)^{-1}$, then B is compact, and for $n_k > n_0$,

$$\mu^{n_k}(B) > 1 - \delta.$$

In other words, (μ^{n_k}) is tight. By Lemmas 2.2 and 2.3, there is a subsequence $(m_k) \subset (n_k)$ such that

$$\mu^{m_{k+1}-m_k} \rightarrow \nu = \nu * \nu \in P(S)$$

weakly as $k \rightarrow \infty$. For $x \in S(\nu)$ and any open set $N(x)$ containing x ,

1174

$$0 < \nu(N(x)) \leq \liminf_{k \rightarrow \infty} \mu^{m_{k+1}-m_k}(N(x)).$$

This implies that $S(\nu) \subset H$. Also note that $\mu * \nu = \nu * \mu$, so that $\text{cl}(S(\mu)S(\nu)) = \text{cl}(S(\nu)S(\mu))$. Since

1175
1176

$$S = \text{cl}\left(\bigcup_{n=1}^{\infty} S(\mu)^n\right),$$

it follows that $\text{cl}[S \cdot S(\nu)] = \text{cl}[S(\nu) \cdot S] \subset H$, where H is an ideal of S . Since $\text{cl}(S \cdot S(\nu))$ is also an ideal of S and H is simple, $\text{cl}(S \cdot S(\nu)) = H$. Thus, the closed completely simple semigroup H is generated by the support of $\mu * \nu$. Also

$$(\mu * \nu)^{m_k+1-m_k} = \mu^{m_k+1-m_k} * \nu \rightarrow \nu * \nu$$

weakly, so that $(\mu * \nu)^n$ does not converge to zero vaguely. By Theorem 2.10, the group factor of H is compact. Then if e is an idempotent element in H , eHe must be a compact group. Since $eSe = eHe$, we have

$$S = [e^{-1}(eHe)]e^{-1},$$

which must be compact by the assumptions in Theorem 2.11. Theorem 2.11 follows easily. $\square$

The problem of determining when the sequence (μ^n) for a given $\mu \in P(S)$ is tight is difficult. An easy way of checking simple conditions necessary and sufficient for the tightness of (μ^n) is still not known (except in some special cases). However, we have the following interesting Theorem 2.12.

Theorem 2.12. *Let $\mu \in P(S)$ and*

$$S = \text{cl} \left[\bigcup_{n=1}^{\infty} S(\mu)^n \right].$$

Then the sequence (μ^n) is tight if and only if S has a closed completely simple minimal ideal H with a compact group factor such that for every open set $V \supset H$,

$$\lim_{n \rightarrow \infty} \mu^n(V) = 1. \quad (2.106)$$

Proof. The only if part follows from Theorem 2.7(i) and Assertion (2.35). The if part is proven in two steps.

Step 1. Suppose there is a positive integer k such that $\mu^k(H) > 0$. In this case, by Proposition 2.7,

$$\lim_{n \rightarrow \infty} \mu^n(H) = 1.$$

Given $\epsilon > 0$, there exists a positive integer m such that for $n \geq m$,

$$\mu^n(H) > 1 - \epsilon.$$

Let us write $H = X \times G \times Y$, the usual product representation of H with G compact. Let $A \subset X$ and $B \subset Y$ be compact subsets such that

$$\mu^k(A \times G \times B) > 1 - \epsilon.$$

Thus if $n \geq m$, then we have

1200

$$\begin{aligned}\mu^{n+2k}(A \times G \times B) &\geq \mu^k(A \times G \times B)\mu^n(X \times G \times Y)\mu^k(A \times G \times B) \\ &> (1 - \epsilon)^3,\end{aligned}$$

since $A \times G \times B = (A \times G \times B)(X \times G \times Y)(A \times G \times B)$. It follows that (μ^n) is tight. 1201
1202

Step 2. Let us now assume that 1203

$$\mu^k(H) = 0, \quad k \geq 1. \quad (2.107)$$

Even though it is a standard result, let us first show that the sequence (μ^n) is tight whenever every weak* limit point of (μ^n) is a probability measure. To this end, we assume that for some subsequence (n_k) of positive integers, 1204
1205
1206

$$\mu^{n_k} \xrightarrow{w^*} \lambda \Rightarrow \lambda \in P(S), \quad (2.108)$$

Since S is second countable and locally compact, there exists an increasing sequence of open sets (V_n) such that $\bar{V}_n$ is compact for each n and 1207
1208

$$S = \bigcup_{n=1}^{\infty} V_n.$$

We claim that given $\epsilon > 0$, there exists a positive integer $k(\epsilon)$ such that for all $m \geq 1$, 1209
1210

$$\mu^m \left(\bigcup_{n=1}^{k(\epsilon)} V_n \right) > 1 - \epsilon. \quad (2.109)$$

If (2.109) is false, then there exist (p_k) and (m_k) such that $p_k < p_{k+1}$ and $m_k < m_{k+1}$, and for each $k \geq 1$, 1211
1212

$$\mu^{m_k} \left(\bigcup_{n=1}^{p_k} V_n \right) \leq 1 - \epsilon. \quad (2.110)$$

Let λ be a weak* limit point of the sequence $[\mu^{m_k}]$. Then $\lambda \in P(S)$ by (2.108). Let $(n_k) \subset (m_k)$ be such that 1213
1214

$$\mu^{n_k} \xrightarrow{w} \lambda$$

and let $(q_k) \subset (p_k)$ be such that 1215

$$\mu^{n_k} \left(\bigcup_{t=1}^{q_k} V_t \right) \leq 1 - \epsilon \quad (2.111)$$

for $k \geq 1$. If s is any positive integer, then

1216

$$\lambda \left(\bigcup_{t=1}^{q_s} V_t \right) \leq \liminf_{k \rightarrow \infty} \mu^{n_k} \left(\bigcup_{t=1}^{q_s} V_t \right) \leq 1 - \epsilon,$$

which implies that $\lambda(S) \leq 1 - \epsilon$. This contradicts $\lambda \in P(S)$ and establishes (2.109).

1217

It follows that (μ^n) is tight, since each V_n has compact closure.

1218

Thus, it is enough to show that each weak* limit point of (μ^n) is in $P(S)$. To

1219

show this, let us suppose

1220

$$\mu^{n_k} \xrightarrow{w^*} \lambda, \quad \lambda(S) = b < 1. \quad (2.112)$$

Choose $p > 0$ such that $b + 2p < 1$. Write

1221

$$H = \bigcap_{n=1}^{\infty} W_n \quad W_n \supset W_{n+1},$$

W_n open and

1222

$$\mu^n(W_n) < p, \quad n \geq 1. \quad (2.113)$$

This is possible since (2.107) holds. Since S is sigma-compact, we can write

1223

$$H = \bigcup_{m=1}^{\infty} K_m, \quad K_m \text{ compact}, \quad K_m \subset K_{m+1}.$$

Let M_i be an open set with compact closure such that $M_{i+1} \supset M_i \supset K_i$. Since we have

1224

1225

$$\limsup_{k \rightarrow \infty} \mu^{n_k}(M_i) \leq \lambda(\overline{M_i}) \leq b$$

there exists a positive integer $k(i)$ such that for $k \geq k(i)$,

1226

$$\mu^{n_k}(M_i) < b + p.$$

This means there is a subsequence (m_i) such that for $i \geq 1$,

1227

$$\mu^{m_i}(M_i) < b + p. \quad (2.114)$$

Write $C_i = M_i \cap W_{m_i}$, and let

1228

$$C = \bigcup_{n=1}^{\infty} C_i.$$

1229

Then C is open and $C \supset H$. For all $i \geq 1$, $C \subset M_i \cup W_{m_i}$ and thus,

1230

$$\mu^{m_i}(C) \leq \mu^{m_i}(M_i) + \mu^{m_i}(W_{m_i}) \leq b + 2p < 1,$$

which contradicts (2.106).

□ 1231

Under the conditions of Theorem 2.11, we saw that either the sequence μ^n converges vaguely to the zero measure or the sequence μ^n is tight. This behavior does not always hold as we already saw in Example 2.2. However, as we will show in a discrete semigroup, the convolution sequence μ^n always has this property; also in such a semigroup, it is much simpler to describe the limit behavior of this sequence with respect to weak convergence.

1232

1233

1234

1235

1236

1237

First, let S be a discrete semigroup and $\mu \in P(S)$ such that

1238

$$S = \bigcup_{n=1}^{\infty} S(\mu)^n.$$

Since $S(\mu)$ is countable, S is countable. Suppose μ^n does not converge vaguely to the zero measure. Then there is a compact set (which in this case must be finite) K such that $\mu^n(K)$ does not converge to 0 as $n \rightarrow \infty$. This means that there exists an element y in K such that

1239

1240

1241

1242

$$\lim_{n \rightarrow \infty} \sup \mu^n(y) > 0. \quad (2.115)$$

Notice that for the element y in (2.115), we have

1243

$$\sum_{n=1}^{\infty} \mu^n(y) = \infty. \quad (2.116)$$

Let

1244

$$H = \left\{ z \in S \mid \sum_{n=1}^{\infty} \mu^n(z) = \infty \right\}.$$

Then from the proof of Theorem 2.11, it follows easily that H is a completely simple ideal of S . Since the set of all z in S that has the property

1245

1246

$$\lim_{n \rightarrow \infty} \sup \mu^n(z) > 0,$$

is an ideal of S and a subset of H , H can also be identified by

1247

$$H = \left\{ z \in S \mid \lim_{n \rightarrow \infty} \sup \mu^n(z) > 0 \right\}. \quad (2.117)$$

Since H is an ideal and S is discrete, it follows from Proposition 2.7 that

1248

$$\lim_{n \rightarrow \infty} \mu^n(H) = 1. \quad (2.118)$$

Using (2.118) and proof of Theorem 2.10 and Lemma 2.4, we can also establish that the completely simple semigroup H must have its group factor compact (and therefore finite). Thus, we can now write

$$H = X \times G \times Y \quad (2.119)$$

(the usual product representation), where G is a finite group. Using (2.118) and (2.119), given $\epsilon > 0$, we can then find a positive integer N and finite subsets $A \subset X$ and $B \subset Y$ such that for $n \geq N$,

$$\mu^n(H) > 1 - \epsilon, \quad \mu^N(A \times G \times B) > 1 - \epsilon.$$

This means that for $n \geq 3N$,

$$\mu^n(A \times G \times B) \geq \mu^N(A \times G \times B) \mu^{n-2N}(H) \mu^N(A \times G \times B) > (1 - \epsilon)^3,$$

which establishes that the sequence μ^n is then tight. Now we use Theorem 2.7. Let $\eta = \eta^2$ be the idempotent weak limit point of (μ^n) . Then we have

$$S(\eta) = X \times H \times Y, \quad (2.120)$$

where H is a normal subgroup of G . Suppose that the sequence μ^n does not converge weakly. Then by Theorem 2.7,

$$\eta * \mu = \mu * \eta \neq \eta$$

and

$$S(\mu * \eta) = X \times gH \times Y, \quad g \in G - H.$$

Since G is finite, there is a smallest positive integer d such that $g^d \in H$. Notice that

$$S(\mu^d * \eta) = X \times H \times Y. \quad (2.121)$$

Write $\lambda \equiv \mu^d * \eta$. Then we have

$$\lambda * \eta = \eta * \lambda = \lambda. \quad (2.122)$$

It follows from the idempotence of η , Theorem 2.2, (2.120)–(2.122) that for $B \subset S$, $x \in S(\lambda)$,

$$\lambda(Bx^{-1}) = \int \eta(Bx^{-1}y^{-1}) \lambda(dy) = \eta(Bx^{-1})$$

and therefore, using (2.122) again,

$$\lambda(B) = \int \lambda(Bx^{-1}) \eta(dx) = \int \eta(Bx^{-1}) \eta(dx) = \eta(B)$$

so that $\mu^d * \eta = \eta$. Now let ν be any weak limit point of (μ^n) . Then there is a subsequence (n_k) of positive integers such that

$$\mu^{n_k} \rightarrow \nu \text{ weakly.} \quad (2.123)$$

Write

$$n_k = m_k \cdot d + s_k, \quad 0 \leq s_k < d.$$

Then there is a subsequence $(p_k) \subset (m_k)$ and some r , $0 \leq r < d$ such that

$$\mu^{p_k d + r} \rightarrow \nu \text{ weakly.}$$

Therefore,

$$\mu^{p_k d + r} * \eta = \mu^r * (\mu^d * \eta)^{p_k} = \mu^r * \eta = \nu * \eta = \nu.$$

In other words, the only weak limit points of (μ^n) , when μ^n does not converge weakly, are $\eta, \mu * \eta, \dots, \mu^{d-1} * \eta$, and

$$S(\mu^k * \eta) = X \times g^k H \times Y, \quad 0 \leq k \leq d. \quad (2.124)$$

Thus, we have proven Theorem 2.13.

Theorem 2.13. Let S be a discrete semigroup and $\mu \in P(S)$ such that $S = \bigcup_{n=1}^{\infty} S(\mu)^n$. Then the following assertions are valid:

- (i) μ^n does not converge to zero vaguely if and only if S has a completely simple (minimal) ideal H as described by (2.117) and (2.119) with its group factor finite (if and only if the sequence μ^n is tight).
- (ii) If μ^n does not converge vaguely to the zero measure, then there is a positive integer d such that for each integer r such that $0 \leq r < d$,

$$\mu^{nd+r} \xrightarrow{w} \nu_r,$$

where $\nu_r = \mu^r * \nu_0$ for $0 < r < d$, the supports of ν_r are pairwise disjoint for $0 \leq r < d$ and given by (2.124), and $\nu_0 = \eta$, with its support given by (2.120).

As we saw in Theorem 2.8 (and in the remark following this theorem), when S is a noncompact group such that S has no proper closed subgroup containing $S(\mu)$ with $\mu \in P(S)$, then exactly one of the following two possibilities exist:

- (i) For each compact subset $K \subset S$,

$$\lim_{n \rightarrow \infty} [\sup \{\mu^n(Kx^{-1}) : x \in S\}] = 0. \quad (2.125)$$

(ii) There exist elements $x_n \in S$ such that $\mu^n * \delta_{x_n}$ converges weakly.

1288

AQ1

(2.126) 1289

In what follows, we discuss when (2.125) occurs. First, let us look at Example 2.6: 1290

Example 2.6. (This is an example of a noncompact group where (2.125) does not 1291
hold.) Let $0 < d < 1$ and S be the multiplicative group of 2×2 matrices given by 1292

$$S = \left\{ \begin{pmatrix} d^k & r \\ 0 & 1 \end{pmatrix} : k \text{ is an integer, } r \text{ real} \right\}.$$

Let S have the usual topology of $\mathbb{R}^4$ and let $\mu \in P(S)$ such that the support of μ is 1293
the set 1294

$$\left\{ \begin{pmatrix} d & r \\ 0 & 1 \end{pmatrix} : r \text{ real} \right\}$$

and $E|X| < \infty$, where $\begin{pmatrix} d & X \\ 0 & 1 \end{pmatrix}$ has distribution μ . Notice that for an i.i.d. sequence 1295

X_n such that $\begin{pmatrix} d & X_n \\ 0 & 1 \end{pmatrix}$ has distribution μ , 1296

$$\begin{pmatrix} d & X_1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} d & X_2 \\ 0 & 1 \end{pmatrix} \cdots \begin{pmatrix} d & X_n \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} d^n & Z_n \\ 0 & 1 \end{pmatrix},$$

where $Z_n = X_1 + dX_2 + \cdots + d^{n-1}X_n$. Since $E|X_i| < \infty$ and $0 < d < 1$, the 1297
sequence Z_n is a Cauchy sequence in L_1 , so there exists Z such that $E|Z| < \infty$ and 1298

$$\lim_{n \rightarrow \infty} E|Z_n - Z| = 0.$$

In other words, the sequence Z_n has a limiting distribution. Since we have 1299

$$\begin{pmatrix} d^n & Z_n \\ 0 & 1 \end{pmatrix} \begin{pmatrix} d^{-n} & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & Z_n \\ 0 & 1 \end{pmatrix},$$

it follows that the sequence $\mu^n * \delta_{a_n}$ converges weakly if we take 1300

$$a_n = \begin{pmatrix} d^{-n} & 0 \\ 0 & 1 \end{pmatrix}.$$

We will show that the nonvalidity of (2.125) in Example 2.6 is due to the non- 1301
connectedness of S there and to the fact that S is neither discrete nor abelian. 1302

The problem when (2.125) holds is a difficult problem and has attracted the at- 1303
tention of many mathematicians. In a group S , the expression 1304

$$f_n(K) = \sup \{ \mu^n(Kx^{-1}) : x \in S \},$$

where $\mu \in P(S)$ and K is a compact set, is usually referred to as the concentration function of μ^n . (This is a term originally introduced by Paul Levy.)

If the support of μ , $\mu \in P(S)$, is contained in a coset of a compact subgroup H of a group S such that $S(\mu) \subset H \cdot x = x \cdot H$, where $x \in S(\mu)$, then it is obvious that $f_n(H) = 1$ for all n . This support condition on μ is also necessary for $f_n(K)$ not to converge to zero for some compact set K , when S is an abelian group, as Proposition 2.9 [31] shows, and also when we assume that

$$\mu * \bar{\mu} = \bar{\mu} * \mu, \text{ where } \bar{\mu}(A) = \mu(A^{-1})$$

instead of assuming that S is abelian, as Proposition 2.10 (due to Riddhi Shah) shows.

Proposition 2.9. *Let S be a noncompact abelian group. Let $\mu \in P(S)$ such that $S = \text{cl} \left(\bigcup_{n=1}^{\infty} [S(\mu) \cup S(\mu)^{-1}]^n \right)$. Then (2.125) does not hold (or equivalently, (2.3) holds) if and only if the following two conditions hold:*

- (i) S is topologically isomorphic to $Z \times H$, where Z is the discrete group of integers and H is a compact abelian group;
- (ii) $S(\mu) = \{1\} \times A$, where A is some compact subset of H such that $H = \text{cl} \left(\bigcup_{n=1}^{\infty} (A \cup A^{-1})^n \right)$.

Proof. The if part is easy. For the only if part, suppose that there exist elements x_n in S such that $\mu^n * \delta_{x_n}$ converges weakly to some λ in $P(S)$. But this means that

$$\delta_{x_n^{-1}} * \bar{\mu}^n \xrightarrow{w} \bar{\lambda},$$

where for $\nu \in P(S)$, $\bar{\nu}(B) = \nu(B^{-1})$. It follows that

$$(\mu * \bar{\mu})^n = (\mu^n * \delta_{x_n}) * (\delta_{x_n^{-1}} * \bar{\mu}^n) \xrightarrow{w} \lambda * \bar{\lambda} \equiv \beta, \text{ say.}$$

Then, $\beta = \beta * \beta$ and by Theorem 2.2, $S(\beta)$ is a compact abelian group; also,

$$\mu * \bar{\mu} * \beta = \beta$$

so that $S(\mu) \subset S(\beta) \cdot x$, for some $x \notin S(\beta)$, and consequently S is compactly generated. By [89, p. 90], S is topologically isomorphic to the direct product $R^n \times Z^m \times H$, where n, m are nonnegative integers, R is the additive group of reals, Z is the additive group of integers, H is a compact abelian group. If m, n are both positive, then identifying S with $R^n \times Z^m \times H$, it follows that

$$S(\mu) \subset \{x_1\} \times \{x_2\} \times H$$

but in this case $S(\mu)$ cannot generate all of $R^n \times Z^m \times H$, unless $n = 0$ and $m = 1$. The rest is clear. $\square$

Proposition 2.10. *Let $\mu \in P(S)$ such that S is the smallest closed group containing $S(\mu)$. Suppose that* 1332
1333

$$\mu * \bar{\mu} = \bar{\mu} * \mu.$$

Then, for any compact set A , $\lim_{n \rightarrow \infty} f_n(A) = 0$ if and only if there does not exist a compact subgroup H such that for each x in $S(\mu)$, $S(\mu) \subset H \cdot x$ and $H \cdot x = x \cdot H$. 1334
1335

Proof. The only if part is trivial. For the if part, notice that if (2.125) does not hold, 1336
then as in the proof of Proposition 2.9, 1337

$$\mu^n * \bar{\mu}^n \left(\equiv (\mu * \bar{\mu})^n, \text{ since } \mu * \bar{\mu} = \bar{\mu} * \mu \right)$$

converges weakly to some $\beta = \beta * \beta$ in $P(S)$. Then, $S(\beta) \left(\equiv H, \text{ say} \right)$ is a compact 1338
subgroup such that 1339

$$(i) \quad \mu * \bar{\mu} * \beta = \beta \quad 1340$$

$$(ii) \quad \mu * \beta * \bar{\mu} = \beta = \bar{\mu} * \beta * \mu. \quad 1341$$

This means that for any x in S_μ , $S_\mu \subset H \cdot x$ and 1342

$$xHx^{-1} \cup x^{-1}Hx \subset H.$$

The proof is now clear. □ 1343

The problem when (2.125) holds was looked into detail by Hofmann and 1344
Mukherjea in [95]. They defined a locally compact group S to be *strange* if: (i) 1345
 S has a co-compact normal subgroup H_1 ; (ii) there is a compact subgroup H of 1346
 H_1 , and an element z such that for any open set V containing H , 1347

$$H_1 \subset \bigcup_{n=1}^{\infty} z^n V z^{-n}.$$

A locally compact group S , which is not strange, was called neat. They proved that if 1348
 S is a neat group, $\mu \in P(S)$ and S is generated by $S(\mu)$ as a closed semigroup, then 1349
(2.125) holds. Hofmann and Mukherjea also showed that the class of neat groups 1350
includes all noncompact groups which are either almost connected, Lie projective, 1351
discrete, or maximally almost periodic. Their main result, which we will not prove 1352
here as the proof is too technical, is the following: 1353

Theorem 2.14. *Let S be either a noncompact abelian group or a group with a 1354
noncompact quotient which is pro-Lie or maximally almost periodic. Let $\mu \in P(S)$ 1355
such that S is the smallest closed semigroup generated by $S(\mu)$. Then (2.125) holds. 1356*

It is relevant to point out that if μ is the unit mass at 1 and S is the discrete 1357
additive group of integers, then (2.125) does *not* hold. 1358

AQ2 Hoffmann and Mukherjea also conjectured that all locally compact groups are 1359
neat. This conjecture has been proved in the affirmative by Jaworski, Rosenblatt, 1360
and Willis. 1361

When S is a noncompact group and $\mu \in P(S)$ is *strictly aperiodic* (that is, the support of μ is not contained in a coset of a proper closed normal subgroup of S), and if μ is also *adapted* (that is, S is the smallest closed subgroup containing $S(\mu)$) and S is either a Lie group or an almost connected group, then (2.125) holds. This was proven by Jaworski, Rosenblatt, and Willis, and earlier in some special cases by Lin and Wittman and also by Bartoszek. This result was also obtained in [60] with the additional condition that μ is spread out (that is, μ^n is not singular for some positive integer n).

Derriennic and Guivarc'h [59] proved the following result: If S is a nonamenable group and μ is adapted on S , then (2.125) holds.

Dani and Riddhi Shah have proven Theorem 2.15, which we only state below.

Let G be a locally compact topological group. Then if $T : G \rightarrow G$ and $\lim_{k \rightarrow \infty} T^k(g) = e = e^2 \in G \forall g \in G$, T is called a contraction on G . G is called a $\mathcal{C}$ -group if it has closed normal subgroups M , N , and H such that $M \subset N \subset H$, satisfying the following conditions:

- (i) M and H/N are compact and G/H is cyclic;
- (ii) There exists $x \in G$ such that for all $y \in N$, the map σ_x , called the conjugation action of x on N/M , defined by $\sigma_x(yM) = xyx^{-1}M$, is a contraction on N/M .

A coset $H \cdot x$, above, where x is such that the conjugation action on N/M is a contraction is called a *contractive coset* of G .

A closed subgroup of a Lie group is called a $\mathcal{C}$ -subgroup if it is a $\mathcal{C}$ -group with respect to the induced topological Lie group structure.

Theorem 2.15. (a) Let S be a real Lie group and let $\mu \in P(S)$ such that the concentration functions of μ do not converge to zero. Then S_1 , the smallest closed subgroup of S containing the support of μ , is a $\mathcal{C}$ -subgroup of S and $S(\mu)$ is contained in a contractive coset of S .
 (b) Let S be a Lie group with finitely many connected components. Suppose also that the center of S^0/R , where R is the radical of S^0 , is finite. Let $\mu \in P(S)$ be such that the concentration functions of μ do not converge to zero. Then there exists a closed subgroup C containing S_1 and a simply connected nilpotent closed normal subgroup U of C such that C/U is a direct product of a compact subgroup K and a (possibly trivial) one-parameter subgroup Φ and for $g \in S(\mu)$ the conjugation action of g on U is a contraction.

Bougerol [14] studied the rate of convergence in (2.125) for spread-out probability measures μ (that is, for some $n \geq 1$, μ^n is not singular) when S is a compact extension of a closed normal compactly generated abelian subgroup of rank r and has no proper closed subgroup containing $S(\mu)$ and such that $S(\mu)$ is not contained in a coset of a proper closed normal subgroup. He showed that for such μ and S , given any compact set $K \subset S$, there exists a constant $M > 0$ such that for $n \geq 1$,

$$\sup \{ \mu^n(xKy) : x, y \in S \} \leq \frac{M}{n^{r/2}}.$$

Clearly, the problem of the rate of convergence, though very interesting, is far from being solved even in compact groups. However, a number of interesting results exist in the literature. The next two results are due to Bhattacharya [12].

Theorem 2.16. *Let $\mu \in P(S)$, where S is a compact group such that for some positive integer p , the absolutely continuous component of μ^p (with respect to the normed Haar measure m on S) has a support whose m measure exceeds half. Then there exists $M > 0$ and $0 < \alpha < 1$ such that for all $n \geq 1$,*

$$\|\mu^n - m\| \leq M\alpha^n.$$

Proof. With no loss of generality, we assume $p = 1$. Let λ be the absolutely continuous component of μ with respect to m and f be the density $d\lambda/dm$.

Let

$$S_f = \bigcup_{n=1}^{\infty} A_n,$$

where

$$A_n = \left\{ x : f(x) \geq \frac{1}{n} \right\}.$$

Then there exists $0 < c < 1$ such that

$$m(A) > \frac{1}{2}, \quad A = \{x : f(x) \geq c\}.$$

Now notice that if $B \subset A$ and $D \subset A^c$, then $(\lambda - cm)(B) \geq 0$ and $(\lambda - cm)(D) \leq 0$. This means that

$$\begin{aligned} \|\mu - cm\| &\leq \|\mu - \lambda\| + \|\lambda - cm\| \\ &\leq 1 - \lambda(S) + \lambda(A) - cm(A) \\ &\quad + cm(A^c) - \lambda(A^c) \\ &= 1 - 2\lambda(A^c) - c[2m(A) - 1] \leq r, \end{aligned}$$

where $r = 1 - c[2m(A) - 1] < 1$. Then, if we take $v = (\mu - cm)/[1 - c]$, $v * m = m$, and

$$\begin{aligned} \mu^n &= [(1 - c)v + cm]^n \\ &= (1 - c)^n v^n + \sum_{k=0}^{n-1} \binom{n}{k} (1 - c)^k c^{n-k} m \\ &= (1 - c)^n v^n - (1 - c)^n m + m \end{aligned}$$

so that since $\|(1 - c)v\| = \|\mu - cm\| \leq r$, we have:

$$\begin{aligned} \|\mu^n - m\| &\leq (1 - c)^n + r^n \\ &\leq 2[\max\{r, 1 - c\}]^n. \end{aligned}$$

□ 1419

Corollary 2.5. Let $\mu \in P(S)$, where S is a compact connected group such that μ has a nonzero absolutely continuous component with respect to m , the normed Haar measure on S . Then, as $n \rightarrow \infty$, $\|\mu^n - m\|$ converges to zero exponentially fast.

Proof. Let f be the density of the absolute continuous component λ of μ . Let A be such that $h \equiv f \cdot I_A$ is bounded and the measure β , given by

$$\beta(B) = \int_B h dm$$

is nonzero. Let $g = h * h$. Then g is continuous (prove it) and

$$\begin{aligned} \mu^2(B) &= \int \mu(Bx^{-1}) \mu(dx) \\ &\geq \int \beta(Bx^{-1}) \beta(dx) \\ &= \int \beta(Bx^{-1}) h(x) m(dx) \\ &= \int h(x) m(dx) \left[\int I_B(yx) h(y) m(dy) \right] \\ &= \int h(x) m(dx) \left[\int I_B(y) h(yx^{-1}) m(dy) \right] \\ &= \int_B \left[\int h(yx^{-1}) h(x) m(dx) \right] m(dy) \\ &= \int g(y) m(dy) \equiv \nu(B), \text{ say.} \end{aligned}$$

Thus, $\mu^2(B) \geq \nu(B)$ and $S(\nu) \supset \{x : g(x) > 0\} \equiv V$, say. Note that g is continuous, so V is open. If B is open and B contains the identity e , then

$$\bigcup_{n=1}^{\infty} (B \cap B^{-1})^n$$

is an open subgroup and, therefore, also closed. Since S is a connected compact group, there is a positive integer p such that

$$S = \bigcup_{n=1}^{\infty} (B \cap B^{-1})^n = B^p.$$

Thus, if we take $z \in V$, then $Vz^{-1} = W$ is an open set containing e and since S is compact, there is an open set $U \subset W$ such that

$$e \in U, \quad x^{-1}Ux \subset W \quad \forall x \in S.$$

Thus, for $n \geq 1$, $z^{-n}Uz^{n+1} \subset V$. By what we proved earlier with B above, there is 1432
a positive integer k such that $U^k = S$ and, consequently, 1433

$$V^k = \underbrace{V \cdot V \cdots V}_k \supset \prod_{n=1}^k (z^{-n}Uz^{n+1}) = z^{-1}U^kz^{k+1} = S.$$

Thus, $S(V^k) = V^k = S$; therefore, S is the support of the absolutely continuous 1434
component of μ^{2k} . Corollary 2.5 follows from Theorem 2.16. $\square$ 1435

Theorem 2.17. Let $\mu \in P(S)$, where S is a compact group such that it has no 1436
proper closed subgroup containing $S(\mu)$. Suppose that there is a positive integer 1437
 k such that $(\mu^k)_a$, the absolutely continuous component of μ^k (with respect to the 1438
normed Haar measure m on S), is nonzero. Then there exists a positive integer s 1439
such that $\|\mu^{sn} - m_H\| \rightarrow 0$, as $n \rightarrow \infty$, exponentially fast, where H is the smallest 1440
closed subgroup containing $S(\mu^s)$. Moreover, the factor group G/H is finite. 1441

Note that if in Theorem 2.17 we also assume that $S(\mu)$ contains the identity e , 1442
then $H = S$ and $\|\mu^n - m\| \rightarrow 0$ exponentially fast. 1443

Proof. We may and do assume that $k = 1$. Let f be the density of μ_a . Let $A \subset S$ 1444
be such that $m(A) > 0$, and f is strictly positive and bounded on A . Let $g =$ 1445
 $(f \cdot I_A) * (f \cdot I_A)$. Then g is continuous. Let $d > 0$ be such that 1446

$$\{x \in S : g(x) > d\} \equiv J_d(g)$$

is nonempty and open. Since S is a compact group, the closed semigroup generated 1447
by $S(\mu)$ is a compact cancellative semigroup (and therefore, a group) so that 1448

$$S = \text{cl} \left[\bigcup_{n=1}^{\infty} S(\mu)^n \right],$$

and thus, there exists $p \geq 1$ and there is at least one element z in 1449

$$[J_d(g)]^{-1} \cap S(\mu)^p.$$

We claim: 1450

$$J_d(g) \subset S(\mu^2). \quad (2.127)$$

To prove (2.127), let $y \in J_d(g)$ and W be an open neighborhood of y . Then as in 1451
the proof of Corollary 2.5, 1452

$$\mu^2(W) \geq \int_W g \, dm \geq dm(W \cap J_d(g)) > 0, \quad (2.128)$$

which establishes (2.127). Let $s = p + 2$ and $V = zJ_d(g)$, with z as before. Then 1453
(2.127) implies that 1454

$$V \subset S(\mu)^p S(\mu^2) \subset S(\mu)^s.$$

Since $z^{-1} \in J_d(g)$, V contains the identity and

1455

$$V \subset \bigcap_{n=1}^{\infty} S(\mu)^{sn}.$$

By assertion (iv) of Theorem 2.7, μ^{sn} converges weakly as $n \rightarrow \infty$ to ν , the normed Haar measure on the compact subgroup

1456

1457

$$H = \text{cl} \left[\bigcup_{n=1}^{\infty} S(\mu)^{ns} \right].$$

Now we claim there is a positive integer r such that for any $y \in H$,

1458

$$\mu^{rs+s-2} \left(y [J_d(g)]^{-1} \right) > 0. \quad (2.129)$$

To prove (2.129), notice that there exist $x_1, x_2, \dots, x_q$ in H such that

1459

$$H = \bigcup_{i=1}^q x_i V.$$

Since

1460

$$e \in S(\mu)^s \subset S(\mu)^{ns} \subset S(\mu)^{(n+1)s},$$

there is a positive integer r such that $x_i \in S(\mu)^{rs}$ for each i , $1 \leq i \leq q$. Since $V = zJ_d(g)$ and $z \in S(\mu)^p$, it follows that

1461

1462

$$H \subset S(\mu)^{rs+p} J_d(g)$$

which implies (2.129).

1463

Using (2.128), for $t = (r+1)s$, $H = S(\mu^t)$, and any Borel set $B \subset H$,

1464

$$\begin{aligned} \mu^t(B) &= \int \mu^2(u^{-1}B) \mu^{t-2}(du) \\ &\geq \int_B \left[\int g(u^{-1}v) \mu^{t-2}(du) \right] m_H(dv) \\ &\geq d \int_B \mu^{t-2} \left(v [J_d(g)]^{-1} \right) m_H(dv) \end{aligned}$$

and because of (2.129), the support of $(\mu^t)_a$ is all of H . Now Theorem 2.17 follows from Theorem 2.16. $\square$

1465

1466

We end this section showing how the weak convergence of a tight sequence (μ^n) on a semigroup can be reduced to a similar problem for an induced sequence $(\tilde{\mu}^n)$ on a group.

1467

1468

1469

In the rest of this section, μ is a probability measure on the Borel subsets of a locally compact second countable semigroup S . Let $S = \text{cl} \left(\bigcup_{n=1}^{\infty} S(\mu)^n \right)$. We assume that (μ^n) is tight. By Theorem 2.7 ((i) and (ii)), S then admits a completely simple minimal ideal K , where

$$K = X \times H \times Y, \quad Y.X \subset H,$$

$H = eSe$ is a compact group with identity e , e is an arbitrarily chosen, but fixed idempotent in K , and X (resp. Y) is the closed left-zero (resp. right-zero) semigroup of idempotents in Ke (resp. eK).

Let N be the compact normal subgroup generated by the set $YX (\subset H)$. Now we define a continuous homomorphism Φ from S to the factor group $\frac{H}{N}$ as follows.

Let $x \in S$, $y \in S$. Set $\Phi(x) = (exe)N$ and $\Phi(y) = (eye)N$. Write: $ex = agb$ and $ye = a'g'b'$, where a and a' belong to X , b and b' belong to Y , and g and g' to H . Then we have:

$$\begin{aligned} e(xy)e &= ea(g(ba')g')b'e = e(g(ba')g')e \\ &= g(ba')g' \in gN.N.g'N \\ &= (exe)N.(eye)N. \end{aligned}$$

This proves that Φ is a homomorphism. With quotient topology in $\frac{H}{N}$, the continuity of Φ follows since it is the composition of the canonical homomorphism from H to $\frac{H}{N}$ and the continuous map $s \rightarrow ese$ from S to H .

Define for any Borel subset $B \subset \frac{H}{N}$,

$$\tilde{\mu}(B) = \mu(\Phi^{-1}(B)).$$

It is verified easily that

$$S(\tilde{\mu}) = \overline{\Phi(S(\mu))},$$

and

$$(\tilde{\mu})^n = (\mu^n)^{\sim}.$$

Here, $(\mu^n)^{\sim}(B) = \mu^n(\Phi^{-1}(B))$. All of these lead to

Theorem 2.18. *The sequence (μ^n) on S converges weakly if and only if the sequence $(\tilde{\mu}^n)$ on $\frac{H}{N}$ converges weakly.*

Proof. Suppose μ^n converges weakly. If $\tilde{\mu}^n$ is not convergent, then, by Theorem 2.7, $S(\tilde{\mu})$ lies in a proper coset $r\left(\frac{H_1}{N}\right)$ of $\frac{H}{N}$. This will then imply that $eS(\mu)e \subset rH_1$, since $\Phi(S(\mu)) \subset S(\tilde{\mu})$, and $\frac{H_1}{N}$ is a proper closed normal subgroup of $\frac{H}{N}$. This contradicts Theorem 2.7(iii).

Conversely, suppose that $\tilde{\mu}^n$ converges weakly. Then, if μ^n is not convergent, by Theorem 2.7(iii), the set $eS(\mu)e$ lies in a proper coset gH_1 of H , where H_1 is a

proper closed normal subgroup of H , $g \notin H_1$, and $YX \subset H_1$. Since $N \subset H_1 \subset H$ and $g \notin H_1$, we must have: $S(\tilde{\mu}) = \Phi(S(\mu)) = eS(\mu)eN \subset gH_1 = \Phi(g) \cdot \frac{H_1}{N}$, which is a proper coset of $\frac{H}{N}$. This is a contradiction by Corollary 2.1. $\square$

Here is an example illustrating Theorem 2.18.

Let I , A , and B be the 2×2 blocks

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \text{ and } \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$

respectively. Consider the set of 4×4 matrices given by

$$S = \{a, b, c, d, e, f, g, h\},$$

where

$$a = \begin{pmatrix} B & I \\ B & I \end{pmatrix}, b = \begin{pmatrix} B & I \\ B & A \end{pmatrix}, c = \begin{pmatrix} B & A \\ B & A \end{pmatrix}, d = \begin{pmatrix} B & A \\ B & I \end{pmatrix},$$

$$e = \begin{pmatrix} I & B \\ I & B \end{pmatrix}, f = \begin{pmatrix} I & B \\ A & B \end{pmatrix}, g = \begin{pmatrix} A & B \\ I & B \end{pmatrix}, h = \begin{pmatrix} A & B \\ A & B \end{pmatrix}.$$

Note that $\{a, d, e, f\}$ is the set of idempotents of S , and $af = h \neq h^2 = e$. Also, choosing a as the distinguished idempotent in S , we have:

$$H \equiv aSa = \{a, c\}, \quad c^2 = a;$$

the set of idempotents in $Sa \equiv X = \{a, d\}$, $ad = a$, $da = d$; the set of idempotents in $aS \equiv Y = \{a, e\}$, $ae = e$, $ea = a$. It is easily verified that S is a completely simple semigroup and $S = XHY$. The subgroup N here turns out to be H . Thus, for any probability measure μ whose support generates S as a semigroup, the sequence μ^n converges weakly since $\tilde{\mu}^n$ converges by Theorem 2.7.

Section 2.3 Exercises

Exercise 2.16. Let S be a locally compact Hausdorff second countable group and $(\mu_n) \subset P(S)$. Suppose that (μ_n) is a periodic sequence (that is, for $k \geq 1$, $\mu_k = \mu_{k+p}$ for some fixed positive integer p). Show that for $k \geq 1$, the sequence $(\mu_{k,n})$ can only have either zero or probability measures as weak*-cluster points.

Exercise 2.17. Let S , (μ_n) and p be as in Problem 2.16. Show that $\mu_1 * \mu_2 * \cdots * \mu_n$ converges to zero vaguely (that is, in the weak*-sense) iff $S(\mu_s * \mu_{s+1} * \cdots * \mu_{s+p-1})$, for each s , $1 \leq s \leq p$, is not contained in a compact subgroup.

Exercise 2.18. Let $\mu \in P(S)$, where $s = \{x \mid 0 \leq x \leq 1\}$ with usual topology and multiplication. Show that either μ is the unit mass at 1 or μ^n converges weakly to the unit mass at 0. 1519
1520
1521

Exercise 2.19. Suppose that 1522

$$S(\mu) = \left\{ \begin{pmatrix} a & 1-a \\ b & 1-b \end{pmatrix}, \begin{pmatrix} c & 1-c \\ d & 1-d \end{pmatrix} \right\},$$

where $0 \leq b < a \leq 1$, $0 \leq d < c \leq 1$. Show that the kernel of the closed semigroup (with respect to multiplication) generated by $S(\mu)$ consists of matrices of the form $\begin{pmatrix} x & 1-x \\ x & 1-x \end{pmatrix}$, $0 \leq x \leq 1$, and thus, the group H in Theorem 2.7, in this case, is a singleton. It follows that μ^n converges weakly as $n \rightarrow \infty$. Show that the same remains true as long as $S(\mu)$ contains at least one matrix of the form $\begin{pmatrix} a & 1-a \\ b & 1-b \end{pmatrix}$, $0 \leq b < a \leq 1$. 1523
1524
1525
1526
1527
1528

Exercise 2.20. Show that $S(\lambda)$, where $\lambda = (w) \lim_{n \rightarrow \infty} \mu^n$, where μ is as in Exercise 2.19 with $a = \frac{1}{3}$, $b = 0$ and $c = 1$, $d = \frac{2}{3}$ is the Cantor subset of $[0, 1]$, where x in $[0, 1]$ is identified with the matrix $\begin{pmatrix} x & 1-x \\ x & 1-x \end{pmatrix}$. 1529
1530
1531

Exercise 2.21. Show that in Exercise 2.19 if for some probability measure λ whose support $S(\lambda)$ consists of 2×2 stochastic matrices with rank one and $\lambda * \mu = \lambda$, then $\lambda = (w) \lim_{n \rightarrow \infty} \mu^n$. 1532
1533
1534

Exercise 2.22. Let μ and λ be in $P(S)$, where S is a locally compact second countable semigroup, such that $\lambda * \lambda = \lambda$ and $\mu * \lambda = \lambda * \mu = \mu$. Let $c > 0$ and $\mu^0 \equiv \lambda$. Consider 1535
1536
1537

$$\beta_c = \sum_{k=0}^{\infty} e^{-c} \cdot \frac{c^k}{k!} \mu^k.$$

Show that β_c is a well-defined probability measure on S such that it is the n th convolution power of $\beta_{(c/n)}$. (Note: β_c is called a compound Poisson probability. It is infinitely divisible, meaning that for each natural number n , there is a probability measure (namely, $\beta_{(c/n)}$), whose n th convolution power is itself.) 1538
1539
1540
1541

Exercise 2.23. Consider the discrete semigroup of rationals under addition. Show that on this semigroup, the unit mass at 1, though an infinitely divisible probability, is not a compound Poisson probability. (See Exercise 2.22.) (Note: It was, however, shown in [142] that a probability measure on a finite semigroup is infinitely divisible iff it is compound Poisson.) 1542
1543
1544
1545
1546

Exercise 2.24. Consider the stochastic matrices

1547

$$e = \begin{pmatrix} a & 1-a & 0 \\ a & 1-a & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad e' = \begin{pmatrix} 1-a & a & 0 \\ 1-a & a & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad g = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ a & 1-a & 0 \end{pmatrix},$$

$$g' = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1-a & a & 0 \end{pmatrix}, \quad \text{and } x = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Here, $0 \leq a \leq 1$. Let $S(\mu) = \{x, g\}$ and S be the semigroup generated by $S(\mu)$. Show that the weak limit $\lambda = (w) \lim_{n \rightarrow \infty} \mu^n$ exists. Further, show that $S(\lambda) = \{e, e', g, g'\}$, and if $\mu\{g\} = b$, $\mu\{x\} = 1 - b$, $0 < b < 1$, then $\lambda(e) + \lambda(e') = \lambda(g) + \lambda(g') = \frac{1}{2}$ and $\lambda(e) = \lambda(g) = \frac{1}{2(2-b)}$.

Exercise 2.25. Let Γ be a group (with respect to convolution) of probability measures on a discrete semigroup S . Let η be the identity of Γ . Then show that

$$S(\Gamma) = \cup \{S(\lambda) : \lambda \in \Gamma\}$$

is a completely simple semigroup, and the product representations of $S(\Gamma)$ and $S(\eta)$ are, respectively, given by

$$S(\Gamma) = X \times G \times Y, \quad S(\eta) = X \times H \times Y.$$

Show also that the elements of Γ are of the form

$$\mu_1 \times \mu_2 \times \mu_3,$$

where $\mu_1 \in P(X)$, $\mu_3 \in P(Y)$, and μ_2 is the uniform distribution on a coset of the finite subgroup H of the group G .

Exercise 2.26. Let $\mu \in P(S)$, where $S = \overline{\bigcup_{n=1}^{\infty} [S(\mu)]^n}$ is a locally compact Hausdorff second countable semigroup. Suppose that S has a kernel which is completely simple and has a product representation $X \times G \times Y$, where G is a compact group and $YX = G$. Show that μ^n converges weakly iff (μ^n) is a tight sequence.

Exercise 2.27. Let μ be a probability measure such that

$$S(\mu) \subset \left\{ \begin{pmatrix} x & x-1 & x-1 \\ x-1 & x & x-1 \\ x-1 & x-1 & x \end{pmatrix} : x \text{ real} \right\}$$

1564

Let

1565

$$e = \begin{pmatrix} \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} \end{pmatrix},$$

Prove that if $\mu^n(e) > 0$ for some $n \geq 1$, then μ^n converges weakly to the unit mass at e . (Hint: Note that $ye = ey = e$ for $y \in S(\mu)$.)

1567

Exercise 2.28. State and prove a result similar to that in Exercise 2.25 when S is nondiscrete, but compact.

1568

1569

Exercise 2.29. Let $\mu \in P(S)$, where S is a commutative semigroup (with respect to multiplication) of $d \times d$ real matrices (with usual topology) such that S is the closure of $\bigcup_{n=1}^{\infty} S(\mu)^n$. Suppose that the sequence (μ^n) is tight. Show that μ^n converges weakly iff there does not exist a closed subgroup H of K , the kernel of S (which is, in this case, a compact abelian group), and an element $g \in K - H$ such that $eS(\mu) \subset gH$, where e is the identity of K . Use this to discuss the weak convergence of the tight sequence (μ^n) , when S consists of 3×3 real circulant matrices; that is, matrices of the form

1570

1571

1572

1573

1574

1575

1576

1577

$$\begin{pmatrix} a & b & c \\ c & a & b \\ b & c & a \end{pmatrix}$$

(Hint: Consider three cases separately – when the rank of the matrices in K is 1, when it is 2, and when it is 3.)

1578

1579

Exercise 2.30. Let μ_1 and μ_2 be two idempotent probability measures on a locally compact second countable topological semigroup S such that $S(\mu_1) = S(\mu_2)$ and $\mu_1 * \mu_2 = \mu_2 * \mu_1 = \mu_1$. Show that $\mu_1 = \mu_2$.

1580

1581

1582

2.4 Weak Convergence of Convolution Products of Nonidentical Probability Measures

1583

1584

In this section, we consider the problem of weak convergence of the convolution products

1585

1586

$$\mu_{k,n} \equiv \mu_{k+1} * \cdots * \mu_n, \quad n > k$$

as $n \rightarrow \infty$, where μ_n s are in $P(S)$ and S is a topological semigroup. This problem is much more complicated than the corresponding problem when the μ_n s are identical and, not unexpectedly, it is far from being solved. However, we present a

1587

1588

1589

number of interesting results in this context when S is a compact or discrete group, 1590
and when S is a compact or discrete semigroup. Our first two results are due to 1591
Kloss [132, 133]. 1592

Theorem 2.19. *Let S be a compact connected group and $(\mu_n) \subset P(S)$. Suppose 1593
there exist $\epsilon > 0$, $\delta > 0$ such that 1594*

$$m(B) < \epsilon \Rightarrow \mu_n(B) \leq 1 - \delta \quad \forall n, \quad (2.130)$$

where m is the normed Haar measure in $P(S)$. Then, 1595

$$(w) \lim_{n \rightarrow \infty} \mu_{k,n} = m, \quad k \geq 1.$$

Proof. First suppose that $\mu \in P(S)$ and μ satisfies (2.130). We claim that μ^n con- 1596
verges weakly to m . To prove this, since 1597

$$\mu^n(B) = \int \mu(Bx^{-1}) \mu^{n-1}(dx),$$

it is clear that for each $n \geq 1$, μ^n satisfies (2.130). If μ^{n_k} converges weakly to ν , 1598
and B is a Borel set with $m(B) < \epsilon$, there is an open set $V \supset B$ such that $m(V) < \epsilon$ 1599
and, consequently, 1600

$$\nu(B) \leq \nu(V) \leq \liminf_{k \rightarrow \infty} \mu^{n_k}(V) \leq 1 - \delta;$$

this means that ν satisfies (2.130). If, furthermore, ν is idempotent, then $S(\nu) \equiv H$ 1601
is a compact subgroup and ν is the normed Haar measure on H . If H is proper, 1602
then $m(H) = 0$, since $m(H) > 0 \Rightarrow H$ has nonempty interior $\Rightarrow H$ is an open 1603
and (therefore, also) a closed subgroup $\Rightarrow H = S$, since S is connected. Since ν 1604
satisfies (2.130), ν cannot be a probability measure if H is proper. Thus, $\nu = m$ 1605
and, consequently, μ^n converges weakly to m . 1606

Now let μ_n s satisfy (2.130). By Theorem 2.8, there exist elements $a_n \in S$ such 1607
that 1608

$$\mu_1 * \mu_2 * \cdots * \mu_n * \delta_{a_n} \xrightarrow{w} \beta \in P(S).$$

Write 1609

$$\beta_k = \delta_{a_{k-1}^{-1}} * \mu_k * \delta_{a_k}, \quad k > 1,$$

and $\beta_1 = \mu_1 * \delta_{a_1}$. Then 1610

$$\beta_1 * \beta_2 * \cdots * \beta_n \xrightarrow{w} \beta.$$

Let λ be a weak limit point of (β_n) . Then $\beta * \lambda = \beta$; therefore, $\beta * \lambda^n = \beta$ for 1611
 $n \geq 1$. Since each μ_n satisfies (2.130), each β_n does and therefore, λ also satisfies 1612
(2.130). By what we have already established, then $\lambda^n \xrightarrow{w} m$, so $\beta = \beta * m = m$. 1613
This means that 1614

$$\mu_1 * \cdots * \mu_n * \delta_{a_n} \xrightarrow{w} m.$$

2.4 Weak Convergence of Convolution Products of Nonidentical Probability Measures 139

For any subsequence (n_k) such that $a_{n_k} \rightarrow a$, 1615

$$\mu_1 * \cdots * \mu_{n_k} * \delta_{a_{n_k}} * \delta_{a_{n_k}}^{-1} \xrightarrow{w} m * \delta_{a^{-1}} = m.$$

Theorem 2.19 now follows. □ 1616

The next result is interesting, though very simple. 1617

Proposition 2.11. *Let S be a compact group and $(\mu_n) \subset P(S)$ such that* 1618

$$\mu_n(B) \geq \alpha_n m(B), \quad \alpha_n > 0$$

for each Borel set $B \subset S$, where m is the normed Haar measure in $P(S)$. Then for any Borel $B \subset S$, 1619
1620

$$|\mu_1 * \mu_2 * \cdots * \mu_n(B) - m(B)| \leq \prod_{i=1}^n (1 - \alpha_i).$$

Proof. Notice that there are probability measures β_n on S such that $\mu_n = \alpha_n m + (1 - \alpha_n) \beta_n$. It is also clear that for $k \geq 0$, 1621
1622

$$\begin{aligned} \mu_{k,n} &\equiv \mu_{k+1} * \cdots * \mu_n \\ &= \prod_{i=k+1}^n (1 - \alpha_i) \cdot \beta_{k,n} \\ &\quad + \left(1 - \prod_{i=k+1}^n (1 - \alpha_i) \right) \cdot m, \end{aligned}$$

which implies Proposition 2.11. □ 1623

We now present two results along the same lines as Theorem 2.16 and Corollary 2.5. 1624
1625

Theorem 2.20. *Let S be a compact group and $\mu_n \in P(S)$, $n \geq 1$. Let $\beta \in P(S)$ and $\alpha_n > 0$ such that* 1626
1627

$$\mu_n \geq \alpha_n \cdot \beta. \quad (2.131)$$

Suppose also that $m(S(\beta_a)) > 1/2$, where m is the normed Haar measure in $P(S)$ and β_a is the absolutely continuous part of β . Then there exists $0 < d < 1$ such that for $n > k \geq 1$, 1628
1629
1630

$$\|\mu_{k,n} - m\| \leq 2 \prod_{r=k+1}^n (1 - \alpha_r \cdot d). \quad (2.132)$$

Proof. Let $f = d\beta_a/dm$. Then there exists $0 < c < 1$ such that

1631

$$m(A) > \frac{1}{2}, \quad A \equiv \{x : f(x) > c\}.$$

Notice that if $B \subset A$, then $(\beta_a - cm)(B) \geq 0$, and if $B \subset A^c$, then $(\beta_a - cm)(B) \leq 0$. Thus,

1632
1633

$$\begin{aligned} \|\mu_n - \alpha_n cm\| &\leq \|\mu_n - \alpha_n \beta_a\| + \alpha_n \|\beta_a - cm\| \\ &\leq 1 - \alpha_n \beta_a(S) + \alpha_n \beta_a(A) - \alpha_n \beta_a(A^c) - \alpha_n c[2m(A) - 1] \\ &\leq 1 - \alpha_n cq, \end{aligned}$$

where $q = 2m(A) - 1 > 0$. Now taking $v_n = [\mu_n - \alpha_n cm] / (1 - \alpha_n c)$, we have:

1634

$$\mu_n = \alpha_n cm + (1 - \alpha_n c) v_n$$

so that for $k < n$,

1635

$$\mu_{k,n} = \prod_{r=k+1}^n (1 - \alpha_r c) v_{k,n} + \left[1 - \prod_{r=k+1}^n (1 - \alpha_r c) \right] m,$$

which implies Theorem 2.20. $\square$

1636

Theorem 2.21. Let S be a compact connected group and $\mu_n \in P(S)$, $n \geq 1$. Let $\beta \in P(S)$, $\beta_a \neq 0$, and $a > 0$ such that $\mu_n \geq a\beta$. Then there exists a positive integer p and $0 < r < 1$ such that

1637
1638
1639

$$\|\mu_1 * \cdots * \mu_n - m\| \leq 2 \cdot r^{[n/p]+1}$$

for $n \geq 1$, where m is the Haar probability on S .

1640

Proof. As in Corollary 2.5, there is a positive integer p such that $S(\beta_a^p) = S$. Let us write $v_k \equiv \mu_{k,k+p}$. Then if $f \equiv (d/dm)(\beta_a^p)$, as in Corollary 2.5, there exists $0 < c < 1$ such that $m(A) > 1/2$, where $A = \{x \in S : f(x) > c\}$, and for $k \geq 0$,

1641
1642
1643

$$\begin{aligned} \|v_k - a^p cm\| &\leq \|v_k - a^p \beta_a^p\| + a^p \|\beta_a^p - cm\| \leq 1 - a^p c[2m(A) - 1] \\ &\leq 1 - a^p c \equiv r. \end{aligned}$$

As before, write

1644

$$\beta_k \equiv (v_k - a^p cm) / (1 - a^p c)$$

so that $|\beta_k| \leq 1$, and

1645

$$v_k = r\beta_k + (1 - r)m, \quad k \geq 0.$$

1646

Thus, 1647

$$\begin{aligned} & \nu_k * \nu_{k+p} * \cdots * \nu_{k+p(s-1)} \\ &= r^s \beta_k * \beta_{k+p} * \cdots * \beta_{k+p(s-1)} + (1 - r^s) m. \end{aligned}$$

This means that if 1648

$$p(s-1) \leq n < ps,$$

then 1649

$$\begin{aligned} & \|\mu_1 * \cdots * \mu_n - m\| \\ & \leq \|\mu_1 * \cdots * \mu_{p(s-1)} - m\| < 2r^s. \end{aligned} \quad \square \quad 1650$$

Note that under the conditions of Theorem 2.20, we see that when 1651

$$\sum_{n=1}^{\infty} \alpha_n = \infty,$$

for $k \geq 1$, $\mu_{k,n} \rightarrow m$ weakly and the speed of convergence is given by (2.132). 1652

In what follows, we present similar results on convergence (rather than the speed of convergence) on semigroups. 1653

We assume that S is a compact semigroup and $\mu_n \in P(S)$ for $n \geq 1$. 1655
However, compactness is not always needed, even though the assumption that 1656
 $\{\mu_{k,n} : 0 \leq k \text{ and } n > n(k) > k\}$, where $\{n(k) \mid k \geq 0\}$ is a subsequence of posi- 1657
tive integers, **is tight is always in force**. By following the proof of Lemma 2.2, we 1658
observe that for any given sequence (p_i) of positive integers, there exists a subse- 1659
quence $(n_i) \subset (p_i)$ such that for $k \geq 0$, 1660

$$\mu_{k,n_i} \rightarrow \nu_k, \quad \nu_{n_i} \rightarrow \nu = \nu^2, \quad \nu_k = \nu_k * \nu, \quad (2.133)$$

where ν_k s and ν are in $P(S)$. In what follows, whenever we write $\nu_k (\equiv \nu_k(n_i))$, 1661
we imply that (2.133) holds. 1662

First we state and prove Lemma 2.5. 1663

Lemma 2.5. *Given any two subsequences of positive integers (m_i) and (n_i) , there 1664
exist subsequences $(p_i) \subset (m_i)$ and $(q_i) \subset (n_i)$ such that if the probability mea- 1665
sures $\nu_k \equiv \nu_k(p_i)$ and $\nu'_k \equiv \nu'_k(q_i)$ are in the sense of (2.133) and such that 1666
 $\nu_{q_i} \rightarrow \lambda$ and $\nu'_{p_i} \rightarrow \lambda'$, then the following convolution equations hold: 1667*

- (i) $\lambda * \nu = \lambda$, where $\nu_{p_i} \rightarrow \nu = \nu * \nu$; 1668
- (ii) $\lambda' * \nu' = \lambda'$, where $\nu'_{q_i} \rightarrow \nu' = \nu' * \nu'$; 1669
- (iii) $\nu * \lambda' = \lambda'$; 1670
- (iv) $\nu' * \lambda = \lambda$; 1671
- (v) $\lambda' * \lambda = \nu$; 1672
- (vi) $\lambda * \lambda' = \nu'$. 1673

Proof. We show the derivation of two of the convolution equations (derivation of the others is similar). As in (2.133), we know there exist subsequences $(p_i) \subset (m_i)$ and $(q_i) \subset (n_i)$ such that for each $k \geq 0$, we have

$$\begin{aligned} \mu_{k,p_i} &\rightarrow \nu_k & \mu_{k,q_i} &\rightarrow \nu'_k \\ \nu_{p_i} &\rightarrow \nu = \nu * \nu & \nu'_{q_i} &\rightarrow \nu' = \nu' * \nu' \\ \nu_{q_i} &\rightarrow \lambda & \nu'_{p_i} &\rightarrow \lambda' \\ \nu_k * \nu &= \nu_k & \nu'_k * \nu' &= \nu'_k. \end{aligned} \quad (2.134)$$

Let j be fixed and $q_i > p_j$. Then

$$\mu_{k,p_j} * \mu_{p_j,q_i} = \mu_{k,q_i}$$

implies that

$$\mu_{k,p_j} * \nu'_{p_j} = \nu'_k$$

for each j . Taking j to infinity, we have

$$\nu_k * \lambda' = \nu'_k, \quad k \geq 0. \quad (2.135)$$

Equation (2.135) immediately gives

$$\lambda * \lambda' = \nu', \quad \nu * \lambda' = \lambda'.$$

Interchanging the roles of p_j and q_i , we obtain (iv) and (v). Equations (i) and (ii) follow from the equations $\nu_k * \nu = \nu_k$ and $\nu'_k * \nu' = \nu'_k$ in (2.30). $\square$

Lemma 2.6, among other things, shows that when S is abelian, local behavior of the convolution products $\mu_{k,n}$, as $n \rightarrow \infty$, at a single point, as described by (2.136) below actually guarantees the weak convergence of these products.

Lemma 2.6. Suppose there exists an element x in S such that given any open set $N(x)$ containing x , there exists a positive integer k_0 and some $\delta > 0$ such that for each $k \geq k_0$,

$$\lim_{n \rightarrow \infty} \inf \mu_{k,n}(N(x)) > \delta. \quad (2.136)$$

Then there is a compact subgroup H such that given any sequence (m_i) of positive integers, if $\nu_k \equiv \nu_k(p_i)$ are as in Lemma 2.5 (and as in (2.134)), then we have

$$H = eS(\nu)e = eS(\lambda)e,$$

where e is the identity of H and λ is any arbitrary weak limit point of the ν_k s. Moreover, the measures $\delta_e * \lambda$ and $\lambda * \delta_e$ are, respectively, ℓ^* -invariant and r^* -invariant.

Furthermore, over and above (2.136) if we also assume that S is abelian (or every completely simple subsemigroup of S is a group, a condition that, for example,

holds when S is the multiplicative semigroup of $d \times d$ bistochastic matrices), then 1697
the sequence $\mu_{k,n}$ converges weakly for all $k \geq 0$ as $n \rightarrow \infty$. 1698

Proof. Assume $\nu_k, \nu'_k, \nu, \nu', \lambda$, and λ' are as in Lemma 2.5. Since ν and ν' are 1699
both idempotent probability measures, their supports are both completely simple 1700
subsemigroups of S . By Condition (2.136), the element x belongs to 1701

$$S(\lambda) \cap S(\lambda') \cap S(\nu) \cap S(\nu').$$

Now $xS(\nu)x$ is a compact subgroup (since $\nu = \nu * \nu$), since $S(\nu)$ is completely 1702
simple. Let e be its identity. Then $xS(\nu)x = eS(\nu)e$. By Lemma 2.5(i), 1703

$$eS(\nu)e = xS(\nu)x \subset xS(\nu) \subset S(\lambda)S(\nu) \subset S(\lambda), \quad (2.137)$$

which implies that 1704

$$eS(\nu)e \subset eS(\lambda)e. \quad (2.138)$$

Similarly, by Lemma 2.5(iii), 1705

$$eS(\nu)e = xS(\nu)x \subset S(\nu)S(\lambda') \subset S(\lambda') \quad (2.139)$$

so that 1706

$$eS(\nu)e \subset eS(\lambda')e. \quad (2.140)$$

Thus, (2.137) and (2.139) imply that $e \in S(\lambda) \cap S(\lambda')$. Now by Lemma 2.5(v), we 1707
have 1708

$$eS(\lambda)e \subset eS(\nu)e \quad (2.141)$$

and 1709

$$eS(\lambda')e \subset eS(\nu)e. \quad (2.142)$$

Then from (2.138), (2.140)–(2.142), similarly we have 1710

$$eS(\nu)e = eS(\lambda)e = eS(\lambda')e = eS(\nu')e. \quad (2.143)$$

In the same manner, we also obtain the following equalities 1711

$$eS(\nu) = eS(\lambda); \quad (2.144)$$

$$eS(\nu') = eS(\lambda'); \quad (2.145)$$

$$S(\nu)e = S(\lambda')e; \quad (2.146)$$

$$S(\nu')e = S(\lambda)e. \quad (2.147)$$

Let $A \subset eS(\lambda) \subset S(\nu)$. Since ν is idempotent, for y, z in $S(\nu)$, 1712

$$\nu(y^{-1}z^{-1}A) = \nu(z^{-1}A).$$

By Lemma 2.5(i),

$$(\delta_e * \lambda) * \nu = \delta_e * \lambda,$$

and thus, for $z \in S(\nu)$, we have

$$\delta_e * \lambda(z^{-1}A) = \int \nu(y^{-1}z^{-1}A) \delta_e * \lambda(dy) = \nu(z^{-1}A).$$

For $z \in eS(\nu)$ and $A \subset eS(\lambda)$, $e^{-1}A \cap eS(\lambda) = A$ and $z = ez_1$ with $z_1 \in S(\nu)$, so that

$$\begin{aligned} \delta_e * \lambda(z^{-1}A) &= \nu(z^{-1}A) \\ &= \nu(z_1^{-1}e^{-1}A), \quad z = ez_1 \\ &= \nu(e^{-1}A) \\ &= \delta_e * \lambda(e^{-1}A) \\ &= \delta_e * \lambda(A). \end{aligned}$$

This means that $\delta_e * \lambda$ is ℓ^* -invariant and, therefore, it is also idempotent. Similarly, $\lambda * \delta_e$ is r^* -invariant and idempotent.

Finally, let us assume that S is a semigroup with the property that $S(\nu)$ and $S(\nu')$, being completely simple, are groups. Then by (2.143), we have

$$S(\nu) = eS(\nu)e = eS(\nu')e = S(\nu') \equiv H.$$

Consequently, $\nu = \nu' \equiv \omega_H$, the Haar probability on the compact group H . By Lemma 2.5(i),

$$\lambda * \delta_e = \lambda * (\nu * \delta_e) = \lambda * \nu = \lambda$$

since $\omega_H * \delta_e = \omega_H$. Thus, $S(\lambda) = S(\lambda) \cdot e = H$ by (2.147). Similarly using Lemma 2.5(ii) and (2.146), it follows that $S(\lambda') = H$. We then have

$$\lambda = \lambda * \nu = \lambda * \omega_H = \omega_H.$$

Similarly, $\lambda' = \omega_H$. Thus, we have

$$\nu = \nu' = \lambda = \lambda' = \omega_H. \quad (2.148)$$

Now it follows from (2.135) and (2.148) that

$$\nu_k * \nu = \nu'_k, \quad k > 0.$$

However, (2.134) gives

$$\nu_k * \nu = \nu_k, \quad k > 0.$$

Thus, $\nu_k = \nu'_k$ for all $k \geq 0$. This means that for $k \geq 0$, $\mu_{k,n}$ converges weakly as $n \rightarrow \infty$. $\square$

Lemma 2.7 below finds conditions that guarantee Condition (2.136). It then leads to the most interesting theorem of this section (Theorem 2.22).

Lemma 2.7. *Suppose there is a compact, completely simple subsemigroup S_0 with the following two properties:*

- (i) *Given any open set $U \supset S_0$, there exists $\delta > 0$ and a positive integer k_0 such that for $k \geq k_0$,*

$$\lim_{n \rightarrow \infty} \inf \mu_{k,n}(U) > \delta.$$

- (ii) *There exists some $s \in S_0$ such that for each $z \in sS_0s$ and any open set $N(z)$ containing z , there exists some $\delta > 0$ such that*

$$\lim_{n \rightarrow \infty} \inf \mu_n(N(z)) > \delta.$$

Then for $z \in sS_0s$ and any open set $N(z)$ containing z , there exists $\delta > 0$ and a positive integer k_0 such that

$$k \geq k_0 \Rightarrow \lim_{n \rightarrow \infty} \inf \mu_{k,n}(N(z)) > \delta.$$

Proof. Let us write the usual product representation of the completely simple subsemigroup S_0 as

$$S_0 = XHY \ (\equiv X \times H \times Y),$$

where $YX \subset H$, H is a compact subgroup of S_0 , $X \subset S_0$, $Y \subset S_0$ and the map $(x, h, y) \rightarrow xhy$ is one-to-one. Write $s = x_1x_2x_3$, where $s \in S_0$, $x_1 \in X$, $x_2 \in H$, $x_3 \in Y$. Let $h \in H$ and $y \in Y$. Then for $x'_2 \in H$ and $u \in X$,

$$(x_1hy) (u \cdot (yu)^{-1}h^{-1}x'_2 \cdot x_3) = x_1x'_2x_3 \equiv z.$$

Then $z \in sS_0s$ and given open $N(z)$ containing z , there exist open sets $N(x_1hy)$ containing x_1hy and also $N(x_1(yx_1)^{-1}h^{-1}x'_2x_3)$ containing the element $x_1(yx_1)^{-1}h^{-1}x'_2x_3$ (in sS_0s) such that

$$N(x_1hy) N(x_1(yx_1)^{-1}h^{-1}x'_2x_3) \subset N(z).$$

Since $sS_0 = x_1HY$ is compact, there exist $h_i \in H$, $y_i \in Y$ ($1 \leq i \leq n$) such that

$$sS_0 = x_1HY \subset \bigcup_{i=1}^n N(x_1h_iy_i) \equiv U_s.$$

Let us write:

$$z_i \equiv x_1(y_ix_1)^{-1}h_i^{-1}x'_2x_3, \quad 1 \leq i \leq n.$$

Then $z_i \in sS_0s$, $1 \leq i \leq n$. Let $z_{i1}, z_{i2}, \dots, z_{ir}$ ($r \leq n$) be all the distinct elements of the set $\{z_1, z_2, \dots, z_n\}$. Choose pairwise disjoint neighborhoods $N(z_{i1}), N(z_{i2}), \dots, N(z_{ir})$ such that

$$N(x_1 h_j y_j) N(z_{ik}) \subset N(z), \quad 1 \leq k \leq r$$

for all those j such that $z_j = z_{ik}$. Thus for $t \in N(z_{ik})$,

$$\bigcup \{N(x_1 h_j y_j) : z_j = z_{ik}\} \subset N(z)t^{-1},$$

it then follows that

$$\mu_{k,n}(N(z)) \geq \sum_{s=1}^r \int_{N(z_{is})} \mu_{k,n-1}(N(z)t^{-1}) \mu_n(dt)$$

so that there exists $\delta > 0$ such that

$$\liminf_{n \rightarrow \infty} \mu_{k,n}(N(z)) \geq \delta \liminf_{n \rightarrow \infty} \mu_{k,n-1}(U_s)$$

since by hypothesis,

$$\liminf \mu_n(N(z_{ik})) > 0, \quad 1 \leq k \leq r.$$

Now $sS_0 \subset U_s$. Since S_0 is compact and U_s is open, there exist open subsets $N(z)$ containing z and $V \supset S_0$ such that

$$N(z)V \subset U_s.$$

Since there exists k_0 and $\delta' > 0$ such that

$$k \geq k_0 \Rightarrow \mu_k(N(z)) > \delta',$$

we now have for $k \geq k_0$, $n > k + 1$,

$$\mu_{k,n-1}(U_s) \geq \mu_{k+1}(N(z))\mu_{k+1,n-1}(V) > \delta' \cdot \mu_{k+1,n-1}(V).$$

It follows that for $k \geq k_0$, $n > k + 2$,

$$\liminf_{n \rightarrow \infty} \mu_{k,n}(N(z)) \geq \delta \cdot \delta' \cdot \liminf_{n \rightarrow \infty} \mu_{k+1,n}(V).$$

Lemma 2.7 follows. □ 1763

Theorem 2.22. *Let S be a compact abelian semigroup and K its kernel. Suppose that for each $x \in K$ and any open set $N(x)$ containing x , we have* 1764 1765

$$\lim_{n \rightarrow \infty} \inf \mu_n(N(x)) > 0. \quad (*)$$

Then for all $k \geq 0$, the sequence $\mu_{k,n}$ converges weakly as $n \rightarrow \infty$. 1766

Note that Theorem 2.22 implies that if for some $\alpha > 0$ and $m \in P(S)$, $\mu_n \geq \alpha m$ for all $n \geq 1$ and if $K \subset S(m)$, then $\mu_{k,n}$ converges weakly as $n \rightarrow \infty$. 1767 1768

Proof. Notice that for $x \in K$, $xK = xS$, so that for any open set $U \supset K$, there exists an open set $N(x)$ containing x such that 1769 1770

$$N(x) \cdot S \subset U. \quad (2.149)$$

(This is possible since S is assumed compact.) By the assumption (*), there exists a positive integer k_0 and $\delta > 0$ such that 1771 1772

$$k \geq k_0 \Rightarrow \mu_k(N(x)) > \delta. \quad (2.150)$$

Then for $n - 1 > k \geq k_0$, 1773

$$\mu_{k,n}(U) \geq \mu_{k+1}(N(x))\mu_{k+1,n}(S) > \delta$$

because of (2.149) and (2.150). Thus, the conditions (i) and (ii) of Lemma 2.7 hold. Theorem 2.22 now follows from Lemmas 2.6 and 2.7. $\square$ 1774 1775

The next result is our convergence theorem in a compact semigroup. As we do not assume the abelian property here, the proof is more involved. It is also relevant to point out that this theorem says that when S is a compact group, then if for each $x \in S$ and every open set $N(x)$ containing x , 1776 1777 1778 1779

$$\lim_{n \rightarrow \infty} \inf \mu_n(N(x)) > 0,$$

then for all $k \geq 0$, $\mu_{k,n}$ converges weakly as $n \rightarrow \infty$. 1780

Theorem 2.23. *Let S be a compact (not necessarily abelian) semigroup with kernel K . Suppose that the following conditions hold:* 1781 1782

- (i) *There exists $x \in K$ such that for each $z \in xKx$ and any open set $N(z)$ containing z ,* 1783 1784

$$\lim_{n \rightarrow \infty} \inf \mu_n(N(z)) > 0.$$

- (ii) *For any closed subset $C \subset \{y \in S : y = y^2 \text{ and } y \in xK\}$ with x as in (i), $\lim_{n \rightarrow \infty} \mu_n(S \cdot C)$ exists.* 1785 1786

Then for all $k \geq 0$, $\mu_{k,n}$ converges weakly as $n \rightarrow \infty$. 1787

Proof. Notice that given any open set $U \supset K$, $xK \subset K \subset U$ so that $xS = xK \subset U$ and, consequently, there exists open $N(x)$ containing x such that

$$N(x) \cdot S \subset U$$

so that for k large and $n > k + 1$, $\exists \delta > 0$ such that

$$\mu_{k,n}(U) \geq \mu_{k+1}(N(x))\mu_{k+1,n}(S) = \mu_{k+1}(N(x)) > \delta.$$

Thus, conditions (i) and (ii) of Lemma 2.7 hold with K replacing S_0 . By Lemma 2.7, for any open set $N(z)$ containing z , an arbitrary element in xKx , there is a $\delta > 0$ such that for k sufficiently large,

$$\lim_{n \rightarrow \infty} \inf \mu_{k,n}(N(z)) > \delta.$$

By Lemma 2.6, if (n_i) and (m_i) are two given sequences of positive integers, then there exist subsequences $(p_i) \subset (n_i)$ and $(q_i) \subset (m_i)$ such that for $k \geq 0$,

$$\begin{aligned} \mu_{k,p_i} &\rightarrow v_k & \mu_{k,q_i} &\rightarrow v'_k, \\ v_{p_i} &\rightarrow v = v * v & v'_{q_i} &\rightarrow v' = v' * v', \\ v_k * v &= v_k & v'_k * v' &= v'_k, \\ v_{q_i} &\rightarrow \lambda & v'_{p_i} &\rightarrow \lambda'. \end{aligned} \quad (2.151)$$

We claim that for $k \geq 0$,

$$S(v_k) \subset K.$$

Let $k \geq 0$ and let U and V be open sets containing K such that $S.U \subset V$. Then given an positive integer m , there exist open subsets V_i , $1 \leq i \leq m$, and closed $W \subset U$ such that $K \subset V_i \subset \bar{V}_i \subset V_{i+1} \subset W$ and $S.V_i \subset V_{i+1}$, $1 \leq i \leq m$. By condition (i), there exists $\delta > 0$ and a positive integer p such that $\mu_n(V_1) > \delta$ for $n > p$. Let $r > \max\{k, p\}$. Then

$$\begin{aligned} \mu_{r,r+m}(V_m^c) &= \int_{V_{m-1}^c} \mu_{r+1}(V_m^c x^{-1}) \mu_{r+1,r+m}(dx) \\ &\leq \mu_{r+1}(V_{m-1}^c) \mu_{r+1,r+m}(V_{m-1}^c) < (1 - \delta)^m. \end{aligned}$$

It follows that $v_r(W^c) = 0$ so that $v_r(U) = 1$.

Since $\mu_{k,r} * v_r = v_k$, $v_k(V) = 1$. This establishes that $S(v_k) \subset K$. It is clear that for some $\beta_k \in P(S)$, $S(\beta_k) \subset K$ and $\mu' \in P(S)$, where

$$\mu_{p_i j} \rightarrow \mu', \quad \mu_{k,p_i j-1} \rightarrow \beta_k$$

(for each $k \geq 0$), we have

$$\beta_k * \mu' = v_k, \quad k \geq 0.$$

Now let C be a closed subset of $E(xK)$ (that is, the set of idempotents in xK). Then $S \cdot C = K \cdot C$ and

$$\nu_k(SC) = \int \beta_k(KCw^{-1})\mu'(dw).$$

Since K , $S(\nu)$, and $S(\nu')$ are all completely simple subsemigroups and ν , ν' are both idempotent, we have

$$\begin{aligned} K &= XGY \approx X \times G \times Y; \\ S(\nu) &= X_1HY_1 \approx X_1 \times H \times Y_1; \\ S(\nu') &= X_2HY_2 \approx X_2 \times H \times Y_2 \\ \nu &\approx \nu_1 \times \omega_H \times \nu_2, \quad \nu' \approx \nu'_1 \times \omega_H \times \nu'_2 \\ \nu &= \nu_1 * \omega_H * \nu_2, \quad \nu' = \nu'_1 * \omega_H * \nu'_2, \end{aligned}$$

where $\nu_1 \in P(X_1)$, $\nu_2 \in P(Y_1)$, $\nu'_1 \in P(X_2)$, $\nu'_2 \in P(Y_2)$, ω_H is the Haar probability on H . These decomposition results have already been established in earlier sections. Now notice that for $C \subset Y$, we have

$$\nu(SC) = \nu(KC) = \nu(XGC) = \nu_2(C)$$

and similarly,

$$\nu'(SC) = \nu'_2(C).$$

This means that $\nu_2 = \nu'_2$. By Lemma 2.5, we have

$$\lambda' * \nu' = \lambda'; \quad (2.152)$$

also by (2.135),

$$\nu_k * \lambda' = \nu'_k, \quad k \geq 0. \quad (2.153)$$

Then using (2.134), for all $k \geq 0$, we have

$$\begin{aligned} \nu'_k &= \nu_k * \lambda' \\ &= (\nu_k * \nu) * (\lambda' * \nu') \\ &= \nu_k * (\nu_1 * \omega_H * \nu_2) * \lambda' * (\nu'_1 * \omega_H * \nu'_2) \\ &= \nu_k * \nu_1 * \omega_H * (\nu_2 * \lambda' * \nu'_1) * \omega_H * \nu'_2 \\ &= \nu_k * \nu_1 * \omega_H * \nu'_2 \\ &= \nu_k * (\nu_1 * \omega_H * \nu_2) = \nu_k * \nu = \nu_k \end{aligned}$$

since $S(\nu_2 * \lambda' * \nu'_1) \subset H$ and $\nu_2 = \nu'_2$. (Note that by Lemma 2.7, $eS(\lambda')$ $e = H$.) It follows that for all $k \geq 0$, $\mu_{k,n}$ converges weakly. $\square$

In what follows, we describe some of the consequences when for $k \geq 0$, the sequence $\mu_{k,n}$ converges weakly to some ν_k as $n \rightarrow \infty$. In this context, our next result describes the structure of the weak limit points of the sequence (ν_k) .

Theorem 2.24. *Suppose that for each $k \geq 0$, $\mu_{k,n}$ converges weakly to $\nu_k \in P(S)$ and $\{\nu_k : k \geq 0\}$ is tight. Let us define the sets $\mathcal{F}$, $\mathcal{J}_1$ and $\mathcal{J}_2$ as follows:*

$$\mathcal{F} = \{v \in P(S) : v \text{ is a weak limit point of } (\nu_k)\},$$

$$\mathcal{J}_1 = \bigcup \{S(v) : v \in \mathcal{F}\},$$

$$\mathcal{J}_2 = \text{cl}(\mathcal{J}_1).$$

Then the following assertions hold:

- (i) $\mathcal{J}_1$ and $\mathcal{J}_2$ are both completely simple subsemigroups of S .
- (ii) Let $\mathcal{J}_1 = XHY \approx X \times H \times Y$ be the usual product representation for the completely simple subsemigroup $\mathcal{J}_1$. Then for any $v \in \mathcal{F}$,

$$S(v) = X_vHY \approx X_v \times H \times Y$$

and

$$v \approx v_1 \times \omega_H \times v_y,$$

where ω_H is the Haar probability on the compact subgroup H , $v_1 \in P(X_v)$ and $v_y \in P(Y)$.

The proof of Theorem 2.24 is omitted. See [26] for a proof.

Now we look into a useful connection between the weak convergence of the sequence $\mu_{k,n}$ for $k \geq 0$ and the almost sure convergence of the random walk $X_1X_2 \cdots X_n$, where (X_i) is an independent S -valued sequence such that X_i has distribution μ_i . To this end, let us assume that S is either abelian or a group, for $k \geq 0$, $\mu_{k,n}$ converges weakly as $n \rightarrow \infty$ to ν_k and furthermore, the sequence $\{\nu_k : k \geq 0\}$ is tight. First we claim that in this case, as $k \rightarrow \infty$, the sequence ν_k converges weakly.

To prove this claim, let (p_i) and (q_i) be sequences of positive integers such that $\nu_{p_i} \rightarrow \nu$ and $\nu_{q_i} \rightarrow \nu'$. For $k < s < n$,

$$\mu_{k,s} * \mu_{s,n} = \mu_{k,n}$$

and it follows after writing $s = p_i$ and then $s = q_i$, respectively,

$$\nu_k * \nu = \nu_k, \quad \nu_k * \nu' = \nu_k, \quad k \geq 0.$$

This means that $\nu = \nu^2$, $\nu' = \nu'^2$; also

$$\nu' * \nu = \nu', \quad \nu * \nu' = \nu. \tag{2.154}$$

By Theorem 2.2, $S(\nu)$ and $S(\nu')$ are both compact subgroups of S . When S is abelian, clearly $\nu' = \nu' * \nu = \nu * \nu' = \nu$, and when S is a group, it is clear from (2.154) that $S(\nu') \cdot S(\nu) = S(\nu')$ and $S(\nu) \cdot S(\nu') = S(\nu)$, so that $S(\nu) = S(\nu')$ and ν, ν' (where both are Haar measures on their supports which are both compact subgroups of S) are equal. Thus, ν_k converges weakly to some $\nu \in P(S)$ as $k \rightarrow \infty$. Let H be the support of ν . Then H is a compact subgroup of S . Consider the collection S/H of cosets $\{gH : g \in S\}$. Note that S/H , when equipped with quotient topology, is an abelian Hausdorff topological semigroup if S is abelian. When S is a group, S/H with quotient topology is a Hausdorff topological space. Since S is assumed to be second countable, it follows easily that S/H is also second countable. Let p be the natural map from S to S/H , i.e., $p(x) = xH$. Then we have Theorem 2.25, which may be skipped after the first reading.

Theorem 2.25. *Let S be either a topological group or a compact abelian semigroup. Let $\mu_n \in P(S)$, $n \geq 1$. Suppose that for each $k \geq 0$, $\mu_{k,n}$ converges weakly to ν_k . Then the following assertions hold:*

- (i) *There is a compact subgroup H such that $\nu_k \rightarrow \omega_H$ weakly as $k \rightarrow \infty$, where ω_H is the Haar probability on H .*
- (ii) *For each open set $U \supset H$,*

$$\sum_{n=1}^{\infty} [1 - \mu_n(UU^{-1})] < \infty.$$

Note that when S is a group, given any open set $V \supset H$, there is an open set $U \supset H$ such that $UU^{-1} \subset V$ and therefore, in this case

$$\sum_{n=1}^{\infty} [1 - \mu_n(V)] < \infty.$$

- (iii) *If $p(x) = xH$ for $x \in S$, then $p(X_1X_2 \cdots X_n)$ almost surely converges, where the X_i s are S -valued independent random variables such that X_i has distribution μ_i .*

Proof. Part (i) was proved earlier. We prove (ii) in several steps.

Step 1. Let S_0 be a compact subsemigroup of S . Let U be an open set containing S_0 . Then there is an open set V such that $S_0 \subset V$, $\bar{V} \cdot \bar{V} \subset U$ and an element $z \in U$ such that

$$VS_0^{-1} \subset Uz^{-1}. \quad (2.155)$$

When S is a group, S_0 is then necessarily a group and (2.155) is then trivial.

Now we prove (2.155) when S is an abelian semigroup. Since $S_0 \subset U$, $S_0 \cdot S_0 \subset U$, and S_0 is compact, there exists an open V such that $S_0 \subset V$, $\bar{V}$ is compact, and $\bar{V} \cdot \bar{V} \subset U$. Now the class of sets

$$\{y \cdot \bar{V} \mid y \in S_0\}$$

is contained in the compact space $S_0 \cdot \bar{V}$ and it has the finite intersection property 1874
since 1875

$$y_1 y_2 \cdots y_n \in \bigcap_{i=1}^n y_i \bar{V}$$

if each y_i is in S_0 . Hence, there exists 1876

$$z \in \bigcap_{y \in S_0} y \cdot \bar{V}.$$

Let $w \in VS_0^{-1}$. Then there exists $w' \in S_0$ such that $w \cdot w' \in V$. Now $z \in w' \bar{V}$, so 1877
that $z = w' w''$ for some $w'' \in \bar{V}$, hence 1878

$$wz = (ww') w'' \in V \cdot \bar{V} \subset U$$

or $w \in Uz^{-1}$, this establishes (2.155). 1879

Step 2. We do not need S to be abelian; however, in this step S must be either a 1880
group or a compact semigroup. For x, y in S , and an open set $N(x)$ containing x , 1881
there exist open sets $N'(x)$ and $N'(y)$ containing x and y , respectively, such that 1882

$$N'(x)N'(y)^{-1} \subset N(x)y^{-1}. \quad (2.156)$$

This is trivial when S is a group, and left as a problem when S is a compact 1883
semigroup. 1884

Step 3. Let S_0 be a compact subset of S and V an open subset such that $V \supset S_0$. 1885
Then there exists an open subset $W \supset S_0$ so that 1886

$$WW^{-1} \subset VS_0^{-1}.$$

The proof of Step 3 follows routinely from Step 2, using the compactness of S_0 . 1887
1888

Step 4. Let $1 \leq i \leq m$. Let $\{A_1, A_2, \dots, A_m\}$ and $\{B_1, B_2, \dots, B_m\}$ be two fam- 1889
ilies of events such that for each i , B_i is independent of each $A_1^c, A_2^c, \dots, A_{i-1}^c, A_i$. 1890
Then we have 1891

$$P \left(\bigcup_{i=1}^m (A_i \cap B_i) \right) \geq \left[\inf_{1 \leq i \leq m} P(B_i) \right] P \left(\bigcup_{i=1}^m A_i \right).$$

To see this, notice that 1892

$$\begin{aligned} P \left(\bigcup_{i=1}^m (A_i \cap B_i) \right) &\geq P \left((A_1 \cap B_1) \cup (A_2 \cap B_2 \cap A_1^c) \cup \cdots \right. \\ &\quad \left. \cup (A_m \cap B_m \cap A_1^c \cap \cdots \cap A_{m-1}^c) \right) \\ &= \left[\sum_{i=2}^m P(B_i) P(A_i \cap A_{i-1}^c \cap \cdots \cap A_1^c) \right] \\ &\quad + P(B_1) P(A_1). \end{aligned}$$

The proof of Step 4 is now clear. 1893

Step 5. Let us assume that for each $k \geq 1$, $\mu_{k,n} \rightarrow \nu_k$ as $n \rightarrow \infty$ and $\nu_k \rightarrow \omega_H$ as $k \rightarrow \infty$, where H is a compact subgroup of S . Given $\epsilon > 0$, let U be an open set with compact closure containing H . Then there exists k_0 such that for $k \geq k_0$,

$$\nu_k(U) > 1 - \epsilon. \quad (2.157)$$

Since for $k < m$, $\nu_k = \mu_{k,m} * \nu_m$, it follows that for $m > k \geq k_0$,

$$\mu_{k,m}(UU^{-1}) > 1 - 2\epsilon. \quad (2.158)$$

Let $X_1, X_2, \dots$ be a sequence of independent random variables with values in S such that $P(X_n \in B) = \mu_n(B)$. From Steps 1–3, there exists $z_0 \in U$ and an open set $V \supset H$ such that

$$VV^{-1} \subset Uz_0^{-1}. \quad (2.159)$$

By the same reason, there exists $z \in V$ and an open set $W \supset H$ such that

$$WW^{-1} \subset Vz^{-1}. \quad (2.160)$$

Notice that

$$\begin{aligned} \bigcup_{i=0}^{m-1} \{X_k X_{k+1} \cdots X_{k+i} \notin VV^{-1} \text{ and } X_{k+i+1} \cdots X_{k+m} \in Vz^{-1}\} \\ \subset \{X_k X_{k+1} \cdots X_{k+m} \notin Vz^{-1}\} \end{aligned} \quad (2.161)$$

since

$$\begin{aligned} X_k X_{k+1} \cdots X_{k+m} \in Vz^{-1}, \quad X_{k+i+1} \cdots X_{k+m} \in Vz^{-1} \\ \Rightarrow X_k X_{k+1} \cdots X_{k+i} \in (Vz^{-1})(Vz^{-1})^{-1} \subset VV^{-1}. \end{aligned}$$

Using (2.160) and the same method as was used in (2.158), there exists k_1 such that $r > k \geq k_1$ implies

$$\mu_{k,r}(Vz^{-1}) > 1 - \epsilon. \quad (2.162)$$

Using Step 4, (2.161) and (2.162), for $k \geq k_1$, we have

$$P\left(\bigcup_{i=0}^m \{X_k X_{k+1} \cdots X_{k+i} \notin VV^{-1}\}\right) \leq \frac{\epsilon}{1 - \epsilon} < 2\epsilon, \quad 0 < \epsilon < \frac{1}{2}.$$

Thus, for $k \geq k_1$, and all $m \geq 1$,

$$P(X_k \in VV^{-1}, X_k X_{k+1} \in VV^{-1}, \dots, X_k X_{k+1} \cdots X_{k+m} \in VV^{-1}) > 1 - 2\epsilon.$$

so that

1909

$$P(X_k \in Uz_0^{-1}, X_k X_{k+1} \in Uz_0^{-1}, \dots, X_k X_{k+1} \cdots X_{k+m} \in Uz_0^{-1}) > 1 - 2\epsilon. \quad (2.163)$$

We can now write

1910

$$P(X_k \in Uz_0^{-1} \cdots X_k X_{k+1} \cdots X_{k+m+1} \in Uz_0^{-1}) \leq \prod_{i=0}^{m+1} \mu_{k+i}(UU^{-1}). \quad (2.164)$$

(Notice that $(Uz_0^{-1})(Uz_0^{-1})^{-1} \subset UU^{-1}$, also

1911

$$\begin{aligned} & P(X_k \in Uz_0^{-1}, X_k X_{k+1} \in Uz_0^{-1}) \\ & \leq P(X_k \in Uz_0^{-1}, X_{k+1} \in (Uz_0^{-1})^{-1}(Uz_0^{-1})) \\ & \leq P(X_k \in Uz_0^{-1}) P(X_{k+1} \in U^{-1}U). \end{aligned}$$

Thus, we can use induction.)

1912

The proof of Part (ii) is now complete from (2.163) and (2.164). This step is also clear in the group case. $\square$

1913

1914

Proof (of part (iii)). There exists a sequence of open sets $V_i \supset V_{i+1}$ such that $\bar{V}_1$ is compact and

1915

1916

$$H = \bigcap_{i=1}^{\infty} V_i. \quad (2.165)$$

Let V be any open set containing H . Then there exists i such that

1917

$$V_i(V_i V_i^{-1}) \subset V. \quad (2.166)$$

Equation (2.166) is trivial when S is a group, however it is not completely obvious when S is a compact abelian semigroup, and is left as an exercise.

1918

1919

Let us now write

1920

$$A(i) = \bigcup_{k=1}^{\infty} \bigcap_{m=0}^{\infty} \{X_k X_{k+1} \cdots X_{k+m} \in V_i V_i^{-1}\},$$

where the V_i s are as in (2.165). By (2.163) and (2.164), $P(A(i)) = 1$ for $i \geq 1$. Let

1921

$$A = \bigcap_{i=1}^{\infty} A(i)$$

so that $P(A) = 1$. We show that for $\omega \in A$,

1922

$$p(X_1(\omega)X_2(\omega) \cdots X_n(\omega))$$

converges as $n \rightarrow \infty$.

1923

Notice $\{X_1(\omega) \cdots X_n(\omega) : n \geq 1\}$ is relatively compact since $V_i V_i^{-1}$ is relatively compact for each $i \geq 1$. Let $n(j)$ be a subsequence of positive integers such that

$$X_1(\omega)X_2(\omega) \cdots X_{n(j)}(\omega) \rightarrow g. \quad (2.167)$$

Let U be an open set containing H . By (2.166), there exist i_0 and $i_1 > i_0$ such that for all $i \geq i_1$,

$$V_{i_0} [V_{i_0} V_{i_0}^{-1}] \subset U$$

and

$$V_i [V_i V_i^{-1}] \subset V_{i_0}.$$

This means that when S is a compact abelian semigroup, for $i \geq i_1$,

$$\begin{aligned} [(V_i V_i^{-1})^{-1} (V_i V_i^{-1})] V_i &= V_i [\{V_i (V_i V_i^{-1})\}^{-1} V_i] \\ &\subset V_{i_0} [V_{i_0}^{-1} V_{i_0}] \\ &= V_{i_0} [V_{i_0} V_{i_0}^{-1}] \\ &\subset U. \end{aligned}$$

When S is a group, it is easy to see that this inequality holds, namely, there exists i_1 such that for $i \geq i_1$,

$$[(V_i V_i^{-1})^{-1} (V_i V_i^{-1})] V_i \subset U. \quad (2.168)$$

Now choose $i \geq i_1$. Then $\omega \in A(i)$, and therefore there exists some k such that for each $m \geq k$,

$$X_k(\omega)X_{k+1}(\omega) \cdots X_m(\omega) \in V_i V_i^{-1}.$$

Choose $m > n(j) > k$. Then we have

$$X_{n(j)+1}(\omega) \cdots X_m(\omega) \in (V_i V_i^{-1})^{-1} (V_i V_i^{-1}). \quad (2.169)$$

By (2.167), given an open set V_g containing g , there exists j_0 such that $n(j_0) > k$, and for $j \geq j_0$,

$$X_1(\omega)X_2(\omega) \cdots X_{n(j)}(\omega) \in V_g.$$

Therefore, for $m > n(j_0) > k$, from (2.169), we have

$$X_1(\omega)X_2(\omega) \cdots X_m(\omega) \in V_g \left([V_i V_i^{-1}]^{-1} (V_i V_i^{-1}) \right),$$

so for $m > n(j_0)$,

$$p(X_1(\omega) \cdots X_m(\omega)) \in p \left(V_g \left((V_i V_i^{-1})^{-1} (V_i V_i^{-1}) \right) \right) \subset p[V_g U].$$

It follows that for each $\omega \in A$, $p(X_1(\omega) \cdots X_n(\omega))$ converges.

□ 1939

In the remainder of this section, we present sufficient conditions for the weak convergence of $\mu_{k,n}$ for $k \geq 0$ when S is either a discrete group or a discrete abelian semigroup.

Lemma 2.8. *Let S be a discrete group and $(\mu_n) \subset P(S)$. Let H be a subgroup of S and $X_1, X_2, \dots$ independent S -valued random variables such that the distribution of X_n is μ_n . Then if p is the natural map (that is, $p(x) = xH$ for $x \in S$), $p(X_1 X_2 \cdots X_n)$ converges almost surely if and only if*

$$\sum_{n=1}^{\infty} \mu_n(S - H) < \infty;$$

furthermore, in this case,

$$\lim_{k \rightarrow \infty} \inf_{n > k} \mu_{k,n}(H) = 1.$$

Proof. The first part of Lemma 2.8 is immediate by the Borel–Cantelli Lemma (and since S is discrete). For the second part, notice that if

$$\sum_{n=1}^{\infty} \mu_n(S - H) < \infty,$$

then by the Borel–Cantelli Lemma,

$$P(X_n \notin H \text{ infinitely often}) = 0$$

so that given $\epsilon > 0$, there exists a positive integer N_0 such that for $n > N > N_0$,

$$P\left(\bigcap_{k=N}^n \{X_k \in H\}\right) > 1 - \epsilon$$

so that

$$\lim_{k \rightarrow \infty} \inf_{k < n} \mu_{k,n}(H) = 1.$$

□

Theorem 2.26. *Let S be a discrete group and $(\mu_n) \subset P(S)$. Then, as $n \rightarrow \infty$, the sequence $\mu_{k,n}$ converges to ν_k in $P(S)$ for each $k \geq 0$ if and only if there exists a finite subgroup H such that the following two conditions hold:*

- (i) $\sum_{n=1}^{\infty} \mu_n(S - H) < \infty;$
- (ii) If H' is a proper subgroup of H and $h_n \in S$, $n \geq 0$, then

$$\sum_{n=1}^{\infty} \mu_n(S - h_{n-1}H'h_n^{-1}) = \infty.$$

Proof. Let us assume the conditions (i) and (ii). We claim: 1959

$$\mu_{k,n} \rightarrow \nu_k \in P(S) \text{ for } k \geq 0, \quad \nu_k \rightarrow \omega_H \quad (2.170)$$

(the uniform distribution on H). 1960

To prove (2.170), notice that by Lemma 2.8 and Theorem 2.8, there exist elements x_n in S such that for $k \geq 0$, 1961
1962

$$\mu_{k,n} * \delta_{x_n} \rightarrow \pi_k \in P(S). \quad (2.171)$$

Let us write 1963

$$\delta_{x_{k-1}^{-1}} * \mu_k * \delta_{x_k} = \beta_k.$$

Then for $k \geq 0$, 1964

$$\beta_{k,n} \rightarrow \delta_{x_k^{-1}} * \pi_k \equiv \gamma_k. \quad (2.172)$$

By Theorem 2.25(i), there is a finite subgroup H_0 such that 1965

$$\gamma_k \xrightarrow{k \rightarrow \infty} \omega_{H_0} \text{ and } \gamma_k * \omega_{H_0} = \gamma_k \text{ for } k \geq 1. \quad (2.173)$$

Let y be a limit point of (x_n) . (Note that from Condition (i) and (2.171), the x_n s must be contained in some finite set, since otherwise π_k is a zero measure.) Then 1966
1967
 $\pi_k * \delta_y^{-1}$ is a weak limit point of $(\mu_{k,n})$, and by (2.171) and Condition (i), 1968

$$\pi_k(Hy) \rightarrow 1, \quad k \rightarrow \infty. \quad (2.174)$$

By (2.172), 1969

$$\gamma_k(x_k^{-1}Hy) \rightarrow 1, \text{ as } k \rightarrow \infty$$

so that by (2.173), 1970

$$\omega_{H_0}(y^{-1}Hy) = 1.$$

Consequently, 1971

$$H_0 \subset y^{-1}Hy \text{ or } yH_0y^{-1} \subset H. \quad (2.175)$$

By (2.172), (2.173), and Theorem 2.25, 1972

$$\sum_{n=1}^{\infty} \beta_n(S - H_0) < \infty,$$

which means that 1973

$$\sum_{n=1}^{\infty} \mu_n \left(S - x_{n-1}y^{-1}(yH_0y^{-1})(x_ny^{-1})^{-1} \right) < \infty.$$

By (2.175) and Condition (ii) in Theorem 2.26, it follows that

1974

$$yH_0y^{-1} = H. \quad (2.176)$$

Similarly, if z is another limit point of (x_n) , then from (2.174),

1975

$$Hy \cap Hz \neq \emptyset.$$

Therefore, $y = hz$ for some $h \in H$. By (2.172), (2.173), and (2.176), we have

1976

$$\pi_k * \omega_{y^{-1}Hy} = \pi_k. \quad (2.177)$$

Since $y = hz$, (2.177) implies that

1977

$$\begin{aligned} \pi_k * \delta_{y^{-1}} &= \pi_k * \delta_{y^{-1}} * \omega_H \\ &= \pi_k * \delta_{y^{-1}} * \omega_H * \delta_h = \pi_k * \delta_{z^{-1}}. \end{aligned} \quad (2.178)$$

It follows from (2.171) and (2.178) that $\mu_{k,n}$ converges to $\pi_k * \delta_y^{-1} \equiv v_k$, say, as $n \rightarrow \infty$. It is clear that

1978

1979

$$v_k * \omega_H = v_k \text{ and } v_k \rightarrow \omega_H.$$

This establishes (2.170).

1980

To prove the converse, let us assume that for $k \geq 0$,

1981

$$\mu_{k,n} \rightarrow v_k, \quad v_k \in P(S). \quad (2.179)$$

Then by Theorem 2.25, there is a finite subgroup H such that $v_k \rightarrow \omega_H$ as $k \rightarrow \infty$ and Condition (i) holds. Now suppose there exist elements h_n in S and a proper subgroup H' of H such that

1982

1983

1984

$$\sum_{n=1}^{\infty} \left[1 - \delta_{h_{n-1}^{-1}} * \mu_n * \delta_{h_n}(H') \right] < \infty.$$

Then by Lemma 2.8, we have

1985

$$\lim_{k \rightarrow \infty} \inf_{n > k} \mu_{k,n}(h_k H' h_n^{-1}) = 1. \quad (2.180)$$

By (2.179) and (2.180), the sequence (h_n) must be finite and, therefore, since $\mu_{k,n} \rightarrow v_k$ and $v_k \rightarrow \omega_H$, there is some h in S such that $h_n = h$ for infinitely many n and

1986

1987

1988

$$\omega_H(h H' h^{-1}) = 1,$$

so that $H \subset hH'h^{-1}$. This is a contradiction, since H' is assumed to be a proper subgroup of H , which means that the cardinality of $hH'h^{-1}$ is less than that of H .
□

Corollary 2.6. *Let S be a discrete group with identity e and $(\mu_n) \subset P(S)$. Suppose there exists $\delta > 0$ such that for each $n \geq 1$, $\mu_n(e) > \delta$. Then the following two conditions are equivalent:*

(i) *There is a finite subgroup $G \subset S$ such that*

$$\sum_{n=1}^{\infty} \mu_n(S - G) < \infty.$$

(ii) *For $k \geq 0$, the sequence $\mu_{k,n}$ converges to some ν_k in $P(S)$.*

Proof. Note that Condition (ii) implies Condition (i) by Theorem 2.26. For the converse, let us assume Condition (i). Then there is a *smallest* integer p with the property that there is a finite subgroup G' with cardinality p such that

$$\sum_{n=1}^{\infty} \mu_n(S - G') < \infty$$

If $\mu_{k,n}$ does not converge for some k , as $n \rightarrow \infty$, then by Theorem 2.26, there exist elements h_n in S and a proper subgroup H of G' such that

$$\sum_{n=1}^{\infty} \mu_n(S - h_{n-1}Hh_n^{-1}) < \infty.$$

Then there exists a positive integer N such that $n \geq N$ implies that $e \in h_{n-1}Hh_n^{-1}$ or $h_{N-1}H = h_NH = h_nH$ for $n > N$; in other words, for $n \geq N$, $h_{n-1}Hh_n^{-1}$ is a subgroup and it is the same subgroup for all such n . But since H is proper in G' , the cardinality of this subgroup is smaller than p . This is a contradiction. □

Our last result in this section concerns the weak convergence of $\mu_{k,n}$ for all $k \geq 0$, when S is a discrete abelian semigroup. For this, we assume the following condition:

(*) For $(\mu_n) \subset P(S)$, $\{\mu_n : n \geq 1\}$ is tight and there exists a sequence $n(k)$ of positive integers such that the set $\{\mu_{k,n} : k \geq 0, n > n(k)\}$ is tight.

Theorem 2.27. *Let S be a discrete abelian semigroup and $(\mu_n) \subset P(S)$. Suppose that condition (*) above holds. Suppose also that*

$$\lim_{n \rightarrow \infty} \inf \mu_n(B) > 0, \quad (2.181)$$

where

$$B = \{x \in S \mid fx = (fx)^2, \text{ whenever } f \text{ is an idempotent in } S\}.$$

Then for $k \geq 0$, $\mu_{k,n}$ converges weakly.

Notice that under condition (*), as in (2.133), given any sequence (p_i) of positive integers, there exists a subsequence $(n_i) \subset (p_i)$ such that for $k \geq 0$,

$$\left. \begin{aligned} \mu_{k,n_i} &\rightarrow \nu_k, \\ \nu_{n_i} &\rightarrow \nu = \nu^2, \\ \nu_k * \nu &= \nu_k. \end{aligned} \right\} \quad (2.182)$$

We refer to ν_k as $\nu_k(n_i)$ and ν as the tail idempotent corresponding to (n_i) . Let ν' be another tail idempotent corresponding to a different subsequence (m_i) such that

$$\mu_{k,m_i} \rightarrow \nu'_k, \quad \nu'_{m_i} \rightarrow \nu' = \nu'^2.$$

Let us now assume that S is abelian. If λ is a weak limit point of (ν_{m_i}) and λ' is a weak limit point of (ν'_{n_i}) , then as in Lemma 2.5,

$$\nu = \lambda' * \lambda = \lambda * \lambda' = \nu', \quad (2.183)$$

and ν is the Haar probability on its support that is a compact group, say, H . Thus, the tail idempotent ν is independent of whatever sequence (n_i) or (m_i) is chosen, and, therefore, unique in this sense. Also,

$$\lambda * \nu = \lambda, \quad \lambda' * \nu = \lambda'. \quad (2.184)$$

If e is the identity of H , then $\nu * \delta_e = \nu$ and

$$\lambda * \delta_e = \lambda * \nu * \delta_e = \lambda * \nu = \lambda. \quad (2.185)$$

For $x \in S(\lambda)$ and $y \in S(\lambda')$, the element $y \cdot x$ is in H by (2.183), and, therefore, there exists z in H such that

$$(zy)x = z(yx) = e. \quad (2.186)$$

Since $zy \in S(\lambda')$ by (2.184), from (2.183) and (2.185), we have

$$S(\lambda) = \overline{S(\lambda) \cdot e} = \overline{S(\lambda)(zy)x} \subset Hx.$$

By (2.184), $xH \subset S(\lambda)$; therefore, for $x \in S(\lambda)$,

$$S(\lambda) = xH. \quad (2.187)$$

Now we claim that

$$ue = u \text{ whenever } u \in S(\lambda). \quad (2.188)$$

To this end, let $u \in S(\lambda)$. By (2.187), there exists $v \in H$ such that $u = vu$. Let $w \in H$ such that $vw = wv = e$. Then

$$wu = w(vu) = eu$$

so that

$$u = vu = (ve)u = (vw)u = eu = ue.$$

Thus, (2.188) is established. It is now clear that if S_1 is the subsemigroup of S generated by

$$A = \bigcup \left\{ S(\lambda) : \begin{array}{l} \text{there is a subsequence } (n_i) \text{ such that (2.182)} \\ \text{holds and } \lambda \text{ is a weak limit point of } (v_k) \end{array} \right\},$$

then S_1 is a group. This follows from (2.183) and (2.188).

We can now prove Lemma 2.9, needed for the proof of Theorem 2.27.

Lemma 2.9. *Let S be abelian and $(\mu_n) \subset P(S)$. Then the subsemigroup S_1 (as just defined) is a group. There is a unique compact subgroup H of S_1 such that ω_H , the Haar probability on H , is the unique tail idempotent (as defined in (2.182)); moreover, if λ is a weak limit point of the v_k s in (2.182) and $\lambda \neq \omega_H$, then there exists $x \in S_1 - H$ such that $\lambda = \omega_H * \delta_x$. Furthermore, the following assertions hold:*

- (i) *Given $\epsilon > 0$ and any open set D such that $D \supset V \cdot S_1$, V is open and $V \supset S_1$, there exists a positive integer k_0 such that for each $k \geq k_0$, there exists a positive integer $n(k)$ such that for $n \geq n(k)$,*

$$\mu_{k,n}(D) > 1 - \epsilon. \quad (2.189)$$

- (ii) *Let $B = \{x \in S \mid fx = (fx)^2 \text{ whenever } f \text{ is an idempotent in } S\}$. If μ' is a weak limit point of (μ_n) , then $e_y = e$ whenever y is an element of $B \cap S(\mu')$, where e is the identity of S_1 .*
- (iii) *Suppose S is discrete. [Note that discreteness is not assumed in (i) or (ii).] Then we have*

$$\sum_{n=1}^{\infty} [1 - \delta_e * \mu_n(S_1)] < \infty,$$

where e is the identity of S_1 .

Proof. Assertion (i): Let $S_1 \subset V$, $V \cdot S_1 \subset D$, where V and D are open subsets of S . We know that given any sequence of positive integers, there is a subsequence (n_i) of positive integers such that as $i \rightarrow \infty$, for $k \geq 0$,

$$\begin{aligned} v_k * \omega_H &= v_k, \\ v_{n_i} &\rightarrow v \equiv \omega_H, \\ \mu_{k,n_i} &\rightarrow v_k. \end{aligned} \quad (2.190)$$

Let $\epsilon > 0$. Then there exists a positive integer k_0 such that

2056

$$k \geq k_0 \Rightarrow \nu_k(V) > 1 - \epsilon \quad (2.191)$$

since all weak limit points of (ν_k) have their supports inside S_1 . Now if (2.189) is false, there exists some $p \geq k_0$ and a subsequence (m_i) of positive integers such that

2057

2058

$$\begin{aligned} \mu_{p,m_i}(D) &\leq 1 - \epsilon, \quad i \geq 1; \\ \mu_{k,m_i} &\rightarrow \nu'_k \quad \forall k \geq 0. \end{aligned} \quad (2.192)$$

It follows from (2.190) and (2.192) that for some weak limit point λ' of (ν'_{n_i}) ,

2059

$$\nu_p * \lambda' = \nu'_p.$$

Since $V \cdot S_1 \subset D$ and $S(\lambda') \subset S_1$, this implies that

2060

$$\nu'_p(D) \geq \nu_p(V) > 1 - \epsilon,$$

which contradicts (2.192). This proves Assertion (i).

2061

Assertion (ii): If μ' is a weak limit point of (μ_n) , then noting (2.190) and observing that

2062

2063

$$\mu_{k,n_i} * \mu_{n_i+1} = \mu_{k,n_i+1}, \quad k < n_i,$$

we have $S(\omega_H * \mu') \subset S_1$. Thus if $x \in B \cap S(\mu')$, assuming that $B \cap S(\mu')$ is nonempty, then ex is an idempotent in S_1 by definition of B and, therefore, $ex = e$.

2064

2065

Assertion (iii): Here, we assume that S is discrete. First, we make the following observations:

2066

2067

By Assertion (i), there exist positive integers k and $n(k)$ such that

2068

$$\mu_{k,n}(S_1) > \frac{2}{3}, \quad n \geq n(k) > k, \quad (2.193)$$

$$y \in S_1 e^{-1} x^{-1}, \quad x \in S_1 e^{-1} \Rightarrow y \in S_1 e^{-1}, \quad (2.194)$$

$$y \in S_1 e^{-1} x^{-1}, \quad y \in S_1 e^{-1} \Rightarrow x \in S_1 e^{-1}. \quad (2.195)$$

Let p be any positive integer and $n > n(k)$. Then, using (2.193), (2.194), and (2.195), we have the following:

2069

2070

$$\begin{aligned} \mu_{k,n+p+1}(S_1 e^{-1}) &= \sum_{x \in S_1 e^{-1}} \mu_{k,n+p}(S_1 e^{-1} x^{-1}) \mu_{n+p+1}(x) \\ &\quad + \sum_{x \notin S_1 e^{-1}} \mu_{k,n+p}(S_1 e^{-1} x^{-1}) \mu_{n+p+1}(x) \\ &\leq \mu_{k,n+p}(S_1 e^{-1}) \mu_{n+p+1}(S_1 e^{-1}) \\ &\quad + [1 - \mu_{k,n+p}(S_1 e^{-1})][1 - \mu_{n+p+1}(S_1 e^{-1})] \end{aligned}$$

$$\begin{aligned} &\leq \mu_{k,n+p}(S_1 e^{-1}) - \frac{1}{3} [1 - \mu_{n+p+1}(S_1 e^{-1})] \\ &\leq \mu_{k,n}(S_1 e^{-1}) - \frac{1}{3} \sum_{r=n+1}^{n+p+1} [1 - \mu_r(S_1 e^{-1})]. \end{aligned}$$

The claim in Assertion (iii) follows from this inequality.

□ 2071

Proof (of Theorem 2.27). By (2.181), there exist a positive integer n_0 and $\delta > 0$ such that

$$n \geq n_0 \Rightarrow \mu_n(B) > \delta.$$

Let e be the identity of the group S_1 . If $\delta_e * \mu_n(e) < \delta/2$ for infinitely many n , then there is a weak limit point μ' of (μ_n) such that $\delta_e * \mu'(e) \leq \delta/2$.

However since (μ_n) is tight, there is a finite subset $B_1 \subset B$ such that $\mu_n(B_1) > 2/3\delta$ for $n \geq n_0$. By Assertion (ii) of Lemma 2.9, this means

$$\begin{aligned} \delta_e * \mu'(e) &= \delta_e * \mu'(e[B_1 \cap S(\mu')]) \\ &\geq \mu'(B_1) \geq \frac{2}{3}\delta, \end{aligned}$$

which is a contradiction. Thus, there exists a positive integer n_1 such that

$$n \geq n_1 \Rightarrow \delta_e * \mu_n(e) \geq \frac{\delta}{2}. \quad (2.196)$$

For $n \geq n_1$, we define the probability measure $\beta_n \in P(S_1)$ such that for $Q \subset S_1$,

$$\beta_n(Q) \cdot \delta_e * \mu_n(S_1) = \delta_e * \mu_n(Q).$$

Notice that for $Q \subset S$, by Assertion (iii) of Lemma 2.9

$$\sum_{n=n_1}^{\infty} |\beta_n(Q) - \delta_e * \mu_n(Q)| \leq \sum_{n=n_1}^{\infty} 2[1 - \delta_e * \mu_n(S_1)] < \infty,$$

since if $Q \cap S_1 = \emptyset$, $\beta_n(Q) = 0$ by definition, and $\sum_{n=1}^{\infty} \delta_e * \mu_n(Q) < \infty$. This means that whenever $\beta_{k,n}$ converges weakly as $n \rightarrow \infty$, the sequence

$$\mu_{k,n} * \delta_e = (\delta_e * \mu_{k+1}) * \cdots * (\delta_e * \mu_n)$$

also converges weakly and, consequently, $\mu_{k,n}$ converges weakly. This last statement follows, since in (2.190)

$$\nu_k * \omega_H = \nu_k \Rightarrow \nu_k * \delta_e = \nu_k$$

by Theorem 2.1. Since for each k , $\{\beta_{k,n} : n > k\}$ is tight and since we can consider the β_k s as probability measures on the group S_1 , it follows from Theorem 2.8 (using arguments similar to those used in (2.171) and (2.172)) that there exist elements x_n in S_1 such that if

$$\pi_k = \delta_{x_k^{-1}} * \beta_k * \delta_{x_k},$$

then for each $k > 0$, $\pi_{k,n}$ converges weakly so that by Theorem 2.25, there is a finite subgroup H_0 such that

$$\sum_{n=1}^{\infty} [1 - \pi_n(H_0)] < \infty,$$

which implies that

$$\sum_{n=1}^{\infty} [1 - \beta_n(H_0 g_n)] < \infty, \quad (2.197)$$

where $g_n = x_n^{-1} x_{n-1}$. By (2.196) and (2.197), $e \in H_0 g_n$ for n large, so that

$$\sum_{n=1}^{\infty} [1 - \beta_n(H_0)] < \infty. \quad (2.198)$$

Theorem 2.27 now follows from (2.196), (2.198) and Corollary 2.6. $\square$

Section 2.4 Exercises

Exercise 2.31. Let m be the Haar probability measure on a compact (Hausdorff) group G and H be a compact subgroup of G such that $m(H) > 0$. Show that H is also an open subgroup. (Hint: Consider the function f given by

$$f(x) = \int I_H(xy) I_H(y) m(dy).$$

Given $\delta > 0$, let V be an open set and N an open set such that $H \subset V$, $m(V) < m(H) + \frac{1}{2} \delta$, and $N.H \subset V$. Use the function g defined by

$$g(x) = \int I_V(xy) I_V(y) m(dy)$$

to show that f is continuous at $x = e$.)

Exercise 2.32. The well-known Birkhoff theorem states that a $d \times d$ bistochastic matrix is a convex combination of $d \times d$ permutation matrices (though this representation is not unique). (For a proof, see [151] or the paper by L. Mirsky, Proc.

Amer. Math. Soc. 9 (1958), 371–374.) Note that if $P \equiv (P_{ij})$ is a $d \times d$ stochastic matrix, then we can write: 2105
2106

$$P = \sum_{\substack{k_1, k_2, \dots, k_d \\ 1 \leq k_i \leq d}} P_{1k_1} P_{2k_2} \cdots P_{dk_d} A_{k_1 k_2 \dots k_d},$$

where $A_{k_1 k_2 \dots k_d}$ is a 0–1 stochastic matrix with 1 appearing on the i th row and the k_i th column for $1 \leq i \leq d$, and zeros elsewhere. In other words, a stochastic matrix is a convex combination of 0–1 stochastic matrices. 2107
2108
2109

If S is the finite multiplicative semigroup of $d \times d$ 0–1 stochastic matrices, and we define the probability measure μ on S by 2110
2111

$$\mu(A_{k_1 k_2 \dots k_d}) = P_{1k_1} P_{2k_2} \cdots P_{dk_d},$$

then we write: $P \leftrightarrow \mu$. If Q is another $d \times d$ stochastic matrix, and $Q \leftrightarrow \nu$, the probability measure corresponding to Q , then show that $P \cdot Q \leftrightarrow \mu * \nu$. Also, show that if $P_k \leftrightarrow \mu_k$, for $k \geq 1$, then the product $P_{k,n} \equiv P_{k+1} \cdots P_n$ converges pointwise if and only if $\mu_{k,n}$ converges weakly (that is, elementwise in this case). 2112
2113
2114
2115

Use the finite group of $d \times d$ permutation matrices, the Birkhoff theorem, and Corollary 2.6 to prove the following: 2116
2117

Let (P_n) be a sequence of $d \times d$ bistochastic matrices such that each of the d diagonal entries of each P_n is greater than some $\delta > 0$. Then the sequence $P_{k,n}$ converges elementwise as $n \rightarrow \infty$ for each $k \geq 1$. 2118
2119
2120

Exercise 2.33. Let (P_n) be a sequence of $d \times d$ stochastic matrices such that for each positive integer k , the sequence $P_{k,n}$ converges elementwise to the matrix Q_k . Show that $\lim_{n \rightarrow \infty} Q_k = Q_\infty$ exists, where Q_∞ is an idempotent bistochastic matrix. (Hint: $P_{k,m} P_{m,n} = P_{k,n}$ implies $P_{k,m} Q_m = Q_k$. Thus if Q' and Q'' are two limit points of (Q_k) , then $Q_k Q' = Q_k$ for $k \geq 1$ and therefore, $Q'' Q' = Q''$, $Q' Q' = Q'$ and $Q' Q'' = Q'$. The probability measure λ' (corresponding to Q') on the finite group of $d \times d$ permutation matrices is idempotent, and thus, the Haar (or uniform) probability measure on a finite subgroup, and so is λ'' , the probability measure corresponding to Q'' . Thus, $\lambda'' * \lambda' = \lambda''$ and $\lambda' * \lambda'' = \lambda'$.) 2121
2122
2123
2124
2125
2126
2127
2128
2129

Exercise 2.34. Let S be a locally compact Hausdorff second countable semigroup and $\mu_n \in P(S)$, $n \geq 1$. Let $\nu_k = (w) \lim_{n \rightarrow \infty} \mu_{k,n}$. Let $\mathcal{F}$ be the set of all weak limit points of $\{\nu_k : k \geq 1\}$. Suppose that $\mathcal{F}$ is nonempty. Show that every element in $\mathcal{F}$ is an idempotent in $P(S)$, and so, the closure of the union of supports of elements in $\mathcal{F}$ is a completely simple subsemigroup of S . 2130
2131
2132
2133
2134

Exercise 2.35. In Problem 2.34, let λ_1 and λ_2 be two elements in $\mathcal{F}$ such that $S(\lambda_1) \cap S(\lambda_2)$ is nonempty. Let the usual product representations of $S(\lambda_1)$ and S_0 be given by: 2135
2136
2137

$$S(\lambda_1) = X_0 \times G_0 \times Y_0, \quad S_0 = X \times G \times Y,$$

where $G_0 = eS(\lambda_1)e$, $G = eS_0e$, $e = e^2 \in S(\lambda_1)$. Show that $S(\lambda_2) = X_1 \times G_0 \times Y_0$, where $Y_0(X_0 \cup X_1) \subset G_0$. 2138
2139

Exercise 2.36. Let G be a compact Hausdorff second countable group and for each $n \geq 1$, $\mu_n \in P(G')$. Suppose that $(w) \lim_{n \rightarrow \infty} \mu_{k,n} = \nu_k$ for $k \geq 1$. Show that $(w) \lim_{k \rightarrow \infty} \nu_k = \lambda_H$ exists, where λ_H is the Haar probability measure on some compact subgroup H , and also H is the largest compact subgroup such that $\nu_k * \lambda_H = \nu_k$ for all $k \geq 1$.

Exercise 2.37. In Exercise 2.36, show also that

$$\lim_{k \rightarrow \infty} \mu_{k,k+m}(V) = 1,$$

uniformly in $m \geq 1$, whenever V is open and $V \supset H$.

Exercise 2.38. Let (X_n) be a sequence of real-valued independent random variables. Set $S_n = X_1 + X_2 + \cdots + X_n$. The classical Lévy equivalence theorem says that S_n converges almost surely iff S_n converges in distribution. This theorem was generalized to group-valued random variables in [50] and in [72]. Write a report on this extension.

Exercise 2.39. Let $X_1, X_2, \dots$ be a sequence of independent random variables taking values in a discrete completely simple semigroup $S = X \times G \times Y$ (usual product representation). Show that the following conditions are equivalent:

- (i) The sequence $Z_n = X_1 X_2 \cdots X_n$ converges almost surely to Z as $n \rightarrow \infty$;
- (ii) There exists y in Y such that

$$\sum_{n=1}^{\infty} [1 - P(X_n \in I_y)] < \infty,$$

where I_y is the set of idempotents in $X \times G \times \{y\}$, that is, elements $(x, (yx)^{-1}, y)$, $x \in X$.

(Hint: (i) implies that there exists y in Y such that $P(Z \in X \times G \times \{y\}) = \delta > 0$. Note that for $k \geq 1$, $Z_k \in X \times G \times \{y\}$ iff $X_k \in X \times G \times \{y\}$, and $Z_n = Z_{n+1} \in X \times G \times \{y\}$ only if $X_{n+1} \in I_y$.)

Exercise 2.40. Consider $S = G \times Y$, where $G = \{e\}$, $Y = \{y_1, y_2\}$ such that

$$(e, y_1)(e, y_2) = (e, y_2) = (e, y_2)(e, y_2)$$

and

$$(e, y_2)(e, y_1) = (e, y_1) = (e, y_1)(e, y_1).$$

Define: for $n \geq 1$, $\mu_n \{(e, y_2)\} = \frac{1}{n}$, $\mu_n \in P(S)$. Show that:

- (i) $\lim_{n \rightarrow \infty} \inf_{m > n} \mu_{n,m}(I_{y_1}) = 1$, I_{y_1} as in Problem 2.39;
- (ii) For $k \geq 1$, $\mu_{k,n}$ converges weakly to the unit mass at (e, y_1) ;
- (iii) Z_n does not converge almost surely (use Problem 2.39).

(Hint: Notice that $I_{y_1} \bar{S}^1 = \emptyset$, $s \notin I_{y_1} = S$, $s \in I_{y_1}$.)

Exercise 2.41. Let S and (X_n) be as in Exercise 2.39. Show that the following are equivalent: 2169
2170

- (i) The sequence Z_n converges almost surely; 2171
- (ii) There exists $y \in Y$ such that 2172

$$\sum_{n=1}^{\infty} [1 - \mu_n(X \times G \times \{y\})] < \infty,$$

and $\lim_{n \rightarrow \infty} \inf_{m > n} \mu_{n,m}(I_y) = 1$, where I_y is as in Exercise 2.39 and for $k \geq 1$, 2173
 $\mu_k(x) = P(X_k = x)$. 2174

Exercise 2.42. Consider Exercise 2.39 in the nondiscrete locally compact second 2175
countable situation. Show that as before, when Z_n converges almost surely to Z , 2176
then there exists $y \in Y$ such that 2177

$$P(Z \in X \times G \times \{y\}) = 1.$$

(This and other relevant results appear in [151].) 2178

Exercise 2.43. Let two sequences (μ_n) and (μ'_n) in $P(S)$, S is a locally compact 2179
second countable semigroup, be called equivalent if 2180

$$\sum_{n=1}^{\infty} \|\mu_n - \mu'_n\| < \infty,$$

where for μ and μ' in $P(S)$, 2181

$$\|\mu - \mu'\| = \sup \{ |\mu(B) - \mu'(B)| : B \text{ a Borel subset of } S \}.$$

Show that $\mu_{k,n}$ converges weakly as $n \rightarrow \infty$ for $k \geq 1$ iff $\mu'_{k,n}$ does the same when 2182

(μ_n) and (μ'_n) are equivalent. (Hint: $\|\mu_{k,k+m} - \mu'_{k,k+m}\| \leq \sum_{n=k+1}^{k+m} \|\mu_n - \mu'_n\|$.) 2183

Exercise 2.44. Let S be a compact second countable semigroup and $(\mu_n) \subset P(S)$. 2184
An element x in S is called *inessential* if there is an open set $N(x)$ containing x 2185
such that 2186

$$\sum_{n=1}^{\infty} \mu_n(N(x)) < \infty.$$

An element, when not inessential, is called *essential*. Show that the set E of *essential* 2187
points is the smallest closed set C such that 2188

$$\sum_{n=1}^{\infty} [1 - \mu_n(V)] < \infty,$$

whenever V is open and $C \subset V$. 2189

Exercise 2.45. Suppose in Exercise 2.44 S is also a group, and that for $k \geq 1$, $\mu_{k,n}$ converges weakly as $n \rightarrow \infty$ to ν_k . By Theorem 2.25, $\nu_k \rightarrow \omega_H$ weakly as $k \rightarrow \infty$. Show that the set E of essential points is a subset of the compact subgroup H . Show also that if the homogeneous space $H/[E]$, $[E]$ being the smallest closed subgroup containing E is finite, then $H = [E]$. Use Example 2.5 to show that $[E]$ can be a proper subgroup of H . (The notion of essential points and the above result are due to [140].)

Exercise 2.46. Prove assertion (2.156) when S is a compact Hausdorff semigroup.

Exercise 2.47. Prove assertion (2.166) when S is a compact Hausdorff abelian semigroup.

2.5 Notes and Comments

Section 2.2

Basic results, such as Propositions 2.1, 2.2, and 2.3, are standard and wellknown (see [176], [186] or [78]). Theorem 2.1 is due to many authors (independently); for instance, see [58] and [221]. A complete analog of this theorem in the nonabelian case is still unsolved. Proposition 2.4 is taken from [147]. Proposition 2.5 (note that μ is not assumed to be finite here) is taken from [172]. Theorem 2.2 in locally compact (or complete metric) semigroups was first given in [172]; earlier in compact semigroups, it was given in [88] and [191]. Theorem 2.3 is also contained in [172]. It was given in [111] for infinite μ when S is a semigroup of matrices. A related result in other special semigroups is given in [108]. It is relevant to point out that an analog of Theorem 2.3, when μ is not finite, is still an unsolved problem. Theorem 2.4, given in [173], sheds some light on this open problem. Theorem 2.5 appears in [174]. For characterizations (and results) concerning other types of invariant measures, we refer the reader to [3] and [9]. For analogs of Theorems 2.2 and 2.3 for semitopological semigroups (where the semigroup multiplication is separately continuous and not necessarily jointly continuous), the reader is referred to [192] and [174].

Section 2.3

Theorem 2.7 is a key theorem in this section. Ideas used in the proof of this theorem are due to many authors. Parts (i) and (ii) of this theorem appeared in [151], and parts (iii) and (iv) in [137]. Parts (i), (ii), and the first half of (iii) were given earlier by Rosenblatt for the case of compact semigroups. (See also the papers of [47, 48] in this context.) Corollary 2.1, most of which was originally due to Kawada and Ito, is

given in this form in [210]. (The part “(i) equivalent to (ii)” was given earlier in [47].) 2224
 Corollaries 2.3 and 2.4 are taken from [149]. Proposition 2.6 is due to Rosenblatt. 2225
 Lemma 2.2 is due to Csiszar. Theorem 2.8, in the case of a group, was given in [50]; 2226
 see also [223]. This version of Theorem 2.8 is taken from [151]. Theorem 2.9 ap- 2227
 peared in [149]; it was also obtained independently in [57]. Theorem 2.10 is taken 2228
 from [150]. Proposition 2.8 appeared in [149] and Theorem 2.11 in [153]. Theo- 2229
 rem 2.12 is due to Högnäs and Mukherjea. Theorem 2.13 first appeared in [142]; 2230
 the proof presented here is somewhat different. Obtaining an analog of this theorem 2231
 in a nondiscrete semigroup is much more difficult and still an unsolved problem. 2232
 Example 2.6 is due to Harry Kesten (personal communication). Theorem 2.16 and 2233
 Corollary 2.5 are given in [12]. Theorem 2.17 was communicated by Mindlin. The- 2234
 orem 2.18 appeared in [113]. It was, however, first initiated by [33]. Exponential 2235
 convergence to Haar measure with respect to other probability metrics have also 2236
 been studied. H. Horst, in Math. Nachrichten 104 (1981) 49–59, considers a com- 2237
 pact abelian group and a wider class of measures than the spread-out ones. He finds 2238
 one particular norm (called the BL-norm) for which there is exponential conver- 2239
 gence. If μ admits a density with respect to the Haar measure, then so does μ^n 2240
 for each positive integer n . O. Johnson and Yu Suhov, in J. Theoret. Probability 13 2241
 (2000) 843–857, investigate the exponential convergence of the successive densities 2242
 to the uniform density in the Kullback–Leibler distance. 2243

Section 2.4

2244

Theorems 2.22 and 2.23 are due to Budzban and Mukherjea. Theorem 2.25 is given 2245
 in [154]. Theorem 2.26 appeared in [31]. See also papers by Maksimov in the 2246
 context of all these results. Corollary 2.6 is a result of Maksimov. Theorem 2.27 2247
 appeared in [26]; however, a related and very similar result, with a completely dif- 2248
 ferent proof, first appeared in [208]. 2249

AUTHOR QUERIES

AQ1. Please check Eq. No. (2.126) without equation.

AQ2. Please check spelling of Hoffmann is Ok.

Uncorrected Proof

Chapter 3

Random Walks on Semigroups

1
2

3.1 Introduction

3

The term *random walk* suggests stochastic motion in space, a succession of random steps combined in some way. In Chap. 3, we interpret the term very narrowly: We require the steps to be *independent* and to have the *same* probability distribution. The walk is then a succession of *products* of those steps. Later on, we apply our results to slightly more general situations, e.g., to cases where the steps depend on each other in a Markovian way. Thus, our study of random walks is synonymous with the study of *products of independent identically distributed random elements* of a semigroup. We study the most basic notions for these processes which are of course discrete-time Markov chains with the semigroups as state spaces. We deal with, for example, communication relations, irreducibility questions, recurrence vs. transience, periodicity and ergodicity. Generally speaking, these probabilistic notions have an algebraic counterpart, in the sense that the probabilistic properties of a random walk cannot be satisfied unless the semigroup supporting the random walk has a certain algebraic structure. The situation is very similar to that in Chap. 2 where we saw, for example, that only completely simple semigroups with a compact group factor support limit points of a tight convolution sequence of measures (see Theorem 2.7).

Let S be a locally compact, second-countable Hausdorff topological semigroup. Let $\mathcal{B}$ denote the Borel sets of S , i.e., $\mathcal{B}$ is the σ -algebra generated by the open sets of S . Our basic measurable space Ω is the set S^∞ of sequences $\omega_1, \omega_2, \dots$ of elements of S , equipped with the σ -algebra $\mathcal{F}$ generated by coordinate mappings $\omega \mapsto \omega_i$ and the Borel sets of S . The sub- σ -algebra generated by the n first coordinate mappings will be denoted by $\mathcal{F}_n$.

We refer to the first two sections of Chap. 2 for the basic properties of probability measures on S and the convolution operation. Let $\mu_1, \mu_2, \dots \in P(S)$, the set of probability measures on S . The infinite product $\mu_1 \times \mu_2 \times \dots$ – the factor indexed by i is μ_i – is well defined on the basic measurable space Ω . The coordinate functions $X_i(\omega) \equiv \omega_i$ are measurable and the X_i 's are independent S -valued random variables with the distribution μ_i .

We almost exclusively take equal μ_i 's for $i \geq 1$. The product measure on Ω described above is then denoted by $\mathbf{P}_\mu$. For the sake of simplicity, the subscript is often dropped. The symbol $\mathbf{E}$ denotes expectation with respect to the probability measure $\mathbf{P}$.

Of paramount importance is the fact that the *product* of independent S -valued random variables is again a random variable: Its distribution being the *convolution* of the distributions of the factors. This follows from the joint continuity of the multiplication and the second countability of S : The set $\{(x, y) | xy \in B\}$ (where $B \in \mathcal{B}$) is a Borel set in $S \times S$ and belongs to the product σ -algebra generated by the rectangles $A_1 \times A_2$, $A_1, A_2 \in \mathcal{B}$. The formulas

$$\begin{aligned} \mu_1 * \mu_2(B) &= \int 1_B(s_1 s_2) d\mu_1(s_1) d\mu_2(s_2) \\ &= \int \mu_2(s_1^{-1} B) d\mu_1(s_1) = \int \mu_1(B s_2^{-1}) d\mu_2(s_2), \end{aligned} \quad (3.1)$$

proved in Proposition 2.2 show how to calculate the distribution of the product of two independent factors with distributions μ_1 and μ_2 , respectively.

Note that it is possible to dispense with the second countability assumption and still obtain a workable definition of random walks on S , see [61].

Definition 3.1. Let $X_1, X_2, \dots$ be independent S -valued random variables with $\mathcal{L}(X_i) = \mu$, $i \geq 1$, where $\mathcal{L} y(X)$, law of X , denotes the distribution of the random variable X . Let $x \in S$ be arbitrary. Then the collection of random variables $Z_n, n = 0, 1, 2, \dots$ defined by

$$Z_0 = x, \quad Z_n = Z_{n-1} X_n \quad (Z_n = X_n Z_{n-1}), \quad n = 1, 2, \dots$$

is called the *right (left) random walk* on S generated by μ with starting point x . To distinguish random walks with different starting point, we also write Z_n^x for the random walk started at x .

If $Y_1, Y_2, \dots$ is another set of random variables of law μ , independent of the X -variables as well as of each other, then we can define the *bilateral walk* generated by μ as the set $W_n, n = 0, 1, 2, \dots$ of random variables satisfying

$$W_0 = x, \quad W_n = X_n W_{n-1} Y_n, \quad n = 1, 2, \dots$$

Note that the basic probability space is $\Omega \times \Omega$ endowed with the σ -algebra $\mathcal{F} \times \mathcal{F}$ and the measure $\mathbf{P} \times \mathbf{P}$.

We occasionally look at a *mixed* random walk $M_n, n = 0, 1, 2, \dots$ where $M_0 = x$ and $M_n = M_{n-1} Y_n$ with probability $\alpha \in (0, 1)$ and $M_n = X_n M_{n-1}$ with probability $1 - \alpha$, the choices are independent of each other and of the X and Y variables.

Consider the right random walk Z_n^x generated by μ with starting point x . Its transition probability function $P_r(x, B)$, $x \in S$, $B \in \mathcal{B}$, is defined by $P_r(x, B) = \mathbf{P}\{Z_1^x \in B\}$. Hence,

$$P_r(x, B) = (\delta_x * \mu)(B) = \mu(x^{-1}B), \quad (3.2)$$

where δ_x is the unit mass at the point x . We see that $P_r(x, B)$ is a measurable function of x for each fixed $B \in \mathcal{B}$ (see Proposition 2.1). On the other hand, $P_r(x, \cdot)$ is a regular probability measure on $\mathcal{B}$. Thus, we can define the *transition probability operator* of the right random walk Z_n by the formula

$$f \mapsto \int P_r(x, dy) f(y) = \int f(xy) d\mu(y) = \mathbf{E} f(Z_1^x), \quad (3.3)$$

where x ranges over the semigroup S and f over the bounded $\mathcal{B}$ -measurable real functions on S . We use the same symbol P_r for the transition probability operator. Notice that the transition probability function $P_r(\cdot, B)$ is the transition probability operator acting on the characteristic function of B : $P_r 1_B$.

Since

$$P_r^2 1_B(x) = \int P_r(x, dy) P_r(y, B) = (\delta_x * \mu * \mu)(B) = (\delta_x * \mu^2)(B)$$

the two-step transition probability $\mathbf{P}\{Z_2^x \in B\}$ is $P_r^2 1_B(x) = (\delta_x * \mu^2)(B)$. Similarly, the n -step transition probabilities are given by iterating the transition probability operator n times or equivalently convolving the measure μ with itself n times.

Definitions are analogous the *left* random walk. The transition probability $P_l(x, B)$, $x \in S$, $B \in \mathcal{B}$, is, for example, given by

$$P_l(x, B) = (\mu * \delta_x)(B). \quad (3.4)$$

The *bilateral* random walk has transition probability function

$$P_b(x, B) = (\mathbf{P} \times \mathbf{P})\{W_1^x \in B\} = (\mu * \delta_x * \mu)(B). \quad (3.5)$$

For the *mixed* random walk, we have simply

$$P_m(x, B) = (1 - \alpha) P_r(x, B) + \alpha P_l(x, B). \quad (3.6)$$

Remark 3.1. Note that $P_b = P_r P_l = P_l P_r$. To see this, consider

$$\begin{aligned} P_b 1_B(x) &= (\mathbf{E} \times \mathbf{E}) 1_B(W_1^x) \\ &= (\mu \times \mu)\{(s, t) | sxt \in B\} \\ &= \int \mu\{s | sy \in B\} P_r(x, dy) \\ &= \int (P_l 1_B)(y) P_r(x, dy) = P_r(P_l 1_B)(x). \end{aligned}$$

The random walks just defined are *Markov chains* on the measurable space $(S, \mathcal{B})$ with the initial point x and the transition probabilities given by (3.2), (3.4), (3.5), and (3.6), respectively. This all-important fact is used extensively throughout the rest of the book.

We show this for the right walk. The other three proofs are similar.

Let the right random walk $Z_0^x = x$, $Z_1^x = xX_1$, $Z_2^x = xX_1X_2$, ... be generated by μ . In the sequel, the superscript x is dropped for ease of notation.

Following [199], Chap. 1, Sect. 2, and [87], p. 2, we consider the expectation

$$\mathbf{E} \prod_{i=1}^n f_i(Z_i),$$

for any bounded measurable real-valued functions $f_1, f_2, \dots, f_n$ and any nonnegative integer n . We obtain

$$\int \dots \int f_1(xx_1) \dots f_{n-1}(xx_1 \dots x_{n-1}) f_n(xx_1 \dots x_n) d\mu(x_1) \dots d\mu(x_n),$$

which by first integrating out x_n , equals

$$\int \dots \int f_1(xx_1) \dots f_{n-1}(xx_1 \dots x_{n-1}) P_r f_n(xx_1 \dots x_{n-1}) d\mu(x_1) \dots d\mu(x_{n-1}).$$

This in turn is

$$\mathbf{E} \prod_{i=1}^{n-1} f_i(Z_i) P_r f_n(Z_{n-1}).$$

In other words, the conditional expectation of $f_n(Z_n)$ given the past, $\mathbf{E}\{f_n(Z_n)|Z_1, \dots, Z_{n-1}\}$, equals the conditional expectation of $f_n(Z_n)$ given the immediate preceding random variable, or *state* in Markov chain terminology, $\mathbf{E}\{f_n(Z_n)|Z_{n-1}\}$, which in turn is $P_r f_n(Z_{n-1})$ $\mathbf{P}$ -almost surely. Actually, one can show slightly more: $\mathbf{E}\{f_n(Z_n)|\mathcal{F}_{n-1}\} = P_r f_n(Z_{n-1})$ $\mathbf{P}$ -almost surely.

The intuitive meaning of the calculation can perhaps be explained more simply as follows. Let B be a Borel set in S and recall that $z^{-1}B$ denotes the set $\{y|zy \in B\}$. Then

$$\begin{aligned} & \mathbf{P}\{Z_{n+1} \in B | Z_0 = x, Z_1 = z_1, \dots, Z_n = z_n\} \\ &= \mathbf{P}\{X_{n+1} \in z_n^{-1}B | Z_0 = x, Z_1 = z_1, \dots, Z_n = z_n\} \\ &= \mathbf{P}\{X_{n+1} \in z_n^{-1}B\} (= P_r 1_B(z_n)) \\ &= \mathbf{P}\{Z_{n+1} \in B | Z_n = z_n\}. \end{aligned}$$

Hence, the right random walk is a Markov chain.

Thus, we have Theorem 3.1.

Theorem 3.1. *The right, left, bilateral, and mixed random walks are Markov chains on $(S, \mathcal{B})$.*

We remark that our random walks are *not* canonical Markov chains in the sense of [199], p. 18. Also, the σ -algebras $\mathcal{F}_n$ are in general strictly larger than the ones generated by the random variables $Z_1^x, Z_2^x, \dots, Z_n^x$ (where Z_n^x is now any one of the random walks defined above). This is due to the fact that in a semigroup, knowledge of the first factor $z_1 = x_1$ and the product $z_2 = x_1 x_2$ does not imply that the second factor x_2 is known as well. In a group, we have of course $z_1^{-1} z_2 = x_2$.

Consider one of the random walks; call its transition probability kernel P . Given x and P there is a unique probability measure $\mathbf{P}'_x$ on $\Omega = S^\infty$ equipped with the σ -algebra $\mathcal{F}' \equiv$ the smallest σ -algebra making the coordinate maps measurable, with the property that

$$\begin{aligned} \mathbf{P}'_x \{ \omega \in \Omega \mid \omega_i \in A_i, i = 1, 2, \dots, n \} = \\ = \int_{A_1} P(x, dx_1) \int_{A_2} P(x_1, dx_2) \dots \int_{A_{n-1}} P(x_{n-2}, dx_{n-1}) P(x_{n-1}, A_n), \end{aligned}$$

where n is any positive integer and the sets A_i are arbitrary Borel sets of S , (see [199], p. 20, and [87], p. 8). The random variables (coordinate maps) $Z'_0(\omega) = x, Z'_n(\omega) = \omega_n, n > 0$, constitute a Markov chain, the *canonical* Markov chain on S with the initial point x and the transition probability kernel P .

The two different approaches to random walks lead to the same thing in the following sense:

$$\mathbf{P}'_x \{ Z'_i \in A_i, i = 1, 2, \dots, n \} = \mathbf{P} \{ Z_i^x \in A_i, i = 1, 2, \dots, n \}. \quad (3.7)$$

In other words, all probabilistic statements about the Z' -chain can be translated into a corresponding one for the Z^x -chain and vice versa, (see [199], pp. 20 and 31). Briefly stated $\mathbf{P}'_x(A) = \mathbf{P} \{ (x, Z_1^x, Z_2^x, \dots) \in A \}$ for any $A \in \mathcal{F}'$.

Both approaches have their advantages (and drawbacks) so we will not hesitate to choose the most convenient one for the particular problem we are trying to solve.

Definition 3.2. Let Z' be a canonical random walk with transition probability kernel P and starting point x . Define the shift or translation operator Θ on Ω by

$$(\Theta(\omega))_n = \omega_{n+1}, n = 0, 1, 2, \dots \quad \text{or} \quad Z'_n \circ \Theta = Z'_{n+1}, n = 0, 1, 2, \dots \quad (3.8)$$

The iterates of Θ are denoted by $\Theta_2, \Theta_3, \dots$

A *stopping time* of the canonical random walk Z' is a random variable (defined on $(\Omega, \mathcal{F}')$ with values in $\bar{\mathbf{N}}$ (the nonnegative integers with $+\infty$ added) such that for every $n \in \mathbf{N}$ the set $\{T \leq n\} \in \mathcal{F}'_n$, where $\mathcal{F}'_n$ is the σ -algebra generated by the first n coordinate maps $Z'_1, Z'_2, \dots, Z'_n$. The σ -algebra $\mathcal{F}'_T$ of events $A \in \mathcal{F}'$ such that $A \cap \{T \leq n\} \in \mathcal{F}'_n$ for all $n \in \mathbf{N}$ is called the σ -algebra associated with T . (Intuitively, these are the events occurring no later than T .)

Z'_T is defined to be Z'_n on the set $\{T = n\}$, $n = 0, 1, 2, \dots$ 138

Similarly, Θ_T is defined to be Θ_n on $\{T = n\}$. 139

For a Borel set $B \in \mathcal{B}$, we define the *first hitting time* T_B of B and the *first return time* S_B of B by 140
141

$$T_B = \inf\{n \geq 0 | Z'_n \in B\} \quad \text{and} \quad S_B = \inf\{n > 0 | Z'_n \in B\},$$

respectively, where the infimum of the empty set is taken to be $+\infty$. 142

Proposition 3.1. *The canonical random walk satisfies the relation (the strong Markov property)* 143
144

$$\mathbf{E}'_x\{X \circ \Theta_T | F'_T\} = \mathbf{E}'_{Z'_T}(X) \quad \text{on} \quad \{T < \infty\} \quad \mathbf{P}'_x - \text{almost surely} \quad \forall x$$

for any random variable X defined on $(\Omega, \mathcal{F}')$. 145

Let T_n be the successive return times to set $B \in \mathcal{B}$. Then the sequence $Y_n \equiv Z'_{T_n}$ 146
is a Markov chain with respect to the σ -algebras $\mathcal{F}'_{T_n}$. 147

For a proof and a detailed discussion of these issues, see [199], Chap. 1, Sect. 3. 148

In view of the almost equivalence, see (3.7), of the two definitions of our random 149
walks, we henceforth drop the primes when using the canonical random walks. 150

The transition probability operators P just defined are *Feller* (in the sense of 151
[199]) or *weakly continuous*, [229], i.e., Pf is bounded and continuous whenever 152
 f is. This fact was used early in Chap. 2; it is an important link in the argument lead- 153
ing to the definition of the convolution operation (see the proof of Proposition 2.1). 154

If S is a group and μ is absolutely continuous with respect to Haar measure, then 155
 Pf is continuous for *all* bounded, measurable f . This will be shown later. For the 156
discussion of these and related questions, we need some additional concepts. 157

Definition 3.3. The transition probability operator P is said to be *strongly con-* 158
tinuous if Pf is bounded and continuous for all bounded measurable real-valued 159
functions f . (Equivalently, we could require $P1_B$ to be continuous for all $B \in \mathcal{B}$.) 160

Let $T(x, B)$ be a nonnegative kernel on $S \times \mathcal{B}$, i.e., $T(\cdot, B)$ is a measurable 161
function on S and $T(x, \cdot)$ is a nonnegative measure, not necessarily a probability 162
measure, on $\mathcal{B}$. T is called a *component* of P if $T(x, B) \leq P(x, B)$ for all $x \in S$ 163
and all $B \in \mathcal{B}$. The component T is said to be *nontrivial at* x if $T(x, S) > 0$ and 164
continuous at x if $T(x, B)$ is lower semi-continuous at x for all B . T is a *continuous* 165
component of P if it is continuous at all $x \in S$. If the measures $T(x, \cdot)$ and $P(x, \cdot)$ 166
are equivalent, then we say that T is *equivalent to* P at x . If this is the case for all 167
 $x \in S$, then T is an *equivalent component* of P . 168

The Markov chain Z_n with transition probability kernel P is said to have T as 169
its continuous component if T is a continuous component of $\sum_{n=1}^{\infty} 2^{-n} P^n$, i.e., 170
 $T(x, B) \leq \sum_{n=1}^{\infty} 2^{-n} P^n(x, B)$, $x \in S$, $B \in \mathcal{B}$. 171

Let S_μ be the support of the measure μ . If the starting point lies in S_μ as well (or more generally, the initial distribution has its support inside S_μ), then the right (or left) random walk will $\mathbf{P}_\mu$ -almost surely never leaves the subsemigroup

$$\text{closure} \left(\bigcup_{n=1}^{\infty} S_\mu^n \right) = \text{closure} \left(\bigcup_{n=1}^{\infty} S_{\mu^n} \right),$$

(see formula (2.8)). Another common assumption is that the random walk starts at the identity element of S (adjoined if need be). In that case too, the effective state space of the walk is the smallest closed subsemigroup containing the supports of all the convolution powers μ^n , $n = 1, 2, \dots$

We will henceforth make the following blanket assumption

$$S = \text{closure} \left(\bigcup_{n=1}^{\infty} S_\mu^n \right). \quad (3.9)$$

In other words, we will restrict our attention to the *closed subsemigroup generated by the support of μ* . This set can also be described, e.g., as the support of the probability measure

$$\sum_{n=1}^{\infty} 2^{-n} \mu^n.$$

In other words, any open set in S has positive μ^n -measure for some n . Note that by restricting the topological space to a closed subspace as above we preserve all the basic properties, i.e., S remains a locally compact second countable topological semigroup. The closure of a set B is also written $\bar{B}$.

Before beginning the detailed analysis of our random walks, let us see how the elementary communication relations from the theory of Markov chains can be formulated algebraically. We use the generic S_n for any of the four random walks introduced in Definition 3.1 and P for any of the corresponding transition probability functions or transition probability operators, as the case may be.

Definition 3.4. Let P be the Feller transition probability operator of a Markov chain on S . In particular, we take P to be the transition probability of one of the walks of Definition 3.1. For $x, y \in S$, we say that x *leads to* y , $x \rightarrow y$ if for all neighborhoods $N(y)$ of y , $P^n(x, N(y)) > 0$ for some $n \geq 1$. If $x \rightarrow y$ and $y \rightarrow x$, then x and y are said to *communicate* (notation: $x \leftrightarrow y$). x is a *return point* if $x \rightarrow x$. The return point x is said to be an *essential* element if $x \rightarrow y$ implies $y \rightarrow x$.

x is said to *lead to y infinitely often*, ($x \rightarrow y$ infinitely often,) if $\mathbf{P}_x\{S_n \in N(y) \text{ infinitely often}\} = 1$ for any neighborhood $N(y)$ of y . If $x \rightarrow x$ infinitely often, then x is said to be *recurrent* for the walk S_n . The element x is *conservative* (*positive*) if for every neighborhood $N(x)$ of x

$$\sum_{n=1}^{\infty} P^n(x, N(x)) = \infty$$

$$(\limsup_{n \rightarrow \infty} P^n(x, N(x)) > 0).$$

If the point x is not recurrent for a random walk, then it is said to be *transient*. 204

If a random walk admits a recurrent point then it is said to be a *recurrent random walk*. If the recurrent random walk admits an invariant probability distribution π , i.e., 205
206
207

$$\pi(A) = \int \pi(dx)P(x, A), \quad A \in \mathcal{B} \quad \text{and} \quad \pi(S) = 1,$$

then it is said to be *positive recurrent*. Otherwise, it is called *null recurrent*. If a random walk admits no recurrent points, it is called *transient*. 208
209

Proposition 3.2. (i) If S satisfies our basic assumption (3.9), the following holds for the random walks on S : 210
211

(a) For the right walk on S , $x \rightarrow y$ if and only if $y \in \overline{xS}$, 212

(b) For the left walk, $x \rightarrow y$ if and only if $y \in \overline{Sx}$, 213

(c) For the bilateral walk, $x \rightarrow y$ if and only if $y \in \overline{\bigcup_{n=1}^{\infty} S_{\mu}^n x S_{\mu}^n}$, 214

(d) For the mixed walk, $x \rightarrow y$ if and only if $y \in \overline{Sx \cup xS \cup SxS}$. 215

(ii) The relation $\rightarrow$ is transitive. The relation $\leftrightarrow$ is an equivalence relation on the set of return points. 216
217

Proof. We first recall that if O is an open set then the sets $x^{-1}O$ and Ox^{-1} are open in S and the set $\{(s, t) | sxt \in O\}$ is open in $S \times S$ (see Proposition 1.16). The sets xS , xS_{μ} , S_{μ}^n are countable unions of compact sets (and thus Borel sets) because S is, by assumption, locally compact and second countable. 218
219
220
221

(i) We begin by proving the Claim (ia). 222

Clearly, $\mathbf{P}_x\{Z_n \notin xS\} = 0$ for all n . So any $z \notin \overline{xS}$ will have a neighborhood $N(z)$ such that $\mathbf{P}_x\{Z_n \in N(z)\} = 0$ for all n . This proves that $x \rightarrow y$ implies $y \in \overline{xS}$. 223
224
225

To prove the converse, let $y \in \overline{xS}$ and let $N(y)$ be any neighborhood of y . $N(y) \cap xS$ is nonempty; thus, we conclude that $x^{-1}N(y)$ is a nonempty open set. By our basic assumption in (3.9), $P_r^n(x, N(y)) = \mu^n(x^{-1}N(y)) > 0$ for some n . This holds for all neighborhoods $N(y)$ of y , hence $x \rightarrow y$. 226
227
228
229

The proof of Claim (ib) is similar and therefore omitted. 230

In Claim (ic), we note that $P_b^n(x, N(y)) = (\mu^n \times \mu^n)\{(s, t) | sxt \in N(y)\} > 0$ for a neighborhood $N(y)$ of y if and only if 231
232

$$N(y) \cap S_{\mu}^n x S_{\mu}^n \neq \emptyset.$$

Reasoning as above, we conclude that $x \rightarrow y$ if and only if y belongs to the closure of the union of the sets $S_{\mu}^n x S_{\mu}^n$. 233
234

The basis of the proof of Claim (id) is the fact $P_m^n(x, N(y)) > 0$ if and only if $N(y)$ has nonempty intersection with any of the sets $xS_{\mu}^n, S_{\mu}^n x, S_{\mu}^k x S_{\mu}^{n-k}, k = 1, 2, \dots, n-1$. 235
236
237

(ii) The transitivity of $\rightarrow$ is a consequence of (i) and the continuity of the multiplication. For the bilateral walk, the argument runs as follows. If $x \rightarrow y$ and $y \rightarrow z$, then y is the limit point of a sequence $s_n x t_n$ and z the limit point of a sequence $s'_m y t'_m$ (where $s_n, t_n \in S_{\mu}^n, s'_m, t'_m \in S_{\mu}^m$). Hence, z is a limit point of $s'_m s_n x t_n t'_m \in S_{\mu}^{n+m} x S_{\mu}^{n+m}, n, m \rightarrow \infty$. The rest of (ii) follows easily. $\square$ 238
239
240
241
242

Remark 3.2. The set $\bigcup_{n=1}^{\infty} S_{\mu}^n x S_{\mu}^n$ is not necessarily a subsemigroup of S , while the sets mentioned in the statements (ia), (ib), and (id) are.

The behavior of Markov chains on general nondiscrete topological spaces is markedly different from the discrete case, see, for example, [205]. As for our random walks on semigroups, we will find in the subsequent sections that only careful analysis reveals to what extent the notions originating from discrete Markov chain theory are transportable to the topological semigroup setting.

Example 3.1. Let S be the closed semigroup $0 \cup \{2^n | n = \dots, -2, -1, 0, 1, 2, \dots\}$ generated by the probability measure μ on the multiplicative semigroup $[0, \infty]$ defined by $\mu\{2\} = \mu\{\frac{1}{2}\} = \frac{1}{2}$. There is only one essential element, 0, for the random walk with transition probability $\mu(x^{-1}B)$. 0 is, of course, also recurrent. The nonzero elements form a communicating class of *nonessential recurrent* elements. This is because the random walk on $S \setminus \{0\}$ is the isomorphic image of a symmetric Bernoulli random walk on the integers, which is known to be recurrent.

We now proceed to a more detailed study of various special types of semigroups. We try to determine the communication structure and characterize the recurrent elements. As we have seen throughout the book, a certain probabilistic behavior may only be possible in a semigroup with a very particular structure. As far as possible, we point out the differences and similarities with classical random walk theory on $\mathbb{R}^d$ and on abelian as well as nonabelian groups.

Exercise 3.1. Prove Theorem 3.1 for the bilateral case.

Exercise 3.2. Prove (ii) of Proposition 3.2 for the mixed walk.

Exercise 3.3. Prove the assertions in Remark 3.2.

Exercise 3.4. Let α_n be a sequence of positive real numbers, $\sum_n \alpha_n = 1$. Assume also that the irreducibility condition (3.9) holds. Show that the support of the probability measure $\sum_n \alpha_n \mu^n$ is all of S .

Exercise 3.5. Let μ be a probability measure on the semigroup S .

Let T be a right or left ideal of S . Prove that $\mu^n(T)$ is a monotone function of n .

Let I be a two-sided ideal of S . Show that $\lim \mu^n(I)$ is either 0 or 1.

Exercise 3.6. Consider a random walk on a locally compact noncompact semigroup. Show that the transition probability operators P do not necessarily map $\mathcal{C}_0$ into itself. ($\mathcal{C}_0$ is the family of continuous functions vanishing at infinity.)

Exercise 3.7. Suppose S is a topological group. Prove that the transition probability operator P_r maps $\mathcal{C}_0$ into $\mathcal{C}_0$.

3.2 Discrete Semigroups

277

If the topology of the semigroup S is discrete, then the support of a regular probability measure μ on S is necessarily a countable set, $S_\mu = \{s \in S \mid \mu(s) > 0\}$. Since S is assumed to be generated by S_μ , see (3.9), we conclude that the semigroup S under study is a countable set. The random walks on S are thus particular Markov chains on the countable set S and we have at our disposal the notions and the results of Markov chain theory. Basic references are, for example, the first chapters of [37, 122, 204].

General definitions from the previous section are considerably simplified since the subsets of S are both open and closed. The neighborhood $N(y)$ of y is interpreted as $\{y\}$ and $\overline{xS}$ as xS , etc. The notation $x^{-1}y$ is used to denote the set $\{s \mid xs = y\} \equiv x^{-1}\{y\}$. Accordingly, the transition probability function for the right random walk will be written $P_r(x, y) = \mu(x^{-1}y)$ instead of $P_r(x, \{y\})$. Similar changes in notation will be made for the other random walks. We emphasize at this point that these notational simplifications will be used in Sect. 3.2 only.

Proposition 3.3. *Let S be a countable semigroup generated by the support of the probability measure μ . Then the following assertions are true*

- (i) x is essential for the right random walk if and only if x belongs to a minimal right ideal.
- (ii) x is essential for the left random walk if and only if x belongs to a minimal left ideal.
- (iii) x is essential for the bilateral walk if and only if x belongs to a minimal two-sided ideal.
- (iv) x is essential for the mixed walk if and only if x belongs to a minimal two-sided ideal. Furthermore, all essential elements communicate.

Proof. Assertion (i): By Proposition 3.2 $x \rightarrow y$ if and only if $y \in xS$. If x is essential and $x \rightarrow y$, then $x \in yS$ and $xS = yS$, i.e., the ideal generated by x is minimal. On the other hand, if the ideal $\{x\} \cup xS$ is minimal, then $x \in xS$ (implying that x is a return point) and $yS = xS \ni x$ for all $y \in xS$ (implying that $y \rightarrow x$ for all y with $x \rightarrow y$).

We omit the proof of Assertion (ii). To see that Assertion (iii) holds, consider a bilateral walk on S and let x be an essential element. If $x \rightarrow y$, then $y \rightarrow x$, i.e., $x \in \bigcup_{n=1}^{\infty} S_\mu^n y S_\mu^n$, so x and y generate the same bilateral ideal of S : $\{x\} \cup xS \cup Sx \cup SxS = \{y\} \cup yS \cup Sy \cup SyS$.

For $u \in S_\mu^k$, the element xu is essential, too. To prove this assertion, let $xu \rightarrow y$. Then $y = s_n x u t_n$ for some appropriate $s_n, t_n \in S_\mu^n$. $x \rightarrow uy$ since $uy \in S_\mu^{n+k} x S_\mu^{n+k}$. x is essential by assumption so $uy \rightarrow x$, i.e., $x = s_m u y t_m$ for some appropriate s_m, t_m . But then $xu = (s_m u) y (t_m u) \in S_\mu^{m+k} y S_\mu^{m+k}$ implying $y \rightarrow xu$.

Similarly, ux and uxu' are essential for any $u, u' \in S$. Hence, the set of essential elements form a two-sided ideal of S .

Let x, x' be two essential elements. Then $x' \rightarrow xx'x$ and $xx'x \rightarrow x'$ which implies that $x' \in SxS$. Hence, all essential elements lie in the ideal generated by x . Consequently, the set C of essential elements is contained in the ideal SxS generated by x . On the other hand, we just saw that C is an ideal so SxS must equal C . Since $SyS = C$ independently of the choice of $y \in SxS$, it must be a minimal ideal. Conversely, let $x \in S$ generate a minimal ideal $I = SxS = IxI = IyI$ (where y in any element of I). Suppose $x \rightarrow y$. We have to show that $y \rightarrow x$. By assumption $x \in SyS$, i.e., $x = s_my_t_n$ for some $s_m \in S_\mu^m \cap I$ and $t_n \in S_\mu^n \cap I$. If $m = n$, then $y \rightarrow x$ and the proof is finished. If not, let $h = n - m$. For definiteness, take $h > 0$.

In the following calculations, we use the convention that a subscripted element s_i belongs to the set S_μ^i .

For some k , $y = a_kxb_k$ since $x \rightarrow y$. Then $x = s_ma_kxb_kt_n$ and $x = (s_ma_k)^\alpha x(b_kt_n)^\alpha$ for all $\alpha = 1, 2, \dots$. Furthermore, there are i and j such that $s_m = c_is_md_j$ whence $s_m = c_i^\beta s_md_j^\beta$ for $\beta = 1, 2, \dots$.

Let $n > \alpha_1 > 0$ and $\alpha_1 + \alpha_2 = \alpha$. Then

$$x = (c_i^\beta s_md_j^\beta a_k)^{\alpha_1} (s_ma_k)^{\alpha_2} x(b_kt_n)^\alpha = (c_i^\beta s_md_j^\beta a_k)^{\alpha_1} (s_ma_k)^{\alpha_2-1} s_my_t_n(b_kt_n)^{\alpha-1}$$

If we now can arrange $\alpha_1, \alpha_2, \beta$ to satisfy

$$\alpha_1[\beta(i+j) + m + k] + (\alpha_2 - 1)(m + k) + m = n + (\alpha - 1)(k + n),$$

or, equivalently,

$$\alpha_1[\beta(i+j)] = \alpha h, \quad (3.10)$$

then $x \in \bigcup_{n \geq 1} S_\mu^n y S_\mu^n$, i.e., $y \rightarrow x$. To satisfy (3.10) we can choose, for example,

$$\beta = 1, \quad \alpha_1 = h, \quad \alpha = i + j \quad \text{if } h \leq i + j$$

and

$$\beta > \frac{h}{i+j}, \quad \alpha_1 = h, \quad \alpha = \beta(i+j) \quad \text{if } h > i + j.$$

This concludes the proof of (iii).

Assertion (iv): Let x be essential for the mixed random walk and let $x \rightarrow y$. Then $y \in xS \cup Sx \cup SxS$ is a return point and $x \in yS \cup Sy \cup SyS$, i.e., x lies in the bilateral ideal generated by y . Thus, the ideal generated by x must be minimal. In particular $Sx, xS \subset SxS$, $yS, Sy \subset SyS$ and $SxS = SyS$.

Conversely, if x belongs to a minimal bilateral ideal then $\{x\} \cup Sx \cup xS \cup SxS = SxS$ and for any $y \in SxS$ we have $SyS = SxS$. Since $x \in SxS$ this means that $x \rightarrow y$ implies $y \rightarrow x$.

The above argument also shows that the essential elements (= the elements of the minimal bilateral ideal) communicate. $\square$

Remark 3.3. Propositions 3.2 and 3.3 show that the communicating essential elements form equivalence classes. These are called *essential classes*. *Inessential classes* are made up of communicating elements that are not essential.

- Corollary 3.1.** (i) Let x be essential for one of the unilateral walks. Then x is essential for the bilateral and mixed walks as well.
(ii) x is essential for both the right and the left random walk if and only if x belongs to the completely simple minimal two-sided ideal K of S .

Proof. (i) Let x be essential for the left random walk. Let I be the set of all essential elements for the left random walk. I is the union of the minimal left ideals of S . To show that I is a minimal two-sided ideal, let us consider the elements si and it ($i \in I, s, t \in S$.) Clearly, $si \in I$, since $Si = Ssi$ is a minimal left ideal. Also, $is \in I$, since is is essential for the left walk whenever i is. ($is \rightarrow t \iff t = uis$ for some $u \in S$. But $i \in Sui$ so $is = (vui)s = vt$ for some v . In other words, $t \rightarrow is$.) Thus, $SiS = I$ for all $i \in I$. ($SiS \neq I$ is impossible because if $j \in I$ then $j \in Sij$ and, further, $Sij \subset SiS$.) Hence, I is a minimal two-sided ideal. The claim then follows from Assertion (iii) of Proposition 3.3.

(ii) Let x be essential for both the right and the left random walks. Then x belongs to a minimal right ideal and a minimal left ideal of S . Let K be the union of the minimal left ideals of S . By (i), K is a minimal two-sided ideal and hence a simple subsemigroup of S . Proposition 1.6(ii) shows that K is completely simple since it contains at least one minimal right ideal.

Conversely, let K be the completely simple minimal ideal of S . Then K is a union of minimal left (right) ideals, so any $x \in K$ is essential for the left and right random walks. $\square$

Example 3.2. (i) Let S be $Q_+ \times Q_+$ endowed with the multiplication

$$(a, b)(a', b') = (aa', ba' + b'),$$

where Q_+ is the set of positive rational numbers. Then S does not admit any minimal left or right ideals. Nevertheless, S is simple since $SxS = S$ for any $x \in S$. Example 3.2 was studied in Sect. 1.4 in a slightly different form.

- (ii) Let S be the set of strictly increasing polygonal lines of the form

$$a(t) = \begin{cases} s_{m-1} + \frac{t-r_{m-1}}{r_m-r_{m-1}}(s_m - s_{m-1}), & \text{if } r_{m-1} \leq t \leq r_m, m = 1, 2, \dots, n, \\ s_n + t - r_n, & \text{if } t > r_n, \end{cases}$$

where $0 = r_0 < r_1 < \dots < r_n$ and $0 < s_0 < s_1 < \dots < s_n$ are rational and $n = 0, 1, 2, \dots$. The multiplication in S is ordinary composition of functions. S is left simple since the equation $xb = c$ is solvable for any $b, c \in S$. One solution is

$$x(t) = \begin{cases} (b^{-1}(t)), & \text{if } b(0) \leq t, \\ \frac{c(0)}{2}(1 + \frac{t}{b(0)}), & \text{if } 0 \leq t \leq b(0). \end{cases}$$

S is not completely simple, however, since it contains no idempotent. To see this, suppose that there is an $e \in S$ with $ee = e$. In particular,

$e(e(0)) = e(0) = s_0 > 0$. But this is impossible since the elements of S are strictly increasing functions. Thus, S cannot contain any minimal right ideals.

Definition 3.5. Let P be the transition matrix of one of the random walks defined in Sect. 3.1. The *period* of a return point $x \in S$ (in the random walk in question) is the greatest common divisor (gcd) of the set

$$\{n \geq 1 | P^n(x, x) > 0\}.$$

The definition of the period is the usual one in Markov chain theory. It also follows from the general theory that communicating elements have the same period.

Proposition 3.4. *Essential classes for the right/left/bilateral/mixed random walk have the same period p_r, p_l, p_b , and p_m , respectively. If both the right and the left walk admit essential elements x , then*

$$p_r = p_l = \gcd\{n - m | \mu^n(x), \mu^m(x) > 0\},$$

for any essential x .

Proof. Let x and y be essential with respect to the right random walk. Call their periods $p_r(x)$ and $p_r(y)$, respectively. If $P_r^n(x, x)$ is strictly positive, then so is $P_r^n(yx, yx)$. This implies that $p_r(y)$ (which is equal to $p_r(yx)$) is a divisor of $p_r(x)$. Similarly, $p_r(x)$ can be shown to be a divisor of $p_r(y)$. Hence, $p_r(x) = p_r(y)$. The case of the left random walk is analogous.

Consider the bilateral random walk on S . Let I be the set of essential elements and assume that it contains more than one communicating class. Let x and y be noncommunicating elements of I . We have $y = sxt$ for some $s \in S_\mu^{n'}$ and $t \in S_\mu^{n''}$ where $n' \neq n''$. Let z belong to the same communicating class as x . Then

$$z = s_1 s x t s_2,$$

for some $s_1 \in S_\mu^{n_1}$ and $s_2 \in S_\mu^{n_2}$ with $n_1 + n' = n_2 + n'' = n$. Furthermore, since $z \rightarrow x$ in the bilateral walk, we have

$$x = uzv = u(s_1 s x t s_2)v = u s_1 y s_2 v,$$

where $u, v \in S_\mu^m$ for some m . Then $y = (s u s_1) y (s_2 v t)$, i.e., $y \in S_\mu^{n+m} y S_\mu^{n+m}$. If $x = a x b$ with $a, b \in S_\mu^k$, we can write y in the form $(s a u s_1) y (s_2 v b t)$ which means that $y \in S_\mu^{k+n+m} y S_\mu^{k+n+m}$.

We conclude that if $P_b^k(x, x) > 0$ then $P_b^{k+m+n}(y, y) > 0$. Hence, the period of y divides that of x .

By interchanging x and y , we obtain the result $p_b(x) = p_b(y)$.

As for the mixed random walk, we note that it admits at most one essential class. The period is, then, the same for all the essential elements.

If both unilateral random walks admit essential elements, then these elements form the completely simple minimal ideal K of S . Let x belong to K and assume that $\mu^k(x) > 0$. Denote the positive integer

$$\gcd\{n - m \mid \mu^n(x), \mu^m(x) > 0\},$$

by q . If $P_r^n(x, x) > 0$ ($P_l^n(x, x) > 0$), then clearly $\mu^{k+n}(x) > 0$. Hence, q is a divisor of $p_r(x)$ ($p_l(x)$).

Since $x \in K$ there is a $y \in K$ with $xyx = x$. $\mu^k(y) > 0$ for some k . Then $\mu^n(x), \mu^m(x) > 0$ implies that $P_r^{k+n}(x, x), P_r^{k+m}(x, x) > 0$. p_r being a divisor of $k+n$ and $k+m$ it is also a divisor of $n-m$. Hence, p_r divides q , which shows that p_r , in fact, equals q . Since p_r is constant on K we conclude that q is independent of the choice of $x \in K$. A similar argument shows that $p_l = q$. $\square$

We saw in Corollary 3.1(i) that the existence of essential elements for both unilateral walks furnishes us with a completely simple minimal ideal K . The very detailed knowledge that we have about the algebraic structure of K (see Sect. 1.4) is immensely helpful as shown in Theorem 3.2.

Theorem 3.2. *Let K be the completely simple minimal ideal of S . Suppose that the unilateral walks are periodic with period p . Define the sets $C_k \equiv \{x \in K \mid \mu^m(x) > 0 \Rightarrow m \equiv k \pmod{p}\}$ for $k = 0, 1, \dots, p-1$. Then*

- (i) *The sets $x C_k (C_k x)$ are the cyclic classes for the right (left) random walk starting at $x \in K$. They satisfy $C_k C_j = C_{k+j \pmod{p}}$ for all j and k .*
- (ii) *If K has the Rees–Suschkewitsch representation $E \times G \times F$ then $C_0 = E \times H \times F$ where H is a normal subgroup of the group G and $C_k = E \times H_k \times F$ for $k = 0, 1, \dots, p-1$ (where $H_0 = H$). The sets H_k are cosets of H and they satisfy $H_k H_j = H_{k+j \pmod{p}}$. H_k may also be written $g^k H$ where g is any element of H_1 .*
- (iii) *The bilateral walk has period $\frac{p}{2}$ if p is even and p if p is odd. The number of essential classes with respect to the bilateral walk is 2 if p is even and 1 if p is odd.*

Recall that a set of communicating states (of a periodic Markov chain with transition probability matrix P and period p) is partitioned into *cyclic classes* by the equivalence relation $x \sim y \Leftrightarrow P^{mp}(x, y) > 0$ for some positive integer m . We also know that $P^{mp}(x, y)$ is in fact strictly positive for any integer m large enough, larger than some $m_{x,y}$, say.

Proof. The completely simple minimal ideal K is exactly the set of essential states (for all walks).

The Proposition 3.4 implies that the sets C_k form a partition of K . Clearly, $C_k C_j \subset C_{k+j \pmod{p}}$. Any idempotent e of K must belong to C_0 , since $\mu^m(e) > 0$ implies $\mu^{2m}(e) > 0$. For any element z of $C_{k+j \pmod{p}}$, there are e idempotent, $x \in C_k$ and $y \in K$ such that $exy = z$. y must belong to C_j . Hence $C_0 C_k C_j = C_k C_j = C_{k+j \pmod{p}}$.

$x \in K$ can be written $x = xe$ for some idempotent e ; hence, $x \in xC_0$. If $y = xs = xt$, then s and t belong to the same C_k since

$$\{n - m | \mu^n(y), \mu^m(y) > 0\},$$

must consist of multiples of p . Then,

$$xS = xK = \bigcup_{k=0}^{p-1} xC_k$$

(disjoint union).

If $y \in xC_k$ and $y \sim z$ (for the right random walk starting from x), then $y = xs, z = xst$ for some $s, t \in K, s \in C_k$. But we have necessarily $t \in C_0$; hence, $z \in xC_k$. This completes the proof of (i) for the right random walk case. The proof for the left random walk is similar.

Random walks generated by μ^p are aperiodic and admit essential elements forming a completely simple subsemigroup K' of the completely simple minimal ideal K of our original semigroup S . Actually, $K' = C_0$. Let the Rees–Suschkewitsch representation of K be $E \times G \times F$ and that of C_0 be $E' \times H \times F'$. Since all idempotents of K belong to C_0 the side factors must coincide: $E = E', F = F'$. We must also have $\phi(F, E) \subset H$, where ϕ is the sandwich function (see Sect. 1.4).

The relation $C_j C_k = C_{k+j}$ shows that C_k is of the form $EH_k F$, using here the product form of the representation.

$H_j F E H_k = H_{j+k}$. Recalling the fact that $FE \subset H$ we see that $HH_k = H_k H = H_k$ and $H_{-k} H_k, H_k H_{-k} \subset H$. The last inclusion is, in fact, an equality because the H_k 's form a partition of G . Similarly, one can then show that $cH = Hc = H_k$ if $c \in H_k$, so H_k is a left and right coset of H . (Clearly $cH \subset H_k$. But any $d \in H_k$ may be written $cc^{-1}d$. There is no other alternative for $c^{-1}d$ than to belong to H .) This is valid for any k , which implies that H is a normal subgroup of

$$G = \bigcup_{k=0}^{p-1} H_k.$$

Writing $H_1 = gH$ for a $g \in H_1$ one sees that $H_k = g^k H$ and $g^p H = H$.

To prove part (iii), notice that the idempotent $e \in K$ may be reached from itself in a multiple of $\frac{p}{2}$ bilateral steps if p is even. On the other hand, only the elements of

$$E \times \bigcup_{k=0}^{\frac{p}{2}-1} H_{2k} \times F$$

are reached from e . Similarly, the elements of $E \times H_k \times F$ for odd k communicate (in the bilateral walk) only with each other.

If p is odd the disjoint union $E \times \bigcup_{k=0}^{p-1} H_{2k} \times F$ equals all of K . Hence, e communicates in the bilateral walk with all elements of K . The p cyclic classes are of the form $E \times H_{2k} \times F$. $\square$

The concept of *recurrence* plays a central role in random walk theory. There are several ways in which one can express the phenomenon and we shall shortly see that, fortunately, a few other possible definitions turn out to be equivalent to our Definition 3.4: A point $x \in S$ is recurrent with respect to a random walk S_n if $x \rightarrow x$ infinitely often, viz., $\mathbf{P}_x\{S_n = x \text{ infinitely often}\} = 1$. In a more general topological setting, this equivalence between the different definitions may no longer hold, however, as will be shown in later sections.

Proposition 3.5. *Let W_n (Z_n) be the right (left) random walk generated by the probability measure μ on the discrete semigroup S . Then the following are equivalent:*

- (a) $x \rightarrow x$ infinitely often for W_n .
- (a') $x \rightarrow x$ infinitely often for Z_n .
- (b) $\mathbf{P}\{W_n = x \text{ infinitely often}\} > 0$.
- (b') $\mathbf{P}\{Z_n = x \text{ infinitely often}\} > 0$.
- (c) $\sum_{n=1}^{\infty} P_r^n(x, x) = \infty$.
- (c') $\sum_{n=1}^{\infty} P_l^n(x, x) = \infty$.
- (d) $\sum_{n=1}^{\infty} \mu^n(x) = \infty$.

Furthermore, any of the above conditions imply that the unilateral, bilateral, and mixed walks all admit essential elements. The essential elements form the completely simple minimal ideal of S .

We remark that x satisfying one of the conditions (b) is said to be *unconditionally recurrent* for the random walk in question.

Proof. To fix the notation, let $W_n = X_1 \cdots X_{n-1} X_n$ be the right random walk generated by μ .

Condition (d) being symmetric, we need only show the equivalence of Conditions (a), (b), and (c).

Conditions (a) and (c) are equivalent. This is a standard fact in Markov chain theory, see [122], p. 66.

Condition (a) implies Condition (b) since $\mathbf{P}\{W_n = x \text{ infinitely often}\}$ is no less than $\mathbf{P}\{W_k = x, x X_{k+1} \cdots X_n = x \text{ infinitely often}\}$. This last probability is simply $\mu^k(x)$ which, by our blanket irreducibility assumption (3.9), is positive for some k .

$\mathbf{P}\{W_n = x \text{ infinitely often}\} = \mathbf{P}\{W_k = x \text{ for some positive integer } k, x X_{k+1} \cdots X_n = x \text{ infinitely often}\} = \mathbf{P}\{W_k = x \text{ for some positive integer } k\} \cdot \mathbf{P}_x\{W_n = x \text{ infinitely often}\}$. According to Markov chain theory, $\mathbf{P}_x\{W_n = x \text{ infinitely often}\}$ is either 0 or 1, see [122], p. 72. If Condition (b) holds so does Condition (a), because then the preceding probability is 1.

For any $x \in S$, there exists k such that $\mu^k(x) > 0$. Then

$$\mu^n(x) \geq \mu^k(x) P_r^{n-k}(x, x), \quad (3.11)$$

if $k < n$. Summation over n shows that Condition (c) implies Condition (d).

Assume that x is nonconservative for W_n , i.e., the series $\sum_{n=1}^{\infty} P_r^n(x, x)$ converges. We can write (see [204], pp. 8 and 15)

$$P_r^n(y, x) = \sum_{k=0}^n f_{y,x}^k P_r^{n-k}(x, x) \quad (3.12)$$

and

$$\sum_{n=1}^{\infty} P_r^n(y, x) = f_{y,x}^* \sum_{n=0}^{\infty} P_r^n(x, x), \quad (3.13)$$

where $f_{y,x}^k$ ($f_{y,x}^*$) is the probability that the right random walk starting from $y \in S$, $y \neq x$ will hit x at exactly the k th step (will ever visit x); $f_{x,x}^*$ and $P_r^0(x, x)$ are defined to be 1.

Using this relation, we obtain

$$\begin{aligned} \sum_{n=1}^{\infty} \mu^n(x) &= \sum_{n=1}^{\infty} \sum_{y \in S} \mu(y) P_r^{n-1}(y, x) \\ &= \sum_{n=0}^{\infty} P_r^n(x, x) \sum_{y \in S} \mu(y) f_{y,x}^* \\ &\leq \sum_{n=0}^{\infty} P_r^n(x, x) \end{aligned} \quad (3.14)$$

This shows that Condition (d) implies Condition (c).

x cannot be recurrent unless it is essential, cf. [37], p. 20. Furthermore, it is recurrent, thus essential, for *both* unilateral walks. Corollary 3.1 shows that the essential elements for all the walks considered are exactly the same, making up the completely simple minimal ideal K of S . $\square$

The existence of essential elements for the bilateral or mixed random walks was not enough to guarantee the essentiality of the unilateral walks (see Example 3.2). Recurrence is a strong enough condition: Existence of recurrent elements for the bilateral or mixed walks is actually equivalent to any of the conditions mentioned in Proposition 3.5.

Proposition 3.6. *x is recurrent for the bilateral or mixed random walks if and only if it is recurrent for any (thus both) of the unilateral ones. The set of recurrent elements then coincides with the completely simple minimal two-sided ideal K of S .*

Proof. If x is transient for the right random walk and $\mu^k(x) > 0$, then

$$\sum_n \mu^k(x) P_b^n(x, x) \leq \sum_n \mu^{2n+k}(x) < \infty,$$

by Proposition 3.5, Condition (d). Hence, x is transient for the bilateral walk. Also 540

$$\sum_n \mu^k(x) P_m^n(x, x) \leq \sum_n \mu^{n+k}(x) < \infty,$$

so the same conclusion can be drawn for the mixed random walk. 541

Reasoning as in the proof of Proposition 3.5, see formulas 3.13 and 3.14, we can 542
write 543

$$\mu^{2n+1}(x) = \sum_y \mu(y) P_b^n(y, x)$$

and 544

$$\sum_{n=1}^{\infty} \mu^{2n+1}(x) = \sum_{n=0}^{\infty} P_b^n(x, x) \sum_y \mu(y) f_{y,x}^* \leq \sum_n P_b^n(x, x) < \infty$$

if x is transient for the bilateral walk. ($f_{y,x}^*$ is the probability that the bilateral walk 545
starting from x will ever reach $y \neq x$.) $xs \in K$ is also transient for the bilateral 546
walk, for any $s \in S$. This will be proved below. Then 547

$$\sum_n \mu^{2n}(x) \mu(s) \leq \sum_n P_b^n(xs, xs) < \infty,$$

which implies that x is transient for the unilateral random walks. 548

Let $x \in K$, $\mu^\alpha(x) > 0$, be transient for the bilateral walk and take $s \in S_\mu$. 549
Then $yxst = x$ for some t, y with $\mu^k(t), \mu^{k+1}(y) > 0$. This is possible since 550
 $x \in K$ is essential for the bilateral walk. Let u, v be such that $\mu^n(u), \mu^n(v) > 0$ and 551
 $uxsv = xs$. Hence, 552

$$uxsv = xs \Rightarrow yuxsvt = x \quad (3.15)$$

and so 553

$$\mu^\alpha(x) \mu(s) P_b^n(xs, xs) \mu^k(t) \mu^{k+1}(y) \leq \mu^\alpha(x) P_b^{n+k+1}(x, x). \quad (3.16)$$

Summing over n proves that xs is transient for the bilateral walk. The same argu- 554
ment goes through for all elements of $K = SxS$. 555

We note that 556

$$\mu^{n+1}(x) = \sum_y \mu(y) P_m^n(y, x)$$

so the same reasoning as for the bilateral case can be employed to prove that if x 557
is transient for the mixed random walk, then it is transient for the unilateral ones as 558
well. 559

Suppose that the right random walk admits a recurrent point x . Then all points 560
of the completely simple minimal ideal K are recurrent for the unilateral walks. To 561
see this, consider 562

$$\sum_n \mu^k(s) \mu^n(x) \leq \sum_n \mu^{n+k}(sx).$$

Points satisfying Condition (d) of Proposition 3.5 form a two-sided ideal of S . On the other hand, we know that points outside the minimal ideal are inessential and hence transient. $\square$

In Definition 3.4, we introduced the concept of a *positive* point. In the context of a Markov chain P on a countable state space x is said to be positive if

$$\limsup_{n \rightarrow \infty} P^n(x, x) > 0.$$

In other words, x is positive if and only if it is *positive recurrent* in the usual Markov chain sense. A point which is recurrent but not positive recurrent is said to be *null recurrent*.

Definition 3.6. The *type* of a point with respect to one of the random walks is the property of being positive recurrent, null recurrent or transient. (These possibilities are exhaustive and mutually exclusive.) In other words, x is positive recurrent iff $\limsup_n P^n(x, x) > 0$, null recurrent iff $\lim_n P^n(x, x) = 0$ and $\sum_n P^n(x, x) = \infty$, transient iff $\lim_n P^n(x, x) < \infty$.

A random walk (or the generating probability measure μ) is called positive (null) recurrent if there are points $x \in S$ which are positive (null) recurrent with respect to the random walk in question. (Theorem 3.3 shows that positive and null recurrent points do not coexist.)

A semigroup S is called (positive) recurrent if there is at least one (positive) recurrent probability measure μ on S (satisfying the irreducibility condition (3.9)).

Justification for the above definitions is given in Theorem 3.3.

Theorem 3.3. (i) Consider the two unilateral, the bilateral, and the mixed walks on S generated by the probability measure μ . An element x of S is of the same type for all the four random walks:

- x is positive recurrent iff $\limsup_n \mu^n(x) > 0$,
- x is null recurrent iff $\lim_n \mu^n(x) = 0$ and $\sum_n \mu^n(x) = \infty$,
- x is transient iff $\sum_n \mu^n(x) < \infty$.

(ii) If one element of S is positive (null) recurrent, then all essential elements are positive (null) recurrent.

(iii) There are positive recurrent points if and only if the semigroup S contains a completely simple minimal ideal K with a finite group factor in its Rees representation. In that case, the type is independent of the generating measure μ as long as the irreducibility condition (3.9) is met.

Proof. Methods developed in the proofs of Propositions 3.5 and 3.6 are used fairly directly.

(i) We first show that x positive recurrent for the right random walk is equivalent to $\limsup \mu^n(x) > 0$. Equation (3.11) shows that the condition is necessary. To show the sufficiency, we use (3.12) to obtain

$$\mu^{n+1}(x) = \sum_y \mu(y) \sum_{k=0}^n f_{y,x}^k P_r^{n-k}(x, x) \leq \sum_{y \in S'} \sum_{k=0}^N f_{y,x}^k P_r^{n-k}(x, x) + \varepsilon, \quad (3.17)$$

where $0 < \varepsilon \ll \limsup \mu^n(x)$, S' is a large finite subset of S with $\mu(S')$ very close to 1, and N is some large fixed number making the tail of the convergent series $\sum_k f_{y,x}^k$ small. Then necessarily $\limsup_n P_r^n(x, x) > 0$.

By symmetry, the same recurrence criterion works for the left random walk as well.

As in the proof of Proposition 3.6, we see that x null recurrent for the right random walk implies x null recurrent for the bilateral one.

All essential elements are positive recurrent for the bilateral walk if one is. This may be demonstrated exactly as in the proof of the corresponding statement for recurrence (see formulas (3.15) and (3.16)). Using (3.17) here as (3.14) was used in the proof of Proposition 3.6, we conclude that if x is positive recurrent for the right random walk, then it is positive recurrent for the bilateral walk, too.

The proof for the mixed random walk is analogous. The mixed walk P_m satisfies (3.11), hence $\limsup \mu^n(x) > 0$ is a necessary condition for positive recurrence. P_m also satisfies (3.17) (if probabilities are interpreted with respect to the mixed random walk), so the condition is also sufficient. Proposition 3.6 showed that if x is recurrent for one walk, x is recurrent for the other three as well. Thus, we have proved Part (i).

(ii) The relation

$$\mu^k(s)\mu^n(x) \leq \mu^{k+n}(sx),$$

can be used to show that the recurrent elements form a two-sided ideal (stated in Proposition 3.6) and the positive recurrent elements form a two-sided ideal. Hence, if one element is positive (null) recurrent, then so are all elements of K , the completely simple minimal ideal of S . Outside the minimal ideal, all the elements are inessential.

(iii) We first note that

$$\mu^n(I) \rightarrow 1, \quad (3.18)$$

if I is any two-sided ideal of S . To prove this assertion, consider $x \in I$ with $\mu^k(x) > 0$ and a right random walk W_n on S . Write

$$W_{nk} = X_0(X_1 \cdots X_k)(X_{k+1}X_{k+2} \cdots X_{2k}) \cdots (\cdots X_{nk}).$$

If $W_{nk} \notin I$, then necessarily none of the n blocks of k factors X each can belong to I . Hence,

$$\mu^{nk}(I^c) \leq (\mu^k(I))^n \leq (1 - \mu^k(x))^n.$$

Let K be the completely simple minimal ideal of S , $K = E \times G \times F$ with G 629
finite. There are a positive integer k and a finite set $E' \times G \times F' = E'GF'$ such 630
that $\mu^k(E'GF') > 0$. (This number can be chosen arbitrarily close to 1, by (3.18).) 631
Our set $E'GF'$ is the intersection of a union of minimal right ideals with a union of 632
minimal left ideals, so 633

$$(E'GF')S(E'GF') = E'GF'KE'GF' = E'GF'.$$

This means that $\mu^{2k+n}(E'GF') \geq \mu^{2k}(E'GF') > 0$ for all n . Since $E'GF'$ is 634
finite this proves (by Part (i)) that positive recurrent points exist. By Part (ii) they 635
constitute the whole minimal ideal K . 636

Note that this argument holds true for any probability measure generating a semi- 637
group S with the same completely simple ideal K . 638

Now assume that there are positive recurrent points. These points form the com- 639
pletely simple minimal ideal $K = EGF$. 640

Put 641

$$\alpha(e) = \lim_n \mu^n(eGF), \quad e \in E \quad \text{and} \quad \beta(f) = \lim_n (EGf), \quad f \in F. \quad (3.19)$$

α and β are welldefined as limits of increasing sequences. (The μ^n -measure of any 642
right (left) ideal is an increasing function of n .) Also, they are *probability measures* 643
on E and F , respectively, because ultimately all the mass is concentrated on the 644
minimal ideal K (see (3.18)). Random walks on S are thus eventually absorbed by 645
the essential classes, with probability 1. 646

Consider the right random walk on S begun at some point $x \in S$. Suppose the 647
walk is periodic with period p . There is a probability distribution ${}_x\pi$ on xK , with 648
support xK such that 649

$${}_x\pi(y) = \lim_n \frac{1}{p} \sum_{k=0}^{p-1} \delta_x * \mu^{np+k}(y).$$

(In fact $\delta_x * \mu^{np+k}(y)$ converges for each k , $k = 0, 1, \dots, p-1$.) This follows from 650
(3.18) using the general theory of positive recurrent Markov chains (see [37], p. 33). 651

Then there is a finite set S' with $\mu(S') > 1 - \varepsilon$ and a finite subset K' of K such 652
that 653

$$\mu^{n+1}(K') > \sum_{x \in S'} \mu(x) \mu^n(x^{-1}K') > (1 - \varepsilon)^2,$$

for n larger than some n_0 . This means that the sequence μ^n , $n = 1, 2, 3, \dots$ is tight. 654

Theorem 2.7 can then be invoked to complete the proof of Part (iii). $\square$ 655

Actually, the theory in Chap. 2, for example, Theorem 2.7 gives us much more: 656

Corollary 3.2. *In the positive recurrent case we have for all $y \in K$* 657

$$\lim_n \frac{1}{p} \sum_{k=0}^{p-1} \mu^{np+k}(y) = \pi(y), \quad (3.20)$$

where π is a probability measure supported by K with

658

$$\mu * \pi = \pi * \mu = \pi$$

and

659

$$\pi(e, g, f) = \frac{\alpha(e)\beta(f)}{|G|}, \quad (e, g, f) \in E \times G \times F \equiv K, \quad (3.21)$$

with $|G|$ denoting the order of the finite group G and the probability measures α 660
and β are defined by (3.19). Furthermore, 661

$$\lim_n \mu^{np+k}(y) = \eta_k(y), \quad (3.22)$$

with η_k a probability measure on $E \times g^k H \times F$, g and H as in Theorem 3.2. 662

$$\mu * \eta_k = \eta_k * \mu = \eta_{k+1} \pmod{p}$$

and

663

$$\eta_k(e, gh, f) = \frac{\alpha(e)\beta(f)}{|H|}, \quad (e, gh, f) \in E \times gH \times F \quad (3.23)$$

Remark 3.4. Note that, in our discrete setting, the pointwise convergence of a se- 664
quence of probability measures to a probability measure is equivalent to weak 665
convergence. 666

Proof. We know from Chap. 2 that the Cesàro limit π of μ^n exists. The probability 667
measure π on $K = E \times G \times F$ is a product measure where the middle factor, χ_G , 668
is the uniform or normed Haar measure on G . Clearly, π must then satisfy (3.21), 669
i.e., $\pi = \alpha \times \chi_G \times \beta$. 670

In the proof of Theorem 3.3, we saw that sequences $\delta_x * \mu^{np+k}$, $n = 1, 2, \dots$ 671
converge for $k = 0, 1, 2, \dots, p-1$. From this, we can deduce that the sequences 672
 μ^{np+k} converge pointwise, thus weakly, for $k = 0, 1, \dots, p-1$. In particular, the 673
Cesàro limit π satisfies (3.20). 674

The weak limit η_0 of $\mu^{np} = (\mu^p)^n$, $n = 1, 2, \dots$ is supported by $E \times H \times F$, 675
where H is a normal subgroup of G (see Theorems 2.7, 2.13 and 3.2). η_0 must 676
satisfy (3.23). In other words, $\eta_0 = \alpha \times \chi_H \times \beta$. 677

The structure of the cyclic classes of the right random walk was studied in 678
Theorem 3.2. The support of $\eta_k = \eta_0 * \mu^k = \eta_0 * \eta_k$ is $E \times g^k H \times F$, the set 679
reached by the right random walk in $np + k$ steps from some point in $E \times H \times F$. 680
Since these supports are disjoint and the average of the η 's is equal to π , they must 681
satisfy (3.23) $\square$ 682

The ergodic theory of Markov chains describes the asymptotic behavior of a 683
typical realization of the chain. In particular if there is only one essential class for a 684
positive recurrent chain, then the time averages of a function h or the integrals of h 685
with respect to the *empirical distributions* converge almost surely to the integral of 686
 h with respect to the unique invariant measure π for the chain (see [37], p. 92). 687

For our positive recurrent unilateral, bilateral, and mixed random walks on S with completely simple minimal ideal $K = EGF$, we have by Proposition 3.3 and Theorem 3.2 only one essential class (at least) in the following cases: For the (left) random walk if $|E| = 1$ ($|F| = 1$) or if the starting point is in K , and for the bilateral walk, if p is odd or if the starting point is in K . There is always just one essential class for the mixed random walk.

We can then state the following Corollary 3.3

Corollary 3.3. *Let μ be a positive recurrent probability measure on S . Let the completely simple minimal ideal K of S have the representation EGF . Denote by S_n (T_n) the right (left) random walk generated by μ and by W_n (M_n) the bilateral (mixed) random walk generated by μ . If E is a singleton, then*

$$\begin{aligned} \frac{1}{n} \sum_{k=1}^n \delta_{S_k} &\rightarrow \chi_G \times \beta \text{ weakly almost surely, or} \\ \frac{1}{n} \sum_{k=1}^n h(S_k) &\rightarrow \int h d(\chi_G \times \beta) \text{ almost surely} \end{aligned} \quad (3.24)$$

where h is any bounded real-valued function on S and δ_x denotes the point mass at x . Equivalently,

$$\frac{1}{n} |\{k \leq n | S_k = (g, f)\}| \rightarrow \frac{\beta(f)}{|G|} \text{ almost surely.} \quad (3.25)$$

Similarly, if F is a singleton then

$$\frac{1}{n} \sum_{k=1}^n \delta_{T_k} \rightarrow \alpha \times \chi_G \text{ weakly almost surely.} \quad (3.26)$$

If p is odd it follows that

$$\frac{1}{n} \sum_{k=1}^n \delta_{W_k} \rightarrow \alpha \times \chi_G \times \beta \text{ weakly almost surely.} \quad (3.27)$$

For every positive recurrent μ , we have

$$\frac{1}{n} \sum_{k=1}^n \delta_{M_k} \rightarrow \alpha \times \chi_G \times \beta \text{ weakly almost surely.} \quad (3.28)$$

Theorem 3.3 shows that surprisingly enough, checking positive recurrence for random walks on a discrete semigroup is actually an easy task. It depends on the measure μ only through its support S_μ , or more exactly through the semigroup S generated by S_μ . One only needs to verify that S admits a minimal two-sided ideal $K = EGF$, where the group factor G is finite. Trivially, any random walk on G is positive recurrent.

The natural question is then whether there is a connection between the type of the semigroup S and that of the group factor G of its kernel $K = EGF$. (The discrete group G is said to be (positive / null) recurrent if there is an irreducible, in the sense of (3.9), (positive / null) recurrent random walk on G . If there is no such walk, G is a transient group.) As we show in Sect. 3.3, there are a number of recurrence criteria for groups that we would like to make use of in the semigroup context. Let us first look at some simple cases.

If S is an abelian semigroup or if S is an inverse semigroup, then its kernel, if it exists, must be a group. If $K = E \times G \times F$ is a *direct product*, i.e., the sandwich function $\phi(f, e) \equiv u$, where u is the identity of the group G , then for any right random walk $S_n = (E_n, G_n, F_n)$ on K , $G_n = uS_nu$ is a right random walk on G . Hence, any probability measure of the form $\mu = \alpha \times \nu \times \beta$ on $E \times G \times F$ has the property that μ is recurrent on K if and only if ν is recurrent on G . If K cannot be written as a direct product, then the projection G_n is not a random walk on the group component G . The relationship between the types of $E \times G \times F$ and G is still not completely understood. We have, however, many partial results.

Theorem 3.4. *Suppose the semigroup S is recurrent. Let $K = E \times G \times F$ be the kernel of S . Then G is a recurrent group.*

Proof. We first note that the positive recurrent case is covered by Theorem 3.3.

Suppose that μ generates a null recurrent right random S_n walk on S . Consider a maximal subgroup of the kernel K of S . This group is isomorphic with the middle factor G in the Rees–Suschkewitsch representation of K (see Sect. 1.4). For notational convenience, this group is also called G .

We consider S_n starting at the identity u of G . Define the stopping times (or optional random variables) $T_k, k = 1, 2, \dots$ successively by

$$T_1 = \min\{n > 0 | S_n \in G\} \quad T_{k+1} = \min\{n > T_k | S_n \in G\}, \quad k = 1, 2, \dots \quad (3.29)$$

(Of course, we define $T_k = \infty$ if the events do not occur.) The times are the successive return times to G and, since the random walk was assumed to be recurrent and G is included in the essential class $uS = uK$, they are all almost surely finite: T_k is no larger than the k th return time to u , which is almost surely finite, see [37] (1967), p. 81.

Define the process Y_n on G by

$$Y_n = S_{T_n}, n = 0, 1, 2, \quad (3.30)$$

(if T_n is infinite we can define Y_n arbitrarily). Define a probability measure ν on G by

$$\nu(g) = \mathbf{P}_u\{Y_1 = g\}. \quad (3.31)$$

Evidently

$$\nu(g) = \sum_{k=1}^{\infty} {}_G P_{u,g}^{(k)} = {}_G P_{u,g}^*,$$

using the *taboo probabilities* of [37], pp. 45–46, i.e., ${}_G P_{u,g}^{(k)}$ is the probability of the right random walk S_n starting from the state u of entering the state g at the k th step under the restriction that none of the states of G is visited in between, and ${}_G P_{u,g}^*$ is the probability of first hitting the set G at the state g . ν is a probability measure since T_1 is almost surely finite and Y_1 is thus a well-defined random variable taking values in G .

Note that $S_n = uX_1X_2 \cdots X_n$ is in G if and only if $gS_n = gX_1X_2 \cdots X_n$ is, for any $g \in G$. In fact, the event $S_n \in g^{-1}h$ has the same probability as the event $gS_n = h$, which means that ${}_G P_{g,h}^* = {}_G P_{u,g^{-1}h}^* = \nu(g^{-1}h)$.

The T s are stopping times; Y_n is, by Proposition 3.1, a Markov chain on G . It has initial state u and transition probabilities

$$Q(g, h) = \mathbf{P}_g\{Y_1 = h\} = \nu(g^{-1}h). \quad (3.32)$$

Y_n visits the state $g \in G$ exactly as many times as S_n does. Hence, the Y -chain is recurrent since the S -chain was assumed to be recurrent. But (3.32) shows that Y_n is actually a *right random walk* on G . We have thus shown that G admits a recurrent random walk. In other words, G is a recurrent group. $\square$

For the converse problem, we quote some of the best available results from the work of Larisse in [135]:

A measure ν on a group G is *symmetric* if $\nu(g) = \nu(g^{-1})$, $g \in G$.

Proposition 3.7. *Let G be a group and let $K = E \times G \times F$ with sandwich function ϕ . Denote by N the normal subgroup of G generated by $\phi(F, E)$. Then*

- (i) *If N is finite, then K is of the same type as G .*
- (ii) *If G is such that any symmetric probability measure ν on any generator of G is recurrent, then K is recurrent.*
- (iii) *If G is the limit of a sequence of finitely generated subgroups G_n with the property that any symmetric probability measure on any generator of G_n is recurrent, then both G and K are recurrent. In particular, if G is abelian then G and K are recurrent if and only if G is of rank at most two.*
- (iv) *If G is locally finite, i.e., any finite subset of G generates a finite subgroup of G , then G and K are recurrent.*

We will close this section with a useful application of the method of proof of Theorem 3.4 to *Markov random walks*.

Proposition 3.8. *Let the discrete semigroup S be given and assume that it admits a kernel K with a finite group factor. Let $X_0, X_1, X_2, \dots$ be a stationary Markov chain on S such that its invariant probability distribution is positive for some $a \in K$. Define*

$$S_0 = X_0 \quad \text{and} \quad S_n = S_{n-1}X_n, \quad n = 1, 2, \dots \quad (3.33)$$

Then S_n is positive recurrent.

Remark 3.5. Strictly speaking, positive recurrence for S_n is not defined since S_n is not a Markov chain. Here, we will adopt the following provisional definition: s is positive recurrent for S_n if, for some a , (a, s) is positive recurrent for the corresponding *Koutsky process* (X_n, S_n) , see Sect. 3.3, i.e.,

$$\limsup_n \mathbf{P}\{X_n = a, S_n = s | X_0 = a, S_0 = s\} > 0. \quad (3.34)$$

Proof. Let $a \in K$ be given such that a is positive recurrent for the stationary Markov chain X_n . Consider the random return times to a :

$$T_0 = 0 \quad \text{and} \quad T_n = \min\{k > T_{n-1} | X_k = a\}, \quad n = 1, 2, \dots \quad (3.35)$$

(and $T_n = \infty$ if the defining set is empty). The return times are all almost surely finite. The mean interarrival time τ is finite.

Let $Y_0 = a$ and

$$Y_{n+1} = X_{T_{n+1}} X_{T_{n+2}} \cdots X_{T_{n+1}-1} \cdot a, \quad n = 0, 1, 2, \dots \quad (3.36)$$

To see what this means in practice take $n = 1$. Y_2 is the block of factors in the product S_n occurring between the first occurrence of a , at time T_1 , and the second occurrence of a at time T_2 . The blocks $Y_1, Y_2, \dots$ are independent and identically distributed since X_n is a Markov chain. A detailed description of this block, or excursion, technique is given, for example, in [8] (see also Proposition 3.1 and our treatment of Markov random walk at the end of Sect. 3.3 as well as [199], Chap. 1, Sect. 3).

Consider the original process S_n started at a , $S_{T_n} = aY_1Y_2 \cdots Y_n$ is a random walk on S . Actually, its state space belongs to the set aSa which is a subset of K and also a finite group. Consequently, the random walk is positive recurrent and its state space is a subgroup H of aSa . Let u be the identity of H (and aSa). Then

$$\limsup_n \mathbf{P}_{(a,b)}\{S_{T_n} = u\} > \delta > 0,$$

for any $b \in H$. Recall that $X_{T_n} = a$. (If the random walk with increments Y_n is aperiodic, the limit exists and is strictly positive.) In particular, $\mathbf{P}\{bX_1X_2 \cdots X_{N-1}X_N = u, X_N = a\} > \delta' > 0$ for some $\delta' \leq \delta$ and some (large) $N = N_b$.

Let us assume, for simplicity, that the Markov chain X_n is aperiodic. Then $\mathbf{P}\{X_n = a\} \geq \varepsilon > 0$ for all n large enough and $\mathbf{P}_{(a,u)}\{X_{n+N_j} = a, S_{n+N_j} = u \text{ some } j \in H\} \geq \sum_{b \in H} \mathbf{P}_{(a,u)}\{X_n = a, S_n = b, X_{n+N_j} = a, S_{n+N_j} = u \text{ some } j \in H\} \geq \varepsilon\delta' > 0$. H is finite so (a, u) is positive recurrent. $\square$

Exercise 3.8. Let S_μ be a subset of a finite group G . Suppose that S satisfies the irreducibility condition (3.9). Prove that S is a subgroup of G .

Exercise 3.9. Show that S in Example 3.2(i) is simple but not completely simple.

Exercise 3.10. Construct transformation semigroups S where the normal subgroup N of G , see Proposition 3.7, is nontrivial. 810
811

Exercise 3.11. Consider the four random walks on the bicyclic semigroup $\mathcal{C}$ (see Definition 1.11). Determine the essential elements, if any. 812
813
Is $\mathcal{C}$ recurrent for any of the random walks? 814

Exercise 3.12. Show that the multiplicative semigroup of nonzero Gaussian integers $\{m + ni \neq 0 | m, n \text{ integers}\}$ is transient. 815
816
Is the additive group of Gaussian integers recurrent? 817

Exercise 3.13. Consider the additive group of integers $\mathbf{Z}$. Let $\mu(1) = \mu(-1) = \frac{1}{2}$. 818
The random walk generated by μ is called the symmetric simple random walk on $\mathbf{Z}$. 819
Calculate $P^n(0, 0)$ and prove that the random walk is recurrent. 820

Exercise 3.14. In $\mathbf{Z}^2$, let μ be the uniform distribution on the four nearest neighbors 821
of the origin: $(\pm 1, 0), (0, \pm 1)$. Prove that the simple symmetric random walk on the 822
additive group $\mathbf{Z}^2$ is recurrent by using the criterion in Proposition 3.5(c). 823

Exercise 3.15. In $\mathbf{Z}^3$, let μ be the uniform distribution on the six nearest neighbors 824
of the origin: $(\pm 1, 0, 0), (0, \pm 1, 0), (0, 0, \pm 1)$. Prove that the simple symmetric 825
random walk on the additive group $\mathbf{Z}^3$ is transient by using the criterion in 826
Proposition 3.5(c). 827

Exercise 3.16. Suppose that the Rees product $K = E \times G \times F$ is direct. Prove that 828
 G recurrent implies K recurrent. 829

Exercise 3.17. Assume that xy^{-1} and $y^{-1}x$ are finite for all x, y in S . Let μ and 830
 ν be probability measures on S and let $\mu * \nu$ be their convolution. Prove that the 831
convolution operation is continuous in the vague topology. 832

Exercise 3.18. Let $0 < a < 1$ and let the probability measure β be given by 833

$$\beta = (1 - a) \sum_{n=1}^{\infty} a^{n-1} \mu^n.$$

Suppose μ is recurrent. Is β also recurrent? 834

Exercise 3.19. Let $G = \{(\varepsilon, n) | \varepsilon = \pm 1, n \in \mathbf{Z}\}$ with the multiplication defined by 835

$$(\varepsilon, n)(\varepsilon', n') = (\varepsilon\varepsilon', n + \varepsilon n').$$

Prove that G is recurrent. (Hint: Use the visits to the subgroup $\{(1, n) | n \in \mathbf{Z}\}$ as in 836
the proof of Theorem 3.4.) 837

3.3 Locally Compact Groups

838

In Sect. 3.3, we answer the basic questions about the asymptotics of the random walks defined in Sect. 3.1 under the assumption that the state space G is a locally compact second-countable group (or a subsemigroup of such a group).

The treatment will be self-contained when general results and general criteria for recurrence/transience are discussed. There are many excellent expositions specializing in probabilistic analysis of random walks on groups, so we will treat particular group theoretical results rather summarily. Relevant literature on the subject include the books by Bougerol and Lacroix [17], Heyer [92], Revuz [199], Guivarc'h et al. [87], and the lecture notes by Roynette [207].

Let μ be a probability measure on the locally compact second countable group G . The random walks (left, right, bilateral, mixed) on G are defined in Definition 3.1. If $Z_n = Z_{n-1}X_n$, $Z_0 = x$ is the right random walk started at x we denote by $Z_m^{-1}Z_n$ ($m < n$) the random variable $X_{m+1} \cdots X_{n-1}X_n$. Note that Z_m and $Z_m^{-1}Z_n$ are independent and distributions w. r. t. $\mathbf{P}_e$ of Z_{n-m} and $Z_m^{-1}Z_n$ are the same. The same remarks hold for the left random walk and the random variables $Z_nZ_m^{-1} \equiv X_nX_{n-1} \cdots X_{m+1}$.

The measure μ is said to be *adapted* to G if G itself is the minimal closed subgroup containing the support S_μ of μ . Equivalently, G is the smallest closed subgroup containing the smallest closed subsemigroup S generated by the support of μ ,

$$S \equiv \overline{\bigcup_{n=1}^{\infty} S_\mu^n}$$

(see the irreducibility condition (3.9)). We shall henceforth assume that the probability measures generating our random walks are adapted to the group in question.

In most cases of interest to us S and G actually coincide:

Proposition 3.9. *If one of the unilateral walks admits essential elements, then S is a group and all elements of S are essential for the right, left, bilateral, and mixed random walks. S then constitutes one essential class w. r. t. the unilateral and mixed walks and at most two essential classes for the bilateral walk.*

Proof. Suppose that x is essential for the right random walk on S . The sets sS are closed for all $s \in S$ (Proposition 1.16) so the proof of Proposition 3.3, Assertion (ia) can be applied without change to show that xS is a minimal closed ideal R of S : $rS = R$ for any $r \in R$.

Clearly, R is a subsemigroup of S . Also, $r^n \in rS$ for $n = 1, 2, \dots$ so $r \in r^nS$. This implies $r^{-n} \in S$ and $r^{-n+1} \in R$. Thus, $r^{-1} \in R$ and R is a group. Since the identity e of G belongs to R we can conclude that $R = eS = S$. By Ellis's theorem (Theorem 1.4), S is a topological group. Since S is closed, it must coincide with G .

If S is a topological group, then $xS = Sx = SxS = S$ so the proof of Proposition 3.3 shows that all elements of S are essential for both the right and the left and the mixed random walks. There is only one essential class, all of S , for these walks.

To prove the assertions for the bilateral walk, we first notice that μ^2 generates a
subsemigroup

$$H \equiv \overline{\bigcup_n S_{\mu^2}^n}$$

of G . By considering that $t_n \in S_{\mu}^n$ with $\lim t_n = e$ implies $\lim t_n^2 = ee = e$ we see
that $e \in H$. Assume that $H \neq S$. For $s, t \in S_{\mu}$ sH is closed and hence

$$S = sS = sH \cup s(S \setminus H) \text{ and } S = stS = stH \cup st(S \setminus H).$$

Since $stH \subset H$ this means that $stH = H$ or $hH = H$ for all $h \in S_{\mu}^2$. Similarly,
 $hH = H$ for all $h \in H$. As above one concludes that H is a group, a closed
subgroup of $S = G$. Furthermore, H is normal in G .

The essential elements for the bilateral walk form an ideal of S . To prove this,
assume that x is essential and let xs be any element of xS . Since

$$s \in \overline{\bigcup_n S_{\mu}^n},$$

we can find a sequence $s_{n'} \in S_{\mu}^{n'}$ converging to s . Let y be such that $xs \rightarrow y$, i.e.,

$$y = \overline{\bigcup_k S_{\mu}^k xs S_{\mu}^k}.$$

Then

$$sy \in \overline{\bigcup_k S_{\mu}^k xs S_{\mu}^k} \in \overline{\bigcup_{n'} \bigcup_k S_{\mu}^{k+n'} x S_{\mu}^{k+n'}}$$

so $x \rightarrow sy$. Since x is essential, $sy \rightarrow x$ or

$$x \in \overline{\bigcup_k S_{\mu}^k sy S_{\mu}^k} \text{ and } xs \in \overline{\bigcup_{n'} \bigcup_k S_{\mu}^{k+n'} y S_{\mu}^{k+n'}}$$

implying $y \rightarrow xs$. In other words, xs is essential. Similarly, sx and sxt are essential
for all $s, t \in S$.

$e \rightarrow h$ in the bilateral walk if and only if $x \in H$. The argument in the preceding
paragraph ($xs \rightarrow y \Rightarrow x \rightarrow sy$) shows that $e = hh^{-1} \rightarrow h$ implies $h \rightarrow h^{-1}h = e$.
Hence, e is essential for the bilateral walk. Since the essential elements form an
ideal all elements of the group $S = SeS$ are essential for the bilateral walk. H is,
of course, an essential class. If $s \notin H$, then still $ss^{-1} \rightarrow h$ and $s \rightarrow s^{-1}h$ for any
 $h \in H$ proving that $sH = S \setminus H$ is the other essential class. $\square$

Note: Example 3.2(i) shows that the bilateral or mixed walks may admit essential
elements without S being a group.

According to Definition 3.4, x is recurrent for the random walk Z_n if $\mathbf{P}_x\{Z_n \in$
 $N(x) \text{ infinitely often}\} = 1$ for every neighborhood $N(x)$ of x .

Lemma 3.1. *If the right (left) random walk admits a recurrent element, then $S = G$. All elements of S are recurrent if one is.*

Proof. Consider the right random walk on S started at $x \in S$. Clearly, $xX_1X_2 \cdots X_n \in N(x)$ if and only if $eX_1X_2 \cdots X_n \in x^{-1}N(x)$. $N(x)$ is a compact neighborhood of x if and only if $x^{-1}N(x)$ is a compact neighborhood of the identity e . Hence, $x \rightarrow x$ infinitely often if and only if $e \rightarrow e$ infinitely often. In particular, the existence of recurrent elements implies $e \in S$. Actually, $s \in S$ is equivalent to $e \rightarrow s$. It remains to show that $s \rightarrow e$, i.e., e is essential and S is in fact a group.

Suppose s does not lead to e (which is equivalent to saying that $s^{-1} \notin S$). Then there are compact neighborhoods $N(s)$ of s , $N(e)$ of e and $V(s^{-1}) \supset N(s)^{-1}N(e)$ of s^{-1} such that $\mu^k(N(s)) > 0$ for some k but

$$\mu^n(N(s)^{-1}N(e)) \leq \mu^n(V(s^{-1})) = 0$$

for all n . Consider the probabilities

$$\begin{aligned} \mathbf{P}_e\{Z_n \in N(e) \text{ finitely often}\} &\geq \mathbf{P}_e\{eX_1X_2 \cdots X_n \notin N(e) \text{ for } n > k\} \\ &\geq \mathbf{P}_e\{eX_1X_2 \cdots X_k \in N(s), \\ &\quad X_k^{-1}X_n \notin N(s)^{-1}N(e)\} \\ &\geq \mu^k(N(s)) \cdot (1 - \mu^{n-k}(N(s)^{-1}N(e))) \\ &= \mu^k(N(s)) \cdot 1 > 0, \end{aligned}$$

contradicting the assumption $e \rightarrow e$ infinitely often. Hence, s^{-1} belongs to S and S coincides with the group G . $\square$

Proposition 3.10. *The left and right random walks are recurrent if and only if the return time to any neighborhood of the identity is $\mathbf{P}_e$ -almost surely finite.*

Proof. Let Z_n be the right random walk generated by μ and beginning at e . Let S_N be the return time to the compact neighborhood N of e : $S_N = \min\{n \geq 1 | Z_n \in N\}$. As usual, S_N is defined to be infinite if the defining set is empty. If e is recurrent, then clearly S_N is finite with probability one.

Conversely, assume that S_N is finite almost surely for all neighborhoods N of the identity. Let K and V be symmetric neighborhoods of e with K compact and V open and $K \subset V$. (A is said to be a *symmetric* subset of G if $A^{-1} = A$.) If V^c denotes the complement of V , then KV^c is a closed set (see Proposition 1.16) not containing e . $U \equiv (KV^c)^c$ is an open neighborhood of e .

Consider the event $\{Z_m \in K, Z_{m+n} \notin V\} \subset \{Z_m \in K, Z_m^{-1}Z_{m+n} \in K^{-1}V^c\}$. Since $Z_m^{-1}Z_{m+n}$, $n = 1, 2, \dots$ are independent of Z_m and have the same distribution as Z_n , $n = 1, 2, \dots$,

$$\begin{aligned} \mathbf{P}_e\{Z_m \in K, Z_{m+n} \in V^c, n = 1, 2, \dots\} \\ \leq \mathbf{P}_e\{Z_m \in K\} \mathbf{P}_e\{Z_m^{-1}Z_{m+n} \in U^c, n = 1, 2, \dots\} \\ = \mu^m(K) \mathbf{P}_e\{Z_n \notin U, n = 1, 2, \dots\}. \end{aligned}$$

The last probability is $\mu^m(K) \mathbf{P}_e\{S_U = \infty\}$, which is 0 by assumption.

Let $K_i, i = 1, 2, \dots$ be an increasing sequence of compact neighborhoods of the origin such that

$$\bigcup_i K_i = V.$$

Since

$$0 = \lim_i \mathbf{P}_e\{Z_m \in K_i, Z_{m+n} \notin V\} = \mathbf{P}_e\{Z_m \in V, Z_{m+n} \notin V\}$$

we conclude that the last exit time T from V is infinite with probability 1: Put $T = \sup\{n \geq 0 | Z_n \in V\}$ and notice that $0 \leq T \leq \infty$.

$$\begin{aligned} \mathbf{P}_e\{T < \infty\} &= \sum_{m=0}^{\infty} \mathbf{P}_e\{T = m\} \\ &= \sum_{m=0}^{\infty} \mathbf{P}_e\{Z_m \in V, Z_{m+n} \notin V, n = 1, 2, \dots\} = 0 \end{aligned}$$

This means that

$$\mathbf{P}_e\{Z_n \in V \text{ finitely often}\} = \mathbf{P}_e\{T < \infty\} = 0$$

so $Z_n \in V$ infinitely often $\mathbf{P}_e$ -almost surely. e is recurrent since for any neighborhood N of e we can find a symmetric open V contained in N . (If N_0 is the interior of N , take $V = N_0 \cap N_0^{-1}$.) $\square$

Theorem 3.5. (i) For unilateral walks either all or none of the elements of S are recurrent.

(ii) The following conditions are equivalent for a left or right random walk Z_n :

- (a) x is recurrent.
- (b) $\mathbf{P}_e\{Z_n \in N(x) \text{ infinitely often}\} > 0$ for any neighborhood $N(x)$ of x (unconditional recurrence).
- (c) e is recurrent.
- (d) Equation (3.37) holds

$$\sum_{n=1}^{\infty} \mu^n(N) = \infty \quad (3.37)$$

for some compact neighborhood N of the identity element e .

(e) Equation (3.37) holds for all neighborhoods N of e .

(iii) The left random walk generated by μ is recurrent if and only if the right random walk is recurrent.

Proof. Part (i) was shown in Lemma 3.1.

Part (ii). The equivalence of Conditions (iia) and (iic) was proved in Lemma 3.1. Clearly, Condition (iie) implies Condition (iid).

Notice that the expression in (3.37) can be written as

955

$$\mathbf{E}_e \left\{ \sum_{n=1}^{\infty} 1_N(Z_n) \right\} = \sum_{n=1}^{\infty} \mathbf{P}_e \{ Z_n \in N \} = \infty. \quad (3.38)$$

The so-called direct half of the Borel–Cantelli lemma (see [19], p. 41) tells us that if the event $\{Z_n \in N \text{ infinitely often}\}$ has positive probability then the sum in (3.37) diverges. Hence, Condition (iic) implies Condition (iie).

Let us assume that (3.37) holds for all compact symmetric neighborhoods K of e . As in the proof of Proposition 3.10 let T be the last exit time from K . We note that $Z_m \in K$ and $Z_m^{-1} Z_{m+n} \notin K^2$ implies $Z_{m+n} \notin K$ (because $K^{-1} \cdot K \subset K^2$). Then

$$\begin{aligned} & \sum_{m=0}^{\infty} \mathbf{P}_e \{ Z_m \in K, Z_m^{-1} Z_{m+n} \notin K^2, n = 1, 2, \dots \} \\ & \leq \sum_{m=0}^{\infty} \mathbf{P}_e \{ Z_m \in K, Z_{m+n} \notin K, n = 1, 2, \dots \} \\ & = \mathbf{P}_e \{ T < \infty \} \leq 1. \end{aligned} \quad (3.39)$$

Further

962

$$\begin{aligned} & \sum_{m=0}^{\infty} \mathbf{P}_e \{ Z_m \in K, Z_m^{-1} Z_{m+n} \notin K^2, n = 1, 2, \dots \} \\ & = \sum_{m=0}^{\infty} \mathbf{P}_e \{ Z_m \in K \} \cdot \mathbf{P}_e \{ Z_n \notin K^2, n = 1, 2, \dots \} \\ & = \sum_{m=0}^{\infty} \mu^m(K) \cdot \mathbf{P}_e \{ S_{K^2} = \infty \}. \end{aligned} \quad (3.40)$$

The divergence of the series in (3.37) shows that the return time S_{K^2} to the compact neighborhood K^2 is $\mathbf{P}_e$ -almost surely finite. Again, for any neighborhood N of e we can construct a compact neighborhood K with $K^2 \subset N$. Hence, S_N is almost surely finite for all neighborhoods N of e . By Proposition 3.10, e is recurrent. We have thus shown that Condition (iie) implies Condition (iic).

Assume that there is a compact set K satisfying $\sum_n \mu^n(K) = \infty$. By compactness, there is a point $x \in K$ such that

$$\sum_{n=1}^{\infty} \mu^n(N(x)) = \infty \quad (3.41)$$

for all neighborhoods $N(x)$ of x . If not, each $k \in K$ has a neighborhood $V(k)$ for which the series converges. $\bigcup_{k \in K} V(k)$ is an open covering of K and contains a finite subcovering $V(k_1), \dots, V(k_m)$. Then

$$\sum_n \mu^n(K) \leq \sum_{i=1}^m \sum_n \mu^n(K_i) < \infty,$$

a contradiction.

We mimic the proof of the implication Condition (iie) $\Rightarrow$ Condition (iic) with K a compact symmetric neighborhood of e and T the last exit time from xK (which is a compact neighborhood of x). Again, $Z_m \in xK, Z_m^{-1}Z_{m+n} \notin K^2$ implies $Z_{m+n} \notin xK$ (because $(xK)^{-1} \cdot xK \subset K^2$). Then the inequalities corresponding to (3.39) and (3.40) yield

$$1 \leq \sum_{m=0}^{\infty} \mathbf{P}_e\{Z_m \in xK\} \mathbf{P}_e\{Z_n \notin K^2, n = 1, 2, \dots\}$$

and $\mathbf{P}_e\{S_{K^2} = \infty\} = 0$ and e is recurrent. This proves that Condition (iid) $\Rightarrow$ Condition (iic).

It remains to prove that Condition (iib) is equivalent to the other assertions in Part (ii).

Condition (iib) $\Rightarrow$ Condition (3.41), again by the Borel–Cantelli Lemma. We just saw that Condition (3.41) implies Condition (iic).

Conversely, let e be recurrent. Using the method of proof of Proposition 3.10 with xK substituted for K and xV for V , we get

$$\mathbf{P}_e\{Z_m \in xK, Z_{m+n} \notin xV\} \leq \mathbf{P}_e\{Z_m \in xK\} \mathbf{P}_e\{Z_m^{-1}Z_{m+n} \in K^{-1}V^c\}.$$

The last exit time from the set xK is infinite $\mathbf{P}_e$ -almost surely and hence $\mathbf{P}_e\{Z_n \in xV \text{ infinitely often}\} = 1$. This completes the proof of Part (ii).

Part (iii) follows from (3.37). $\square$

Corollary 3.4. *The left and right random walks are transient if and only if $\sum_n \mu^n$ is a Radon measure on S . In particular, any left or right random walk on a compact group is recurrent.*

Proof. If one of the random walks is recurrent, then $\sum_n \mu^n(K) = \infty$ for many compact sets K . In other words, $\sum_n \mu^n$ is not a Radon measure.

Conversely, $\sum_n \mu^n(K) = \infty$ for a compact set K implies the existence of a point satisfying (3.41), which in turn implies that e is recurrent.

$\sum_n \mu^n(S)$ always diverges, so if S is compact then the random walk is necessarily recurrent. $\square$

Corollary 3.5. *The left and right random walks are recurrent if and only if the hitting time of a relatively compact open subset of S is $\mathbf{P}_e$ -almost surely finite.*

Proof. This was shown above in the proof of the equivalence of Conditions (iib) 1001
and (iic) in Theorem 3.5. $\square$ 1002

A topological group G is a *recurrent group* if there is a probability measure μ 1003
adapted to G satisfying (3.37). Hence, a group is recurrent if and only if it admits 1004
an irreducible (in the sense of (3.9)) recurrent right or left random walk. 1005

Remark 3.6. (i) A compact group is recurrent (see Corollary 3.4). An open sub- 1006
group H of a recurrent group G is recurrent, since the open neighborhoods of 1007
the identity in H are open sets in G as well and thus satisfy (3.37). 1008

(ii) $\mathbb{R}^1$, $\mathbb{R}^2$, $\mathbb{Z}$ (the additive group of integers) and $\mathbb{Z}^2$ are recurrent. 1009

It turns out that on $\mathbb{R}^1$ or $\mathbb{Z}^1$ the random walk generated by μ is recurrent if 1010
 $\int |x| d\mu(x) < \infty$ and $\int x d\mu(x) = 0$. For $\mathbb{R}^2$ or $\mathbb{Z}^2$ or $\mathbb{R}^1 \oplus \mathbb{Z}^1$ (where $\oplus$ 1011
stands for direct sum), a sufficient condition for recurrence is $\int |x|^2 d\mu(x)$ and 1012
 $\int x d\mu(x) = 0$. For a comprehensive treatment of recurrence criteria like these, 1013
the reader is referred to [199]. 1014

(iii) The direct product of a recurrent group and a compact group is again recurrent. 1015
The multiplicative group C^* of nonzero complex numbers is isomorphic to the 1016
direct product of circle group and the multiplicative group of strictly positive 1017
real numbers $\mathbb{R}_+^*$, which in turn is isomorphic to the recurrent additive group 1018
 $\mathbb{R}^1$. Hence, C^* is recurrent. 1019

(iv) The semidirect product $G_2 = T \times C$ of the circle group T and the additive 1020
group of complex numbers, where the multiplication is given by 1021

$$(e^{i\theta}, c)(e^{i\theta'}, c') = (e^{i(\theta+\theta')}, c + e^{i\theta}c')$$

is recurrent. For the probability measure μ on G_2 to be recurrent only the sec- 1022
ond moment needs to be finite. μ need not be centered. For a proof, see [49]. 1023

The smoother μ the more precise are the results one expects to derive for random 1024
walks on the group G . Such a smoothness condition is the nonsingularity of μ , or 1025
some power of it, with respect to a Haar measure on G . A continuity condition on the 1026
transition probability kernel also works. Indeed, these two approaches, developed in 1027
[199], Chap. 3, and [229], respectively, can be shown to be equivalent in the context 1028
of random walks on groups. 1029

Generally speaking, it is very difficult to extend the methods, and even some of 1030
the most basic ideas, from the theory of Markov chains on a discrete state space to 1031
chains on a topological state space. In the final analysis, the concepts of spreadout- 1032
ness and continuous components (Definition 3.3.) are successful precisely because 1033
they make the extension from discrete to topological possible. 1034

Definition 3.7. A probability measure μ on the topological group G is said to be 1035
spread out if some convolution power μ^n is not singular with respect to the (right) 1036
Haar measure m on G . A random walk on G is *Harris recurrent* if the return time 1037
to any m -positive set is almost surely finite, $\mathbf{P}_x\{S_A < \infty\} = 1$ for all $x \in G$ and all 1038
 $A \in \mathcal{B}$ with $m(A) > 0$. 1039

Example 3.3. (i) μ may be singular but nevertheless spread out. Consider, e.g., the group G of upper triangular matrices

$$\begin{pmatrix} 1 & x & z \\ 0 & 1 & y \\ 0 & 0 & 1 \end{pmatrix} \quad x, y, z \in \mathfrak{R}.$$

Lebesgue measure on $\mathfrak{R}^3$ is a Haar measure on the group. Let μ be an absolutely continuous (w. r. t. two-dimensional Lebesgue measure) measure on $\{(x, y, 0) | x, y \in \mathfrak{R}\}$ with everywhere positive density of the form $f(x)g(y)$. Clearly, μ is singular but μ^2 is nonsingular with respect to Haar measures on G .
(ii) The left and right Haar measures on the affine group

$$\begin{pmatrix} x & y \\ 0 & 1 \end{pmatrix}, \quad x > 0, y \in \mathfrak{R}$$

are different. For an f defined on $\{(x, y) | x > 0\}$ (f continuous with compact support, say), the integral

$$\int_{-\infty}^{\infty} \int_0^{\infty} \frac{f(x, y)}{x} dx dy$$

is invariant with respect to multiplication from the right by an element of the affine group (see [115], Chap. 6). The corresponding left invariant integral is

$$\int_{-\infty}^{\infty} \int_0^{\infty} \frac{f(x, y)}{x^2} dx dy.$$

A product measure of the form $\mu = \alpha \times \delta_b$ is singular with respect to the Haar measures on the group. If α is absolutely continuous (w. r. t. one-dimensional Lebesgue measure) and $b \neq 0$, then μ^2 is nonsingular and μ^3 has a density with respect to the Haar measures $\frac{dx dy}{x}$ and $\frac{dx dy}{x^2}$.

Lemma 3.2 gives some elementary but important properties of absolutely continuous and spread-out measures. For a detailed account of Haar measures and convolutions, the reader is referred to, for example, [115], Chap. 6, or [89], Chaps. 4 and 5.

Lemma 3.2. (i) Suppose μ has compact support and density f with respect to a right Haar measure m on G . Then the transition probability kernel of the right random walk generated by μ is

$$P_r(x, B) = \mu(x^{-1}B) = \int 1_B(xy) f(y) m(dy) = \Delta(x)(1_B * \check{f})(x), \quad (3.42)$$

where $x \in G$, $B \in \mathcal{B}$, and Δ is the modular function of G and $\check{f}$ is the function defined by $\check{f}(x) = f(x^{-1})$, $x \in G$.

Also,

1063

$$\mu^2(dx) = (f * f)m(dx) \text{ or } \mu^2(B) = \int_B (f * f)(x)m(dx). \quad (3.43)$$

- (ii) μ is spread out if and only if for some k and some nonempty open set A μ^k dominates a multiple of right Haar measure on A . 1064
- (iii) Let D_μ be the union of those open sets A on which μ^k for some k dominates a multiple of right Haar measure m . If D_μ is nonempty, it is a two-sided open ideal of S . 1065
1066
1067
1068

Proof. For a proof of Part (i), see [89], Chaps. 4 and 5. Note that the convolution of a bounded measurable function and an integrable function is *continuous* on G . The modular function Δ is a continuous homomorphism from G into the positive real numbers. 1069
1070
1071
1072

Part (ii): The condition is obviously sufficient. Suppose, then, that $\mu^k = a_k + b_k$ where a_k is the absolutely continuous part and b_k is the singular part of μ^k w. r. t. m . If f_k is the density of a_k on a suitable compact set K , then μ^{2k} dominates the measure $(f_k * f_k)m$. Since the density is continuous and $\neq 0$, we can then find a positive c and an open set A such that $\mu^{2k} \geq cm$ on A . 1073
1074
1075
1076
1077

Part (iii): D_μ is an open set contained in S , the smallest closed subsemigroup of G generated by the support of μ . Take $s \in D_\mu$ and $t \in S$. Let $A \ni s$ be an open set such that $\mu^k \geq cm$ on A (c is a positive constant). We can find a symmetric neighborhood V of the identity as well as neighborhoods $N(t)$ of t and $N(st)$ of st such that 1078
1079
1080
1081
1082

$$N(t)N(t)^{-1} = V, \quad N(st)V \subset At \text{ and } \mu^n(N(t)) \geq \delta > 0$$

for some k . Then for any $B \subset N(st)$ we have $BN(t)^{-1} \subset A$ and hence 1083

$$\mu^{k+n}(B) \geq \int_{N(t)} \mu^k(By^{-1})\mu^n(dy) \geq cm(B)\delta.$$

Hence, μ^{k+n} dominates a multiple of m on $N(st)$ so $st \in D_\mu$. □ 1084

Proposition 3.11. Let μ be an adapted probability measure on G . The right (left) random walk generated by μ has a nontrivial continuous component if and only if μ is spread out. 1085
1086
1087

Proof. Let ν be the probability measure $\sum_{n=1}^{\infty} 2^{-n} \mu^n$ and $T(x, B)$ the continuous component of 1088
1089
1090

$$\sum_{n=1}^{\infty} 2^{-n} P_r^n(x, B) = \nu(x^{-1}B), \quad x \in S, B \in \mathcal{B}.$$

If μ is spread out, then ν dominates a multiple of a right Haar measure on some nonempty open set A . A can be chosen to be relatively compact. Equation (3.42) 1091
1092

can be applied to the measure μ' with density $f = c1_A$. The modular function Δ 1093
and the convolution $1_B * \check{f}$ are continuous so $T(x, B) \equiv \mu'(x^{-1}B)$ is a continu- 1094
ous component of the right random walk. It is an everywhere nontrivial continuous 1095
component because 1096

$$T(x, xA) = \mu'(x^{-1}(xA)) = \mu'(A) = cm(A) > 0.$$

Conversely, suppose the random walk has a nontrivial continuous component T 1097
at x_0 and μ is not spread out. There is a set B with $\nu(x_0^{-1}B) = 1$ and $m(B) = 0$. 1098
Since T is nontrivial at x_0 , there is a compact set $C \subset B$ such that $T(y, C) > 0$ for 1099
all y in some neighborhood N' of x_0 . Hence, there is a δ such $\nu(y^{-1}C) \geq \delta > 0$ 1100
for all y in some (smaller) open set N . 1101

$$\begin{aligned} m(C) &= \int m(Cz^{-1})\nu(dz) \\ &= \int m(dy)\nu(y^{-1}C) \geq \int_N m(dy)\nu(y^{-1}C) \\ &\geq m(N)\delta > 0, \end{aligned}$$

contradicts the assumption $m(B) = 0$. 1102

The assertion for the left random walk follows since the left and right Haar mea- 1103
sures are equivalent. In fact, if m is a right Haar measure then $m(dx)\Delta(x)$ is a left 1104
Haar measure. $\square$ 1105

Theorem 3.6. *Let μ be an adapted probability measure on G . Suppose that the* 1106
right random walk generated by μ is recurrent. Then it is Harris recurrent if and 1107
only if μ is spread out. 1108

Proof. Suppose μ is not spread out. Then there is a set B with $\nu(B) = 1$ and 1109
 $m(B) = 0$ where $\nu = \sum_n 2^{-n}\mu^n$. Let Z_n be the right random walk generated 1110
by μ . $\mathbf{P}_e\{Z_n \in B^c\} = 0$ for all n so $\mathbf{P}_e\{S_{B^c} < \infty\} = 0$, i.e., Z_n is not Harris 1111
recurrent. 1112

Suppose on the contrary μ is spread out. The random walk is recurrent so S 1113
coincides with G and hence, by Lemma 3.2, Part (iii), $D_\mu = G$. 1114

Let A be a compact set with positive Haar measure. It suffices to show that 1115
 $\mathbf{P}_e\{S_A < \infty\} = 1$, since $m(xA) > 0 \Leftrightarrow m(A) > 0$ for all $x \in S$. 1116

Choose n such that $\sum_1^n \mu^i$ dominates a multiple of m on A . (To find n consider 1117
an open covering of A with sets O_x on which μ^{n_x} dominates some multiple of 1118
a right Haar measure.) As in Lemma 3.2, Part (i), the kernel $\sum_1^n \mu^i(x^{-1}A)$ has a 1119
continuous component ($\neq 0$) at the identity e , so 1120

$$\sum_1^n \mu^i(x^{-1}A) \geq \delta > 0$$

on some neighborhood N of the identity. This means

1121

$$\mathbf{P}_x\{Z_i \in A, \text{ some } i, 1 \leq i \leq n\} \geq \frac{\delta}{n}.$$

for $x \in N$.

1122

Let $T_1, T_2, \dots$ be the almost surely finite successive return times to N separated by at least n steps:

1123

1124

$$T_{i+1} = \inf\{j > T_i + n | Z_j \in N\}.$$

Then

1125

$$\mathbf{P}_e\{S_A > T_2\} \leq \mathbf{P}_{Z_{T_1}}\{Z_1, Z_2, \dots, Z_n \notin A\} \leq 1 - \frac{\delta}{n}$$

(by the strong Markov property, Proposition 3.1) and

1126

$$\mathbf{P}_e\{S_A > T_{i+1}\} \leq (1 - \frac{\delta}{n})^i.$$

Hence, $\mathbf{P}_e\{S_A < \infty\} = 1$ and the random walk is Harris recurrent.

□ 1127

It is shown in [229] that recurrent and conservative are essentially equivalent properties in the presence of a continuous component. We formulate this important result as Proposition 3.12

1128

1129

1130

Proposition 3.12. *Let Z_n be a Markov chain on the locally compact space S equipped with its Borel σ -algebra $\mathcal{B}$. Call its transition probability kernel P . Assume*

1131

1132

1133

- (i) *There is a point x_0 in S such that $\sum_n P^n(x, N) > 0$ for all $x \in S$ and all neighborhoods N of x_0 and*
- (ii) *The Markov chain has a nontrivial continuous component at x_0 .*

1134

1135

1136

Then the chain is

1137

- (i) *Recurrent if and only if x_0 is conservative and*
- (ii) *Positive recurrent if and only if x_0 is positive.*

1138

1139

For a proof see [229], Theorem 4.1. As an application, we obtain Theorem 3.7

1140

Theorem 3.7. *Suppose that μ is spread out. Then the bilateral and mixed random walks on S are recurrent if and only if (3.37) holds. In particular, the four random walks are either all recurrent or all transient.*

1141

1142

1143

Proof. As in the proof of Proposition 3.11, the bilateral and mixed walks have non-trivial continuous components everywhere on S . (Note that the powers P_m^k dominate certain multiples of P_r^k and P_l^k and that $P_b^k = P_l^k P_r^k$; detailed calculations are omitted.)

1144

1145

1146

1147

Let us first assume that (3.37) holds. Then S is the whole group G (Lemma 3.1).
 $e \rightarrow x$ in the bilateral walk if and only if $x \in \overline{\bigcup_n S_{\mu^2}^n} = H$, a closed (and open)
 normal subgroup of G (see Proposition 3.9 and its proof). As far as the mixed walk
 is concerned, all elements of G communicate.

Equation (3.37) implies that $\sum_n \mu^{2n}(V)$ diverges for any compact neighborhood
 V of the identity. Assume the contrary. Then $\sum_n \mu^{2n+1}(V)$ has to diverge and so
 does $\sum_n \mu^{2n+2}(VK)$, where K is a suitable compact subset of S_μ , since the terms
 are $\geq \mu^{2n+1}(V)\mu(K)$. VK is compact so this means that

$$\sum_n \mu^{2n}(N(x)) = \infty$$

for all neighborhoods $N(x)$ of a point $x \in VK$. (3.41) is equivalent to (3.37).
 Replacing μ with μ^2 in (3.37) and (3.41), we conclude that the identity is con-
 servative for the bilateral walk. Recall that $P_b^n(e, B) = \mu^{2n}(B)$ and that x is
 conservative if

$$\sum_n P_b^n(x, N(x)) = \infty$$

for all neighborhoods $N(x)$ of x . e communicates with all elements of H and the
 bilateral walk has a nontrivial continuous component at e , so by Proposition 3.12,
 e is recurrent. The same argument shows that all elements of H are recurrent.

To prove that all elements of the essential class H^c are recurrent, it suffices to
 prove that an $x \in S_{\mu^d}$, d odd, is conservative. Without loss of generality, we can
 assume that μ^d has a continuous positive density f at x (see Lemma 3.2).

Let $N(x)$ be any neighborhood of x . (In the calculations that follow, a set with
 subscript m is an open set with positive μ^m -measure.)

Then there are open sets V_{k+d} and W_k with $V_{k+d}xW_k \subset N(x)$ and
 $\mu^{k+d}(V_{k+d}), \mu^k(W_k) > 0$. Actually, there is an open set $U(x^2) \ni x^2$ such that
 $V_{k+d}U(x^2)W_k \subset N(x)$. Consider pairs of open sets A_n, B_n such that $A_nx^2B_n$
 $\subset U(x^2)$.

We have

$$A_nx^2B_n \subset U(x^2) \implies V_{k+d}A_nx^2B_nW_k \subset N(x),$$

which permits us to compare $P_b^n(x^2, U(x^2))$ with $P_b^{n+k+1}(x, N(x))$ (see the proof
 of Proposition 3.6). We get the following relation between the densities:

$$\mu^{k+d}(V_{k+d})\mu^k(W_k)\frac{d\mu^{2d}}{dm}(x^2)P_b^n(x^2, U(x^2)) \leq fP_b^{n+k+d}(x, N(x))$$

Summing over n , we conclude that x has to be conservative since $x^2 \in H$ is. We
 have thus proved that (3.37) implies that all elements of $S = G$ are recurrent for the
 bilateral walk.

Suppose now that some x_0 is recurrent for the bilateral walk. Since the ran-
 dom walk has a nontrivial continuous component at x_0 , it follows that all elements

communicating with x_0 are recurrent. Since μ is spread out the essential class containing x_0 has nonempty interior. Let N be a relatively compact open subset of the essential class of x_0 . Then $\sum_n P_b^n(\cdot, N)$ diverges on the open set N . $\mu^k(N) > 0$ for some k so $\sum_n \mu^{2n+k}(N) \geq \sum_n \int_N \mu^k(y) P_b^n(y, N) = \infty$, i.e., (3.41) holds for some $x \in N$. Hence the equivalent statement (3.37) holds as well.

Proofs for the mixed walk, which are similar, are omitted here. $\square$

There is a vast literature on random walks on groups. We will briefly mention some highlights on the recurrence/transience question for these walks. In general, our probability measures will *not* be assumed spread out.

If a random walk generated by μ is recurrent, the question arises as to how much the generating measure μ may be altered. For concrete groups, we have sometimes precise and easily verifiable criteria (as given, e.g., in Remark 3.6) for recurrence.

For a measure μ on the locally compact group S we define $\check{\mu}$ to be the image of μ under the inversion $x \mapsto x^{-1}$. If μ coincides with $\check{\mu}$, it is said to be *symmetric*.

Lemma 3.3. *If the probability measure μ is recurrent, i.e., satisfies (3.37), so is the symmetric measure $\frac{1}{2}(\mu + \check{\mu})$. Let ν be a symmetric probability measure on S and ϕ any symmetric nonnegative function on S with compact support and strictly positive on some open set. Define the probability measures $\gamma \equiv \sum_{n=0}^{\infty} 2^{-n-1} \nu^n$ (with $\nu^0 \equiv \delta_e$) and $\gamma_1 \equiv c\phi\gamma$, i.e., the measure with density ϕ with respect to γ normed to be a probability measure. The support of γ is all of S and that of γ_1 is a compact set. If ν is recurrent, γ and γ_1 are recurrent as well.*

For a proof see [4] where it is an important ingredient in the derivation of the main result – Theorem 3.8

Theorem 3.8. *A connected Lie group is recurrent if and only if it is of polynomial growth of order at most 2.*

Theorem 3.8 is historically speaking one of the latest links in a development of recurrence criteria for groups. For abelian groups, the criteria came much earlier. For these groups, a Fourier transform theory is available and we have the following precise characterization, originating with [39] and [138]:

Theorem 3.9. *Let μ be a probability measure on S and let ϕ be its Fourier transform. μ is recurrent if and only if*

$$\limsup_{t \rightarrow 1} \int_N \Re(1 - t\phi)^{-1} = \infty$$

for all neighborhoods N of 0 in the dual group of S . (The integration is understood to be with respect to Haar measure on the dual group.)

Corollary 3.6. $\mathbb{R}^{d'} \oplus \mathbb{Z}^{d''}$ is recurrent if and only if $d' + d'' \leq 2$.

For proofs and further discussion, see, for example, [199], Chap. 3.

Definition 3.8. Let P be the transition probability operator of a right random walk Z_n on the group S , generated by the probability measure μ . Let f be a positive or bounded real-valued measurable function on S . Then f is *harmonic* (with respect to the random walk) if $Pf = f$. In other words, f is harmonic if $\mathbf{E}f(Z_1^x) = f(x)$ (see (3.3)). Similarly, f is *superharmonic* (*subharmonic*) if $Pf \leq f$ ($Pf \geq f$).

For positive Radon measures ρ on S , we define the measure ρP by the formula

$$\rho P(A) = \int_S \rho(dx) P(x, A) = \int_S \rho(dx) \mu(x^{-1}A)$$

for $A \in \mathcal{B}$. ρ is said to be *harmonic* with respect to the random walk if $\rho P = \rho$. (Using the terminology from Sect. 3.1, ρ can also be termed an *invariant* measure for P .) ρ is *superharmonic* (*subharmonic*) if $\rho P \leq \rho$ ($\rho P \geq \rho$).

A proof of the following recurrence criterion may be found in [87] and [207]:

Theorem 3.10. *The following conditions are equivalent*

- (i) *The right random walk generated by μ is recurrent.*
- (ii) *Every positive continuous superharmonic function is constant.*
- (iii) *Every superharmonic measure is a right Haar measure.*
- (iv) *Every positive superharmonic function is constant μ -almost everywhere.*

Another useful result in Proposition 3.13 connects recurrence with the left-right invariance properties of the group.

Proposition 3.13. *A recurrent group is unimodular, i.e., the right and left Haar measures coincide.*

Proposition 3.13 is also proved in [87] and [207].

An immediate consequence is that the affine group (see Example 3.3(ii)) is transient.

Before going over to positive recurrence, we refer to Proposition 3.14, which is due to Brunel and Revuz [22].

Proposition 3.14. *If all probability measures on S are recurrent, then S is compact.*

A convolution sequence μ^n of probability measures on S is said to converge vaguely to 0 if $\mu^n(K) \rightarrow 0$ for all compact K . Theorem 2.9 of Chap. 2 tells us that μ^n does not converge vaguely to 0 if and only if the group S is compact. Theorem 3.10 implies that the invariant measure is a Haar measure and thus a probability measure only in the case of a compact S . Hence, we can summarize the positive recurrent case in a

Theorem 3.11. *The right or left random walks are positive recurrent if and only if the group S is compact. This is the case if and only if $\limsup \mu^n(K) > 0$ for some compact K .*

Corollary 3.7. *If μ is spread out, Theorem 3.11 holds for the bilateral and mixed random walks as well.*

Proof. Consider the bilateral walk on S . If S is compact, then the normal subgroup H is compact, too. The Haar measure $\chi = \chi_H$ on H is an invariant probability measure for the bilateral walk (which is obviously recurrent). Hence, the walk is positive recurrent on H . Similarly, it is positive recurrent on the compact set $H^c = gH = Hg$, where a translate $\delta_g * \chi_H = \chi_H * \delta_g$ of the Haar measure on H serves as invariant probability measure:

$$\delta_g * \chi_{P_b} = \mu * \delta_g * \chi * \mu = \delta_g * \chi.$$

If the bilateral walk is positive recurrent on one essential class, then an argument similar to the one given in the proof of Theorem 3.7 shows that it is positive recurrent on all of S . If the point e is positive recurrent, then it is positive (see Proposition 3.12). It follows that

$$\limsup \mu^{2n}(K) > 0$$

for some compact set K . By Theorem 3.11, S is compact.

The proof for the mixed random walk is omitted. $\square$

By Corollary 2.3 and (2.54) of Chap. 2, the Cesàro average

$$\frac{1}{n} \sum_{k=1}^n \mu^k$$

converges weakly to the Haar measure χ_S on S . In terms of convergence of empirical distributions, we can formulate Theorem 3.12. Its proof is deferred to Sect. 3.4 (see Corollary 3.11).

Theorem 3.12. Let μ be a probability measure on the compact group S . Denote by Z_n (T_n) the right (left) random walk generated by μ and by W_n (M_n) the bilateral (mixed) random walk generated by μ . Let H be the compact normal subgroup of S generated by μ^2 . χ_S and χ_H are the Haar measures on S and H , respectively. Then the random walks are equidistributed on S in the following sense

$$\frac{1}{n} \sum_{k=1}^n f(Z_k) \rightarrow \int f d\chi_S \text{ almost surely}$$

for any continuous real-valued function on S . The same limit holds for T_n and M_n . It also holds for W_n in case H and S coincide. If not,

$$\frac{1}{n} \sum_{k=1}^n f(W_k) \rightarrow \begin{cases} \int f d\chi_H \text{ almost surely} & \text{if } W_0 \in H; \\ \int f d(\delta_g * \chi_H) \text{ almost surely,} & \text{otherwise} \end{cases}$$

where $g \in S \setminus H$.

A Markov random walk was defined for discrete semigroups in Sect. 3.2 in a very intuitive way. The idea is to study the behavior of the product of random elements which are no longer independent but are drawn from a Markov chain. For topological groups and semigroups, we need to define these objects more carefully.

Definition 3.9. Let S be a locally compact second countable semigroup and let C be a locally compact second countable Hausdorff space. Let X_n be a Markov chain on C , equipped with its Borel σ -algebra $\mathcal{C}$, with transition kernel P , which we will assume to be Feller. $\mathbf{P}_x$ will denote the probability on the canonical probability space (path space) C^∞ induced by P and the initial condition $X_0 = x$. Let f be a continuous function from C to S . The process

$$S_n \equiv \prod_{k=1}^n f(X_k), \quad n = 1, 2, 3, \dots$$

is called a *Markov random walk* on S .

A noteworthy special case is

$$S_n = \prod_{k=1}^n X_k,$$

that is, $C \subset S$ and f is the identity map. When the factors X_k are independent and identically distributed, S_n is an ordinary random walk on S .

Example 3.4. (i) If X_k is a Markov chain on the real line, then

$$S_n \equiv X_1 + X_2 + X_3 + \dots + X_n, \quad n = 1, 2, 3, \dots$$

is a Markov random walk on $\mathbb{R}$.

(ii) Let C be

$$\{-2, 0\} \cup \left\{\frac{1}{n} \mid n = 1, 2, 3, \dots\right\},$$

endowed with the topology induced by that of the real line. Define the transition probabilities

$$P(0, 1) = P(-2, 1) = 1, 0 < P(1, -2) < 1,$$

$$P(1, \frac{1}{2}) = 1 - P(1, -2), P(n^{-1}, (n+1)^{-1}) = \alpha_n,$$

$$P(n^{-1}, 1) = 1 - \alpha_n,$$

where $0 < \alpha_n < 1$ and $\lim \alpha_n = 0$.

(iii) Let μ be a probability distribution on the set of idempotent matrices of the form

$$(1 + \alpha\beta)^{-1} \begin{pmatrix} 1 & \alpha \\ \beta & \alpha\beta \end{pmatrix},$$

where α and β are real numbers, $\alpha\beta \neq -1$. The behavior of a product 1299

$$Y_1 Y_2 Y_3 \cdots Y_n$$

of independent μ -distributed random matrices can be studied conveniently by 1300
considering the (1,1) entry of the product 1301

$$S_n = (1 + \alpha_1 \beta_1)^{-1} (1 + \alpha_1 \beta_2) \cdots (1 + \alpha_{n-1} \beta_n) (1 + \alpha_n \beta_n)^{-1},$$

where the subscript i corresponds to the random matrix Y_i . Under mild condi- 1302
tions on μ , S_n is a Markov random walk on the multiplicative group of nonzero 1303
reals. 1304

- (iv) Let C be a finite set of affine contractions from $\mathfrak{R}^d$ to itself, a $w \in C$ is of the 1305
form $w(x) = Ax + b$ where A is a matrix with Euclidean norm less than 1 and 1306
 b is a vector in $\mathfrak{R}^d$. Let $W_1, W_2, \dots$ be a Markov chain on C . The process 1307

$$S_n = W_n \circ W_{n-1} \circ \cdots \circ W_2 \circ W_1, \quad n = 1, 2, \dots$$

is a Markov random walk on the semigroup of affine contractions from $\mathfrak{R}^d$ 1308
into itself. Its action on $\mathfrak{R}^d$ $S_n x$ is very much studied in the context of iterated 1309
function systems. We return to this application in Chap. 4. 1310

In our analysis of the Markov random walks, we make use of the fact that an 1311
auxiliary process, the *Koutsky process* $Z_n = (X_n, S_n)$, $n = 1, 2, \dots$, is a Markov 1312
chain on $C \times S$ with transition kernel U defined by 1313

$$U((c, s), A \times B) = \int 1_A(z) 1_B(sf(z)) P(c, dz), \quad c \in C, \quad s \in S, \quad A \in \mathcal{C}, \quad B \in \mathcal{S},$$

where $\mathcal{C}$ and $\mathcal{S}$ are the Borel σ -algebras of C and S , respectively. U can be shown 1314
to be a Feller kernel. 1315

Definition 3.10. We call a $(n+1)$ -tuple $(c_0, c_1, c_2, \dots, c_n)$ a *chain* of length n from 1316
 x to y if $x = c_0$, $y = c_n$ and c_{i+1} is in the support of the transition probability mea- 1317
sure $P(c_i, \cdot)$, $i = 0, 1, \dots, n-1$. Denote the chain by k_x and define the functions 1318
 p and q by 1319

$$p(k_x) = f(c_1) f(c_2) \cdots f(c_n) \quad \text{and} \quad q(k_x) = c_n.$$

Suppose a sequence of chains $(k_x^n)_{n=1}^\infty$ has the property that the endpoint sequence 1320
 $q(k_x^n)$ converges to some $u \in C$. Let w be a limit point of the sequence $p(k_x^n)$. 1321
The latter is of the form $v_n f(q(k_x^n))$ for some $v_n \in S$. (If S is a group, this im- 1322
plies that w has the form $w = v f(u)$ for some $v \in S$.) The set of all such w s is 1323
denoted by 1324

$$\Pi(x, u) \equiv \{w \in S | q(k_x^n) \rightarrow u, \quad w = \lim p(k_x^{n_i}) \text{ for some subsequence } n_i\}.$$

If there are chains k_x^n such that $\lim q(k_x^n) = u$, then we say that x leads to u , $x \rightarrow u$. If $x \rightarrow u$ and $u \rightarrow x$, then x and u communicate. If $x \rightarrow u$ implies $u \rightarrow x$, then x is said to be *essential*. Note that $x \rightarrow u$ is equivalent to

$$u \in \overline{\bigcup_{n=1}^{\infty} \text{supp}(P^n(x, \cdot))},$$

where P^n is the n -step transition kernel for our Markov chain X on C . We remark that these communication relations are defined for the Koutsky process (X_n, S_n) , see Definition 3.4, since it is a Markov chain with a Feller transition probability operator on a locally compact state space.

The following assumption about the Markov chain X on C guarantees that all elements of C communicate:

$$C = \overline{\bigcup_{n=1}^{\infty} \text{supp}(P^n(x, \cdot))} \text{ for every } x \in C. \quad (3.44)$$

The statement

$$C = \bigcup_{n=1}^{\infty} \text{supp}(P^n(x, \cdot)) \quad (3.45)$$

means that there is a *finite* chain from x to every u .

In our Example 3.4(i), $\Pi(0, 0)$ is the set of all sums $x_1 + x_2 + \dots + x_{n-1} + 0$ such that the transition probabilities $P(0, x_1), P(x_1, x_2), \dots, P(x_{n-1}, 0)$ are all strictly positive. If the random walk only visits nonnegative integer points starting from 0, then $\Pi(0, 0)$ is a semigroup of nonnegative integers. Otherwise, it is all $\mathbb{Z}$ or possibly a lattice of the form $d\mathbb{Z}$ where d is a positive integer.

In (ii) $\Pi(1, 1)$ includes the sums

$$-2 + 1, \frac{1}{2} + 1, \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{n} + 1$$

and all sums of such terms. Thus, $\Pi(1, 1)$, which is closed, is all of $\mathfrak{R}$. $\Pi(0, 0)$ is empty however, since a chain ending in an element $q(k_0^n) = 1/m_n$ close to 0 contributes at least the following sum to $p(k_0^n)$:

$$1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{m_n}.$$

This Markov chain satisfies (3.44). On the other hand, no element of C satisfies (3.45).

In the remainder of Sect. 3.3, we assume that S is a locally compact group. $\Pi(x, x)$ is either empty or a closed subsemigroup of S , the *invariance semigroup*

of $x \in S$. This follows from the fact that $(x, e) \rightarrow (x, g)$ (in the Koutsky process) 1349
if and only if $g \in \Pi(x, x)$. The reader is referred to [242] and [106] for the details. 1350

Suppose there is an essential element (c, g) in $C \times S$ with respect to the Koutsky 1351
process. Then one sees that the identity e belongs to $\Pi(c, c)$ (because $(c, g) \rightarrow$ 1352
 (c, g)). It is also the case that essentiality means $(c, g) \rightarrow (c, gh)$ implies $(c, gh) \rightarrow$ 1353
 (c, g) . That statement can in turn be translated into the implication 1354

$$h \in \Pi(c, c) \Rightarrow h^{-1} \in \Pi(c, c).$$

$\Pi(c, c)$ is a closed set which is algebraically a subgroup of S . By Ellis's Theorem 1.4. 1355
 $\Pi(c, c)$ is a closed topological subgroup of S . 1356

Clearly, $\Pi(c, u) \neq \emptyset$ implies that (u, s) is essential for some s . Hence, $\Pi(u, u)$ is 1357
also a subgroup. It can be proved to be conjugate to $\Pi(c, c)$: Consider the transitions 1358

$$(u, e) \rightarrow (c, l) \rightarrow (u, e)$$

or equivalently, 1359

$$(u, e) \rightarrow (c, l) \text{ and } (c, e) \rightarrow (u, l^{-1}).$$

If $s \in \Pi(c, c)$, 1360

$$lsl^{-1} \in \Pi(u, u).$$

Conversely, every element of $\Pi(u, u)$ can be represented in this way. Hence, 1361

$$l\Pi(c, c)l^{-1} = \Pi(u, u).$$

The preceding discussion is summarized in Proposition 3.15 1362

Proposition 3.15. Assume that (3.44) holds. If there is an essential element (c, s) 1363
for the Koutsky process, the set of essential elements contains the set 1364

$$\bigcup_{\Pi(c, u) \neq \emptyset} \{u\} \times S.$$

The invariance semigroup $\Pi(c, c)$ is a closed subgroup of S . Furthermore, for u 1365
such that $\Pi(c, u) \neq \emptyset$, $\Pi(u, u)$ is a group conjugate to $\Pi(c, c)$. If in addition 1366
(3.45) holds for all $c \in C$, then either all or none of the invariance semigroups 1367
 $\Pi(c, c)$ are groups. Then all or none, respectively, of the elements of $C \times S$ are 1368
essential for the Koutsky process. 1369

Proof. The above discussion contains the main points of the proof. Remaining 1370
details may be found in [106, 242]. $\square$ 1371

We now turn to the question of recurrence/transience for the Markov random 1372
walk. To facilitate our analysis, we assume that the Markov chain X_n on C is *sta-* 1373
tionary; i.e., X_n has an invariant probability distribution π as its initial distribution. 1374

We also assume a very strong form of Harris recurrence: There is a probability measure ν on $(C, \mathcal{C})$ such that $\mathbf{P}_x\{X_n \in A \text{ infinitely often}\} = 1$ for all $x \in C$ and all $A \in \mathcal{C}$ with $\nu(A) > 0$. In addition

$$P(x, A) \geq h(x)\nu(A), \quad x \in C, \quad A \in \mathcal{C}, \quad (3.46)$$

where $h > 0$ is a function on C with strictly positive integral $\int h d\nu$ with respect to ν . For example, a positive recurrent aperiodic Markov chain on a discrete space C satisfies (3.46). Let c be a positive recurrent point and set $h = 1_{\{c\}}(x)$ and $\nu = P(c, \cdot)$. A Markov chain on $\mathfrak{X}$ with a continuous strictly positive transition density with respect to Lebesgue measure,

$$P(x, dy) = p(x, y)dy,$$

satisfies (3.46) with $h = 1_N$ and $\nu = \alpha P(c, M \cap \cdot)$, where N and M are suitable compact neighborhoods of c and $\alpha < 1$ a constant. Harris recurrent chains and domination conditions such as (3.46) are discussed extensively in [181].

Using the splitting technique of [181] (also see [180] and [8]), we can decompose the chain $X = (X_1, X_2, \dots)$ by almost surely finite random times $\tau(1), \tau(2), \dots$ in such a way that the blocks

$$X_{\tau(i-1)+1}, \dots, X_{\tau(i)} \quad \text{and} \quad X_{\tau(i)+1}, \dots, X_{\tau(i+1)}$$

are independent and identically distributed (i. i. d.).

Note that for a Harris recurrent chain there is always a $k \geq 1$ such that P^k satisfies (3.46). The choice of $k = 1$ here is necessary to ensure that the above procedure results in a decomposition of $X_1, X_2, \dots$, into independent blocks.

When C is discrete we work essentially as in the proof of Proposition 3.8. X_n is assumed to be stationary and aperiodic on C . We saw above that the domination condition (3.46) is satisfied for $c \in C$ positive recurrent. The times $\tau(1), \tau(2), \dots$ are the successive returns of X_n to c . They are almost surely finite: $\mathbf{E}_c \tau(1) = (\pi(c))^{-1}$.

$$f(X_{\tau(1)+1})f(X_{\tau(1)+2}) \cdots f(X_{\tau(n)}) = S_{\tau(1)}^{-1} S_{\tau(n)}, \quad n = 1, 2, \dots$$

are products of i. i. d. random elements of $\Pi(c, c) \subset S$.

In other words, $S_{\tau(1)}^{-1} S_{\tau(n)}$, $n = 1, 2, \dots$ is a random walk on $\Pi(c, c)$, generating all of $\Pi(c, c)$.

Remark 3.7. From the point of view of applications, it is worthwhile to note the simple but important fact that $\Pi(c, c)$ is not necessarily a discrete object even if C is discrete. Then the random walk $S_{\tau(1)}^{-1} S_{\tau(n)}$, $n = 1, 2, \dots$ is transient if $\Pi(c, c)$ is a transient group. If the invariance semigroup $\Pi(c, c)$ is not a group at all, then the random walk is transient, too. This follows from the fact that any recurrent random walk on a group automatically generates a subgroup (see Lemma 3.1). The process $S_{\tau(n)}$, $n = 1, 2, \dots$ will also be transient since it differs from $S_{\tau(1)}^{-1} S_{\tau(n)}$,

$n = 1, 2, \dots$ only by its initial distribution. Extending the Main lemma of [8] to the group case, we conclude that the Markov random walk S_n itself is transient.

In the general case, the times $\tau(i)$, $i = 1, 2, \dots$, form a subset of the returns to the set where the function h of (3.46) is positive. The blocks $S_{\tau(n-1)}^{-1}S_{\tau(n)}$, $n = 1, 2, \dots$ again generate a random walk on S , which transient if and only if S_n is transient. Our goal is to formulate some sufficient transience criteria on the distribution of the blocks $S_{\tau(n-1)}^{-1}S_{\tau(n)}$.

If the distribution of blocks generates a transient group, or a semigroup that is not a group at all, then S_n is transient. In particular, if the group generated is an open subgroup of S and S is transient, then S_n is transient.

We have thus proved Theorem 3.13

Theorem 3.13. *Let the Markov chain X_n be positive Harris satisfying condition (3.46). Then*

- (i) *If C is discrete and the invariance semigroups $\Pi(c, c)$, $c \in C$, are either no groups at all or transient groups or*
 - (ii) *If the blocks $S_{\tau(n-1)}^{-1}S_{\tau(n)}$ generate a transient group or semigroup,*
- then the Markov random walk S_n is transient.*

Random walks on $\mathfrak{R}^d$, $d \geq 3$, are transient. Hence, we have Corollary 3.8.

Corollary 3.8. *If the invariance semigroups, or the groups generated by the blocks, are isomorphic to $\mathfrak{R}^d$, $d \geq 3$, then S_n is transient.*

$\mathfrak{R}^d$, $d \leq 2$, admit recurrent random walks. As shown in [199] a random walk S_n on $\mathfrak{R}^2$ is recurrent if and only if $\mathbf{E}S_1 = 0$ and $\mathbf{E}S_1^2 < \infty$. Dekking [56] proves that zero expectation of S_1 with respect to the stationary probability measure π is necessary and sufficient for the recurrence of a Markov random walk on $\mathfrak{R}$:

$$\mathbf{E}_\pi S_1 \equiv \int \mathbf{E}_x f(X_1) d\pi(x) = 0.$$

Using this information [106] proves the result in Proposition 3.16.

Proposition 3.16. *Let (X_n, S_n) be a Markov random walk on $C \times S$. Let X_n be a stationary positive Harris chain on C satisfying (3.46). Call its invariant distribution π . If*

- (i) *C is discrete, the invariance semigroups $\Pi(c, c)$ are isomorphic to $\mathfrak{R}$ or $\mathfrak{R}^2$ and $\mathbf{E}_\pi S_1 = 0$ and, in addition, in the $\mathfrak{R}^2$ -case,*

$$\mathbf{E}_c \left| \sum_{i=1}^{\tau} f(X_i) \right|^2 < \infty, \text{ for some } c \in C,$$

where τ is the first return time to c ; or

- (ii) *$S = \mathfrak{R}$ and $\mathbf{E}_\pi S_1 = 0$; or*

(iii) $S = \mathfrak{R}^2$, $\mathbf{E}_\pi S_1 = 0$ and

1439

$$\mathbf{E}_v \left| \sum_{i=1}^{\tau(1)} f(X_i) \right|^2 < \infty,$$

where $\tau(1), \tau(2), \dots$, are the times discussed above decomposing the chain 1440

$X_1, X_2, \dots$ into i. i. d. blocks, 1441

then the Markov random walk S_n is recurrent. 1442

Remark 3.8. The conditions given are also necessary provided the blocks do not 1443

generate a proper closed subgroup of the group in question. If S , or $\Pi(c, c)$ in the 1444

discrete case, is compact, then S_n is of course always recurrent. 1445

Exercise 3.20. Let S_μ be a subset of a compact group G . Prove: If S satisfies the 1446

irreducibility condition (3.9), then S is a closed subgroup of G . 1447

Exercise 3.21. Prove that the multiplicative group of nonzero complex numbers 1448

$\mathbf{C} \setminus \{0\}$ is recurrent. 1449

Exercise 3.22. Use Theorem 3.9 to verify that the simple symmetric random walk 1450

on $\mathfrak{R}$ is recurrent. 1451

Exercise 3.23. Prove Corollary 3.7 for the case of the mixed random walk. 1452

Exercise 3.24. Suppose that the matrices in Example 3.4 (iii) are projection matri- 1453

ces. Find criteria for the recurrence/transience of S_n . 1454

Exercise 3.25. Let $G_1 = \{(\varepsilon, c) | \varepsilon = \pm 1, c \in \mathfrak{R}\}$ with the multiplication defined by 1455

$$(\varepsilon, c)(\varepsilon', c') = (\varepsilon\varepsilon', c + \varepsilon c').$$

Prove that G_1 is recurrent. (Hint: Use the visits to the subgroup of translations 1456

$\{(1, c) | c \in \mathfrak{R}\}$ as in the proof of Theorem 3.4.) 1457

Exercise 3.26. Let $\mathbf{C}$ denote the complex numbers. Let $G_2 = \{(t, c) | |t| = 1, t, c \in$ 1458

$\mathbf{C}\}$ with the multiplication defined by 1459

$$(t, c)(t', c') = (tt', c + tc').$$

Prove that G_2 is recurrent. (Hint: Use Proposition 3.16.) 1460

3.4 Compact Semigroups

1461

Throughout Sect. 3.4, S represents a compact topological semigroup. In most in- 1462

stances it is generated, in the sense of (3.9), by a probability measure μ . Our 1463

program is to study the communication relations, recurrence/transience properties, 1464

and ergodicity for the right, left, bilateral, and mixed random walks introduced in Sect. 3.1. Compactness is a strong assumption, but we are able to give a complete account of many aspects of the asymptotic behavior of these walks without introducing further assumptions on μ or the transition probability operators P . The transition probability operators were defined in Sect. 3.1, (3.2)–(3.6).

Communication relations are easily handled for random walks on compact semigroups. General statements (see Proposition 3.2) are simplified. Since the multiplication operation is (jointly) continuous and S is compact, the sets xS , Sx , SxS , SS , etc. are compact subsets of S . Hence, Proposition 3.2 takes the form in Proposition 3.17

Proposition 3.17. *Let μ be a probability measure the compact semigroup S . Under the irreducibility condition (3.9), the following holds for the random walks on S generated by μ :*

- (i) *For the right walk on S , $x \rightarrow y$ if and only if $y \in xS$.*
- (ii) *For the left walk on S , $x \rightarrow y$ if and only if $y \in Sx$.*
- (iii) *For the bilateral walk, $x \rightarrow y$ if and only if*

$$y \in \overline{\bigcup_{n=1}^{\infty} S_{\mu}^n x S_{\mu}^n}.$$

- (iv) *For the mixed walk, $x \rightarrow y$ if and only if $y \in Sx \cup xS \cup SxS$.*

Many arguments from the discrete semigroup case can be employed in the study of compact semigroups. Lemma 3.4 is the analog of Proposition 3.3.

Lemma 3.4. *Assume S and μ as in Proposition 3.17. Then*

- (i) *x is essential for the right random walk if and only if x belongs to a minimal right ideal of S .*
- (ii) *x is essential for the left random walk if and only if x belongs to a minimal left ideal of S .*
- (iii) *x is essential for the bilateral walk if and only if x belongs to the minimal two-sided ideal of S .*
- (iv) *x is essential for the mixed random walk if and only if x belongs to the minimal two-sided ideal of S . Furthermore, all essential elements communicate.*

Proof. Assertions (i) and (ii) are proved exactly as the corresponding parts of Proposition 3.3.

The proof of Assertion (iii) follows that of Proposition 3.4(iic). If x is essential and $x \rightarrow y$, then clearly x and y generate the same bilateral ideal of S .

The essential elements are shown to form an ideal. We use the convention that a subscripted element v_k belongs to the compact set S_{μ}^k . (Recall that if μ and ν are two probability measure on the compact semigroup S then their supports satisfy the relation $S_{\mu * \nu} = S_{\mu} S_{\nu}$; see (2.8) of Chap. 2.) Let x be essential. To see that xu_k is essential, too, assume $xu_k \rightarrow y$. Then $y = \lim s_n x u_k t_n$ for some appropriate s_n, t_n . $u_k y$ is a limit point of a sequence in $S_{\mu}^{n+k} x S_{\mu}^{n+k}$ implying $x \rightarrow u_k y$. By the

essentiality of x , $x = \lim a_m u_k y b_m$ for some appropriate a_m, b_m . But then xu_k is a limit point of the sequence $(a_m u_k) y (b_m u_k)$. In other words, $y \rightarrow xu_k$, so the latter element is essential. The joint continuity of the multiplication and the compactness of S make it possible to extend this argument to any $xu = \lim xu_k$. We omit the details.

Similarly, ux and uxu' are essential for any $u, u' \in S$. Hence, the essential elements form a two-sided ideal of S . Exactly as in the proof of Proposition 3.3 we see that this ideal is a minimal one.

It remains to prove that the set of essential elements is nonempty. Consider the family $\mathcal{F}$ of topologically closed (thus compact) nonempty sets which are also closed for bilateral walks: $F \in \mathcal{F}$ if and only if $F \neq \emptyset$ is compact and

$$\overline{\bigcup_{n=1}^{\infty} S_{\mu}^n F S_{\mu}^n} \subset F.$$

$\mathcal{F}$ is a nonempty family of sets (S is a member), partially ordered by ordinary set inclusion $\subset$. By compactness, every linearly ordered subfamily $\mathcal{G}$ has a smallest element, the intersection of all the sets in $\mathcal{G}$. Zorn's lemma then guarantees the existence of a minimal element B in $\mathcal{F}$. The points of B are essential for the bilateral walk, because

$$B = \overline{\bigcup_{n=1}^{\infty} S_{\mu}^n b S_{\mu}^n} \quad \forall b \in B.$$

This ends the proof of Assertion (iii).

Assertion (iv) is proved as in Proposition 3.3

The proof of existence of essential elements for the bilateral walk is similar to the proof of Proposition 1.17 in Chap. 1, which actually is an existence proof of essential elements for the left (right) walk. Furthermore, we learn from Theorem 1.6 and its proof that the union of the minimal right ideals equals the union of the minimal left ideals which in turn constitutes the minimal two-sided ideal, the completely simple kernel of S . The structure theory of completely simple semigroups is developed in Sects. 1.4 and 1.5.

Theorem 3.14. *The set of essential elements for any of the four random walks on the compact semigroup S is the completely simple kernel K of S .*

Example 3.5. (i) The $d \times d$ stochastic matrices form a compact semigroup with respect to the ordinary matrix multiplication and the topology inherited from $\mathbb{R}^{d^2}$. Its kernel is the set of matrices with identical rows; see also the treatment of this semigroup in Sect. 1.5. For $d = 2$, this semigroup is made up of matrices of the form

$$\begin{pmatrix} a & 1-a \\ b & 1-b \end{pmatrix}, \quad 0 \leq a, b \leq 1.$$

These matrices may conveniently be put into a one-to-one correspondence with points (a, b) in the unit square. The product of two matrices corresponding to (a, b) and (a', b') , respectively, translates into

$$(aa' + b'(1 - a'), ba' + b'(1 - b)).$$

The kernel is the diagonal $a = b$.

- (ii) The $d \times d$ matrices of Euclidean norm at most 1 form a compact semigroup with kernel $\{0\}$.
 - (iii) Compact abelian semigroups have a kernel which is a compact abelian group.
- The 2×2 bistochastic matrices

$$\begin{pmatrix} a & 1-a \\ 1-a & a \end{pmatrix}, \quad 0 \leq a \leq 1$$

form a compact abelian semigroup. The kernel consists of the single element

$$\begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}.$$

We used the semigroup of probability measures with convolution $*$ as operation in Chap. 2. This semigroup is compact in the weak topology. The monothetic subsemigroup

$$\overline{\{\mu^n | n = 1, 2, \dots\}}$$

generated by a single element μ is a compact abelian subsemigroup.

- (iv) Consider the semigroup of affine contractions on $\mathfrak{N}^d$. Such a mapping may be written $x \mapsto Ax + b$, where A is a $d \times d$ matrix of Euclidean norm less than 1 and b is a vector in $\mathfrak{N}^d$. It can also be represented as a $(d + 1) \times (d + 1)$ matrix of the form

$$\begin{pmatrix} A & b \\ 0 & 1 \end{pmatrix},$$

where A is a $d \times d$ matrix, b is a $d \times 1$ column vector, 0 stands for a $1 \times d$ row vector (1 is the number 1).

With the topology of element-wise convergence (the topology of $\mathfrak{N}^{2(d+1)}$) and ordinary matrix multiplication, this structure is a compact topological semigroup. Its kernel is obtained when A is taken to be the 0 matrix, i.e., the elements of the kernel correspond to constant mappings $x \mapsto b$.

Remark 3.9. In most of our examples the kernel K is small compared to the semigroup S in the sense that it has empty interior and μ^k -measure 0 for all k .

Let us now study recurrence/transience properties of these random walks. We first dispose of elements outside the kernel.

Lemma 3.5. *If $x \notin K$, then x is transient.*

Proof. Recall that S and K are compact and that $SKS = K$. Using the proof method in Lemma 1.4, given any open $U \supset K$, find an open set $G \supset K$ such that $SG, GS, SGS \subset U$.

If $x \notin K$, then there exist open set $V \ni x$ and $U \supset K$ with $U \cap V = \emptyset$. Let S_n be any of the four random walks on S . For any $v \in V$ and any m ,

$$\mathbf{P}_v\{S_m \in V\} \leq \mu^m(G^c),$$

which decreases geometrically fast (see Proposition 2.6). Hence, by the Borel–Cantelli Lemma,

$$\mathbf{P}_x\{S_n \in V \text{ infinitely often}\} = 0$$

so x is transient.

Proposition 3.18. (i) For any point $x \in K$ and any neighborhood $N(x)$ of x , we have

$$\limsup \mu^n(N(x)) > 0.$$

(ii) In the sense of Definition 3.4, the points of K are positive for all four walks.

Proof. (i) is immediate from the much stronger statement in Theorem 2.7(i) that the Cesàro average

$$\frac{1}{n} \sum_{k=1}^n \mu^k \quad (3.47)$$

converges weakly to a probability measure having all K as its support.

(ii) Let x be a point in K and take e to be an idempotent in K such that $exe = x$. For example, let e be the identity of the group containing x (see Theorem 1.1 and Corollary 1.1). Given a neighborhood $N(x)$ of x , there is an open set U containing e such that $UxU \subset N(x)$.

$$P_b^n(x, N(x)) \geq (\mu^n(U))^2.$$

By part (i) of Proposition 3.18, the right hand side is $\geq \delta > 0$ for infinitely many n , proving that

$$\limsup P_b^n(x, N(x)) > 0.$$

Hence, x is positive for the bilateral walk.

The proofs for the unilateral walks are quite similar, and we leave the details to the reader.

To prove the assertion for the mixed walk, we note that for any neighborhood $N(e)$ of an idempotent $e \in K$ and some suitable $\delta > 0$, the set

$$\{n | \mu^n(N(e)) > \delta^2\}$$

contains all multiples of a positive integer d . To see this, choose another neighborhood $N'(e)$ with

$$N'(e)N'(e) \subset N(e).$$

Then

$$\mu^k(N'(e)), \mu^m(N'(e)) > \delta$$

implies $\mu^{k+m}(N(e)) > \delta^2$. From (i), we know that this happens for infinitely many k and m .

Let $P = P_m$ be the transition probability kernel for the mixed random walk. Take an $x \in K$ and a neighborhood $N(x)$ of x . Consider $exe = x$. There exist neighborhoods $N(e), N'(x)$ of e and x , respectively, such that

$$N(e)N'(x)N(e) \subset N(x).$$

The mixed random walk M beginning at x (or from some point in $N'(x)$) hits the set $N(x)$ at time nd with probability at least δ^4 on condition that the factors to the right and left are both multiples of d :

$$M_{nd} = X_{kd}X_{kd-1} \cdots X_2X_1xY_1Y_2 \cdots Y_{(n-k)d-1}Y_{(n-k)d}.$$

For large n , the probability of the number of factors on either side being a multiple of d is approximately $1/d$. Hence,

$$P^{nd}(x, N(x)) > \frac{\delta^4}{2d}, \quad n \geq n_0.$$

Thus, the averages satisfy

$$\frac{1}{nd} \sum_{k=1}^{nd} P^k(x, N(x)) > \frac{\delta^4}{2d^2}$$

for large enough n . (The same relation holds when x is replaced by some other element of $N'(x)$.) This proves that x is positive for the mixed random walk. $\square$

From Theorem 2.7, we know that the limit ν of (3.47) is an idempotent probability measure that is invariant with respect to μ

$$\nu = \nu * \nu = \nu * \mu = \mu * \nu. \quad (3.48)$$

Hence, it is invariant with respect to the transition probability operators of the unilateral, bilateral, and mixed random walks,

$$\nu = \nu P_r = \nu P_l = \nu P_b = \nu P_m. \quad (3.49)$$

We hasten to point out that the walks may admit many invariant measures; for example, for any $x \in K$, the probability measure $\delta_x * \nu$ is invariant with respect to the right random walk and its support is the minimal right ideal xS .

Corollary 3.9. *Let S be a compact semigroup with 0 . Then the random walks are positive recurrent. The nonzero points are transient, 0 is the only recurrent point. δ_0 is the only invariant probability measure for random walks on S .*

Remark 3.10. In a general locally compact semigroup positive and nonpositive conservative points can coexist, see Example 3.1. In a compact semigroup however, this is not possible, since all conservative points are necessarily positive by Lemma 3.5 and Proposition 3.18.

We begin our investigation of recurrence with Lemma 3.6

Lemma 3.6. (i) *Let $x \in S$ and let $N(y)$ be a neighborhood of y . Then there are neighborhoods $U(x)$, $V(y)$ of x and y , respectively, such that*

$$\bigcup_{z \in U(x)} z^{-1}V(y) \equiv U(x)^{-1}V(y) \subset x^{-1}N(y).$$

The corresponding statement is true for $N(y)x^{-1}$.

(ii) *Let $x \in K$ and let $N(y)$ be a neighborhood of y . Then there are neighborhoods $U(x)$, $V(y)$ of x and y , respectively, such that if $szt \in V(y)$ for some $z \in U(x)$, $s, t \in S_\mu$ then $sxt \in N(y)$.*

Proof. (i) If $x^{-1}N(y)$ is empty; i.e., $xS \cap N(y) = \emptyset$ so there are open sets $U(x)$, $V(y) \subset N(y)$ such that $U(x)S \in \overline{V(y)}^c$ (see Lemma 1.4). This implies that $z^{-1}V(y)$ is empty for all $z \in U(x)$. In other words, $U(x)^{-1}V(y) = x^{-1}N(y)$.

Suppose that $x^{-1}N(y)$ is nonempty and that (i) is false. Then there is a sequence z_n converging to x and $s_n \in S$ such that z_ns_n converges to y but $xs_n \notin N(y)$. If s is a limit point of s_n we have on the one hand, $xs = y$ and on the other hand, $xs \notin N(y)$, which is a contradiction.

(ii) If $y \notin S_\mu x S_\mu$, then we can find $U(x)$ and $V(y)$ such that

$$S_\mu U(x) S_\mu \subset V(y)^c$$

or $szt \notin V(y)$, for all $z \in U(x)$, $s, t \in S_\mu$.

If $S_\mu x S_\mu \cap N(y)$ is nonempty, we need to show that if z_n converges to x and $s_n z_n t_n$ converges to y then $s_n x t_n$ converges to y . If this is not the case, then a subsequence $s_m x t_m$ remains outside a neighborhood $W(y)$ of y . Consider the limit points $s, t \in S_\mu$ of s_m and t_m , respectively. These limit points satisfy $sxt \in W(y)^c$. If $w = sxt$ then, by the joint continuity of the multiplication, a subsequence of $s_n z_n t_n$ converges to $w \neq y$, contradicting the assumption that $s_n z_n t_n$ converges to y . $\square$

Note that S_μ in Part (ii) of the statement can be replaced by $S_{\mu^k} = S_\mu^k$ for any k . We can also replace it by S itself which is the support of the measure $\sum_k 2^{-k} \mu^k$.

Proposition 3.19. *The points of K are recurrent for all four walks.*

Proof. Let S_n be one of the random walks generated by μ . We show that

1643

$$\mathbf{P}_x\{S_n \in N(x) \text{ infinitely often}\} = 1$$

for all $x \in K$ and all neighborhoods $N(x)$ of x .

1644

Let $Z_n \equiv x X_1 X_2 \cdots X_n$ be the right random walk on S starting at a point $x \in K$.

1645

For an arbitrary neighborhood $N(x)$ of x , consider $\mathbf{P}\{Z_n \in N(x) \text{ finitely often}\}$.

1646

Let k be an arbitrary positive integer. Then for any m , $1 \leq m \leq k$

1647

$$\begin{aligned} & \mathbf{P}\{Z_n \in N(x) \text{ finitely often}\} \\ & \geq \sum_{i=0}^{\infty} \mathbf{P}\{Z_{m+ik} \in N(x), Z_n \notin N(x), \text{ for all } n \geq m+ik+k\} \end{aligned} \quad (3.50)$$

$$\begin{aligned} & \geq \sum_{i=0}^{\infty} \mathbf{P}\{Z_{m+ik} \in N(x), X_{m+ik+1} \cdots X_n \notin N(x)^{-1}N(x) \text{ for all } n \geq m+ik+k\} \\ & \quad (3.51) \end{aligned}$$

$$\begin{aligned} & \geq \sum_{i=0}^{\infty} \mathbf{P}\{X_{m+ik+1} \cdots X_n \notin N(x)^{-1}N(x) \text{ for all } n \geq m+(i+1)k\} \mathbf{P}\{Z_{m+ik} \in N(x)\} \\ & \quad (3.52) \end{aligned}$$

$$\geq \mathbf{P}\{X_1 \cdots X_n \notin N(x)^{-1}N(x) \text{ for all } n \geq k\} \sum_{i=0}^{\infty} \mathbf{P}\{Z_{m+ik} \in N(x)\} \quad (3.53)$$

We know that $\limsup \mu^n(U)$ is strictly positive for all open sets U containing a point of K . Hence, the sum in (3.53) is divergent for some m . Thus

1648

1649

$$\mathbf{P}\{X_1 \cdots X_n \notin N(x)^{-1}N(x) \text{ for all } n \geq k\} = 0. \quad (3.54)$$

Given any neighborhood $N'(x)$ of x we can find a neighborhood $N(x)$ of x such that

1650

$$N(x)^{-1}N(x) \subset x^{-1}N'(x).$$

This follows from Lemma 3.6 for the case $x = y$. Then

1651

$$\begin{aligned} & \mathbf{P}\{Z_n = x X_1 \cdots X_n \notin N'(x) \quad \forall n \geq k\} = 0 \\ & \iff \mathbf{P}_x\{Z_n \in N'(x) \text{ infinitely often}\} = 1. \end{aligned} \quad (3.55)$$

Since this is true for any neighborhood $N'(x)$ of x , x is recurrent for the right random walk Z_n .

1652

1653

The statement for the left random walk is proved similarly.

1654

Let $W_n \equiv X_n \cdots X_1 x Y_1 \cdots Y_n$ be the bilateral walk, starting at x , a point of K .

1655

Let $N(x)$ be an open set containing x .

1656

As above, we obtain inequalities

1657

$$\begin{aligned} & \mathbf{P}\{W_n \in N(x) \text{ finitely often}\} \geq \\ & \sum_i \mathbf{P}\{X_n \cdots X_{m+ik+1} N(x) Y_{m+ik+1} \cdots Y_n \notin N(x), \\ & \quad n \geq m + (i+1)k\} \cdot \mathbf{P}\{W_{m+ik} \in N(x)\} \\ & \geq \mathbf{P}\{X_n \cdots X_1 N(x) Y_1 \cdots Y_n \notin N(x), n \geq k\} \times \sum_i \mathbf{P}\{W_{m+ik} \in N(x)\} \quad (3.56) \end{aligned}$$

The sum in (3.56) equals

1658

$$\sum_i P_b^{m+ik}(x, N(x)).$$

It is divergent for some m because x is a positive point for the bilateral walk (Proposition 3.18(ii)). Using 3.6(ii), we conclude that

1659

1660

$$\mathbf{P}_x\{W_n \in N'(x) \text{ infinitely often}\} = 1 \quad (3.57)$$

for any neighborhood $N'(x)$ of $x \in K$. Thus, x is recurrent for the bilateral walk.

1661

The proof for the mixed walk is similar to those already given, so we omit the details. $\square$

1662

1663

As (3.48) and (3.49) show, there is a probability measure ν on K which is invariant for all our random walks. The preceding results are summarized in Theorem 3.15

1664

1665

1666

Theorem 3.15. *The four random walks are positive recurrent. x is a recurrent and positive point for any of the walks if and only if $x \in K$.*

1667

1668

Corollary 3.10. *Let S be a compact semigroup generated by the probability measure μ . Let K be the kernel of S . Then the following are equivalent:*

1669

1670

- (i) $\sum_{n=1}^{\infty} \mu^n(N(x)) = \infty$ for all open sets $N(x)$ containing the point x ; 1671
- (ii) $x \in K$; 1672
- (iii) $\sum_{n=1}^{\infty} \mu^n(N(x)x^{-1}) = \infty$ for all open sets $N(x)$ containing x ; 1673
- (iv) $\sum_{n=1}^{\infty} \mu^n(x^{-1}N(x)) = \infty$ for all open sets $N(x)$ containing x ; 1674
- (v) $\mathbf{P}\{S_n \in N(x) \text{ infinitely often}\} > 0$ for all open sets $N(x)$, where S_n is any of the four random walks. 1675

1676

Proof. If $x \notin K$, then the proof of Lemma 3.5 shows that the sums in (i), (iii), and (iv) converge for a neighborhood of x which is disjoint from the open covering U of K . Then (v) does not hold, either, by the Borel–Cantelli Lemma.

1677

1678

1679

If $x \in K$, then $N(x)$, $N(x)x^{-1}$, $x^{-1}N(x)$ are open sets, not disjoint from K . By Proposition 3.18(i), our statements (i), (iii), and (iv) hold.

1680

1681

Let Z_n be a right random walk started at $x \in K$. Then $Z_n \in C$ i. o., where C is chosen to be compact neighborhood $C \ni x$ with $C \subset N(x)$. Then $x(x^{-1}C) \subset N(x)$.

1682

1683

By Lemma 1.4, $O(x)U \subset N(x)$ where $O(x) \ni x$ and $U \supset x^{-1}C$ are open sets. Since $Z_n \in C$ infinitely often we have $X_1 \cdots X_n \in x^{-1}C$ infinitely often with probability 1. Consequently, $S_{k+n} = S_k X_1 \cdots X_n \in N(x)$ infinitely often with probability no less than $\mathbf{P}\{S_k \in O(x)\} \cdot 1$. The semigroup S being generated by μ , there is a k such that $\mathbf{P}\{S_k \in O(x)\} > 0$.

The proof for the left random walk is similar.

Reasoning as in the proof of Lemma 1.4 (or in the proof of the equicontinuity of the transition operators in Lemma 3.7) one can show that there exist neighborhoods $N'(x)$ and $O(x)$ of x such that $sxt \in N'(x)$ implies $sO(x)t \subset N(x)$. Combining these facts with the ideas in the proof for the right random walk Z_n , we can prove that $\mathbf{P}\{S_n \in N(x) \text{ infinitely often}\} > 0$ for the bilateral and mixed walks as well. $\square$

Remark 3.11. The proof also shows that under our compactness assumption (i), (iii), and (iv) can be written as

- (i)' $\limsup \mu^n(N(x)) > 0$ for all open $N(x)$ containing x ,
- (iii)' $\limsup \mu^n(N(x)x^{-1}) > 0$ for all open $N(x)$ containing x , and
- (iv)' $\limsup \mu^n(x^{-1}N(x)) > 0$ for all open $N(x)$ containing x .

It is clear from the discussion in Lemma 3.4 that the essential classes for the unilateral walks are the minimal right and left ideals, respectively. It was also shown that the kernel K constitutes one essential class for the mixed random walk. Comparison with the discrete case suggests that the bilateral walk admits at most two essential classes. Before we go into the detailed proofs, let us review some important facts. Recall from Sect. 1.5 that the completely simple kernel K of a compact semigroup S is algebraically and topologically isomorphic to a Rees product $X \times G \times Y$, where X (Y) is a compact left (right) zero subsemigroup of S and G is a compact topological subgroup of S . The multiplication in $X \times G \times Y$ is given by

$$(x, g, y)(x', g', y') = (x, gyx'g', y').$$

Recall from Theorem 2.7 and Corollary 2.3 that the limit of the Cesàro averages ν has the form

$$\alpha \times \chi_G \times \beta, \quad (3.58)$$

where χ_G is the normed Haar measure on the group G and α and β are probability measures on X and Y , respectively. Actually, (3.48) shows that $\alpha(A) = \nu(AS)$, $A \subset X$, A compact and $\beta(B) = \nu(SB)$, $B \subset Y$, B compact.

Let A be compact and consider an open set $U \supset AS$. Then $VS \subset U$ for some open $V \supset AS$. Since VS is a right ideal, $\mu^n(VS)$ is increasing and its limit is strictly positive and lies between the weak Cesàro limits $\nu(AS)$ and $\nu(\overline{U})$. Hence, we can write for compact $A \subset X$, $B \subset Y$

$$\alpha(A) = \inf_{V \supset AS} \lim_n \mu^n(VS) \quad \text{and} \quad \beta(B) = \inf_{V \supset SB} \lim_n \mu^n(SV). \quad (3.59)$$

So far we have mostly described our random walks in terms of products of random semigroup elements. The random walks are of course just as well described by

their transition probability operators. Since we want to make use of this approach, we must introduce some further concepts. These are more fully developed in [204], Chaps. 4 and 5.

Definition 3.11. Let $\mathcal{C}$ be a family of continuous functions from S to $\mathfrak{R}$. $\mathcal{C}$ is said to be an *equicontinuous* family if, for every $\varepsilon > 0$, there exists a finite covering $O_1, O_2, \dots, O_n$ of S such that

$$\forall f \in \mathcal{C} : \sup_{s, t \in O_j} |f(s) - f(t)| < \varepsilon, \quad j = 1, 2, \dots, n.$$

Let P be a bounded linear operator mapping the family $\mathcal{C}(S)$ of continuous functions on S into itself. (The norm of such a function f is understood to be $\sup_{s \in S} |f(s)|$.) P is said to be an *equicontinuous operator* on $\mathcal{C}(S)$ if the family $P^n f, n = 1, 2, 3, \dots$, is equicontinuous for all $f \in \mathcal{C}(S)$.

Lemma 3.7. Let P_r, P_l, P_b, P_m be the transition probability operators of the right, left, bilateral, and mixed random walks on the compact semigroup S generated by μ ; i.e., for $f \in \mathcal{C}(S)$ and $x \in S$,

$$\begin{aligned} P_r f(x) &= \int f(xs) d\mu(s), \\ P_l f(x) &= \int f(sx) d\mu(s), \\ P_b f(x) &= \int \int f(sxt) d\mu(s) d\mu(t), \\ P_m f(x) &= c \int f(xs) d\mu(s) + (1-c) \int f(sx) d\mu(s), \quad 0 < c < 1. \end{aligned}$$

- (i) P_r, P_l, P_b, P_m are equicontinuous.
- (ii) Let P be any of P_r, P_l, P_b, P_m . Then the Cesàro sum

$$\frac{1}{n} \sum_{k=1}^n P^k$$

converges to an operator $\overline{P}$ on $\mathcal{C}(S)$ in the following sense

$$\sup_{s \in S} \frac{1}{n} \sum_{k=1}^n P^k f(s) - \overline{P} f(s) \rightarrow 0$$

as $n \rightarrow \infty$ for all $f \in \mathcal{C}(S)$. Furthermore, $\overline{P}P = P\overline{P} = \overline{P}^2 = \overline{P}$.

Proof. (i) P_r and P_l are shown to map continuous functions into continuous functions in Sect. 2.2. As pointed out in connection with the definition of P_b , in Sect. 3.1 it

is the product of P_r and P_l . Hence, $P_b f = P_r P_l f$ is continuous if f is. Similarly, $P_m f$ is continuous, being a convex combination of $P_r f$ and $P_l f$.

Let f be a continuous function from S to $\Re$ and let $s, x, t \in S$. By the joint continuity of the multiplication in S , there are open sets $O_1 \ni s, O_2 \ni x, O_3 \ni t$ such that

$$\sup_{(s', x', t') \in O} |f(s' x' t') - f(s x t)| < \varepsilon,$$

where $O = O_1 \times O_2 \times O_3$. The sets $O = O_1 \times O_2 \times O_3$ cover the compact set $S \times S \times S$ and so there is a finite covering $O^i = O_1^i \times O_2^i \times O_3^i$, containing the points (s^i, x^i, t^i) , $i = 1, 2, \dots, n$, such that

$$|f(s x t) - f(s^i x^i t^i)| < \varepsilon \text{ if } (s, x, t) \in O_1^i \times O_2^i \times O_3^i.$$

Let

$$U = \bigcap_{x \in O_2^i} O_2^i,$$

i.e., the intersection of those members of the open covering O_2^i that contain a fixed element $x \in S$. Then we have for $x' \in U$

$$|f(s x t) - f(s x' t)| \leq |f(s x t) - f(s^i x^i t^i)| + |f(s^i x^i t^i) - f(s x' t)|$$

and hence $|f(s x t) - f(s x' t)| < 2\varepsilon$ for every $s, t \in S$.

The set $U = U_x$ is an open set containing x . The family of sets U_x forms an open covering of S . Hence, there is a finite subcovering $U_i, i = 1, 2, \dots, n$, such that

$$\sup_{x, x' \in U_i} |f(s x t) - f(s x' t)| < 2\varepsilon \quad \forall s, t \in S. \quad (3.60)$$

Consider the transition probability operator P_b for the bilateral random walk. We have

$$P_b^k f(x) = \int \int f(s x t) \mu^k(ds) \mu^k(dt)$$

and clearly for all k ,

$$|P_b^k f(x) - P_b^k f(x')| < 2\varepsilon \text{ if } x, x' \in U_i.$$

Hence, P_b is an equicontinuous operator.

Proofs for the other transition probability operators are similar, so we omit the details.

Assertion (ii), which is valid for all positive equicontinuous operators, is proved in [204], pp. 134–135. $\square$

Definition 3.12. Let S_n be one of the random walks. Call the transition probability operator P . Suppose that S_n is stationary; i.e., the initial distribution π is invariant with respect to P : $\pi = \pi P$ meaning

$$\int f d\pi = \int (P f) d\pi \quad \text{or} \quad \int_S f(s) \pi(ds) = \int_S \int_S \pi(ds) P(s, dt) f(t)$$

S_n and P are said to be *ergodic* if for every bounded $\mathcal{B}$ -measurable function f 1765

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} P^k f(x) = \int f d\pi \quad \pi - \text{almost surely.} \quad (3.61)$$

Equivalently (see the discussion in [204], Sect. 2 of Chap. 4), S_n is ergodic if for 1766
every pair of Borel sets A and B 1767

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \int_A \pi(ds) P^k(s, B) = \pi(A)\pi(B). \quad (3.62)$$

Remark 3.12. Rosenblatt [204] also shows that P is ergodic with respect to the 1768
stationary probability measure π if and only if the translation or *shift* operator 1769
 Θ (see Definition 3.2) is ergodic on the space of trajectories $\Omega = S^\infty$ with re- 1770
spect to the measure $\mathbf{P}_\pi$. Thus, for every bounded $\mathbf{P}_\pi$ -measurable function F on Ω , 1771
we have 1772

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} F(\Theta_k(\omega)) = \int F d\mathbf{P}_\pi \quad \mathbf{P}_\pi - \text{almost surely.} \quad (3.63)$$

Let F be a function depending only on one coordinate: $F(\omega) = f(\omega_1) =$ 1773
 $f(S_1(\omega))$, say. We then get 1774

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n f(S_k(\omega)) = \int f d\pi \quad \mathbf{P}_\pi - \text{almost surely} \quad (3.64)$$

for every bounded Borel function f on S . 1775

Theorem 3.16 characterizes the form of the stationary distributions for the four 1776
random walks. 1777

Theorem 3.16. (i) *Stationary probability distributions for the right random walk* 1778
are supported on the closed right ideals of K . A closed right ideal of K may be 1779
written $A \times G \times Y$, where A is a closed subset of X . The stationary probability 1780
measure then takes the form $\alpha' \times \chi_G \times \beta$, where α' is a probability measure on 1781
 X with support A . 1782
(ii) *Stationary probability distributions for the left random walk are supported on* 1783
the closed left ideals of K . A closed left ideal of K may be written $X \times G \times B$, 1784
where B is a closed subset of Y . The stationary probability measure then takes 1785
the form $\alpha \times \chi_G \times \beta'$, where β' is a probability measure on Y with support B . 1786
(iii) *For the bilateral walk, stationary probability distributions are convex combi-* 1787
nations of $\alpha \times \chi_H \times \beta$ supported on $X \times H \times Y$ and $\alpha \times \chi_{gH} \times \beta$ supported on 1788
 $X \times gH \times Y$. Here, $X \times H \times Y$ is the Rees–Suschkewitsch representation of 1789

the kernel of the compact semigroup generated by μ^2 , χ_H is the normed Haar
measure on the compact group H and $\chi_{gH} = \delta_g * \chi_H$ denotes its translate
by an element $g \in G \setminus H$.

- (iv) The only stationary probability distribution for the mixed random walk is $\nu =$
 $\alpha \times \chi_G \times \beta$.

Proof. (i) If $x \in K$, then $\delta_x * \nu$ is an invariant probability distribution for the right
random walk. Its support is the minimal right ideal xK .

Let π be an invariant distribution for the right random walk on S . $\pi P_r = \pi$ is
equivalent to $\pi * \mu = \pi$. The assertion then follows from Corollary 2.3.

(ii) Is proved in a similar way.

(iii) ν is an invariant probability measure for the bilateral walk since $\mu * \nu *$
 $\mu = \nu$.

By Theorem 2.7, all limit points of the sequence μ^n have their support inside K .
If π is an invariant probability measure, then $\mu^n * \pi * \mu^n = \pi$ for all n . Passing
to convergent subsequences, we then obtain $S_\pi \subset K$. Using the support of the
idempotent limit point η (see the proof of Theorem 2.7), we get

$$(X \times H_1 \times Y)S_\pi(X \times H_1 \times Y) = S_\pi.$$

Here, H_1 is a closed normal subgroup of G , and $H_1 = G$ if and only if μ^n con-
verges (to $\eta = \nu$). Since the middle factor of the product representation of ν is the
normalized Haar measure on G , the equation $\nu * \pi * \nu = \pi$ implies that $\pi = \nu$,
when $H_1 = G$. If $\eta \neq \nu$, by considering the limit point $\mu^k * \eta = \eta * \mu^k$, we see
that

$$(X \times g^k H_1 \times Y)S_\pi(X \times g^k H_1 \times Y) = S_\pi,$$

for some group element $g \notin H_1$. It also follows that π is a product measure $\alpha \times$
 $\gamma \times \beta$ where the middle factor γ has a support which is either the closed group H
generated by $g^2 H_1$ or a coset gH of H or their union.

H can also be characterized as the group factor of the kernel of the semigroup
generated by μ^2 . H is a normal subgroup of G . We have, then, two cases: the odd
case with $H = G$ and the even case with $H \neq G$; then $G = H \cup gH$.

Consider the limit points of the sequence of measures $\mu^n * \delta_x * \mu^n$ for an $x \in K$.
Such limit points are, for example, $\eta * \delta_x * \eta$ and $(\mu * \eta)^n * \eta * \delta_x * \eta * (\eta * \mu)^n$.
Since the middle factor of η is a Haar measure on H_1 analysis of the above expres-
sion reveals that $\mu^n * \eta * \delta_x * \eta * \mu^n = \eta * \delta_x * \eta * \mu^{2n}$. The Cesàro limit of
this expression is either $\alpha \times \chi_H \times \beta$ or $\alpha \times \chi_{gH} \times \beta$ depending on whether or not
the middle component is supported on H . Then, if π is invariant with respect to the
bilateral walk, it is necessarily a convex combination of $\alpha \times \chi_H \times \beta$ and $\alpha \times \chi_{gH} \times \beta$.

- (iv) π is invariant with respect to a mixed random walk iff $\pi = c\pi * \mu +$
 $(1 - c)\mu * \pi$ for some $c \in (0, 1)$. Clearly, $\nu = \alpha \times \chi_G \times \beta$ satisfies this criterion.

Iterating the preceding criterion, we obtain

$$\pi = \sum_{k=0}^n \binom{n}{k} c^k (1 - c)^{n-k} \mu^{n-k} * \pi * \mu^k, \quad (3.65)$$

where $\pi * \mu^0$ ($\mu^0 * \pi$) is interpreted as just π .

Let M_n be a mixed random walk beginning at $x \in S$. The transition probabilities are given by

$$P_m^n(x, \cdot) = \sum_{k=0}^n \binom{n}{k} c^k (1-c)^{n-k} \mu^{n-k} * \delta_x * \mu^k, \quad n = 1, 2, 3, \dots \quad (3.66)$$

Lemma 3.7(ii) tells us that the Cesàro averages of the terms (3.66), operating on a continuous f , converge to a $\bar{P}f(x)$. For given x , $f \mapsto \bar{P}f(x)$ is a continuous linear functional on $\mathcal{C}(S)$. Furthermore, the constant function 1 is mapped onto itself. In terms of measures, this means that the Cesàro averages of the measures (3.66) converge weakly to a probability measure π_x , say. We show that the limit is independent of x and invariant with respect to P_m and furthermore that any invariant measure is necessarily equal to ν .

The proof of Lemma 3.5 shows that π_x is supported on K . Take any compact subset C of K^c and approximate its indicator function from above by an $f \in \mathcal{C}(S)$ that is 0 on K . For any positive ε , $\mu^n\{s \in S \mid f(s) \geq \varepsilon\}$ decreases geometrically fast since the set is compact and disjoint from K . Thus, $P^n f(x) < 2\varepsilon$ for large n and so $\int f d\pi_x = 0$ and $\pi_x(C) = 0$. Consequently $\pi_x(K^c) = 0$.

Let f be an arbitrary continuous real-valued function on S and assume that $\bar{P}f$ is nonconstant on K , with maximum a and minimum $b < a$. Let $s_1(s_2)$ be such that $\bar{P}f(s_1) = a$ ($\bar{P}f(s_2) = b$). Since $\bar{P}^2 f = \bar{P}f$, we can actually choose s_1 and s_2 in K :

$$\bar{P}f(s) = \bar{P}^2 f(s) = \int \pi_s(dt)(\bar{P}f)(t) = \int_K \pi_s(dt)(\bar{P}f)(t),$$

where the last expression lies between the extreme values of $\bar{P}f$ on K .

By continuity, there is an open set $O \ni s_2$ such that

$$\bar{P}f(t) < b + (a - b)/2 \text{ for } t \in O.$$

We know that all elements of K are positive (Theorem 3.15). Indeed the proof of Proposition 3.18(ii) showed that, for $x \in K$, the Cesàro averages of $P^n(x, N(x))$ are bounded below by a positive constant $\delta^4/2d^2$. We also noted that this inequality holds true for $P^n(y, N(x))$ for y in a neighborhood $N'(x)$ of x . In our context, this means that the Cesàro averages of $P^n(\cdot, O)$ are bounded below by a $\gamma > 0$ on a neighborhood N' of s_2 . The following equation

$$P^{n+k}(s_1, O) = \int P^k(s_1, dy) P^n(y, O)$$

implies that the Cesàro averages of $P^{n+k}(s_1, O)$ are bounded below by $P^k(s_1, N')$ γ which is strictly positive for some k . (Recall that K is a communicating class for the mixed walk, see Lemma 3.4(iv).) Hence $\pi_{s_1}(O) > 0$.

Then

$$\bar{P} \bar{P} f(s_1) = \int \pi_{s_1}(dt) \bar{P} f(t) < a = \bar{P} f(s_1),$$

contradicting $\bar{P}^2 = \bar{P}$. Thus, there is a common π_0 such that

$$\bar{P} f(x) = \int f d\pi_x = \int f d\pi_0 \quad \forall x \in S.$$

π_0 is invariant:

$$\int f d\pi_0 = \pi_x P_m f = \bar{P} P_m f(x) = \bar{P} f(x) = \int f d\pi_x = \int f d\pi_0.$$

Consider the expression (3.66) operating on a continuous f and integrated with respect to an invariant probability π . Then

$$\begin{aligned} \int \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \pi(dx) P^k f(x) &= \int \int \pi(dx) \pi_x(dy) f(y) \\ &= \int \int \pi(dx) \pi_0(dy) f(y) \\ &= \int f d\pi_0. \end{aligned}$$

On the other hand, (3.65) implies that the above expression equals $\int f d\pi$.

Hence, the invariant probability measure is unique and equals ν . $\square$

Corollary 3.11. (i) *The right random walk is ergodic on a minimal right ideal of the semigroup generated by μ .*

(ii) *The left random walk is ergodic on a minimal left ideal of the semigroup generated by μ .*

(iii) *The bilateral walk is ergodic on the minimal ideal of the semigroup generated by μ^2 .*

(iv) *The mixed random walk is ergodic on the minimal ideal of the semigroup generated by μ .*

Proof. Note first that all the walks admit a *unique* invariant probability distribution on the ideal in question: For P_r , it is $\delta_s * \nu = \delta_x \times \chi_G \times \beta$ on the minimal right ideal generated by $s = (x, g, y) \in K$, which is $\{x\} \times G \times Y$. For P_l , it is $\nu * \delta_s = \alpha \times \chi_G \times \delta_y$ on the minimal left ideal generated by $s = (x, g, y) \in K$, which is $X \times G \times \{y\}$. For P_b , it is $\nu * \delta_s * \nu = \alpha \times \chi_H \times \beta$ for $s = (x, g, y)$ in $X \times H \times Y$. For P_m , it is $\nu = \alpha \times \chi_G \times \beta$ for $s \in K = X \times G \times Y$. Furthermore, the unique invariant probability measure is the weak Cesàro limit of $P^k(s, \cdot)$ in all four cases. This enables us to deduce immediately that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \int f(s) \pi(ds) P^k g(s) = \int f d\pi \int g d\pi, \quad (3.67)$$

where π denotes the invariant probability measure for one of the random walks on the corresponding minimal ideal and f, g are continuous functions on S .

To prove ergodicity we need to verify (3.67) with f and g arbitrary measurable functions instead of continuous functions. It suffices to do it for measurable indicator functions $f = 1_A$, $g = 1_B$ (see (3.62)). π is a regular measure so there are continuous positive f and g , bounded by 1, such that $\int |f - 1_A| d\pi$ and $\int |g - 1_B| d\pi$ are $< \varepsilon$. Choose first B open and $g \leq 1_B$. Consider

$$\begin{aligned} & \left| \int_A \pi(dx) P^k(x, B) - \int \int f(x) \pi(dx) P^k(x, dy) g(y) \right| \\ & \leq \left| \int_A \pi(dx) P^k(x, B) - \int f(x) \pi(dx) P^k(x, B) \right| \\ & \quad + \left| \int f(x) \pi(dx) P^k(x, B) - \int f(x) \pi(dx) P^k g(x) \right| \\ & = \left| \int (1_A(x) - f(x)) \pi(dx) P^k(x, B) \right| + \left| \int \pi(dx) f(x) (P^k(x, B) - P^k g(x)) \right| \\ & \leq \varepsilon + \int \pi(dx) (P^k(x, B) - P^k g(x)) = \varepsilon + \int \pi(dx) (1_B(x) - g(x)) \leq 2\varepsilon \end{aligned}$$

For B compact and $g \geq 1_B$, the calculation yields the same result. For an arbitrary Borel set B , we have

$$\liminf \frac{1}{n} \sum_{k=1}^n \int_A \pi(dx) P^k(x, B) \geq \pi(A) \pi(C)$$

for any compact $C \subset B$. In a corresponding way

$$\limsup \frac{1}{n} \sum_{k=1}^n \int_A \pi(dx) P^k(x, B) \leq \pi(A) \pi(O),$$

where $O \supset B$ is open. Since $\pi(B)$ can be arbitrarily well approximated from below and from above by $\pi(C)$, C compact, and $\pi(O)$, O open, respectively, the Cesàro sum on the left is in fact convergent to the limit $\pi(A) \pi(B)$.

Corollary 3.12. (i) Suppose that K is right simple. Let Z_n be a right random walk on S starting at x . Then the sequence of empirical measures

$$\frac{1}{n} \sum_{k=0}^{n-1} \delta_{Z_k}$$

almost surely converges weakly to the invariant measure $\chi_G \times \beta$.

- (ii) Suppose K is left simple. Let S_n be a left random walk on S starting at x . Then the sequence of empirical measures

$$\frac{1}{n} \sum_{k=0}^{n-1} \delta_{S_k}$$

almost surely converges weakly to the invariant measure $\alpha \times \chi_G$.

- (iii) Suppose μ^2 generates all of S . Let W_n be a bilateral random walk on S starting at x . Then the sequence of empirical measures

$$\frac{1}{n} \sum_{k=0}^{n-1} \delta_{W_k}$$

almost surely converges weakly to the invariant measure $\alpha \times \chi_G \times \beta$.

- (iv) Let M_n be the mixed random walk on S starting at x . Then the sequence of empirical measures

$$\frac{1}{n} \sum_{k=0}^{n-1} \delta_{M_k}$$

almost surely converges weakly to the invariant measure $\alpha \times \chi_G \times \beta$.

Proof. We prove the assertions using a coupling argument. A walk beginning at an arbitrary x after a finite time almost coalesces with an ergodic process. The argument will be given for the right random walk in Assertion (i), only. The ideas underlying the proofs for the other walks are quite similar.

Let $x \in S$ and let Z_n be the right random walk generated by μ with starting point x .

$$Z_n = xX_1X_2 \cdots X_n, \quad n = 1, 2, \dots,$$

where $X_1, X_2, \dots$ are mutually independent μ -distributed random variables in S . Assertion (i) of Corollary 3.12 means that for arbitrary continuous f

$$\frac{1}{n} \sum_{k=0}^{n-1} f(Z_k)$$

almost surely converges with limit $\int f d\nu$.

Let $U_i, i = 1, 2, \dots, N$, be any open covering of $K = G \times Y$ [since the right ideal structure was assumed to be trivial in Assertion (i)]. We may take the U_i to have a nonempty intersection with K , thus $\nu(U_i) > 0$ for all i . $\nu = \chi_G \times \beta$ is the unique stationary probability distribution for a right random walk on S . By Corollary 3.11, the right random walk with initial distribution ν is ergodic.

Let Y^i be a right random walk on K generated by μ with initial measure ν_i the restriction of ν to U_i , appropriately normed to be a probability measure.

Then Y_n^i is ergodic in the sense that the sum in (3.64) converges to $\int f d\nu$ for all bounded measurable f $\mathbf{P}_{\nu_i}$ -almost surely. If not, then the set of exceptional trajectories would have positive measure with respect to $\mathbf{P}_\nu$ as well, contradicting the ergodicity of the right random walk with initial measure ν .

Let f be a continuous function on S and take an $\varepsilon > 0$. Then there exists a covering $U_i, i = 1, 2, \dots, N$ of K such that (3.60) holds. Let O be their union $\cup_{i=1}^n U_i$. $O \supset K$ and from Proposition 2.6 we know that for some M and $\rho < 1$ and all x ,

$$\mu^n(x^{-1}O^c), \mu^n(O^c) \leq M\rho^n. \quad (3.68)$$

Let T be the time when the walk Z_n first enters the open set O . Formally,

$$T = \inf\{n \geq 0 | Z_n \in O\},$$

where T is defined to be infinite if Z_n never enters O . (3.68) shows that

$$\mathbf{E}_x T = \sum_n \mathbf{P}_x\{T > n\}$$

is finite. Hence, T is finite with probability 1.

Let us now define the *parallel* right random walks Z_n^i on K by letting Z_0^i be distributed according to ν_i and

$$Z_n^i = Z_{n-1}^i X_{T+n}, \quad n = 1, 2, \dots$$

In other words, increments of parallel walks are the same as those of the original process Z_n from the (random) time T onwards. The Z_n^i process is similar to the Y_n^i process just discussed, apart from the numbering of the increments.

By construction, Z_T belongs to some $U_i, i = 1, 2, \dots, N$, U_1 , say, and then $f(Z_{T+n})$ differs from $f(Z_n^1)$ by less than ε for all n . Then

$$\frac{1}{n} \sum_{k=T}^{T+n-1} f(Z_k)$$

differs by less than ε from

$$\frac{1}{n} \sum_{k=T}^{T+n-1} f(Z_k^1).$$

The latter sum was seen to converge almost surely to $\int f d\nu$. Therefore repeating the argument for a decreasing sequence of values of ε , we finally obtain that $\mathbf{P}_x$ -almost surely

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} f(Z_k) = \int f d\nu.$$

□

Remark 3.13. Theorem 3.12 is a special case of Corollary 3.12. The convergence of the Cesàro sums in Theorem 3.12 are to hold $\mathbf{P}_e$ -a. s. We cannot use Corollary 3.11 directly since the equidistribution for arbitrary measurable f holds only $\mathbf{P}_\pi$ -almost surely and certainly not $\mathbf{P}_e$ -almost surely. Indeed, if μ is deterministic or supported on a finite set of points, the statement of Theorem 3.12 does not hold for noncontinuous f . (Our stopping time T is identically 0 in Theorem 3.12.)

Corollary 3.12 does not quite cover the case in Theorem 3.12, where the group generated by μ^2 is a proper subgroup H of G ; however, the proof of Corollary 3.12 can easily be extended. The proof relies on the fact that the bilateral random walk confined to H (gH) admits a unique invariant probability distribution χ_H (χ_{gH}).

Note also that the mixed walk *always* can be used as a means of experimental determination of K and the idempotent measure $\nu = \alpha \times \chi_G \times \beta$ on K . In the general non-ergodic case, the empirical measures of Z_n , S_n , and W_n converge to some (sample-dependent) stationary distribution.

Corollary 3.13. *Suppose S is a locally compact semigroup with a compact ideal I such that $\mu^k(I)$ is strictly positive for some k . Then S admits a completely simple kernel K and the assertions of all the theorems in Sect. 3.4 are valid for the random walks generated by μ .*

Proof. Exactly as in Proposition 2.6, one shows, using the relation $SIS \subset I$, the analog of (3.68)

$$P^n(x, I^c), \mu^n(I^c) \leq M\rho^n, \quad n = 1, 2, \dots, \quad (3.69)$$

where P stands for any of the four transition probability operators considered.

Asymptotic and ergodic properties of the random walks are thus entirely determined by the process on the compact subsemigroup I . Limit points of all sequences of measures are supported on subsets of I and hence on the compact minimal ideal K of I . Equation (3.68) is again satisfied for any open set O containing K . After a random almost surely finite time, trajectories of all our random walks enter the ideal I and hence any relatively open $O \cap I \supset K$. We leave the detailed verification to the reader. $\square$

Spreadoutness and random walks with continuous components were useful in the study of random walks on locally compact groups. In a compact semigroup situation, such conditions affect the structure profoundly.

Proposition 3.20. *Suppose that the right random walk on the compact semigroup S has a nontrivial continuous component for every $k \in K$. Let $X \times G \times Y$ be the Rees–Suschkewitsch representation of the kernel K . Then X is finite.*

Proof. If k is an element of K , then kS has μ^n -measure 1 for all n . Also

$$P_r^n(\cdot, kS) = 1$$

on kS . Consequently, if $T(\cdot, \cdot)$ denotes the continuous component, the set

$$\{s | T(s, kS) > 0\}$$

is an open set containing every element of kS . The compact kernel K is the union of these minimal right ideals kS . Since the minimal right ideals form an open covering of K , there is in fact only a finite number of minimal right ideals. $\square$

The intersections of the minimal right ideals and the minimal left ideals are isomorphic groups ($=kKk$). If the left walk also satisfies the conditions of Proposition 3.20, then we can postulate Corollary 3.14.

Corollary 3.14. *If both unilateral walks have continuous components which are nontrivial at the points of K , then the subgroups kKk are open and K is a finite union of open groups.*

In Sect. 3.3, terminology for Markov random walks was developed, see the Definitions 3.9 and 3.10. We use the terminology here to present an equidistribution theorem on a compact semigroup S . Define a Markov random walk on S as in Definition 3.9 with $C \subset S$ and the function f the identity function on C . Then S_n is simply $X_1 X_2 \cdots X_n$. Assume further that X_n is ergodic on C and also that C intersects the kernel K of S . Proposition 3.21 is a special case drawn from the theory developed in [242].

Proposition 3.21. *Suppose that the kernel of S is a compact group G . Then the set of essential elements for the Koutsky process (X_n, S_n) is precisely $C \times G$. If $C \times G$ is the only essential class for the Koutsky process, then S_n is ergodic in the sense that it satisfies (3.64) with respect to the Haar measure χ_G*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n F(S_k) = \int F d\chi_G, \text{ almost surely, } F \text{ bounded measurable}$$

Another example with applications to iterated function systems, see Sect. 4.2, is the following.

We assume that the set C is finite. Let X_n be stationary on C , call its invariant distribution π . Assume further that the transition matrix P for the X chain is strictly positive; i.e., for any $c, c' \in C$ there is a positive probability for a transition from c to c' in one step.

Our aim is to investigate the left Markov random walk

$$Y_n = f(X_n) f(X_{n-1}) \cdots f(X_2) f(X_1)$$

on a compact semigroup S that we assume generated by $f(C)$; i.e.,

$$S = \overline{\bigcup_{n=1}^{\infty} (f(C))^n}.$$

Suppose S is equipped with a kernel $K = A \times G$ that is left simple; i.e., there is a unique minimal left ideal, K itself. We also make another restrictive condition on S : sK is a minimal right ideal for every $s \in S$ (and not just for the elements of K).

The invariance semigroups $\Pi(c, c)$ (see Definition 3.10) have the form $f(c)S$. This is an immediate consequence of the strong assumption made on the transition probability matrix P . Then, using the ideas behind the proof of Lemma 3.4, it is not difficult to show that the set of essential elements for the Koutsky process (X_n, Y_n) is $C \times K$. In addition, the elements of $C \times K$ communicate.

The transition probability operator U for the Koutsky process is equicontinuous and hence ergodic on the essential class $C \times K$. The invariant measure is $\pi \times \nu$, where $\nu = \alpha \times \chi_G$ is idempotent on K . The measure α on the minimal right ideals is given by $\alpha(xK) = \pi\{c \in C \mid f(c)K = xK\}$.

To complete Sect. 3.4, let us look at a case, where the transition probability matrix P for the Markov chain on C is extremely structured in contrast with the above situation, where all transitions were possible in one step.

Take C to be the set $\{0, 1/4, -1/4, 1/6, 1/3, -1/2\}$ and take $f(c) = \exp(2\pi ic)$ in the circle group T . The state space of the Markov random walk is thus (a subset of) the compact abelian group T .

Let P be the following 6×6 matrix where the above ordering of the elements of C is preserved:

$$\begin{pmatrix} .2 & .4 & 0 & .4 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

P is aperiodic and irreducible. Nevertheless, the invariance groups $\Pi(c, c)$ are all reduced to the singleton $\exp(2\pi i0) = 1$.

Exercise 3.27. Consider the semigroup S of Example 3.5 (i) represented as points in the unit square. Suppose μ charges only two particular points $(a, a + d)$ and $(a', a' + d)$, where $d < \frac{1}{3}$. Show that the kernel of the semigroup generated by μ is a Cantor-like compact subset of the diagonal of the unit square.

Exercise 3.28. Let f be a continuous function on the compact semigroup S . Suppose f is 0 on the kernel K of S . Let X_n be a random walk on S . Prove that

$$\lim_{n \rightarrow \infty} f(X_n) = 0$$

almost surely. Is it true that

$$\sum_{n=1}^{\infty} f(X_n) < \infty$$

almost surely?

Exercise 3.29. Let S be the semigroup of $d \times d$ stochastic matrices. Suppose the support of μ contains a transition matrix of an irreducible Markov chain. Show that

the left random walk generated by μ converges. Describe the asymptotic behavior of the right, bilateral, and mixed walks.

Exercise 3.30. Let μ be a probability measure on the semigroup of $d \times d$ doubly stochastic matrices. Suppose the support of μ^k contains an all-positive matrix, for some k . Prove that the random walks generated by μ converge.

Exercise 3.31. (Continuation of Exercise 3.27) Consider matrices of the form

$$\begin{pmatrix} p & 1-p & a \\ q & 1-q & b \\ 0 & 0 & c \end{pmatrix},$$

where $0 < p, q < 1$, a and b are real, and $c = \pm 1$.

3.5 Completely Simple Semigroups

The importance of the completely simple minimal ideal of a locally compact semigroup has been stressed throughout this book. For example, it was shown in Chap. 2 that any idempotent probability measure is supported on a completely simple subsemigroup, with a compact group factor. In Sect. 3.2 we showed, for a discrete semigroup S , that if a random walk admits recurrent points, these constitute the completely simple minimal ideal K of S . It is natural then to single out completely simple semigroup for special consideration.

In Sect. 3.5, we assume from the start that the underlying state space is a completely simple, locally compact topological semigroup S . Recall that S is algebraically and topologically isomorphic to a Rees–Suschkewitsch product $X \times G \times Y$ with multiplication

$$(x, g, y)(x', g', y') = (x, g\phi(y, x')g', y'),$$

where X and Y are locally compact spaces, G a locally compact group and the *sandwich function* ϕ a continuous map from $Y \times X$ into G . ϕ is not uniquely determined (see Proposition 1.10). One way of choosing the factors in the Rees–Suschkewitsch product is the following (see Theorem 1.1): Let e be an idempotent of S . Then eSe is a closed (thus locally compact) subgroup of S . Let X be the idempotents of Se and Y the idempotents of eS . With $\phi(y, x)$ taken to be the product yx , $X \times eSe \times Y$ is topologically and algebraically isomorphic to the completely simple semigroup S . (Notice that there is considerable freedom in the choice of the representation. e can in fact be chosen arbitrarily among the idempotents of S .) When dealing with completely simple subsemigroups of S , as D in Propositions and Corollaries below, we choose $e \in D$, so that $eDe \subset eSe$, $E(De) \subset E(Se)$ and $E(eD) \subset E(eS)$ (where $E(\cdot)$ denotes the idempotents of the set in question.) Thus, the Rees–Suschkewitsch factors of D are (chosen to be) *subsets* of the corresponding factors of S .

Let μ be a probability measure on S generating a subsemigroup D :

2075

$$D = \overline{\bigcup_{n=1}^{\infty} S_{\mu}^n}. \quad (3.70)$$

We will work mainly with the problem of recurrence/transience for our random walks generated by μ . This theory is developed for the unilateral walks only. Let us start out though by making some general observations on essentiality.

Proposition 3.22. *Let S be a completely simple locally compact semigroup and μ a measure on S generating the closed subsemigroup D . Then either all or none of the elements of D are essential for the right (left) walk. If the elements of D are essential for one of the unilateral walks, they are also essential for the other unilateral walk. In that case, D is a completely simple subsemigroup of S .*

Proof. Suppose x is essential for the right random walk on D , i.e., $\overline{x\overline{D}}$ is a minimal closed right ideal and $x \in \overline{y\overline{D}} = \overline{x\overline{D}}$ for all $y \in \overline{x\overline{D}}$.

In particular, let $y = x^3$. Then there is a sequence d_n in D such that $x = \lim x^3 d_n = x^2 \lim x d_n$. We have further $x^2 = x^2 \lim x d_n x$. In the group xSx , the equation $x^2 = x^2 y$ has a unique solution, however. Thus, $\lim x d_n x$ is necessarily equal to e , the identity of xSx . (Recall that D is closed.) Hence, $e \in D$ and also $\overline{x\overline{D}} = \overline{e\overline{D}} = e\overline{D} = x\overline{D}$.

Let I be the set of all essential elements for the right random walk. Just as in the proof of Corollary 3.1 it is seen that I is a minimal ideal of D .

I is a simple subsemigroup of S , containing idempotents. All idempotents of S are primitive and hence I is a completely simple subsemigroup of D and S (see Definitions 1.5 and 1.6).

Also, I necessarily coincides with D . To see this, let $d = (a, g, b) \in D$. Since I is completely simple, a union of groups $ili, i \in I$, then $dId \subset I$ contains an idempotent f (= the identity of the group dSd) such that $df = fd = d$. Hence $d \in I$.

D being completely simple, the sets $Dx, x \in D$, are also closed and minimal left ideals. In other words, the left random walk on D admits essential elements. Actually, all elements of D are essential since D is the union of minimal left ideals. $\square$

Corollary 3.15. *If D is completely simple, then all elements of D are essential for the bilateral and mixed walks. There is one essential class for the mixed walk and at most two essential classes for the bilateral walk.*

Proof. The assertions for the mixed random walk follow trivially from the fact that D is simple.

The essential elements for the bilateral walk form a two-sided ideal of D . The verification of this fact can be made along the lines of the corresponding proof in Proposition 3.9. Thus, it remains to show that there are essential elements for the bilateral walk.

To this end, take e to be an idempotent of D . There is a sequence $s_n \in S_\mu^n$ with s_n 2113
converging to e . Then $s_n e s_n$ converges to $eee = e$. We can deduce that e is a return 2114
state for the bilateral walk. Let $e \rightarrow h$. Then, as in the proof of Proposition 3.9, 2115
 $hk = e$ implies $e \rightarrow kh$ and $jh = e$ implies $e \rightarrow hj$. 2116

For all s, t such that $ths = e$ we have $h \rightarrow sht$. There is considerable freedom in 2117
the choice of s and t to satisfy $ths = e$. It is possible for example to take $t = e$ and 2118
 $s \in eDe$. Making use of the explicit product representation (with $G = eDe$) we can 2119
then show that $sht = e$ for that particular choice of s and t . We have thus proved that 2120
 $e \rightarrow h$ implies $h \rightarrow e$. 2121

Next, we sketch the proof of the fact that there are at most two essential classes 2122
for the bilateral walk. 2123

Write D in product form $A \times G \times B = AGB$. Consider the decomposition 2124

$$D = \overline{\bigcup_{n=1}^{\infty} S_\mu^n} = \overline{\bigcup_{n=1}^{\infty} S_\mu^{2n}} \cup \overline{\bigcup_{n=0}^{\infty} S_\mu^{2n+1}}. \quad (3.71)$$

The first term on the right-hand side is the subsemigroup generated by μ^2 . It con- 2125
tains all the squares d^2 of the elements of D and hence all the idempotents of D and 2126
consequently the range of ϕ . We can write it AHB where H is a normal subgroup 2127
of G . 2128

If $H \neq G$, then $G = H \cup gH$ for a $g \in G \setminus H$. 2129

It remains to prove that the identity e of H (and G) communicates with all of 2130
 AHB . In the calculation to follow the subscripted element, c_n indicates an element 2131
of S_μ^n . The elements of D are written in the product form AGB . 2132

Let $s_n \cdot s'_n = agb \cdot a'g'b'$ be an element of S_μ^{2n} written as a product of two 2133
elements in S_μ^n . Then the product is, in fact, unaltered if we expand it to 2134

$$agb \cdot a_{2k} e a_{2k} \cdot a'g'b',$$

where a_{2k} is assumed to equal the idempotent $a' \in E(De)$. This shows that $s_n s'_n$ 2135
is reachable from e by the bilateral walk in $n + 2k$ steps. If $\lim a_{2k} = a'$ instead, 2136
we get the same result, $e \rightarrow s_n s'_n$ and, by continuity, e communicates with all limit 2137
points of sequences $s_n s'_n$, i.e., with all elements of AHB . $\square$ 2138

The starting point for the study of recurrence is the following analog of 2139
Theorem 3.5. 2140

Proposition 3.23. *If $x \in D$ is recurrent for the right (left) random walk, then it is* 2141
essential for the right (left) walk. 2142

The element $x \in D$ is recurrent for the right (left) random walk if and only if 2143

$$\sum_{n=1}^{\infty} \mu^n(N(x)) = \infty \quad (3.72)$$

for all neighborhoods $N(x)$ of x .

For a proof we refer to the detailed treatment in Mukherjea and Tserpes [176], Chap. 2, Sect. 3. The primary difficulty in going from locally compact groups to locally compact completely simple semigroups is the fact that sets of the form $K^{-1}K'$ (where K and K' are compact) are no longer compact.

Propositions 3.22 and 3.23 enable us to draw the following conclusion.

Theorem 3.17. *The points of D are either all or none recurrent with respect to the unilateral walks.*

The right (left) random walk admits recurrent points if and only if

$$\sum_{n=1}^{\infty} \mu^n(K) = \infty \quad (3.73)$$

for some compact $K \subset D$.

Under (3.73), D is a completely simple subsemigroup of S .

The only assertion left to prove is the relationship between (3.73) and recurrence. A compactness argument shows that (3.73) holds for some compact K if and only if (3.72) holds for some $x \in D$, i.e., if and only if the set of recurrent points is nonempty.

Corollary 3.16. *The right (left) random walk is recurrent if and only if the $\sum_{n=1}^{\infty} \mu^n$ is not a Radon measure.*

Remark 3.14. If the Rees–Suschkewitsch product happens to be *direct*, then the behavior of the random walks can essentially be read off from that of the group factor.

Let μ_2 be the marginal distribution of the middle factor g . Then $\sum_n \mu^n$ is a Radon measure if and only if $\sum_n \mu_2^n$ is.

We go on to characterize positive recurrence for the random walks on the completely simple semigroup.

Proposition 3.24. *The set of elements x satisfying*

$$\limsup \mu^n(N(x)) > 0 \quad (3.74)$$

for all neighborhoods $N(x)$ of x , is either empty or consists of all D .

Proof. The points x satisfying (3.74) form a two-sided ideal D' of D : If $N(xs)$ is a neighborhood of xs ($s \in S$) we choose neighborhoods $N(x)$ and $N(s)$ of x and s , respectively, with $N(x)N(s) \subset N(xs)$. Then $\mu^k(N(s)) > 0$ for some k and $\mu^{n+k}(N(xs)) \geq \mu^n(N(x))\mu^k(N(s))$. Thus, xs satisfies (3.73) if x does. The left ideal property is proved similarly.

If x satisfies (3.74), it is obviously recurrent and so D is completely simple (Propositions 3.22 and 3.23). It is easily seen that any bilateral ideal D' of D must necessarily coincide with D . $\square$

Remark 3.15. (i) In the same way as in Theorem 3.17, we see that (3.74) is satisfied for some $x \in D$ if and only if μ^n does not converge to zero vaguely, i.e.,

$$\limsup_{n \rightarrow \infty} \mu^n(K) > 0$$

for some compact subset of $K \subset D$.

(ii) The right random walk has a nontrivial continuous component on D if and only if some power of μ is nonsingular with respect to the measure $\mu_1 \times \chi \times \mu_3$. For a proof see [100], Sect. 5. Note that the topology of the first factor A is necessarily discrete.

Theorem 3.18. *If the group factor G in the Rees–Suschkewitsch representation is compact then all elements of D are positive recurrent.*

If the right random walk has a nontrivial continuous component on D , then the elements of D are positive recurrent if and only if (3.74) is satisfied for some $x \in D$. This is the case if and only if G is compact.

AQ1 *Proof.* If G is compact then (3.74), or rather the equivalent formulation in Remark 3.15(i), is certainly satisfied since

$$\begin{aligned} \mu^{n+2}(A' \times G \times B') &= \mu(A' \times G \times B) \mu^n(A \times G \times B) \mu(A \times G \times B') \\ &= \mu(A' \times G \times B) \mu(A \times G \times B') \end{aligned}$$

for compact subset A', B' of A and B , respectively.

Hence, the random walk is recurrent. Consider the measure $\pi \equiv \mu_1 \times \chi \times \mu_3$, where $\mu_1(\mu_3)$ is the marginal distribution of the first (third) component a (b) and χ is the normed Haar measure on G . (Formally, $\mu_1(A') = \mu(A' \times G \times B)$ for $A' \subset A$.) Clearly, π is the required probability measure invariant with respect to the right and left random walks.

By Proposition 3.12, a positive recurrent point x satisfies the positivity condition

$$\limsup_{n \rightarrow \infty} \mu^n(x^{-1}N(x)) > 0$$

for all neighborhoods $N(x)$ of x .

Take $N(x)$ to be compact and let the $\limsup$ above be $\geq 2\delta > 0$. The set $x^{-1}N(x)$ is noncompact in general but the Rees–Suschkewitsch product structure $D = A \times G \times B$ will enable us to use almost the same arguments as in the group case. Write $x = (a, g, b)$ and take $N(x)$ to be $A' \times G' \times B'$. Let A'' be a compact subset of A with $\mu_1(A'') = \mu(A'' \times G \times B) > 1 - \delta$. Then

$$\begin{aligned} x^{-1}N &= (a, g, b)^{-1}(A' \times G' \times B') \\ &\subset A'' \times (g\phi(b, A''))^{-1}G' \times B' \cup (A \setminus A'') \times G \times B. \end{aligned}$$

Since $\mu^n((A \setminus A'') \times G \times B) = \mu_1(A \setminus A'') < \delta$, we conclude that the μ^n -measure of the compact set $A'' \times (g\phi(b, A''))^{-1}G' \times B'$ is at least δ for infinitely many n .

Thus, (3.74) is satisfied for all $x \in D$.

Conversely, assume that (3.74) holds for the elements of D . Take an idempotent $e \in D$. Then $e^{-1}N(e)$ contains a neighborhood $N'(e)$ of e . Thus, $\limsup \mu^n(e^{-1}N(e))$ is strictly positive. In other words, e is a positive point for the right random walk and hence positive recurrent. If e has the Rees–Suschkewitsch representation $(a, (\phi(b, a))^{-1}, b)$ we see that $\delta_a \times \chi \times \mu_3$ is invariant with respect to the right random walk on $eD = \{a\} \times G \times B$. By positive recurrence, this invariant measure is a probability measure. Thus, the group factor G has to be compact. Clearly all the elements of D are then positive recurrent. $\square$

Remark 3.16. Unfortunately, the previous theorems do not carry over to the completely 0-simple case. Let μ generate the multiplicative locally compact semigroup $S = [0, \infty) = \{0\} \cup (0, \infty)$. Then 0 is essential, but the nonzero elements are not, in general. μ can be chosen to generate a null-recurrent random walk on the open interval $(0, \infty)$ – take μ to be a distribution on a small interval around 1, such that its geometric mean is 1. Then 0 will be the only essential, positive recurrent point, while the other elements of S are inessential and null-recurrent.

Exercise 3.32. Prove the assertion in Remark 3.14: If the Rees–Suschkewitsch product is direct, then the random walk on S is recurrent if and only if the random walk on the middle factor G is.

Exercise 3.33. Let S be the semigroup of 2×2 matrices of the form

$$\begin{pmatrix} ka & a \\ kac & ac \end{pmatrix},$$

where k, a , and c are real numbers (see Example 1.2). S is a nonabelian, completely 0-simple semigroup. Construct a probability measure μ in such a way as to make the nonzero elements of S null-recurrent. Of course, 0 is always positive recurrent.

Exercise 3.34. Let S be the semigroup of 3×3 matrices of the form

$$\begin{pmatrix} 1 & 2 & x \\ \frac{1}{3} & \frac{2}{3} & y \\ 0 & 0 & \varepsilon \end{pmatrix},$$

where x and y are real and ε is ± 1 . Show that S supports null-recurrent left and right random walks.

Exercise 3.35. Let S be the semigroup of 3×3 matrices of the form

$$\begin{pmatrix} p & 1-p & x \\ q & 1-q & y \\ 0 & 0 & \varepsilon \end{pmatrix},$$

where $0 \leq p, q \leq 1$, $x, y \in \mathfrak{R}$ and ε is ± 1 . Show that one can construct a null-recurrent right random walk on S . Does S support a recurrent left random walk as well?

3.6 Notes and Comments

2236

There is certainly no clear distinction between the study of convolution sequences 2237
on semigroups (the subject of Chap. 2) and random walks as in this chapter. After all, 2238
the distribution of the random walk X_n is the n -fold convolution of the generating 2239
measure μ . These very brief Notes and comments will, however, show a preference 2240
for work on path-wise and almost sure properties, such as communication structure, 2241
recurrence/transience, and ergodicity. Chapter 2 of the lecture notes by Mukherjea 2242
and Tserpes [176] opens up with a more detailed historical survey of the develop- 2243
ment of theory of random walks on groups and semigroups. The volume edited by 2244
Cohen et al. [44] is a survey of the study of random matrices. It also contains a large 2245
bibliography. The recent book by Woess [241] treats random walks on graphs and 2246
discrete groups. The recurrence/transience problem and its relation to the structure 2247
of the underlying space is the focus of Chap. 1. Guivarc'h in [81] studies random 2248
walks on large classes of classical groups, notably groups of matrices. He offers a 2249
deep survey of current work on limit theorems and also Gaussian laws (central limit 2250
theorems) on such groups. 2251

Section 3.1

2252

The book by Revuz [199] gives a solid introduction to Markov chains and random 2253
walks. 2254

The concept of continuous component was introduced Tuominen and Tweedie in 2255
[229]. They also saw the (essential) equivalence between their concept and that of 2256
spreadoutness (*étalement*), a nonsingularity condition of great utility in the study of 2257
random walks on groups. 2258

The bilateral random walk seems to have been first studied by Larisse [135] 2259
(in discrete semigroups) and the mixed random walk by Högnäs [97] (in compact 2260
semigroups). 2261

Section 3.2

2262

The pioneering paper by Furstenberg and Kesten [70] on random matrices deals with 2263
a semigroup structure but does not primarily view it in semigroup terms. The very 2264
influential and inspiring book by Grenander [78] insists on the importance of work- 2265
ing with the underlying algebraic structure. Grenander's student Martin-Löf [142] 2266
highlights the interplay between algebra and probability theory in his very clearly 2267
written paper. Martin-Löf deals exclusively with discrete semigroups as does Larisse 2268
[135], who develops and deepens the theory. Our treatment owes very much to these 2269
two authors. 2270

There is a deep connection between algebraic coding theory and semigroup theory. This has been recognized especially in France, where the school of Schützenberger has developed both theories in parallel. A particularly nice probabilistic example is provided by the study of *densities* of code words.

Let A be an alphabet and A^* the set of all words on the alphabet A . A^* is a semigroup under concatenation of words, the empty word being the identity element. A code over A is a subset X of A^* such that any nonempty word in X^* can be written uniquely as a product of words in X (see [11], Chap. 1).

Given a probability distribution π on the alphabet A , it is easily extended to a probability distribution on longer words by the recursion $\pi(aw) = \pi(a)\pi(w)$, where $a \in A$ and $w \in A^*$. Such a distribution is called a *Bernoulli distribution* on A^* . By definition, it is also a homomorphism from A^* into the unit interval. The Bernoulli distribution π is *positive* if it is positive on all letters of A . In the terminology of this book, we are studying an irreducible random walk on A^* .

If L is a subset of A^* , then it has a *density* if the sequence $\pi(L \cap A^n)$ converges in the Cesàro sense (where A^n consists of the n -letter words in A^*). If L is highly structured then it turns out that the density of L may be shown to exist.

Let M be a *well-founded monoid*, i.e., a semigroup with identity admitting a completely simple minimal ideal with finite group factor. If $\phi : A^* \rightarrow M$ is a homomorphism, then it may be shown that all sets of the form $\phi^{-1}(m)$ do have a density. Furthermore, this density is nonzero if and only if $m \in K$, the completely simple kernel of M (see [11], Theorem 2.4 of Chap. 6).

As shown in [11], Chap. 6, the applications in coding theory emanate from the fact that certain codes X are naturally associated with particular monoids M . Hence if the completely simple kernel of such a monoid is known, then the inverse images of its elements, minimal ideals and maximal groups have densities which are expressible in terms of the Rees–Suschkewitsch product representation of the kernel (in fact, it is the order of the group factor that is the most important feature). Among other things, an expression for the density of X^* , the interesting quantity in this connection, is obtained.

The ergodicity results of Corollary 3.3 indicate ways of experimentally determining the idempotent measure $\alpha \times \chi_G \times \beta$. One of the reasons the mixed random walk was devised in [97] was precisely the wish to find a simple way to simulate the idempotent measure (see formula (3.28) and Corollary 3.12(iv)).

The quotient of G by the normal subgroup N generated by the sandwich function in Proposition 3.7 defines the maximal homomorphism from S to G , see [112, 135, 214].

The Markov random walks, cf. Proposition 3.8 and Definition 3.9, were introduced by Koutsky [134] and Cigler [40] in the finite group case. Schmetterer [211], Muthsam [178], and Högnäs [102] study Markov random walks on discrete semigroups and obtain structure theorems, which are not restated here. The group case and the compact semigroup case are, however, treated in some detail in Sects. 3 and 4, respectively. The main sources are [242] and [106]. The same or similar processes have been widely studied under different names, such as Markov walks (see [80]), semi-Markov processes and Markov renewal processes, e.g., in [127] and [2].

For Exercise 3.17 and further discussion on the continuity of convolution, see [142], Sect. 3. Exercise 3.18 is taken from [135], Part II. Exercise 3.25: Recurrence and transience of random walks on the semidirect product of a compact group with Z^d or $\mathfrak{N}^d$ is the subject of [49].

Section 3.3

2320

Pólya [190] seems to be the first to treat the recurrence problem, for random walks on the d -dimensional lattice. Other early papers on recurrence are [39, 138].

Theorem 3.8 was the so-called Kesten conjecture for connected Lie groups. The conjecture goes back to [126], where the relationship between the growth and recurrence properties is explored. (The more precise version could be called the Guivarc'h-Keane conjecture, explored in a series of papers by these two authors, e.g., [82, 83].) The growth of a connected locally compact group G is said to be polynomial of order α if the Haar measure $\chi(V^n)$ of the n th power of a neighborhood V of the identity e is bounded by a constant times n^α for all n . The growth is termed exponential if it is faster than polynomial of order α for all α . The connected Lie group case may be considered solved by 1977: The proof of Baldi [4] combines an observation by Baldi et al. [5] and the results in the book by Guivarc'h et al. [87].

Originally, the Kesten conjecture was formulated for discrete groups (where V is then taken to be a finite set generating the group): a discrete group with exponential growth is transient. The problem of recurrence criteria for (finitely generated) discrete groups was solved by Varopoulos [232, 233]. The monograph by Woess [241] gives an account in the context of random walks on graphs.

The groups G_1 and G_2 in Exercises 3.25 and 3.26 are studied by Crépel in [49].

Section 3.4

2339

If the underlying space is a group, the study of random walks has a rich and well-known structure at its disposal. In the semigroup situation, the first task is to find the right structure to work with. It turns out that random walks with interesting properties will force certain structure at least on some subsemigroups. In particular, any stationary distribution will have to be supported by a completely simple subsemigroup (see Chap. 2).

Early interest in the recurrence/transience of the random walks on a compact semigroup is shown by Brissaud [20], drawing on papers by Tortrat [223–225] as well as on the book by Grenander [78]. These authors have published papers on this question in the early 1970s with co-authors Sun and Tserpes: [98, 99, 171, 172, 177, 218], see also the lecture notes [176].

We follow largely the method of Rosenblatt [202, 204] exploiting the equicontinuity of the transition probability operators. In addition to being instrumental in

proofs of existence theorems of invariant probability measures, they are efficient tools when establishing ergodic theorems as well. Rosenblatt's results were considerably deepened by Raugi in [195]. For equicontinuous transition probability operators, he obtains a decomposition of the absorbing part of the state space into compact sets corresponding to ergodic classes. Those classes are in turn decomposed into cyclic subclasses. For our right random walks, the ergodic classes are the closed minimal right ideals $\{x\} \times G \times Y$ of the compact semigroup generated by μ . The cyclic classes are closed subsets of the form $\{x\} \times gH \times Y$ for some compact subgroup H of G and a $g \notin H$.

Examples 3.5(i) and 3.27 are treated more thoroughly in [204], pp. 159–160. See also [217]. More information on bistochastic matrices, part (iii) of the same Example, can be found in [137] while the affine contractions, part (iv), are also discussed in Chap. 4.

The exact form (3.59) (and the corresponding (3.19) in the discrete case) are new as far as we know.

The parallel random walks, essentially a coupling argument, used in the proof of Corollary 3.12 follow [96, 97]. Corollary 3.13 is the subject of [99].

Section 3.5

The recurrence results are due to Sun et al. [218].

Referring to Theorem 3.18 we note that the obvious conjecture – positive recurrence implies that the group factor is compact – has been proved only for special cases.

There have been numerous attempts to study recurrent random walks on locally compact semigroups S . The natural starting point has been formula (3.72), i.e., it is assumed that for some $x \in S$

$$\sum_{n=1}^{\infty} \mu^n(N(x)) = \infty$$

for all neighborhoods $N(x)$ of x . This assumption alone has not been very successful, results are obtained only when some structure is imposed on the semigroups or the random walks.

If the random walks are assumed to admit continuous components, then (3.72) implies the existence of a completely simple kernel, see [103].

On the other hand, if the semigroup itself satisfies some compactness conditions then, again, (3.72) guarantees the existence of a completely simple kernel, where the random walks are ultimately attracted.

Such a compactness condition is:

AB^{-1} and $A^{-1}B$ are compact whenever A and B are.

A weaker set of conditions (CR) and (CL) were introduced by Sun and used for example in [149]:

(CR) $x \notin Sy$ implies that there exist neighborhoods N_x, N_y of x and y such that $N_x N_y^{-1} = \emptyset$.

(CL) is defined dually.

Matrix semigroups do not satisfy (CR) and (CL) in general. Being fairly concrete objects something can be said about them nevertheless. The assumption (3.72) together with some additional assumptions implies the existence of a completely 0-simple structure within S , see [111].

For the details of Exercise 3.33, we refer to [111].

The Rees–Suschkewitsch structure appears quite naturally in the study of random walks and convolution sequences on groups. Bougerol in [16] sets out to study measures supported by a group of invertible matrices. Under a tightness condition, the semigroup generated turns out to have a completely simple kernel (with a compact group factor) consisting of matrices of lower rank. The semigroup is a direct product. The limit points of the convolution sequence essentially live on this kernel.

Guivarc’h in [81] discusses central limit theorems for random walks on groups. If the group $G = SL(V)$, V a vector space, one can use the polar decomposition of G to obtain a description of the limiting Gaussian law on a completely simple semigroup with a direct product structure. Interestingly, if G is a group of affine maps on the real line, an “ $ax + b$ ” group, then the limiting structure is an affine semigroup under the MaxPlus algebra, where the semigroup operation may be written

$$(\beta, \alpha)(\beta', \alpha') = (\max\{\beta, \alpha + \beta'\}, \alpha + \alpha'),$$

cf. the Notes and Comments to Sect. 1.6. and [81], Chaps. 2 and 7.

AUTHOR QUERY

- AQ1. Please check if the Remark 3.15(i) given in the sentence “If G is compact then...” is appropriate.

Uncorrected Proof

Chapter 4

Random Matrices

4.1 Introduction

In Chap. 4, we apply methods and results from Chaps. 2 and 3 to random matrices. To include only results that are reasonably complete, we restrict our attention to the class of nonnegative matrices (that is, matrices whose entries are all nonnegative). Although there have been a great deal of results in this area in different directions (see [17] or [44]), here we restrict ourselves only to problems involving recurrence, tightness, invariant measures, and laws of large numbers for products of random matrices. The reason is, of course, our own bias and also the desire to avoid duplicating work already available in books or well-known papers.

In Sect. 4.2, we study the structure of the set of recurrent states for left and right random walks in nonnegative matrices. Here, the reader finds similarities with classical results for random walks on R^d . In the context of positive recurrence and also the existence of invariant measures for such walks, the concept of tightness for the convolution powers of a probability measure/distribution (namely, the distribution of individual matrices in independent and identically distributed (i.i.d.) products) is unavoidable, and as such, we examine thoroughly this concept in Sect. 4.3. In Sect. 4.4, we consider invariant measures for mixed random walks and use these measures to obtain laws of large numbers for such walks. In Appendix C, we give a complete description of the asymptotic behavior in the growth of products of random nonnegative matrices in the almost sure sense. In Appendix B, an interesting example of a sequence of i.i.d. $d \times d$ stochastic matrices (where the products are almost surely convergent) is given, and here, the limit distribution has been computed in details in a special case. In Appendix A, we examine whether convergence in distribution of products of i.i.d. random stochastic matrices can be affected by the skeletons of these matrices.

4.2 Recurrent Random Walks in Nonnegative Matrices

28

In this section, we use techniques and results established in Chaps. 2 and 3 to obtain results on recurrence for products of i.i.d. random nonnegative matrices, then apply these results to obtain laws of large numbers for i.i.d. random affine maps from $(R^+)^d$ into $(R^+)^d$. (Here, nonnegative matrices are matrices whose entries are nonnegative.)

The main results in this section are Theorems 4.2, 4.3, and 4.5.

Consider the i.i.d. sequence (X_i) , $i \geq 0$, of random variables with values in $d \times d$ nonnegative matrices with distribution μ . Let S be the closed (with usual topology for matrices) multiplicative semigroup generated by S_μ , the support¹ of μ . As in Chap. 3, we define the right random walk by the sequence (Z_n) , where $Z_n = X_0 X_1 \dots X_n$, and the left random walk by the sequence (W_n) , where $W_n = X_n X_{n-1} \dots X_0$. Let J_c denote the set of all matrices in S that have at least one zero column (that is, a column consisting of only zeros), J_r the set of all matrices in S that have at least one zero row, and let $J = J_c \cup J_r$. As in Chap. 2, we use set notations

$$A^{-1}B = \{x \in S : yx \in B \text{ for some } y \text{ in } A\}$$

and

$$AB^{-1} = \{x \in S : xy \in A \text{ for some } y \in B\},$$

where $A \subset S$, $B \subset S$.

Our first result on recurrence is Theorem 4.1.

Theorem 4.1. *Suppose there exists y in $S - J_c$ such that for every open set N_y containing y ,*

$$\sum_{n=1}^{\infty} \mu^n(N_y) = \infty. \quad (4.1)$$

Let $K = \{x \in S : x \text{ has property (4.1)}\}$. Then K is an ideal of S such that $K \cap J_c^c$ is a completely simple (minimal) ideal of $S \cap J_c^c$. A similar result holds if (4.1) holds for some y in $S - J_r$.

Proof. It follows from the proof of Proposition 3.19 in Chap. 3 that for any open set N_y containing y , we have

$$P(Z_n \in N_y^{-1}N_y \text{ infinitely often}) = 1. \quad (4.2)$$

Let (N_n) be a sequence of relatively compact decreasing open sets forming a local base at y . Then there are elements y_n and z_n such that $y_n \in N_n$, $y_n z_n \in N_n$ and

¹ Note that the support of μ has also been denoted by $S(\mu)$.

$z_n \in N_n^{-1}N_n$. Since y_n converges to y and y has no zero column, there exists a positive integer n_0 and $\delta > 0$ such that for $n > n_0$,

$$\sum_{i=1}^d (y_n)_{ij} > \delta, \quad 1 \leq j \leq d,$$

so that for some $m_0 > 0$ and $n > n_0$,

$$\begin{aligned} \delta \cdot (z_n)_{jk} &\leq \sum_{i=1}^d (y_n)_{ij} (z_n)_{jk} \\ &\leq \sum_{i=1}^d (y_n z_n)_{ik} \leq m_0, \end{aligned}$$

whenever $1 \leq j \leq d$ and $1 \leq k \leq d$. This means the sequence (z_n) is bounded so that there exists z in S such that it is a subsequential limit of the z_n s and $y = yz$. Since y has no zero column, z also cannot have a zero column. Thus, we have shown that

$$y \in y(S - J_c). \quad (4.3)$$

Notice that $S - J_c$ is a semigroup and let $u \in y(S - J_c)$. Then $u = ys$, $s \in S - J_c$. Let N_s be an open set containing s . There exists a positive integer k such that

$$\begin{aligned} 0 &< P(Z_k \in N_s) \\ &= P(Z_k \in N_s, Z_n \in N_y^{-1}N_y \text{ infinitely often}) \\ &\leq P(Z_k \in N_s, X_{k+1} \dots X_n \in N_s^{-1}[N_y^{-1}N_y] \text{ infinitely often}) \\ &= P(Z_k \in N_s) P(X_{k+1} \dots X_n \in N_s^{-1}[N_y^{-1}N_y] \text{ infinitely often}). \end{aligned}$$

This means that given any open set N_u containing u , $N_u^{-1}N_y \neq \emptyset$ if N_y is chosen so that $N_y N_s \subset N_u$ for some open set N_s containing s . Again, by using a similar argument to that used above in establishing (4.3), it follows that there exists v in S such that $y = u \cdot v$ and $v \in S - J_c$. Thus, we have

$$u \in y(S - J_c) \Rightarrow y \in u(S - J_c).$$

This means $y(S - J_c)$ is a minimal right ideal of $S - J_c$. Since K is an ideal (which can be easily verified), it follows that

$$K - J_c = \cup \{y(S - J_c) : y \in K - J_c\}. \quad (4.4)$$

Now if $I \subset K - J_c$ and I is an ideal of $S - J_c$, then for any $y \in K - J_c$, yI is a right ideal of $S - J_c$ and $yI \subset y(S - J_c)$. Since $y(S - J_c)$ is a minimal right ideal containing y , it follows that

$$y \in y(S - J_c) = yI \subset I$$

so that $K - J_c$ is simple.

It remains to show that $K - J_c$ has at least one primitive idempotent element. To this end, let $y \in K - J_c$. Since $y^2 \in y(S - J_c)$ and $y(S - J_c)$ is a minimal right ideal, $y \in y^2(S - J_c)$, so there exists $z \in y(S - J_c)$ such that $y = y \cdot z$. Then for each positive integer m , $y = yz^m$. Since y has no zero column, the sequence (z^m) is a bounded sequence. Notice also that whenever a sequence yz_n converges to some element z_0 , the sequence (z_n) is bounded (since y has no zero column), so $z_0 \in yS$. This means yS is a closed subset of S . Thus, the closure of $\{z^m : m \geq 1\}$, which is a compact subsemigroup, contains an idempotent element e (see Theorem 1.6) and $y = y \cdot e$. Now e has no zero column, since y has no zero column. Consequently, since $z \in yS$,

$$e \in yS - J_c = y(S - J_c),$$

which is now a right group. Now we show that every idempotent in $K - J_c$ is primitive. Let e_1, e_2 be any two idempotents in $K - J_c$ such that $e_1 e_2 = e_2 e_1 = e_1$. Then e_1 is an idempotent in the right group $e_2(S - J_c)$. This means $e_1 = e_2$. This proves that every idempotent in $K - J_c$ is primitive, and therefore, $K - J_c$ is completely simple. $\square$

Corollary 4.1. *Suppose there exists $y \notin J$ such that (4.1) holds for y . Then, $K - J$, K as in Theorem 4.1, is a completely simple (minimal) ideal of $S - J$.*

Proof. Notice that if $w \in z(S - J_c)$ and $z \in K - J$, then by Theorem 4.1, $z \in w(S - J_c)$. This means $w \notin J_r$, since $z \notin J_r$. Thus, it follows from (4.4) that

$$\cup \{z(S - J_c) : z \in K - J\} \subset K - J.$$

By Theorem 4.1, for each $z \in K - J$, $z \in z(S - J_c)$. Then the preceding inclusion is an equality, and $K - J$ is a union of right groups. Since $K - J$ is an ideal of $S - J$, the corollary follows. $\square$

Our next result shows that results similar to those for random walks with values in R^d also hold for random walks in the present context.

Theorem 4.2. *Let $y \notin J_c$. Then the following results are equivalent:*

- (i) $\sum_{n=1}^{\infty} \mu^n(N_y) = \infty$ for every open set N_y containing y .
- (ii) $P(Z_n \in N_y \text{ infinitely often} \mid Z_0 = y) = 1$ for every open set N_y containing y .
- (iii) $\sum_{n=1}^{\infty} \mu^n(y^{-1}N_y) = \infty$ for every open set N_y containing y .
- (iv) $P(Z_n \in N_y \text{ infinitely often}) > 0$ for every open set N_y containing y .

If $y \notin J_r$, then results similar to the preceding also hold for (W_n) with $y^{-1}N_y$ above replaced by $N_y y^{-1}$. If $y \notin J$, then each of the preceding four statements is also equivalent to each of the following:

- (i) $P(W_n \in N_y \text{ infinitely often} \mid W_0 = y) = 1$ for every open set N_y containing y .
- (ii) $P(W_n \in N_y \text{ infinitely often}) > 0$ for every open set N_y containing y .

(iii) $\sum_{n=1}^{\infty} \mu^n (N_y, y^{-1}) = \infty$ for every open set N_y containing y . 109

Proof. (i) $\Rightarrow$ (ii). (i) implies (as in Theorem 4.1) that for every open set N_y containing y , we have 110
111

$$\Pr (Z_n \in N_y^{-1} N_y \text{ infinitely often}) = 1. \quad (4.5)$$

Since $y \in J_c^c$, it can easily be verified that given an open set N_y containing y , there is always an open set N_{0y} containing y such that 112
113

$$N_{0y}^{-1} N_{0y} \subset y^{-1} N_y.$$

Then (ii) follows from (4.5). 114

(ii) $\Rightarrow$ (iii). Follows by the Borel–Cantelli lemma. 115

(iii) $\Rightarrow$ (i). Assume statement (iii). Let $(N_{my}), \text{cl}(N_{m+1,y}) \subset N_{my}$, be a countable local base at y such that each is relatively compact. Since $y \in J_c^c$, the set $y^{-1}(\text{cl}(N_{my}))$ is compact for each $m \geq 1$. Now utilizing this along with (iii) with N_{my} (in place of N_y), we see there exist z_m such that the sequence yz_m converges to y and for every open set N_{z_m} containing z_m , 116
117
118
119
120

$$\sum_{n=1}^{\infty} \mu^n (N_{z_m}) = \infty. \quad (4.6)$$

Since $y \in J_c^c$, it follows that the z_m s are bounded and as such there exists z in S such that z is a subsequential limit of the z_m s and $y = yz$. It follows from (4.6) that for every open set N_z containing z , we also have 121
122
123

$$\sum_{n=1}^{\infty} \mu^n (N_z) = \infty. \quad (4.7)$$

Since the set of all points of S with the property (4.7) is an ideal of S , it follows that $y = yz$ also has the property in (4.7). Thus (i) follows. 124
125

(iv) $\Rightarrow$ (i) [$\Rightarrow$ (ii)] follows by the Borel–Cantelli lemma. 126

(ii) $\Rightarrow$ (iv). We use an argument similar to the one in (iii) $\Rightarrow$ (i). By (ii), there is a countable local base N_{my} such that for $m \geq 1$, 127
128

$$\Pr (Z_n \in y^{-1}(\text{cl}(N_{my})) \text{ infinitely often}) = 1.$$

Again by the compactness of $y^{-1}(\text{cl}(N_{my}))$, there exist z_m as before such that yz_m converges to y , z_m s are bounded and 129
130

$$\Pr (Z_n \in N_{z_m} \text{ infinitely often}) > 0 \text{ for every open set } N_{z_m} \text{ containing } z_m.$$

If z is a subsequential limit of the z_m s, then $yz = y$ and 131

$$\Pr (Z_n \in N_z \text{ infinitely often}) > 0 \text{ for every open set } N_z \text{ containing } z.$$

Now given an open set N_y containing y , there exist open subsets N_{1y} containing y 132
and N_{1z} containing z such that $N_{1y}N_{1z} \subset N_y$. Let k be a positive integer such that 133
 $\Pr(Z_k \in N_{1y}) > 0$. Then we have 134

$$\begin{aligned} & \Pr(Z_n \in N_y \text{ infinitely often}) \\ & \leq \Pr(Z_k \in N_{1y}, X_{k+1} \dots X_n \in N_{1z} \text{ infinitely often}) \\ & = \Pr(Z_k \in N_{1y}) \cdot \Pr(Z_n \in N_{1z} \text{ infinitely often}) > 0. \end{aligned}$$

This establishes (iv). 135

The rest of Theorem 4.2 follows by similar dual arguments. $\square$ 136

Remark 4.1. We call an element x in S positive recurrent if 137

$$\lim_{n \rightarrow \infty} \sup \mu^n(N_x) > 0$$

for every open set N_x containing x . Suppose the set I of positive recurrent states is 138
nonempty and $I \cap J^c \neq \emptyset$. Suppose also $\mu(J) = 0$. Then it is easily verified that 139
for each positive integer n , $\mu^n(J) = 0$, and as such, we can restrict μ to J^c (since 140
 J^c is open) when we are considering only elements in S that have the property in 141
(4.1). It then follows from Theorem 2.11, see also Proposition 4.1, that μ^n is tight. 142
By Theorem 2.7, S has a completely simple (minimal) ideal, which is precisely the 143
set I . Thus, we have proven that whenever $I \cap J^c \neq \emptyset$ and $\mu(J) = 0$, I is the 144
kernel of S , and by Theorem 4.1 and Corollary 4.1, 145

$$I \cap J^c = K \cap J^c, I \cap J_r^c = K \cap J_r^c, \text{ and } I \cap J_c^c = K \cap J_c^c,$$

where K is the set of all elements in S with the property in (4.1). 146

Let us now formally define the following recurrent sets: 147

$$\begin{aligned} R &= \left\{ x \in S : \sum_{n=1}^{\infty} \mu^n(N_x) = \infty \text{ for every open set } N_x \text{ containing } x \right\}, \\ R(Z) &= \{x \in S : P(Z_n \in N_x \text{ infinitely often}) > 0 \\ & \quad \text{for every open set } N_x \text{ containing } x\}, \\ R(W) &= \{x \in S : P(W_n \in N_x \text{ infinitely often}) > 0 \\ & \quad \text{for every open set } N_x \text{ containing } x\}. \end{aligned}$$

Note that $R \supset R(Z) \cup R(W)$, $R(Z)$ is a left ideal of S , and $R(W)$ is a right ideal 148
of S , while R is a (two-sided) ideal of S . Theorems 4.1 and 4.2 did not deal with 149
elements in the set J in the context of recurrence. In what follows, we show that 150
under mild conditions, we can say more about the recurrent sets $R(Z)$ and $R(W)$. 151
First, we need the two following technical lemmas. 152

Lemma 4.1. *Let H be a (multiplicative) semigroup of $d \times d$ nonnegative matrices and $\mathcal{J}$ the semigroup of all $d \times d$ strictly positive matrices such that*

$$\mathcal{J}_1 \equiv H \cap \mathcal{J} \neq \phi, m(H - \{0\}) \cap J_r^c \neq \phi.$$

Write:

$$X \equiv m(H - \{0\}) \cup \{0\} \text{ and } Y \equiv m(H - \{0\}) \cap J_r^c (= m(H \cap J_r^c)).$$

As in Chap. 1, $m(A) = \{x \in A \mid \text{rank}(x) \leq \text{rank}(y) \text{ for all } y \in A\}$. Then X is a completely 0-simple subsemigroup if and only if Y is a completely simple subsemigroup. When $0 \notin H$, the same result holds with obvious modifications.

Proof. Let us assume Y is a completely simple subsemigroup. Write $H_1 \equiv H - \{0\}$. Observe that

$$\mathcal{J}_1 \cdot H_1 \mathcal{J}_1 \subset J$$

so that

$$\mathcal{J}_1 \cdot m(H_1) \cdot \mathcal{J}_1 \subset m(H_1) \cap J.$$

Let $y \in m(H_1) \cap \mathcal{J}$. Then $y \in Y$, and by our assumption $y \cdot Y \cdot y$ is a group. Let e be the identity of this group. Then $e \in \mathcal{J}$. Since $e \in \mathcal{J}$,

$$e \cdot [m(H_1) \cap J_r] e \subset \mathcal{J} \subset J_r^c.$$

Then it follows that

$$\begin{aligned} & e \cdot m(H_1) \cdot e \\ &= e [m(H_1) \cap J_r] e \cup e [m(H_1) \cap J_r^c] e \\ &= e \cdot Y \cdot e (= yYy), \end{aligned}$$

which is a group. Since $e \in \mathcal{J}$,

$$e [m(H_1) \cap J_r] \subset J_r^c.$$

Therefore, $e \cdot m(H_1) = eY$, which is a minimal right ideal of Y . This means $eY \cup \{0\}$ is a nonzero minimal right ideal of X . Now let us consider $m(H_1) \cdot e$. Let $x \in m(H_1)$. Then there exists z in $m(H_1)$ such that

$$(eze)(exe) = e.$$

Consequently,

$$[m(H_1) e] x e \supset [m(H_1) eze] exe = m(H_1) e;$$

then $m(H_1)e \cup \{0\}$ is a nonzero minimal left ideal of X . Write:

170

$$A \equiv e \cdot m(H_1) \cup \{0\}, L \equiv m(H_1)e \cup \{0\}.$$

Then $L \cdot A \supset L \cup A$, and $L \cdot A$ is an ideal of X . It is actually a nonzero minimal ideal of X , since if $I \subset L \cdot A$ and I is a nonzero ideal of X , then for any $w \in A$, $w \neq \{0\}$, Lw is a nonzero minimal left ideal of X and

171

172

173

$$Lw = I \cdot Lw \subset I.$$

Thus, $L \cdot A$ is a completely 0-simple subsemigroup (see Proposition 1.14). We now show that $L \cdot A = X$. Let $x \in m(H_1)$. Since $e \in \mathcal{J}$, $xex \neq 0$; also, x and xex have the same (minimal) rank, and thus as linear transformations from R^d into itself, they have the same range space. Since $xex \in L \cdot A$ (e is in $L \cdot A$), there exists an idempotent e_0 in $L \cdot A$ such that

174

175

176

177

178

$$e_0 \cdot xex = xex.$$

Now let $u \in R^d$. Then there exists $v \in R^d$ such that $x(u) = [xex](v)$. Thus

179

$$[e_0 \cdot x](u) = [e_0 \cdot (xex)](v) = [xex](v) = x(u).$$

This means $x = e_0x \in L \cdot A$. This proves that $x = L \cdot A$, and X is a completely 0-simple subsemigroup.

180

181

Now suppose X is a completely 0-simple subsemigroup. Let $y \in Y$. Then by Proposition 1.14, $yXy - \{0\}$ is a group containing y . Since $y \notin J_r$, every element of this group is in

182

183

184

$$J_r^c \cap m(H_1) \equiv Y$$

so

185

$$yXy - \{0\} = y[yXy - \{0\}]y \subset y \cdot Y \cdot y,$$

which is a group containing y . We show that Y is simple. If I is an ideal of Y , then for any $z \in Y$, $I \supset zIz$. Since $I \cap zYz \neq \phi$, $z \in zYz$, which is a group, it follows that $I \supset zYz$. Thus, Y is simple.

186

187

188

It is easily verified that for $y \in Y$, $Y \cdot y$ is a minimal left ideal and yY is a minimal right ideal. If $L_0 \subset Y \cdot y$ and L_0 is a left ideal of Y , then for $z \cdot y \in L_0$ and $z \in Y$, $yz \cdot y \in L_0 \cap yYy$. $y \in yYy \subset L_0$. Hence $Y \cdot y \subset L_0$. Also, if $R_0 \subset yY$ and R_0 is a right ideal of Y , then for $y \cdot z \in R_0$ and $z \in Y$, we have $yz \cdot y \in R_0 \cap yYy$. Then $y \in yYy \subset R_0$ and $yY \subset R_0$. Hence, Y is completely simple. $\square$

189

190

191

192

193

To consider recurrence for affine maps f on $(R^+)^d$, $f(x) = Ax + B$ with A and B as in (4.8), we need Lemma 4.2.

194

195

Lemma 4.2. *Let H be a (multiplicative) semigroup of $(d+1) \times (d+1)$ nonnegative matrices of the form* 196
197

$$\begin{pmatrix} A & B \\ 0 & 1 \end{pmatrix}, \quad (4.8)$$

where A is a $d \times d$ matrix, B is a $d \times 1$ vector, $0 \equiv (0, \dots, 0)$ is a $1 \times d$ all-zero vector. Let $\mathcal{M}$ be the set of all such matrices such that 198
199

$$\begin{pmatrix} A & B \\ 0 & 1 \end{pmatrix} \in \mathcal{M} \text{ if and only if } \max \left\{ \min_{1 \leq j \leq d} A_{ij}, B_i \right\} > 0,$$

for each i , $1 \leq i \leq d$. Suppose $H \cap \mathcal{M} \neq \emptyset$ and $m(H) \cap J_r^c \neq \emptyset$, where J_r is the 200
set of all $(d+1) \times (d+1)$ nonnegative matrices with at least one zero row. Then 201
 $m(H)$ is a completely simple (minimal) ideal of H if and only if 202

$$m(H) \cap J_r^c (= m(H \cap J_r^c))$$

is a completely simple (minimal) ideal of $H \cap J_r^c$. 203

Proof. The proof follows the same lines as that of Lemma 4.1. The only point we 204
need to note here is that 205

$$\mathcal{M} \cdot (H - \{w^*\}) \cdot \mathcal{M} \subset \mathcal{M},$$

where w^* is the special element (the $(d+1) \times (d+1)$ matrix), where the only 206
nonzero element is 1 and it is in the last row and the last column. $\square$ 207

Now we can present a version of Theorem 4.2 even for elements in J , but with 208
the additional assumptions: 209

$$\mu(J) = 0 \text{ and } S \cap \mathcal{J} \neq \emptyset. \quad (4.9)$$

Theorem 4.3. *Let μ , S , Z_n , and W_n be defined as before. (Here, we are considering 210
general $d \times d$ nonnegative matrices as in Theorems 4.1 and 4.2.) Assume (4.9). 211
Suppose $R(Z)$ is nonempty. Then it must either be $\{0\}$ (in case $0 \in S$) or 212*

$$R(Z) = R(W) = m(S - \{0\}) \cup \{0\}.$$

The same is true if $R(W)$ is assumed nonempty instead of $R(Z)$. 213

Proof. Recall that R is defined by 214

$$R = \left\{ y \in S : \sum_{n=1}^{\infty} \mu^n(N(y)) = \infty \text{ for every open set } N(y) \text{ containing } y \right\}.$$

Then $R \supset R(Z) \cup R(W)$. Suppose $R \neq \phi$ and R contains a nonzero element of S . 216
If $x \neq 0$ and $x \in R$, then since R is an ideal for $y \in \mathcal{J}$ 217

$$yx \in R \cap J_r^c, \quad xy \in R \cap J_c^c.$$

Therefore, by Theorem 4.1, $yx \in R(W)$ and $xy \in R(Z)$. By Theorem 4.1, $R \cap J_r^c$ 218
is a completely simple (minimal) ideal of $S \cap J_r^c$, and the same is true for $R \cap J_c^c$, 219
which is a completely simple (minimal) ideal of $S \cap J_c^c$. Thus 220

$$R \cap J_r^c = m(S \cap J_r^c), \quad R \cap J_c^c = m(S \cap J_c^c). \quad (4.10)$$

By Lemma 4.1,

$$m(S - \{0\}) \cup \{0\}$$

is a completely 0-simple ideal of S . We claim that

$$\begin{cases} R \cap J_c^c \subset \text{cl}(R \cap J_r^c), \\ R \cap J_r^c \subset \text{cl}(R \cap J_c^c). \end{cases} \quad (4.11)$$

Let us prove only the first statement in (4.11). The proof of the second statement 223
is similar. Let $z \in R \cap J_c^c$. Then by Theorem 4.2, $z \in R(Z)$. Let V be an open set 224
containing z such that $0 \notin \bar{V}$ and V has compact closure. Suppose 225

$$\bar{V} \cap \mathcal{J} \cdot R = \phi. \quad (4.12)$$

Then for any $x \in \mathcal{J}$, $x^{-1}\bar{V}$ is compact and $x^{-1}\bar{V} \cap R = \phi$. This means

$$\sum_{n=1}^{\infty} \mu^n(x^{-1}\bar{V}) < \infty. \quad (4.13)$$

Let W be an open set containing z such that $\bar{W} \subset V$. Then for each $x \in \mathcal{J}$ and 227
any $y \in \bar{W}$, given any open set $N(y)$ containing y , there exist open sets $N(x)$ 228
containing x and $N_x(y)$ containing y such that 229

$$N(x)^{-1}N_x(y) \subset x^{-1}N(y). \quad (4.14)$$

It follows from (4.13) and (4.14) that given any compact set $A \subset \mathcal{J}$,

$$\sum_{n=1}^{\infty} \mu^n(A^{-1}W) < \infty. \quad (4.15)$$

Notice that $z \in R(Z)$ and

$$\lim_{n \rightarrow \infty} \mu^n(\mathcal{J}) = 1,$$

where $\mathcal{J}$ is an open ideal of S . The reason is the following:

232

$$\begin{aligned}\mu^{n+1}(\mathcal{J}^c) &= \int_{\mathcal{J}^c} \mu^n(\mathcal{J}^c x^{-1}) \mu(dx) \\ &\leq \mu^n(\mathcal{J}^c) \mu(\mathcal{J}^c) \leq [\mu(\mathcal{J}^c)]^n,\end{aligned}$$

and therefore,

233

$$\sum_{n=1}^{\infty} \mu^n(\mathcal{J}^c) < \infty,$$

since some $k \geq 1$, $\mu^k(\mathcal{J}) > 0$. Thus, there is a compact set $A \subset \mathcal{J}$ such that

234

$$P(Z_k \in A, Z_n \in W \text{ infinitely often}) > 0. \quad (4.16)$$

However, (4.16) implies that

235

$$\begin{aligned}0 &< P(Z_k \in A, Z_k \cdot X_{k+1} \dots X_n \in W \text{ infinitely often}) \\ &\leq P(X_{k+1} \dots X_n \in A^{-1}W \text{ infinitely often}) \\ &= P(Z_n \in A^{-1}W \text{ infinitely often}),\end{aligned}$$

which contradicts (4.15). Thus, (4.12) is impossible, and there exist elements $y_1 \in \mathcal{J}$, $y_2 \in R$ such that

236

237

$$y_1 \cdot y_2 \in \overline{V}.$$

Since $0 \notin \overline{V}$, $y_2 \neq 0$ and $y_1 \cdot y_2 \in J_r^c \cap R$. This establishes (4.11). Notice that the preceding proof also demonstrates

238

239

$$R(Z) \subset \text{cl}(R \cap J_r^c) \subset R(W), \quad (4.17)$$

and similarly,

240

$$R(W) \subset \text{cl}(R \cap J_c^c) \subset R(Z). \quad (4.18)$$

Therefore, $R(Z) = R(W)$. Since $x \in m(S - \{0\})$, $\mathcal{J}x \subset m(S \cap J_r^c)$, it is clear that

241

242

$$m(S \cap J_r^c) = m(S - \{0\}) \cap J_r^c \subset m(S - \{0\}),$$

Therefore,

243

$$R(Z) \subset \text{cl}(R \cap J_r^c) = \text{cl}[m(S \cap J_r^c)] \subset m(S - \{0\}) \cup \{0\}. \quad (4.19)$$

Now we prove the final part of Theorem 4.3, namely,

244

$$R(Z) = m(S - \{0\}) \cup \{0\} = R(W). \quad (4.20)$$

Let $x \in m(S - \{0\})$. Then $Sx \neq \{0\}$. Since

245

$$m(S - \{0\}) \cup \{0\}$$

is completely 0-simple, $xSx - \{0\}$ is a group with an identity e and $Sx = Se$.
Notice that

247

$$\mathcal{J}(R - \{0\}) \mathcal{J} \subset \mathcal{J} \cap R \subset R(Z) \cap R(W).$$

Let $y \in R(Z) \cap \mathcal{J}$. Then $ye \in R \cap J_r^c$. By (4.11), $ye \in R(Z)$ since $R(Z)$ is closed
and $R \cap J_c^c \subset R(Z)$ by Theorem 4.2. Since $R(Z)$ is a left ideal, $S \cdot ye \in R(Z)$.
Since $x \in Sx$ and $Sx = Se = Sye \in R(Z)$, $x \in R(Z)$. Thus,

250

$$R(Z) = m(S - \{0\}) \cup \{0\} [= R(W)].$$

□ 251

To examine the situation for affine maps, let us now consider (as in (4.8)) the
 $(d + 1) \times (d + 1)$ random nonnegative matrix

253

$$X_n \equiv \begin{pmatrix} A_n & B_n \\ 0 & 1 \end{pmatrix}, \quad (4.21)$$

where $X_0, X_1, \dots$ are i.i.d. with distribution μ . Then

254

$$Z_n \equiv X_0 X_1 \dots X_n = \begin{pmatrix} A_0 A_1 \dots A_n & C_n \\ 0 & 1 \end{pmatrix},$$

where $C_n = B_0 + A_0 B_1 + \dots + A_0 A_1 \dots A_{n-1} B_n$. Notice that each C_n is a $d \times 1$
vector and for $1 \leq i \leq d$

256

$$(C_{n+1})_i \geq (C_n)_i$$

then $\lim_{n \rightarrow \infty} (C_n)_i$ exists but is possibly infinite. We also have

257

$$W_n \equiv X_n X_{n-1} \dots X_0 = \begin{pmatrix} A_n A_{n-1} \dots A_0 & D_n \\ 0 & 1 \end{pmatrix},$$

where

258

$$D_n = A_n A_{n-1} \dots A_1 B_0 + A_n \dots A_2 B_1 + \dots + A_n B_{n-1} + B_n.$$

Notice that for each $n \geq 0$,

259

$$P(C_n \in \mathcal{B}) = P(D_n \in \mathcal{B}). \quad (4.22)$$

Furthermore if $R(Z) \neq \phi$, then there exists $u > 0$ such that

260

$$P((C_n)_i \leq u \text{ for } 1 \leq i \leq d \text{ eventually}) > 0.$$

This implies that there exists $\delta > 0$ and a positive integer N such that

261

$$n \geq N \Rightarrow \mu^n \{x \in S : x_{i,d+1} \leq u \text{ for } 1 \leq i \leq d\} > \delta. \quad (4.23)$$

Note that $R(W) \neq \phi$ does not immediately lead to an inequality of the form (4.23), since, unlike C_n s, D_n s cannot be assumed to be an increasing sequence. However, if W_n s are assumed to have a positive recurrent state, that is, when

262

263

264

$$\lim_{n \rightarrow \infty} \sup P(D_n \in \mathcal{B}) > 0$$

for some bounded open set B , we must then also have by (4.22),

265

$$\lim_{n \rightarrow \infty} \sup P(C_n \in \mathcal{B}) > 0. \quad (4.24)$$

This implies an inequality of the form (4.23).

266

At this time, it is not known exactly when $R(W) \neq \phi$ implies $R(Z) \neq \phi$. However, when the rank of the matrices in $m(S)$, where S is the closed multiplicative semigroup generated by S_μ is greater than 1, then $R \neq \phi$ implies (and therefore $R(W) \neq \phi$ also implies) that $R(Z) \neq \phi$ if S contains at least one matrix of the form

267

268

269

270

271

$$\begin{pmatrix} A & B \\ 0 & 1 \end{pmatrix},$$

where A is strictly positive. The reason is if we take

272

$$\begin{pmatrix} C & D \\ 0 & 1 \end{pmatrix}$$

in R , then $C \neq 0$, and consequently if A is strictly positive, then

273

$$\begin{pmatrix} C & D \\ 0 & 1 \end{pmatrix} \begin{pmatrix} A & B \\ 0 & 1 \end{pmatrix} \in R \cap J_c^c = R(Z) \cap J_c^c.$$

In what follows, we present a number of interesting results for affine maps, namely, that under very mild conditions, $R(Z) \neq \phi$ (in fact an assumption even weaker than this) implies the tightness of the sequence (μ^n) and therefore the existence of a μ -invariant probability measure ν where

274

275

276

277

$$\nu = (\text{weak}) \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \mu^k,$$

and $S(\nu) = m(S)$. These results will also help us identify the sets $R(Z)$ and $R(W)$.

278

Theorem 4.4. Let μ be a Borel probability measure on $(d+1) \times (d+1)$ nonnegative matrices of the form (4.8). Let S be the closed (with usual topology) multiplicative semigroup of such matrices generated by S_μ . Suppose the following three conditions hold:

279

280

281

282

(i) There exist $q > 0$ and a subset B_z with $z > 0$, defined by 283

$$B_z = \{x \in S : x_{i,d+1} \leq z \text{ for } 1 \leq i \leq d+1\},$$

such that for some subsequence (n_k) of positive integers $\mu^{n_k}(B_z) > q$, $k \geq 1$; 284

(ii) The set $\mathcal{J}_0$ defined by 285

$$\mathcal{J}_0 = \{x \in S : \text{for each } i, 1 \leq i \leq d, x_{i,d+1} > 0\}$$

is nonempty; and 286

(iii) $\mu(J) = 0$. 287

Then the sequence (μ^n) is tight. 288

Proof. We defer the proof to Sect. 4.3, Theorem 4.13. 289

Theorem 4.5. Let $X_0, X_1, \dots$ be a sequence of $(d+1) \times (d+1)$ i.i.d. (random) 290
nonnegative matrices of the form given in (4.8). Suppose μ is the distribution of 291
 X_0 and S is the closed (with usual topology) multiplicative semigroup generated 292
by $S(\mu)$, the support of μ . Let J_c (respectively, J_r) be the set of all matrices in S 293
with at least one zero column (respectively, zero row) and $J = J_c \cup J_r$. Let us also 294
define the set I by 295

$$I = \left\{ x \in S : \max \left\{ \min_{1 \leq j \leq d} x_{ij}, x_{i,d+1} \right\} > 0 \text{ for each } i, 1 \leq i \leq d \right\}. \quad (4.25)$$

Assume 296

$$\mu(J) = 0, \quad I \neq \phi \quad (4.26)$$

and there exists in S a matrix 297

$$\begin{pmatrix} A & B \\ 0 & 1 \end{pmatrix}$$

for which $B \neq 0$. (Without this last assumption, the situation is exactly as in 298
Theorem 4.3.) Let 299

$$Z_n \equiv X_0 X_1 \dots X_n \text{ and } W_n \equiv X_n X_{n-1} \dots X_0.$$

Suppose $R(Z) \neq \phi$. Then every point in $m(S)$ (matrices in S with the minimal 300
rank) is positive recurrent with respect to both (Z_n) and (W_n) , and 301

$$R(Z) = m(S) \subset R(W).$$

Moreover, the sequence (μ^n) is tight, and there is a unique μ -invariant probability 302
measure ν on S such that $\mu * \nu = \nu * \mu = \mu$ and $S(\nu) = m(S)$, which is a 303
completely simple subsemigroup of S . 304

Proof. Suppose $R(Z) \neq \phi$. Suppose also

305

$$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \in S.$$

Let us call this special element w . Notice that $R(Z)$ is a left ideal of S , and

306

$$w \cdot R(Z) = \{w\}.$$

This means that if $R(Z) \neq \phi$, then $w \in R(Z)$. Let

307

$$\begin{pmatrix} A & B \\ 0 & 1 \end{pmatrix} \in S, \quad B \neq 0.$$

Call this element w_1 . Then $w_1 w \neq w$ and $w_1 w \in R(Z)$. If $x \in I$ and $y \in S$ with $y \neq w$, then $xy \in J_r^c$. Thus, $R(Z) \cap J_r^c$ is a nonempty set containing $I \cdot (w_1 w)$.

308

309

Notice that when $w \in S$, $w \in m(S)$, and $w_1 w \in m(S)$. In other words, $m(S)$ always contains an element other than w . Since

310

311

$$I [m(S) - \{w\}] \subset m(S) \cap J_r^c,$$

it follows from Theorem 4.1 that

312

$$m(S \cap J_r^c) = m(S) \cap J_r^c (= R(W) \cap J_r^c).$$

Let us now show that $R(Z) \subset m(S)$. Let $x \in R(Z) - m(S)$ if possible. Since $m(S)$ is closed, there is an open set V containing x such that $\bar{V}$ is compact and $\bar{V} \cap m(S) = \phi$. Exactly as in the proof of Theorem 4.3, we then have

313

314

315

$$\bar{V} \cap IR \neq \phi. \quad (4.27)$$

(To establish this, we need to observe only that $\mu^n(J) = 0$ for all $n \geq 1$. We can verify that I is a nonempty open ideal of $S - J$, and as such, $\lim_{n \rightarrow \infty} \mu^n(I) = 1$.)

316

317

Then it follows from (4.27) that there exists $z_1 \in I$, $z_2 \in R$ such that $z_1 z_2 \in \bar{V}$. Since $\bar{V} \cap m(S) = \phi$, z_2 must be different from w . This means

318

319

$$z_1 z_2 \in J_r^c \cap R \cap \bar{V}.$$

In other words,

320

$$R(Z) - m(S) \subset \text{cl}[R \cap J_r^c] = \text{cl}[m(S) \cap J_r^c] \subset m(S),$$

which is a contradiction. This proves

321

$$R(Z) \subset m(S). \quad (4.28)$$

By Lemma 4.2, $m(S)$ is a completely simple subsemigroup of S . Now we claim that the rank of the matrices in $m(S)$ is 1. Suppose this rank is greater than 1. In this case when this rank is greater than 1, we can verify that

$$I \cdot m(S) \cdot I \subset m(S) \cap I. \quad (4.29)$$

Thus, $m(S) \cap I \neq \emptyset$. Let $x \in m(S) \cap I$. Since $m(S)$ is a completely simple subsemigroup, $x \in xm(S)x$ and $xm(S)x$ is a group with an identity. Let e be this identity. By (4.29),

$$e \in e \cdot m(S) \cdot e = x \cdot m(S) \cdot x \subset I \cap m(S)$$

and since e is an idempotent, it must then take the form

$$\begin{pmatrix} A_0 & 0 \\ 0 & 1 \end{pmatrix},$$

where A_0 is a strictly positive $d \times d$ matrix. The reason is the following: If for some i , $1 \leq i \leq d$, $e_{i,d+1} > 0$, then for some other j , $1 \leq j \leq d$, $e_{j,d+1} = 0$ means that $e_{ji} > 0$ (since e is in I). Since $e = e^2$,

$$e_{j,d+1} \geq e_{ji}e_{i,d+1} > 0.$$

Then, either the $(d+1)$ th column of e is strictly positive, in which case e must have rank 1 (since e is an idempotent), or else the first d elements on the $(d+1)$ th column of e are all zeros. We assumed e has rank greater than 1 and $e \in I$; it follows that e must have the preceding form.

Since there exists $y \in xm(S)x$ such that $xy = e$, it is clear that x must have the form

$$\begin{pmatrix} C & 0 \\ 0 & 1 \end{pmatrix}.$$

Thus, every matrix in $m(S) \cap I$ has this form where $C \neq 0$. But we assumed that there is an element in S of the form

$$\begin{pmatrix} A & B \\ 0 & 1 \end{pmatrix}, \quad B \neq 0.$$

Notice that

$$\begin{pmatrix} A & B \\ 0 & 1 \end{pmatrix} \begin{pmatrix} C & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} AC & B \\ 0 & 1 \end{pmatrix} \equiv y', \text{ say,}$$

which is an element of $m(S)$ (and where AC cannot be zero), if

$$\begin{pmatrix} C & 0 \\ 0 & 1 \end{pmatrix} \in m(S).$$

Notice that taking

342

$$\begin{pmatrix} A_1 & B_1 \\ 0 & 1 \end{pmatrix} \text{ in } I,$$

we see that

343

$$\begin{pmatrix} A_1 & B_1 \\ 0 & 1 \end{pmatrix} \cdot y' \cdot \begin{pmatrix} A_1 & B_1 \\ 0 & 1 \end{pmatrix} \in m(S) \cap I$$

and this element is not of the form

344

$$\begin{pmatrix} C & 0 \\ 0 & 1 \end{pmatrix},$$

since the matrix B in y' is nonzero. This contradicts the form of the elements in $m(S)$. Thus, we have proven that the rank of the matrices in $m(S)$ is 1 and consequently every matrix in $m(S)$ has the form

345
346
347

$$\begin{pmatrix} 0 & B \\ 0 & 1 \end{pmatrix}.$$

Thus for $x, y \in m(S)$, $xy = x$, and consequently since $R(Z)$ is a left ideal, it follows from (4.28) that $R(Z) = m(S)$.

348
349

Now as we pointed out earlier in (4.23), $R(Z) \neq \phi$ implies that Condition (i) in Theorem 4.4 holds. Also notice that the element $w_1 w$ (considered earlier in the proof) $\neq w$, $w_1 w \in m(S)$, and

350
351
352

$$I \cdot w_1 w \cdot I \subset I \cap m(S).$$

Since the matrices in $m(S)$ have the form

353

$$\begin{pmatrix} 0 & B \\ 0 & 1 \end{pmatrix},$$

it is clear that matrices in $I \cap m(S)$ have the same form

354

$$\begin{pmatrix} 0 & B \\ 0 & 1 \end{pmatrix}$$

with B strictly positive. Consequently, Condition (ii) in Theorem 4.4 holds also. It follows then that (μ^n) is tight. Thus, the weak limit

355
356

$$\nu = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \mu^k$$

exists, and we also have:

357

$$u * v = v * \mu = v \text{ and } S(v) = m(S).$$

(See Theorem 2.7.)

□ 358

Remark 4.2. Under the assumptions of Theorem 4.5, for any bounded, real, continuous function f on $(R^+)^d$ and any u in $(R^+)^d$, 359
360

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n f(W_k(u)) = \int f d\lambda$$

almost surely, where $\lambda = \nu * \delta_u$. (The proof is deferred to Sect.4.4, where we present a law of large numbers for the more general mixed random walks.) 361
362

Now we present conditions when $R(Z) \neq \phi$ actually implies $R(Z) = R(W) = m(S)$. 363
364

Theorem 4.6. Let Z_n, W_n, S, μ be as in Theorem 4.5. Let us replace the sets I, J_c , and J_r in Theorem 4.5 by the sets I^*, J_c^* and J_r^* , respectively, which are defined as follows: 365
366
367

$$\begin{aligned} I^* &= \left\{ \begin{pmatrix} A & B \\ 0 & 1 \end{pmatrix} \in S : A \text{ is strictly positive} \right\}, \\ J_c^* &= \left\{ \begin{pmatrix} A & B \\ 0 & 1 \end{pmatrix} \in S : A \text{ has at least one zero column} \right\} (\equiv J_c), \\ J_r^* &= \left\{ \begin{pmatrix} A & B \\ 0 & 1 \end{pmatrix} \in S : A \text{ has at least one zero row} \right\} (\supset J_r). \end{aligned}$$

Suppose there exists $\begin{pmatrix} A & B \\ 0 & 1 \end{pmatrix} \in S$ with $B \neq 0$ and 368

$$I^* \neq \phi, \quad \mu(J_c^* \cup J_r^*) = 0. \quad (4.30)$$

Then

369

$$R(Z) \neq \phi \Rightarrow R(Z) = R(W) = m(S).$$

Proof. Suppose $R(Z) \neq \phi$. Then as in the proof of Theorem 4.5, $R(Z) \cap J_r^c$ is a nonempty subset of $m(S)$, and thus, using Theorem 4.2, we have: 370
371

$$R(W) \cap m(S) \neq \phi.$$

It also follows that $m(S)$ is a completely simple subsemigroup of S .

372

Since I^* is an open ideal of $S - (J_c^* \cup J_r^*)$, it follows from arguments used earlier that

$$\lim_{n \rightarrow \infty} \mu^n(I^*) = 1. \quad (4.31)$$

If the rank of the matrices in $m(S)$ is more than 1, there does not exist an

$$\begin{pmatrix} A & B \\ 0 & 1 \end{pmatrix}$$

in S for which $A = 0$. Then we can easily verify that

$$I^* \cdot m(S) \cdot I^* \subset I^* \cap m(S).$$

Using this inequality instead of (4.29), it follows easily as in the proof of Theorem 4.5 that the rank of the matrices in $m(S)$ must be 1 and also

$$R(Z) = m(S) \subset R(W). \quad (4.32)$$

Now let $x \in R(W)$, $x \notin m(S)$. Again following an argument similar to the one used earlier and using (4.31), we can prove that for any open set V containing x such that

$$\overline{V} \cap m(S) = \phi,$$

we must have

$$\overline{V} \cap RI^* \neq \phi.$$

Thus, there exist $y \in R$, $z \in I^*$ such that $yz \in \overline{V}$. Since

$$\overline{V} \cap m(S) = \phi,$$

y does not have the form

$$\begin{pmatrix} 0 & B \\ 0 & 1 \end{pmatrix}.$$

Therefore, as is easily verified,

$$yz \in J_c^c \cap R$$

(since R is an ideal $y \in R$ and $z \in I^*$). It follows that

$$\overline{V} \cap (R \cap J_c^c)$$

is nonempty. This means that

$$x \in \text{cl}[R \cap J_c^c] = \text{cl}[R(Z) \cap J_c^c] \subset R(Z).$$

This contradicts (4.32).

□

Let us now consider the sets $R(Z)$ and $R(W)$ in the special case when μ is a probability measure on the set of nonnegative 2×2 matrices of the form

$$\begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix}.$$

As before, let $X_0, X_1, \dots$ be a sequence of i.i.d. nonnegative random matrices of the form

$$\begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix}.$$

Let

$$\begin{aligned} Z_n &\equiv X_0 X_1 \dots X_n, \\ W_n &\equiv X_n \dots X_1 X_0. \end{aligned}$$

Write:

$$X_n \equiv \begin{pmatrix} \xi_n & \eta_n \\ 0 & 1 \end{pmatrix}.$$

Let μ be the distribution of X_0 and S the closed (with usual topology) multiplicative semigroup generated by $S(\mu)$, the support of μ .

The case when

$$S \subset \left\{ \begin{pmatrix} a & 0 \\ 0 & 1 \end{pmatrix} : a \geq 0 \right\}$$

can easily be taken care of, since S can then be regarded as a closed multiplicative semigroup of $[0, \infty)$. When

$$S \subset \left\{ \begin{pmatrix} 0 & b \\ 0 & 1 \end{pmatrix} : b \geq 0 \right\},$$

it is then immediate that

$$S = S(\mu) = R(Z) = R(W),$$

since then for $x, y \in S$, $x \cdot y = x$.

Let us now consider the case when

$$S \cap \left\{ \begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix} : a > 0, b > 0 \right\} \neq \emptyset. \quad (4.33)$$

Let J_r (respectively, J_c) as usual be the set of those matrices in S that have at least one zero row (respectively, one zero column). Let $J_c \cup J_r \equiv J$.

Suppose first that

$$\mu(J) > 0. \quad (4.34)$$

Then we observe that

406

$$J = \left\{ \begin{pmatrix} 0 & b \\ 0 & 1 \end{pmatrix} \in S \mid b \geq 0 \right\},$$

$$\sum_{n=1}^{\infty} \mu^n(J^c) < \infty.$$

If K is compact, $K \subset J$ and $\mu(K) > 0$, then since $K \cdot S = K$, it follows that for all positive integers n ,

407
408

$$\mu^n(K) \geq \mu(K) > 0.$$

It also follows that

409

$$R = R(Z) = R(W) = J$$

and every point in J is positive recurrent.

410

Now suppose

411

$$\mu(J) = 0. \quad (4.35)$$

If $J = \phi$, then

412

$$S \subset \left\{ \begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix} : a > 0, b \geq 0 \right\}.$$

In this case, $m(S) = S$ and by Theorem 4.2,

413

$$R = R(Z) = R(W).$$

If $R \neq \phi$, $m(S) = S$ is a completely simple semigroup that must be a group (since S is cancellative). In this case, the inverse of

414
415

$$\begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix}$$

can exist only when $b = 0$, and then it is

416

$$\begin{pmatrix} 1/a & 0 \\ 0 & 1 \end{pmatrix}.$$

In other words, since S is closed, S must be a singleton (the identity matrix).

417

Let us then assume (4.33), and also, let

418

$$\mu(J) = 0 \text{ and } J \neq \phi. \quad (4.36)$$

In this case,

419

$$m(S) = \left\{ \begin{pmatrix} 0 & b \\ 0 & 1 \end{pmatrix} \in S : b \geq 0 \right\},$$

and by Theorem 4.5,

$$R(Z) \neq \phi \Rightarrow R(Z) = m(S). \quad (4.37)$$

Suppose now that (4.33) and (4.36) hold and also

$$R \neq \phi. \quad (4.38)$$

Since by Theorem 4.1,

$$R \cap J_c^c = R(Z) \cap J_c^c, \quad (4.22)$$

it is clear from (4.37) that $R \subset m(S)$. Since R is an ideal, for x, y in $m(S)$,

$$x \cdot y = x, \quad R = m(S).$$

If

$$m(S) \neq \left\{ \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \right\}, \quad (4.24)$$

then $R \cap J_r^c \neq \phi$. By Theorem 4.1,

$$R \cap J_r^c = R(W) \cap J_r^c, \quad (4.25)$$

so that

$$m(S) \supset R(W) \supset m(S) - \left\{ \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \right\}. \quad (4.39) \quad (4.26)$$

Note that by assumption (4.33), there exist $a > 0, b > 0$ such that

$$\begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix} \in S,$$

Since

$$\begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & b \\ 0 & 1 \end{pmatrix}, \quad (4.28)$$

it is clear that

$$m(S) \neq \left\{ \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \right\} \quad (4.29)$$

so that $R(W) \neq \phi$. We now claim that under (4.33), (4.36), and (4.38),

$$R = R(W) = m(S). \quad (4.40)$$

To prove (4.40), we need to show only that when

$$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \in S, \quad (4.31)$$

then it must be in $R(W)$. Let us write:

432

$$w \equiv \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad w_n \equiv \begin{pmatrix} a_n & b_n \\ 0 & 1 \end{pmatrix}.$$

First suppose w is not an isolated point in S . Then there exist $w_n \in S$ such that 433
 $(a_n, b_n) \neq (0, 0)$ and $w_n \rightarrow w$ as $n \rightarrow \infty$. Now there exists $b > 0$ such that 434

$$\begin{pmatrix} 0 & b \\ 0 & 1 \end{pmatrix} \in m(S).$$

Since

435

$$\begin{pmatrix} a_n & b_n \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & b \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & a_n b + b_n \\ 0 & 1 \end{pmatrix},$$

and also since $(a_n, b_n) \neq (0, 0)$, $a_n b + b_n \neq 0$, and therefore by (4.39), for 436
each $n \geq 1$, 437

$$\begin{pmatrix} 0 & a_n b + b_n \\ 0 & 1 \end{pmatrix} \in R(W).$$

Since $R(W)$ is closed and $a_n b + b_n \rightarrow 0$, it follows that $w \in R(W)$. Now suppose 438
 w is an isolated point in S . Now $w \cdot S = \{w\}$; therefore, there is a positive integer k 439
such that 440

$$\mu^k(\{w\}) > 0 \text{ and } \mu^n(\{w\}) \geq \mu^k(\{w\}) \text{ for } n \geq k.$$

It follows that $w \in R(W)$. This proves (4.40).

441

Recall that

442

$$X_n \equiv \begin{pmatrix} \xi_n & \eta_n \\ 0 & 1 \end{pmatrix}.$$

Let us now mention that the assumptions

443

$$P(\xi_0 = 0) = 0, \quad (4.41)$$

$$-\infty \leq E(\log \xi_0) < 0, \quad (4.42)$$

$$E \log \max \{\eta_0, 1\} < \infty, \quad (4.43)$$

imply

444

$$R(Z) \neq \phi, \quad R(Z) = m(S) \text{ and } (\mu^n) \text{ is tight.} \quad (4.44)$$

Also, under (4.41),

445

$$0 < E(\log \xi_0) \leq \infty \Rightarrow R(Z) = R(W) = \phi, \quad (4.45)$$

and

446

$$E(\log \xi_0) > -\infty \text{ and } R(Z) \neq \phi \Rightarrow E \log \max \{\eta_0, 1\} < \infty. \quad (4.46)$$

The proof of (4.45) is immediate by the law of large numbers, since $E(\log \xi_0) > 0$ 447
implies 448

$$P\left(\lim_{n \rightarrow \infty} \xi_0 \xi_1 \dots \xi_n = \infty\right) = 1.$$

Proofs of assertions (4.44) and (4.46) are deferred until Sect. 4.3, where they are 449
part of Theorem 4.17. It is also relevant to point out (and very easy to show) that if 450

$$E(\log \xi_0) = 0, P(\xi_0 = 1) < 1, \text{ and}$$

$$P(\eta_0 \geq \varepsilon) = 1 \text{ for some } \varepsilon > 0,$$

then $R(Z) = \phi$. 451

In the rest of Sect. 4.2, we look briefly into the connection of the set $R(W)$ with 452
what has been called an attractor in the literature. 453

Let $\mathcal{F}$ be a family of affine maps from $(R^+)^d$ into $(R^+)^d$ such that each $f \in \mathcal{F}$ 454
can be represented as 455

$$\begin{pmatrix} f(x) \\ 1 \end{pmatrix} = \begin{pmatrix} A_f & B_f \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ 1 \end{pmatrix}.$$

where $x \in (R^+)^d$, and the nonnegative $(d+1) \times (d+1)$ matrix on the right has 456
the form of (4.8). Notice that for f and g in $\mathcal{F}$, 457

$$\begin{pmatrix} f \circ g(x) \\ 1 \end{pmatrix} = \left[\begin{pmatrix} A_f & B_f \\ 0 & 1 \end{pmatrix} \begin{pmatrix} A_g & B_g \\ 0 & 1 \end{pmatrix} \right] \begin{pmatrix} x \\ 1 \end{pmatrix}.$$

This means that if (X_n) , $n \geq 0$, is a sequence of i.i.d. $\mathcal{F}$ -valued random variables 458
with common distribution μ and $W_n = X_n X_{n-1} \dots X_0$, then the sequence (X_n) 459
can be considered a sequence of i.i.d. matrices of the form of (4.8) and the W_n s as 460
products of such matrices. 461

Let us now define the attractor $\mathcal{A}(u)$, $u \in (R^+)^d$, as the set 463

$$\mathcal{A}(u) = \left\{ y \in (R^+)^d : \text{for every open } N_y \text{ containing } y, \right. \\ \left. P(W_n \cdot u \in N_y \text{ infinitely often}) > 0 \right\},$$

where $W_n \cdot u = W_n(u)$ and W_n is the affine map $X_n X_{n-1} \dots X_0$. Using matrix 464
notation, we can then write 465

$$\mathcal{A}(u) = \left\{ y \in (R^+)^d : \text{for every open } N(y^*) \text{ containing } y^* = \begin{pmatrix} y \\ 1 \end{pmatrix}, \right. \\ \left. P\left(W_n \cdot \begin{pmatrix} u \\ 1 \end{pmatrix} \in N(y^*) \text{ infinitely often}\right) > 0 \right\},$$

where W_n is the product of $X_n \dots X_0$, X_i s now regarded as nonnegative $(d+1) \times$ 466
 $(d+1)$ matrices, as mentioned. 467

The next result is useful in identifying the attractor as the support of a μ -invariant probability measure, as we show in Theorem 4.7. 468 469

Lemma 4.3. Let (X_n) be a sequence of $d \times d$ i.i.d. nonnegative matrices with common distribution μ . Then for $x \in (R^+)^d$, $x > 0$ (that is, every entry of x is positive), 470 471

$$\mathcal{A}(x) = R(W) \cdot x (\equiv \{C \cdot x : C \in R(W)\}),$$

where $R(W)$ is the recurrent set as defined earlier. 472

Proof. Let $x > 0$, $C \in R(W)$, and $y = Cx$. Notice that given an open set $N(y)$ 473 containing y , there exists an open set $N(C)$ containing C such that $N(C)x \subset N(y)$. 474 This means $y \in \mathcal{A}(x)$. For the converse, let $y \in \mathcal{A}(x)$. Let $N_n(y)$, $1 \leq n < \infty$, be 475 a countable open (and relatively compact) local base at y such that 476

$$\text{cl}(N_{n+1}(y)) \subset N_n(y) \quad \forall n.$$

For each n , define the set 477

$$U(N_n(y)) = \{D : D \text{ is a } d \times d \text{ nonnegative matrix and } Dx \in N_n(y)\}.$$

Then $U(N_n(y))$ is an open set; since $x > 0$, it can be verified that $U(N_n(y))$ is 478 relatively compact. Now notice that for each positive integer m , 479

$$\Pr(W_n \in \text{cl}(U(N_m(y))) \text{ infinitely often}) > 0.$$

Since $U(N_m(y))$ is relatively compact, there exists for each m a matrix C_m in 480 $R(W)$ such that 481

$$C_m \in \text{cl}(U(N_m(y))).$$

Since for each m , 482

$$C_{m+1}x \subset \text{cl}(N_m(y)) \subset N_1(y),$$

it follows that the C_m s are bounded (since $x > 0$). If C_0 is a limit point of these 483 C_m s, then $C_0 \in R(W)$, since $R(W)$ is a closed set, and $C_0x = y$. $\square$ 484

Theorem 4.7 identifies the attractor as the support of a μ -invariant probability measure. 485 486

Theorem 4.7. Let μ , S , (Z_n) , (W_n) , I , J_c , J_r be as in Theorem 4.5. Let ν be the 487 μ -invariant probability measure in Theorem 4.5. Consider the hypothesis of this 488 theorem. Then for $u \in (R^+)^d$, $u > 0$, the attractor $\mathcal{A}(u)$, defined by 489

$$\mathcal{A}(u) = \left\{ y \in (R^+)^d : \text{for every open set } N(y^*) \text{ containing } y^*, \right. \\ \left. P(W_n u^* \in N(y^*) \text{ infinitely often}) \text{ is positive} \right\},$$

where

491

$$y^* = \begin{pmatrix} y \\ 1 \end{pmatrix}, \quad u^* = \begin{pmatrix} u \\ 1 \end{pmatrix},$$

turns out to be given by the set $A \equiv \{y \in (R^+)^d : y^* \in S(v) \cdot u^*\}$, except for a set 492
with d -dimensional Lebesgue measure zero. With the hypothesis of Theorem 4.6, the 493
attractor $\mathcal{A}(u)$ is actually equal to the set A , which is the support of the μ -invariant 494
probability measure $v * \delta_u$, where δ_u is the unit mass at u . 495

Proof. Note that by Lemma 4.3, for $u \in (R^+)^d$, $u > 0$, 496

$$\{y^* : y \in \mathcal{A}(u)\} = \{x \cdot u^* : x \in R(W)\}.$$

Under the hypothesis of Theorem 4.5, 497

$$R(W) \cap J_r^c \subset m(S) = S_v \subset R(W),$$

and under the hypotheses of Theorem 4.6, $R(W) = S_v$, where 498

$$v = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \mu^k,$$

as in Theorem 4.5. The proof therefore follows immediately, since for $x \in R(W)$ 499
 $\cap J_r$, 500

$$x \cdot u^* \in \{y^* : y_i = 0 \text{ for some } i, 1 \leq i \leq d\} \equiv M,$$

and 501

$$\left\{y \in (R^+)^d : y^* \in M\right\}$$

has d dimensional Lebesgue measure zero. $\square$ 502

Now we use the formula given in Lemma 4.3 and Theorems 4.5–4.7 to describe 503
completely what is wellknown as a Sierpinski gasket. We also describe one of its un- 504
bounded versions. Even though the examples below are simple, they well illustrate 505
the concepts. 506

Sierpinski gasket. Consider the three functions on the plane given by 507

$$f(x, y) = \left(\frac{x}{2}, \frac{y}{2}\right), g(x, y) = \left(\frac{x+1}{2}, \frac{y}{2}\right) \text{ and } h(x, y) = \left(\frac{x}{2}, \frac{y+1}{2}\right).$$

Note that the semigroup (with respect to the composition of maps) generated by these three functions can be identified with the multiplicative semigroup generated by matrices (called by the same names, for convenience) given by

$$f = \begin{pmatrix} 1/2 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad g = \begin{pmatrix} 1/2 & 0 & 1/2 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$h = \begin{pmatrix} 1/2 & 0 & 0 \\ 0 & 1/2 & 1/2 \\ 0 & 0 & 1 \end{pmatrix}.$$

Take any (this is quite interesting, though the logic is very simple) distribution μ such that $S(\mu) = \{f, g, h\}$. Then S , the smallest closed (with usual topology for matrices) multiplicative semigroup containing $S(\mu)$ is a compact semigroup. It is easy to see that $m(S)$ consists of all rank 1 matrices in S . Since $R(W) = m(S)$, where S is a compact semigroup (see Chap. 3), it follows by Lemma 4.3 that for any point (x, y) , the attractor set $\mathcal{A}(x, y)$ is given by

$$\mathcal{A} \equiv \mathcal{A}(x, y) = \left((a, b) : \begin{pmatrix} 0 & 0 & a \\ 0 & 0 & b \\ 0 & 0 & 1 \end{pmatrix} \text{ belongs to } m(S) \right).$$

Recall that $m(S)$ is the set of all those matrices in S that have minimal rank (in this case, rank 1). Now compute $m(S)$ to identify the set $\mathcal{A}$ as the Sierpinski gasket with vertices at $(0, 0)$, $(0, 1)$, and $(1, 0)$.

Since f, g, h are substochastic, it is clear that

$$(a, b) \in \mathcal{A} \Rightarrow \{a, b\} \subset [0, 1].$$

Also notice that for any positive integer m ,

$$f^m \cdot \begin{pmatrix} a \\ b \\ 1 \end{pmatrix} = \begin{pmatrix} a/2^m \\ b/2^m \\ 1 \end{pmatrix}, \quad g^m \cdot \begin{pmatrix} a \\ b \\ 1 \end{pmatrix} = \begin{pmatrix} a/2^m + (1/2 + \cdots + 1/2^m) \\ b/2^m \\ 1 \end{pmatrix}$$

and

$$h^m \cdot \begin{pmatrix} a \\ b \\ 1 \end{pmatrix} = \begin{pmatrix} a/2^m \\ b/2^m + (1/2 + \cdots + 1/2^m) \\ 1 \end{pmatrix}.$$

Since $m(S) \cdot m(S) = m(S)$, by the simplicity of $m(S)$ as an ideal, $m(S)$ maps any point $(x, y, 1)$ into the point $(a, b, 1)$, where (a, b) is on or within the triangle Δ bounded by the lines $a = 0$, $b = 0$, and $a + b = 1$. Observe that h maps the points $(x, y, 1)$, where $(x, y) \in \Delta$, to the points $(a, b, 1)$, where (a, b) is on or within the triangle with vertices at $(0, 1/2)$, $(1/2, 1/2)$, and $(0, 1)$. g maps the same

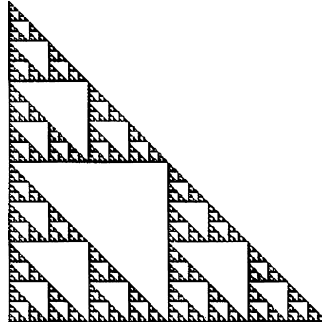


Fig. 4.1

set of points to the points $(a, b, 1)$, where (a, b) is on or within the triangle with 528
vertices at $(1/2, 1/2)$, $(1/2, 0)$, and $(1, 0)$. f maps those points to points $(a, b, 1)$, 529
where (a, b) is on or within the triangle with vertices at $(0, 0)$, $(0, 1/2)$, and $(1/2, 0)$. 530
Since $m(S) = m(S)m(S)m(S)$, it follows that $m(S)$ maps the points in $R^2 \times (1)$ 531
to the points $(a, b, 1)$, where (a, b) is on or inside the triangle Δ but outside the 532
interior of the triangle with vertices at $(0, 1/2)$, $(1/2, 1/2)$, and $(1/2, 0)$. Repeating 533
this argument and noting that $m(S) = [m(S)]^m$ for every positive integer m , the 534
attractor set is easily seen to be the known Sierpinski gasket (see Fig. 4.1) and its 535
graph is the closure of the graph of $E_0 \cup E_1 \cup E_2 \dots$, where $E_0 = (0, 0)$. For $n \geq 0$, 536

$$\begin{aligned} E_{n+1} = & \{(a/2^m, b/2^m + (1/2 + \dots + 1/2^m)) : m \geq 1 \text{ and } (a, b) \in E_n\} \\ & \cup \{(a/2^m + (1/2 + \dots + 1/2^m), b/2^m) : m \geq 1 \text{ and } (a, b) \in E_n\} \\ & \cup \{(a/2^m, b/2^m) : m \geq 1 \text{ and } (a, b) \in E_n\}. \end{aligned}$$

These formulas describe the attractor completely. 537

Unbounded attractor for an expansive iterated function system. Consider the 538
functions (f, g, h, k) from the plane into itself, where the functions f , g , and h 539
are the same as in the example of the Sierpinski gasket and functions k is given by 540
 $k(x, y) = (2x, y)$. 541

Take a probability distribution μ such that $S(\mu) = \{f, g, h, k\}$ and such that the 542
average contractivity (discussed in Example 4.7, condition (4.187) or any other con- 543
dition that implies the tightness of (μ^n) . As before, we identify these functions with 544
the corresponding 3×3 matrices and let S be the smallest closed matrix semigroup 545
(with usual topology and multiplication) containing $S(\mu)$. It is then immediate from 546
the example of the Sierpinski gasket that the set $m(S)$ of all matrices in S with min- 547
imal rank consists of those matrices in S with rank 1, and this set is given by 548

$$m(S) = \left\{ \begin{pmatrix} 0 & 0 & 2^m \cdot a \\ 0 & 0 & b \\ 0 & 0 & 1 \end{pmatrix} : \begin{array}{l} m \geq 0 \text{ and} \\ (a, b) \in \text{the Sierpinski triangle} \\ \text{of the previous example} \end{array} \right\}.$$

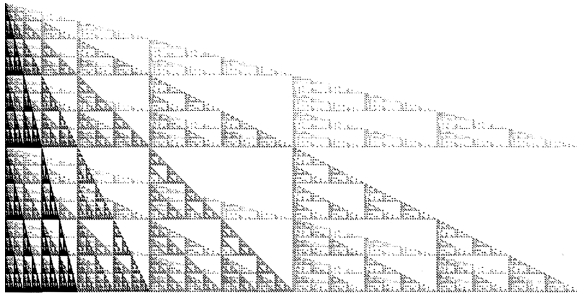


Fig. 4.2

The attractor set $\mathcal{A}(x, y)$ is the set (graphed in Fig.4.2), the graph actually extend- 549
ing to infinity, and given by 550

$$\mathcal{A}(x, y) = \left\{ (2^m \cdot a, b) : \begin{array}{l} m \geq 0 \text{ and} \\ (a, b) \in \text{the Sierpinski triangle} \\ \text{of the previous example} \end{array} \right\}. \quad \square \quad 551$$

Other relevant results for attractors for general metric spaces are outlined in the 552
problems. 553

Section 4.2. Exercises 554

Exercise 4.1. Verify the inclusion (4.29). 555

Exercise 4.2. Why is the set $R(Z)$, in Theorem 4.5, contained in the set $R(W)$? 556

Exercise 4.3. Why does the condition $R(Z) \neq \emptyset$ imply in Theorem 4.6 that 557
 $R(W) \cap m(S) \neq \emptyset$? Explain in details. 558

Exercise 4.4. The statement 559

$$S = S(\mu) = R(Z) = R(W)$$

appears just preceding (4.33). Prove it. 560

Exercise 4.5. Prove that in the presence of the condition (4.33) and (4.34) the fol- 561
lowing inequality holds: $\sum_{n=1}^{\infty} \mu^n(J^c) < \infty$. 562

Exercise 4.6. Prove that under (4.33) and (4.34), the following is true: $R =$ 563
 $R(Z) = R(W) = J$, and also, every point in J is positive recurrent with respect to 564
 (Z_n) as well as (W_n) . 565

Exercise 4.7. Let $X_1, X_2, \dots$ be a sequence of independent random variables with 566
values in the interval $[0, 1]$ such that $P(X_i \in B) = \mu_i(B)$ for any Borel set 567

$B \subset [0, 1]$, where $\mu_1, \mu_2, \dots$ is a sequence of probability measures on the Borel subsets of $[0, 1]$. Write: $X_{k,n} = X_{k+1} \dots X_n$ for $k < n$. Show that $Y_k = \lim_{n \rightarrow \infty} X_{k,n}$ exists almost surely, and so does $Y = \lim_{k \rightarrow \infty} Y_k$. Show also that either $Y = 0$ almost surely or $Y = 1$ almost surely, and that $Y = 1$ almost surely iff $\sum_{n=1}^{\infty} [1 - X_n]$ converges almost surely.

Exercise 4.8. Describe the attractor set $\mathcal{A}(x, y)$ determined by the functions

$$f(x, y) = \left(\frac{x}{2}, \frac{y}{2}\right), \quad g(x, y) = \left(\frac{x}{2}, \frac{y+1}{2}\right), \quad h(x, y) = \left(\frac{x+1}{2}, \frac{y+1}{2}\right).$$

Exercise 4.9. Do the previous exercise replacing the function f in it by the function $f_0(x, y) = \left(\frac{x+1}{2}, \frac{y}{2}\right)$.

Exercise 4.10. Let S be a complete metric space with metric d and $\{f_i : 1 \leq i \leq n\}$ be a finite family of contraction maps from S into itself (that is, there exists $0 < r < 1$ such that $d(f_i(x), f_i(y)) \leq rd(x, y)$ for any x, y in S , $1 \leq i \leq n$). Show that there is a unique compact set $K \subset S$, called the attractor of the contracting system $\{f_i : 1 \leq i \leq n\}$ such that $K = \bigcup_{i=1}^n f_i(K)$.

(Hint: Define for A, B in the family $\mathcal{F}$ of nonempty compact subsets of S the metric ρ by

$$\rho(A, B) = \inf \{k > 0 : A \subset N_k(B), B \subset N_k(A)\},$$

where for $C \in \mathcal{F}$,

$$N_k(C) = \{y \in S : d(y, x) < k \text{ for some } x \in C\}.$$

Show that $(\mathcal{F}, \rho)$ is a complete metric space. This ρ is called the Hausdorff metric. Next, define $\Phi : \mathcal{F} \rightarrow \mathcal{F}$ by $\Phi(A) = \bigcup_{i=1}^n f_i(A)$. Then Φ is a contraction with a unique fixed point K .)

Exercise 4.11. Let (S, d) be a metric space and $\beta > 0$. Define $\mathcal{F}_\beta$ by

$$\mathcal{F}_\beta = \{f : S \rightarrow R \mid \text{for all } x, y \in S, \\ |f(x) - f(y)| \leq d(x, y) \text{ and } |f(x)| \leq \beta\}.$$

Consider the set $P_1(S)$ of all regular (with respect to compact sets from inside) probability measures on the Borel subsets of S . For μ_1, μ_2 in $P_1(S)$, define

$$d_\beta(\mu_1, \mu_2) = \sup \left\{ \left| \int h d\mu_1 - \int h d\mu_2 \right| : h \in \mathcal{F}_\beta \right\}$$

Show that d_β is a metric on $P_1(S)$. This metric is known as the Monge–Kantorovich metric.

Exercise 4.12. In the previous exercise, show that if (μ_n) is a sequence in $P_1(S)$, $\mu \in P_1(S)$ and $\beta > 0$, then $\mu_n \rightarrow \mu$ weakly iff $\int h d\mu_n \rightarrow \int h d\mu$ for each bounded real Lipschitz function h on S (that is, $|h(x) - h(y)| \leq d(x, y)$ for x, y in S and $|h(x)| \leq M$ for all x in S for some $M > 0$) iff $d_\beta(\mu_n, \mu) \rightarrow 0$.

Exercise 4.13. In Exercise 4.11, show that a Cauchy sequence in $(P_1(S), d_\beta)$, when S is complete, is tight. Thus, using Exercise 4.12 and the fact that $B_1(S) = \{\mu \mid \mu \text{ is a nonnegative regular measure, } \mu(S) \text{ not exceeding } 1\}$ is weak*-compact, it follows that every subsequence of a Cauchy sequence in $(P_1(S), d_\beta)$ must have a weakly convergent subsequence, which must also converge in the metric d_β . This means that $(P_1(S), d_\beta)$ is complete for any $\beta > 0$, when S is complete.

Exercise 4.14. Let S be a complete separable metric space, and f_i , $1 \leq i \leq n$, be n contraction maps from S into itself such that $d(f_i(x), f_i(y)) \leq r d(x, y)$, $0 < r < 1$, $1 \leq i \leq n$, for all x, y in S . Let $p_1, p_2, \dots, p_n$ be nonnegative numbers such that $\sum_{i=1}^n p_i = 1$. Then show that there is a unique invariant probability measure μ on the Borel subsets of S such that for all Borel subsets $A \subset S$,

$$\mu(A) = \sum_{i=1}^n p_i \mu(f_i^{-1}(A)).$$

Show also that the support of μ is the compact set K of Exercise 4.13.

(Hint: Let $f_i(x_i) = x_i$, $1 \leq i \leq n$, $s = \max_{1 \leq i \leq n} d(x_i, x_1)$, $t = \frac{(1+r)s+1}{1-r}$. Show that if $u \geq t$, then for each i , $1 \leq i \leq n$, $f_i(B_u(x_1)) \subset B_{u-1}(x_1)$, where

$$B_u(x_1) = \{y \in S : d(y, x_1) \leq u\}.$$

Define $\Phi : P(S) \rightarrow P(S)$ by

$$\Phi(\mu)(A) = \sum_{i=1}^n p_i \mu[f_i^{-1}(A)].$$

Show that for $\beta > t$, for μ_1, μ_2 in B_t ,

$$d_\beta(\Phi(\mu_1), \Phi(\mu_2)) \leq r d_\beta(\mu_1, \mu_2).$$

Exercise 4.15. Let S be a complete metric space and $f_1, f_2, \dots, f_n$ be n contraction maps from S into itself such that $d(f_i(x), f_i(y)) \leq r d(x, y)$ for some r , $0 < r < 1$, for all $x, y \in S$, $1 \leq i \leq n$. Let K be the attractor given as in

Exercise 4.10. Let $X_1, X_2, \dots$ be i.i.d. random variables with values in $\{f_1, f_2, \dots, f_n\}$ such that $P(X_1 = f_i) = p_i$, $1 \leq i \leq n$, $\sum_{i=1}^n p_i = 1$. Let $x_0 \in S$ and $x_n = X_n X_{n-1} \dots X_1(x_0)$. Show that

$$K = \{x \in S \mid P(X_n X_{n-1} \dots X_1(x_0) \in N(x) \text{ i.o.}) = 1 \text{ for each open set } N(x) \text{ containing } x\}.$$

4.3 Tightness of Products of I.I.D. Random Matrices: Weak Convergence

Consider the discrete time Markov process (η_n) on the state space R^d , d finite, where the transition rule is governed by a $d \times d$ random matrix Y with real entries, so that

$$\eta_{n+1} = Y_{n+1} \eta_n, \quad n \geq 0, \quad (4.47)$$

where $Y_1, Y_2, \dots$ are an i.i.d. sequence of $d \times d$ random matrices (which are copies of Y). We are interested in the limiting distribution of η_n as $n \rightarrow \infty$, when it exists.

Let μ be the distribution of Y and β that of η_0 , so that μ is a probability measure on $d \times d$ real matrices and β on R^d . Let S be the closed (in usual matrix topology) multiplicative semigroup generated by the support S_μ of μ ; in other words,

$$S = \text{cl} \left(\bigcup_{n=1}^{\infty} S_\mu^n \right).$$

Assuming the Y_i s are independent of η_0 , the distribution β_n of η_n is given by

$$\beta_n = \mu^n * \beta, \quad n \geq 1. \quad (4.48)$$

Here for any probability measure λ on R^d , the convolution product $\mu * \lambda$ is defined by

$$\mu * \lambda(\mathcal{B}) = \int \mu\{A : Ay \in \mathcal{B}\} \lambda(dy).$$

Note that for μ_1, μ_2 in $P(S)$,

$$\mu_1 * (\mu_2 * \lambda) = (\mu_1 * \mu_2) * \lambda,$$

and the map $(\mu, \lambda) \rightarrow \mu * \lambda$ is continuous with respect to the weak topology. It is also clear that if β = the unit mass at the zero vector, then for each $n \geq 1$, β_n is also (no matter what μ is). In this context, Theorem 4.8 is interesting.

Theorem 4.8. Suppose S contains only matrices with all entries nonnegative and at least one matrix with all entries strictly positive. Let $\beta \in P((R^+)^d)$ such that

is not the unit mass at the zero vector. Then in (4.48), the sequence (β_n) is tight if 638
and only if the sequence (μ^n) is tight. 639

Proof. The “if” part is immediate since 640

$$\beta_n\{A \cdot x \mid A \in \mathcal{B}, x \in \mathcal{C}\} \geq \mu^n(\mathcal{B}) \cdot \beta(\mathcal{C}).$$

For the only if part, let us suppose the sequence $\beta_n = \mu^n * \beta$ is tight. Let $\mathcal{J}_k$ be 641
the set defined by 642

$$\mathcal{J}_k = \left\{ x = (x_1, x_2, \dots, x_d) : x_i > \frac{1}{k}, i = 1, 2, \dots, d \right\}.$$

By the theorem hypotheses, there is a positive integer m such that 643

$$\mu^m(\text{the strictly positive matrices in } S) > 0.$$

Notice that if A is a strictly positive $d \times d$ matrix, $x \in (R^+)^d$ and $x \neq 0$, then 644
 $(Ax)_i > 0$ for $i = 1, 2, \dots, d$. This means there exists $\delta > 0$ such that 645

$$\mu^m * \beta \left\{ x \in (R^+)^d : x_i > 0 \text{ for } i = 1, 2, \dots, d \right\} = 2\delta. \quad (4.49)$$

Then there is a positive integer k such that 646

$$\mu^m * \beta(\mathcal{J}_k) > \delta. \quad (4.50)$$

Notice that if $\mathcal{C}$ is a compact subset of $(R^+)^d$, then the set $\mathcal{C} \mathcal{J}_k^{-1}$ defined by 647

$$\mathcal{C} \mathcal{J}_k^{-1} = \{A \in S : A \cdot x \in \mathcal{C} \text{ for some } x \in \mathcal{J}_k\}$$

is relatively compact. 648

Let us now suppose the sequence (μ^n) is not tight. Then we have 649

$$\sup_{\mathcal{C} \subset (R^+)^d, \mathcal{C} \text{ compact}} \lim_{n \rightarrow \infty} \inf \mu^n(\mathcal{C} \mathcal{J}_k^{-1}) = 1 - 2u < 1. \quad (4.51)$$

It follows from (4.50) and (4.51) that for any compact subset $\mathcal{C} \subset (R^+)^d$, 650

$$\begin{aligned} \mu^{n+m} * \beta(\mathcal{C}) &= \int \mu^n(\mathcal{C} x^{-1}) \mu^m * \beta(dx) \\ &\leq (1-u) \mu^m * \beta(\mathcal{J}_k) + \mu^m * \beta(\mathcal{J}_k^c) \\ &\leq \mu^m * \beta(\mathcal{J}_k) + \mu^m * \beta(\mathcal{J}_k^c) - u \cdot \mu^m * \beta(\mathcal{J}_k) \\ &\leq 1 - u\delta < 1, \end{aligned}$$

for infinitely many n . This contradicts the assumption that $(\mu^n * \beta)$ is a tight 651
sequence. $\square$ 652

Let us now give a convergence theorem for the sequence (β_n) in (4.48). We write δ_0 to denote the unit mass at the zero vector in R^d as well as the unit mass at the zero matrix, that is, the matrix with all entries zeros. (Its meaning should be clear from the context.)

Theorem 4.9. *Let S consist of only $d \times d$ nonnegative matrices (that is, matrices with nonnegative entries). Let $\beta \in P((R^+)^d)$ and suppose (μ^n) is tight. If S contains the zero matrix, then μ^n converges weakly to δ_0 , and consequently β_n converges weakly to δ_0 . When S does not contain the zero matrix and S contains a strictly positive matrix, then S is necessarily compact and the sequence μ^n converges weakly to some $\nu \in P(S)$ such $\nu(\mathcal{J}) > 0$, where $\mathcal{J}$ is the set of all strictly positive matrices in S . When $\beta \neq \delta_0$, the sequence β_n converges weakly to some $\lambda \in P((R^+)^d)$ such that*

$$\lambda \left\{ x \in (R^+)^d : x_i > 0 \text{ for } 1 \leq i \leq d \right\} > 0.$$

If S contains a matrix with rank 1 instead of a strictly positive matrix, it remains true that the sequence μ^n converges weakly.

Proof. Suppose (μ^n) is tight. By Theorem 2.7 (Chap. 2), the sequence

$$\frac{1}{n} \sum_{k=1}^n \mu^k$$

converges weakly to some ν in $P(S)$ and S_ν , the support of ν , is the completely simple minimal ideal of S with a compact group factor. If $0 \in S$, then $S_\nu = \{0\}$ and $\nu = \delta_0$. It follows from Theorem 2.7(iii) that μ^n converges weakly to δ_0 .

Now suppose $0 \notin S$ and S contains a strictly positive matrix. Let ν be as just defined. Then for $n \geq 1$,

$$\mu^n * \nu * \mu^n = \nu. \quad (4.52)$$

Since $\mathcal{J} \supset \mathcal{J} \cdot S \cdot \mathcal{J} \equiv \{A_1 A_2 A_3 : A_1 \text{ and } A_3 \text{ belong to } \mathcal{J} \text{ and } A_2 \in S\}$, it follows from (4.52) that

$$\nu(\mathcal{J}) \geq \mu^n(\mathcal{J})\nu(S)\mu^n(\mathcal{J}), \quad n \geq 1$$

so that $\nu(\mathcal{J}) > 0$. Thus, $S_\nu \cap \mathcal{J} \neq \emptyset$. Let $x \in S_\nu \cap \mathcal{J}$. Since S_ν is completely simple, the set $x \cdot S_\nu \cdot x$ is a compact group (see Theorems 1.7 and 2.7(i)). Let e be the identity of this group. Since $x \in \mathcal{J}$ and $0 \notin S_\nu$, it follows (by direct multiplication of matrices) that

$$e \in x S_\nu x \subset \mathcal{J}.$$

Since $eSe = eS_v e = xS_v x$ is a compact group of strictly positive matrices, it is a singleton set (see Corollary 1.8). Thus, there exists $M > 0$ such that if $A \in S$ and $1 \leq i \leq d, 1 \leq j \leq d$, then

$$e_{ii} A_{ij} e_{jj} \leq (eAe)_{ij} \leq M.$$

Since $e \in \mathcal{J}$, it follows that S is compact. It follows from Theorem 2.7 that μ^n converges weakly to ν (since eSe is a singleton set). To prove the second part, observe that when $0 \notin S$ and S contains a rank 1 matrix, then S_v (the minimal ideal of S) consists of all rank 1 matrices in S (see Proposition 1.9). Then eSe , a compact group of nonnegative rank 1 matrices, is necessarily a singleton. The convergence of μ^n then follows from Theorem 2.7. $\square$

Theorem 4.10 (essentially due to Kesten and Spitzer [128]) deals with the problem of weak convergence of (μ^n) .

Consider the following condition:

$$\mu(J) = 0, \quad (4.53)$$

where $J = J_c \cup J_r$, J_c (respectively, J_r) is the set of all $d \times d$ matrices in S with at least one column (respectively, row) containing only zeros. Since (4.53) implies

$$\mu^{+1}(J^c) = \int_{J^c} \mu^n(J^c x^{-1}) \mu(dx) \geq \mu^n(J^c) \mu(J^c),$$

it follows from (4.53) that for $n \geq 1$,

$$\mu^n(J) = 0. \quad (4.54)$$

Also if S contains only nonnegative matrices and $S \cap \mathcal{J} \neq \emptyset$, where $\mathcal{J}$ is the set of all strictly positive matrices, then since $\mathcal{J}$ is an ideal of J^c , it follows from (4.53), (4.54), and Proposition 2.7 that

$$\mu^n(\mathcal{J}) \uparrow 1. \quad (4.55)$$

Note that $\mathcal{J}$ is an open set in S and thus, for some positive integer $m \geq 1$, $\mu^m(\mathcal{J}) > 0$.

Theorem 4.10. Let μ be a probability measure on nonnegative $d \times d$ matrices and S the closed multiplicative semigroup generated by S_μ . Suppose $S \cap \mathcal{J} \neq \emptyset$ and (4.53) holds. Then the following statements are equivalent:

- (i) $\mu^n \rightarrow \nu \in P(S)$ weakly and $S_v \cap J^c \neq \emptyset$.
- (ii) For each $u > 0$, there exists $k_u > 0$ such that for each positive integer n , $\mu^n(\mathcal{B}_u) > 1 - u$, where

$$\mathcal{B}_u = \left\{ A \in S : \min_{i,j} A_{ij} \leq k_u \right\},$$

and there exists $v > 0$ and a subsequence (n_t) such that for $t \geq 1$, $\mu^{n_t}(\mathcal{A}_v) > v$, where

$$\mathcal{A}_v = \left\{ B \in S : \max_{i,j} B_{ij} > v \right\}.$$

(iii) (μ^n) is tight and μ^n does not converge weakly to δ_0 , the unit mass at the zero matrix.

(iv) S is compact and also the second condition in (ii) holds.

Proof. We prove only that (ii) $\Rightarrow$ (iii). The rest of the proof can easily be completed following arguments in Theorem 4.9. Assume that (ii) holds. Write:

$$\mathcal{J}_k = \left\{ A \in S : A_{ij} \geq \frac{1}{k} \text{ for all } i, j \right\}.$$

Let $\varepsilon > 0$. From (4.55), there exists a positive integer N and an integer $m \geq 1$ such that

$$\mu^N(\mathcal{J}_m) > 1 - \varepsilon. \quad (4.56)$$

Let (X_i) be i.i.d. matrices in S with distribution μ . Write:

$$Y_n \equiv X_1 X_2 \dots X_n.$$

Notice that for any positive integer s ,

$$\begin{aligned} & \{Y_{2N+s} \in \mathcal{B}_u, Y_N \in \mathcal{J}_m, X_{N+s+1} \dots X_{2N+s} \in \mathcal{J}_m\} \\ & \subset \{X_{N+1} \dots X_{N+s} \in \mathcal{J}_m^{-1}(\mathcal{B}_u \mathcal{J}_m^{-1})\} \end{aligned}$$

and

$$\begin{aligned} \mathcal{J}_m^{-1}(\mathcal{B}_u \mathcal{J}_m^{-1}) &= \{C : ACB \in \mathcal{B}_u \text{ for some } A, B \text{ in } \mathcal{J}_m\} \\ &\subset \left\{ C : \max_{i,j} C_{ij} \leq k_u m^2 \right\}. \end{aligned}$$

It follows that

$$\mu^s \left\{ C : \max_{i,j} C_{ij} \leq k_u m^2 \right\} \geq 1 - 2\varepsilon - u.$$

Remark 4.3. Under the conditions of Theorem 4.10, we can easily show that the following statements are also equivalent:

- (i) (μ^n) is tight and $0 \notin S$;
- (ii) μ^n converges weakly to some $\nu \in P(S)$ and $\nu(\mathcal{J}) = 1$;
- (iii) S is compact and $0 \notin S$;
- (iv) $m(S)$ is bounded and $m(S) \neq \{0\}$, where

$$m(S) = \{A \in S : \text{rank } A \leq \text{rank } B \text{ for all } B \text{ in } S\}.$$

Note that the condition (iv) implies that $0 \notin S$ and as such,

725

$$\mathcal{J} \cap m(S) \supset (\mathcal{J} \cap S)m(S)(\mathcal{J} \cap S).$$

It is also easily shown that (iv) implies $m(S)$ is a compact ideal of S and it has a completely simple kernel (see Theorem 1.6), which must be itself (see Proposition 1.9). Thus if $x \in \mathcal{J} \cap m(S)$, then $xm(S)x$, contained in $\mathcal{J}$, is a compact group containing an identity e , and $eSe = em(S)e$ is compact (in fact, a singleton). As in Theorem 4.9, it follows immediately that S is compact.

730

It is clear from the preceding discussion that the question of when the sequence (μ^n) is tight is fundamental in the context of weak convergence. Theorem 2.12 characterizes this tightness property. As we show below, we can be more specific in the context of matrices.

734

First we need a general lemma (which can be easily proven following the proof of Theorem 2.12).

736

Lemma 4.4. *Let (μ^n) be a sequence of probability measures on a locally compact Hausdorff second-countable topological semigroup. Suppose there is a closed subset F such that $\lim_{n \rightarrow \infty} \mu_n(F) = 0$. If for every open set $W \supset F$, $\lim_{n \rightarrow \infty} \mu_n(W) = 1$, then the sequence (μ_n) is tight.*

740

We now present Theorem 4.11 characterizing the property of tightness of products of i.i.d. random matrices.

742

Theorem 4.11. *Let μ be a (Borel) probability measure on $d \times d$ real matrices with usual topology and support S_μ . Let S be the closed multiplicative semigroup generated by S_μ . Let $m(S)$ be the set defined by*

745

$$m(S) = \{x \in S : \text{rank } x \leq \text{rank } y \text{ for all } y \in S\}.$$

Let a be the rank of the matrices in $m(S)$. Then the following results hold:

746

- (i) Suppose $a = d$. Then the sequence of convolution powers (μ^n) is tight if and only if S is a compact group.
- (ii) Suppose $a = 0$. Then the sequence (μ^n) is tight if and only if μ^n converges weakly to δ_0 , the unit mass at the zero matrix.
- (iii) Suppose $0 < a < d$. Then the sequence (μ^n) is tight if and only if the following two conditions hold:
 - a) There is a compact group $\mathcal{G}$ of invertible $a \times a$ matrices and an invertible $d \times d$ matrix y such that for any x in S , the matrix $y^{-1}xy$ can be uniquely represented in the form

755

$$y^{-1}xy = \begin{pmatrix} A & BD \\ DC & D \end{pmatrix}, \quad (4.57)$$

- where D is an element of $\mathcal{G}$, C is an $a \times (d - a)$ matrix, B is a $(d - a) \times a$ matrix, and A is a $(d - a) \times (d - a)$ matrix; 756
 b) For any open set V containing the set of matrices given by 758

$$\mathcal{M} = \left\{ \begin{pmatrix} BDC & BD \\ DC & D \end{pmatrix} \middle| \begin{array}{l} \text{there exists } A \text{ such that the matrix} \\ \text{in (4.57) is an element of } y^{-1}Sy \end{array} \right\},$$

$$\lim_{n \rightarrow \infty} \mu^n(yVy^{-1}) = 1.$$

In the only if case, the set $\mathcal{M}$ coincides with the set $m(y^{-1}Sy)$. Therefore when the sequence (μ^n) is tight, (μ^n) converges weakly if and only if there does not exist a proper normal subgroup $\mathcal{H}$ of $\mathcal{G}_0$, where $\mathcal{G}_0$ is the compact group with identity e_0 of $d \times d$ matrices given by

$$\left\{ \begin{pmatrix} 0 & 0 \\ 0 & D \end{pmatrix} : D \in \mathcal{G} \right\}$$

with $\mathcal{G}$ as in Condition a), such that

$$E(e_0 \cdot y^{-1}Sy) E(y^{-1}Sy \cdot e_0) \subset \mathcal{H}$$

and

$$e_0(y^{-1}S(\mu)y)e_0 \subset g\mathcal{H}$$

for some $g \in \mathcal{G}_0 - \mathcal{H}$. (Here, $E(W)$ stands for the set of all idempotent elements in the set W .)

Proof. Let us prove (i). The “if” part is obvious. For the “only if” part, notice that from Theorem 2.7,

$$\frac{1}{n} \sum_{k=1}^n \mu^k \text{ converges weakly to a probability measure } \lambda,$$

$$\lambda * \mu = \mu * \lambda = \lambda = \lambda * \lambda.$$

Here, the support S_λ of λ is the minimal (completely simple) ideal of S . If e is an idempotent matrix in S_λ , then $e \cdot S_\lambda \cdot e$ is a compact group. Since matrices in S are assumed to be of full rank and S_λ is an ideal of S , it follows that e is the usual identity matrix and

$$S = S \cdot e \subset S_\lambda = e \cdot S_\lambda \cdot e.$$

Let us prove (ii). The “if” part is obvious. For the “only if” part, notice that when $0 \in S$, the support S_λ , the minimal ideal of S , must be the singleton $\{0\}$, so that $\lambda = \delta_0$. Then the weak convergence of μ^n to λ follows from Theorem 2.7.

For the “if” part of (iii), suppose that Conditions (a) and (b) hold. If for each positive integer k ,

$$\mu^k(m(S)) = 0,$$

then it follows from Lemma 4.4 that the sequence (μ^n) is tight. Note that condition (b) is used here. Let us then suppose that there is a positive integer k such that

$$\mu^k(m(S)) > 0.$$

Since $m(S)$ is an ideal of S , it follows from Chap. 2 that

$$\lim_{n \rightarrow \infty} \mu^n(m(S)) = 1. \quad (4.58)$$

By Condition (a), there is an invertible matrix y such that for each x in S , $y^{-1}xy$ has representation given in (4.57). Let us define the set S_1 and the measure λ as follows:

$$\begin{aligned} S_1 &= \{y^{-1}xy \mid x \in S\}, \\ \lambda(y^{-1}\mathcal{B}y) &= \mu(\mathcal{B}), \mathcal{B} \subset S. \end{aligned} \quad (4.59)$$

It is easily verified that for $n \geq 1$,

$$\lambda^n(y^{-1}\mathcal{B}y) = \mu^n(\mathcal{B})$$

and also

$$m(S_1) = y^{-1} \cdot m(S) \cdot y.$$

It follows from (4.48) that

$$\lim_{n \rightarrow \infty} \lambda^n(m(S_1)) = 1. \quad (4.60)$$

it is also clear that the sequence (μ^n) is tight if and only if the sequence (λ^n) is tight.

Let $\varepsilon > 0$. Then it follows from (4.59) that there is a positive integer N and a compact subset $\mathcal{A} \subset m(S_1)$ such that

$$\lambda^N(\mathcal{A}) > \sqrt[3]{1 - \varepsilon}. \quad (4.61)$$

By Condition (a), every matrix in S_1 has the form in (4.57); that is, it looks like

$$\begin{pmatrix} A & BD \\ DC & D \end{pmatrix}, \quad (4.62)$$

where D is an element of the compact group $\mathcal{G}$ of $a \times a$ matrices. Note that the matrix D itself has rank a . Thus if the matrix in (4.62) belongs to $m(S_1)$, then it has rank a and it must look like

$$\begin{pmatrix} BDC & BD \\ DC & D \end{pmatrix}, \quad (4.63)$$

where $D \in \mathcal{G}$. In fact every matrix in $m(S_1)$ has the form (4.63). Now we claim that the set

$$\mathcal{A} \cdot m(S_1) \cdot \mathcal{A} \equiv \{x_1 \cdot x_2 \cdot x_3 : x_1, x_3 \in \mathcal{A} \text{ and } x_2 \in m(S_1)\}$$

is a compact subset of S_1 . To prove this claim, let us consider

$$\begin{pmatrix} B_1 D_1 C_1 & B_1 D_1 \\ D_1 C_1 & D_1 \end{pmatrix} \in \mathcal{A}, \quad \begin{pmatrix} B_2 D_2 C_2 & B_2 D_2 \\ D_2 C_2 & D_2 \end{pmatrix} \in \mathcal{A}$$

and

$$\begin{pmatrix} BDC & BD \\ DC & D \end{pmatrix} \in m(S_1).$$

Notice that

$$\begin{aligned} & \begin{pmatrix} B_1 D_1 C_1 & B_1 D_1 \\ D_1 C_1 & D_1 \end{pmatrix} \begin{pmatrix} BDC & BD \\ DC & D \end{pmatrix} \begin{pmatrix} B_2 D_2 C_2 & B_2 D_2 \\ D_2 C_2 & D_2 \end{pmatrix} \\ &= \begin{pmatrix} B_1 D_1 C_1 & B_1 D_1 \\ D_1 C_1 & D_1 \end{pmatrix} \begin{pmatrix} BD^* C_2 & BD^* \\ D^* C_2 & D^* \end{pmatrix} \\ &= \begin{pmatrix} B_1 D^{**} C_2 & B_1 D^{**} \\ D^{**} C_2 & D^{**} \end{pmatrix}, \end{aligned} \quad (4.64)$$

where

$$D^* = D(CB_2 + I_a)D_2 \in \mathcal{G},$$

I_a the $a \times a$ identity matrix, and

$$D^{**} = D_1(C_1 B + I_a)D^* \in \mathcal{G}.$$

Now observe that the sets

$$\left\{ B \mid \begin{pmatrix} BDC & BD \\ DC & D \end{pmatrix} \in \mathcal{A} \text{ for some } D \in \mathcal{G} \text{ and for some } C \right\}$$

and

$$\left\{ C \mid \begin{pmatrix} BDC & BD \\ DC & D \end{pmatrix} \in \mathcal{A} \text{ for some } D \in \mathcal{G} \text{ and for some } B \right\}$$

are both compact subsets in the usual topology of $(d - a) \times a$ and $a \times (d - a)$ matrices, respectively, since $\mathcal{A}$ is a compact subset of $d \times d$ matrices. The reason is the following: For every matrix B in the first set, entries in the matrix $B \cdot D$ ($\equiv B_1$, say), which is the block in the upper right-hand corner of the matrix

$$\begin{pmatrix} BDC & BD \\ DC & D \end{pmatrix},$$

which belongs to $\mathcal{A}$, must be bounded, since $\mathcal{A}$ is a compact set. Since the matrix D belongs to the compact group $\mathcal{G}$ of matrices with full rank, it is clear that entries in all such matrices $B = B_1 D^{-1}$ must be bounded. A similar argument applies to the second set. This observation along with the form of the product in (4.64) implies that the set $\mathcal{A} \cdot m(S_1) \cdot \mathcal{A}$ is a compact subset of S_1 . This establishes our claim and consequently it follows from (4.60) and (4.61) that

$$\begin{aligned} & \lim_{n \rightarrow \infty} \lambda^{n+2N} (\mathcal{A} \cdot m(S_1) \cdot \mathcal{A}) \\ & \geq \lim_{n \rightarrow \infty} \lambda^N (\mathcal{A}) \cdot \lambda^n (m(S_1)) \cdot \lambda^N (\mathcal{A}) \\ & > 1 - \varepsilon. \end{aligned}$$

It follows that the sequence (λ^n) and therefore the sequence (μ^n) is tight.

Let us now prove the “only if” part in (iii). Assume the sequence (μ^n) is tight. Then the sequence

$$\frac{1}{n} \sum_{k=1}^n \mu^k$$

converges weakly to some probability measure ν such that S_ν , the support of ν , is the (completely simple) minimal ideal of S , and consequently $S_\nu = m(S)$. Now we must exploit the algebraic structure of $m(S)$. Note that $m(S)$ is a completely simple subsemigroup, and thus, it has at least some idempotent e . Then e has rank a , and there is an invertible $d \times d$ matrix y such that

$$y^{-1}ey = \begin{pmatrix} 0 & 0 \\ 0 & I_a \end{pmatrix}, \quad (4.65)$$

where I_a is the $a \times a$ identity matrix.

Let $x \in S$. Write $y^{-1}xy$ as follows:

$$y^{-1}xy = \begin{pmatrix} A & B \\ C & D \end{pmatrix}, \quad (4.66)$$

where D is $a \times a$, B is $(d-a) \times a$, A is $(d-a) \times (d-a)$, and C is $a \times (d-a)$. Now since $m(S)$ is completely simple and the support of an idempotent probability measure ν , the set $eSe = em(S)e$ is a compact group (see Theorem 2.7). Observe that

$$y^{-1}(exe)y = \begin{pmatrix} 0 & 0 \\ 0 & I_a \end{pmatrix} \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & I_a \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & D \end{pmatrix}.$$

Let

$$\mathcal{G} = \left\{ D \mid \text{there exist } A, B, C \text{ as in (4.66) such that } \begin{pmatrix} A & B \\ C & D \end{pmatrix} \in y^{-1}Sy \right\}.$$

Then it is clear that $\mathcal{G}$ is a compact group of $a \times a$ matrices. Now recall from Chap. 1 that

$$m(S) = E(m(S)e) \cdot [em(S)e] \cdot E(em(S)). \quad (4.67)$$

where $E(W)$ denotes the set of all idempotent elements in the set W . A typical element in $y^{-1}[m(S)e]y$ has the form

$$y^{-1}(xe)y = (y^{-1}xy)(y^{-1}ey) = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & I_a \end{pmatrix} = \begin{pmatrix} 0 & B \\ 0 & D \end{pmatrix}.$$

If this element is idempotent, then

$$\begin{pmatrix} 0 & B \\ 0 & D \end{pmatrix} = \begin{pmatrix} 0 & B \\ 0 & D \end{pmatrix} \begin{pmatrix} 0 & B \\ 0 & D \end{pmatrix} = \begin{pmatrix} 0 & BD \\ 0 & D^2 \end{pmatrix}$$

so D must be I_a . Thus, elements in

$$y^{-1} \cdot E(m(S)e)y$$

have the form

$$\begin{pmatrix} 0 & B \\ 0 & I_a \end{pmatrix}.$$

Similarly, elements in

$$y^{-1}E(em(S))y$$

have the form

$$\begin{pmatrix} 0 & 0 \\ C & I_a \end{pmatrix}.$$

Then it follows from (4.67) that a typical element in $y^{-1}m(S)y$ has the form

$$\begin{pmatrix} 0 & B \\ 0 & I_a \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & D \end{pmatrix} \begin{pmatrix} 0 & 0 \\ C & I_a \end{pmatrix} = \begin{pmatrix} BDC & BD \\ DC & D \end{pmatrix}, \quad (4.68)$$

where D belongs to the compact group $\mathcal{G}$. Consider an arbitrary element in $y^{-1}Sy$. Let it be $y^{-1}zy$, which looks like

$$\begin{pmatrix} A_1 & B_1 \\ C_1 & D_1 \end{pmatrix}, \quad D_1 \in \mathcal{G}, \quad (4.69)$$

expressed as in (4.66). Then notice that

$$y^{-1}(ze)y \in y^{-1}m(S)y;$$

therefore,

843

$$\begin{pmatrix} A_1 & B_1 \\ C_1 & D_1 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & I_a \end{pmatrix} = \begin{pmatrix} 0 & B_1 \\ 0 & D_1 \end{pmatrix}$$

must have the same form as in (4.68), so that $B_1 = BD_1$ for some $(d - a) \times a$ 844
matrix B . Notice that this B is unique when B_1 and D_1 are given, since $D_1 \in \mathcal{G}$ 845
and $\mathcal{G}$ is a group. Similarly, 846

$$y^{-1}(ez)y \in y^{-1}m(S)y.$$

Therefore,

847

$$\begin{pmatrix} 0 & 0 \\ 0 & I_a \end{pmatrix} \begin{pmatrix} A_1 & B_1 \\ C_1 & D_1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ C_1 & D_1 \end{pmatrix}$$

must have the form in (4.68). This means that $C_1 = D_1 C$ for some unique $a \times (d-a)$ 848
matrix C . Thus, it follows from (4.69) that every matrix in $y^{-1}Sy$ has the form 849

$$\begin{pmatrix} A_1 & BD_1 \\ D_1 C & D_1 \end{pmatrix}, \quad (4.70)$$

where D_1 is an element of the compact group $\mathcal{G}$ of $a \times a$ matrices and (4.70) has 850
the same form described in (4.66). 851

The proof of (iii) is complete once we show that whenever the matrix 852

$$\begin{pmatrix} A & BD \\ DC & D \end{pmatrix},$$

with the same form as in (4.70) is in $y^{-1}Sy$, then the matrix 853

$$\begin{pmatrix} BDC & BD \\ DC & D \end{pmatrix}$$

must also be an element of $y^{-1}Sy$ and therefore in $y^{-1}m(S)y$. 854

To prove this part, let us take x in S such that 855

$$y^{-1}xy = \begin{pmatrix} A & BD \\ DC & D \end{pmatrix}$$

with the same form as in (4.70). Notice that $eSe = em(S)e$ is a group and as such, 856
(exe)⁻¹, the inverse of exe in this group, belongs to S . If 857

$$z \equiv y^{-1}(x(exe)^{-1}x)y,$$

then

858

$$z = \begin{pmatrix} A & BD \\ DC & D \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & D^{-1} \end{pmatrix} \begin{pmatrix} A & BD \\ DC & D \end{pmatrix} = \begin{pmatrix} BDC & BD \\ DC & D \end{pmatrix}.$$

Thus, we have proven that the set $m(S)$ consisting of all matrices in S with the minimal rank coincides with the set $y\mathcal{M}y^{-1}$. Since

$$\frac{1}{n} \sum_{k=1}^n \mu^k$$

converges weakly to the probability measure ν and the support of ν is $m(S)$, the Condition (b) now follows immediately from Theorem 2.7. $\square$

Next we consider the case of 2×2 real matrices. In this case, as we show in Theorem 4.12, we can be more specific.

Theorem 4.12. *Let μ be a (Borel) probability measure on 2×2 real matrices and S the closed (with respect to the usual topology) multiplicative semigroup generated by $S(\mu)$, the support of μ . Suppose the sequence (μ^n) is tight and the rank of the matrices in $m(S)$ is 1. (When this rank is 0 or 2, exactly what happens is clear from Theorem 4.9.)*

Suppose also that $m(S)$ does not contain a group of the form $\{1, -1\}$. Then either there is a common right nonzero eigenvector for every matrix in S with common eigenvalue 1 or there is a common left nonzero eigenvector for every matrix in S with common eigenvalue 1. In particular, there is an invertible 2×2 matrix y such that in case of the first possibility

$$y^{-1}Sy \subset \left\{ \begin{pmatrix} a & 0 \\ c & 1 \end{pmatrix} : a, c \text{ scalars} \right\}.$$

In case of the second possibility,

$$y^{-1}Sy \subset \left\{ \begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix} : a, b \text{ scalars} \right\}.$$

Suppose now that $m(S)$ contains a group of the form $\{1, -1\}$. Then either $m(S)$ consists of exactly eight elements or there is a common right eigenvector for every matrix in S with eigenvalue 1 or -1 , or there is a common left eigenvector for every matrix in S with eigenvalue 1 or -1 . In case of the last two possibilities, there is an invertible 2×2 matrix y such that in case of the second possibility,

$$y^{-1}Sy \subset \left\{ \begin{pmatrix} a & 0 \\ c & b \end{pmatrix} : a, c \text{ scalars and } b = \pm 1 \right\},$$

and in case of the third possibility,

$$y^{-1}Sy \subset \left\{ \begin{pmatrix} a & c \\ 0 & b \end{pmatrix} : a, c \text{ scalars and } b = \pm 1 \right\}.$$

Proof. Suppose (μ^n) is tight and the rank of the matrices in $m(S)$ is 1. Then as in the proof of Theorem 4.11, $m(S)$ is a completely simple subsemigroup of S with a compact group factor. Let $e = e^2$ be a fixed idempotent element of $m(S)$. Since $\text{rank}(e) = 1$, there is an invertible 2×2 matrix y such that

$$y^{-1}ey = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

Consider the set

$$y^{-1}(m(S)e)y.$$

Since the rank of the matrices in $y^{-1}m(S)y$ is 1, a typical element in the set $y^{-1}(m(S)e)y$ has the form

$$\begin{pmatrix} a_1b_1 & a_1b_2 \\ a_2b_1 & a_2b_2 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & a_1b_2 \\ 0 & a_2b_2 \end{pmatrix}.$$

If this is idempotent, then $a_2b_2 = 1$, so the set

$$E(y^{-1}m(S)e)y = y^{-1}(E(m(S)e))y,$$

where $E(W)$ is the set of idempotent elements in W , consists of elements of the form

$$\begin{pmatrix} 0 & a \\ 0 & 1 \end{pmatrix}.$$

Similarly, the set

$$E(y^{-1}em(S)y) = y^{-1}E(em(S))y$$

consists of elements of the form

$$\begin{pmatrix} 0 & 0 \\ b & 1 \end{pmatrix}.$$

Now suppose

$$E(m(S)e) \neq e \text{ and } E(em(S)) \neq e. \quad (4.71)$$

Then there exist elements x_1, x_2 in $m(S)$ such that $x_1e = (x_1e)^2$, $ex_2 = (ex_2)^2$ and

$$y^{-1}(x_1e)y = \begin{pmatrix} 0 & a \\ 0 & 1 \end{pmatrix}, \quad a \neq 0,$$

$$y^{-1}(ex_2)y = \begin{pmatrix} 0 & 0 \\ b & 1 \end{pmatrix}, \quad b \neq 0.$$

Notice that

$$y^{-1}(ex_2x_1e)y = \begin{pmatrix} 0 & 0 \\ b & 1 \end{pmatrix} \begin{pmatrix} 0 & a \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & ab + 1 \end{pmatrix}, \quad (4.72)$$

where $ab + 1 \neq 1$, since $ab \neq 0$.

But since $y^{-1}(em(S)e)y$ is a compact group of matrices of the form $\begin{pmatrix} 0 & 0 \\ 0 & c \end{pmatrix}$, 900
where c is a nonzero scalar, it is clear that in (4.72), $ab + 1 \in \{1, -1\}$. Since 901
 $ab \neq 0$, $ab + 1 = -1$ or $ab = -2$. 902

It follows that if (4.71) does occur, then we must have 903

$$E[y^{-1}(em(S))y] = \left\{ \begin{pmatrix} 0 & 0 \\ -2/a & 1 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \right\} \quad (4.73)$$

and 904

$$E[y^{-1}(m(S)e)y] = \left\{ \begin{pmatrix} 0 & a \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \right\}$$

for some $a \neq 0$. It is also clear that when $m(S)$ does not contain a group of the 905
form $\{1, -1\}$, the element $ab + 1$ must be 1, so that $ab = 0$. This means that (4.71) 906
cannot occur in this case. 907

Observe that the set $y^{-1}(em(S)e)y$, as a compact group must be either 908

$$\left\{ \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \right\} \text{ or } \left\{ \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix} \right\}.$$

As a consequence, since $m(S)$ is completely simple, it follows easily that every 909
element in $m(S)$ is idempotent when $y^{-1}(em(S)e)y$ is a singleton. When this set 910
is not a singleton, $m(S) = -m(S)$, so that every element in $m(S)$ is either an 911
idempotent or the negative of an idempotent. 912

Since we have the set equality 913

$$m(S) = E(m(S)e) \cdot em(S)e \cdot E(em(S)),$$

it follows that when (4.73) occurs, $y^{-1}m(S)y$ consists of exactly the following eight 914
elements; that is, there is $a \neq 0$ such that 915

$$\begin{aligned} y^{-1}m(S)y = & \left\{ \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}, \begin{pmatrix} 0 & a \\ 0 & 1 \end{pmatrix}, \right. \\ & \left. \begin{pmatrix} 0 & -a \\ 0 & -1 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ -2/a & 1 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 2/a & -1 \end{pmatrix}, \right. \\ & \left. \begin{pmatrix} 2 & -a \\ 2/a & -1 \end{pmatrix}, \begin{pmatrix} -2 & a \\ -2/a & 1 \end{pmatrix} \right\}. \end{aligned} \quad (4.74)$$

Now we look at the situation when $m(S)$ does not contain a group of the form 916
 $\{1, -1\}$. In this case, (4.71) does not occur, and then we have either 917

$$Se = m(S)e = E(m(S)e) = \{e\}, \quad (4.75)$$

4.3 Tightness of Products of I.I.D. Random Matrices: Weak Convergence

299

or

$$eS = em(S) = E(em(S)) = \{e\}. \quad (4.76)$$

918

Suppose (4.75) holds. Then for any $s \in S$, let

919

$$y^{-1}sy = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Then,

920

$$y^{-1}(se)y = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

in which case, $b = 0$ and $d = 1$. This means that when (4.75) occurs

921

$$y^{-1}Sy \subset \left\{ \begin{pmatrix} a & 0 \\ c & 1 \end{pmatrix} : a, c \text{ scalars} \right\}.$$

Similarly when (4.76) occurs,

922

$$y^{-1}Sy \subset \left\{ \begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix} : a, b \text{ scalars} \right\}.$$

Finally, $m(S)$ may contain a group of the form $\{1, -1\}$ and (4.71) also does not occur. In this case, as is clear from the preceding, we must have either

923

924

$$Se = m(S) \cdot e = \{e, -e\}$$

or

925

$$eS = e \cdot m(S) = \{e, -e\}.$$

The rest of the details are clear and left to the reader.

□ 926

In Theorem 4.10, we studied weak convergence of the sequence (μ^n) in nonnegative matrices. We now present a sufficient condition for weak convergence in real matrices. First, we need Lemma 4.5.

927

928

929

Lemma 4.5. Suppose $x = x^2$ is a $d \times d$ real matrix with no zero row. Suppose there are two rank 1 $d \times d$ real matrices y and z such that $xy = x$ and $xz = x$. Then we must have: $y = y^2$, $z = z^2$, and $yz = y$.

930

931

932

Proof. Notice that x has rank 1. Write x as

933

$$x = (x_{ij}) = (a_i x_j),$$

where each a_i is nonzero (since x has no zero row) and not all the x_j s are zero. Write y as

934

935

$$y = (y_{ij}) = (b_i y_j).$$

Since $xy = x$, for $1 \leq i, j \leq d$ we have

936

$$a_i x_j = \sum_{k=1}^d a_i x_k b_k y_j. \quad (4.77)$$

Since $x = x^2$ and $x \neq 0$,

937

$$\sum_{i=1}^d a_i x_i = 1. \quad (4.78)$$

It follows from (4.77) that

938

$$\sum_{i=1}^d a_i x_i = \left(\sum_{i=1}^d a_i y_i \right) \cdot \left(\sum_{k=1}^d b_k x_k \right). \quad (4.79)$$

Canceling a_i and then multiplying by b_j in (4.77), we have

939

$$b_j x_j = b_j y_j \cdot \sum_{k=1}^d b_k x_k$$

so that

940

$$\sum_{j=1}^d b_j x_j = \left(\sum_{j=1}^d b_j y_j \right) \left(\sum_{k=1}^d b_k x_k \right). \quad (4.80)$$

It follows from (4.78)–(4.80) that

941

$$\sum_{k=1}^d b_k x_k \neq 0 \text{ and } \sum_{j=1}^d b_j y_j = 1. \quad (4.81)$$

This means $y = y^2$. Similarly, we write z as

942

$$z = (z_{ij}) = (c_i z_j).$$

Since $xz = x$, it follows as before that $z = z^2$. It follows from (4.77) that

943

$$y_j = 0 \Rightarrow x_j = 0; \quad (4.82)$$

also

944

$$x_j / y_j = \sum_{k=1}^d b_k x_k.$$

945

4.3 Tightness of Products of I.I.D. Random Matrices: Weak Convergence

301

Let us now prove that $yz = y$. First notice that

946

$$(yz)_{ij} = b_i z_j \cdot \sum_{k=1}^d c_k y_k. \quad (4.83)$$

As in (4.82), we also have

947

$$x_j = 0, \text{ whenever } z_j = 0; \quad (4.84)$$

and,

948

$$x_j / z_j = \sum_{k=1}^d c_k x_k,$$

when $z_j \neq 0$. Suppose $y_j = 0$. Then $x_j = 0$ by (4.82). Since $xz = x$,

949

$$a_i x_j = a_i z_j \cdot \sum_{k=1}^d b_k x_k.$$

Since by (4.81),

950

$$\sum_{k=1}^d b_k x_k \neq 0,$$

we have $z_j = 0$. Then $(yz)_{ij} = 0$ by (4.83). Thus when $y_j = 0$,

951

$$(yz)_{ij} = y_{ij} = (b_i y_j).$$

Now suppose $y_j \neq 0$. By (4.77) and (4.81),

952

$$x_j = y_j \cdot \sum_{k=1}^d b_k x_k \neq 0. \quad (4.85)$$

Therefore by (4.84), $z_j \neq 0$ and

953

$$0 \neq x_j = z_j \cdot \sum_{k=1}^d c_k x_k. \quad (4.86)$$

It also follows from (4.77) that

954

$$\sum_{s=1}^d c_s x_s = \left(\sum_{s=1}^d c_s y_s \right) \cdot \left(\sum_{k=1}^d b_k x_k \right). \quad (4.87)$$

By (4.85)–(4.87),

955

$$y_j = z_j \cdot \sum_{s=1}^d c_s y_s. \quad (4.88)$$

By (4.83) and (4.88), we have

956

$$(yz)_{ij} = b_i y_j = y_{ij}.$$

This proves that $yz = y$.

□ 957

In Theorem 4.10, the strictly positive matrices played an important role. A similar 958
role is played by the set of matrices of rank 1 in Theorem 4.13, which provides a 959
sufficient condition for the weak convergence of (μ^n) in real matrices. 960

Theorem 4.13. *Let μ be a probability measure on $d \times d$ real matrices and let S be 961
the closed multiplicative semigroup generated by S_μ . Let I be the set of all matrices 962
in S with rank 1. Suppose for every open set $G \supset I$, 963*

$$\lim_{n \rightarrow \infty} \mu^n(G) = 1. \quad (4.89)$$

(Note that this condition holds when S does not contain the zero matrix and for 964
some positive integer m , $\mu^m(I) > 0$, since then I is an ideal of S .) (Suppose for 965
some $x \equiv (x_1, x_2, \dots, x_d)$, not all x_i s zero, $x \cdot S_\mu = x$ or $S_\mu \cdot x^T = x^T$. (In other 966
words, suppose x is a common left (or respectively, right) eigenvector for every 967
matrix in S_μ with common eigenvalue 1.) Then there is a probability measure λ on 968
 S such that $\mu^n \rightarrow \lambda$ weakly as $n \rightarrow \infty$ and $S_\lambda \subset I$. 969

Proof. First, we show that (μ^n) is tight. Let us assume that $xS_\mu = x$ and $x \neq 0$. 970
Since $xS_\mu = x$, we have $x \cdot S = x$ so that S does not contain the zero matrix. If 971
 $\mu^m(I) = 0$ for every $m \geq 1$, then since I is a closed subset of S , it follows from 972
(4.89) and Lemma 4.4 that (μ^n) is tight. Suppose now that for some positive integer 973
 m , $\mu^m(I) > 0$. It follows from Proposition 2.7 that 974

$$\lim_{n \rightarrow \infty} \mu^n(I) = 1. \quad (4.90)$$

Since $xS = x$ for the nonzero vector x , we can form a rank 1 $d \times d$ idempotent 975
matrix x_0 such that x is its first row; then $x_0S = x_0$. It follows from Lemma 4.5 976
that 977

$$y \in I \text{ and } z \in I \Rightarrow y \cdot z = y. \quad (4.91)$$

Let $p > 0$. It follows from (4.90) that there exists $q > 0$ and a positive integer N 978
such that 979

$$n \geq N \Rightarrow \mu^n(I) > 1 - q > \sqrt{1 - p}. \quad (4.92)$$

980

Let K be a compact subset of I such that $\mu^N(K) > 1 - q$. Let $n \geq N$. Then

981

$$\begin{aligned}\mu^{n+N}(K) &= \int \mu^N(Ky^{-1}) \mu^n(dy) \\ &\geq \int_I \mu^N(Ky^{-1}) \mu^n(dy) \\ &\geq \mu^N(K) \mu^n(I) > (1 - q)^2 > 1 - p.\end{aligned}$$

(Notice that for $y \in I$, $Ky^{-1} \supset K$, since $K \subset I$ and $K \cdot y = K$ by (4.91).) This proves that (μ^n) is tight.

983

By Theorem 2.7,

984

$$\frac{1}{n} \sum_{k=1}^n \mu^k$$

converges weakly to some $\lambda \in P(S)$, where S_λ is a complete simple (minimal) ideal of S with a compact group factor. Thus, $S_\lambda \subset I$. Since a compact group of rank 1 $d \times d$ matrices is either a singleton or a group of the form $\{1, -1\}$ and since $xS = x$, it is clear that the group factor of S_λ is a singleton. (If A is a rank 1 $d \times d$ matrix in the compact group factor of S_λ , then if the i th row of A is $(b_1 a_1, b_1 a_2, \dots, b_1 a_d)$, the matrix $A^{n+1} = \alpha^n \cdot A$, where $\alpha = b_1 a_1 + b_2 a_2 + \dots + b_d a_d$, and consequently, $|\alpha|$ must be 1 so that $A^2 = A$ or $-A$.) It follows from Theorem 2.7 that $\mu^n \rightarrow \lambda$ weakly. $\square$

992

Let us now present some examples illustrating the above theory.

993

Example 4.1. We show that Theorem 4.12 cannot be extended to $d \times d$ matrices with $d > 2$.

994

995

Consider the semigroup S of 3×3 strictly positive matrices given by $S = \{e_0, e_1, e_2, e_3\}$, where

996

997

$$\begin{aligned}e_0 &= \begin{pmatrix} 1/2 & 1/6 & 1/3 \\ 1/2 & 1/6 & 1/3 \\ 1/2 & 1/6 & 1/3 \end{pmatrix}, & e_1 &= \begin{pmatrix} 3/4 & 1/12 & 1/6 \\ 3/4 & 1/12 & 1/6 \\ 3/4 & 1/12 & 1/6 \end{pmatrix}, \\ e_2 &= \begin{pmatrix} 1/2 & 1/6 & 1/3 \\ 1 & 1/3 & 2/3 \\ 1/4 & 1/12 & 1/6 \end{pmatrix}, & e_3 &= \begin{pmatrix} 3/4 & 1/12 & 1/6 \\ 3/2 & 1/6 & 1/3 \\ 3/8 & 1/24 & 1/12 \end{pmatrix}.\end{aligned}$$

Then the multiplication table for S is given by

998

$$\begin{array}{cccc} \times & e_0 & e_1 & e_2 & e_3 \\ e_0 & e_0 & e_1 & e_0 & e_1 \\ e_1 & e_0 & e_1 & e_0 & e_1 \\ e_2 & e_2 & e_3 & e_2 & e_3 \\ e_3 & e_2 & e_3 & e_2 & e_3 \end{array}$$

Consider the probability measure μ on S given by

999

$$\mu(e_i) = \frac{1}{4}, \quad i = 0, 1, 2, 3.$$

Then by Theorem 4.10, μ^n converges weakly to some probability measure on S . 1000
However, unlike Theorem 4.12, there does not exist a nonzero common eigenvector 1001
for every matrix in S with common eigenvalue 1. Suppose there exists a nonzero 1002
vector $x = (x_1, x_2, x_3)$ such that $xe_i = x$ for $i = 0, 1, 2, 3$. Then we can choose 1003
 $a \neq 0, b \neq 0$, and $c \neq 0$ such that $ax_1 + bx_2 + cx_3 = 1$ and 1004

$$x_0 = \begin{pmatrix} ax_1 & bx_2 & cx_3 \\ ax_1 & bx_2 & cx_3 \\ ax_1 & bx_2 & cx_3 \end{pmatrix}$$

is an idempotent matrix and $x_0e_i = x_0, i = 0, 1, 2, 3$. It follows from Lemma 4.5 1005
that for y, z in S , $yz = y$. But this contradicts the multiplication table for S . Simi- 1006
larly, it can be shown that there is no nonzero x such that 1007

$$e_ix^T = x^T, \quad i = 0, 1, 2, 3.$$

Example 4.2. This example was considered in [128]. Consider the probability mea- 1008
sure μ such that $S_\mu = \{A, B\}$, where A and B are the 2×2 matrices given by 1009

$$A = \begin{pmatrix} 1 & p \\ 0 & q \end{pmatrix}, \quad B = \begin{pmatrix} r & 0 \\ s & 1 \end{pmatrix},$$

where $p > 0, q > 0, r > 0$, and $s > 0$. Assume $0 < q < 1, 0 < r < 1$, and 1010
 $(1 - q)(1 - r) = ps$. Consider the vector 1011

$$x = \left(1, \frac{p}{1 - q}\right).$$

Then $xA = x$ and $xB = x$. It follows that $x \cdot S = x$, so S must be compact 1012
and S does not contain the zero matrix. Notice that AB is strictly positive, and it 1013
follows by Theorem 4.10 that μ^n converges weakly to some probability measure λ 1014
and the strictly positive matrices in S have λ measure 1. It is relevant to point out 1015
that $S_\lambda \cap J \neq \emptyset$, since 1016

$$\lim_{n \rightarrow \infty} A^n \in J \text{ and } S_\lambda = m(S).$$

Example 4.3. This example illustrates Theorem 4.13. Consider the probability mea- 1017
sure μ such that $S_\mu = \{A, B, C\}$, where A, B, C are the 2×2 matrices given by 1018

$$A = \begin{pmatrix} 1 & p \\ 0 & q \end{pmatrix}, \quad B = \begin{pmatrix} r & 0 \\ s & 1 \end{pmatrix}, \quad C = \begin{pmatrix} 0 & 0 \\ (1 - q)/p & 1 \end{pmatrix},$$

where $q \neq 1$, p, q, r, s are nonzero real numbers such that $ps = (1 - q)(1 - r)$. It is then easily verified that the vector

$$x = \left(1, \frac{p}{1 - q}\right)$$

satisfies the equations $xA = x$, $xB = x$, and $xC = x$. It follows by Theorem 4.13 then μ^n converges weakly to some λ , where λ (the rank 1 matrices in S) = 1.

Example 4.4. This example illustrates Theorem 4.13 when S is a noncompact semigroup of 2×2 nonnegative matrices. Consider the sets K_1 and K_2 defined by

$$K_1 = \left\{ \begin{pmatrix} 1 & m \\ 0 & n \end{pmatrix} : m, n \text{ are nonnegative integers} \right\}$$

and

$$K_2 = \left\{ \begin{pmatrix} 1 & m \\ 0 & 0 \end{pmatrix} : m \text{ is a nonnegative integer} \right\}.$$

Consider a probability measure μ on $S_0 = K_1 \cup K_2$, which is a semigroup with respect to matrix multiplication such that

$$S_0 \supset S_\mu \supset \left\{ \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \right\}.$$

Notice that if $x = (1, 0)$, then

$$\begin{pmatrix} 1 & m \\ 0 & n \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

so that $S_0 x^T = x^T$. Then it follows from Theorem 4.13 that μ^n converges weakly to some probability measure λ such that $S_\lambda = K_2$.

Now we appeal to Theorem 2.11 for an interesting result on tightness for nonnegative matrices.

Proposition 4.1. *Let μ be a probability measure on the Borel subsets of $d \times d$ ($0 < d < \infty$) nonnegative matrices. Suppose S is the closed multiplicative semigroup generated by S_μ and $S \subset J^c$; that is, no matrix in S has a zero row or a zero column. If $\mu^n(K)$ does not converge to zero as $n \rightarrow \infty$ for some compact set K , then the sequence (μ^n) is tight.*

Proof. The proof follows immediately from Theorem 2.11 and the remark immediately preceding Theorem 2.8. $\square$

As we saw in the preceding examples, we do not often have the situation when $S \subset J^c$, and as such, the next result, due to Kesten and Spitzer, is interesting. The proof is taken from [153].

Theorem 4.14. Let μ and S be as in Proposition 4.1. Instead of assuming $S \subset J^c$, suppose

$$\mu(J) = 0 \text{ and } S \cap \mathcal{J} \neq \emptyset,$$

where $\mathcal{J}$ is the set of all $d \times d$ strictly positive matrices in S and J is the set of all $d \times d$ matrices in S that have at least one zero row or one zero column. Then either μ^n converges to zero in the weak* sense (or vaguely) in the nonzero matrices of S or the sequence (μ^n) is tight.

Proof. We prove slightly more than is necessary.

Suppose there exist $q_1 > 0$, $q_2 > 0$, $\varepsilon > 0$ and a subsequence (n_k) of positive integers such that for $k \geq 1$,

$$\mu^{n_k}(B(q_1, q_2)) > \varepsilon, \quad (4.93)$$

where

$$B(q_1, q_2) = \left\{ A \in S : \max_{i,j} A_{ij} \geq q_2 \text{ and for each } s, \min_i A_{is} \leq q_1 \right\}.$$

We show that if (4.93) and the assumptions

$$\mu(J) = 0, \quad S \cap \mathcal{J} \neq \emptyset \quad (4.94)$$

hold, then the sequence (μ^n) is tight.

AQ2 *Step 6.* There is a positive number q_0 , a positive integer N , and a compact subset $W \subset \mathcal{J}$ such that for large k ,

$$\mu^{n_k+N}(W) > q_0. \quad (4.95)$$

To prove (4.95), take ε be as in (4.93). It follows from (4.94), (4.54), and (4.55) that there is a positive integer N and $\nu > 0$ such that

$$\mu^N(\mathcal{J}_\nu) > 1 - \frac{\varepsilon}{4}, \quad (4.96)$$

where

$$\mathcal{J}_\nu = \left\{ A \in S : \frac{1}{\nu} \leq A_{ij} \leq \nu \text{ for every } i, j \right\}.$$

It follows from (4.93) and (4.96) that for $n_k > N$,

$$\begin{aligned} \varepsilon &< \mu^{n_k}(B(q_1, q_2)) \\ &= \int \mu^{n_k-N}(x^{-1}B(q_1, q_2)) \mu^N(dx) \\ &\leq \mu^{n_k-N}(\mathcal{J}_\nu^{-1}B(q_1, q_2)) + \frac{\varepsilon}{4}. \end{aligned}$$

It is easily verified that the product set

1061

$$\mathcal{J}_v [\mathcal{J}_v^{-1} B(q_1, q_2)] \mathcal{J}_v$$

is contained in a compact set $W \subset \mathcal{J}$. Notice that

1062

$$\begin{aligned} & \mathcal{J}_v^{-1} B(q_1, q_2) \\ & \subset F(q_1, q_2, v) = \left\{ A \in S : \frac{q_2}{vd} \leq \max_{i,j} A_{ij} \leq q_1 v \right\} \end{aligned}$$

and there exist $a > 0, b > 0$ such that

1063

$$\mathcal{J}_v F(q_1, q_2, v) \mathcal{J}_v \subset \{A \in S : a \leq A_{ij} \leq b \text{ for all } i, j\}.$$

It follows that for $n_k > N$,

1064

$$\begin{aligned} \mu^{n_k+N}(W) & \geq \mu^N(\mathcal{J}_v) \mu^{n_k-N}(\mathcal{J}_v^{-1} B(q_1, q_2)) \mu^N(\mathcal{J}_v) \\ & > \frac{(1 - \varepsilon/4)^2 \cdot 3\varepsilon}{4}. \end{aligned}$$

This proves (4.95).

1065

Step 7. If there exists $q > 0$, a compact set $W \subset \mathcal{J}$, and a subsequence (m_k) of positive integers such that

1066

1067

$$\mu^{m_k}(W) > q, \quad k \geq 1, \quad (4.97)$$

then (μ^{m_k}) is tight, and moreover given $\varepsilon > 0$, there are positive integers n_0 and k_0 and a compact set $W_\varepsilon \subset \mathcal{J}$ such that for $k \geq k_0$,

1068

1069

$$\mu^{m_k+n_0}(W_\varepsilon) > 1 - \varepsilon. \quad (4.98)$$

To prove this, write for $u > 0, 1 \leq i \leq d$,

1070

$$B_u = \left\{ A \in S : u \leq \sum_{j=1}^d A_{ij} \leq \frac{1}{u} \right\}.$$

Then we have

1071

$$\sup_{u>0} \lim_{n \rightarrow \infty} \sup_{x \in S} \mu^n(x^{-1} B_u) = 1. \quad (4.99)$$

To see this, notice that the set W in (4.97) is contained in some B_u and the set $B_u \mathcal{J}_v^{-1}, \mathcal{J}_v$ as in Step 1 is contained in B_w for some $w > 0$. Now let the sup

1072

1073

in (4.99) be the number b . Suppose $b < 1$. Because of (4.97), $b > 0$. Let $b < c < 1$ and $c(1 + c)/2 < b$. Let $u > 0$. There exists a positive integer N , $v > 0$, and $0 < \varepsilon < c/4$ such that

$$\mu^N(S - \mathcal{J}_v) < \varepsilon, \quad \sup_{z \in S} \mu^N(z^{-1}B_u) < c - \varepsilon.$$

Let $E = B_u \mathcal{J}_v^{-1}$. Then $E \subset B_w$ for some $w > 0$. For $n > N$ and $x \in S$,

$$\begin{aligned} \mu^n(x^{-1}B_u) &= \int \mu^N(y^{-1}x^{-1}B_u) \mu^{n-N}(dy) \\ &= \int_{x^{-1}E} + \int_{S-x^{-1}E} \\ &\leq (c - \varepsilon)\mu^{n-N}(x^{-1}E) + \left(\frac{c}{2} - \varepsilon\right) \left[1 - \mu^{n-N}(x^{-1}E)\right] \\ &\leq \frac{c}{2}\mu^{n-N}(x^{-1}E) + \frac{c}{2} \leq \frac{c + c^2}{2} < b, \end{aligned}$$

which is a contradiction. This establishes (4.99). It follows from (4.97) and (4.99) that given $0 < \varepsilon < q$, there exist $u > 0$ and elements x_n in S such that for all sufficiently large n and each $k \geq 1$,

$$\mu^{m_k}(W) > q, \quad \mu^n(x_n^{-1}B_u) > 1 - \frac{\varepsilon}{2}.$$

Thus, for sufficiently large k ,

$$x_{m_k}^{-1}B_u \cap W \neq \emptyset$$

so that $x_{m_k} \in B_u W^{-1} \equiv T_u$, say.

We can now verify easily that there exist $a_\varepsilon > 0$, $b_\varepsilon > 0$ such that

$$T_u^{-1}B_u \subset L_\varepsilon \subset D_\varepsilon, \quad (4.100)$$

where sets L_ε and D_ε are defined by

$$L_\varepsilon = \left\{ A \in S : \max_{i,j} A_{ij} \geq a_\varepsilon \text{ and } \min_{i,j} A_{ij} \leq b_\varepsilon \right\}$$

and

$$D_\varepsilon = \left\{ A \in S : \min_{i,j} A_{ij} \leq b_\varepsilon \right\}.$$

It follows that given $\varepsilon > 0$, there exists k_0 such that for $k \geq k_0$,

$$\mu^{m_k}(D_\varepsilon) \geq \mu^{m_k}(L_\varepsilon) > 1 - \varepsilon. \quad (4.101)$$

Now we use the proof of (ii) $\Rightarrow$ (iii) in Theorem 4.10. Let the X_i s and Y_i s be as in that proof. Let $k \geq k_0$. Then, for $m_k > 2N$ with N as before, we have

$$\begin{aligned} 1 - 2\varepsilon &\leq \Pr \{Y_{2N+(m_k-2N)} \in D_\varepsilon, Y_N \in \mathcal{J}_v, X_{m_k-N+1} \cdots X_{m_k} \in \mathcal{J}_v\} \\ &\leq \Pr \{Y_N \in \mathcal{J}_v, X_{N+1} \cdots X_{m_k-N} \in \mathcal{J}_v^{-1}(D_\varepsilon \mathcal{J}_v^{-1}), \\ &\quad X_{m_k-N+1} \cdots X_{m_k} \in \mathcal{J}_v\} \\ &\leq \mu^{m_k-2N} (\mathcal{J}_v^{-1}(D_\varepsilon \mathcal{J}_v^{-1})). \end{aligned}$$

It follows as in the proof of Theorem 4.10 that the sequence (μ^{m_k}) is tight. It is easily verified that

$$\mathcal{J}_v \cdot L_\varepsilon \cdot \mathcal{J}_v \subset \mathcal{J} \quad (4.102)$$

so that (4.98) follows immediately from the tightness of (μ^{m_k}) and (4.102). This completes Step 2.

Step 8. It follows from Steps 1 and 2 that when (4.93) occurs, the sequence (μ^n) has a weak* limit point in $P(S)$ that is not the unit mass at the zero matrix. Thus, there is a subsequence (p_k) of positive integers such that

$$\mu^{p_k} \rightarrow \lambda_0 \neq \delta_0$$

weakly. Notice that for any positive integer j , given $\varepsilon > 0$, taking compact sets $E_1 \subset S$ and $E_2 \subset S - J$ such that

$$\mu^{p_k}(E_1) > 1 - \varepsilon, \quad \mu^j(E_2) > 1 - \varepsilon$$

we have

$$\begin{aligned} 1 - \varepsilon &< \mu^{p_k}(E_1) = \int \mu^{p_k-j}(E_1 x^{-1}) \mu^j(dx) \\ &\leq \int_{E_2} \mu^{p_k-j}(E_1 x^{-1}) \mu^j(dx) + \varepsilon \end{aligned}$$

so that

$$\mu^{p_k-j}(E_1 E_2^{-1}) \geq 1 - 2\varepsilon.$$

Since $E_2 \subset S - J$, the set $E_1 E_2^{-1}$ is compact, and consequently (μ^{p_k-j}) is a tight sequence. Following the proof of Lemma 2.2, we observe that there is a further subsequence $(t_k) \subset (p_k)$ such that for each $j \geq 1$,

$$\mu^{t_k-j} \rightarrow \lambda_j \in P(S),$$

weakly. We can also assume that (4.98) holds with (t_k) replacing (m_k) . Now we show that the sequence (λ_{t_k}) is tight. Since for $j \geq 1$,

$$\mu^j * \lambda_j = \lambda_j * \mu^j = \lambda_0.$$

Then for $k \geq 1$, we have

1105

$$\mu^{t_k} * \lambda_{t_k} = \lambda_{t_k} * \mu^{t_k} = \lambda_0.$$

Since (4.98) and (4.101) hold with (t_k) replacing (m_k) , it follows from (4.102) that 1106
given $\varepsilon > 0$, there exists a compact set $F \subset \mathcal{J}$ and a positive integer N such that 1107
for $k \geq 1$, 1108

$$\mu^{t_k+2N}(F) > 1 - \varepsilon, \quad \lambda_0 * \mu^{2N}(F) > 1 - \varepsilon.$$

(Note that $\lambda_0 * \mu = \mu * \lambda_0$ and because of (4.101), $\lambda_0(L_\varepsilon) \geq 1 - \varepsilon$.) Thus for 1109
 $k \geq 1$, 1110

$$\begin{aligned} 1 - \varepsilon &< \mu^{2N} * \lambda_0(F) = \int \lambda_{t_k}(x^{-1}F) \mu^{t_k+2N}(dx) \\ &= \int_F + \int_{S-F} \leq \lambda_{t_k}(F^{-1}F) + \varepsilon. \end{aligned}$$

Using a similar argument, it follows that for $k \geq 1$,

1111

$$\lambda_{t_k}(F^{-1}F \cap FF^{-1}) > 1 - 6\varepsilon. \quad (4.103)$$

Note that $F^{-1}F \cap FF^{-1}$ is compact and contained in $S - J$. Thus, it follows that 1112
the sequence (λ_{t_k}) is tight and every weak* limit point of (λ_{t_k}) is in $P(S)$ and has 1113
zero mass for the set J . Now following the proof of Lemma 2.2, one observes that 1114
if λ is a weak* limit point of (λ_{t_k}) , then 1115

$$\lambda \in P(S), \quad \lambda * \lambda = \lambda, \quad \lambda(S - J) = 1 \quad (4.104)$$

and there is a subsequence $(s_k) \subset (t_k)$ such that

1116

$$s_{k+1} - s_k \rightarrow \infty \text{ as } k \rightarrow \infty$$

and

1117

$$\mu^{s_{k+1}-s_k} \rightarrow \lambda \text{ weakly.} \quad (4.105)$$

We show that $\lambda(\mathcal{J}) = 1$. Write

1118

$$r_k = s_{k+1} - s_k \text{ and } \mathcal{E}_v = \left\{ A \in S : \max_{i,j} A_{ij} > v \right\}.$$

By (4.104) and (4.105), there are positive u, v, w such that for large k ,

1119

$$\mu^{r_k}(\mathcal{E}_v) > 1 - \varepsilon, \quad \mu^N(\mathcal{J}_u) > 1 - \varepsilon$$

and

1120

$$\mu^{2N}(\mathcal{J}_w) > 1 - \varepsilon$$

for some positive integer N . It is easily seen that $\mathcal{J}_w^{-1}\mathcal{E}_v \subset \mathcal{E}_z$ for some $z > 0$ and for k large, 1121

$$\mu^{rk-2N}(\mathcal{J}_w^{-1}\mathcal{E}_v) > 1 - 2\varepsilon. \quad 1122$$

Since the closure of the set $\mathcal{J}_u \cdot \mathcal{E}_z \cdot \mathcal{J}_u \subset \mathcal{J}$, it follows that given $\varepsilon > 0$, there is a closed set $D \subset \mathcal{J}$ such that for large k , 1123

$$\mu^{rk}(D) > (1 - \varepsilon)^2(1 - 2\varepsilon). \quad 1124$$

It follows from (4.105) that $\lambda(\mathcal{J}) = 1$. 1125

Now (4.93) gives (4.97) and by (4.97), there exists $x \in \mathcal{J}$ such that for any open set U containing x , 1126

$$\mu^n(U) \not\rightarrow 0. \quad (4.106) \quad 1127$$

Notice that if $y \in S - J$ such that $x \notin Sy$, then there exist open sets U_x, U_y such that 1128

$$x \in U_x \subset S - J, \quad y \in U_y \subset S - J, \text{ and } U_x U_y^{-1} = \emptyset. \quad 1129$$

A similar result holds when $x \notin yS$. Following the proof of Theorem 2.11, it follows that if (4.106) holds for $x \in \mathcal{J}$, then 1130

$$x \in Sx \text{ and } x \in xS; \quad (4.107)$$

$$y \in Sx \cap J^c \Rightarrow Sx = Sy; \quad (4.108)$$

$$y \in xS \cap J^c \Rightarrow xS = yS. \quad (4.109)$$

Let $S_1 \equiv S - \{0\}$. Then xS_1x is a subsemigroup of $\mathcal{J}$. Let $y \in xS_1x$. Then 1132

$$yx \in xS \cap J^c \text{ and } xy \in Sx \cap J^c.$$

It follows from (4.108) and (4.109) that 1133

$$y(xS_1x) = xS_1x = (xS_1x)y$$

so that xS_1x is a group contained in $\mathcal{J}$. Define the set $\mathcal{K}_0$ by 1134

$$\mathcal{K}_0 = \{y \in \mathcal{J} : \mu^n(U_y) \not\rightarrow 0 \text{ as } n \rightarrow \infty \text{ for every open } U_y \text{ containing } y\}.$$

Then $\mathcal{K}_0$ is a nonempty ideal of $S \cap J^c$. Since it is clear from (4.108) and (4.109) that for each $y \in \mathcal{K}_0$, S_1y is a minimal left ideal of $\mathcal{K}_0$ and yS_1 is a minimal right ideal of $\mathcal{K}_0$, it follows from Propositions 1.7 and 1.9 that the set $m(\mathcal{K}_0)$ of matrices with minimal rank from $\mathcal{K}_0$ is the completely simple minimal ideal of $\mathcal{K}_0$. By (4.107), 1135

$$y \in \mathcal{K} \Rightarrow y \in yS_1 \cap S_1y. \quad (4.110) \quad 1139$$

Since the union of all minimal right ideals of $\mathcal{K}_0$ is the minimal ideal of $\mathcal{K}_0$ (see the proof of Proposition 1.7), it follows from (4.110) that $m(\mathcal{K}_0) = \mathcal{K}_0$. 1140

1141

To complete the proof, consider the probability measure λ in (4.105). Note that $\lambda(\mathcal{J} \cap S_\lambda) = 1$, so if $x \in \mathcal{J} \cap S_\lambda$ and U_x is an open set containing x , $\lambda(U_x) > 0$. It follows from (4.105) that

$$\mu^n(U_x) \not\rightarrow 0 \text{ as } n \rightarrow \infty.$$

It follows that $\mathcal{J} \cap S_\lambda \subset \mathcal{H}_0$ so $\lambda(\mathcal{H}_0) = 1$. Note that $\mu * \lambda = \lambda * \mu$ and $\lambda = \lambda * \lambda$, so $(\mu * \lambda)^n = \mu^n * \lambda$ and

$$\begin{aligned} (\mu * \lambda)^n(\mathcal{H}_0) &\geq \mu^n * \lambda((S \cap J^c) \mathcal{H}_0) \\ &\geq \mu^n(S \cap J^c) \lambda(\mathcal{H}_0) = 1 \quad \forall n \geq 1. \end{aligned}$$

This allows us to restrict probability measures λ and $\mu * \lambda$ to the completely simple semigroup $\mathcal{H}_0$ (which is a closed subsemigroup of S) with relative topology. Note that the support of $\mu * \lambda = \lambda * \mu$ restricted to $\mathcal{H}_0$ generate the closed semigroup $\text{cl}(S \cdot S_\lambda) \cap \mathcal{H}_0$, which is an ideal of $\mathcal{H}_0$ since

$$\text{cl}(S \cdot S_\lambda) = \text{cl}(S_\lambda \cdot S).$$

Since $\mathcal{H}_0$ is simple, the support of $\mu * \lambda$ restricted to $\mathcal{H}_0$ generates $\mathcal{H}_0$. Notice that if

$$x \in \mathcal{J} \cap S_\lambda \subset \mathcal{H}_0,$$

then $x^2 \in \mathcal{H}_0$ and for any open set U containing x^2 , if U_x is an open set containing x such that $u_x \cdot U_x \subset U$, then

$$(\mu * \lambda)^n(U) \geq \mu^n(U_x) \lambda(U_x) \not\rightarrow 0$$

as $n \rightarrow \infty$. It follows from Theorem 2.10 that the group factor of $\mathcal{H}_0$ is compact. Since $\mathcal{H}_0 \subset J$, the group factor of $\mathcal{H}_0$ must be a singleton. (See Corollary 1.8.) Then Theorem 2.7(iii) implies

$$\mu^n * \lambda = (\mu * \lambda)^n$$

converges weakly as $n \rightarrow \infty$. Thus given $\varepsilon > 0$, there exist a compact set $E_1 \subset \mathcal{H}_0 \subset \mathcal{J}$ and a compact set $E_2 \subset \mathcal{J}$ such that for $n \geq 1$,

$$\mu^n * \lambda(E_1) > 1 - \varepsilon \text{ and } \lambda(E_2) > 1 - \varepsilon.$$

Thus for $n \geq 1$,

$$\begin{aligned} 1 - \varepsilon &< \int \mu^n(E_1 x^{-1}) \lambda(dx) \\ &\leq \mu^n(E_1 E_2^{-1}) + \varepsilon. \end{aligned}$$

Since $E_1 E_2^{-1}$ is compact, it follows that the sequence (μ^n) is tight. $\square$

Now we recall from Sect. 4.2 that some of the problems involving certain affine maps from $(R^+)^d$ into itself can be conveniently handled using the corresponding nonnegative $(d + 1) \times (d + 1)$ matrices that represent them. But these nonnegative matrices are never strictly positive, since they take the form

$$\begin{pmatrix} A & B \\ 0 & 1 \end{pmatrix}, \quad (4.111)$$

where A is $d \times d$, B is $d \times 1$, $0 \equiv (0, \dots, 0)$ is $1 \times d$, and 1 is 1×1 . Therefore, it is natural to ask in this case whether there can be analogs of Theorems 4.14 and 4.10. The answers in both cases are positive as Theorems 4.15 and (4.30) show. (Note that Theorem 4.15 is the restatement of Theorem 4.4, the proof of which was postponed until now.)

Theorem 4.15. *Let μ be a Borel probability measure on $(d + 1) \times (d + 1)$, d is a finite positive integer, nonnegative matrices as in (4.111). Let S be the closed multiplicative semigroup of such matrices generated by S_μ (with usual topology). Suppose the following conditions hold:*

- (i) *There exist $q > 0$ and a subset B_z with $z \geq 1$, defined by*

$$B_z = \{x \in S : x_{i,d+1} \leq z \text{ for } 1 \leq i \leq d + 1\},$$

such that for some subsequence (n_k) of positive integers

$$\mu^{n_k}(B_z) > q, \quad k \geq 1.$$

- (ii) *The set $\mathcal{J}_0$ defined by*

$$\mathcal{J}_0 = \{x \in S : \text{for each } i, 1 \leq i \leq d, x_{i,d+1} > 0\}$$

is nonempty.

- (iii) $\mu(J) = 0$, where J is the set of all matrices in S that have at least one zero row or zero column.

Then the sequence (μ^n) is tight.

Proof. Notice that Condition (iii) implies that $\mu^n(J) = 0$ for $n \geq 1$ (see (4.53) and (4.54)). It is also easily verified that $\mathcal{J}_0$ is a nonempty open ideal of J_r^c , where $J_r = \{x \in J \mid x \text{ has at least one zero row}\}$. It follows as in (4.55) that

$$\lim_{n \rightarrow \infty} \mu^n(\mathcal{J}_0) = 1. \quad (4.112)$$

Define for each $u > 0$, the set $\mathcal{H}_u$ by

1186

$$\mathcal{H}_u = \left\{ x \in S : \min_{1 \leq i \leq d} x_{i,d+1} \geq u, \min_{1 \leq i \leq d} \sum_{j=1}^d x_{ji} \geq u, \max_{1 \leq i, j \leq d+1} x_{ij} \leq \frac{1}{u} \right\}.$$

Then

1187

$$\mathcal{I}_0 \cap J_c^c = \bigcup_{m=1}^{\infty} \mathcal{H}_{\frac{1}{m}}. \quad (4.113)$$

For each $u \geq 1$, define the set B_u by

1188

$$B_u = \{x \in S : x_{i,d+1} \leq u \text{ for } 1 \leq i \leq d+1\}.$$

Notice that for any $x \in S$,

1189

$$B_u x^{-1} \subset B_u;$$

then, for $k \geq 1, n \geq 1$,

1190

$$\mu^{n+k}(B_u) \leq \mu^n(B_u)$$

so that $\lim_{n \rightarrow \infty} \mu^n(B_u)$ exists.

1191

We claim that

1192

$$b \equiv \sup_{u \geq 1} \lim_{n \rightarrow \infty} \mu^n(B_u) = 1. \quad (4.114)$$

By Condition (i) of Theorem 4.15, $b > 0$. To prove (4.114), we proceed as we did in establishing (4.99). Let $b < 1$. Choose c such that

1194

$$b < c < \sqrt{2b + \frac{1}{4}} - \frac{1}{2}.$$

Then $b < c < 1$ and $c + c^2 < 2b$. Let $s \geq 1$. Choose $q_1 > 0$ such that

1195

$$q_1 < c - b, \quad q_1 < q \text{ and } q_1 < \frac{c}{2}.$$

From (4.112) and (4.113), there exist positive integers M and $t > 0$ such that

1196

$$\left. \begin{aligned} \mu^M(\mathcal{H}_t) &> 1 - (c/2 - q_1), \\ \mu^M(B_s) &< c - q_1 \end{aligned} \right\}. \quad (4.115)$$

It is easily verified that

1197

$$\mathcal{H}_t^{-1} B_s \subset B_{s_0}, \quad s_0 = ds/t.$$

Consequently,

1198

$$B_s y^{-1} \cap \mathcal{H}_t \neq \emptyset \Rightarrow y \in \mathcal{H}_t^{-1} B_s \subset B_{s_0}.$$

Let $n > M$. Then

1199

$$\begin{aligned}\mu^n(B_s) &= \int \mu^M(B_s y^{-1}) \mu^{n-M}(dy) \\ &= \int_{B_{s_0}} + \int_{S-B_{s_0}} \\ &\leq (c - q_1) \mu^{n-M}(B_{s_0}) + \left(\frac{c}{2} - q_1\right) \left[1 - \mu^{n-M}(B_{s_0})\right] \\ &= \frac{c}{2} - q_1 + \frac{c}{2} \cdot \mu^{n-M}(B_{s_0}).\end{aligned}$$

It follows that for any $s \geq 1$,

1200

$$\lim_{n \rightarrow \infty} \mu^n(B_s) \leq \frac{c}{2} + \frac{c^2}{2} < b,$$

contradicting the definition of b . Thus $b = 1$, and given $\varepsilon > 0$, there exists $u \geq 1$ 1201
such that for all $n \geq 1$, 1202

$$\mu^n(B_u) > 1 - \varepsilon. \quad (4.116)$$

Choose the positive integers N and $\mu_0 > 0$ such that 1203

$$\mu^N(\mathcal{H}_{u_0}) > 1 - \varepsilon. \quad (4.117)$$

Now we can easily verify that the set $B_u \mathcal{H}_{u_0}^{-1}$ is relatively compact, and also for 1204
 $n > N$, 1205

$$\begin{aligned}1 - \varepsilon < \mu^n(B_u) &= \int \mu^{n-N}(B_u y^{-1}) \mu^N(dy) \\ &\leq \mu^{n-N}(B_u \mathcal{H}_{u_0}^{-1}) + \varepsilon.\end{aligned}$$

It follows that (μ^n) is tight. □

1206

Theorem 4.16. Let μ , S , and J be as in Theorem 4.15. Assume 1207

- (i) $\mu(J) = 0$. 1208
- (ii) $\mathcal{H} \neq \emptyset$, where $\mathcal{H}$ is given by 1209

$$\mathcal{H} = \left\{ x \in S : \text{for each } i, 1 \leq i \leq d, \max \left\{ x_{i,d+1}, \min_{1 \leq j \leq d} x_{ij} \right\} > 0 \right\}.$$

(iii) The sequence (μ^n) is tight. 1210

Then the sequence (μ^n) converges weakly to a probability measure λ , where $S_\lambda =$ 1211
 $m(S)$ ($\equiv$ the set of all matrices in S with the minimal rank) and $\lambda(\mathcal{H}) = 1$. 1212

Proof. Notice that for $n \geq 1$, $\mu^n(J) = 0$. $\mathcal{H}$ is easily verified to be a nonempty open ideal of $S - J$. It follows as in (4.55) that

$$\mu^n(\mathcal{H}) \uparrow 1 \text{ as } n \rightarrow \infty.$$

Assume that $w \notin S$, where

$$w = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

Notice that since $w \notin S$,

$$\mathcal{H} \cdot m(S) \cdot \mathcal{H} \subset \mathcal{H} \cap m(S). \quad (4.118)$$

By Condition (iii), the sequence

$$\frac{1}{n} \sum_{k=1}^n \mu^k$$

converges weakly to a probability measure λ on S such that $\lambda = \mu * \lambda = \lambda * \mu$ and $S_\lambda = m(S)$ (see Theorem 2.7 and Proposition 1.9). Since for each positive integer n ,

$$\lambda = \mu^n * \lambda * \mu^n,$$

it follows from (4.118) that

$$\lambda(\mathcal{H} \cap m(S)) \geq \mu^n(\mathcal{H})\lambda(m(S))\mu^n(\mathcal{H}) \rightarrow 1$$

as $n \rightarrow \infty$. It follows that $\lambda(\mathcal{H}) = 1$. Note that S_λ is a completely simple sub-semigroup (Theorem 2.7), and for $x \in \mathcal{H} \cap m(S)$, the set $xS_\lambda x = xm(S)x$ is a group containing x . Let e be the identity of this group. By (4.118),

$$e \in xm(S)x \subset \mathcal{H} \cap m(S). \quad (4.119)$$

Suppose that the rank of the matrices in $m(S)$ is greater than 1. Then e has rank greater than 1, and as such, e , a nonnegative idempotent matrix, cannot have a strictly positive column (see Theorem 1.11). Now if for some i , $1 \leq i \leq d$, $e_{i,d+1} > 0$, then since there exists j , $1 \leq j \leq d$, such that $e_{j,d+1} = 0$, it follows that $e_{ji} > 0$ since $e \in \mathcal{H}$ (see the definition of $\mathcal{H}$) and

$$e_{j,d+1} = (e \cdot e)_{j,d+1} \geq e_{ji}e_{i,d+1} > 0,$$

which is impossible. Thus, the first d elements in the $(d+1)$ th column of e are all zeros. Since $e \in \mathcal{H}$, it follows that e has the form

$$\begin{pmatrix} A & 0 \\ 0 & 1 \end{pmatrix}, \quad (4.120)$$

where A is a strictly positive $d \times d$ matrix. Since $e = e^2$, $A = A^2$; therefore by Theorem 1.11, the matrix has rank 1, and consequently e has rank 2. Now notice that if $x \in S$ and

$$x = \begin{pmatrix} B & C \\ 0 & 1 \end{pmatrix}, \quad C \neq 0.$$

Then exe is given by

$$exe = \begin{pmatrix} ABA & AC \\ 0 & 1 \end{pmatrix},$$

where $AC \neq 0$, since A is strictly positive and $C \neq 0$. Since

$$exe \in eSe = em(S)e = xm(S)x,$$

which is a group with identity e , it follows that there exists y in this group such that

$$y = \begin{pmatrix} D & E \\ 0 & 1 \end{pmatrix}$$

and $(exe)y = e$. Then

$$(exe)y = \begin{pmatrix} ABAD & ABAE + AC \\ 0 & 1 \end{pmatrix},$$

where $ABAE + AC \neq 0$, since $AC \neq 0$. This contradicts the representation of e as given in (4.120). Thus, every element in S has the form

$$\begin{pmatrix} B & 0 \\ 0 & 1 \end{pmatrix}, \quad B \neq 0, \quad (4.121)$$

since we have assume the element $w \notin S$.

Now $em(S)e = eS_\lambda e$, the group factor of S_λ , the support of an idempotent probability measure, must be a compact group. It is a compact group of strictly positive matrices, since if A is a strictly positive $d \times d$ matrix and B is a nonzero nonnegative $d \times d$ matrix, then ABA is a strictly positive $d \times d$ matrix. By Corollary 1.8, the set $eS_\lambda e$ must be a singleton. By Theorem 2.7, it follows that the sequence (μ^n) converges weakly to λ .

If the rank of the matrices in $m(S)$ is 1, then for $x, y \in m(S)$, $xy = x$; as a result, the set $em(S)e$ is again a singleton. As before it follows that $\mu^n \rightarrow \lambda$ weakly.

If we assume that the element $w \in S$, then the rank of the matrices in $m(S)$ is 1, and therefore, as we just argued, the sequence μ^n converges weakly to λ . $\square$

It is clear from Theorem 4.12, if μ is a probability measure on 2×2 nonnegative matrices, then (μ^n) is tight if and only if there exists either a common left eigenvector

$$x \equiv (x_1, x_2) \text{ with } x_1 > 0, \quad x_2 > 0$$

for every matrix in S_μ with common eigenvalue 1 or a common *right* eigenvector 1256

$$x \equiv (x_1, x_2) \text{ with } x_1 > 0, \quad x_2 > 0$$

for every matrix in S_μ with common eigenvalue 1. Notice that $xS_\mu = x$, where 1257
 $x_1 > 0, x_2 > 0$, implies $xS = x$ and S is compact, where 1258

$$S = \text{cl} \left(\bigcup_{n=1}^{\infty} S_\mu^n \right).$$

No such result holds for general 2×2 *real* matrices, though Theorem 4.17 is a result 1259
in that direction for real matrices. 1260

Let (X_n) be a sequence of 2×2 i.i.d. real matrices such that the distribution of 1261
 X_1 is μ . Write 1262

$$X_n = \begin{pmatrix} y_n & z_n \\ 0 & w_n \end{pmatrix}, \quad n \geq 1,$$

where $w_n = \pm 1$. According to Theorem 4.12, this is the most general situation, in 1263
the context of tightness for 2×2 real matrices if we ignore one exceptional situation. 1264

If $\Pr(y_1 = 0) > 0$, then the sequence (μ^n) is tight. The reason is the following: 1265
Let 1266

$$S = \text{cl} \left(\bigcup_{n=1}^{\infty} S_\mu^n \right)$$

and I be the set defined by 1267

$$I = \left\{ x \in S : x = \begin{pmatrix} 0 & a \\ 0 & b \end{pmatrix} \text{ for } b = 1 \text{ or } -1 \text{ and some real } a \right\}.$$

Notice that I is a closed ideal of S and if 1268

$$\mu(I) = \Pr(X_1 \in I) > 0,$$

then by Proposition 2.7, 1269

$$\lim_{n \rightarrow \infty} \mu^n(I) = 1.$$

Therefore given $\varepsilon > 0$, there is a positive integer N and a positive number r such 1270
that 1271

$$1 - \varepsilon < \mu^N \left\{ x \in S : x = \begin{pmatrix} 0 & a \\ 0 & b \end{pmatrix}, |a| \leq r \text{ and } b = \pm 1 \right\}.$$

Let us call the preceding set on the right B . Note that if $b = \pm 1, d = \pm 1$, then 1272

$$\begin{pmatrix} 0 & a \\ 0 & b \end{pmatrix} \begin{pmatrix} 0 & c \\ 0 & d \end{pmatrix} = \begin{pmatrix} 0 & ad \\ 0 & bd \end{pmatrix}$$

so it is clear that $\mathcal{B} \cdot I \subset \mathcal{B}$. This implies

1273

$$\mu^{k+N}(\mathcal{B}) \geq \mu^N(\mathcal{B})\mu^k(I).$$

The tightness of (μ^n) follows from this.

1274

Let us now assume

1275

$$\left. \begin{aligned} \Pr(y_1 = 0) &= 0, \\ -\infty \leq E(\log |y_1|) &< 0, \\ E \log \max\{|z_1|, 1\} &< \infty. \end{aligned} \right\} \quad (4.122)$$

We claim that (μ^n) is tight by (4.122). To prove this claim, notice that

1276

$$\begin{aligned} X_1 X_2 \dots X_n &= \begin{pmatrix} y_1 y_2 \dots y_n z_1 \pm y_1 z_2 \pm \dots \pm y_1 y_2 \dots y_{n-1} z_n \\ 0 & w_1 w_2 \dots w_n \end{pmatrix} \\ &\equiv \begin{pmatrix} \xi_n & \eta_n \\ 0 & \zeta_n \end{pmatrix}, \text{ say.} \end{aligned}$$

It follows from (4.122) and the law of large numbers that for any $\beta > 0$,

1277

$$\Pr\left(\sum_{i=1}^n \log |y_i| < -\beta \text{ eventually}\right) = 1,$$

so that for any $\varepsilon > 0$,

1278

$$\Pr(|y_1 y_2 \dots y_n| < \varepsilon \text{ eventually}) = 1. \quad (4.123)$$

Let us now write

1279

$$\Phi_n = \log \max\{|z_n|, 1\}$$

and let

1280

$$F(x) = \Pr(\Phi_n \leq x).$$

Then for any $A > 0$, it follows from (4.122) that

1281

$$A \cdot \sum_{n=1}^{\infty} \Pr(\Phi_n > An)$$

$$= A \cdot \sum_{n=1}^{\infty} [1 - F(An)] \leq \int_0^{\infty} [1 - F(x)] (dx) = E(\Phi_1) < \infty.$$

Using the Borel–Cantelli lemma,

1282

$$\Pr\left(\frac{\Phi_n}{n} \leq A \text{ eventually}\right) = 1.$$

1283

This means

$$\lim_{n \rightarrow \infty} \sup \frac{\Phi_n}{n} = 0 \text{ almost surely.}$$

Thus with probability 1,

$$\begin{aligned} & \lim_{n \rightarrow \infty} \sup \sqrt[n]{|y_1 y_2 \dots y_{n-1} z_n|} \\ & \leq \lim_{n \rightarrow \infty} \sup e^{\frac{1}{n} \Phi_n} \cdot \exp \left(\frac{1}{n} \sum_{i=1}^{n-1} \log |y_i| \right) \\ & = e^{E \log |y_i|} < 1, \end{aligned}$$

by (4.122). Thus, $\Pr(\eta_n \text{ converges as } n \rightarrow \infty) = 1$. This and (4.123) immediately imply the tightness of the sequence (μ^n) .

Now notice that if we have

$$0 < E(\log |y_1|) \leq \infty, \quad (4.124)$$

then by the law of large numbers,

$$\Pr \left(\lim_{n \rightarrow \infty} |y_1 y_2 \dots y_n| = \infty \right) = 1,$$

so that (μ^n) cannot be tight. Let us now suppose that

$$\begin{aligned} & E |\log y_i| > -\infty, \\ & E \log \max \{|z_1|, 1\} = \infty. \end{aligned} \quad (4.125)$$

Then for any $A > 0$, with F and Φ_n as before,

$$\begin{aligned} & A + A \sum_{n=1}^{\infty} [1 - F(An)] \\ & \geq \int_0^{\infty} [1 - F(x)] dx = E(\Phi_1) = \infty \end{aligned}$$

so that for any $A > 0$,

$$\Pr \left(\frac{\Phi_n}{n} > A \text{ infinitely often} \right) = 1.$$

This implies

$$\lim_{n \rightarrow \infty} \sup \frac{\Phi_n}{n} = \infty \text{ almost surely} \quad (4.126)$$

and consequently from (4.125) and 4.126,

1295

$$\lim_{n \rightarrow \infty} \sup \sqrt[n]{|z_n y_1 y_2 \dots y_{n-1}|} \\ \geq \left(\lim_{n \rightarrow \infty} \sup e^{\frac{1}{n} \Phi_n} \right) \cdot e^{E \log |y_1|} = \infty \text{ almost surely.}$$

This means that given any $A > 0$,

1296

$$\Pr(|z_n y_1 y_2 \dots y_{n-1}| = |\eta_n - \eta_{n-1}| > A \text{ infinitely often}) = 1.$$

If (μ^n) is tight, then given $0 < \varepsilon < 1/2$, there exists $A > 0$ such that

1297

$$\Pr(|\eta_n| \leq A) > 1 - \varepsilon, \quad n \geq 1,$$

so that for $n > 1$,

1298

$$\Pr(|\eta_n - \eta_{n-1}| \leq 2A) > 1 - 2\varepsilon > 0,$$

which is a contradiction. Thus, we have proven Theorem 4.17.

1299

Theorem 4.17. *Let (X_n) be a sequence of 2×2 i.i.d. real matrices with common distribution μ . Suppose*

1300

1301

$$X_n = \begin{pmatrix} y_n & z_n \\ 0 & w_n \end{pmatrix}, \quad n \geq 1,$$

where $w_n = \pm 1$ almost surely. Then the sequence (μ^n) is tight if one of the following two conditions hold:

1302

1303

(i) $\Pr(y_1 = 0) > 0$;

1304

(ii) $\Pr(y_1 = 0) = 0$, $-\infty \leq E \log |y_1| < 0$ and $E \log \max\{|z_1|, 1\} < \infty$.

1305

Also if $E \log |y_1| > -\infty$ and the sequence (μ^n) is tight, then we must have:

1306

$$E \log \max\{|z_1|, 1\} < \infty.$$

The rest of Sect. 4.3 is devoted to connections between tightness for the sequence of products of i.i.d. random matrices with distribution μ and the existence of μ -invariant probability measures. It is of course known to us (see Theorem 2.7) that if (μ^n) is tight, then the sequence

1307

1308

1309

1310

$$\frac{1}{n} \sum_{k=1}^n \mu^k$$

converges weakly to a probability measure λ such that $\mu * \lambda = \lambda * \mu = \lambda$. It is the converse of this result that we are more interested in. Let us first look at the following general lemma, Lemma 4.6. This is a result of Guivarc'h and Raugi [85], generalizing an earlier result of Furstenberg [69].

1311

1312

1313

1314

Lemma 4.6. Let (X_n) be a sequence of i.i.d. random variables taking values in a topological semigroup S with distribution $\mu \in P(S)$. Let Y be a locally compact second-countable space such that there is a continuous map $(s, y) \rightarrow s \cdot y$ from $S \times Y \rightarrow Y$. Suppose that $\lambda \in P(Y)$ such that $\mu * \lambda = \lambda$. Then for almost all ω (in the sample space), there is a probability measure $\lambda_\omega \in P(Y)$ such that the two sequences of probability measures given by $(X_1(\omega)X_2(\omega) \dots X_n(\omega)s\lambda)$ and $(X_1(\omega)X_2(\omega) \dots X_n(\omega)\lambda)$ both converge weakly to λ_ω as $n \rightarrow \infty$ for almost all s in S with respect to the probability measure

$$\sum_{n=1}^{\infty} \frac{1}{2^n} \cdot \mu^n.$$

Here for any s in S ,

$$s\lambda(B) = \lambda\{y \in Y : s \cdot y \in B\} \text{ for } B \subset Y.$$

Moreover if f is any bounded Borel measurable real function on Y , then

$$\int f d\lambda = E \left[\int f d\lambda_\omega \right].$$

Proof. Let f be a bounded continuous function on Y . Define the function Φ on S by

$$\Phi(s) = \int f(sy)\lambda(dy). \quad (4.127)$$

Write:

$$Z_n \equiv X_1 X_2 \dots X_n.$$

Notice that

$$\begin{aligned} E(\Phi(Z_{n+1}) \mid X_1, X_2, \dots, X_n) &= \int \Phi(Z_n \cdot s) \mu(ds) \\ &= \int \left[\int f(Z_n s y) \lambda(dy) \right] \mu(ds) \\ &= \int f(Z_n y) \mu * \lambda(dy) \\ &= \int f(Z_n y) \lambda(dy) = \Phi(Z_n), \end{aligned}$$

so that $(\Phi(Z_n))$ is a martingale. Since Φ is bounded, $\Phi(Z_n)$ converges almost surely to some random variable $W(\Phi)$ and

$$E(W(\Phi)) = E(\Phi(Z_1)) = \int f d\lambda. \quad (4.128)$$

Now for any positive integers m, n ,

1330

$$\begin{aligned} & E |\Phi(Z_{m+n}) - \Phi(Z_m)|^2 \\ &= E [\Phi(Z_{m+n})]^2 + E [\Phi(Z_m)]^2 - 2E [\Phi(Z_{m+n}) \Phi(Z_m)] \\ &= E [\Phi(Z_{m+n})]^2 - E [\Phi(Z_m)]^2, \end{aligned}$$

since

1331

$$\begin{aligned} & E (\Phi(Z_{m+n}) \Phi(Z_m)) \\ &= E [E (\Phi(Z_{m+n}) \Phi(Z_m) | \Phi(Z_m))] \\ &= E \left[\Phi(Z_m) \int \Phi(Z_m s) \mu^n(ds) \right] \\ &= E \left[\Phi(Z_m) \int \left\{ \int f(Z_m s y) \lambda(dy) \right\} \mu^n(ds) \right] \\ &= E \left[\Phi(Z_m) \int f(Z_m y) \mu^n * \lambda(dy) \right] \\ &= E \left[\Phi(Z_m) \int f(Z_m y) \lambda(dy) \right] \\ &= E [\Phi(Z_m)]^2. \end{aligned}$$

Thus for any positive integer k ,

1332

$$\begin{aligned} & \sum_{m=1}^k E |\Phi(Z_{m+n}) - \Phi(Z_m)|^2 \\ &= \sum_{m=1}^k [E (\Phi(Z_{m+n}))^2 - E (\Phi(Z_m))^2] \\ &\leq 2n \cdot \|f\|, \end{aligned} \tag{4.129}$$

after canceling appropriate terms. Notice that for any real measurable function h

1333

$$\begin{aligned} & E [h(Z_{k+m})] \\ &= E (E (h(Z_{k+m}) | Z_k)) \\ &= E \left[\int h(Z_k s) \mu^m(ds) \right] \end{aligned}$$

so that for $n \geq 1$, we have

1334

$$\begin{aligned} & \sum_{m=1}^{\infty} E \left[\int |\Phi(Z_m s) - \Phi(Z_m)|^2 \mu^n(ds) \right] \\ &= \sum_{m=1}^{\infty} E |\Phi(Z_{m+n}) - \Phi(Z_m)|^2 \leq 2n \|f\|. \end{aligned}$$

This means

$$\sum_{m=1}^{\infty} |\Phi(Z_m s) - \Phi(Z_m)|^2 < \infty, \quad (4.130)$$

almost everywhere with respect to the measure

$$P \times \left(\sum_{n=1}^{\infty} \frac{1}{2^n} \mu^n \right),$$

where P is the probability measure. It follows from (4.129) and (4.130) that almost everywhere

$$\lim_{m \rightarrow \infty} \Phi(Z_m s) = \lim_{m \rightarrow \infty} \Phi(Z_m) \equiv I(f),$$

say, exists. Since the space of continuous functions with compact support on Y is a separable metric space in the usual sup norm, if (f_i) is dense in $C_c(Y)$, then

$$\lim_{m \rightarrow \infty} \int f_i(Z_m s y) d\lambda(y) = I(f_i), \quad i \geq 1, \text{ almost everywhere} \quad (4.131)$$

with respect to the measure

$$\left(P \times \sum_{n=1}^{\infty} \frac{1}{2^n} \mu^n \right).$$

Define $\lambda_{m,s}(\omega) \in P(Y)$ by

$$\lambda_{m,s}(\omega)(B) = \lambda \left((Z_m(\omega)s)^{-1} B \right).$$

It follows from (4.131) that almost everywhere, the weak* limit of $\lambda_{m,s}(\omega)$ is independent of s , and if we call this limit $\lambda(\omega)$, then from (4.130), we have

$$\lim_{m \rightarrow \infty} \int f_i d\lambda_{m,s}(\omega) = \int f_i d\lambda(\omega), \quad i \geq 1.$$

By (4.128), for $i \geq 1$,

$$E \left(\int f_i d\lambda(\omega) \right) = \int f_i d\lambda.$$

Then for any bounded Borel function f on Y , we have

$$E \left(\int f(y) d\lambda(\omega)(y) \right) = \int f(y) d\lambda(y),$$

which implies that the measure $\lambda(\omega)$ must be almost surely a probability measure.

□

Proposition 4.2 is due to P. Bougerol.

1349

Proposition 4.2. *Let (X_n) be a sequence of $d \times d$ i.i.d. random real matrices with distribution μ . Suppose*

1350
1351

$$S = \text{cl} \left(\bigcup_{n=1}^{\infty} S_{\mu}^n \right).$$

Suppose also that $\lambda \in P(R^d)$, $\mu * \lambda = \lambda$, and for any proper subspace V of R^d , $\lambda(V) < 1$. Then the sequence (μ^n) is tight. Moreover, if there is no nontrivial proper subspace V of R^d such that $x \cdot V \subset V$ for each x in S , then S is compact.

1352

1353

1354

Proof. Consider the given $\lambda \in P(R^d)$. Then $\mu * \lambda = \lambda$. By Lemma 4.6, if

1355

$$Z_n \equiv X_1 X_2 \dots X_n,$$

then the sequence of probability measures $(Z_n(\omega)\lambda)$ weakly converges; almost surely, to a random probability measure $\lambda(\omega)$ as $n \rightarrow \infty$. Suppose for some ω ,

1356

1357

$$(w) \lim_{n \rightarrow \infty} Z_n(\omega)\lambda = \lambda(\omega),$$

but

1358

$$\sup \{ \|Z_n(\omega)\| : n \geq 1 \} = \infty.$$

Since every element in the sequence $\|Z_n(\omega)\|^{-1} Z_n(\omega)$ has norm 1, there is a subsequence (n_k) such that

1359

1360

$$\lim_{k \rightarrow \infty} \|Z_{n_k}(\omega)\| = \infty \text{ and } \lim_{k \rightarrow \infty} \|Z_{n_k}(\omega)\|^{-1} Z_{n_k}(\omega) = H(\omega),$$

where $H(\omega)$ is a nonzero matrix with norm 1.

1361

Let $u \in R^d$ such that $H(\omega)u \neq 0$. Then since

1362

$$\lim_{k \rightarrow \infty} \|Z_{n_k}(\omega)\|^{-1} \|Z_{n_k}(\omega) \cdot u\| = \|H(\omega)u\|$$

is positive,

1363

$$\lim_{k \rightarrow \infty} \|Z_{n_k}(\omega) \cdot u\| = \infty.$$

Thus for any continuous f with compact support on R^d ,

1364

$$\begin{aligned} \int f(y) d\lambda(\omega)(y) &= \lim_{k \rightarrow \infty} \int f(y) d(Z_{n_k}(\omega)\lambda)(y) \\ &= \lim_{k \rightarrow \infty} \int f(Z_{n_k}(\omega)y) d\lambda(y) \\ &= \lim_{k \rightarrow \infty} \int I_{K_\omega}(y) \cdot f(Z_{n_k}(\omega)y) d\lambda(y), \end{aligned}$$

where

$$K_\omega = \{y \in R^d : H(\omega)y = 0\}.$$

This implies $\lambda(K_\omega) = 1$. But this contradicts our assumption on λ , since K_ω is a proper subspace of R^d . Thus, it follows that almost surely the sequence $(Z_n(\omega))$ is bounded. Then the sequence (μ^n) must be tight.

Now by Lemma 4.6, for almost all s in S with respect to $\sum_{n=1}^{\infty} \frac{1}{2^n} \mu^n$,

$$(w) \lim_{n \rightarrow \infty} Z_n(\omega)s\lambda = \lambda(\omega), \quad (4.132)$$

almost surely. Since $(Z_n(\omega))$ is bounded almost surely, (4.132) holds for every $s \in S$ almost surely. If $\|s\| < 1$ for some $s \in S$, then the zero matrix is in S and it follows from (4.132) that $\lambda(\omega)$, almost surely, is the unit mass at the zero element in R^d , which contradicts $\lambda = E(\lambda(\omega))$, (see (4.128)).

To prove the last part of Proposition 4.2, assume there is no proper subspace V of R^d such that for each $x \in S$, $x \cdot V \subset V$. Suppose S is not compact. Then there exists $s_n \in S$ such that

$$\lim_{n \rightarrow \infty} \|s_n\| = \infty,$$

but

$$\lim_{n \rightarrow \infty} \|s_n\|^{-1} s_n = s_0,$$

where s_0 is a nonzero matrix. Then almost surely

$$\lambda(\omega) = Z(\omega)s s_n \lambda,$$

for $n \geq 1$, where $Z(\omega)$ is a limit point of the sequence $(Z_n(\omega))$. Suppose $y \in R^d$ such that $Z(\omega) \cdot s s_0 y \neq 0$. Then

$$\lim_{n \rightarrow \infty} \|Z(\omega)s s_n y\| = \infty.$$

Then if f is a continuous function with compact support in R^d ,

$$\begin{aligned} \int f(y) \lambda(\omega)(dy) &= \int f(Z(\omega)s s_n y) \lambda(dy) \\ &= \lim_{n \rightarrow \infty} \int I_V(y) f(Z(\omega)s s_n y) \lambda(dy), \end{aligned}$$

where

$$V = \{y \mid Z(\omega)s s_0 y = 0\}.$$

It follows that $\lambda(V) = 1$ and $V = R^d$. If $y_0 \in R^d$ is such that $s_0 y_0 \neq 0$, then

$$\{s s_0 y_0 \mid s \in S\}$$

spans a subspace W of R^d such that

1384

$$W \subset \{y \mid Z(\omega)y = 0\}.$$

Since $sW \subset W$ for $s \in S$, $W = R^d$ and consequently $Z(\omega)$ must be the zero matrix. This contradicts that S does not contain the zero matrix. $\square$

1385

1386

We conclude Sect.4.3 by considering the preceding problem in nonnegative matrices.

1387

1388

Proposition 4.3. *Let μ be a probability measure on $d \times d$ nonnegative matrices and let*

1389

1390

$$S = \text{cl} \left(\bigcup_{n=1}^{\infty} S_{\mu}^n \right).$$

Suppose there exists a probability measure λ on $(R^+)^d$ such that: (i) $\lambda(\{0\}) < 1$ and (ii) $\mu * \lambda = \lambda$, where 0 is the zero vector in $(R^+)^d$. If S contains a strictly positive matrix, then the sequence (μ^n) is tight and $\lambda = \beta * \sigma$, where

1391

1392

1393

$$\beta = (w) \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \mu^k \text{ and } \sigma \in P \left((R^+)^d \right).$$

Proof. Suppose

1394

$$0 < b \equiv \lambda(\{0\}) < 1.$$

Consider the probability measure

1395

$$\lambda_0 = \frac{(\lambda - b\delta_0)}{(1 - b)}.$$

Then $\lambda(\{0\}) = 0$ and $\mu * \lambda_0 = \lambda_0$. Thus, we may assume $\lambda(\{0\}) = 0$. Because of this, given $\delta > 0$, there exists a positive integer N such that

1396

1397

$$\lambda(F_N) > 1 - \delta, \tag{4.133}$$

where

1398

$$F_N = \left\{ x \in (R^+)^d : \text{for some } i, 1 \leq i \leq d, \frac{1}{N} < x_i < N \right\}.$$

Since $\mu^n * \lambda = \lambda$ for $n \geq 1$,

1399

$$\begin{aligned} \lambda(F_N) &= \int \mu^n(F_N x^{-1}) \lambda(dx) \\ &\leq \mu^n(F_N F_N^{-1}) + \delta \end{aligned}$$

so that for $n \geq 1$,

$$\mu^n (F_N F_N^{-1}) > 1 - 2\delta. \quad (4.134) \quad 1400$$

It follows that given $\delta > 0$, there exists $u > 0$ such that for each positive integer n , 1401

$$\mu^n (H(u)) > 1 - \delta, \quad (4.135) \quad 1402$$

where

$$H(u) = \left\{ A \in S : \min_{i,j} A_{ij} \leq u \right\},$$

since

$$F_n F_n^{-1} \subset H(n^2), \quad n \geq 1. \quad 1403$$

The strictly positive matrices form an open subset of S , and thus, there exist 1404
positive integers k, m and a positive number t such that 1405

$$\mu^k (\mathcal{J}_m) = t, \quad (4.136)$$

where

$$\mathcal{J}_m = \left\{ A \in S : \min_{i,j} A_{ij} > \frac{1}{m} \right\}. \quad 1406$$

Let s be any positive integer and let (X_n) be a sequence of i.i.d. matrices with 1407
distribution μ . Write 1408

$$X_1 X_2 \dots X_n \equiv Z_n.$$

Choose $\delta > 0$ such that $\delta < t^2$. For this δ , let u be such that (4.135) holds. Then we 1409
have 1410

$$\begin{aligned} t^2 - \delta &\leq \Pr(Z_{2k+s} \in H(u), Z_k \in \mathcal{J}_m, X_{k+s+1} \dots X_{2k+s} \in \mathcal{J}_m) \\ &\leq \Pr(Z_k \in \mathcal{J}_m, X_{k+1} \dots X_{k+s} \in \mathcal{J}_m^{-1} H(u) \mathcal{J}_m^{-1}, X_{k+s+1} \dots X_{2k+s} \in \mathcal{J}_m) \\ &= t^2 \cdot \mu^s (\mathcal{J}_m^{-1} H(u) \mathcal{J}_m^{-1}), \end{aligned}$$

where

$$\begin{aligned} &\mathcal{J}_m^{-1} H(u) \mathcal{J}_m^{-1} \\ &= \{A \in S : BAC \in H(u) \text{ for some } B \text{ and } C \text{ in } \mathcal{J}_m\}, \end{aligned} \quad 1411$$

which is easily verified to be contained in the compact set 1412

$$\left\{ A \in S : \max_{i,j} A_{ij} \leq u \cdot m^2 \right\}.$$

It follows that the sequence (μ^n) is tight. The rest of the proof is left to the reader. $\square$ 1413

Our last result connecting μ invariance to tightness of (μ^n) is given in Proposition 4.4. 1414 1415

Proposition 4.4. *Let μ be a probability measure on nonnegative $d \times d$ matrices and* 1416 1417

$$S = \text{cl} \left(\bigcup_{n=1}^{\infty} S_{\mu}^n \right).$$

Suppose λ is a probability measure on $d \times d$ nonnegative matrices such that $\mu * \lambda = \lambda$. (We do not assume that $S_{\lambda} \subset S$.) Consider the following conditions: 1418 1419

- (i) $\lambda(J_r) = 0$, where J_r is the set of all matrices that have at least one zero row; 1420
- (ii) $S_{\lambda} S \subset S$; 1421
- (iii) $S_{\lambda} \subset S$. 1422

(Note that (iii) $\Rightarrow$ (ii). But (ii) in general is weaker than (iii).) If (i) holds, the sequence (μ^n) is tight. If (i) and (ii) hold, then the solution λ of $\mu * \lambda = \lambda$ must be an idempotent probability measure. If (i) and (iii) hold, then both λ_0 and λ , where 1423 1424 1425

$$\lambda_0 = (w) \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{m=1}^n \mu^m,$$

being idempotent probability measures, have product representations (see Theorem 2.8) that can be described as follows: Let e be an idempotent matrix of minimal rank in S . Then there exist $\alpha, \beta, \gamma, \gamma_1$ in $P(S)$ such that 1426 1427 1428

$$\lambda_0 = \alpha * \beta * \gamma$$

and 1429

$$\lambda = \alpha * \beta * \gamma_1,$$

where $S_{\beta} = eSe$ is a finite group and β is the uniform distribution on it, $S_{\gamma_1} \subset S_{\gamma} \subset$ the set of all idempotent matrices in $e \cdot S$, and $S_{\alpha} \subset$ the set of all idempotent matrices in $S \cdot e$. 1430 1431 1432

Proof. Suppose (i) holds. Given $\varepsilon > 0$, there is a compact set $K \subset J_r^c$ such that $\lambda(K) > 1 - \varepsilon$. Then $\mu^n * \lambda = \lambda$ implies 1433 1434

$$\lambda(K) \leq \int_K \mu^n(Kx^{-1}) \lambda(dx) + \varepsilon$$

so that $\mu^n(KK^{-1}) > 1 - \varepsilon$ for $n \geq 1$. Since $K \subset J_r^c$ and K is compact, it is easily verified that KK^{-1} is a compact set. This implies that (μ^n) is tight and 1435 1436

$$\lambda_0 = (w) \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{m=1}^n \mu^m$$

exists. Since

$$\mu * \lambda_0 = \lambda_0 * \mu = \lambda_0 = \lambda_0 * \lambda_0,$$

we have $\lambda_0 * \lambda = \lambda$ and S_{λ_0} is the completely simple (minimal) ideal of S .

Now suppose (i) and (ii) hold. Then by (ii),

$$S_{\lambda} S_{\lambda_0} \subset S_{\lambda} (\overline{S S_{\lambda_0}}) \subset S_{\lambda_0}.$$

Let us write: $\lambda * \lambda_0 \equiv \lambda_1$. Then it easily follows that

$$\left. \begin{aligned} \lambda_1 * \lambda_0 &= \lambda_1 = \lambda_0 * \lambda_1, \\ S_{\lambda_1} &\subset S_{\lambda_0} \end{aligned} \right\}. \quad (4.137)$$

Then (4.137) implies

$$\begin{aligned} S_{\lambda_1} \cdot S_{\lambda_1} &\subset S_{\lambda_0} \cdot S_{\lambda * \lambda_0} \\ &\subset \overline{S_{\lambda_0} S_{\lambda} \cdot S_{\lambda_0}} \subset \overline{S_{\lambda_0 * \lambda} S_{\lambda_0}} \\ &\subset \overline{S_{\lambda} \cdot S_{\lambda_0}} = S_{\lambda_1}. \end{aligned}$$

It follows that S_{λ_1} is an ideal of S_{λ_0} . Therefore, since S_{λ_0} is simple, $S_{\lambda_1} = S_{\lambda_0}$.

Using (4.137) and Proposition 2.4, for any Borel set $B \subset S$ and $x \in S_{\lambda_0}$, we have

$$\begin{aligned} \lambda_1(Bx^{-1}) &= \int \lambda_0(Bx^{-1}y^{-1}) \lambda_1(dy) \\ &= \int \lambda_0(Bx^{-1}) \lambda_1(dy) = \lambda_0(Bx^{-1}). \end{aligned}$$

Consequently,

$$\begin{aligned} \lambda_1(B) &= \int \lambda_1(Bx^{-1}) \lambda_0(dx) \\ &= \int \lambda_0(Bx^{-1}) \lambda_0(dx) = \lambda_0(B), \end{aligned}$$

which means $\lambda * \lambda_0 = \lambda_1 = \lambda_0$. Thus,

$$\lambda_0 * \lambda = \lambda, \quad \lambda * \lambda_0 = \lambda_0 \text{ and } \lambda_0 = \lambda_0 * \lambda_0.$$

This implies

$$\begin{aligned} \lambda &= \lambda_0 * \lambda = (\lambda * \lambda_0) * (\lambda_0 * \lambda) \\ &= \lambda * (\lambda_0 * \lambda) = \lambda * \lambda. \end{aligned}$$

Thus, we have proven that under (i) and (ii), the solution λ of $\mu * \lambda = \lambda$ must be an idempotent probability measure. We leave the rest of the proof to the reader. $\square$

Section 4.3. Exercises

1449

Exercise 4.16. Establish (4.49).

1450

Exercise 4.17. Explain why the set $\mathcal{C} \mathcal{J}_k^{-1}$ in the proof of Theorem 4.8 is relatively compact.

1451

1452

Exercise 4.18. Let μ be a probability measure on $d \times d$ real matrices such that μ^m {the zero matrix} is positive for some positive integer m . Show that μ^n converges weakly to the unit mass at the zero matrix. Does this mean $\mu\{0\} > 0$? If not, give an example.

1453

1454

1455

1456

Exercise 4.19. Let (X_n) be a sequence of i.i.d. nonnegative random variables such that either $P(X_1 = 0) > 0$ holds or both conditions $P(X_1 = 0) = 0$ and $E(\log X_1) < 0$ hold. Show that the sequence $X_1 X_2 \dots X_n$ must then converge to zero almost surely.

1457

1458

1459

1460

Exercise 4.20. Let μ be a probability measure on $d \times d$ nonnegative matrices and S be the closed multiplicative semigroup generated by $S(\mu)$. Define the sets A_c and A_r as follows:

1461

1462

1463

Let

1464

$$D = \{1, 2, \dots, d\}, \quad A_c \subset D, \quad A_r \subset D$$

such that

1465

$$A_c = \{i \in D \mid \text{if } x \in m(S), \text{ then the } i\text{th column of } x \text{ is a zero column}\},$$

and

1466

$$A_r = \{j \in D \mid \text{if } x \in m(S), \text{ then the } j\text{th column of } x \text{ is a zero row}\}.$$

Let $A_c \cap A_r = A_{cr}$, $A_c \setminus A_r = A_{co}$, $A_r \setminus A_c = A_{ro}$, $B = D - (A_c \cup A_r)$. Suppose that the zero matrix is not in S . Show that then $A_c \cup A_r$ is a proper subset of D , and if (μ^n) is tight, then every element in S has the following form:

1467

1468

1469

	A_{co}	B	A_{cr}	A_{ro}
A_{co}				
B	0	$< M$	0	
A_{cr}	0	0		
A_{ro}	0	0	0	

1470

Specifically, there exists $M > 0$ such that if $x \in S$, then the $B \times B$ block of x has each of its entry bounded above by M . (This result is taken from [158].)

1471

1472

(Hint: Let $\lambda = (w) \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \mu^k$. Then $S(\lambda) = m(S)$ is completely simple 1473
with a compact group factor, and as such, for any x, y in $m(S)$, $xSy = xm(S)y$ 1474
is compact. Also, note that for any x in $m(S)$ and $s \in S$, $xs \in m(S)$, and thus, for 1475
 $i \in A_c$, $(xs)_{ui} = 0$ for all u in D .) 1476

Exercise 4.21. Let (X_n) be a sequence of $d \times d$ random i.i.d stochastic matrices 1477
such that the products $X_n X_{n-1} \dots X_1$ converge in distribution to a random stochas- 1478
tic matrix Z with identical rows. Let μ and λ , respectively, be the distributions of 1479
 X_1 and Z (that is, $\mu(B) = P(X_1 \in B)$ and $\lambda(B) = P(Z \in B)$, for any Borel set 1480
 B of $d \times d$ stochastic matrices). Show that if β is a probability measure on $d \times d$ 1481
stochastic matrices such that $\beta * \mu = \beta$, then β must be λ . 1482

Exercise 4.22. Let (X_n) be as in Exercise 4.21. Suppose that $P(X_1$ is strictly 1483
positive) > 0 . Then show that the sequence $X_n X_{n-1} \dots X_1$ converges almost surely 1484
to a $d \times d$ random stochastic matrix with identical rows. This result is due to 1485
Chamayou and Letac. 1486

(Hint: Let s be a $d \times 1$ positive vector and let $s_n = X_n X_{n-1} \dots X_1 s$. Write: 1487
 $v_n = \max_i (s_n)_i$, $u_n = \min_i (s_n)_i$ and $M_n = \min_{i,j} (X_n)_{ij}$. Note that almost surely, 1488
 $\sum_{k=1}^{\infty} M_k = \infty$ (using the fact that $P(M_1 > \delta) > 0$ for some $\delta > 0$). Show that 1489

$$u_{n-1} + M_n (v_{n-1} - u_{n-1}) \leq u_n \leq v_n \leq v_{n-1} - M_n (v_{n-1} - u_{n-1}).$$

This leads to $\lim_{n \rightarrow \infty} v_n - u_n = 0$ almost surely.) 1490

Exercise 4.23. Suppose that $d (> 1)$ committee members are evaluating an appli- 1491
cant for a teaching position at a university. Let η_0 be a $d \times 1$ vector, $(\eta_0)_i > 0$ 1492
for each i , representing their initial evaluations. The members then start discussing 1493
the applicant in a number of subsequent meetings to influence each other, to revise 1494
their scores from the previous meeting, and to reach a consensus. These result in a 1495
sequence of i.i.d. $d \times d$ stochastic matrices (X_n) such that $\eta_n = X_n \eta_{n-1}$, $n \geq 1$, 1496
where $(\eta_n)_i$ is the score that the i th member arrives at after the n th meeting. If 1497
 $P(X_1 > 0) > 0$, show that $\eta = \lim_{n \rightarrow \infty} \eta_n$ exists, and η_i is independent of i . 1498

Exercise 4.24. Let $X_1, X_2, \dots$ be a sequence of i.i.d. 2×2 random nonnegative 1499
matrices such that $X_1 X_2 \dots X_n \equiv Y_n$, X_1 has distribution μ , 1500

$$X_n = \begin{pmatrix} A_n & B_n \\ 0 & 1 \end{pmatrix}, \quad Y_n = \begin{pmatrix} Z_n & W_n \\ 0 & 1 \end{pmatrix},$$

and μ is not the unit mass at the identity matrix. Show that (μ^n) is tight iff the 1501
following two conditions hold: (i) Z_n converges to zero in probability, and (ii) W_n 1502
converges almost surely. 1503

Exercise 4.25. Show that in the previous exercises, when $P(A_1 = 0) = 0$ and $-\infty < E(\log A_1) < 0$, then the sequence W_n is almost surely convergent iff $E(\log \sup \{1, B_1\}) < \infty$. Show also that when $P(A_1 = 0) > 0$, then the sequence Z_n converges to 0 almost surely. (See [79].)

Exercise 4.26. Consider i.i.d. 2×2 random nonnegative matrices (X_n) given by

$$X_n = \begin{pmatrix} A_n & B_n \\ C_n & D_n \end{pmatrix},$$

such that the distribution of X_1 is μ . Let S be, as usual, the closed semigroup generated by $S(\mu)$. Let $Y_n = X_1 X_2 \dots X_n$, $n \geq 1$. Suppose that (μ^n) is tight. Show that one of the following possibilities must then occur:

- (i) S is compact;
- (ii) Y_n converges in distribution to the unit mass at a matrix which is either $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ or $\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$ or $\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$;
- (iii) Either $P(A_n = 1, B_n = 0) = 1$ or $P(A_n = 1, C_n = 0) = 1$ or $P(C_n = 0, D_n = 1) = 1$ or $P(B_n = 0, D_n = 1) = 1$.

(This is taken from [158].)

(Hint: Use Exercises 4.20 and 4.24.)

4.4 Invariant Measures for Random Walks in Nonnegative Matrices: Laws of Large Numbers

Mixed random walks in the general context are discussed in Chap. 3. Here, we study mixed random walks with values in nonnegative matrices. First, we consider the existence and uniqueness of invariant measures for such walks, then we apply these results to obtain laws of large numbers for such walks. The main results in this section are Theorems 4.19–4.23.

Let (Y_n) be a sequence of i.i.d. random variables such that for each $n \geq 1$ and for some α , $0 \leq \alpha \leq 1$,

$$P(Y_n = 1) = \alpha \text{ and } P(Y_n = 0) = 1 - \alpha.$$

Consider a mixed walk (M_n) , where

$$\begin{aligned} M_{n+1} &= M_n X_{n+1}, \text{ if } Y_{n+1} = 1; \\ &= Z_{n+1} \cdot M_n, \text{ if } Y_{n+1} = 0, \end{aligned} \tag{4.138}$$

for $n \geq 0$, where

$$(X_i) \text{ and } (Z_i)$$

are two i.i.d. sequences of $d \times d$ random nonnegative matrices. The X_i s, Y_i s, and Z_i s are independent of one another with $M_0 \equiv X_0$. Let μ_1 be the distribution of X_i , $i \geq 0$ and μ_2 be the distribution of Z_i , $i \geq 1$. Let us assume the following conditions:

- (i) $S_{\mu_1} = S_{\mu_2}$;
- (ii) $\mu_1(J) = 0 = \mu_2(J)$, where J is the set of all $d \times d$ matrices with at least one zero row or one zero column;
- (iii) $S \cap \mathcal{J} \neq \emptyset$, where

$$S = \text{cl} \left(\bigcup_{n=1}^{\infty} S_{\mu_1}^n \right)$$

- and $\mathcal{J}$ is the set of all $d \times d$ strictly positive matrices;
- (iv) (μ_1^n) and (μ_2^n) are both tight.

We are considering the usual topology for matrices. It follows from Theorem 4.10 that under the preceding conditions, the (multiplicative) semigroup S defined in Condition (iii) is compact when S does not contain the zero matrix; however, in any case, there are probability measure β_1 and β_2 such that

$$\mu_1^n \rightarrow \beta_1 \text{ and } \mu_2^n \rightarrow \beta_2 \quad (4.139)$$

weakly as $n \rightarrow \infty$. Let K be the kernel (that is, the smallest ideal) of S . Then we know (Theorem 2.7) that K is closed and it is the support of both β_1 and β_2 .

Let $P^n(x, A)$ be the n -step transition probability function for the mixed random walk (M_n) . Then it is clear that for x in S ,

$$\left. \begin{aligned} P(x, A) &= \alpha \delta_x * \mu_1(A) + (1 - \alpha) \mu_2 * \delta_x(A) \text{ and for } n > 1, \\ P^n(x, A) &= \sum_{k=0}^n \binom{n}{k} \alpha^k (1 - \alpha)^{n-k} \mu_2^{n-k} * \delta_x * \mu_1^k(A) \end{aligned} \right\}. \quad (4.140)$$

Let us define the Markov operator T by

$$Tf(x) = \int f(y) P(x, dy),$$

where $f \in C(S) \equiv$ the space of all real continuous functions on S . Then for $n \geq 1$,

$$T^n f(x) = \sum_{k=0}^n \binom{n}{k} \alpha^k (1 - \alpha)^{n-k} \iint f(yxz) \mu_2^{n-k}(dy) \mu_1^k(dz). \quad (4.141)$$

Now suppose $0 \notin S$. Since then S is compact, $\mu_1^n \rightarrow \beta_1$ weakly and $\mu_2^n \rightarrow \beta_2$ weakly, it follows that given $\varepsilon > 0$, there exists a positive integer N such that for $n > N$, for each $y \in S$, we have

$$\left| \int f(yxz) \mu_1^n(dz) - \int f(yxz) \beta_1(dz) \right| < \varepsilon$$

and for each $z \in S$,

1553

$$\left| \int f(yxz) \mu_2^n(dy) - \int f(yxz) \beta_2(dy) \right| < \varepsilon.$$

We can also choose a positive integer N_2 such that for $n > N_2$,

1554

$$\sum_{k=N}^{n-N} \binom{n}{k} \alpha^k (1-\alpha)^{n-k} > 1 - \varepsilon.$$

Then it follows from (4.141) that for each x in S and $f \in C(S)$,

1555

$$\lim_{n \rightarrow \infty} T^n f(x) = \bar{T} f(x), \quad (4.142)$$

where

1556

$$\begin{aligned} \bar{T} f(x) &= \iint f(yxz) \beta_2(dy) \beta_1(dz) \\ &= \int f(w) \beta_2 * \delta_x * \beta_1(dw). \end{aligned}$$

Let us write $\lambda_x \equiv \beta_2 * \delta_x * \beta_1$. Then $S_{\lambda_x} = S_{\beta_2} \cdot x \cdot S_{\beta_1} = K \cdot x \cdot K$ for each x 1557
in S . Since $KxK \subset K$, it is clear that $KxK = K$. Now for $f \in C(S)$, 1558

$$\int f(y) \lambda_x(dy) = \lim_{n \rightarrow \infty} \int f(y) P^n(x, dy). \quad (4.143)$$

Notice that for $f \in C(S)$,

1559

$$\bar{T} f = \bar{T}(\bar{T} f), \quad \bar{T} f = \bar{T}(Tf). \quad (4.144)$$

Then for $f \in C(S)$,

1560

$$\int f(y) \lambda_x(dy) = \int \left[\int f(z) P(y, dz) \right] \lambda_x(dy). \quad (4.145)$$

Since a probability measure on a compact (or a complete separable) metric space is 1561
always regular, the probability measure ν defined by 1562

$$\nu(A) = \int P(y, A) \lambda_x(dy)$$

is regular. Therefore $\nu = \lambda_x$, since for $f \in C(S)$,

1563

$$\int f d\nu = \int \left[\int f(z) P(y, dz) \right] \lambda_x(dy) = \int f d\lambda_x.$$

Thus for any $x \in S$ and any Borel set $A \subset S$,

1564

$$\lambda_x(A) = \int P(y, A) \lambda_x(dy). \quad (4.146)$$

Now let us show that for any x and y in K , $\lambda_x = \lambda_y$. To prove this, it is enough to show that for any $f \in C(S)$, $\overline{T}f$ is a constant on K . Let $a = b + \delta$, $\delta > 0$,

1565

1566

$$a = \max_{x \in K} \overline{T}f(x) = \overline{T}f(x_1), \quad x_1 \in K,$$

and

1567

$$b = \min_{x \in K} \overline{T}f(x) = \overline{T}f(x_2), \quad x_2 \in K.$$

Then

1568

$$\begin{aligned} a &= \overline{T}f(x_1) = \overline{T}^2 f(x_1) = \int \overline{T}f(y) \lambda_{x_1}(dy) \\ &= \int_{N(x_2)} + \int_{K_N(x_2)}, \end{aligned}$$

where $N(x_2)$ is an open set containing x_2 such that for $y \in N(x_2)$,

1569

$$\overline{T}f(y) < \frac{1}{2}(a + b).$$

(Recall that K is the support of λ_x , for each $x \in S$.) Now it follows that

1570

$$a \leq \frac{a+b}{2} \lambda_{x_1}(N(x_2)) + [1 - \lambda_{x_1}(N(x_2))] \cdot a$$

or

1571

$$a \leq a - \lambda_{x_1}(N(x_2)) \left[\frac{a-b}{2} \right] < a,$$

since $\lambda_{x_1}(N(x_2)) > 0$. This means $\overline{T}f(x)$ (and therefore the measure λ_x) is the same for all $x \in K$.

1572

1573

Now using the second relationship in (4.144), for $x \in S$, we have

1574

$$\begin{aligned} \overline{T}f(x) &= \int \overline{T}f(y) \lambda_x(dy) \\ &= \int_K \overline{T}f(y) \lambda_x(dy), \end{aligned}$$

since K is the support of λ_x . It follows that $\overline{T}f(x)$ is the same for all $x \in S$, and consequently for any x and y in S , $\lambda_x = \lambda_y$.

1575

1576

Now let λ be any invariant probability measure, so that for any Borel set $A \subset S$, 1577

$$\lambda(A) = \int P(y, A) \lambda(dy). \quad (4.147)$$

It follows from (4.147) that for any f in $C(S)$ and any positive integer n , 1578

$$\int f d\lambda = \int \left[\int f(x) P^n(y, dz) \right] \lambda(dy)$$

and by (4.143), taking n to infinity, we have 1579

$$\begin{aligned} \int f d\lambda &= \int \left[\int f(z) \lambda_y(dz) \right] \lambda(dy) \\ &= \int f d\lambda_y, \end{aligned}$$

since for all $y \in S$, λ_y is the same. It follows that $\lambda = \lambda_y$. Thus, we have proven 1580
Theorem 4.18, when $0 \notin S$. 1581

Theorem 4.18. Let 1582

$$(X_n)_{n=0}^{\infty} \text{ and } (Z_n)_{n=1}^{\infty}$$

be two sequences of i.i.d. random $d \times d$ nonnegative matrices such that the distribu- 1583
tion of each X_n is μ_1 and that of each Z_n is μ_2 . Let S_{μ_1} be the support of μ_1 and S 1584
the closed (multiplicative) semigroup generated by S_{μ_1} with usual matrix topology. 1585
Let us assume Conditions (i)–(iv) given at the beginning of this section. Let $(Y_n)_{n=1}^{\infty}$ 1586
be a sequence of i.i.d. random variables such that for each $n \geq 1$, 1587

$$P(Y_n = 1) = \alpha, \quad P(Y_n = 0) = 1 - \alpha,$$

where $0 \leq \alpha \leq 1$. Assume the X_i s, Y_i s and Z_i s are all independent of one another. 1588
Consider the mixed walk $(M_n)_{n=1}^{\infty}$ as defined in (4.138). Then the mixed walk has a 1589
unique invariant probability measure λ given by 1590

$$\lambda = \beta_2 * \delta_x * \beta_1,$$

where $x \in S$ and $\mu_i^n \rightarrow \beta_i$ ($i = 1, 2$) weakly as $n \rightarrow \infty$. $\square$ 1591

Theorem 4.20 treats the case when $0 \in S$. But first we have the following 1592
theorem. 1593

Theorem 4.19. Let μ be a probability measure on a locally compact Hausdorff 1594
second-countable topological semigroup S such that 1595

$$S = \text{cl} \left(\bigcup_{n=1}^{\infty} S_{\mu}^n \right).$$

Suppose (μ^n) is tight and in this case, the weak limit β , given by

1596

$$\beta = (w) \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \mu^k,$$

exists. Consider the S -valued mixed random walk (S_n) induced by an i.i.d. sequence of S -valued random variables (X_i) with distribution μ as introduced in (4.138), with $Z_i = X_i$ almost surely for each i . The transition probability function $P(x, B)$ of this random walk is given by

1597
1598
1599
1600

$$P(x, B) = \alpha \mu(x^{-1}B) + (1 - \alpha) \mu(Bx^{-1}), \quad (4.148)$$

where

1601

$$0 \leq \alpha \leq 1, \quad Bx^{-1} = \{y \in S : yx \in B\} \text{ and } x^{-1}B = \{y \in S : xy \in B\}.$$

Let λ be an invariant probability measure on S for (S_n) , so that for $B \subset S$,

1602

$$\lambda(B) = \int P(x, B) \lambda(dx). \quad (4.149)$$

Then λ satisfies the following conditions:

1603

- (i) S_λ , the support of λ , is the smallest ideal of S , so that $S_\lambda = S_\beta$ and
- (ii) $\beta = \beta * \lambda = \lambda * \beta$ and λ^n converges weakly to β .

1604
1605

Proof. Let us prove (i) first. From (4.148) and (4.149),

1606

$$\lambda = \alpha \lambda * \mu + (1 - \alpha) \mu * \lambda. \quad (4.150)$$

This implies that supports of $\lambda * \mu$ and $\mu * \lambda$ are both contained in S_λ . Therefore,

1607

$$S_\lambda \cdot S_\mu \cup S_\mu \cdot S_\lambda \subset S_\lambda, \quad (4.151)$$

so that

1608

$$S_\lambda \cdot S \cup S \cdot S_\lambda \subset S_\lambda$$

and S_λ is an ideal of S . Since S_β is the smallest ideal of S ,

1609

$$S_\beta \subset S_\lambda. \quad (4.152)$$

To prove the converse inclusion, let us observe that by repeated applications of (4.150), for $n \geq 1$, we have

1610
1611

$$\lambda = \sum_{k=0}^n \binom{n}{k} (1 - \alpha)^{n-k} \alpha^k \mu^{n-k} * \lambda * \mu^k. \quad (4.153)$$

4.4 Invariant Measures for Random Walks in Nonnegative Matrices: Laws of Large Numbers 339

Let $\varepsilon > 0$ and V be any open set containing S_β . Since (μ^n) is tight, there is a compact set $E \subset S$ such that for $n \geq 1$,

$$\mu^n(E) > 1 - \varepsilon. \quad (4.154)$$

Now notice that

$$E \cdot S_\lambda \cdot S_\beta \cdot E \subset S_\beta \subset V,$$

since S_β is an ideal of S . Let A be a compact subset of S_λ so that $\lambda(A) > 1 - \varepsilon$ and W be an open set containing S_β so that

$$E \cdot A \cdot W \cdot E \subset V.$$

Since

$$\beta = (w) \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \mu^k,$$

W is open and $W \supset S_\beta$, there exists a positive integer N such that

$$\mu^N(W) > 1 - \varepsilon.$$

Then for integers k and s greater than N ,

$$\begin{aligned} & \mu^k * \lambda * \mu^s(V) \\ & \leq \mu^k(E) \lambda(A) \mu^N(W) \mu^{s-N}(E) \\ & > (1 - \varepsilon)^4. \end{aligned} \quad (4.155)$$

Now for a sufficiently large n and $n > 2N + 2$,

$$\sum_{k=N+1}^{n-N-1} \binom{n}{k} \alpha^k (1 - \alpha)^{n-k} > 1 - \varepsilon. \quad (4.156)$$

It follows from (4.153), (4.155), and (4.156) that

$$\lambda(V) > (1 - \varepsilon)^6.$$

Since $\varepsilon > 0$ is arbitrary, this proves

$$S_\lambda \subset S_\beta. \quad (4.157)$$

Assertions (4.152) and (4.157) establish (i).

To prove (ii), from (4.150) we have, for $n \geq 1$,

$$\lambda * \mu^n = \alpha \lambda * \mu^{n+1} + (1 - \alpha) \mu * (\lambda * \mu^n).$$

Taking average and then weak limit as n tends to infinity, we then have

1625

$$\lambda * \beta = \alpha \lambda * \beta + (1 - \alpha) \mu * (\lambda * \beta)$$

or

1626

$$\lambda * \beta = \mu * (\lambda * \beta). \quad (4.158)$$

Similarly for $n \geq 1$, we can have

1627

$$\mu^n * \lambda = \alpha (\mu^n * \lambda) * \mu + (1 - \alpha) \mu^{n+1} * \lambda$$

and as above

1628

$$\beta * \lambda = (\beta * \lambda) * \mu. \quad (4.159)$$

Now (4.158) gives

1629

$$\lambda * \beta = \mu^n * (\lambda * \beta)$$

and consequently $\lambda * \beta = \beta * (\lambda * \beta)$. Similarly from (4.159),

1630

$$\beta * \lambda = (\beta * \lambda) * \beta.$$

Therefore,

1631

$$\lambda * \beta = \beta * \lambda = \beta * \lambda * \beta. \quad (4.160)$$

Since $\beta = \beta * \beta$, for $B \subset S$,

1632

$$\beta(Bx^{-1}) = \beta(Bx^{-1}y^{-1}), \quad (4.161)$$

whenever x, y are in S_β (see Proposition 2.4). Since $S_\beta = S_\lambda$, it follows from (4.160) and (4.161) that for $x \in S_\beta$,

1633

1634

$$\beta * \lambda(Bx^{-1}) = \int \beta(Bx^{-1}y^{-1}) \lambda(dy) = \beta(Bx^{-1}).$$

Therefore,

1635

$$\begin{aligned} \beta(B) &= \beta * \beta(B) = \int \beta(Bx^{-1}) \beta(dx) \\ &= \int \beta * \lambda(Bx^{-1}) \beta(dx) \\ &= \beta * \lambda * \beta(B). \end{aligned}$$

Thus, we have

1636

$$\beta = \beta * \lambda = \lambda * \beta. \quad (4.162)$$

4.4 Invariant Measures for Random Walks in Nonnegative Matrices: Laws of Large Numbers 341

Now we prove the final assertion: λ^n weakly converges to β as $n \rightarrow \infty$. We first observe that $S_\lambda = S_\beta$ is the completely simple (minimal) ideal of S with a compact group factor (see Theorem 2.2). Thus, we can write

$$S_\lambda = X \cdot G \cdot Y, \quad Y \cdot X \subset G,$$

where G is a compact group. If $A \subset X$ and $B \subset Y$ are compact subsets such that

$$\lambda(AGB) > 1 - \varepsilon,$$

then for $n \geq 3$,

$$\lambda^n(A \cdot G \cdot B) \geq \lambda(AGB)\lambda^{n-2}(XGY) \cdot \lambda(AGB) \geq (1 - \varepsilon)^2,$$

which means that (λ^n) is tight. Consider the compact semigroup

$$\{\lambda^n : n \geq 1\}^-,$$

where the semigroup operation is convolution and the bar denotes weak closure. Such a semigroup always has an idempotent element, say, λ_0 , so that for some subsequence (n_k) of positive integers

$$\lambda^{n_k} \xrightarrow{w} \lambda_0 = \lambda_0 * \lambda_0. \quad (4.163)$$

It follows from (4.162) and (4.163) that

$$\beta * \lambda_0 = \lambda_0 * \beta = \beta. \quad (4.164)$$

Since from (4.150), we have

$$\lambda^{n_k} = \alpha \lambda * \mu * \lambda^{n_k-1} + (1 - \alpha) \mu * \lambda^{n_k},$$

it is clear that

$$\lambda_0 \geq (1 - \alpha) \mu * \lambda_0, \quad (4.165)$$

and similarly

$$\lambda_0 \geq \alpha \lambda_0 * \mu. \quad (4.166)$$

The assertions (4.165) and (4.166) imply

$$S_\mu S_{\lambda_0} \cup S_{\lambda_0} S_\mu \subset S_{\lambda_0}$$

and therefore S_{λ_0} is an ideal of S , and consequently

$$S_\beta \subset S_{\lambda_0}.$$

Since S_{λ_0} is completely simple, it follows that $S_\beta = S_{\lambda_0}$. From (4.164), for $B \in S$ and $x \in S_\beta$, 1652
1653

$$\begin{aligned}\beta(Bx^{-1}) &= \int \lambda_0(Bx^{-1}y^{-1})\beta(dy) \\ &= \int \lambda_0(Bx^{-1})\beta(dy) = \lambda_0(Bx^{-1})\end{aligned}$$

Since $\lambda_0 = \lambda_0 * \lambda_0$,

1654

$$\begin{aligned}\lambda_0(B) &= \int \lambda_0(Bx^{-1})\lambda_0(dx) \\ &= \int \beta(Bx^{-1})\lambda_0(dx) \\ &= \beta * \lambda_0(B) = \beta(B).\end{aligned}$$

Thus, $\lambda_0 = \beta$. Since $S_\lambda = S_\beta = S_{\lambda_0}$, it follows immediately that the sequence λ^n converges weakly to $\lambda_0 = \beta$. 1655
□ 1656

Let us now state Theorem 4.20, the first part of which is contained in Theorem 4.18, and present the proof of its converse part. 1657
1658

Theorem 4.20. Let $(X_i)_{i=1}^\infty$ be a sequence of $d \times d$ i.i.d. random nonnegative matrices with distribution μ . Let $(Y_i)_{i=1}^\infty$ be an i.i.d. sequence, independent of the X_i s such that 1659
1660
1661

$$P(Y_i = 1) = \alpha, \quad P(Y_i = 0) = 1 - \alpha, \quad 0 \leq \alpha \leq 1.$$

Define the mixed random walk (S_n) by

1662

$$\begin{aligned}S_{n+1} &= S_n X_{n+1}, \text{ if } Y_{n+1} = 1; \\ &= X_{n+1} S_n, \text{ if } Y_{n+1} = 0.\end{aligned}$$

Let S be the smallest closed multiplicative semigroup generated by S_μ . Suppose μ satisfies the following conditions: 1663
1664

- (i) (μ^n) is tight; 1665
- (ii) $\mu(J) = 0$ and $S \cap \mathcal{J} \neq \emptyset$, where J and $\mathcal{J}$ are as in Theorem 4.18. 1666

Then the mixed random walk has a unique invariant probability measure λ ; and the support of λ is the smallest ideal of S with $\lambda(\mathcal{J}) = 1$, when S does not contain the zero matrix, and $\lambda = \delta_0$, when S contains the zero matrix. 1667
1668
1669

Conversely under (ii), the existence of an invariant probability measure (such that it has no mass on the zero matrix) for the mixed random walk (S_n) implies the tightness of the sequence (μ^n) . 1670
1671
1672

4.4 Invariant Measures for Random Walks in Nonnegative Matrices: Laws of Large Numbers 343

The proof of the converse part of Theorem 4.20, not given earlier, follows. 1673

If λ is an invariant probability measure for (S_n) , then as we have seen in (4.153), 1674
that for $n \geq 1$, λ must satisfy: 1675

$$\lambda = \sum_{k=0}^n \binom{n}{k} (1-\alpha)^{n-k} \alpha^k \mu^{n-k} * \lambda * \mu^k. \quad (4.167)$$

Notice that if x and z are strictly positive matrices and y is a nonzero nonnegative 1676
matrix, then the product xyz is a strictly positive matrix and as such, 1677

$$\begin{aligned} & \mu^{n-k} * \lambda * \mu^k(\mathcal{J}) \\ & \geq \mu^{n-k}(\mathcal{J}) \lambda(S - \{0\}) \mu^k(\mathcal{J}). \end{aligned} \quad (4.168)$$

It follows from (ii) in the theorem that $\lim_{n \rightarrow \infty} \mu^n(\mathcal{J}) = 1$, (see (4.55)). It follows 1678
from (4.156), (4.167), and (4.168) that $\lambda(\mathcal{J}) = 1$. This observation is crucial in 1679
establishing the tightness of (μ^n) as we will observe. 1680

Let $\varepsilon > 0$. Let A be a compact set such that $A \subset \mathcal{J}$ and $\lambda(A) > 1 - \varepsilon$. Let $\delta > 0$ 1681
such that 1682

$$\delta < \min \{x_{ii} : 1 \leq i \leq d \text{ and } x \in A\}.$$

Let $s \in A$ and $x \in S$ and $y \in S$. Then if $1 \leq i \leq d$ and $1 \leq \ell \leq d$, 1683

$$xsy \in A \Rightarrow \sum_j x_{ij} s_{jj} y_{j\ell} \leq (xsy)_{i\ell} \leq M$$

for some $M > 0$. This means 1684

$$xsy \in A \text{ and } s \in A \Rightarrow \delta \sum_j x_{ij} y_{j\ell} = \delta (xy)_{i\ell} \leq M$$

so there is a compact set B such that 1685

$$I_A(xsy) I_A(s) \leq I_B(xy).$$

From (4.167), for $n > 1$ 1686

$$\begin{aligned} & 1 - \varepsilon < \lambda(A) \\ & = \sum_{k=0}^n \binom{n}{k} (1-\alpha)^{n-k} \alpha^k \iiint I_A(xsy) \mu^{n-k}(dx) \lambda(ds) \mu^k(dy) \\ & \leq \varepsilon + \sum_{k=0}^n \binom{n}{k} (1-\alpha)^{n-k} \alpha^k \iint I_B(xy) \mu^{n-k}(dx) \mu^k(dy) \\ & = \varepsilon + \mu^n(B). \end{aligned}$$

This establishes that (μ^n) is tight. Thus, we have completely proved Theorem 4.20. 1687
□ 1688

Now we consider mixed random walks taking values in a set of affine maps from $(R^+)^d$ into itself. Such affine maps can be identified with $(d + 1) \times (d + 1)$ nonnegative matrices of the form

$$\begin{pmatrix} x & y \\ 0 & 1 \end{pmatrix}, \quad (4.169)$$

where x is a $d \times d$ matrix, y is a $d \times 1$ vector, $0 \equiv (0, 0, \dots, 0)$ is a $1 \times d$ vector. For a mixed random walk on such matrices, (ii) with $S \cap \mathcal{J} \neq \emptyset$ in Theorem 4.20 becomes meaningless. However if we replace (ii) with the assumption that S contains a rank 1 matrix, then the conclusion of Theorem 4.20 holds. The reason is the special multiplication property for such matrices given by

$$\begin{pmatrix} 0 & z \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x & y \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & z \\ 0 & 1 \end{pmatrix} \quad (4.170)$$

so that a rank 1 matrix of the form in (4.169) multiplied by another matrix of the form in (4.169) (not necessarily rank 1) remains unchanged. Now we need to use Theorem 4.19. Note that in Theorem 4.19, $S_\lambda = S_\beta =$ the kernel of S , which, in this particular situation, is the set of all rank 1 matrices of S (see Proposition 1.9), and moreover $\beta = \beta * \lambda = \lambda * \beta$, which now implies that for any Borel set $B \subset S$,

$$\begin{aligned} \beta(B \cap S_\beta) &= \beta(B) = \int \beta(w^{-1}B) \lambda(dw) \\ &= \int_{S_\lambda} \beta(w^{-1}B) \lambda(dw) \\ &= \lambda(B \cap S_\lambda), \end{aligned}$$

since

$$w^{-1}B = \{u \in S \mid w \cdot u \in B\} = \{u \in S \mid w \in B\},$$

where w has rank 1 and consequently $w \cdot u$ equals w . Thus $\lambda = \beta$ and the invariant probability measure is unique.

When S contains no rank 1 matrix, then (ii) in Theorem 4.20 can be replaced by

(ii)' $\mu(J) = 0$ and $S \cap I \neq \emptyset$,

where J is as in Theorem 4.20 and I as in (4.25) to reach the same conclusion. The proof is similar (see (4.120), (4.121), and Theorem 4.10) and omitted.

Thus, we can state Theorem 4.21.

Theorem 4.21. Consider the mixed random walk (S_n) , defined as in Theorem 4.20, with the corresponding i.i.d. sequence (X_i) taking values in $(d + 1) \times (d + 1)$ nonnegative matrices in the form of (4.169). Let μ , S , and J be as in Theorem 4.20. Suppose (μ^n) is tight and one of the following assumptions hold:

- (i) S contains a matrix with rank 1;
- (ii) $\mu(J) = 0$ and $S \cap I \neq \emptyset$, where I is as in (4.25).

Then the mixed random walk has a unique invariant probability measure whose support is the smallest ideal of S . Conversely under (ii) existence of a non-trivial (that is, not a unit mass) invariant measure implies the tightness of the sequence (μ^n) .

Now we present a law of large numbers for the mixed random walk. The following theorem of Furstenberg and Kifer is useful in this context.

Theorem 4.22 (Furstenberg–Kifer). Let $\{Z_n : n = 0, 1, 2, \dots\}$ be a Markov process with values in a compact metric space M with transition probability μ_x such that

$$P\{Z_{n+1} \in A \mid Z_0, \dots, Z_n\} = \mu_{Z_n}(A),$$

where μ_x is a continuous map from M into $P(M)$ and $P(M)$ is the compact metric space of probability measures on the Borel sets of M with respect to weak convergence. Then for any continuous real function f on M , the following results hold:

$$(i) \lim_{n \rightarrow \infty} \sup \frac{1}{n} \sum_{k=0}^n f(Z_k) \sup \left\{ \int f d\nu : \nu \in P(M) \text{ and for } A \subset M, \right. \\ \left. \nu(A) = \int \mu_x(A) \nu(dx) \right\};$$

(ii) If the set inside the parenthesis in (i) is a singleton and its value is $\bar{f}$, then

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^n f(Z_k) = \bar{f} \text{ almost surely.}$$

Proof. The proof is given in several steps. *Step 1.* Consider the Markov operator P defined on $C(M)$, the real continuous functions on M with the usual sup norm, by

$$Pf(x) = \int f(y) \mu_x(dy). \quad (4.171)$$

Then P maps $C(M)$ into $C(M)$. Let $f = g - Pg$ with $g \in C(M)$. Then we claim

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^n f(Z_k) = 0 \text{ almost surely.} \quad (4.172)$$

To prove (4.172), consider the sequence (U_n) given by

$$U_{n+1} = \sum_{k=0}^n \frac{1}{(k+1)} [Pg(Z_k) - g(Z_{k+1})].$$

Then the following assertions hold:

1735

(i)

$$\begin{aligned} & E(U_{n+1} \mid Z_0, Z_1, \dots, Z_n) \\ &= U_n + \frac{1}{n+1} [Pg(Z_n) - E\{g(Z_{n+1}) \mid Z_0, \dots, Z_n\}] = U_n, \end{aligned}$$

since $Pg(Z_n) = E\{g(Z_{n+1}) \mid Z_0, \dots, Z_n\}$;

1736

(ii) For $m < n$,

1737

$$\begin{aligned} & E\{[Pg(Z_m) - g(Z_{m+1})][Pg(Z_n) - g(Z_{n+1})]\} \\ &= E\{E[(Pg(Z_m) - g(Z_{m+1}))(Pg(Z_n) - g(Z_{n+1})) \mid Z_m, Z_{m+1}, Z_n]\} \\ &= 0. \end{aligned}$$

It follows that

1738

$$E(U_{n+1}^2) \leq 4\|g\|^2 \cdot \sum_{k=1}^{\infty} \frac{1}{k^2} < \infty.$$

By the Martingale convergence theorem (see [38]), the sequence (U_n) converges almost surely. By Kronecker's lemma (see [38, page 123]),

1739

1740

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^n [Pg(Z_k) - g(Z_k)] = 0$$

almost surely. This establishes (4.172).

1741

Step 2. Let $f \in C(M)$, $f \geq 0$. Then given $\varepsilon > 0$, we can write

1742

$$f = Pg - g + h, \quad (4.173)$$

where g and h are in $C(M)$ and

1743

$$\|h\| \leq \sup \left\{ \int f d\nu : \nu \in P(M) \text{ and } P * \nu = \nu \right\} + \varepsilon,$$

where

1744

$$P * \nu(A) = \int \mu_x(A) \nu(dx).$$

To prove (4.173), let $V \subset C(M)$ be the subspace given by

1745

$$V = \{Pg - g : g \in C(M)\}$$

and let $\text{dist}(f, V) = \delta$. If $\delta = 0$, then (4.173) is obvious. Let $\delta > 0$. Then by the Hahn–Banach theorem (of functional analysis), there is a bounded linear functional x^* on $C(M)$ such that

1746

1747

1748

$$\|x^*\| = 1, \quad x^*(f) = \delta \text{ and } x^*(h) = 0$$

for $h \in V$. By the Riesz-representation theorem, there is a bounded signed measure β such that

$$x^*(h) = \int h d\beta, \quad h \in C(M).$$

Since $x^*(h) = x^*(Ph)$ for $h \in C(M)$,

$$\int h d\beta = \int Ph d\beta = \int h d(P * \beta),$$

where

$$P * \beta(A) = \int \mu_x(A) \beta(dx).$$

Decomposing β into its positive and negative parts, we have

$$\beta = \beta^+ - \beta^-, \quad P * \beta = P * \beta^+ - P * \beta^-.$$

Notice that if A is a positive set so that $\beta(A) = \beta^+(A)$ and

$$\begin{aligned} P * \beta^+(A) &= \int \mu_x(A) \beta^+(dx) \\ &\geq \int \mu_x(A) \beta^+(dx) - \int \mu_x(A) \beta^-(dx) \\ &= P * \beta(A) = \beta(A) = \beta^+(A). \end{aligned}$$

This means $P * \beta^+ \geq \beta^+$. Similarly, $P * \beta^- \geq \beta^-$. It follows that $P * \beta^+ = \beta^+$ and $P * \beta^- = \beta^-$. Since $\int f d\beta = \delta$, $f \geq 0$, and $\|\beta\| = 1$, it is clear that $\|\beta^+\| > 0$. Let

$$\lambda = \frac{\beta^+}{\|\beta^+\|}.$$

Then,

$$P * \lambda = \lambda \text{ and } \int f d\lambda \geq \int f d\beta = \delta.$$

Assertion (4.173) now follows immediately.

Step 3. We observe that (i) in Theorem 4.22 follows immediately by proving (i) first for $C + f$, where $C = \|f\|$, so that $C + f \geq 0$. By considering (i) first for f then for $-f$, the Assertion (ii) also follows. $\square$

Note that the Furstenberg–Kifer theorem exploited the compactness of the state space M . However, even in the absence of compactness, we can use the preceding proof to obtain a law of large numbers for the mixed random walk with some modifications.

In what follows, we consider the mixed random walk of Theorem 4.20 (with the assumption that the zero matrix is not in S and under (i) and (ii) of this theorem) or the walk of Theorem 4.21 (under condition (ii) given in this theorem). Note that we still assume (μ^n) is tight and define P as in (4.148) and T as in (4.141). It fol-

lows that the unique invariant probability measure λ (for P) satisfies the following properties:

- (i) Under conditions of Theorem 4.20, $\lambda(\mathcal{J}) = 1$; furthermore, every element in S_λ is an idempotent, since S_λ is a completely simple subsemigroup, where its group factor (the set $eS_\lambda e$, $e = e^2 \in S_\lambda$) is a singleton (see the proof of Theorem 4.9).
- (ii) Under conditions of Theorem 4.21,

$$\lambda\{s \in S : \text{the last column of } s \text{ is strictly positive}\} = 1$$

if S_λ consists of all rank 1 matrices in S . If there is no matrix in S with rank 1, then S consists of only matrices of the form

$$\begin{pmatrix} B & 0 \\ 0 & 1 \end{pmatrix}, \quad B \neq 0, \quad B \text{ is } d \times d, \quad (4.174)$$

(as shown in (4.121)), so that this case is the same as in Theorem 4.20 and as such, the results in (i) are valid in this case, identifying all $(d+1) \times (d+1)$ matrices of the form given in (4.174) with simply the matrices B there.

Note that when (i) holds, then given any compact subset $K \subset S_\lambda$ such that $\lambda(K) > 0$, there is a function h in $C_0(S)$, the space of real continuous functions on S vanishing at infinity, such that $0 \leq h \leq 1$ and $h = 1$ on $K \cdot K$ and since $K \subset K \cdot K$ (see Property (i) above),

$$\int h d\lambda > 0; \quad (4.175)$$

also for any $x \in K$, since $K \cdot x \subset K \cdot K$,

$$\int h(yx) \lambda(dy) > 0. \quad (4.176)$$

When (ii) holds and S_λ consists of only rank 1 matrices (under conditions of Theorem 4.21), then

$$x, y \in S_\lambda \Rightarrow y \cdot x = y.$$

This means that given any compact subset $K \subset S_\lambda$ with $\lambda(K) > 0$, we can again find $h \in C_0(S)$ satisfying (4.175) and (4.176).

Let us now observe that the operator T , defined as in (4.141), is a bounded linear operator from $C_0(S)$ into $C_0(S)$. The reason follows.

Given $\varepsilon > 0$, let A be a compact subset of S such that $A \subset J^c$ and $\mu(A) > 1 - \varepsilon$. If $f \in C_0(S)$, B is compact and

$$|f(s)| < \varepsilon \text{ for } s \notin B,$$

then for $y \in A$,

$$|f(xy)| < \varepsilon$$

if $x \notin BA^{-1}$, and

$$|f(xy)| < \varepsilon$$

if $x \notin A^{-1}B$. Since A does not contain a matrix which has a zero row or a zero column and since S contains matrices with only nonnegative entries, the sets BA^{-1} and $A^{-1}B$ are both compact. Then it follows easily that

$$T : C_0(S) \rightarrow C_0(S).$$

Now following the proof of Theorem 4.22, we observe that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^n h(M_k) = 0 \text{ almost surely,} \quad (4.177)$$

where $h = Tf - f$, $f \in C_0(S)$, and (M_k) is the mixed random walk.

Let $f \in C_0(S)$ and $\int f d\lambda = 0$, where λ is the *unique* invariant probability measure for (M_n) . Again considering the subspace V as in the proof of Theorem 4.22, we see that either the measure β^+ or β^- there must be the zero measure (because of the uniqueness of the invariant measure for (M_n)) and consequently the normalized measure $\beta/\beta(S)$ obtained there must be either λ or $-\lambda$. Since $\int f d\lambda = 0$, it follows that $f \in \text{closure}(V)$, and consequently from (4.177),

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^n f(M_k) = 0 \text{ almost surely.} \quad (4.178)$$

Now let $f \in C_0(S)$ and $\int f d\lambda > 0$. Again using the same argument as before and noting that λ is the unique invariant probability measure for (M_n) , it follows that either the measure β^+ or β^- (obtained in the proof of Theorem 4.22) again must be the zero measure. Since $\int f d\lambda > 0$, β^- must be the zero measure, $\beta = \beta^+$ and $\lambda = \beta/\beta(S)$. Since the functional x^* there has norm 1 and the measure β is induced by x^* , $\beta(S) = 1$ and consequently $\lambda = \beta$, so we have almost surely

$$\lim_{n \rightarrow \infty} \sup \frac{1}{n} \sum_{k=0}^n f(M_k) \leq \int f d\lambda. \quad (4.179)$$

Note that

$$\begin{aligned} & E \left(\lim_{n \rightarrow \infty} \sup \frac{1}{n} \sum_{k=0}^n f(M_k) \right) \\ & \geq \lim_{n \rightarrow \infty} \sup E \left(\frac{1}{n} \sum_{k=0}^n f(M_k) \right) \\ & = \lim_{n \rightarrow \infty} \sup \frac{1}{n} \sum_{k=0}^n \iint f(y) P^k(x, dy) \mu(dx). \end{aligned}$$

Under the conditions we have assumed, it follows from Theorems 4.9 and 4.16 that μ^n converges to a probability measure β_0 weakly. Following the proof of Theorem 4.18 (which works here even though S is not assumed to be compact, since we can restrict y and z in (4.142) with no loss of generality to a compact set using the tightness of (μ^n)), it follows that for each x in S and f in $C_0(S)$, as in the proof of Theorem 4.18,

$$\lim_{n \rightarrow \infty} T^n f(x) = \bar{T} f(x);$$

that is,

$$\lim_{n \rightarrow \infty} \int f(y) P^n(x, dy) = \int f(y) \beta_0 * \delta_x * \beta_0(dy). \quad (4.180)$$

This means

$$\begin{aligned} & \lim_{n \rightarrow \infty} \iint f(y) P^n(x, dy) \mu(dx) \\ &= \iint f(y) \beta_0 * \delta_x * \beta_0(dy) \mu(dx) \\ &= \iiint f(yxz) \beta_0(dy) \mu(dx) \beta_0(dz) \\ &= \int f(y) \beta_0 * \mu * \beta_0(dy) = \int f(y) \beta_0(dy), \end{aligned}$$

since $\beta_0 * \mu * \beta_0 = \beta_0 * \beta_0 = \beta_0$. It follows that

$$E \left(\lim_{n \rightarrow \infty} \sup \frac{1}{n} \sum_{k=0}^n f(M_k) \right) \geq \int f d\beta_0. \quad (4.181)$$

Since $\beta_0 = \mu * \beta_0 = \beta_0 * \mu$, it follows from Proposition 2.5 that for each $x \in S$, $\beta_0 * \delta_x * \beta_0 = \beta_0$. Since

$$\begin{aligned} & \int P(x, A) \beta_0(dx) \\ &= \alpha \beta_0 * \mu(A) + (1 - \alpha) \mu * \beta_0(A) \\ &= \alpha \beta_0(A) + (1 - \alpha) \beta_0(A) = \beta_0(A), \end{aligned}$$

it follows that $\lambda = \beta_0$, where the invariant measure is unique.

Now we observe that the inequality in (4.179) must actually be an equality almost surely. The reason is: If there is strict inequality in (4.179) on a set of positive probability, then

$$\int f d\lambda > E \left(\lim_{n \rightarrow \infty} \sup \frac{1}{n} \sum_{k=0}^n f(M_k) \right),$$

which contradicts (4.181), since $\lambda = \beta$. Thus, for $f \in C_0(S)$ and $\int f d\lambda > 0$, 1831

$$\lim_{n \rightarrow \infty} \sup \frac{1}{n} \sum_{k=0}^n f(M_k) = \int f d\lambda, \text{ almost surely.} \quad (4.182)$$

Now let X be a nonnegative random matrix, independent of the X_i s and the Y_i s 1832
such that the distribution of X is λ . Then the process $f(M_k X)$, $1 \leq k < \infty$ 1833
and $f \in C_0(S)$, is a stationary process, so by the classical law of large numbers, 1834
 $P(B) = 1$, where 1835

$$B = \left\{ \omega : \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^n f(M_k X) \text{ exists} \right\}.$$

Then

$$1 = \int P(B \mid X = x) \lambda(dx). \quad 1836$$

This means given any subset $A \subset S_\lambda$ with $\lambda(A) > 0$ and given any $f \in C_0(S)$, 1837
there exists $x \in A$ such that there exists almost surely 1838

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^n f(M_k x). \quad (4.183)$$

Let K be a compact subset of S_λ such that $\lambda(K) > 0$ and $K \subset \{s \in S : \text{the last} \quad 1839$
column of s is strictly positive\}. Note that this is possible because of Observations 1840
(i) and (ii) just preceding (4.174). Choose h in $C_0(S)$ for this K as in (4.175). By 1841
(4.183), there exists $x \in K$ such that there exists almost surely 1842

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^n h(M_k x). \quad (4.184)$$

Define $h_1(y) = h(yx)$ for $y \in S$. Since the last column of x is strictly positive, 1843
 $h_1 \in C_0(S)$ and $\int h_1 d\lambda > 0$ by (4.176). Now let $g \in C_0(S)$ such that $\int g d\lambda \neq 0$. 1844
Choose a real number p such that 1845

$$\int (p \cdot h_1 - g) d\lambda = 0.$$

Using results from (4.178) and (4.184), it follows that there exists almost surely 1846

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^n g(M_k).$$

It follows from (4.178) and (4.182) that if $g \in C_0(S)$ and $\int g d\lambda > 0$, then 1847

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^n g(M_k) = \int g d\lambda \text{ almost surely} \quad (4.185)$$

Replacing g by $-g$ when $\int g d\lambda < 0$, it follows that (4.185) holds for any $g \in C_0(S)$. 1848
1849

Finally, let g be a bounded continuous nonnegative function on S . Then given $\varepsilon > 0$, there is some g_0 in $C_0(S)$ such that 1850
1851

$$0 \leq g_0 \leq g \text{ and } \int (g - g_0) d\lambda < \varepsilon.$$

Thus, 1852

$$\begin{aligned} & \lim_{n \rightarrow \infty} \inf \frac{1}{n} \sum_{k=0}^n g(M_k) \\ & \geq \lim_{n \rightarrow \infty} \inf \frac{1}{n} \sum_{k=0}^n g_0(M_k) = \int g_0 d\lambda \\ & > \int g d\lambda - \varepsilon, \end{aligned}$$

almost surely. When g is an arbitrary bounded continuous function (such that $|g| \leq M$, a constant), then applying the preceding result on the functions $M - g$ and $M + g$, we have 1853
1854
1855

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^n g(M_k) = \int g d\lambda, \text{ almost surely.} \quad (4.186)$$

Thus, we have proven Theorem 4.23. 1856

Theorem 4.23. Consider the mixed random walk (M_n) in Theorem 4.20 (under the assumptions there) or the mixed random walk in Theorem 4.21 (under the assumption (ii) there). Assume that S does not contain the zero matrix. Then for any bounded continuous function f on S , 1857
1858
1859
1860

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^n f(M_k) = \int f d\lambda,$$

almost surely, where λ is the unique invariant probability for (M_k) . 1861

4.4 Invariant Measures for Random Walks in Nonnegative Matrices: Laws of Large Numbers 353

We conclude Sect. 4.4 with a number of examples illustrating conditions used in Theorem 4.23 in the context of affine maps from R^d into R^d (or equivalently, $(d + 1) \times (d + 1)$ matrices of the form (4.169)).

Example 4.5. Consider maps f , g , and h given at the end of Sect. 4.2. Any probability measure μ such that $S_\mu = \{f, g, h\}$ meets all the conditions of Theorem 4.23.

Example 4.6. Consider the family $\mathcal{F}$ of affine maps from R^d into R^d such that each $f \in \mathcal{F}$ has the form

$$f(x) = A(f)x + B(f), \quad x \in R^d,$$

where f can be identified with the matrix

$$\begin{pmatrix} A(f) & B(f) \\ 0 & 1 \end{pmatrix}$$

(which has the form (4.169)) and there are positive numbers a with $0 < a < 1$ and b such that for each $f \in \mathcal{F}$,

$$\|A(f)\| < a \text{ and } \|B(f)\| < b.$$

If x is a d -column vector, then

$$\begin{pmatrix} f(x) \\ 1 \end{pmatrix} = \begin{pmatrix} A(f) & B(f) \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ 1 \end{pmatrix}$$

and

$$\begin{pmatrix} f \circ g(x) \\ 1 \end{pmatrix} = \begin{pmatrix} A(f) & B(f) \\ 0 & 1 \end{pmatrix} \begin{pmatrix} A(g) & B(g) \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ 1 \end{pmatrix}.$$

It is easily verified that if $S_\mu \subset \mathcal{F}$, then

$$S \left(\equiv \text{cl} \left(\bigcup_{n=1}^{\infty} S_\mu^n \right) \right)$$

is compact.

Example 4.7. Consider an i.i.d. sequence of random $(d + 1) \times (d + 1)$ real matrices (X_n) such that

$$X_n = \begin{pmatrix} A_n & B_n \\ 0 & 1 \end{pmatrix}$$

is of the type described in (4.169). Suppose that almost surely, each $\|B_n\| < M < \infty$, and there are numbers $\alpha_0 < \alpha_1 < \cdots < \alpha_N$ such that

$$P(\alpha_{j-1} < \|A_n\| \leq \alpha_j) = \beta_j, \quad 1 \leq j \leq N,$$

where the α_i s and β_i s satisfy conditions

1880

$$\sum_{j=1}^N \beta_j = 1 \text{ and } \sum_{j=1}^N \beta_j \log \alpha_j (\equiv \log r, \text{ say}) < 0. \quad (4.187)$$

Note that unlike Examples 4.5 and 4.6, the X_n s here do not need to be almost surely contraction maps, since some of the α_i s can be much larger than 1 as long as (4.187) holds. 1881 1882 1883

Let μ be the distribution of each X_n . We show that the convolution sequence (μ^n) is tight, even though S_μ is not necessarily contained in a compact matrix semigroup. Let us write 1884 1885 1886

$$B(j, n) = \{\omega : \alpha_{j-1} < \|A_n(\omega)\| \leq \alpha_j\}. \quad (4.188)$$

By the strong law of large numbers, then we have almost surely,

1887

$$y_{nj} \equiv \frac{1}{n} \sum_{m=1}^n I_{B(j,m)}$$

converges to β_j . Choose $s > 0$ such that

1888

$$\log r_1 \equiv \log r + s \sum_{j=1}^N |\log \alpha_j| < 0. \quad (4.189)$$

Let $0 < t < s$. There exists a positive integer-valued random variable $n_0(\omega)$ such that $P(n_0 < \infty) = 1$ and $n > n_0(\omega)$ implies 1889 1890

$$\left| \frac{n(j, \omega)}{n} - \beta_j \right| < t, \quad 1 \leq j \leq N, \quad (4.190)$$

where $n(j, \omega)$ is the cardinality of the set

1891

$$\{m \leq n : \alpha_{j-1} < \|A_m(\omega)\| \leq \alpha_j\}.$$

Here,

1892

$$n_0(\omega) = \inf \{m : |y_{mj}(\omega) - \beta_j| < s \ \forall n \geq m\}.$$

Let L be a positive integer such that $P(n_0 < L) > 1-t$. Let $n > L$ and $n_0(\omega) < L$. Then 1893 1894

$$\begin{aligned} \|A_1(\omega)A_2(\omega) \dots A_n(\omega)\| &\leq \prod_{j=1}^N (\alpha_j)^{n(j,\omega)} \\ &\leq \prod_{j=1}^N (\alpha_j)^{n(\beta_j+s)} = r_1^n. \end{aligned}$$

Since the first row of $X_1 X_2 \dots X_n$ is

1895

$$(A_1 A_2 \dots A_n, B_1 + A_1 B_2 + A_1 A_2 \dots A_{n-1} B_n),$$

it is clear there exists $C > 0$ such that

1896

$$\{n_0(\omega) < L\} \subset \{\|X_1(\omega) \dots X_n(\omega)\| \leq C \ \forall n > L\}.$$

Section 4.4. Exercises

1897

Exercise 4.27. Establish (4.140) and (4.142).

1898

Exercise 4.28. Prove the inequality just preceding (4.142).

1899

Exercise 4.29. Establish (4.142).

1900

Exercise 4.30. Discuss the difficulties in the development of the results without the assumption $S_{\mu_1} = S_{\mu_2}$ at the beginning of this section.

1901
1902

Exercise 4.31. Give versions of Theorem 4.18 in full details for the left random walk as well as for the right random walk.

1903
1904

Exercise 4.32. Give the details of the proof of Theorem 4.21.

1905

Exercise 4.33. Explain the difficulties in the proof of Theorem 4.23 if S is allowed to contain the zero matrix.

1906
1907

4.5 Notes and Comments

1908

Section 4.2

1909

Theorems 4.1 and 4.2 and Corollary 4.1 were originally results of an effort to understand how much of the classical theory due to Chung and Fuchs (see [19]) remains valid for random matrices. These results are taken from [160]. The reader may note that nonnegativity of the entries in the matrices has been crucial in proving these results. The recurrent sets $R(Z)$ and $R(W)$ are of course the same in R^d ; but the picture is less clear for products of random matrices where we do not have commutativity of multiplication, and as such, natural questions arise regarding the structures of the sets $R(Z)$ and $R(W)$ and how they are related. Theorems 4.3–4.6 answer some of these questions, and these are taken from [161]. The result that $R(Z)$ is nonempty in (4.46) is from [79].

1910
1911
1912
1913
1914
1915
1916
1917
1918
1919

The concept of recurrence for matrices is used in the last part of this section to define attractors in a manner consistent with works of M. Barnsley and colleagues in [6]. Attractors are also introduced by various authors as supports of certain invariant probability measures. Theorem 4.7, taken from [161], shows how the two definitions are connected. Finally in this section, we show how the structure of the recurrent set $R(W)$ can be exploited to give a complete description of the well-known attractor, the Sierpinski Gasket, and one of its unbounded versions. This was also given in [157].

Quite a few results in this section are less clear for matrices which are not necessarily nonnegative, even though many of the results in Sect. 4.3 are given for the general case of real matrices. A number of related results for real matrices can be found in [111]. Related results also appear in [104, 109, 110]. It is worthwhile to obtain analogs of all results in this section for random real matrices.

Section 4.3

Many of the results in this section are obtained in the presence of one or both of the following conditions: (1) the set of matrices with at least one zero row or one zero column has zero measure, and (2) the set of strictly positive matrices in S is nonempty. As far as we know, these conditions were first explored by Kesten and Spitzer in [128] to avoid some pathologies in their study of convergence in distribution of products of i.i.d. nonnegative matrices. While some results in this section are for nonnegative matrices, a number of results such as Theorems 4.11–4.17 and Proposition 4.2 are given for real matrices. It will be of course interesting to find whenever possible what the results are in the general situation. It is relevant to point out that papers by Bougerol [16] and Mukherjea [163] give results on tightness for the general case of real matrices. Theorems 4.11 and 4.12 are contained in [163] and Theorem 4.13 in [153]. A short readable account of some of the problems and results in this context is given in [107]. Theorem 4.17 follows from the work in [79], Theorem 4.8 appeared in [159], and Proposition 4.1 and Theorem 4.9 in [153]. Both Theorems 4.10 and 4.14 are due to Kesten and Spitzer [128]. Theorem 4.14 was only stated there. Proofs of Theorems 4.10 and 4.14 are taken from [153]. Analogs of these two results for affine maps from $(R^+)^d$ into $(R^+)^d$ are in Theorems 4.15 and 4.16, which appeared in [161]. Propositions 4.2–4.4 appeared in [16], [160], and [159], respectively.

It is worthwhile pointing out the following interesting result (contained in Lemma 2 of [128]).

Let (X_i) be i.i.d. $d \times d$ nonnegative matrices with distribution μ , $Y_n = X_n X_{n-1} \dots X_1$, $\mu(J) = 0$ and $S \cap \mathcal{J} \neq \emptyset$, where J , S and $\mathcal{J}$ are as in Theorem 4.10. Then the sequence $Q_n = \frac{Y_n}{\|Y_n\|}$ has a limit distribution which is concentrated on the strictly positive matrices of rank 1.

A “semigroup” proof of this result can be given along the following lines.

Note that if we define for matrices A and B , $A \cdot B = \frac{AB}{\|AB\|}$, then $(J^c, \cdot)$ is a locally compact Hausdorff topological semigroup and μ^n , the n th convolution power of μ with respect to this new multiplication, is the distribution of Q_n . Let $T = \{x \mid \|x\| = 1\}$. Then $\{\mu^n, n \geq 2\} \subset P(T)$ and it is easy to see that the averages $\frac{1}{n} \sum_{k=1}^n \mu^k$ converges weakly to ν , $\mu^k * \nu * \mu^k = \nu$ for $k \geq 1$, $\mathcal{J} \cdot S_\nu \cdot \mathcal{J} \subset \mathcal{J}$ and $\nu(\mathcal{J}) = 1$ (since (4.55) holds). Also, $\nu * \nu = \nu$. Restricting μ to J^c , it follows that the support of the ν_0 , the restriction of ν to J^c , is a completely simple semigroup $X \times G \times Y$ by Theorem 2.8, where G is a compact group (with respect to $\cdot$) of strictly positive matrices. (Note that for any $x \in S_{\nu_0}$, the set $x \cdot S_1 \cdot x$, where S_1 is the closed semigroup generated by S_{ν_0} , can be taken as G .) Now we claim that G is a singleton set so that S_{ν_0} consists of idempotent (with respect to $\cdot$) matrices. To see this, we use the Frobenius theory. By Theorem 2.3 on page 546 of Karlin and Taylor (A First Course in Stochastic Processes, 2nd edition), if e is the identity of G , then for some constant c , $\frac{e^n}{c^n}$ converge pointwise to a rank 1 idempotent strictly positive matrix f . This means that since $\|e^n\| = 1$, $e = \frac{e^n}{c^n} \frac{c^n}{\|e^n\|} \rightarrow \frac{f}{\|f\|}$, so that e has rank 1. Since for $x \in G$, $x = x \cdot e$, x has rank 1. Thus, for $x \in G$, $x^2 = rx$ for some constant r so that $x \cdot x = \frac{x^2}{\|x^2\|} = x$, since $\|x\| = 1$. Hence, G is a singleton, since every element G is idempotent.

Next, considering $\{\mu^n, n \geq 2\} \subset P(T)$, we claim that for any weak limit point λ of (μ^n) , $\lambda(J^c) = 1$. The reason is the following: $J = J_r \cup J_c$ (see the definition on page 287). If $\lambda(J_r) > 0$, then since $\lambda * \nu = \nu = \nu * \lambda$ with

$$\nu = (w) \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n \mu^k,$$

it follows that $\nu(J_r) \geq \lambda(J_r) > 0$ which contradicts that $\nu(\mathcal{J}) = 1$. Thus, $\lambda(J_r) = 0$. Similarly, $\lambda(J_c) = 0$ and therefore $\lambda(J^c) = 1$. This means that the weak limit points of μ^n in $P(T)$ can be regarded as weak limit points of μ^n in $P(J^c)$. It follows by Theorem 2.7 that μ^n converges weakly to ν_0 , $\nu_0(\mathcal{J}) = 1$, and S_{ν_0} is a completely simple semigroup of rank 1 matrices with norm 1 and such matrices are easily seen to be of the form $r\ell^T$, where r is a strictly positive column vector and ℓ^T is a strictly positive probability row vector. Note that since G is a singleton, we can regard $S_{\nu_0} = XY$ and $YX = e = e \cdot e$.

If (X_i) is a stationary, ergodic sequence of $d \times d$ matrices and $Y_n = X_n X_{n-1} \dots X_1$, then it follows from Kingman's subadditive ergodic theorem (see [131]) that when $\mathbf{E}[\log^+ \|X_1\|] < \infty$, $\lim_{n \rightarrow \infty} \frac{1}{n} \log \|Y_n\| = E \in [-\infty, \infty)$, almost surely. The limit E is usually called the maximal Lyapunov exponent. Eric Key (see [130]) has used the Kesten and Spitzer result proved above to show that when (X_i) are i.i.d. with distribution μ , $\mu(J) = 0$, $S \cap \mathcal{J} \neq \emptyset$ and $\mathbf{E}[\log^+ \|X_1\|] < \infty$, then $\lim_{n \rightarrow \infty} \frac{1}{n} \log \rho(Y_n) = E$ almost surely, where $\rho(Y_n)$ is the spectral radius of Y_n . This result was earlier conjectured by Joel Cohen in [44]. In the context of computing the maximal Lyapunov exponent, let us mention [104, 129], and [189].

Finally, in the context of weak convergence in 2×2 stochastic matrices, it is relevant to point out the following result:

Let μ be a probability measure on 2×2 stochastic matrices. Let (a, b) denote the matrix whose first column elements are a and b . Then μ^n converges weakly if and only if μ is not the unit mass at $(0, 1)$. If $S(\mu)$, the support of μ , contains some (a, b) , where either $0 < a < 1$ or $0 < b < 1$, then the support of the weak limit of μ^n contains only matrices with identical rows.

For a proof of this, see [151].

Section 4.4

As far as we know, mixed walks in compact semigroups were first studied by Högnäs (see Chap. 3); left and right random walks are special cases of these walks. The emphasis in this section is results showing the existence of a unique invariant probability measure for such walks. Theorems 4.18 through 4.21 are results in this direction; they are given in [112]. These results and Theorem 4.22, which is due to Furstenberg and Kifer [71], are used to obtain Theorem 4.23, which appeared in [112]. Example 4.7 is taken from [157]. This example describes an average contractivity condition (see (4.187)) for affine maps. It is shown here that this condition is really the tightness condition in disguise at least in the i.i.d. situation. A very similar condition was first introduced by Elton (see [6, 67]).

AUTHOR QUERIES

- AQ1. Please provide appropriate caption for the Figs. 4.1 and 4.2.
AQ2. Please confirm step environment start from 6. Number is ok

Uncorrected Proof

Appendix A

Products of I.I.D. Random Stochastic Matrices: Their Skeletons and Convergence in Distribution

Let μ_1 and μ_2 be two probability measures on the Borel subsets of $d \times d$ stochastic matrices with the usual Euclidean topology. Let S_1 and S_2 be, respectively, the compact multiplicative semigroups of $d \times d$ stochastic matrices generated by the supports $S(\mu_1)$ and $S(\mu_2)$. For x in $S_1 \cup S_2$, we denote by $\bar{x}$ the idempotent (identity) element in the group of limit points of $\{x^n \mid n \geq 1\}$. Throughout our discussion here, we make the following “skeleton” assumption.

Assumption*. Let $D = \{1, 2, \dots, d\}$. For each $x \in S(\mu_1)$, there is some $y \in S(\mu_2)$ such that x and y have the same skeleton (that is, for all i, j in D , $x_{ij} > 0$ iff $y_{ij} > 0$). Similarly, for each y in $S(\mu_2)$, there is some x in $S(\mu_1)$ such that x and y have the same skeleton. $\square$

Our main purpose here is to show that under the above assumption, μ_1^n converges weakly as $n \rightarrow \infty$ iff μ_2^n does so. Note that the skeleton assumption (*) easily extends to $\bigcup_{m=1}^{\infty} S(\mu_1)^m$ and $\bigcup_{m=1}^{\infty} S(\mu_2)^m$, but may not extend to their closures, that is, to S_1 and S_2 . For $i = 1, 2$, let us define the sets K_i by

$$K_i = \{x \in S_i \mid x \text{ has the minimal rank among all matrices in } S_i\}.$$

Then it is known (see Chap. 1) that K_i is the kernel of S_i , $i = 1, 2$.

We claim that under assumption (*), the matrices in $K_1 \cup K_2$ have the same rank; that is, if $x \in K_1$ and $y \in K_2$, then $\text{rank } x = \text{rank } y$.

To prove this claim, let e be an idempotent element of K_1 with basis $\{T, C_1, C_2, \dots, C_p\}$ such that among all idempotent elements in K_1 , e has the minimum number of zero columns. See Theorem 1.11 for the definition of basis.

First, suppose that $e \in \bigcup_{m=1}^{\infty} S(\mu_1)^m$. Then, of course, there is some $z \in \bigcup_{m=1}^{\infty} S(\mu_2)^m$ such that e and z have the same skeleton. Thus, for $1 \leq i \leq p$, $1 \leq j \leq p$, each $C_i \times T$ block, each $C_i \times C_j$ ($i \neq j$) block, and the $T \times T$ block of z is an all zero block, and each $C_i \times C_i$ block of z is a strictly positive stochastic matrix. 28

AQ1 The same is true for all powers of z , and therefore, for the element $\bar{z}$, an idempotent element in S_2 . Note that since

$$\bar{z} |_{C_i \times C_i} = \lim [z |_{C_i \times C_i}]^n, \quad 1 \leq i \leq p,$$

$\bar{z} |_{C_i \times C_i}$ has rank one. Also, it is easily verified that if $z |_{T \times C_i}$ is strictly positive, for some i , $1 \leq i \leq p$, then the same is true for $\bar{z}$. Thus, e and $\bar{z}$ have the same basis, and also the same skeleton.

Next we assume that $e \in K_1 - \bigcup_{m=1}^{\infty} S(\mu_1)^m$, and e has basis $\{T, C_1, C_2, \dots, C_p\}$. Then, with no loss of generality, we can assume that there are elements e_n in $\bigcup_{m=1}^{\infty} S(\mu_1)^m$ such that $e_n \rightarrow e$ as $n \rightarrow \infty$, for each i , $1 \leq i \leq p$, $e_n |_{C_i \times C_i}$ is strictly positive for each n , for all j in $D = \{1, 2, \dots, d\}$ and $t \in T$, $(e_n)_{jt} < \frac{1}{2d^2}$, and $e_n |_{T \times C_k}$ is strictly positive whenever $e |_{T \times C_k}$ is so, for each n . Let G_n be the group of limit points of $\{e_n^m \mid m \geq 1\}$. Then $\bar{e}_n$ is the identity of G_n . Choose $y_n \in G_n$ such that

$$y_n (e_n \bar{e}_n) = (e_n \bar{e}_n) y_n = \bar{e}_n.$$

But since $y_n \bar{e}_n = y_n = \bar{e}_n y_n$ and $e_n \bar{e}_n = \bar{e}_n e_n$, we then have:

$$y_n e_n = e_n y_n = \bar{e}_n. \quad (\text{A.1})$$

Let the basis of $\bar{e}_n$ be $\{T', C'_1, C'_2, \dots, C'_q\}$. Note that $\bar{e}_n \in S_1$, but $e \in K_1$, and thus $q \geq p$. Now we claim that e and $\bar{e}_n$ have the same basis and the same skeleton.

Let $j \in C'_i$, $1 \leq i \leq q$, and $t \in T'$. Then we have

$$\begin{aligned} 0 &= (e_n \bar{e}_n)_{jt} = (\bar{e}_n e_n)_{jt} \\ &= \sum_{k \in C'_i} (\bar{e}_n)_{jk} (e_n)_{kt}; \end{aligned}$$

this implies that $e_n |_{C'_i \times T'}$ is an all zero block. Corresponding to y_n in (A.1) by Theorem 1.14, there is a permutation π of $\{1, 2, \dots, q\}$ such that the $C'_i \times C'_j$ of y_n is a block with all entries strictly positive or zero accordingly as $\pi(i) = j$ or $\pi(i) \neq j$. Notice also that $y_n e_n = \bar{e}_n$, and thus, for $u \in C'_k$, we have:

$$1 = \sum_{j \in C'_k} (\bar{e}_n)_{uj} = \sum_{\ell \in C'_{\pi(k)}} (y_n)_{u\ell} \sum_{j \in C'_k} (e_n)_{\ell j},$$

and this implies that $e_n |_{C'_{\pi(k)} \times C'_k}$ is a stochastic matrix, whenever $1 \leq k \leq q$. We will now show that the permutation π corresponding to y_n must be the identity permutation. If not, then there is some j , $1 \leq j \leq q$, such that $j \neq \pi(j)$, and consequently, $\pi(\pi(j)) \neq \pi(j)$. Write: $\pi(j) = i$ and $\pi(k) = j$. Then $i \neq j$, $j \neq k$. It follows that $e_n |_{C'_i \times C'_i}$ as well as $e_n |_{C'_j \times C'_j}$ is an all zero block. Thus,

$C'_i \cup C'_j \subset T$, because of the way e_n was chosen, and $C'_i \times C'_j \subset T \times T$. But this contradicts the fact that $e_n |_{C'_i \times C'_j}$ is a stochastic matrix, since for $t_1 \in T, t_2 \in T$, $(e_n)_{t_1 t_2} < \frac{1}{2d^2}$, because of the way e_n was chosen. Thus, π must be the identity permutation, and thus, y_n must be the element $\bar{e}_n$ (see Chap. 1), and this means that for $k \geq 1$,

$$\bar{e}_n e_n^k = e_n^k \bar{e}_n = \bar{e}_n.$$

It follows that for any y in G_n , $\bar{e}_n y = y \bar{e}_n = \bar{e}_n$ so that G_n is a singleton and $\lim_{k \rightarrow \infty} e_n^k = \bar{e}_n$. Also, since $e_n \bar{e}_n = \bar{e}_n e_n = \bar{e}_n$, e_n must be of the form

	T'	C'_1	C'_2
T'			
C'_1	0	Stochastic	0
C'_2	0	0	Stochastic

61

Recall that $e_n |_{T \times T}$ is strictly sub-stochastic and thus, no C'_i can be completely contained in T , and no C_i can intersect two different C'_j (because of the way e_n was chosen). Thus, each C'_j is completely contained in $C_i \cup T$, for some i ($i = i(j)$, i depending on j), and $i(j) \neq i(k)$ for $j \neq k$. It follows that $p = q$, $T' \subset T$, and each $C_i \subset$ some C'_j . Now e , by choice, has the minimum number of zero columns among all idempotent elements of K_1 ; this means that $T' = T$, and consequently, each C_i is some C'_j . In other words, e and $\bar{e}_n$ have the same basis; in fact, they must also have the same skeleton since $e_n \bar{e}_n = \bar{e}_n e_n = \bar{e}_n$.

Now since $e_n \in \bigcup_{m=1}^{\infty} S(\mu_1)^m$, by our skeleton assumption, there must exist $z_n \in \bigcup_{m=1}^{\infty} S(\mu_2)^m$ such that z_n and e_n have the same skeleton for each n . Thus, z_n , for each n , must be of the form

	T	C_1	C_2
T			
C_1	0	Stochastic (and all entries positive)	0
C_2	0	0	Stochastic (and all entries positive)

73

It is clear that each row sum in the $z_n |_{T \times T}$ block is less than 1 since the same is true for e_n . It follows that $\bar{z}_n$ must have the same basis $\{T, C_1, \dots, C_p\}$ and also the same skeleton as e . Thus, the minimal rank of the matrices in S_2 is at most p . Reversing the roles of S_1 and S_2 , we have proven the following theorem.

Theorem A.1. *Under Assumption (*), the minimal rank of the matrices in K_1 is the same as that of the matrices in K_2 . Furthermore, there are idempotent elements*

e in K_1 and f in K_2 such that e and f have the same basis and the same skeleton. 80
□ 81

The kernels K_1 and K_2 are completely simple subsemigroups of S_1 and S_2 , 82
respectively, and as such, they have their usual Rees–Suskevitch decompositions 83
of the form $X_1 \times G_1 \times Y_1$ and $X_2 \times G_2 \times Y_2$. Let us first describe how we choose 84
the sets X_1, X_2, G_1, G_2 , and Y_1, Y_2 . 85

Theorem A.1 gives us idempotent elements $e \in K_1$ and $f \in K_2$ such that e 86
and f have the same skeleton and the same basis. We take the set $E(K_1e)$ as X_1 , 87
 $E(K_2f)$ as X_2 , $E(eK_1)$ as Y_1 , $E(fK_2)$ as Y_2 , the group eK_1e as G_1 , and the 88
group fK_2f as G_2 . Here, $E(A)$ denotes the set of all idempotent elements in A . It 89
is wellknown that K_1 is isomorphic to $X_1 \times G_1 \times Y_1$ and K_2 to $X_2 \times G_2 \times Y_2$ (see 90
Chap. 1). 91

We claim that under Assumption (*), the group factors G_1 and G_2 are isomorphic. 92

To prove this claim, let us suppose that the basis of the idempotent elements e 93
and f is given by 94

$$\{T, C_1, C_2, \dots, C_p\}.$$

Let $x \in K_1$. Then $exe \in G_1$. There exists a sequence $x_n \in \bigcup_{m=1}^{\infty} S(\mu_1)^m$ such 95
that $x_n \rightarrow x$. Then $ex_n e \rightarrow exe$ as $n \rightarrow \infty$. We choose n sufficiently large so that 96
whenever $(exe)_{ij} > 0$, then $(ex_n e)_{ij}$ is also positive. This means that exe and $ex_n e$ 97
(for n large), both elements in G_1 , must correspond to the same permutation π on 98
 $\{1, 2, \dots, p\}$ in the following sense: $C_i \times C_j$ block of exe has all elements strictly 99
positive iff $j = \pi(i)$. 100

Since G_1 is compact (in fact, finite being a compact group of stochastic matrices), 101
the correspondence $g \rightarrow \pi$ from G_1 to the group of permutations on $\{1, 2, \dots, p\}$ is 102
an isomorphism (though possibly not onto). This means that there are elements y_n 103
in $\bigcup_{m=1}^{\infty} S(\mu_2)^m$ such that for each n , $ex_n e$ and $fy_n f$ have the same skeleton so that the 104
element exe in K_1 and the element fyf in K_2 (when y is a limit point of the sequence 105
 y_n) must correspond to the same permutation π on $\{1, 2, \dots, p\}$. Reversing the roles 106
of G_1 and G_2 , it is clear that the groups G_1 and G_2 are isomorphic. Thus, we have 107
proven the following theorem. 108

Theorem A.2. *In the presence of Assumption (*), the group factors G_1 and G_2 in 109
the Rees–Suskevitch product representations of K_1 and K_2 are isomorphic. □ 110*

Consider again the Rees–Suskevitch product representations $X_1 \times G_1 \times Y_1$ 111
of K_1 . Let N_1 be the smallest normal subgroup of G_1 such that $Y_1 X_1 \subset N_1$. Define 112
the map $\Phi_1 : S_1 \rightarrow \frac{G_1}{N_1}$, the factor group of G_1 by N_1 as follows: 113

$$\Phi_1(s) = ese N_1,$$

where e is the fixed idempotent in K_1 such that $G_1 = eK_1e$, $X_1 \equiv E(K_1e)$, and $Y_1 \equiv E(eK_1)$. As is shown in Chap. 2, Theorem 2.18, Φ_1 is a continuous homomorphism onto, and if we define the probability measure $\tilde{\mu}_1$ on $\frac{G_1}{N_1}$ by $\tilde{\mu}_1(B) = \mu_1(\Phi_1^{-1}(B))$, then the sequence μ_1^n on S_1 converges weakly if the sequence $\tilde{\mu}_1^n$ on $\frac{G_1}{N_1}$ converges weakly.

As we defined N_1 and Φ_1 above, we similarly define N_2 and Φ_2 . Now we claim that the homomorphic group images $\frac{G_1}{N_1}$ and $\frac{G_2}{N_2}$ are isomorphic under Assumption (*).

To prove this claim, we consider again the Rees–Suskevitch decomposition $X_1 \times G_1 \times Y_1$ and $X_2 \times G_2 \times Y_2$ of K_1 and K_2 (that we considered earlier), where $G_1 = eK_1e$ and $G_2 = fK_2f$, as shown in Theorem A.2, are isomorphic. Notice that $Y_1X_1 \subset G_1$ and $Y_2X_2 \subset G_2$. If we can show that the sets Y_1X_1 and Y_2X_2 both correspond to the same set of permutations under the usual isomorphisms from the groups G_1 and G_2 to the group of permutations on $\{1, 2, \dots, p\}$, then it follows immediately that the factor groups $\frac{G_1}{N_1}$ and $\frac{G_2}{N_2}$ are also isomorphic, where N_1 and N_2 are the smallest normal subgroups containing Y_1X_1 and Y_2X_2 , respectively. To show this, let $y_1 \in Y_1$ and $x_1 \in X_1$. Let $y_{1n} \in \bigcup_{m=1}^{\infty} S(\mu_1)^m$, $x_{1n} \in \bigcup_{m=1}^{\infty} S(\mu_1)^m$, $y_{1n} \rightarrow y_1$, $x_{1n} \rightarrow x_1$, and such that whenever $(y_1)_{ij} > 0$, $(y_{1n})_{ij}$ is also so, and whenever $(x_1)_{ij} > 0$, $(x_{1n})_{ij}$ is also so.

Choose x_{2n} and y_{2n} in $\bigcup_{m=1}^{\infty} S(\mu_2)^m$ such that for each n , x_{2n} and x_{1n} have the same skeleton, and y_{2n} and y_{1n} have the same skeleton. Let g_n be the identity in the group of limit points of the sequence $\{(fy_{2n})^m : m \geq 1\}$, and h_n be the identity in the group of limit points of the sequence $\{(x_{2n}f)^m : m \geq 1\}$. Then, $fg_n = g_n$ and $h_nf = h_n$. Choose M and L sufficiently large so that $[(fy_{2n})^M]_{ij}$ is positive, whenever $(g_n)_{ij} > 0$, and $[(x_{2n}f)^L]_{ij} > 0$, whenever $(h_n)_{ij} > 0$.

Note that the elements $e(e y_{1n})^M(x_{1n}e)^L e$ and $f(f y_{2n})^M(y_{1n}f)^L f$ have the same skeleton and, therefore, represent the group elements in G_1 and G_2 that correspond to the same permutation on $\{1, 2, \dots, p\}$. Also, $[e(e y_{1n})^M(x_{1n}e)^L e]_{ij}$ is positive whenever the element $[e(e y_1)^M(x_1e)^L e]_{ij} = (y_1x_1)_{ij}$ is positive. Similarly, the element $[f(f y_{2n})^M(x_{2n}f)^L f]_{ij} > 0$, whenever the element $[fg_nh_nf]_{ij} = (g_nh_n)_{ij} > 0$, and thus, both represent the same group element g_nh_n . Notice that $g_n^2 = g_n = fg_n \in fK_2$ so that $g_n \in Y_2$, and $h_n^2 = h_n = h_nf \in K_2f$ so that $h_n \in X_2$. Thus, with every element y_1x_1 in Y_1X_1 , we have associated an element g_nh_n in Y_2X_2 that corresponds to the same permutation on $\{1, 2, \dots, p\}$. Reversing the roles of K_1 and K_2 , it is clear that the sets Y_1X_1 and Y_2X_2 both correspond to the same set of permutation under the usual isomorphism from the groups G_1 and G_2 to the group of permutations on $\{1, 2, \dots, p\}$. Thus, we have proven:

Theorem A.3. *Under Assumption (*), the homomorphic group images $\frac{G_1}{N_1}$ and $\frac{G_2}{N_2}$ are isomorphic.* 153
□ 154

Now we are in a position to draw the promised conclusion of this appendix. 155

Theorem A.4. *Under Assumption (*), the sequence μ_1^n converges weakly iff μ_2^n does so.* 156
157

Proof. Let $S_1, S_2, K_1, K_2, G_1, G_2, N_1$ and N_2 be as before. Let $e \in K_1$ (and also the identity of G_1) and $f \in K_2$ (and also the identity of G_2) be idempotents with the same skeleton (and the same basis). For $B \subset \frac{G_1}{N_1}$, define the probability measure $\tilde{\mu}_1$ by 158
159
160
161

$$\tilde{\mu}_1(B) = \mu_1(\Phi_1^{-1}(B)),$$

where $\Phi_1 : S \rightarrow \frac{G_1}{N_1}$ is defined by 162

$$\Phi_1(x) = exe N_1.$$

We define Φ_2 and $\tilde{\mu}_2$ similarly. 163

Notice that $S(\tilde{\mu}_1) = \{exe N_1 : x \in S(\mu_1)\}$ and $S(\tilde{\mu}_2) = \{fyf N_2 : y \in S(\mu_2)\}$. It is known (Theorem 2.18 in Chap. 2) that μ_1^n (respectively, μ_2^n) converges weakly iff $\tilde{\mu}_1^n$ (respectively, $\tilde{\mu}_2^n$) does. 164
165
166

Notice that for $x \in S(\mu_1)$, there exists $y \in S(\mu_2)$ such that x and y (and therefore, exe and fyf) have the same skeleton. It follows from the proof of Theorem A.3 that the cosets $exe N_1$ and $fyf N_2$ correspond to the same set of permutations under the isomorphisms from the groups G_1 and G_2 to the group of permutations on $\{1, 2, \dots, p\}$. Similarly, for each y in $S(\mu_2)$, there exists $x \in S(\mu_1)$ such that a similar assertion is again true for the sets $fyf N_2$ and $exe N_1$. As a result, it follows that under the isomorphism from $\frac{G_1}{N_1}$ to $\frac{G_2}{N_2}$, observed in the proof of Theorem A.3, $S(\tilde{\mu}_2)$ is the isomorphic image of $S(\tilde{\mu}_1)$. Thus, $\tilde{\mu}_1^n$ converges weakly iff $\tilde{\mu}_2^n$ also does so. The proof of Theorem A.4 is now complete. 167
168
169
170
171
172
173
174
□ 175

For more details and generalizations, see [115]. 176

Appendix B

An Example Due to Chamayou and Letac

Here, we present an example of a sequence of i.i.d. $d \times d$ (random) stochastic matrices (X_n) , where each row of X_1 has Dirichlet density, the rows of X_1 are independent such that the sequence of products $X_n X_{n-1} \dots X_1$ converges almost surely to a (random) stochastic matrix Z with identical rows where each row of Z has also Dirichlet density.

The Dirichlet density, a generalization of Beta density, with (positive) parameters $\alpha_1, \alpha_2, \dots, \alpha_d$, is given by:

$$\frac{\Gamma\left(\sum_{i=1}^d \alpha_i\right)}{\Gamma(\alpha_1) \Gamma(\alpha_2) \dots \Gamma(\alpha_d)} x_1^{\alpha_1-1} x_2^{\alpha_2-1} \dots x_{d-1}^{\alpha_{d-1}-1} (1 - x_1 - x_2 - \dots - x_{d-1})^{\alpha_d-1}, \quad (\text{B.1})$$

where for $1 \leq i \leq d-1$, $0 \leq x_i \leq 1$. This density is zero if any of the x_i s is less than 0 or greater than 1 or if $x_1 + x_2 + \dots + x_{d-1} > 1$.

The lemma below is then easily verified.

Lemma B.1. Let $Y_1, Y_2, \dots, Y_k, k \geq 2$, be independent Gamma random variables so that their joint density function is given by

$$\mathcal{L}(y_1, y_2, \dots, y_k) = \prod_{i=1}^k \frac{1}{\Gamma(\alpha_i)} y_i^{\alpha_i-1} e^{-y_i} I_{(0,\infty)}(y_i).$$

Let $W_i = Y_i / \sum_{j=1}^k Y_j$, $1 \leq i \leq k-1$, $W_k = \sum_{j=1}^k Y_j$. Then the random vector $(W_1, W_2, \dots, W_{k-1})$ is independent of W_k , such that $(W_1, W_2, \dots, W_{k-1})$ has Dirichlet density with parameters $\alpha_1, \alpha_2, \dots, \alpha_k$ given by

$$\frac{\Gamma\left(\sum_{j=1}^k \alpha_j\right)}{\Gamma(\alpha_1) \Gamma(\alpha_2) \dots \Gamma(\alpha_k)} w_1^{\alpha_1-1} w_2^{\alpha_2-1} \dots w_{k-1}^{\alpha_{k-1}-1} (1 - w_1 - w_2 - \dots - w_{k-1})^{\alpha_k-1}$$

for $w_1 > 0, w_2 > 0, \dots, w_{k-1} > 0, w_1 + w_2 + \dots + w_{k-1} \leq 1$, and $= 0$, otherwise, 18
and W_k has Gamma density with parameter $\sum_{j=1}^k \alpha_j$. $\square$ 19

Now let $A = (a_{ij})$ be a $d \times d$ matrix with all entries strictly positive. Let 20
 $Y \equiv (Y_1, Y_2, \dots, Y_d)$ be a $1 \times d$ random probability vector and $X \equiv (X_{ij})$ 21
be a $d \times d$ random stochastic matrix such that Y is independent of X . Suppose 22
that for $1 \leq i \leq d$, $(X_{i1}, X_{i2}, \dots, X_{i(d-1)})$ has Dirichlet density with parameters 23
 $(a_{i1}, a_{i2}, \dots, a_{id})$, and that $Y_1, Y_2, \dots, Y_{d-1}$ has Dirichlet density with parameters 24
 $\sum_{j=1}^d a_{ij}, i = 1, 2, \dots, d$. Then we claim the following: 25

Write $W = YX$. Then W has Dirichlet density with parameters $\sum_{i=1}^d a_{ij}, j =$ 26
 $1, 2, \dots, d$. 27

We prove this here using an idea due to Steffen Lauritzen (see [34]). Let Z be 28
a $d \times d$ random nonnegative matrix with independent entries such that the random 29
variable Z_{ij} has Gamma density with parameter a_{ij} . 30

Write: $L_i = \sum_{j=1}^d Z_{ij}, K_j = \sum_{i=1}^d Z_{ij}, T = \sum_{i=1}^d L_i = \sum_{j=1}^d K_j$. Define: $X_{ij}^* =$ 31
 $\frac{Z_{ij}}{L_i}, Y_i^* = \frac{L_i}{T}, W_j^* = \frac{K_j}{T}$. Then we have: $W^* = Y^* X^*$. Also, it is clear that each 32

K_j has Gamma density with parameter $\sum_{i=1}^d a_{ij}, 1 \leq j \leq d$. It also follows from 33
Lemma B.1 that $(W_1^*, W_2^*, \dots, W_{d-1}^*)$ (that is to say, W^*) has Dirichlet density 34
with parameters $\sum_{i=1}^d a_{ij}, j = 1, 2, \dots, d$, whereas Y^* , that is $(Y_1^*, \dots, Y_{d-1}^*)$ has 35

Dirichlet density with parameters $\sum_{j=1}^d a_{ij}, i = 1, 2, \dots, d$. The Lemma also says 36

that for each i , $(X_{i1}^*, X_{i2}^*, \dots, X_{i(d-1)}^*)$ is independent of L_i (and also, independent 37
of all $L_j, j \neq i$). Thus, Y^* and X^* are independent. Note also that the rows of the 38
stochastic matrix X^* are independent, and the i th row of X^* has Dirichlet density 39
with parameters $a_{i1}, a_{i2}, \dots, a_{id}, 1 \leq i \leq d$. This proves the claim. 40

The following theorem will be now relevant and it is given in [34]. See also 41
Exercise 4.22. The proof is not difficult. 42

Proposition B.1. Let $X_1, X_2, \dots$ be a sequence of $d \times d$ i.i.d. random stochastic 43
matrices such that $P(X_i \text{ is strictly positive}) > 0$. Then the sequence $X_n X_{n-1} \dots X_1$ 44
converges almost surely to a random $d \times d$ stochastic matrix Y with identical rows 45
such that if μ is the distribution of X_1 and λ that of Y so that for any Borel set B , 46
 $\mu(B) = P(X_1 \in B)$ and $\lambda(B) = P(Y \in B)$, then $\lambda * \mu = \lambda$ and λ is the only 47
solution of this convolution equation. $\square$ 48

The following is the last and the main result of this appendix.

49

Theorem B.1. *Let $X_1, X_2, \dots$ be a sequence of i.i.d. random $d \times d$ stochastic matrices such that the rows of X_1 are independent, and the i th row of X_1 has Dirichlet density with parameters $a_{i1}, a_{i2}, \dots, a_{id}$. Suppose that the*

52

$$\sum_{j=1}^d a_{ij} = \sum_{j=1}^d a_{ji} = r_i, \quad 1 \leq i \leq d.$$

Then, if $Z = \lim_{n \rightarrow \infty} X_n X_{n-1} \dots X_1$, almost surely, Z has identical rows and each row of Z has Dirichlet density with parameters $r_1, r_2, \dots, r_d$.

□ 54

Proof. By Proposition B.1, Z exists almost surely, Z has identical rows, and if λ is the distribution of Z , then λ is the *unique* solution of $\lambda * \mu = \lambda$, where μ is the distribution of X_1 . The discussion preceding Proposition B.1 shows that W^* and Y^*

there will have the same distribution iff $\sum_{i=1}^d a_{ij} = \sum_{i=1}^d a_{ji}$ for each i , $1 \leq i \leq d$. The proof follows.

□ 59

Let us mention the paper of Van Assche [231], where similar ideas (with different proofs) appeared first for 2×2 stochastic matrices.

60
61

Appendix C

Asymptotic Behavior of $\|X_n X_{n-1} \dots X_0 u\|$ for I.I.D. Random Nonnegative Matrices

Here, we follow the method of Furstenberg and Kifer [71] and use their arguments to describe the asymptotic behavior of $\|X_n X_{n-1} \dots X_0 u\|$, where (X_n) is a sequence of i.i.d. $d \times d$ nonnegative random matrices and u is a $d \times 1$ nonnegative vector. It is natural to do this here for nonnegative matrices as much of the Chap. 4 deals with such matrices. In [71], matrices were invertible real matrices, while here they are nonnegative, but not necessarily invertible. We show that results very similar to those in [71] also hold here, under slightly different conditions. The main theorem here is Theorem C.2.

Let $X_0, X_1, X_2, \dots$ be i.i.d. $d \times d$ random nonnegative (that is, with nonnegative entries) matrices with distribution μ . Let S_μ be the support of μ and let S be the closed (with usual matrix topology) multiplicative semigroup generated by S_μ , so that

$$S = \text{cl} \left(\bigcup_{n=1}^{\infty} S_\mu^n \right). \quad (\text{C.1})$$

We consider throughout the following conditions:

$$\mu(J_c) = 0, \quad (\text{C.2})$$

where J_c is the set of all matrices in S with at least one column with only zero entries;

$$\int [\log \|s\| + |\log k(s)|] \mu(ds) < \infty, \quad (\text{C.3})$$

where

$$k(s) = \min \left\{ \sum_{i=1}^d s_{ij} : 1 \leq j \leq d \right\}. \quad (\text{C.4})$$

Let us define the compact set $Y \subset (R^+)^d$, with usual topology, by

$$Y = \{y = (y_1, y_2, \dots, y_d) : y_i \geq 0, 1 \leq i \leq d, y_1^2 + y_2^2 + \dots + y_d^2 = 1\}. \quad (\text{C.5})$$

370 C Asymptotic Behavior of $\|X_n X_{n-1} \dots X_0 u\|$ for I.I.D. Random Nonnegative Matrices

Let A be a $d \times d$ nonnegative matrix with no zero columns. For $y \in Y$, let 21

$$A \cdot y = (Ay) / \|Ay\|,$$

where 22

$$(Ay)_i = \sum_{j=1}^d A_{ij} y_j, \quad \|y\| = \sqrt{\sum y_i^2}.$$

This action is then welldefined and 23

$$(AB) \cdot y = A \cdot (B \cdot y),$$

whenever A and B are two nonnegative matrices with no zero columns. It is also 24
easily verified that for β_1, β_2 in $P(S)$, the set of all (Borel) probability measures 25
on S , and $\lambda \in P(Y)$, 26

$$[\beta_1 * (\beta_2 * \lambda)](\mathcal{A}) = [(\beta_1 * \beta_2) * \lambda](\mathcal{A}),$$

for any Borel subset $\mathcal{A} \subset Y$, where 27

$$\beta_2 * \lambda(\mathcal{A}) = \int \beta_2\{s \in S: s \cdot y \in \mathcal{A}\} \lambda(dy).$$

Notice that $P(Y)$ is compact in the weak* topology (by Banach–Alaoglu’s theorem, 28
see [168]), since Y is compact. Therefore, since 29

$$\|\mu * \beta_n - \beta_n\| \rightarrow 0 \text{ as } n \rightarrow \infty,$$

where 30

$$\beta_n \equiv \left(\frac{1}{n} \sum_{k=1}^n \mu^k \right) * \beta, \quad \beta \in P(Y),$$

any weak* cluster point of (β_n) [and (β_n) has at least one weak* cluster point] is a 31
 μ -invariant probability measure in $P(Y)$. 32

From now on, we assume throughout condition (C.2). 33

Let $u \in Y$ and $W_n \equiv X_n X_{n-1} \dots X_0$. Write: 34

$$Z_0^* \equiv (X_0, u), \dots, Z_{n+1}^* \equiv (X_{n+1}, W_n \cdot u).$$

Then (Z_n^*) is a Markov process on the set $M \equiv S_\mu \times Y$ with a transition function 35
given by 36

$$P(Z_{n+1}^* \in \mathcal{D} \mid Z_n^* = (s, y)) = \mu\{t \in S_\mu : (t, s \cdot y) \in \mathcal{D}\}. \quad (\text{C.6})$$

C Asymptotic Behavior of $\|X_n X_{n-1} \dots X_0 u\|$ for I.I.D. Random Nonnegative Matrices 371

Let λ be an invariant probability measure for (Z_n^*) on M so that 37

$$\lambda(\mathcal{D}) = \int P((s, y), \mathcal{D}) \lambda(d(s, y)). \quad (\text{C.7})$$

Let us define λ_0 in $P(Y)$ by 38

$$\lambda_0(\mathcal{B}) = \lambda\{(s, y): s \cdot y \in \mathcal{B}\}. \quad (\text{C.8})$$

It follows from (C.6) and (C.7) that for $\mathcal{D} = \mathcal{A} \times \mathcal{B} \subset M$, 39

$$\lambda(\mathcal{D}) = \mu(\mathcal{A}) \lambda_0(\mathcal{B}) \quad (\text{C.9})$$

so that λ_0 uniquely determines λ . Also if f is any bounded continuous function 40
on Y , then for the function g on M defined by 41

$$g(s, y) = f(s \cdot y),$$

we have 42

$$\begin{aligned} \int f d\lambda_0 &= \int f(s \cdot y) \lambda(d(s, y)) \\ &= \int g(s, y) \lambda(d(s, y)) \\ &= \iint g(t, s \cdot y) \mu(dt) \lambda(d(s, y)) \\ &= \iint f(ts \cdot y) \mu(dt) \lambda(d(s, y)) \\ &= \int \left[\int f(t \cdot (s \cdot y)) \mu(dt) \right] \lambda(d(s, y)) \\ &= \int \left[\int f(t \cdot y) \mu(dt) \right] \lambda_0(dy), \end{aligned}$$

which implies 43

$$\mu * \lambda_0 = \lambda_0. \quad (\text{C.10})$$

Now assume that S_μ is compact; then M is compact. It is easily verified that 44
condition (C.2) implies that $\mu^n(J_c) = 0$ for all positive integers n . Let us define 45
the function h on M by 46

$$h(s, y) \equiv -\log \left(\frac{\|y\|}{\|sy\|} \right) = \log \|sy\|; \quad (\text{C.11})$$

372 C Asymptotic Behavior of $\|X_n X_{n-1} \dots X_0 u\|$ for I.I.D. Random Nonnegative Matrices

then h is continuous and

47

$$\begin{aligned} h(s_2, s_1 \cdot y) &= -\log \left(\frac{\|s_1 \cdot y\|}{\|s_2(s_1 \cdot y)\|} \right) \\ &= -\log \left(\frac{\|s_1 y\|}{\|(s_2 s_1) y\|} \right). \end{aligned}$$

By Theorem 4.22, with probability 1,

48

$$\lim_{n \rightarrow \infty} \sup \frac{1}{n} \sum_{k=0}^n g(Z_k^*) \leq \sup \left\{ \int g d\lambda : \lambda \text{ satisfies (C.7)} \right\}. \quad (\text{C.12})$$

This means that for any $u \in Y$, almost surely

49

$$\lim_{n \rightarrow \infty} \sup \frac{1}{n} \log \|W_n u\| \leq \sup_{\lambda} \left\{ \int \log \|sy\| d\lambda(s, y) \right\}, \quad (\text{C.13})$$

where the supremum is taken over all λ satisfying (C.7).

50

Let us now drop the assumption of compactness on S_μ and instead, assume condition (C.3). We also continue to assume condition (C.2). Consider the set $M^* = S_\mu^* \times Y$, where S_μ^* is the one point compactification of S_μ . With the functions h , defined as in (C.11), we define h_m on M^* as follows:

54

$$h_m = h \text{ for } y \in Y \text{ and } S \in S_\mu \text{ with } \|s\| \leq m \text{ and } k(s) \geq \frac{1}{m},$$

where $k(s)$ is as defined in (C.4). Let h_m be extended continuously to all of M^* and still bounded by $\log m + \log d$. Let us define the sets A_m , B_m , and C_m by

56

$$\begin{aligned} A_m &= \{s \in S_\mu : \|s\| > m\}, \\ C_m &= \left\{s \in S_\mu : k(s) < \frac{1}{m}\right\} \text{ and } B_m = A_m \cup C_m. \end{aligned}$$

Note that for any $s \in S$ and $y \in Y$,

57

$$\frac{1}{d} \cdot k(s) \leq \|sy\| \leq \|s\|. \quad (\text{C.14})$$

The second inequality is obvious. For the first, notice that for any $y \in Y$, $y = (y_1, y_2, \dots, y_d)$, there is some j , $1 \leq j \leq d$, such that

59

$$y_j \geq \frac{1}{\sqrt{d}}$$

C Asymptotic Behavior of $\|X_n X_{n-1} \dots X_0 u\|$ for I.I.D. Random Nonnegative Matrices 373

and for this j ,

$$\|sy\|^2 \geq \frac{1}{d} \sum_{i=1}^d s_{ij}^2 \geq \frac{1}{d^2} \left(\sum_{i=1}^d s_{ij} \right)^2. \quad 60$$

The last inequality follows from the observation that for any n real numbers $a_1, a_2, \dots, a_n$, 61

$$\left(\sum_{i=1}^n a_i \right)^2 \leq n \left(\sum_{i=1}^n a_i^2 \right), \quad 62$$

which can be established immediately by induction on n . This establishes (C.14). It follows from (C.14) that 63 64

$$|\log \|sy\|| \leq \log d + |\log k(s)| + |\log \|s\||. \quad (\text{C.15})$$

It follows that we have almost surely

$$\begin{aligned} & \left| \frac{1}{n} \sum_{k=0}^n g(Z_k^*) - \frac{1}{n} \sum_{k=0}^n g_m(Z_k^*) \right| \\ & \leq \frac{1}{n} \sum_{k=0}^n [\log d + |\log k(X_k)| + |\log \|X_k\|| + \log m] I_{B_m}(X_k), \end{aligned} \quad 65$$

which converges as $n \rightarrow \infty$, by the ergodic theorem to

$$\int_{B_m} [|\log k(s)| + |\log \|s\||] \mu(ds) + [\log d + \log m] \mu(B_m).$$

This expression goes to zero as $m \rightarrow \infty$ because of our assumption of condition (C.3). Condition (C.3) also implies 67 68

$$\lim_{m \rightarrow \infty} \left| \int g d\lambda - \int g_m d\lambda \right| = 0.$$

Since the inequality (C.20) holds for the function g_m defined on all of the compact space M^* , it follows from the preceding considerations that (C.13) holds under conditions (C.2) and (C.3) (and in this case, for (C.13) to hold, the compactness of S_μ is not needed). 69 70 71 72

Let us now write 73

$$b(\mu) \equiv \sup\{a(\mu, \nu): \mu * \nu = \nu, \nu \in P(Y)\},$$

where

$$\begin{aligned} a(\mu, \nu) & \equiv \int \log \|sy\| d\lambda(s, y) \\ & = \iint \log \|sy\| \mu(ds) \nu(dy), \end{aligned} \quad 74$$

374 C Asymptotic Behavior of $\|X_n X_{n-1} \dots X_0 u\|$ for I.I.D. Random Nonnegative Matrices

if ν is the measure λ_0 in (C.8). Notice that it follows from condition (C.3) and 75
inequality (C.15) that $b(\mu)$ is finite, so that for each positive integer n , there exists 76
 ν_n in $P(Y)$ such that 77

$$b(\mu) < a(\mu, \nu_n) + \frac{1}{n}.$$

Since $\nu_n \in P(Y)$ and Y is compact, there is a subsequence (ν_{n_i}) such that 78
 $\nu_{n_i} \rightarrow \nu_0$ weakly so that 79

$$\int h_1 d\nu_{n_i} \rightarrow \int h_1 d\nu_0,$$

where

$$h_1(y) = \int \log \|sy\| \mu(ds),$$

which is a bounded continuous function. Thus,

$$b(\mu) = a(\mu, \nu_0).$$

This means there exists $\nu_0 \in P(Y)$, $\mu * \nu_0 = \nu_0$ such that for any $u \in Y$, we have 82
almost surely 83

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log \|W_n u\| \leq a(\mu, \nu_0). \quad (\text{C.16})$$

Notice that if for each ν with $\mu * \nu = \nu$, $\nu \in P(Y)$, $a(\mu, \nu)$ is the same, then 84
besides (C.13) we have almost surely, 85

$$\begin{aligned} & \limsup_{n \rightarrow \infty} \frac{1}{n} [-\log \|W_n u\|] \\ & \leq \iint -\log \|sy\| \mu(ds) \nu(dy) \end{aligned}$$

so that

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log \|W_n u\| \geq a(\mu, \nu). \quad (\text{C.17})$$

This means that if ν is the unique μ -invariant probability measure on Y (or if $a(\mu, \nu)$ 87
is the same for all μ -invariant ν in $P(Y)$), then for any $u \in Y$, we have almost surely, 88

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \|W_n u\| = a(\mu, \nu). \quad (\text{C.18})$$

Now let U be a Y -valued random variable with distribution $\nu \in P(Y)$, $\mu * \nu = \nu$, 89
such that U is independent of the X_i s. Then if 90

$$h(s, y) = \log(\|y\| / \|sy\|),$$

the process

$$h(X_{n+1}, W_n \cdot U), \quad n \geq 1,$$

91

C Asymptotic Behavior of $\|X_n X_{n-1} \dots X_0 u\|$ for I.I.D. Random Nonnegative Matrices 375

is stationary. It follows from the ergodic theorem that 92

$$\frac{1}{n} \log \|W_n U\| \rightarrow H, \text{ almost surely,} \quad (\text{C.19})$$

where H is a random variable such that $E(H) = a(\mu, \nu)$. This means 93

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log \|W_n\| \geq a(\mu, \nu) \quad (\text{C.20})$$

on a set of positive probability, since $H \geq E(H)$ on a set of positive probability. 94

Notice that for given N , $1 \leq N < n$, 95

$$\|W_n\| \leq \|X_n \dots X_{N+1}\| \|W_n\|$$

so that 96

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log \|W_n\| \leq \liminf_{n \rightarrow \infty} \frac{1}{n} \log \|X_n \dots X_{N+1}\|.$$

This means 97

$$\begin{aligned} & \left\{ \liminf_{n \rightarrow \infty} \frac{1}{n} \log \|W_n\| < a(\mu, \nu) \right\} \\ & \supset \left\{ \liminf_{n \rightarrow \infty} \frac{1}{n} \log \|X_n \dots X_{N+1}\| < a(\mu, \nu) \right\} \end{aligned}$$

Then these two events are equal almost surely, since the (X_i) are i.i.d. so they have 98
the same probability. By the Kolmogorov zero-one law, these two events (see (C.20)) 99
have probability zero. It follows that 100

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log \|W_n\| \geq a(\mu, \nu), \quad (\text{C.21})$$

almost surely. Since the left side of (C.21) is independent of ν , it follows from (C.16) 101
that 102

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \|W_n\| = a(\mu, \nu) = b(\mu), \text{ almost surely.} \quad (\text{C.22})$$

We call $b(\mu)$ the rate of growth. Note that 103

$$\max_i \|sy^{(i)}\|,$$

where 104

$$y^{(i)} = (0, \dots, 0, 1, 0, \dots, 0),$$

the i th entry of $y^{(i)}$ being 1, is a norm; therefore, (C.16) and (C.20) lead to (C.21). 105

If in (C.19), the random variable U has its distribution ν chosen as one of the 106
extreme points of the compact convex set of μ -invariant probability measures in 107

376 C Asymptotic Behavior of $\|X_n X_{n-1} \dots X_0 u\|$ for I.I.D. Random Nonnegative Matrices

$P(Y)$, then the process $h(X_{n+1}, W_n \cdot U)$ is ergodic (see [19, p. 118, problem 6.11]).
Consequently in (C.19), the random variable H becomes a constant almost surely,
and

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \|W_n U\| = a(\mu, \nu), \text{ almost surely.} \quad (\text{C.23})$$

It follows from (C.23) that

$$\begin{aligned} 1 &= P \left(\lim_{n \rightarrow \infty} \frac{1}{n} \log \|W_n U\| = a(\mu, \nu) \right) \\ &= \int P \left(\lim_{n \rightarrow \infty} \frac{1}{n} \log \|W_n U\| = a(\mu, \nu) \mid U = u \right) \nu(du), \end{aligned}$$

which implies

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \|W_n u\| = a(\mu, \nu), \text{ almost surely,} \quad (\text{C.24})$$

for ν almost all u , where ν is an extremal μ -invariant measure in $P(Y)$.

Notice that

$$\begin{aligned} &\sup\{a(\mu, \nu) : \mu * \nu = \nu, \nu \in P(Y)\} \\ &= \sup\{a(\mu, \nu) : \mu * \nu = \nu, \nu \text{ is one of the extreme points} \\ &\quad \text{in the set of } \mu\text{-invariant measures in } P(Y)\}. \end{aligned}$$

By the Krein–Milman Theorem of functional analysis, it follows from (C.18) that
for all $u \neq 0$ in $(R^+)^d$,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \|W_n u\| = b(\mu), \text{ almost surely,} \quad (\text{C.25})$$

if for each μ -invariant extremal measure ν , $a(\mu, \nu) = b(\mu)$.

Suppose there exists $u \neq 0$ in $(R^+)^d$ such that (C.25) is not valid. Then there
exists an extremal μ -invariant measure ν such that

$$a(\mu, \nu) < b(\mu).$$

Then for this ν , (C.24) holds and therefore

$$\nu(L_\nu) = 1,$$

where L_ν is defined by

$$L_\nu = \left\{ u \in R^d : \lim_{n \rightarrow \infty} \sup \frac{1}{n} \log \|W_n u^+\| \leq a(\mu, \nu) \text{ almost surely} \right\} \cup \{0\}.$$

Here,

$$u^+ = (|u_1|, |u_2|, \dots, |u_d|),$$

if $u = (u, u, \dots, u_d)$. It is easily verified that L_v is a subspace.

Since there exists a μ -invariant extremal measure β such that $a(\mu, v) < a(\mu, \beta)$ and since (C.24) holds for β replacing v there, it follows that L_v is a proper subspace. Because of (C.24), $v(L_v) = 1$. Let L_v^* be the smallest subspace such that $v(L_v^*) = 1$. Note:

$$L_v^* = \cap \{V : V \text{ is a subspace with } v(V) = 1\}$$

so that $(L_v^*)^c$ is the union (and therefore, a countable union) of open sets with zero v -measure. Now

$$\begin{aligned} 1 &= v(L_v^*) = \mu * v(L_v^*) \\ &= \int v(x^{-1} L_v^*) \mu(dx), \end{aligned}$$

where $x^{-1} L_v^* = \{y \in Y : x \cdot y \in L_v^*\}$. This means that for μ -almost all x ,

$$v(x^{-1} L_v^*) = 1.$$

Since $x^{-1} L_v^*$ is a subspace, because of the minimality of L_v^* ,

$$L_v^* \subset x^{-1} L_v^* \text{ or } x L_v^* \subset L_v^*, \text{ almost everywhere } (\mu).$$

Thus, we have proved Theorem C.1.

Theorem C.1. Suppose $X_0, X_1, X_2, \dots$ are i.i.d. $d \times d$ matrices with nonnegative entries with distribution μ such that conditions (C.2) and (C.3) hold. Then for any $u \in (R^+)^d$, $u \neq 0$, either

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \|X_n X_{n-1} \dots X_0 u\| = b(\mu), \text{ almost surely,}$$

or for some proper μ -invariant subspace L (that is, $xL \subset L$ for μ almost all x) $\subset R^d$, for every $u \neq 0$ in L ,

$$\lim_{n \rightarrow \infty} \sup \frac{1}{n} \log \|X_n X_{n-1} \dots X_0 u\| \leq \alpha, \text{ almost surely,}$$

where α is a constant less than $b(\mu)$.

Notice that if L is a μ -invariant subspace of R^d , then every matrix g in the support of μ has the form

$$g = \begin{pmatrix} g_1 & g_2 \\ 0 & g_3 \end{pmatrix}, \quad (\text{C.26})$$

378 C Asymptotic Behavior of $\|X_n X_{n-1} \dots X_0 u\|$ for I.I.D. Random Nonnegative Matrices

by choosing a basis of R^d whose initial vectors form a basis of L , where g_1 represents the restriction of g to L . Note that 141
142

$$\begin{pmatrix} g_1 & g_2 \\ 0 & g_3 \end{pmatrix} \begin{pmatrix} h_1 & h_2 \\ 0 & h_3 \end{pmatrix} = \begin{pmatrix} g_1 h_1 & g_1 h_2 + g_2 h_3 \\ 0 & g_3 h_3 \end{pmatrix}.$$

This means μ induces a probability measure μ_L on the set of matrices 143

$$\{g_1 : g \in S_\mu, g \text{ as in (C.26)}\}$$

and a probability measure $\mu' (\equiv \mu_{R^d|L})$ on the set of matrices 144

$$\{g_3 : g \in S_\mu, g \text{ as in (C.26)}\}$$

Let us now establish that 145

$$b(\mu) = \max \{b(\mu_L), b(\mu')\}, \quad (\text{C.27})$$

where $b(\mu_L)$ and $b(\mu')$ are defined for μ_L and μ' , respectively, in the same way 146
 $b(\mu)$ was defined earlier for the measure μ . Notice that the result in (C.22) can 147
be established under weaker conditions, namely, when $E(\log^+ \|X_1\|) < \infty$, using 148
Kingman's subadditive ergodic theorem (see [131]). Indeed letting 149

$$W_{mn} = \log \|X_{n-1} X_{n-2} \dots X_m\|, \quad m < n,$$

we see that for $m < n < p$, $W_{mp} \leq W_{mn} + W_{np}$. The distribution of W_{mn} depends 150
only on $n - m$ and $E(W_{01}^+) < \infty$, so that it follows from Kingman's subadditive 151
ergodic theorem that 152

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \|W_n\|$$

exists with probability 1. This limit is almost surely a constant by the Kolmogorov 153
zero-one law, as we explained when we established (C.22). Thus for the limit in 154
(C.22) to exist, condition (C.3) alone is sufficient. 155

Using the representation (C.26), X_1 can be written as 156

$$X_1 = \begin{pmatrix} X_{11} & X_{12} \\ 0 & X_{13} \end{pmatrix}.$$

Then 157

$$\max \{\log^+ \|X_{11}\|, \log^+ \|X_{13}\|\} \leq \log^+ \|X_1\|$$

so that such results as (C.22) also hold for μ_L and μ' . To establish (C.27), it is clear 158
that 159

$$\max \{b(\mu_L), b(\mu')\} \leq b(\mu).$$

C Asymptotic Behavior of $\|X_n X_{n-1} \dots X_0 u\|$ for I.I.D. Random Nonnegative Matrices 379

Suppose that the inequality here is strict. Using an argument from Furstenberg and Kifer [71], we consider the product 160
161

$$\begin{aligned} W_{2N} &= (X_{2N} X_{2N-1} \dots X_{N+1}) W_N \\ &= \begin{pmatrix} W'_{N1} & W'_{N2} \\ 0 & W'_{N3} \end{pmatrix} \begin{pmatrix} W_{N1} & W_{N2} \\ 0 & W_{N3} \end{pmatrix}. \end{aligned}$$

Now given $\delta > 0$, $\varepsilon > 0$, there exists N_0 such that for $N > N_0$ with probability greater than $1 - \delta$, the following inequalities hold: 162
163

$$\begin{aligned} \|W_{N1}\| &\leq e^{(1+\varepsilon)Nb(\mu_L)}, \\ \|W_{2N}\| &> e^{(1-\varepsilon)2Nb(\mu)}, \\ \|W_{N2}\| &\leq e^{(1+\varepsilon)Nb(\mu)}, \\ \|W_{N3}\| &\leq e^{(1+\varepsilon)Nb(\mu')}. \end{aligned} \tag{C.28}$$

These inequalities (not counting the second one) hold also for $\|W'_{N1}\|$, $\|W'_{N2}\|$, and $\|W'_{N3}\|$, respectively. Hence with probability greater than $1 - 7\delta$, 164
165

$$\begin{aligned} \|W_{2N}\| &\leq e^{(1+\varepsilon)2Nb(\mu_L)} + e^{(1+\varepsilon)2Nb(\mu')} \\ &+ e^{(1+\varepsilon)N[b(\mu') + b(\mu)]} + e^{(1+\varepsilon)N[b(\mu_L) + b(\mu)]}, \end{aligned}$$

which contradicts (C.28). This proves (C.27). 166

It follows from (C.27) that if L_1 and L_2 are two μ -invariant subspaces of R^d , then $L_1 + L_2$ is also a μ -invariant subspace of R^d and 167
168

$$b(\mu_{L_1+L_2}) = \max \{b(\mu_{L_1}), b(\mu_{L_2})\}.$$

Thus, there is a unique *maximal* μ -invariant proper subspace L_1 such that $b(\mu_{L_1}) < b(\mu)$. 169
170

Now as in (C.26), we can assume with no loss of generality 171

$$X_n = \begin{pmatrix} X_{n1} & X_{n2} \\ 0 & X_{n3} \end{pmatrix}. \tag{C.29}$$

By (C.27), 172

$$\max \{b(\mu_{L_1}), b(\mu_{R^d/L_1})\} = b(\mu);$$

then 173

$$b(\mu_{R^d/L_1}) = b(\mu).$$

Let $v \in R^d/L_1$. Then by Theorem C.1, if 174

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \|X_{n3} X_{n-1,3} \dots X_{0,3} v^+\|$$

380 C Asymptotic Behavior of $\|X_n X_{n-1} \dots X_0 u\|$ for I.I.D. Random Nonnegative Matrices

is not $b(\mu)$ with probability 1, then there is a proper subspace L_2 of R^d/L_1 with a 175
rate of growth less than 176

$$b(\mu_{R^d/L_1}) = b(\mu).$$

But this means the rate of growth of the proper subspace $L_1 + L_2$, which is also 177
 μ -invariant, is less than $b(\mu)$. This contradicts the maximality of L_1 . Thus for rep- 178
resentation (C.29), for any $v \in R^d/L_1$, we have almost surely 179

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \|X_{n3} X_{n-1,3} \dots X_0 v^+\| = b(\mu). \quad (\text{C.30})$$

Now let $w \notin L_1$. Then we can write 180

$$w = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix},$$

where $w_1 \in L_1$ and $w_2 \in R^d/L_1$. Writing X_n as in (C.29), we have 181

$$\begin{aligned} W_n &= \begin{pmatrix} W_{n1} & W_{n2} \\ 0 & W_{n3} \end{pmatrix}, \\ \|W_n w^+\| &= \left\| \begin{pmatrix} W_{n1} w_1^+ + W_{n2} w_2^+ \\ W_{n3} w_2^+ \end{pmatrix} \right\| \\ &\geq \|W_{n3} w_2^+\|. \end{aligned}$$

Thus, it follows from (C.30) and (C.16) that 182

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \|W_n w^+\| = b(\mu), \text{ almost surely.} \quad (\text{C.31})$$

By repeating this procedure, we have now proven Theorem C.2. 183

Theorem C.2. Let (X_n) , $n \geq 0$, be a sequence of i.i.d. $d \times d$ nonnegative matrices 184
with distribution μ satisfying conditions (C.2) and (C.3). There is a sequence of 185
subspaces 186

$$\{0\} \subset L_r \subset \dots \subset L_2 \subset L_1 \subset R^d$$

and a sequence of constants 187

$$b(\mu) > b_1(\mu) > \dots > b_r(\mu)$$

such that if $v \in L_i \setminus L_{i+1}$, then with probability 1, 188

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \|X_n X_{n-1} \dots X_0 v^+\| = b_i(\mu).$$

We remark that in the context of Theorem C.2 above the most well-known re- 189
sult of this type for nonsingular matrices was proven by Valery Oseledets in [182]. 190

In 1979, Raghunathan [194] gave another proof of the Oseledets' multiplicative ergodic theorem in [182], using Kingman's subadditive ergodic theorem in [131]. Ruelle in (Ergodic theory of differentiable dynamical systems, Inst. des Hautes Etudes Scient., Publ. Math. 50 (1979) 275–306) extended Oseledets' theorem to Hilbert spaces, and R. Mañé in "Geometric Dynamics" (Lecture Notes in Math., vol. 1007, Springer, 1983, 522–577) extended it to Banach spaces.

If (X_n) is a stationary Markov chain with state space S and f is a Borel measurable map from S to the $d \times d$ invertible matrices $GL(d, R)$, then applying the Oseledets' theorem [112] on the sequence $(f(X_n))$, we can find a sequence of Lyapunov constants $b_\alpha \leq b_{\alpha-1} \leq \dots \leq b_1$ and a sequence of decreasing subspaces L_i , $1 \leq i \leq d$, of R^d , such that

$$L_i = \left\{ v \in R^d : \lim_{n \rightarrow \infty} \frac{1}{n} \log \|f(X_n) f(X_{n-1}) \dots f(X_1) v\| \leq b_i \right\}.$$

In [77], Goldsheid and Margulis proved the following theorem: Let (X_n) be i.i.d. random matrices in $GL(d, R)$ with distribution μ . Let S be the semigroup generated by $S(\mu)$, the support of μ . If the algebraic closure of S in $GL(d, R)$, then the Lyapunov exponents b_i , $1 \leq i \leq d$, of the matrix products $\{X_n, X_{n-1} \dots X_1 : n \geq 1\}$ are all distinct.

Also, in this context, the paper of Guivarc'h ("Sur les exposants de Liapunoff des marches aleatoires", C. R. A. S., t. 292 (1981) 327–329), and also the paper [85] are relevant. Finally, let us mention here that Kaimanovich (in J. Soviet Math. 47 (1989) 2387–2398) proved a multiplicative ergodic theorem for semisimple Lie groups, and that Karlsson and Ledrappier (Ann. Prob. 34 (2006) 1693–1706) proved a multiplicative ergodic theorem for the group $ISOM(X)$ of a proper metric space X .

Appendix D

Continuous Singularity of the Limit Distribution of Products of I.I.D. Stochastic Matrices

Let μ be a probability measure on the Borel subsets of 2×2 stochastic matrices (with the usual topology) and have support $S(\mu)$. Let S be the smallest closed multiplicative semigroup generated by $S(\mu)$, and (a, b) denote the stochastic matrix whose first column is (a, b) . Then the sequence μ^n converges weakly iff μ is not the unit mass at $(0, 1)$. See [151].

In what follows, we make the following assumption:

- (*) Each matrix in $S(\mu)$ is invertible, and there is at least one matrix in $S(\mu)$, which has a power with a strictly positive column.

It is clear that under condition (*), μ^n converges weakly, and its weak limit λ has its support consisting of matrices with rank one, since condition (*) guarantees that there is at least one matrix S which has rank one, this matrix being the pointwise limit of the matrix in $S(\mu)$, which has a power with a strictly positive column.

Lemma D.1. *The measure λ is continuous (that is, $\lambda\{x\} = 0$ for each $x \in S(\lambda)$) iff λ is not degenerate (that is, its support is not a singleton) iff there are at least two points in $S(\mu)$ which, as points on the plane are not collinear with the point $(1, 0)$.*

Proof. Suppose λ is not degenerate. If it is not continuous, then

$$b \equiv \sup \{\lambda(\{P\}) : P \in S(\lambda)\}$$

is positive. After noticing that the set $\{P \in S(\lambda) : \lambda(\{P\}) > b/2\}$ is finite, it is clear that the supremum is attained, that is, for some $P \in S(\lambda)$, $\lambda(\{P\}) = b$, where this point P has rank one. For each $y \in S(\mu)$, y being invertible, the set

$$\{P\}y^{-1} \equiv \{Q \in S : Qy = P\}$$

is a singleton. Now, $\lambda * \mu = \lambda$, and for each $y \in S(\mu)$,

$$\lambda(\{P\}) \geq \lambda(\{P\}y^{-1}).$$

Thus,

$$\int [\lambda(\{P\}) - \lambda(\{P\}y^{-1})] \mu(dy) = 0,$$

and therefore, for μ -almost all y ,

$$\lambda(\{P\}y^{-1}) = \lambda(\{P\}) = b > 0.$$

The map $y \rightarrow \lambda(\{P\}y^{-1})$ is upper semicontinuous, and therefore, for all $y \in S(\mu)$,

$$\lambda(\{P\}y^{-1}) = b > 0.$$

Similarly, since $\lambda * \mu^{(n)} = \lambda$, we can also show that for any $y \in S(\mu)$ and any positive integer n , $y^n \in S(\mu^{(n)})$, and therefore,

$$\lambda(\{P\}y^{-n}) = b > 0.$$

Now notice that if $y \notin \{(1, 0), (0, 1)\}$, then $\{P\} = \{P\}y^{-1}$ if and only if y^n converges to P if and only if the points P , y and $(1, 0)$ are collinear. If for each point $y \in S(\mu)$, $\{P\} = \{P\}y^{-1}$, then S will be contained in the line joining the points P and $(1, 0)$, and consequently, $S(\lambda) = \{P\}$, which implies that λ is degenerate. Thus, there is at least one point $y \in S(\mu)$ such that $\{P\}$ is different from $\{P\}y^{-1}$; this means that the points P , y and $(1, 0)$ are not collinear. We also know that for each positive integer $m > 1$, the points y , y^m and $(1, 0)$ are collinear; this means that for any $m > 1$, the points y^m , P and $(1, 0)$ are also not collinear. Therefore, for any two positive integers m and n ,

$$\{P\}y^{-m} \neq \{P\}y^{-n},$$

but this will contradict the fact that λ is a probability measure.

The following theorem, though known to experts, does not seem to be available in standard literature. The proof here follows that given in [164].

Theorem D.1. *The measure λ has exactly one of the following three properties:*

- (i) λ is discrete;
- (ii) λ is continuous singular with respect to m , the Lebesgue measure on $[0, 1]$;
- (iii) λ is absolutely continuous with respect to m .

(Note that the set of rank one stochastic matrices $\{(x, x) : 0 \leq x \leq 1\}$ is here identified with the interval $[0, 1]$).

Proof. Let $E = \{x \in S(\lambda) : \lambda\{x\} > 0\}$, and for $A \subseteq S(\lambda)$, $\lambda_1(A) = \lambda(A \cap E)$, $\lambda_2(A) = \lambda(A \cap E^c)$. Then there exist constants α, β such that $0 \leq \alpha \leq 1$, $0 \leq \beta \leq 1$, $\alpha + \beta = 1$, and probability measures λ_{11} and λ_{12} such that

$$\lambda = \alpha\lambda_{11} + \beta\lambda_{12},$$

where $\lambda_{11} = \frac{1}{\lambda(E)}\lambda_1$ and $\lambda_{12} = \frac{1}{\lambda(E^c)}\lambda_2$. In case $\lambda(E) = 0$, $\lambda = \lambda_{12}$ is continuous, 51
and when $\lambda(E^c) = 0$, $\lambda = \lambda_{11}$ is discrete. Since each y in $S(\mu)$ is assumed 52
invertible, the set 53

$$\{x\} y^{-1} = \{z \in S : zy = x\}$$

is a singleton for each x in S , and consequently, λ_{12} being continuous, $\lambda_{12} * \mu$ is 54
also continuous. Also, since $\lambda = \lambda * \mu$, we have: 55

$$\alpha\lambda_{11} + \beta\lambda_{12} = \alpha\lambda_{11} * \mu + \beta\lambda_{12} * \mu.$$

We can also write: 56

$$\lambda_{11} * \mu = \gamma\lambda_{21} + \delta\lambda_{22},$$

where $0 \leq \gamma \leq 1$, $0 \leq \delta \leq 1$, $\gamma + \delta = 1$, λ_{21} is a discrete probability measure, and 57
 λ_{22} is a continuous probability measure. Then we have: 58

$$\alpha[\lambda_{11} - \gamma\lambda_{21}] = \alpha\delta\lambda_{22} + \beta(\lambda_{12} * \mu - \lambda_{12}).$$

If $\alpha = 0$, then λ is continuous. If $\alpha > 0$, then $\gamma = 1$, $\delta = 0$, $\lambda_{11} = \lambda_{21} = \lambda_{11} * \mu$ 59
and $\lambda_{12} = \lambda_{12} * \mu$. Similarly, it follows that $\lambda_{11} = \mu * \lambda_{11}$ and $\lambda_{12} = \mu * \lambda_{12}$. 60
Thus, if $\alpha > 0$, 61

$$\lambda_{11} = \lambda_{11} * \mu = \mu * \lambda_{11},$$

and this means that 62

$$\lambda_{11} = \lambda_{11} * \mu^n = \mu^n * \lambda_{11}, \quad n \geq 1,$$

so that 63

$$\lambda_{11} = \lambda_{11} * \lambda = \lambda * \lambda_{11}.$$

First, notice that $S(\lambda_{11}) \subset S(\lambda)$, $S(\lambda_{11})$ is an ideal of $S(\lambda)$, and $S(\lambda)$ is simple 64
since $\lambda = \lambda * \lambda$. Thus, $S(\lambda_{11}) = S(\lambda)$. Also, by Proposition 2.4 in Chap. 2, it 65
follows that for $B \subset S(\lambda)$ and $x \in S(\lambda)$, 66

$$\begin{aligned} \lambda_{11}(Bx^{-1}) &= \int \lambda(Bx^{-1}y^{-1})\lambda_{11}(dy) \\ &= \lambda(Bx^{-1}), \end{aligned}$$

and 67

$$\begin{aligned} \lambda(B) &= \lambda * \lambda(B) \\ &= \int \lambda(Bx^{-1})\lambda(dx) = \int \lambda_{11}(Bx^{-1})\lambda(dx) \\ &= \lambda_{11} * \lambda(B) = \lambda_{11}(B), \end{aligned}$$

386 D Continuous Singularity of the Limit Distribution of Products of I.I.D. Stochastic Matrices

so that $\lambda = \lambda_{11}$ and consequently, if $\alpha > 0$ then λ is discrete. Suppose now that λ is continuous. Consider $S(\lambda)$ (topologically) as a subspace of $[0, 1]$. Let m be the Lebesgue measure on $[0, 1]$. By the Lebesgue decomposition theorem, we can write:

$$\lambda = \alpha_3 \lambda_{31} + \beta_3 \lambda_{32},$$

where $0 \leq \alpha_3 \leq 1, 0 \leq \beta_3 \leq 1, \lambda_{31} \perp m$ and $\lambda_{32} \ll m$. Notice that

$$S(\lambda_{32}) \subset S(\lambda) \subset \{x : x \text{ is a stochastic matrix with rank } 1\}.$$

This means that $\mu * \lambda_{32} = \lambda_{32}$ since for $x \in S(\lambda_{32}), \mu(Bx^{-1}) = 1$ iff $x \in B$, and when $x \notin B, Bx^{-1} = \emptyset$. Also, we see that $\lambda_{32} * \mu \ll m$. The reason is the following: Let $m(B) = 0, B \subset [0, 1]$, and $x \in S(\mu)$. Let $\epsilon > 0$. Then there is an open subset $C \subset [0, 1]$ such that $B \subset C$ and $m(C) < \delta |x_1 - x_2|$, where $x = (x_1, x_2)$, and $\delta > 0$ is chosen such that whenever $m(A) < \delta, A \subset [0, 1]$, then $\lambda_{32}(A) < \epsilon$. Notice that

$$\begin{aligned} m(Cx^{-1}) &= m\{y : y(x_1 - x_2) + x_2 \in C\} \\ &< \delta, \end{aligned}$$

and this means that $\lambda_{32}(Cx^{-1}) < \epsilon$ or $\lambda_{32}(Bx^{-1}) < \epsilon$ for $x \in S(\mu)$. In other words, $\lambda_{32} * \mu(B) = 0$. We have seen above that $\lambda_{32} \ll m$, and also, $\lambda_{32} * \mu \ll m, \mu * \lambda_{32} \ll m$. We now show that $\lambda_{31} * \mu \perp m$ and $\mu * \lambda_{31} \perp m$. Since $\lambda_{31} \perp m$, there are disjoint sets E and F such that

$$[0, 1] = E \cup F, \quad \lambda_{31}(E) = 1 = m(F).$$

Since $\lambda = \alpha_3 \lambda_{31} + \beta_3 \lambda_{32}, \lambda(E) = \alpha_3$. Also, $\lambda = \lambda * \mu = \alpha_3 \lambda_{31} * \mu + \beta_3 \lambda_{32} * \mu$. Since $\lambda_{32} * \mu \ll m, \lambda_{32} * \mu(E) = 0$, and therefore, $\alpha_3 = \lambda(E) = \alpha_3 \lambda_{31} * \mu(E)$. Thus, if $\alpha_3 = 0$, then $\lambda \ll m$, and if $\alpha_3 > 0, \lambda_{31} * \mu(E) = 1$ so that $\lambda_{31} * \mu \perp m$. Similarly, since $\lambda = \mu * \lambda$, it also follows that $\mu * \lambda_{31} \perp m$. It follows that

$$\alpha_3 \lambda_{31} + \beta_3 \lambda_{32} = \alpha_3 \lambda_{31} * \mu + \beta_3 \lambda_{32} * \mu$$

or

$$\alpha_3 [\lambda_{31} - \lambda_{31} * \mu] = \beta_3 [\lambda_{32} * \mu - \lambda_{32}]$$

so that if $\alpha_3 > 0$,

$$\lambda_{31} = \lambda_{31} * \mu \quad (= \mu * \lambda_{31}, \text{ similarly}).$$

In other words, for $n \geq 1$,

$$\lambda_{31} = \lambda_{31} * \mu^n = \mu^n * \lambda_{31}$$

or

$$\lambda_{31} = \lambda_{31} * \lambda = \lambda * \lambda_{31}.$$

As before, like in the equation $\lambda_{11} = \lambda_{11} * \lambda = \lambda * \lambda_{11}$ presented earlier, it follows again that when $\alpha_3 > 0$, $\lambda = \lambda_{31}$, and consequently, $\lambda \perp m$; otherwise, $\lambda \ll m$. This completes the proof of the theorem.

We remark that if μ is a probability measure with its support consisting of the two stochastic matrices $\{(a, b), (c, d)\}$, $a > b, c > d$, $\frac{b}{1-a} < \frac{d}{1-c}$ such that $(a - b) + (c - d) < 1$, then $S(\lambda)$ is a Cantor-type subset of $\{(x, x) : \frac{b}{1-a+b} \leq x \leq \frac{d}{1-c+d}\}$. In particular if $S(\mu) = \{(\frac{1}{3}, 0), (1, \frac{2}{3})\}$, then $S(\lambda)$ is the Cantor subset of the diagonal $\{(x, x) : 0 \leq x \leq 1\}$. If $S(\mu) = \{(a, b), (c, d)\}$, $a > b, c > d$, $\frac{b}{1-a} < \frac{d}{1-c}$ and $(a - b) + (c - d) = 1$, then λ is the uniform distribution on the interval $\{(x, x) : \frac{b}{1-a+b} \leq x \leq \frac{d}{1-c+d}\}$ iff $\mu\{(a, b)\} = a - b$. For these and other related results, we refer the reader to [176].

The main aim of this appendix is to prove the following general theorem, which first appeared in Mukherjea and Restrepo [169].

Theorem D.2. *Let μ be a probability measure on 2×2 stochastic matrices with a finite or countably infinite support such that*

$$S(\mu) = \{(x_i, y_i) : 1 \leq i < \infty, x_i \neq y_i, 0 < x_i < 1, 0 < y_i < 1\},$$

where

$$\mu\{(x_i, y_i)\} = p_i, \quad 0 < p_i < 1 \quad \text{and} \quad \sum_{i=1}^{\infty} p_i = 1.$$

Let $\pi_i = |x_i - y_i|$ for each i such that one of the following three conditions hold:

- (i) $\sum_{i=1}^{\infty} \pi_i < 1$;
- (ii) $\sum_{i=1}^{\infty} \pi_i = 1$, and for at least one i , $\pi_i \neq p_i$;
- (iii) $\prod_{i=1}^{\infty} \left(\frac{\pi_i}{p_i}\right)^{p_i} < 1$.

We also assume:

$$(iv) \sum_{i=1}^{\infty} p_i \log p_i > -\infty.$$

Then $\lambda = (w) \lim_{n \rightarrow \infty} \mu^n$ is continuous singular with respect to the Lebesgue measure m on $[0, 1]$, where $[0, 1]$ is identified with the set of all 2×2 stochastic matrices with rank 1.

We prove this theorem in four steps.

115

Step 1:

116

Let (S, d) be a compact metric space and let $\mathfrak{K}(S)$ be the class of all closed subsets of S . For A, B in $\mathfrak{K}(S)$ define $D(A, B) \equiv \inf\{s > 0 : A \subseteq N_s(B) \text{ and } B \subseteq N_s(A)\}$ where $N_s(B) \equiv \{y \in S : \text{for some } x \in B, d(x, y) < s\}$. Then $(\mathfrak{K}(S), D)$ is a complete metric space.

117

118

119

120

Let $\{f_i : S \rightarrow S\}_{i \in \Lambda}$, with Λ a countably infinite set, be a family of functions such that $d(f_i(x), f_i(y)) \leq \pi_i d(x, y)$ for all $x, y \in S$ and some $\pi_i > 0$ and let $\pi = \sup\{\pi_i : i \in \Lambda\}$. Then, as is well known for finite Λ , here also if $\pi < 1$ then there is a unique compact subset $A \subseteq S$ such that

121

122

123

124

$$A = \overline{\bigcup_{i \in \Lambda} f_i(A)}. \quad (\text{D.1})$$

The idea of the proof is the same as is wellknown in the case when Λ is finite:

125

Define for each $A \in \mathfrak{K}(S)$ the function $F : \mathfrak{K}(S) \rightarrow \mathfrak{K}(S)$ such that $F(A) = \overline{\bigcup_{i \in \Lambda} f_i(A)}$. It is then easy to verify that for $A, B \in \mathfrak{K}(S)$,

126

127

$$D(F(A), F(B)) \leq \pi D(A, B) \quad (\text{D.2})$$

and therefore (D.1) follows from the classical contraction mapping theorem and the completeness of the space $(\mathfrak{K}(S), D)$.

128

129

Along the same lines, let $\{p_i\}_{i \in \Lambda}$ be a discrete probability measure on Λ and let I be a Λ -valued random variable distributed according to $\{p_i\}_{i \in \Lambda}$. Assume that $\mathbb{E}[\pi_I] < 1$. Then there is also a unique invariant probability measure λ on the Borel subsets of S such that

130

131

132

133

$$\lambda(A) = \sum_{i \in \Lambda} p_i \lambda(f_i^{-1}(A)) \quad (\text{D.3})$$

for all Borel subsets $A \subseteq S$. Briefly, the idea of the proof is the following: Since S is compact, there exists $\beta > 0$ such that for any $x, y \in S$, $d(x, y) < \beta$. Define the set $\mathfrak{F}_\beta$ by

134

135

136

$$\mathfrak{F}_\beta \equiv \{h : S \rightarrow \mathbb{R} : |h(x) - h(y)| \leq d(x, y) \text{ and } |h(x)| \leq \beta \text{ for all } x, y \in S\}$$

and define in $\mathbf{P}(S)$ (the space of all the probability measures defined on the Borel subsets of S) the metric

137

138

$$\rho(\mu_1, \mu_2) \equiv \sup \left\{ \left| \int h d\mu_1 - \int h d\mu_2 \right| : h \in \mathfrak{F}_\beta \right\}.$$

Then $(\mathbf{P}(S), \rho)$ is a complete metric space. To establish (D.3), we define $\Phi : \mathbf{P}(S) \rightarrow \mathbf{P}(S)$ such that for any Borel set $A \subseteq S$,

139

140

$$\Phi(\mu)(A) = \sum_{i \in \Lambda} p_i \mu(f_i^{-1}(A)). \quad (\text{D.4})$$

Now, taking a fixed point $z \in S$, it follows that for $h \in \mathfrak{F}_\beta$, $\mu_1, \mu_2 \in \mathbf{P}(S)$, 141

$$\begin{aligned} & \left| \int h d\Phi(\mu_1) - \int h d\Phi(\mu_2) \right| \\ &= \left| \sum_{i \in \Lambda} p_i \left(\int [h \circ f_i(x) - h \circ f_i(z)] d\mu_1(x) - \int [h \circ f_i(x) - h \circ f_i(z)] d\mu_2(x) \right) \right| \\ &\leq \left| \sum_{i \in \Lambda} p_i \pi_i \left[\int h_i d\mu_1 - \int h_i d\mu_2 \right] \right|, \text{ where } h_i(x) = \frac{1}{\pi_i} [h \circ f_i(x) - h \circ f_i(z)] \\ &\leq \mathbb{E}[\pi_I] \rho(\mu_1, \mu_2), \text{ since } h_i \in \mathfrak{F}_\beta. \end{aligned}$$

Now (D.3) follows by applying the contraction mapping theorem on Φ . Notice that, 142
in particular, $\mathbb{E}[\pi_I] < 1$ if π defined earlier (just preceding (D.1)) is less than 1. 143

Step 2: The model map. 144

To proceed, we need a “model” map like the one presented in the book by Edgar 145
[64]. Let S , Λ , $\{f_i, \pi_i : i \in \Lambda\}$ and π be defined as before. Consider the product 146
space $\Lambda^{(\omega)} = \prod_{i \in \Lambda} \Lambda_i$, where $\Lambda_i \equiv \Lambda$ for each i . The elements σ in $\Lambda^{(\omega)}$ can 147

be considered as a point $(e_1, e_2, e_3, \dots)$, $e_i \in \Lambda_i$ or as an infinite word $e_1 e_2 e_3 \dots$ 148
where the letters are taken from the alphabet Λ . If $\sigma = e_1 e_2 \dots e_k \sigma_k$ where $e_1, e_2,$ 149
 $\dots, e_k$ are the first k letters of σ and $\sigma_k \in \Lambda^{(\omega)}$, then we denote the finite word 150
 $e_1 e_2 \dots e_k$ by simply $\sigma_{\upharpoonright k}$, and by $[\sigma_{\upharpoonright k}]$ we denote the set $\{\tau \in \Lambda^{(\omega)} : \tau_{\upharpoonright k} = \sigma_{\upharpoonright k}\}$, a 151
natural topology over $\Lambda^{(\omega)}$ is the one generated by the sets of the form $[\sigma_{\upharpoonright k}]$, we will 152
denote this topology by $\mathcal{T}$. Also, in $\Lambda^{(\omega)}$ we are considering the distance function 153

$$\rho(\sigma, \tau) \equiv \prod_{i=1}^k \pi_{e_i}, \text{ where } k \text{ is the largest number such that } \tau_{\upharpoonright k} = \sigma_{\upharpoonright k}. \quad 154$$

Given a probability distribution $\{p_i\}_{i \in \Lambda}$ for Λ , it induces the product probability 155
measure $\mathbb{P}^{(\infty)}$ on $\Lambda^{(\omega)}$. In particular, it follows that for any Borel subset $B \subseteq \Lambda^{(\omega)}$, 156
we have that 157

$$\mathbb{P}^{(\infty)}(B) = \sum_{i \in \Lambda} p_i \mathbb{P}^{(\infty)} \left\{ \sigma \in \Lambda^{(\omega)} \mid i \sigma \in B \right\} \quad (\text{D.5})$$

We will need (D.5) after we construct a model map $h : \Lambda^{(\omega)} \rightarrow S$, fashioned after 158
[64], page 211, in Proposition D.1. 159

In what follows, let I be a Λ -valued random variable distributed according to 160
 $\{p_i\}_{i \in \Lambda}$. 161

Remark D.1. Notice that if $\pi \equiv \sup_{i \in \Lambda} \pi_i < 1$, then $\rho(\sigma, \tau)$ is a metric over $\Lambda^{(\infty)}$ 162
that generates the topology T and that agrees with $P^{(\infty)}$, meaning that all Borel 163
sets of $\Lambda^{(\omega)}$ under ρ are measurable. Under the less restrictive condition $\pi \leq 1$, we 164
should include also the assumption $-\infty \leq E[\log \pi_I] < 0$, that is trivially attained 165
in the cases $\pi < 1$ or $E[\pi_I] < 1$. Under this condition, we have that for $P^{(\infty)}$ - 166

almost all $\sigma = e_1 e_2 \dots \in \Lambda^{(\omega)}$, $\prod_{i=1}^k \pi_{e_i} \downarrow 0$ as $k \rightarrow \infty$ i.e., $\rho(\sigma, \sigma) = 0$. Redefining 167

ρ over those σ that do not satisfy $\rho(\sigma, \sigma) = 0$, we obtain a metric on $\Lambda^{(\omega)}$ such that 168

for $\mathbb{P}^{(\infty)}$ -almost all $\sigma \in \Lambda^{(\omega)}$, $\rho(\sigma, \tau) = \prod_{i=1}^k \pi_{e_i}$ for all $\tau \in \Lambda^{(\omega)}$ that coincide with σ in exactly the first k letters. 169 170

Proposition D.1. *Let S be a compact metric space and for each $i \in \Lambda$, a countably infinite set, let f_i be a map from S to S such that $d(f_i(x), f_i(y)) \leq \pi_i d(x, y)$ for all $x, y \in S$, where $\pi_i > 0$ and $-\infty \leq \mathbb{E}[\log \pi_I] < 0$. Consider the distance function ρ on $\Lambda^{(\omega)}$ and the product probability measure $\mathbb{P}^{(\infty)}$ induced by $\{\pi_i\}_{i \in \Lambda}$. Then, there exists a unique measurable function $h : \Lambda^{(\omega)} \rightarrow S$ such that $h(i\sigma) = f_i(h(\sigma))$ for each $\sigma \in \Lambda^{(\omega)}$ and each $i \in \Lambda$. This h also satisfies the following conditions:* 171 172 173 174 175 176 177

1. The measure $\lambda \in \mathbf{P}(S)$ such that for each Borel set $A \subseteq S$, $\lambda(A) = \mathbb{P}^{(\infty)}(h^{-1}(A))$ is the unique invariant probability measure given in (D.3). 178 179
2. $\overline{h(\Lambda^{(\omega)})} = \bigcup_{i \in \Lambda} f_i(\overline{h(\Lambda^{(\omega)})})$. 180
3. For $\mathbb{P}^{(\infty)}$ -almost all $\sigma, \tau \in \Lambda^{(\omega)}$, $d(h(\sigma), h(\tau)) \leq \beta \rho(\sigma, \tau)$ where $\beta = \sup\{d(x, y) : x \in S, y \in S\}$. 181 182
4. If $\pi = \sup\{\pi_i : i \in \Lambda\} < 1$ then h is continuous and $\overline{h(\Lambda^{(\omega)})}$ coincides with the unique invariant set A given in (D.1). 183 184

The proof follows the same lines as in [64]. For completeness, we include it. 185

Proof. Fix an arbitrary $a \in S$. We define recursively a sequence of continuous functions g_k , $1 \leq k < \infty$, $g_k : \Lambda^{(\omega)} \rightarrow S$ such that $g_0(\sigma) = a$ for each $\sigma \in \Lambda^{(\omega)}$ and $g_{k+1}(i\sigma) = f_i(g_k(\sigma))$ for each $\sigma \in \Lambda^{(\omega)}$, $i \in \Lambda$ and $k \geq 0$. Notice that g_0 is continuous on $\Lambda^{(\omega)}$. Now, let us assume that for $k \geq 0$, g_k is also continuous in $(\Lambda^{(\omega)}, \mathcal{T})$. Let $i\sigma \in \Lambda^{(\omega)}$. Given $\epsilon > 0$, let l be such that if $\tau \in [\sigma]_l$ then $d(g_k(\sigma), g_k(\tau)) < \frac{\epsilon}{\pi_i}$. Now, for any $i\tau \in [(i\sigma)]_{k+1}$, 186 187 188 189 190 191

$$d(g_{k+1}(i\sigma), g_{k+1}(i\tau)) = d(f_i(g_k(\sigma)), f_i(g_k(\tau))) \leq \pi_i d(g_k(\sigma), g_k(\tau)) < \epsilon,$$

proving that g_{k+1} is continuous in $(\Lambda^{(\omega)}, \mathcal{T})$. Similarly, it is possible to show that if $\beta \equiv \sup\{d(x, y) : x \in S, y \in S\}$, then for any $\sigma = e_1 e_2 \dots \in \Lambda^{(\omega)}$ and $k \geq 1$, $d(g_{k+m}(\sigma), g_k(\sigma)) \leq \beta \prod_{i=1}^k \pi_{e_i}$, but the condition $\mathbb{E}[\log \pi_I] < 0$ implies (by the strong law of large numbers) that for $\mathbb{P}^{(\infty)}$ -almost all $\sigma \in \Lambda^{(\omega)}$, 192 193 194 195

$$\limsup_{k \rightarrow \infty} \prod_{i=1}^k \pi_{e_i} = \limsup_{k \rightarrow \infty} \exp \left\{ \sum_{i=1}^k \log \pi_{e_i} \right\} = 0, \quad (\text{D.6})$$

proving that the sequence $\{g_k(\sigma)\}_{k \geq 1}$ is Cauchy for $\mathbb{P}^{(\infty)}$ -almost all $\sigma \in \Lambda^{(\omega)}$. Therefore, there exists a measurable function $h(\sigma)$ such that for $\mathbb{P}^{(\infty)}$ -almost all 196 197

$\sigma \in \Lambda^{(\omega)}$, $\lim_{k \rightarrow \infty} g_k(\sigma) = h(\sigma)$. Notice that if $\pi \equiv \sup\{\pi_i \mid i \in \Lambda\} < 1$, then 198
 $d(g_{k+m}(\sigma), g_k(\sigma)) \leq \beta\pi^k$ implying that the convergence is uniform and therefore 199
 that $h(\sigma)$ is continuous. Now, since $g_{k+1}(i\sigma) = f_i(g_k(\sigma))$ for each $\sigma \in \Lambda^{(\omega)}$, 200
 $i \in \Lambda$ and $k \geq 1$, it follows that for $\mathbb{P}^{(\infty)}$ -almost all $\sigma \in \Lambda^{(\omega)}$, $h(i\sigma) = f_i(h(\sigma))$. 201
 Also, from (D.5) it follows that for any Borel subset $A \subseteq S$, we have: 202

$$\begin{aligned} \mathbb{P}^{(\infty)}(h^{-1}(A)) &= \sum_{i \in \Lambda} p_i \mathbb{P}^{(\infty)} \left\{ \sigma \in \Lambda^{(\omega)} \mid i\sigma \in h^{-1}(A) \right\} \\ &= \sum_{i \in \Lambda} p_i \mathbb{P}^{(\infty)} \left\{ \sigma \in \Lambda^{(\omega)} \mid f_i(h(\sigma)) \in A \right\} \\ &= \sum_{i \in \Lambda} p_i (\mathbb{P}^{(\infty)} \circ h^{-1})(f^{-1}(A)), \end{aligned}$$

establishing property (1) in the proposition. It is now clear that $\overline{h(\Lambda^{(\omega)})} =$ 203
 $\bigcup_{i \in \Lambda} f_i \left(\overline{h(\Lambda^{(\omega)})} \right)$, and in the case that $\pi < 1$, $\overline{h(\Lambda^{(\omega)})}$ coincides with the unique 204
 invariant set given in (D.1). Finally, to prove property (3), let k be the greatest 205
 integer such that $\tau_{\uparrow k} = \sigma_{\uparrow k}$. If $\tau = e_1 e_2 \dots e_k \tau_k$ and $\sigma = e_1 e_2 \dots e_k \sigma_k$, then 206

$$\begin{aligned} d(h(\tau), h(\sigma)) &= d(f_{e_1} \circ f_{e_2} \circ \dots \circ f_{e_k} \circ h(\tau_k), f_{e_1} \circ f_{e_2} \circ \dots \circ f_{e_k} \circ h(\sigma_k)) \\ &\leq \prod_{i=1}^k \pi_{e_i} d(h(\tau_k), h(\sigma_k)) \\ &\leq \beta \rho(\tau, \sigma). \end{aligned}$$

Step 3: A bound for the upper packing dimension. 207

Let λ be a Borel probability measure on a metric space S . Then the essential 208
 supremum (with respect to λ) of $\overline{L}_\lambda(x)$, $x \in S$ defined by 209

$$\overline{L}_\lambda(x) \equiv \limsup_{\delta \rightarrow 0} \frac{\log \lambda \left(\overline{B_\delta(x)} \right)}{\log \delta},$$

where $B_\delta(x)$ is the open ball with center x and radius δ , is called the upper packing 210
 dimension of the measure λ . When $S = \mathbb{R}^d$ and λ -ess sup $\overline{L}_\lambda(x) < d$, then λ is 211
 singular with respect to the Lebesgue measure m on $\mathbb{R}^d$. To see this suppose that 212
 $\lambda \ll m$ where m is the Lebesgue measure in $\mathbb{R}^d$ and let $f(x) = \frac{d\lambda}{dm}$. Define 213

$$\alpha_r(x) \equiv \frac{\int_{B_r(x)} f(y) dy}{m(B_r(x))}.$$

Notice that for almost all $x \in \mathbb{R}^d$ (and therefore for λ -almost all $x \in \mathbb{R}^d$), $f(x) < +\infty$ and $\lim_{r \rightarrow 0} \alpha_r(x) = f(x)$. Now, let $S \equiv \{x \in \mathbb{R}^d : f(x) > 0\}$. For λ -almost all $x \in S$,

$$\begin{aligned} \bar{L}_\lambda(x) &= \limsup_{r \rightarrow 0} \frac{\log \left(\frac{\int_{B_r(x)} f(y) dy}{B_r(x)} \right)}{\log r} \\ &= \limsup_{r \rightarrow 0} \frac{\log \left(\int_{B_r(x)} f(y) dy \right)}{\log r} \\ &= \limsup_{r \rightarrow 0} \left(\frac{\log \alpha_r(x)}{\log r} + \frac{\log m(B_r(x))}{\log r} \right) \\ &= d. \end{aligned}$$

But $\lambda(S) = \int_S f(x) dx = \int_{\mathbb{R}^d} f(x) dx = \lambda(\mathbb{R}^d)$, so that for λ -almost all $x \in \mathbb{R}^d$,

$\bar{L}_\lambda(x) = d$ and therefore, λ -ess sup $\bar{L}_\lambda(x) = d$, which is a contradiction.

Now, let $S, \Lambda, \{f_i : i \in \Lambda\}, \{p_i : i \in \Lambda\}, \{\pi_i : i \in \Lambda\}$ be as in Proposition D.1. The following proposition holds.

Proposition D.2. *Let λ be the unique invariant probability measure satisfying (see (D.3)) such that $\lambda(A) = \sum_{i \in \Lambda} p_i \lambda(f_i^{-1}(A))$ for all Borel subsets $A \subseteq S$, and let I be a Λ -valued random variable distributed according to $\{p_i : i \in \Lambda\}$. Suppose that*

1. $-\infty \leq \mathbb{E}[\log \pi_I] < 0$.
2. $\mathbb{E}[|\log p_I|] < \infty$.

Then,

$$\lambda - \text{ess sup } \bar{L}_\lambda(x) \leq \frac{\mathbb{E}[\log p_I]}{\mathbb{E}[\log \pi_I]} = \frac{\sum_{i \in \Lambda} p_i \log p_i}{\sum_{i \in \Lambda} p_i \log \pi_i}. \quad (\text{D.7})$$

Proof. Consider the product space $\Lambda^{(\omega)}$ together with the product measure $\mathbb{P}^{(\infty)}$ induced by $\{p_i : i \in \Lambda\}$. Using the strong law of large numbers we have that for $\mathbb{P}^{(\infty)}$ -almost all $\sigma \in \Lambda^{(\omega)}$, $\sigma = e_1 e_2 e_3 \dots$,

$$\lim_{k \rightarrow \infty} \frac{\sum_{i=1}^k \log p_{e_i}}{k} = \mathbb{E}[\log p_I] \quad (\text{D.8})$$

and

$$\lim_{k \rightarrow \infty} \frac{\sum_{i=1}^k \log \pi_{e_i}}{k} = \mathbb{E}[\log \pi_I]. \quad (\text{D.9})$$

D Continuous Singularity of the Limit Distribution of Products of I.I.D. Stochastic Matrices 393

Also, we have from (D.6) that the condition $\mathbb{E}[\log \pi_I] < 0$ implies that for $\mathbb{P}^{(\infty)}$ - 231
almost all $\sigma \in \Lambda^{(\omega)}$, $\sigma = e_1 e_2 e_3 \dots$, 232

$$\lim_{k \rightarrow \infty} \prod_{i=1}^k \pi_{e_i} = 0. \quad (\text{D.10})$$

Let $F \subseteq \Lambda^{(\omega)}$ with $\mathbb{P}^{(\infty)}(F) = 1$ be such that (D.8), (D.9), (D.10) and the property 233
(3) of Proposition D.1 hold, for all σ in F . Fix $x = h(\sigma)$ where $\sigma \in F$, $\sigma = 234$
 $e_1 e_2 e_3 \dots$. For each $\epsilon > 0$, define 235

$$k_\epsilon \equiv \max \left\{ k : \prod_{i=1}^k \pi_{e_i} < \frac{\epsilon}{\beta} \leq \prod_{i=1}^{k-1} \pi_{e_i} \right\}.$$

If $\sigma' \in [\sigma]_{k_\epsilon} \cap F$, then we have by the property (3) of Proposition D.1, that 236

$$d(h(\sigma'), x) \leq \beta \rho(\sigma', \sigma) \leq \beta \prod_{i=1}^{k_\epsilon} \pi_{e_i} < \epsilon$$

and therefore, $h([\sigma]_{k_\epsilon} \cap F) \subseteq B_\epsilon(x)$, so that 237

$$\begin{aligned} \log \lambda(\overline{B_\epsilon(x)}) &\geq \log \lambda(h([\sigma]_{k_\epsilon} \cap F)) \\ &= \log \mathbb{P}^{(\infty)}(h^{-1}(h([\sigma]_{k_\epsilon} \cap F))) \\ &\geq \log \mathbb{P}^{(\infty)}([\sigma]_{k_\epsilon} \cap F) \\ &= \sum_{i=1}^{k_\epsilon} \log p_{e_i}. \end{aligned}$$

Now, we have: 238

$$\begin{aligned} \limsup_{\epsilon \rightarrow 0} \frac{\log \lambda(\overline{B_\epsilon(x)})}{\log \epsilon} &\leq \lim_{\epsilon \rightarrow 0} \frac{\frac{1}{k_\epsilon} \sum_{i=1}^{k_\epsilon} \log p_{e_i}}{\frac{1}{k_\epsilon - 1} \sum_{i=1}^{k_\epsilon - 1} \log \pi_{e_i}} \left(\frac{k_\epsilon}{k_\epsilon - 1} \right) \\ &= \lim_{k \rightarrow \infty} \frac{\frac{1}{k} \sum_{i=1}^k \log p_{e_i}}{\frac{1}{k-1} \sum_{i=1}^{k-1} \log \pi_{e_i}} \left(\frac{k}{k-1} \right) \\ &= \frac{\mathbb{E}[\log p_I]}{\mathbb{E}[\log \pi_I]}. \end{aligned}$$

This means that for all $x \in h(F)$, 239

$$\overline{L}_\lambda(x) \leq \frac{\sum_{i \in \Lambda} p_i \log p_i}{\sum_{i \in \Lambda} p_i \log \pi_i},$$

but $\lambda(h(F)) = \mathbb{P}^{(\infty)}(h^{-1}(h(F))) \geq \mathbb{P}^{(\infty)}(F) = 1$, which implies that 240

$$\lambda\text{-ess sup } \bar{L}_\lambda(x) \leq \frac{\sum_{i \in \Lambda} p_i \log p_i}{\sum_{i \in \Lambda} p_i \log \pi_i}.$$

Final step: Continuous singularity of λ .

First, notice that $\lambda = \lim_{n \rightarrow \infty} \mu^{(n)}$ satisfies the convolution equation $\lambda = \lambda * \mu$ and is the self-similar probability measure satisfying

$$\lambda = \sum_{i=1}^{\infty} p_i \lambda \circ f_i^{-1}, \quad (\text{D.11})$$

where $f_i(x) = y_i + (x_i - y_i)x$, $x \in [0, 1]$.

In the present context, $f_i(h(\sigma)) = h(\sigma) \cdot (x_i, y_i)$, where $h(\sigma)$ represents a stochastic matrix where the first element in each row is $h(\sigma)$, and (x_i, y_i) represents a stochastic matrix where the first elements in the two rows are, respectively, x_i and y_i and the multiplication is matrix multiplication. With this observation, it is easy to verify that the last equation in step 2 involving $\mathbb{P}^{(\infty)} \circ h^{-1}$ can be written as

$$(\mathbb{P}^{(\infty)} \circ h^{-1}) = (\mathbb{P}^{(\infty)} \circ h^{-1}) * \mu,$$

implying that the invariant measure λ , being the unique solution of the convolution equation $\lambda = \lambda * \mu$, must be the measure $(\mathbb{P}^{(\infty)} \circ h^{-1})$.

Note that as is known, λ can be considered as a probability measure on the closed unit interval $[0, 1]$, and its support K will then satisfy (as was shown earlier) the equality

$$K = \overline{\bigcup_{i=1}^{\infty} f_i(K)}. \quad (\text{D.12})$$

Suppose now that m is the Lebesgue measure on $[0, 1]$. Let us assume one of the three conditions (i), (ii), and (iii), and also, condition (iv). If one of the conditions (i) and (ii) holds, then Jensen's inequality implies that

$$\begin{aligned} \sum_{i=1}^{\infty} p_i \log \frac{\pi_i}{p_i} &\leq \log \left(\sum_{i=1}^{\infty} \pi_i \right) \\ &\leq 0, \end{aligned}$$

with equality if and only if $\sum_{i=1}^{\infty} \pi_i = 1$ and $\pi_i = p_i$ for all $i \geq 1$. Therefore,

$$-\sum_{i=1}^{\infty} p_i \log p_i < -\sum_{i=1}^{\infty} p_i \log \pi_i.$$

Thus, in case (i), (ii), or (iii) holds, we have:

$$\frac{\sum_{i=1}^{\infty} p_i \log p_i}{\sum_{i=1}^{\infty} p_i \log \pi_i} < 1$$

D Continuous Singularity of the Limit Distribution of Products of I.I.D. Stochastic Matrices 395

and therefore, it follows from Proposition D.2 that on $[0, 1]$, 260

$$\lambda - \text{ess sup } \overline{L}_\lambda(x) < 1. \quad (\text{D.13})$$

It follows that λ is then singular with respect to m . Let us point out that the 261
corresponding result when $S(\mu)$ is finite appeared earlier in [164]. 262

Uncorrected Proof

Appendix E

Rate of Decay of Concentration Functions on Discrete Groups, by T. M. Retzlaff

E.1 Preliminaries

Given a probability measure μ on a locally compact group G , one defines the *concentration functions* to be

$$f_n(K) = \sup_{g \in G} \mu^n(Kg),$$

where K is a compact subset of G and μ^n is the n th convolution power of μ . Jaworski [118] completed the work in Jaworski et al. [120], which had originally started in Hofmann and Mukherjea [95], to find necessary and sufficient conditions for the concentration functions to converge to zero. These conditions depend only on the support of μ , which will be denoted by S throughout. One says that μ is *adapted* if the smallest closed subgroup containing S is G . A measure is called *irreducible* if the smallest closed subsemigroup containing S is G . For our purposes, the smallest closed subgroup containing a set A will be denoted by $gp(A)$ and the smallest closed subsemigroup will be called *semigp*(A). We will analogously call a set or a probability density function adapted or irreducible as well.

We say that a unimodular group G , with Haar measure λ , has volume growth rate which is at least polynomial with degree D (abbreviated $V_G(m) \geq cm^D$) if there exists a compact symmetric neighborhood of the identity, call it K , and a constant c such that

$$\lambda(K^m) \geq cm^D,$$

for all $m \geq 0$. Note that λ can be taken to be counting measure when H is discrete.

When our measure is irreducible, we will show that the volume growth rate of the group determines the rate of decay of the concentration functions. In particular, we will prove:

Theorem E.1 ([197]). *Let G be a discrete group. Let F be a nonnegative function of G such that $\int F = 1$ and $S = \text{supp}(F)$ generates G as a semigroup. Then, if $V_G(m) \geq cm^D$, we have*

$$\|F^{(k)}\|_\infty = O(k^{-D/2}),$$

where $F^{(k)}$ is the k th convolution power of F .

Let N_μ be the smallest closed normal subgroup a coset of which contains S . If the measure we are interested in is merely adapted (i.e., its support generates G as a subgroup, but not as a subsemigroup), then it is not the volume growth rate of G which will determine the rate of decay of the concentration functions, but rather the volume growth rate of N_μ . In particular:

Theorem E.2 ([198]). *Let G be an infinite discrete group, and let F be an adapted probability density function with associated probability measure μ . If N_μ satisfies $V_{N_\mu}(m) \geq cm^D$, then $\|F^{(k)}\|_\infty = O(k^{-D/2})$. Thus, the concentration functions associated with μ also decay as $O(k^{-D/2})$.*

Thanks to Derrienic and Lin [60] we have a convenient generating set for N_μ .

Theorem E.3 ([60]). $N_\mu = gp(\bigcup (S^n S^{-n} \cup S^{-n} S^n))$.

Proof. Let $S \subset gN_\mu$. Then, because N_μ is normal, we have $S^n \subset g^n N_\mu$. Therefore, $S^n S^{-n} \subset N_\mu$. Likewise, we have $S^{-n} S^n \subset N_\mu$ since $S^{-1} \subset N_\mu g^{-1} = g^{-1} N_\mu$ and then a similar argument applies. Therefore, $gp(\bigcup (S^n S^{-n} \cup S^{-n} S^n)) \subset N_\mu$.

It is easy to see that N_μ is the closed normal subgroup generated by $S^{-n} S^n$ or $S^n S^{-n}$. So we need to show that $gp(\bigcup (S^n S^{-n} \cup S^{-n} S^n))$ is a normal subgroup. Let $s \in S$. Then

$$s S^{-n} S^n s^{-1} = s S^{-1} S^{-(n-1)} S^{n-1} S s^{-1} \subset gp\left(\bigcup (S^n S^{-n} \cup S^{-n} S^n)\right).$$

Also, it is clear that

$$s S^n S^{-n} s^{-1} \subset S^{n+1} S^{-(n+1)} \subset gp\left(\bigcup (S^n S^{-n} \cup S^{-n} S^n)\right).$$

Therefore,

$$s \left(gp\left(\bigcup (S^n S^{-n} \cup S^{-n} S^n)\right) \right) s^{-1} \subset gp\left(\bigcup (S^n S^{-n} \cup S^{-n} S^n)\right).$$

A similar argument shows that

$$s^{-1} \left(gp\left(\bigcup (S^n S^{-n} \cup S^{-n} S^n)\right) \right) s \subset gp\left(\bigcup (S^n S^{-n} \cup S^{-n} S^n)\right).$$

Since S is adapted, this shows that

$$g \left(gp\left(\bigcup (S^n S^{-n} \cup S^{-n} S^n)\right) \right) g^{-1} \subset gp\left(\bigcup (S^n S^{-n} \cup S^{-n} S^n)\right)$$

for all $g \in G$. Thus $gp(\bigcup (S^n S^{-n} \cup S^{-n} S^n))$ is a normal closed subgroup containing $S^{-1} S$, so $N_\mu \subset gp(\bigcup (S^n S^{-n} \cup S^{-n} S^n))$. $\square$

When G is a discrete group and μ is adapted, Jaworski's result reduces to the fact that the concentration functions converge to zero if and only if N_μ is infinite. This

was actually proven previously by Bartoszek [7]. Bartoszek's proof relies on the following Lemmas:

Lemma E.1 ([60]). *Let G be a discrete group. For any finite set $A \subset G$ and a probability measure μ on G , the following are equivalent*

1. $\lim_{n \rightarrow \infty} \|T^n 1_A\|_\infty = 0$
2. $\lim_{n \rightarrow \infty} \|T^n 1_A\|_2 = 0$
3. $\lim_{n \rightarrow \infty} f_n(A) = 0$ (i.e., the concentration functions go to zero).

Proof. The fact that (1) is equivalent to (3) is obvious. Likewise, the fact that (2) implies (1) is also obvious since G is a discrete group. The fact that (1) implies (2) follows from Holder's Inequality:

$$\|T^n f\|_2^2 \leq \|T^n f\|_\infty \|T^n f\|_1 \leq \|T^n f\|_\infty \|f\|_1.$$

□

Lemma E.2 (See [7, 220] p. 40). *Let μ be a probability measure on a discrete group G . Let T be the operator of right convolution by μ and $\check{T}$ its adjoint. Then there exists a measure ρ on G such that*

$$\lim_{n \rightarrow \infty} T^n \check{T}^n = T_\rho,$$

in the strong operator topology on $L^2(G)$ and where T_ρ is the operator of right convolution by ρ . The measure ρ satisfies $\check{\mu} * \rho * \mu = \rho$.

Proof. Let T be the operator of right convolution by μ and $\check{T}$ be its adjoint. The operator T is contracting on $L^2(G)$. Thus,

$$\|f\|_2 \geq \|Tf\|_2 \geq \dots \geq \|T^n f\|_2 \geq \dots \geq 0.$$

It follows that

$$I \geq T\check{T} \geq \dots \geq T^n \check{T}^n \geq \dots \geq 0.$$

Therefore, $S = \lim_{n \rightarrow \infty} T^n \check{T}^n$ exists in the strong operator topology. Furthermore $S = \lim_{n \rightarrow \infty} T^{n+1} \check{T}^{-(n+1)} = TS\check{T}$.

Bartoszek's result now is:

Theorem E.4 ([7]). *Let μ be an adapted probability measure on a discrete group G . Let T be the operator of right convolution by μ and $\check{T}$ its adjoint. Then the following conditions are equivalent:*

1. The concentration functions do not converge to zero (i.e., $f_n(K) \not\rightarrow 0$ for all compact sets K).
2. There exists a function $f \in L^2(G)$ such that $\lim_{n \rightarrow \infty} \|T^n f\|_2 > 0$.
3. There exists a probability measure $\tilde{\rho}$ on G such that $\check{\mu} * \tilde{\rho} * \mu = \tilde{\rho}$.

$$4. \lim_{n \rightarrow \infty} |S^n| = d_\mu < \infty. \quad 84$$

$$5. N_\mu \text{ is finite.} \quad 85$$

Proof. (1) $\Rightarrow$ (2) follows from Lemma E.1. 86

(2) $\Rightarrow$ (3). If there exists a function $f \in L^2(G)$ such that 87

$$\lim_{n \rightarrow \infty} \|T^n f\|_2^2 = \lim_{n \rightarrow \infty} (T^n \check{T}^n f, f) > 0,$$

then the measure ρ in Lemma E.2 is not zero. Then the measure $\tilde{\rho} = \rho/\rho(G)$ 88
satisfies (iii). 89

(3) $\Rightarrow$ (4). Let $\check{\mu} * \tilde{\rho} * \mu = \tilde{\rho}$. Because G is discrete and ρ is a probability 90
measure, there is a $g_0 \in G$ such that $\tilde{\rho}(g_0) = \sup_{g \in G} \tilde{\rho} > 0$. For every positive 91
integer n , we have 92

$$\tilde{\rho}(g_0) = \sum_{h_1, h_2} \tilde{\rho}(h_1 g_0 h_2) \mu^n(h_1) \check{\mu}^n(h_2).$$

Thus, $\tilde{\rho}(h_1 g_0 h_2) = \tilde{\rho}(g_0)$ for all $h_1 \in S^n$ and $h_2 \in S^{-n}$. If we now fix an $h_2 \in S^{-n}$ 93
we have 94

$$1 \geq \sum_{h_1 \in S^n} \tilde{\rho}(h_1 g_0 h_2) = |S^n| \tilde{\rho}(g_0).$$

Therefore, 95

$$d_\mu = \lim_{n \rightarrow \infty} |S^n| \leq \frac{1}{\tilde{\rho}(g_0)}.$$

(4) $\Rightarrow$ (5). Since $|S^n|$ is monotonic there is an N_1 such that $|S^n| = d_\mu$ for all 96
 $n \geq N_1$. Thus, for all $g \in S$ we have 97

$$gS^n = S^n g = S^{n+1},$$

for all $n \geq N_1$. From this, we see that for all $n \geq N_1$ we have 98

$$S^{(-n+k)} S^{-k} = S^{-k} S^{(n+k)} = S^n$$

and 99

$$S^{-(n+k)} S^k = S^k S^{-(n+k)} = S^{-n}.$$

Clearly, the sets $R_{n,1} = S^{-n} S^n$ and $R_{n,2} = S^n S^{-n}$ are nondecreasing, symmetric 100
and $|R_{n,j}| \leq d_\mu^2$. Hence, $R_{n,j} = R_{N_0,j} =: R_j$ for $j = 1, 2$ and for all $n \geq N_0$ 101
where $N_0 \geq N_1$ is sufficiently large. Then 102

$$\begin{aligned} R_1 R_2 &= S^{-n} S^n S^n S^{-n} = S^{-n} S^{2n} S^{-n} \\ &= S^{-n} S^n = S^n S^{-n} = R_1 = R_2 \end{aligned}$$

Since $R_{n,j}$ is symmetric, this then implies that $R := R_1 = R_2$ is a subgroup of G . 103

Furthermore, if $g \in S$, then $g^{-1}Rg = g^{-1}R_{n,1}g = R_{n+1,1} = R$ for $n \geq N_0$.
 Since S is adapted this implies that R is a normal subgroup of G . Finally, we have
 $N_\mu = R = S^{-n}S^n$ by Lemma E.3.

(5) $\Rightarrow$ (1). Because N_μ is normal for any $g \in S$ we have

$$S^n \subset g^n N_\mu.$$

Therefore, $\mu^n(g^n N_\mu) = 1$ for all $n \geq 1$.

The following lemma will be useful to achieve our desired results.

Lemma E.3. Suppose $X \leq C_0 Y k^{-\frac{D}{2}} + kZ$, that $Y/Z \geq 1$, and that

$$k \leq (2Y/Z)^{\frac{2}{D+2}} < k + 1,$$

with $X, Y, Z, C_0, D, k > 0$. Then $X^{1+\frac{2}{D}} \leq C' Y^{\frac{2}{D}} Z$ for some constant C' which is independent of X, Y, Z , and k .

Proof.

$$\begin{aligned} X &\leq C_0 Y k^{-\frac{D}{2}} + kZ \leq C_0 Y k^{-\frac{D}{2}} + Z (2Y/Z)^{\frac{2}{D+2}} \\ &= C_0 k^{-\frac{D}{2}} Y + Z^{\frac{D}{D+2}} (2Y)^{\frac{2}{D+2}} \\ &\leq C_0 Y \left((2Y/Z)^{\frac{2}{D+2}} - 1 \right)^{-\frac{D}{2}} + Z^{\frac{D}{D+2}} (2Y)^{\frac{2}{D+2}}. \end{aligned}$$

Since $1 < 2^{\frac{2}{D+2}} \leq (2Y/Z)^{\frac{2}{D+2}}$ we see that

$$\frac{(2Y/Z)^{2/D+2}}{(2Y/Z)^{2/D+2} - 1} = 1 + \frac{1}{(2Y/Z)^{2/D+2} - 1} \leq 1 + \frac{1}{2^{2/D+2} - 1}.$$

So, setting $C_1 = 1 + \frac{1}{2^{2/D+2} - 1}$ we have

$$\frac{1}{\{2Y/Z\}^{\frac{2}{D+2}} - 1} \leq \frac{C_1}{\{2Y/Z\}^{\frac{2}{D+2}}}.$$

Therefore,

$$X \leq C_0 C_1^{D/2} Y \left(\{2Y/Z\}^{\frac{2}{D+2}} \right)^{-\frac{D}{2}} + Z^{\frac{D}{D+2}} (2Y)^{\frac{2}{D+2}} = C_2 Y^{\frac{2}{D+2}} Z^{\frac{D}{D+2}},$$

where $C_2 = C_0 C_1^{D/2} 2^{-D/D+2} + 2^{2/D+2}$. Finally,

$$X^{\frac{D+2}{D}} \leq C' Z Y^{\frac{2}{D}},$$

where $C' = C_2^{D+2/D}$.

Throughout the remainder of this appendix, unless otherwise stated, we will be assuming that G is an infinite discrete group and that F is an adapted probability density function on G with associated probability measure μ and support S . Let $\tilde{F}(x) = F(x^{-1})$. We will also let $\|\cdot\|_{p \rightarrow q}$ denote the operator norm from $L^p(G)$ to $L^q(G)$. The proof of the following Lemma by Varopoulos et al. lies outside of the scope of this appendix.

Lemma E.4 (Varopoulos et al. [233] VII.2.6). *Let T_k , $k = 1, \dots$ be a sequence of operators which contract all the spaces $L^p(G)$, $1 \leq p \leq \infty$. Let us assume that these operators satisfy the family of inequalities:*

$$\|T_j f\|_2^{2+4/n} \leq C_1 (\|T_j f\|_2^2 - \|T_k T_j f\|_2^2) \|f\|_1^{4/n},$$

for all $f \in L^1(G) \cap L^2(G)$, $j, k \geq 1$ and for some $n > 0$. Then we have

$$\|T_k \dots T_1\|_{2 \rightarrow \infty} \leq (C C_1 n / k)^{n/4} \quad \text{for all } k \geq 1.$$

Moreover, if the adjoint operators $\check{T}_k$ also satisfy the above hypotheses, we have

$$\|T_k \dots T_1\|_{1 \rightarrow \infty} \leq \left(\frac{C' C_1 n}{k} \right)^{n/2} \quad \text{for all } k \geq 2,$$

for some constants C and C'

Lemma E.5. *If T and T_0 are contracting on $L^2(G)$, $\|T\|_{1 \rightarrow 2} < 1$, $\|T_0^k\|_{1 \rightarrow 2} = O(k^{-D/4})$ and there exists a C' such that $\|f\|_2^2 - \|T_0 f\|_2^2 \leq C' (\|f\|_2^2 - \|T f\|_2^2)$ for all $f \in L^1(G) \cap L^2(G)$, then*

$$\|T f\|_2^{2+4/D} \leq C'' (\|T f\|_2^2 - \|T T f\|_2^2) \|f\|_1^{4/D}.$$

Proof. First notice that

$$\begin{aligned} \|(T_0)^j f\|_2^2 - \|(T_0)^{j+1} f\|_2^2 &= \|(I - T_0^2)^{1/2} (T_0)^j f\|_2^2 \\ &\leq \|(I - T_0^2)^{1/2} f\|_2^2 = \|f\|_2^2 - \|T_0 f\|_2^2. \end{aligned}$$

We know that $\|(T_0)^k\|_{1 \rightarrow 2} \leq C_0 k^{-D/4}$ for some C_0 . So,

$$\begin{aligned} \|f\|_2^2 &= \|(T_0)^k f\|_2^2 + \sum_{j=0}^{k-1} (\|(T_0)^j f\|_2^2 - \|(T_0)^{j+1} f\|_2^2) \\ &\leq C_0^2 k^{-D/2} \|f\|_1^2 + k (\|f\|_2^2 - \|T_0 f\|_2^2). \end{aligned}$$

Changing f to Tf in the above yields

135

$$\|Tf\|_2^2 \leq C_0^2 k^{-D/2} \|f\|_1^2 + k (\|Tf\|_2^2 - \|T_0 Tf\|_2^2).$$

Thus,

136

$$\|Tf\|_2^2 - \|T_0 Tf\|_2^2 \leq \|Tf\|_2^2 \leq \|f\|_1^2.$$

Therefore, since C_0 can be taken to be bigger than or equal to one, we can choose $k \in \mathbb{N}^*$ such that

137

138

$$k \leq \{2C_0^2 \|f\|_1^2 / (\|Tf\|_2^2 - \|T_0 Tf\|_2^2)\}^{2/(D+2)} < k + 1$$

The choice of k will depend on f , but as it turns out, the particular value of k will not matter. Indeed, using Lemma E.3 we see that there exists a C' such that, for all $f \in L^1(G) \cap L^2(G)$,

139

140

141

$$\|Tf\|_2^{2+4/D} \leq C' (\|Tf\|_2^2 - \|T_0 Tf\|_2^2) \|f\|_1^{4/D}.$$

By assumption, we have $\|f\|_2^2 - \|T_0 f\|_2^2 \leq C (\|f\|_2^2 - \|Tf\|_2^2)$. This gives us

142

$$\|Tf\|_2^{2+4/D} \leq C'' (\|Tf\|_2^2 - \|T_0 Tf\|_2^2) \|f\|_1^{4/D},$$

for some constant C'' .

□ 143

Lemma E.6. Let F be a nonnegative function on a discrete group G such that $\int F = 1$. Let T be right convolution by F . Then $\|Tf\|_2^2 \leq \|f\|_1^2$.

144

145

Proof.

$$\begin{aligned} \|Tf\|_2^2 &= \int |F * f|^2 \leq \|F\|_\infty \|f\|_1 \int |F * f| \\ &= \|F\|_\infty \|f\|_1 \|F * f\|_1 \\ &\leq \|F\|_\infty \|F\|_1 \|f\|_1^2 \\ &= \|F\|_\infty \|f\|_1^2 \leq \|f\|_1^2. \end{aligned}$$

□ 146

Lemma E.7. Let F , F_0 and F_1 be probability density functions on G such that $F^{(n)} * \check{F}^{(n)} \geq CF_0$ and $\check{F}^{(n)} * F^{(n)} \geq CF_1$ for some constant C and such that F_0 and F_1 are symmetric. Let T , T_0 , and T_1 be the operators of right convolution by $F^{(n)}$, F_0 and F_1 , respectively. Also let $\check{T}$ be the operator of right convolution by $\check{F}^{(n)}$ (equivalently, the adjoint to T). Then there exists a constant C' such that $\|f\|_2^2 - \|T_0^{1/2} f\|_2^2 \leq C' (\|f\|_2^2 - \|Tf\|_2^2)$ and $\|f\|_2^2 - \|T_1^{1/2} f\|_2^2 \leq C' (\|f\|_2^2 - \|Tf\|_2^2)$ for all $f \in L^1(G) \cap L^2(G)$. Furthermore, the operators T , $T_0^{1/2}$, and $T_1^{1/2}$ are contracting on $L^2(G)$.

147

148

149

150

151

152

153

154

Proof. Given an $f \in L^1(G) \cap L^2(G)$, we have

155

$$\begin{aligned} \|f\|_2^2 - \|T_0^{1/2} f\|_2^2 &= (f, (I - (T_0^{1/2})^2) f) = (f, (I - T_0) f) \\ &= \frac{1}{2} \iint (f(xh) - f(x))^2 F_0(h) dh dx \\ &\leq \frac{1}{C} \frac{1}{2} \iint (f(xh) - f(x))^2 F^{(n)} * \check{F}^{(n)}(h) dh dx \\ &= \frac{1}{C} (\|f\|_2^2 - \|Tf\|_2^2) \end{aligned}$$

Likewise for F_1 and T_1 .

156

Because the integrand $(f(xh) - f(x))^2 F_0(h) \geq 0$ it follows that $\|f\|_2^2 - \|T_0^{1/2} f\|_2^2 \geq 0$ and therefore $T_0^{1/2}$ is contracting on $L^2(G)$. Likewise for T and $T_1^{1/2}$. $\square$

159

Corollary E.1. Let F, F_0, F_1, T, T_0 , and T_1 be as in Lemma E.7 except that now let $F^{(n)} * \check{F}^{(n)} \geq CF_0^{(2)}$ and $\check{F}^{(n)} * F^{(n)} \geq CF_1^{(2)}$. Then there exists a constant C' such that $\|f\|_2^2 - \|T_0 f\|_2^2 \leq C' (\|f\|_2^2 - \|Tf\|_2^2)$ and $\|f\|_2^2 - \|T_1 f\|_2^2 \leq C' (\|f\|_2^2 - \|Tf\|_2^2)$ for all $f \in L^1(G) \cap L^2(G)$. Furthermore, T_0 and T_1 are contracting on $L^2(G)$.

164

E.2 Irreducible Measures

165

In this section, we will consider the case where the measure is irreducible (i.e., its support generates G as a subsemigroup). We will show that the volume growth rate of the group determines the rate of decay of the concentration functions. The following lemma will help us to accomplish this goal.

Lemma E.8 ([197] Lemma 5). Let G be a discrete group, F and F_0 probability density functions on G . Assume that $\|F_0^{(k)}\|_\infty = O(k^{-D/2})$ and that F_0 is symmetric. Further assume that $F \geq cF_0$. Then $\|F^{(k)}\|_\infty = O(k^{-D/2})$.

172

Proof. First notice that $F * \check{F} \geq c' F_0 * \check{F}_0 = c' F_0^{(2)}$ and $F * \check{F}$ is symmetric. Letting T and T_0 be the operators of right convolution by F and F_0 , respectively, we have by Corollary E.1 that

175

$$\|f\|_2^2 - \|T_0 f\|_2^2 \leq C' (\|f\|_2^2 - \|Tf\|_2^2),$$

for all $f \in L^1(G) \cap L^2(G)$ and that T and T_0 are contracting on $L^2(G)$.

176

By Lemma E.6 and the fact that $\|F_0^{(k)}\|_\infty = O(k^{-D/2})$, we know that

177

$$\|T_0^{(k)}\|_{1 \rightarrow 2} \leq C_0 (k^{-D/4}).$$

Also by Lemma E.6 we also have $\|T\|_{1 \rightarrow 2} \leq 1$. Therefore, by Lemma E.5 we see that for all $f \in L^1(G) \cap L^2(G)$

$$\|Tf\|_2^{2+4/D} \leq C'' (\|Tf\|_2^2 - \|TTf\|_2^2) \|f\|_1^{4/D}.$$

Note that the same argument works for the adjoint operator $\check{T}$. So by Lemma E.4 we get

$$\|F^{(k)}\|_\infty = O(k^{-D/2}).$$

□

We will now require the following result from Varopolous et al., the proof of which again lies outside the scope of this appendix:

Theorem E.5 (Varopoulos et al. [233] VII.1.1). *Let G be a unimodular compactly generated group. Let U be an open, generating (as a semigroup) neighborhood of the origin $e \in G$, and let F be a nonnegative function such that $\int F = 1$, $\|F\|_\infty < C$ and $\inf_U F \geq \varepsilon > 0$. Then, if $V(m) \geq cm^D$, we have*

$$\|F^{(k)}\|_\infty = O(k^{-D/2}).$$

We will also need the following, which is modified from Proposition 6.21 from [188].

Proposition E.1. *Let H be a normal subgroup of finite index in a discrete group G . If $V_G(n) \geq cn^D$ then $V_H(m) \geq c'm^D$.*

Proof. Let $Q : G \rightarrow G/H$ be the quotient map. Since G/H is finite we can find a finite set B such that $Q(B) = G/H$, $Q|_B$ is one-to-one, and with $e \in B$. In other words, B is a transversal of H in G containing the identity.

Let $A \ni e$ be a finite subset of G and $A_0 = H \cap (B^{-1}A)$. Then A_0 is a finite subset of H and $e \in A_0$. Since $G = BH$,

$$A \subset BA_0.$$

Define the following subsets of H as follows

$$D = (B^{-1}B^2) \cap H, \quad D_0 = \cup_{x \in B} (x^{-1}A_0x), \quad A_1 = DD_0.$$

Since B and A are finite, A_0 and D are finite. Therefore D_0 and A_1 are also finite. It is easy to see that all of these subsets contain the identity. To prove the proposition it suffices to show that

$$|A^n| \leq k|A_1^n|,$$

where k is a constant (in this case $k = |G/H|$) and where $|A|$ denotes the number of elements in A . It can be shown that

$$(BA_0)^n \subset B(A_1)^n.$$

We will illustrate this in the case where $n = 3$, the proof for other n 's being similar. 204
Let $b_1, b_2, b_3 \in B$ and $a_1, a_2, a_3 \in A_0$. Find $b_4, b_5 \in B$ and $h, h' \in H$ such that 205

$$b_2b_3 = b_4h, \quad b_1b_2b_3 = b_1b_4h = b_5h'h.$$

This is possible since $G = BH$. Then both h, h' belong to $(B^{-1}B^2) \cap H \subset D$. So, 206

$$\begin{aligned} (b_1a_1)(b_2a_2)(b_3a_3) &= (b_1b_2b_3)[(b_2b_3)^{-1}a_1(b_2b_3)][b_3^{-1}a_2b_3]a_3 \\ &= b_5h'h(h^{-1}b_4^{-1}a_1b_4h)(b_3^{-1}a_2b_3)a_3 \\ &= b_5(h'b_4^{-1}a_1b_4)(hb_3^{-1}a_2b_3)(eea_3e^{-1}) \in B(A_1)^3. \end{aligned}$$

Consider $xH \cap B(A_1)^n = xH \cap \{b_1A_1^n \cup \dots \cup b_kA_1^n\}$. Since $A_1^n \subset H$ and B is a 207
transversal of H in G , exactly one $b_jA_1^n$ intersects xH . Thus, 208

$$|xH \cap B(A_1)^n| = |A_1^n|,$$

and therefore 209

$$|B(A_1)^n| = k|A_1^n|.$$

Since $A^n \subset (BA_0)^n \subset B(A_1)^n$, 210

$$|A^n| \leq |B(A_1)^n| = k|A_1^n|.$$

□ 211

The following lemma will be useful in eliminating the requirement in Theorem E.5 212
that U contain the identity. 213

Lemma E.9 ([197]). *Let S generate a discrete group G as a semigroup. Let k be 214
such that $e \in S^k$. Then S^k generates a normal subgroup H of G as a semigroup. If 215
 G is infinite, then $gp(S^k)$ is infinite as well. 216*

Proof. Let $x \in S^k$. Since S generates G as a semigroup, $x^{-1} \in S^\ell$ for some ℓ . So 217
 $e \in S^{k+\ell}$. So $x^{-1} \in S^{m(k+\ell)+\ell}$ for all integers $m \geq 0$. Let $m = k - 1$. Then $x^{-1} \in$ 218
 $S^{(k-1)(k+\ell)+\ell} = S^{k(k+\ell-1)} \subset \text{semigp}(S^k)$. Thus, $\text{semigp}(S^k) = gp(S^k) = H$. 219

To prove that H is normal, we shall show that $aha^{-1} \in H$ for all $a \in G$ and 220
 $h \in H$. Because S generates G as a semigroup, and by the definition of H it is 221
possible to find $t_{i,j}, a_i, b_i \in S$ such that 222

$$\begin{aligned} h &= t_{1,1} \cdots t_{1,k} \cdots t_{n,1} \cdots t_{n,k} \\ a &= a_1 \cdots a_l \\ a^{-1} &= b_1 \cdots b_m. \end{aligned}$$

Thus, 223

$$aha^{-1} = a_1 \cdots a_l t_{1,1} \cdots t_{n,k} b_1 \cdots b_m.$$

Notice that there are $kn + l + m$ terms from S on the right-hand side. Also notice that $e = a_1 \cdots a_l b_1 \cdots b_m$ which has $l + m$ terms from S on the right-hand side. Multiplying aha^{-1} by the identity element $k - 1$ times yields

$$aha^{-1} = a_1 \cdots a_l t_{1,1} \cdots t_{n,k} b_1 \cdots b_m (a_1 \cdots a_l b_1 \cdots b_m)^{k-1}.$$

Notice that the right-hand side of this equation contains $k(n + m + l)$ terms from S . Thus, $aha^{-1} \in H$.

Note that $|S^i|$ is a monotonically increasing sequence. If G is infinite, the $|S^i|$ must go to infinity and therefore $|S^{k^j}|$ must go to infinity as well. Thus $gp(S^k)$ must be infinite. $\square$

Proposition E.2 ([197]). Assume the discrete group G satisfies $V_G(m) \geq cm^D$ and that the set S generates G as a semigroup. Then there exists some finite subset $S_0 \subset S$ that generates a subgroup G_0 as a semigroup with $V_{G_0}(m) \geq cm^D$.

Proof. Because G has $V_G(m) \geq cm^D$, G contains a finite set $K = \{e, x_1, \dots, x_k\}$ for which $|K^n| \geq cn^D$. Then K generates a subgroup $H \subset G$ with $V_H(m) \geq cm^D$. Only finitely many elements of S are need to express the x_i 's and their inverses. Call this set $S_0 \subset S$. Now we will show that the semigroup generated by S_0 is in fact a group. Let $s \in S_0$. Then $x_i = a_1 a_2 \cdots a_j s a_{j+1} \cdots a_l$ for some $1 \leq i \leq k$ and $a_1, \dots, a_l \in S_0$. Also, $x_i^{-1} = b_1 \cdots b_m$ for $b_1, \dots, b_m \in S_0$. So $e = a_1 \cdots a_j s a_{j+1} \cdots a_l b_1 \cdots b_m$, and thus by conjugation we get $e = s a_{j+1} \cdots a_l b_1 \cdots b_m a_1 \cdots a_j$. Thus, the inverse of s is $a_{j+1} \cdots a_l b_1 \cdots b_m a_1 \cdots a_j \in \text{semigp}(S_0)$. Thus, $\text{semigp}(S_0) = gp(S_0) = G_0$. Clearly G_0 satisfies $V_{G_0} \geq cm^D$ since it contains H . $\square$

It is now possible to relate the rate of decay of the concentration functions to the volume growth rate of the group when the probability measure is irreducible.

Proof of Theorem E.1: By Proposition E.2, we can choose a finite $S_0 \subset S$ which generates a subgroup G_0 as a semigroup with $V_{G_0}(m) \geq cm^D$. Now $e \in S_0^m$ for some integer m . By Lemma E.9, S_0^m generates, as a semigroup, a finitely generated normal subgroup of G_0 , call it $G' = gp(S_0^m)$. G' is of finite index in G_0 since S_0 is finite (there are at most $|S_0|^m$ cosets of G' in G_0). Since G' is normal and of finite index in G_0 , by Proposition E.1, we have $V_{G'}(m) \geq c'm^D$.

Now consider $\tilde{F} := \alpha 1_{S_0^m \cup S_0^{-m}}$ where α is chosen so that $\tilde{F}$ is normalized. Then $\tilde{F}, G'$ and S_0^m satisfy the hypotheses of Theorem E.5. So $\|\tilde{F}^{(k)}\|_\infty = O(k^{-D/2})$. Let $\{x_1, \dots, x_k\}$ be a generating set for G' . Since $e \in S_0^m$, we know that $x_1^\pm, \dots, x_n^\pm \in S_0^{\ell}$ for some ℓ . So $F^{(m\ell)} \geq c\tilde{F}$ for some c (since $\text{supp}(F^{(m\ell)}) = (\text{supp}(F))^{(m\ell)} = S_0^{m\ell} \supset S_0^{\ell} \supset \{e, x_1^\pm, \dots, x_k^\pm\}$ and $\tilde{F}$ has finite support). $\tilde{F}$ is symmetric, so by Lemma E.8 we have $\|F^{(mkl)}\|_\infty = O(k^{-D/2})$. By monotonicity then $\|F^{(k)}\|_\infty = O(k^{-D/2})$. $\square$

E.3 Adapted Measures

260

If the measure we are interested in is merely adapted (i.e., its support generates G as a subgroup, but not as a subsemigroup), then the rate of decay of the concentration functions is governed by the volume growth rate of the subgroup N_μ , as we will show in this section. The following lemma will be necessary to our investigation.

Lemma E.10 ([198]). *Let F be an adapted probability density function on a discrete group G with support S . Let $A \subset S^n S^{-n}$ [$B \subset S^{-n} S^n$] for some $n > 0$ with $V_{gp(A)}(m) \geq cm^D$ [$V_{gp(B)}(m) \geq cm^D$]. Then there exists a symmetric probability density function F_0 [F_1] such that $F^{(n)} * \check{F}^{(n)} \geq CF_0$ [$\check{F}^{(n)} * F^{(n)} \geq CF_1$] for some constant C and $\|F_0^{(k)}\|_\infty = O(k^{-D/2})$ [$\|F_1^{(k)}\|_\infty = O(k^{-D/2})$].*

Proof. By the definition of volume growth, there exists a symmetric finite set $K = \{k_1, \dots, k_N\} \subset gp(A)$ such that $|K^m| \geq cm^D$. It is possible to find a set $\{a_{i,j}\} \subset A \cup A^{-1}$ such that $k_i = a_{i,1} \cdots a_{i,n_i}$ for each i . Let $A_0 = \{a_{i,j} \mid 1 \leq i \leq N, 1 \leq j \leq n_i\}$. Let $\Omega = A_0 \cup A_0^{-1}$. Notice that $\Omega \subset S^n S^{-n}$ since $S^n S^{-n}$ is symmetric. Also notice that $V_{gp(A_0)}(m) \geq cm^D$ since $K \subset gp(A_0)$. Then $F_0 = \frac{1}{|\Omega|} 1_\Omega$ satisfies the hypothesis of Theorem E.1 for $gp(A_0)$ and therefore $\|F_0^{(k)}\|_\infty = O(k^{-D/2})$. Because $\Omega \subset S^n S^{-n}$ and since Ω is a finite set, we can find a constant C such that $F^{(n)} * \check{F}^{(n)} \geq CF_0$. Similar arguments work for the $B \subset S^{-n} S^n$ case. $\square$

Theorem E.6 ([198]). *Let G be a discrete group and let F be an adapted probability density function on G with $S = \text{supp}(F)$ such that for some positive integer n and some sets $A \subset S^n S^{-n}$ and $B \subset S^{-n} S^n$ we have $V_{gp(A)}(m) \geq cm^D$ and $V_{gp(B)}(m) \geq cm^D$. Then $\|F^{(k)}\|_\infty = O(k^{-D/2})$.*

Proof. Using Lemma E.10, we can find a symmetric probability density function F_0 such that $F^{(n)} * \check{F}^{(n)} \geq CF_0$. Likewise, we can find a symmetric probability density function F_1 such that $\check{F}^{(n)} * F^{(n)} \geq CF_1$. Let T, T_0 , and T_1 be the operators of right convolution by $F^{(n)}$, F_0 and F_1 , respectively. Also by Lemma E.10 we know that $\|F_0^{(k)}\|_\infty = O(k^{-D/2})$ and therefore $\|T_0^k\|_{1 \rightarrow 2} = \|(T_0^{1/2})^2 k\|_{1 \rightarrow 2} = O(k^{-D/4})$. Since $T_0^{1/2}$ is contracting on $L^2(G)$ (see Lemma E.7), we see that $\|(T_0^{1/2})^k\|_{1 \rightarrow 2} = O(k^{-D/4})$. Also by Lemma E.7 we have

$$\|f\|_2^2 - \|T_0^{1/2} f\|_2^2 \leq C' (\|f\|_2^2 - \|Tf\|_2^2),$$

for all $f \in L^1(G) \cap L^2(G)$ and for some constant C' . By Lemma E.6 we see that $\|T\|_{1 \rightarrow 2} \leq 1$. Putting all of this together with Lemma E.5 we see that

$$\|Tf\|_2^{2+4/D} \leq C'' (\|Tf\|_2^2 - \|T Tf\|_2^2) \|f\|_1^{4/D},$$

for all $f \in L^1(G) \cap L^2(G)$.

This entire argument works for the adjoint operator $\check{T}$ using $\check{F}^{(n)} * F^{(n)} \geq F_1$ and T_1 . Therefore, by Lemma E.4 we see that

$$\|F^{(nk)}\|_\infty = O(k^{-D/2}).$$

Finally, by monotonicity we get

$$\|F^{(k)}\|_\infty = O(k^{-D/2}).$$

□ 296

Since the sets $S^n S^{-n}$ and $S^{-n} S^n$ are obviously so important to our investigation, we will now consider these sets and certain of their subsets more closely.

Lemma E.11 ([198]). *Given a fixed integer $m > 0$ the following are equivalent:*

1. *There exists an integer $n > 0$ such that $S^m \subset S^n S^{-n}$*
2. *There exists an integer $k > 0$ such that $S^m \cap S^k S^{-k} \neq \emptyset$*
3. *There exists an integer $k > 0$ such that $S^{m+k} \cap S^k \neq \emptyset$*
4. *There exists an integer $k > 0$ such that $S^m \cap S^{-k} S^k \neq \emptyset$*
5. *There exists an integer $l > 0$ such that $S^m \subset S^{-l} S^l$*

Proof. The equivalences of (2), (3), and (4) are obvious, as are (1) $\Rightarrow$ (2) and (5) $\Rightarrow$ (4). To show (2) $\Rightarrow$ (1): Let $n = k + m$. Then

$$\begin{aligned} S^m \cap S^k S^{-k} \neq \emptyset &\Leftrightarrow S^m \cap S^{n-m} S^{-(n-m)} \neq \emptyset \\ &\Rightarrow \exists y \text{ such that } y \in S^m \text{ and } y \in S^{n-m} S^{-(n-m)} \\ &\Rightarrow e = yy^{-1} \in S^{n-m} S^{-(n-m)} S^{-m} = S^{n-m} S^{-n} \\ &\Rightarrow S^m \subset S^n S^{-n}. \end{aligned}$$

Similarly, (4) $\Rightarrow$ (5).

□ 307

Proposition E.3 ([206]). *Let G be a group and let $A \subset G$ be some nonempty subset. When there exists $a, b \in G$ with $aA \cup bA \subset A$, then either a and b generates a free subsemigroup or there exists z in the subsemigroup generated by a and b such that $zA \subset aA \cap bA$.*

Proof. Suppose a and b do not generate a free subsemigroup. Then there exist two formal positive power words $W(x, y) \neq V(x, y)$ such that $W(a, b) = V(a, b)$. Without loss of generality we may assume that the two words do not have the same first entry on the left. If one of the words, say $W(x, y)$, is empty, then $A = eA = V(a, b)A$. Note that since $V(x, y) \neq W(x, y)$ we know that $V(x, y)$ is not also empty, so it must have a first entry, let us call it v . Now, using $aA \cup bA \subset A$ we get $V(a, b)A \subset vA$. Let $z = a$ if $v = b$ or vice versa. Then $zA \subset A = V(a, b)A \subset vA$. Therefore $zA \subset aA \cap bA$.

If both $V(x, y)$ and $W(x, y)$ are nonempty then let v and w be the first entries on the left of $V(a, b)$ and $W(a, b)$, respectively. Let $z = W(a, b) = V(a, b)$. Now, since $aA \cup bA \subset A$ we have $zA \subset vA$ and $zA \subset wA$. However, $V(x, y)$ and $W(x, y)$ were assumed to start with different entries, so $v \neq w$ yielding the result. $\square$

By taking inverses, we get the following corollary:

Corollary E.2. *Let G be a group and let $A \subset G$ be some nonempty subset. When there exists $a, b \in G$ with $Aa \cup Ab \subset A$, then either a and b generates a free subsemigroup or there exists z in the subsemigroup generated by a and b such that $Az \subset Aa \cap Ab$.*

Lemma E.12 ([198]). *Let G be a discrete group and S be an adapted subset such that $S^m \cap S^n = \emptyset$ when $m \neq n$. Then either $\text{semigp}(S)$ contains a free subsemigroup on two generators or both $S^m S^{-m} \subset \bigcup S^{-k} S^k$ and $S^{-m} S^m \subset \bigcup S^k S^{-k}$ for all m .*

Proof. Let $A = \bigcup_{k=1}^{\infty} S^k$. If $a_1 \cdots a_m A \cap b_1 \cdots b_m A = \emptyset$ for some $a_i, b_i \in S$ then, by Proposition E.3, $\alpha = a_1 \cdots a_m$ and $\beta = b_1 \cdots b_m$ would generate a free subsemigroup. We may then assume that $\alpha A \cap \beta A \neq \emptyset$ for all $\alpha, \beta \in S^m$. For fixed $\alpha, \beta \in S^m$, $\alpha A \cap \beta A \neq \emptyset$ implies $\alpha s_1 \cdots s_k = \beta t_1 \cdots t_l$ for some $s_i, t_i \in S$. Furthermore, since the powers of S are pairwise disjoint, we have $k = l$. Thus, $\beta^{-1} \alpha = t_1 \cdots t_k s_k^{-1} \cdots s_1^{-1}$, and therefore $\beta^{-1} \alpha \in S^k S^{-k}$. Since α and β are arbitrary elements of S^m , $S^{-m} S^m \subset \bigcup S^k S^{-k}$. We get $S^m S^{-m} \subset \bigcup S^{-k} S^k$ in a similar fashion using Corollary E.2 instead of Proposition E.3. $\square$

Thus, there are three cases we will need to consider:

1. $S^m \subset S^n S^{-n}$ and $S^m \subset S^{-n} S^n$ for some $m, n > 0$ (equivalently $S^n \cap S^k \neq \emptyset$ for some $n, k > 0$ by Lemma E.11);
2. $S^{-m} S^m \subset \bigcup S^k S^{-k}$ and $S^m S^{-m} \subset \bigcup S^{-k} S^k$ for all $m > 0$;
3. G contains a free subsemigroup on two generators and $S^m \cap S^n = \emptyset$ for all $m \neq n$.

We will now consider each of the three cases individually to prove Theorem E.2:

Proof of Theorem E.2:

Case I: $S^m \subset S^n S^{-n}$ and $S^m \subset S^{-n} S^n$ for some $m, n > 0$

By Theorem E.3 we know that $N_\mu = gp(\bigcup (S^k S^{-k} \cup S^{-k} S^k))$ and therefore $gp(S^m) \subset N_\mu$. Since the sets $S^k S^{-k}$ and $S^{-k} S^k$ are increasing, we also have $N_\mu = gp(\bigcup (S^{mk} S^{-mk} \cup S^{-mk} S^{mk}))$. Thus $N_\mu \subset gp(S^m)$ and thus $N_\mu = gp(S^m)$. Now using Theorem E.6 with $A = B = S^m$ we get the desired result.

Case II: $S^{-m} S^m \subset \bigcup S^k S^{-k}$ and $S^m S^{-m} \subset \bigcup S^{-k} S^k$ for all $m > 0$

Again by Theorem E.3 we know that

$$N_\mu = gp\left(\bigcup\left(S^k S^{-k} \cup S^{-k} S^k\right)\right) = gp\left(\bigcup S^k S^{-k}\right) = gp\left(\bigcup S^{-k} S^k\right)$$

Since $V_{N_\mu}(m) \geq cm^D$ there exists a finitely generated subgroup $G_0 \subset N_\mu$ with $V_{G_0}(m) \geq cm^D$. Since G_0 is finitely generated, $G_0 \subset gp(S^n S^{-n})$ and $G_0 \subset gp(S^{-n} S^n)$ for some n . Letting $A = S^n S^{-n}$ and $B = S^{-n} S^n$ in Theorem E.6, we again obtain $\|F^{(k)}\|_\infty = O(k^{-D/2})$.

Case III: G contains a free subsemigroup on two generators and $S^m \cap S^n = \emptyset$ for all $m \neq n$

Since in this case we are assuming that G contains a free subsemigroup, $V_G(m) \geq cm^D$ for all $D > 0$. Therefore we need to show that $\|F^{(k)}\|_\infty = O(k^{-D/2})$ for all $D > 0$.

Consider the probability density function $\left(\frac{F+F^{(2)}}{2}\right)^{(n)}$. This is clearly adapted since its support is $S \cup S^2 \supset S$. Also, by Theorem E.3, $N_\mu = G$ since $S \subset (S \cup S^2)(S \cup S^2)^{-1}$. Furthermore, it is clear that $(S \cup S^2)^2 \cap (S \cup S^2) \neq \emptyset$. Therefore, using Lemma E.11 we see that Case I of this proof applies yielding

$$\text{for all } D > 0 \text{ there exists a } C_1 \text{ such that } \left\| \left(\frac{F + F^{(2)}}{2} \right)^{(n)} \right\|_\infty \leq \frac{C_1}{n^{D/2}}.$$

Now, since in this case we are assuming that $\text{supp}(F^{(m)}) = S^m$ and $\text{supp}(F^{(\tilde{m})}) = S^{\tilde{m}}$ have empty intersection for $m \neq \tilde{m}$, we have the equality

$$\begin{aligned} \left\| \left(\frac{F + F^{(2)}}{2} \right)^{(n)} \right\|_\infty &= \left\| \sum_{\ell=0}^n \frac{n!}{\ell!(n-\ell)!} \frac{1}{2^n} F^{(n+\ell)} \right\|_\infty \\ &= \max_{\ell=0, \dots, n} \left\{ \frac{n!}{\ell!(n-\ell)!} \frac{1}{2^n} \|F^{(n+\ell)}\|_\infty \right\}. \end{aligned}$$

Letting $n = 2k$ and considering the term with $\ell = k$ we get

$$\begin{aligned} \frac{(2k)!}{k!k!} \frac{1}{2^{2k}} \|F^{(3k)}\|_\infty &\leq \max_{\ell=0, \dots, 2k} \left\{ \frac{(2k)!}{\ell!(2k-\ell)!} \frac{1}{2^{2k}} \|F^{(2k+\ell)}\|_\infty \right\} \\ &= \left\| \left(F + F^{(2)} \right)^{2k} \right\|_\infty \leq \frac{C_1}{(2k)^{D/2}}. \end{aligned}$$

But by a simple calculation involving Stirling's formula we see that $\frac{(2k)!}{k!k!} \frac{1}{2^{2k}} \sim \frac{1}{\sqrt{\pi k}}$. So,

$$\frac{(2k)!}{k!k!} \frac{1}{2^{2k}} \geq \frac{C_2}{\sqrt{\pi k}} = \frac{C_3}{\sqrt{k}}$$

Thus,

412

E Rate of Decay of Concentration Functions on Discrete Groups, by T. M. Retzlaff

for each $D > 0$ there exists a C_1 such that $\frac{C_3 \|F^{(3k)}\|_\infty}{\sqrt{k}} \leq \frac{C_1}{(2k)^{D/2}},$

whence

377

for each $D > 0$ there exists a C_4 such that $\|F^{(3k)}\|_\infty \leq \frac{C_4}{k^{(D-1)/2}}.$

Since this is true for all $D > 0$ we can replace D by $D + 1$ (notice that the constant 378
may change in so doing since the constant may depend on the specific value of D) 379
to get 380

for each $D > 0$ there exists a C_5 such that $\|F^{(3k)}\|_\infty \leq \frac{C_5}{k^{D/2}}.$

By monotonicity, we finally arrive at

381

for each $D > 0$ there exists a C_6 such that $\|F^{(k)}\|_\infty \leq \frac{C_6}{k^{D/2}}.$

This is the desired result.

□ 382

AUTHOR QUERIES

AQ1. Kindly check the suppressed running head is ok.

AQ2. Please check if the references given in the text can be moved to the list.

Uncorrected Proof

References

1

AQ1

1. Argabright, L.N.: A note on invariant integrals on locally compact semigroups. *Proc. Am. Math. Soc.* **17**, 377–382 (1966) 2 3
2. Babillot, M.: Théorie du renouvellement pour des chaînes semi-markoviennes transientes. *Ann. Inst. Henri Poincaré (B), Probabilités et Statistiques* **24**, 507–569 (1988) 4 5
3. Baker, J.W.: Measure algebras on semigroups. In: *The Analytical and Topological Theory of Semigroups. de Gruyter Exp. Math.*, vol. 1, pp. 221–252. de Gruyter, Berlin (1990) 6 7
4. Baldi, P.: Caractérisation des groupes de lie connexes récurrents. *Ann. Inst. Henri Poincaré (B), Probabilités et Statistiques* **17**, 281–308 (1981) 8 9
5. Baldi, P., Lohoué, N., Peyrière, J.: Sur la classification des groupes récurrents. *C. R. Acad. Sci. Paris, Sér. A–B* **285**, 1103–1104 (1977) 10 11
6. Barnsley, M.F., Demko, S.G., Elton, J.H., Geronimo, J.S.: Invariant measures for Markov processes arising from iterated function systems with place-dependent probabilities. *Ann. Inst. H. Poincaré Probab. Statist.* **24**(3), 367–394 (1988) 12 13 14
7. Bartoszek, W.: On concentration functions on discrete groups. *Ann. Probab.* **22**(3), 1596–1599 (1994) 15 16
8. Berbee, H.: Recurrence and transience for random walks with stationary increments. *Z. Wahrsch. Verw. Gebiete* **56**, 531–536 (1981) 17 18
9. Berglund, J.F., Hofmann, K.H.: Compact semitopological semigroups and weakly almost periodic functions. *Lecture Notes in Math.*, vol. 42. Springer, Berlin (1967) 19 20
10. Berglund, J.F., Junghenn, H.D., Milnes, P.: *Analysis on semigroups. Function Spaces, Compactifications, Representations.* Wiley, New York (1989) 21 22
11. Berstel, J., Perrin, D.: *Theory of Codes.* Academic, Orlando (1985) 23
12. Bhattacharya, R.N.: Speed of convergence of the n -fold convolution of a probability measure on a compact group. *Z. Wahrsch. Verw. Gebiete* **25**, 1–10 (1972/1973) 24 25
13. Bougerol, P.: Fonctions de concentration sur les extensions compactes de groupes abéliens. *C. R. Acad. Sci. Paris, Sér. A* **283**, 527–529 (1976) 26 27
14. Bougerol, P.: Fonctions de concentration sur certains groupes localement compacts. *Z. Wahrsch. Verw. Gebiete* **45**(2), 135–157 (1978) 28 29
15. Bougerol, P.: Une majoration universelle des fonctions de concentration sur les groupes localement compacts non compacts. In: *Probability measures on groups (Proc. Fifth Conf., Oberwolfach, 1978).* *Lecture Notes in Math.*, vol. 706, pp. 36–40. Springer, Berlin (1979) 30 31 32
16. Bougerol, P.: Tightness of products of random matrices and stability of linear stochastic systems. *Ann. Probab.* **15**(1), 40–74 (1987) 33 34
17. Bougerol, P., Lacroix, J.: Products of random matrices with applications to Schrödinger operators. *Progress in Probability and Statistics*, vol. 8. Birkhäuser, MA (1985) 35 36
18. Bowen, R.M., Wang, C.C.: Introduction to vectors and tensors, vol. 2. *Vector and tensor analysis. Mathematical Concepts and Methods in Science and Engineering.* Plenum, London (1976) 37 38 39
19. Breiman, L.: *Probability.* Addison-Wesley, MA (1968) 40

20. Brissaud, M.: Sur les marches aléatoires dans les demi-groupes topologiques. C. R. Acad. Sci. Paris, Sér. A **268**, 1286–1289 (1969) 41
21. Brown, D.: On clans of nonnegative matrices. Proc. Am. Math. Soc. **15**, 671–674 (1964) 42
22. Brunel, A., Revuz, D.: Un critère probabiliste de compacité des groupes. Ann. Prob. **2**, 745–746 (1974) 43
23. Budzban, G.: Necessary and sufficient conditions for the convergence of convolution products of non-identical distributions on finite abelian semigroups. J. Theoret. Probab. **7**(3), 635–646 (1994) 44
24. Budzban, G., Feinsilver, P.: Quaternionic algebras with applications to the road coloring problem. In: Proceedings of the Workshop on Semigroups and Automata, ICALP 2005, Lisbon, pp. 23–42 (2005) 45
25. Budzban, G., Feinsilver, P.: A vector space approach to the road coloring problem. In: Workshops on Words and Automata, St Petersburg, Russia, June 2006 46
26. Budzban, G., Mukherjea, A.: Convolution products of nonidentical distributions on a topological semigroup. J. Theoret. Probab. **5**(2), 283–307 (1992) 47
27. Budzban, G., Mukherjea, A.: A semigroup approach to the road coloring problem. In: Probability on Algebraic Structures. *Contemporary Mathematics*, vol. 261, pp. 195–207. American Mathematical Society, RI (2000) 48
28. Byczkowski, T., Woś, J.: On infinite products of independent random elements on metric semigroups. Colloq. Math. **37**(2), 271–285 (1977) 49
29. Carruth, J.H., Hildebrandt, J.A., Koch, R.J.: The Theory of Topological Semigroups. Marcel Dekker, New York (1983) 50
30. Carruth, J.H., Hildebrandt, J.A., Koch, R.J.: The Theory of Topological Semigroups, vol. 2. Marcel Dekker, New York (1986) 51
31. Center, B., Mukherjea, A.: More on limit theorems for iterates of probability measures on semigroups and groups. Z. Wahrsch. Verw. Gebiete **46**(3), 259–275 (1979) 52
32. Chakraborty, S., Mukherjea, A.: Completely simple semigroups of real $d \times d$ matrices and recurrent random walks. In: M.A.G. Viana, H.P. Wynn (eds.) Algebraic Methods in Statistics and Probability II. *Contemporary Mathematics*, vol. 516. American Mathematical Society, RI (2010) (to appear) 53
33. Chakraborty, S., Rao, B.V.: Convolution powers of probabilities on stochastic matrices. J. Theoret. Probab. **14**(2), 599–603 (2001) 54
34. Chamayou, J.F., Letac, G.: A transient random walk on stochastic matrices with Dirichlet distributions. Ann. Probab. **22**(1), 424–430 (1994) 55
35. Choquet, G., Deny, J.: Sur l'équation de convolution $\mu = \mu * \sigma$. C. R. Acad. Sci. Paris **250**, 799–801 (1960) 56
36. Choy, S.T.L.: Idempotent measures on compact semigroups. Proc. Lond. Math. Soc. (3) **20**, 717–733 (1970) 57
37. Chung, K.L.: Markov Chains with Stationary Transition Probabilities, 2nd edn. Springer, Berlin (1967) 58
38. Chung, K.L.: A course in probability theory. *Probability and Mathematical Statistics*, vol. 21, 2nd edn. Academic, New York (1974) 59
39. Chung, K.L., Fuchs, W.H.: On the distribution of values of sums of random variables. Mem. Am. Math. Soc. **1951**(6), 12 (1951) 60
40. Cigler, J.: Über die Grenzverteilung von Summen Markowscher Ketten auf endlichen Gruppen i, ii. Z. Wahrsch. Verw. Gebiete **1**, 415–432 (1963) 61
41. Clark, W.E.: Remarks on the kernel of a matrix semigroup. Czechoslovak Math. J. **15**(90), 305–309 (1965) 62
42. Clifford, A.H., Preston, G.B.: The Algebraic Theory of Semigroups, vol. II. *Mathematical Surveys*, vol. 7. American Mathematical Society, RI (1967) 63
43. Clifford, A.H., Preston, G.B.: The Algebraic Theory of Semigroups, vol. I, 2nd edn. reprinted with corrections. *Mathematical Surveys*, vol. 7. American Mathematical Society, RI (1977) 64
44. Cohen, J.E., Kesten, H., Newman, C.M. (eds.): Random matrices and their applications. *Contemporary Mathematics*, vol. 50. American Mathematical Society, RI (1986) 65

AQ2

45. Collins, H.S.: Primitive idempotents in the semigroup of measures. *Duke Math. J.* **27**, 397–400 (1960) 95
46. Collins, H.S.: The kernel of a semigroup of measures. *Duke Math. J.* **28**, 387–392 (1961) 96
47. Collins, H.S.: Convergence of convolution iterates of measures. *Duke Math. J.* **29**, 259–264 (1962) 97
48. Collins, H.S.: Idempotent measures on compact semigroups. *Proc. Am. Math. Soc.* **13**, 442–446 (1962) 98
49. Crépel, P.: Marches aléatoires sur le groupe des déplacements du plan. *C. R. Acad. Sci. Paris, Série A* **278**, 961–964 (1974) 99
50. Császár, I.: On infinite products of random elements and infinite convolutions of probability distributions on locally compact groups. *Z. Wahrsch. Verw. Gebiete* **5**, 279–295 (1966) 100
51. Császár, I.: On the weak* continuity of convolution in a convolution algebra over an arbitrary topological group. *Studia Sci. Math. Hungar.* **6**, 27–40 (1971) 101
52. Dani, S.G., Shah, R.: Collapsible probability measures and concentration functions on Lie groups. *Math. Proc. Camb. Philos. Soc.* **122**(1), 105–113 (1997) 102
53. Darling, R.W.R., Mukherjea, A.: Stochastic flows on a countable set: convergence in distribution. *J. Theoret. Probab.* **1**(2), 121–147 (1988) 103
54. Darling, R.W.R., Mukherjea, A.: Probability measures on semigroups of nonnegative matrices. In: K.H. Hofmann, J.D. Lawson, J.S. Pym (eds.) *The Analytical and Topological Theory of Semigroups. de Gruyter Exp. Math.*, vol. 1, pp. 361–377. de Gruyter, Berlin (1990) 104
55. Darling, R.W.R., Mukherjea, A.: Discrete time voter models: a class of stochastic automata. In: H. Heyer (ed.) *Probability Measures on Groups, X* (Oberwolfach, 1990), pp. 83–94. Plenum, New York (1991) 105
56. Dekking, F.M.: On transience and recurrence of generalized random walks. *Z. Wahrsch. Verw. Gebiete* **61**, 459–465 (1982) 106
57. Derriennic, Y.: Lois “zéro ou deux” pour les processus de Markov. Applications aux marches aléatoires. *Ann. Inst. H. Poincaré Sect. B (N.S.)* **12**(2), 111–129 (1976) 107
58. Derriennic, Y.: Sur le théorème de point fixe de Brunel et le théorème de Choquet-Deny. *Ann. Sci. Univ. Clermont-Ferrand II Probab. Appl.* **87**(4), 107–111 (1985) 108
59. Derriennic, Y., Guivarc’h, Y.: Théorème de renouvellement pour les groupes non moyennables. *C. R. Acad. Sci. Paris Sér. A–B* **277**, A613–A615 (1973) 109
60. Derriennic, Y., Lin, M.: Convergence of iterates of averages of certain operator representations and of convolution powers. *J. Funct. Anal.* **85**(1), 86–102 (1989) 110
61. Dudley, R.M.: Pathological topologies and random walks on abelian groups. *Proc. Am. Math. Soc.* **15**, 231–238 (1964) 111
62. Durbin, J.: *Modern Algebra*. Wiley, New York (1979) 112
63. Edgar, G.A.: *Measure, Topology, and Fractal Geometry*. Springer, New York (1990) 113
64. Edgar, G.A.: *Integral, Probability and Fractal Measures*. Springer, New York (1998) 114
65. Ellis, R.: A note on the continuity of the inverse. *Proc. Am. Math. Soc.* **8**, 372–373 (1957) 115
66. Ellis, R.: Locally compact transformation groups. *Duke Math. J.* **24**, 119–125 (1957) 116
67. Elton, J.H.: An ergodic theorem for iterated maps. *Ergodic Theor. Dynam. Syst.* **7**(4), 481–488 (1987) 117
68. Flor, P.: On groups of non-negative matrices. *Comp. Math.* **21**, 376–382 (1969) 118
69. Furstenberg, H.: Noncommuting random products. *Trans. Am. Math. Soc.* **108**, 377–428 (1963) 119
70. Furstenberg, H., Kesten, H.: Products of random matrices. *Ann. Math. Stat.* **31**, 457–469 (1960) 120
71. Furstenberg, H., Kifer, Y.: Random matrix products and measures on projective spaces. *Israel J. Math.* **46**(1–2), 12–32 (1983) 121
72. Galmarino, A.R.: The equivalence theorem for compositions of independent random elements on locally compact groups and homogeneous spaces. *Z. Wahrsch. Verw. Gebiete* **7**, 29–42 (1967) 122
73. Gard, J.R., Mukherjea, A.: On the convolution iterates of a probability measure. *Semigroup Forum* **10**(2), 171–184 (1975) 123

74. Gaubert, S., Plus, M.: Methods and applications of $(\max, +)$ linear algebra. In: Reischuk, 149
Morvan (eds.) STACS '97 Proceedings. *Lecture Notes in Computer Science*, vol. 1200, 150
pp. 261–282. Springer, Berlin (1997) 151
75. Gibert, S., Mukherjea, A.: Results in semigroups in the context of nonhomogeneous Markov 152
chains: tail idempotents and their structure for infinite-dimensional nonnegative matrices. In: 153
Probability measures on groups, VII (Oberwolfach, 1983). *Lecture Notes in Math.*, vol. 1064, 154
pp. 147–160. Springer, Berlin (1984) 155
76. Glicksberg, I.: Convolution semigroups of measures. *Pacific J. Math.* **9**, 51–67 (1959) 156
77. Gol'dsheid, I.Y., Margulis, G.A.: Lyapunov exponents of a product of random matrices. *Us- 157
pekhi Mat. Nauk* **44**(5), 13–59 (1989) 158
78. Grenander, U.: Probabilities on Algebraic Structures. Wiley, New York (1963) 159
79. Grincevičius, A.K.: The continuity of the distribution of a certain sum of dependent variables 160
that is connected with independent walks on lines. *Teor. Verojatnost. i Primenen.* **19**, 163–168 161
(1974) 162
80. Guivarc'h, Y.: Application d'un théorème limite local à la transience et à la récurrence des 163
marches de Markov. In: *Théorie du potentiel* (Orsay 1983), pp. 301–332. *Lecture Notes in* 164
Math., vol. 1096. Springer, Berlin (1984) 165
81. Guivarc'h, Y.: Limit theorems for random walks and products of random matrices. In: S.G. 166
Dani, P. Graczyk (eds.) *Probability Measures on Groups: Recent Directions and Trends*, 167
pp. 255–330. Tata Institute of Fundamental Research, Mumbai (2006) 168
82. Guivarc'h, Y., Keane, M.: Marches aléatoires sur les groupes nilpotents. *C. R. Acad. Sc. Paris,* 169
Sér. A–B **272**, A1462–A1463 (1971) 170
83. Guivarc'h, Y., Keane, M.: Transience of random walks on nilpotent groups. In: *Papers from* 171
the Open House for Probabilists. Various Publ. Ser., vol. 21, pp. 92–100. Aarhus Universitet, 172
Mat. Inst., Denmark (1972) 173
84. Guivarc'h, Y., Raja, C.R.E.: Recurrence and Ergodicity of Random Walks on Linear Groups 174
and Homogeneous Spaces (to appear) 175
85. Guivarc'h, Y., Raugi, A.: Frontière de Furstenberg, propriétés de contraction et théorèmes de 176
convergence. *Z. Wahrsch. Verw. Gebiete* **69**(2), 187–242 (1985) 177
86. Guivarc'h, Y., Raugi, A.: Products of random matrices: convergence theorems. In: *Random* 178
Matrices and their Applications (Brunswick, Maine, 1984). *Contemp. Math.*, vol. 50, 179
pp. 31–54. American Mathematical Society, RI (1986) 180
87. Guivarc'h, Y., Keane, M., Roynette, B.: Marches aléatoires sur les groupes de Lie. *Lecture* 181
Notes in Mathematics, vol. 624. Springer, Berlin (1977) 182
88. Heble, M., Rosenblatt, M.: Idempotent measures on a compact topological semigroup. *Proc.* 183
Am. Math. Soc. **14**, 177–184 (1963) 184
89. Hewitt, E., Ross, K.A.: Abstract harmonic analysis, vol. I: Structure of topological groups. 185
Integration Theory, Group Representations. Die Grundlehren der mathematischen Wis- 186
senschaften, Bd. 115. Academic, NY (1963) 187
90. Heyer, H.: Über Haarsche Masse auf lokalkompakten Gruppen. *Arch. Math. (Basel)* **17**, 188
347–351 (1966) 189
91. Heyer, H.: Probabilistic Characterization of Certain Classes of Locally Compact Groups, 190
pp. 315–355. Academic, London (1975) 191
92. Heyer, H.: Probability measures on locally compact groups. *Ergebnisse der Mathematik und* 192
ihrer Grenzgebiete, vol. 94. Springer, Berlin (1977) 193
93. Hofmann, K.H., Mostert, P.S.: Elements of Compact Semigroups. Charles E. Merrill Books, 194
Columbus (1966) 195
94. Hofmann, K.H., Mukherjea, A.: On the density of the image of the exponential function. 196
Math. Ann. **234**(3), 263–273 (1978) 197
95. Hofmann, K.H., Mukherjea, A.: Concentration functions and a class of noncompact groups. 198
Math. Ann. **256**(4), 535–548 (1981) 199
96. Högnäs, G.: A note on random walks on a compact semigroup. *Acta Acad. Aboensis Ser. B* 200
33(10), 1–8 (1973) 201
97. Högnäs, G.: An ergodic random walk on a compact semigroup. *Acta Acad. Aboensis Ser. B* **33** 202

AQ3

AQ4

98. Högnäs, G.: Marches aléatoires sur un demi-groupe compact. *Ann. Inst. Henri Poincaré* **10**, 203–204 (1974)
99. Högnäs, G.: Remarques sur les marches aléatoires dans un demi-groupe compact avec un idéal compact ayant une probabilité positive. *Ann. Inst. Henri Poincaré* **10**, 345–354 (1974)
100. Högnäs, G.: On random walks with continuous components. *Tech. rep.* (1976/1977)
101. Högnäs, G.: Random semigroup acts on a finite set. *J. Austral. Math. Soc. Ser. A* **23**, 481–498 (1977)
102. Högnäs, G.: A note on the product of random elements of a semigroup. *Monatsh. Math.* **85**, 317–321 (1978)
103. Högnäs, G.: On random walks with continuous components. *Semigroup Forum* **17**, 75–93 (1979)
104. Högnäs, G.: On products of random projections. *Acta Acad. Åbo. Ser. B* **44**(5), 18 (1984)
105. Högnäs, G.: Recurrence and transience of markov random walks. *Acta Acad. Aboensis Ser. B* **44**(4), 5 (1984)
106. Högnäs, G.: Markov random walks on groups. *Math. Scand.* **58**, 35–45 (1986)
107. Högnäs, G.: A note on products of random matrices. *Statist. Probab. Lett.* **5**(5), 367–370 (1987)
108. Högnäs, G.: A note on the semigroup of analytic mappings with a common fixed point. In: *Probability Measures on Groups, IX (Oberwolfach, 1988). Lecture Notes in Math.*, vol. 1379, pp. 135–140. Springer, Berlin (1989)
109. Högnäs, G.: Invariant measures and random walks on the semigroup of matrices. In: H. Langer (ed.) *Markov Processes and Control Theory (Gaußig, 1988). Math. Res.*, vol. 54, pp. 77–84. Akademie, Berlin (1989)
110. Högnäs, G.: Sequences of random transformations. *Theor. Probab. Appl.* **37**, 54–60 (1992)
111. Högnäs, G., Mukherjea, A.: Recurrent random walks and invariant measures on semigroups of $n \times n$ matrices. *Math. Z.* **173**(1), 69–94 (1980)
112. Högnäs, G., Mukherjea, A.: A mixed random walk on nonnegative matrices: a law of large numbers. *J. Theoret. Probab.* **8**(4), 973–990 (1995)
113. Högnäs, G., Mukherjea, A.: Maximal homomorphic group image and convergence of convolution sequences on a semigroup. *J. Theoret. Probab.* **16**(4), 847–854 (2003)
114. Högnäs, G., Mukherjea, A.: Products of i. i. d. random nonnegative matrices: Their skeletons and convergence in distribution. *Sankhya (The Indian J. of Statistics)* **67**(4), 615–633 (2005)
115. Husain, T.: *Introduction to Topological Groups*. W. B. Saunders Company, London (1966)
116. Itô, K., Nisio, M.: On the convergence of sums of independent Banach space valued random variables. *Osaka J. Math.* **5**, 35–48 (1968)
117. Jaworski, W.: The density of the image of the exponential function and spacious locally compact groups. *J. Lie Theor.* **5**(1), 129–134 (1995)
118. Jaworski, W.: Contractive automorphisms of locally compact groups and the concentration function problem. *J. Theoret. Probab.* **10**(4), 967–989 (1997)
119. Jaworski, W.: On shifted convolution powers and concentration functions in locally compact groups. In: *Probability on Algebraic Structures (Gainesville, FL, 1999). Contemp. Math.*, vol. 261, pp. 23–41. American Mathematical Society, RI (2000). English summary
120. Jaworski, W., Rosenblatt, J., Willis, G.: Concentration functions in locally compact groups. *Math. Ann.* **305**(4), 673–691 (1996)
121. Kari, J.: Synchronizing finite automata on Eulerian digraphs. *Theoret. Comp. Sci.* **295**, 223–232 (2003)
122. Karlin, S., Taylor, H.M.: *A first course in stochastic processes*, 2nd edn. Academic, New York (1975)
123. Kawada, Y., Itô, K.: On the probability distribution on a compact group. I. *Proc. Phys. Math. Soc. Jpn.* **22**(3), 977–998 (1940)
124. Kelley, J.H.: *General Topology*. Van Nostrand, Princeton (1955)
125. Kelley, J.L.: Averaging operators on $C_\infty(X)$. *Illinois J. Math.* **2**, 214–223 (1958)
126. Kesten, H.: The Martin boundary of recurrent random walks on countable groups. In: *Proc. 5th Berkeley Symp. on Mathematical Statistics and Probability, II*, p. 51 (1967)

127. Kesten, H.: Random difference equations and renewal theory for products of random matrices. *Acta Math.* **131**, 207–248 (1973) 257
128. Kesten, H., Spitzer, F.: Convergence in distribution of products of random matrices. *Z. Wahrsch. Verw. Gebiete* **67**(4), 363–386 (1984) 259
129. Key, E.: Computable examples of the maximal Lyapunov exponent. *Probab. Theor. Relat. Field* **75**(1), 97–107 (1987) 261
130. Key, E.S.: Lower bounds for the maximal Lyapunov exponent. *J. Theoret. Probab.* **3**(3), 477–488 (1990) 262
131. Kingman, J.F.C.: Subadditive ergodic theory. *Ann. Probab.* **1**, 883–909 (1973); With discussion by D. L. Burkholder, Daryl Daley, H. Kesten, P. Ney, Frank Spitzer and J. M. Hammersley, and a reply by the author 265
132. Kloss, B.M.: Probability distributions on bicomact topological groups. *Theor. Probab. Appl.* **4**, 237–270 (1959) 266
133. Kloss, B.M.: Limit distributions on bicomact Abelian groups. *Teor. Veroyatnost. i Primenen* **6**, 392–421 (1961) 267
134. Koutsky, Z.: Einige Eigenschaften der modulo k addierten Markowschen Ketten. In: *Transactions of the 2nd Prague Conference on Information Theory, Random Functions and Statistical Decision Theory*, pp. 263–278. Czechoslovak Academy of Sciences, Prague (1959) 272
135. Larisse, J.: Marches au hasard sur les demi-groupes discrets, i, ii, iii. *Ann. Inst. Henri Poincaré* **8**, 107–125, 127–173, 229–240 (1972) 273
136. Ljapin, E.S.: Semigroups. *Translations of Mathematical Monographs*, vol. 3, American Mathematical Society, RI (1974) 274
137. Lo, C.C., Mukherjea, A.: Convergence in distribution of products of $d \times d$ random matrices. *J. Math. Anal. Appl.* **162**(1), 71–91 (1991) 275
138. Loynes, R.: Probability distributions on a topological group. *Z. Wahrsch. Verw. Gebiete* **5**, 446–455 (1966) 276
139. Maksimov, V.M.: Necessary and sufficient conditions for the convergence of the convolution of non-identical distributions on a finite group. *Teor. Veroyatnost. i Primenen.* **13**, 295–307 (1968) 277
140. Maksimov, V.M.: Convolution-convergent sequences of measures on compact groups. *Teor. Veroyatnost. i Primenen.* **16**, 52–66 (1971) 278
141. Maksimov, V.M.: A generalized Bernoulli scheme and its limit distributions. *Teor. Veroyatnost. i Primenen.* **18**, 547–556 (1973) 279
142. Martin-Löf, P.: Probability theory on discrete semigroups. *Z. Wahrsch. Verw. Gebiete* **4**, 78–102 (1965) 280
143. Meyer, C.D.: The role of the group generalized inverse in the theory of finite Markov chains. *SIAM Rev.* **17**, 443–464 (1975) 281
144. Mindlin, D.S.: Convolutions of random measures on a compact topological semigroup. *J. Theoret. Probab.* **3**(2), 181–197 (1990) 282
145. Mindlin, D.S.: The rate of convergence of convolutions of random measures on a compact group. *Teor. Veroyatnost. i Primenen.* **35**(2), 361–367 (1990) 283
146. Mindlin, D.S., Rubshtein, B.A.: Convolutions of random measures on a compact group. *Teor. Veroyatnost. i Primenen.* **33**(2), 376–379 (1988) 284
147. Mukherjea, A.: On the convolution equation $P = P * Q$ of Choquet and Deny for probability measures on semigroups. *Proc. Am. Math. Soc.* **32**, 457–463 (1972) 285
148. Mukherjea, A.: On the equation $P(B) = \int P(Bx^{-1})P(dx)$ for infinite P . *J. Lond. Math. Soc.* **6**(2), 224–230 (1973) 286
149. Mukherjea, A.: Limit theorems for probability measures on non-compact groups and semigroups. *Z. Wahrsch. Verw. Gebiete* **33**(4), 273–284 (1975/1976) 287
150. Mukherjea, A.: Limit theorems for convolution iterates of a probability measure on completely simple or compact semigroups. *Trans. Am. Math. Soc.* **225**, 355–370 (1977) 288
151. Mukherjea, A.: Limit theorems: stochastic matrices, ergodic Markov chains, and measures on semigroups. In: *Probabilistic Analysis and Related Topics*, vol. 2, pp. 143–203. Academic, New York (1979) 289

152. Mukherjea, A.: Completely simple semigroups of matrices. *Semigroup Forum* **33**, 405–429 (1986) 311
153. Mukherjea, A.: Convergence in distribution of products of random matrices: a semigroup approach. *Trans. Am. Math. Soc.* **303**(1), 395–411 (1987) 313
154. Mukherjea, A.: Convolution products of nonidentical distributions on a compact abelian semigroup. In: *Probability Measures on Groups, IX* (Oberwolfach, 1988). *Lecture Notes in Math.*, vol. 1379, pp. 217–241. Springer, Berlin (1989) 315
155. Mukherjea, A.: The role of nonnegative idempotent matrices in certain problems in probability. In: *Matrix Theory and Applications* (Phoenix, AZ, 1989). *Proc. Sympos. Appl. Math.*, vol. 40, pp. 199–232. American Mathematical Society, RI (1990) 318
156. Mukherjea, A.: Multiplicative semigroups of infinite-dimensional matrices. *Semigroup Forum* **33**, 115–122 (1991) 321
157. Mukherjea, A.: Semigroups, attractors, and products of random matrices. In: H. Heyer (ed.) *Probability Measures on Groups, X* (Oberwolfach, 1990), pp. 303–313. Plenum, New York (1991) 323
158. Mukherjea, A.: Tightness of products of i.i.d. random matrices. *Probab. Theor. Relat. Field* **87**(3), 389–401 (1991) 326
159. Mukherjea, A.: Convergence in distribution of a Markov process generated by i.i.d. random matrices. In: M.A. Pinsky, V. Wihstutz (eds.) *Diffusion Processes and Related Problems in Analysis*, vol. II (Charlotte, NC, 1990). *Progr. Probab.*, vol. 27, pp. 171–200. Birkhäuser Boston, MA (1992) 328
160. Mukherjea, A.: Recurrent random walks in nonnegative matrices: attractors of certain iterated function systems. *Probab. Theor. Relat. Field* **91**(3–4), 297–306 (1992) 332
161. Mukherjea, A.: Recurrent random walks in nonnegative matrices. II. *Probab. Theor. Relat. Field* **96**(4), 415–434 (1993) 333
162. Mukherjea, A.: Some remarks on products of random affine maps on $(\mathbf{R}^+)^d$. In: Doeblin and modern probability (Blaubeuren, 1991). *Contemp. Math.*, vol. 149, pp. 321–330. American Mathematical Society, RI (1993) 334
163. Mukherjea, A.: Tightness of products of i.i.d. random matrices. II. *Ann. Probab.* **22**(4), 2223–2234 (1994) 335
164. Mukherjea, A.: Topics in products of random matrices. *Tata Institute of Fundamental Research Lectures on Mathematics*, vol. 87. Narosa Publishing House, New Delhi (2000) 336
165. Mukherjea, A., Nakassis, A.: On the limit of the convolution iterates of a probability measure on $n \times n$ stochastic matrices. *J. Math. Anal. Appl.* **60**(2), 392–397 (1977) 337
166. Mukherjea, A., Pothoven, K.: Real and functional analysis. *Mathematical Concepts and Methods in Science and Engineering*, vol. 6. Plenum, New York (1978) 338
167. Mukherjea, A., Pothoven, K.: Real and functional analysis. Part A. *Mathematical Concepts and Methods in Science and Engineering*, vol. 27, 2nd edn. Plenum, New York (1984). Real analysis 339
168. Mukherjea, A., Pothoven, K.: Real and functional analysis. Part B. *Mathematical Concepts and Methods in Science and Engineering*, vol. 28, 2nd edn. Plenum, New York (1986). Functional analysis 350
169. Mukherjea, A., Restrepo, R.: Upper packing dimension of a measure and the limit distribution of products of i. i. d. stochastic matrices. *Proc. Indian Acad. Sci. (Math. Sci.)* **119**(5), 669–677 (2009) 351
170. Mukherjea, A., Saff, E.B.: Behavior of convolution sequences of a family of probability measures on $[0, \infty)$. *Indiana Univ. Math. J.* **24**, 221–226 (1974/1975) 352
171. Mukherjea, A., Sun, T.C.: Convergence of products of independent random variables with values in a discrete semigroup. *Z. Wahrsch. Verw. Gebiete* **46**(2), 227–236 (1978/1979) 353
172. Mukherjea, A., Tserpes, N.A.: Idempotent measures on locally compact semigroups. *Proc. Am. Math. Soc.* **29**, 143–150 (1971) 354
173. Mukherjea, A., Tserpes, N.A.: A problem on r^* -invariant measures on locally compact semigroups. *Indiana Univ. Math. J.* **21**, 973–977 (1971/1972) 355
174. Mukherjea, A., Tserpes, N.A.: Invariant measures and the converse of Haar's theorem on semitopological semigroups. *Pacific J. Math.* **44**, 251–262 (1973) 356

175. Mukherjea, A., Tserpes, N.A.: Bilateral random walks on compact semigroups. *Proc. Am. Math. Soc.* **47**, 457–466 (1975) 366
176. Mukherjea, A., Tserpes, N.A.: Measures on topological semigroups: convolution products and random walks. *Lecture Notes in Mathematics*, vol. 547. Springer, Berlin (1976) 368
177. Mukherjea, A., Sun, T.C., Tserpes, N.A.: Random walks on compact semigroups. *Proc. Am. Math. Soc.* **39**, 599–605 (1973) 370
178. Muthsam, H.: Über die Summe Markoffscher Ketten auf Halbgruppen. *Monatsh. Math.* **76**, 43–54 (1972) 372
179. Nakassis, A.: Limit behavior of the convolution iterates of a probability measure on a semigroup of matrices. *J. Math. Anal. Appl.* **70**(2), 337–347 (1979) 374
180. Niemi, S., Nummelin, E.: Central limit theorems for Markov random walks. *Soc. Sci. Fenn. Comment. Phys. Math.* **54** (1982) 376
181. Nummelin, E.: General Irreducible Markov Chains and Nonnegative Operators. Cambridge University Press, Cambridge (1984) 377
182. Oseledec, V.I.: A multiplicative ergodic theorem. Characteristic Ljapunov exponents of dynamical systems. *Trudy Moskov. Mat. Obšč.* **19**, 179–210 (1968) 380
183. Owen, W.S.J.: A Characterization of the Semigroup of 2×2 Real Matrices. PhD Thesis, University of Georgia, Athens (1972) 382
184. Owen, W.S.J.: The Rees theorem for locally compact semigroups. *Semigroup Forum* **6**, 133–152 (1973) 384
185. Paalman-de Miranda, A.B.: Topological Semigroups, 2nd edn. Mathematisch Centrum, Amsterdam (1970) 386
186. Parthasarathy, K.R.: Probability measures on metric spaces. *Probability and Mathematical Statistics*, vol. 3. Academic, New York (1967) 388
187. Paterson, A.L.T.: Invariant measure semigroups. *Proc. Lond. Math. Soc.* (3) **35**(2), 313–332 (1977) 389
188. Paterson, A.L.T.: Amenability. American Mathematical Society, RI (1988) 390
189. Pincus, S.: Strong laws of large numbers for products of random matrices. *Trans. Am. Math. Soc.* **287**(1), 65–89 (1985) 391
190. Pólya, G.: Über eine Aufgabe der Wahrscheinlichkeitsrechnung betreffend die Irrfahrt im Straßennetz. *Math. Ann.* **84**, 149–160 (1921) 392
191. Pym, J.S.: Idempotent measures on semigroups. *Pacific J. Math.* **12**, 685–698 (1962) 393
192. Pym, J.S.: Idempotent probability measures on compact semitopological semigroups. *Proc. Am. Math. Soc.* **21**, 499–501 (1969) 394
193. Radjavi, H., Rosenthal, P.: Simultaneous Triangularization. Universitext. Springer, Berlin (2000) 395
194. Raghunathan, M.S.: A proof of Oseledec's multiplicative ergodic theorem. *Israel J. Math.* **32**(4), 356–362 (1979) 396
195. Raugi, A.: Théorie spectral d'un opérateur de transition sur un espace métrique compact. *Ann. Inst. Henri Poincaré* **28**, 281–309 (1992) 397
196. Rees, D.: On semigroups. *Proc. Camb. Philos. Soc.* **36**, 387–400 (1940) 398
197. Retzlaff, T.M.: Rate of decay of concentration functions on locally compact groups. *J. Theoret. Probab.* **16**, 391–399 (2003) 399
198. Retzlaff, T.M.: Decay of concentration functions for adapted probabilities on discrete groups. *J. Theoret. Probab.* **17**, 1031–1040 (2004) 400
199. Revuz, D.: Markov Chains. North Holland, Amsterdam (1975) 401
200. Rosen, W.G.: On invariant means over compact semigroups. *Proc. Am. Math. Soc.* **7**, 1076–1082 (1956) 402
201. Rosenblatt, M.: Limits of convolution sequences of measures on a compact topological semigroup. *J. Math. Mech.* **9**, 293–305 (1960) 403
202. Rosenblatt, M.: Equicontinuous Markov operators. *Teor. Veroyatnost. i Primenen.* **9**, 205–222 (1964) 404
203. Rosenblatt, M.: Products of independent identically distributed stochastic matrices. *J. Math. Anal. Appl.* **11**, 1–10 (1965) 405

204. Rosenblatt, M.: Markov processes. Structure and asymptotic behavior. Die Grundlehren der 420
mathematischen Wissenschaften, Band 184. Springer, New York (1971) 421
205. Rosenblatt, M.: Invariant and subinvariant measures of transition probability functions acting 422
on continuous functions. *Z. Wahrsch. Verw. Gebiete* **25**, 209–221 (1973) 423
206. Rosenblatt, M.: Invariant measures and growth conditions. *Trans. Am. Math. Soc.* **193**, 33–53 424
(1974) 425
207. Roynette, B.: Marches aléatoires sur les groupes de Lie. In: *Ecole d'Été de Probabilités de* 426
Saint-Flour VII - 1977. Lecture Notes in Mathematics, vol. 678. Springer, Berlin (1978) 427
208. Ruzsa, I.Z.: Infinite Convolution of Distributions on Discrete Commutative Semigroups, 428
pp. 365–376. Plenum, New York (1991) 429
209. Ruzsa, I.Z., Székely, G.J.: Algebraic probability theory. *Wiley Series in Probability and Math-* 430
ematical Statistics: Probability and Mathematical Statistics. Wiley, Chichester (1988) 431
210. Sazonov, V.V., Tutubalin, V.N.: Probability distributions on topological groups. *Teor. Vero-* 432
jatnost. i Primenen. **11**, 3–55 (1966) 433
211. Schmetterer, L.: Über die Summe Markov'scher Ketten auf Halbgruppen. *Monatsh. Math.* 434
71, 223–230 (1967) 435
212. Schwartz, S.: Convolution semigroup of measures on compact noncommutative semigroups. 436
Czech. Math. J. **14**, 95–115 (1964) 437
213. Simon, M.: Automata Theory. World Scientific, Singapore (1999) 438
214. Stoll, R.R.: Homomorphisms of a semigroup onto a group. *Am. J. Math.* **73**, 475–481 (1951) 439
215. Stromberg, K.: Probabilities on a compact group. *Trans. Am. Math. Soc.* **94**, 295–309 (1960) 440
216. Sun, T.C.: Limits of convolutions of probability measures on the set of 2×2 stochastic ma- 441
trices. *Bull. Inst. Math. Acad. Sinica* **3**(2), 235–248 (1975) 442
217. Sun, T.C.: Random walks on semigroups. In: J.E. Cohen, H. Kesten, C.M. Newman (eds.) 443
Random Matrices and their Applications (Brunswick, Maine, 1984). *Contemp. Math.*, vol. 50, 444
pp. 221–239. American Mathematical Society, RI (1986) 445
218. Sun, T.C., Mukherjea, A., Tserpes, N.A.: On recurrent random walks on semigroups. *Trans.* 446
Am. Math. Soc. **185**, 213–227 (1973) 447
219. Suschkewitsch, A.K.: Über die endlichen Gruppen ohne das Gesetz der eindeutigen 448
Umkehrbarkeit. *Math. Ann.* **99**, 30–50 (1928) 449
220. Sz.-Nagy, B., Foias, C.: Harmonic Analysis of Operators on Hilbert Space. North-Holland, 450
Amsterdam (1970) 451
221. Székely, G.J., Zeng, W.B.: The Choquet-Deny convolution equation $\mu = \mu * \sigma$ for probability 452
measures on abelian semigroups. *J. Theoret. Probab.* **3**(2), 361–365 (1990) 453
222. The New Encyclopaedia Britannica, 15th edn. The Encyclopaedia Britannica, Chicago (1982) 454
223. Tortrat, A.: Lois de probabilité sur un espace topologique complètement régulier et produits 455
infinis à termes indépendants dans un groupe topologique. *Ann. Inst. H. Poincaré Sect. B* **1**, 456
217–237 (1964/1965) 457
224. Tortrat, A.: Lois de probabilité dans les semi-groupes topologiques complètement réguliers. 458
C. R. Acad. Sci. Paris, Sér. A **261**, 3941–3944 (1965) 459
225. Tortrat, A.: Lois tendues μ sur un demi-groupe topologique complètement simple X . 460
Z. Wahrsch. Verw. Gebiete **6**, 145–160 (1966) 461
226. Trahtman, A.N.: On the road coloring problem. *Israel J. Math.* **172**, 51–60 (2009) 462
227. Tserpes, N., Sun, T.C.: Idempotent probability measures on locally compact abelian semi- 463
groups. *J. Math. Mech.* **19**, 1113–1116 (1969/1970) 464
228. Tserpes, N.A., Kartsatos, A.: Measure semiinvariants sur un semigroups localement compact. 465
C. R. Acad. Sci. Paris. Ser. A **267**, 507–509 (1968) 466
229. Tuominen, P., Tweedie, R.L.: Markov chains with continuous components. *Proc. Lond. Math.* 467
Soc. **38**(3), 89–114 (1979) 468
230. Urbanik, K.: On the limiting probability distribution on a compact topological group. *Fund.* 469
Math. **44**, 253–261 (1957) 470
231. Van Assche, W.: Products of 2×2 stochastic matrices with random entries. *J. Appl. Probab.* 471
23(4), 1019–1024 (1986) 472
232. Varopoulos, N.T.: Théorie du potentiel sur des groupes et des variétés. *C. R. Acad. Sc. Paris,* 473
Sér. I **302**, 203–205 (1986) 474

233. Varopoulos, N.T., Saloff-Coste, L., Coulhon, T.: Analysis and Geometry on Groups. 475
Cambridge University Press, Cambridge (1992) 476
234. Wallace, A.D.: The Rees-Suschkewitsch structure theorem for compact simple semigroups. 477
Proc. Nat. Acad. Sci. **42**, 430–432 (1956) 478
235. Wendel, J.G.: Haar measure and the semigroup of measures on a compact group. Proc. Am. 479
Math. Soc. **5**, 923–929 (1954) 480
236. Williamson, J.H.: Harmonic analysis on semigroups. J. Lond. Math. Soc. **42**(2), 1–41 (1967) 481
237. Willis, G.: The structure of totally disconnected, locally compact groups. Math. Ann. **300**(2), 482
341–363 (1994) 483
238. Willis, G.: Totally disconnected groups and proofs of conjectures of Hofmann and Mukherjee. 484
Bull. Austral. Math. Soc. **51**(3), 489–494 (1995) 485
239. Willis, G.A.: Probability measures on groups and some related ideals in group algebras. 486
J. Funct. Anal. **92**(1), 202–263 (1990) 487
240. Willis, G.A.: An extension of a non-commutative Choquet-Deny theorem. Proc. Am. Math. 488
Soc. **128**(1), 111–118 (2000) 489
241. Woess, W.: Random Walks on Infinite Graphs and Groups. Cambridge University Press, 490
Cambridge (2000) 491
242. Wolff, M.: Über Produkte abhängiger zufälliger Veränderlicher mit Werten in einer kompak- 492
ten Halbgruppe. Z. Wahrsch. Verw. Gebiete **35**, 253–264 (1976) 493

AUTHOR QUERIES

- AQ1. The references [1, 13, 15, 23, 27, 28, 30, 35, 36, 45, 46, 51, 52, 54, 55, 63, 73, 76, 86, 91, 94, 105, 114, 116, 117, 119, 139, 141, 144–146, 162, 165, 166, 170, 175, 179, 200, 201, 203, 209, 212, 215, 216, 227, 228, 230, 236–240] are not cited in the text. please provide citation or delete from the list.
- AQ2. Please update the reference [32].
- AQ3. Please update the reference [84].
- AQ4. Please provide complete details for the reference [97].
- AQ5. Please provide complete details for the reference [180].

Uncorrected Proof

Index

1

A

Abelian group, 32, 50, 81, 126, 127, 137, 169, 210, 222, 240
 Abelian semigroup, 2, 26, 37, 88, 147, 151, 154–156, 168, 194, 222
 discrete, 63, 156, 159
 Absolutely continuous component, 129–131
 Absolutely continuous measure, 131, 139, 176, 205
 Act, 6, 53, 54
 Adapted, 24, 128, 198, 204, 206, 207, 397–399, 401, 402, 408–412
 Affine contraction, 214, 222, 250
 Affine maps, 251, 254, 260, 264, 265, 276, 313, 344, 353, 356, 358
 Almost sure convergence, 150
 Antihomomorphism, 6, 44, 45, 58, 59
 Antiisomorphism, 6
 Attractor, 276–283, 356
 Average contractivity condition, 280

B

Basis of a matrix, 38, 40, 42, 45–49, 55, 57, 59
 Bernoulli distribution, 248
 Bicancellative, 72, 78, 79
 Bicyclic semigroup, 20–22, 59, 197
 Bilateral, 173, 175
 ideal, 2, 26, 180, 181, 220, 224
 identity, 209
 Bilateral random walk, 173, 183, 187, 208, 211, 212, 220, 224, 229, 230, 236, 238, 247
 Bilateral walk, 172, 178, 180, 183–185, 188, 190, 193, 198, 199, 209, 212, 220, 221, 223, 226–228, 231, 232, 234, 242, 243
 Birkhoff's theorem, 48, 164, 165

Bistochastic matrix, 48, 49, 59, 61, 143, 164, 165, 222, 250
 Bistochastic stochastic matrix, 48

C

Cancellative semigroup, 5, 32, 95, 131, 273
 Cell, 9
 group, 11, 18, 19, 31, 32, 35
 null, 11, 18, 19, 30, 31, 35
 C-group, 128
 Chain, 48, 175, 192, 195, 204, 208, 214, 215, 217–219, 239
 Circulant matrix, 5, 50, 137
 Code, 248
 Coding theory, 60, 248
 Coloring semigroup, 51
 Communicate, 177, 180, 181, 185, 209, 215, 220, 240, 243
 Communication relation, 171, 177, 215, 219, 220
 Commutative semigroup. *See* Abelian semigroup
 Commute, 2
 Compact group, 32, 33, 49, 63, 74–77, 82, 87, 88, 90, 94, 95, 106, 113, 115, 119, 127, 129–131, 133, 136, 139, 144, 147, 160, 198–219, 232, 238, 239, 241, 244, 249, 286, 287, 289–291, 293–295, 297, 298, 303, 317, 332, 341, 357, 362, 397
 connected, 130, 138, 140
 of nonnegative matrices, 42, 45, 46, 61
 Compact group element, 32
 Compact group factor, 33, 87, 88, 90, 94, 106, 113, 119, 241, 286, 297, 303, 332, 341
 Compactness condition, 33, 50, 99, 102, 103, 105, 106, 115, 116, 250

Compact semigroup, 26–29, 32, 63, 100, 141, 152, 219–242, 246–250, 279, 341, 358
 abelian, 26, 50, 63, 81, 88, 126, 137, 147, 151, 154, 222, 240
 Completely simple minimal ideal. *See* Kernel
 Completely 0-simple semigroup, 17–21, 28, 30, 33, 60, 246
 Completely simple (sub)semigroup, 145, 241
 of nonnegative matrices, 316
 Complete set of representatives, 3, 4
 Component of a transition probability operator, 173, 176, 177, 179, 211, 215, 220, 224, 229, 230, 238, 240, 249, 250
 continuous, 176, 179
 equivalent, 173, 176
 nontrivial, 176, 238
 Composition of functions, 61, 182
 Composition of relations, 3, 5
 Concentration functions, 126, 128, 397–412
 (CL) condition, 251
 (CR) condition, 251
 Congruence, 6, 7, 9, 51
 Conservative element, 177, 209
 Constant mapping, 3, 12, 222
 Continuous component, 129–131, 176, 204, 206–209, 238, 239, 245, 247, 250
 Continuous measure, 176, 383, 385
 Continuous singular measure, 205, 206, 384, 387
 Contractive coset, 128
 Convergence
 almost sure, 156, 166, 167, 192, 235–237, 253, 282, 322, 325, 331–333, 346, 366
 vague, 66, 84, 99, 105
 to 0, 85, 101, 107–109, 115, 116, 119, 122, 124, 134, 211, 245, 306
 weak, 63, 64, 66, 67, 83–168, 192, 284–333, 345, 358
 of convolution sequences, 122
 of empirical distributions, 192
 of products of affine maps, 356, 358
 of products of nonnegative matrices, 356
 of products of real matrices, 299, 302
 Convolution, 63–68, 72, 83–168, 172, 176, 197, 206, 207, 211, 222, 247, 249, 284, 341, 366, 394, 399, 403, 404, 408
 power, 68, 87, 135, 177, 204, 253, 289, 357, 397
 Convolution sequence, 121, 171, 211, 247, 251, 354

tight, 83, 87–89, 96, 97, 101, 122, 123, 132, 133, 137, 150, 171, 286, 289–291, 293, 296, 305, 306, 309, 310, 312, 315, 354
 weak convergence of, 63, 87, 101, 122, 150, 192, 289, 299
 Cross section of a partition, 8, 9, 16, 20
 Cyclic classes, 184, 185, 192, 250

D

Density, 129–131, 169, 205–207, 209, 210, 217, 248, 365–367, 397, 398, 402–404, 408, 411
 logarithmic, AQ1
 Direct product, 7–11, 17, 22, 27, 126, 128, 194, 204, 251
 Dirichlet density, 365–367
 Discrete group, 63, 126, 138, 156, 159, 194, 247, 249, 397–412
 Discrete measure, 63, 138, 180, 186, 192, 385, 388, 397–400, 402
 Discrete semigroup, 122, 124, 135, 136, 138, 169, 180–197, 213, 220, 241, 247, 248
 abelian, 63, 159
 Distinguished subgroup, 134
 Divisor of zero, 10–12, 19, 27–30, 37
 Doubly stochastic matrix. *See* Bistochastic stochastic matrix

E

Edge coloring, 51
 Effective act, 6
 Ellis's theorem, 24, 26–28, 30, 42, 198, 216
 Embedding, 52, 81 AQ2
 Empirical distributions
 convergence of, 192, 212
 Equicontinuous operator, 229, 230
 Equidistributed random walk, 212
 Equivalent component, 176
 Ergodic, 192, 231, 234, 236–240, 250, 357, 373, 375, 376, 378, 381
 Essential class, 181, 183, 184, 191–194, 198, 199, 209, 210, 212, 228, 239, 240, 242, 243
 Essential element, 177, 179–187, 189, 190, 197–199, 216, 220, 221, 239, 240, 242
 Exponential growth, 249

F

Factor, 2, 7, 11, 15, 20, 22, 30, 171, 172, 175, 185, 190, 192, 196, 213, 224, 241, 245
 group, 123, 124, 189, 193–195, 232, 244–246, 248, 250, 312, 317, 348, 362
 compact, 33, 87, 88, 90, 94, 106, 113, 119, 171, 241, 251, 286, 297, 303, 332, 341
 finite, 124, 131, 189, 193, 195, 248
 left, 224
 middle, 192, 194, 232, 244, 246
 right, 15, 224
 Factor semigroup, 131, 133, 362, 363
 Factor semigroup, 6, 7, 20
 Feller operator, 177, 215
 Finite group factor, 124, 131, 189, 193, 195, 248
 First hitting time, 176
 First return time, 176, 218
 Free subsemigroup, 409–412
 Frobenius norm, 33
 Full transformation semigroup, 3, 6, 8, 20
 Furstenberg-Kifer theorem, 345–352, 358

G

Gamma random variable, 365
 General linear group, 32, 34, 41
 Generating set, 398, 407
 Group, 2, 63, 171, 256, 359, 381, 397–412
 with 0, 10, 12, 18, 19, 29, 31
 abelian, 32, 50, 81, 126, 127, 137, 169, 210, 222, 240
 compact, 32, 33, 45, 49, 63, 74–77, 82, 87, 88, 90, 94, 95, 106, 113, 115, 119, 127, 129–131, 133, 136, 139, 144, 147, 160, 198–219, 232, 238, 239, 241, 244, 249, 286, 287, 289–291, 293–295, 297, 298, 303, 317, 332, 341, 357, 362, 397
 connected, 130, 138, 140, 249
 of nonnegative matrices, 46, 61
 discrete, 63, 126, 138, 156, 159, 194, 247, 249, 397–412
 factor, 131, 133, 362, 363
 general linear, 32, 34, 41
 locally compact, 45, 49, 127, 198–219, 238, 241, 244, 249, 397
 neat, 127
 noncompact, 109, 124–128
 of nonnegative matrices, 42, 45, 46
 permutation, 8, 44, 45, 58, 59, 362–364

recurrent, 194, 195, 204, 211
 strange, 127
 symmetric, 8, 9
 topological, 24, 26–34, 42, 45, 72, 75, 81, 94–96, 128, 151, 179, 198, 204, 213
 transient, 194, 217, 218
 Group cell, 11, 18, 19, 31, 32, 35
 Group factor, 133, 194, 232, 244, 348, 362, 363
 compact, 33, 87, 88, 90, 94, 106, 113, 119, 123, 171, 241, 245, 246, 250, 251, 286, 297, 303, 312, 317, 332, 341
 finite, 124, 131, 189, 193, 195, 248
 Growth, 210, 249, 253, 375, 380, 397, 398, 404, 407, 408
 Guivarc'h-Keane conjecture, 249

H

Haar measure, 76, 78, 79, 129–132, 138, 139, 169, 176, 192, 204–207, 210–212, 228, 232, 239, 245, 249
 unimodular, 72, 81, 211, 397
 Harmonic, 211
 Harris recurrent, 204, 207, 208, 217
 Hitting time, 176, 203
 Homomorphism, 6–10, 14, 17, 33–35, 58, 80, 81, 133, 206, 248, 363

I

Ideal, 1, 75, 179–191, 254–264, 385
 completely simple minimal (*see* Kernel)
 0-minimal, 10, 12, 18
 Idempotent, 2, 64–83, 182, 256, 359
 primitive, 12–14, 16–18, 28–30, 34, 35, 242, 256
 Idempotent matrix, 16, 34, 36, 38, 42, 48, 49, 52, 56, 57, 290, 302, 304, 316, 329
 Idempotent probability measure, 63–83, 87, 88, 96, 115, 137, 143, 241, 293, 329, 330
 Identity, 2–6, 8–14, 18–21, 23, 25–27, 29, 31, 32, 34–36, 40–43, 45, 46, 50, 52–54, 57–59, 75, 87–90, 93, 94, 108, 130–133, 136, 137, 142, 143, 159–161, 163, 177, 194, 196, 198, 200, 201, 204, 206–209, 213, 216, 223, 239, 242, 243, 248, 249, 259, 264, 268, 273, 286, 289, 290, 292, 293, 316, 317, 332, 357, 360, 361, 363, 364, 397, 405–407

AQ4

AQ5

AQ3

Inessential class, 181
 Invariance semigroup, 215–218, 240
 Invariant measure, 63, 192, 211, 224, 233, 235, 236, 240, 246, 253, 333–355, 376, 394. *See also* Probability measure
 I^* , 72, 75, 76, 79, 142, 144
 left, 81
 r^* , 72, 77
 right, 80, 81
 Inverse, 2–4, 14, 19–21, 23, 24, 27, 28, 34, 42, 44, 46, 53, 54, 61, 194, 248, 273, 295, 407, 410
 of a transition probability matrix, 48, 61
 Irreversibility,
 Isomorphism, 6, 8, 14, 28, 41, 42, 362–364
 topological, 27, 79, 81, 87
 Iterated function system, 214, 239, 280

AQ6

J

Jointly continuous multiplication, 24, 30, 33, 35, 168, 220
 Juxtaposition, 1, 20

K

Kernel, 15, 26, 27, 32, 33, 50–54, 58, 60, 75, 87–89, 94, 97, 100, 116, 118, 135–137, 147, 175, 176, 194, 195, 204, 205, 207, 208, 213, 214, 221, 222, 224, 227, 228, 232, 238–240, 248, 250, 251, 258, 289, 334, 344, 359, 362
 Kesten conjecture, 249
 Koutsky process, 196, 214–216, 239, 240

L

Law of a random variable, 172
 Laws of large numbers, 270, 276, 319, 320, 345, 347, 351, 354, 390, 392
 Leads to, 65, 145, 215, 332
 infinitely often, 107, 108, 177, 217
 Left cancellative, 7, 72
 Left factor, 15
 Left group, 7, 14, 20, 51, 52, 75–79, 83
 representation, 77
 Left ideal, 2, 3, 13, 15, 16, 26, 27, 31, 75, 118, 179, 180, 182, 191, 220, 221, 231, 234, 238, 239, 242, 244, 258, 260, 264, 267, 269, 311
 0-minimal, 18
 Left identity, 2, 4, 5, 8, 10, 13, 27

Left invariant measure, 81
 Left quotient,
 Left random walk, 172, 173, 177, 180, 182–186, 190, 193, 198, 200, 201, 204, 206, 207, 211, 212, 220, 226, 228, 231, 234, 236, 241–246, 254, 355
 Left simple, 2, 5, 7, 8, 13, 14, 32, 59, 182, 236, 239
 Left 0-simple, 10, 11
 Left translation, 6, 22, 23
 Left zero, 2, 3
 Left zero semigroup, 3, 9, 14, 133
 I^* -invariant measure, 72, 75, 76, 79, 142, 144
 Locally compact group, 45, 49, 127, 198–219, 238, 241, 244, 249, 397
 Locally compact semigroup, 27, 29, 225, 238, 241, 242, 246, 250
 Logarithmic density,

AQ7

M

Markov chain, 48, 49, 171, 174–177, 179, 180, 183, 184, 186, 189, 191, 192, 195, 196, 204, 208, 213–218, 240, 247, 381
 canonical, 175
 Markov property, 176, 208
 Markov random walk, 195, 196, 213, 214, 216, 218, 219, 239, 240, 248
 Matrix semigroup, 27, 37, 59, 60, 251, 280, 354
 MaxPlus algebra, 251
 Measure. *See also* Probability measure
 Haar, 76, 78, 79, 129–132, 138, 139, 169, 176, 192, 204–207, 210–212, 228, 232, 239, 245, 249
 unimodular, 72, 81, 211, 397
 left invariant, 81
 I^* -invariant, 72, 75, 76, 79, 142, 144
 Radon, 203, 211, 244
 right invariant, 80, 81
 r^* -invariant, 72, 77
 symmetric, 210
 Middle factor, 11, 192, 194, 232, 244, 246
 Minimal ideal, 2, 12, 15, 16, 22, 28, 37, 54, 119, 124, 133, 181, 182, 184–191, 193, 234, 235, 238, 241, 242, 248, 254, 256, 258, 260–262, 286, 287, 290, 293, 303, 311, 330, 341
 completely simple (*see* Kernel)
 Minimal rank, 12, 15–17, 35, 53, 260, 266, 279, 296, 311, 315, 329, 359, 361

AQ8

Mixed random walk, 172, 173, 175, 181,
183, 187, 188, 190, 193, 198, 208,
211, 212, 219, 220, 224, 228, 229,
232–234, 236, 242, 247, 248, 253,
270, 333, 334, 338, 342, 344, 345,
347, 349, 352
Mod (Modulo), 6, 7, 11, 184, 192
Model map, 389
Monoid, 248
Monothetic subsemigroup, 26, 222
Multiplication, 1–3, 5–9, 11, 13, 14, 20, 22–24,
27–30, 32, 33, 46, 50, 51, 58, 61,
85–87, 95, 135, 137, 168, 182, 197,
204, 205, 219, 221, 222, 225, 228,
230, 241, 280, 286, 303–305, 344,
355, 357, 394
jointly continuous multiplication, 24, 30,
33, 35, 168, 220
separately continuous multiplication, 22,
24

N

Neat group, 127
Nonzero elements, 10, 18, 19, 48, 179, 246,
261, 262
Norm, asymptotic behavior of,
Null cell, 11, 18, 19, 30, 31, 35
Null recurrent, 178, 189, 190, 194, 246
Null semigroup, 10, 18
Null space, 16, 17, 33, 34, 36, 41

O

Open translations, 81

P

Parallel random walks, 250
Partition of a transformation, 20
Period of random walk, 183
Permutation group, 8
Permutation matrix, 6, 7
Polynomial growth, 210
Positive elements, 190, 233
Positive Harris, 218
Positive recurrent, 178, 189–196, 208, 211,
212, 217, 225, 227, 244–246, 250,
253, 258, 265, 266, 273, 281
Primitive idempotent, 12–14, 16–18, 20, 21,
28–30, 256
Principal ideal, 2, 26
left, 2
right, 2

Probability measure

absolutely continuous component of, 129,
131
adapted, 206, 207, 399
convergence of, 83–168 (*see also*
Convergence)
convolution of, 83–168
idempotent, 63–81, 87, 88, 96, 115, 137,
143, 224, 241, 293, 317, 329, 330
invariant, 61, 77, 83, 97, 212, 225, 232,
234, 235, 250, 265, 266, 277, 278,
283, 321, 337, 338, 342–345, 348,
356, 358, 370, 371, 374, 375, 388,
390, 392
for mixed random walks on nonnegative
matrices, 333
for products of I.I.D. random matrices,
284–333
recurrent, 189, 193
null, 189
positive, 189
regular, 64, 65, 75, 173, 180
spread-out, 128
strictly aperiodic, 128
support of, 177, 179, 180
tightness of, 119

Product measure, 74, 172, 192, 205, 232, 392

Product σ -algebra, 172

Projection matrix, 29, 34, 219

Proper divisors of zero, 10–12, 19, 27, 28

Pseudoinvertible, 22, 54

Q

Quotient semigroup, 6, 7, 34

R

Radon measure, 203, 211, 244
Random walk, 108, 150, 171–251, 253–284,
333–355, 358. *See also* Bilateral
random walk; Left random walk;
Mixed random walk; Right random
walk
equidistributed, 212, 238, 239
Harris, 204, 207, 208, 217
parallel, 237
period of, 183
recurrent, 178, 194, 195, 217, 218, 246,
250, 254–284
null, 178, 194, 246
positive, 189–196, 208, 211, 212, 217,
225, 227, 245, 246
positive Harris, 218

AQ9

AQ10

AQ11

transient, 178, 187, 188, 197, 203, 217
 unilateral, 184, 188
 Range
 of a matrix, 33
 of a transformation,
 Rank. *See also* Minimal rank
 of a matrix, 16
 of a transformation, 20
 Rate of decay of the concentration function,
 397–412
 Recurrent
 Harris, 204, 207, 208, 217
 null, 178, 189, 190, 194, 246
 positive, 178, 189–196, 208, 211, 212, 217,
 225, 227, 244–246, 250, 253, 258,
 265, 266, 273, 281
 Recurrent sets (R , $R(Z)$, $R(W)$) of random
 walks on matrices, 258, 355
 Rees factor (quotient) semigroup, 7
 Rees product, 9, 11, 12, 14, 15, 17, 18, 22,
 27–29, 35, 49, 61, 197, 228
 over a group with 0, 11, 12
 representation, 15
 Rees–Suschkewitsch representation, 12–22,
 26, 35, 184, 185, 194, 231, 238,
 245, 246
 Regular probability measures, 64, 65, 75, 180
 Regular sandwich function, 10
 Regular semigroup,
 Return point, 177, 178, 180, 181, 183
 Return time, 176, 194, 196, 200, 202, 204,
 208, 218
 Right cancellative, 2, 7, 72, 77, 81
 Right convolution operator, 399, 403, 404
 Right factor, 15
 Right group, 7–10, 16, 20, 27, 31, 75, 76, 256
 Right ideal, 2, 3, 9–11, 13–16, 18, 22, 26, 31,
 34, 53, 117, 118, 180, 182, 183,
 191, 220, 221, 224, 228, 231, 232,
 234, 236, 239, 240, 242, 250, 255,
 256, 258–260, 311
 0-minimal, 10
 Right identity, 2, 6, 10, 14
 Right-invariant measure, 80, 81
 Right random walk, 173, 174, 180, 182, 183,
 185–188, 190–192, 194, 195, 198,
 200, 201, 203, 205, 207, 211, 220,
 224, 226–228, 231, 232, 234–238,
 242, 245, 246, 250, 253, 254, 355,
 358
 Right reversible, 81
 Right simple, 2, 7, 8, 24, 26, 31, 72, 235
 Right 0-simple, 10–12
 Right translation, 6, 22, 81

AQ12

Right zero, 2, 3, 34
 Right zero semigroup, 7, 9–11, 14, 27, 29, 31,
 42, 93, 133
 r^* -invariant measure, 72, 77
 Road-coloring problem, 61

S
 Sandwich function, 9, 11, 12, 14, 17, 18, 22,
 35, 185, 194, 195, 241, 248
 regular, 11
 trivial, 9
 Semigroup, 1–251, 253–256, 259–261, 265,
 266, 268, 270, 272, 273, 279, 280,
 284, 287, 289, 293, 296, 297, 302,
 303, 305, 311–313, 316, 322, 331,
 333, 334, 337, 341, 342, 348, 354,
 356–359, 362, 369, 381, 383, 397,
 398, 404–411
 abelian, 2, 26, 63, 147, 151, 154–156, 159,
 168, 194, 222
 discrete, 63, 156, 159
 of affine maps, 214, 222, 251
 bicyclic, 20–22, 59, 197
 cancellative, 32, 95, 131
 commutative (*see* Abelian)
 compact, 26–29, 32, 63, 100, 141, 147,
 152, 168, 179, 219–242, 246–250,
 279, 305, 341, 358
 adelian, 26, 63, 147, 151, 154, 155, 222
 completely simple, 13–17, 26, 31, 42, 46,
 47, 49, 52, 59, 60, 63, 82, 87, 88,
 91, 106, 109, 115, 119, 123, 134,
 136, 166, 171, 221, 241–246, 251,
 273, 312, 357
 of nonnegative matrices, 46, 47
 completely 0-simple, 17–19, 21, 30, 33, 60,
 246
 direct product of, 7, 251
 discrete, 63, 122, 124, 135, 136, 138, 169,
 180–197, 213, 220, 241, 247, 248
 abelian, 63, 156, 159
 factor, 6, 7, 20
 invariance, 215–218, 240
 left zero, 3, 9, 14
 locally compact, 27, 29, 225, 238, 241,
 242, 246, 250
 of matrices, 16, 42, 53, 168
 of nonnegative matrices, 46, 47, 50, 105,
 259, 261
 null, 10, 18
 positive recurrent, 189
 pseudoinvertible, 54
 quotient, 34

- recurrent, 189, 194
 - positive, 189
 - Rees factor (quotient), 7, 20
 - regular, 11, 12
 - right reversible, 81
 - right zero, 7, 9–11, 14, 27, 29, 31, 42, 93, 133
 - semitopological, 22, 168
 - topological, 22–34, 48, 60, 63–169, 171, 177, 179, 219, 222, 241, 289, 322, 337, 357
 - of transformations, 1, 3, 4, 6–8, 12, 15, 16, 19, 20, 52, 197
 - full, 3, 6, 8, 20
 - transient, 218
 - with 0, 10, 12, 17, 27, 29, 225
 - 0-simple, 11, 12, 17–19, 21, 28–30, 60, 246, 260
 - Semigroup homomorphism, 6
 - Semigroup isomorphism, 6, 14, 27
 - Semigroup monothetic, 3, 29
 - Semigroup of relations, 4, 5, 60
 - Separately continuous multiplication, 22, 24, 168
 - Shift operator, 175, 231
 - Sierpinski gasket, 278–280, 356
 - Simple, 2
 - Skeleton of a matrix, 253
 - Solvable transformation semigroup, 52
 - Spread-out, 128, 169, 204–208, 210, 211
 - State space, 52, 55, 171, 177, 189, 196, 198, 204, 215, 240, 241, 250, 284, 347, 381
 - Stationary distributions, 231, 238, 249
 - Stochastic matrix, 27, 48, 165, 332, 360, 361, 365, 366, 383, 386, 394
 - Stopping time, 175, 194, 195, 238
 - Strange group, 127
 - Strictly aperiodic, 128
 - Strongly continuous operator, 176
 - Strong Markov property, 176, 208
 - Subgroup, 2
 - distinguished, 134
 - Subharmonic, 211
 - Subsemigroup, 2–6, 8, 9, 11–13, 15, 19, 21–23, 25–30, 32, 33, 35, 37, 54, 61, 68, 74, 75, 79, 81, 87, 88, 90, 92, 94, 95, 97, 98, 113, 142, 143, 145, 149–151, 161, 165, 177, 179, 182, 185, 198, 199, 206, 215, 222, 228, 238, 241–244, 249, 256, 259, 260, 266, 268, 270, 293, 297, 311, 312, 316, 348, 362, 397, 398, 404, 408–411. *See also* Semigroup monothetic
 - Substochastic matrix, 27, 48
 - Superharmonic, 211
 - Support of a measure, 177, 225
 - Suschkewitsch, 12–22, 26, 29, 30, 35, 60, 184, 185, 194, 231, 238, 241, 244–246, 248, 251
 - Symmetric group, 8, 9
 - Symmetric measure, 210
- T**
- Taboo probability, 195
 - Tail-idempotent, 160, 161
 - Tight, 83, 87–89, 96–99, 101, 105, 106, 110–113, 118–124, 132, 133, 136, 137, 141, 150, 159, 163, 164, 171, 191, 258, 266, 269, 275, 283, 285, 286, 288–291, 293, 296, 302, 303, 305–307, 309, 310, 312, 313, 315, 317–321, 325–329, 331–334, 338, 339, 341–344, 347, 354
 - Tightness criterion
 - for products of affine maps, 250, 254, 313
 - for products of random matrices, 284–333
 - for products of random 2×2 matrices, 318
 - Topological group, 24, 26–34, 42, 45, 72, 75, 81, 94–96, 128, 151, 179, 198, 204, 213
 - Topological isomorphism, 27, 79, 81, 87
 - Topological semigroups, 22–34, 48, 60, 63–169, 171, 177, 179, 219, 222, 241, 289, 322, 337, 357
 - Trace, 36, 40, 46, 52
 - Trace of transformation, 52
 - Transformation, 1, 3–8, 12, 15, 16, 19–21, 33, 34, 48, 52, 61, 197, 260
 - semigroup of, 3, 4, 20
 - Transformation matrix, 7, 48
 - Transient, 49, 178, 187–189, 194, 197, 203, 208, 211, 217, 218, 222, 223, 225, 249
 - Transition probability function, 173, 177, 180, 334, 338
 - Transition probability matrix, 61, 184, 240
 - inverse, 61
 - Transition probability operator, 173, 176, 177, 179, 211, 215, 220, 224, 229, 230, 238, 240, 249, 250
 - Translation, 6, 22, 23, 81, 175, 219, 231
 - open, 81

Trivial sandwich function, 9
 Two sided. *See* Bilateral
 Type of a point, 189

U

Unconditionally recurrent, 186
 Unilateral walk, 182, 184, 187, 188, 198, 201,
 223, 238, 239, 242, 244
 Unimodular, 72, 81, 211, 397, 405
 Upper packing dimension, 391

V

Vague convergence, 66, 84, 85, 99, 101, 105,
 107, 108, 116, 122, 124, 211
 to 0, 85, 101, 105, 107, 108, 116, 122, 134,
 211
 Volume growth rate, 28, 398, 404, 407, 408

W

Weak convergence
 of convolution sequences, 63, 64, 83–168
 of empirical distributions, 192
 of products of affine maps, 313
 of products of nonnegative matrices, 287
 of products of real matrices, 302
 Weak*-convergence, 67, 101
 Weakly continuous operator, 176

Z

Zero, 2, 10
 divisor of, 10, 28
 proper, 10
 Zero semigroup, 3, 7, 9–11, 14, 27, 29, 31, 42,
 93, 133
 0-minimal ideal, 10
 0-minimal left (right) ideal, 10, 18
 0-simple, 10–12, 17–22, 28–30, 33–35, 37, 60,
 246, 251, 259, 260, 262, 264

AUTHOR QUERIES

AQ1: The entry “logarithmic” not found. Please check.	2
AQ2: Found as 'Embedded' in text. Please check.	3
AQ3: The entry “Group” occurs more than 700 instances. We have picked first in- stance page number from each chapter.	4 5
AQ4: The entry “Ideal” occurs more than 300 instances. We have picked first instance page number from each chapter.	6 7
AQ5: The entry “Item potent” occurs more than 300 instances. We have picked first instance page number from each chapter.	8 9
AQ6: The entry “Irreversibility” not found please check.	10
AQ7: The entry “Left quotient” not found please check.	11
AQ8: The entry “Logarithmic density” not found please check.	12
AQ8: The entry “Norms, asymptotic behavior of” not found please check.	13
AQ10: Exact term not found please check.	14
AQ11: The entry “of a transformation” not found. Please check.	15
AQ12: The entry “Regular semigroup” not found. Please check.	16

Uncorrected Proof

