Lectures on Stochastic Stability

Sergey FOSS

Heriot-Watt University

Lecture 5

Monotonicity and Saturation Rule

1 Introduction

The paper of Loynes [8] was the first to consider a system (single server queue, and, later, queues in tandem) with stationary-ergodic driver. The classical recursion $X_{n+1} = (X_n + \xi_n)^+$ studied by Loynes is monotone. There is a variety of monotone models (including queues in tandem, multi-server queues, Jackson-type networks, etc.) for which one can develop a unified approach for stability study. Here we provide a short survey, see the references for more details.

2 Single-server queue revisited

Consider a single server queue with interarrival times σ_n and service times t_n. Assume that a sequence $\{(\sigma_n, t_n)\}$ is stationary with finite means $b = \mathbf{E}\sigma_1$ and $a = \mathbf{E}t_1$ and satisfies the SLLN, i.e.

$$\frac{1}{n}\sum_{-n}^{-1}\sigma_i = b \quad \text{and} \quad \frac{1}{n}\sum_{-n}^{-1} t_i = a \quad \text{a.s.} \tag{1}$$

A sufficient condition for (1) to hold (but not necessary in general – see, e.g., a discussion in [6]) is that a sequence $\{(\sigma_n, t_n)\}$ is ergodic (see, e.g., Lecture 1 for definitions).

Assume further that $b < a$. Assume also, for simplicity, that customer 1 arrives in an empty queue. Let W_n be a waiting time of customer n. Then $W_1 = 0$ and

$$W_n = \max_{1\le j\le n}\sum_{i=j}^{n-1}(\sigma_i - t_i).$$

Also, $\{W_n\}$ satisfy the following recursion

$$W_{n+1} = \max(0, W_n + \sigma_n - t_n). \tag{2}$$

Note that W_n coincides in distribution with

$$W_n \circ \theta^{-n} = \max_{-n+1 \le j \le 0} \sum_{i=j}^{-1} (\sigma_i - t_i).$$

The latter sequence increases a.s. and "couples" with an a.s. finite limit

$$W^0 = \sup_{j \le 0} \sum_{i=j}^{-1} (\sigma_i - t_i)$$

due to the SLLN (1). One can also define, for any m,

$$W^m = \sup_{j \le m} \sum_{i=j}^{-1} (\sigma_i - t_i).$$

Then $\{W^m\}$ is a stationary sequence and

$$W^{m+1} = \max(0, W^m + \sigma_m - t_m).$$

Thus, this is a stationary solution to recursive equation (2).

Exercise 1. Show that this is the only stationary solution.

3 Tandem of two single-server queues

Tandem queue will be our toy example.

Consider an open network with two single-server stations in tandem. Customers arrive to the first station with interarrival times $\{t_n\}$ and form a queue there. In this example, it is convenient to assume that t_n is an interarrival time between customers $n-1$ and n. A server serves them in order of arrival with service times $\{\sigma_n^{(1)}\}$. Upon service completion, customers go to the second station where are served also in order of arrival with service times $\{\sigma_n^{(2)}\}$. Assume that customer 1 arrives in an empty system. Denote by Z_n a sojourn time of customer n. Then

$$Z_1 = \sigma_1^{(1)} + \sigma_1^{(2)}$$

and, more generally, for any $n \ge 1$,

$$Z_n = \max_{1 \le k \le m \le n} \left(\sum_{k}^{m} \sigma_j^{(1)} + \sum_{m}^{n} \sigma_j^{(2)} - \sum_{k+1}^{n} t_j \right) \tag{3}$$

Exercise 2. Prove formula (3).

It is known that if a sequence $\{(\sigma_n^{(1)}, \sigma_n^{(2)}, t_n)\}$ is stationary ergodic and if

$$\frac{\max(b^{(1)}, b^{(2)})}{a} < 1 \tag{4}$$

where $b^{(i)} = \mathbf{E}\sigma_1^{(i)}$ and $a = \mathbf{E}t_1$, then a distribution of Z_n converges to a limiting stationary distrribuion in the total variation norm.

Exercise 3. Show that if there is the opposite strict inequality in (4), then a sequence $\{Z_n\}$ tends to infinity a.s.

In order to introduce a recursive scheme, let $Z_n^{(1)}$ be a sojourn time of customer n in the first system (i.e. a sum of a waiting time and of service time $\sigma_n^{(1)}$). Then $Z_1^{(1)} = \sigma_1^{(1)}$ and we get the following recursive relations:

$$\begin{aligned} Z_{n+1}^{(1)} &= \max(0, Z_n^{(1)} - t_{n+1}) + \sigma_{n+1}^{(1)}, \\ Z_{n+1} &= \max(Z_{n+1}^{(1)}, Z_n - t_{n+1}) + \sigma_{n+1}^{(2)}. \end{aligned}$$

In other words, introduce a function $f : \mathbb{R}^5 \to R^2$ as follows:

$$f(x_1, x_2, y_1, y_2, y_3) = (\max(0, x_1 - y_3) + y_1, \max(\max(0, x_1 - y_3) + y_1, x_2 - y_3) + y_2).$$

Then f is monotone non-decreasing in (x_1, x_2), and we get a recursion

$$(Z_{n+1}^{(1)}, Z_{n+1}) = f((Z_n^{(1)}, Z_n), \xi_{n+1})$$

where $\xi_{n+1} = (\sigma_{n+1}^{(1)}, \sigma_{n+1}^{(2)}, t_{n+1})$.

4 General statements on monotone and homogeneous recursions

There are many applications, especially in queueing networks, where monotonicity in the dynamics can be exploited to prove existence and uniqueness of stationary solutions. Although the theory can be presented in the very general setup of a partially ordered state space (see Brandt *et al.* [6]) we will only focus on the case where the state is $\mathbb{R}^d$. Consider then the SRS

$$X_{n+1} = f(X_n, \xi_{n+1}) =: \varphi_{n+1}(X_n)$$

and assume that $\varphi_0 : \mathbb{R}_+^d \to \mathbb{R}_+^d$ is increasing and right-continuous, where the ordering is the standard component-wise ordering on $\mathbb{R}_+^d$. Let θ be stationary and ergodic flow on $(\Omega, \mathscr{F}, P)$ and assume that $\varphi_n = \varphi_0 \circ \theta^n$, $n \in \mathbb{Z}$. In other words, $\{\varphi_n\}$ is a stationary-ergodic sequence of random elements of the space of right-continuous increasing functions on $\mathbb{R}_+^d$. We first explain Loynes' method. Define

$$\Phi_n := \varphi_n \cdots \varphi_1.$$

Thus, $\Phi_n(Y)$ is the solution of the SRS at $n \geq 0$ when $X_0 = Y$, a.s. Since 0 is the least element of $(R_+^d, \leq)$, we have $\Phi_n(0) \leq \Phi_n(Y)$, a.s., for any R_+^d–valued r.v. Y. Next consider

$$\Phi_{m+n}(0) \circ \theta^{-m} = \varphi_n \cdots \varphi_{-m+1}(0), \quad n \geq -m,$$

and interpret $\Phi_{m+n}(0)$ as the solution of the SRS at time $n \geq -m$, starting with 0 at time $-m$. Clearly, $\Phi_{m+n}(0)$ increases as m increases, because:

$$\begin{aligned} \Phi_{(m+1)+n}(0) \circ \theta^{-(m+1)} &= \varphi_n \cdots \varphi_{-m+1} \varphi_{-(m+1)}(0) \\ &= \varphi_n \cdots \varphi_{-m+1}(\varphi_{-(m+1)}(0)) \\ &\geq \varphi_n \cdots \varphi_{-m+1}(0) = \Phi_{m+n}(0) \circ \theta^{-m}. \end{aligned}$$

Finally define

$$\tilde{X}_n := \lim_{m\to\infty} \Phi_{m+n}(0)\circ\theta^{-m}, \quad n \in \mathbb{Z}.$$

The r.v. X_n is either finite a.s., or is infinite a.s., by ergodicity. Assuming that the first case holds, we further have

$$\begin{aligned}
\tilde{X}_{n+1} &= \lim_{m\to\infty} \Phi_{m+(n+1)}(0)\circ\theta^{-m} \\
&= \lim_{m\to\infty} \varphi_{m+n+1}\varphi_{m+n}\cdots\varphi_1(0)\circ\theta^{-m} \\
&= \lim_{m\to\infty} \varphi_{n+1}\varphi_n\cdots\varphi_{-m+1}(0) \\
&= \lim_{m\to\infty} \varphi_{n+1}(\varphi_n\cdots\varphi_{-m+1}(0)) \\
&= \lim_{m\to\infty} \varphi_{n+1}(\Phi_{m+n}(0)\circ\theta^{-m+1}) \\
&= \varphi_{n+1}(\tilde{X}_n).
\end{aligned}$$

Provides then that we have a method for proving $\mathbf{P}(X_0 < \infty) > 0$, Loynes' technique results in the construction of a stationary-ergodic solution $\{\tilde{X}_n\}$ of the SRS.

For a tandem queue, we get

$$\begin{aligned}
\tilde{X}_n &= (Z_n^{(1)}, Z_n) \circ \theta^{-n} \\
&= \left(\sigma_0^{(1)} + \max_{-n+1\le i\le 0} \sum_{j=i}^{-1} (\sigma_j^{(1)} - t_j), \max_{-n+1\le i\le k\le 0} \left(\sum_i^k \sigma_j^{(1)} + \sum_k^0 \sigma_j^{(2)} - \sum_{i+1}^0 t_j\right)\right)
\end{aligned}$$

and clearly this sequence increases a.s. Here, by convention, $\sum_i^r \ldots = 0$ is $i > r$.

Consider another example of a multiserver queue.

Example. A multiserver queue $G/G/s$ with s servers and FCFS service discipline.

Customers arrive with interarrival times $\{t_n\}$ and have service times $\{\sigma_n\}$ (service times are associated with customers, not with servers. Upon arrival to the system, a customer is immediately sent to a server (one of servers) with a minimal workload. At each server, customers are served in order of arrival. Denote by $V_{n,i}$ a workload of server i just before arrival of nth customer. Let R be an operator that orders coordinates of a vector in the non-decreasing order. Let

$$W_n = R(V_{n,1}, \ldots\ V_{n,s}).$$

Then vectors $\{W_n\}$ satisfy the following recursive equation ("Kiefer-Wolfowitz"):

$$W_{n+1} = R(W_n + \mathbf{e}_1\sigma_n - \mathbf{i}t_{n+1})^+$$

Here $x^+ = \max(x, 0)$ (coordinatewise), and $\mathbf{e}_1 = (1, 0, \ldots, 0)$ and $\mathbf{i} = (1, 1, \ldots, 1)$.

Exercise 4. Show the monotonicity of the latter recursion.

Return to the general setting. Without further assumptions and structure, not much can be said. Assume next that, in addition, φ_0 is homogeneous, i.e.,

$$\varphi_0(x + c\mathbf{1}) = \varphi_0(x) + c\mathbf{1},$$

for all $x \in \mathbb{R}_+^d$ and all $c \in \mathbb{R}$. Such is the case, e.g., with the usual Lindley function $\varphi_0 : \mathbb{R}_+ \to \mathbb{R}_+$, with $\varphi_0(x) = \max(x + \xi_0, 0)$. The homogeneity assumption is quite frequent in queueing theory. It is easy to see that

$$|\varphi_0(x) - \varphi_0(y)| \leq |x - y|,$$

where $|x| := \max(|x_1|, \ldots, |x_d|)$. Suppose then that $\{X_n\}$, $\{Y_n\}$ are two stationary solutions of the SRS. Then

$$|X_{n+1} - Y_{n+1}| = |\varphi_{n+1}(X_n) - \varphi_{n+1}(Y_n)| \leq |X_n - Y_n|, \tag{5}$$

for all n, a.s., and since $\{|X_n - Y_n|, n \in \mathbb{Z}\}$ is stationary and ergodic, this a.s. monotonicity may only hold if $|X_n - Y_n| = r$, for some constant $r \geq 0$. Thus, a necessary and sufficient condition for the two solutions to coincide is that,

$$\mathbf{P}(|\varphi_1(X_0) - \varphi_1(Y_0)| < |X_0 - Y_0|) > 0. \tag{6}$$

Remark. Conditions (5) and (6) form the basic for the so-called *contraction* approach.

A classical example where, in general, (5) holds but (6) fails is the $G/G/s$ queue, that is, the s-server queue with stationary-ergodic data. Let λ, μ be the arrival and service rates, respectively. Here, there is a minimal and a maximal stationary solution which, provided that $\lambda < s\mu$, may not coincide. For details see Brandt et al [6]. But condition (6) holds if the "driving" sequences are i.i.d. (or satisfy a weaker "mixing" condition). Also condition (6) holds for the tandem queue with stationary and ergodic driving sequences. This is **Exercise 5** to you.

5 The Monotone-Homogeneous-Separable (MHS) framework

Consider a recursion of the form

$$W_{n+1} = f(W_n, \xi_{n+1}, \tau_{n+1}),$$

where ξ_n are general marks, and $\tau_n \geq 0$. The interpretation is that τ_n is the interarrival time between the $n-1$-th and n-th customer, and W_n is the state just after the arrival of the n-th customer. We consider arrival epochs $\{T_n\}$ such that $T_{n+1} - T_n = \tau_n$. We write $W_{m,n}$ for the solution of the recursion at index n when we start with a specific state, say 0, at $m \leq n$. Finally we consider a functions of the form

$$X_{[m,n]} = f_{m+n-1}(W_{m,n}; T_m, \ldots, T_n; \xi_{m+1}, \ldots, \xi_n),$$

which will be thought of as epochs of last activity in the system. For instance, when we have an s-server queue, $X[m,n]$ represents the departure time of the last customer when the queue is fed only by customers with indices from m to n. Correspondingly, we define the quantity

$$Z_{[m,n]} := X_{[m,n]} - T_n,$$

the time elapsed between the arrival of the last customer and the departure of the last customer. The framework is formulated in terms of the $X_{[m,n]}$, $Z_{[m,n]}$ and their dependence on the $\{T_n\}$. For $c \in \mathbb{R}$, let $\{T_n\} + c = \{T_n + c\}$. For $c > 0$, let $c\{T_n\} = \{cT_n\}$. Define $\{T_n\} \leq \{T'_n\}$ if $T_n \leq T'_n$ for all n. We require a set of four assumptions:

(A1) $Z_{[m,n]} \geq 0$

(A2) $\{T_n\} \leq \{T'_n\} \Rightarrow X_{[m,n]} \leq X'_{[m,n]}$.

The first assumption is natural. In the second one, $X'_{[m,n]}$ are the variables obtained by replacing each T_n by T'_n; it says that delaying the arrival epochs results in delaying of the last activity epochs.

(A3) $\{T'_n\} = \{T_n\} + c \Rightarrow X'_{[m,n]} = X_{[m,n]} + c$.

This is a time-homogeneity assumption.

(A4) For $m \leq \ell < \ell + 1 \leq n$, $X_{[m,\ell]} \leq T_{\ell+1} \Rightarrow X_{[m,n]} = X_{[\ell+1,n]}$.

If the premise $X_{[m,\ell]} \leq T_{\ell+1}$ of the last assumption holds, we say that we have separability at index ℓ. It means that the last activity due to customers with indices in $[m, \ell]$ happens prior to the arrival of the $\ell + 1$-th customer, and so the last activity due to customers with indices in $[m, n]$ is not influenced by those customers with indices in $[m, \ell]$. Basic consequences of the above assumptions are summarized in:

Lemma 1. *(i) The response $Z_{[m,n]}$ depends on $T_m, \ldots, T_n$ only through the differences $\tau_m, \ldots, \tau_{n-1}$.*
(ii) Let $a \leq b$ be integers. Let $T'_n = T_n + Z_{[a,b]}\mathbf{1}(n > b)$, $T''_n = T_n - Z_{[a,b]}\mathbf{1}(n \leq b)$. And let $X'_{[m,n]}$, $X''_{[m,n]}$ be the corresponding last activity epochs. Then both of them exhibit separability at index b.
(iii) The variables $X_{[m,n]}$, $Z_{[m,n]}$ increase when m decreases.
(iv) For $a \leq b < b + 1 \leq c$, $Z_{[a,c]} \leq Z_{[a,b]} + Z_{[b+1,c]}$.

Proof. (i) Follows from the definition $Z_{[m,n]} = X_{[m,n]} - T_n$ and the homogeneity assumption (A3).
(ii) Obviously, $Z_{[a,b]} \leq \tau_b + Z_{[a,b]}$, and so $X_{[a,b]} - T_b \leq \tau_b + Z_{[a,b]}$, which implies $X_{[a,b]} \leq T_{b+1} + Z_{[a,b]}$. The right-hand side is T'_{b+1}, by definition. The left-hand side is equal to $X'_{[a,b]}$ because $T'_n = T_n$ for $n \leq b$. So $X'_{[a,b]} \leq T'_{b+1}$ and this is separability at index b. Similarly for the other variable.
(iii) Let $a = b = m$ in (ii). Since we have separability at index m, we conclude that $X''_{[m,n]} = X''_{[m+1,n]}$. But $T''_k = T_k$ for $k \in [m + 1, n]$ and so $X''_{[m+1,n]} = X_{[m+1,n]}$. On the other hand, $\{T''_k\} \leq \{T_k\}$ and so, by (A2), $X''_{[m,n]} \leq X_{[m,n]}$. Thus $X_{[m,n]} \geq X_{[m+1,n]}$. And so $Z_{[m,n]} \geq Z_{[m+1,n]}$ also.
(iv) Apply (ii) again. Since $\{T_k\} \leq \{T'_k\}$, (A2) gives $X_{[a,c]} \leq X'_{[a,c]}$. By separability at index b, as proved in (ii), we have $X'_{[a,c]} = X'_{[b+1,c]}$. Because $T'_k = T_k + Z_{[a,b]}$ for all $k \in [b + 1, c]$, we have, by (A3), $X'_{[b+1,c]} = X_{[b+1,c]} + Z_{[a,b]}$. Thus, $X_{[a,c]} \leq X_{[b+1,c]} + Z_{[a,b]}$. Subtracting T_c from both sides gives the desired. □

Introduce next the usual stationary-ergodic assumptions. Namely, consider $(\Omega, \mathscr{F}, \mathbf{P})$ and a stationary-ergodic flow θ. Let $\xi_n = \xi_0 \circ \theta^n$, $\tau_n = \tau_0 \circ \theta^n$, set $T_0 = 0$, and suppose $\mathbf{E}\tau_0 = \lambda^{-1} \in (0, \infty)$, $\mathbf{E}Z_{0,0} < \infty$. Stability of the original system can, in specific but important cases, be translated in a stability statement for $Z_{[m,n]}$. Hence we shall focus on it. Note that $Z_{[m,n]} \circ \theta^k = Z_{[m+k,n+k]}$ for all $k \in \mathbb{Z}$. For any $c \geq 0$, introduce the epochs

$c\{T_n\} = \{cT_n\}$ and let $X_{[m,n]}(c)$, $Z_{[m,n]}(c)$ be the quantities of interest. The subadditive ergodic theorem gives that

$$\gamma(c) := \lim_{n\to\infty} \frac{1}{n} Z_{[-n,-1]}(c) = \lim_{n\to\infty} \frac{1}{n} \mathbf{E} Z_{[-n,-1]}(c)$$

is a nonnegative, finite constant. The previous lemma implies that $\gamma(c) \geq \gamma(c')$ when $c > c'$. Similarly, $\lim n^{-1} X_{[1,n]}(c) = \gamma(c) + \lambda^{-1} c$, and the latter quantity increases as c increases. Monotonicity implies that $Z_{[-n,-1]}(c)$ increases as n increases, and let $\tilde{Z}(c)$ be the limit. Ergodicity implies that $\mathbf{P}(\tilde{Z}(c) < \infty) \in \{0, 1\}$. Put $\tilde{Z} = \tilde{Z}(1)$. The stability theorem[1] is:

Theorem 1. *If* $\lambda\gamma(0) < 1$ *then* $\mathbf{P}(\tilde{Z} < \infty) = 1$. *If* $\lambda\gamma(0) > 1$ *then* $\mathbf{P}(\tilde{Z} < \infty) = 0$.

Proof. Assume first that $\lambda\gamma(0) > 1$. Fix $n \geq 1$. Define $T'_k = T_{-n}$ for all $k \in \mathbb{Z}$. Hence $X'_{[-n,0]}(1) \leq X_{[-n,0]}(1) = Z_{[-n,0]}(1)$, by (A2). On the other hand, by (A3), $X'_{[-n,0]}(1) = X_{[-n,0]}(0) + T_{-n} = Z_{[-n,0]}(0) + T_{-n}$. Thus, $n^{-1} Z_{[-n,0]}(1) \geq n^{-1} Z_{[-n,0]}(0) + n^{-1} T_{-n}$, and, taking limits as $n \to \infty$, we conclude $\liminf n^{-1} Z_{[-n,0]}(1) \geq \gamma(0) - \lambda^{-1} > 0$, a.s.

Assume next that $\lambda\gamma(0) < 1$. Let $\gamma_n(0) := \mathbf{E} Z_{[-n+1,0]}(0)/n$. Since $\gamma(0) = \lim_{n\to\infty} \gamma_n(0) = \inf_n \gamma_n(0)$, we can find an integer K such that $\lambda\gamma_K(0) < 1$. Consider next an auxiliary single server queue with service times $\sigma^*_n := Z_{[-Kn+1,-K(n-1)]}(0)$ and interarrival times $t^*_n := \sum_{i=-Kn+1}^{-K(n-1)} t_i$. Notice that $\{(t^*_n, s^*_n), n \in \mathbb{Z}\}$ is a stationary sequence which satisfies the SLLN. Consider the waiting time W_n of this auxiliary system: $W_{n+1} = (W_n + s^*_n - t^*_n)^+$. Since $\mathbf{E} s^*_n = \gamma_K < \lambda^{-1} = E t^*_n$, the auxiliary queue is stable. Since the separability property holds, we have the following domination:

$$Z_{[-nK+1,0]}(1) \leq W_n \circ \theta^{-n} + s^*_0, \text{ a.s.},$$

where W_n here is the waiting time of the n-th customer if the queue starts empty. By the Loynes' scheme, $W_n \circ \theta^{-n}$ converges (increases) to an a.s. finite random variable. Hence $\tilde{Z} = \lim_n Z_{[-nK+1,0]}(1)$ is also a.s. finite. □

Example Consider a tandem queue and find $\gamma(0)$. Let $b = \max(b^{(1)}, b^{(2)})$. Then

$$\begin{aligned}
\gamma(0) &= \lim_{n\to\infty} \frac{1}{n} \max_{0\leq m\leq n} \left(\sum_{-n}^{-m} \sigma_j^{(1)} + \sum_{-m}^{0} \sigma_j^{(2)} \right) \\
&\geq \lim \frac{1}{n} \max \left(\sum_{-n}^{0} \sigma_i^{(1)}, \sum_{-n}^{0} \sigma_i^{(2)} \right) \\
&= b.
\end{aligned}$$

¿From the other side, assume that, say, $b = b^{(1)} > b^{(2)}$. Then

$$\begin{aligned}
\gamma(0) &= \lim \frac{1}{n} \left(\sum_{-n}^{0} \sigma_i^{(1)} + \max_{0\leq m\leq n} \sum_{-m}^{0} (\sigma_i^{(2)} - \sigma_i^{(1)}) \right) \\
&\leq \lim \frac{1}{n} \sum_{-n}^{0} \sigma_i^{(1)} + \frac{1}{n} \sup_{m\geq 0} \sum_{-m}^{0} (\sigma_i^{(2)} - \sigma_i^{(1)})
\end{aligned}$$

[1] This is known as the "saturation rule"

where the supremum in the RHS is finite a.s. and does not depend on n. Therefore, the second term tends to 0 a.s. Thus, $\gamma(0) = b$. The same conclusion holds if $b^{(1)} < b^{(2)}$ (by the symmetry) and if $b^{(1)} = b^{(2)}$ (Why ?? – this is **Exercise 6** for you !)

Exercise 7. Using Theorem 1, find stability conditions for $G/G/s$ queue.

6 Saturation rule for large deviations

The proposed construction of an upper single-server queue may be of use not only for staibility study, but also for study of large deviations.

We consider here only an example of a tandem queue. In this section, we assume that three driving sequences $\{\sigma_n^{(1)}\}$, $\{\sigma_n^{(2)}\}$, and $\{t_n\}$ are mutually independent and each of them consists of i.i.d.r.v.'s. We assume also that the stability condition $a > b$ holds. We are interested in the asymptotics for $\mathbf{P}(Z > x)$ as $x \to \infty$. For the simplicity,
– we consider only the case when both distributions of random variables $\sigma^{(1)}$ and $\sigma^{(2)}$ are light-tailed, and
– we study only the logarithmic asymptotics: $\log \mathbf{P}(Z > x) \sim \ldots$.

For $i = 1, 2$, let $\varphi^{(i)}(u) = \mathbf{E} \exp(u\sigma_1^{(i)})$ and $\varphi_\tau(u) = \mathbf{E} \exp(ut_1)$. Let

$$\gamma^{(i)} = \sup\{u \ : \ \varphi^{(i)}(u)\varphi_\tau(-u) \le 1\}.$$

Theorem 2. *If both $\gamma^{(1)}$ and $\gamma^{(2)}$ are positive, then*

$$-\log \mathbf{P}(Z > x) \sim \gamma x$$

where $\gamma = \min(\gamma^{(1)}, \gamma^{(2)})$.

Sketch of Proof. Since $Z \ge Z^{(1)}$,

$$\limsup_{x\to\infty} \frac{-\log \mathbf{P}(Z > x)}{x} \le \gamma^{(1)}.$$

Similarly, consider an auxiliary system where service times in the first queue are replaced by zeros. Then we get a single server queue with service times $\{\sigma_n^{(2)}\}$ and interarrival times $\{t_n\}$. If we denote by $Z^{(2)}$ a stationary service time in this system, then $Z^{(2)} \le Z$ and, therefore,

$$\limsup_{x\to\infty} \frac{-\log \mathbf{P}(Z > x)}{x} \le \gamma^{(2)}.$$

Thus,

$$\limsup \frac{-\log \mathbf{P}(Z > x)}{x} \le \gamma.$$

To obtain the lower bound, we take a sufficiently large K (see the proof of Theorem 1) and an auxiliary upper single server queue. If we let

$$\gamma^* = \sup\{u \ : \ \varphi_{\sigma^*}(u)\varphi_{t^*}(-u) \le 1\},$$

then one can show that
(a)$\liminf \frac{-\log \mathbf{P}(Z>x)}{x} \ge \gamma^*$,
(b) one can choose γ^* as close to γ as possible.

Exercise 8. Complete the proof of Theorem 2.

Remark. The exact asymptotics for $\mathbf{P}(Z > x)$ in the heavy tail case have been found in [3].

References

[1] Baccelli, F. and Brémaud, P. (2003) *Elements of Queueing Theory.* Springer, Berlin.

[2] Baccelli, F. and Foss, S. (1995) On the Saturation Rule for the Stability of Queues. *J. Appl. Probab.* **23**, n.2, 494-507.

[3] Baccelli F, and Foss, S. (2004) Moments and tails in monotone-separable stochastic networks. *Ann. of Appl. Probab.* **14**, 612-650.

[4] Borovkov, A.A. (1998) *Ergodicity and Stability of Stochastic Processes.* Wiley, New York.

[5] Borovkov, A.A. and Foss, S.G. (1992) Stochastically recursive sequences and their generalizations. *Sib. Adv. Math.* **2**, n.1, 16-81.

[6] Brandt, A., Franken, P. and Lisek, B. (1992) *Stationary Stochastic Models.* Wiley, New York.

[7] Diaconis, P. and Freedman, P. (1999) Iterated random functions. *SIAM Review* **41** 45-76.

[8] Loynes, R.M. (1962) The stability of queues with non-independent interarrival and service times. *Proc. Cambridge Phil. Soc.* **58**, 497-520.

[9] Meyn, S. and Tweedie, R. (1993) *Markov Chains and Stochastic Stability.* Springer, New York.

[10] Whitt, W. (2002) *Stochastic-Process Limits.* Springer, New York.