J.F. Gentleman and R.C. Mullin (1989): The Distribution of Occurrence of Nucleotide Subsequences, Based on Their Overlap Capability. *Biometrics*, 45, pp. 35–52.

L.J. Guibas and O.M. Odlyzko (1981): String Overlaps, Pattern Matching and Nontransitive Games. *Journal of Combinatorial Theory, Ser. A*, pp. 183–208.

J. Kleffe and M. Borodovsky (1992): First and second moment counts of words in random texts generated by Markov chains. *Computer Applications in Biological Sciences* (CABIOS), 8, pp. 433–441.

P.A. Pevzner (1992): Nucleotide sequences versus Markov models. *Computers and Chemistry*, 16, pp. 103–106.

B. Prum, F. Rodolphe and E. de Turckheim (1995): Finding Words with Unexpected Frequencies in Deoxyribonucleic Acid Sequences. *Journal of the Royal Statistical Society*, **B**, 57, pp. 205–220.

G. Reinert, S. Schbath and M. Waterman (2000): Probabilistic and Statistical Properties of Words: An Overview. *Journal of Computational Biology*, vol. 7, pp. 1–46.

S. Robin and J.J. Daudin (1999): Exact Distribution of Word Occurrences in a Random Sequence of Letters. *Journal of Applied Probability*, 36, pp. 179–193.

J. Rudander (1996): *On the first occurrence of a given pattern in a semi-Markov process.* Uppsala Dissertations in Mathematics 2. Department of mathematics, Uppsala university.

- Estimating probabilities from scarce data

A. Nádas (1991): Good, Jelinek, Mercer, and Robbins on Turing's Estimate of Probabilities. *American Journal of Mathematical and Management Sciences*, 11, pp. 299–308.

Chapter 10

Hidden Markov Models: an Overview

10.1 Introduction

10.1.1 The HMM

A *hidden Markov model* (HMM) (earlier also known as a *probabilistic function of a Markov chain*, or as a *Markov source*, or as a *Markov regime model*) is a stochastic process generated by two interrelated probabilistic mechanisms. These are an underlying Markov chain with a finite number of states, and a set of random functions, each associated with its respective state. At discrete instants of time the process is assumed to be in some state and an observation is generated by the random function corresponding to the current state. The underlying Markov chain then changes its state according to its transition matrix.

The observer sees only the output of the random functions associated with each state and cannot directly observe the states of the underlying Markov chain. Hence the Markov chain is hidden but the name for the model family is *hidden Markov model.*

In principle, the outputs from the states of the hidden Markov chain may be multivariate random processes having some continuous joint probability density function. In this text we shall, however, limit the treatment to output processes with a discrete finite alphabet.

10.1.2 The Thumb Tack Again

Let us recall modelling of a sequence of flips of a thumb tack proposed and discussed by D. Heckerman (see Chapter 3). The model entertained there was a sequence of Bernoulli random variables that were conditionally independent given the chance of success. Next we make a variation of the theme as in (Rabiner 1989).

Let us now imagine a sequence of flips performed by a person who stands behind a barrier (like a curtain) through which we cannot see what is happening. The individual behind the curtain does not tell us what is exactly going on there but reports only the results of each flip of the thumb tack.

Thus a sequence of *hidden* flips of the thumb tack is being performed with the observation sequence consisting of a sequence of reported heads (H) and tails (T). Let us suppose that a series reported to us would be

$$HTHHHTTTTTTTTTTTHHHTTTTTTTTTTHH.$$

We might pay attention to the long uninterrupted runs of tails. That might be modeled by means of a scenario, where we think of two different techniques of tossing a thumb tack, the first suitably adjusted to produce mostly T and the other one more or less a random flip of thumb tack. The choice between the techniques might be done by some method of randomization before every flip.

10.2 Standard HMM

10.2.1 Definition and notations

An HMM can be seen as a model family for a sequence of symbols from an alphabet $\mathcal{O} = \{o_1, o_2, \ldots o_K\}$. The model is based on the idea of a hidden sequence of Markovian state transitions. More formally an HMM is characterized by (I)−(III) in the following:

(I) **Hidden Markov Chain** A Markov chain $\{X_n\}_{n=0}^{\infty}$ assuming values in a finite state space $S = \{1, 2, \ldots, J\}$ with J states. The conditional probabilities

$$a_{i|j} = P(X_n = j | X_{n-1} = i), \qquad n \geq 1, \qquad i, j \in S \tag{2.1}$$

are assumed to be time-homogeneous. The transition matrix is designated by

$$A = \left(a_{i|j}\right)_{i=1,j=1}^{J,J} \tag{2.2}$$

with the familiar constraints

$$a_{i|j} \geq 0, \sum_{j=1}^{J} a_{i|j} = 1.$$

At time $n = 0$ the state X_0 is specified by the probability distribution $\pi_j(0) = P(X_0 = j)$ with

$$\pi(0) = (\pi_1(0), \ldots, \pi_J(0)).$$

(II) **Observable Random Process** A random process $\{Y_n\}_{n=0}^{\infty}$ with a finite state space $\mathcal{O} = \{o_1, o_2, \ldots o_K\}$, where K need not equal J. The processes $\{Y_n\}_{n=0}^{\infty}$ and $\{X_n\}_{n=0}^{\infty}$ are for any fixed n related by the conditional probability distributions

$$b_j(k) = P(Y_n = o_k | X_n = j).$$

We set

$$B = \{b_j(k)\}_{j=1,k=1}^{J,K}.$$

We shall call this the *emission probability matrix*. This is another stochastic matrix in the sense that

$$b_j(k) \geq 0, \qquad \sum_{k=1}^{K} b_j(k) = 1.$$

(III) **Conditional Independence**

Assumption 10.2.1 For any sequence of states $j_0 j_1 \ldots j_n$ the probability of the sequence $o_0 o_1 \ldots o_n$ is

$$P(Y_0 = o_0, \ldots, Y_n = o_n \mid X_0 = j_0, \ldots, X_n = j_n, B) = \prod_{l=0}^{n} b_{j_l}(l). \tag{2.3}$$

In other words
The emitted symbols are conditionally independent given the state sequence. ∎

The postulate (III) is crucial for all the mathematical developments that are to follow. The detailed consequences of combining (III) and the Markov property will be established in Chapter 13. The assumption (III) is sometimes made fairly tacitly in tutorials and applied literature on HMM. The process $\{Y_n\}_{n=0}^{\infty}$ can always be represented as function of a Markov chain (see exercise) and is in general not a Markov chain.

By virtue of these assumptions we may write the *joint probability* of $o_0 \dots o_n$ and $j_0 \dots j_n$ as

$$\begin{aligned} P(Y_0 &= o_0, \dots, Y_n = o_n, X_0 = j_0, \dots, X_n = j_n; A, B, \pi(0)) \\ &= P(Y_0, \dots, Y_n \mid X_0, \dots, X_n, B) \cdot P(X_0, \dots, X_n, A, \pi(0)) \\ &= \pi_{j_0}(0) \cdot \prod_{l=0}^{n} b_{j_l}(l) \cdot \prod_{l=1}^{n} a_{j_{l-1}|j_l}, \end{aligned}$$

where the last equality follows by (2.3) above and by (1.9) from Chapter 7. Rearranging the last expression gives

$$\begin{aligned} P(Y_0 &= o_0, \dots, Y_n = o_n, X_0 = j_0, \dots, X_n = j_n; A, B, \pi(0)) \\ &= \pi_{j_0}(0) b_{j_0}(0) \cdot \prod_{l=1}^{n} a_{j_{l-1}|j_l} b_{j_l}(l). \end{aligned}$$

Then we obtain the joint probability of $o_0 \dots o_n$ as a marginal distribution by summing over all possible paths of the state sequence. This is written as:

$$P(Y_0, \dots, Y_n; A, B, \pi(0)) = \sum_{j_o=1}^{J} \cdots \sum_{j_n=1}^{J} \pi_{j_0}(0) b_{j_0}(0) \prod_{l=1}^{n} a_{j_{l-1}|j_l} b_{j_l}(l). \quad (2.4)$$

Hence the finite dimensional distributions of $\{Y_n\}_{n=1}^{\infty}$ are **fully specified** by our choice of the stochastic matrices A, B and of the initial distribution $\pi(0)$. Therefore we may use the compact notation for the model

$$\lambda = (A, B, \pi(0)).$$

Then we have

THE MODEL FAMILY:

CONDITIONED ON $\lambda = (A, B, \pi(0))$,

THE STRING $\mathbf{o} = o_0 \dots o_n$ HAS THE PROBABILITY

$$\begin{aligned} P(\mathbf{o}) &= P(Y_0 = o_0, \dots, Y_n = o_n; \lambda) \\ &= \sum_{j_o=1}^{J} \cdots \sum_{j_n=1}^{J} P(Y_0 = o_0 \dots, Y_n = o_n, X_0 = j_0, \dots, X_n = j_n; \lambda), \end{aligned}$$

where

$$\begin{aligned} P(Y_0 &= o_0, \dots, Y_n = o_n, X_0 = j_0, \dots, X_n = j_n; \lambda) \\ &= \pi_{j_0}(0) \cdot \prod_{l=0}^{n} b_{j_l}(l) \cdot \prod_{l=1}^{n} a_{j_{l-1}|j_l}. \end{aligned}$$

■

Remark 10.2.1 (Initial Emission) The various contributions and tutorials on HMM contain one minor discrepancy in detail with respect to the the formulation of $P(\mathbf{o})$ in (2.5). Levinson et al. (1983) and Baum's and Petrie's papers in the 1960's write

$$P(Y_1 \dots, Y_n, X_0, \dots, X_n; \lambda) = \pi_{j_0}(0) \cdot \prod_{l=1}^{n} b_{j_l}(l) \prod_{l=1}^{n} a_{j_{l-1}|j_l}, \quad (2.5)$$

which means that there is no emission in the initial state or that the initial state is a *silent state*. Another way to interpret this is that the initial symbol is always known to be the same. Our formula (2.5) follows the presentation in Rabiner's tutorial (1989). Thus care is required in reading and in detailed comparison of various formulas in the sequel and in the references. ■

Remark 10.2.2 Jelinek (1999) expresses the hidden Markov models in terms of an emission distribution that is a function of the state transition. Formally this means that $b_j(k) = P(Y_n = o_k | X_n = j)$ is replaced by

$$b_{i|j}(k) = P(Y_n = o_k | X_{n-1} = i, X_n = j).$$

This constitutes, however, a completely equivalent formulation to that used above, as is easily seen (Jelinek 1999, p. 36).

10.2.2 Generative Modelling of Sequences

Phrases like 'probabilistic mechanisms' and similar used above can be seen as evocative of the following generator of sequences in $\mathcal{O}^{N+1}$:

1) Choose an initial state $X_0 = j_0$ according to the distribution $\pi(0)$;

2) Set $n = 0$.

3) Choose $Y_n = o_k$ according to the symbol probability distribution in the state j_n which is $b_{j_n}(k)$;

4) Transit to a new state $X_{n+1} = j_{n+1}$ according to the state probability distribution $a_{j_n|j_{n+1}}$;

5) Set $n \to n+1$ and return to step 3) if $n \leq N$, otherwise terminate.

The above procedure can be used as both generator of observations and as a model for how a given sequence $\mathbf{o} = o_0 \dots o_n \in \mathcal{O}^{N+1}$ was generated by an appropriate HMM.

10.3 Examples and Applications

Example 10.3.1 (Finite Mixtures) The marginal distribution for any Y_n is by the preceding definition of HMM

$$P(Y_n = o_k) = \sum_{j=1}^{J} P(Y_n = o_k | X_n = j) \cdot \pi_j(n).$$

Assume that all the rows of A are identical, that is

$$a_{i|j} = w_j \text{ for } j = 1, 2, \dots, J, \text{ for all } i.$$

Then $\pi(n) = (w_1, w_2, \dots, w_J)$ for all $n \geq 1$ and for any $\pi(0)$, as is immediately seen, recalling $\pi(n) = \pi(0) \cdot A^n$ from Chapter 7. In other words, for $n \geq 1$

$$P(Y_n = o_k) = \sum_{j=1}^{J} P(Y_n = o_k | X_n = j) \cdot w_j,$$

where $P(Y_n = o_k | X_n = j)$ does not depend on n. The assumption of identical rows of A means that X_n are independent random variables and the assumption of conditional independence in (2.3) is the assumption of pairwise independence for (X_n, Y_n). But this is nothing else than the family of finite mixture of distribution models presented in Chapter 4. ∎

Example 10.3.2 (Burst Noise) Consider a Markov chain $\{X_n\}_{n=0}^{\infty}$ with the binary state space $S = \{b, g\}$ and the transition matrix

$$A = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}.$$

This Markov chain is hidden in another process $\{Y_n\}_{n=0}^{\infty}$ with the state space $\mathcal{O} = \{0, 1\}$ and such that the assumption (2.3) is valid. The process $\{Y_n\}_{n=0}^{\infty}$ has the emission probability matrix

$$B = \begin{pmatrix} \rho & 1-\rho \\ 1 & 0 \end{pmatrix},$$

where $0 \leq \rho \leq 1$. This example is owed to Gilbert (1960). Gilbert considered this scheme for modelling burst noise in binary communications channels. In that context we think of Z_n as a received bit, S_n as a transmitted bit and Y_n as a noise bit with

$$Z_n = S_n +_2 Y_n,$$

where $+_2$ is addition modulo two, ($1 +_2 1 = 0 +_2 0 = 0$ and $1 +_2 0 = 0 +_2 1 = 1$). If the Markov chain assumes the state g (for 'good'), then the emission probabilities are such that only digital zeroes will be emitted and there will be no transmission errors. In the state b (bad) the errors will be interspersed with error-free transmissions. A graphical recapitulation of the model generating Y_n is Figure 10.1. ∎

Example 10.3.3 (Noisy Markov Chains) A Markov chain $\{X_n\}_{n=0}^{\infty}$, with the binary digits $S = \{0, 1\}$ as the state space, has the transition matrix

$$A = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}. \tag{3.1}$$

We observe another process $\{Y_n\}_{n=0}^{\infty}$ with the state space $\mathcal{O} = \{0, 1\}$ such that

$$\begin{aligned} P(Y_0 &= o_0, \dots, Y_n = o_n | X_0 = j_0, \dots, X_n = j_n; B) \\ &= \prod_{l=0}^{n} \epsilon^{|o_l - j_l|} \cdot (1-\epsilon)^{1-|o_l - j_l|}. \end{aligned}$$

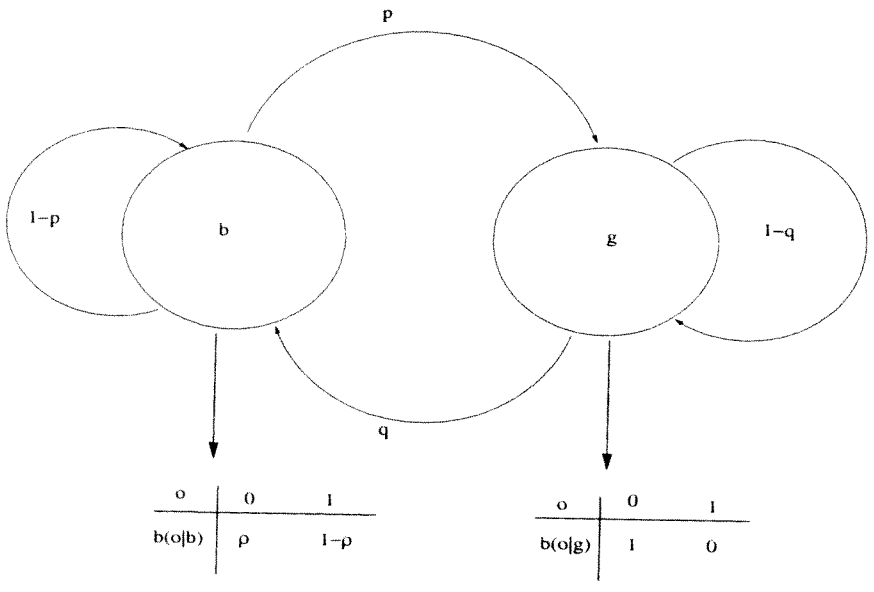

Figure 10.1: Burst noise

Clearly the assumption 10.2.1 is satisfied and when this setting is completed by some initial distribution $\pi(0)$, we have specified a hidden Markov model. The HMM has the emission probability matrix

$$B = \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix}.$$

This is a Markov chain observed in noise, where the probability of error in observation is ϵ. A system of random variables (X_n, Y_n) having these distributions is

$$X_{n+1} = X_n +_2 W_n$$
$$Y_n = X_n +_2 V_n,$$

where $+_2$ is addition modulo two and $\{W_n\}_{n=0}^{\infty}$ is a sequence of independent Bernoulli random variables with probability of success $= p$, and $\{V_n\}_{n=0}^{\infty}$ is a sequence of independent Bernoulli random variables with probability of success $= \epsilon$. We assume here that $\{W_n\}_{n=0}^{\infty}$ and $\{V_n\}_{n=0}^{\infty}$ are independent of each other. ∎

Example 10.3.4 (A Null Model in Sequence Analysis) In biological sequence analysis the following model turns out to play a certain role to be clarified later. A Markov chain $\{X_l\}_{l=0}^{\infty}$ has the state space $\{G, F\}$ and the following transition matrix

$$A = \begin{pmatrix} 1-\alpha & \alpha \\ 0 & 1 \end{pmatrix}.$$

The chain starts always in state G and ends (is absorbed) in the state F. The state F is *a silent state*, which means that no symbol is emitted there. For all l with $X_l = G$, a symbol o_l from a finite alphabet $\mathcal{O}$ is emitted. ∎

10.4 Influence Diagrams, Nonstandard HMM

A standard hidden Markov model can be regarded as the graph in Figure 10.2. This kind of graph is known as an *influence diagram* or a *belief network* (Smith 1989). Each node in the graph represents a random variable describing the state X_n or the observation Y_n at some time n. This is quite different from the state diagrams exemplified by, e.g., Figure 10.1. In Figure 10.2 the edges (arrows) represent direct influences. The graph expresses the model assumptions. The Markov assumption states that if we know which state j is visited at time n, no other information from the past is relevant about the future. In the Figure the variable X_n separates the previous variables from the future: by removing X_n from the graph, the variables X_m $(m > n)$ become disconnected from X_m $(m < n)$.

In terms of influence diagrams one can easily grasp and/or develop other than standard hidden Markov models. Examples of non-standard models are given in Figure 10.3, which is an autoregressive HMM, Figure 10.4, a coupled HMM, and Figure 10.5, a factor HMM. The key to reading the Figures is to see circles as containing states of a hidden chain and rectangles as containing the emitted symbols at respective times. A detailed study of factor HMMs as in Figure in 10.5 is given in (Ghahramani and Jordan 1997). Boys et al. (2000) apply autoregressive HMM in biological sequence analysis.

Given the standard hidden Markov model family above, there are evidently three basic problems of interest that must be solved for the model to be useful in applications to sequence data. These problems are the following (Rabiner 1989, Juang & Rabiner 1991):

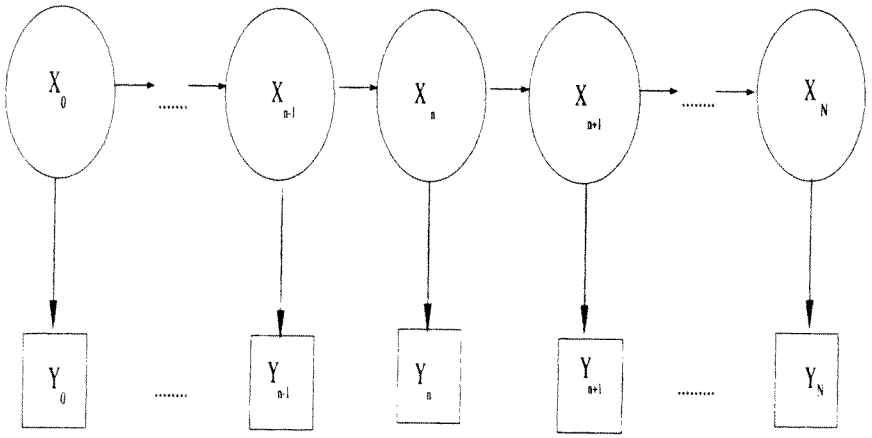

Figure 10.2: Standard HMM: Influence diagram

(1) **The Evaluation or Scoring Problem**
The real concern in the evaluation problem is computational efficiency. Without complexity constraints one can straightforwardly evaluate

$$P(Y_0 = o_0, \ldots, Y_n = o_n; \lambda)$$

directly as this probability is defined in (2.5). Since the summation involves J^{n+1} possible sequences, the total computational requirements are of the order $2(n+1) \cdot J^{n+1}$ operations. The need to compute without the exponential growth in computation is the first basic problem of HMM. The solution is known as the *forward–backward procedure* derived in Chapter 14 using the results in Chapter 13.

(2) **The Decoding or Alignment Problem**
We are often interested in uncovering the most likely state sequence that led to the observed sequence $(o_0 \ldots o_n)$. In the context of hidden Markov models for biological sequence families, see Chapter 12, this is the *alignment problem*. There are, of course, several ways to define the decoding criterion. Finding the sequence $j_0^* \ldots j_n^*$ that maximizes

$$P(X_0 = j_0, \ldots, X_n = j_n, Y_0 = o_0, \ldots, Y_n = o_n; \lambda)$$

for a fixed observed sequence $o_0 \ldots o_n$ is in practice the most frequently used criterion, since it can be implemented by the *Viterbi algorithm*,

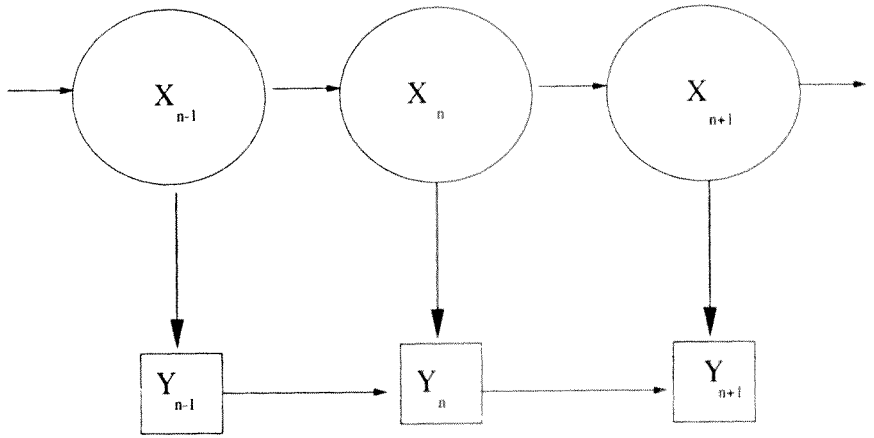

Figure 10.3: Autoregressive HMM: Influence diagram

actually already present the computation of the evolutionary distance in Chapter 5 and derived independently from scratch again in Chapter 14.

(3) **The Estimation or Training Problem**
Given an observed sequence $\mathbf{o} = o_0 \ldots o_n$, the estimation problem involves finding the model

$$\lambda = (A, B, \pi(0))$$

that specify the model most likely to generate the given sequence $o_0 \ldots o_n$ now also called a *training sequence*. In solving the estimation problem we shall analyse:

1) Maximum likelihood $\widehat{\lambda}_{\mathrm{ML}}$ is defined by.

$$\widehat{\lambda}_{\mathrm{ML}} = \arg\max_{\lambda} P(Y_0 = o_0, \ldots, Y_n = o_n; \lambda).$$

The probability in the right hand side is given in (2.5). This expression can have various degrees of complexity depending on the actual expressions for the conditional distributions in B. $\widehat{\lambda}_{\mathrm{ML}}$ is for standard HMM computed by using the *Baum–Welch algorithm*,

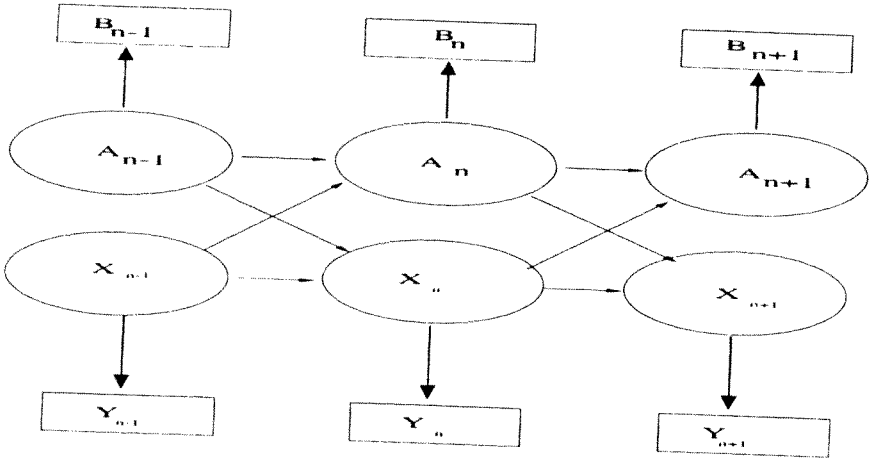

Figure 10.4: Coupled HMM: Influence diagram

which is based on the same procedures as the EM algorithm for finite mixtures in Chapter 4.

The properties of the Baum–Welch algorithm for computing $\widehat{\lambda}_{\mathrm{ML}}$ are treated in extensive detail in Chapters 15 and 16. The property of *asymptotic consistency*, i.e., that of proving the convergence of $\widehat{\lambda}_{\mathrm{ML}}$ as $n \to \infty$ to the postulated 'true model' within the family, is proved for ergodic HMM in Chapter 17.

The quantities computed in the Baum–Welch learning are fundamental even for the other techniques (i.e., **2**)-**4**) below) for estimating λ.

2) MAP learning

$$\widehat{\lambda}_{\mathrm{MAP}} = \arg\max_{\lambda} P\left(Y_0 = o_0, \ldots, Y_n = o_n; \lambda\right) \phi\left(\lambda\right)$$

where $\phi\left(\lambda\right)$ is a prior density. The version of MAP using the Dirichlet priors is given in Chapter 16.

3) Minimum discrimination information is owed to (Ephraim et al. 1989) and is based on minimization of a suitable Kullback distance.

4) Smooth learning algorithms A smooth reparametrization of λ is made to use the method of gradient descent. The technique

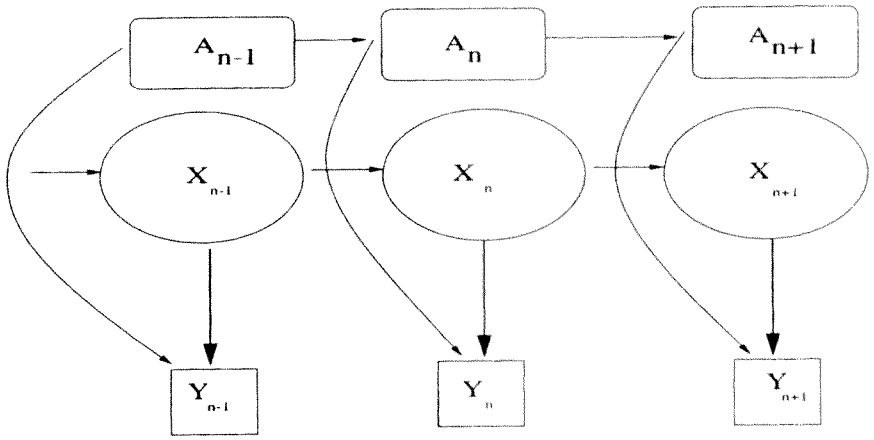

Figure 10.5: Factor HMM: Influence diagram

is developed in (Baldi et al. 1994) and will be presented in an exercise to Chapter 16.

5) Viterbi training If we set

$$V^{\star}(\lambda) = \max_{j_0 \cdots j_n} P\left(X_0 = j_0, \ldots, X_n = j_n; Y_0 = o_0 \ldots, Y_n = o_n; \lambda\right)$$

and try to find λ so as to maximize $V^{\star}(\lambda)$, then we may talk about Viterbi training. The method has been developed and analyzed under a different name in (Juang and Rabiner 1990) and will be treated by exercises in Chapters 15 and 16.

Against this background the hidden Markov models in Figures 10.3–10.5 are of real interest only if there exist effective algorithmic solutions for the three problems stated above. A condition for computable evaluation and learning in standard and nonstandard HMM is found in (Lucke 1996).

Smyth et al. (1997) introduce a general framework of graphical models, 'probabilistic independence networks', by means of which one can derive the computational algorithms for scoring and alignment problems in a quite different manner than done in this text. The technique of probabilistic independence networks can also be used to treat scoring and alignment for a number of non-standard HMM's.

In (Kung 1993) the *pattern completion problem* is also introduced and discussed. By this is meant the estimation of a randomly missing symbol in $\mathbf{o}$. The solution can be found by means of the Viterbi algorithm.

10.4.1 A Fourth Problem

If the states of the Markov chain in HMM have no physical or biological interpretation, there would seem to exist a fourth problem. This is that HMM's with different 'topologies' or architectures might be used to model the same sequence of data. How an architecture could be learned from data does not seem to have been widely studied and will not be discussed in this text. However, in principle, the method of Bayesian model selection should be applicable up to a certain degree. There are heuristics for trimming of architecture in HMM implemented in some of the software used in bioinformatics.

10.5 Finite State Stochastic Machines

10.5.1 Notations and Terminology

The definitions of probabilistic automata or finite state stochastic machines are variants or generalizations of HMM. References for basic properties of stochastic machines are (Ericson 1972) or (Paz 1971). The data compression model of (Ott 1967) is based on finite state stochastic machines. In fact, the models in (Churchill 1989) are stochastic sequential machines.

An (autonomous) *stochastic Mealy machine* $\mathbf{M_e}$ is the triple

$$\mathbf{M_e} = \{\mathcal{S}, \mathcal{O}, f_{Y_n,X_n|X_{n-1}}(o, s^{'}|s)\}, \tag{5.2}$$

where

$$s, s^{'} \in \mathcal{S} = \{1, 2, \cdots, J\}$$

is the *state alphabet* of $\mathbf{M_e}$ and

$$z \in \mathcal{O} = \{z_1, z_2, \cdots, z_K\}$$

is the *output alphabet* of $\mathbf{M_e}$. The probability

$$f_{Y_n,X_n|X_{n-1}}(o, s^{'}|s) = P(Y_n = o, X_n = s^{'}|X_{n-1} = s), \tag{5.3}$$

is in this context called the 'state $\mapsto$ output and state $\mapsto$ next state' map of the machine, where Y_n is the *output* of $\mathbf{M_e}$, X_n is the *state* of $\mathbf{M_e}$.

A stochastic machine can in this way be thought of being 'realized' by a factorization of the map (5.3) rewritten by the chain rule of conditional probability (2.16) in Chapter 1 as

$$f_{Y_n,X_n|X_{n-1}}(o, s^{'}|s) = f_{Y_n|X_{n-1}}(o|s) \cdot f_{X_n|Y_n,X_{n-1}}(s^{'}|o, s). \tag{5.4}$$

A *stochastic machine with probabilistic output* is obtained assuming in (5.4) that the current output symbol does not influence the next state, or

$$f_{Y_n,X_n|X_{n-1}}(o, s^{'}|s) = f_{Y_n|X_{n-1}}(o|s) \cdot f_{X_n|X_{n-1}}(s^{'}|s). \tag{5.5}$$

The state process X_n in a stochastic machine with probabilistic output is a Markov chain. Note that this differs formally from our definition of an HMM in the appearance of the one step delay in the emission probability distributions. The formula for the joint probability of the output of a stochastic machine is

$$P(Y_1 = o_{i_1}, \cdots, Y_m = o_{i_m}) = \sum_{i_0 \in \mathcal{S}} \sum_{j_1 \in \mathcal{S}} \cdots \sum_{j_m \in \mathcal{S}} \pi_{i_0} \cdot \mathbf{\Phi}, \tag{5.6}$$

where

$$\mathbf{\Phi} := \prod_{k=1}^{m} f_{Y_n,X_{k+1}|X_k}(o_{i_k}, j_k|j_{k-1}). \tag{5.7}$$

10.5.2 Carlyle's Matrix Formula

We introduce, for any finite state stochastic machine with a probabilistic output, cf. (5.5), the matrix $\mathbf{P}_Y(o_k)$ with the elements

$$(\mathbf{P}_Y(o_k))_{ij} := f_{Y_n|X_{n-1}}(o_k|j) \cdot f_{X_n|X_{n-1}}(j|i). \tag{5.8}$$

Hence, as $a_{i|j} = f_{X_n|X_{n-1}}(j|i)$,

$$\mathbf{P}_Y(o_k) = A\mathbf{F}(o_k), \tag{5.9}$$

where $\mathbf{F}(z_k)$ is the diagonal matrix

$$\mathbf{F}(o_k) := \begin{pmatrix} f_{Y_n|X_{n-1}}(o_k|1) & 0 & \cdots & 0 \\ 0 & f_{Y_n|X_{n-1}}(o_k|2) & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & f_{Y_n|X_{n-1}}(o_k|J) \end{pmatrix}. \tag{5.10}$$

Thus (5.6) becomes the Carlyle matrix formula

$$P(Y_1 = o_{i_1}, \cdots, Y_m =_{i_m}) = \pi(0) \prod_{l=1}^{m} \mathbf{P}_Y(o_{i_l})\mathbf{1} = \pi(0) \prod_{l=1}^{m} A\mathbf{F}(o_{i_l})\mathbf{1}, \quad (5.11)$$

where $\mathbf{1}$ is the $J|1$ column vector of ones, see (Carlyle 1963). We shall in the sequel prove a version of this formula for HMM by means of the forward and backward variables.

10.6 Exercises

1. Consider Gilbert's channel, Example 10.3.2 above. Assume that the Markov chain $\{X_n\}_{n=0}^{\infty}$ is *stationary*. Show that the probability of error $P(e)$ in transmission, i.e.,

$$P(e) = P(Z_n \neq S_n)$$

is

$$P(e) = (1-\rho) \cdot \frac{q}{q+p}.$$

2. The process $\{Y_n\}_{n=0}^{\infty}$ is a function of a Markov chain $\{X_n\}_{n=0}^{\infty}$

$$Y_n = g(X_n), \quad (6.1)$$

where $g(\cdot)$ is a map from S to $\mathcal{O}$, where the number of symbols in $\mathcal{O}$ is smaller than the number of symbols in S. Show that $\{X_n, Y_n\}_{n=0}^{\infty}$ is an HMM in the sense of the definition above. What is the emission probability matrix B in this case ?

3. **HMM as a Function of a Markov Chain**
Let $\{X_n\}_{n=0}^{\infty}$ and $\{Y_n\}_{n=0}^{\infty}$ be an HMM and let $\{X_n\}_{n=0}^{\infty}$ be stationary. We define a state space Z using the Cartesian product

$$Z = \mathcal{S} \times \mathcal{O} = \{1, \ldots, J\} \times \{o_1, o_2, \ldots o_K\}.$$

Let us also define the new Markov process $\{Z_n\}_{n=0}^{\infty}$ with values in Z by means of the transition probability matrix (a $J \cdot K \times J \cdot K$ matrix)

$$a_{(i,o_k)|(j,o_l)} = a_{i|j} \cdot b_j(o_l) = P(Z_{n+1} = (j, o_l) | Z_n = (i, o_k))$$

Let us next define a map g from Z onto $\mathcal{O}$ as

$$g((i, o_k)) = o_k$$

and let us define a new random process $\{Y_n'\}_{n=0}^{\infty}$, which is a function of a Markov chain, by

$$Y_n' = g(Z_n), \qquad n = 0, \ldots, .$$

Show that $\{Y_n'\}_{n=0}^{\infty}$ and $\{Y_n\}_{n=0}^{\infty}$ are probabilistically equivalent in the sense that

$$P(Y_0 = o_0, Y_1 = o_1, \ldots, Y_n = o_n) = P\left(Y_0' = o_0, Y_1' = o_1, \ldots, Y_n' = o_n\right)$$

for any sequence $o_0 o_1 \ldots o_n$.

4. **Naive Rule of Decoding**
Markov chain $\{X_n\}_{n=0}^{\infty}$, with the binary digits $S = \{0, 1\}$ as the state space, has the transition matrix

$$A = \begin{pmatrix} \frac{1+p}{2} & \frac{1-p}{2} \\ \frac{1-p}{2} & \frac{1+p}{2} \end{pmatrix}, \quad (6.2)$$

where $0 \leq p < 1$. The initial distribution is the distribution

$$\pi = \left(\frac{1}{2}, \frac{1}{2}\right).$$

We observe another process $\{Y_n\}_{n=0}^{\infty}$ with the state space $\mathcal{O} = \{0, 1\}$ and the emission probability matrix

$$B = \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix}$$

and assuming conditional independence:

$$\begin{aligned} P(Y_0 &= o_{j_0}, \ldots, Y_n = o_{j_n} | X_0 = j_0, \ldots, X_n = j_n; B) \\ &= \prod_{l}^{n} \epsilon^{|o_{j_l} - j_l|} (1-\epsilon)^{1-|o_{j_l} - j_l|}. \end{aligned}$$