Notes on Sequence Modelling

G.A. Jarrad

July 24, 2015

1 Random Sequence Processes

Consider a random process R, graphically depicted in Figure 1.1, that generates arbitrary sequences of values of the form $\vec{r}_n = (r_1, r_2, \ldots, r_n)$, where the length of any particular sequence is governed by a random variable N. Let $\vec{R}_N = (R_1, R_2, \ldots, R_N)$ denote the corresponding sequence of random variables, where R_t denotes the t-th discrete stage in the sequence.

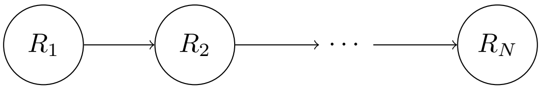

Figure 1.1: *A random process R for generating sequences of arbitrary length N. The arrows indicate transitions from one stage in the sequence to the next.*

We assume that each R_t randomly takes some discrete or continuous value $r_t \in \mathcal{R}$, and hence the probability (or probability density) of observing a particular sequence $\vec{r}_n$ of length n is given by

$$p(\vec{R}_N = \vec{r}_n) \quad = \quad p(N = n)\, p(\vec{R}_n = \vec{r}_n)\,, \tag{1.1}$$

where

$$p(\vec{R}_n = \vec{r}_n) \quad = \quad p(R_1 = r_1, \ldots, R_n = r_n)\,. \tag{1.2}$$

In practice, this definition presupposes that we know we have observed a *complete* sequence that was initiated at stage 1 and terminated at stage n. Suppose instead that the sequence $\vec{r}_n$ was observed one stage at a time. How do we know if the underlying process has actually terminated, or will instead continue to generate another observed value r_{n+1}, leading to the extended sequence $\vec{r}_{n+1}$? Similarly, how do we know that the first observed value r_1 was not in fact part of a longer, unobserved sequence of values $(\ldots, r_0, r_1, \ldots)$?

In order to handle such difficulties, we consider any arbitrary sequence $\vec{r}_n$ to be *incomplete*, and explicitly denote the corresponding, complete sequence as $\langle\vec{r}_n\rangle$. Additionally, we introduce the notion of *partially complete* sequences, defining a *start sequence* to be a sequence that has a definite start but an indefinite end, denoted by $\langle\vec{r}_n]$, and futher defining an *end sequence* to be a sequence that has a definite end but an indefinite start, denoted by $[\vec{r}_n\rangle$.

Under this augmented notation, knowledge about the start of a sequence can be encapsulated in a random indicator variable ι_t, which takes on the value 1 if R_{t+1} is definitely the first stage in the sequence, or the value 0 if it is not. Similarly, the random indicator variable τ_t takes on the value 1 if R_{t-1} is definitely the last stage in the sequence, or the value 0 if it is not. We assume that the random process R only ever produces complete sequences described by $\langle\vec{R}_N\rangle$, independently of the observation process, which might provide partial or complete sequences of values. Notionally, the indicators ι_0 and τ_{N+1} can be thought to correspond to pseudo-stages 0 and $N+1$, such that the generated sequence $\langle\vec{R}_N\rangle$ is initiated at stage 0 and terminated at stage $N+1$. This augmented random process is depicted in Figure 1.2.

The probability of a given complete sequence $\langle\vec{r}_n\rangle$ is now defined as

$$p(\langle\vec{r}_n\rangle) \quad = \quad p(\iota_0 = 1, \tau_1 = 0, R_1 = r_1, \ldots, \tau_n = 0, R_n = r_n, \tau_{n+1} = 1)\,, \tag{1.3}$$

such that

$$p(N = n) \quad = \quad p(\iota_0 = 1, \tau_1 = 0, \ldots, \tau_n = 0, \tau_{n+1} = 1)\,. \tag{1.4}$$

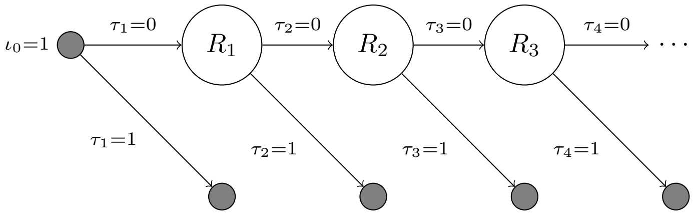

Figure 1.2: *A random process for generating complete sequences of arbitrary length, with explicit stages for sequence initiation and termination. Multiple arrows exiting from a node indicate different possible (mutually exclusive) stage transition pathways.*

This takes the form of a generalised Bernoulli sequence. Likewise, the probability of a given start sequence $\langle\vec{r}_n]$ is defined as

$$p(\langle\vec{r}_n]) \quad = \quad p(\iota_0\!=\!1, \tau_1\!=\!0, R_1 = r_1, \ldots, \tau_n\!=\!0, R_n = r_n)\,, \tag{1.5}$$

and the probability of the end sequence $[\vec{r}_n\rangle$ is

$$p([\vec{r}_n\rangle) \quad = \quad p(\tau_1\!=\!0, R_1 = r_1, \ldots, \tau_n\!=\!0, R_n = r_n, \tau_{n+1}\!=\!1)\,. \tag{1.6}$$

In the special case where we know in advance that a start sequence definitely does not terminate at stage $n+1$, we may instead write

$$p(\langle\vec{r}_n!) \quad = \quad p(\iota_0\!=\!1, \tau_1\!=\!0, R_1 = r_1, \ldots, \tau_n\!=\!0, R_n = r_n, \tau_{n+1}\!=\!0)\,. \tag{1.7}$$

Likewise, if an end sequence definitely does not initiate at stage 0, then

$$p(!\vec{r}_n\rangle) \quad = \quad p(\iota_0\!=\!0, \tau_1\!=\!0, R_1 = r_1, \ldots, \tau_n\!=\!0, R_n = r_n, \tau_{n+1}\!=\!1)\,. \tag{1.8}$$

The four remaining types of sequences, namely $[\vec{r}_n]$, $!\vec{r}_n]$, $[\vec{r}_n!$ and $!\vec{r}_n!$, can be similarly defined.

2 Markov Sequence Processes

In Section 1 we defined a random process R and the sequences it generates. We now assume that the process is also *causal*, meaning that each stage of a sequence, including the termination stage, depends only on the preceding stages. This causal process, depicted in Figure 2.1, is simply the random process from Figure 1.2 with additional, explicit dependencies (in the form of dashed arrows). Hence, under the Markov assumption of conditional independence, the causal sequence process leads to the fully-dependent, conditional model

$$\begin{aligned} p(\langle\vec{r}_n\rangle) \quad = \quad & p(\iota_0 = 1) \\ & \times \prod_{t=1}^{n} \Big\{ p(\tau_t\!=\!0 \mid \iota_0\!=\!1, \vec{\tau}_{t-1}\!=\!\vec{0}, \vec{R}_{t-1}\!=\!\vec{r}_{t-1})\, p(R_t\!=\!r_t \mid \iota_0\!=\!1, \vec{\tau}_t\!=\!\vec{0}, \vec{R}_{t-1}\!=\!\vec{r}_{t-1}) \Big\} \\ & \times p(\tau_{n+1}\!=\!1 \mid \iota_0\!=\!1, \vec{\tau}_n\!=\!\vec{0}, \vec{R}_n\!=\!\vec{r}_n)\,. \end{aligned} \tag{2.1}$$

The related models for partially complete or incomplete sequences can be similarly obtained by suitably modifying the corresponding boundary conditions for ι_0 and τ_{n+1} — refer to Section 1.

In practice, the causal model is usually simplified further by dropping some of the explicit (dashed) dependencies. For example, one might limit the conditionality on past values to a maximum number m of depenencies. This leads to the so-called *m-th order Markov model*. An example from the realm of natural language understanding is the lexicographical analysis of the character sequences of words using bigrams (pairs of adjacent characters, corresponding to $m = 1$), and trigrams (triples of adjacent characters, corresponding to $m = 2$), et cetera.

In the special case of $m = 1$, the first-order Markov model takes on the restricted conditional form

$$\begin{aligned} p(\langle\vec{r}_n\rangle) \quad = \quad & p(\iota_0 = 1)\, p(\tau_1\!=\!0 \mid \iota_0\!=\!1) p(R_1\!=\!r_1 \mid \iota_0\!=\!1, \tau_1\!=\!0) \\ & \times \prod_{t=2}^{n} \{ p(\tau_t\!=\!0 \mid R_{t-1}\!=\!r_{t-1})\, p(R_t\!=\!r_t \mid \tau_t\!=\!0, R_{t-1}\!=\!r_{t-1}) \} \\ & \times p(\tau_{n+1}\!=\!1 \mid R_n\!=\!r_n)\,. \end{aligned} \tag{2.2}$$

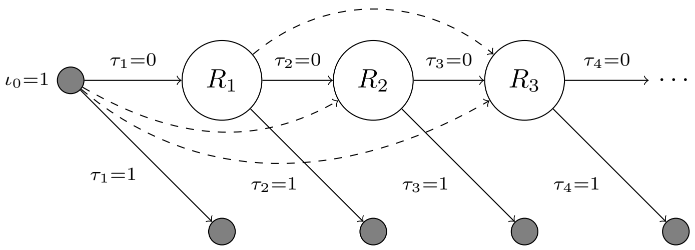

Figure 2.1: *A fully-dependent, causal process for generating complete sequences of arbitrary length. Solid arrows indicate possible stage transitions. Both dashed arrows and solid arrows indicate parent–child dependencies, such that the child node is conditionally dependent on the parent and all other previous nodes.*

This is just the strict Markov interpretation of the random process depicted in Figure 1.2, where each stage directly depends only on the previous stage *and* on the transition path between the two adjacent stages.

3 Stateful Markov Sequence Processes

Consider the first-order Markov process R depicted in Figure 1.2. Suppose now that the random variable R_t at stage t can be decomposed into the tuple $R_t = (S_t, X_t)$, where S_t is a random variable taking values $s_t \in \mathcal{S}$, and X_t is a random variable taking values $x_t \in \mathcal{X}$. We may call S_t the *state* of the process at stage t, and X_t its *value*. As is usual, we presuppose that the stage transitions in the sequence generating process are primarily between states, e.g. from S_{t-1} to S_t, and hence it follows that the value is generated after the state has been determined, i.e. X_t depends upon S_t. Keeping to the first-order Markov interpretation of stage-to-stage dependencies leads to the *stateful* process depicted in Figure 3.1, with full cross-dependencies between (S_t, X_t) and (S_{t+1}, X_{t+1}).

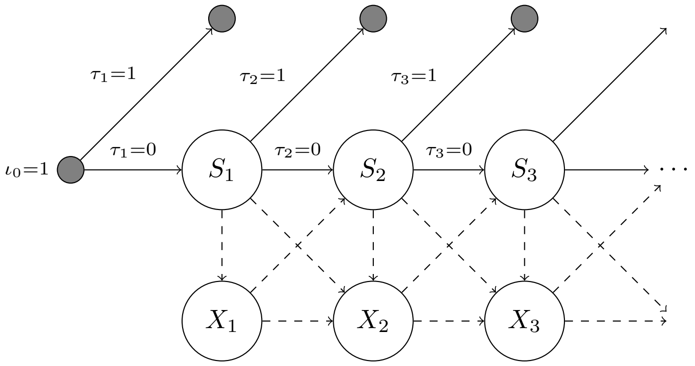

Figure 3.1: *A random process for generating complete, stateful sequences of arbitrary length, with explicit cross-dependencies between the states and values of adjacent stages.*

Hence, the fully-structured stateful model is now given by

$$\begin{aligned} p(\iota_0, \vec{S}_n, \vec{X}_n, \vec{\tau}_{n+1}) \;\; &= \;\; p(\iota_0)\, p(\tau_1 \,|\, \iota_0)\, p(S_1 \,|\, \iota_0, \tau_1)\, p(X_1 \,|\, S_1) \\ & \quad \times \prod_{t=2}^{n} \left\{ p(\tau_t \,|\, S_{t-1})\, p(S_t \,|\, \tau_t, S_{t-1}, X_{t-1})\, p(X_t \,|\, S_t, S_{t-1}, X_{t-1}) \right\} \\ & \quad \times p(\tau_{n+1} \,|\, S_n)\,. \end{aligned} \tag{3.1}$$

Conditioning the state S_t on both the previous state S_{t-1} and its value X_{t-1} can be useful in some circumstances, e.g. in sequence classification problems. However, due to the increased complexity of such models, it is more usual to further restrict the stateful process by also imposing the first-order Markov assumption at the level of the state–value dependencies themselves. In terms of the process depicted in Figure 3.1, this means retaining only direct node-to-node dependencies, rather than stage-to-stage dependencies. Hence, this restricted process, depicted in Figure 3.2, corresponds to the sequence model

$$\begin{aligned} p(\iota_0, \vec{S}_n, \vec{X}_n, \vec{\tau}_{n+1}) &= p(\iota_0)\, p(\tau_1 \,|\, \iota_0)\, p(S_1 \,|\, \iota_0)\, p(X_1 \,|\, S_1) \\ &\quad \times \prod_{t=2}^{n} \left\{ p(\tau_t \,|\, S_{t-1})\, p(S_t \,|\, S_{t-1})\, p(X_t \,|\, S_t) \right\}\, p(\tau_{n+1} \,|\, S_n)\,. \end{aligned} \tag{3.2}$$

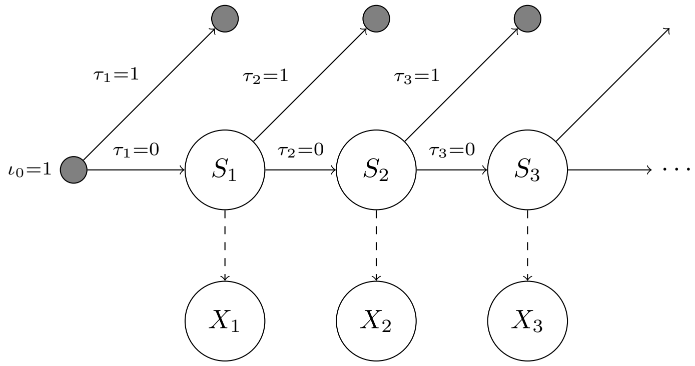

Figure 3.2: *A first-order Markov process for generating complete, stateful sequences of arbitrary length.*

4 Discrete-state Sequence Models

Consider the stateful, first-order Markov process depicted by Figure 3.2. Let us now restrict our attention to the class of models where the state S_t at any stage t may now only take *discrete* values in the set $\mathcal{S} = \{\sigma_1, \sigma_2, \ldots, \sigma_S\}$. Thus, the sequence model (3.2) may be parameterised in terms of arbitrary states $\vec{s}_n = (\sigma_{i_1}, \sigma_{i_2}, \ldots, \sigma_{i_n})$, where the factors modelling the probabilities within stage t depend only upon σ_{i_t}, and the factors modelling the transition between stage t and stage $t+1$ depend only upon σ_{i_t} and $\sigma_{i_{t+1}}$. For convenience, let us further assume that the process does not generate zero-length sequences[1], so that $p(\tau_1 = 1 \,|\, \iota_0 = 1) = 1$. Consequently, model (3.2) can now be explicitly rewritten in terms of the parameters $\theta = (\vec{\pi}, \vec{\bar{\pi}}, \Gamma, \vec{\omega})$, such that the probability of an arbitrary, complete sequence is given by

$$p(\langle \vec{s}, \vec{x} \rangle \,|\, \theta) = \pi_{1,i_1}\, o_{1,i_1}(x_1) \left\{ \prod_{t=1}^{n-1} \bar{\omega}_{t,i_t}\, \Gamma_{t,i_t,i_{t+1}}\, o_{t,i_{t+1}}(x_{t+1}) \right\} \omega_{n,i_n}\,, \tag{4.1}$$

where $n = |\vec{x}| = |\vec{s}|$. The initial state S_1 of the sequence at stage $t = 1$ is governed by the general parameter

$$\pi_{t,i} = p(\iota_{t-1}{=}1 \,|\, \theta)\, p(S_t{=}\sigma_i \,|\, \iota_{t-1}{=}1, \theta)\,, \tag{4.2}$$

and the terminal and non-terminal transitions between stages t and $t+1$ are governed by

$$\omega_{t,i} = p(\tau_{t+1} = 1 \,|\, S_t{=}\sigma_i, \theta)\,, \tag{4.3}$$

$$\bar{\omega}_{t,i} = p(\tau_{t+1} = 0 \,|\, S_t{=}\sigma_i, \theta) = 1 - \omega_{t,i}\,, \tag{4.4}$$

respectively. The possible transitions between the states S_t and S_{t+1} of consecutive stages t and $t+1$ are governed by

$$\Gamma_{t,i,j} = p(S_{t+1}{=}\sigma_j \,|\, S_t{=}\sigma_i, \theta)\,. \tag{4.5}$$

Clearly, a state transition from S_t to S_{t+1} implies a non-terminal transition from stage t to $t+1$ in the sequence. This is explicitly modelled by the presence of the $\bar{\omega}_{t,i_t}$ term, which is interpreted to mean that S_t is a non-terminal state. It then follows that S_{t+1} is a non-initial state which, if modelled explicitly, would be represented by a corresponding $\bar{\pi}_{t+1,i_{t+1}}$ term, defined by

$$\bar{\pi}_{t,i} = p(\iota_{t-1}{=}0 \,|\, \theta)\, p(S_t{=}\sigma_i \,|\, \iota_{t-1}{=}0, \theta)\,. \tag{4.6}$$

[1] Zero-length sequences are usually undetectable in practice, unless the generating process explicitly indicates the start and end of such sequences. Even if clearly delimited, zero-length sequences may typically be ignored in most applications.

However, $\bar{\pi}_{t+1,i_{t+1}}$ is already implicit the explicit transition term $\Gamma_{t,i_t,i_{t+1}}$, since we cannot count state S_{t+1} twice. This fact is important when it comes to parameter estimation (see Section **??**).

Lastly, the probabilities of observed values at stage t are governed by

$$o_{t,i}(x) \quad = \quad p(X_t\!=\!x \mid S_t\!=\!\sigma_i,\theta) \qquad \forall x \in \mathcal{X}\,. \tag{4.7}$$

We do not, however, explicitly declare the parameterisation structure of this likelihood model (see Section **??** for a plausible model if X_t takes discrete values). It suffices for our calculations that each $o_{t,i_t}(x_t)$ is available when required.

Note that all other types of incomplete sequences (see Section 1) are handled similarly to complete sequences by modifying the first and/or last terms of model (4.1). For example, $!\vec{s},\vec{x}\rangle$ is modelled by replacing π_{1,i_1} by $\bar{\pi}_{1,i_1}$, whereas $[\vec{s},\vec{x}\rangle$ is modelled by replacing π_{1,i_1} by $\breve{\pi}_{1,i_1}$, where

$$\begin{aligned}
\breve{\pi}_{t,i} &= \pi_{t,i} + \bar{\pi}_{t,i} \\
&= p(\iota_{t-1}\!=\!1)\,p(S_t\!=\!\sigma_i \mid \iota_{t-1}\!=\!1,\theta) + p(\iota_{t-1}\!=\!0)\,p(S_t\!=\!\sigma_i \mid \iota_{t-1}\!=\!0,\theta) \\
&= p(\iota_{t-1}\!=\!1, S_t\!=\!\sigma_i \mid \theta) + p(\iota_{t-1}\!=\!0, S_t\!=\!\sigma_i \mid \theta) \\
&= p(S_t\!=\!\sigma_i \mid \theta)\,.
\end{aligned} \tag{4.8}$$

Similarly, $\langle\vec{s},\vec{x}!$ is modelled by replacing ω_{n,i_n} by $\bar{\omega}_{n,i_n}$, whereas $\langle\vec{s},\vec{x}]$ is modelled by replacing ω_{n,i_n} by $\breve{\omega}_{n,i_n}$, where

$$\begin{aligned}
\breve{\omega}_{t,i} &= \omega_{t,i} + \bar{\omega}_{t,i} \\
&= p(\tau_{t+1}\!=\!1 \mid S_t\!=\!\sigma_i,\theta) + p(\tau_{t+1}\!=\!0 \mid S_t\!=\!\sigma_i,\theta) \\
&= 1\,.
\end{aligned} \tag{4.9}$$

4.1 Posterior State Prediction

A *hidden-state* Markov model (or HMM) results from a stateful Markov sequence process like Figure 3.2, from which the values $\vec{x}$ are observed but the states $\vec{s}$ are not. The true state values are then said to be *missing* or *hidden*, and must be estimated from the observed data. In particular, a known problem is to deduce the state S_t given $\vec{x}$, at each stage $t = 1, 2, \ldots, n$. This is accomplished via the *forward–backward algortithm*, which uses the causal nature of the process to notionally partitition the sequence into past and present stages $1, 2, \ldots, t$ and future stages $t+1, t+2, \ldots, n$. The standard algorithm is modified here to include the stage-by-stage probabilities of sequence termination. Thus, the posterior state probabilities at stage t for a complete sequence $\langle\vec{x}\rangle$ are given by

$$\begin{aligned}
\gamma_{t,i} &= p(S_t\!=\!\sigma_i \mid \langle\vec{x}\rangle,\theta) \\
&= \frac{p(S_t\!=\!\sigma_i, \langle\vec{x}\rangle \mid \theta)}{p(\langle\vec{x}\rangle \mid \theta)} \\
&= \frac{p(S_t\!=\!\sigma_i, \langle\vec{x}_t] \mid \theta)\,p([\overleftarrow{x}_{t+1}\rangle \mid S_t\!=\!\sigma_i,\theta)}{\sum_{i'=1}^{S} p(S_t\!=\!\sigma_{i'}, \langle\vec{x}_t] \mid \theta)\,p([\overleftarrow{x}_{t+1}\rangle \mid S_t\!=\!\sigma_{i'},\theta)} \\
&= \frac{\alpha_{t,i}\,\beta_{t,i}}{\sum_{i'=1}^{S} \alpha_{t,i}\,\beta_{t,i}}\,,
\end{aligned} \tag{4.10}$$

where we have defined $\overleftarrow{x}_t = (x_t, x_{t+1}, \ldots, x_n)$ for all $t = 1, 2, \ldots, n$, with $n = |\vec{x}|$.

The *forward step*, which incorporates information about the initiation of the sequence, is recursively defined via

$$\begin{aligned}
\alpha_{t,i} &= p(S_t\!=\!\sigma_i, \langle\vec{x}_t] \mid \theta) \\
&= \sum_{j=1}^{S} p(S_{t-1}\!=\!\sigma_j, \langle\vec{x}_{t-1}] \mid \theta)\,p(\tau_t\!=\!0 \mid S_{t-1}\!=\!\sigma_j,\theta) \\
&\qquad \times\, p(S_t\!=\!\sigma_i \mid S_{t-1}\!=\!\sigma_j,\theta)\,p(X_t\!=\!x_t \mid S_t\!=\!\sigma_i,\theta) \\
&= \left\{\sum_{j=1}^{S} \alpha_{t-1,j}\,\bar{\omega}_{t-1,j}\,\Gamma_{t-1,j,i}\right\} o_{t,i}(x_t)\,,
\end{aligned} \tag{4.11}$$

for $t = 2, 3, \ldots, n$. The forward step commences with

$$\begin{aligned} \alpha_{1,i} &= p(S_1 = \sigma_i, \langle x_1] \,|\, \theta) \\ &= p(\iota_0 = 1)\, p(S_1 = \sigma_i \,|\, \iota_0 = 1, \theta)\, p(X_1 = x_1 \,|\, S_1 = \sigma_i, \theta) \\ &= \pi_{1,i}\, o_{1,i}(x_1)\,. \end{aligned} \tag{4.12}$$

Note that incompletely–initiated sequences such as $!\vec{x}]$ and $[\vec{x}]$ can also be handled by substituting $\bar{\pi}$ and $\breve{\pi}$ for π in α, thereby obtaining $\bar{\alpha}$ and $\breve{\alpha}$ respectively.

Consequently, we may predict S_t from a partially observed sequence $\langle\vec{x}_t]$ via

$$\begin{aligned} p(S_t = \sigma_i \,|\, \langle\vec{x}_t], \theta) &= \frac{p(S_t = \sigma_i, \langle\vec{x}_t] \,|\, \theta)}{\sum_{i'=1}^{S} p(S_t = \sigma_{i'}, \langle\vec{x}_t] \,|\, \theta)} \\ &= \frac{\alpha_{t,i}}{\sum_{i'=1}^{S} \alpha_{t,i'}}\,. \end{aligned} \tag{4.13}$$

Similarly, we may predict the next observation X_{t+1} via

$$\begin{aligned} p(X_{t+1} = x \,|\, \langle\vec{x}_t], \theta) &= \frac{p(\langle\vec{x}_t, x] \,|\, \theta)}{p(\langle\vec{x}_t] \,|\, \theta)} \\ &= \frac{\sum_{j=1}^{S} p(S_{t+1} = \sigma_j, \langle\vec{x}_t, x] \,|\, \theta)}{\sum_{i=1}^{S} p(S_t = \sigma_i, \langle\vec{x}_t] \,|\, \theta)} \\ &= \frac{\sum_{j=1}^{S} \left\{\sum_{i=1}^{S} \alpha_{t,i}\, \bar{\omega}_{t,i}\, \Gamma_{t,i,j}\right\} o_{t+1,j}(x)}{\sum_{i=1}^{S} \alpha_{t,i}}\,. \end{aligned} \tag{4.14}$$

The *backward step*, which incorporates information about the termination of the sequence, is now also recursively defined via

$$\begin{aligned} \beta_{t,i} &= p([\overleftarrow{x}_{t+1}\rangle \,|\, S_t = \sigma_i, \theta) \\ &= \sum_{j=1}^{S} p([\overleftarrow{x}_{t+2}\rangle \,|\, S_{t+1} = \sigma_j, \theta)\, p(X_{t+1} = x_{t+1} \,|\, S_{t+1} = \sigma_j, \theta) \\ &\qquad \times p(S_{t+1} = \sigma_j \,|\, S_t = \sigma_i, \theta)\, p(\tau_{t+1} = 0 \,|\, S_t = \sigma_i, \theta) \\ &= \left\{\sum_{j=1}^{S} \beta_{t+1,j}\, o_{t+1,j}(x_{t+1})\, \Gamma_{t,i,j}\right\} \bar{\omega}_{t,i}\,, \end{aligned} \tag{4.15}$$

for $t = n-1, n-2, \ldots, 1$. The backward step commences with

$$\beta_{n,i} = p(\tau_{n+1} = 1 \,|\, S_n = \sigma_i, \theta) = \omega_{n,i}\,. \tag{4.16}$$

Note that incompletely–terminated sequences such as $[\vec{x}!$ and $[\vec{x}]$ can also be handled by substituting $\bar{\omega}$ and $\breve{\omega}$ for ω in β, thereby obtaining $\bar{\beta}$ and $\breve{\beta}$ respectively.

The combination of the forward step with the backward step now enables us to use all of the information contained in the observed sequence, including its possible initiation and/or termination. Particularly, we can compute the joint probability of any observed sequence, as a prelude to estimating the state of each stage from equation (4.10). For example, observe that

$$\begin{aligned} p(\langle\vec{x}! \,|\, \theta) &= \sum_{i=1}^{S} p(S_t = \sigma_i, \langle\vec{x}! \,|\, \theta) \\ &= \sum_{i=1}^{S} p(S_t = \sigma_i, \langle\vec{x}_t] \,|\, \theta)\, p([\overleftarrow{x}_{t+1}! \,|\, S_t = \sigma_i, \theta) \\ &= \sum_{i=1}^{S} \alpha_{t,i} \bar{\beta}_{t,i}\,, \end{aligned} \tag{4.17}$$

for all $t = 1, 2, \ldots, n$.

Finally, the forward–backward calculations also enable us to compute the posterior probabilities of the joint states of stages t and $t+1$. For example, given the observed, complete sequence $\langle\vec{x}\rangle$, we obtain

$$\begin{aligned} \xi_{t,i,j} &= p(S_t\!=\!\sigma_i, S_{t+1}\!=\!\sigma_j \mid \langle\vec{x}\rangle, \theta) \\ &= \frac{p(S_t\!=\!\sigma_i, S_{t+1}\!=\!\sigma_j, \langle\vec{x}\rangle \mid \theta)}{p(\langle\vec{x}\rangle \mid \theta)}\,, \end{aligned} \tag{4.18}$$

where

$$\begin{aligned} p(S_t\!=\!\sigma_i, S_{t+1}\!=\!\sigma_j, \langle\vec{x}\rangle \mid \theta) &= p(S_t\!=\!!\sigma_i, \langle\vec{x}_t] \mid \theta)\, p(\tau_{t+1}\!=\!0 \mid S_t\!=\!\sigma_i, \theta) \\ &\quad \times\, p(S_{t+1}\!=\!\sigma_j \mid S_t\!=\!\sigma_i, \theta)\, p([\overleftarrow{x}_{t+1}] \mid S_{t+1}\!=\!\sigma_j, \theta) \\ &= \alpha_{t,i}\, \bar{\omega}_{t,i}\, \Gamma_{t,i,j}\, \beta_{t+1,j}\,, \end{aligned} \tag{4.19}$$

and thus

$$\begin{aligned} p(\langle\vec{x}\rangle \mid \theta) &= \sum_{i'=1}^{S} \sum_{j'=1}^{S} p(S_t\!=\!\sigma_i, S_{t+1}\!=\!\sigma_j, \langle\vec{x}\rangle \mid \theta) \\ &= \sum_{i'=1}^{S} \sum_{j'=1}^{S} \alpha_{t,i'}\, \bar{\omega}_{t,i'}\, \Gamma_{t,i',j'}\, \beta_{t+1,j'} \\ &= \sum_{i'=1}^{S} \alpha_{t,i'}\, \beta_{t,i'}\,, \end{aligned} \tag{4.20}$$

as expected.

5 Hidden-state Parameter Estimation

Suppose now that the hidden-state Markov model (**??**) implicitly depends upon some parameter θ, the value of which needs to be estimated from observed data. In particular, let us assume that $\theta = (\Pi, \Gamma, \Omega)$, where $\Pi = (\vec{\pi}^{+}, \vec{\pi}^{-})$ and $\Omega = (\vec{\omega}^{+}, \vec{\omega}^{-})$ respectively specify the possible distributions of the initial and final states of an arbitrary sequence (defined in further detail below), and Γ represents the *stationary* distribution of state transitions between stages.

For convenience, we now suppose that the discrete set of possible states is given by $\mathcal{S} = \{\sigma_1, \sigma_2, \ldots, \sigma_S\}$. Then we may define the initial distributions of states via $\vec{\pi} = (\pi_1, \pi_2, \ldots, \pi_S)$ where

$$\pi_i^{+} = p(\iota_0\!=\!1 \mid \theta)\, p(S_1\!=\!\sigma_i \mid \iota_0\!=\!1, \theta)\,, \tag{5.1}$$

$$\pi_i^{-} = p(\iota_0\!=\!0 \mid \theta)\, p(S_1\!=\!\sigma_i \mid \iota_0\!=\!0, \theta)\,, \tag{5.2}$$

such that $\pi_i^{*} = \pi_i^{+} + \pi_i^{-} = p(S_1 = \sigma_i \mid \theta)$. Similarly, the final distributions of states are defined by $\vec{\omega} = (\omega_1, \ldots, \omega_S)$ where

$$\omega_i^{+} = p(\tau_{n+1} = 1 \mid S_n\!=\!\sigma_i, \theta)\,, \tag{5.3}$$

$$\omega_i^{-} = p(\tau_{n+1} = 0 \mid S_n\!=\!\sigma_i, \theta)\,, \tag{5.4}$$

such that $\omega_i^{*} = \omega_i^{+} + \omega_i^{-} = 1$. Note that Π and Ω describe the initial and terminal state distributions of the random process itself, not those of any observed sequences.

The last parameter of interest, specified by the matrix $\Gamma = [\Gamma_{i,j}]_{i,j=1}^{S}$, defines the state transitions

$$\Gamma_{i,j} = p(S_{t+1}\!=\!\sigma_j \mid S_t\!=\!\sigma_i, \theta)\,. \tag{5.5}$$

Observe that the assumption of stationarity implies that Γ is constant for all t.

Now, since the observed value sequence $\vec{x}_n$ is always here assumed to be known, we may for convenience define

$$o_{t,i} = p(X_t\!=\!x_t \mid S_t\!=\!\sigma_i, \theta)\,, \tag{5.6}$$

although we are not concerned here with the internal parameterisation structure of $o_{t,i}$ itself. Thus, the explicitly parameterised version of model (**??**) is given by

$$p(\iota_0, \vec{S}_n\!=\!\vec{s}_n, \vec{X}_n\!=\!\vec{x}_n, \tau_{n+1} \mid \theta) = \pi_{i_1} \prod_{t=1}^{n-1} \Gamma_{i_t,i_{t+1}} \prod_{t=1}^{n} o_{t,i_t}\, \omega_{i_n}\,, \tag{5.7}$$

where the unknown state sequence $\vec{s}_n$ corresponding to $\vec{x}_n$ is arbitrarily specified by $\vec{s}_n = (\sigma_{i_1}, \sigma_{i_2}, \ldots, \sigma_{i_n})$.

Let us now suppose that we have observed an ordered set of value sequences $\mathbb{X} = \{(\iota_0^{(d)}, \vec{x}^{(d)}, \tau_{n^{(d)}+1}^{(d)})\}_{d=1}^{D}$. Notionally, we may also define the correspondingly ordered set $\mathbb{S} = \{\vec{s}^{(d)}\}_{d=1}^{D}$ of arbitrary state sequences. Hence, under the assumption that the observed sequences are independent, the joint log-likelihood of the data is given by

$$\begin{aligned} L(\theta) &= \log p(\mathbb{S}, \mathbb{X} \mid \theta) \\ &= \log \prod_{d=1}^{D} p(\iota_0^{(d)}, \vec{s}^{(d)}, \vec{x}^{(d)}, \tau_{n^{(d)}+1}^{(d)} \mid \theta) \\ &= \sum_{d=1}^{D} \log p(\iota_0^{(d)}, \vec{s}^{(d)}, \vec{x}^{(d)}, \tau_{n^{(d)}+1}^{(d)} \mid \theta) \\ &= \sum_{d=1}^{D} L^{(d)}(\theta)\,, \end{aligned} \tag{5.8}$$

where

$$L^{(d)}(\theta) \;=\; \log \pi_{i_1^{(d)}}^{(d)} + \sum_{t=1}^{n^{(d)}-1} \log \Gamma_{i_t^{(d)}, i_{t+1}^{(d)}} + \sum_{t=1}^{n^{(d)}} \log o_{t, i_t^{(d)}} + \log \omega_{i_{n^{(d)}}^{(d)}}^{(d)}\,, \tag{5.9}$$

and $n^{(d)} = |\vec{x}^{(d)}|$.

However, recall that $\mathbb{S}$ is actually uknown. Hence, we take an expectation of the log-likelihood over all possible values of $\mathbb{S}$, namely[2]

$$\begin{aligned} Q(\theta) &= E_{\mathbb{S} \mid \mathbb{X}, \theta} \left[\log p(\mathbb{S}, \mathbb{X} \mid \theta)\right] \\ &= E_{\mathbb{S} \mid \mathbb{X}\theta} \left[\sum_{d=1}^{D} L^{(d)}(\theta)\right] \\ &= \sum_{d=1}^{D} E_{\mathbb{S} \mid \mathbb{X}, \theta} \left[L^{(d)}(\theta)\right] \\ &= \sum_{d=1}^{D} \sum_{i_1^{(d)}=1}^{S} \cdots \sum_{i_{n^{(d)}}^{(d)}}^{S} p(\vec{s}^{(d)} \mid \iota_0^{(d)}, \vec{x}^{(d)}, \tau_{n^{(d)}+1}^{(d)}, \theta)\, L^{(d)}(\theta)\,. \end{aligned} \tag{5.10}$$

In practice, it is difficult to optimise this nonlinear expression analytically. A feasible alternative is to iteratively apply the *expectation–maximisation* (EM) algorithm:

1. *Expectation step:* Compute the expected log-likelihood conditioned on a known parameter estimate $\hat{\theta}_k$, namely

$$\begin{aligned} Q(\theta, \hat{\theta}_k) &= E_{\mathbb{S} \mid \mathbb{X}, \hat{\theta}_k} \left[\log p(\mathbb{S}, \mathbb{X} \mid \theta)\right] \\ &= \sum_{d=1}^{D} \sum_{i_1^{(d)}=1}^{S} \cdots \sum_{i_{n^{(d)}}^{(d)}}^{S} p(\vec{s}^{(d)} \mid \iota_0^{(d)}, \vec{x}^{(d)}, \tau_{n^{(d)}+1}^{(d)}, \hat{\theta}_k)\, L^{(d)}(\theta)\,. \end{aligned} \tag{5.11}$$

2. *Maximisation step:* Obtain the optimal parameter estimate $\hat{\theta}_{k+1}$ that maximises the conditional expected log-likehood, namely

$$\hat{\theta}_{k+1} \;=\; \arg\max_{\theta} Q(\theta, \hat{\theta}_k)\,. \tag{5.12}$$

These two steps are iterated until $\hat{\theta}_k$ has converged to a value $\hat{\theta}^*$ that maximises $L(\hat{\theta}^*) = Q(\hat{\theta}^*, \hat{\theta}^*)$.

blah about additivity

[2] Other expectations are possible, e.g. over the joint distribution $\mathbb{S}, \mathbb{X} \mid \theta$. This latter produces macro-averaged parameter estimates of the form $\sum_{d=1}^{D} \phi^{(d)} / \sum_{d=1}^{D} \psi^{(d)}$, whereas the discriminative distribution $\mathbb{S} \mid \mathbb{X}, \theta$ often leads to micro-averaged estimates of the form $\sum_{d=1}^{D} \phi^{(d)} / \psi^{(d)} / D$.

$$\begin{aligned}
\frac{\partial Q}{\partial \Gamma_{i,j}} &= \frac{\partial}{\partial \Gamma_{i,j}} \sum_{d=1}^{D} \sum_{i_1^{(d)}=1}^{S} \cdots \sum_{i_{n^{(d)}}^{(d)}}^{S} p(\vec{s}^{(d)} \mid \iota_0^{(d)}, \vec{x}^{(d)}, \tau_{n^{(d)}+1}^{(d)}, \hat{\theta}') \sum_{t=1}^{n^{(d)}-1} \log \Gamma_{i_t^{(d)}, i_{t+1}^{(d)}} \\
&= \sum_{d=1}^{D} \sum_{t=1}^{n^{(d)}-1} \sum_{i_1^{(d)}=1}^{S} \cdots \sum_{i_{n^{(d)}}^{(d)}}^{S} \delta(i_t^{(d)} = i)\delta(i_{t+1}^{(d)} = j) \frac{p(\vec{s}^{(d)} \mid \iota_0^{(d)}, \vec{x}^{(d)}, \tau_{n^{(d)}+1}^{(d)}, \hat{\theta}')}{\Gamma_{i,j}} \\
&= \sum_{d=1}^{D} \sum_{t=1}^{n^{(d)}-1} \sum_{i_1^{(d)}=1}^{S} \cdots \sum_{i_{n^{(d)}}^{(d)}}^{S} \delta(i_t^{(d)} = i)\delta(i_{t+1}^{(d)} = j) \frac{p(\vec{s}^{(d)} \mid \iota_0^{(d)}, \vec{x}^{(d)}, \tau_{n^{(d)}+1}^{(d)}, \hat{\theta}')}{\Gamma_{i,j}} \\
&= \frac{\sum_{d=1}^{D} \sum_{t=1}^{n^{(d)}-1} p(S_t = \sigma_i, S_{t+1} = \sigma_j \mid \iota_0^{(d)}, \vec{x}^{(d)}, \tau_{n^{(d)}+1}^{(d)}, \hat{\theta}')}{\Gamma_{i,j}} \\
&= \frac{\sum_{d=1}^{D} \sum_{t=1}^{n^{(d)}-1} \xi_t^{(d)}(\sigma_i, \sigma_j; \hat{\theta}')}{\Gamma_{i,j}} \qquad (5.13)
\end{aligned}$$

from equation (**??**). Now, subject to the constraint that $\sum_{j=1}^{S} \Gamma_{i,j} = 1$, we induce the appropriate Lagrangian multiplier to provide the proper normalisation, and hence derive that the optimal parameter estimate is given by

$$\hat{\Gamma}_{i,j}^{*} = \frac{\sum_{d=1}^{D} \sum_{t=1}^{n^{(d)}-1} \xi_t^{(d)}(\sigma_i, \sigma_j; \hat{\theta}')}{\sum_{j=1}^{S} \sum_{d=1}^{D} \sum_{t=1}^{n^{(d)}-1} \xi_t^{(d)}(\sigma_i, \sigma_j; \hat{\theta}')} = \frac{\sum_{d=1}^{D} \sum_{t=1}^{n^{(d)}-1} \xi_t^{(d)}(\sigma_i, \sigma_j; \hat{\theta}')}{\sum_{d=1}^{D} \sum_{t=1}^{n^{(d)}-1} \gamma_t^{(d)}(\sigma_i; \hat{\theta}')} \qquad (5.14)$$

from equation (**??**).