Notes on Sequence Modelling

G.A. Jarrad

January 28, 2018

1 Random Sequence Processes

Consider a random process R that generates arbitrary-length sequences of the form $\vec{R} = (R_1, R_2, \ldots, R_N)$, where N is a random variable governing the length of a sequence, and R_t is a random variable governing the value at *stage* t of the sequence. This sequence process is graphically depicted in Figure 1.1.

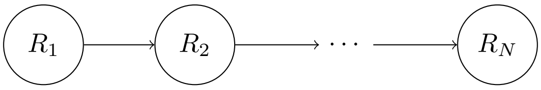

Figure 1.1: *A random process R for generating sequences of random length N. The arrows indicate transitions from one stage in the sequence to the next.*

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 $\vec{r}$ of length $n = |\vec{r}|$ is given by

$$p(\vec{R}{=}\vec{r}) \quad = \quad p(N = n)\, p(R_1{=}r_1, \ldots, R_n{=}r_n \mid N = n)\,. \tag{1.1}$$

In practice, this definition presupposes that we know we have observed a *complete* sequence that started at stage 1 and ended at stage n. Suppose instead that the sequence $\vec{r}$ 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}? Similarly, how do we know that the first observed value r_1 was not in fact part of a longer, unobserved sequence of values? We assume that the random process R only ever produces complete sequences, independently of the observation process, which might provide partial or complete sequences of values. Furthermore, if the random process does not signal the start and end of generated sequences, then an observed sequence might actually comprise multiple, contiguously generated subsequences.

In order to handle such difficulties, we consider any arbitrary sequence $\vec{r}$ by default to be *incomplete*, and explicitly denote the corresponding, complete sequence by $\langle\vec{r}\rangle$. We can now introduce the notion of *partially complete* sequences. Thus, a *start sequence* is a generated sequence with an observed (or definite) start (at stage 1) but an unobserved (or indefinite) end, i.e. it might or might not terminate at stage n. This is denoted by $\langle\vec{r}$ if we are truly uncertain as to the termination, or by $\langle\vec{r}]$ if we actually know that the generated sequence does not terminate at stage n. Similarly, an *end sequence* is a generated sequence with an observed end (at stage n) but an unobserved start, i.e. it might or might not have initiated at the observed stage 1. This is denoted by $\vec{r}\rangle$ if we are truly uncertain as to sequence initiation, or by $[\vec{r}\rangle$ if we actually know that the generated sequence was not initiated at stage 1. Clearly, we may also specify the remaining incomplete sequences, namely $[\vec{r}]$, $\vec{r}]$ and $[\vec{r}$.

Under this augmented notation, knowledge about the start of a sequence can be encapsulated in a random indicator variable ι_{t-1}, which takes on the value 1 if some observed r_t is definitely the first stage in the generated sequence, or the value 0 if it is not. Similarly, the random indicator variable τ_{t+1} takes on the value 1 if r_t is definitely the last stage in the generated sequence, or the value 0 if it is not. In general, these indicators allow us to handle the observation of possibly concatenated, multiple, generated sequences. From now on, however, we shall assume (unless otherwise stated) that we are dealing with a single, contiguous sequence. Thus, notionally, the indicators ι_0 and τ_{n+1} can be thought to correspond to pseudo-stages 0 and $n{+}1$, such that an arbitrary generated sequence is initiated at stage 0 and terminated at some random stage $N + 1$. This augmented random process is depicted in Figure 1.2.

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

$$p(\langle\vec{r}\rangle) \quad = \quad p(\iota_0{=}1, \tau_1{=}0, R_1 = r_1, \tau_2{=}0, \ldots, \tau_n{=}0, R_n = r_n, \tau_{n+1}{=}1)\,, \tag{1.2}$$

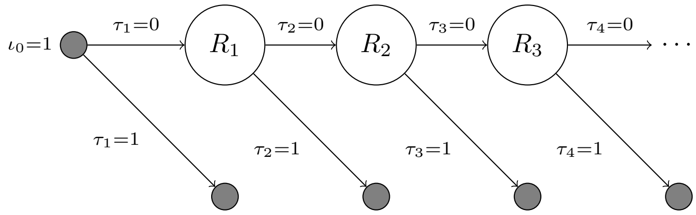

Figure 1.2: *A random process for generating complete sequences of random length, with explicit stages for sequence initiation and termination. Multiple arrows exiting from a node indicate different possible (mutually exclusive) stage transition pathways.*

such that

$$p(N\!=\!n) \quad = \quad p(\iota_0\!=\!1, \tau_1\!=\!0, \ldots, \tau_n\!=\!0, \tau_{n+1}\!=\!1)\,. \tag{1.3}$$

This has the form of a generalised Bernoulli sequence.

Note that when the context is clear, we may for convenience drop explicit mention of the random variable R_t. Similarly, we may denote $\iota_t = 1$ by ι_t^+, on the understanding that ι_t^- denotes the negation $\iota_t = 0$. Likewise, we may denote $\tau_t = 1$ by τ_t^+ and $\tau_t = 0$ by τ_t^-. Hence, it is plausible to simplify equation (1.2) as

$$p(\langle\vec{r}\rangle) \quad = \quad p(\iota_0^+, \tau_1^-, r_1, \ldots, \tau_n^-, r_n, \tau_{n+1}^+)\,. \tag{1.4}$$

Consequently, we may simplify the explicitly terminated process of Figure 1.2 to more resemble the implicitly terminatel process of Figure 1.1; the result is shown in Figure 1.3.

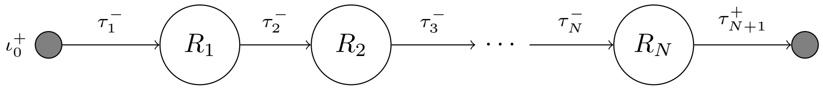

Figure 1.3: *A simplified represenation of a random process for generating complete sequences of random length N, with explicit stages for sequence initiation and termination, and explicit labelling of non-terminating transitions.*

We can now handle both completely observed and partially observed sequences by introducing indicators $\underline{\iota}$ and $\underline{\tau}$ to correspond to the start-of-sequence and end-of-sequence symbols, respectively. In particular, $\underline{\iota} = 1$ corresponds to '⟨' and $\underline{\iota} = 0$ corresponds to '['; when the start of the sequence is unknown, we let $\underline{\iota} = *$ (see Section 1.1). Likewise, $\underline{\tau} = 1$ corresponds to '⟩', $\underline{\tau} = 0$ corresponds to ']', and $\underline{\tau} = *$ corresponds to unknown sequence termination. Hence, in general, we write

$$\begin{aligned} p(\underline{\iota}, \vec{r}, \underline{\tau}) \quad &= \quad p(\iota_0\!=\!\underline{\iota}, \tau_1\!=\!0, R_1\!=\!r_1, \ldots, \tau_n\!=\!0, R_n\!=\!r_n, \tau_{n+1}\!=\!\underline{\tau}) \\ &= \quad p(\underline{\iota}, \tau_1^-, r_1, \ldots, \tau_n^-, r_n, \underline{\tau})\,. \end{aligned} \tag{1.5}$$

Note that, formally, this is equivalent to redefining the sequence as $\vec{r} = (\tau_1^-, r_1, \tau_2^-, r_2, \ldots, \tau_n^-, r_n)$. Informally, we take the specification of a non-empty vector $\vec{r}$ of contiguous values to automatically imply the existence of non-terminating transitions between stages. Thus, by convention, the τ_t^- terms are kept implicit when dealing with functions of $\vec{r}$ (e.g. the left-hand side of the above equation), and are only made explicit when dealing directly with functions of the expanded values $r_1, r_2, \ldots, r_n$ (e.g. the right-hand side of the above equation).

1.1 Missing Values

The main difference between a complete, generated sequence $\vec{r} = (r_1, \ldots, r_n)$ and the observed sequence of values, say[1] $\underline{\vec{r}} = (\underline{r}_1, \ldots, \underline{r}_n)$, is the possibility that some values were unobserved, i.e. either arbitrarily *missing* or systematically *hidden*. For convenience, let $\underline{r}_t = *$ denote the case where the value of the t-th stage is unobserved; recall from above that $\underline{\iota} = *$ or $\underline{\tau} = *$ if we do not kow whether or not we observed the start or end of the generated sequence, respectively. The '*' symbol is just a representational device

[1] We are ignoring the very real problem of aligning the observed values with the generated stages. This difficulty can be partially alleviated under the assumption of stationary distributions, such that each stage behaves like the previous one.

– its presence has no effect on the computed probabilities, other than to indicate that any associated variable should be marginalised out. Thus, for example:

$$p(\vec{r}) \;=\; p(*,\tau_1^-,r_1,\ldots,\tau_n^-,r_n,*) \;=\; p(\tau_1^-,r_1,\ldots,\tau_n^-,r_n)\,. \tag{1.6}$$

In practice, we allow for both observed values and missing values by introducing an indicator function $\delta(\cdot)$, where $\delta(x{=}y)=1$ if $x=y$ and $\delta(x{=}y)=0$ if $x\neq y$; by definition, we take $\delta(x{=}*)=1$. Hence, we obtain

$$p(\underline{\iota},\vec{r},\underline{\tau}) \;=\; \sum_{\iota_0=0}^{1}\delta(\iota_0{=}\underline{\iota})\sum_{\tau_{n+1}=0}^{1}\delta(\tau_{n+1}{=}\underline{\tau})\;p(\iota_0,\tau_1^-,r_1,\ldots,\tau_n^-,r_n,\tau_{n+1})\,. \tag{1.7}$$

In general, if the domain $\mathcal{R}$ is discrete, then the likelihood of an observed sequence $\underline{\vec{r}}$ is given by

$$\begin{aligned} p(\underline{\iota},\underline{\vec{r}},\underline{\tau}) \;=\; &\sum_{\iota_0=0}^{1}\delta(\iota_0{=}\underline{\iota})\sum_{r_1\in\mathcal{R}}\delta(r_1{=}\underline{r}_1)\cdots\sum_{r_n\in\mathcal{R}}\delta(r_n{=}\underline{r}_n)\sum_{\tau_{n+1}=0}^{1}\delta(\tau_{n+1}{=}\underline{\tau})\\ &p(\iota_0,\tau_1^-,r_1,\ldots,\tau_n^-,r_n,\tau_{n+1})\,. \end{aligned} \tag{1.8}$$

Alternatively, if $\mathcal{R}$ is continuous, then the likelihood becomes

$$\begin{aligned} p(\underline{\iota},\underline{\vec{r}},\underline{\tau}) \;=\; &\sum_{\iota_0=0}^{1}\delta(\iota_0{=}\underline{\iota})\sum_{\tau_{n+1}=0}^{1}\delta(\tau_{n+1}{=}\underline{\tau})\int_{\mathcal{R}}\delta(r_1{-}\underline{r}_1)\cdots\int_{\mathcal{R}}\delta(r_n{-}\underline{r}_n)\\ &p(\iota_0,\tau_1^-,r_1,\ldots,\tau_n^-,r_n,\tau_{n+1})\;dr_1\,dr_2\cdots dr_n\,, \end{aligned} \tag{1.9}$$

where $\delta(\cdot)$ is now the Dirac delta function, and where, by extension, we define $\delta(x-*)=1$. On the understanding that $\sum$ and $\delta(x=y)$ must be swapped respectively for $\int$ and $\delta(x-y)$ as needed for a continuous or semi-continuous domain, we may henceforth simply utilise the discrete form (1.8) without loss of generality.

1.2 Generic Forward–Backward Algorithm

The likelihood (1.8) of an observed sequence $\underline{\vec{r}}$ has been written in a computationally inefficient form, but can in practice be efficiently evaluated by nesting the summations, using a modification of the *forward–backward algorithm* to include knowledge of sequence initiation and termination. The precise details of these calculations depend upon the chosen factorisation of the probability model, which is itself a function of the explicit dependencies between various stages in the sequence. Such dependency modelling is dealt with further in Section 2.

Despite not knowing these dependencies in advance, however, the basic form of the forward–backward algorithm can still be formulated. The first requirement is that the sequence process be *causal*, meaning that each stage of a sequence depends only on preceding stages, and never on future stages. This causality allows us to partition a generated sequence into two parts at some arbitrary *pivot* stage t, as shown in Figure 1.4. The second requirement is that the dependence on past stages can be limited in scope to some arbitrary *historical* stage s, as also shown.

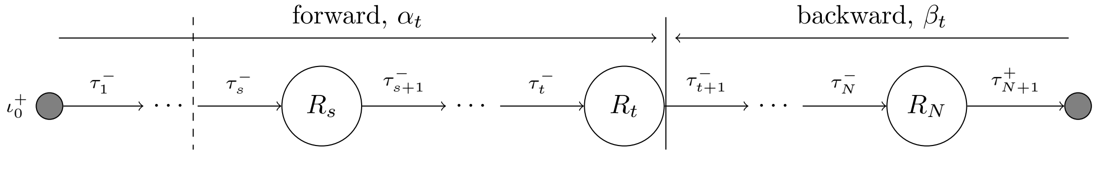

Figure 1.4: *Causality allows the sequence to be partitioned at some pivot stage t, thereby dividing the sequence into past and future stages. Further limitation of past dependencies to some historical stage s defines the active window for one step of the forward–backward algorithm.*

Let us now define the sub-sequence $\vec{r}_{s,t}=(r_s,r_{s+1},\ldots,r_t)$; by definition, $\vec{r}_{s,t}=(\,)$ if $s>t$. Furthermore, consider a concatenation operator '$\circ$', such that $\vec{r}_{s,k}\circ\vec{r}_{k+1,t}=\vec{r}_{s,t}$. Then observe, for a sufficiently long sequence (defined in Section 2), that

$$\begin{aligned}
p(\underline{\iota},\underline{\vec{r}},\underline{\tau}) &= p(\underline{\iota},\underline{\vec{r}}_{1,s-1}\circ\underline{\vec{r}}_{s,t}\circ\underline{\vec{r}}_{t+1,n},\underline{\tau}) \\
&= \sum_{r_s\in\mathcal{R}}\delta(r_s\!=\!\underline{r}_s)\cdots\sum_{r_t\in\mathcal{R}}\delta(r_t\!=\!\underline{r}_t)\,p(\underline{\iota},\underline{\vec{r}}_{1,s-1}\circ\vec{r}_{s,t}\circ\underline{\vec{r}}_{t+1,n},\underline{\tau}) \\
&= \sum_{r_s\in\mathcal{R}}\delta(r_s\!=\!\underline{r}_s)\cdots\sum_{r_t\in\mathcal{R}}\delta(r_t\!=\!\underline{r}_t)\,p(\underline{\iota},\underline{\vec{r}}_{1,s-1}\circ\vec{r}_{s,t})\,p(\underline{\vec{r}}_{t+1,n},\underline{\tau}\,|\,\underline{\iota},\underline{\vec{r}}_{1,s-1}\circ\vec{r}_{s,t}) \\
&= \sum_{r_s\in\mathcal{R}}\delta(r_s\!=\!\underline{r}_s)\cdots\sum_{r_t\in\mathcal{R}}\delta(r_t\!=\!\underline{r}_t)\,\alpha_t(\vec{r}_{s,t})\,\beta_t(\vec{r}_{s,t})\,, \qquad (1.10)
\end{aligned}$$

where

$$\alpha_t(\vec{r}_{s,t}) \;=\; p(\underline{\iota},\underline{\vec{r}}_{1,s-1}\circ\vec{r}_{s,t}) \;=\; p(\underline{\iota},\tau_1^-,\underline{r}_1,\ldots,\tau_{s-1}^-,\underline{r}_{s-1},\tau_s^-,r_s,\ldots,\tau_t^-,r_t) \qquad (1.11)$$

is the foward factor, and

$$\begin{aligned}
\beta_t(\vec{r}_{s,t}) &= p(\underline{\vec{r}}_{t+1,n},\underline{\tau}\,|\,\underline{\iota},\underline{\vec{r}}_{1,s-1}\circ\vec{r}_{s,t}) \\
&= p(\tau_{t+1}^-,\underline{r}_{t+1},\ldots,\tau_n^-,\underline{r}_n,\underline{\tau}\,|\,\underline{\iota},\tau_1^-,\underline{r}_1,\ldots,\tau_{s-1}^-,\underline{r}_{s-1},\tau_s^-,r_s,\ldots,\tau_t^-,r_t) \qquad (1.12)
\end{aligned}$$

is the backward factor. The entire forward pass of the forward–backward algorithm starts from some initial, historical stage t_0 and progressively computes α_t forward along the sequence for each applicable stage $t_0 \leq t \leq n$. Likewise, the backward pass starts at termination stage n, and computes β_t backwards along the sequence for each applicable stage $t_0 \leq t \leq n$. The precise details of these calculations rely upon the nature of the fine-grained dependencies, as disdcussed in the next section.

2 Markov Sequence Processes

In Section 1, we defined a causal random sequence process R, such 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 1.3 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

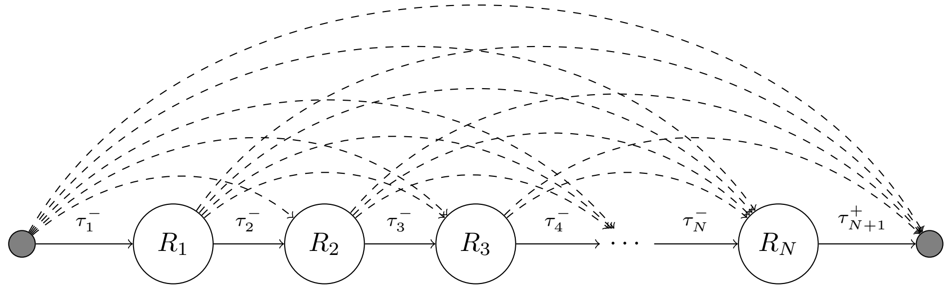

Figure 2.1: *A fully-dependent, causal process for generating complete, random sequences of random length N. Solid arrows indicate stage transitions. Both dashed arrows and solid arrows indicate parent–child dependencies, such that each stage is conditionally dependent on the preceding stages.*

$$\begin{aligned}
p(\iota_0,\vec{r},\tau_{n+1}) &= p(\iota_0,\tau_1^-,r_1,\tau_2^-,r_2,\ldots,\tau_n^-,r_n,\tau_{n+1}) \\
&= p(\iota_0)\,p(\tau_1^-\,|\,\iota_0)\,p(r_1\,|\,\iota_0,\tau_1^-)\,p(\tau_2^-\,|\,\iota_0,\tau_1^-,r_1)\,p(r_2\,|\,\iota_0,\tau_1^-,r_1,\tau_2^-) \\
&\quad\;\cdots p(\tau_n^-\,|\,\iota_0,\ldots,\tau_{n-1}^-,r_{n-1})p(r_n\,|\,\iota_0,\ldots,\tau_n^-)\,p(\tau_{n+1}\,|\,\iota_0,\ldots,r_n) \\
&= p(\iota_0)\left\{\prod_{t=1}^{n} p(\tau_t^-,r_t\,|\,\iota_0,\vec{r}_{1,t-1})\right\}p(\tau_{n+1}\,|\,\iota_0,\vec{r}_{1,n})\,. \qquad (2.1)
\end{aligned}$$

In practice, this fully-dependent model is considerably simplified by dropping some or even most of the explicit (dashed) dependencies. For example, one might limit the conditionality on past stages to a

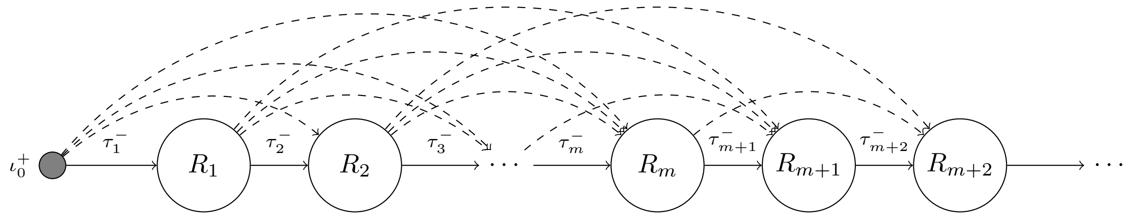

Figure 2.2: *An m-th order Markov sequence process of arbitrary length (here $n \geq m+2$).*

maximum of m depenencies. This leads to the so-called *m-th order Markov model*, shown in Figure 2.2. The corresponding likelihood model is given by

$$p(\iota_0, \vec{r}, \tau_{n+1}) \;=\; p(\iota_0) \left\{ \prod_{t=1}^{m} p(\tau_t^-, r_t \,|\, \iota_0, \vec{r}_{1,t-1}) \right\} \left\{ \prod_{t=m+1}^{n} p(\tau_t^-, r_t \,|\, \vec{r}_{t-m,t-1}) \right\} p(\tau_{n+1} \,|\, \vec{r}_{n-m+1,n}) \,, \tag{2.2}$$

for $n \geq m$.

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 $m = 2$), etc. The second-order Markov sequence process, for example, is depicted in Figure 2.3.

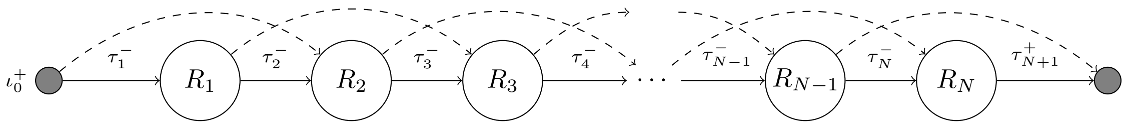

Figure 2.3: *A second-order Markov sequence process of random length N.*

In the special case of $m = 1$, the first-order Markov model, depicted in Figure 1.3, takes on the especially-simple conditional form of

$$p(\iota_0, \vec{r}, \tau_{n+1}) \;=\; p(\iota_0)\, p(\tau_1^-, r_1 \,|\, \iota_0) \left\{ \prod_{t=2}^{n} p(\tau_t^-, r_t \,|\, \tau_{t-1}^-, r_{t-1}) \right\} p(\tau_{n+1} \,|\, \tau_n^-, r_n) \,, \tag{2.3}$$

for $n > 0$.

2.1 Markov Forward–Backward Algorithm

The basic description of the generic forward–backward algorithm in Section 1.2 can now be refined under the restriction of the causal sequence process to an m-th order Markov process. Specifically, for any stage $1 \leq t \leq n$, we take the limiting historical stage to be $s = \max(1, t-m+1)$. Then, for a sufficiently long sequence (i.e. $n \geq m$), the forward factor (1.11) may be computed for stages $t = m, m+1, \ldots, n$ via

$$\begin{aligned} \alpha_t(\vec{r}_{t-m+1,t}) \;&=\; p(\underline{\iota}, \tau_1^-, \underline{r}_1, \ldots, \tau_{t-m}^-, \underline{r}_{t-m}, \tau_{t-m+1}^-, r_{t-m+1}, \ldots, \tau_t^-, r_t) \\ &=\; \sum_{\iota_0=0}^{1} \delta(\iota_0 = \underline{\iota}) \sum_{r_1 \in \mathcal{R}} \delta(r_1 = \underline{r}_1) \cdots \sum_{r_{t-m} \in \mathcal{R}} \delta(r_{t-m} = \underline{r}_{t-m}) \\ &\qquad p(\iota_0) \left\{ \prod_{i=1}^{m} p(\tau_i^-, r_i \,|\, \iota_0, \vec{r}_{1,i-1}) \right\} \left\{ \prod_{i=m+1}^{t} p(\tau_i^-, r_i \,|\, \vec{r}_{i-m,i-1}) \right\} , \end{aligned} \tag{2.4}$$

from equation (2.2). Furthermore, if $m \leq t < n$ then we may simplify the forward pass by observing that

$$\begin{aligned}
\alpha_{t+1}(\vec{r}_{t-m+2,t+1}) &= p(\underline{\iota}, \tau_1^-, \underline{r}_1, \ldots, \tau_{t-m+1}^-, \underline{r}_{t-m+1}, \tau_{t-m+2}^-, r_{t-m+2}, \ldots, \tau_{t+1}^-, r_{t+1}) \\
&= \sum_{\iota_0=0}^{1} \delta(\iota_0 = \underline{\iota}) \sum_{r_1 \in \mathcal{R}} \delta(r_1 = \underline{r}_1) \cdots \sum_{r_{t-m+1} \in \mathcal{R}} \delta(r_{t-m+1} = \underline{r}_{t-m+1}) \\
&\qquad p(\iota_0) \left\{ \prod_{i=1}^{m} p(\tau_i^-, r_i \,|\, \iota_0, \vec{r}_{1,i-1}) \right\} \left\{ \prod_{i=m+1}^{t+1} p(\tau_i^-, r_i \,|\, \vec{r}_{i-m,i-1}) \right\} \\
&= \sum_{r_{t-m+1} \in \mathcal{R}} \delta(r_{t-m+1} = \underline{r}_{t-m+1})\, \alpha_t(\vec{r}_{t-m+1,t})\, p(\tau_{t+1}^-, r_{t+1} \,|\, \vec{r}_{t-m+1,t}) \,. \qquad (2.5)
\end{aligned}$$

Effectively, this recursive relation comes from moving the size-m active window from stage t to stage $t+1$ and thereby marginalising over the observation $\underline{r}_{t-m+1}$ that has now left the window. The forward pass commences from stage m (the last stage that still depends directly on ι_0) by first computing the factor

$$\alpha_m(\vec{r}_{1,m}) \;=\; p(\underline{\iota}, \tau_1^-, r_1, \ldots, \tau_m^-, r_m) \;=\; \sum_{\iota_0=0}^{1} \delta(\iota_0 = \underline{\iota})\, p(\iota_0) \prod_{i=1}^{m} p(\tau_i^-, r_i \,|\, \iota_0, \vec{r}_{1,i-1}) \,. \qquad (2.6)$$

Observe that for the special case of $m = 1$, the entire forward pass reduces to

$$\begin{aligned}
\alpha_1(r_1) &= \sum_{\iota_0=0}^{1} \delta(\iota_0 = \underline{\iota})\, p(\iota_0)\, p(\tau_1^-, r_1 \,|\, \iota_0) \,, \\
\alpha_t(r_t) &= \sum_{r_{t-1} \in \mathcal{R}} \delta(r_{t-1} = \underline{r}_{t-1})\, \alpha_{t-1}(r_{t-1})\, p(\tau_t^-, r_t \,|\, \tau_{t-1}^-, r_{t-1}) \quad \text{for } t = 2, \ldots, n \,. \qquad (2.7)
\end{aligned}$$

Similarly, the backward pass is also well-defined for stages $t \geq m$, such that the backward factor (1.12) becomes

$$\begin{aligned}
\beta_t(\vec{r}_{t-m+1,t}) &= p(\tau_{t+1}^-, \underline{r}_{t+1}, \ldots, \tau_n^-, \underline{r}_n, \underline{\tau} \,|\, \tau_{t-m+1}^-, r_{t-m+1}, \ldots, \tau_t^-, r_t) \\
&= \sum_{r_{t+1} \in \mathcal{R}} \delta(r_{t+1} = \underline{r}_{t+1}) \cdots \sum_{r_n \in \mathcal{R}} \delta(r_n = \underline{r}_n) \sum_{\tau_{n+1}=0}^{1} \delta(\tau_{n+1} = \underline{\tau}) \\
&\qquad \left\{ \prod_{i=t+1}^{n} p(\tau_i^-, r_i \,|\, \vec{r}_{i-m,i-1}) \right\} p(\tau_{n+1} \,|\, \vec{r}_{n-m+1,n}) \,. \qquad (2.8)
\end{aligned}$$

Likewise, if $m < t \leq n$ then we may move the window backwards from stage t to stage $t-1$, thereby obtaining the recursive relation

$$\begin{aligned}
\beta_{t-1}(\vec{r}_{t-m,t-1}) &= p(\tau_t^-, \underline{r}_t, \ldots, \tau_n^-, \underline{r}_n, \underline{\tau} \,|\, \tau_{t-m}^-, r_{t-m}, \ldots, \tau_{t-1}^-, r_{t-1}) \\
&= \sum_{r_t \in \mathcal{R}} \delta(r_t = \underline{r}_t) \cdots \sum_{r_n \in \mathcal{R}} \delta(r_n = \underline{r}_n) \sum_{\tau_{n+1}=0}^{1} \delta(\tau_{n+1} = \underline{\tau}) \\
&\qquad \left\{ \prod_{i=t}^{n} p(\tau_i^-, r_i \,|\, \vec{r}_{i-m,i-1}) \right\} p(\tau_{n+1} \,|\, \vec{r}_{n-m+1,n}) \\
&= \sum_{r_t \in \mathcal{R}} \delta(r_t = \underline{r}_t)\, p(\tau_t^-, r_t \,|\, \vec{r}_{t-m,t-1})\, \beta_t(\vec{r}_{t-m+1,t}) \,. \qquad (2.9)
\end{aligned}$$

Effectively, moving the size-m active window from stage t to stage $t-1$ allows us to marginalise over the observation $\underline{r}_t$ that has now left the window. The backward pass commences from stage n, such that stage $n-m+1$ (the last stage upon which τ_{n+1} directly depends) is the first stage in the sliding window. The inital backward factor is then computed as

$$\beta_n(\vec{r}_{n-m+1,n}) \;=\; p(\underline{\tau} \,|\, \tau_{n-m+1}^-, r_{n-m+1}, \ldots, \tau_n^-, r_n) \;=\; \sum_{\tau_{n+1}=0}^{1} \delta(\tau_{n+1} = \underline{\tau})\, p(\tau_{n+1} \,|\, \vec{r}_{n-m+1,n}) \,. \; (2.10)$$

Observe for the special case of $m = 1$ that the entire backward pass reduces to

$$\begin{aligned}
\beta_n(r_n) &= \sum_{\tau_{n+1}=0}^{1} \delta(\tau_{n+1} = \underline{\tau})\, p(\tau_{n+1} \mid \tau_n^-, r_n)\,, \\
\beta_t(r_t) &= \sum_{r_{t+1}\in\mathcal{R}} \delta(r_{t+1} = \underline{r}_{t+1})\, p(\tau_{t+1}^-, r_{t+1} \mid \tau_t^-, r_t)\, \beta_{t+1}(r_{t+1}) \quad \text{for } t = n-1, \ldots, 1\,. \qquad (2.11)
\end{aligned}$$

Note that for a short sequence of length $n < m$ we must evaluate the observation likelihood (1.8) directly using equation (2.2). The forward–backward algorithm is of no help in such a case, since the active window overlaps both ends of the sequence, namely $\underline{\iota}$ and $\underline{\tau}$, making the dependencies highly stage specific. Likewise, even for $n \geq m$, all stages for $t < m$ must be computed indivudually without recourse to the forward factors, due to the changing number of dependencies on the start of the sequence.

3 Stateful Markov Sequence Processes

Consider the fully-dependent causal process R depicted in Figure 2.1. 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 *state* variable taking values $s_t \in \mathcal{S}$, and X_t is a random *value* variable taking values $x_t \in \mathcal{X}$. Thus, we may define $\vec{r} = ((s_1, x_1), (s_2, x_2), \ldots, (s_n, x_n))$.

We now make the common presumption that the stage transitions in the sequence generating process are entirely between states, e.g. from S_{t-1} to S_t. It follows from causation that the value x_t is generated after the state s_t has been determined, i.e. X_t depends upon S_t. Consequently, the fully-dependent stateful sequence process, shown in Figure 3.1, is derived from Figure 2.1 by replacing each node R_t by the pair of nodes S_t and X_t with a dependency from S_t to X_t, such that every *afferent* dependency pointing to R_t becomes two dependencies pointing to S_t and X_t, and every *efferent* dependency pointing from R_t becomes two dependencies pointing from S_t and X_t, respectively.

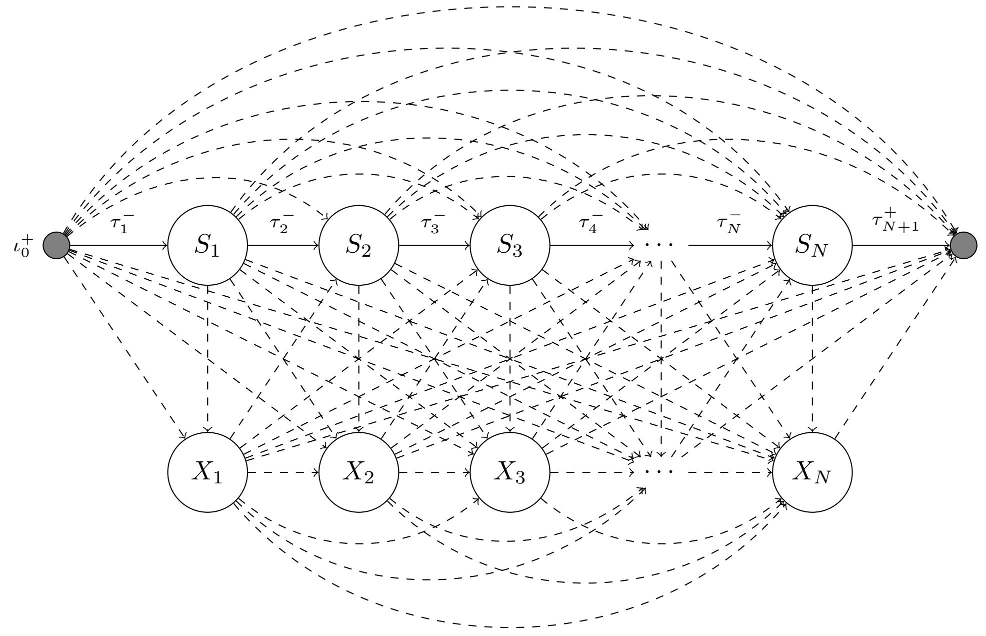

Figure 3.1: *A fully-dependent, causal, stateful process for generating complete, random sequences of random length N, consisting of pairs of states and values.*

For convenience, we may notionally separate the states from the corresponding values at each stage by informally defining $\vec{s} = (s_1, s_2, \ldots, s_n)$ and $\vec{x} = (x_1, x_2, \ldots, x_n)$. More formally, since the non-terminating transitions are associated with the states, we might write $\vec{s} = (\tau_1^-, s_1, \tau_2^-, s_2, \ldots, \tau_n^-, s_n)$. Hence, the joint likelihood of a fully-dependent state–value sequence is now derived from model (2.1) as

$$\begin{aligned} p(\iota_0,\vec{s},\vec{x},\tau_{n+1}) &= p(\iota_0,\tau_1^-,s_1,x_1,\tau_2^-,s_2,x_2,\ldots,\tau_n^-,s_n,x_n,\tau_{n+1}) \\ &= p(\iota_0)\left\{\prod_{t=1}^{n} p(\tau_t^-,s_t\,|\,\iota_0,\vec{s}_{1,t-1},\vec{x}_{1,t-1})\,p(x_t\,|\,\iota_0,\vec{s}_{1,t},\vec{x}_{1,t-1})\right\} \\ &\quad \times p(\tau_{n+1}\,|\,\iota_0,\vec{s}_{1,n},\vec{x}_{1,n})\,. \end{aligned} \tag{3.1}$$

The dependence of S_t and X_t on X_{t-1} (and further historical values) induces a regression model that can be useful in some circumstances, e.g. in sequence classification problems. However, the increased complexity of such models can be difficult to manage in practice, especially if the set $\mathcal{X}$ of values is continuous and/or multi-dimensional. Also, such a dependency causes difficulties handling missing values (see Section 3.1). Lastly, as we discussed earlier, the general 'spirit' of stateful processes is that the state drives the value, not the other way around. Hence, common practice is to limit the dependicies of X_t to only the current and previous states, as shown in Figure 3.2. The joint likelihood of a state–value sequence now reduces from the fully-dependent model (3.1) to the partially-dependent model

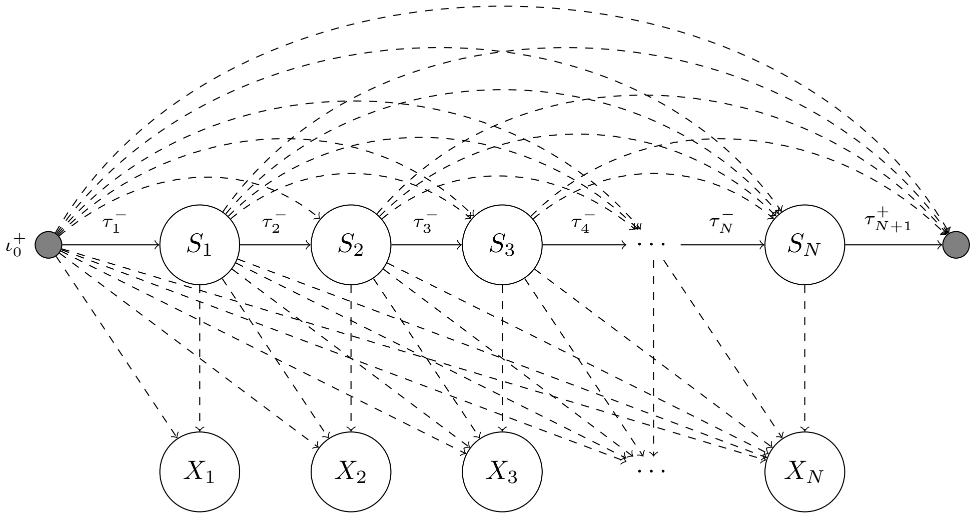

Figure 3.2: *A partially-dependent, stateful process of random length N, with strong causal dependencies from states to values.*

$$p(\iota_0,\vec{s},\vec{x},\tau_{n+1}) = p(\iota_0)\left\{\prod_{t=1}^{n} p(\tau_t^-,s_t\,|\,\iota_0,\vec{s}_{1,t-1})\,p(x_t\,|\,\iota_0,\vec{s}_{1,t})\right\} p(\tau_{n+1}\,|\,\iota_0,\vec{s}_{1,n})\,. \tag{3.2}$$

It is further common practice to restrict the state dependencies of the partially-dependent model to a maximum of m previous states (i.e. the states form an m-th order Markov sequence), and to restrict the value dependencies to a maximum of ℓ states (past and present), as shown in Figure 3.3. The corresponding likelihood model of this order-(m,ℓ) state–value sequence process is

$$\begin{aligned} p(\iota_0,\vec{s},\vec{x},\tau_{n+1}) &= p(\iota_0)\left\{\prod_{t=1}^{m} p(\tau_t^-,s_t\,|\,\iota_0,\vec{s}_{1,t-1})\right\}\left\{\prod_{t=1}^{\ell-1} p(x_t\,|\,\iota_0,\vec{s}_{1,t})\right\} \\ &\quad \times\left\{\prod_{t=m+1}^{n} p(\tau_t^-,s_t\,|\,\vec{s}_{t-m,t-1})\right\}\left\{\prod_{t=\ell}^{n} p(x_t\,|\,\vec{s}_{t-\ell+1,t})\right\} \\ &\quad \times p(\tau_{n+1}\,|\,\vec{s}_{n-m+1,n})\,, \end{aligned} \tag{3.3}$$

for $n \geq m$ and $n \geq \ell - 1$. Note that adherence to the assumption of causality, and in particular the idea that values are driven by states, suggests that $\ell \leq m+1$, since otherwise X_t would depend upon more past states than does S_t. Hence, we take the model to hold for $n \geq m \geq \ell - 1$.

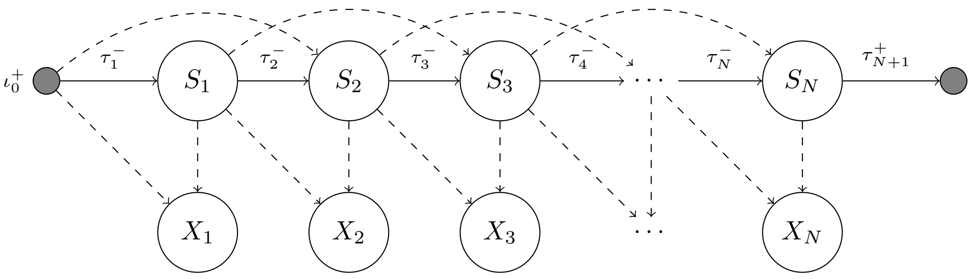

Figure 3.3: *An example ($m = 2, \ell = 2$) of an order-(m, ℓ) stateful Markov process of random length N, where state S_t depends on (at most) m previous states, and value X_t depends upon S_t and (at most) $\ell - 1$ previous states.*

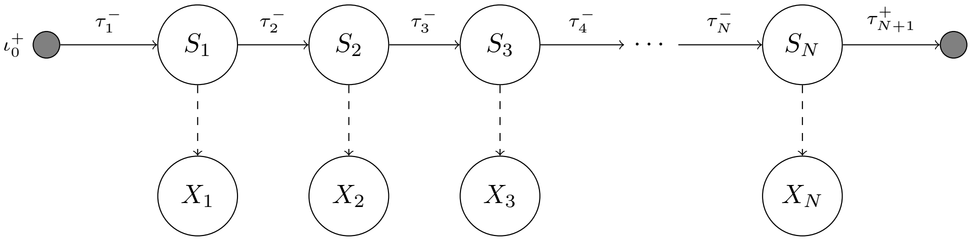

Figure 3.4: *A first-order ($m = 1$, $\ell = 1$) Markov process for generating complete state–value sequences of random length N.*

Finally, note that it is common to further restrict the stateful process depicted in Figure 3.2 by also imposing the first-order Markov assumption at the level of the state–value dependencies themselves, as well as the stage-to-stage dependencies; this restricted process is depicted in Figure 3.4. The corresponding order-$(1, 1)$ sequence model is then given by

$$\begin{aligned} p(\iota_0, \vec{s}, \vec{x}, \tau_{n+1}) &= p(\iota_0)\, p(\tau_1^-, s_1 \mid \iota_0)\, p(x_1 \mid \tau_1^-, s_1) \\ &\quad \times \left\{ \prod_{t=2}^{n} p(\tau_t^-, s_t \mid \tau_{t-1}^-, s_{t-1})\, p(x_t \mid \tau_t^-, s_t) \right\} p(\tau_{n+1} \mid \tau_n^-, s_n)\,. \end{aligned} \tag{3.4}$$

3.1 Hidden State Sequences

A special case of the stateful Markov sequence process is the so-called *hidden Markov model* (HMM), where the values of the state sequence $\vec{s}$ are entirely unobserved (and perhaps unobservable). However, more generally we have to consider the possibility that the values of any of the random variables ι_0, S_t, X_t and τ_{n+1} might or might not have been observed in practice. The extension of equation (1.8) is straightforward for discrete[2] states and values, namely

$$\begin{aligned} p(\underline{\iota}, \underline{\vec{s}}, \underline{\vec{x}}, \underline{\tau}) &= \sum_{\iota_0=0}^{1} \delta(\iota_0 = \underline{\iota}) \sum_{s_1 \in \mathcal{S}} \delta(s_1 = \underline{s}_1) \sum_{x_1 \in \mathcal{X}} \delta(x_1 = \underline{x}_1) \cdots \sum_{s_n \in \mathcal{S}} \delta(s_n = \underline{s}_n) \sum_{x_n \in \mathcal{X}} \delta(x_n = \underline{x}_n) \\ &\quad \sum_{\tau_{n+1}=0}^{1} \delta(\tau_{n+1} = \underline{\tau})\, p(\iota_0, \tau_1^-, s_1, x_1, \ldots, \tau_n^-, s_n, x_n, \tau_{n+1})\,. \end{aligned} \tag{3.5}$$

[2] Recall from Section 1.1 that the continuous analogue is readily derivable from the discrete model.

Upon substitution of model (3.3), we then rearrange the order of summation to obtain

$$
\begin{aligned}
p(\underline{\iota}, \underline{\vec{s}}, \underline{\vec{x}}, \underline{\tau}) \quad = \quad & \sum_{s_1 \in \mathcal{S}} \delta(s_1 = \underline{s}_1) \cdots \sum_{s_n \in \mathcal{S}} \delta(s_n = \underline{s}_n) \\
& \left[\sum_{\iota_0 = 0}^{1} \delta(\iota_0 = \underline{\iota})\, p(\iota_0) \left\{ \prod_{t=1}^{m} p(\tau_t^-, s_t \,|\, \iota_0, \vec{s}_{1,t-1}) \right\} \left\{ \prod_{t=1}^{\ell-1} \sum_{x_t \in \mathcal{X}} \delta(x_t = \underline{x}_t)\, p(x_t \,|\, \iota_0, \vec{s}_{1,t}) \right\} \right] \\
& \times \left\{ \prod_{t=m+1}^{n} p(\tau_t^-, s_t \,|\, \vec{s}_{t-m,t-1}) \right\} \left\{ \prod_{t=\ell}^{n} \sum_{x_t \in \mathcal{X}} \delta(x_t = \underline{x}_t)\, p(x_t \,|\, \vec{s}_{t-\ell+1,t}) \right\} \\
& \times \left\{ \sum_{\tau_{n+1}=0}^{1} \delta(\tau_{n+1} = \underline{\tau})\, p(\tau_{n+1} \,|\, \vec{s}_{n-m+1,n}) \right\} .
\end{aligned}
\tag{3.6}
$$

Suppose now that some value x_t is unobserved, i.e. $\underline{x}_t = *$. This is easily handled, since variable X_t depends only on S_t and possibly earlier stages. Specifically, the marginalisation over $x_t \in \mathcal{X}$ affects exactly one term in the model (as shown in the derivation above). Consequently, we use the fact that

$$
p(\underline{x}_t \,|\, \iota_0, \vec{s}_{1,t}) \quad = \quad \sum_{x_t \in \mathcal{X}} \delta(x_t = \underline{x}_t)\, p(x_t \,|\, \iota_0, \vec{s}_{1,t}) \,, \tag{3.7}
$$

$$
p(\underline{x}_t \,|\, \vec{s}_{t-\ell+1,t}) \quad = \quad \sum_{x_t \in \mathcal{X}} \delta(x_t = \underline{x}_t)\, p(x_t \,|\, \vec{s}_{t-\ell+1,t}) \,, \tag{3.8}
$$

to deduce that $p(X_t = * \,|\, \ldots) = \sum_{x_t \in \mathcal{X}} p(x_t \,|\, \ldots) = 1$, and so the marginalisation of each term in x_t is always well defined. Note that this marginalisation is only applicable when X_t does not depend upon any earlier X_{t-k}; such higher-order dependencies will require the more general handling of equation (3.5).

Similarly, if the termination marker τ_{n+1} is unobserved, i.e. $\underline{\tau} = *$, then the termination probability

$$
p(\underline{\tau} \,|\, \vec{s}_{n-m+1,n}) \quad = \quad \sum_{\tau_{n+1}=0}^{1} \delta(\tau_{n+1} = \underline{\tau})\, p(\tau_{n+1} \,|\, \vec{s}_{n-m+1,n}) \,, \tag{3.9}
$$

or more precisely

$$
p(\tau_{n+1} = \underline{\tau} \,|\, \vec{s}_{n-m+1,n}) \quad = \quad \sum_{\tau=0}^{1} \delta(\tau = \underline{\tau})\, p(\tau_{n+1} = \tau \,|\, \vec{s}_{n-m+1,n}) \,, \tag{3.10}
$$

is also well defined, since likewise $p(\tau_{n+1} = * \,|\, \vec{s}_{n-m+1,n}) = \sum_{\tau=0}^{1} p(\tau_{n+1} = \tau \,|\, \vec{s}_{n-m+1,n}) = 1$.

However, if the initiation marker ι_0 is unobserved, i.e. $\underline{\iota} = *$, then the marginalisation is more difficult because ι_0 potentially conditions both S_t and X_t, as shown by equation (3.6). To allow for this, note that

$$
p(\underline{\iota}, \vec{s}_{1,m}, \underline{\vec{x}}_{1,\ell-1}) \quad = \quad \sum_{\iota_0=0}^{1} \delta(\iota_0 = \underline{\iota})\, p(\iota_0) \left\{ \prod_{t=1}^{m} p(\tau_t^-, s_t \,|\, \iota_0, \vec{s}_{1,t-1}) \right\} \left\{ \prod_{t=1}^{\ell-1} p(\underline{x}_t \,|\, \iota_0, \vec{s}_{1,t}) \right\} \,, \tag{3.11}
$$

since $\ell - 1 \leq m$. Hence, the marginalised observation probability as a function of unknown states takes the form

$$
p(\underline{\iota}, \vec{s}, \underline{\vec{x}}, \underline{\tau}) \;=\; p(\underline{\iota}, \vec{s}_{1,m}, \underline{\vec{x}}_{1,\ell-1}) \left\{ \prod_{t=m+1}^{n} p(\tau_t^-, s_t \,|\, \vec{s}_{t-m,t-1}) \right\} \left\{ \prod_{t=\ell}^{n} p(\underline{x}_t \,|\, \vec{s}_{t-\ell+1,t}) \right\} p(\underline{\tau} \,|\, \vec{s}_{n-m+1,n}) \,, \tag{3.12}
$$

such that the (hidden) state marginalisation becomes

$$
p(\underline{\iota}, \underline{\vec{s}}, \underline{\vec{x}}, \underline{\tau}) \quad = \quad \sum_{s_1 \in \mathcal{S}} \delta(s_1 = \underline{s}_1) \cdots \sum_{s_n \in \mathcal{S}} \delta(s_n = \underline{s}_n)\, p(\underline{\iota}, \vec{s}, \underline{\vec{x}}, \underline{\tau}) \,. \tag{3.13}
$$

Note that the conventional HMM formulation is obtained by dropping the $\underline{\iota}$, $\underline{\vec{s}}$ and $\underline{\tau}$ terms and the $\delta(\cdot)$ indicators.

3.2 Stateful Forward–Backward Algorithm

Under the conventions discussed in the previous section for handling missing observations, the forward–backward algorithm for order–(m,ℓ) stateful Markov sequences now follows from the derivation of Section 2.1. In particular, the forward pass commences with the analogue of factor (2.6), namely

$$\alpha_m(\vec{s}_{1,m}) \;=\; p(\underline{\iota},\vec{s}_{1,m},\underline{\vec{x}}_{1,m}) \;=\; p(\underline{\iota},\vec{s}_{1,m},\underline{\vec{x}}_{1,\ell-1})\prod_{t=\ell}^{m} p(\underline{x}_t\,|\,\vec{s}_{t-\ell+1,t})\,, \tag{3.14}$$

utilising equation (3.11). Note that the last factor vanishes in the special case that $\ell = m+1$. Next, the analogue of the recusive forward relation (2.5) is

$$\begin{aligned}\alpha_t(\vec{s}_{t-m+1,t}) &= p(\underline{\iota},\underline{\vec{s}}_{1,t-m},\vec{s}_{t-m+1,t},\underline{\vec{x}}_{1,t})\\ &= \sum_{s_{t-m}\in\mathcal{S}}\delta(s_{t-m}\!=\!\underline{s}_{t-m})\,\alpha_{t-1}(\vec{s}_{t-m,t-1})\\ &\quad\times p(\tau_t^-,s_t\,|\,\vec{s}_{t-m,t-1})\,p(\underline{x}_t\,|\,\vec{s}_{t-\ell+1,t})\,,\end{aligned} \tag{3.15}$$

for $t = m+1,\ldots,n$. Note that for the special case of $m=\ell=1$, the forward pass reduces to

$$\begin{aligned}\alpha_1(s_1) &= \sum_{\iota_0=0}^{1}\delta(\iota_0\!=\!\underline{\iota})\,p(\iota_0)\,p(\tau_1^-,s_1\,|\,\iota_0)\,p(\underline{x}_1\,|\,\tau_1^-,s_1)\,,\\ \alpha_t(s_t) &= \sum_{s_{t-1}\in\mathcal{S}}\delta(s_{t-1}\!=\!\underline{s}_{t-1})\,\alpha_{t-1}(s_{t-1})\,p(\tau_t^-,s_t\,|\,\tau_{t-1}^-,s_{t-1})\,p(\underline{x}_t\,|\,\tau_t^-,s_t)\;\text{ for } t=2,\ldots,n\,.\end{aligned} \tag{3.16}$$

Similarly, the backward pass commences with the analogue of factor (2.10), namely

$$\beta_n(\vec{s}_{n-m+1,n}) \;=\; p(\underline{\tau}\,|\,\vec{s}_{n-m+1,n})\,, \tag{3.17}$$

utilising equation (3.9). Then the backward recursive relation is the analogue of equation (2.9), namely

$$\begin{aligned}\beta_{t-1}(\vec{s}_{t-m,t-1}) &= p(\underline{\vec{s}}_{t,n},\underline{\vec{x}}_{t,n},\underline{\tau}\,|\,\vec{s}_{t-m,t-1})\\ &= \sum_{s_t\in\mathcal{S}}\delta(s_t\!=\!\underline{s}_t)\,p(\tau_t^-,s_t\,|\,\vec{s}_{t-m,t-1})\,p(\underline{x}_t\,|\,\vec{s}_{t-\ell+1,t})\,\beta_t(\vec{s}_{t-m+1,t})\,,\end{aligned} \tag{3.18}$$

for $t = n,\ldots,m+1$. For the special case of $m=\ell=1$, the backward pass reduces to

$$\begin{aligned}\beta_n(s_n) &= p(\underline{\tau}\,|\,s_n)\,,\\ \beta_{t-1}(s_{t-1}) &= \sum_{s_t\in\mathcal{S}}\delta(s_t\!=\!\underline{s}_t)\,p(\tau_t^-,s_t\,|\,\tau_{t-1}^-,s_{t-1})\,p(\underline{x}_t\,|\,\tau_t^-,s_t)\,\beta_t(s_t)\;\text{ for } t=n,\ldots,2\,.\end{aligned} \tag{3.19}$$

Finally, for the entire sequence of length $n\geq m$, we have

$$p(\underline{\iota},\underline{\vec{s}},\underline{\vec{x}},\underline{\tau}) \;=\; \sum_{s_{t-m+1}\in\mathcal{S}}\delta(s_{t-m+1}\!=\!\underline{s}_{t-m+1})\cdots\sum_{s_t\in\mathcal{S}}\delta(s_t\!=\!\underline{s}_t)\,\alpha_t(\vec{s}_{t-m+1,t})\,\beta_t(\vec{s}_{t-m+1,t})\,, \tag{3.20}$$

for all $t=m,\ldots,n$, from equations (1.10) and (3.13).

3.3 Posterior Prediction

The order–(m,ℓ) forward–backward algorithm of the previous section can now be utilised to predict unknown states and values from an observed sequence. Let $\underline{\nu}=(\underline{\iota},\underline{\vec{s}},\underline{\vec{x}},\underline{\tau})$ represent a (possibly incomplete) sequence of length $n\geq m$, such that equation (3.20) becomes

$$p(\underline{\nu}) \;=\; \sum_{\vec{s}_{t-m+1,t}\in\mathcal{S}^m}\underline{\delta}(\vec{s}_{t-m+1,t})\,\alpha_t(\vec{s}_{t-m+1,t})\,\beta_t(\vec{s}_{t-m+1,t})\,, \tag{3.21}$$

for all $t=m,\ldots,n$, where, for convenience, we now define $\underline{\delta}(\vec{s}_{t_1,t_2})=\prod_{t=t_1}^{t_2}\delta(s_t\!=\!\underline{s}_t)$.

Next, in order to notationally distinguish between predicted and observed states, we continue to let $s_t\in\mathcal{S}$ be an arbitrary state value of S_t, and $\underline{s}_t$ be the (possibly missing) observed state value; however, we now let $\sigma_t \in \mathcal{S}$ represent some specific, predicted state value. Thus, for example, the posterior probability of sequence $\underline{\nu}$ having specific states $\vec{\sigma}_{t-m+1,t}$ is given by

$$\begin{aligned}
\gamma_t(\vec{\sigma}_{t-m+1,t}) &= p(\vec{\sigma}_{t-m+1,t}\,|\,\underline{\nu}) \\
&= \frac{p(\vec{\sigma}_{t-m+1,t},\underline{\iota},\underline{\vec{s}},\underline{\vec{x}},\underline{\tau})}{p(\underline{\nu})} \\
&= \underline{\delta}(\vec{\sigma}_{t-m+1,t})\,\frac{p(\underline{\iota},\underline{\vec{s}}_{1,t-m}\circ\vec{\sigma}_{t-m+1,t}\circ\underline{\vec{s}}_{t+1,n},\underline{\vec{x}}_{1,t}\circ\underline{\vec{x}}_{t+1,n},\underline{\tau})}{p(\underline{\nu})} \\
&= \frac{\underline{\delta}(\vec{\sigma}_{t-m+1,t})\,\alpha_t(\vec{\sigma}_{t-m+1,t})\,\beta_t(\vec{\sigma}_{t-m+1,t})}{\sum_{\vec{s}_{t-m+1,t}\in\mathcal{S}^m}\underline{\delta}(\vec{s}_{t-m+1,t})\,\alpha_t(\vec{s}_{t-m+1,t})\,\beta_t(\vec{s}_{t-m+1,t})}\,. \qquad (3.22)
\end{aligned}$$

For the $m=\ell=1$ case, this simplifies to

$$\gamma_t(\sigma_t) = \frac{\delta(\sigma_t\!=\!\underline{s}_t)\,\alpha_t(\sigma_t)\,\beta_t(\sigma_t)}{\sum_{s_t\in\mathcal{S}}\delta(s_t\!=\!\underline{s}_t)\,\alpha_t(s_t)\,\beta_t(s_t)}\,, \qquad (3.23)$$

which further reduces to $\gamma_t(\sigma_t)=\delta(\sigma_t\!=\!\underline{s}_t)$ if $\underline{s}_t$ is known, and to the standard formula if $\underline{s}_t$ is unknown (recall that $\delta(\sigma_t\!=\!*)=1$ by definition).

Similarly, we may predict the next state S_{n+1} in a partial sequence, provided we have not observed the sequence to terminate at stage n, via

$$\begin{aligned}
p(\tau^-_{n+1},\sigma_{n+1}\,|\,\underline{\nu}) &= \frac{p(\tau^-_{n+1},\sigma_{n+1},\underline{\iota},\underline{\vec{s}},\underline{\vec{x}},\underline{\tau})}{p(\underline{\nu})} \\
&= \delta(\underline{\tau}\!=\!0)\,\frac{p(\underline{\iota},\underline{\vec{s}}\circ\sigma_{n+1},\underline{\vec{x}})}{p(\underline{\nu})} \\
&= \delta(\underline{\tau}\!=\!0)\,\frac{\sum_{\vec{s}_{n-m+1,n}\in\mathcal{S}^m}\underline{\delta}(\vec{s}_{n-m+1,n})\,\alpha_n(\vec{s}_{n-m+1,n})\,p(\tau^-_{n+1},\sigma_{n+1}\,|\,\vec{s}_{n-m+1,n})}{\sum_{\vec{s}_{n-m+1,n}\in\mathcal{S}^m}\underline{\delta}(\vec{s}_{n-m+1,n})\,\alpha_n(\vec{s}_{n-m+1,n})\,\beta_n(\vec{s}_{n-m+1,n})} \\
&= \delta(\underline{\tau}\!=\!0)\,\frac{\sum_{\vec{s}_{n-m+2,n}\in\mathcal{S}^{m-1}}\underline{\delta}(\vec{s}_{n-m+2,n})\,\alpha_{n+1}(\vec{s}_{n-m+2,n}\circ\sigma_{n+1};*)}{\sum_{\vec{s}_{n-m+1,n}\in\mathcal{S}^m}\underline{\delta}(\vec{s}_{n-m+1,n})\,\alpha_n(\vec{s}_{n-m+1,n})\,\beta_n(\vec{s}_{n-m+1,n})}\,, \qquad (3.24)
\end{aligned}$$

where $\alpha_{n+1}(\vec{s}_{n-m+2,n}\circ\sigma_{n+1};*)$ is computed from the forward relation (3.15) by setting $\underline{x}_{n+1}=*$, following from the fact that $p(*\,|\,\vec{s}_{n-\ell+2,n}\circ\sigma_{n+1})=1$ by definition.

By marginalising out state σ_{n+1}, we now see that the posterior probability of sequence non-termination is

$$\begin{aligned}
p(\tau^-_{n+1}\,|\,\underline{\nu}) &= \sum_{\sigma_{n+1}\in\mathcal{S}} p(\tau^-_{n+1},\sigma_{n+1}\,|\,\underline{\nu}) \\
&= \delta(\underline{\tau}\!=\!0)\,\frac{\sum_{\vec{s}_{n-m+1,n}\in\mathcal{S}^m}\underline{\delta}(\vec{s}_{n-m+1,n})\,\alpha_n(\vec{s}_{n-m+1,n})\,p(\tau^-_{n+1}\,|\,\vec{s}_{n-m+1,n})}{\sum_{\vec{s}_{n-m+1,n}\in\mathcal{S}^m}\underline{\delta}(\vec{s}_{n-m+1,n})\,\alpha_n(\vec{s}_{n-m+1,n})\,p(\underline{\tau}\,|\,\vec{s}_{n-m+1,n})}\,, \qquad (3.25)
\end{aligned}$$

from the backward factor (3.17). As expected, this reduces to $p(\tau^-_{n+1}\,|\,\underline{\nu})=0$ for $\underline{\tau}=1$, and to $p(\tau^-_{n+1}\,|\,\underline{\nu})=1$ for $\underline{\tau}=0$; recall that if $\underline{\tau}=*$ then $\delta(\underline{\tau}\!=\!0)=1$ by definition, and $p(*\,|\,\vec{s}_{n-m+1,n})=1$.

Similarly, we may also predict the future value of X_{n+1} via

$$\begin{aligned}
p(\tau^-_{n+1},x_{n+1}\,|\,\underline{\nu}) &= \sum_{\sigma_{n+1}\in\mathcal{S}} p(\tau^-_{n+1},\sigma_{n+1},x_{n+1}\,|\,\underline{\nu}) \\
&= \delta(\underline{\tau}\!=\!0)\,\frac{\sum_{\sigma_{n+1}\in\mathcal{S}}p(\underline{\iota},\underline{\vec{s}}\circ\sigma_{n+1},\underline{\vec{x}}\circ x_{n+1})}{p(\underline{\nu})} \\
&= \delta(\underline{\tau}\!=\!0)\,\frac{\sum_{\vec{s}_{n-m+2,n}\in\mathcal{S}^{m-1}}\underline{\delta}(\vec{s}_{n-m+2,n})\,\alpha_{n+1}(\vec{s}_{n-m+2,n};x_{n+1})}{\sum_{\vec{s}_{n-m+1,n}\in\mathcal{S}^m}\underline{\delta}(\vec{s}_{n-m+1,n})\,\alpha_n(\vec{s}_{n-m+1,n})\,\beta_n(\vec{s}_{n-m+1,n})}\,, \qquad (3.26)
\end{aligned}$$

where $\alpha_{n+1}(\vec{s}_{n-m+2,n};x_{n+1})\equiv\sum_{\sigma_{n+1}\in\mathcal{S}}\alpha_{n+1}(\vec{s}_{n-m+2,n}\circ\sigma_{n+1};x_{n+1})$, with $\alpha_{n+1}(\vec{s}_{n-m+2,n}\circ\sigma_{n+1};x_{n+1})$ computed from the foward relation (3.15) by setting $\underline{x}_{n+1}=x_{n+1}$.

****

4 Discrete-state Sequence Models

The sequence model (3.4) may now be explicitly conditioned on a general parameter θ that governs the various discrete state distributions. Each term in the model depends directly on the stage index t and indirectly on the state index i_t. Furthermore, each term represents either the initial state, the terminal state, or the non-terminal transitions between states at adjacent stages. Hence, let $\theta = (\Pi, \Gamma, \Omega)$, such that the probability of an arbitrary, observed[3] sequence (with no hidden states) is given by

$$p(\underline{\iota}, \vec{s}, \vec{x}, \underline{\tau} \,|\, \theta) \quad = \quad \pi_{\underline{\iota},1,i_1}\, o_{1,i_1}(x_1) \left\{ \prod_{t=1}^{n-1} \omega_{0,t,i_t}\, \Gamma_{t,i_t,i_{t+1}}\, o_{t+1,i_{t+1}}(x_{t+1}) \right\} \omega_{\underline{\tau},n,i_n}\,. \tag{4.1}$$

The initial state S_1 of the sequence at stage $t = 1$ is governed by the parameter $\vec{\pi}$, where

$$\begin{aligned} \pi_{0,t,i} &= p(\iota_{t-1}{=}0 \,|\, \theta)\, p(\tau_t{=}0 \,|\, \iota_{t-1}{=}0, \theta)\, p(S_t{=}\sigma_i \,|\, \iota_{t-1}{=}0, \tau_t{=}0, \theta)\,, && (4.2)\\ \pi_{1,t,i} &= p(\iota_{t-1}{=}1 \,|\, \theta)\, p(\tau_t{=}0 \,|\, \iota_{t-1}{=}1, \theta)\, p(S_t{=}\sigma_i \,|\, \iota_{t-1}{=}1, \tau_t{=}0, \theta)\,, && (4.3) \end{aligned}$$

and

$$\begin{aligned} \pi_{*,t,i} &= p(\iota_{t-1}{=}* \,|\, \theta)\, p(\tau_t{=}0 \,|\, \iota_{t-1}{=}*, \theta)\, p(S_t{=}\sigma_i \,|\, \iota_{t-1}{=}*, \tau_t{=}0, \theta) \\ &= p(\tau_t{=}0, S_t{=}\sigma_i \,|\, \theta) \;=\; \pi_{0,t,i} + \pi_{1,t,i}\,. && (4.4) \end{aligned}$$

Observe that each state S_t for $t > 1$ is a non-initial state, governed by π_{0,t,i_t}. However, such terms do not explicitly appear in model (4.1), except if $\underline{\iota} \neq 1$, since they are already accounted for by the state transitions. These implicit terms become important when it comes to parameter estimation (see Section 4.2).

The terminal state S_n at stage $t = n$ is likewise governed by the parameter $\vec{\omega}$, where

$$\begin{aligned} \omega_{0,t,i} &= p(\tau_{t+1}{=}0 \,|\, S_t{=}\sigma_i, \theta)\,, && (4.5)\\ \omega_{1,t,i} &= p(\tau_{t+1}{=}1 \,|\, S_t{=}\sigma_i, \theta)\,, && (4.6) \end{aligned}$$

and

$$\omega_{*,t,i} \quad = \quad p(\tau_{t+1}{=}* \,|\, S_t{=}\sigma_i, \theta) \;=\; \omega_{0,t,i} + \omega_{1,t,i} \;=\; 1\,. \tag{4.7}$$

Observe that each state S_t for $t < n$ is a non-terminal state, and is explicitly modelled by the term ω_{0,t,i_t}.

Lastly, the permissible transitions between the states S_t and S_{t+1} of consecutive stages t and $t + 1$ are governed by the parameter Γ, where

$$\Gamma_{t,i,j} \quad = \quad p(S_{t+1}{=}\sigma_j \,|\, S_t{=}\sigma_i, \tau_{t+1}{=}0, \theta)\,. \tag{4.8}$$

Note that the model also includes the likelihood of each observed value x_t at stage t, for $t = 1, 2, \ldots, n$. This so-called *data likelihood* is governed by the separate model

$$o_{t,i}(x) \quad = \quad p(X_t{=}x \,|\, S_t{=}\sigma_i, \theta) \qquad \forall x \in \mathcal{X}\,. \tag{4.9}$$

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.

Finally, note that in the situtation where any state in the observed state sequence $\vec{s}$ is hidden, we have to marginalise model (4.1) over each such missing state. Hence, in general, we may define

$$\begin{aligned} p(\underline{\iota}, \vec{s}, \vec{x}, \underline{\tau} \,|\, \theta) \quad = \quad & \sum_{i'_1=1}^{S} \delta(i'_1{=}i_1) \sum_{i'_2=1}^{S} \delta(i'_2{=}i_2) \cdots \sum_{i'_n=1}^{S} \delta(i'_n{=}i_n) \\ & \pi_{\underline{\iota},1,i'_1}\, o_{1,i'_1}(x_1) \left\{ \prod_{t=1}^{n-1} \omega_{0,t,i'_t}\, \Gamma_{t,i'_t,i'_{t+1}}\, o_{t+1,i'_{t+1}}(x_{t+1}) \right\} \omega_{\underline{\tau},n,i'_n}\,, && (4.10) \end{aligned}$$

where $\delta(\cdot)$ is an indicator function taking the value 1 (or 0) if its argument is true (or false). Note that if S_t is a hidden state, then $i_t = *$ and $\delta(i'_t{=}*) = 1$ for all $i'_t \in \{1, 2, \ldots, S\}$; otherwise, the summation over i'_t collapses to the observed value i_t. The observation likelihood given by model (4.10) can be efficiently computed by an extension of the forward–backward algorithm, described in the next section.

[3] We assume that all observed sequences are non-zero in length, since zero-length sequences are typically unobservable unless the generating process explicitly signals the start and end of each sequence. The modelling of zero-length sequences will require an extra parameter.

4.1 Posterior Prediction

Given an observed sequence with one or more missing values, it is useful to be able to predict the probable values of the missing variables. For stateful Markov sequences, this typically means predicting the state S_t at some (or each) stage t. Alternatively, one might wish to predict a future value of S_{t+1} or X_{t+1} given a partially observed sequence. The foward–backward algorithm of Section **??** enables all of these calculations.

For instance, from equation (**??**), the posterior probabilities of state S_t given an observed sequence are computed as

$$\begin{aligned}
\gamma_{t,i} &= p(S_t\!=\!\sigma_i \,|\, \underline{\iota}, \vec{s}, \vec{x}, \underline{\tau}, \theta) \\
&= \frac{p(S_t\!=\!\sigma_i, \underline{\iota}, \vec{s}, \vec{x}, \underline{\tau} \,|\, \theta)}{p(\underline{\iota}, \vec{s}, \vec{x}, \underline{\tau} \,|\, \theta)} \\
&= \frac{\delta(i = i_t)\, \alpha_{t,i}\, \beta_{t,i}}{\sum_{i'=1}^{S} \delta(i' = i_t)\, \alpha_{t,i'}\, \beta_{t,i'}} \,.
\end{aligned} \tag{4.11}$$

Observe that $\gamma_{t,i}$ reduces to $\delta(i = i_t)$ in the special case where $s_t = \sigma_{i_t}$ is known.

Similarly, we may predict the next state S_{n+1} in a given observed sequence of length $n = |\vec{x}|$ via

$$\begin{aligned}
p({\downarrow}\sigma_i \,|\, \underline{\iota}, \vec{s}, \vec{x}, \underline{\tau}, \theta) &= p(\tau_{n+1}\!=\!0, S_{n+1}\!=\!\sigma_i \,|\, \underline{\iota}, \vec{s}, \vec{x}, \underline{\tau}, \theta) \\
&= \frac{p(\tau_{n+1}\!=\!0, S_{n+1}\!=\!\sigma_i, \underline{\iota}, \vec{s}, \vec{x}, \underline{\tau} \,|\, \theta)}{p(\underline{\iota}, \vec{s}, \vec{x}, \underline{\tau} \,|\, \theta)} \\
&= \delta(\underline{\tau}\!=\!0)\, \frac{p(\underline{\iota}, \vec{s} \circ \sigma_i, \vec{x} \,|\, \theta)}{p(\underline{\iota}, \vec{s}, \vec{x}, \underline{\tau} \,|\, \theta)} \\
&= \delta(\underline{\tau}\!=\!0)\, \frac{\bar{\alpha}_{n+1,i}}{\sum_{i'=1}^{S} \delta(i'\!=\!i_n)\, \alpha_{n,i'}\, \beta_{n,i'}} \,,
\end{aligned} \tag{4.12}$$

from equation (**??**). Consequently, we may also predict the future value of X_{t+1} via

$$\begin{aligned}
p({\downarrow}x \,|\, \underline{\iota}, \vec{s}, \vec{x}, \underline{\tau}, \theta) &= \sum_{i=1}^{S} p({\downarrow}\sigma_i, x \,|\, \underline{\iota}, \vec{s}, \vec{x}, \underline{\tau}, \theta) \\
&= \sum_{i=1}^{S} p({\downarrow}\sigma_i \,|\, \underline{\iota}, \vec{s}, \vec{x}, \underline{\tau}, \theta)\, p(X_{n+1}\!=\!x \,|\, S_{n+1}\!=\!\sigma_i, \theta) \\
&= \delta(\underline{\tau}\!=\!0)\, \frac{\sum_{i=1}^{S} \bar{\alpha}_{n+1,i}\, o_{n+1,i}(x)}{\sum_{i'=1}^{S} \delta(i'\!=\!i_n)\, \alpha_{n,i'}\, \beta_{n,i'}} \,.
\end{aligned} \tag{4.13}$$

Proceding to predicting stage transitions, the forward–backward calculations also enable us to compute the posterior probabilities of the joint states of stages t and $t+1$ via

$$\begin{aligned}
\xi_{t,i,j} &= p(S_t\!=\!\sigma_i, S_{t+1}\!=\!\sigma_j \,|\, \underline{\iota}, \vec{s}, \vec{x}, \underline{\tau}, \theta) \\
&= \frac{p(S_t\!=\!\sigma_i, S_{t+1}\!=\!\sigma_j, \underline{\iota}, \vec{s}, \vec{x}, \underline{\tau} \,|\, \theta)}{p(\underline{\iota}, \vec{s}, \vec{x}, \underline{\tau} \,|\, \theta)} \\
&= \delta(i = i_t)\delta(j = i_{t+1})\, \frac{p(\underline{\iota}, \vec{s}_{t-1} \circ \sigma_i \circ \sigma_j \circ \overleftarrow{s}_{t+2}, \vec{x}, \underline{\tau} \,|\, \theta)}{p(\underline{\iota}, \vec{s}, \vec{x}, \underline{\tau} \,|\, \theta)} \\
&= \delta(i = i_t)\delta(j = i_{t+1})\, \frac{\alpha_{t,i}\, \omega_{0,t,i}\, \Gamma_{t,i,j}\, o_{t+1,j}(x_{t+1})\, \beta_{t+1,j}}{\sum_{i'=1}^{S} \delta(i'\!=\!i_t)\, \alpha_{t,i'}\, \beta_{t,i'}} \,,
\end{aligned} \tag{4.14}$$

since

$$\begin{aligned}
p(\underline{\iota}, \vec{s}_{t-1} \circ \sigma_i \circ \sigma_j \circ \overleftarrow{s}_{t+2}, \vec{x}, \underline{\tau} \,|\, \theta) &= p(\underline{\iota}, \vec{s}_{t-1} \circ \sigma_i, \vec{x}_t \,|\, \theta)\, p({\downarrow}\sigma_j, x_{t+1} \,|\, S_t\!=\!\sigma_i, \theta) \\
&\quad \times p({\downarrow}\overleftarrow{s}_{t+2}, \overleftarrow{x}_{t+2}, \underline{\tau} \,|\, S_{t+1}\!=\!\sigma_j, \theta) \\
&= \alpha_{t,i}\, \omega_{0,t,i}\, \Gamma_{t,i,j}\, o_{t+1,j}(x_{t+1})\, \beta_{t+1,j} \,,
\end{aligned} \tag{4.15}$$

from the forward pass (**??**) and the backward pass (**??**). Observe that

$$\begin{aligned}
\gamma_{t,i} &= p(S_t\!=\!\sigma_i \,|\, \underline{\iota}, \vec{s}, \vec{x}, \underline{\tau}, \theta) \\
&= \sum_{j=1}^{S} p(S_t\!=\!\sigma_i, S_{t+1}\!=\!\sigma_j \,|\, \underline{\iota}, \vec{s}, \vec{x}, \underline{\tau}, \theta) \;=\; \sum_{j=1}^{S} \xi_{t,i,j} \,,
\end{aligned} \tag{4.16}$$

from equations (4.11) and (4.14).

Finally, the modified forward–backward algorithm also allows us to predict the start and/or end of partially observed sequences. For instance, at the start of a sequence we can predict

$$\begin{aligned} p(\iota_0 = \underline{\iota}', S_1 = \sigma_i \mid \underline{\iota}, \vec{s}, \vec{x}, \underline{\tau}, \theta) &= \frac{p(\iota_0 = \underline{\iota}', S_1 = \sigma_i, \underline{\iota}, \vec{s}, \vec{x}, \underline{\tau} \mid \theta)}{p(\underline{\iota}, \vec{s}, \vec{x}, \underline{\tau} \mid \theta)} \\ &= \delta(\underline{\iota}' = \underline{\iota})\, \delta(i = i_1)\, \frac{p(\underline{\iota}', \sigma_i \circ \vec{s}_2, \vec{x}, \underline{\tau} \mid \theta)}{p(\underline{\iota}, \vec{s}, \vec{x}, \underline{\tau} \mid \theta)} \\ &= \delta(\underline{\iota}' = \underline{\iota})\, \delta(i = i_1)\, \frac{\pi_{\underline{\iota}',1,i}\, o_{1,i}(x_1)\, \beta_{1,i}}{\sum_{i'=1}^{S} \delta(i' = i_1)\, \alpha_{1,i'}\, \beta_{1,i'}} \\ &= \delta(\underline{\iota}' = \underline{\iota})\, \gamma_{1,i}\, \frac{\pi_{\underline{\iota}',1,i}}{\bar{\alpha}_{1,i}} \;=\; \gamma_{1,i}\, \kappa_{\underline{\iota}',1,i} \,, \end{aligned} \tag{4.17}$$

where

$$\kappa_{\underline{\iota}',1,i} \;=\; p(\iota_0 = \underline{\iota}' \mid S_1 = \sigma_i, \underline{\iota}, \vec{s}, \vec{x}, \underline{\tau}, \theta) \;=\; \begin{cases} \delta(\underline{\iota}' = \underline{\iota}) & \text{if } \underline{\iota} = 0 \text{ or } 1 \\ \frac{\pi_{\underline{\iota}',1,i}}{\pi_{0,1,i} + \pi_{1,1,i}} & \text{if } \underline{\iota} = * \end{cases} \,, \tag{4.18}$$

from equations (??), (??) and (4.4). It then follows that

$$p(\iota_0 = \underline{\iota}' \mid \underline{\iota}, \vec{s}, \vec{x}, \underline{\tau}, \theta) \;=\; \sum_{i=1}^{S} p(\iota_0 = \underline{\iota}', S_1 = \sigma_i \mid \underline{\iota}, \vec{s}, \vec{x}, \underline{\tau}, \theta) \;=\; \sum_{i=1}^{S} \gamma_{1,i}\, \kappa_{\underline{\iota}',1,i} \,. \tag{4.19}$$

Likewise, at the end of a sequence we can predict

$$\begin{aligned} p(\tau_{n+1} = \underline{\tau}', S_n = \sigma_i \mid \underline{\iota}, \vec{s}, \vec{x}, \underline{\tau}, \theta) &= \frac{p(\tau_{n+1} = \underline{\tau}', S_n = \sigma_i, \underline{\iota}, \vec{s}, \vec{x}, \underline{\tau} \mid \theta)}{p(\underline{\iota}, \vec{s}, \vec{x}, \underline{\tau} \mid \theta)} \\ &= \delta(\underline{\tau}' = \underline{\tau})\, \delta(i = i_n)\, \frac{p(\underline{\iota}, \vec{s}_{n-1} \circ \sigma_i, \vec{x}, \underline{\tau}' \mid \theta)}{p(\underline{\iota}, \vec{s}, \vec{x}, \underline{\tau} \mid \theta)} \\ &= \delta(\underline{\tau}' = \underline{\tau})\, \delta(i = i_n)\, \frac{\alpha_{n,i}\, \omega_{\underline{\tau}',n,i}}{\sum_{i'=1}^{S} \delta(i' = i_n)\, \alpha_{n,i'}\, \beta_{n,i'}} \\ &= \delta(\underline{\tau}' = \underline{\tau})\, \gamma_{n,i} \frac{\omega_{\underline{\tau}',n,i}}{\beta_{n,i}} \;=\; \gamma_{n,i}\, \zeta_{\underline{\tau}',n,i} \,, \end{aligned} \tag{4.20}$$

where

$$\zeta_{\underline{\tau}',n,i} \;=\; p(\tau_{n+1} = \underline{\tau}' \mid S_n = \sigma_i, \underline{\iota}, \vec{s}, \vec{x}, \underline{\tau}, \theta) \;=\; \begin{cases} \delta(\underline{\tau}' = \underline{\tau}) & \text{if } \underline{\tau} = 0 \text{ or } 1 \\ \omega_{\underline{\tau}',n,i} & \text{if } \underline{\tau} = * \end{cases} \,, \tag{4.21}$$

from equations (??) and (??). It then follows that

$$p(\tau_{n+1} = \underline{\tau}' \mid \underline{\iota}, \vec{s}, \vec{x}, \underline{\tau}, \theta) \;=\; \sum_{i=1}^{S} p(\tau_{n+1} = \underline{\tau}', S_n = \sigma_i \mid \underline{\iota}, \vec{s}, \vec{x}, \underline{\tau}, \theta) \;=\; \sum_{i=1}^{S} \gamma_{n,i}\, \zeta_{\underline{\tau}',n,i} \,. \tag{4.22}$$

An example of the use of these posterior predictions is given in Section 4.3, when estimating the model parameters from observations with missing data.

4.2 Posterior Parameter Estimation with Known Data

We desire to estimate the model parameter $\theta = (\Pi, \Gamma, \Omega)$ given an ordered set $\mathbb{V} = \{\vec{v}^{(d)}\}_{d=1}^{D}$ of observed state and value sequences, where each observation takes the form of $\vec{v}^{(d)} = (\underline{\iota}^{(d)}, \vec{s}^{(d)}, \vec{x}^{(d)}, \underline{\tau}^{(d)})$. As before, we assume that $\vec{x}^{(d)}$ is a contiguous sequence of observed values with no missing values, whereas each 'observed' state s_t might either be known, i.e. $s_t = \sigma_{i_t}$, or missing, i.e. $s_t = *$ and $i_t = *$. Similarly, the sequence initiation and termination markers, $\underline{\iota}^{(d)}$ and $\underline{\tau}^{(d)}$ respectively, might also be known or unknown. In this section, let us suppose that each $\vec{v}^{(d)}$ is entirely known. The case of hidden data is analysed in the next section.

Due to the typical shortage of observed data, let us additionally assume that the distributions for the sub-parameters are stationary in time; that is, $\Gamma_{t,i,j} \equiv \Gamma_{i,j}$ for any stage t, and likewise $\pi_{\underline{\iota},t,i} \equiv \omega_{\underline{\iota},i}$,

$\omega_{\underline{\tau},t,i} \equiv \omega_{\underline{\tau},i}$ and $o_{t,i}(x) \equiv o_i(x)$. Then, from equation (4.1), we obtain the likelihood of the d-th observed sequence as

$$p(v^{(d)}\,|\,\theta) \;\;=\;\; \pi_{\underline{\iota}^{(d)},i_1^{(d)}}\, o_{i_1^{(d)}}(x_1^{(d)}) \left\{ \prod_{t=1}^{n^{(d)}-1} \omega_{0,i_t^{(d)}}\, \Gamma_{i_t^{(d)},i_{t+1}^{(d)}}\, o_{i_{t+1}^{(d)}}(x_{t+1}^{(d)}) \right\} \omega_{\underline{\tau}^{(d)},i_{n^{(d)}}^{(d)}}\,, \tag{4.23}$$

where $n^{(d)} = |\vec{x}^{(d)}|$, and the log-likelihood as

$$\ell(v^{(d)}\,|\,\theta) \;\;=\;\; \log \pi_{\underline{\iota}^{(d)},i_1^{(d)}} + \sum_{t=1}^{n^{(d)}-1} \log \omega_{0,i_t^{(d)}} \Gamma_{i_t^{(d)},i_{t+1}^{(d)}} + \sum_{t=1}^{n^{(d)}} \log o_{i_t^{(d)}}(x_{i_t^{(d)}}) + \log \omega_{\underline{\tau}^{(d)},i_{n^{(d)}}^{(d)}}\,. \tag{4.24}$$

Now, under the assumption that the observed sequences are independent, the log-likelihood of the observed data is given by

$$L(\theta) \;\;=\;\; \log p(\mathbb{V}\,|\,\theta) \;\;=\;\; \log \prod_{d=1}^{D} p(v^{(d)}\,|\,\theta) \;\;=\;\; \sum_{d=1}^{D} \ell(v^{(d)},\theta)\,. \tag{4.25}$$

Hence, to estimate θ we maximise the log-likelihood subject to the necessary (Lagrangian) constraints on the sub-parameters. Starting with the state transitions, we maximise

$$F_\Gamma(\theta) \;\;=\;\; \sum_{d=1}^{D} \sum_{t=1}^{n^{(d)}-1} \log \Gamma_{i_t^{(d)},i_{t+1}^{(d)}} - \sum_{i=1}^{S} \lambda_i \left(\sum_{j=1}^{S} \Gamma_{i,j} - 1 \right) \tag{4.26}$$

$$\Rightarrow \frac{\partial F_\Gamma(\theta)}{\partial \Gamma_{i,j}} \;\;=\;\; \sum_{d=1}^{D} \sum_{t=1}^{n^{(d)}-1} \delta(i = i_t^{(d)})\, \delta(j = i_{t+1}^{(d)})\, \frac{1}{\Gamma_{i,j}} - \lambda_i \;\;=\;\; 0 \text{ when } \theta = \hat{\theta}$$

$$\Rightarrow \hat{\lambda}_i \;\;=\;\; \sum_{j=1}^{S} \sum_{d=1}^{D} \sum_{t=1}^{n^{(d)}-1} \delta(i = i_t^{(d)})\, \delta(j = i_{t+1}^{(d)}) \;\;=\;\; \sum_{d=1}^{D} \sum_{t=1}^{n^{(d)}-1} \delta(i = i_t^{(d)})$$

$$\Rightarrow \hat{\Gamma}_{i,j} \;\;=\;\; \frac{\sum_{d=1}^{D} \sum_{t=1}^{n^{(d)}-1} \delta(i = i_t^{(d)})\, \delta(j = i_{t+1}^{(d)})}{\sum_{d=1}^{D} \sum_{t=1}^{n^{(d)}-1} \delta(i = i_t^{(d)})}\,. \tag{4.27}$$

Observe that this estimate corresponds to counting all the transitions from state i to state j across all the data, and then normalising these counts by the sum over j.

Similarly, for sequence termination or non-termination, we maximise

$$F_\Omega(\theta) \;\;=\;\; \sum_{d=1}^{D} \left\{ \sum_{t=1}^{n^{(d)}-1} \log \omega_{0,i_t^{(d)}} + \log \omega_{\underline{\tau}^{(d)},i_{n^{(d)}}^{(d)}} \right\} - \sum_{i=1}^{S} \lambda_i \left(\omega_{0,i} + \omega_{1,i} - 1 \right) \tag{4.28}$$

$$\Rightarrow \frac{\partial F_\Omega(\theta)}{\partial \omega_{0,i}} \;\;=\;\; \sum_{d=1}^{D} \left\{ \sum_{t=1}^{n^{(d)}-1} \frac{\delta(i = i_t^{(d)})}{\omega_{0,i}} + \frac{\delta(\underline{\tau}^{(d)}\!=\!0)\, \delta(i = i_{n^{(d)}}^{(d)})}{\omega_{0,i}} \right\} - \lambda_i\,,$$

$$\frac{\partial F_\Omega(\theta)}{\partial \omega_{1,i}} \;\;=\;\; \sum_{d=1}^{D} \left\{ \frac{\delta(\underline{\tau}^{(d)}\!=\!1)\, \delta(i = i_{n^{(d)}}^{(d)})}{\omega_{1,i}} \right\} - \lambda_i\,. \tag{4.29}$$

Hence, by multiplying the two derivatives by $\omega_{0,i}$ and $\omega_{1,i}$, respectively, adding the terms and setting the result to zero, we obtain

$$\hat{\lambda}_i \;\;=\;\; \sum_{d=1}^{D} \sum_{t=1}^{n^{(d)}} \delta(i = i_t^{(d)})$$

$$\Rightarrow \hat{\omega}_{0,i} \;\;=\;\; \frac{\sum_{d=1}^{D} \left\{ \sum_{t=1}^{n^{(d)}-1} \delta(i = i_t^{(d)}) + \delta(\underline{\tau}^{(d)}\!=\!0)\, \delta(i = i_{n^{(d)}}^{(d)}) \right\}}{\sum_{d=1}^{D} \sum_{t=1}^{n^{(d)}} \delta(i = i_t^{(d)})}\,,$$

$$\hat{\omega}_{1,i} \;\;=\;\; \frac{\sum_{d=1}^{D} \delta(\underline{\tau}^{(d)}\!=\!1)\, \delta(i = i_{n^{(d)}}^{(d)})}{\sum_{d=1}^{D} \sum_{t=1}^{n^{(d)}} \delta(i = i_t^{(d)})}\,. \tag{4.30}$$

Observe that this latter estimate corresponds to counting the various terminal states over all observed sequences, and then normalising these counts by the overall count of each state. Also note that we have assumed that $\underline{\tau}^{(d)}$ is known; unfortunately, these estimates will be inaccurate if $\underline{\tau}^{(d)}$ is unknown, since they ascribe equal weight to $\underline{\tau}^{(d)} = 0$ and $\underline{\tau}^{(d)} = 1$ regardless of $v^{(d)}$. The correct estimates in the case of missing data will be analysed in the next section.

Finally, for sequence initiation or non-initiation, we recall the comment made in Section 4 that each stage transition is both explicitly a non-terminal transition and implicitly a non-initial transtion; that is, each state transition $\Gamma_{t,i,j}$ also implies a sequence non-initiation $\pi_{0,t+1,j}$. Hence, from equation (4.24), we maximise the function

$$\begin{aligned} F_\Pi(\theta) &= \sum_{d=1}^{D} \left\{ \log \pi_{\underline{\iota}^{(d)},i_1^{(d)}} + \sum_{t=2}^{n^{(d)}} \log \pi_{0,i_t^{(d)}} \right\} - \lambda \left(\sum_{i=1}^{S} \{\pi_{0,i} + \pi_{1,i}\} - 1 \right) && (4.31) \\ \Rightarrow \frac{\partial F_\Pi(\theta)}{\partial \pi_{0,i}} &= \sum_{d=1}^{D} \left\{ \frac{\delta(\underline{\iota}^{(d)}=0)\,\delta(i_1^{(d)}=i)}{\pi_{0,i}} + \sum_{t=2}^{n^{(d)}} \frac{\delta(i_t^{(d)}=i)}{\pi_{0,i}} \right\} - \lambda\,, \\ \frac{\partial F_\Pi(\theta)}{\partial \pi_{1,i}} &= \sum_{d=1}^{D} \left\{ \frac{\delta(\underline{\iota}^{(d)}=1)\,\delta(i_1^{(d)}=i)}{\pi_{1,i}} \right\} - \lambda\,. && (4.32) \end{aligned}$$

Thus, by multiplying the two derivatives by $\pi_{0,i}$ and $\pi_{1,i}$, respectively, adding and summing the terms over i, and setting the result to zero, we obtain

$$\begin{aligned} \hat{\lambda} &= \sum_{i=1}^{S}\sum_{d=1}^{D}\sum_{t=1}^{n^{(d)}} \delta(i_t^{(d)}=i) \;=\; \sum_{d=1}^{D} n^{(d)} \\ \Rightarrow \hat{\pi}_{0,i} &= \frac{\sum_{d=1}^{D} \left\{ \delta(\underline{\iota}^{(d)}=0)\,\delta(i_1^{(d)}=i) + \sum_{t=2}^{n^{(d)}} \delta(i_t^{(d)}=i) \right\}}{\sum_{d=1}^{D} n^{(d)}}\,, \\ \hat{\pi}_{1,i} &= \frac{\sum_{d=1}^{D} \delta(\underline{\iota}^{(d)}=1)\,\delta(i_1^{(d)}=i)}{\sum_{d=1}^{D} n^{(d)}}\,. && (4.33) \end{aligned}$$

Observe that this latter estimate corresponds to counting the various initial states over all observed sequences, and then normalising these counts by the overall count of all states. Also note that these estimates are inaccurate if $\underline{\iota}$ is unknown; the correct estimates are derived in the next section.

4.3 Posterior Parameter Estimation with Missing Data

In contrast to Section 4.2, suppose now that any or all values of $\underline{\iota}^{(d)}$, $\underline{\tau}^{(d)}$ and $\vec{s}^{(d)}$ may be unknown when observing the d-th sequence $v^{(d)}$. The basic procedure is then to first estimate these missing values from the observed data $\mathbb{V}$, and then to estimate the most likely model parameter value $\hat{\theta}$ given $\mathbb{V}$ and the missing values. This is the principle of the *expectation–maximisation* (EM) algorithm, which underlies the modified *Baum–Welch* parameter estimation algorithm derived here.

Suppose we let $\mathbb{Z} = \{z^{(d)}\}_{d=1}^{D}$ denote the ordered set of missing values corresponding to the observed values $\mathbb{V} = \{v^{(d)}\}_{d=1}^{D}$, where $z^{(d)} = (\overline{\underline{\iota}^{(d)}}, \overline{\vec{s}^{(d)}}, \overline{\underline{\tau}^{(d)}})$; that is, notionally $\mathbb{Z}$ contains the true (but still unknown) values missing from $\mathbb{V}$. Hence, we take an expectation of the log-likelihood over all possible values of $\mathbb{Z}$, namely[4]

$$\begin{aligned} Q(\theta) &= E_{\mathbb{Z}\,|\,\mathbb{V},\theta}\left[\log p(\mathbb{Z},\mathbb{V}\,|\,\theta)\right] \\ &= E_{\mathbb{Z}\,|\,\mathbb{V},\theta}\left[\sum_{d=1}^{D}\log p(z^{(d)},v^{(d)}\,|\,\theta)\right] \\ &= \sum_{d=1}^{D} E_{\mathbb{Z}\,|\,\mathbb{V},\theta}\left[\ell(\overline{\underline{\iota}^{(d)}},\overline{\vec{s}^{(d)}},\vec{x}^{(d)},\overline{\underline{\tau}^{(d)}};\theta)\right] \\ &= \sum_{d=1}^{D}\sum_{\overline{\underline{\iota}}=0}^{1}\sum_{\overline{i_1}=1}^{S}\cdots\sum_{\overline{i_{n^{(d)}}}}^{S}\sum_{\overline{\underline{\tau}}=0}^{1} p(\overline{\underline{\iota}},\overline{\vec{s}},\overline{\underline{\tau}}\,|\,\underline{\iota}^{(d)},\vec{s}^{(d)},\vec{x}^{(d)},\underline{\tau}^{(d)},\theta)\,\ell(\overline{\underline{\iota}},\overline{\vec{s}},\vec{x}^{(d)},\overline{\underline{\tau}};\theta) \\ &= \sum_{d=1}^{D}\sum_{\overline{\underline{\iota}}=0}^{1}\sum_{\overline{i_1}=1}^{S}\cdots\sum_{\overline{i_{n^{(d)}}}}^{S}\sum_{\overline{\underline{\tau}}=0}^{1} p(z\,|\,v^{(d)},\theta)\,\ell(\overline{v^{(d)}};\theta)\,, \end{aligned} \tag{4.34}$$

where $z=(\overline{\underline{\iota}},\overline{\vec{s}},\overline{\underline{\tau}})$ and $\overline{v^{(d)}}=(\overline{\underline{\iota}},\overline{\vec{s}},\vec{x}^{(d)},\overline{\underline{\tau}})$. In principle, the optimal parameter value $\hat{\theta}$ is estimated by maximising this expected log-likelihood subject to parameter constraints.

In practice, it is difficult to optimise this nonlinear expression analytically. A feasible alternative is to iteratively apply the EM algorithm:

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

$$\begin{aligned} Q(\theta,\hat{\theta}_k) &= E_{\mathbb{Z}\,|\,\mathbb{V},\hat{\theta}_k}\left[\log p(\mathbb{Z},\mathbb{V}\,|\,\theta)\right] \\ &= \sum_{d=1}^{D}\sum_{\overline{\underline{\iota}}=0}^{1}\sum_{\overline{i_1}=1}^{S}\cdots\sum_{\overline{i_{n^{(d)}}}}^{S}\sum_{\overline{\underline{\tau}}=0}^{1} p(z\,|\,v^{(d)},\hat{\theta}_k)\,\ell(\overline{v^{(d)}};\theta)\,. \end{aligned} \tag{4.35}$$

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

$$\hat{\theta}_{k+1} = \arg\max_{\theta} Q(\theta,\hat{\theta}_k)\,. \tag{4.36}$$

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

Following the methodology of Section 4.2, we now break the optimisation of $Q(\theta,\hat{\theta})$ down into separate maximisation problems over each sub-parameter. For instance, we iteratively estimate the state transitions Γ by optimising

$$\begin{aligned} Q_\Gamma(\theta,\hat{\theta}) &= \sum_{d=1}^{D}\sum_{\overline{\underline{\iota}}=0}^{1}\sum_{\overline{i_1}=1}^{S}\cdots\sum_{\overline{i_{n^{(d)}}}=1}^{S}\sum_{\overline{\underline{\tau}}=0}^{1}\sum_{t=1}^{n^{(d)}-1} p(z\,|\,v^{(d)},\hat{\theta})\,\log\Gamma_{\overline{i_t},\overline{i_{t+1}}} \\ &= \sum_{d=1}^{D}\sum_{t=1}^{n^{(d)}-1}\sum_{i=1}^{S}\sum_{j=1}^{S} p(S_t=\sigma_i,S_{t+1}=\sigma_j\,|\,v^{(d)},\hat{\theta})\,\log\Gamma_{i,j} \\ &= \sum_{d=1}^{D}\sum_{t=1}^{n^{(d)}-1}\sum_{i=1}^{S}\sum_{j=1}^{S}\hat{\xi}_{t,i,j}^{(d)}\,\log\Gamma_{i,j}\,, \end{aligned} \tag{4.37}$$

subject to the appropriate contraints. Note that use has been made of equation (4.14). Hence, borrowing

[4] Other expectations are possible, e.g. over the joint distribution $\mathbb{Z},\mathbb{V}\,|\,\theta$. 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 $\mathbb{Z}\,|\,\mathbb{V},\theta$ often leads to micro-averaged estimates of the form $\sum_{d=1}^{D}[\phi^{(d)}/\psi^{(d)}]/D$.

the Lagrangian constraints from equation (4.26), we estimate the value $\hat{\Gamma}'$ that maximises

$$\begin{aligned} F_\Gamma(\theta,\hat{\theta}) &= \sum_{d=1}^{D}\sum_{t=1}^{n^{(d)}-1}\sum_{i=1}^{S}\sum_{j=1}^{S}\hat{\xi}_{t,i,j}^{(d)}\,\log\Gamma_{i,j}-\sum_{i=1}^{S}\lambda_i\left(\sum_{j=1}^{S}\Gamma_{i,j}-1\right) & (4.38)\\ \Rightarrow\frac{\partial F_\Gamma(\theta,\hat{\theta})}{\partial\Gamma_{i,j}} &= \sum_{d=1}^{D}\sum_{t=1}^{n^{(d)}-1}\frac{\hat{\xi}_{t,i,j}^{(d)}}{\Gamma_{i,j}}-\lambda_i \;=\; 0 \text{ when } \theta=\hat{\theta}' \\ \Rightarrow\hat{\lambda}_i' &= \sum_{d=1}^{D}\sum_{t=1}^{n^{(d)}-1}\sum_{j=1}^{S}\hat{\xi}_{t,i,j} \;=\; \sum_{d=1}^{D}\sum_{t=1}^{n^{(d)}-1}\hat{\gamma}_{t,i} \\ \Rightarrow\hat{\Gamma}_{i,j}' &= \frac{\sum_{d=1}^{D}\sum_{t=1}^{n^{(d)}-1}\hat{\xi}_{t,i,j}^{(d)}}{\sum_{d=1}^{D}\sum_{t=1}^{n^{(d)}-1}\hat{\gamma}_{t,i}^{(d)}} & (4.39)\end{aligned}$$

from equation (4.16).

Similarly, we iteratively estimate the sequence initiation distributions $\pi_{0,i}$ and $\pi_{1,i}$ by optimising

$$\begin{aligned} Q_\Pi(\theta,\hat{\theta}) &= \sum_{d=1}^{D}\sum_{\underline{\iota}=0}^{1}\sum_{\overline{i_1}=1}^{S}\cdots\sum_{\overline{i_{n^{(d)}}}=1}^{S}\sum_{\overline{\underline{\iota}}=0}^{1}p(z\,|\,v^{(d)},\hat{\theta})\left\{\log\pi_{\underline{\iota},\overline{i_1}}+\sum_{t=2}^{n^{(d)}}\log\pi_{0,\overline{i_t}}\right\} \\ &= \sum_{d=1}^{D}\left\{\sum_{\underline{\iota}=0}^{1}\sum_{\overline{i_1}=1}^{S}p(\iota_0=\underline{\iota},S_1=\sigma_{\overline{i_1}}\,|\,v^{(d)},\hat{\theta})\,\log\pi_{\underline{\iota},\overline{i_1}}\right. \\ &\qquad\left.+\sum_{t=2}^{n^{(d)}}\sum_{\overline{i_t}=1}^{S}p(S_t=\sigma_{\overline{i_t}}\,|\,v^{(d)},\hat{\theta})\,\log\pi_{0,\overline{i_t}}\right\} \\ &= \sum_{d=1}^{D}\sum_{i=1}^{S}\left\{\sum_{\underline{\iota}'=0}^{1}\hat{\gamma}_{1,i}^{(d)}\,\hat{\kappa}_{\underline{\iota}',1,i}^{(d)}\log\pi_{\underline{\iota}',i}+\sum_{t=2}^{n^{(d)}}\hat{\gamma}_{t,i}^{(d)}\log\pi_{0,i}\right\} & (4.40)\end{aligned}$$

subject to the appropriate contraints. Note that we have utilised equations (4.14) and (4.17). Hence, borrowing the Lagrangian constraint of equation (4.31), we maximise

$$\begin{aligned} F_\Pi(\theta,\hat{\theta}) &= \sum_{d=1}^{D}\sum_{i=1}^{S}\left\{\sum_{\underline{\iota}'=0}^{1}\hat{\gamma}_{1,i}^{(d)}\,\hat{\kappa}_{\underline{\iota}',1,i}^{(d)}\log\pi_{\underline{\iota}',i}+\sum_{t=2}^{n^{(d)}}\hat{\gamma}_{t,i}\,\log\pi_{0,i}\right\} \\ &\quad-\lambda\left(\sum_{i=1}^{S}\{\pi_{0,i}+\pi_{1,i}\}-1\right) & (4.41)\\ \Rightarrow\frac{\partial F_\Pi(\theta,\hat{\theta})}{\partial\pi_{0,i}} &= \sum_{d=1}^{D}\left\{\frac{\hat{\gamma}_{1,i}^{(d)}\,\hat{\kappa}_{0,1,i}^{(d)}}{\pi_{0,i}}+\sum_{t=2}^{n^{(d)}}\frac{\hat{\gamma}_{t,i}}{\pi_{0,i}}\right\}-\lambda \;=\; 0 \text{ when } \theta=\hat{\theta}'\,, \\ \frac{\partial F_\Pi(\theta,\hat{\theta})}{\partial\pi_{1,i}} &= \sum_{d=1}^{D}\frac{\hat{\gamma}_{1,i}^{(d)}\,\hat{\kappa}_{1,1,i}^{(d)}}{\pi_{1,i}}-\lambda \;=\; 0 \text{ when } \theta=\hat{\theta}' \\ \Rightarrow\hat{\lambda}' &= \sum_{d=1}^{D}\sum_{i=1}^{S}\left\{\hat{\gamma}_{1,i}^{(d)}\,\hat{\kappa}_{0,1,i}^{(d)}+\hat{\gamma}_{1,i}^{(d)}\,\hat{\kappa}_{1,1,i}^{(d)}+\sum_{t=2}^{n^{(d)}}\hat{\gamma}_{t,i}^{(d)}\right\} \;=\; \sum_{d=1}^{D}\sum_{i=1}^{S}\sum_{t=1}^{n^{(d)}}\hat{\gamma}_{t,i}^{(d)} \;=\; \sum_{d=1}^{D}n^{(d)}, & (4.42)\end{aligned}$$

which leads to

$$\begin{aligned} \hat{\pi}_{0,i}' &= \frac{\sum_{d=1}^{D}\left\{\hat{\gamma}_{1,i}^{(d)}\,\hat{\kappa}_{0,1,i}^{(d)}+\sum_{t=2}^{n^{(d)}}\hat{\gamma}_{t,i}^{(d)}\right\}}{\sum_{d=1}^{D}n^{(d)}}\,, \\ \hat{\pi}_{1,i}' &= \frac{\sum_{d=1}^{D}\hat{\gamma}_{1,i}^{(d)}\,\hat{\kappa}_{1,1,i}^{(d)}}{\sum_{d=1}^{D}n^{(d)}}\,. & (4.43)\end{aligned}$$

Finally, we iteratively estimate the sequence termination distributions $\omega_{0,i}$ and $\omega_{1,i}$ by optimising

$$
\begin{aligned}
Q_\Omega(\theta,\hat{\theta}) &= \sum_{d=1}^{D}\sum_{\underline{\tau}=0}^{1}\sum_{\overline{i_1}=1}^{S}\cdots\sum_{\overline{i_{n^{(d)}}}=1}^{S}\sum_{\overline{\underline{\tau}}=0}^{1} p(z\,|\,v^{(d)},\hat{\theta})\left\{\sum_{t=1}^{n^{(d)}-1}\log\omega_{0,\overline{i_t}}+\log\omega_{\overline{\underline{\tau}},\overline{i_{n^{(d)}}}}\right\} \\
&= \sum_{d=1}^{D}\left\{\sum_{t=1}^{n^{(d)}-1}\sum_{\overline{i_t}=1}^{S} p(S_t=\sigma_{\overline{i_t}}\,|\,v^{(d)},\hat{\theta})\,\log\omega_{0,\overline{i_t}}\right. \\
&\qquad\left.+\sum_{\overline{i_n^{(d)}}=1}^{S}\sum_{\overline{\underline{\tau}}=0}^{1} p(\tau_{n^{(d)}+1}=\overline{\underline{\tau}},S_{n^{(d)}}=\sigma_{\overline{i_{n^{(d)}}}}\,|\,v^{(d)},\hat{\theta})\,\log\omega_{\overline{\underline{\tau}},\overline{i_{n^{(d)}}}}\right\} \\
&= \sum_{d=1}^{D}\sum_{i=1}^{S}\left\{\sum_{t=1}^{n^{(d)}-1}\hat{\gamma}_{t,i}^{(d)}\log\omega_{0,i}+\sum_{\underline{\tau}'=0}^{1}\hat{\gamma}_{n^{(d)},i}^{(d)}\,\hat{\zeta}_{\underline{\tau}',n^{(d)},i}^{(d)}\log\omega_{\underline{\tau}',i}\right\} \qquad (4.44)
\end{aligned}
$$

subject to the appropriate contraints. Note that we have utilised equations (4.14) and (4.20). Hence, borrowing the Lagrangian constraint of equation (4.28), we maximise

$$
\begin{aligned}
F_\Omega(\theta,\hat{\theta}) &= \sum_{d=1}^{D}\sum_{i=1}^{S}\left\{\sum_{t=1}^{n^{(d)}-1}\hat{\gamma}_{t,i}^{(d)}\log\omega_{0,i}+\sum_{\underline{\tau}'=0}^{1}\hat{\gamma}_{n^{(d)},i}^{(d)}\,\hat{\zeta}_{\underline{\tau}',n^{(d)},i}^{(d)}\log\omega_{\underline{\tau}',i}\right\} \\
&\quad -\sum_{i=1}^{S}\lambda_i\left(\omega_{0,i}+\omega_{1,i}-1\right) \qquad (4.45) \\
\Rightarrow\frac{\partial F_\Omega(\theta,\hat{\theta})}{\partial\omega_{0,i}} &= \sum_{d=1}^{D}\left\{\sum_{t=1}^{n^{(d)}-1}\frac{\hat{\gamma}_{t,i}^{(d)}}{\omega_{0,i}}+\frac{\hat{\gamma}_{n^{(d)},i}^{(d)}\,\hat{\zeta}_{0,n^{(d)},i}^{(d)}}{\omega_{0,i}}\right\}-\lambda_i \;=\; 0 \text{ when } \theta=\hat{\theta}'\,, \\
\frac{\partial F_\Omega(\theta,\hat{\theta})}{\partial\omega_{1,i}} &= \sum_{d=1}^{D}\left\{\frac{\hat{\gamma}_{n^{(d)},i}^{(d)}\,\hat{\zeta}_{1,n^{(d)},i}^{(d)}}{\omega_{1,i}}\right\}-\lambda_i \;=\; 0 \text{ when } \theta=\hat{\theta}' \\
\Rightarrow\hat{\lambda}_i' &= \sum_{d=1}^{D}\left\{\sum_{t=1}^{n^{(d)}-1}\hat{\gamma}_{t,i}^{(d)}+\hat{\gamma}_{n^{(d)},i}^{(d)}\,\hat{\zeta}_{0,n^{(d)},i}^{(d)}+\hat{\gamma}_{n^{(d)},i}^{(d)}\,\hat{\zeta}_{1,n^{(d)},i}^{(d)}\right\} \;=\; \sum_{d=1}^{D}\sum_{t=1}^{n^{(d)}}\hat{\gamma}_{t,i}^{(d)} \\
\Rightarrow\hat{\omega}_{0,i}' &= \frac{\sum_{d=1}^{D}\left\{\sum_{t=1}^{n^{(d)}-1}\hat{\gamma}_{t,i}^{(d)}+\hat{\gamma}_{n^{(d)},i}^{(d)}\,\hat{\zeta}_{0,n^{(d)},i}^{(d)}\right\}}{\sum_{d=1}^{D}\sum_{t=1}^{n^{(d)}}\hat{\gamma}_{t,i}^{(d)}}\,, \\
\hat{\omega}_{1,i}' &= \frac{\sum_{d=1}^{D}\hat{\gamma}_{n^{(d)},i}^{(d)}\,\hat{\zeta}_{1,n^{(d)},i}^{(d)}}{\sum_{d=1}^{D}\sum_{t=1}^{n^{(d)}}\hat{\gamma}_{t,i}^{(d)}} \;=\; 1-\hat{\omega}_{0,i}'\,. \qquad (4.46)
\end{aligned}
$$

Observe in comparison to equation (4.30) for known data that each certainty, represented by a $\delta(\cdot)$ term, has now been replaced by a corresponding posterior probability.