# Notes on Sequence Modelling

G.A. Jarrad

July 31, 2015

## 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 $N = |\vec{R}|$ 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.

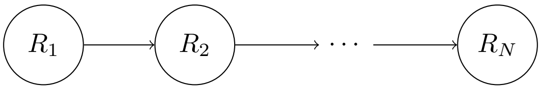

Figure 1.1: *A random process $R$ for generating sequences of arbitrary length $N$. The arrows indicate transitions from one stage in the sequence to the next.*

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

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

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 a subsequence of multiple, contiguously generated sequences.

In order to handle such difficulties, we consider any arbitrary sequence $\vec{r}$ 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: let $\langle\vec{r}$ be a *start sequence* that has a definite start but an indefinite end; and let $\vec{r}\rangle$ be an *end sequence* that has a definite end but an indefinite start. Furthermore, if we know that all of the values of $\vec{r}$ are contiguous values of the same sequence, then we can denote this by introducing additional, paired delimiters. Thus, we use the symbol $\uparrow$ to indicate that the true sequence definitely does not end at the observed value $r_n$, e.g. $\langle\vec{r}\uparrow$, and the symbol $]$ to indicate that we are uncertain as to whether or not the sequence ends at $r_n$, but definitely does not end at an earlier stage, e.g. $\langle\vec{r}]$. Similarly, let $\downarrow$ indicate that the true sequence starts at an earlier stage than $r_1$, e.g. $\downarrow\vec{r}\rangle$, and let $[$ indicate that the sequence might start at $r_1$ or at an earlier stage, e.g. $[\vec{r}\rangle$. Clearly, we may also specify the remaining partial, contiguous sequences $[\vec{r}]$, $\downarrow\vec{r}]$, $[\vec{r}\uparrow$ and $\downarrow\vec{r}\uparrow$.

Under this augmented notation, knowledge about the start of a sequence can be encapsulated in a random indicator variable $\iota_{t-1}$, which takes on the value 1 if some observed $r_t$ is definitely the first stage in the true sequence, or the value 0 if it is not. Similarly, the random indicator variable $\tau_{t+1}$ takes on the value 1 if $r_t$ is definitely the last stage in the true sequence, or the value 0 if it is not. Notionally, the indicators $\iota_0$ and $\tau_{n+1}$ can be thought to correspond to pseudo-stages 0 and $n+1$, such that the generated sequence 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 $\langle\vec{r}\rangle$ is now defined as

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

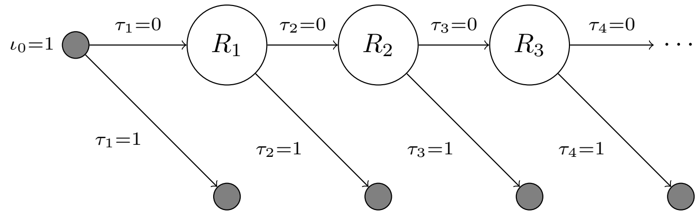

Figure 1.2: *A random process for generating complete sequences of arbitrary length, with explicit stages for sequence initiation and termination. Multiple arrows exiting from a node indicate different possible (mutually exclusive) stage transition pathways.*

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. Conversely, the probability of the start sequence $\langle\vec{r}\!\uparrow$ is

$$p(\langle\vec{r}\!\uparrow) \quad = \quad p(\iota_0\!=\!1, \tau_1\!=\!0, R_1\!=\!r_1, \ldots, \tau_n\!=\!0, R_n\!=\!r_n, \tau_{n+1}\!=\!0)\,, \tag{1.4}$$

and the probability of the end sequence $\downarrow\!\vec{r}\rangle$ is

$$p(\downarrow\!\vec{r}\rangle) \quad = \quad p(\iota_0\!=\!0, \tau_1\!=\!0, R_1\!=\!r_1, \ldots, \tau_n\!=\!0, R_n\!=\!r_n, \tau_{n+1}\!=\!1)\,. \tag{1.5}$$

We may also write the probability of the ambiguous start sequence $\langle\vec{r}]$ as

$$p(\langle\vec{r}]) \quad = \quad p(\iota_0\!=\!1, \tau_1\!=\!0, R_1\!=\!r_1, \ldots, \tau_n\!=\!0, R_n\!=\!r_n, \tau_{n+1}\!=\!*)\,, \tag{1.6}$$

where $\tau_{n+1} = *$ is just a shorthand to indicate that we are uncertain of the true value of $\tau_{n+1}$; probabilistically, the term has no effect and may be dropped. Likewise, the probability of the end sequence $[\vec{r}\rangle$ is

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

The likelihood of the other types of partial sequences can similarly be defined. Generically, we can write the likelihood of any complete or partially complete sequence as

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

where the observed sequence-start indicator $\underline{\iota} \in \{0, 1, *\}$ corresponds to the delimiters $\downarrow$ ,$\langle$, and [, respectively, and the observed sequence-end indicator $\bar{\tau} \in \{0, 1, *\}$ corresponds to the delimiters $\uparrow$, $\rangle$, and ], respectively.

## 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 full-dependency conditional model

$$\begin{aligned} p(\underline{\iota}, \vec{r}, \underline{\tau}) \quad = \quad & p(\iota_0\!=\!\underline{\iota})\, p(\tau_1\!=\!0 \mid \iota_0\!=\!\underline{\iota})\, p(R_1\!=\!r_1 \mid \iota_0\!=\!\underline{\iota}, \tau_1\!=\!0)\, p(\tau_2\!=\!0 \mid \iota_0\!=\!\underline{\iota}, \tau_1\!=\!0, R_1\!=\!r_1) \\ & \times p(R_2\!=\!r_2 \mid \iota_0\!=\!\underline{\iota}, \tau_1\!=\!0, R_1\!=\!r_1, \tau_2\!=\!0) \cdots p(R_n\!=\!r_n \mid \iota_0\!=\!1, \ldots, \tau_n\!=\!0) \\ & \times p(\tau_{n+1}\!=\!1 \mid \iota_0\!=\!1, \ldots, R_n\!=\!r_n)\,. \end{aligned} \tag{2.1}$$

We can generalise this model by first defining a forward observation sub-sequence $\vec{r}_t = (r_1, r_2, \ldots, r_t)$, for $t = 1, 2, \ldots, n$, where $\vec{r}_0$ is taken to be the empty sub-sequence by definition. For later convenience, we

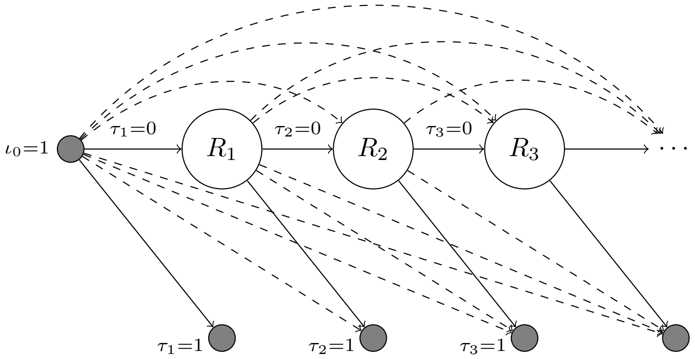

Figure 2.1: *A fully-dependent, causal process for generating complete sequences of arbitrary length. Solid arrows indicate possible stage transitions. Both dashed arrows and solid arrows indicate parent–child dependencies, such that the child node is conditionally dependent on the parent and all other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

also define the backward observation sub-sequence $\overleftarrow{r}_t = (r_t, r_{t+1}, \ldots, r_n)$, for $t = n, n-1, \ldots, 1$; likewise, $\overleftarrow{r}_{n+1}$ is an empty sub-sequence. Hence, we obtain

$$\begin{aligned} p(\underline{\iota}, \vec{r}, \underline{\tau}) \quad = \quad & p(\iota_0 \!=\! \underline{\iota}) \prod_{t=1}^{n} \Big\{ p(\tau_t \!=\! 0 \mid \iota_0 \!=\! \underline{\iota}, \vec{\tau}_{t-1} \!=\! \vec{0}_{t-1}, \vec{R}_{t-1} \!=\! \vec{r}_{t-1}) \\ & \qquad\qquad \times\, p(R_t \!=\! r_t \mid \iota_0 \!=\! \underline{\iota}, \vec{\tau}_t \!=\! \vec{0}_t, \vec{R}_{t-1} \!=\! \vec{r}_{t-1}) \Big\} \\ & \times p(\tau_{n+1} \!=\! \bar{\tau} \mid \iota_0 \!=\! \underline{\iota}, \vec{\tau}_n \!=\! \vec{0}, \vec{R}_n \!=\! \vec{r}_n)\,. \end{aligned} \tag{2.2}$$

In practice, the full-dependency model is usually considerably simplified by dropping some of the explicit (dashed) dependencies. For example, one might limit the conditionality on past values to a maximum of $m$ 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. The second-order Markov sequence process is depicted in Figure 2.2.

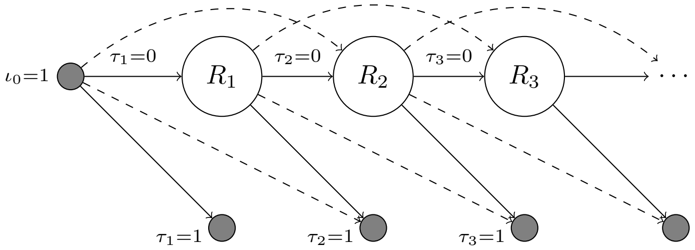

Figure 2.2: *A second-order Markov process for generating complete sequences of arbitrary length.*

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

$$\begin{aligned} p(\underline{\iota}, \vec{r}, \underline{\tau}) \quad = \quad & p(\iota_0 \!=\! \underline{\iota})\, p(\tau_1 \!=\! 0 \mid \iota_0 \!=\! \underline{\iota}) p(R_1 \!=\! r_1 \mid \iota_0 \!=\! \underline{\iota}, \tau_1 \!=\! 0) \\ & \times \prod_{t=2}^{n} \{ p(\tau_t \!=\! 0 \mid R_{t-1} \!=\! r_{t-1})\, p(R_t \!=\! r_t \mid \tau_t \!=\! 0, R_{t-1} \!=\! r_{t-1}) \} \\ & \times p(\tau_{n+1} \!=\! \underline{\tau} \mid R_n \!=\! r_n)\,. \end{aligned} \tag{2.3}$$

This is just a 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 $S_t$. 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 $(S_{t+1}, X_{t+1})$.

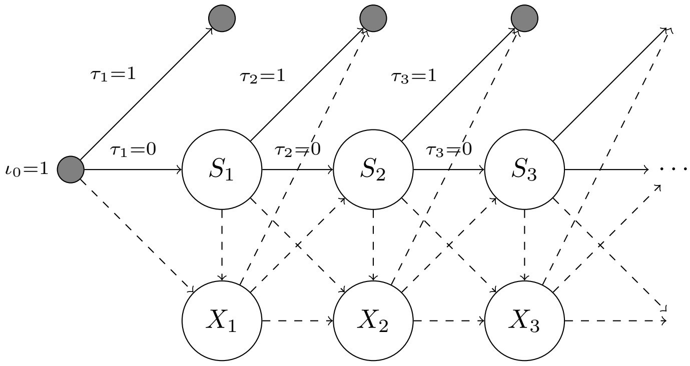

Figure 3.1: *A random process for generating complete, stateful sequences of arbitrary length, with explicit cross-dependencies between adjacent stages.*

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

$$\begin{aligned} p(\underline{\iota}, \vec{s}, \vec{x}, \underline{\tau}) \;\; = \;\; & p(\iota_0 = \underline{\iota})\, p(\tau_1 = 0 \mid \iota_0 = \underline{\iota})\, p(S_1 = s_1 \mid \iota_0 = \underline{\iota}, \tau_1 = 0) \\ & \times\; p(X_1 = x_1 \mid S_1 = s_1, \iota = \underline{\iota}, \tau_1 = 0) \\ & \times \prod_{t=2}^{n} \{ p(\tau_t = 0 \mid S_{t-1} = s_{t-1}, X_{t-1} = x_{t-1}) \\ & \qquad \times\; p(S_t = s_t \mid \tau_t = 0, , S_{t-1} = s_{t-1}, X_{t-1} = x_{t-1}) \\ & \qquad \times\; p(X_t = x_t \mid S_t = s_t, S_{t-1} = s_{t-1}, X_{t-1} = x_{t-1}) \} \\ & \times\, p(\tau_{n+1} = \overline{\tau} \mid S_n = s_n, X_n = x_n)\,. \end{aligned} \tag{3.1}$$

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. This restricted process is depicted in Figure 3.2.

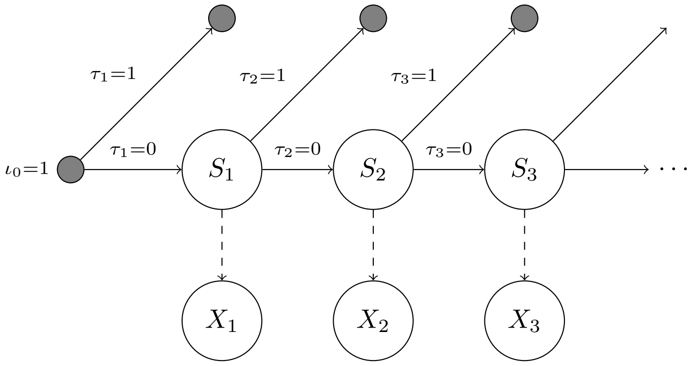

Figure 3.2: *A first-order Markov process for generating complete, stateful sequences of arbitrary length.*

The corresponding sequence model is now given by

$$\begin{aligned} p(\underline{\iota}, \vec{s}, \vec{x}, \underline{\tau}) &= p(\iota_0 = \underline{\iota})\, p(\tau_1 = 0 \mid \iota_0 = \underline{\iota})\, p(S_1 = s_1 \mid \iota_0 = \underline{\iota}, \tau_1 = 0)\, p(X_1 = x_1 \mid S_1 = s_1) \\ & \quad \times \prod_{t=2}^{n} \{ p(\tau_t = 0 \mid S_{t-1} = s_{t-1})\, p(S_t = s_t \mid \tau_t = 0, , S_{t-1} = s_{t-1}) \\ & \qquad \times\ p(X_t = x_t \mid S_t = s_t) \}\ p(\tau_{n+1} = \bar{\tau} \mid S_n = s_n)\,. \end{aligned} \tag{3.2}$$

So far, we have made no assumption as to whether the states take continuous values or discrete values. In the next section, we consider the important sub-class of models where the states are discrete-valued.

# 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 corresponding sequence models where the state $S_t$ at any stage $t$ may now only take *discrete* values in the set $\mathcal{S} = \{\sigma_1, \sigma_2, \ldots, \sigma_S\}$. Hence, the sequence of states may arbitrarily be specified as $\vec{s} = (\sigma_{i_1}, \sigma_{i_2}, \ldots, \sigma_{i_n})$, where each $i_t \in \{1, 2, \ldots, S\}$. In the event that a particular state $S_t$ is unobserved, we say that the state is *missing* or *hidden*, and denote $i_t = *$ and $s_t = *$. In the situation where all values of $\vec{s}$ are unobserved, the sequence model (3.2) is known as a *hidden-state Markov model* (HMM).

The sequence model (3.2) may now be explicitly conditioned on a general parameter $\theta$ 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[1] sequence (with no hidden states) is given by

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

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

$$\pi_{0,t,i} = p(\iota_{t-1} = 0 \mid \theta)\, p(\tau_t = 0 \mid \iota_{t-1} = 0, \theta)\, p(S_t = \sigma_i \mid \iota_{t-1} = 0, \tau_t = 0, \theta)\,, \tag{4.2}$$

$$\pi_{1,t,i} = p(\iota_{t-1} = 1 \mid \theta)\, p(\tau_t = 0 \mid \iota_{t-1} = 1, \theta)\, p(S_t = \sigma_i \mid \iota_{t-1} = 1, \tau_t = 0, \theta)\,, \tag{4.3}$$

and

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

Observe that each state $S_t$ for $t > 1$ is a non-initial state, governed by $\pi_{0,t,i_t}$. 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 4.3).

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

$$\omega_{0,t,i} = p(\tau_{t+1} = 0 \mid S_t = \sigma_i, \theta)\,, \tag{4.5}$$

$$\omega_{1,t,i} = p(\tau_{t+1} = 1 \mid S_t = \sigma_i, \theta)\,, \tag{4.6}$$

and

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

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

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 $\Gamma$, where

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

[1] 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.

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

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

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.

## 4.1 Modified Forward–Backward Algorithm

The sequence model (4.10) can be efficiently evaluated by marginalising over the state of each stage in turn, using a modification of the *forward–backward algorithm* to include knowledge of sequence initiation and termination. The forward pass involves first summing over all terms containing $i'_1$, then over all remaining terms containing $i'_2$, and so on up to summing over $i'_n$. This is equivalent to evaluating the reordered model

$$\begin{aligned} p(\underline{\iota}, \vec{s}, \vec{x}, \underline{\tau} \,|\, \theta) \quad = \quad & \left\{ \sum_{i'_n=1}^{S} \delta(i'_n\!=\!i_n) \cdots \left\{ \sum_{i'_2=1}^{S} \delta(i'_2\!=\!i_2) \right.\right. \\ & \qquad \times \left\{ \sum_{i'_1=1}^{S} \delta(i'_1\!=\!i_1)\, \pi_{\underline{\iota},1,i'_1}\, o_{1,i'_1}(x_1)\, \omega_{0,1,i'_1}\, \Gamma_{1,i'_1,i'_2} \right\} \\ & \left.\left. o_{2,i'_2}(x_2) \omega_{0,2,i'_2}\, \Gamma_{2,i'_2,i'_3} \right\} \cdots\; o_{n,i'_n}(x_n)\, \omega_{\underline{\tau},n,i'_n} \right\}\,. \end{aligned} \tag{4.11}$$

Conversely, the backaward pass reverses the order of evaluation, first summing over all terms containing $i'_n$, and then over all remaining terms containing $i'_{n-1}$, and so on down to summing over $i'_1$. This is equivalent to evaluating the reordered model

$$\begin{aligned} p(\underline{\iota}, \vec{s}, \vec{x}, \underline{\tau} \,|\, \theta) \quad = \quad & \left\{ \sum_{i'_1=1}^{S} \delta(i'_1\!=\!i_1)\, \pi_{\underline{\iota},1,i'_1}\, o_{1,i'_1}(x_1)\, \omega_{0,1,i'_1} \cdots \right. \\ & \times \left\{ \sum_{i'_{n-1}=1}^{S} \delta(i'_{n-1}\!=\!i_{n-1})\, \Gamma_{n-2,i'_{n-2},i'_{n-1}}\, o_{n-1,i'_{n-1}}(x_{n-1})\, \omega_{0,n-1,i'_{n-1}} \right. \\ & \quad \left.\left. \times \left\{ \sum_{i'_n=1}^{S} \delta(i'_n\!=\!i_n)\, \Gamma_{n-1,i'_{n-1},i'_n}\, o_{n,i'_n}(x_n)\, \omega_{\underline{\tau},n,i'_n} \right\} \right\} \cdots \right\}\,. \end{aligned} \tag{4.12}$$

A more efficient mechanism for evaluation comes from making use of the first-order Markov dependencies. Notionally, from the process depicted in Figure 3.2, we may arbitrarily consider the transition from some stage $t$ to stage $t+1$, and partition the sequence into: (i) past values from the initial node up to and including $S_t$ and $X_t$; and (ii) future values from $S_{t+1}$ and $X_{t+1}$ up to and including the terminal node. Note that the termination or non-termination of stage $t$ is governed by $\tau_{t+1}$, which is therefore a future value. The Markov dependency then implies that the future values are conditioned only on state $S_t$ via $\tau_{t+1}$ and $S_{t+1}$. Hence, remembering that notionally $s_t = \sigma_{i_t}$, model (4.10) reduces to

$$
\begin{aligned}
p(\underline{\iota}, \vec{s}, \vec{x}, \underline{\tau} \,|\, \theta) &= \sum_{i=1}^{S} p(S_t = \sigma_i, \underline{\iota}, \vec{s}, \vec{x}, \underline{\tau} \,|\, \theta) \\
&= \sum_{i=1}^{S} \delta(i = i_t)\, p(\underline{\iota}, \vec{s}_{t-1} \circ \sigma_i \circ \overleftarrow{s}_{t+1}, \vec{x}, \underline{\tau} \,|\, \theta) \\
&= \sum_{i=1}^{S} \delta(i = i_t)\, p(\underline{\iota}, \vec{s}_{t-1} \circ \sigma_i, \vec{x}_t \,|\, \theta)\, p(\downarrow \overleftarrow{s}_{t+1}, \overleftarrow{x}_{t+1}, \underline{\tau} \,|\, S_t = \sigma_i, \theta) \\
&= \sum_{i=1}^{S} \delta(i = i_t)\, \alpha_{t,i}\, \beta_{t,i}\,, \qquad (4.13)
\end{aligned}
$$

where $\circ$ represents sequence concatenation. The forward step $\alpha_{t,i}$ is defined as

$$
\begin{aligned}
\alpha_{t,i} &= p(\underline{\iota}, \vec{s}_{t-1} \circ \sigma_i, \vec{x}_t \,|\, \theta) \\
&= p(\underline{\iota}, \vec{s}_{t-1} \circ \sigma_i, \vec{x}_{t-1} \,|\, \theta)\, p(X_t = x_t \,|\, S_t = \sigma_{i'_t}, \theta) \\
&= \bar{\alpha}_{t,i}\, o_{t,i}(x_t)\,, \qquad (4.14)
\end{aligned}
$$

where $\bar{\alpha}_{t,i}$ is recursively defined as

$$
\begin{aligned}
\bar{\alpha}_{t,i} &= p(\underline{\iota}, \vec{s}_{t-1} \circ \sigma_i, \vec{x}_{t-1} \,|\, \theta) \\
&= \sum_{j=1}^{S} p(S_{t-1} = \sigma_j, \underline{\iota}, \vec{s}_{t-1} \circ \sigma_i, \vec{x}_{t-1} \,|\, \theta) \\
&= \sum_{j=1}^{S} \delta(j = i_{t-1})\, p(\underline{\iota}, \vec{s}_{t-2} \circ \sigma_j \circ \sigma_i, \vec{x}_{t-1} \,|\, \theta) \\
&= \sum_{j=1}^{S} \delta(j = i_{t-1})\, p(\underline{\iota}, \vec{s}_{t-2} \circ \sigma_j, \vec{x}_{t-1} \,|\, \theta)\, p(\tau_t = 0 \,|\, S_{t-1} = \sigma_j, \theta) \\
&\qquad \times\ p(S_t = \sigma_i \,|\, \tau_t = 0, S_{t-1} = \sigma_j, \theta) \\
&= \sum_{j=1}^{S} \delta(j = i_{t-1})\, \alpha_{t-1,j}\, \omega_{0,t-1,j}\, \Gamma_{t-1,j,i}\,, \qquad (4.15)
\end{aligned}
$$

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

$$
\begin{aligned}
\alpha_{1,i} &= p(\underline{\iota}, S_1 = \sigma_i, X_1 = x_1 \,|\, \theta) \\
&= p(\iota_0 = \underline{\iota})\, p(\tau_1 = 0 \,|\, \iota = \underline{\iota}, \theta)\, p(S_1 = \sigma_i \,|\, \iota_0 = \underline{\iota}, \tau_1 = 0, \theta)\, p(X_1 = x_1 \,|\, S_1 = \sigma_i, \theta) \\
&= \pi_{\underline{\iota},1,i}\, o_{1,i}(x_1)\,. \qquad (4.16)
\end{aligned}
$$

Hence, observe that $\alpha_{2,i'_2}$, is just the entire summation over $i'_1$ from the forward model (4.11). Also note that the standard forward pass derivation commences with the equivalent of $\pi_{*,1,i}$ and does not include the $\delta(\cdot)$ or $\omega$ terms.

Conversely to the forward pass, the backward step $\beta_{t,i}$ is defined as

$$
\begin{aligned}
\beta_{t,i} &= p(\downarrow \overleftarrow{s}_{t+1}, \overleftarrow{x}_{t+1}, \underline{\tau} \,|\, S_t = \sigma_i, \theta) \\
&= p(\tau_{t+1} = 0 \,|\, S_t = \sigma_i, \theta)\, p(\overleftarrow{s}_{t+1}, \overleftarrow{x}_{t+1}, \underline{\tau} \,|\, \tau_{t+1} = 0, S_t = \sigma_i, \theta) \\
&= \omega_{0,t,i}\, \bar{\beta}_{t,i}\,, \qquad (4.17)
\end{aligned}
$$

where

$$
\begin{aligned}
\bar{\beta}_{t,i} &= p(\overleftarrow{s}_{t+1}, \overleftarrow{x}_{t+1}, \underline{\tau} \,|\, \tau_{t+1}=0, S_t=\sigma_i, \theta) \\
&= \sum_{j=1}^{S} p(S_{t+1}=\sigma_j, \overleftarrow{s}_{t+1}, \overleftarrow{x}_{t+1}, \underline{\tau} \,|\, \tau_{t+1}=0, S_t=\sigma_i, \theta) \\
&= \sum_{j=1}^{S} \delta(j=i_{t+1})\, p(\sigma_j \circ \overleftarrow{s}_{t+2}, x_{t+1} \circ \overleftarrow{x}_{t+2}, \underline{\tau} \,|\, \tau_{t+1}=0, S_t=\sigma_i, \theta) \\
&= \sum_{j=1}^{S} \delta(j=i_{t+1})\, p(S_{t+1}=\sigma_j \,|\, \tau_{t+1}=0, S_t=\sigma_i, \theta)\, p(X_{t+1} = x_{t+1} \,|\, S_{t+1}=\sigma_j, \theta) \\
&\qquad \times p(\downarrow \overleftarrow{s}_{t+2}, \overleftarrow{x}_{t+2}, \underline{\tau} \,|\, S_{t+1}=\sigma_j, \theta) \\
&= \sum_{j=1}^{S} \delta(j=i_{t+1})\, \Gamma_{t,i,j}\, o_{t+1,j}(x_{t+1})\, \beta_{t+1,j}\,, \qquad (4.18)
\end{aligned}
$$

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

$$
\beta_{n,i} \;=\; p(\tau_{n+1}=\underline{\tau} \,|\, S_n=\sigma_i, \theta) \;=\; \omega_{\underline{\tau},n,i}\,. \qquad (4.19)
$$

Observe that $\bar{\beta}_{n-1,i'_{n-1}}$ is just the entire summation over $i'_n$ from the backward model (4.12). Also note that the standard backward pass derivation commences with the equivalent of $\omega_{*,n,i} = 1$, and does not include the $\delta(\cdot)$ or $\omega$ terms.

## 4.2 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 4.1 enables all of these calculations.

For instance, from equation (4.13), the posterior probabilities of state $S_t$ given an observed sequence are computed as

$$
\begin{aligned}
\gamma_{t,i} &= p(S_t=\sigma_i \,|\, \underline{\iota}, \vec{s}, \vec{x}, \underline{\tau}, \theta) \\
&= \frac{p(S_t=\sigma_i, \underline{\iota}, \vec{s}, \vec{x}, \underline{\tau} \,|\, \theta)}{p(\underline{\iota}, \vec{s}, \vec{x}, \underline{\tau} \,|\, \theta)} \\
&= \frac{\delta(i = i_t)\, \alpha_{t,i}\, \beta_{t,i}}{\sum_{i'=1}^{S} \delta(i' = i_t)\, \alpha_{t,i'}\, \beta_{t,i'}}\,. \qquad (4.20)
\end{aligned}
$$

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, partially observed sequence via

$$
\begin{aligned}
p(\downarrow \sigma_i \,|\, \underline{\iota}, \vec{s}_n, \vec{x}_n, \underline{\tau}, \theta) &= p(\tau_{n+1}=0, S_{n+1}=\sigma_i \,|\, \underline{\iota}, \vec{s}_n, \vec{x}_n, \underline{\tau}, \theta) \\
&= \frac{p(\tau_{n+1}=0, S_{n+1}=\sigma_i, \underline{\iota}, \vec{s}_n, \vec{x}_n, \underline{\tau} \,|\, \theta)}{p(\underline{\iota}, \vec{s}_n, \vec{x}_n, \underline{\tau} \,|\, \theta)} \\
&= \delta(\underline{\tau}=0) \frac{p(\underline{\iota}, \vec{s}_n \circ \sigma_i, \vec{x}_n \,|\, \theta)}{p(\underline{\iota}, \vec{s}_n, \vec{x}_n, \underline{\tau} \,|\, \theta)} \\
&= \delta(\underline{\tau}=0) \frac{\bar{\alpha}_{n+1,i}}{\sum_{i'=1}^{S} \delta(i=i_n)\, \alpha_{n,i'}\, \beta_{n,i'}}\,, \qquad (4.21)
\end{aligned}
$$

from equation (4.15). Consequently, we may also predict the future value of $X_{t+1}$ via

$$
\begin{aligned}
p(\downarrow x \,|\, \underline{\iota}, \vec{s}_n, \vec{x}_n, \underline{\tau}, \theta) &= \sum_{i=1}^{S} p(\downarrow \sigma_i, x \,|\, \underline{\iota}, \vec{s}_n, \vec{x}_n, \underline{\tau}, \theta) \\
&= \sum_{i=1}^{S} p(\downarrow \sigma_i \,|\, \underline{\iota}, \vec{s}_n, \vec{x}_n, \underline{\tau}, \theta)\, p(X_{n+1}=x \,|\, S_{n+1}=\sigma_i, \theta) \\
&= \delta(\underline{\tau}=0) \frac{\sum_{i=1}^{S} \bar{\alpha}_{n+1,i}\, o_{n+1,i}(x)}{\sum_{i'=1}^{S} \delta(i=i_n)\, \alpha_{n,i'}\, \beta_{n,i'}}\,. \qquad (4.22)
\end{aligned}
$$

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

$$
\begin{aligned}
\xi_{t,i,j} &= p(S_t\!=\!\sigma_i, S_{t+1}\!=\!\sigma_j \,|\, \underline{\iota}, \vec{s}, \vec{x}, \underline{\tau}, \theta) \\
&= \frac{p(S_t\!=\!\sigma_i, S_{t+1}\!=\!\sigma_j, \underline{\iota}, \vec{s}, \vec{x}, \underline{\tau} \,|\, \theta)}{p(\underline{\iota}, \vec{s}, \vec{x}, \underline{\tau} \,|\, \theta)} \\
&= \frac{\sum_{i=1}^{S}\sum_{j=1}^{S} \delta(i=i_t)\delta(j=i_{t+1})\, p(\underline{\iota}, \vec{s}_{t-1}\circ\sigma_i\circ\sigma_j\circ\overleftarrow{s}_{t+2}, \vec{x}, \underline{\tau} \,|\, \theta)}{p(\underline{\iota}, \vec{s}, \vec{x}, \underline{\tau} \,|\, \theta)} \\
&= \frac{\sum_{i=1}^{S}\sum_{j=1}^{S} \delta(i=i_t)\delta(j=i_{t+1})\, \alpha_{t,i}\, \omega_{0,t,i}\, \Gamma_{t,i,j}\, o_{t+1,j}(x_{t+1})\, \beta_{t+1,j}}{\sum_{i'=1}^{S} \delta(i\!=\!i_n)\, \alpha_{n,i'}\, \beta_{n,i'}}\,,
\end{aligned}
\tag{4.23}
$$

since

$$
\begin{aligned}
p(\underline{\iota}, \vec{s}_{t-1}\circ\sigma_i\circ\sigma_j\circ\overleftarrow{s}_{t+2}, \vec{x}, \underline{\tau} \,|\, \theta) &= p(\underline{\iota}, \vec{s}_{t-1}\circ\sigma_i, \vec{x}_t \,|\, \theta)\, p({\downarrow}\sigma_j, x_{t+1} \,|\, S_t\!=\!\sigma_i, \theta) \\
&\quad \times p({\downarrow}\overleftarrow{s}_{t+2}, \overleftarrow{x}_{t+2}, \underline{\tau} \,|\, S_{t+1}\!=\!\sigma_j, \theta) \\
&= \alpha_{t,i}\, \omega_{0,t,i}\, \Gamma_{t,i,j}\, o_{t+1,j}(x_{t+1})\, \beta_{t+1,j}\,,
\end{aligned}
\tag{4.24}
$$

from the forward pass (4.14) and the backward pass (4.17). Observe that

$$
\begin{aligned}
\gamma_{t,i} &= p(S_t\!=\!\sigma_i \,|\, \underline{\iota}, \vec{s}, \vec{x}, \underline{\tau}, \theta) \\
&= \sum_{j=1}^{S} p(S_t\!=\!\sigma_i, S_{t+1}\!=\!\sigma_j \,|\, \underline{\iota}, \vec{s}, \vec{x}, \underline{\tau}, \theta) \;=\; \sum_{j=1}^{S} \gamma_{t,i,j}\,,
\end{aligned}
\tag{4.25}
$$

from equation (4.20).

## 4.3 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

$$
p(v^{(d)} \,|\, \theta) \;=\; \pi_{\underline{\iota}^{(d)}, i_1^{(d)}}\, o_{i_1^{(d)}}(x_1^{(d)}) \left\{ \prod_{t=1}^{n^{(d)}-1} \omega_{0, i_t^{(d)}}\, \Gamma_{i_t^{(d)}, i_{t+1}^{(d)}}\, o_{i_{t+1}^{(d)}}(x_{t+1}^{(d)}) \right\} \omega_{\underline{\tau}^{(d)}, i_{n^{(d)}}^{(d)}}\,,
\tag{4.26}
$$

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

$$
\ell(v^{(d)} \,|\, \theta) \;=\; \log \pi_{\underline{\iota}^{(d)}, i_1^{(d)}} + \sum_{t=1}^{n^{(d)}-1} \log \omega_{0, i_t^{(d)}} \Gamma_{i_t^{(d)}, i_{t+1}^{(d)}} + \sum_{t=1}^{n^{(d)}} \log o_{i_t^{(d)}}(x_{i_t^{(d)}}) + \log \omega_{\underline{\tau}^{(d)}, i_{n^{(d)}}^{(d)}}\,.
\tag{4.27}
$$

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

$$
L(\theta) \;=\; \log p(\mathbb{V} \,|\, \theta) \;=\; \log \prod_{d=1}^{D} p(v^{(d)} \,|\, \theta) \;=\; \sum_{d=1}^{D} \ell(v^{(d)}, \theta)\,.
\tag{4.28}
$$

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

$$\begin{aligned}
F_\Gamma(\theta) &= \sum_{d=1}^{D}\sum_{t=1}^{n^{(d)}-1} \log\Gamma_{i_t^{(d)},i_{t+1}^{(d)}} - \sum_{i=1}^{S}\lambda_i\left(\sum_{j=1}^{S}\Gamma_{i,j}-1\right) &\quad (4.29)\\
\Rightarrow \frac{\partial F_\Gamma(\theta)}{\partial\Gamma_{i,j}} &= \sum_{d=1}^{D}\sum_{t=1}^{n^{(d)}-1}\delta(i=i_t^{(d)})\,\delta(j=i_{t+1}^{(d)})\,\frac{1}{\Gamma_{i,j}}-\lambda_i \;=\; 0 \text{ when } \theta=\hat{\theta}\\
\Rightarrow \hat{\lambda}_i &= \sum_{j=1}^{S}\sum_{d=1}^{D}\sum_{t=1}^{n^{(d)}-1}\delta(i=i_t^{(d)})\,\delta(j=i_{t+1}^{(d)}) \;=\; \sum_{d=1}^{D}\sum_{t=1}^{n^{(d)}-1}\delta(i=i_t^{(d)})\\
\Rightarrow \hat{\Gamma}_{i,j} &= \frac{\sum_{d=1}^{D}\sum_{t=1}^{n^{(d)}-1}\delta(i=i_t^{(d)})\,\delta(j=i_{t+1}^{(d)})}{\sum_{d=1}^{D}\sum_{t=1}^{n^{(d)}-1}\delta(i=i_t^{(d)})}\,. &\quad (4.30)
\end{aligned}$$

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

$$\begin{aligned}
F_\Omega(\theta) &= \sum_{d=1}^{D}\left\{\sum_{t=1}^{n^{(d)}-1}\log\omega_{0,i_t^{(d)}}+\log\omega_{\underline{\tau}^{(d)},i_{n^{(d)}}^{(d)}}\right\}-\sum_{i=1}^{S}\lambda_i\left(\omega_{0,i}+\omega_{1,i}-1\right) &\quad (4.31)\\
\Rightarrow \frac{\partial F_\Omega(\theta)}{\partial\omega_{0,i}} &= \sum_{d=1}^{D}\left\{\sum_{t=1}^{n^{(d)}-1}\frac{\delta(i=i_t^{(d)})}{\omega_{0,i}}+\frac{\delta(\underline{\tau}^{(d)}\!=\!0)\,\delta(i=i_{n^{(d)}}^{(d)})}{\omega_{0,i}}\right\}-\lambda_i\,,\\
\frac{\partial F_\Omega(\theta)}{\partial\omega_{1,i}} &= \sum_{d=1}^{D}\left\{\frac{\delta(\underline{\tau}^{(d)}\!=\!1)\,\delta(i=i_{n^{(d)}}^{(d)})}{\omega_{1,i}}\right\}-\lambda_i\,. &\quad (4.32)
\end{aligned}$$

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

$$\begin{aligned}
\hat{\lambda}_i &= \sum_{d=1}^{D}\sum_{t=1}^{n^{(d)}}\delta(i=i_t^{(d)})\\
\Rightarrow \hat{\omega}_{0,i} &= \frac{\sum_{d=1}^{D}\left\{\sum_{t=1}^{n^{(d)}-1}\delta(i=i_t^{(d)})+\delta(\underline{\tau}^{(d)}\!=\!0)\,\delta(i=i_{n^{(d)}}^{(d)})\right\}}{\sum_{d=1}^{D}\sum_{t=1}^{n^{(d)}}\delta(i=i_t^{(d)})}\,,\\
\hat{\omega}_{1,i} &= \frac{\sum_{d=1}^{D}\delta(\underline{\tau}^{(d)}\!=\!1)\,\delta(i=i_{n^{(d)}}^{(d)})}{\sum_{d=1}^{D}\sum_{t=1}^{n^{(d)}}\delta(i=i_t^{(d)})}\,. &\quad (4.33)
\end{aligned}$$

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 (4.27), we maximise the function

$$\begin{aligned}
F_\Pi(\theta) &= \sum_{d=1}^{D}\left\{\log\pi_{\underline{\iota}^{(d)},i_1^{(d)}}+\sum_{t=2}^{n^{(d)}}\log\pi_{0,i_t^{(d)}}\right\}-\lambda\left(\sum_{i=1}^{S}\{\pi_{0,i}+\pi_{1,i}\}-1\right) &\quad (4.34)\\
\Rightarrow \frac{\partial F_\Pi(\theta)}{\partial\pi_{0,i}} &= \sum_{d=1}^{D}\left\{\frac{\delta(\underline{\iota}^{(d)}\!=\!0)\,\delta(i_1^{(d)}\!=\!i)}{\pi_{0,i}}+\sum_{t=2}^{n^{(d)}}\frac{\delta(i_t^{(d)}\!=\!i)}{\pi_{0,i}}\right\}-\lambda\,,\\
\frac{\partial F_\Pi(\theta)}{\partial\pi_{1,i}} &= \sum_{d=1}^{D}\left\{\frac{\delta(\underline{\iota}^{(d)}\!=\!1)\,\delta(i_1^{(d)}\!=\!i)}{\pi_{1,i}}\right\}-\lambda\,. &\quad (4.35)
\end{aligned}$$

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

$$
\begin{aligned}
\hat{\lambda} &= \sum_{i=1}^{S}\sum_{d=1}^{D}\sum_{t=1}^{n^{(d)}} \delta(i_t^{(d)}\!=\!i) \;=\; \sum_{d=1}^{D} n^{(d)} \\
\Rightarrow \hat{\pi}_{0,i} &= \frac{\sum_{d=1}^{D}\left\{\delta(\underline{\iota}^{(d)}\!=\!0)\,\delta(i_1^{(d)}\!=\!i) + \sum_{t=2}^{n^{(d)}}\delta(i_t^{(d)}\!=\!i)\right\}}{\sum_{d=1}^{D} n^{(d)}}\,, \\
\hat{\pi}_{1,i} &= \frac{\sum_{d=1}^{D}\delta(\underline{\iota}^{(d)}\!=\!1)\,\delta(i_1^{(d)}\!=\!i)}{\sum_{d=1}^{D} n^{(d)}}\,. \qquad (4.36)
\end{aligned}
$$

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.

## 4.4 Posterior Parameter Estimation with Missing Data

In contrast to Section 4.3, 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[2]

$$
\begin{aligned}
Q(\theta) &= E_{\mathbb{Z}\,|\,\mathbb{V},\theta}\left[\log p(\mathbb{Z},\mathbb{V}\,|\,\theta)\right] \\
&= E_{\mathbb{Z}\,|\,\mathbb{V},\theta}\left[\sum_{d=1}^{D}\log p(z^{(d)}, v^{(d)}\,|\,\theta)\right] \\
&= \sum_{d=1}^{D} E_{\mathbb{Z}\,|\,\mathbb{V},\theta}\left[\ell(\overline{\underline{\iota}^{(d)}}, \overline{\vec{s}^{(d)}}, \vec{x}^{(d)}, \overline{\underline{\tau}^{(d)}}\,|\,\theta)\right] \\
&= \sum_{d=1}^{D}\sum_{\overline{\underline{\iota}}=0}^{1}\sum_{\overline{i_1}=1}^{S}\cdots\sum_{\overline{i_{n^{(d)}}}}^{S}\sum_{\overline{\underline{\tau}}=0}^{1} p(\overline{\underline{\iota}}, \overline{\vec{s}}, \overline{\underline{\tau}}\,|\,\underline{\iota}^{(d)}, \vec{s}^{(d)}, \vec{x}^{(d)}, \underline{\tau}^{(d)}, \theta)\,\ell(\overline{\underline{\iota}}, \overline{\vec{s}}, \vec{x}^{(d)}, \overline{\underline{\tau}}\,|\,\theta) \\
&= \sum_{d=1}^{D}\sum_{\overline{\underline{\iota}}=0}^{1}\sum_{\overline{i_1}=1}^{S}\cdots\sum_{\overline{i_{n^{(d)}}}}^{S}\sum_{\overline{\underline{\tau}}=0}^{1} p(z\,|\,v^{(d)},\theta)\,\ell(\overline{v^{(d)}}\,|\,\theta)\,, \qquad (4.37)
\end{aligned}
$$

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

$$
\begin{aligned}
Q(\theta,\hat{\theta}_k) &= E_{\mathbb{Z}\,|\,\mathbb{V},\hat{\theta}_k}\left[\log p(\mathbb{Z},\mathbb{V}\,|\,\theta)\right] \\
&= \sum_{d=1}^{D}\sum_{\overline{\underline{\iota}}=0}^{1}\sum_{\overline{i_1}=1}^{S}\cdots\sum_{\overline{i_{n^{(d)}}}}^{S}\sum_{\overline{\underline{\tau}}=0}^{1} p(z\,|\,v^{(d)},\hat{\theta}_k)\,\ell(\overline{v^{(d)}}\,|\,\theta)\,. \qquad (4.38)
\end{aligned}
$$

[2] 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 $\sum_{d=1}^{D}[\phi^{(d)}/\psi^{(d)}]/D$.

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

$$\hat{\theta}_{k+1} \quad = \quad \arg\max_{\theta} Q(\theta, \hat{\theta}_k)\,. \tag{4.39}$$

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

Hence, following the derivation of equation (4.29), we iteratively estimate the state transitions by maximising

$$\begin{aligned}
F_\Gamma(\theta, \hat{\theta}) &= \sum_{d=1}^{D} \sum_{\underline{\iota}=0}^{1} \sum_{\overline{i_1}=1}^{S} \cdots \sum_{\overline{i_{n^{(d)}}}=1}^{S} \sum_{\overline{\tau}=0}^{1} \sum_{t=1}^{n^{(d)}-1} p(z \mid v^{(d)}, \hat{\theta}) \, \log \Gamma_{\overline{i_t}, \overline{i_{t+1}}} - \sum_{i=1}^{S} \lambda_i \left( \sum_{j=1}^{S} \Gamma_{i,j} - 1 \right) \\
&= \sum_{d=1}^{D} \sum_{t=1}^{n^{(d)}-1} \sum_{i=1}^{S} \sum_{j=1}^{S} p(S_t = \sigma_i, S_{t+1} = \sigma_j \mid v^{(d)}, \hat{\theta}) \, \log \Gamma_{i,j} - \sum_{i=1}^{S} \lambda_i \left( \sum_{j=1}^{S} \Gamma_{i,j} - 1 \right) \\
&= \sum_{d=1}^{D} \sum_{t=1}^{n^{(d)}-1} \sum_{i=1}^{S} \sum_{j=1}^{S} \hat{\xi}_{t,i,j}^{(d)} \, \log \Gamma_{i,j} - \sum_{i=1}^{S} \lambda_i \left( \sum_{j=1}^{S} \Gamma_{i,j} - 1 \right) \,, \qquad (4.40)
\end{aligned}$$

from equation (4.23). It follows that

$$\begin{aligned}
\frac{\partial F_\Gamma(\theta, \hat{\theta})}{\partial \Gamma_{i,j}} &= \sum_{d=1}^{D} \sum_{t=1}^{n^{(d)}-1} \frac{\hat{\xi}_{t,i,j}^{(d)}}{\Gamma_{i,j}} - \lambda_i \; = \; 0 \text{ when } \theta = \hat{\theta}^* \\
\Rightarrow \hat{\Gamma}_{i,j}^* &= \frac{\sum_{d=1}^{D} \sum_{t=1}^{n^{(d)}-1} \hat{\xi}_{t,i,j}^{(d)}}{\sum_{d=1}^{D} \sum_{t=1}^{n^{(d)}-1} \hat{\gamma}_{t,i}^{(d)}} \qquad (4.41)
\end{aligned}$$

from equation (4.25).