Notes on Sequence Modelling

G.A. Jarrad

July 25, 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.

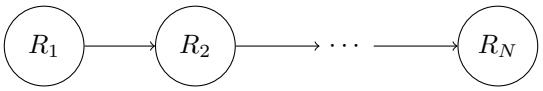

Figure 1.1: *A random process R for generating sequences of arbitrary length N. The arrows indicate transitions from one stage in the sequence to the next.*

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 ι_{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 τ_{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 ι_0 and τ_{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 $\langle\vec{r}_n\rangle$ is now defined as

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

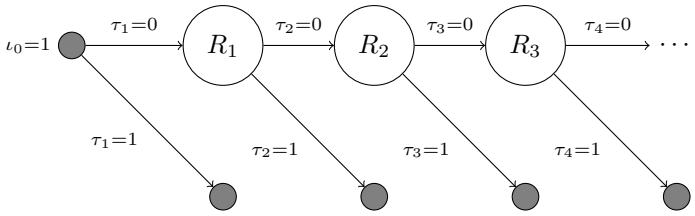

Figure 1.2: *A random process for generating complete sequences of arbitrary length, with explicit stages for sequence initiation and termination. Multiple arrows exiting from a node indicate different possible (mutually exclusive) stage transition pathways.*

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 $\langle\vec{r}_n]$ as

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

where $\tau_{n+1} = *$ is just a shorthand to indicate that we are uncertain of the true value of τ_{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

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

We can generalise this model by defining the forward observation sub-sequence $\vec{r}_t = (r_1, r_2, \ldots, r_t)$, for $t = 1, 2, \ldots, n$. Hence, we obtain

$$\begin{aligned} p(\underline{\iota}, \vec{r}, \underline{\tau}) \quad = \quad & p(\iota_0{=}\underline{\iota})\, p(\tau_1{=}0 \,|\, \iota_0{=}\underline{\iota})\, p(R_1{=}r_1 \,|\, \iota_0{=}\underline{\iota}, \tau_1{=}0) \\ & \times \prod_{t=2}^{n} \Big\{ p(\tau_t{=}0 \,|\, \iota_0{=}\underline{\iota}, \vec{\tau}_{t-1}{=}\vec{0}, \vec{R}_{t-1}{=}\vec{r}_{t-1})\, p(R_t{=}r_t \,|\, \iota_0{=}\underline{\iota}, \vec{\tau}_t{=}\vec{0}, \vec{R}_{t-1}{=}\vec{r}_{t-1}) \Big\} \\ & \times p(\tau_{n+1}{=}\bar{\tau} \,|\, \iota_0{=}\underline{\iota}, \vec{\tau}_n{=}\vec{0}, \vec{R}_n{=}\vec{r}_n)\,. \end{aligned} \tag{2.2}$$

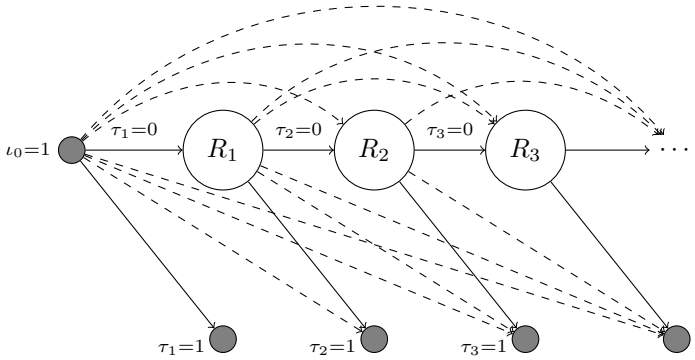

Figure 2.1: *A fully-dependent, causal process for generating complete sequences of arbitrary length. Solid arrows indicate possible stage transitions. Both dashed arrows and solid arrows indicate parent–child dependencies, such that the child node is conditionally dependent on the parent and all other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

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.

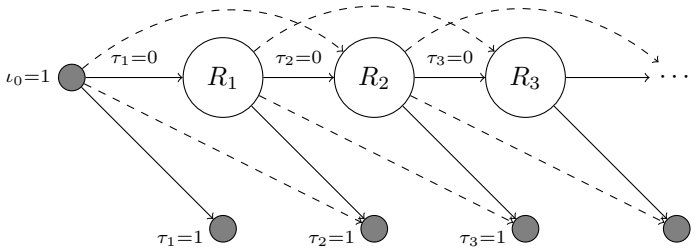

Figure 2.2: *A second-order Markov process for generating complete sequences of arbitrary length.*

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

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

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}).

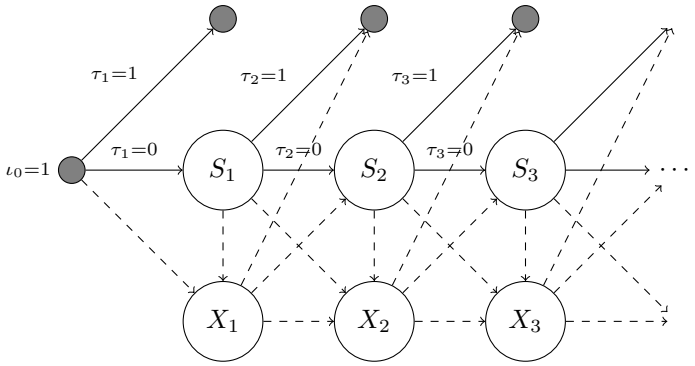

Figure 3.1: *A random process for generating complete, stateful sequences of arbitrary length, with explicit cross-dependencies between adjacent stages.*

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

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

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.

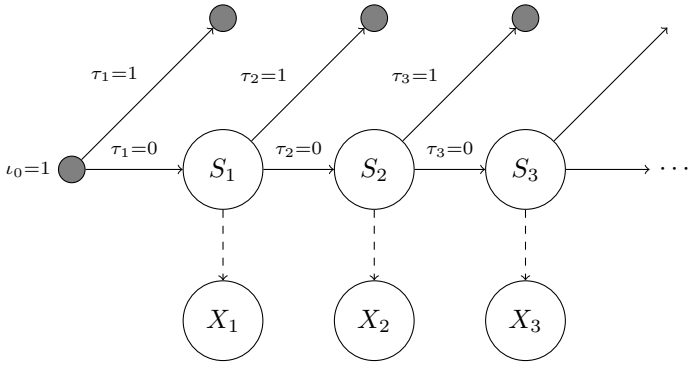

Figure 3.2: *A first-order Markov process for generating complete, stateful sequences of arbitrary length.*

The corresponding sequence model is now given by

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

Observe that the τ_1 dependency can additionally be dropped in practice, since zero-length sequences convey little information, and are in fact undetectable if the generating process does not explicitly indicate the start and end of sequences.

4 Discrete-state Sequence Models

Consider the stateful, first-order Markov process depicted by Figure 3.2, and suppose for convenience that zero-length sequences are impossible. 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), with the τ_1 dependency dropped, 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 $\theta = (\vec{\pi}, \Gamma, \vec{\omega})$, such that the probability of an arbitrary, observed sequence (with no hidden states) is given by

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

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

$$\underline{\pi}_{1,i} \;=\; p(\iota_0 = \underline{\iota} \,|\, \theta)\, p(S_1 = \sigma_i \,|\, \iota_0 = \underline{\iota}, \theta) \,. \tag{4.2}$$

More generally, we define

$$\pi^{-}_{t,i} \;=\; p(\iota_{t-1} = 0 \,|\, \theta)\, p(S_t = \sigma_i \,|\, \iota_{t-1} = 0, \theta) \,, \tag{4.3}$$

$$\pi^{+}_{t,i} \;=\; p(\iota_{t-1} = 1 \,|\, \theta)\, p(S_t = \sigma_i \,|\, \iota_{t-1} = 1, \theta) \,, \tag{4.4}$$

and

$$\pi*_{t,i} \;=\; p(\iota_{t-1} = * \,|\, \theta)\, p(S_t = \sigma_i \,|\, \iota_{t-1} = *, \theta) \;=\; p(S_t = \sigma_i \,|\, \theta) \;=\; \pi^{-}_{t,i} + \pi^{+}_{t,i} \,. \tag{4.5}$$

Observe that each state S_t for $t > 1$ is a non-initial state, governed by π^{-}_{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 **??**).

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

$$\underline{\omega}_{n,i} \;=\; p(\tau_{n+1} = \underline{\tau} \,|\, S_n = \sigma_i, \theta) \,, \tag{4.6}$$

where

$$\omega^{-}_{t,i} \;=\; p(\tau_{t+1} = 0 \,|\, S_n = \sigma_i, \theta) \,, \tag{4.7}$$

$$\omega^{+}_{t,i} \;=\; p(\tau_{t+1} = 1 \,|\, S_n = \sigma_i, \theta) \,, \tag{4.8}$$

and

$$\omega^{*}_{t,i} \;=\; p(\tau_{t+1} = * \,|\, S_n = \sigma_i, \theta) \;=\; \omega^{-}_{t,i} + \omega^{+}_{t,i} \;=\; 1 \,. \tag{4.9}$$

Observe that each state S_t for $t < n$ is a non-terminal state, and is explicitly modelled by the term ω^{-}_{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

$$\Gamma_{t,i,j} \;=\; p(S_{t+1} = \sigma_j \,|\, S_t = \sigma_i, \tau_{t+1} = 0, \theta) \,. \tag{4.10}$$

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

$$o_{t,i}(x) \;=\; p(X_t = x \,|\, S_t = \sigma_i, \theta) \qquad \forall x \in \mathcal{X} \,. \tag{4.11}$$

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

$$\begin{aligned} p(\underline{\iota}, \vec{s}, \vec{x}, \underline{\tau} \,|\, \theta) \;\; = \;\; & \sum_{i'_1=1}^{S} \delta(i'_1 = i_1) \sum_{i'_2=1}^{S} \delta(i'_2 = i_2) \cdots \sum_{i'_n=1}^{S} \delta(i'_n = i_n) \\ & \underline{\pi}_{1,i'_1} \, o_{1,i'_1}(x_1) \left\{ \prod_{t=1}^{n-1} \omega^{-}_{t,i'_t} \, \Gamma_{t,i'_t,i'_{t+1}} \, o_{t,i'_{t+1}}(x_{t+1}) \right\} \underline{\omega}_{n,i'_n} \,, \end{aligned} \tag{4.12}$$

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.

4.1 Posterior State Prediction

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

$$\begin{aligned} \gamma_{t,i} \;\; &= \;\; p(S_t = \sigma_i \,|\, \langle \vec{x} \rangle, \theta) \\ &= \;\; \frac{p(S_t = \sigma_i, \langle \vec{x} \rangle \,|\, \theta)}{p(\langle \vec{x} \rangle \,|\, \theta)} \\ &= \;\; \frac{p(S_t = \sigma_i, \langle \vec{x}_t] \,|\, \theta) \, p([\overleftarrow{x}_{t+1} \rangle \,|\, S_t = \sigma_i, \theta)}{\sum_{i'=1}^{S} p(S_t = \sigma_{i'}, \langle \vec{x}_t] \,|\, \theta) \, p([\overleftarrow{x}_{t+1} \rangle \,|\, S_t = \sigma_{i'}, \theta)} \\ &= \;\; \frac{\alpha_{t,i} \, \beta_{t,i}}{\sum_{i'=1}^{S} \alpha_{t,i} \, \beta_{t,i}} \,, \end{aligned} \tag{4.13}$$

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

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

$$\begin{aligned} \alpha_{t,i} \;\; &= \;\; p(S_t = \sigma_i, \langle \vec{x}_t] \,|\, \theta) \\ &= \;\; \sum_{j=1}^{S} p(S_{t-1} = \sigma_j, \langle \vec{x}_{t-1}] \,|\, \theta) \, p(\tau_t = 0 \,|\, S_{t-1} = \sigma_j, \theta) \\ &\qquad \times \, p(S_t = \sigma_i \,|\, S_{t-1} = \sigma_j, \theta) \, p(X_t = x_t \,|\, S_t = \sigma_i, \theta) \\ &= \;\; \left\{ \sum_{j=1}^{S} \alpha_{t-1,j} \, \bar{\omega}_{t-1,j} \, \Gamma_{t-1,j,i} \right\} o_{t,i}(x_t) \,, \end{aligned} \tag{4.14}$$

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

$$\begin{aligned} \alpha_{1,i} \;\; &= \;\; p(S_1 = \sigma_i, \langle x_1] \,|\, \theta) \\ &= \;\; p(\iota_0 = 1) \, p(S_1 = \sigma_i \,|\, \iota_0 = 1, \theta) \, p(X_1 = x_1 \,|\, S_1 = \sigma_i, \theta) \\ &= \;\; \pi_{1,i} \, o_{1,i}(x_1) \,. \end{aligned} \tag{4.15}$$

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

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

$$\begin{aligned} p(S_t{=}\sigma_i \,|\, \langle \vec{x}_t], \theta) &= \frac{p(S_t{=}\sigma_i, \langle \vec{x}_t] \,|\, \theta)}{\sum_{i'=1}^{S} p(S_t{=}\sigma_{i'}, \langle \vec{x}_t] \,|\, \theta)} \\ &= \frac{\alpha_{t,i}}{\sum_{i'=1}^{S} \alpha_{t,i'}} \,. \end{aligned} \tag{4.16}$$

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

$$\begin{aligned} p(X_{t+1}{=}x \,|\, \langle \vec{x}_t], \theta) &= \frac{p(\langle \vec{x}_t, x] \,|\, \theta)}{p(\langle \vec{x}_t] \,|\, \theta)} \\ &= \frac{\sum_{j=1}^{S} p(S_{t+1}{=}\sigma_j, \langle \vec{x}_t, x] \,|\, \theta)}{\sum_{i=1}^{S} p(S_t{=}\sigma_i, \langle \vec{x}_t] \,|\, \theta)} \\ &= \frac{\sum_{j=1}^{S} \left\{ \sum_{i=1}^{S} \alpha_{t,i}\, \bar{\omega}_{t,i}\, \Gamma_{t,i,j} \right\} o_{t+1,j}(x)}{\sum_{i=1}^{S} \alpha_{t,i}} \,. \end{aligned} \tag{4.17}$$

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

$$\begin{aligned} \beta_{t,i} &= p([\overleftarrow{x}_{t+1}\rangle \,|\, S_t{=}\sigma_i, \theta) \\ &= \sum_{j=1}^{S} p([\overleftarrow{x}_{t+2}\rangle \,|\, S_{t+1}{=}\sigma_j, \theta)\, p(X_{t+1} = x_{t+1} \,|\, S_{t+1}{=}\sigma_j, \theta) \\ &\qquad \times p(S_{t+1}{=}\sigma_j \,|\, S_t{=}\sigma_i, \theta)\, p(\tau_{t+1}{=}0 \,|\, S_t{=}\sigma_i, \theta) \\ &= \left\{ \sum_{j=1}^{S} \beta_{t+1,j}\, o_{t+1,j}(x_{t+1})\, \Gamma_{t,i,j} \right\} \bar{\omega}_{t,i} \,, \end{aligned} \tag{4.18}$$

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

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

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

The combination of the forward step with the backward step now enables us to use all of the information contained in the observed sequence, including its possible initiation and/or termination. Particularly, we can compute the joint probability of any observed sequence, as a prelude to estimating the state of each stage from equation (4.13). For example, observe that

$$\begin{aligned} p(\langle \vec{x}! \,|\, \theta) &= \sum_{i=1}^{S} p(S_t{=}\sigma_i, \langle \vec{x}! \,|\, \theta) \\ &= \sum_{i=1}^{S} p(S_t{=}\sigma_i, \langle \vec{x}_t] \,|\, \theta)\, p([\overleftarrow{x}_{t+1}! \,|\, S_t{=}\sigma_i, \theta) \\ &= \sum_{i=1}^{S} \alpha_{t,i} \bar{\beta}_{t,i} \,, \end{aligned} \tag{4.20}$$

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

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

$$\begin{aligned} \xi_{t,i,j} &= p(S_t{=}\sigma_i, S_{t+1}{=}\sigma_j \,|\, \langle \vec{x} \rangle, \theta) \\ &= \frac{p(S_t{=}\sigma_i, S_{t+1}{=}\sigma_j, \langle \vec{x} \rangle \,|\, \theta)}{p(\langle \vec{x} \rangle \,|\, \theta)} \,, \end{aligned} \tag{4.21}$$

where

$$\begin{aligned} p(S_t{=}\sigma_i, S_{t+1}{=}\sigma_j, \langle \vec{x} \rangle \,|\, \theta) &= p(S_t{=}!\sigma_i, \langle \vec{x}_t] \,|\, \theta)\, p(\tau_{t+1}{=}0 \,|\, S_t{=}\sigma_i, \theta) \\ &\qquad \times\; p(S_{t+1}{=}\sigma_j \,|\, S_t{=}\sigma_i, \theta)\, p([\overleftarrow{x}_{t+1}] \,|\, S_{t+1}{=}\sigma_j, \theta) \\ &= \alpha_{t,i}\, \bar{\omega}_{t,i}\, \Gamma_{t,i,j}\, \beta_{t+1,j} \,, \end{aligned} \tag{4.22}$$

and thus

$$\begin{aligned} p(\langle \vec{x} \rangle \,|\, \theta) &= \sum_{i'=1}^{S} \sum_{j'=1}^{S} p(S_t = \sigma_i, S_{t+1} = \sigma_j, \langle \vec{x} \rangle \,|\, \theta) \\ &= \sum_{i'=1}^{S} \sum_{j'=1}^{S} \alpha_{t,i'} \, \bar{\omega}_{t,i'} \, \Gamma_{t,i',j'} \, \beta_{t+1,j'} \\ &= \sum_{i'=1}^{S} \alpha_{t,i'} \, \beta_{t,i'} \,, \end{aligned} \tag{4.23}$$

as expected.

4.2 Modelling Observed Data

Suppose that we have observed an ordered set of sequences $\mathbb{X} = \{\vec{v}^{(d)}\}_{d=1}^{D}$, where each observation takes the form of $\vec{v}^{(d)} = (\iota^{(d)}, \vec{s}^{(d)}, \vec{x}^{(d)}, \tau^{(d)})$. Here $\iota^{(d)}$ takes one of the values 0, 1, or $*$, corresponding to the sequence initiation marker $\langle$, ! or [, respectively. Note that $\iota^{(d)} = *$ is taken to mean that the true value of ι_0 is unknown. Similarly, $\tau^{(d)}$ takes one of the values 0, 1, or $*$, corresponding to the sequence termination marker $\rangle$, ! or], respectively, where $\tau^{(d)} = *$ means that the true value of $\tau_{|\vec{x}^{(d)}|+1}$ is unknown. The observed values $\vec{x}^{(d)}$ are assumed to be known and to form a non-empty, contiguous sequence (or sub-sequence). Hence, we may always assume here that $\vec{\tau}_{|\vec{x}^{(d)}|} = \vec{0}$. Finally, the values of the state sequence $\vec{s}^{(d)}$ might or might not be specified. In some circumstances, $\vec{s}^{(d)}$ is known; in others, some or all of these states might remain hidden.

In order to model the ambiguities that might be present in $\vec{v}^{(d)}$, we need some notation. For convenience, we define the data-specified initial state distribution of the d-th observation as

$$\pi_{1,i}^{(d)} = \pi_{1,i} \delta(\iota_0^{(d)} = 1) + \bar{\pi}_{1,i} \delta(\iota_0^{(d)} = 0) \,, \tag{4.24}$$

where $\delta(\cdot) = 1$ (or 0) if its argument is true (or false). Note that $\iota^{(d)} = *$ is taken to match both $\iota_0 = 1$ and $\iota_0 = 0$. Similarly, we define the data-specified terminal state distribution of the d-th observation as

$$\omega_{1,i}^{(d)} = \omega_{1,i} \delta(\iota_0^{(d)} = 1) + \bar{\omega}_{1,i} \delta(\iota_0^{(d)} = 0) \,. \tag{4.25}$$

Finally, if the state S_t of stage t is known, then we define $s_t^{(d)} = \sigma_{i_t^{(d)}}$, for $i_t^{(d)} \in \{1, 2, \ldots, S\}$. Otherwise, we define both $s_t^{(d)} = *$ and $i_t^{(d)} = *$. Hence, from model (4.1), we obtain

$$p(\vec{v}^{(d)} \,|\, \theta) = \sum_{i_1=1}^{S} \delta(i_1^{(d)} = i_1) \cdots \sum_{i_n^{(d)}=1}^{S} \delta(i_{n^{(d)}}^{(d)} = i_{n^{(d)}}) \tag{4.26}$$

where we have marginalised over any missing data.

********** Notionally, we may also define the correspondingly ordered set $\mathbb{S} = \{\vec{s}^{(d)}\}_{d=1}^{D}$ of arbitrary state sequences. Hence, under the assumption that the observed sequences are independent, the joint log-likelihood of the data is given by

$$\begin{aligned} L(\theta) &= \log p(\mathbb{S}, \mathbb{X} \,|\, \theta) \\ &= \log \prod_{d=1}^{D} p(\iota_0^{(d)}, \vec{s}^{(d)}, \vec{x}^{(d)}, \tau_{n^{(d)}+1}^{(d)} \,|\, \theta) \\ &= \sum_{d=1}^{D} \log p(\iota_0^{(d)}, \vec{s}^{(d)}, \vec{x}^{(d)}, \tau_{n^{(d)}+1}^{(d)} \,|\, \theta) \\ &= \sum_{d=1}^{D} L^{(d)}(\theta) \,, \end{aligned} \tag{4.27}$$

where

$$L^{(d)}(\theta) = \log \pi_{i_1^{(d)}}^{(d)} + \sum_{t=1}^{n^{(d)}-1} \log \Gamma_{i_t^{(d)}, i_{t+1}^{(d)}} + \sum_{t=1}^{n^{(d)}} \log o_{t, i_t^{(d)}} + \log \omega_{i_{n^{(d)}}^{(d)}}^{(d)} \,, \tag{4.28}$$

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

However, recall that $\mathbb{S}$ is actually uknown. Hence, we take an expectation of the log-likelihood over all possible values of $\mathbb{S}$, namely[1]

$$\begin{aligned} Q(\theta) &= E_{\mathbb{S}\,|\,\mathbb{X},\theta}\left[\log p(\mathbb{S},\mathbb{X}\,|\,\theta)\right] \\ &= E_{\mathbb{S}\,|\,\mathbb{X}\theta}\left[\sum_{d=1}^{D} L^{(d)}(\theta)\right] \\ &= \sum_{d=1}^{D} E_{\mathbb{S}\,|\,\mathbb{X},\theta}\left[L^{(d)}(\theta)\right] \\ &= \sum_{d=1}^{D}\sum_{i_1^{(d)}=1}^{S}\cdots\sum_{i_{n^{(d)}}^{(d)}}^{S} p(\vec{s}^{(d)}\,|\,\iota_0^{(d)},\vec{x}^{(d)},\tau_{n^{(d)}+1}^{(d)},\theta)\,L^{(d)}(\theta)\,. \end{aligned} \tag{4.29}$$

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

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

$$\begin{aligned} Q(\theta,\hat{\theta}_k) &= E_{\mathbb{S}\,|\,\mathbb{X},\hat{\theta}_k}\left[\log p(\mathbb{S},\mathbb{X}\,|\,\theta)\right] \\ &= \sum_{d=1}^{D}\sum_{i_1^{(d)}=1}^{S}\cdots\sum_{i_{n^{(d)}}^{(d)}}^{S} p(\vec{s}^{(d)}\,|\,\iota_0^{(d)},\vec{x}^{(d)},\tau_{n^{(d)}+1}^{(d)},\hat{\theta}_k)\,L^{(d)}(\theta)\,. \end{aligned} \tag{4.30}$$

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

$$\hat{\theta}_{k+1} = \arg\max_{\theta} Q(\theta,\hat{\theta}_k)\,. \tag{4.31}$$

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

blah about additivity

$$\begin{aligned} \frac{\partial Q}{\partial \Gamma_{i,j}} &= \frac{\partial}{\partial \Gamma_{i,j}}\sum_{d=1}^{D}\sum_{i_1^{(d)}=1}^{S}\cdots\sum_{i_{n^{(d)}}^{(d)}}^{S} p(\vec{s}^{(d)}\,|\,\iota_0^{(d)},\vec{x}^{(d)},\tau_{n^{(d)}+1}^{(d)},\hat{\theta}')\sum_{t=1}^{n^{(d)}-1}\log\Gamma_{i_t^{(d)},i_{t+1}^{(d)}} \\ &= \sum_{d=1}^{D}\sum_{t=1}^{n^{(d)}-1}\sum_{i_1^{(d)}=1}^{S}\cdots\sum_{i_{n^{(d)}}^{(d)}}^{S}\delta(i_t^{(d)}=i)\delta(i_{t+1}^{(d)}=j)\frac{p(\vec{s}^{(d)}\,|\,\iota_0^{(d)},\vec{x}^{(d)},\tau_{n^{(d)}+1}^{(d)},\hat{\theta}')}{\Gamma_{i,j}} \\ &= \sum_{d=1}^{D}\sum_{t=1}^{n^{(d)}-1}\sum_{i_1^{(d)}=1}^{S}\cdots\sum_{i_{n^{(d)}}^{(d)}}^{S}\delta(i_t^{(d)}=i)\delta(i_{t+1}^{(d)}=j)\frac{p(\vec{s}^{(d)}\,|\,\iota_0^{(d)},\vec{x}^{(d)},\tau_{n^{(d)}+1}^{(d)},\hat{\theta}')}{\Gamma_{i,j}} \\ &= \frac{\sum_{d=1}^{D}\sum_{t=1}^{n^{(d)}-1} p(S_t=\sigma_i,S_{t+1}=\sigma_j\,|\,\iota_0^{(d)},\vec{x}^{(d)},\tau_{n^{(d)}+1}^{(d)},\hat{\theta}')}{\Gamma_{i,j}} \\ &= \frac{\sum_{d=1}^{D}\sum_{t=1}^{n^{(d)}-1}\xi_t^{(d)}(\sigma_i,\sigma_j;\hat{\theta}')}{\Gamma_{i,j}} \end{aligned} \tag{4.32}$$

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

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

estimate is given by

$$\hat{\Gamma}^*_{i,j} = \frac{\sum_{d=1}^{D}\sum_{t=1}^{n^{(d)}-1}\xi_t^{(d)}(\sigma_i,\sigma_j;\hat{\theta}')}{\sum_{j=1}^{S}\sum_{d=1}^{D}\sum_{t=1}^{n^{(d)}-1}\xi_t^{(d)}(\sigma_i,\sigma_j;\hat{\theta}')} = \frac{\sum_{d=1}^{D}\sum_{t=1}^{n^{(d)}-1}\xi_t^{(d)}(\sigma_i,\sigma_j;\hat{\theta}')}{\sum_{d=1}^{D}\sum_{t=1}^{n^{(d)}-1}\gamma_t^{(d)}(\sigma_i;\hat{\theta}')} \tag{4.33}$$

from equation (**??**).