Notes on Sequence Modelling

G.A. Jarrad

July 17, 2015

1 Random Sequence Processes

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

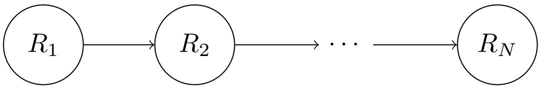

Figure 1.1: *A random process R for generating sequences of arbitrary length N. The arrows indicate transitions from one stage in the sequence to the next.*

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

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

where

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

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

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

Under this augmented notation, knowledge about the start of a sequence can be encapsulated in a random indicator variable ι_t, which takes on the value

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

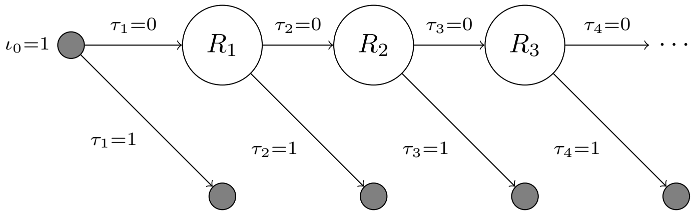

Figure 1.2: *A random process for generating complete sequences of arbitrary length, with explicit stages for sequence initiation and termination. Multiple arrows exiting from a node indicate different possible (mutually exclusive) stage transition pathways.*

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

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

such that

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

Likewise, the probability of a given start sequence $\langle\vec{r}_n]$ is defined as

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

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

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

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

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

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

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

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

2 Markov Sequence Processes

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

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

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

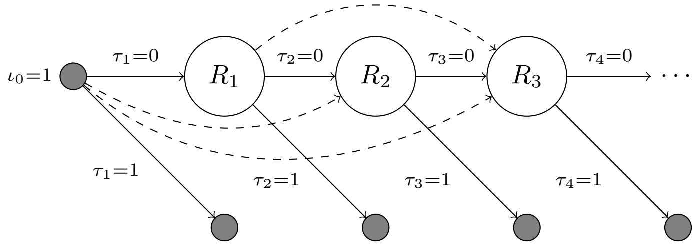

Figure 2.1: *A fully-dependent, causal process for generating complete sequences of arbitrary length. Solid arrows indicate possible stage transitions. Both dashed arrows and solid arrows indicate parent–child dependencies, such that the child node is conditionally dependent on the parent and all other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

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

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

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

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

3 Stateful Markov Sequence Processes

Consider the first-order Markov process R depicted in Figure 1.2. Suppose now that the random variable R_t at stage t can be decomposed into the tuple $R_t = (S_t, X_t)$, where S_t is a discrete random variable taking values $s_t \in \mathcal{S}$, and X_t is a discrete or continuous 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*. Hence, as is usual, we may suppose that X_t depends upon S_t, such that the stage transitions are between states. Keeping to the first-order Markov interpretation of stage-to-stage dependencies leads to the *stateful* process depicted in Figure 3.1, with full cross-dependencies between (S_t, X_t) and (S_{t+1}, X_{t+1}).

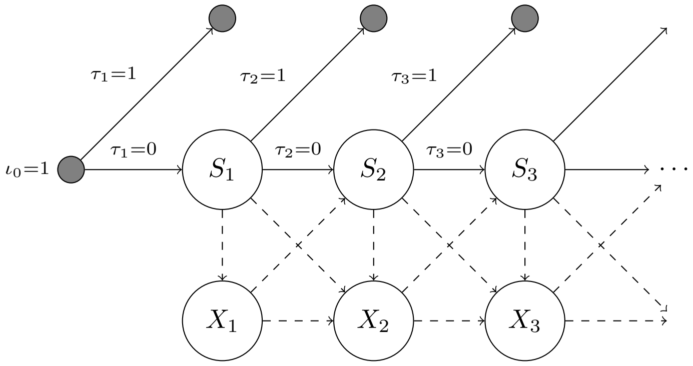

Figure 3.1: *A random process for generating complete, stateful sequences of arbitrary length, with explicit cross-dependencies between the states and values of adjacent stages.*

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

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

It is more usual, however, to further restrict the complexity of the stateful process by also imposing the first-order Markov assumption at the level of the state–value dependencies themselves. In terms of the process depicted in Figure 3.1, this means retaining only direct node-to-node dependencies (rather than stage-to-stage dependencies). Hence, this restricted process, depicted in

Figure 3.2, corresponds to the sequence model

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

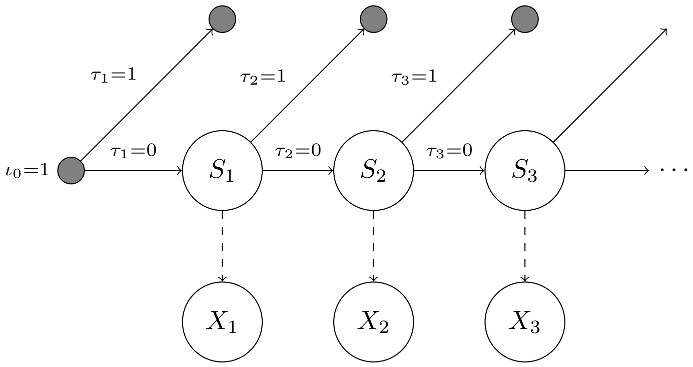

Figure 3.2: *A first-order Markov process for generating complete, stateful sequences of arbitrary length.*

4 Hidden-state Markov Sequence Processes

Consider the stateful, first-order Markov process depicted by Figure **??**. Suppose now that the value of state S_t at any stage t is never observed, only the value of X_t. Then the model (**??**) may be considered to be a *hidden-state* Markov model (or HMM). As such, the state of S_t must be deduced from knowledge of the observed sequence $\vec{x}_n$. This is accomplished via the forward–backward algortithm. The forward step commences with stages 0 and 1, by defining

$$\begin{aligned} \alpha_1(s_1) &= p(\iota_0, X_1 = x_1, S_1 = s_1) \\ &= p(\iota_0)\, p(S_1 = s_1 \,|\, \iota_0)\, p(X_1 = x_1 \,|\, S_1 = s_1) \\ &= p(\iota_0)\, p(s_1 \,|\, \iota_0)\, p(x_1 \,|\, s_1)\,, \end{aligned} \tag{4.1}$$

from equation (**??**), where the explicit variables S_t and X_t may be dropped for convenience when the context is unambiguous. Then it follows that

$$\begin{aligned} \alpha_2(s_2) &= p(\iota_0, X_1 = x_1, X_2 = x_2, S_2 = s_2) \\ &= \sum_{s_1 \in \mathcal{S}} p(\iota_0, x_1, s_1)\, p(s_2 \,|\, s_1)\, p(x_2 \,|\, s_2) \\ &= \sum_{s_1 \in \mathcal{S}} \alpha_1(s_1)\, p(s_2 \,|\, s_1)\, p(x_2 \,|\, s_2)\,, \end{aligned} \tag{4.2}$$

and in general that

$$\begin{aligned}\alpha_t(s_t) &= p(\iota_0, \vec{X}_t = \vec{x}_t, S_t = s_t) \\ &= \left\{ \sum_{s_{t-1} \in \mathcal{S}} \alpha_{t-1}(s_{t-1})\, p(s_t \,|\, s_{t-1}) \right\} p(x_t \,|\, s_t)\,, \end{aligned} \tag{4.3}$$

for $t = 2, 3, \ldots, n$. Consequently, we may predict S_t from a partially observed sequence $\vec{x}_t$ via

$$p(S_t = s_t \,|\, \iota_0, \vec{X}_t = \vec{x}_t) = \frac{p(\iota_0, \vec{X}_t = \vec{x}_t, S_t = s_t)}{p(\iota_0, \vec{X}_t = \vec{x}_t)} = \frac{\alpha_t(s_t)}{\sum_{s'_t \in \mathcal{S}} \alpha_t(s'_t)}\,. \tag{4.4}$$

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

$$\begin{aligned} p(X_{t+1} = x_{t+1} \,|\, \iota_0, \vec{X}_t = \vec{x}_t) &= \frac{p(\iota_0, \vec{X}_t = \vec{x}_t, X_{t+1} = x_{t+1})}{p(\iota_0, \vec{X}_t = \vec{x}_t)} \\ &= \frac{\sum_{s_{t+1} \in \mathcal{S}} \sum_{s_t \in \mathcal{S}} p(\iota_0, \vec{x}_t, s_t)\, p(s_{t+1} \,|\, s_t)\, p(x_{t+1} \,|\, s_{t+1})}{p(\iota_0, \vec{x}_t)} \\ &= \frac{\sum_{s_{t+1} \in \mathcal{S}} \sum_{s_t \in \mathcal{S}} \alpha_t(s_t)\, p(s_{t+1} \,|\, s_t)\, p(x_{t+1} \,|\, s_{t+1})}{\sum_{s_t \in \mathcal{S}} \alpha_t(s_t)} \end{aligned} \tag{4.5}$$

The backward step now commences with stage $n+1$, by defining

$$\beta_n(s_n) = p(\tau_{n+1} \,|\, S_n = s_n)\,, \tag{4.6}$$

and

$$\begin{aligned} \beta_{n-1}(s_{n-1}) &= p(X_n = x_n, \tau_{n+1} \,|\, S_{n-1} = s_{n-1}) \\ &= \sum_{s_n \in \mathcal{S}} p(\tau_{n+1} \,|\, s_n)\, p(x_n \,|\, s_n)\, p(s_n \,|\, s_{n-1}) \\ &= \sum_{s_n \in \mathcal{S}} \beta_n(s_n)\, p(x_n \,|\, s_n)\, p(s_n \,|\, s_{n-1})\,. \end{aligned} \tag{4.7}$$

In general, we let $\overleftarrow{x}_t = (x_t, x_{t+1}, \ldots, x_n)$, and then recursively define

$$\begin{aligned} \beta_t(s_t) &= p(\overleftarrow{X}_{t+1} = \overleftarrow{x}_{t+1}, \tau_{n+1} \,|\, S_t = s_t) \\ &= \sum_{s_{t+1} \in \mathcal{S}} p(\overleftarrow{x}_{t+2}, \tau_{n+1} \,|\, s_{t+1})\, p(x_{t+1} \,|\, s_{t+1})\, p(s_{t+1} \,|\, s_t) \\ &= \sum_{s_{t+1} \in \mathcal{S}} \beta_{t+1}(s_{t+1})\, p(x_{t+1} \,|\, s_{t+1})\, p(s_{t+1} \,|\, s_t)\,. \end{aligned} \tag{4.8}$$

Consequently, for an observed sequence $\vec{x}_n$, the forward–backward algorithm gives the stage t probability

$$\begin{aligned} p(\iota_0, \vec{X}_n = \vec{x}_n, S_t = s_t, \tau_{n+1}) &= p(\iota_0, \vec{X}_t = \vec{x}_t, S_t = s_t) \\ &\quad\; p(\overleftarrow{X}_{t+1} = \overleftarrow{x}_{t+1}, \tau_{n+1} \,|\, S_t = s_t) \\ &= \alpha_t(s_t)\beta_t(s_t)\,. \end{aligned} \tag{4.9}$$

Thus, the probability of $\vec{x}_n$ is

$$\begin{aligned} p(\iota_0, \vec{X}_n\!=\!\vec{x}_n, \tau_{n+1}) &= \sum_{s_t\in\mathcal{S}} p(\iota_0, \vec{X}_n\!=\!\vec{x}_n, S_t\!=\!s_t, \tau_{n+1}) \\ &= \sum_{s_t\in\mathcal{S}} \alpha_t(s_t)\beta_t(s_t)\,, \end{aligned} \tag{4.10}$$

and the posterior prediction of the state S_t given $\vec{x}_n$ is subsequently given by

$$\begin{aligned} \gamma_t(s_t) &= p(S_t\!=\!s_t \,|\, \iota_0, \vec{X}_n\!=\!\vec{x}_n, \tau_{n+1}) \\ &= \frac{p(\iota_0, \vec{X}_n\!=\!\vec{x}_n, S_t\!=\!s_t, \tau_{n+1})}{p(\iota_0, \vec{X}_n\!=\!\vec{x}_n, \tau_{n+1})} \\ &= \frac{\alpha_t(s_t)\beta_t(s_t)}{\sum_{s'_t\in\mathcal{S}} \alpha_t(s'_t)\beta_t(s'_t)}\,. \end{aligned} \tag{4.11}$$

Likewise, the posterior prediction of the state transition from stage t to stage $t+1$ is given by

$$\begin{aligned} \xi_t(s_t, s_{t+1}) &= p(S_t\!=\!s_t, S_{t+1}\!=\!s_{t+1} \,|\, \iota_0, \vec{X}_n\!=\!\vec{x}_n, \tau_{n+1}) \\ &= \frac{p(\iota_0, \vec{X}_n\!=\!\vec{x}_n, S_t\!=\!s_t, S_{t+1}\!=\!s_{t+1}, \tau_{n+1})}{p(\iota_0, \vec{X}_n\!=\!\vec{x}_n, \tau_{n+1})} \\ &= \frac{p(\iota_0, \vec{x}_t, s_t)\, p(s_{t+1} \,|\, s_t)\, p(x_{t+1} \,|\, s_{t+1})\, p(\overleftarrow{x}_{t+2}, \tau_{n+1} \,|\, s_{t+1})}{p(\iota_0, \vec{x}_n, \tau_{n+1})} \\ &= \frac{\alpha_t(s_t)\, p(s_{t+1} \,|\, s_t)\, p(x_{t+1} \,|\, s_{t+1})\, \beta_{t+1}(s_{t+1})}{\sum_{s'_t\in\mathcal{S}} \alpha_t(s'_t)\beta_t(s'_t)}\,. \end{aligned} \tag{4.12}$$

5 Hidden-state Parameter Estimation

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

where

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

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

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

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

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

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

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

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

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

blah about additivity

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

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

from equation (4.12). Now, subject to the constraint that $\sum_{j=1}^{S} \Gamma_{i,j} = 1$, we induce the appropriate Lagrangian multiplier to provide the proper normalisation, and hence derive that the optimal parameter estimate is given by

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

from equation (4.11).