Notes on Sequence Modelling

G.A. Jarrad

July 7, 2015

1 Introduction

blah, blah, blah

2 Random Sequence Processes

Consider a random process $\vec{R}=(R_1,R_2,R_3,\ldots)$ 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 determined by a random variable N, and the random variable R_t denotes the t-th discrete stage in the sequence. 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

$$p(\vec{R}=\vec{r}_n) \quad = \quad p(N=n)\,p(\vec{R}=\vec{r}_n\,|\,N=n)\,, \tag{2.1}$$

where

$$p(\vec{R}=\vec{r}_n\,|\,N=n) \quad = \quad p(\vec{R}_n=\vec{r}_n) \;=\; p(R_1=r_1,\ldots,R_n=r_n)\,. \tag{2.2}$$

In practice, this definition presupposes that we know we have 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_0, which takes on the value 1 if R_1 is definitely the first stage in the sequence, or the value 0 if it is not. Similarly, the random indicator variable τ_{n+1} takes on the value 1 if R_n is definitely the last stage in the sequence, or the value 0 if it is not. Notionally, these indicators can be thought to correspond to pseudo-stages 0 and $N+1$, such that the sequence is initiated at stage 0 and terminated at stage $N+1$. This augmented random process is depicted in Figure 2.1.

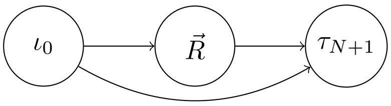

Figure 2.1: A random process for generating both complete and incomplete sequences of random length N, with explicit stages for sequence initiation and termination.

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

$$p(\langle\vec{r}_n\rangle) \;=\; p(\iota_0=1, R_1=r_1\ldots, R_n=r_n, \tau_{n+1}=1)\,, \tag{2.3}$$

such that $p(N=n)\equiv p(\iota_0=1,\tau_{n+1}=1)$. Likewise, the probability of a given start sequence $\langle\vec{r}_n]$ is defined as

$$p(\langle\vec{r}_n]) \;=\; p(\iota_0=1, R_1=r_1\ldots, R_n=r_n)\,, \tag{2.4}$$

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

$$p([\vec{r}_n\rangle) \;=\; p(R_1=r_1\ldots, R_n=r_n, \tau_{n+1}=1)\,. \tag{2.5}$$

In the special case where we know in advance that a start sequence definitely does not terminate at stage $n+1$ (i.e. $\tau_{n+1}=0$), we may instead write

$$p(\langle\vec{r}_n!) \;=\; p(\iota_0=1, R_1=r_1\ldots, R_n=r_n, \tau_{n+1}=0)\,. \tag{2.6}$$

Likewise, if an end sequence definitely does not initiate at stage 0 (i.e. $\iota_0=0$), then

$$p(!\vec{r}_n\rangle) \;=\; p(\iota_0=0, R_1=r_1\ldots, R_n=r_n, \tau_{n+1}=1)\,. \tag{2.7}$$

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

3 Markov Sequence Processes

In Section 2 we defined a random process $\vec{R}$ and the sequences it generates. We now assume that the process is also *causal*, meaning that each stage of a sequence, including the initiation stage and the termination stage, depends only on the preceding stages. Hence, under the Markov assumption of conditional independence, the process depicted in Figure 2.1 leads to the conditional model

$$p(\langle\vec{r}_n\rangle) \;=\; p(\iota_0=1)\,p(\vec{R}_n=\vec{r}_n\,|\,\iota_0=1)\,p(\tau_{n+1}=1\mid\iota_0=1,\vec{R}_n=\vec{r}_n)\,. \tag{3.1}$$

We can further decompose the model for $\vec{R}$, since the distribution of values for variable R_t, at stage t, depends directly upon the values generated previously in the sequence at stages $t-1, t-2, \ldots, 1$. This expanded causal process is depicted in Figure 3.1. Hence, the probability of a complete, causal sequence is

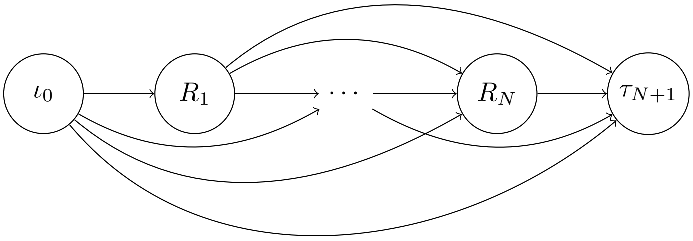

Figure 3.1: A fully-connected causal process for generating temporal sequences of random length N.

taken here to be

$$\begin{aligned} p(\langle\vec{r}_n\rangle) &= p(\iota_0 = 1) \prod_{t=1}^{n} p(R_t = r_t \,|\, \vec{R}_{t-1} = \vec{r}_{t-1}, \iota_0 = 1) \\ & \quad\; p(\tau_{n+1} = 1 \,|\, \vec{R}_n = \vec{r}_n, \iota_0 = 1)\,. \end{aligned} \tag{3.2}$$

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 2. In general, all types of sequences can be handled by a slight change of notation. Let V denote an arbitrary node variable, such that $V_0 = \iota_0$, $V_t = R_t$ for $t = 1, 2, \ldots, N$, and $V_{N+1} = \tau_{N+1}$, and consider $\vec{V} = (V_0, \ldots, V_{N+1})$. Likewise, let $\vec{v}$ denote an observed sequence of values, e.g. $\vec{v} = \langle\vec{r}_n\rangle$, or $\vec{v} = [\vec{r}_n\rangle$, et cetera. Then the causal process model (3.2) reduces to

$$p(\vec{v}) = \prod_{t=0}^{n+1} p(V_t = v_t \,|\, \vec{\Pi}_t(\vec{V}) = \vec{\pi}_t(\vec{v}))\,, \tag{3.3}$$

where $\vec{\Pi}_t(\vec{V}) = (V_0, V_1, \ldots, V_{t-1})$ denotes the predecessor nodes upon which node V_t is conditionally dependent, and $\vec{\pi}_t(\vec{v}) = (v_0, v_1, \ldots, v_{t-1})$ similarly denotes the observed values of those predecessor nodes.

In practice, the causal model is usually simplified further by limiting the conditional dependency on past values to a maximum number m of terms. Hence, this so-called *m-th order Markov model* is given by

$$p(\vec{v}) = \prod_{t=0}^{n+1} p(V_t = v_t \,|\, \vec{\Pi}_t^{(m)}(\vec{V}) = \vec{\pi}_t^{(m)}(\vec{v}))\,, \tag{3.4}$$

where the predecessor nodes are now given by

$$\vec{\Pi}_t^{(m)}(\vec{V}) = \left\{ \begin{array}{ll} (V_0, V_1, \ldots, V_{t-1}) & \text{if } t \leq m\,, \\ (V_{t-m}, V_{t-m+1}, \ldots, V_{t-1}) & \text{if } t > m \end{array} \right. . \tag{3.5}$$

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 even simpler form

$$p(\vec{v}) \quad = \quad \prod_{t=0}^{n+1} p(V_t = v_t \,|\, V_{t-1} = v_{t-1})\,, \tag{3.6}$$

or, closer to the original notation:

$$\begin{aligned} p(\iota_0, \vec{r}_n, \tau_{n+1}) \quad &= \quad p(\iota_0) p(R_1 = r_1 \,|\, \iota_0) \prod_{t=2}^{n} p(R_t = r_t \,|\, R_{t-1} = r_{t-1}) \\ & \qquad p(\tau_{n+1} \,|\, R_n = r_n)\,. \end{aligned} \tag{3.7}$$

This process is depicted in Figure 3.2, and will henceforth be taken as the basis of our analyses.

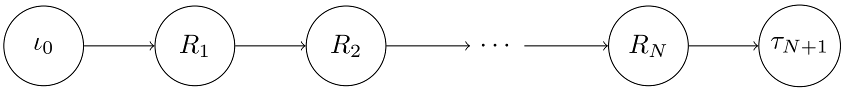

Figure 3.2: A first-order Markov process for generating causal sequences of random length N.

4 Stateful Markov Sequence Processes

Consider the first-order Markov process $\vec{R}$ depicted in Figure 3.2. 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*. The joint state–value model then takes the form

$$\begin{aligned} p(\iota_0, \vec{S}_n, \vec{X}_n, \tau_{n+1}) \quad &= \quad p(\iota_0)\, p(S_1, X_1 \,|\, \iota_0) \\ & \qquad \prod_{t=2}^{n} p(S_t, X_t \,|\, S_{t-1}, X_{t-1}) p(\tau_{n+1} \,|\, S_n, X_n)\,, \end{aligned} \tag{4.1}$$

corresponding to the *stateful* process is depicted in Figure 4.1.

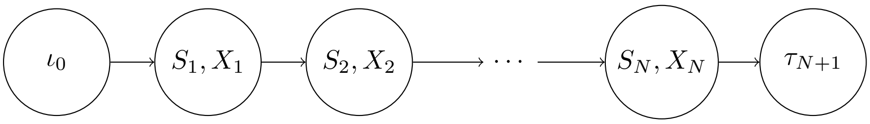

Figure 4.1: A first-order Markov process for generating stateful sequences of random length N.

We now impose futher structure on the process by specifying the relationship within stage t, and also expanding on the relationship between stage t and stage $t+1$. Firstly, it is commonly supposed that the process determines the state S_t based on available information, and then from S_t selects the value X_t. Next, in the general case both S_{t+1} and X_{t+1} may depend upon S_t and X_t. Hence, the structured stateful model is now given by

$$\begin{aligned} p(\iota_0, \vec{S}_n, \vec{X}_n, \tau_{n+1}) &= p(\iota_0)\, p(S_1 \mid \iota_0)\, p(X_1 \mid S_1, \iota_0) \\ &\quad \prod_{t=2}^{n} p(S_t \mid S_{t-1}, X_{t-1})\, p(X_t \mid S_t, S_{t-1}, X_{t-1}) \\ &\quad p(\tau_{n+1} \mid S_n, X_n)\,, \end{aligned} \tag{4.2}$$

corresponding to the process is depicted in Figure 4.2.

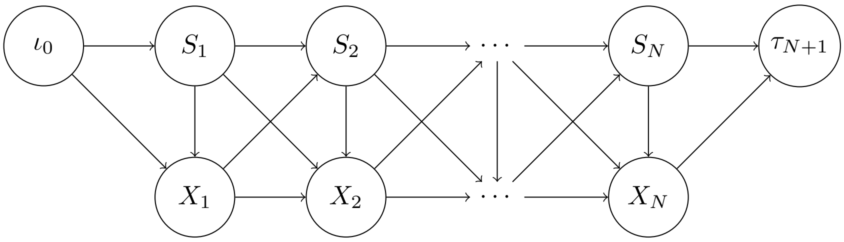

Figure 4.2: A general interpretation of the first-order stateful Markov process for generating sequences of random length N.

It is more usual, however, to further restrict the complexity of the process by also imposing the first-order Markov assumption at the level of the state–value transitions themselves, leading to the restricted stateful model

$$\begin{aligned} p(\iota_0, \vec{S}_n, \vec{X}_n, \tau_{n+1}) &= p(\iota_0)\, p(S_1 \mid \iota_0)\, p(X_1 \mid S_1, \iota_0) \\ &\quad \prod_{t=2}^{n} p(S_t \mid S_{t-1})\, p(X_t \mid S_t) \\ &\quad p(\tau_{n+1} \mid S_n)\,, \end{aligned} \tag{4.3}$$

with the corresponding process shown in Figure 4.3.

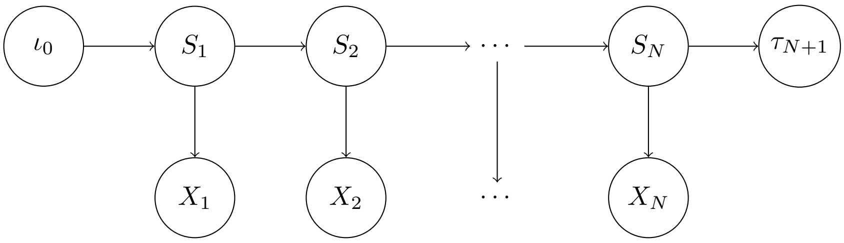

Figure 4.3: A restricted interpretation of the first-order stateful Markov process for generating sequences of random length N.

5 Hidden-state Markov Sequence Processes

Consider the stateful, first-order Markov process depicted by Figure 4.3. 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 (4.3) 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

$$\begin{aligned} \alpha_1(s_1) &= p(\iota_0, x_1, s_1) \\ &= p(\iota_0)\, p(S_1 = s_1 \,|\, \iota_0)\, p(X_1 = x_1 \,|\, S_1 = s_1) \\ &= p(\iota_0)\, p(s_1 \,|\, \iota_0)\, p(x_1 \,|\, s_1)\,, \end{aligned} \tag{5.1}$$

from equation (4.3), where the explicit variables S_t and X_t may now be dropped for convenience when the context is unambiguous. Then it follows that

$$\begin{aligned} \alpha_2(s_2) &= p(\iota_0, x_1, x_2, s_2) \\ &= \sum_{s_1 \in \mathcal{S}} p(\iota_0, x_1, s_1)\, p(s_2 \,|\, s_1)\, p(x_2 \,|\, s_2) \\ &= \sum_{s_1 \in \mathcal{S}} \alpha_1(s_1)\, p(s_2 \,|\, s_1)\, p(x_2 \,|\, s_2)\,, \end{aligned} \tag{5.2}$$

and in general that

$$\begin{aligned} \alpha_t(s_t) &= p(\iota_0, \vec{x}_t, s_t) \\ &= \left\{ \sum_{s_{t-1} \in \mathcal{S}} \alpha_{t-1}(s_{t-1})\, p(s_t \,|\, s_{t-1}) \right\} p(x_t \,|\, s_t)\,, \end{aligned} \tag{5.3}$$

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

$$p(s_t \,|\, \iota_0, \vec{x}_t) = \frac{p(\iota_0, \vec{x}_t, s_t)}{p(\iota_0, \vec{x}_t)} = \frac{\alpha_t(s_t)}{\sum_{s'_t \in \mathcal{S}} \alpha_t(s'_t)}\,. \tag{5.4}$$

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

$$\begin{aligned} p(x_{t+1} \,|\, \iota_0, \vec{x}_t) &= \frac{p(\iota_0, \vec{x}_t, x_{t+1})}{p(\iota_0, \vec{x}_t)} \\ &= \frac{\sum_{s_{t+1} \in \mathcal{S}} \sum_{s_t \in \mathcal{S}} p(\iota_0, \vec{x}_t, s_t)\, p(s_{t+1} \,|\, s_t)\, p(x_{t+1} \,|\, s_{t+1})}{p(\iota_0, \vec{x}_t)} \\ &= \frac{\sum_{s_{t+1} \in \mathcal{S}} \sum_{s_t \in \mathcal{S}} \alpha_t(s_t)\, p(s_{t+1} \,|\, s_t)\, p(x_{t+1} \,|\, s_{t+1})}{\sum_{s_t \in \mathcal{S}} \alpha_t(s_t)}\,. \end{aligned} \tag{5.5}$$

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

$$\beta_n(s_n) = p(\tau_{n+1} \,|\, s_n)\,, \tag{5.6}$$

and

$$\begin{aligned} \beta_{n-1}(s_{n-1}) &= p(x_n, \tau_{n+1} \,|\, s_{n-1}) \\ &= \sum_{s_n \in \mathcal{S}} p(\tau_{n+1} \,|\, s_n)\, p(x_n \,|\, s_n)\, p(s_n \,|\, s_{n-1}) \\ &= \sum_{s_n \in \mathcal{S}} \beta_n(s_n)\, p(x_n \,|\, s_n)\, p(s_n \,|\, s_{n-1})\,. \end{aligned} \tag{5.7}$$

In general, we recursively define

$$\begin{aligned}
\beta_t(s_t) &= p(x_{t+1}, \ldots, x_n, \tau_{n+1} \,|\, s_t) \\
&= \sum_{s_{t+1} \in \mathcal{S}} p(x_{t+2}, \ldots, x_n, \tau_{n+1} \,|\, s_{t+1})\, p(x_{t+1} \,|\, s_{t+1})\, p(s_{t+1} \,|\, s_t) \\
&= \sum_{s_{t+1} \in \mathcal{S}} \beta_{t+1}(s_{t+1})\, p(x_{t+1} \,|\, s_{t+1})\, p(s_{t+1} \,|\, s_t)\,. \qquad (5.8)
\end{aligned}$$

Consequently, we may now calculate the probability of an entirely observed sequence $\vec{x}_n$ as

$$\begin{aligned}
p(\iota_0, \vec{x}_n, \tau_{n+1}) &= \sum_{s_n \in \mathcal{S}} p(\iota_0, \vec{x}_n, s_n)\, p(\tau_{n+1} \,|\, s_n) \\
&= \sum_{s_n \in \mathcal{S}} \alpha_n(s_n) \beta_n(s_n)\,, \qquad (5.9)
\end{aligned}$$

and retrospectively predict S_t given $\vec{x}_n$ via

$$\begin{aligned}
p(s_t \,|\, \iota_0, \vec{x}_n, \tau_{n+1}) &= \frac{p(\iota_0, \vec{x}_n, s_t, \tau_{n+1})}{p(\iota_0, \vec{x}_n, \tau_{n+1})} \\
&= \frac{p(\iota_0, \vec{x}_t, s_t)\, p(x_{t+1}, \ldots, x_n, \tau_{n+1} \,|\, s_t)}{p(\iota_0, \vec{x}_n, \tau_{n+1})} \\
&= \frac{\alpha_t(s_t)\beta_t(s_t)}{\sum_{s'_t \in \mathcal{S}} \alpha_t(s'_t)\beta_t(s'_t)}\,. \qquad (5.10)
\end{aligned}$$