Notes on Sequence Modelling

G.A. Jarrad

July 3, 2015

1 Introduction

blah, blah, blah

2 Random Process Sequences

Consider, in general terms, a random process R that generates a sequence of variables, $R_1, R_2, R_3, \ldots$, where the index i gives the discrete *stage* in the sequence, and each variable R_i randomly takes some value $r_i \in \mathcal{R}$. Then, for some arbitrary sequence length n, we define

$$\vec{R}_n \quad = \quad (R_1, R_2, \ldots, R_n) \tag{2.1}$$

to be a length-n process sequence, i.e. $|\vec{R}_n| = n$, and further define

$$\vec{r}_n \quad = \quad (r_1, r_2, \ldots, r_n) \in \mathcal{R}^n \tag{2.2}$$

to be a corresponding length-n observation sequence of values. The probability (for a discrete-value process) or probability density (for a continuous-value process) of a given sequence $\vec{r}_n$ is then defined as

$$p(\vec{R}_n = \vec{r}_n) \quad = \quad p(R_1 = r_1, \ldots, R_n = r_n)\,. \tag{2.3}$$

Hence, note that if R is a discrete-value process then we must have

$$\sum_{\vec{r}_n \in \mathcal{R}^n} p(\vec{R}_n = \vec{r}_n) \quad = \quad \sum_{r_1 \in \mathcal{R}} \cdots \sum_{r_n \in \mathcal{R}} p(R_1 = r_1, \ldots, R_n = r_n) \quad = \quad 1\,. \tag{2.4}$$

Similarly, if R is a continuous-value process, then we must instead have

$$\int_{\mathcal{R}^n} p(\vec{R}_n = \vec{r}_n)\, d\vec{r}_n \;=\; \int_{\mathcal{R}} \cdots \int_{\mathcal{R}} p(R_1 = r_1, \ldots, R_n = r_n)\, dr_1 \cdots dr_n = 1\,. \tag{2.5}$$

In other words, given the sequence length n, the set $\mathcal{R}^n$ of all possible sequences $\vec{r}_n$ covers the entire probability space.

This summation property causes modelling problems if we do not know in advance the exact length of a sequence. For example, if the set of all length-1 sequences already covers the entire probability space, and so too does the set of all length-2 sequences, then the set of all length-1 and length-2 sequences covers the space twice over. This problem, however, only exists due to our ambiguous notion of a sequence. In practice, suppose we have observed a given sequence $\vec{r}_n$. How do we know if the underlying process R has 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 $\ldots, r_{-2}, r_{-1}, r_0$?

In order to handle such difficulties, we distinguish between a so-called *incomplete* sequence $\vec{R}_n$, and a *complete* sequence $\langle\vec{R}_n\rangle$ that has definite stages of initiation and termination[1]. A complete sequence may be specified by introducing indicator variables that define the start and end of the sequence. Thus, indicator ι_0 takes a value of 1 if the sequence starts at stage 0 (i.e. just prior to R_1), or a value of 0 if it does not. Similarly, indicator τ_{n+1} takes a value of 1 if the sequence terminates at stage $n+1$ (i.e. immediately after R_n), or a value of 0 if it does not. The probability (or probability density) of a given complete sequence $\langle\vec{r}_n\rangle$ is then defined as

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

The augmented random process is depicted in Figure 2.1.

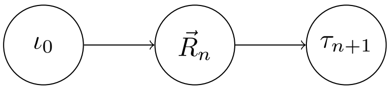

Figure 2.1: A random process for generating both complete and incomplete sequences of length n.

We may also consider *partially complete* sequences, namely the *start sequence* $\langle\vec{R}_n]$, which was initiated at stage 0 but not yet definitely terminated (i.e. $\iota_0 = 1$ but τ_{n+1} is unknown), and the *end sequence* $[\vec{R}_n\rangle$, which was terminated at stage $n+1$ but not definitely initiated at stage 0 (i.e. $\tau_{n+1} = 1$ but ι_0 is unknown). The probability of a given start sequence $\langle\vec{r}_n]$ is then defined as

$$p(\langle\vec{r}_n]) \quad = \quad p(\iota_0 = 1, R_1 = r_1 \ldots, R_n = r_n)\,, \tag{2.7}$$

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

$$p([\vec{r}_n\rangle) \quad = \quad p(R_1 = r_1 \ldots, R_n = r_n, \tau_{n+1} = 1) \tag{2.8}$$

for an end sequence. 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

$$p(\langle\vec{r}_n!) \quad = \quad p(\iota_0 = 1, R_1 = r_1 \ldots, R_n = r_n, \tau_{n+1} = 0)\,. \tag{2.9}$$

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

$$p(!\vec{r}_n\rangle) \quad = \quad p(\iota_0 = 0, R_1 = r_1 \ldots, R_n = r_n, \tau_{n+1} = 1)\,. \tag{2.10}$$

[1] Thus, we see that each length-2 incomplete sequence $\vec{r}_2$ starts with a length-1 incomplete sequence $\vec{r}_1$, so that measuring the set of all length-1 and length-2 incomplete sequences amounts to double counting.

Hence, for a discrete-value process, the probability mass of all complete length-n sequences is given by

$$\sum_{\vec{r}_n \in \mathcal{R}^n} p(\langle\vec{r}_n\rangle) \quad = \quad p(\iota_0 = 1, \tau_{n+1} = 1) \;\doteq\; P(N = n)\,, \tag{2.11}$$

where we have introduced the random variable N to denote the length of an arbitrary complete sequence. For the corresponding continuous-value process, we likewise deduce that

$$\int_{\mathcal{R}^n} p(\langle\vec{r}_n\rangle)\; d\vec{r}_n \quad = \quad p(\iota_0 = 1, \tau_{n+1} = 1) \;=\; P(N = n)\,. \tag{2.12}$$

We therefore deduce that the set $\mathcal{R}^* = \bigcup_{n=0}^{\infty} \mathcal{R}^n$ of all complete sequences of arbitrary length covers the entire probability space exactly once, since for the discrete-value case we have

$$\sum_{\vec{r}_* \in \mathcal{R}^*} p(\langle\vec{r}_*\rangle) \quad = \quad \sum_{n=0}^{\infty} \sum_{\vec{r}_n \in \mathcal{R}^n} p(\langle\vec{r}_n\rangle) \;=\; \sum_{n=0}^{\infty} P(N = n) \;=\; 1\,, \tag{2.13}$$

and for the continuous-value case we have

$$\int_{\mathcal{R}^*} p(\langle\vec{r}_*\rangle)\; d\vec{r}_* \quad = \quad \sum_{n=0}^{\infty} \int_{\mathcal{R}^n} p(\langle\vec{r}_n\rangle)\; d\vec{r}_n \;=\; \sum_{n=0}^{\infty} P(N = n) \;=\; 1\,. \tag{2.14}$$

3 Markov Process Sequences

In Section 2 we defined random processes and the sequences they generate. We now assume that the random process R is also *causal*, meaning that the distribution of values for variable R_t, at stage t, depends entirely upon the values generated previously in the sequence at stages $t-1, t-2, \ldots, 1$. In addition, for a complete sequence the distribution of the variable R_1 at the initial stage depends strongly upon being first in the sequence, and likewise the distribution of the variable R_n, for a length-n sequence, depends strongly upon the past values in the sequence and weakly on the fact that it is the final stage. Hence, the probability of a complete, causal sequence is taken here to be

$$\begin{aligned} p(\langle\vec{r}_n\rangle) \quad = \quad & p(\iota_0 = 1) \prod_{t=1}^{n} p(R_t = r_t \mid \vec{R}_{t-1} = \vec{r}_{t-1}, \iota_0 = 1) \\ & p(\tau_{n+1} = 1 \mid \vec{R}_n = \vec{r}_n, \iota_0 = 1)\,. \end{aligned} \tag{3.1}$$

The causal process is depicted in Figure 3.1.

The related models for partially complete or incomplete sequences may 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 may 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 $\vec{v}$ denoted 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 (3.1) reduces to

$$p(\vec{v}) \quad = \quad \prod_{t=0}^{n+1} p(V_t = v_t \mid \vec{\Pi}_t(\vec{V}) = \vec{\pi}_t(\vec{v}))\,, \tag{3.2}$$

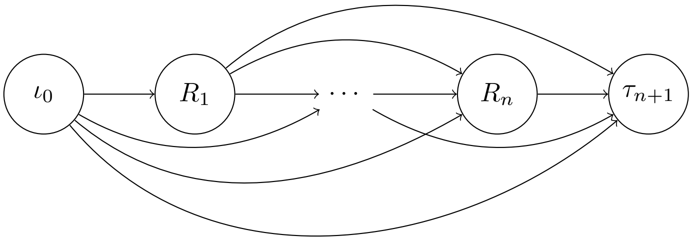

Figure 3.1: A fully-connected causal process for generating temporal sequences of length n.

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

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

where the predecessor nodes are now given by

$$\vec{\Pi}_t^{(m)}(\vec{V}) \quad = \quad \begin{cases} (V_0, V_1, \ldots, V_{t-1}) & \text{if } t \leq m\,, \\ (V_{t-m}, V_{t-m+1}, \ldots, V_{t-1}) & \text{if } t > m \end{cases} \,. \tag{3.4}$$

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

$$p(\vec{v}) \quad = \quad \prod_{t=0}^{n+1} p(V_t = v_t \mid V_{t-1} = v_{t-1})\,. \tag{3.5}$$

This process is depicted in Figure 3.2.

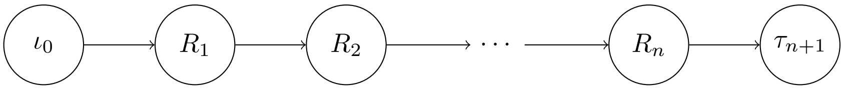

Figure 3.2: A first-order Markov process for generating causal sequences of length n.