Appendix A Notation

Processes

$\mathcal{X}$	Alphabet (set of symbols), canonically $\{0, 1, \ldots, m-1\}$.
$\mathcal{X}^{\mathbb{Z}}$	The set of bi-infinite sequences of symbols in $\mathcal{X}$.
$\mathcal{X}^-$, $\mathcal{X}^+$	The history and future sequence spaces, which are sets of semi-infinite sequences.
x	A bi-infinite sequence, that is, an element of $\mathcal{X}^{\mathbb{Z}}$.
x^-, x^+	Semi-infinite history and future sequences, which are elements of $\mathcal{X}^-$ and $\mathcal{X}^+$ respectively.
x_i	ith element in sequence x, x^+, or x^-.
$\mathcal{X}^*$	The set of all (finite-length) words of symbols in $\mathcal{X}$.
w, s	Words in $\mathcal{X}^*$ or subsequences, especially history suffixes, which are subsequences with end time -1, or next words, which are subsequences with end time 0.
$\|w\|$	The length of a word or subsequence w.
A_w	The set of (bi-infinite) sequences which match the word w.
$\mathbb{X}$	The σ-field on $\mathcal{X}^{\mathbb{Z}}$ generated by the cylinder sets.
$\mathbb{F}$	The future σ-field, subset of $\mathbb{X}$.
$\mathbb{H}$	The history σ-field, subset of $\mathbb{X}$.
$\mathbf{P}$	A probability distribution on measureable space $\left(\mathcal{X}^{\mathbb{Z}}, \mathbb{X}\right)$.
$\mathcal{P}$	A process, which is a measure space $\left(\mathcal{X}^{\mathbb{Z}}, \mathbb{X}, \mathbf{P}\right)$ in which $\mathbf{P}$ is stationary.
$\mathcal{N}$	The set of bad histories, on which we do not condition.
R	The set of non-null history suffixes — that is, the set of words with positive probability — on which conditioning is well defined.
$\mathbf{P}(\cdot\|s)$	The conditional distribution on the future induced by a word s. If $s \in R$, $\mathbf{P}(\cdot\|s)$ is a reachable process state.
$\mathbf{P}\left(\cdot\|\mathrm{x}^-\right)$	The conditional distribution on the future induced by a history x^-. If $s \notin \mathcal{N}$, $\mathbf{P}\left(\cdot\|\mathrm{x}^-\right)$ is an infinitely preceded process state.
$\mathbf{A}$	A process state, which is a conditional distribution on the future induced by conditioning on a history, a history suffix, or both.

Hidden Markov Models

V	Set of presentation states or Markov chain states.
$\lvert V\rvert$	The size — that is, number of elements — of V.
π	The initial distribution of a Markov chain or an HMM, which we always take to be stationary.
P	The transition matrix of a Markov Chain (V, P, π).
i, j	Indices which refer to specific presentation states.
k	An index which refers to a symbol in the alphabet $\mathcal{X}$.
T^k	A joint matrix of a Hidden Markov Model, also sometimes referred to as a transition matrix. Note that the superscript k is an index, not an exponent.
T^k_{ij}	An entry in a joint matrix: if the HMM is in the presentation state i, T^k_{ij} is the probability that the HMM will make a transition to presentation state j and emit the symbol k.
T^w	For any word $w = w_1 \ldots w_n$, T^w is the product $T^{w_1}T^{w_2}\cdots T^{w_n}$.
$\left(V, \mathcal{X}, \left\{T^k\right\}, \pi\right)$	A Hidden Markov Model. The of matrices $\left\{T^k\right\}$ contains one matrix for each symbol $k \in \mathcal{X}$.
B	The output matrix for an HMM which emits symbols from states rather than from transitions. Such an HMM may be converted to joint matrix form by assigning $T^k_{ij} = P_{ij}B_{jk}$.
$\vec{1}$	A column vector containing all 1s. Size is implied by context.
N	The normalization operator: if v is a row vector, $N(v)$ is v multiplied by a scalar so that its entries sum to 1.
$\eta(s)$, $\eta\left(\mathbf{x}^-\right)$	The mixed state induced by a history suffix $s \in R$ or a history $\mathbf{x}^- \in \mathcal{N}$. For s, we have $\eta(s) = N(\pi T^s)$.
μ, ν	Mixed states of an HMM, which are row vectors satisfying $\mu\vec{1} = 1$.
$\mathcal{W}$	The space of all signed measures on the future.
$\mathcal{U}$	The span of the reachable process states, which is a subspace of $\mathcal{W}$.

Generalized Hidden Markov Models

M	Conjugation matrix, which is an invertible unit-sum matrix.
$\mathcal{H}$, $\mathcal{F}$	History and future vector spaces, which are spaces of row and column vectors.
$K_{\mathcal{F}}$	The subspace of $\mathcal{H}$ consisting of all vectors which are sent to zero by multiplication on the right by every vector in $\mathcal{F}$.
$K_{\mathcal{H}}$	The subspace of $\mathcal{F}$ consisting of all vectors which are sent to zero by multiplication on the left by every vector in $\mathcal{H}$.
H	A matrix, the rows of which are mixed states and form a basis for $\mathcal{H}/K_{\mathcal{F}}$.
F	A matrix, the columns of which have the form $T^s\vec{1}$ and form a basis for $\mathcal{F}/K_{\mathcal{H}}$.
W, S	History and future wordlists. The rows of H are the mixed states induced by the words in W, and the columns of F are the vectors $T^s\vec{1}$ for the words $s \in S$.
B^k	A joint matrix for the standard presentation. B^k solves $HT^kF = B^kHF$.
γ	The initial vector of the standard presentation. γ solves $\pi F = \gamma HF$.

Reconstruction

q_i The ith word in a fixed ordering of $\mathcal{X}^*$.

P The $\mathbb{N} \times \mathbb{N}$ matrix with entries $P_{ij} = \mathbf{P}(q_j|q_i)$.

$\overline{P}$ The $|W'| \times |S'|$ truncation of P containing those rows and columns which correspond to words in the large wordlists W' and S' respectively.

$\pi(s|w)$ The frequency-count estimate of $\mathbf{P}(s|w)$ estimated from sample data.

$\hat{P}$ A $|W'| \times |S'|$ approximation to $\overline{P}$ estimated from sample data.

$r(w)$ The row of P which corresponds to the word w.

$c(s)$ The column of P which corresponds to the word s.

$r'(w)$ The sub-row of $r(w)$ containing only those columns corresponding to words in S

$c'(s)$ The sub-column of $c(s)$ containing only those rows corresponding to words in W.

G The submatrix of P containing those rows and columns which correspond to words in the wordlists W and S respectively. G is normally chosen to be invertible.

C^k The $|W| \times |S|$ matrix with entries $C^k_{ij} = \mathbf{P}(ks_j|w_i)$.

B^k A joint matrix of the reconstructed presentation, given by $B^k = C^k G^{-1}$

Appendix B Selected Probability Theory

This appendix outlines selected elements of probability theory that are used in this dissertation. For a thorough presentation of this material, see any text on the subject, for example [14,13].

B.1 Kolmogorov's Extension Theorem and Process Existence

The purpose of this section is to prove the following theorem, which is our tool for showing that processes exist.

Theorem B.1.1. Given a map $f : \mathcal{X}^* \to [0,1]$ statisfying

1. $f(\lambda) = 1$, and
2. For all words $w \in \mathcal{X}^*$, $f(w) = \sum\limits_{z \in \mathcal{X}} f(zw) = \sum\limits_{z \in \mathcal{X}} f(wz)$,

there is a unique (stationary) process $\mathcal{P} = \left(\mathcal{X}^{\mathbb{Z}}, \mathbb{X}, \mathbf{P}\right)$ such that for all $w \in \mathcal{X}^*$, $\mathbf{P}(w) = f(w)$.

This result is derived from Kolmogorov's extension theorem, which appears in the literature in several forms. None of the forms the author has seen, however, can be transformed into the form we need without an unreasonable amount of manipulation, thus this section. We will use $\mathcal{R}^n$ and $\mathcal{R}^{\mathbb{N}}$ to refer to the Borel σ-fields on $\mathbb{R}^n$ and $\mathbb{R}^{\mathbb{N}}$, respectively.

Theorem B.1.2. (Kolmogorov's Extension Theorem [13 p. 428]) Suppose that we are given probability measures μ_n on $(\mathbb{R}^n, \mathcal{R}^n)$ that are consistent; that is,

$$\mu_{n+1}((a_1, b_1] \times \ldots \times (a_n, a_n] \times \mathbf{R}) = \mu_n((a_1, b_1] \times \ldots \times (a_n, a_n]). \tag{B.1}$$

Then there is a unique probability measure $\overline{\mathbf{P}}$ on $\left(\mathbb{R}^{\mathbb{N}}, \mathcal{R}^{\mathbb{N}}\right)$ with

$$\overline{\mathbf{P}}(x | x \in (a_i, b_i], 1 \leq i \leq n) = \mu_n((a_1, b_1] \times \ldots \times (a_n, b_n]). \tag{B.2}$$

There are three diferences between the probability measure $\left(\mathbb{R}^{\mathbb{N}}, \mathcal{R}^{\mathbb{N}}, \overline{\mathbf{P}}\right)$ shown to exist by Kolmogorov's Extension Theorem and a stationary process $\left(\mathcal{X}^{\mathbb{Z}}, \mathbb{X}, \mathbf{P}\right)$. We will need to surmount all three to prove theorem B.1.1. First, $\mathbb{R}^{\mathbb{N}}$ is a product of copies of the real numbers and $\mathcal{X}^{\mathbb{Z}}$ is a product of copies of the finite discrete set $\mathcal{X}$. We will deal with this by an injective map $g : \mathcal{X} \to \mathbb{R}$. Second, and most troublesome, elements of $\mathbb{R}^{\mathbb{N}}$ are semi-infinite sequences and elements of $\mathcal{X}^{\mathbb{Z}}$ are bi-infinite sequences. Our trick for working around this difficulty makes use of a bijective map $h : \mathbb{N} \to \mathbb{Z}$, and involves considering the integers in the order $0, 1, -1, 2, -2, \ldots$. This has an unfortunate effect on the readability of the proofs. Last, $\overline{\mathbf{P}}$ need not be stationary nor does it need to satisfy any similar condition. We will use Kolmogorov's Extension Theorem to show that $\mathbf{P}$ exists, and then prove separately that it is stationary.

Before we begin, we will introduce several functions and some notation. First, we define g to be any injective map $g : \mathcal{X} \to \mathbb{R}$. It will not matter what the images of particular symbols $x \in \mathcal{X}$ are, as long as they are different.

Second, we define the bijective map $h : \mathbb{N} \to \mathbb{Z}$ mentioned above by

$$h(n) = \begin{cases} n/2 & n \text{ even} \\ (1-n)/2 & n \text{ odd.} \end{cases} \tag{B.3}$$

Its inverse is given by

$$h^{-1}(y) = \begin{cases} 2y & y > 0 \\ 1 - 2y & y \leq 0. \end{cases} \tag{B.4}$$

In effect, h alternately returns positive and negative integers. It maps $1, 2, 3, 4, 5$ to $0, 1, -1, 2, -2$ respectively, and h^{-1} maps $-2, -1, 0, 1, 2$ to $5, 3, 1, 2, 4$ in that order. Several more functions are defined in terms of h: $J(n)$ is a set-valued function $J(n) = \left\{y \in \mathbb{Z} | h^{-1}(y) \leq n\right\}$, and $d(n)$ and $c(n)$ are respectively the largest and smallest values in $J(n)$. Because we will use $J(n)$ primarily as an index set, we will consider its elements in a particular order, namely increasing numerical order. These functions can be characterized by the following equations:

$$c(n) = \text{min}(h(n), h(n-1)) = \begin{cases} \frac{2-n}{2} & n \text{ even} \\ \frac{1-n}{2} & n \text{ odd} \end{cases} \tag{B.5}$$

$$d(n) = \text{max}(h(n), h(n-1)) = \begin{cases} \frac{n}{2} & n \text{ even} \\ \frac{n-1}{2} & n \text{ odd} \end{cases} \tag{B.6}$$

$$J(n) = \{c(n), \ldots, d(n)\}. \tag{B.7}$$

The reader may wish to verify a few facts which will be needed presently. If n is even, we have $h(n+1) < 0$, $c(n+1) = c(n) - 1$ and $d(n+1) = d(n)$. If n is odd, we have $h(n+1) > 0$, $c(n+1) = c(n)$ and $d(n+1) = d(n) + 1$.

Now we move on to sets of subsequences. We will use $\mathcal{X}^{[a,b]}$ to denote the set of all subsequences $x_a \ldots x_b$ with start time a and end time b. Similarly, we will use $\mathcal{X}^{J(n)}$ to denote the set of all subsequences $x_{c(n)} \ldots x_{d(n)}$ with start time $c(n)$ and end time $d(n)$. Because $|J(n)|$ is always equal to n, $\mathcal{X}^{J(n)}$ is a product of n copies of $\mathcal{X}$. In fact, $\mathcal{X}^{J(n)}$ is identical to $\mathcal{X}^n$ except that the coordinates of $\mathcal{X}^{J(n)}$ are labeled with $c(n), c(n+1), \ldots, d(n)$ instead of $1, 2, \ldots, n$. Thus, if n is odd, $\mathcal{X}^{J(n+1)} = \mathcal{X}^{J(n)} \times \mathcal{X}$, whereas if n is even, $\mathcal{X}^{J(n+1)} = \mathcal{X} \times \mathcal{X}^{J(n)}$.

In addition, we need to define a function that takes subsequences in $\mathcal{X}^{J(n)}$ to subsequences in $\mathbb{R}^n$ and a closely related function that takes bi-infinite sequences in $\mathcal{X}^{\mathbb{Z}}$ to semi-infinite sequences in $\mathbb{R}^{\mathbb{N}}$. We will denote both of these functions by H. They will reorder their argument's coordinates and map them into $\mathbb{R}$. If $x \in \mathcal{X}^{J(n)}$, then there is a subsequence $v \in \mathbb{R}^n$ that satisfies $v_i = g\big(x_{h(i)}\big)$ for all $i \in 1, \ldots, n$. We define $H : \mathcal{X}^{J(n)} \rightarrow \mathbb{R}^n$ by $H(x) = v$. For example, if $x = x_{-1}x_0x_1x_2 \in \mathcal{X}^{J(4)}$, then

$$H(x_{-1}x_0x_1x_2) = g(x_0)g(x_1)g(x_{-1})g(x_2). \tag{B.8}$$

For an arbitrary $x = x_{c(n)} \cdots x_{d(n)} \in \mathcal{X}^{J(n)}$, we have

$$H\big(x_{c(n)}x_{c(n)+1} \ldots x_{d(n)}\big) = g(x_0)g(x_1) \ldots g\big(x_{c(n)+d(n)+1}\big). \tag{B.9}$$

Similarly, if $\mathrm{x} \in \mathcal{X}^{\mathbb{Z}}$, then there exists $\mathrm{v} \in \mathbb{R}^{\mathbb{N}}$ such that for all $i \in \mathbb{Z}$, the coordinate v_i satisfies $v_i = g\big(x_{h(i)}\big)$. We define $H : \mathcal{X}^{\mathbb{Z}} \rightarrow \mathbb{R}^{\mathbb{N}}$ by $H(\mathrm{x}) = \mathrm{v}$. Neither version of H is invertible, because g is not invertible. However, they do have set inverses. For instance, if $A \subset \mathbb{R}^n$ then $H^{-1}(A) = \Big\{x \in \mathcal{X}^{J(n)} | H(x) \in A\Big\}$.

Next, we define an *indexed product set* S to be a subset of $\mathcal{X}^{[a,b]}$ of the form $S = S_a \times \ldots \times S_b$, where $S_i \subset \mathcal{X}$. If $S = S_a \times \ldots \times S_b$ is an index product set, and $S' \subset \mathcal{X}$, then $S \times S'$ is an indexed product set contained in $\mathcal{X}^{[a,b+1]}$. We define the cylinder map Cyl which takes indexed product sets to sets of sequences as follows:

$$Cyl(S) = \Big\{\mathrm{x} \in \mathcal{X}^{\mathbb{Z}} | x_i \in S_i,\ a \leq i \leq b\Big\}. \tag{B.10}$$

That is, $Cyl(S)$ is the set of all sequences in $\mathcal{X}^{\mathbb{Z}}$ that match a subsequence in S. We will also define the shift map T on indexed product sets by

$$T(S) = \Big\{x \in \mathcal{X}^{[a-1,b-1]} | x_i \in S_{i+1},\ a \leq i+1 \leq b\Big\}. \tag{B.11}$$

(The reader may have noticed that we are using the same notation for the shift map on indexed product sets as we use on sequences.)

Lemma B.1.3. The following facts about Cyl and T may be easily verified, and we will give no proof.

1. $Cyl(S) = Cyl(S \times \mathcal{X}) = Cyl(\mathcal{X} \times S)$
2. $T(S \times \mathcal{X}) = T(S) \times \mathcal{X} = \mathcal{X} \times T(S)$
3. $T(Cyl(S)) = Cyl(T(S))$

One special interaction is worth noting. Let $S \subset \mathcal{X}^{J(n)}$ be an indexed product set $S = S_{c(n)} \times \ldots \times S_{d(n)}$ and let

$$A = Cyl(S) = \Big\{ \mathbf{x} \in \mathcal{X}^{\mathbb{Z}} | x_i \in S_i \text{ for all } i,\ c(n) \leq i \leq d(n) \Big\}. \tag{B.12}$$

Then we have

$$T(S) = \Big\{ \mathbf{x} \in \mathcal{X}^{[c(n)-1, d(n)-1]} | x_i \in S_{i+1},\ c(n) \leq i+1 \leq d(n) \Big\}, \tag{B.13}$$

$$T(A) = \Big\{ \mathbf{x} \in \mathcal{X}^{\mathbb{Z}} | x_i \in S_{i+1},\ c(n) \leq i+1 \leq d(n) \Big\}, \tag{B.14}$$

and $T(A) = Cyl(T(S))$ and thus by lemma B.1.3, $T(A) = Cyl(T(S) \times \mathcal{X})$.

Finally, notice that for any indexed product set $S \in \mathcal{X}^{[a,b]}$ there is a set $\overline{S} \subset \mathcal{X}^{J(n)}$, where $n = \min(-2a, 2b+1)$, such that $Cyl(S) = Cyl(\overline{S})$. Since every cylinder set can be written as $Cyl(S)$ for some $S \in \mathcal{X}^{[a,b]}$, this means that every cylinder set can be written as $Cyl(\overline{S})$ for some $\overline{S} \in \mathcal{X}^{J(n)}$.

At last, we are ready to state and prove a result. This is Kolmogorov's extension theorem in a form which applies to processes.

Theorem B.1.4 (Bidirectional Discrete version of Kolmogorov's Extension Theorem). Suppose we have a sequence of measures ν_n on $\mathcal{X}^{J(n)}$ which satisfy the following conditions for all n and for all indexed product sets $S \subset \mathcal{X}^{J(n)}$. If n is odd,

$$\nu_n(S) = \nu_{n+1}(S \times \mathcal{X}) \tag{B.15}$$

and if n is even,

$$\nu_n(S) = \nu_{n+1}(\mathcal{X} \times S) \text{ and} \tag{B.16}$$

$$\nu_n(S) = \nu_{n+1}(T(S) \times \mathcal{X}). \tag{B.17}$$

Then there exists a unique stationary process $\mathcal{P} = \left(\mathcal{X}^{\mathbb{Z}}, \mathbb{X}, \mathbf{P}\right)$ such that, for all n and for all $S \subset \mathcal{X}^{J(n)}$,

$$\mathbf{P}(Cyl(S)) = \nu_n(S). \tag{B.18}$$

The proof of this statement is in two parts. In the first, we show that $\mathbf{P}$ exists using the awkward mapping tricks defined above, theorem B.1.2 and equations B.15 and B.16. In the second, we use equations B.15, B.16, and B.17 to show that $\mathbf{P}$ is stationary.

Proof. We define a sequence of measures μ_n on $\mathbb{R}^n$ as follows: if $A \subset \mathbb{R}^n$, we define $\mu_n(A) = \nu_n\left(H^{-1}(A)\right)$, that is

$$\mu_n(A) = \nu_n\left\{x \in \mathcal{X}^{J(n)} | H(x) \in A\right\} \tag{B.19}$$

The following calculation establishes that the μ_ns are consistent. By definition,

$$\mu_{n+1}(A \times \mathbb{R}) = \nu_{n+1}\left(H^{-1}(A \times \mathbb{R})\right). \tag{B.20}$$

Now, $H^{-1}(A \times \mathbb{R})$ can be rewritten as

$$H^{-1}(A \times \mathbb{R}) = \begin{cases} H^{-1}(A) \times \mathcal{X} & n \text{ odd} \\ \mathcal{X} \times H^{-1}(A) & n \text{ even,} \end{cases} \tag{B.21}$$

so we have

$$\mu_{n+1}(A \times \mathbb{R}) = \begin{cases} \nu_{n+1}\left(H^{-1}(A) \times \mathcal{X}\right) & n \text{ odd} \\ \nu_{n+1}\left(\mathcal{X} \times H^{-1}(A)\right) & n \text{ even.} \end{cases} \tag{B.22}$$

And by applying equations B.15 and B.16 to the right-hand sides, we get

$$\begin{aligned} \mu_{n+1}(A \times \mathbb{R}) &= \nu_n\left(H^{-1}(A)\right) \\ &= \mu_n(A). \end{aligned} \tag{B.23}$$

Thus, we can apply Kolmogorov's Extension Theorem to the measures μ_n, which gives us a unique measure $\overline{\mathbf{P}}$ on $\left(\mathbb{R}^{\mathbb{N}}, \mathcal{R}^{\mathbb{N}}\right)$ which agrees with the μ_n: if A_i is a Borel set for all $i \in 1, \ldots, n$ and $A = A_1 \times \ldots \times A_n$, then

$$\overline{\mathbf{P}}(x | x_i \in A_i, \; i \in 1, \ldots, n). \tag{B.24}$$

Now we define $\mathbf{P}$. For all $S \in \mathbb{X}$,

$$\mathbf{P}(S) = \overline{\mathbf{P}}(H(x) | x \in S). \tag{B.25}$$

This $\mathbf{P}$ is in fact an extension of the ν_ns. For all $S \in \mathcal{X}^{J(n)}$, we have

$$\begin{aligned} \nu_n(S) &= \mu_n(H(S)) \\ &= \overline{\mathbf{P}}\Big(a \in \mathbb{R}^{\mathbb{N}} | a_1 \dots a_n \in H(S)\Big) \\ &= \mathbf{P}\Big(\mathbf{x} \in \mathcal{X}^{\mathbb{Z}} | x_{c(n)} \dots x_{d(n)} \in S\Big) = \mathbf{P}(Cyl(S)) \end{aligned} \tag{B.26}$$

Thus, we have shown that $\mathbf{P}$ exists.

To show that $\mathbf{P}$ is stationary, we need to show that $\mathbf{P}(T(A)) = \mathbf{P}(A)$ for a sufficiently rich set of $A \in \mathbb{X}$. Any collection which contains all the cylinder sets will suffice. The collection we choose is

$$\mathcal{A} = \Big\{Cyl(S) | \text{ indexed product sets } S \subset \mathcal{X}^{J(n)} \text{ for some } n\Big\}. \tag{B.27}$$

Thus, every $A \in \mathcal{A}$ has an associated n and an associated S. (Of course, there will be more than one suitable n, S pair, but there will be a smallest n and a unique associated S, and these are the n and S to which we refer.)

If n is even, equation B.17 gives us

$$\begin{aligned} \mathbf{P}(A) &= \nu_n(S) \\ &= \nu_{n+1}(T(S) \times \mathcal{X}) \\ &= \mathbf{P}(Cyl(T(S) \times \mathcal{X})). \end{aligned} \tag{B.28}$$

And since $Cyl(T(S) \times \mathcal{X}) = Cyl(T(S)) = T(A)$, this becomes $\mathbf{P}(A) = \mathbf{P}(T(A))$.

If n is odd, we expand S by one and then do a similar calculation.

$$\begin{aligned} \mathbf{P}(A) &= \nu_n(S) \\ &= \nu_{n+1}(S \times \mathcal{X}) \\ &= \nu_{n+2}(T(S \times \mathcal{X}) \times \mathcal{X}) \\ &= \mathbf{P}(Cyl(T(S \times \mathcal{X}) \times \mathcal{X})) \end{aligned} \tag{B.29}$$

Here we can again apply lemma B.1.3 to get $Cyl(T(S \times \mathcal{X}) \times \mathcal{X}) = T(A)$, and thus $\mathbf{P}(A) = \mathbf{P}(T(A))$.■

Now we are ready to prove theorem B.1.1, which we restate here:

Theorem B.1.1. Given a map $f : \mathcal{X}^* \to [0,1]$ statisfying

1. $f(\lambda) = 1$, and
2. For all words $w \in \mathcal{X}^*$,

$$f(w) = \sum_{z \in \mathcal{X}} f(zw) = \sum_{z \in \mathcal{X}} f(wz), \tag{B.30}$$

there is a unique stationary process $\mathcal{P} = \left(\mathcal{X}^{\mathbb{Z}}, \mathbb{X}, \mathbf{P}\right)$ such that for all $w \in \mathcal{X}^*$, $\mathbf{P}(w) = f(w)$.

The proof will proceed as follows: we will construct measures ν_n which satisfy equations B.15, B.16, and B.17, and then apply Theorem B.1.4 to get the result.

Proof. For all $w \in \mathcal{X}^*$ and $|w| = n$, let $S = \{w\}$ and define

$$\nu_n(S) = f(w). \tag{B.31}$$

For a general $S \subset \mathcal{X}^{J(n)}$, S is a disjoint union of sets of the form $S_w = \{w\}$. Thus we may safely define the measure ν_n on all subsets of $\mathcal{X}^{J(n)}$ by

$$\nu_n(S) = \sum_{w \in S} \nu_n(S_w) = \sum_{w \in S} f(w). \tag{B.32}$$

We need to show that ν_n is a probability measure; that is, we need to show that $\nu_n\left(\mathcal{X}^{J(n)}\right) = 1$. We will do this by induction on n. If $n = 0$ then $\mathcal{X}^{J(n)} = \{\lambda\}$ and we are given $f(\lambda) = 1$, so ν_1 is a probability measure. The induction step depends on equation B.30, and the odd and even cases must be done separately. If n is odd and ν_n is a probability distribution, then we have

$$\begin{aligned} \nu_{n+1}\left(\mathcal{X}^{J(n+1)}\right) &= \sum_{w \in \mathcal{X}^{J(n+1)}} f(w) \\ &= \sum_{w \in \mathcal{X}^{J(n)}} \sum_{x \in \mathcal{X}} f(wx) \\ &= \sum_{w \in \mathcal{X}^{J(n)}} f(w) \\ &= \nu_n\left(\mathcal{X}^{J(n)}\right) = 1. \end{aligned} \tag{B.33}$$

If n is even and ν_n is a probability measure, then the calculation is the same except that we write $f(xw)$ in place of $f(wx)$ and we use the other half of equation B.30.

Now we must show that equations B.15, B.16, and B.17 are satisfied. We will establish each of them using equation B.30 much as before. First equation B.15, assuming n is odd:

$$\begin{aligned} \nu_n(S) = \sum_{w \in S} f(w) &= \sum_{w \in S} \sum_{x \in \mathcal{X}} f(wx) \\ &= \sum_{z \in (S \times \mathcal{X})} f(z) \\ &= \nu_{n+1}(S \times \mathcal{X}). \end{aligned} \tag{B.34}$$

Second, equation B.16, assuming n is even:

$$\begin{aligned}\nu_n(S) = \sum_{w \in S} f(w) &= \sum_{w \in S} \sum_{x \in \mathcal{X}} f(xw) \\ &= \sum_{z \in (\mathcal{X} \times \mathcal{S})} f(z) \\ &= \nu_{n+1}(\mathcal{X} \times S)\end{aligned} \tag{B.35}$$

Lastly, equation B.17, again assuming n is even:

$$\begin{aligned}\nu_n(S) = \sum_{w \in S} f(w) &= \sum_{w \in S} \sum_{x \in \mathcal{X}} f(wx) \\ &= \sum_{z \in (S \times \mathcal{X})} f(z).\end{aligned} \tag{B.36}$$

But this time, $S \times \mathcal{X} \not\subset \mathcal{X}^{J(n)}$, so the expression $\nu_{n+1}(S \times \mathcal{X})$ does not make sense. However, we do have $T(S \times \mathcal{X}) = T(S) \times \mathcal{X} \subset \mathcal{X}^{J(n)}$, and f does not depend on time indices, so we have

$$\begin{aligned}\nu_n(S) &= \sum_{z \in (S \times \mathcal{X})} f(z) = \sum_{z \in (T(S) \times \mathcal{X})} f(z) \\ &= \nu_{n+1}(T(S) \times \mathcal{X}).\end{aligned} \tag{B.37}$$

We have now shown that the ν_ns satisfy all of the conditions of theorem B.1.4. Thus, applying this theorem, we have shown that there exists a unique stationary process $\mathcal{P} = \left(\mathcal{X}^{\mathbb{Z}}, \mathbb{X}, \mathbf{P}\right)$ such that for all n and for all $S \subset \mathcal{X}^{J(n)}$, we have $\mathbf{P}(Cyl(S)) = \nu_n(S)$. Therefore, if we let $S = \{w\}$ for any length n word w, we have

$$f(w) = \nu_n(S) = \mathbf{P}(Cyl(S)) = \mathbf{P}(w). \tag{B.38}$$

So we are done.■

B.2 Martingales

This section presents the martingale convergence results needed in chapters 2 and 3. Let X be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, and let $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. That is, $\mathcal{G}$ is a σ-field and $\mathcal{G} \subset \mathcal{F}$ as sets of sets.

Definition B.2.1. The *conditional expectation* of X with respect to $\mathcal{G}$ is any random variable Y that satisfies

1. Y is measureable with respect to $\mathcal{G}$, and
2. for all $A \in \mathcal{G}$,

$$\int_A X d\mathbf{P} = \int_A Y d\mathbf{P}. \tag{B.39}$$

Several facts about conditional expectation are worth noting. First, conditional expectations exist on all the probability spaces considered in this dissertation. Second, conditional expectations are unique up to sets of measure zero — if both Y and Y' satisfy definition B.2.1, then $Y = Y'$ almost everywhere. Third, a conditional expectation is a variable, not a constant. That is, Y is a function on Ω. However, Y is constant almost everywhere on atoms of $\mathcal{G}$. (A set $A \in \mathcal{G}$ is an atom if the only sets in $\mathcal{G}$ which are subsets of A are the empty set and A itself.)

The connection between conditional expectation and the conditional probability of elementary probability is as follows. If Z is a random variable on $(\Omega, \mathcal{F}, \mathbf{P})$, then Z induces a σ-field $\sigma(Z)$ on Ω, namely the smallest σ-field containing all sets of the form $Z^{-1}(B)$ for Borel sets B. Let A be a set in $\mathcal{F}$ with indicator function (characteristic function) 1_A. If we fix a constant $c \in \mathbb{R}$ and evaluate the conditional expectation $\mathbf{E}(1_A|\sigma(Z))$ on $x \in Z^{-1}(c)$, we find that

$$\mathbf{E}(1_A|\sigma(Z))(x) = \mathbf{P}(x \in A|Z = c) \tag{B.40}$$

for almost every such x.

A *filtration* is an increasing sequence of σ-fields $\{\mathcal{F}\} = \mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$.

Definition B.2.2. A sequence $\{X\} = X_1, X_2, \ldots$ is a *martingale* with respect to the filtration $\{\mathcal{F}\}$ if

1. for all i, X_i is measureable with respect to $\mathcal{F}_i$, and
2. for all $s, t \in \mathbb{N}$ such that $t > s$, $X_s = \mathbf{E}(X_t|\mathcal{F})$

Martingale convergence theorems are commonly stated like this: if $\{X_n\}$ is a martingale and $\mathbf{E}(|X_n|) < \infty$ for all n, then there exists a random variable X such that X_n converges almost surely to X with $\mathbf{E}(|X|) < \infty$. Note that this statement establishes that the limit X exists, not what X is.

The result needed in this disseratation is this. Given a filtration $\{\mathcal{F}\}$, let $\mathcal{G}$ be the smallest σ-field that contains every $\mathcal{F}_i$. Let A be a set in $\mathcal{F}$, and define X_n to be the conditional expectation $\mathbf{E}(1_A|\mathcal{F}_n)$. Then $\{X\}$ is a martingale with respect to $\{\mathcal{F}\}$. The fact we need is $X_n \to \mathbf{E}(1_A|\mathcal{G})$ almost surely.

The following result appears as theorem 4.3 in [36].

Theorem B.2.3. (Doob's Martingale Convergence Theorem) If $\{\mathcal{F}_n\}$ is an increasing sequence of σ-fields (that is, for all $n \geq 0$, $\mathcal{F}_n$ is a sub-σ-field of $\mathcal{F}_{n+1}$), and X is a measureable function such that $\mathbf{E}(|X|) < \infty$, then $\lim_{n \to \infty} \mathbf{E}(X|\mathcal{F}_n)$ converges almost surely to $\mathbf{E}(X|\mathcal{F})$, where $\mathcal{F}$ is the smallest σ-field which contains all of the $\mathcal{F}_n$s.

Now the result we need is simply Doob's Martingale convergence theorem restricted to the case in which X is an indicator function.

Corollary B.2.4. If $\{\mathcal{F}_n\}$ is an increasing sequence of σ-fields and A is an event, then $\mathbf{P}(A|\mathcal{F}_n) \to \mathbf{P}(A|\mathcal{F})$ almost surely, where $\mathcal{F}$ is the smallest σ-field which contains all of the $\mathcal{F}_n$s.

Bibliography

[1] D. Blackwell, “The entropy of functions fo finite-state markov chains,” in *Transactions of the First Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, (Prague), pp. 13–20, Publishing house of the Czechoslovak Academy of Sciences, 1957.

[2] D. Blackwell and L. Koopmans, “On the identifiability problem for functions of markov chains,” *Annals of Mathematical Statistics*, vol. 28, pp. 1011–1015, 1957.

[3] L. R. Rabiner and B. H. Juang, “An introduction to hidden markov models,” *IEEE Acoustics, Speech, and Signal Processing Magazine*, vol. 3, pp. 4–16, January 1986.

[4] M. J. Russell and R. K. Moore, “Explicit modeling of state occupancy in hidden markov models for speech signals,” in *Proceedings ICASSP-85 International Conference on Acoustics, Speech, and Signal Processing*, (New York), pp. 5–8, Institute of Electrical and Electronic Engineers, IEEE, 1985.

[5] K. S. Fu, *Statistical Pattern Classification Using Contextual Information*. Research Studies Press, 1980.

[6] K. Kobayashi, *A Coding-Theoretic Study on Finitary Information Systems*. University of Tokyo, 1987. Dr. Thesis.

[7] H. Ito, S. ichi Amari, and K. Kobayashi, “Identifiability of hidden markov processes and their minimum degrees of freedom,” *IEEE Transactions on Information Theory*, vol. 38, pp. 324–333, March 1992. A version of this paper appeared in Electronics and Communications in Japan, vol. 74, pp. 77-84.

[8] V. Balasubramanian, “Equivalence and reduction of hidden markov models,” Tech. Rep. A.I. Techhnical Report No. 1370, Massachusetts Institute of Technology, January 1993.

[9] J. P. Crutchfield and K. Young, “Inferring statistical complexity,” *Physics Review Letters*, vol. 63, pp. 105–108, 1989.

[10] S. W. Dharmadhikari, “Functions of finite markov chains,” *Annals of Mathematical Statistics*, vol. 34, pp. 1022–1032, 1963.

[11] J. P. Crutchfield, “The calculi of emergence: Computation, dynamics, and induction,” *Physica D*, vol. 75, pp. 11 – 54, 1994.

[12] A. M. Fraser and A. Dimitriadis, “Forcasting probability densities using hidden markov models with mixed states,” in *Time Series Prediction: Predicting the Future and Understanding the Past* (A. Weigend and N. Gershenfeld, eds.), Addison-Wesley, 1993.

[13] R. Durrett, *Probability: Theory and Examples*. Belmont, California: Duxbury, 1991.

[14] K. L. Chung, *A Course in Probability Theory*, vol. 21 of *Probability and mathematical statistics*. Academic Press, 2nd ed., 1974.

[15] J. P. Crutchfield, “Semantics and thermodynamics,” in *Nonlinear Modeling and Forecasting* (M. Casdagli and S. Eubank, eds.), vol. XII of *Santa Fe Institute Studies in the Sciences of Complexity*, (Reading, Massachusetts), p. 317, Addison-Wesley, 1991.

[16] B. H. Juang and L. R. Rabiner, “Hidden markov models for speech recognition,” *Technometrics*, vol. 33, pp. 251–272, August 1991.

[17] L. R. Rabiner, “A tutorial on hidden markov models selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, pp. 257–286, February 1989.

[18] D. Des Jardins, “-.” personal communication, 1995.

[19] J. P. Crutchfield, “Reconstructing language hierarchies,” in *Information Dynamics* (H. A. Atmanspracher and H. Scheingraber, eds.), (New York), p. 45, Plenum, 1991.

[20] E. J. Gilbert, “On the identifiability problem for functions of finite markov chains,” *Annals of Mathematical Statistics*, vol. 30, pp. 688–697, 1959.

[21] A. Heller, “On stochastic processes derived from markov chains,” *Annals of Mathematical Statistics*, vol. 36, pp. 1286–1291, 65.

[22] L. T. Niles and H. F. Silverman, “Combining hidden markov model and neural network classifiers,” in *Proceedings ICASSP-90 International Conference on Acoustics, Speech and Signal Processing*, (New York), pp. 417–420, Institute of Electrical and Electronic Engineers, IEEE, 1990.

[23] B. G. Leroux, “Maximum-likelihood estimation for hidden markov models,” *Stochastic Processes and their Applications*, vol. 40, pp. 127–143, 1992.

[24] L. E. Baum and T. Petrie, “Statistical inference for probabilistic functions of finite markov chains,” *Annals of Mathematical Statistics*, vol. 37, pp. 1554–1563, 1966.

[25] W. Qian and D. M. Titterington, “Estimation of parameters in hidden markov models,” *Philosophcal Transactions of the Royal Society, Series A Physical Sciences and Engineering*, vol. 337, pp. 407–28, 16 December 1991.

[26] A. Kehagias, “The local backward-forward algorithm,” in *IJCNN 91 Seattle: International Joint Conference on Neural Networks*, (New York), pp. 321–326 in Volume 2 of 2, IEEE, IEEE, 1991.

[27] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[28] S. W. Dharmadhikari, “Sufficient conditions for a stationary process to be a function of finite markov chain,” *Annals of Mathematical Statistics*, vol. 34, pp. 1033–1041, 1963.

[29] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipies in C*. Cambridge University Press, first ed., 1988.

[30] Y. Ephraim, A. Dembo, and L. R. Rabiner, “A minimum discrimination information approach for hidden markov modeling,” *IEEE Transactions on Information Theory*, vol. 35, pp. 1001–1013, September 1989.

[31] B. H. Juang and L. R. Rabiner, “The segmental k-means algorithm for estimating parameters of hidden markov models,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, pp. 1639–1641, September 1990.

[32] N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. S. w, “Geometry from a time series,” *Phys. Rev. Let.*, vol. 45, p. 712, 1980.

[33] F. Takens, “Detecting strange attractors in fluid turbulence,” in *Symposium on Dynamical Systems and Turbulence* (D. A. Rand and L. S. Young, eds.), vol. 898, (Berlin), p. 366, Springer-Verlag, 1981.

[34] J. J. Birch, “Approximations for the entropy for functions of markov chains,” *Annals of Mathematical Statistics*, vol. 33, pp. 930–938, 1962.

[35] J. J. Birch, “On information rates for finite-state channels,” *Information and Control*, vol. 6, pp. 372–380, 63.

[36] J. L. Doob, *Stochastic Processes*. Wiley publications in statistics, Wiley, 1953.

Index

alphabet, 7

bad histories, 17
band-merging process, 19
base word, 8
Baum-Welsh, 83

Cantor, 46
conditional expectation, 123
conditional future distribution, 14
conjugate, 59
cutpoint, 95
cylinder set, 8

degeneralization problem, 108
deterministic, 35

elusive, 20, 22
equivalence class, 14
equivalent, 59

fair coin, 11
filtration, 124
forward-backward, 83
future, 12
  events, 12
  σ-field, 12
  matrix, 65
  space, 12
  wordlist, 65

Generalized Hidden Markov Model, 2, 55, 57
GHMM, 55, 57
Golden Mean Process, 43
good histories, 17

Hidden Markov Model, 1, 28
history, 12
  events, 12
  σ-field, 12
  matrix, 65
  space, 12
  wordlist, 65
HMM, 28

identifiability, 72
indexed product set, 118
induced, 39
inducing set of histories, 18
infinitely preceded, 21
irreducible, 27

joint matrices, 28

labeled directed graph, 33

Markov Chain, 25
martingale, 124
match, 8
minimal, 65
  for a process, 86
minimization, 72
mixed state, 36
mixed state representation, 41
mixed state version, 37
modified nested parentheses process, 34

next symbol, 8
normalize, 36
null history, 13
null word, 13

order, 100

presentation, 25
presentation states, 28
process, 7, 10
  defined by an HMM presentation, 32
process state, 15, 17
process state graph, 20
Proto-Generalized Hidden Markov Model, 56
Proto-GHMM, 56

quasi-initial vector, 76
quasi-presentation, 76
quasi-transition matrix, 76

rank, 87
reachable, 21
realization, 13
reconstruction from a sample, 94
reconstruction problem, 83
recurrent, 21
recurrent component, 27
reducible, 27
reverse state, 85

SDFA, 35
shift map, 9
Simple Nondeterministic Source (SNS), 44
standard initial vector, 73
standard presentation, 72, 73
standard transition matrices, 73
stationary, 9
stationary distribution, 28
stochastic, 53
Stochastic Deterministic Finite Automaton, 35
stochastic finite automaton, 35
strictly transient, 22
subcolumn, 87
subsequence, 8
sufficient, 65
  for a process, 86
symbol, 2, 7
synchronizing word, 24

transient, 22
transition matrix, 29
Two Biased Coins (2BC), 47

underlying Markov Chain, 28
unit-sum, 52

valid, 56

word, 8
wordlist, 65