Santa Fe Institute Working Paper 13-XX-XXX
arxiv.org:13XX.XXXX [physics.gen-ph]

Mixed States of Hidden Markov Processes and Their Presentations: What and how to calculate

James P. Crutchfield[1, 2, *]

[1] *Complexity Sciences Center and Department of Physics, University of California at Davis, One Shields Avenue, Davis, CA 95616*
[2] *Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501*

(Dated: April 28, 2013)

Brief notes on state distributions and mixed states of hidden Markov Models, how to calculate them and their transition dynamic.

Keywords:

PACS numbers: 02.50.-r 89.70.+c 05.45.Tp 02.50.Ey 02.50.Ga

I. INTRODUCTION

This is a brief resource note covering methods to calculate basic features of ergodic hidden Markov processes (HMPs) generated by hidden Markov models (HMMs). The emphasis is on unifilar HMMs, such as ϵ-machines.

Sources for these notes can be found in the following publications:

- Ref. [1]: Short introduction that motivates the mixed state in terms of interesting statistical properties of HMPs.
- Ref. [2]: The theory behind it. See Sec. "Alternative Presentations".
- Refs. [3] and [4]: A useful application.

II. PRESENTATION

We will discuss generative models and the observed processes they produce.

The generative models will be given as a unifilar presentation: a class of state-based models in which observed symbols are emitted on state-to-state transitions. (The alternative, that symbols are emitted on entering a state, are not considered. There is no loss of generality, since these HMM presentations are equivalent to "edge-labeled" HMM presentations here.)

A *presentation* $M = \left\{\mathcal{A}, \boldsymbol{\mathcal{S}}, \{T^{(x)}\}\right\}$ of a HMP consists of:

1. Alphabet of observed symbols: $x \in \mathcal{A}$. The associated random variable at time t is denoted X_t.
2. Set of internal (hidden) states, $\sigma \in \boldsymbol{\mathcal{S}} = \{0, 1, \ldots, N-1\}$, with random variable $\mathcal{S}_t$ at time t; and
3. Set of symbol-labeled state transition matrices: $\{T^{(x)}_{\sigma,\sigma'} : \sigma, \sigma' \in \boldsymbol{\mathcal{S}}, x \in \mathcal{A}\}$. These determine the probability $T^{(x)}_{\sigma,\sigma'} = \Pr(\sigma'|\sigma, x)$ of transitioning from state σ' starting in state σ and seeing x.

The internal process is described by a Markov chain with transition matrix $T = \sum_{x\in\mathcal{A}} T^{(x)}$. It is a *stochastic matrix*: The transition probabilities leaving a state sum to one: $\sum_{\sigma'} T_{\sigma,\sigma'} = 1$. That is, a transition is always made on each step. The individual $T^{(x)}$ are referred to as *substochastic* matrices.

Unifilarity is a strong constraint on the $\{T^{(x)}\}$: the transition to a next state, if allowed, is uniquely determined by the current state and measurement symbol. One result is that the matrices are sparse.

For a given process, unifilar presentations need not be unique.

Since we consider only ergodic HMPs, there is a single strongly connected set of recurrent states that is visited asymptotically. Here, we assume $\boldsymbol{\mathcal{S}}$ consists only of these recurrent states.

III. JOINT PROCESS

Given a presentation of a process, there is a most prosaic description. This is the *joint process* over internal states and generated symbols:

$$\{\ldots(\sigma, x)_{t-1}, (\sigma, x)_t, (\sigma, x)_{t+1}, \ldots\} \quad (1)$$

where $\sigma_t \in \boldsymbol{\mathcal{S}}$ and $x \in \mathcal{A}$. The presentation is simply a list of all allowed realizations. Their probabilities are specified by the joint distribution $\Pr\left(\overleftrightarrow{\mathcal{S}}, \overleftrightarrow{X}\right)$.

* chaos@ucdavis.edu

IV. WORD DISTRIBUTION

The observed process is the marginal distribution of the joint process:

$$\Pr(\overleftrightarrow{X}) = \sum_{\overleftrightarrow{\sigma} \in \overleftrightarrow{\mathcal{S}}} \Pr\left(\overleftrightarrow{\mathcal{S}}{=}\overleftrightarrow{\sigma}, \overleftrightarrow{X}\right) \ .$$

We will assume a stationary observed process:

$$\Pr\left(\overleftrightarrow{X}_t\right) = \Pr\left(\overleftrightarrow{X}_0\right) \ ,$$

for all t. Due to this we drop the absolute time index where convenient.

The distribution of words $w = x_0 x_1 \ldots x_{L-1}$ of length L is a marginal of the observed process:

$$\begin{aligned} &\Pr(X^L = w) \\ &= \sum_{x_L, x_{L+1}, \ldots \in \mathcal{A}} \Pr(X^L = w, X_L = x_L, X_{L+1} = x_{L+1}, \ldots) \ . \end{aligned} \tag{2}$$

Being infinite, the explicit process presentation of Eq. (1) is often not helpful. Except, perhaps, for the case of periodic processes. This is especially true when we come to practical questions related to calculating word probabilities and other properties. In a sense, this is one of the main reasons we use models—presentations of a process. So, how do we directly calculate these from a presentation? The following sections outline what things there are to know and how to calculate them.

V. INTERNAL STATE DISTRIBUTION

Consider the internal Markov chain. We represent the state distribution $\mu_t \equiv \Pr(S_t)$ at time t as a row vector:

$$\mu_t = \Pr(\mathcal{S}_t = 0, \ldots, \mathcal{S}_t = N - 1) \ . \tag{3}$$

Then, in vector notation, the state distribution at the next time is:

$$\mu_{t+1} = \mu_t T \ . \tag{4}$$

The time-asymptotic probability of visiting states is denoted π. It is calculated as the principal left eigenvector of the internal state transition matrix:

$$\pi = \pi T \ .$$

Its components are normalized in probability: $\sum_\sigma \pi_\sigma = 1$. Since T is a stochastic matrix, the principal eigenvector is unity. Operationally, this means that Eq. (4) preserves probability: μ_{t+1} is a normalized distribution, if μ_t is.

VI. WORD DISTRIBUTION FROM PRESENTATION

The symbol probability starting from a given state is specified in the symbol-labeled transition matrices:

$$\Pr(x|\sigma) = \begin{cases} T^{(x)}_{\sigma,\sigma'}, & \text{if } \sigma' \text{ is allowed,} \\ 0, & \text{if disallowed.} \end{cases}$$

Recall that presentation's unifilarity means at most one transition is allowed on each symbol leaving a state. (At most one $T^{(x)}_{\sigma,\sigma'}$ is positive, for a given symbol $x \in \mathcal{A}$ and state $\sigma \in \mathcal{A}$.) This is an import subclass of HMM presentations. One that allows us to calculate a number of properties of the observed process in terms of the internal Markov chain. In short, unifilarity provides a mapping between sequences of internal state transitions and sequences of observed symbols.

Given a presentation of a process, the probability of seeing symbol $x \in \mathcal{A}$ is the average probability that it is generated from each state:

$$\Pr(X = x) = \sum_{\sigma \in \boldsymbol{\mathcal{S}}} \Pr(\sigma) \Pr(X = x|\sigma) \ .$$

If at time t we have $\Pr(\mathcal{S}_t) = \mu_t$, then this translates to the vector notation:

$$\Pr(X_t = x) = \mu_t T^{(x)} \mathbf{1} \ , \tag{5}$$

where $\mathbf{1}$ is the all-1s column vector. It sums over the states in which one ends after seeing x.

If we believe the process is in statistical equilibrium—that the state probabilities are given by π—then the expected symbol probability is:

$$\Pr(X = x) = \pi T^{(x)} \mathbf{1} \ . \tag{6}$$

This probability is stationary (and so the t index was dropped), since the states are distributed according to π.

These calculations extend directly to words $w \in \mathcal{A}^+$:

$$\Pr(w) = \sum_{\sigma \in \boldsymbol{\mathcal{S}}} \Pr(\sigma) \Pr(w|\sigma) \ . \tag{7}$$

The conditional probability on the RHS is the word probability starting from a given state:

$$\begin{aligned} &\Pr(X^L = w|\mathcal{S} = \sigma_0) \\ &\quad = \Pr(\sigma_0) \Pr(x_0|\sigma_0) \Pr(x_1|\sigma_1) \cdots \Pr(x_{L-1}|\sigma_{L-1}) \ . \end{aligned}$$

Here, the sequence of states visited in generating w is denoted $\sigma_0, \ldots, \sigma_{L-1}$.

Again, assuming state probabilities are stationary, in vector notation Eq. (7) becomes:

$$\begin{aligned}\Pr(X^L = w) &= \pi T^{(s_0)} \cdots T^{(s_{L-1})} \mathbf{1} \\ &= \pi T^{(w)} \mathbf{1} \ . \end{aligned} \tag{8}$$

These word probabilities are stationary since the states are distributed according to π.

When we say that we have a presentation M of a process $\mathcal{P}$, we mean that M generates all of and exactly $\mathcal{P}$'s word distribution. That is, the distributions of Eqs. (2) and (8) are the same.

VII. MIXED STATE EVOLUTION

Above, partly as an aid to compare a given stationary process to the process generated by a presentation, we assumed the presentation started in asymptotic state distribution π. We can take any initial distribution μ, though. To emphasize the geometric view of a state distribution as a vector, as in Eq. (3), we refer to μ as a *mixed state*. That is, μ is a *mixture* of state probabilities. Individual state probabilities expressed as mixed states are $\Pr(\mathcal{S} = i) = (0, \ldots, 1, \ldots, 0)$. These *pure states* are vertices of the mixed-state simplex. And so, given a presentation M, we can work with an alternative model called the *mixed state presentation*, denoted $\mathcal{M}$. By construction, $\mathcal{M}$ is unifilar, even if M is not. Here, we show how this alternative works.

Nothing else said, the internal state evolution is given by Eq. (4). However, we are also interested in how a given state distribution $\Pr(\mathcal{S})$ changes when observing a particular symbol. This is denoted by the conditional probability $\Pr(\mathcal{S}|x)$. Using simple probability identities, we see that it is:

$$\begin{aligned}\Pr(\mathcal{S}|x) &= \frac{\Pr(\mathcal{S}, x)}{\Pr(x)} \\ &= \frac{\Pr(\mathcal{S}) \Pr(x|\mathcal{S})}{\Pr(x)} \ . \end{aligned}$$

In vector notation, we have the mixed state $\mu_t = \Pr(\mathcal{S})$. Then, if we observe $X_t = x$, we have more information about the update of internal states than assumed in Eq. (4). The result is that we transition to a new state distribution or mixed state $\mu_{t+1} \equiv \Pr(\mathcal{S}_{t+1}|X_t = x)$ given by:

$$\mu_{t+1} = \frac{\mu_t T^{(x)}}{\mu_t T^{(x)} \mathbf{1}} \ . \tag{9}$$

This is a row vector representing the normalized state probability. The numerator, though similar to how Eq. (4) evolves the state distribution, is only a partial distribution. The denominator normalizes that vector back to a distribution. Note that the denominators in this—$\mu_t T^{(x)} \mathbf{1}$—and the immediately preceding nonvector version—$\Pr(x)$—are related by Eq. (5).

That is, the probability to see x starting in mixed state μ_t was given by Eq. (5). Rewriting this slightly to indicate the current mixed state explicitly, we have:

$$\Pr(x|\mu_t) = \mu_t T^{(x)} \mathbf{1} \ .$$

In Eq. (6) we assumed π and obtained a stationary symbol probability. Here, the probability need not be stationary.

A key point is that due to unifilarity of the mixed state presentation, $\Pr(x)$ is also the *transition probability* of going from mixed state μ_t to mixed state μ_{t+1} on seeing x. That is,

$$\Pr(\mu_t \to_x \mu_{t+1}) = \mu_t T^{(x)} \mathbf{1} \ .$$

VIII. SYNOPSIS

Here's a reprise of the main mixed-state (vector) expressions. However, as just noted, one needs to be careful to specify on which state distribution we condition and the word that induces a mixed state distribution. And so, we introduce some slightly modified notation.

Evolve state distribution $\mathcal{S} \sim \mu$:

$$\mu_{t+1} = \mu_t T \ . \tag{10}$$

Word probability given a mixed state:

$$\Pr(X^L = w|\mathcal{S} \sim \mu) = \mu T^{(w)} \mathbf{1} \ . \tag{11}$$

Mixed state evolution induced by symbol x:

$$\mu_{t+1}(x) = \frac{\mu_t T^{(x)}}{\mu_t T^{(x)} \mathbf{1}} \ .$$

Mixed state transition on generating symbol x:

$$\Pr(\mu_t \to_x \mu_{t+1}) = \mu_t T^{(x)} \mathbf{1} \ .$$

ACKNOWLEDGMENTS

This work was partially supported by ARO grants W911NF-12-1-0288 and W911NF-12-1-0234.

[1] J. P. Crutchfield, C. J. Ellison, and J. R. Mahoney. Time's barbed arrow: Irreversibility, crypticity, and stored information. *Phys. Rev. Lett.*, 103(9):094101, 2009.

[2] C. J. Ellison, J. R. Mahoney, and J. P. Crutchfield. Prediction, retrodiction, and the amount of information stored in the present. *J. Stat. Phys.*, 136(6):1005–1034, 2009.

[3] J. R. Mahoney, C. J. Ellison, and J. P. Crutchfield. Information accessibility and cryptic processes. *J. Phys. A: Math. Theo.*, 42:362002, 2009.

[4] J. R. Mahoney, C. J. Ellison, and J. P. Crutchfield. Information accessibility and cryptic processes: Linear combinations of causal states. 2009. arxiv.org:0906.5099 [cond-mat].