Prediction and Generation of Binary Markov Processes: Can a Finite-State Fox Catch a Markov Mouse?

Joshua Ruebeck
Ryan G. James
John R. Mahoney
James P. Crutchfield

SFI WORKING PAPER: 2017-08-027

SFI Working Papers contain accounts of scientific work of the author(s) and do not necessarily represent the views of the Santa Fe Institute. We accept papers intended for publication in peer-reviewed journals or proceedings volumes, but not papers that have already appeared in print. Except for papers by our external faculty, papers must be based on work done at SFI, inspired by an invited visit to or collaboration at SFI, or funded by an SFI grant.

©NOTICE: This working paper is included by permission of the contributing author(s) as a means to ensure timely distribution of the scholarly and technical work on a non-commercial basis. Copyright and all rights therein are maintained by the author(s). It is understood that all persons copying this information will adhere to the terms and constraints invoked by each author’s copyright. These works may be reposted only with the explicit permission of the copyright holder.

www.santafe.edu
Prediction and Generation of Binary Markov Processes: Can a Finite-State Fox Catch a Markov Mouse?

Joshua Ruebeck

Department of Physics and Astronomy, Carleton College, One North College Street, Northfield, MN 55057

Ryan G. James,† John R. Mahoney,‡ and James P. Crutchfield§

Complexity Sciences Center and Physics Department, University of California at Davis, One Shields Avenue, Davis, CA 95616

(Dated: August 1, 2017)

Understanding the generative mechanism of a natural system is a vital component of the scientific method. Here, we investigate one of the fundamental steps toward this goal by presenting the minimal generator of an arbitrary binary Markov process. This is a class of processes whose predictive model is well known. Surprisingly, the generative model requires three distinct topologies for different regions of parameter space. We show that a previously proposed generator for a particular set of binary Markov processes is, in fact, not minimal. Our results shed the first quantitative light on the relative (minimal) costs of prediction and generation. We find, for instance, that the difference between prediction and generation is maximized when the process is approximately independently, identically distributed.

Keywords: stochastic process, hidden Markov model, ϵ-machine, causal states, mutual information.

PACS numbers: 89.75.Kd 89.70.+c 05.45.Tp 02.50.Ey 02.50.-r 02.50.Ga

I. INTRODUCTION

Imagine a mouse being chased by a fox. Survival suggests that the mouse should generate a path that is difficult for the fox to predict. We might imagine that the mouse brain is designed or trained to maximize the fox’s difficulty and, similarly, that the fox somehow has optimized the task of predicting the mouse’s path. Are these two tasks actually distinct? If so, do there exist escape paths that are easier to generate than predict? Every animal has limited computational resources and we might reasonably suppose that the mouse has fewer than the fox. Given that mice clearly continue to survive, we can ask whether this disparity in resources exists in tension with the disparity in task-complexity—path-generation versus path-prediction.

In lieu of mouse paths, we consider the space of discrete stationary stochastic processes—objects consisting of temporal sequences that span the range from perfectly ordered to completely random. We then frame resource questions quantitatively via hidden Markov model (HMM) representations of these processes. We focus on two particular HMM representations of any given process: the minimal predictive HMM—its computational mechanics’ ϵ-machine [1]—and its minimal generative HMM. We then find two primary measures of memory resource: $C_\mu$—defined as the ϵ-machine’s state-entropy—quantifies the cost of prediction, while $C_g$—the state entropy of the generative machine—quantifies the cost of generation. Introduced over two and a half decades ago, the ϵ-machine predictive representation is well studied and can be constructed for arbitrary processes [2]. The generative machine offers more challenges, as it involves a nonconvex constrained minimization over high-dimensional spaces. While there are several known bounds on $C_g$ and restrictions on the construction of generative HMMs [3–6], they have received significantly less attention than the predictive case and, as a consequence, are markedly less well understood.

The following presents the first construction of the minimal generators for an arbitrary stationary binary Markov process. This allows for the analytic calculation of $C_g$, as well as other properties of generative models. These models elucidate the differences between the tasks of generation and prediction. The techniques introduced here should also lead to minimal generators for other process classes.

II. MODELS

We represent stochastic processes using edge-emitting (Mealy) hidden Markov models (HMMs). Such a representation is specified by a set of states, a set of output
some instance of the past $x_0$, the task is to yield the exact conditional probability distribution $\Pr(X_{0:t}|x_0)$ for any length $t$.

### A. $\epsilon$-Machine Construction

The minimal predictive model of a process $P$ is known as its $\epsilon$-machine and its construction is straightforward. The theory of computational mechanics provides a framework for the detailed characterization of $\epsilon$-machines in topological and information-theoretic terms [1].

The kernel underlying this construction is the causal equivalence relation $\sim_\epsilon$. This is a relation over the set $\{x_0\}$ of semi-infinite pasts such that two pasts, $x_0$ and $x_0'$, belong to the same equivalence class if their conditional futures agree:

$$x_0 \sim_\epsilon x_0' \iff \Pr(X_0|x_0) = \Pr(X_0|x_0') .$$

Each equivalence class is a state of the system, encapsulating in minimal form the degree to which the past influences the future. Thus, we refer to the classes as causal states and denote by $S_t$ the causal state at time $t$. The memory required by the $\epsilon$-machine to implement the act of prediction is $C_\mu = H[S]$—the statistical complexity [1].

Then, transitions between these states follow directly from the equivalence relation:

$$T_{i,j}^k = \Pr(X_0 = k, S_1 = j|S_0 = i) .$$

As previously stated, the excess entropy $E$ is the amount of information shared between past and future. The causal equivalence relation induces a particular random variable $S$ that “captures” $E$. Importantly, $E$ is not itself the entropy of a random variable. Thus, the causal-state random variable cannot generally be of size $E$ bits. We might then think of the difference $\chi = C_\mu - E$, also known as the crypticity, as the predictive overhead [10]. It is an interesting fact that a nonzero predictive overhead $\chi$ is generic in the space of all processes.

### B. Binary Markov Processes

Let us now narrow our focus and construct the predictive models for the particular class of binary Markov
Markovity. This leads directly to the causal state is completely determined by the previous single symbol, a simple consequence of the process’ independence of the future given the past. Applying the causal equivalence relation, we find that its stationary state distribution is:

$$
\pi = \begin{bmatrix} 1 - q & 1 - p \\ 2 - p - q & 2 - p - q \end{bmatrix}.
$$

The informational properties of this class of processes—entropy rate, excess entropy, and statistical complexity—can be stated in closed form:

$$
\begin{align*}
 h_\mu &= \pi_A H(p) + \pi_B H(q), \\
 E &= H(\pi) - h_\mu, \\
 C_\mu &= E + h_\mu,
\end{align*}
$$

where \( H(p) = -(p \log p) - ((1 - p) \log (1 - p)) \) denotes Shannon’s binary entropy function \([8]\). The simple relation among these measures follows from the fact that any (nontrivial) binary Markov process is also equivalent to a spin chain—a restricted class of Markov chains \([10]\).

This class of binary Markov processes spans a variety of structured processes, summarized in Fig. 3. At the extremes of either \( p = 0 \) or \( q = 0 \), we have a period-1 (constant) process. If either \( p = 1 \) or \( q = 1 \), we have Golden Mean Processes, where \( 0s \) or \( 1s \), respectively, occur in isolation. If \( p = 1 - q \), the process loses its dependence on the prior symbol and it becomes a biased coin.

**IV. GENERATIVE MODELS**

Let’s now return to our original topic and describe the second type of process representation—generative models. The only requirement of a generative model is that it be able to correctly sample from the distribution \( \text{Pr}(X_0) \) over futures. More specifically, we require that, given any instance \( x_0 \) of the past, the generative model yields a next symbol \( X_0 \) with the same probability distribution \( \text{Pr}(X_0|X_0 = x_0) \) as specified by the process.

Note that, on the one hand, it may seem obvious that prediction subsumes generation. On the other, it is not so obvious how these two tasks might prefer different mechanisms.

Like the ε-machine causal state, a generative state \( R \) must also render past and future conditionally independent. Importantly, as a consequence of the causal equivalence relation ε-machines are unifilar which, when paired with their minimality, implies that the causal states are functions of the prior observables. Generative models, however, need not have this restriction. Consequently, a given sequence of past symbols (finite or semi-infinite) may induce more than one generative state.

Generative models are much less well understood than their predictive cousins. This is due in large part to the lack of constructive methods for working with and otherwise constructing them. This is why our results here, though addressing only on a relatively simple class of processes, mark a substantial step forward.
V. LöHR EXAMPLE

Let us now focus on a subclass of binary Markov processes, those for which $0 < p = q < 1/2$; refer to the orange line in Fig. 3. Reference [4] offers up a three-state HMM generator for this class, which we refer to as the Löhr model; see Fig. 4. We see from the HMM that when probability $p$ is near $1/2$, the process is nearly independent, identically distributed (IID). An IID process has only a single causal state and therefore zero statistical complexity, $C_\mu = 0$. However, for any deviation from $p = 1/2$, the statistical complexity is a full bit, $C_\mu = 1$. Why is it that a generator of a nearly IID process—that is, a nearly memoryless process—still needs a full bit of memory?

The motivation for constructing this three-state model is that it might concentrate the IID behavior into a single state and use the other states only for those infrequent deviations that “make up the difference”. And so, the state-entropy may be reduced even though there are three states instead of two. A priori it is not obvious that it is possible to yield the correct process in this construction. It is, however, straightforward to check that the Löhr model produces the correct conditional statistics. It is a generator of the processes. Note that in general it is sufficient to check these probabilities for all words of length $2N - 1$ where $N = \max(|S|, |R|)$. Of course, this result is useless on its own: $|X_0|$ and $|X_0|$ are generically infinite.

A recent result shows that the maximum number of states in an entropically minimal channel $Z$ is $|Z| \leq \min \{|X||Y|, 2^{\min\{|X||Y|\}} - 1\}$, where $X$ and $Y$ are the channel input and output processes [12]. Since a generative model is a form of communication channel from the past to the future, we find that the number of states of the minimal generative model is bounded by $|R| \leq \min \{|X_0||X_0|, 2^{\min\{|X_0||X_0|\}} - 1\}$. Of course, this result is useless on its own: $|X_0|$ and $|X_0|$ are generically infinite.

This bound can be made practical by combining the data processing inequality for exact common information $G[X : Y]$ [12] with the existence of the following two Markov chains [13]:

\[ X_0 - S^+ - S^- - X_0, \]

\[ S^+ - X_0 - X_0 - S^-. \]
We denote forward- and reverse-time causal states $S^+$ and $S^-$, respectively. Combined, these tell us that $G[X_0 : X_1] = G[S^+ : S^-]$. Therefore, the bound can be tightened to $|R| \leq \min\{2|S^+| + |S^-|, 2^{\min\{|S^+|, |S^-| \}} - 1\}$. This is a particularly helpful application of causal states.

**B. Binary Markov Chains**

In the particular case of processes represented by binary Markov chains, the reverse process is also represented by a binary Markov chain. And so, both $|S^+| = 2$ and $|S^-| = 2$. From the above bounds, we find that $|R| \leq 3$. Closely following the proof in Ref. [12], one can then show that no three-state representation is minimal. And, since a single state model can only represent IID processes, this leaves only models with two-states as the possible minimal representations.

We begin with the assumption that an observation $X_0$ maps stochastically to a state $R_0$, which then stochastically maps to a symbol $X_1$. Constraining this pair of channels to produce observations $X_0$ and $X_1$ consistent with the binary Markov chain yields the parametrized hidden Markov model found in Fig. 5. (Appendix III gives the background calculations.)

For each point $(p, q)$ in the binary Markov process-space, we now have a two-parameter model-space of HMMs, specified by $(\alpha, \beta)$. The constraint that conditional probabilities be between zero and one restricts our model-space parameters to a rectangle $\alpha \leq \min\{q, 1 - p\}$ and $\beta \geq \max\{q, 1 - p\}$. One can now compute the state entropy within this constrained model-space and identify the minima.

Since the entropy is concave in $\alpha$ and $\beta$ and the allowable regions in $(\alpha, \beta)$-space are convex (rectangles), it is sufficient to search for local minima along the boundary. Figure 6 illustrates this for three different points in process space. We find that at each of the points $(p = \ldots)$
$1/4, q = 1/2)$ and $(p = 1/2, q = 1/4)$, there is a single global minimum. For the point $(p = 3/4, q = 3/4)$, we find that there are two minima equivalent in value, but corresponding to nonisomorphic HMMs. Both representations are biased toward producing a periodic sequence, with fluctuations interjected at different phases of the period.

In this way, one can discover the minimal generator for any binary Markov chain. Examining these minimal topologies at each point, we find that process-space is divided into three triangular regions with topologically-distinct generators. This is a somewhat surprising contrast with the fact that this model class requires only one predictive topology.

Let us briefly return to the restricted process previously considered—the perturbed coin. We may now quantitatively compare the three state-entropies of interest. In Fig. 7 we see that the statistical complexity $C_\mu = 1$ everywhere, but at $p = 1/2$, where it vanishes, $C_\mu = 0$. The Löhr model’s state-entropy $C_L$ falls below $C_\mu$, but only for a subset of $p$ values. However, the generative complexity $C_g$ (a smooth function) is everywhere less than both $C_\mu$ and $C_L$. (The generative models for $p < 1/2$ and $p > 1/2$ are shown above.) This shows that the proposed Löhr model is not the generative model for any value of $p$.

As implied by the conditional independence requirement, the excess entropy $E$ remains a lower bound on each of these state-entropies. Löhr [4] constructed a tighter lower bound, denoted $L$, on any model of the perturbed coin. We see that $C_g$ is slightly larger than this bound, but only $L$ everywhere. $C_g$ may be useful to generalize this lower bound for other processes.

The minimal generators are defined over all of $(p, q)$-space. We can compare the cost $C_\mu$ of prediction with the cost $C_g$ of generation and the information necessarily captured by a model—the excess entropy $E$. This comparison is seen in Fig. 8.

Focusing on the upper two panels of Fig. 8, we see that both $C_\mu$ and $C_g$ display $p \leftrightarrow q$ symmetry. Furthermore, $C_g$ has a discontinuous derivative along this line of symmetry, but only in the southwest (SW).

For $C_\mu$, the line $p + q = 1$ is special in that this lines marks a causal-state collapse—two causal states merge into one under the equivalence relation. For $C_g$, however, this line marks a qualitative change in behavior (SW versus NE). Since the generative complexity is lower semi-continuous [3], we know that a predictive gap $C_\mu - C_g$ must exist around this line.

The lower two panels of Fig. 8 suggest that the costs of generation and of prediction may have different causes. The parameters for which $C_g - E$ is high are disjoint from those where $C_\mu - C_g$ is high. $C_g$ is high when $p$ and $q$ are correlated (near the $p$-$q$ symmetry line), but only for $p q < 1/2$. In the other half of parameter space, $C_g$ is high when $p$ and $q$ are anti-correlated and away from the causal collapse. In contrast, $C_\mu$ is high exclusively near the line of causal collapse.

VI. CONCLUSION

We presented the minimal generators of binary Markov stochastic processes. Curiously, the literature appears to contain no other examples of generative models, for processes with finite-state $\epsilon$-machines. And so, our contribution here is a substantial step forward. It allows us to begin to understand the difference between prediction and generation through direct calculation. It also opens these new models to analysis by a host of previously developed techniques including the information diagrams presented here.

To put the results in a larger setting, we note that HMMs have found application in many diverse settings, ranging from speech recognition to bioinformatics. And so, there are many reasons to care about the states and information-theoretic properties of these models, some
The concept of model state is central, for example, in model selection. A simple and common method for selecting one model over another is through application of a penalty related to the number of states (or entropy thereof) [14]. Since the predictive model will never have a lower entropy than the corresponding generative one, an entropic penalty should never yield the predictive model, however a state-number penalty might. Similarly, in model parameter inference, if one distinguishes between the predictive and generative classes, the maximum likelihood estimated parameters will differ between the two classes.

Finally, we close by drawing out the consequences for fundamental physics. Understanding states bears directly on thermodynamics. Landauer’s Principle states that erasing memory comes at a minimum, unavoidable cost—a heat dissipation proportional to the size of the memory erased [15]. One can consider HMMs as abstract representations of processes with memory (the state) that must be modified or erased as time progresses. Applying Landauer’s Principle assigns thermodynamic consequences to the HMM time evolution. Which HMM (and corresponding states) is appropriate, though? We now see that prediction and generation, two very natural tasks for a thermodynamic system to perform, actually deliver two different answers. It is important to understand how physical circumstances relate to this choice of task—it will be expressed in terms of heat.

**ACKNOWLEDGMENTS**

We thank the Santa Fe Institute for its hospitality during visits, where JPC is an External Faculty member. This material is based upon work supported by, or in part by, John Templeton Foundation grant 52095, Foundational Questions Institute grant FQXi-RFP-1609, and the U. S. Army Research Laboratory and the U. S. Army Research Office under contracts W911NF-13-1-0390 and W911NF-13-1-0340. JR was funded by the 2016 NSF Research Experience for Undergraduates program.


Supplementary Materials

Prediction and Generation of Binary Markov Processes: Can a Finite-State Fox Catch a Markov Mouse?
Joshua Ruebeck, Ryan G. James, John R. Mahoney, and James P. Crutchfield

These supplementary materials give additional information on the processes and models analyzed in the main body. They cover the relationship between causal states and Löhr states, process information diagram analysis, including crypticity and oracular information, and the derivation of the general, parametrized binary Markov process model itself.

I. CAUSAL STATES AND LÖHR STATES

Since the process considered in the main article is Markov order $R = 1$ and its causal states are in one-to-one correspondence with the symbol last seen, we can compactly represent the relation between pasts, causal states, and Löhr states. The mapping from causal states $A$ and $B$ (or observation symbols 0 and 1) to states $C$, $D$, and $E$ of the Löhr presentation is given by:

$$
\Pr(R|S) = \begin{cases} 
A/0 & \begin{pmatrix} C & D & E \\
1 - 2p & 2p & 0 \end{pmatrix} \\
B/1 & \begin{pmatrix} 0 & 2p & 1 - 2p \end{pmatrix}
\end{cases}
$$

(S1)

For example, if the last symbol was $x = 0$, this induces causal state $A$. The corresponding Löhr states occur with respective probabilities $\Pr(C, D, E) = (1 - 2p, 2p, 0)$.

II. INFORMATION DIAGRAM ANALYSIS

The information diagram [7] is a tool that has become increasingly useful for analyzing and characterizing the information content of process presentations [13]. It gives a visual representation of multivariate information-theoretic dependencies. It arises from a duality between information measures and set theory: set-intersection corresponds to mutual information and set-union corresponds to joining distributions. Here, we analyze several information measures introduced in Ref. [9] for the processes considered in the main article; the associated I-diagram is depicted in Fig. 9.

A generic model state $R$ has nonzero intersection (mutual information) with both the process’ past $X_0$ and future $X_0$: $I[R : X_0] > 0$ and $I[R : X_0] > 0$. Importantly, the state information of a model that correctly generates a process—a process presentation—must intersect the past-future union, which completely contains the past-future intersection. We might say that the state of the model “captures” this atom—the excess entropy $E = I[X_0 : X_0]$. Any model that does not capture at least $E$ of the past and future cannot correctly generate the process. (Recall that if a model is predictive, then it is also generative.)

The $\epsilon$-machine is simple not only in the sense that it is constructible, but it is informationally simple as it appears in the information diagram as well. This simplicity follows from the fact that causal states are functions of the past. As a result, $\epsilon$-machine state information has no intersection with the future beyond that of that past’s—it contains no, what we call, oracular information $\zeta = I[R : X_0|X_0]$. It also contains no information outside the past-future union—it contains not gauge information $\phi = H[R|X_0,X_0]$. Beyond the excess entropy atom, it has only one potential region—what we call the process’ crypticity $\chi_{\epsilon\text{-machine}} = C_\mu - E$, where $C_\mu = H[S]$ is the process’ statistical complexity.

The situation is richer for general representations, including generative ones; denote them $g$. The more general definition of crypticity is $\chi_g = I[R : X_0|X_0]$. A general representation may have oracular and gauge information. And, it may have a crypticity greater than, equal to, or less than the $\epsilon$-machine crypticity.

Generative models are restricted in two ways. First, their crypticity is never greater than the $\epsilon$-machine’s crypticity: $\chi_g \leq \chi_{\epsilon\text{-machine}}$. This follows straightforwardly since if it were greater, then the $\epsilon$-machine would necessarily have smaller state entropy, leading to a contradiction. Further, the sum of the generative model’s crypticity, gauge information, and oracular information must be less than...
FIG. 10. Decomposition of the generative state information $C_g$ into excess entropy $E$, crypticity $\chi$, oracular information $\zeta$, and gauge information $\varphi$ over process space $(p, q)$.

The $\epsilon$-machine’s crypticity:

$$\chi_g + \varphi_g + \zeta_g \leq \chi_{eM} \quad (S2)$$

Second, appealing to our newly introduced generative models for binary Markov chains, we can compare the generative-model crypticity to that of the $\epsilon$-machine. A generative model distinct from the $\epsilon$-machine must have nonzero oracular information [16]. Also, a generative model may only have nonzero gauge information if it has both crypticity and oracular information [16]:

$$\varphi_g > 0 \implies \begin{cases} \chi_g > 0 \\ \zeta_g > 0 \end{cases} \quad (S3)$$

while still satisfying Eq. (S2). Effectively, this means that gauge information can be “minimized” away, unless it supports the existence of both cryptic and oracular information.

For our parametrized generative models of binary Markov chains we find that $C_g$ decomposes as shown in the two mosaics in Figs. 10 and 11. The gauge information $\varphi_g$ is, generally, a rather small portion of $C_g$, though it is largest in the same regions as $C_g - E$. Both the crypticity $\chi_g$ and the oracular information $\zeta_g$ are large along the $q = 1$ and $p = 1$ edges, respectively. This implies that $\mathcal{R}$ “leans left” when $q \approx 1$ and “leans right” when $p \approx 1$. This relationship flips if we use the alternative, equivalent model in the $p > 1 - q$ half of the process space.

Note that while gauge information is not required of generative models, it is present in all of the binary Markov chain generators, except along boundary and causal collapse lines. Surveys of other processes suggest that the presence of gauge information may be somewhat rare.

III. PARAMETRIZED BINARY MARKOV PROCESS HMM: DERIVATION

We derive the most generic parametrized, two-state HMM of binary Markov processes. The target binary Markov process is represented by the conditional probability matrix:

$$\Pr(X_t|X_{t-1}) = \begin{pmatrix} 0 & 1 \\ 1 - p & p \end{pmatrix} \quad (S4)$$

As an intermediate step to deriving the full HMM, we find the conditional probabilities involving the machine states relative to the symbols emitted on the transitions to and from a state. We write down two template matrices, denoting the states $A$ and $B$ and the random variable $q = 1/2$, $q = 3/4$. P = q = 1/4.
FIG. 12. As an intermediate step, we construct a skeleton for the fully general HMM. r, s, t, u are temporary variables to be solved for eventually. The other transitions are labeled using the fact that, for example, the probabilities of the paths coming from state A and emitting a 0 must sum to α in order to agree with Eq. S5.

over them \(\mathcal{R}\):

\[
\Pr(X_t|\mathcal{R}) = \begin{bmatrix} 0 & 1 \\ A & B \end{bmatrix} \begin{bmatrix} \alpha & 1 - \alpha \\ \beta & 1 - \beta \end{bmatrix}
\]

and:

\[
\Pr(\mathcal{R}|X_{t-1}) = \begin{bmatrix} 0 & 1 \\ A & B \end{bmatrix} \begin{bmatrix} \gamma & 1 - \gamma \\ \delta & 1 - \delta \end{bmatrix}
\] (S6)

there is ambiguity here in labeling the states. If we switch the rows of Eq. (S5) with each other and the columns of Eq. (S6) with each other, then we obtain different matrices that describe the same model. Only the the state labels have been swapped. To avoid double-counting such machines, we restrict \(\alpha \leq \beta\).

Multiplying out this factorization and requiring it to be consistent with the target provides two constraints:

\[
\gamma = \frac{\beta + p - 1}{\beta - \alpha}, \quad \delta = \frac{\beta - q}{\beta - \alpha}.
\] (S7)

(We keep the denominators as \(\beta - \alpha \) so that both the numerator and denominator are positive.) We ask that all these be valid probabilities. Requiring \(0 \leq \gamma, \delta \leq 1\) and substituting Eq. (S7) yields the constraints:

\[
\alpha \leq \min\{q, 1 - p\}, \quad \beta \geq \max\{q, 1 - p\}.
\]

Now, consider the incomplete HMM that defines the helper variables \(r, s, t, u\) in Fig. 12. To determine the latter’s values, we evaluate the probabilities of a string (a) being generated by this incomplete HMM:

\[
\Pr(01) = \Pr(01|A) \Pr(A) + \Pr(01|B) \Pr(B)
\]

\[
= [s(1 - \alpha) + (\alpha - s)(1 - \beta)] \Pr(A)
\]

\[
+ [(\beta - r)(1 - \alpha) + r(1 - \beta)] \Pr(B) \quad (S8)
\]

and (b) being generated by the Markov process described in Eq. (S4):

\[
\Pr(01) = \Pr(1|0) \Pr(0)
\]

\[
= p \Pr(0). \quad (S9)
\]

We used the string 01 as our first case. And, \(\Pr(0), \Pr(1), \Pr(A), \) and \(\Pr(B)\) are given by the stationary distributions over the symbols and states, respectively:

\[
\Pr(0) = \frac{q}{p + q}
\]

\[
\Pr(1) = \frac{p}{p + q}
\]

\[
\Pr(A) = \frac{q\gamma + p\delta}{p + q}
\]

\[
\Pr(B) = \frac{q(1 - \gamma) + p(1 - \delta)}{p + q}.
\]

Setting Eqs. (S8) and Eq. (S9) equal to one another constrains \(r, s, t, u\). To fully specify all four, though, we need three more independent equations. We obtain these by evaluating the probabilities of the strings 11, 110, and 000 in a similar fashion. Respectively, these yield:

\[
(1 - q) \Pr(1) = [u(1 - \alpha) + (1 - \alpha - u)(1 - \beta)] \Pr(A)
\]

\[
+ [t(1 - \beta) + (1 - \beta - t)(1 - \alpha)] \Pr(B),
\]

\[
(1 - q) \Pr(1) = [u((1 - \alpha) - u)\beta + u\alpha
\]

\[
+ (1 - \alpha - u)(t\beta + (1 - \beta - t)\alpha)] \Pr(A)
\]

\[
+ [t(t\beta + (1 - \beta - t)\alpha
\]

\[
+ (1 - \beta - t)((1 - \alpha - u)\beta + u\alpha)] \Pr(B), \quad (1 - p)^2 \Pr(0) = [s((\alpha - s)\beta + s\alpha
\]

\[
+ (\alpha - s)(r\beta + (\beta - r)\alpha)] \Pr(A)
\]

\[
+ [r((\beta - r)\alpha + r\beta
\]

\[
+ (\beta - r)((\alpha - s)\beta + s\alpha)] \Pr(B).
\]
Solving this system yields:

\[ r = \frac{\beta(1 - p - \alpha)}{\beta - \alpha}, \]
\[ s = \frac{\alpha(\beta + p - 1)}{\beta - \alpha}, \]
\[ t = \frac{(1 - \beta)(q - \alpha)}{\beta - \alpha}, \text{ and} \]
\[ u = \frac{(1 - \alpha)(\beta - q)}{\beta - \alpha}. \]

Finally, we substitute these into Fig. 12 to recover the parametrized HMM the main article introduced in Fig. 5.