Computational mechanics of "syncable" nonunifilar Hidden Markov Models

Sarah Marzen
Physics Department, U.C. Berkeley
smarzen@berkeley.edu

June 21, 2013

Abstract

The biggest obstacle to our ability to make sense of quantitative biological data is our inability to infer large underlying state spaces from highly subsampled data, i.e. inference of nonunifilar Hidden Markov Models. The epsilon machine presentation and/or information theoretic quantities easily calculable using the unifilar epsilon machine presentation might be useful additions to the traditional nonunifilar HMM inference toolbox. I calculate the epsilon-machine presentations in forward and reverse time of some simple syncable nonunifilar word generators, where I use syncable to mean that there is a direct mapping from observed symbols to internal states for observation of all but one of the letters. All of the word generators investigated here have a countable infinity of causal states but finite statistical complexities in both forward and reverse time. I show that, as might have been expected, it is impossible to infer the number of hidden states if all the hidden states are identical; however, some preliminary results suggest that it may be possible to infer the number of hidden states if something is known about how the hidden states' transition probabilities are generated. I also show that the syncable binary word generators considered here are all causally reversible, but preliminary results suggest that the typical syncable word generators that emit three or more letters are causally irreversible.

1 Introduction

At present, there are essentially two disjoint quantitative modeling approaches to understanding neurobiology. The first approach involves painstakingly modeling the detailed components of many neurons and throwing everything and the kitchen sink into a gargantuan simulation, which is then run for hours, days or weeks on a supercomputer. This, for instance, is the approach pursued in Europe's Human Brain Project. This approach is quite useful for answering the questions of *how* neurons fire and *how* neurons increase or decrease connection strength, but this approach makes it difficult to understand *why* any particular set of neurons in any particular part of the brain would connect in some particular way. In other words, whole-scale simulations of cortical regions bury the question of what computation the brain is performing in the details of the simulation and hope that computation will "emerge" from their framework. So far, the state-of-the-art large-scale cortical simulations can say that the brain is chaotic and that alpha and gamma waves exist [1]. Both are true statements, but it is difficult to understand why the simulations produced these results and what purpose the emergent phenomena serve from the simulations alone.

The alternative approach to quantitative modeling of neurobiological data involves studying minimal models that are consistent with some subset of gathered data. These approaches generate tractable models that replicate only some of the available data in hopes that these minimal models have captured the essence of the brain's computation. Underlying such modeling approaches is the assumption that microscopic details don't matter in the same way that critical exponents in statistical physics are functions solely of the system's symmetries. Some notable examples of this approach to neurobiological modeling include Maximum Entropy (MaxEnt) models of neuronal spiking [2, 3], sparse coding [4], and maximization of mutual information (InfoMax) models of receptive fields [5, 6]. There is, of course, an obvious flaw to this class of models, which is that the usefulness of this model is completely determined by which gathered data is taken to be relevant. In MaxEnt approaches, the constraints constitute a value judgement on what data is relevant, and emphasis on matching pairwise correlations yields MaxEnt models that do not capture local cortical computation [7] and do not have sparse neuronal activity [8]. In InfoMax approaches, the specification of input and output constitute value judgements on which input is relevant and which output is readable. Approaches like sparse coding are more direct attempts at manifold learning of the input via a union of planes. However, in trying to make an equivalence between the sparse coding models and brain function, there is still a value judgement in order to constrain an underconstrained learning problem– namely, there is a value judgement that the neuronal activity be sparse. And sparse coding and InfoMax are still limited to predicting receptive fields, which are highly stimulus dependent [9] and therefore unlikely to be the "right" relevant variable for prediction. In other words, the applicability of these models is inherently limited by the prior intuitions of the modeler about computation in the brain.

In short, there are two approaches to neurobiological models, with two disjoint sets of pros and cons and no obvious bridge between the two. What the Brain Activity Map needs in order to succeed is some bridge between the two approaches that allows one to leverage the tractability and intuition provided by minimal models without assuming that the minimal modeler has to know, in advance, the computation being performed by the brain. The ϵ-Machine is a user-proof minimal model of stationary time-series, and as far as I know, it is the *only* user-proof minimal model of stationary time-series. As such, it is entirely worth investigating how to infer ϵ-Machines and how to recognize the nonunifilar word generator that produced a particular (unifilar) ϵ-Machine, essentially unpacking an observed time series into its causal states and repacking those causal states into a "minimal nonunifilar word generator" that might be isomorphic to the underlying system.

In some underlying state space (which could involve many higher order derivatives of the more natural state space variable), the observed data is generated by a nonunifilar Hidden Markov Model (HMM). Therefore, in this report, I considered a class of nonunifilar HMMs with particularly tractable epsilon-machine presentations. These nonunifilar HMMs were generated by taking some underlying state space and partitioning the underlying states into m groups, $m-1$ of which consisted of a single underlying state. The last group contained all remaining underlying states. Transitions to group k resulted in emission of the letter k. The ϵ-Machine presentations of these word generators were particularly simple because these word generators were "syncable", in that seeing all but one of the letters would result in syncing to one of the internal states.

The report's layout is as follows. In Section 2, I go through the example of the already studied simple nonunifilar source. In Section 3, I introduce a variation on the simple nonunifilar source that allow me to flirt with extensions to continuous time in Section 3.1. In Section 4, I consider binary nonunifilar HMMs that are syncable in the sense described in the paragraph above. In Section 4.1, I again flirt with continuous time; in Section 4.2, I show that inference of the number of hidden underlying states is impossible if all hidden underlying states are identical; and in Section 4.3, I prove that these binary syncable nonunifilar HMMs are causally reversible. In Section 5, I provide formulas for statistical complexity and entropy rate for syncable nonunifilar HMMs that emit at least three or more symbols. Interestingly, it seems that these syncable nonunifilar HMMs are causally irreversible in general, unlike their binary counterparts.

2 Simple nonunifilar source

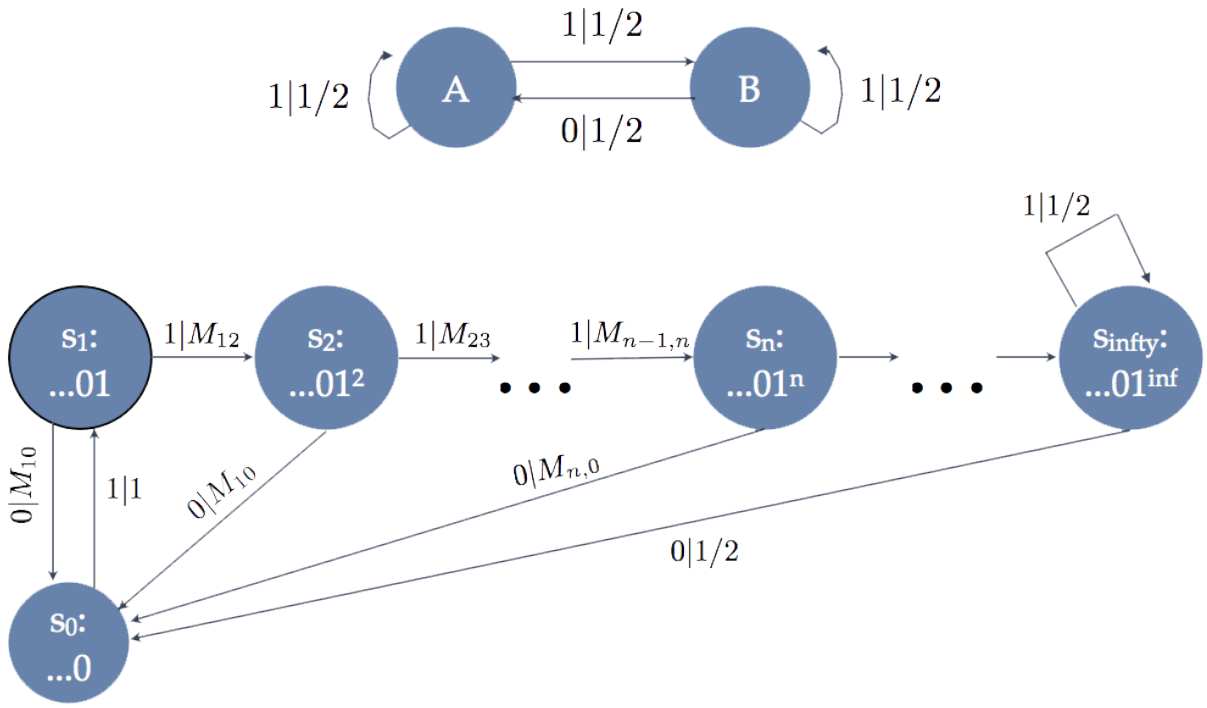

Figure 1: At top is the nonunifilar HMM and at bottom is its epsilon machine presentation.

See Figure 1. The causal states of the SNS are denoted here as s_i with $i = 0, ...$ The causal state s_0 captures all histories that end in a 0 (are on state A) and the causal state s_i, $i \geq 1$, captures all histories that end in 01^i. Each of these are causal states because until a 0 is seen, nothing is synced– we could be in state A or in state A or B with varying mixed state probabilities. The transition probabilities between causal states are given as follows. Since a 0 is always followed by a 1,

$$M^{(1)}_{s_1,s_0} = p(s_0 \to_1 s_1) = 1, \;\; M^{(0)}_{s_1,s_0} = p(s_0 \to_0 s_1) = 0. \tag{1}$$

Causal state s_i can only be followed by s_{i+1} (if a 1 is emitted) and s_0 (if a 0 is a emitted). The transition probabilities are found via the usual

$$M^{(1)}_{s_n,s_{n-1}} = p(s_{n-1} \to_1 s_n) = \frac{p(...01^n)}{p(...01^{n-1})} = \frac{1^\top \left(T^{(1)}\right)^n T^{(0)} \pi}{1^\top \left(T^{(1)}\right)^{n-1} T^{(0)} \pi}, \tag{2}$$

where

$$\pi = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} \tag{3}$$

and

$$T^{(0)} = \begin{pmatrix} 0 & \frac{1}{2} \\ 0 & 0 \end{pmatrix},\ T^{(1)} = \begin{pmatrix} \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}. \tag{4}$$

This gives

$$M^{(1)}_{s_n,s_{n-1}} = \frac{(n+1)/2^n}{n/2^{n-1}} = \frac{1}{2}\frac{n+1}{n} = \frac{1}{2}\left(1+\frac{1}{n}\right). \tag{5}$$

It follows that

$$M^{(0)}_{s_n,s_0} = 1 - M^{(1)}_{s_n,s_{n-1}} = \frac{1}{2}\left(1-\frac{1}{n}\right). \tag{6}$$

This constitutes a complete characterization of the forward epsilon machine of the SNS.

2.1 Calculation of stationary distribution and statistical complexity of the forward epsilon machine

The stationary distribution has that the probability flowing into each of the causal states is equivalent to the probability flowing out. For causal state s_n with $n \geq 2$, as probability can flow into s_n from s_{n-1} only and probability always flows out (either to s_{n+1} or s_0), this yields the equation

$$p(s_n|s_{n-1})\pi_{n-1} = \pi_n \rightarrow \frac{\pi_n}{\pi_{n-1}} = \frac{1}{2}\left(\frac{n+1}{n}\right). \tag{7}$$

This recursive equation gives a simple expression for the probability π_n in terms of π_1:

$$\frac{\pi_n}{\pi_1} = \prod_{i=2}^{n}\frac{\pi_i}{\pi_{i-1}} = \prod_{i=2}^{n}\frac{1}{2}\left(\frac{i+1}{i}\right) = \frac{1}{2^{n-1}} \times \frac{n+1}{2} = \frac{n+1}{2^n}. \tag{8}$$

Balancing the probability flow into and out of causal state s_1 is given similarly by

$$p(s_0 \rightarrow_1 s_1)\pi_0 = \pi_1 \rightarrow \pi_1 = \pi_0. \tag{9}$$

Finally normalization of probabilities means

$$\sum_{i=0}^{\infty}\pi_i = 1 \rightarrow \pi_1 + \pi_1 \cdot \sum_{i=1}^{\infty}\frac{i+1}{2^i} = 1. \tag{10}$$

The infinite sum is easily found via

$$\begin{aligned}
\sum_{i=1}^{\infty}\frac{i+1}{2^i} &= \sum_{i=1}^{\infty}2^{-i} + \sum_{i=1}^{\infty}i2^{-i} && (11)\\
&= \sum_{i=1}^{\infty}2^{-i} + \sum_{i=1}^{\infty}2^{-i} + \sum_{i=2}^{\infty}2^{-i} + ... && (12)\\
&= \sum_{i=1}^{\infty}2^{-i} + \sum_{j=1}^{\infty}\sum_{i=j}^{\infty}2^{-i} && (13)\\
&= 1 + \sum_{j=1}^{\infty}\frac{2^{-j}}{1-\frac{1}{2}} && (14)\\
&= 1 + 2\sum_{j=1}^{\infty}2^{-j} = 1 + 2(1) = 3. && (15)
\end{aligned}$$

Therefore,

$$\pi_1 = \pi_0 = \frac{1}{4},\ \pi_n = \frac{n+1}{4 \cdot 2^n}. \tag{16}$$

Finally, just as a consistency check, probability flow into and out of causal state s_0 must balance, yielding

$$\pi_0 = \sum_{i=1}^{\infty} p(s_i \to_0 s_0)\pi_i = \sum_{i=1}^{\infty} \frac{1}{2}\left(1 - \frac{1}{i+1}\right) \times \frac{i+1}{2^i}\pi_1 \to 1 = \frac{1}{2}\sum_{i=1}^{\infty} \frac{i}{2^i}. \tag{17}$$

This equality is indeed true. The resulting statistical complexity is just (with the aid of Mathematica's summation function so as to avoid fucking up)

$$\begin{aligned} C_\mu &= H[\pi_0, \pi_1, ...] = -\sum_{i=0}^{\infty} \pi_i \log_2 \pi_i & (18) \\ &= \frac{1}{2} + \frac{7}{2} - 1.28853 = 2.71147 \text{ bits.} & (19) \end{aligned}$$

2.2 Calculation of $h_\mu(L)$ for the forward epsilon machine

We can easily calculate the entropy rate h_μ from the unifilar presentation as

$$h_\mu = \sum_i h_{s_i} \pi_i. \tag{20}$$

The entropy rate for causal state s_0 is $H[1] = 0$; the entropy rate for causal state s_i, $i \geq 1$, is $H[\frac{1}{2}\left(1 + \frac{1}{i+1}\right)] = H[\frac{i}{2(i+1)}]$. Thus

$$\begin{aligned} h_\mu &= \sum_{i=1}^{\infty} H[\frac{i}{2(i+1)}]\frac{i+1}{4 \cdot 2^i} & (21) \\ &= -\sum_{i=1}^{\infty} \left(\frac{i}{2(i+1)} \log_2 \frac{i}{2(i+1)} + \frac{i+2}{2(i+1)} \log_2 \frac{i+2}{2(i+1)} \right) \times \frac{i+1}{4 \cdot 2^i} & (22) \\ &\simeq 0.678 \text{ bits.} & (23) \end{aligned}$$

Our goal, however, is to calculate $h_\mu(L)$ from causal shielding arguments. From an explicit formula for $h_\mu(L)$ we will be able to calculate E as the sum of $h_\mu(L) - h_\mu$, for instance.

The formula for $h_\mu(L)$ from the mixed state presentation or (in this case, as they are equivalent) causal states is

$$h_\mu(L) = H[X_L | R_L, R_0 = \mu_0]. \tag{24}$$

Here, the initial mixed state μ_0 is that all the weight is on s_1; R_L is obtained by multiplying μ_0 by the transition matrix over causal state space L times. Each causal state has a particular entropy rate given by the entropy of its outgoing transition probabilities. Superficially, it seems impossible that we would be able to calculate $h_\mu(L)$ even using causal shielding because there are a countable infinity of causal states. However, the number of populated causal states grows only linearly with the number of iterations of our dynamics, as s_n is only connected to s_0 and s_{n+1}. As such, if we want to calculate $h_\mu(L)$, we only need to consider the transition matrix over causal states $s_i,\ i = 0, ..., L+1$:

$$h_\mu(L) = h^\top M^L_{L+1 \times L+1} \mu_0 \tag{25}$$

where $M = M^{(1)} + M^{(0)}$ given previously and h is a list of the entropy generation of each causal state, $h(s_n) = H[\frac{n-1}{2n}]$. Matlab code presented in the Appendix can calculate $h_\mu(L)$ for various L. The figure below shows that $h_\mu(L)$ rapidly decreases as a function of L to its asymptotic value given in eqn. 23.

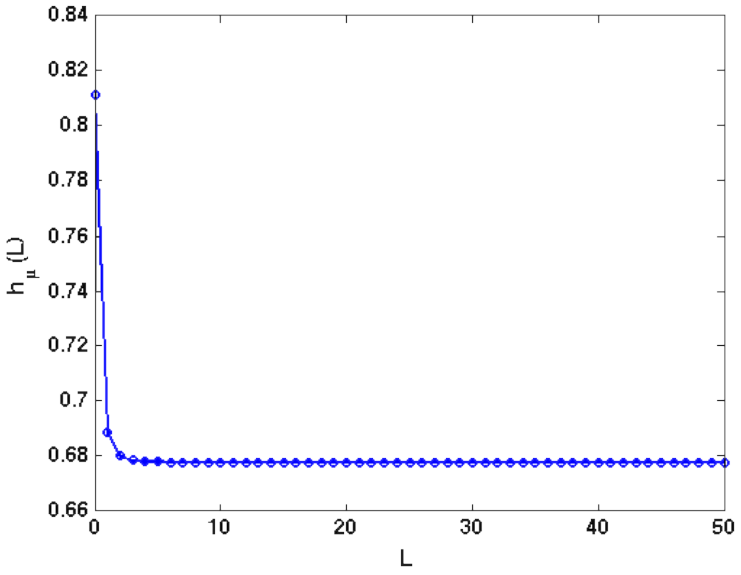

Figure 2: Entropy rate estimates of the simple nonunifilar source from Matlab code listed above. Note that the x-axis starts at $L = 0$, corresponding to my definition that $h_\mu(L) = H(L+1) - H(L)$ (sorry) and also note the quick decay to the entropy rate given in eqn. 23. Estimate of $E = 0.14723$ bits, giving a forward crypticity estimate of $\chi^+ = C_\mu^+ - E = 2.5642$ bits.

2.3 Time-reversed epsilon machine

Calculation of the time reversed epsilon machine can proceed by first calculating the time-reversed nonunifilar presentation of the SNS, and then generating the corresponding epsilon-machine. The time-reversed nonunifilar presentation is clearly isomorphic to its forward-time nonunifilar presentation, modulo switching the labels of the states A and B. This can be seen from the equations

$$\tilde{T}^{(1)}_{yx} = T^{(1)}_{xy} \frac{\pi_y}{\tilde{\pi}_x} = \frac{1}{2}\delta_{xy\neq BA} \times \frac{1/2}{1/2} = \frac{1}{2}\delta_{yx\neq AB} \tag{26}$$

and

$$\tilde{T}^{(0)}_{yx} = T^{(0)}_{xy} \frac{\pi_y}{\tilde{\pi}_x} = \frac{1}{2}\delta_{xy=BA} \times \frac{1/2}{1/2} = \frac{1}{2}\delta_{yx=AB}. \tag{27}$$

Therefore, the reverse-time epsilon machine is equivalent to that of the forward epsilon machine. Both have statistical complexity $C_\mu^- = C_\mu^+ = 2.71147$ bits, making this a causally reversible process:

$$\chi^+ = C_\mu^+ - E = 2.5642 \text{ bits} = C_\mu^- - E = \chi^- \Rightarrow \Xi = \Delta\chi = 0. \tag{28}$$

2.4 Bidirectional epsilon machine

Time reversing M^+ was quite easy, but unifilarizing it was not, and so I will have to leave this calculation to someone else with more patience.

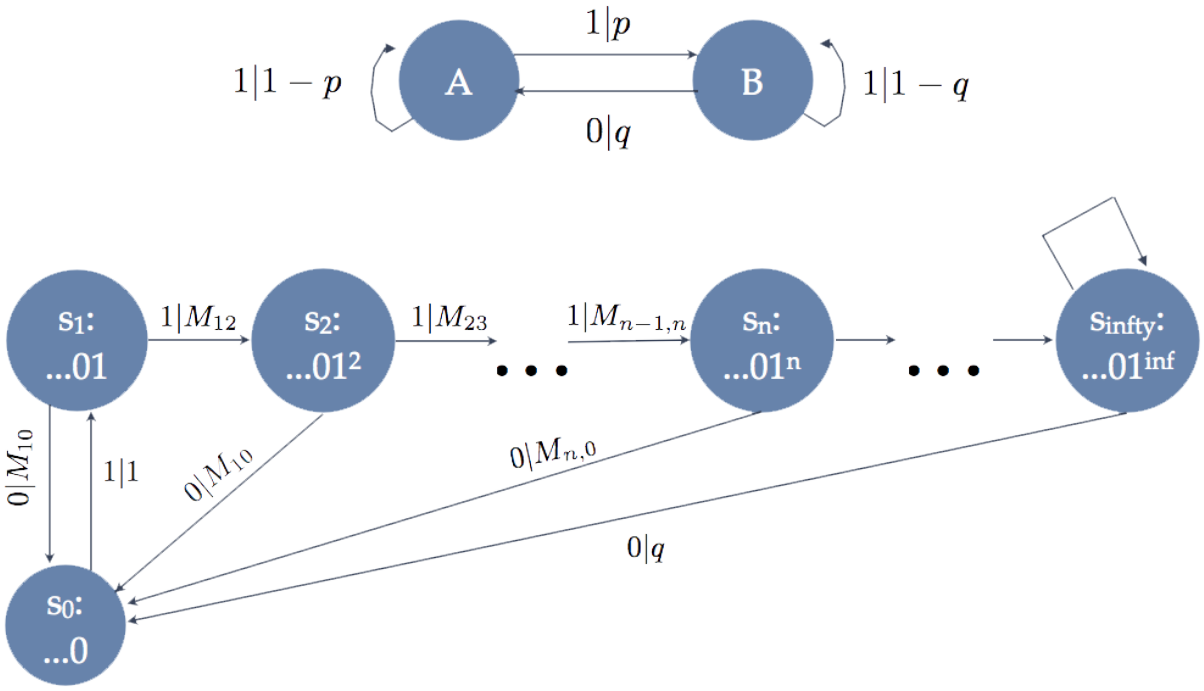

Figure 3: At top is the nonunifilar HMM, variation on the simple nonunifilar source, and at bottom is its epsilon machine presentation. p and q are adjustable parameters.

3 Variation on the simple nonunifilar source

Now we investigate a slight variation on the simple nonunifilar source in which the transition probabilities are adjustable:

$$T^{(1)} = \begin{pmatrix} 1-p & 0 \\ p & 1-q \end{pmatrix}, \; T^{(0)} = \begin{pmatrix} 0 & q \\ 0 & 0 \end{pmatrix}. \tag{29}$$

See Figure 3. The corresponding stationary distribution over the original states is

$$\pi = \begin{pmatrix} \frac{q}{p+q} \\ \frac{p}{p+q} \end{pmatrix}. \tag{30}$$

The mixed state presentation and the causal states, just as in the nonunifilar presentation, can be identified as s_0 (all histories that end in a 0, syncing to state A in the nonunifilar presentation) and s_i (all histories that end in a 01^i, some mix of probability over state A and B in the nonunifilar presentation) with $i \geq 1$. Then, similar to the simple nonunifilar source, we find that[1]

$$M^{(x)}_{n-1,n} = \delta_{x,1} \frac{1^\top \left(T^{(1)}\right)^n T^{(0)} \pi}{1^\top \left(T^{(1)}\right)^{n-1} T^{(0)} \pi} = \delta_{x,1} \frac{p(1-q)^n - q(1-p)^n}{p(1-q)^{n-1} - q(1-p)^{n-1}} \tag{31}$$

and

$$M^{(x)}_{n,0} = \delta_{x,0} \frac{1^\top T^{(0)} \left(T^{(1)}\right)^n T^{(0)} \pi}{1^\top \left(T^{(1)}\right)^n T^{(0)} \pi} = \delta_{x,0} \frac{pq\left((1-q)^n + (1-p)^n\right)}{p(1-q)^n - q(1-p)^n} \tag{32}$$

[1]Found with the aid of Mathematica, but could just has easily been done using matrix diagonalization.

for $n \geq 1$. Causal state s_0 transitions only to s_1 and emits a 1. This implies, using steps similar to those introduced for the simple nonunifilar source, that the stationary distribution on causal states is

$$\pi_n = \frac{\prod_{k=2}^{n} M_{k-1,k}}{2 + \sum_{j=2}^{\infty} \prod_{k=2}^{j} M_{k-1,k}}, \ n \geq 2 \tag{33}$$

and

$$\pi_1 = \pi_0 = \frac{1}{2 + \sum_{j=2}^{\infty} \prod_{k=2}^{j} M_{k-1,k}}. \tag{34}$$

These expressions can be simplified by noting that

$$\prod_{k=2}^{j} M_{k-1,k} = \prod_{k=2}^{j} \frac{1^\top \left(T^{(1)}\right)^k T^{(0)} \pi}{1^\top \left(T^{(1)}\right)^{k-1} T^{(0)} \pi} = \frac{1^\top \left(T^{(1)}\right)^j T^{(0)} \pi}{1^\top T^{(1)} T^{(0)} \pi} = \frac{p\,(1-q)^j - q\,(1-p)^j}{p-q} \tag{35}$$

and therefore

$$\pi_0 = \frac{1}{2 + \sum_{j=2}^{\infty} \frac{p(1-q)^j - q(1-p^j)}{p-q}} = \frac{1}{2 + \frac{p}{p-q}\frac{(1-q)^2}{1-(1-q)} - \frac{q}{p-q}\frac{(1-p)^2}{1-(1-p)}} = \frac{1}{2 + \frac{p+q}{pq} - 2} = \frac{pq}{p+q}, \tag{36}$$

which is akin to the number of transitions in a unit time step made from state B. This also gives

$$\pi_n = \frac{p\,(1-q)^n - q\,(1-p)^n}{p-q} \times \frac{pq}{p+q}. \tag{37}$$

From this, we can calculate the statistical complexity:

$$C_\mu^+(p,q) = -\sum_{n=0}^{\infty} \left(\frac{p\,(1-q)^n - q\,(1-p)^n}{p-q} \times \frac{pq}{p+q} \right) \log_2 \left(\frac{p\,(1-q)^n - q\,(1-p)^n}{p-q} \times \frac{pq}{p+q} \right). \tag{38}$$

The figure below shows $C_\mu^+(p,q)$ for various values of p and q. It is similarly easy to calculate the entropy rate by noting that the entropy accorded to causal state s_n is just $H[\frac{pq((1-q)^n + (1-p)^n)}{p(1-q)^n - q(1-p)^n}]$, which means that

$$h_\mu(p,q) = \sum_{n=0}^{\infty} \left(H[\frac{pq\left((1-q)^n + (1-p)^n\right)}{p(1-q)^n - q(1-p)^n}] \right) \left(\frac{p\,(1-q)^n - q\,(1-p)^n}{p-q} \times \frac{pq}{p+q} \right). \tag{39}$$

The entropy rate as a function of p and q is shown also in the figure below. We can easily find the statistical complexity of the reverse time process by time-reversing the nonunifilar presentation. Then, the transition from B to A has probability

$$\tilde{T}_{BA}^{(x)} = \delta_{x,1} T_{AB} \frac{\pi_A}{\pi_B} = \delta_{x,1} p \ \frac{q/p + q}{p/p + q} = q\delta_{x,1} \tag{40}$$

and the transition from A to B has probability

$$\tilde{T}_{AB}^{(x)} = \delta_{x,0} T_{BA} \frac{\pi_B}{\pi_A} = \delta_{x,0} q \ \frac{p/p + q}{q/p + q} = p\delta_{x,0}. \tag{41}$$

In other words, the reverse time nonunifilar presentation merely switches the labels on states A and B. This implies that

$$C_\mu^-(p,q) = C_\mu^+(q,p) \tag{42}$$

and therefore

$$\Xi(p,q) = C_\mu^+(p,q) - C_\mu^-(p,q) = C_\mu^+(p,q) - C_\mu^+(q,p). \tag{43}$$

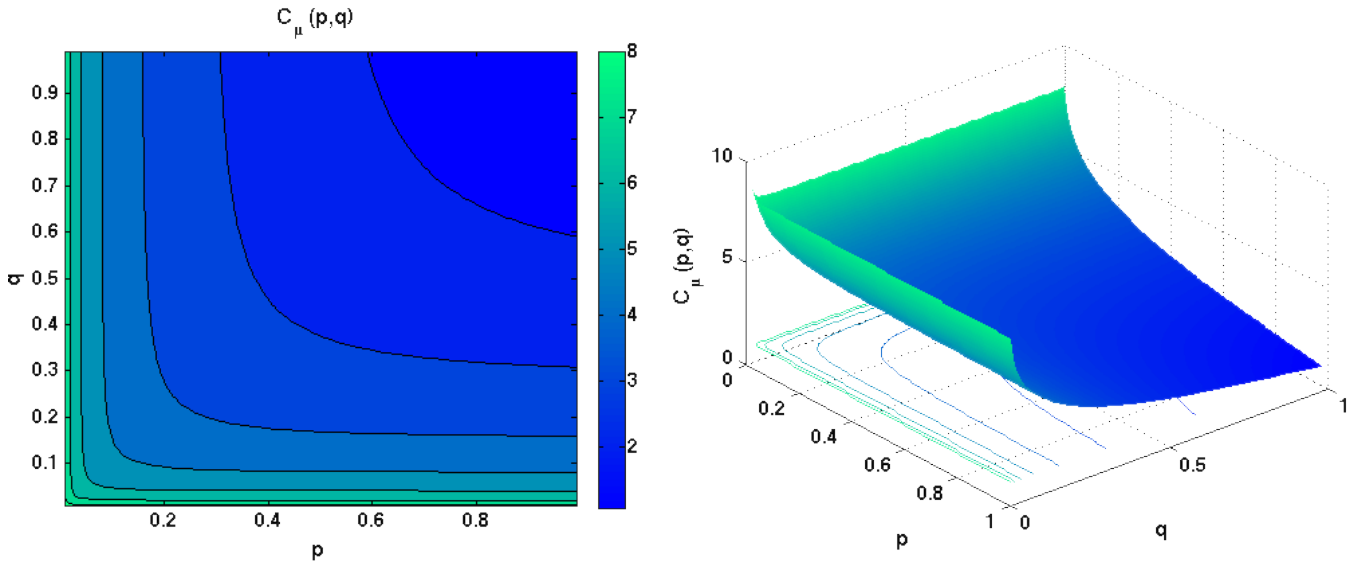

Figure 4: $C_\mu(p,q)$ in bits from eqn. 38 plotted as a contour plot (left) and three-dimensional plot (right). Statistical complexity decreases with both increasing p and increasing q.

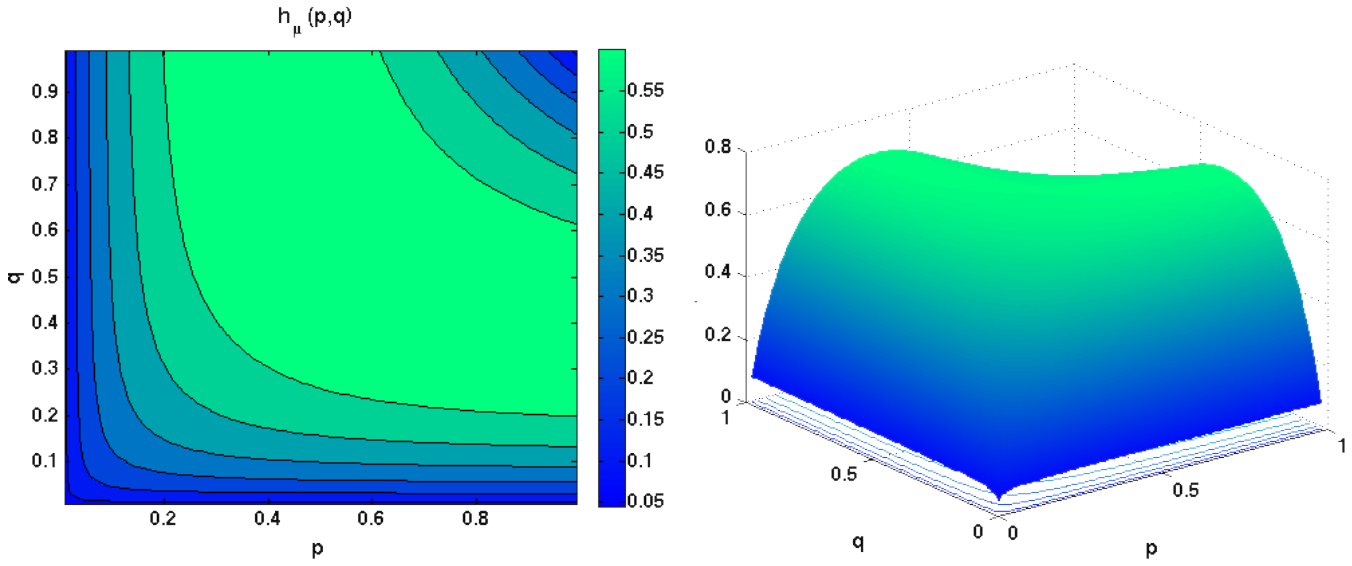

Figure 5: $h_\mu(p,q)$ in bits from eqn. 39 plotted as a contour plot (left) and three-dimensional plot (right). Entropy rate is largest for $p,\ q \sim \frac{1}{2}$.

Since $C_\mu^+(p,q)$ is symmetric with respect to p and q by inspection of eqn. 38,

$$\Xi(p,q) = 0. \tag{44}$$

This class of nonunifilar word generators is causally reversible. Therefore, for the remainder of this section, we will drop the superscripts in $C_\mu^\pm$ and just denote the statistical complexity as C_μ.

As might have been expected, the hardest calculation is that of excess entropy. I again do not have the patience to unifilarize the infinite state time-reversed forward epsilon machine, so I content myself with

approximations of excess entropy using the code in the Appendix. When calculating the excess entropy for the general p, q case, one must consider the infinite set of transient states that correspond to observing word distributions of the form 1^n, since the initial mixed state presentation has full weight on one of these transient states. Approximations to the excess entropy and crypticity, $\xi = C_\mu - E$, as functions of p and q are shown in the figure below. Notice that excess entropy is anti-correlated with statistical complexity over the parameter space, i.e. predictability need not be correlated with the amount of memory required for prediction.

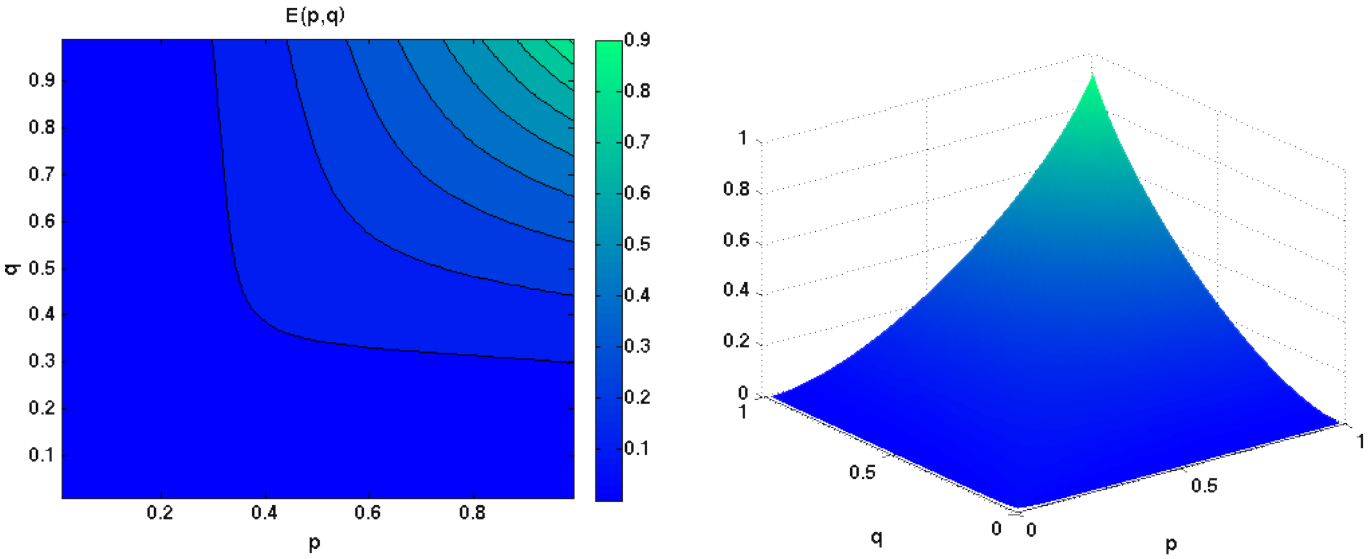

Figure 6: $E(p, q)$ in bits calculated using the code above, plotted as a contour plot (left) and three-dimensional plot (right). Excess entropy increases with both increasing p and increasing q.

3.1 Extension to continuous time

Supposing that the discrete time stochastic transition matrices $T^{(1)}$ and $T^{(0)}$ were derived from some continuous time dynamics, in which

$$\frac{d}{dt}\begin{pmatrix} p(A) \\ p(B) \end{pmatrix} = \begin{pmatrix} -k_{AB} & k_{BA} \\ k_{AB} & -k_{BA} \end{pmatrix} \begin{pmatrix} p(A) \\ p(B) \end{pmatrix}, \tag{45}$$

where k_{AB} and k_{BA} are "kinetic rates" that might have to do with the system's intrinsic characteristics, e.g. activation energies or scattering cross sections. The corresponding discrete time dynamics can be approximated for small time steps Δt as

$$\begin{pmatrix} p(A, t+\Delta t) \\ p(B, t+\Delta t) \end{pmatrix} = \begin{pmatrix} 1 - k_{AB}\Delta t & k_{BA}\Delta t \\ k_{AB}\Delta t & 1 - k_{BA}\Delta t \end{pmatrix} \begin{pmatrix} p(A, t) \\ p(B, t) \end{pmatrix}. \tag{46}$$

Matching the transition matrix elements in the equation above to $T^{(1)}$ and $T^{(0)}$ given in eqn. 29, we see that

$$p = k_{AB}\Delta t, \;\; q = k_{BA}\Delta t. \tag{47}$$

My goal in this next few sets of equations will be to determine how entropy rate, statistical complexity, excess entropy, and crypticity scale with Δt. Hopefully, this will allow me to break up these information theoretic quantities that characterize time series data into a quantity that depends on the physical parameters of the

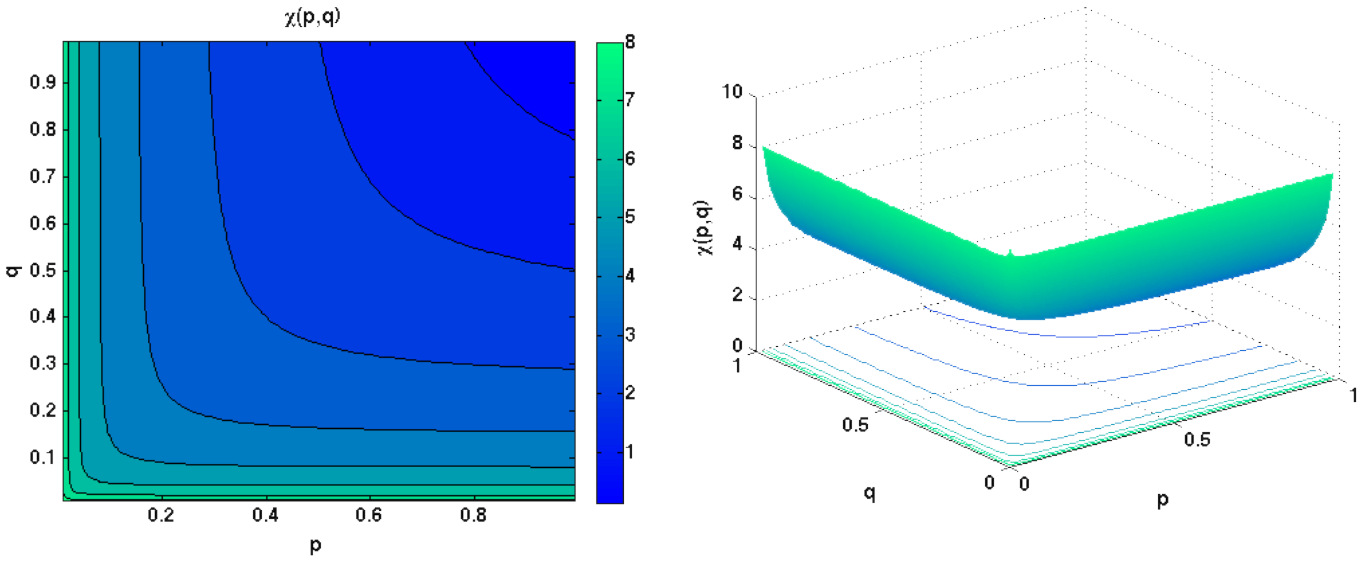

Figure 7: $\xi(p,q)$ in bits, calculated as $C_\mu - E$, plotted as a contour plot (left) and three-dimensional plot (right). Crypticity increases as p and q decrease.

underlying system (i.e., kinetic rates k_{AB} and k_{BA}) and a quantity that depends on the time resolution of the measuring instrument Δt.

We start with the statistical complexity. The π_n for $p = q = k\Delta t$ are given by

$$\pi_n = \frac{p}{2}(1-p)^{(n-1)}\left(1+(n-1)p\right) = \frac{k\Delta t}{2}(1-k\Delta t)^{n-1}\left(1+(n-1)k\Delta t\right) \tag{48}$$

with $\pi_0 = \pi_1 = \frac{k\Delta t}{2}$. As n increases, π_n is the product of a term that increases linearly with n and another that decreases exponentially with n. The "timescale" on which the exponential term decreases is $\frac{1}{k\Delta t}$. We are supposed to calculate $-\sum_{n=1}^{\infty}\pi_n \log_2 \pi_n$, but maybe, we should index π_n by a continuous variable π_t, which will have a probability distribution defined by π_n in the limit of $\Delta t \to 0$:

$$\pi_t = \lim_{\Delta t\to 0,\ t=n\Delta t} \pi_n = \frac{k}{2}(1+kt)e^{-kt}. \tag{49}$$

This happens to be a correctly normalized distribution. In the more general case in which p and q are not identical,

$$\pi_t = \lim_{\Delta t\to 0,\ t=n\Delta t} \pi_n = \frac{k_{AB}k_{BA}}{k_{AB}+k_{BA}}\,\frac{k_{AB}e^{-k_{BA}t} - k_{BA}e^{-k_{AB}t}}{k_{AB}-k_{BA}}, \tag{50}$$

which also happens to be a correctly normalized distribution. (The fact that these are correctly normalized without any legwork is a sign that the limit to continuous was taken correctly.) Now we calculate

$$C_\mu = -\int_0^\infty \pi_t \log_2 \pi_t dt. \tag{51}$$

In the case that $k_{AB} = k_{BA}$, we can calculate this quite easily:

$$\begin{aligned} C_\mu &= -\int_0^\infty \frac{1}{2} k(1+kt) e^{-kt} \left(\log_2 \frac{k}{2} + \log_2(1+kt) - \frac{kt}{\log 2} \right) dt & (52) \\ &= \log_2 \frac{2}{k} + \frac{3}{2\log 2} - \frac{k}{2} \int_0^\infty (1+kt) e^{-kt} \log_2(1+kt) dt & (53) \\ &= \log_2 \frac{2}{k} + \frac{3}{2\log 2} - \frac{k}{2} \int_0^\infty (1+x) e^{-x} \log_2(1+x) \frac{dx}{k} & (54) \\ &= \log_2 \frac{2}{k} + \frac{3}{2\log 2} - \frac{1}{2} \int_0^\infty (1+x) e^{-x} \log_2(1+x) dx & (55) \\ &\simeq 1.013 + \log_2 \frac{2}{k} \text{ bits.} & (56) \end{aligned}$$

It is not lost on me that $\frac{2}{k}$ is the natural relaxation time of the original system and an eigenvalue of the matrix $T^{(1)} + T^{(0)}$. This formula will be explored more generally in a later section. I wasn't able to do a similarly cute manipulation for the case in which $k_{AB} \neq k_{BA}$ due to a pesky difference in the logarithm, but I evaluated C_μ numerically as a function of k_{AB} and k_{BA} and the results are shown below. There are a few differences between the contour plot shown below and the contour plot shown previously– for instance, when $k_{AB} = k_{BA}$, the statistical complexity decreases in a discontinuous manner. (The continuous time formalism forces us to deal with the degeneracy in the eigenvalue spectrum of the continuous time transition matrix $\begin{pmatrix} -k & k \\ k & -k \end{pmatrix}$. Still pondering why this discontinuity would not manifest itself at all in the discrete case.) But as was true for the discrete case, C_μ is still larger when k is smaller, basically because the relaxation time scale increases and the statistical complexity increases as the log of that relaxation time. What this limit actually means is that the countable infinity of causal states becomes an uncountable infinity in the limit of continuous time.

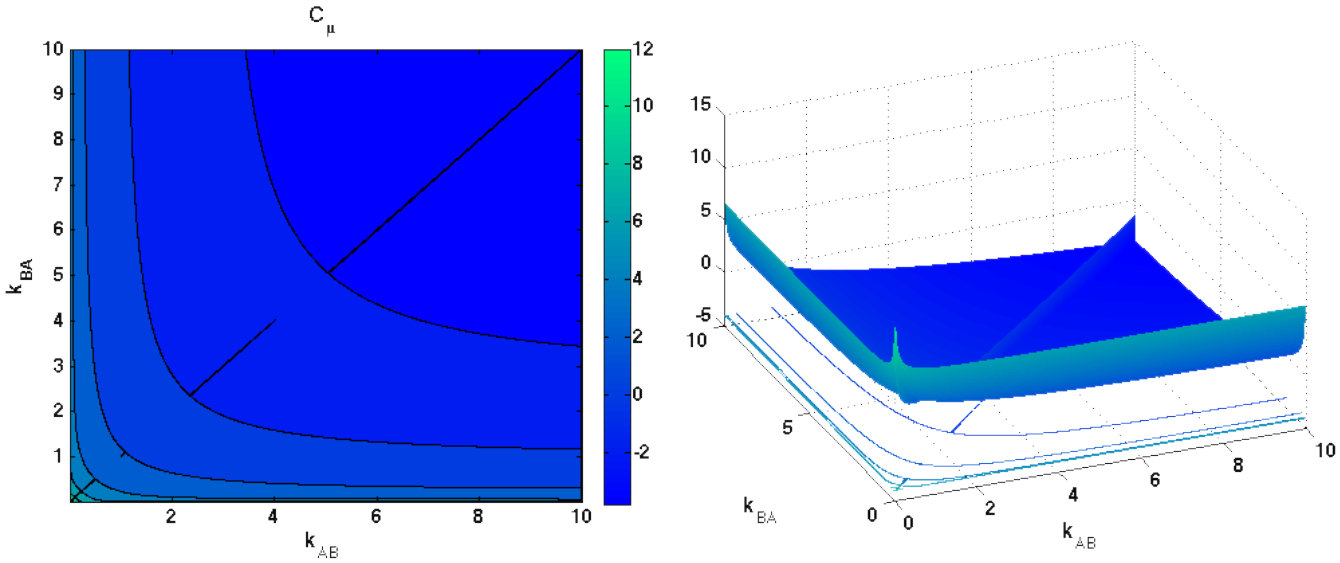

Figure 8: C_μ in bits calculated using the code above. Due to $\log_2 \frac{2}{k}$, C_μ can be negative for large k. This is one of those weird things about differential entropy that I am unsure how to deal with.

The next obvious step would be to compute h_μ in closed form as an integral over all time. There is,

however, a problem. The entropy rate for each causal state is

$$h_t = \lim_{\Delta t << 1, t = n\Delta t} H[M_{n,0}] = H[\frac{k_{AB}k_{BA}(e^{-k_{AB}t} + e^{-k_{BA}t})}{k_{AB}e^{-k_{BA}t} - k_{BA}e^{-k_{AB}t}}\Delta t]. \tag{57}$$

In order to avoid showing that $h_t \to 0$ in the limit that $\Delta t \to 0$, I only took the limit as Δt grew smaller and smaller. A simple expansion shows

$$H[\alpha\Delta t] = -\frac{1}{\ln 2}\left(\alpha\Delta t \ln \alpha\Delta t - (1 - \alpha\Delta t)\ln(1 - \alpha\Delta t)\right) \simeq -\frac{\alpha\Delta t}{\ln 2}\left(\ln \alpha\Delta t + 1\right) \tag{58}$$

which means that the entropy rate of each causal state is proportional to Δt and that $h_t \to 0$ in the limit of continuous time, $\Delta t \to 0$. This would then imply that

$$h_\mu = \int_0^\infty \pi_t h_t dt \to 0 \tag{59}$$

but there should be a nontrivial definition of entropy rate for a continuous process that results in a finite value. If we assume that really, the entropy rate coming from each causal state in the limit of continuous time should be $\frac{H[\alpha\Delta t]}{\Delta t}$ (to rescale the time axis to be in units of the time resolution) then the term $\ln \alpha\Delta t$ would cause the expression for entropy rate to increase without bound. Hopefully, someone else knows the answer to the question of whether or not entropy rate should/will increase without bound for some continuous time processes, and if there is any meaning as to how it increases without bound as $\Delta t \to 0$.

So what about E? Unfortunately, I did not unifilarize the reverse time mixed state presentation to get a closed form solution for E as a function of p and q, which makes taking the limit of small Δt difficult. The numerical results in the section above suggested strongly that E does not increase without bound as Δt decreased to 0, although it's also the case that I found it numerically impossible to evaluate E for very small Δt. My only explanation for the shape of E– that it decreases as kinetic rate decreases, opposite to the structural complexity– is that the mutual information between the past and future is increased when the time between successive 0's (syncing to internal states) is decreased. This is reflected in the following estimation. Clearly, one of the main contributions to the mutual information between the past and the future is the "no successive zeros" rule: if you see a 0, you must see a 1. This manifests itself in the following way:

$$I(x_{t+\Delta t}; x_t) = H[x_{t+\Delta t}] - H[x_{t+\Delta t}|x_t] = H[x_{t+\Delta t}] - p(1)H[x_{t+\Delta t}|x_t = 1]. \tag{60}$$

Going back to using p and q, we have

$$H[x_{t+1}] = H[\frac{pq}{p+q}], \tag{61}$$

$$p(1) = 1 - p(0) = 1 - \frac{pq}{p+q}, \tag{62}$$

and to find $H[x_{t+\Delta t}|x_t = 1] = H[p(0|1)]$,

$$p(0|1) = \frac{p(1,0)}{p(1)} = \frac{\pi_A p(ABA) + \pi_B p(BBA)}{1 - p(0)} = \frac{\frac{q}{p+q}pq + \frac{p}{p+q}(1-q)q}{1 - \frac{pq}{p+q}} = \frac{pq}{p + q - pq} \tag{63}$$

Surprisingly, for $p = k_{AB}\Delta t$, $q = k_{BA}\Delta t$, and $\Delta t << 1$, the mutual information scales as Δt^2 according to Mathematica's series expansion (I got lazy):

$$I(x_{t+\Delta t}; x_t) \approx \left(\frac{k_{AB}k_{BA}}{k_{AB} + k_{BA}}\Delta t\right)^2. \tag{64}$$

In reality, if we see a 0, then we know that the next $\sim \frac{1}{(k_{AB}+k_{BA})\Delta t}$ symbols are likely to be a 1 also, which means that a rough estimate for how E should scale is

$$E \sim \frac{1}{(k_{AB} + k_{BA})\Delta t} I(x_{t+\Delta t}; x_t) \approx \frac{k_{AB}^2 k_{BA}^2 \Delta t}{(k_{AB} + k_{BA})^3} = \pi_A^2 \pi_B^2 (k_{AB} + k_{BA})\Delta t. \tag{65}$$

So as $\Delta t \to 0$, E would tend towards 0, which is (in fact) what the numerical results seem to suggest. And the dependence of E on k_{AB} and k_{BA} according to the equation above is similar to the dependence observed numerically– E increases with kinetic rates. See Figure below, which shows $I(x_{t+1}; x_t)$ as a function of p and q.

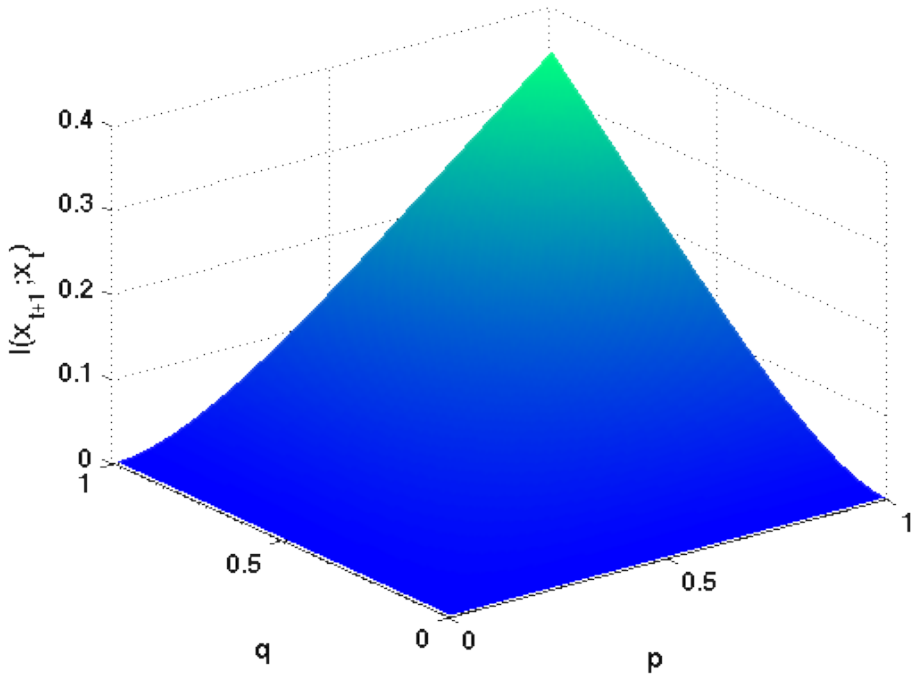

Figure 9: $I(x_{t+1}; x_t)$ in bits as a function of p and q. Compare to E as a function of p and q.

Preliminary conclusions: continuous time is interesting, and results (i.e., discontinuities) can appear in continuous time that did not appear in discrete time. It seems like E and C_μ are anti-correlated over parameter space even when we are talking about continuous time processes. For this particular system, the two constitute very different measures of underlying "structure". The excess entropy is maximized for processes that look closer to periodic (highly predictable) and the statistical complexity is maximized for processes that have long relaxation times.

4 Another simple binary nonunifilar source

Here we consider a different class of simple binary nonunifilar sources, but the topology of connections of the corresponding epsilon machine and mixed state presentation are the same as for the simple nonunifilar source. See Figure 10. Hence, much of the machinery developed previously carries over exactly. The state space is given by $T^{(0)}$ and $T^{(1)}$, with π the eigenvector with eigenvalue 1 of $T = T^{(0)} + T^{(1)}$. The recurrent causal states are (similar to the SNS) the groupings of histories defined by 01^n, $n = 0, ...$ and we let s_n denote the causal state 01^n. Transitions to state s_0 emit a 0; transitions to any other causal state emit a 1. The transition probabilities between these causal states are given by

$$M_{s_n \to s_0} = M_{n,0} = \frac{1^\top T^{(0)} \left(T^{(1)}\right)^n T^{(0)} \pi}{1^\top \left(T^{(1)}\right)^n T^{(0)} \pi} \tag{66}$$

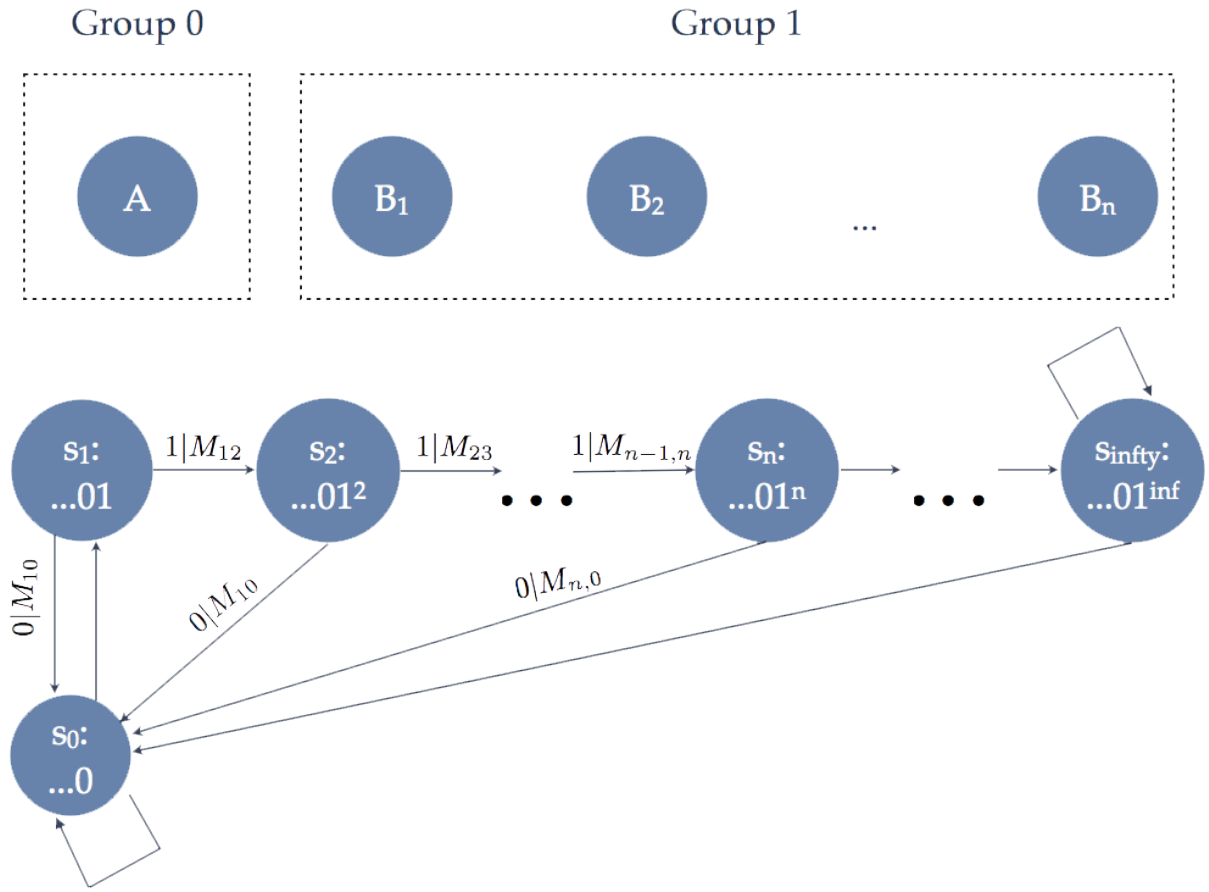

Figure 10: At top is a schematic of the nonunifilar HMM. Transitions to a state in group 0 emit a 0 and transitions to state in group 1 emit a 1; the state space is fully connected. At bottom is its epsilon machine presentation. The self-loop on causal state s_0 was not present in the epsilon machine presentation of the simple nonunifilar source.

and

$$M_{s_n \to s_{n+1}} = M_{n,n+1} = 1 - M_{n,0} = \frac{1^\top \left(T^{(1)}\right)^{n+1} T^{(0)} \pi}{1^\top \left(T^{(1)}\right)^{n} T^{(0)} \pi}. \tag{67}$$

The transitions from s_0 to s_1 are given by

$$M_{s_0 \to s_1} = \frac{1^\top T^{(1)} T^{(0)} \pi}{1^\top T^{(0)} \pi} \tag{68}$$

and

$$M_{s_0 \to s_0} = \frac{1^\top \left(T^{(0)}\right)^{2} \pi}{1^\top T^{(0)} \pi}. \tag{69}$$

From these transition probabilities, we can solve for the statistical complexity in full generality. Again, in equilibrium, the probability flow into and out of causal state s_n must balance:

$$\pi_n = M_{n-1,n} \pi_{n-1} \to \frac{\pi_n}{\pi_{n-1}} = M_{n-1,n} = \frac{1^\top \left(T^{(1)}\right)^{n} T^{(0)} \pi}{1^\top \left(T^{(1)}\right)^{n-1} T^{(0)} \pi}. \tag{70}$$

Therefore,

$$\frac{\pi_n}{\pi_0} = \prod_{k=1}^{n} \frac{\pi_k}{\pi_{k-1}} = \frac{1^\top \left(T^{(1)}\right)^n T^{(0)} \pi}{1^\top T^{(0)} \pi}. \tag{71}$$

Normalizing the probability distribution gives

$$\begin{aligned} 1 &= \sum_{n=0}^{\infty} \pi_n & (72)\\ &= \sum_{n=0}^{\infty} \frac{1^\top \left(T^{(1)}\right)^n T^{(0)} \pi}{1^\top T^{(0)} \pi} \pi_0 & (73)\\ &= \frac{1^\top \left(1 - T^{(1)}\right)^{-1} T^{(0)} \pi}{1^\top T^{(0)} \pi} \pi_0 & (74)\\ \pi_0 &= \frac{1^\top T^{(0)} \pi}{1^\top \left(1 - T^{(1)}\right)^{-1} T^{(0)} \pi}. & (75) \end{aligned}$$

We can therefore write all of the π_n as

$$\pi_n = \frac{1^\top \left(T^{(1)}\right)^n T^{(0)} \pi}{1^\top \left(1 - T^{(1)}\right)^{-1} T^{(0)} \pi}. \tag{76}$$

The statistical complexity is just

$$C_\mu = -\sum_{n=0}^{\infty} \pi_n \log_2 \pi_n = -\sum_{n=0}^{\infty} \frac{1^\top \left(T^{(1)}\right)^n T^{(0)} \pi}{1^\top \left(1 - T^{(1)}\right)^{-1} T^{(0)} \pi} \log_2 \frac{1^\top \left(T^{(1)}\right)^n T^{(0)} \pi}{1^\top \left(1 - T^{(1)}\right)^{-1} T^{(0)} \pi}. \tag{77}$$

This expression can be simplified if we note that each of these expressions are actually scalars, and so

$$C_\mu = \log_2 \left(1^\top \left(1 - T^{(1)}\right)^{-1} T^{(0)} \pi\right) - \sum_{n=0}^{\infty} \frac{1^\top \left(T^{(1)}\right)^n T^{(0)} \pi}{1^\top \left(1 - T^{(1)}\right)^{-1} T^{(0)} \pi} \log_2 \left(1^\top \left(T^{(1)}\right)^n T^{(0)} \pi\right). \tag{78}$$

And that's about the best I can do with that expression. The entropy rate can similarly be calculated in closed form, since the entropy production of each recurrent causal state is

$$h_{s_n} = H[M_{n,0}] = H[\frac{1^\top \left(T^{(1)}\right)^{n+1} T^{(0)} \pi}{1^\top \left(T^{(1)}\right)^n T^{(0)} \pi}]. \tag{79}$$

Therefore,

$$h_\mu = \sum_{n=0}^{\infty} h_{s_n} \pi_n = \sum_{n=0}^{\infty} \frac{1^\top \left(T^{(1)}\right)^n T^{(0)} \pi}{1^\top \left(1 - T^{(1)}\right)^{-1} T^{(0)} \pi} H[\frac{1^\top \left(T^{(1)}\right)^{n+1} T^{(0)} \pi}{1^\top \left(T^{(1)}\right)^n T^{(0)} \pi}]. \tag{80}$$

To evaluate the excess entropy and crypticity, we need to calculate the mixed state presentation, which involves (as before) the transient mixed states corresponding to observations of words 1^n. The mixed state s_{-n} denotes observing 1^n. The transitions between these mixed states (emitting a 1) and to s_0 (emitting a 0) are given by

$$M_{-n,0} = \frac{1^\top T^{(0)} \left(T^{(1)}\right)^n \pi}{1^\top \left(T^{(1)}\right)^n \pi} \tag{81}$$

and

$$M_{-n,-(n+1)} = \frac{1^\top \left(T^{(1)}\right)^{n+1} \pi}{1^\top \left(T^{(1)}\right)^n \pi}. \tag{82}$$

See code in the Appendix.

4.1 Continuous limit

To take this to the continuous limit, we note that for this particular set of nonunifilar word generators, the matrices $T^{(1)}$ and $T^{(0)}$ can be written as

$$T^{(0)} = \begin{pmatrix} 1 - k_A \Delta t & v_{1\times N} \\ 0_{N\times 1} & 0_{N\times N} \end{pmatrix},\ T^{(1)} = \begin{pmatrix} 0 & 0_{1\times N} \\ w_{N\times 1} & I_{N\times N} + M\Delta t \end{pmatrix} \tag{83}$$

where N is the number of states. The subscripts indicate the size of the matrix where appropriate, but will be dropped from here on out. The matrices $w,\ v,\ M$ have a list of the kinetic rates into and out of the various states in the nonunifilar system. In terms of these matrices, in the limit of small Δt, π_n adopts the simple form

$$\pi_{n+1} = \frac{1^\top e^{nM\Delta t} w}{1^\top M^{-1} w}\Delta t,\ \pi_0 = \frac{1}{1^\top M^{-1} w}\Delta t. \tag{84}$$

This suggests that the continuous state space implementation of the causal states would have

$$\pi_t = \frac{1^\top e^{Mt} w}{1^\top M^{-1} w} \tag{85}$$

and

$$C_\mu = -\int_0^\infty \pi_t \log_2 \pi_t dt = -\int_0^\infty \frac{1^\top e^{Mt} w}{1^\top M^{-1} w} \log_2 \frac{1^\top e^{Mt} w}{1^\top M^{-1} w} dt \tag{86}$$

$$= \log_2 \left(1^\top M^{-1} w\right) - \int_0^\infty \frac{1^\top e^{Mt} w}{1^\top M^{-1} w} \log_2 \left(1^\top e^{Mt} w\right) dt. \tag{87}$$

It is no surprise that the statistical complexity should be related to the natural relaxation times of the system, given that the information about relaxation times is contained in M and w. In particular, if we diagonalize $M = SDS^{-1}$, then the eigenvalues $-\vec{\lambda}$ that sit on the diagonal of D are the negative inverse relaxation times, so

$$C_\mu = \log_2(1^\top S^{-1} D^{-1} S w) - \int_0^\infty \frac{1^\top S\ \mathrm{diag}(e^{-\vec{\lambda}t}) S^{-1} w}{1^\top S^{-1} D^{-1} S w} \log_2(1^\top S\ \mathrm{diag}(e^{-\vec{\lambda}t}) S^{-1} w) dt. \tag{88}$$

This expression is still rather intractable, so suppose that all of the eigenvalues $\lambda_i = \lambda$ are identical. In that case, the eigenvectors of M can be chosen without loss of generality to be the basis vectors $\hat{e}_i$. Then C_μ collapses to the particularly simple

$$C_\mu = \log_2(\frac{1}{\lambda}\sum_i w_i) - \int_0^\infty \lambda e^{-\lambda t} \log_2(e^{-\lambda t} \sum_i w_i) dt \tag{89}$$

$$= \log_2 \frac{1}{\lambda} + \frac{1}{\ln 2} \tag{90}$$

As expected, the statistical complexity grows as the logarithm of the relaxation time $\frac{1}{\lambda}$, plus an additional constant that (it seems) has only to do with the topological structure of the causal states.

I attempted to calculate h_μ but ran into the same problems that I did for the variation on the simple nonunifilar source with the time resolution.

4.2 Can we infer the N hidden states?

One of the biggest issues in nonunifilar HMM inference is identifying the number of hidden states that correspond to each observed symbol. Even with perfect noiseless data, this could be a difficult problem for certain pathological cases. Let us suppose that the state $x(t)$, in which

$$x(t) = \begin{pmatrix} p(A,t) \\ p(B_1,t) \\ \vdots \\ p(B_N,t) \end{pmatrix} \tag{91}$$

evolves according to

$$\frac{dx}{dt} = Mx. \tag{92}$$

The notation used here is that $1_{m\times n}$ denotes a matrix of dimension m by n in which all of the entries are 1's, and I_n denotes the (square) identity matrix of column length (or row length) of N. The solution to this equation in general is merely

$$x(t) = e^{Mt}x(0) \tag{93}$$

and when M can be diagonalized as $M = SDS^{-1}$, this is solved as a sum of exponentials with $N+1$ different time constants, one of which is 0 (corresponding to the stationary distribution.) So in general, unless some of the hidden states impose a degeneracy of relaxation times, fitting the autocorrelation function alone to a sum of exponentials will allow us to interpret the number of hidden states.

But what if there is degeneracy in the eigenspectrum of M? Let us first consider a particular instantiation of this problem in which the N hidden states are identical and in which the state space is fully connected:

$$M = \begin{pmatrix} -Nk_A & k_B 1_{1\times N} \\ k_A 1_{N\times 1} & k_B 1_{N\times N} - (N+1)k_B I_N \end{pmatrix}. \tag{94}$$

For this particular matrix M, there are only three distinct eigenvalues: 0, with eigenvector

$$\pi = \begin{pmatrix} \frac{k_B}{k_B+Nk_A} \\ \frac{k_A}{k_B+Nk_A} \\ \vdots \\ \frac{k_A}{k_B+Nk_A} \end{pmatrix} \tag{95}$$

i.e. stationary distribution; $-(k_B + Nk_A)$, with eigenvector

$$v_0 = \begin{pmatrix} -N \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \tag{96}$$

and eigenvalue $-(N+1)\,k_B$ with multiplicity $N-2$. The autocorrelation function will now be a linear combination of three exponentials, and it is difficult to see how one would infer the presence of N hidden states.

The epsilon-machine in general provides far more information about the system than just the autocorrelation function, but for this particular nonunifilar HMM, the epsilon-machine is almost a recasting of the autocorrelation function. It, too, is insensitive to the number of hidden states if we hold π_A and N_{ab} constant, i.e. if we hold the probability of observing a 0 constant. To hold these two variables constant, it is sufficient to set $k_A \to k_A/N$ and to keep k_B constant and independent of N. Additionally using the continuous to discrete transformation $T = I + M\Delta t$ introduced previously, we have

$$T = \begin{pmatrix} 1 - k_A\Delta t & k_B\Delta t 1_{1\times N} \\ \frac{k_A}{N}\Delta t 1_{N\times 1} & k_B\Delta t 1_{N\times N} + (1-(N+1)k_B\Delta t)\, I_N \end{pmatrix}. \tag{97}$$

When we transition to state A, we emit 0; otherwise we emit a 1; and this is shown in the matrices $T^{(1)}$ and $T^{(0)}$ below:

$$T^{(0)} = \begin{pmatrix} 1 - k_A\Delta t & k_B\Delta t 1_{1\times N} \\ 0_{N\times 1} & 0_{N\times N} \end{pmatrix} \tag{98}$$

and

$$T^{(1)} = \begin{pmatrix} 0 & 0_{1\times N} \\ \frac{k_A}{N}\Delta t 1_{N\times 1} & k_B\Delta t 1_{N\times N} + (1-(N+1)k_B\Delta t)\, I_N \end{pmatrix}. \tag{99}$$

The transition probabilities between recurrent causal states are given by $\frac{1^T\left(T^{(1)}\right)^{n+1}T^{(0)}\pi}{1^T\left(T^{(1)}\right)^{n}T^{(0)}\pi}$ and the transition probabilities between transient mixed states are given by $\frac{1^T\left(T^{(1)}\right)^{n+1}\pi}{1^T\left(T^{(1)}\right)^{n}\pi}$.

Claim: If π_A and N_{ab} are held constant for N by varying kinetic rates $k_A \to k_A/N$, $k_B \to k_B$, then the epsilon machine, and all information theoretic quantities calculable from it will be independent of the number of hidden states N.

Proof: If I can show that

$$f_n = \frac{1^T\left(T^{(1)}\right)^n T^{(0)}\pi}{1^T T^{(0)}\pi} \tag{100}$$

is independent of N, then it follows that the transition probabilities in the epsilon machine presentation are independent of N, from which it also will follow that the statistical complexity and entropy rate are independent of N.

First looking at f_n, we see that

$$T^{(0)}\pi = \left((1-k_A\Delta t)\frac{k_B}{k_B+k_A} + Nk_B\Delta t\frac{k_A/N}{k_B+k_A}\right)\begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix} = \frac{k_B}{k_A+k_B}\begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix}, \tag{101}$$

$$T^{(1)}T^{(0)}\pi = \frac{k_A\Delta t}{N}\begin{pmatrix}0\\1\\\vdots\\1\end{pmatrix}, \tag{102}$$

and

$$\left(T^{(1)}\right)^m T^{(1)}T^{(0)}\pi = \frac{k_A\Delta t}{N}R^m 1_{N\times 1} \tag{103}$$

where

$$R = k_B\Delta t\ 1_{N\times N} + (1-(N+1)k_B\Delta t)\ I_N. \tag{104}$$

Clearly $1_{N\times 1}$ is an eigenvector of R since it is an eigenvector of $1_{N\times N}$ with eigenvalue N and an eigenvector of the identity matrix I_N with eigenvalue 1; hence it is an eigenvector of R with eigenvalue $Nk_B\Delta t + (1-(N+1)k_B\Delta t) = 1-k_B\Delta t$. Therefore,

$$R^m 1_{N\times 1} = (1-k_B\Delta t)^m\, 1_{N\times 1}, \tag{105}$$

which then allows us to say that, for $n \geq 2$

$$f_n = \frac{\frac{k_B}{k_A+k_B}\frac{k_A\Delta t}{N}(1-k_B\Delta t)^{n-1}(N)}{\frac{k_B}{k_A+k_B}} = k_A\Delta t\,(1-k_B\Delta t)^{n-1}, \tag{106}$$

which is independent of N. Hence transition probabilities between causal states are independent of N as well, and the epsilon machine cannot distinguish between different numbers of identical hidden states if π_A and N_{ab} are held constant.

Excess entropy is also independent of N, using similar steps to the proof above to find transition probabilities between transient mixed states. So now we can rephrase our original question. We cannot detect the number of hidden states with π_A and N_{ab} held constant when hidden states are all identical, but we can ask how $C_\mu,\ E,\ \chi,\ h_\mu$ vary as functions of k_A and k_B in the case of identical hidden states. The answers are shown in Figures 11-14. Code that was used to calculate this is in the Appendix.

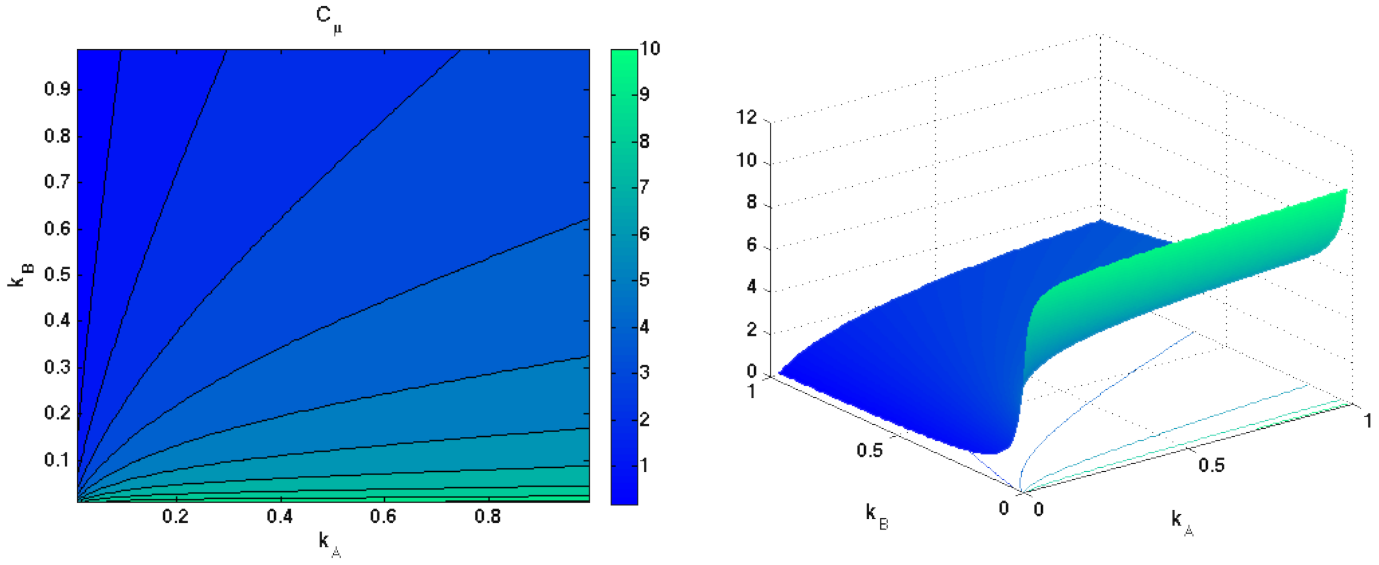

Figure 11: Statistical complexity is highest when probability flows quickly out of A into the hidden states and slowly amongst the hidden states.

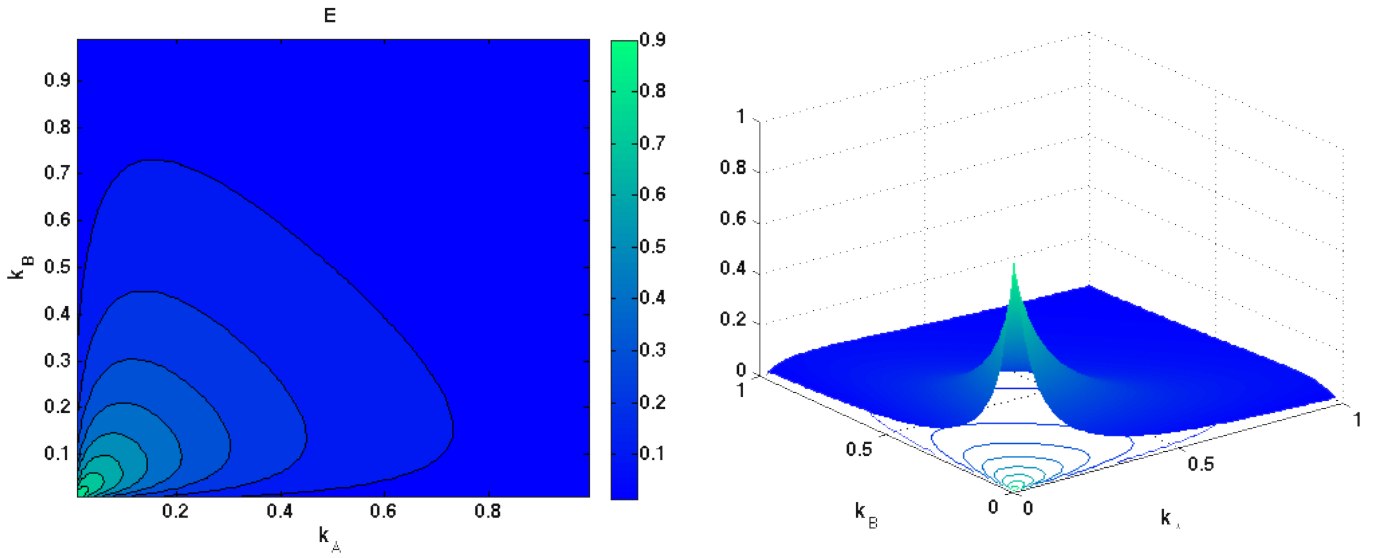

Figure 12: Excess entropy is largest when probability flows slowly both from the syncing state to the hidden state and amongst the hidden states.

If there is a definite bias amongst the hidden states, then C_μ, E, χ, and h_μ change noticeably as a function of the number of hidden states, even when controlling for π_A and N_{ab}. As a particular example, I choose to allow the outgoing kinetic rates of the hidden states adopt the form

$$k_{B_i \to B_j} = f(\frac{i}{N})k_B, \tag{107}$$

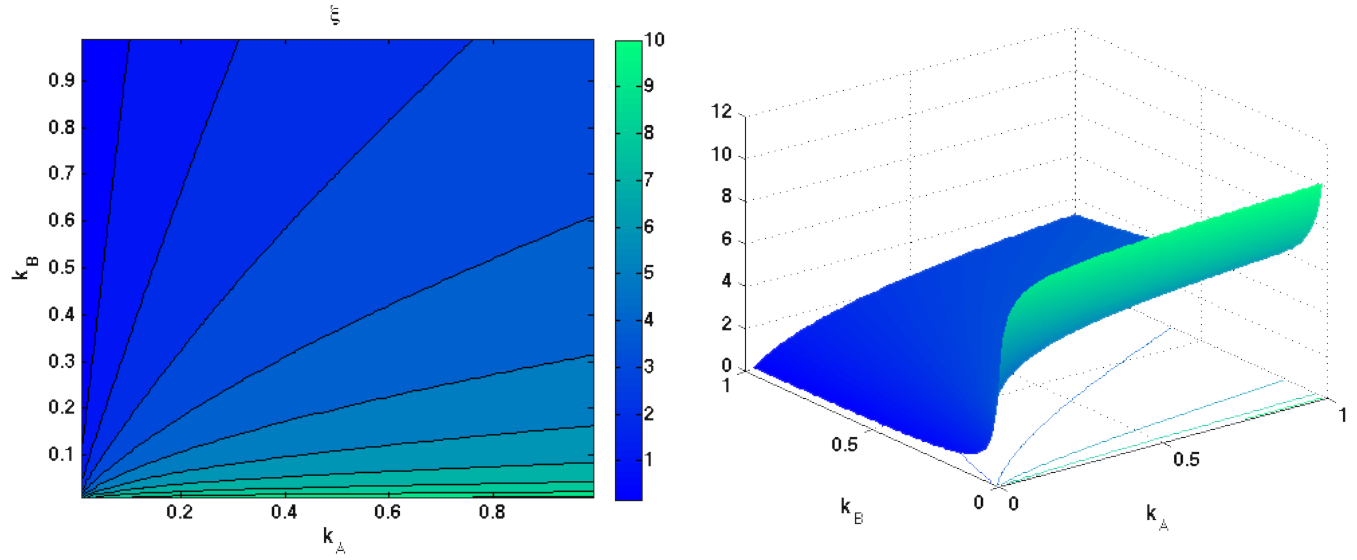

Figure 13: The crypticity looks very much like the statistical complexity.

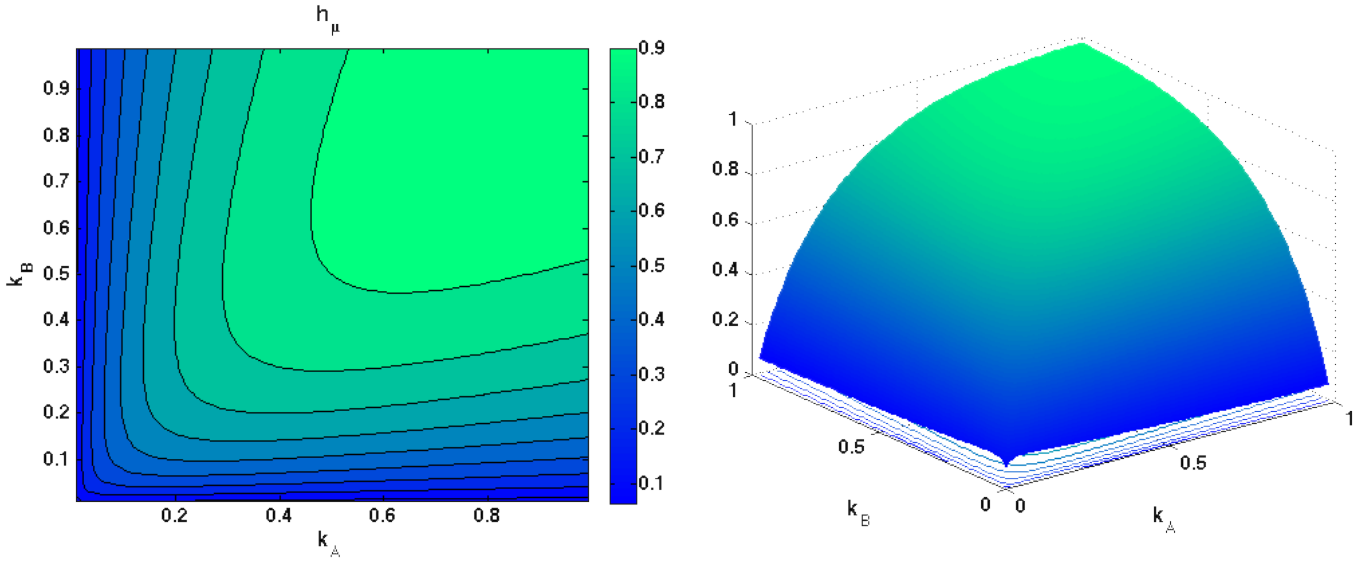

Figure 14: The entropy rate is largest when the probability flows quickly amongst the syncing state A and the hidden states B_i.

and then adjust k_B so as to keep π_A and N_{ab} constant. Detailed balance at stationarity establishes that

$$\pi_i k_{B_i \to B_j} = \pi_j k_{B_j \to B_i} \to \pi_i f(i/N) = \pi_j f(j/N). \tag{108}$$

Then, normalizing the distribution so that $\sum_{i=1}^{N} \pi_i + \pi_A = 1$, we find that

$$\pi_A = \frac{1/k_A}{\frac{1}{k_A} + \sum_{i=1}^{N} \frac{1}{k_B \ f(i/N)}} \tag{109}$$

and therefore that

$$N_{ab} = k_A \pi_A = \frac{1}{\frac{1}{k_A} + \sum_{i=1}^{N} \frac{1}{k_B \ f(i/N)}}. \tag{110}$$

Therefore, we adjust $k_A \to k_A/N$ and $k_B \to \frac{k_B}{N} \sum_{i=1}^{N} \frac{1}{f(i/N)}$ to control for π_A and N_{ab}. Now we find that different functions $f(i/N)$ lead to different forms of these information theoretic quantities as a function of N, as shown in Figures below. Interestingly, when noise is added to the kinetic rates so that $k_{A \to B_i} = k_A \left(1 + \eta_i\right)$ and $k_{B_i \to B_j} = k_{B_i} \left(1 + \eta_i\right)$ where $\eta_i \sim \mathcal{N}(0, \frac{1}{100})$ is uncorrelated white noise, the effect of N on entropy rate is most robust. The systematic effect of N on the other information theoretic quantities that relate more strongly to intrinsic structure– E, χ, C_μ– is swamped by the noise. This is, in some ways, great news. It is much easier to calculate entropy rate than it is to calculate C_μ, for instance, from a data stream given that the epsilon-machine has a countable infinity of causal states.

4.3 Causal irreversibility

These simple nonunifilar binary word generators appear (numerically) to be causally reversible, surprisingly. In fact, I claim something stronger– the forward and reverse time epsilon machines are exactly identical. Recall that the transition probabilities between causal states are of the form $M_{n,n+1} = \frac{1^\top \left(T^{(1)}\right)^{n+1} T^{(0)} \pi}{1^\top \left(T^{(1)}\right)^{n} T^{(0)} \pi}$. So if we can show that these transition matrices are equivalent for the forward and reverse time processes, then we will have shown that the forward and reverse time epsilon machines are exactly equivalent, which implies causal reversibility. Or, if you like, $\pi_n = \frac{1^\top \left(T^{(1)}\right)^{n} T^{(0)} \pi}{\sum_{k=0}^{\infty} 1^\top \left(T^{(1)}\right)^{k} T^{(0)} \pi}$, and so if we can show that these are equivalent in forward and reverse time, causal reversibility follows. Thus we must prove this claim:

Claim: $1^\top \left(T^{(1)}\right)^n T^{(0)} \pi = 1^\top \left(\tilde{T}^{(1)}\right)^n \tilde{T}^{(0)} \tilde{\pi}$, where $\tilde{T}^{(x)}$ and $\tilde{\pi}$ are the transition matrices and stationary distribution for the time-reversed process.

Proof: First, as described in class,

$$T\pi = \pi = \tilde{T}\tilde{\pi}. \tag{111}$$

Let $\hat{e}_{1,1}$ be the matrix with a 1 in the top left corner and nowhere else. Then $T^{(0)}$ can be written succinctly as

$$T^{(0)}_{ij} = \delta_{i,1} T_{ij} \Rightarrow T^{(0)} = \hat{e}_{1,1} T \tag{112}$$

and therefore, for this system,

$$T^{(1)} = T - T^{(0)} = \left(I - \hat{e}_{1,1}\right) T. \tag{113}$$

The time-reversed transition matrices turn out to have similar relationships:

$$\tilde{T}^{(0)}_{ij} = \delta_{j,1} \tilde{T}_{ij} \Rightarrow \tilde{T}^{(0)} = \tilde{T} \hat{e}_{1,1} \tag{114}$$

and

$$\tilde{T}^{(1)} = \tilde{T} - \tilde{T}^{(0)} = \tilde{T} \left(I - \hat{e}_{1,1}\right). \tag{115}$$

Finally, let $D = \mathrm{diag}(\pi)$. For a time reversed process, based on the prescription described in class,

$$\tilde{T}_{ij} = \pi_i T_{ij} \pi_j^{-1} = \delta_{i,i'} \pi_{i'} T_{i',j'} \pi_{j'}^{-1} \delta_{j,j'} \Rightarrow \tilde{T} = D T^\top D^{-1}. \tag{116}$$

Our goal is to prove that $f_n = \tilde{f}_n$, where

$$f_n = 1^\top \left(T^{(1)}\right)^n T^{(0)} \pi. \ \tilde{f}_n = 1^\top \left(\tilde{T}^{(1)}\right)^n \tilde{T}^{(0)} \tilde{\pi}. \tag{117}$$

Concentrating on the time-reversed $\tilde{f}_n$ and substituting in the expressions for $\tilde{T}^{(x)}$ written above, we see that we must reduce the expression

$$\tilde{f}_n = 1^\top \left(D T^\top D^{-1} (I - \hat{e}_{1,1})\right)^n \left(D T^\top D^{-1} \hat{e}_{1,1}\right) \pi. \tag{118}$$

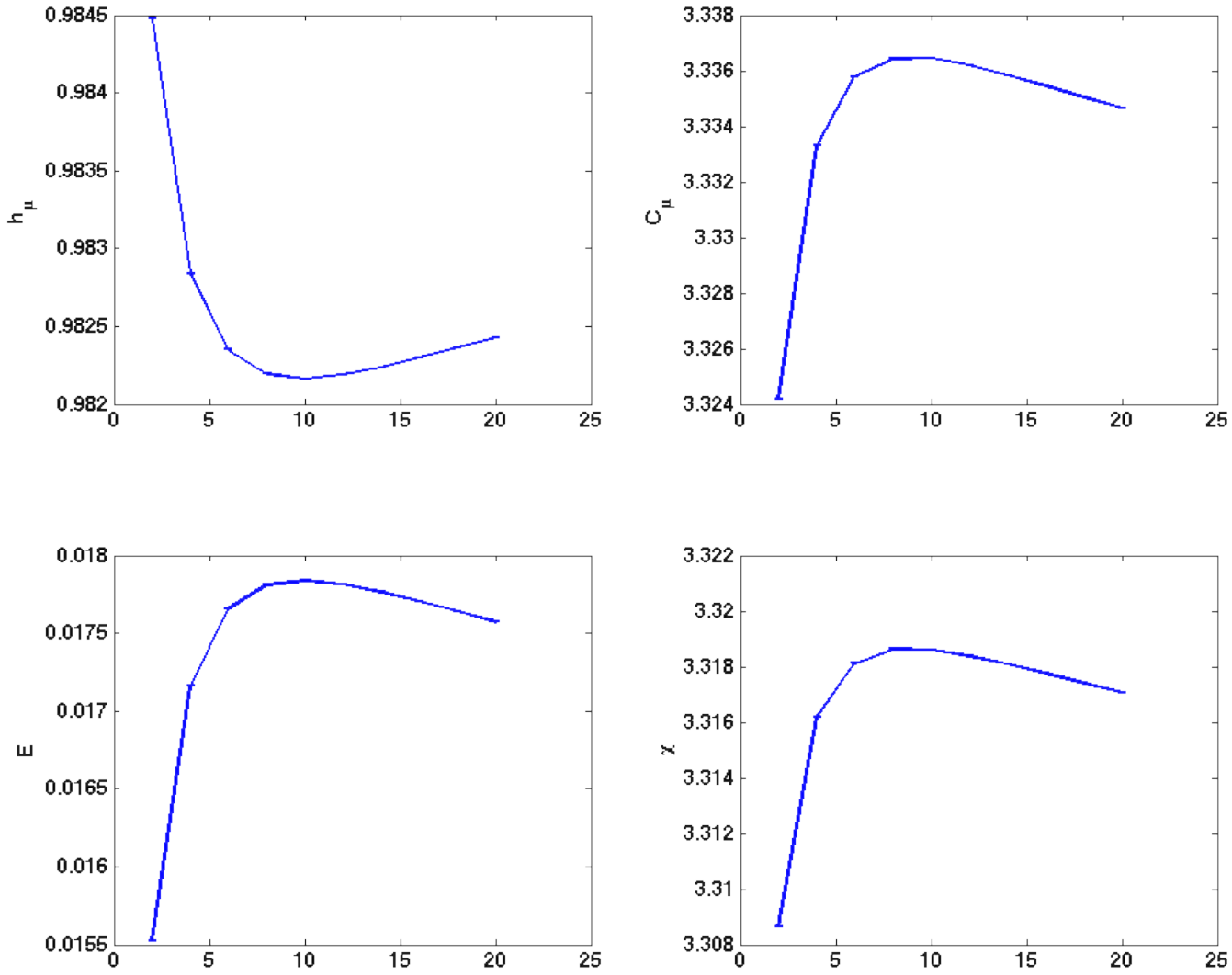

Figure 15: Again, π_A and N_{ab} are held constant and the stationary distribution across the hidden states B_i varies according to $\pi_{B_i} \propto \frac{i}{N}$. (This function was chosen randomly for illustrative purposes; other functions revealed similar curve shapes.) The hidden state space is fully connected and each state has equal kinetic rates to every other state. The information theoretic quantities E, h_μ, χ, C_μ vary systematically as a function of the number of hidden states N. The x-axis shows N and each plot shows one of these information theoretic quantities. E and h_μ are anti-correlated and peak/trough at finite $N = 10$ hidden states; the variation of C_μ and χ with N parallel that of E.

Clearly, one of the recurring expressions in this chain is $D^{-1}\left(I - \hat{e}_{1,1}\right)D$, which is

$$\left(D^{-1}\left(I - \hat{e}_{1,1}\right)D\right)_{ij} = \sum_{k,k'} D^{-1}_{ik}\left(I - \hat{e}_{1,1}\right)_{kk'} D_{k'j} \tag{119}$$

$$= \sum_{k,k'} \pi_i^{-1}\delta_{i,k}\delta_{k,k'}(1 - \delta_{k',1})\pi_j\delta_{k',j} \tag{120}$$

$$= (1 - \delta_{i,1})\delta_{i,j} \tag{121}$$

$$\Rightarrow D^{-1}(I - \hat{e}_{1,1})D = I - \hat{e}_{1,1}. \tag{122}$$

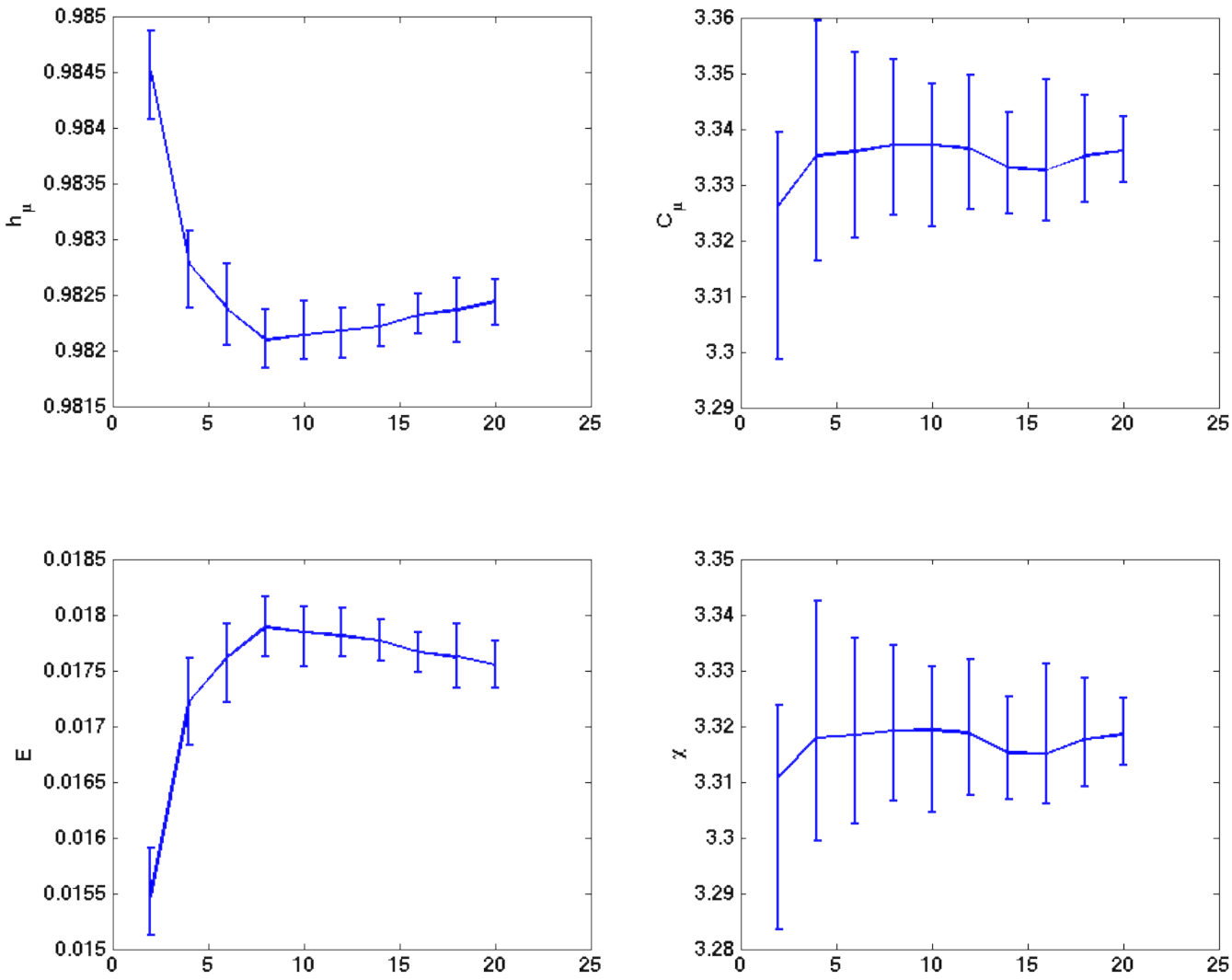

Figure 16: Again, π_A and N_{ab} are held constant and the stationary distribution across the hidden states B_i varies according to $\pi_{B_i} \propto \frac{i}{N}$. (This function was chosen randomly for illustrative purposes; other functions revealed similar curve shapes.) The hidden state space is fully connected and each state has equal kinetic rates to every other state. However, there is no noise as described in the text above in the kinetic rates. The information theoretic quantities $E,\ h_\mu,\ \chi,\ C_\mu$ vary systematically as a function of the number of hidden states N. The x-axis shows N and each plot shows one of these information theoretic quantities. The small amount of noise corrupts the signal from C_μ and χ more than it does for E or h_μ. Errorbars are not the typical standard deviations but denote, out of 15 samples, the lowest and highest observed values of C_μ and so on.

This relationship allows us to greatly simplify the expression for $\tilde{f}_n$ to

$$\tilde{f}_n = 1^\top \left(DT^\top D^{-1}(I - \hat{e}_{1,1})\right)^n \left(DT^\top D^{-1}\hat{e}_{1,1}\right)\pi = 1^\top D \left(T^\top(1 - \hat{e}_{1,1})\right)^n T^\top D^{-1}\hat{e}_{1,1}\pi. \qquad (123)$$

We can further simplify this using

$$\left(1^\top D\right)_i = \sum_j \pi_i \delta_{i,j} = \pi_i \Rightarrow 1^\top D = \pi^\top \tag{124}$$

and

$$\left(D^{-1}\hat{e}_{1,1}\pi\right)_i = \sum_{k,k'} \pi_k^{-1}\delta_{k,k'}\delta_{k',i}(1-\delta_{k',1})\pi_i = \hat{e}_1 = \hat{e}_{1,1}1. \tag{125}$$

This allows us to rewrite scalar $\tilde{f}_n$ as

$$\tilde{f}_n = \pi^\top \left(T^\top(1-\hat{e}_{1,1})\right)^n T^\top \hat{e}_{1,1}1. \tag{126}$$

Note that since this is a scalar, it is equal to its transpose:

$$\begin{aligned}
\tilde{f}_n = \tilde{f}_n^\top &= \left(\pi^\top \left(T^\top(I-\hat{e}_{1,1})\right)^n \left(T^\top \hat{e}_{1,1}\right) 1\right)^\top && (127)\\
&= 1^\top \left(T^\top \hat{e}_{1,1}\right)^\top \left(\left(T^\top(I-\hat{e}_{1,1})\right)^\top\right)^n \pi && (128)\\
&= 1^\top \left(\hat{e}_{1,1}T\right)\left((I-\hat{e}_{1,1})T\right)^n \pi && (129)\\
&= 1^\top T^{(0)} \left(T^{(1)}\right)^n \pi. && (130)
\end{aligned}$$

This is a very compact expression that is almost exactly equal to the expression for f_n, which is

$$f_n = 1^\top \left(T^{(1)}\right)^n T^{(0)}\pi. \tag{131}$$

The final trick to the proof is to show that, within this inner product, $\left(T^{(1)}\right)^n$, $T^{(0)}$ "commute". Noting that $T^{(0)} = T - T^{(1)}$,

$$\tilde{f}_n = 1^\top \left(T - T^{(1)}\right)\left(T^{(1)}\right)^n \pi = 1^\top T \left(T^{(1)}\right)^n \pi - 1^\top \left(T^{(1)}\right)^{n+1} \pi. \tag{132}$$

Since T is a stochastic transition matrix, $1^\top T = 1^\top$ (the columns must sum to 1) and also note that π can be just as easily replaced with $T\pi$, as π is the right eigenvector of T with eigenvalue 1:

$$\tilde{f}_n = 1^\top \left(T^{(1)}\right)^n T\pi - 1^\top \left(T^{(1)}\right)^{n+1} \pi = 1^\top \left(T^{(1)}\right)^n \left(T - T^{(1)}\right)\pi = 1^\top \left(T^{(1)}\right)^n T^{(0)}\pi. \tag{133}$$

But this is just the same as the expression for f_n! So

$$f_n = \tilde{f}_n \tag{134}$$

and equivalence between forward and reverse time epsilon machines follows trivially:

$$\pi_n = \tilde{\pi}_n,\ M_{n,n+1} = \tilde{M}_{n,n+1}. \tag{135}$$

5 Syncable non-binary nonunifilar HMMs

Another case to which the same basic methodology introduced in previous sections does apply almost trivially is that of non-binary, syncable nonunifilar word generators. In particular, assume now that there are many hidden states in grouping 0 and one state in groupings 1 through m. Transitions to a particular grouping lead to emission of that grouping's symbol. Therefore observation of any symbol that is not 0 leads to syncing, but observing different numbers of 0's correspond to a different mixed state, and the causal states are denoted as $y0^n$ where $y = 1, ..., m$ and n is any nonnegative integer. Let π denote the stationary distribution over

the internal (hidden) states and $T^{(x)}$ the transition matrix corresponding to emission of letter x. Let $\pi_{y,n}$ correspond to the stationary probability of causal state $...y0^n$. Any particular causal state $y0^n$ can transition to $y0^{n+1}$ or any of y', $y' \neq 0$. The probability of these transitions are

$$P(y0^n \to y0^{n+1}) = \frac{1^\top \left(T^{(0)}\right)^{n+1} T^{(y)}\pi}{1^\top \left(T^{(0)}\right)^{n} T^{(y)}\pi} \tag{136}$$

and

$$P(y0^n \to y') = \frac{1^\top T^{(y')} \left(T^{(0)}\right)^{n} T^{(y)}\pi}{1^\top \left(T^{(0)}\right)^{n} T^{(y)}\pi}. \tag{137}$$

To find the stationary probability distribution, we first consider the causal state $...y0^n$ where $n \geq 1$. All probability flows out of the state, and the only way to get to the state is to come from $...y0^{n-1}$, so probability flow balance implies

$$\pi_{y,n} = \frac{1^\top \left(T^{(0)}\right)^{n} T^{(y)}\pi}{1^\top \left(T^{(0)}\right)^{n-1} T^{(y)}\pi}\pi_{y,n-1} \tag{138}$$

which therefore implies that

$$\frac{\pi_{y,n}}{\pi_{y,0}} = \prod_{k=1}^{n} \frac{\pi_{y,k}}{\pi_{y,k-1}} = \prod_{k=1}^{n} \frac{1^\top \left(T^{(0)}\right)^{k} T^{(y)}\pi}{1^\top \left(T^{(0)}\right)^{k-1} T^{(y)}\pi} = \frac{1^\top \left(T^{(0)}\right)^{n} T^{(y)}\pi}{1^\top T^{(y)}\pi}. \tag{139}$$

The probability flow into and out of causal state $...y$ is more complicated:

$$\sum_{y' \neq 0} \sum_{n=0}^{\infty} P(y'0^n \to y)\pi_{y',n} = \pi_{y,0}. \tag{140}$$

Using eqns. 136-139, we can simplify the left-hand side of the above expression to

$$\begin{aligned}
\sum_{y' \neq 0} \sum_{n=0}^{\infty} P(y'0^n \to y)\pi_{y',n} &= \sum_{y' \neq 0} \sum_{n=0}^{\infty} \frac{1^\top T^{(y)} \left(T^{(0)}\right)^{n} T^{(y')}\pi}{1^\top \left(T^{(0)}\right)^{n} T^{(y')}\pi} \frac{1^\top \left(T^{(0)}\right)^{n} T^{(y')}\pi}{1^\top T^{(y')}\pi}\pi_{y',0} && (141)\\
&= \sum_{y' \neq 0} \sum_{n=0}^{\infty} \frac{1^\top T^{(y)} \left(T^{(0)}\right)^{n} T^{(y')}\pi}{1^\top T^{(y')}\pi}\pi_{y',0} && (142)\\
&= \sum_{y' \neq 0} \frac{1^\top T^{(y)} \left(I - T^{(0)}\right)^{-1} T^{(y')}\pi}{1^\top T^{(y')}\pi}\pi_{y',0} && (143)
\end{aligned}$$

and therefore eqn. 140 becomes

$$\pi_{y,0} = \sum_{y' \neq 0} \frac{1^\top T^{(y)} \left(I - T^{(0)}\right)^{-1} T^{(y')}\pi}{1^\top T^{(y')}\pi}\pi_{y',0}. \tag{144}$$

This equation holds for all $y \neq 0$, and hence we have a solvable linear system of m equations with m independent variables $\pi_{y,0}$, $y = 1, ..., m$. These probabilities are subject to a normalization constraint that

$$\begin{aligned}
1 &= \sum_{y \neq 0} \sum_{n=0}^{\infty} \pi_{y,n} = \sum_{y \neq 0} \pi_{y,0} \sum_{n=0}^{\infty} \frac{1^\top \left(T^{(0)}\right)^{n} T^{(y)}\pi}{1^\top T^{(y)}\pi} && (145)\\
&= \sum_{y \neq 0} \frac{1^\top \left(I - T^{(0)}\right)^{-1} T^{(y)}\pi}{1^\top T^{(y)}\pi}\pi_{y,0}. && (146)
\end{aligned}$$

From the solution, we can find C_μ^+ as the entropy of the stationary distribution:

$$C_\mu^+ = \sum_{y \neq 0} \sum_{n=0}^{\infty} \pi_{y,n} \log_2 \pi_{y,n}. \tag{147}$$

The entropy rate is also calculable from the fact that the entropy out of causal state $y0^n$ is

$$h_{y0^n} = H[\frac{1^\top \left(T^{(0)}\right)^{n+1} T^{(y)} \pi}{1^\top \left(T^{(0)}\right)^n T^{(y)} \pi}, \{\frac{1^\top T^{(y')} \left(T^{(0)}\right)^n T^{(y)} \pi}{1^\top \left(T^{(0)}\right)^n T^{(y)} \pi}\}_{y' \neq 0}] \tag{148}$$

and the total entropy rate is of course

$$h_\mu = \sum_{y=1,\ldots,x} \sum_{n=0}^{\infty} h_{y0^n} \pi_{y,n}. \tag{149}$$

These expressions are a bit difficult to work with. However, one quite interesting conclusion that arises from these type of formulae is that even though syncable nonunifilar binary word generators are causally reversible, these syncable nonunifilar ternary-or-greater word generators are not. For the purposes of elucidation, let's work through where the previously given proof fails for ternary-or-greater word generators. The reverse-time statistical complexity is the entropy of the reverse-time causal state stationary distribution $\tilde{\pi}_{y0^n}$ where

$$\tilde{\pi}_{y,n} = \frac{1^\top \left(\tilde{T}^{(0)}\right)^n \tilde{T}^{(y)} \pi}{1^\top T^{(y)} \pi} \tilde{\pi}_{y,0} \tag{150}$$

where

$$\tilde{\pi}_{y,0} = \sum_{y' \neq 0} \frac{1^\top \tilde{T}^{(y)} \left(I - \tilde{T}^{(0)}\right)^{-1} \tilde{T}^{(y')} \pi}{1^\top \tilde{T}^{(y')} \pi} \tilde{\pi}_{y',0}. \tag{151}$$

Let

$$f_{y,n} = 1^\top \left(T^{(0)}\right)^n T^{(y)} \pi, \;\; \tilde{f}_{y,n} = 1^\top \left(\tilde{T}^{(0)}\right)^n \tilde{T}^{(y)} \pi. \tag{152}$$

If we can show that $f_{y,n} = \tilde{f}_{y,n}$, then it might be that these ternary-or-greater word generators might be causally reversible. We would also need to show that the transition probabilities between causal states y and y' are equal to complete the proof, but this initial step fails. As discussed in Sec. 4.3, the formulae for $\tilde{T}^{(x)}$ are

$$\tilde{T}^{(x)} = D T^\top D^{-1} P_{(x)} \tag{153}$$

where $P_{(x)}$ is the projection matrix onto the subspace corresponding to emission of the letter x, e.g. $\hat{e}_{1,1}$ for the example shown previously in Sec. 4.3, and $D = \mathrm{diag}(\pi)$. It still holds that

$$D^{-1} P_{(x)} D = P_{(x)}, \;\; 1^\top D = \pi^\top, \;\; D^{-1} P_{(x)} \pi = P_{(x)} 1. \tag{154}$$

for the same reasons described in detail in Sec. 4.3. Therefore, for reasons also described in Sec. 4.3,

$$\tilde{f}_{y0^n} = \tilde{f}_{y0^n}^\top = 1^\top T^{(y)} \left(T^{(0)}\right)^n \pi. \tag{155}$$

Previously, we were able to claim that $T^{(y)} = T - T^{(0)}$. This is no longer true; now $\sum_{y \neq 0} T^{(y)} = T - T^{(0)}$. As such, the entire proof falls apart and there is no longer an argument for these systems being causally reversible in general.

6 Preliminary conclusions and future research

I guess I should point out that I haven't seen these closed form analytic results given anywhere else– now, there are closed form analytic expressions for the statistical complexity, entropy rate, crypticity, and (approximations and code for) the excess entropy of some simple binary nonunifilar word generators. And some of the work hints at an extension of the causal states to continuous time and the "specialness" of a binary partition with respect to causal irreversibility.

The obvious next step would be to consider a system in which grouping A has m states and grouping B has n states. Additionally we can assume that the hidden state space is fully connected. Unfortunately, for a system like this, each past has a different probability distribution over futures, which means that the minimal maximally predictive model involves storing all of the past. The reason for this is that, when m and n are both greater than 1, you can never sync to an internal mixed state unless you see an infinity of the same symbol. (At that point, you sync to the stationary probability distribution within group A or B.) If the hidden state space is not fully connected, then it is possible that two different pasts will lead to the same conditional probability distribution over futures, and therefore lead to some simplification in the causal state presentation.

References

[1] E. M. Izhikevich and G. M. Edelman, "Large-scale model of mammalian thalamocortical systems," *Proc. Nat. Acad. Sci*, **105** pp. 3593-3598 (2009).

[2] E. Schneidmann et al, "Weak pairwise correlations imply strongly correlated network states in a neural population," *Nature*, **440** pp. 1007-1012 (2006).

[3] F-C. Yeh et al, "Maximum Entropy Approaches to Living Neural Networks," *Entropy*, **12** pp. 89-106 (2010).

[4] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, **381** pp. 607-609 (1996).

[5] A. J. Bell and T. J. Sejnowski, "The 'independent components' of natural scenes are edge filters," *Vis. Res.*, **37** pp. 3327-3338 (1997).

[6] Y. Karklin and M. S. Lewicki, "Emergence of complex cell properties by learning to generalize in natural scenes," *Nature*, **457** pp. 83-86 (2009).

[7] I. E. Ohiorhenuan et al, "Sparse coding and high-order correlations in fine-scale cortical networks," *Nature*, **466** pp. 617-621 (2010).

[8] J. H. Macke et al, "Common input explains higher-order correlations and entropy in a simple model of neural population activity," *Phys. Rev. Lett.*, **106** (2011).

[9] J. Fournier et al, "Adaptation of the simple or complex nature of V1 receptive fields to visual statistics," *Nature Neuro.*, **14** pp. 1053-1060 (2011).

7 Appendix

I do not make any claims as to the efficiency or readability of the following Matlab code.

7.1 Code for the variation on the SNS

```
function [E,Cmu,hmu,xi]=SNS2_entropyrate_v2(p,q,L)
% more computational efficient version than v1.
% calculate entropy rate so you can subtract it off
```

```
% \pi_n.  you can make n a vector but p and q should be numbers
pin=@(n) p*q*(p*(1-q).^n-q*(1-p).^n)/(p^2-q^2);
pin2=@(n) 0.5*p*(1-p).^(n-1).*(1+(n-1)*p);

% entropy function
h=@(x) -x.*log2(x)-(1-x).*log2(1-x);

% transition probs amongst recurrent states, from sn-1 to sn:
T=@(n) (p*(1-q).^n-q*(1-p).^n)./(p*(1-q).^(n-1)-q*(1-p).^(n-1));
Tb=@(n) (1-p)*(1+(n-1)*p)./(1+(n-2)*p);
% transition probs amongst transient states, from tilde(s)n-1 to tilde(s)n:
T2=@(n) (p^2*(1-q).^n-q^2*(1-p).^n)./(p^2*(1-q).^(n-1)-q^2*(1-p).^(n-1));
T2b=@(n) (1-p)*(2+(n-2)*p)./(2+(n-3)*p);

x1=0:10000;
if p==q
    foo=h(Tb(x1+2));
    foo(isnan(foo))=0;
    hmu=sum(foo.*pin2(x1+1));

    foo=-[pin2(1) pin2(x1+1)].*log2([pin2(1) pin2(x1+1)]);
    foo(isnan(foo))=0;
    Cmu=sum(foo);
else
    foo=h(T(x1+2));
    foo(isnan(foo))=0;
    hmu=sum(foo.*pin(x1+1));

    foo=-[pin(1) pin(x1+1)].*log2([pin(1) pin(x1+1)]);
    foo(isnan(foo))=0;
    Cmu=sum(foo);
end

% calculate entropy rate estimates and sum them

x0=1; % \tilde{s}_{-1} initially has that much probability
%entrate=zeros(1,L-1);
xL=x0;
%plot(xL);
%hold on;
if p==q
    entrate(1)=h(T2b(1))*xL;
else
    entrate(1)=h(T2(1))*xL;
end

E=entrate(1)-hmu;

for i=2:L
    % hsig = [entropy of transient state s_{-i}, entropy rate of s_{0:i-2}]
    if p==q
```

```
        hsig=h([T2b(i) Tb(1:i-1)]);
        hsig(isnan(hsig))=0;
    else
        hsig=h([T2(i) T(1:i-1)]);
        hsig(isnan(hsig))=0;
    end

    x0=xL;
    %cla;
    %plot(xL);
    %pause(0.1);
    % calculate new mixed state
    if p==q
        % the top entry is now p(s_{-(i+1)})
        xL(1)=x0(1)*T2b(i-1);
        if i>2
            % the next entry is now p(s_{1:i})
            xL(3:i)=x0(2:i-1).*Tb(1:i-2);
            % the next entry is now p(s_0)
            xL(2)=x0(1)*(1-T2b(i-1));
            if i>3
                xL(2)=xL(2)+sum(x0(3:i-1).*(1-Tb(2:i-2)));
            end
        else
            % the next entry is now p(s_0)
            xL(2)=x0(1)*(1-T2b(i-1));
        end
    else
        % the top entry is now p(s_{-(i+1)})
        xL(1)=x0(1)*T2(i-1);
        if i>2
            % the next entry is now p(s_{1:i})
            xL(3:i)=x0(2:i-1).*T(1:i-2);
            % the next entry is now p(s_0)
            xL(2)=x0(1)*(1-T2(i-1));
            if i>3
                xL(2)=xL(2)+sum(x0(3:i-1).*(1-T(2:i-2)));
            end
        else
            % the next entry is now p(s_0)
            xL(2)=x0(1)*(1-T2(i-1));
        end
    end

    % calculate next point on entropy rate curve
    %entrate(i)=hsig*xL';
    %E=E+(entrate(i)-hmu);
    E=E+(hsig*xL'-hmu);
end
E=real(E);
%
```

```
% plot(0:L-1,entrate,'-o','LineWidth',2);
% set(gca,'FontSize',15,'FontName','Helvetica','FontWeight','bold');
% xlabel('L','FontSize',16,'FontName','Helvetica','FontWeight','bold');
% ylabel('h_{\mu}(L)','FontSize',16,'FontName','Helvetica','FontWeight','bold');

% crypticity estimate
xi=Cmu-E;

% hidden information estimate
% smu=E+hmu-h0;
```

7.2 Code for the sync-able nonunifilar binary word generators

```
function [hmu,Cmu,E,xi,smu]=GeneralSNS(kA,kB,N)
% N is number of hidden states
% dt is the time resolution
% hmu is entropy rate, Cmu is stat comp, E is excess entropy,
% xi is crypticity, Xi is causal reversibility
noisefrac=0.01;

% how the hidden state kinetic rates scale
f=@(i) i/N;
kB=kB*sum(f(1:N).^-1);

% set up the transition matrices T0 and T1
M=zeros(N+1);
% fill in A going to B's
M(:,1)=M(:,1)+(kA/N)*(1+noisefrac*randn(N+1,1));
%M(1:N+1,2:N+1)=M(1:N+1,2:N+1)+kB*(1+noisefrac*randn(N+1,N));
for i=2:N+1
    M(:,i)=(kB/N)*f(i-1)*(1+noisefrac*randn(N+1,1));
end
% to make sure the noise doesn't drive to negative kinetic rates
M(M<0)=0;
% to fill in the diagonals properly
for i=1:N+1
    M(i,i)=0;
    M(i,i)=-sum(M(:,i));
end

% get the full transition matrix
dt=1;
T=expm(M*dt);
% get T0 and T1
T0=zeros(N+1);
T1=zeros(N+1);
%
T0(1,:)=T(1,:);
T1(2:N+1,:)=T(2:N+1,:);
% break up into w and P
% w=T(2:N+1,1);
% P=T(2:N+1,2:N+1);
```

```
% get stationary distribution
[V,D]=eig(T);
[~,ind]=min((sum(D)-1).^2);
Pi=V(:,ind);
Pi=Pi/sum(Pi);

v=T0*Pi;
Z=sum(inv(eye(N+1)-T1)*v);
pin=@(n) sum(T1^n*v)/Z;
T00=T(1,1);
Tin=@(n) sum(T1^n*v)/sum(T1^(n-1)*v);
T2in=@(n) sum(T1^n*Pi)/sum(T1^(n-1)*Pi);

% how far do you want to go out?
% keep on getting values until you find that the entropy is 1/100th the
% contribution of pi0

ent=@(t) -t.*log2(t)-(1-t).*log2(1-t);

x=ent(pin(0));
Cmu=x;

hmu=pin(0)*ent(T00);

% build up a lookup table of transition probabilities
trans=1-T00;

k=1;
while x>10^-9*Cmu
    x=ent(pin(k));
    if isnan(x)
        x=0;
    end
    Cmu=Cmu+x;

    trans=[trans Tin(k+1)];
    foo=ent(trans(end));
    if isnan(foo)
        foo=0;
    end
    hmu=hmu+pin(k)*foo;

    k=k+1;
end

% Finding excess entropy
% calculate entropy rate estimates and sum them

x0=1; % \tilde{s}_{-1} initially has that much probability
%entrate=zeros(1,k);
```

```
xL=x0;
hsig=ent(T2in(1));
h0=hsig*xL;

E=h0-hmu;

for i=2:k
    % hsig = [entropy of transient state s_{-i}, entropy rate of s_{0:i-2}]

    hsig(1)=ent(T2in(i));
    hsig(i)=ent(Tin(i-1));
    hsig(isnan(hsig))=0;

    x0=xL;
    % calculate new mixed state

    % the top entry is now p(s_{-(i+1)})
    xL(1)=x0(1)*T2in(i-1);
    % the next entry is now p(s_0)
    %xL(2)=x0(1)*(1-T2in(i-1))
    xL(2)=x0(1)*(1-T2in(i-1))+sum(x0(2:i-1).*(1-trans(1:i-2)));
%     if i>2
%         %xL(2)=xL(2)+x0(2)*(1-Tin(1));
%         %xL(3)=x0(2)*Tin(1);
%         for j=3:i-1
%             xL(2)=xL(2)+x0(j).*(1-Tin(j-1));
%             xL(j)=x0(j-1)*Tin(j-2);
%         end
%         xL(i)=x0(i-1)*Tin(i-2);
%     end
    xL(3:i)=x0(2:i-1).*trans(1:i-2);

    % calculate next point on entropy rate curve
    % entrate(i)=hsig*xL';
    E=E+(hsig*xL'-hmu);
end
E=real(E);
%
% figure, plot(0:k-1,entrate,'-o','LineWidth',2);
% set(gca,'FontSize',15,'FontName','Helvetica','FontWeight','bold');
% xlabel('L','FontSize',16,'FontName','Helvetica','FontWeight','bold');
% ylabel('h_{\mu}(L)','FontSize',16,'FontName','Helvetica','FontWeight','bold');

% crypticity estimate
xi=Cmu-E;

% hidden information estimate
smu=E+hmu-h0;

% Switch to the time-reversed transition matrices
% Tback=T;
```

```
% for i=1:N+1
%     for j=1:N+1
%         Tback(i,j)=T(j,i)*Pi(i)/Pi(j);
%     end
% end
% T0=zeros(N+1);
% T1=zeros(N+1);
% T0(:,1)=Tback(:,1);
% T1(:,2:N+1)=Tback(:,2:N+1);
%
% v=T0*Pi;
% Z=sum(inv(eye(N+1)-T1)*v);
% pin=@(n) sum(T1^n*v)/Z;
%
% x=ent(pin(0));
% Cmu2=x;
%
% k=1;
% while x>10^-9*Cmu2
%     x=ent(pin(k));
%     if isnan(x)
%         x=0;
%     end
%     Cmu2=Cmu2+x;
%
%     k=k+1;
% end
```