Beyond the typical set:
Fluctuations in Intrinsic Computation

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

(Dated: June 11, 2015)

We show how to calculate the spectrum of statistical fluctuations in structured nonequilibrium steady states (SNESSs)—viz., memoryful, stationary processes generated by hidden Markov models—using their ϵ-machine presentations. We review basic fluctuation theory, drawing out parallels between statistical mechanics, information theory, and large deviations. To analyze the interaction between statistical fluctuations and process structure, we introduce the thermodynamic spectra of statistical complexity and excess entropy—structural properties that complement the oft-used spectrum of Renyi entropy rate that monitors fluctuations in information production. We show that Renyi entropy itself decomposes into two spectra that monitor rates of information loss and accumulation. In particular, we fully characterize the range of possible SNESS thermodynamic entropy-energy functions, giving new criteria for when a process's ground state has positive entropy rate. We explore fluctuations in SNESSs that are (i) maximum entropy rate, (ii) causally irreversible, (iii) nonergodic, or (iv) infinite memory. The result is a constructive and comprehensive picture of fluctuations in information processing and the balance of information generation and storage that a given stochastic process achieves.

I. INTRODUCTION

Statistical fluctuations are to be expected in finite-size, finite-dimensional processes and so, too, in SNESS. Thus, the methods and results will be especially helpful as one scales down to implement computing on increasingly smaller-scale physical substrates. Moreover, they introduce a new kind of time-series prediction that extrapolates a process's typical behaviors to determine the likelihood of extremely low probability of events—events not seen before.

Recall Bennett's "Thermodynamics of Computation" [1], which is largely an exposition, with little or no theoretical development, on why it is an important and long-standing goal to view information processing in natural systems. Here, we lay out one part of a statistical mechanics framework for this that allows one to analyze a wide class of complex nonlinear processes in terms of the range or fluctuations in information processing they exhibit.

A complementary statistical mechanics approach identifies emergent macrostates [2–4].

Recall Bowen, Ruelle, and others' work in abstract dynamical systems to understand the complex and strange invariant sets they generate via the "thermodynamic formalism" [5–8]. The idea there was to adapt statistical physics to describe the complex temporal behavior and associated invariant state-sets. In a sense, a dynamical system unfolding in time is considered to be a spin-like system in space—the lattice of time becomes the latter's spatial lattice.

Here, we'd like to provide a constructive framework for that formalism. Refs. [9], [10], and [11], and [12] are readable introductions, in this spirit. The particular emphasis that is new here is the focus on measures of organization and structure and how these affect the thermodynamics of information processing, especially how fluctuations are affected in processes with memory.

This particular emphasis goes under the rubric of *intrinsic computation*—that all systems store, transform, and dissipate information [13].

Real world connection: Technology builds ever-smaller computational devices commandeering nonlinear physical phenomena on increasingly smaller spatial scale and shorter time scale processes.

We assume we have a canonical model of the behavior of such an engineered device. We will use ϵ-machines as the canonical representation [13, 14]. It is the (unifilar) minimal open predictor and it allows one, as we will demonstrate, to calculate many of the device's information processing properties.

Or, we may wish to analyze a naturally occurring nanoscale thermodynamic system—such as a "smart" biomolecule that does useful work, like kinesin's role in intracellular transport—in terms of how it stores and processes information.

In either case, we start with the system's ϵ-machine.

Say what we mean by "fluctuation": Sample variation in event probabilities. The reference for this are the *typical sets* of information theory: The events that one typically sees. Events and their probabilities lying outside this set are fluctuations or, sometimes, "deviations".

Real world connection: Small scale thermodynamic systems are typically dominated by fluctuations. Here, we consider intrinsically generated fluctuations that are

* caghamohammadi@ucdavis.edu

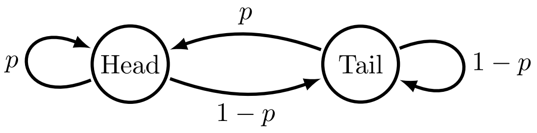

FIG. 1. Markov chain presentation for the Biased Coin Process with bias p.

part and parcel of the stochastic process.

What tools are available to analyze statistical fluctuations?

Recently, the fluctuation relations of Gallavotti, Cohen, Jarzynski, and Crooks [15–19].

Large deviation theory [20] is a relatively new and necessary tool that gives insight into the full range of statistical fluctuations, in particular those well outside the domain of the Law of Large Numbers.

Presaged by the theory of types (Shannon-McMillman-Breiman theory) in information theory [21].

And by the thermodynamic formalism [6] of dynamical systems theory.

But, in truth, the basic ideas reside in Gibb's original ensemble formulation of statistical mechanics. And this was pointed out in Ref. [22].

Given that an ϵ-machine is the minimal sufficient statistic for a given process, in principle every quantity is calculable from it. Here, we show how to calculate the full spectrum of fluctuations from a process's ϵ-machine. The practical result is a set of methods that allow one to efficiently and directly calculate a process's spectrum of fluctuations. Conceptually, the result is a new view of how process information and structure are two, and necessarily complementary, aspects of fluctuation phenomena.

II. WHAT'S THE ISSUE?

Consider the lowly biased coin. This process is generated by the Markov chain shown in the Fig. 1.

After running the machine, it generates a sequence of Heads and Tails. Here we would like to study that the sequences and fluctuations in the probabilities. One simple way to study this is to plot word distribution histograms for ℓ-length words.

To study the words with the length n in the time series let's consider a one to one map between y_i^n and q_i^n which is defined by:

$$q_i^n = \sum_{j=0}^{n} x_{i+j} 2^j$$

With this definition we study q_i^n instead of y_i^n without losing any generality. Figure (2) shows histogram estimated from generated time series.

Give as thorough a direct introduction to thermodynamics as one can for the Biased Coin.

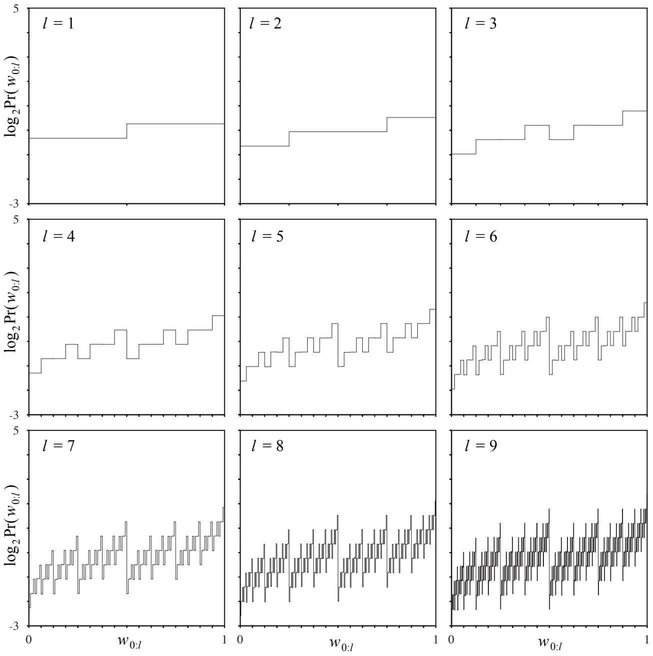

FIG. 2. Biased Coin Process word distributions $\Pr(w_{0:\ell})$ for $\ell = 1, 2, \ldots, 9$. Histograms of word counts are plotted by on the unit interval by metrizing sequences: Heads and Tails mapped to 1s and 0s, respectively, and translated to $x \in [0, 1]$ via $x = \sum_0^{\ell-1} w_i 2^{-i-1}$. Bias is $p = 0.6$.

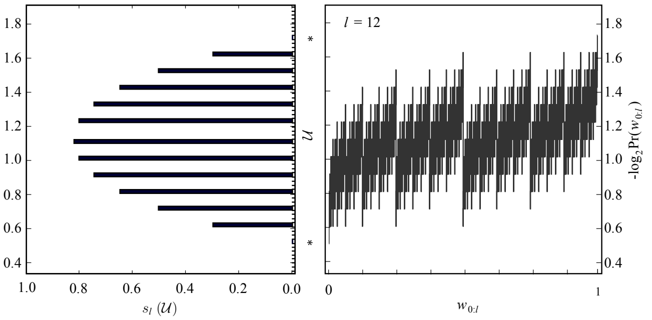

FIG. 3. Biased Coin Process statistical fluctuation histogram construction: energy levels and how histogram envelope approximates $S(U)$. Bias is $p = 0.3$, word length $\ell = 12$. Although there are $4096 = 2^\ell$ sequences, there are only 11 distinct sequence probabilities: $p^n(1-p)^{(\ell-n)}$, $n = 1, \ldots, 11$. The lower asterisk locates the single most probable sequence of all Tails, which has the lowest energy. The upper one, the least likely sequence of all Heads, which has the highest energy.

The biased coin, however, is an unstructured process. As all IID processes are.

Lead into structured processes: Use the Golden Mean Process. See the Markov chain in Fig. 4.

See the word distributions in Fig. 5.

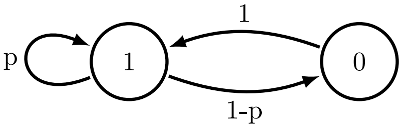

FIG. 4. Markov chain that generates the Golden Mean Process—a binary process with no consecutive 0s.

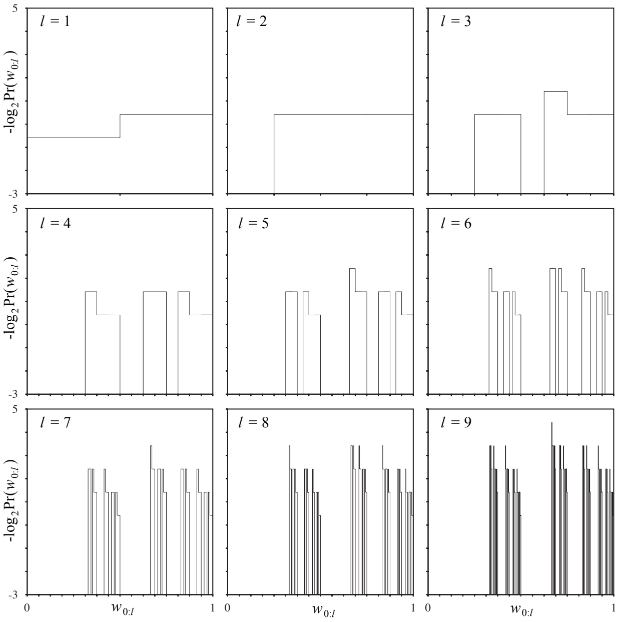

FIG. 5. Golden Mean Process word distributions $\Pr(w_{0:\ell})$ for $\ell = 1, 2, \ldots, 9$. 10^7 iterates and self-loop transition probability $p = 0.6$.

Structure corresponds to restrictions. In the case of the Golden Mean Process consecutive 0s are not allowed, otherwise all sequences are generated.

Thus, a central question here is how this kind of structure interacts with that seen in the sequence probabilities fluctuations. Fractal structure occurs in both.

The development is broken into four parts: intrinsic computation, thermodynamics of computation, fluctuation spectra, and large deviations. It ends with extensions and concluding remarks.

III. INTRINSIC COMPUTATION

A. Background

We assume the reader has basic knowledge of thermodynamics and statistical mechanics, information theory, and large deviations, such as found in the introductory chapters of Refs. [23], [21, esp. Ch. 11], and [20], respectively. The extension of basic information theory to complex processes is reviewed in Ref. [24]. Our development makes particular use of ϵ-machines, a canonical representation of a process that makes many properties directly and easily calculable; for a review see Ref. [14, and citations therein]. The current development reviews and then updates and extends the analysis of fluctuations previously developed in Ref. [25].

1. Processes

We denote subsequences in a time series as $X_{a:b}$, where $a \leq b$, to refer to the random variable chain $X_a X_{a+1} X_{a+2} \cdots X_{b-1}$, which has length $b - a$. We drop an index when it is infinite. For example, the *past* $X_{-\infty:0}$ is denoted $X_{:0}$ and the *future* $X_{0:\infty}$ is $X_{0:}$. We generally use $w = w_0 w_1 \ldots w_{\ell-1}$ to denote a particular realization or *word*—a sequence of symbols w_i drawn from a finite alphabet $w_i \in \mathcal{A}$. All of the words w of length ℓ are those $w \in \mathcal{A}^\ell$. We place two words, u and v, adjacent to each other to denote concatenation: $w = uv$.

Let's define the distribution of words $\Pr(w)$ over $\mathcal{A}^\infty$, assuming we have infinite number of measurements. Given a specific word w, we have its associated set of infinite sequences, its *ℓ-cylinder*:

$$s_w = \{x_: : x_0 = w_0, ..., x_{\ell-1} = w_{\ell-1} \ , \ x_: \in \mathcal{A}^\infty\} \ .$$

The collection of all cylinders at length ℓ is:

$$s^\ell = \bigcup_{w \in \mathcal{A}^\infty} s_w \ .$$

The probability of specific word w, then, is:

$$\Pr(w) = \frac{\|s_w\|}{\|s^\ell\|} \ . \qquad (1)$$

Given a finite data stream $x_{0:k-1}$, $k \gg \ell$, one estimator for $\Pr(w)$ is:

$$\Pr(w) \approx \frac{N(w)}{k - \ell + 1} \ ,$$

where $N(w)$ is the number of occurrences of w in x.

Informally, a *process* is a joint probability distribution $\Pr(X_:)$ over the bi-infinite chain $X_: = X_{:0} X_{0:}$. Formally, it is the probability space $\left(\mathcal{A}^\infty, \Sigma, \mathbb{P}\right)$, where Σ is the σ-algebra generated by the cylinder sets in $\mathcal{A}^\infty$ and $\mathbb{P}$ is the measure defined by Eq. (1).

The finitary processes we consider are useful descriptions of a wide range of systems: spin systems, symbolic dynamics of chaotic dynamical systems, and coarse-grained continuous systems, to mention a few broad classes.

2. ϵ-Machine Presentations

A *presentation* of a given process is any state-based representation that *generates* the process: it produces all of and only the process's words and their probabilities. In the following we consider processes generated by finite hidden Markov models (HMMs). For a given process, while there may be many alternative HMMs, there is a unique, canonical presentation—the process's ϵ*-machine*. The recurrent states $\boldsymbol{\mathcal{S}}$ of a process's ϵ-machine are known as the *causal states* $\sigma \in \boldsymbol{\mathcal{S}}$ and, at time t, the associated random variable is $\mathcal{S}_t$. The causal states are the minimal sufficient statistic of the past $X_{:0}$ for predicting the future $X_{0:}$.

An ϵ-machine is a type of HMM satisfying three conditions: unifilarity, probabilistically distinct states, and irreducibility. *Unifilarity* means that from each state σ there is at most one next state reached on a given symbol x [27]. *Probabilistically distinct states* means that for every pair of states—say, σ_1 and σ_2—there is at least one word w for which the probabilities of observing w starting from those states differ: $\Pr(w|\sigma_1) \neq \Pr(w|\sigma_2)$. *Irreducibility* implies that the internal Markov chain over the causal states is strongly connected and minimal in the sense that it is not possible to make a smaller unifilar HMM that generates the process. For a thorough treatment on presentations and ϵ-machines see Ref. [28].

When an ϵ-machine has a finite or countable number of causal states it is often helpful to represent it in a state transition diagram, as shown in Fig. 6, consisting of a set of states connected by directed, labeled transitions. There is a unique initial (or *start*) state depicted with two concentric circles, a set of transient states that have asymptotic probability zero, and set of recurrent states with positive asymptotic probability. For each transition between states σ_i and σ_j, there is a transition probability $\Pr(\sigma_j|\sigma_i)$ that gives the probability of going from state σ_i to state σ_j. On making the transition, the process emits symbol $x_{ij} \in \mathcal{A}$. A realization $x_0 x_1 \ldots x_t \ldots$ is generated starting in state σ_0 and following transitions according to the specified probabilities and emitting the symbols x_t labeling the transitions visited.

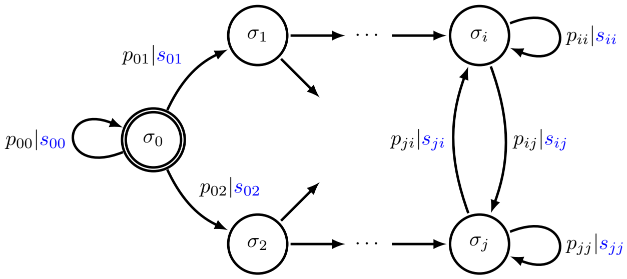

FIG. 6. State transition diagram for an ϵ-machine depicting the unique start state (double circled σ_0), a set of transient states $\{\sigma_0, \sigma_1, \sigma_2, \ldots\}$, and a set of recurrent states $\{\sigma_i, \sigma_j, \ldots\}$. Transition labels $p|x$ denote the probability p of taking the transition and emitting symbol x.

At time t our knowledge of a process's internal state is given by the state distribution:

$$\langle \eta_t | = \langle \Pr(\sigma_0), \Pr(\sigma_1), \ \ldots \ , \Pr(\sigma_N) | \ .$$

As an alternative to starting in σ_0, we can also specify an initial state distribution η_0. A stochastic transition matrix describes the state-to-state transitions as a Markov chain:

$$T := \sum_{x \in \mathcal{A}} T^{(x)} \ , \tag{2}$$

where the transitions on a given symbol x are:

$$T^{(x)}_{ij} = \Pr(\sigma_j, x|\sigma_i) \in [0,1] \ ,$$

for $i,j = 1, \ldots, N$. The transition probabilities are normalized. That is, the transition matrix T is *row-stochastic* [29]:

$$\sum_{j=1}^{N} \sum_{x \in \mathcal{A}} \Pr(\sigma_j, x|\sigma_i) = 1 \ .$$

Its component matrices $T^{(x)}_{ij}$ are said to be *substochastic*. Unifilarity means that at most one component in a row of $T^{(x)}_{ij}$ is nonzero and that $\Pr(\sigma_j, x|\sigma_i) = \Pr(x|\sigma_i)$.

An ϵ-machine's *connection matrix* $\mathbf{T_0}$ has components $\left(\mathbf{T_0}\right)_{ij}$ that give the number of transitions from state σ_i to σ_j.

By way of summarizing, we have the main object that generates a process.

Definition 1. *The* ϵ-machine M *is the set* $\big\{\boldsymbol{\mathcal{S}}, \{T^{(x)}, x \in \mathcal{A}\}, \langle \eta_0 | \, \big\}$.

Our goal is to understand the statistical structure of a process's sequences in terms of its ϵ-machine's calculable properties.

3. Stationarity

Since T is stochastic its principal eigenvalue is unity and so state distributions evolve according to:

$$\langle \eta_t | = \langle \eta_{t-1} | \, T \ , \tag{3}$$

where T is the time evolution operator of Eq. (2). Often we are interested in the asymptotic invariant solution $\langle \pi |$ of Eq. (3)—the eigenvector of T associated with eigenvalue $\lambda = 1$. For general such transition matrices, the all eigenvalues have magnitude less than or equal to one.

At large times and for the class of processes here, starting with any initial $\langle \eta_0 |$, the system approaches the invariant distribution $\langle \pi |$. If one starts an ϵ-machine in state distribution $\langle \pi |$, then the process generated is stationary: $\Pr(X_{t:t+\ell}) = \Pr(X_{0:\ell})$, for all t and ℓ.

For a given process, we calculate the word probabilities in terms of it's ϵ-machine:

$$\begin{aligned} \Pr(w) &= \Pr(x_{0:\ell}) \\ &= \langle \pi | \, T^{(x_0)} T^{(x_1)} \cdots T^{(x_{\ell-1})} \, | 1 \rangle \ , \end{aligned} \tag{4}$$

where $|1\rangle$ is a vector whose elements are one.

B. Information Measures

Now let's review several important information measures that elucidate the various kinds of complexity generated by SNESSs.

1. Topological Entropy

Let $N(\ell)$ denote the number of length-ℓ words $w \in \mathcal{A}^\ell$ a given process generates: $\Pr(w) > 0$. If all possible distinct words of length of ℓ are generated, then $N(\ell) = |\mathcal{A}|^\ell$. For typical processes there are restrictions and so not all words are generated. In this case, $N(\ell)$ scales differently and this is captured by its growth rate—the *topological entropy*:

$$h := \lim_{\ell \to \infty} \frac{\log_2 N(\ell)}{\ell} \ . \tag{5}$$

Simply said, we have the scaling $N(\ell) \propto |\mathcal{A}|^{h\ell}$.

2. Shannon Information Measures

Rather than simply counting words, we can examine how much information they carry. Shannon defined the *self-information* of an event, such as the occurrence of word w, as $-\log_2 \Pr(w)$. This leads to the total amount of information in length-ℓ words, the *block entropy*:

$$\mathrm{H}(\ell) := - \sum_{\{w \in \mathcal{A}^\ell\}} \Pr(w) \log_2 \Pr(w) \ . \tag{6}$$

Note that if the words that occur are equally likely, then $\mathrm{H}(\ell) = \log_2 N(\ell)$. A process's *Shannon entropy rate* is speed at which the block entropy grows:

$$h_\mu := \lim_{\ell \to \infty} \frac{\mathrm{H}(\ell)}{\ell} \ . \tag{7}$$

Paralleling the topological entropy, the block entropy scales as $\mathrm{H}(\ell) \propto h_\mu \ell$. h_μ measures a process's rate of information production and, so, its degree of randomness. If the process is generated by a dynamical system, then h_μ is the *Kolmogorov-Sinai entropy* or metric entropy [31–33].

The Shannon-McMillan-Breimen theorem indicates why the Shannon entropy rate is an important process characteristic [21]. It governs the exponential decay of the probability of typical realizations: $\Pr(w) \propto |\mathcal{A}|^{-\mathrm{H}(\ell)} \propto |\mathcal{A}|^{-h_\mu \ell}$. We return to this topic later in when discussing the typical behaviors of processes and deviations from them.

3. Predictable Information

Complementary to production or randomness rate h_μ, another key Shannon measure is the amount of information that a process communicates from its past to its future. In other words, we view a process as a communication channel: The system in the present moment communicates information from its past to its future. Information-theoretically we measure this as a mutual information between the past and future with the *excess entropy*:

$$\mathbf{E} = \mathrm{I}[X_{:0}; X_{0:}] \ . \tag{8}$$

$\mathbf{E}$ is an information transmission rate through a system when the past is taken as the channel input.

4. Information Relations

Together, the various information measures give a rather complete description of the kinds of intrinsic computation embedded in a process. However, they are not independent. First, note that entropy rate h_μ is equivalent to:

$$h_\mu = \mathrm{H}[X_0 | X_{:0}] \ .$$

And, this form makes clear it's interpretation as the instantaneous Shannon information (surprise) generated in the present.

Recently, using this form Ref. [34] introduced a new and functional decomposition of h_μ into a component—the *ephemeral information* r_μ—giving the part of the generated information dissipation and a component—the *bound information* b_μ—giving the part that is actively stored. In short:

$$h_\mu = r_\mu + b_\mu \ ,$$

where $r_\mu = H[X_0 | X_{:0}, X_{1:}]$ and $b_\mu = \mathrm{I}[X_0; X_{1:} | X_{:0}]$. Thus, some of the information generated (h_μ) in the present is dissipated (r_μ) and some (b_μ) is actively stored. In particular, the excess entropy decomposes into the atoms:

$$\mathbf{E} = b_\mu + q_\mu + \sigma_\mu \ ,$$

giving a more refined understanding of its constituents.

C. The ϵ-Machine Mesoscale: Where Information and Structure Meet

We now show that the preceding macroscopic measures can be directly calculated from a process's ϵ-machine. We state the main results, relegating their proofs to Appendices A 1.

General points: ϵ-machine causal states capture the emergent structure in microstates. In that sense, one can interpret them as capturing macroscopic properties. We prefer and find it less terminologically confusing to think of the ϵ-machine as describing an intermediate—*mesoscopic*—level of system organization. This allows us room to still make distinctions with the traditional macroscale of temperature, pressure, and free energies. There will also be new macroscopic variables: C_μ, $\mathbf{E}$, and the like.

It can be shown that [35] a process's topological entropy is:

$$h = \log_2 \lambda_{\max} \ , \tag{9}$$

where $\lambda_{\max}$ is the largest eigenvalue of its ϵ-machine's connection matrix $\mathbf{T_0}$.

Theorem 1. *A process' Shannon entropy rate can be directly calculated from its ϵ-machine using:*

$$h_\mu = - \sum_{i=1}^{N} \Pr(\sigma_i) \sum_{x \in \mathcal{A}} \Pr(x|\sigma_i) \log_2 \Pr(x|\sigma_i) \ . \tag{10}$$

Proof. *See App. A 1.*

Although we can now calculate them from a process's ϵ-machine, these rates were initially defined in terms of process realizations and their probabilities. They are, in this sense, statistics of the observable process. However, a process's ϵ-machine also gives insight into a process' internal structural organization. For example, there is a certain amount of memory stored by a process in its causal states related to how much if its past it remembers. This is measured by the *statistical complexity* C_μ:

$$C_\mu = - \sum_{i=1}^{N} \Pr(\sigma_i) \log_2 \Pr(\sigma_i) \ , \tag{11}$$

which is the amount of Shannon information in the causal state distribution.

Finally, the past-future mutual information, which otherwise is a complicated quantity involving the semi-finite chain of past $X_{:0}$ and future $X_{0:}$ random variables, receives a particularly elegant form when expressed in terms of the ϵ-machine. Specifically, it was recently shown that $\mathbf{E}$ is the mutual information between the forward-process and reverse-process ϵ-machines [36–38]:

$$\mathbf{E} = \mathrm{I}[\mathcal{S}^-;\mathcal{S}^+] \ , \tag{12}$$

where, in effect, we substituted the forward and reverse causal states for the semi-infinite past and future in Eq. 8.

Finally, we use the ϵ-machine to calculate the ephemeral and bound informations:

$$r_\mu = H[X_0|\mathcal{S}_0^-,\mathcal{S}_1^+] \tag{13}$$

and:

$$b_\mu = \mathrm{I}[X_0;\mathcal{S}_1^+|\mathcal{S}_0^-] \ , \tag{14}$$

where X_0 is the present random variable, $\mathcal{S}_1^+$ is the causal state at time $t = 1$, and the other (reverse-time) causal-state variable is anchored at time $t = 0$. These information measures require the joint distribution $\Pr(\mathcal{S}_0^-, X_0, \mathcal{S}_1^+)$, which is readily determined from a process' *bimachine* [39].

Although we don't avail ourselves directly of it, a new approach was recently developed that gives closed-form expressions for the excess entropy and the other informational measures in terms of the ϵ-machine's spectral decomposition [40].

IV. THERMODYNAMICS OF INTRINSIC COMPUTATION

In contrast to the informational view of a system, thermodynamics largely concerns how various kinds of macroscopic system energy are transformed to and from uncontrollable, unmeasurable thermal energy at microscopic scales. A given system has a total amount U of energy, only some of which is thermal (TS). To the extent there is a difference—some of the total energy is not thermalized—then there is a possibility of recovering some for use on the macroscale. This difference is measured by the *free energy* $F = U - TS$.

The abiding question in classical thermodynamics then boils down to describing thermalized energy. The main answer is that it is, in some way, energy in degrees of freedom that behave in a disordered or randomized manner. (However, those descriptors rather beg the question.)

The mapping we make to temporal systems is that individual sequences are system configurations and configuration energies are directly determined by sequence probabilities. Then, we ask for the corresponding macroscopic quantities—such as total energy, entropy density, and free energy. System size is sequence length ℓ and sequences w are microstates—microscopic configurations.

A. Energy and Entropy Density

To each word $w \in \mathcal{A}^\ell$ one associates an energy density:

$$U_w^\ell := \frac{-\log_2 \Pr(w)}{\ell} \ , \tag{15}$$

mirroring the Boltzmann weight common in statistical physics: $\Pr(w) \propto e^{-U(w)}$. The total energy in a system of size ℓ is $U^\ell = \sum_{w:\ \Pr(w)>0} U^\ell_w$. In this setting, disallowed words ($\Pr(w) = 0$) have infinite energy.

Naturally, different words w and v may lead to same energy density, $U^\ell_w = U^\ell_v$. And so, in the set $U^\ell = \left\{U^\ell_w : w \in \mathcal{A}^\ell\right\}$, energy values may appear repeatedly. Let's denote the frequency of equal U^ℓ_ws by $N(U^\ell_w)$. Then, for the thermodynamic macrostate at energy U, we define the *thermodynamic entropy density*:

$$S(U) := \lim_{\ell\to\infty} \frac{\log_2 N(U^\ell_w = U)}{\ell} \qquad (16)$$

to monitor the range and likelihood of accessible energies (or allowed words). This definition closely mirrors that in the standard statistical physics, where the thermodynamic entropy is proportional to the logarithm of number of accessible microstates. Energy in this formal setting is a proxy for parametrizing *classes* of equal-probability sequences.

B. Renyi Entropy Rate and Partition Function

Shannon block entropy is a linear average of the self-informations $-\log_2 \Pr(w)$. The *Renyi block entropy* is an extension that is the most general entropy that is both additive over independent distributions and a geometric average [41]:

$$\mathrm{H}_\beta(\ell) := \mathrm{H}_\beta[X_{0:\ell}] \qquad (17)$$

$$:= \frac{1}{1-\beta} \log_2 \sum_{\{w\in\mathcal{A}^\ell\}} (\Pr(w))^\beta \ , \qquad (18)$$

where β is an arbitrary real number that "focuses" on word subsets parametrized by probability (or energies $-\log \Pr(w)$). In this, we see that β is analogous to inverse temperature and we can interpret the sum as the *partition function*:

$$\mathcal{Z}(\beta) = \sum_{\{w\in\mathcal{A}^\ell\}} e^{-\beta(-\ln \Pr(w))} \ . \qquad (19)$$

Paralleling h_μ, we have the block entropy growth rate, the *Renyi entropy rate*:

$$\mathsf{h}(\beta) := \lim_{L\to\infty} \frac{\mathrm{H}_\beta(\ell)}{\ell} \ . \qquad (20)$$

C. Thermodynamic Relations

The thermodynamic measures are closely interrelated and those relations tell us much about a process's structure and randomness. We review well known properties, adapted to processes. Appendices A 2, A 3, and A 4 give proofs.

Lemma 1. *The topological and Shannon entropy rates are special cases of the Renyi entropy rate:*

$$h = \mathsf{h}(\beta = 0)$$

and

$$h_\mu = \mathsf{h}(\beta \to 1) \ ,$$

respectively.

Proof. *See App. A 2.*

Lemma 2. *The Renyi entropy rate* $\mathsf{h}(\beta)$ *and thermodynamic entropy density* $S(U)$ *are related by:*

$$S(U(\beta)) = \beta U(\beta) - (\beta - 1)\mathsf{h}(\beta) \ , \qquad (21)$$

where:

$$U(\beta) = \underset{u\in U^\infty}{\operatorname{argmax}} \ (S(u) - \beta u) \ . \qquad (22)$$

Proof. *See App. A 3.*

Lemma 3. *The energy density and Renyi entropy are related by:*

$$U(\beta) = \frac{\partial}{\partial\beta}((\beta - 1)\mathsf{h}(\beta)) \ . \qquad (23)$$

Proof. *See App. A 4.*

There are two important cases that help interpret the entropy $S(U)$. First:

$$S(U(\beta = 0)) = \mathsf{h}(0)$$

and second:

$$S(U(\beta = 1)) = h_\mu \ .$$

Later we show that $\mathrm{d}U/\mathrm{d}\beta \leq 0$, concluding that the inverse function $\beta(U)$ exists. And so, we can define free energy density via a Legendre transform of the thermodynamic entropy density:

$$-\beta F(\beta) = S(U) - U(\beta(U)) \ . \qquad (24)$$

immediately the direct consequence of 2 would be

$$F(\beta) = -\beta^{-1} \log \widehat{\lambda}_\beta \ . \qquad (25)$$

D. How to calculate fluctuation spectra

For a given ϵ-*machine* calculating fluctuation spectra from eq.16 generally is a hard thing to do. Here we build a tool to make this task much easier.

In section IV C we parameterized S and U in terms of β and that means for a given process for every real β there exist a unique U and it's S.

In a next lemma we will show that $S(U(\beta))$ is nothing other than metric entropy over the twisted distribution.

Lemma 4. *$S(U(\beta))$ in Eq. (2) can be written as:*

$$S(U(\beta)) = -\lim_{\ell\to\infty}\frac{1}{\ell}\sum_{\{w\in\mathcal{A}^\ell\}}\mathcal{Q}_\beta(w)\log\mathcal{Q}_\beta(w)\ .$$

with twisted distribution*:*

$$\mathcal{Q}_\beta(w) = \frac{(\Pr(w))^\beta}{\mathcal{Z}(\beta)}\ .$$

Proof. *Using Eqs. (23) and (A5) we have:*

$$U(\beta) = -\lim_{\ell\to\infty}\frac{1}{\ell}\frac{\sum_{\{w\in\mathcal{A}^\ell\}}(\Pr(w))^\beta\log(\Pr(w))}{\mathcal{Z}(\beta)}\ .$$

Using Eqs. (20) and (19) we rewrite Renyi entropy as:

$$\mathsf{h}(\beta) := \frac{1}{1-\beta}\lim_{\ell\to\infty}\frac{1}{\ell}Z(\beta)\ .$$

Using this relation and Eq. (21) we can rewrite $S(U(\beta))$ as:

$$S(U(\beta)) = -\lim_{\ell\to\infty}\frac{1}{\ell}\sum_{\{w\in\mathcal{A}^\ell\}}\frac{(\Pr(w))^\beta}{\mathcal{Z}(\beta)}\log\frac{(\Pr(w))^\beta}{\mathcal{Z}(\beta)}\ .$$

Identifying the twisted distribution completes the proof.

The result is that $S(U(\beta))$ basically is a metric entropy for a new twisted distribution $\mathcal{Q}_\beta(.)$. Now, if we could find and introduce a new process that could generate the words with this new distribution, then we could easily calculate the $S(U(\beta))$ by the result of theorem 1.

Now for every β we introduce a new process that generate twisted distribution. The key step is to define β-parametrized transition matrix:

$$\begin{aligned}\left(T_\beta^{(x)}\right)_{ij} &= e^{\beta\ln\Pr(x|\sigma_i)}\\ &= \left(\Pr(x|\sigma_i)\right)^\beta\ .\end{aligned}$$

Then the associated causal-state transition matrix is:

$$\mathbf{T}_\beta = \sum_{x\in\mathcal{A}}T_\beta^{(x)}\ .$$

Taking $\mathbf{l}_\beta$ ($\mathbf{r}_\beta$) as the left (right) eigenvector of $\mathbf{T}_\beta$, associated with λ_β:

$$\mathbf{l}_\beta\mathbf{T}_\beta = \lambda_\beta\mathbf{l}_\beta \tag{26}$$
$$\mathbf{T}_\beta\mathbf{r}_\beta = \lambda_\beta\mathbf{r}_\beta\ . \tag{27}$$

If the eigenvectors are chosen such that:

$$\mathbf{l}_\beta\cdot\mathbf{r}_\beta = 1\ , \tag{28}$$

then:

$$(\mathbf{T}_\beta)_{ij} = \lambda_\beta(\mathbf{r}_\beta)_i(\mathbf{l}_\beta)_j\ . \tag{29}$$

In this form, the Renyi entropy rate for the process generated by an ϵ-machine is simply [42]:

$$\mathsf{h}(\beta) = \frac{\log\widehat{\lambda}_\beta}{1-\beta}, \tag{30}$$

where $\widehat{\lambda}_\beta$ is the maximum eigenvalue of $\mathbf{T}_\beta$.

$\mathbf{T}_\beta$ is not a stochastic matrix, but one may renormalize it to define a new matrix that is right-stochastic by a mapping $\mathcal{M}_\beta : \mathbf{T}\to\mathbf{S}_\beta$ given by:

$$(\mathbf{S}_\beta)_{ij} = \frac{(\mathbf{T}_\beta)_{ij}(\widehat{\mathbf{r}}_\beta)_j}{\widehat{\lambda}_\beta(\widehat{\mathbf{r}}_\beta)_i}\ , \tag{31}$$

where the component (substochastic) symbol-labeled transition matrices, for each $x\in\mathcal{A}$ map to a new substochastic matrices by $\mathcal{M}_\beta^x : \mathbf{T}^x\to\mathbf{S}_\beta^x$:

$$\left(\mathbf{S}_\beta^{(x)}\right)_{ij} = \frac{\left(\mathbf{T}_\beta^{(x)}\right)_{ij}(\widehat{\mathbf{r}}_\beta)_j}{\widehat{\lambda}_\beta(\widehat{\mathbf{r}}_\beta)_i}\ . \tag{32}$$

Lemma 5. *$\boldsymbol{S}_\beta$ is row-stochastic:*

$$\sum_j(\boldsymbol{S}_\beta)_{ij} = 1\ .$$

Proof.

$$\begin{aligned}\sum_j(\boldsymbol{S}_\beta)_{ij} &= \sum_j\frac{(\mathbf{T}_\beta)_{ij}(\widehat{\mathbf{r}}_\beta)_j}{\widehat{\lambda}_\beta(\widehat{\mathbf{r}}_\beta)_i} = \frac{\sum_j(\mathbf{T}_\beta)_{ij}(\widehat{\mathbf{r}}_\beta)_j}{\widehat{\lambda}_\beta(\widehat{\mathbf{r}}_\beta)_i}\\ &= \frac{\widehat{\lambda}_\beta(\widehat{\mathbf{r}}_\beta)_i}{\widehat{\lambda}_\beta(\widehat{\mathbf{r}}_\beta)_i} = 1\ .\end{aligned}$$

So each row of $\mathbf{S}_\beta$ sums to 1. $\mathbf{S}_\beta$ is a transition matrix and the magnitudes its eigenvalues are less than or equal to one. It has at least one eigenvalue equal to one and the corresponding eigenvector is:

$$(\mathbf{P}_\beta)_i = (\widehat{\mathbf{r}}_\beta)_i(\widehat{\mathbf{l}}_\beta)_i\ . \tag{33}$$

Definition 2. *The* thermodynamic ϵ-machine *at inverse temperature β is the family $M(\beta) = \big\{\boldsymbol{\mathcal{S}}_\beta, \{\mathbf{S}_\beta^{(x)}, x\in\mathcal{A}\}, \eta_0\big\}$.*

Lemma 6. *If the probability of an arbitrary word generated by process M is* $\Pr(w)$*, the probability of the same* *word when generated by $M(\beta)$, as defined by def. 2, is:*

$$\mathcal{Q}_\beta(w) = \frac{(\Pr(w))^\beta}{\mathcal{Z}(\beta)} \ . \qquad (34)$$

Proof. *Let us consider two arbitrary words w_1 and w_2 with length ℓ. The probabilities of generating these words by process S_β, $\Pr_{S_\beta}(w_1)$, and $\Pr_{S_\beta}(w_2)$ can be written as:*

$$\Pr_{S_\beta}(w_1) = (\mathbf{P}_\beta)_{i1} \prod_{a=1}^{n-1} (\boldsymbol{S}_\beta)_{i_a i_{a+1}} \ ,$$

$$\Pr_{S_\beta}(w_2) = (\mathbf{P}_\beta)_{j1} \prod_{b=1}^{n-1} (\boldsymbol{S}_\beta)_{j_b j_{b+1}} \ .$$

Using Eq. (31) we can write the ratio of these two as:

$$\begin{aligned}\frac{\Pr_{S_\beta}(w_1)}{\Pr_{S_\beta}(w_2)} =& \frac{(\mathbf{P}_\beta)_{i_1} \prod_{a=1}^{n-1} \frac{(\boldsymbol{T}_\beta)_{i_a i_{a+1}} (\widehat{\mathbf{r}}_\beta)_{i_{a+1}}}{\widehat{\lambda}_\beta (\widehat{\mathbf{r}}_\beta)_{i_a}}}{(\mathbf{P}_\beta)_{j_1} \prod_{b=1}^{n-1} \frac{(\boldsymbol{T}_\beta)_{j_b j_{b+1}} (\widehat{\mathbf{r}}_\beta)_{j_{b+1}}}{\widehat{\lambda}_\beta (\widehat{\mathbf{r}}_\beta)_{j_b}}} \\ =& \frac{(\mathbf{P}_\beta)_{i_1} \frac{(\widehat{\mathbf{r}}_\beta)_{i_n}}{(\widehat{\mathbf{r}}_\beta)_{i_1}} \prod_{a=1}^{n-1} (\boldsymbol{T}_\beta)_{i_a i_{a+1}}}{(\mathbf{P}_\beta)_{j_1} \frac{(\widehat{\mathbf{r}}_\beta)_{j_n}}{(\widehat{\mathbf{r}}_\beta)_{j_1}} \prod_{b=1}^{n-1} (\boldsymbol{T}_\beta)_{j_b j_{b+1}}} .\end{aligned}$$

Now, using Eq. (33) we have:

$$\frac{\Pr_{S_\beta}(w_1)}{\Pr_{S_\beta}(w_2)} = \frac{(\widehat{\mathbf{l}}_\beta)_{i_1} (\widehat{\mathbf{r}}_\beta)_{i_n} \prod_{a=1}^{n-1} (\boldsymbol{T}_\beta)_{i_a i_{a+1}}}{(\widehat{\mathbf{l}}_\beta)_{j_1} (\widehat{\mathbf{r}}_\beta)_{j_n} \prod_{b=1}^{n-1} (\boldsymbol{T}_\beta)_{j_b j_{b+1}}} \ .$$

We also could rewrite $(\widehat{\mathbf{l}}_\beta)_{i_1} (\widehat{\mathbf{r}}_\beta)_{i_n}$ as:

$$(\widehat{\mathbf{l}}_\beta)_{i_1} (\widehat{\mathbf{r}}_\beta)_{i_n} = \frac{(\mathbf{T}_\beta)_{i_1 i_n}}{\lambda_\beta}$$

$$(\widehat{\mathbf{l}}_\beta)_{j_1} (\widehat{\mathbf{r}}_\beta)_{j_n} = \frac{(\mathbf{T}_\beta)_{j_1 j_n}}{\lambda_\beta} \ ,$$

a direct result of Eq. (29). Using these we have:

$$\begin{aligned}\frac{\Pr_{S_\beta}(w_1)}{\Pr_{S_\beta}(w_2)} =& \frac{(\mathbf{T}_\beta)_{i_1 i_n} \prod_{a=1}^{n-1} (\boldsymbol{T}_\beta)_{i_a i_{a+1}}}{(\mathbf{T}_\beta)_{j_1 j_n} \prod_{b=1}^{n-1} (\boldsymbol{T}_\beta)_{j_b j_{b+1}}} \\ =& \left\{ \frac{(\mathbf{T})_{i_1 i_n} \prod_{a=1}^{n-1} (\boldsymbol{T})_{i_a i_{a+1}}}{(\mathbf{T})_{j_1 j_n} \prod_{b=1}^{n-1} (\boldsymbol{T})_{j_b j_{b+1}}} \right\}^\beta \ .\end{aligned}$$

For $\beta = 1$ we have:

$$\frac{\Pr_T(w_1)}{\Pr_T(w_2)} = \frac{(\mathbf{T})_{i_1 i_n} \prod_{a=1}^{n-1} (\boldsymbol{T})_{i_a i_{a+1}}}{(\mathbf{T})_{j_1 j_n} \prod_{b=1}^{n-1} (\boldsymbol{T})_{j_b j_{b+1}}} \ .$$

And, this means:

$$\frac{\Pr_{S_\beta}(w_1)}{\Pr_{S_\beta}(w_2)} = \left(\frac{\Pr_T(w_1)}{\Pr_T(w_2)} \right)^\beta \ .$$

Since this is true for every pair of arbitrary w_1 and w_2 the proof is complete.

The result of this theorem is when we use $\mathcal{M}_\beta : \mathbf{T} \to \mathbf{S}_\beta$ to map our process T to a new process S_β we twist the distribution of words generated by T in the way that we wanted in lemma. 4.

Theorem 2.

$$S(U(\beta)) = h_\mu(M(\beta)) \ , \qquad (35)$$

Proof. *The proof is a straight forward result of lemma. 4 , theorem. 6 and eq. 7.*

Using theorem 2 from now on we have a powerful tool to calculate fluctuation spectra for a given process.

E. ϵ-Machine Thermodynamics

Section III C concerned average macroscopic informational quantities and how to express them in terms of a process's ϵ-machine—a more detailed, structural mesoscale description. As such, it ignored deviations or fluctuations in these quantities. The preceding section, however, define the spectra of fluctuations in terms of $S(U)$ which allows us to we analyze a process' temporal fluctuations using its ϵ-machine, following the approach introduced in Ref. [25].

As we showed the thermodynamic entropy density is the entropy rate of the new renormalized ϵ-machine at inverse temperature β. As β is varied we monitor the relative "size" of the fluctuation process at that temperature.

In a similar way, we monitor the stored information in the fluctuation processes by calculating the ϵ-machine's statistical complexity C_μ at each β. Via Eqs. (11) and (33), we have the statistical complexity fluctuation spectrum:

$$\begin{aligned}C_\beta &= C_\mu(M(\beta)) \\ &= -\sum_{i=1}^{N} (\mathbf{P}_\beta)_i \log_2 (\mathbf{P}_\beta)_i \ .\end{aligned}$$

In parallel, we have the spectrum of predictable information $\mathbf{E}$ fluctuations:

$$\begin{aligned}\mathbf{E}_\beta &= \mathbf{E}(M(\beta)) \\ &= \mathrm{I}[\mathcal{S}_\beta^- ; \mathcal{S}_\beta^+] \ ,\end{aligned}$$

where we replaced the semi-infinite past and future in Eq. 12 with $M(\beta)$'s forward and reverse causal states, respectively. Finally, a similar substitution leads us to the spectra of ephemeral and bound informations:

$$\begin{aligned}r_\beta &= r_\mu(M(\beta)) \\ &= H[X_{0,\beta} | \mathcal{S}_{0,\beta}^-, \mathcal{S}_{1,\beta}^+]\end{aligned}$$

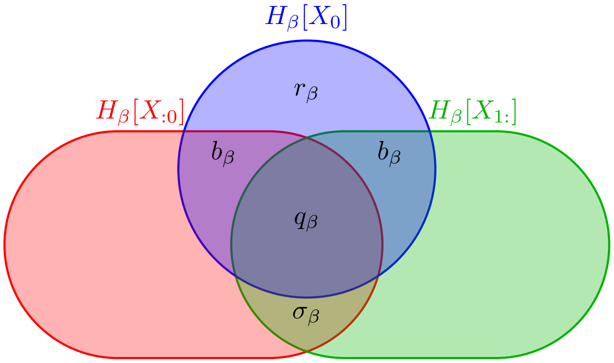

FIG. 7. Thermodynamic information diagram—the analog of the information diagram of Fig. 8 of Ref. [34]–but for the family of processes generated by the inverse-temperature β-parametrized ϵ-machine.

and:

$$\begin{aligned} b_\beta &= b_\mu(M(\beta)) \\ &= \mathrm{I}[X_{0,\beta}; \mathcal{S}^+_{1,\beta} | \mathcal{S}^-_{0,\beta}] \ , \end{aligned}$$

where $X_{0,\beta}$ is the present observed variable generated by $M(\beta)$. Here, $\mathcal{S}^+_{1,\beta}$ is its causal state at time $t = 1$, whereas the other, $\mathcal{S}^-_{0,\beta}$, is the causal state anchored at time $t = 0$ of the time-reversed ϵ-machine.

Figure 7 summarizes the relationship between these various measures in what is called an *information diagram*; cf. Ref. [34]'s Fig. 8. Information diagrams, similar to Venn diagrams, aid in interpreting information relation between variables. There are two differences between them. First, instead of set size the measure is Shannon entropy and Second, here an overlap interprets as an mutual information rather than set intersection. Those intersections which are irreducible are called *atoms* and size of them reflect the Shannon information measures. The name *atoms* is used because they are elementary atoms of sigma-algebra over the random variable space.

F. Stability and Convexity

In thermodynamics the condition of stability of macroscopic states [23] for an ideal gas, for example, is that:

$$\left(\frac{\partial^2 S}{\partial U^2}\right)_{N,V} \leq 0 \ . \tag{36}$$

This means that for stable thermodynamic states, S should be a convex function of U. This comes from the assumption that entropy is maximized for the stable equilibrium state (among other possible macrostates) and that the entropy of a composite system is the sum of the entropies of its constituent subsystems.

For systems with finite degrees of freedom the convexity condition of entropy leads to a maximum for the entropy, if approaching the maximum for positive temperature $T \to \infty$. There is also a maximum value for energy, which corresponds to a region of negative temperature.

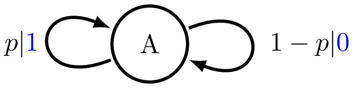

FIG. 8. Biased Coin Process ϵ-machine.

Let's check whether these parallels are meaningful and useful.

Theorem 3. *$S(U)$ is a convex function of U, where the former is given by Eq. (16), the latter by Eq. (22).*

Proof. *See App. A 5.*

Thus, for positive (negative) β, S is an increasing (decreasing) function of U. And, there is unique maximum for S, where β vanishes. If the support of S is finite there is a region $[U_{\min}, U_{\max}]$, for which S is nonzero. This means a typical shape for S(U) is seen in Fig. V A 1.

V. FLUCTUATION SPECTRA

A. Examples

Let's investigate SNESSs that reveal what is captured by fluctuation spectra—several are well known processes and we compare the results with the analytical ones.

1. Biased Coin Process

Recall the Biased Coin Process from the Introduction and its Markov chain presentation in Fig. 1. The latter should be compared to the ϵ-machine (unifilar minimal hidden Markov chain) presentation in Fig. 8, which has only one state.

Now using the method introduced in section 5, let's study the fluctuation spectral density (FSD) for biased coin by investigating its ϵ-machine structure. Figure (V A 1) shows the results.

Spectrum of statistical complexity C_β, Trivially, vanishes, due to ϵ-machine having a single states.

Spectrum of excess entropy $\mathbf{E}_\beta$ vanishes trivially, due to Biased Coin Process being IID and does not have any correlation.

All IID processes have such spectra.

2. Golden Mean Process

This process is define by the Markov chain presentation in Fig. (4 or the ϵ-machine presentation Fig. (10).

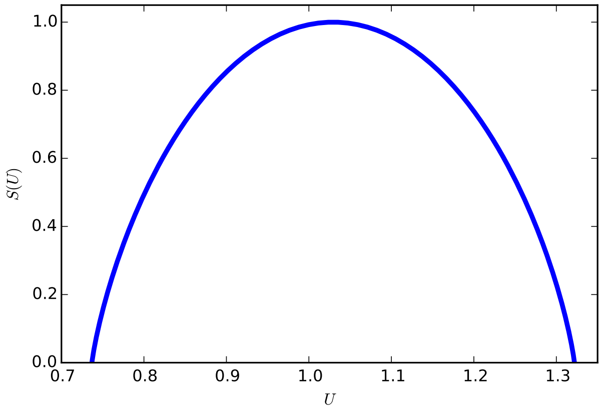

FIG. 9. FSD for biased coin, p=0.4. Add $U_{\min}$, $U_{\max}$, h, h_μ, and lines for unity and zero slope.

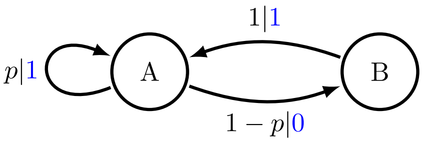

FIG. 10. Golden Mean Process ϵ-machine.

After running the machine, It generates a time series. Here we do the same approach and figure (5) and (11) show the results. As we see, the pattern 00 is absent in the time series (e.g. see the case $\ell = 2$).

As it is seen, when we decrease P, first the set of energies with non zero S decrease till reach the minimum set and then increase. Let's consider this model analytically.

$$A_{n+1} = pA_n + B_n, \tag{37}$$

$$B_{n+1} = qA_n. \tag{38}$$

The matrix T is a left stochastic matrix defined as

$$T := \begin{pmatrix} p & q \\ 1 & 0 \end{pmatrix}, \tag{39}$$

At nth step one may arrive at

$$A_n = \frac{1-(-q)^{n+1}}{1+q}, \tag{40}$$

$$B_n = \frac{q+(-q)^{n+1}}{1+q}. \tag{41}$$

At large times where the system approaches to its stationary values

$$\lim_{n\to\infty} A_n \sim \frac{1}{1+q}, \tag{42}$$

$$\lim_{n\to\infty} B_n \sim \frac{q}{1+q}. \tag{43}$$

Now let's define

$$(\mathbf{T}_\beta)_{ij} := \exp(\beta \ln P_{v_i \to v_j}) \Rightarrow \begin{pmatrix} p^\beta & q^\beta \\ 1 & 0 \end{pmatrix}, \tag{44}$$

Eigenvalues and eigenvectors of T_β are

$$\mathbf{l}_\beta \mathbf{T}_\beta = \lambda_\beta \mathbf{l}_\beta \tag{45}$$

$$\mathbf{T}_\beta \mathbf{r}_\beta = \lambda_\beta \mathbf{r}_\beta. \tag{46}$$

This matrix is not an stochastic matrix, but one may construct a right stochastic $\mathbf{S}_\beta$ from it. It can be easily shown that

$$\begin{aligned} \lambda_{\beta,\max} = \lambda_1 &:= \frac{1}{2}[p^\beta + \sqrt{p^{2\beta}+4q^\beta}], \\ \lambda_2 &:= \frac{1}{2}[p^\beta - \sqrt{p^{2\beta}+4q^\beta}], \end{aligned}$$

$$\begin{aligned} \widehat{\mathbf{l}}_\beta &= \frac{1}{\lambda_1-\lambda_2}\begin{pmatrix} 1 & -\lambda_2 \end{pmatrix}, \\ \widehat{\mathbf{r}}_\beta &= \begin{pmatrix} \lambda_1 \\ 1 \end{pmatrix}, \end{aligned}$$

$$S_\beta = \begin{pmatrix} \frac{\lambda_1+\lambda_2}{\lambda_1} & -\frac{\lambda_2}{\lambda_1} \\ 1 & 0 \end{pmatrix}.$$

and

$$\begin{aligned} S(U(\beta)) =& \frac{1}{\lambda_1-\lambda_2}\left\{(\lambda_1+\lambda_2)\log(\lambda_1+\lambda_2) - \right. \\ & \left. \lambda_1 \log\lambda_1 - \lambda_2 \log(-\lambda_2)\right\} \\ U(\beta) =& \frac{1}{\beta}\left\{S(U(\beta)) - \log\lambda_1\right\} \end{aligned} \tag{47}$$

B. Fluctuations at Zero Temperature

Bet we have not fully characterized the shape of fluctuation spectra. In particular, it is possible and even common to have a huge multiplicity of ground states at zero temperature. In this case, $S(U_{\min}) > 0$. Interestingly, this can also occur at the negative temperature extreme, giving $S(U_{\max}) > 0$. Here, we explore this, first analytically and then through examples.

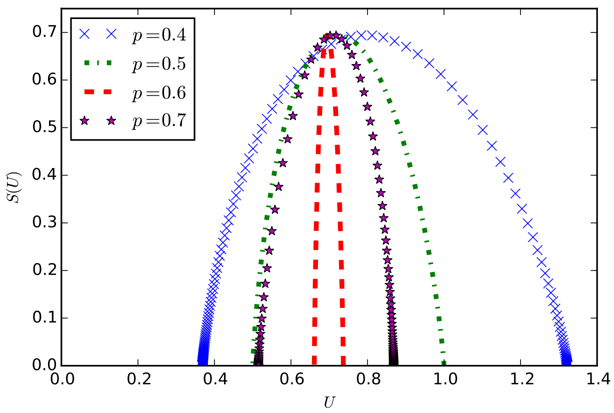

FIG. 11.
FSD for Golden Mean Process

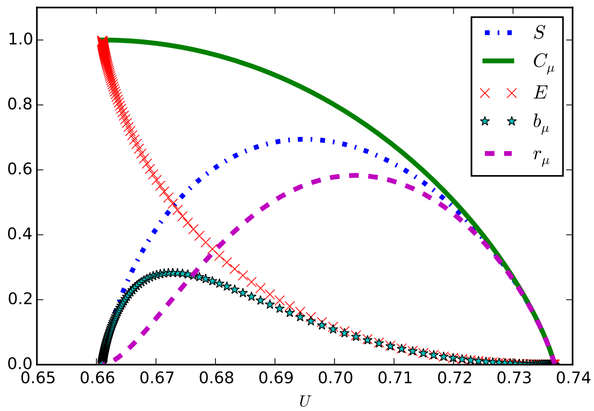

FIG. 12.
Information measures versus Energy for Golden Mean Process with $p = 0.6$

1. Submachine Dynamics

Lemma 7. *A deterministic edge stays deterministic under temperature variation. (See fig. 14.) That is, if for $\beta = 1$: $(\boldsymbol{T}_\beta)_{ij} = 1$, we have $(\boldsymbol{S}_\beta)_{ij} = 1$ and $(\boldsymbol{S}_\beta)_{i,k\neq j} = 0$, for all β.*

Proof.

$$\begin{aligned} &\textit{If for } k \neq j : (\boldsymbol{T}_{\beta=1})_{i,k} = 0 \\ &\quad \to (\boldsymbol{T}_\beta)_{i,k} = (0)^\beta = 0 \\ &\quad \to (\boldsymbol{S}_\beta)_{i,k} = 0 \ . \end{aligned} \tag{48}$$

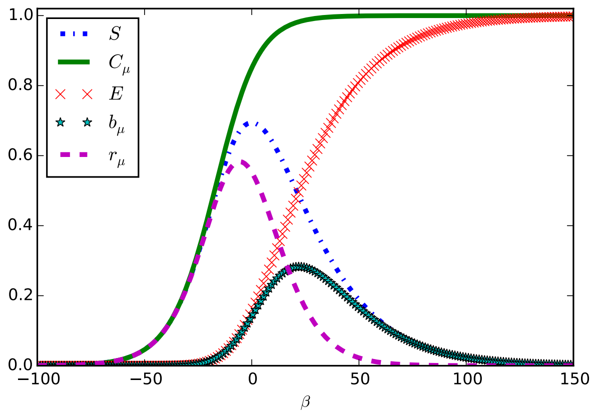

FIG. 13.
Information measures versus β for Golden Mean Process with $p = 0.6$

Also we have for all β:

$$\sum_l (\boldsymbol{S}_\beta)_{il} = 1 \ . \tag{49}$$

These lead to:

$$(\boldsymbol{S}_\beta)_{ij} = 1 \ . \tag{50}$$

Lemma 8. *Splitting isolated branching does not change thermodynamic entropy density.*

Consider an n-state machine, whose general shape is similar to that in fig. 15 with transfer matrix T satisfying condition:

$$T_{k,n} = 0 \ , \tag{51}$$

for all $k \neq i, n$.

Let's concentrate on the states i and n. We call this part of machine as isolated branching. (See fig. 16.) Now a new $n+1$ state machine may be introduced (see figure 17) where the state n is split to two distinct states n and $n+1$ with transfer matrix T'. There exists a T' in such a way that the two machines produce the same time series. So:

$$S'(U(\beta)) = S_1(U(\beta)) \ . \tag{52}$$

Proof. *To guarantee the same thermodynamic entropy*

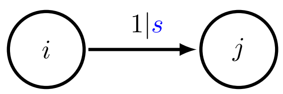

FIG. 14. Deterministic edge.

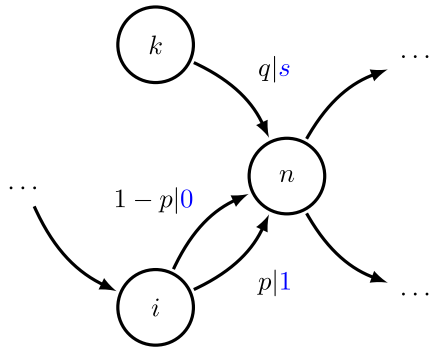

FIG. 15. Portion of an ϵ-machine.

density define:

$$T' = \begin{cases} T'_{in} & = p \\ T'_{i,n+1} & = 1-p \\ T'_{a\neq i, b\neq(n,n+1)} & = T_{ab} \\ T'_{a\neq i, b=(n,n+1)} & = 0 \\ T'_{a=(n,n+1),b} & = T_{nb} \\ T'_{k,b\neq n} & = 0 \\ T'_{k,n} & = T_{kn} \end{cases} . \tag{53}$$

Lemma 9. *Half and half isolated branching (fig. 16 with $p = \frac{1}{2}$) stays the same with temperature changes.*

Proof. *Considering lemma 8, one splits the state n to two state n and $n+1$, going from see fig. 15 to fig. 17, without changing the thermodynamic entropy density:*

$$j \neq n, n+1 : T_{ij} = 0 \ .$$

So, we have:

$$\forall \beta, j \neq n, n+1 : S_{2_{ij}} = 0 \ ,$$

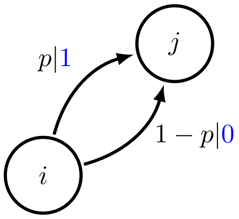

FIG. 16. Isolated branching.

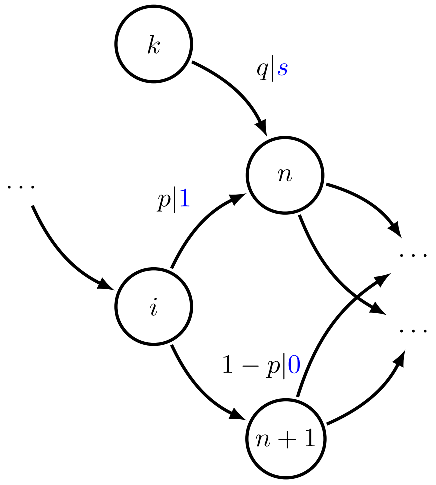

FIG. 17. HMM 2.

And, this means:

$$\forall \beta : S_{2_{in}} + S_{2_{i,n+1}} = 1 \ .$$

Now, due to the symmetry we have in Eq. (53) for state n and $n+1$ we have:

$$\forall \beta \in Z : S_{2_{in}} = S_{2_{i,n+1}} = \frac{1}{2} \ .$$

Conjecture. *For $T_{ij} = p$ and $T_{ik} = 1-p$ which i and k are different state, two things could happen. First there exists a unique c which:*

$$\begin{cases} p < c : & S_{ij_{\beta\to\infty}} = 1 \\ p > c : & S_{ij_{\beta\to-\infty}} = 0 \end{cases} . \tag{54}$$

And, this also means:

$$\begin{cases} p < c : & S_{ik_{\beta\to\infty}} = 0 \\ p > c : & S_{ik_{\beta\to-\infty}} = 1 \end{cases} . \tag{55}$$

second (Hidden symmetry case)

$$S_{ik_{\beta\to\infty}} = S_{ik_{\beta\to-\infty}} = \frac{1}{2} \tag{56}$$

2. Ground States Without Fluctuations

As an example let's look at the biased coin process. This process at $\beta \to \infty$ and $\beta \to -\infty$ turns to machines which are shown in figure 18 and 19.

For the next example let's consider the golden mean process. This machine has a unique ground state. As β

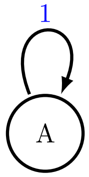

FIG. 18. Biased coin process ϵ-machine, when β converge to negative infinity and $p < 0.5$, or when β converge to infinity and $p > 0.5$ a deterministic process

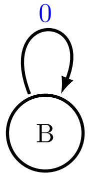

FIG. 19. Biased coin process ϵ-machine, when β converge to negative infinity and $p > 0.5$, or when β converge to infinity and $p < 0.5$ a deterministic process

approaches infinity, one may arrive at

$$S_{0_{\beta\to\infty}} := \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \tag{57}$$

$$S_{1_{\beta\to\infty}} := \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \tag{58}$$

Golden mean process at $\beta \to \infty$ and $\beta \to -\infty$ turns to machines which are shown in figure 20 and 21. As it is seen these machine are completely deterministic. For this process (and as we will show for similar processes) the matrix elements of $S_{0_{\beta\to\infty}}$ and $S_{1_{\beta\to\infty}}$ are zero or one. For any element which is equal to one, all the other elements of the corresponding rows and columns are zero. This means that in the limit $\beta \to \infty$ if we have a unique ground state, machine converge to deterministic machine.

3. Nemo Process

Our third example is the Nemo Process, shown in fig. 22. Since the recurrent states simply permute upon observing a 0, the word 0000... never reveals the current state. This once again means that the process is non-Markovian and has $R = \infty$.

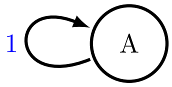

FIG. 20. Golden Mean process ϵ-machine, when β converge to negative infinity and $p < \frac{\sqrt{5}-1}{2}$,or when β converge to infinity and $p > \frac{\sqrt{5}-1}{2}$, a deterministic process

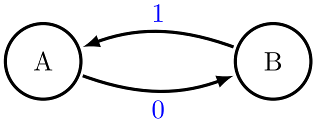

FIG. 21. Golden Mean process ϵ-machine, when β converge to negative infinity, $p > \frac{\sqrt{5}-1}{2}$,or when β converge to infinity, $p < \frac{\sqrt{5}-1}{2}$, a deterministic process

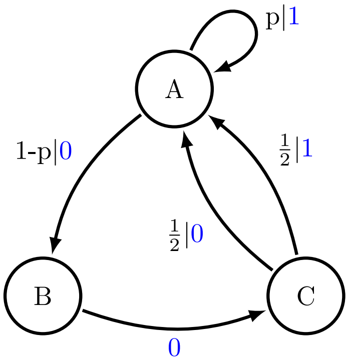

FIG. 22. ϵ-Machine that generates the non-Markovian Nemo Process; its Markov order is infinite. The Nemo Process makes this perhaps clearer, however, since the recurrent states permute into each other upon observing a 0. The transient structure captures this explicitly: ABC maps back to itself on a 0.

This process is define by the machine shown in the fig. 22. After running the machine, It generates a time series. The results are summarized in fig. 23. The interesting point of this process is it's nonzero $S(U_{\min})$. This means the process has (highly) non-unique ground state. Figure 24 shows energy versus β for this process.

Table I shows the analytical result for tree different example.

Considering Fig. 22 using Lemma7 and 8, edge BC and two edges from C to A stay the same when we change temperature. Using the conjecture (Eqs. (54) and (55)) for edge AA and AB there should be a unique c which depending on the case ($p > c$ or $c > p$) one of the edges survive with probability 1 in one of the limits and the other one survive in the other limits. This means depends on p ($p > c$ or $c > p$) we could have non uniqueness in different limits. Using simulation one could find that $0.58 < c < 0.6$ and there will be a sharp phase transition in function $S(U)$ when we change c around 0.59. (See fig. 25.)

For $p < c$ one obtains:

$$S_{0_{\beta\to\infty}} := \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \frac{1}{2} & 0 & 0 \end{pmatrix}, \tag{59}$$

TABLE I. Spectral Properties of Biased Coin, Golden Mean, and Nemo Processes

Process	C_μ	h_μ	$\mathfrak{h}$	$U_{\min}$	$U_{\max}$	$S(U_{\min})$	$S(U_{\max})$
Biased Coin	0.0000	1.0000	0.9710	0.7370	1.3219	0	0
Golden Mean	0.8631	0.6942	0.6935	0.6610	0.7370	0	0
Nemo ($p = \frac{1}{2}$)	1.5000	0.75	1.5000	0.6669	1.0000	0.3347	0

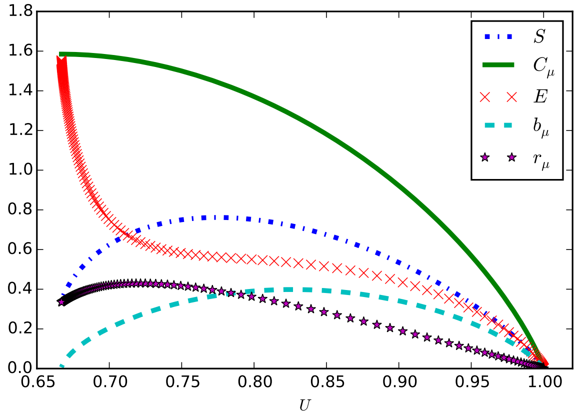

FIG. 23.
Information measures versus energy U for Nemo Process.

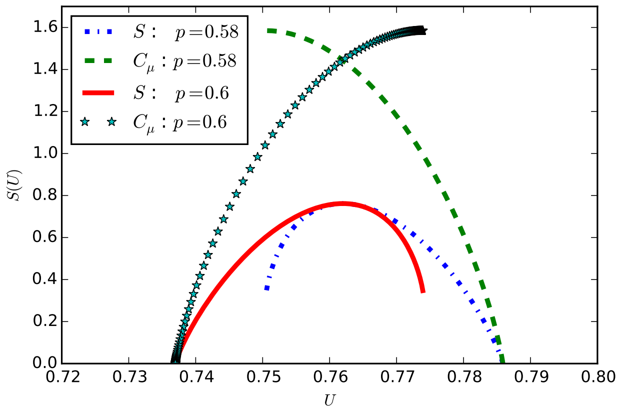

FIG. 25.
Phase transition in the Nemo Process with changing p.

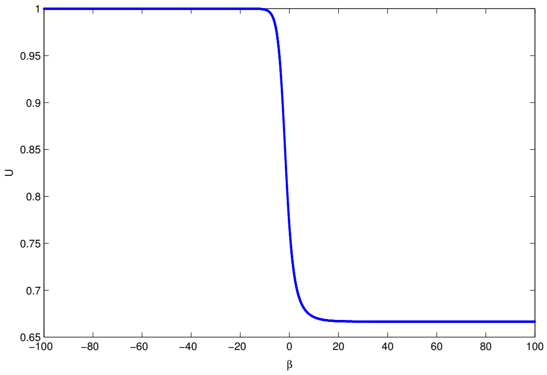

FIG. 24.
$U(\beta)$ for Nemo Process.

$$S_{1_{\beta\to\infty}} := \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 \end{pmatrix} . \qquad (60)$$

(See fig. 27). This does not satisfy the above-mentioned condition that leads to non unique ground states.

4. Persistent Symmetry

For an n state machine, whose general shape is something like fig. 15 and assuming $p = \frac{1}{2}$ (having half and half isolated branching), because of lemma 3, we will have for all $\beta \in Z$:

$$S_{2_{ik}} = S_{2_{il}} = \frac{1}{2}$$

Now, from eq. (**??**):

$$\begin{aligned} S_2(U(\beta)) \geqslant &(\mathbf{P}_{2_\beta})_i((\mathbf{S}_{2_\beta})_{ik} \log{(\mathbf{S}_{2_\beta})_{ik}} \\ &+ (\mathbf{S}_{2_\beta})_{il} \log{(\mathbf{S}_{2_\beta})_{il}}) \\ &= (\mathbf{P}_{2_\beta})_i \log(2) = (\mathbf{P}_{2_\beta})_i \ . \end{aligned}$$

From eq. (52) we have for all β:

$$S_1(U(\beta)) \geqslant (\mathbf{P}_{1_\beta})_i \ .$$

This means that as β sends to positive or negative infinity if the probability of being in the state i does not vanish, there will be non unique ground states.

The simplest machine with lowest number of state which have non unique ground state is shown in figure(28).

These general examples give us the idea that symmetry in the limits of β is the reason behind non uniqueness.

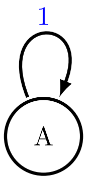

FIG. 26. Nemo Process ϵ-machine, when $p < c \approx 0.5898$ and β converge to negative infinity,or when $p > c \approx 0.5898$ and β converge to infinity, a deterministic process

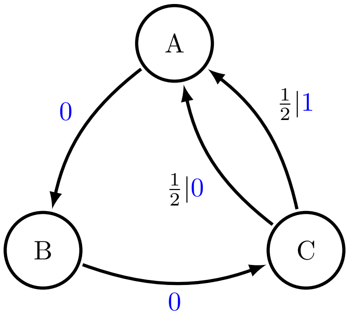

FIG. 27. Nemo Process ϵ-machine, when $p > c \approx 0.59$ and β converge to negative infinity,or when $p < c \approx 0.59$ and β converge to infinity.

For those cases which are shown in figs 15, 27 and 28 it is easy to see their symmetry. In these cases we have symmetry in $\beta = 1$ in the ϵ-machine, and these symmetries survive for every finite β, although the other parts of ϵ-machine change when β changes. If these symmetries survive in the limits where β tends to positive or negative infinity then the system ends up with non unique ground states in that limit. We call these type of symmetries, *persistent symmetries*, which survive without any change for every finite changes in temperature. One of the reasons behind non unique ground states are persistent symmetries.

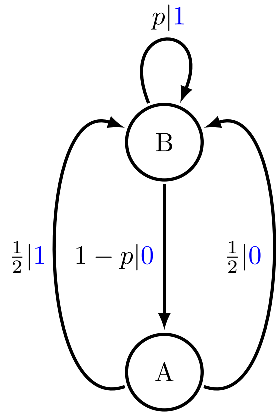

FIG. 28. Two-state ϵ-machine, with nonunique ground states.

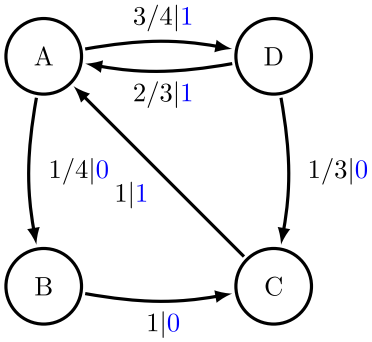

FIG. 29. RRIP Process ϵ-machine.

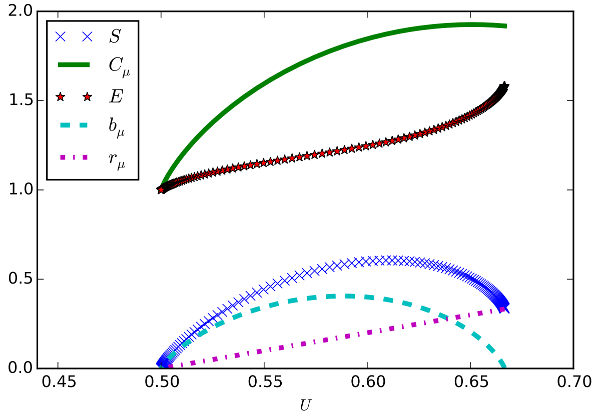

FIG. 30.
Information measures versus Energy for RRIP Process.

5. *Hidden Symmetry*

As it is discussed there should be symmetry at the limits of β to have non unique ground states. In the case of *Persistent Symmetry* not only we should have symmetry at the limit but also we have it for all finite β's. More complex cases are those which do not have symmetry for any finite β but surprisingly symmetry appears at the limit of β. These type of processes are hard to find. An example is the RRIP process which is defined by the machine shown in Fig. 29. Information measures for this process summarized in Fig.30. As it is clear from Fig. 29 we don't have any symmetry for finite β but at the limits, when β approaches to negative infinity the system end up with the Fig. 31 which has the symmetry. So we say this process has *hidden Symmetry*.

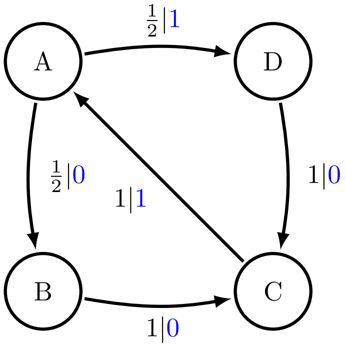

FIG. 31. RRIP ϵ-machine when $\beta \to -\infty$.

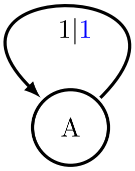

FIG. 32. RRIP ϵ-machine when $\beta \to \infty$.

6. *Nonuniqueness at Positive and Negative Temperatures*

$S(U)$ is a nonzero convex function and U is positive quantity. This means considering the behavior of $S(U)$ in two limits for the general shape of this function we could have five cases: zero-zero, zero-non zero, non zero-zero, symmetric non zero, non zero, asymmetric non zero, non zero. For the first three cases we have biased coin, Nemo ($p < c$), Nemo ($p > c$). For the forth case one could find the example in figure (33)) and it's $S(U)$ in figure (34) and for the fifth case one could find the example in fig. (35)) and it's $S(U)$ in fig. (36).

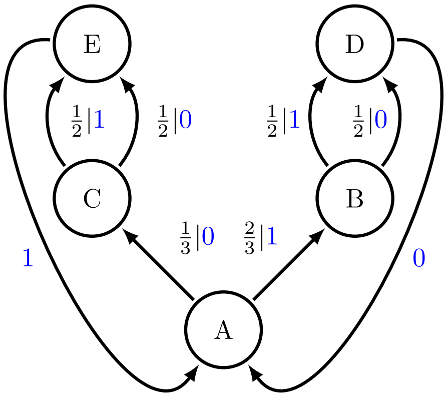

FIG. 33. ϵ-machine with symmetric nonuniqueness in two limits.

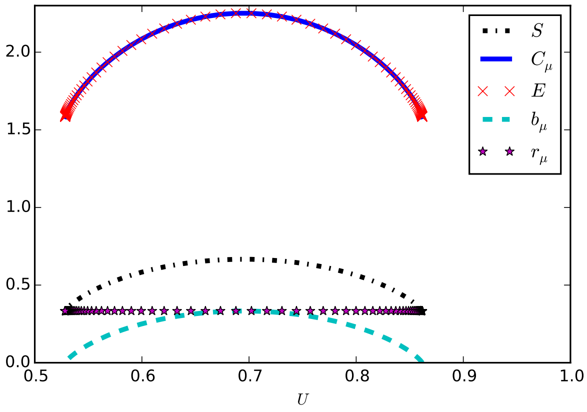

FIG. 34.
FSD for ϵ-machine with nonuniqueness in two limits.

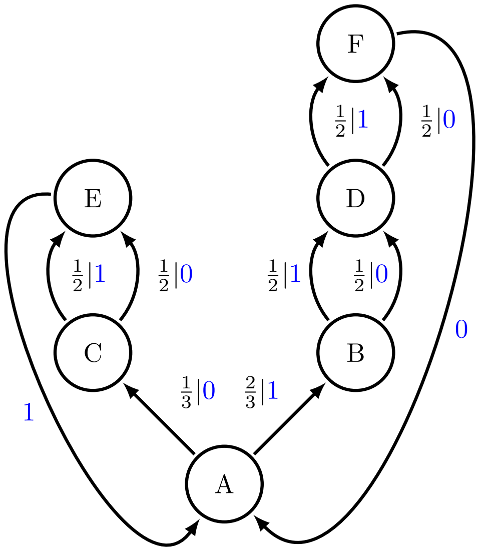

FIG. 35. ϵ-machine with asymmetric nonuniqueness in two limits.

C. Causally Irreversible Processes

Instead of prediction, one may be interested in retrodiction: using the future to predict the past. To do this, the formalism that we built is essentially unchanged and one should just look at the measurements in the reverse time direction. The causal states for revers process $\boldsymbol{\mathcal{S}}^-$ could be different from causal states for forward process $\boldsymbol{\mathcal{S}}$. We call a process causally irreversible if $C_\mu^- \neq C_\mu^+$.

As an example let's look at RIP process. This process

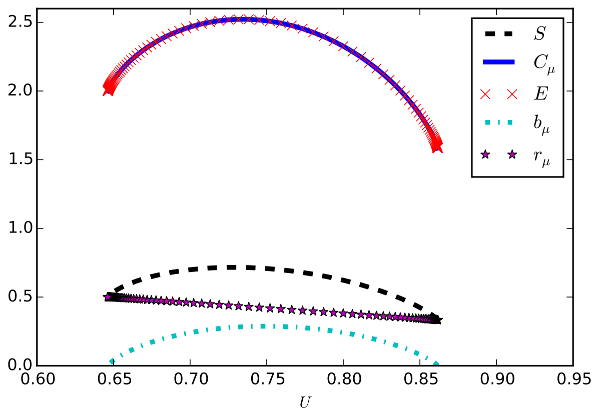

FIG. 36.
FSD for the ϵ-machine with asymmetric nonuniqueness in two limits.

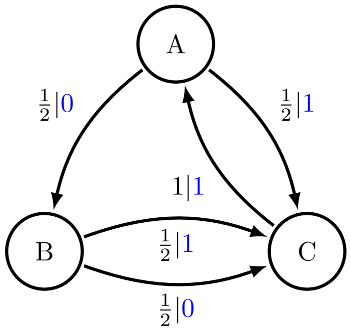

FIG. 37. RIP Process ϵ-machine

is defined by the machine which is shown in fig.37. RRIP process which is shown in fig.29 describe this process in the reverse time direction. Here forward process has three causal states and reversed process has four causal states. This process is causally irreversible. All information measures for forward and reverse process except the statistical complexity are equal as they are shown in figs. 38 .

Lemma 10. *All the information measures which only depend on probabilities of words* $\Pr(w^L)$*'s that are generated by process are equal for forward and reveres processes. e.g.* $\mathbf{E}$*,* h_μ*,* b_μ *and* r_μ

Proof. *Considering an arbitrary process* T *and it's reverse process* T^R*, the probability of every word* w *which is generated by* T *is equal to the probability of it's reversed word* w^R *which is generated by* T^R*. This means we will have the same probability distributions over set of words for* T *and* T^R *by just changing the labels. Now that they have the same distributions all functions of these distributions would be equal for* T *and* T^R*.*

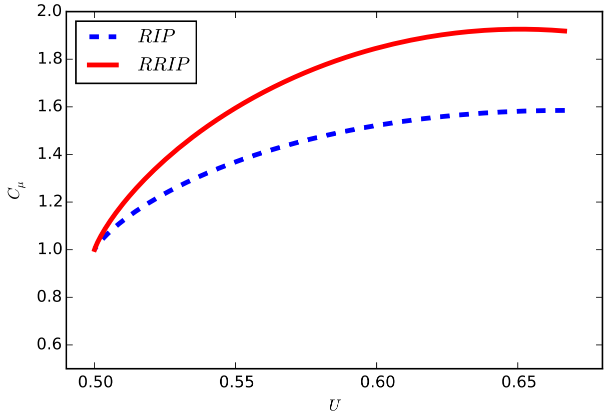

FIG. 38.
Statistical complexity versus Energy for RIP and RRIP Process.

VI. LARGE DEVIATIONS ... BEYOND THE TYPICAL SET

Quick motivation of large deviations.

Start with the biased coin example: The most probable sequence of all Heads is never seen for even only moderate length sequences. Why? Answering that leads to a different notion of what are called "typical sets"—those sets of events that are both most numerous and capture lots of the sequence probability density.

Call refer back to the word distribution for the biased coin as you talk through this motivation.

1. *Asymptotic Equipartition*

Let's consider a sequence of random variables, $X_1, X_2, ...$ which is independent and identically distributed (i.i.d.); means that all random variables are mutually independent and have the same probability distribution. Asymptotic equipartition property states that the joint probability $P(X_1, ..., X_n)$ satisfies

$$\lim_{n\to\infty} \frac{\log_2 P(X_1, X_2, ..., X_n)}{n} = H(X) \qquad (61)$$

where $H(X)$ is the entropy associated with the random variable X. One could divide the set of all sequences with the length n to two partitions. First typical set A_ϵ^n defined through

$$\begin{aligned} A_\epsilon^n =& \{(x_1, x_2, ..., x_n) : 2^{-n(H(X)+\epsilon)} \\ & \leq P(x_1, x_2, ..., x_n) \leq 2^{-n(H(X)-\epsilon)}\} \end{aligned} \qquad (62)$$

The remaining is non typical set. Then for any typical set

$$\forall \epsilon \geq 0, \exists\, n_0 : n > n_0,$$
$$\Pr\{| - \frac{\log_2 P(x_1, x_2, ..., x_n)}{n} - H(X)| \leq \epsilon\} \geq 1 - \epsilon. \tag{63}$$

This means that for large n, typical set is most probable, and the probability of each sequence in the typical set, A^n_ϵ, have almost the same value $2^{-nH(X)}$. It is important to note that although the size of typical set size is really smaller than the whole set of sequences' size, but it contains most of the probability.

It should be noted that, There is also another theorem called Shannon-McMillan-Breiman theorem, which states that one may release discrete-time i.i.d. condition for AEP. It is only need to have discrete-time stationary ergodic process.

Let's consider biased coin process. A typical set with length n forms of sequences containing of np, 1, and $n(1-p)$, 0. A sequence whose all elements are 1, is an example of non-typical set.

2. Infinite Partition Idea

As we discussed, in a classical point of view, $\mathcal{A}^\infty$ could be divided to two subsets, typical and non typical set and most of the information measures are defined on the typical one. For example metric entropy is the exponent of the decay for any realizations of data in typical set, but one could ask what is the exponent of the decay for other part of $\mathcal{A}^\infty$ or many other questions. The problem with studding non typical set is those realizations of data which belong to non typical set are so rare. Here we propose a method to study the whole $\mathcal{A}^\infty$ instead of typical set. We divide the $\mathcal{A}^\infty$ to infinite independent subsets A_β's and another subset called FW which has all the forbidden words (words with zero probabilities) in it. For a given process T we define A^T_β by

$$A^T_\beta := \{w : w \in \mathcal{A}^\infty, \frac{-\lim_{L\to\infty} \Pr(w)}{L} = U(\beta)\}, \tag{64}$$

We label subsets with parameter β. Typical set is one of these subsets with $\beta = 1$. To study A^T_β we do a trick. We use a map $\mathcal{M}_\beta$ to map our process T to a new process S_β which its' typical set is equal to A^T_β.

$$A^T_\beta = A^{S_\beta}_1 \tag{65}$$

Now we could study typical set of S_β which is A^T_β. Changing β we could cover all the set and study all A^T_β. Now all information measures could be parametrized with β and we could study all parts of $\mathcal{A}^\infty$. But how we should partition $\mathcal{A}^\infty$? Because A^T_β is equal to typical set of a S_β, all it's members should have the same probability that means they have the same energy density, so we also could parametrize the set with U instead of β. The result is, we have infinite subsets which every subset include words with the same probability and we could label them with U.

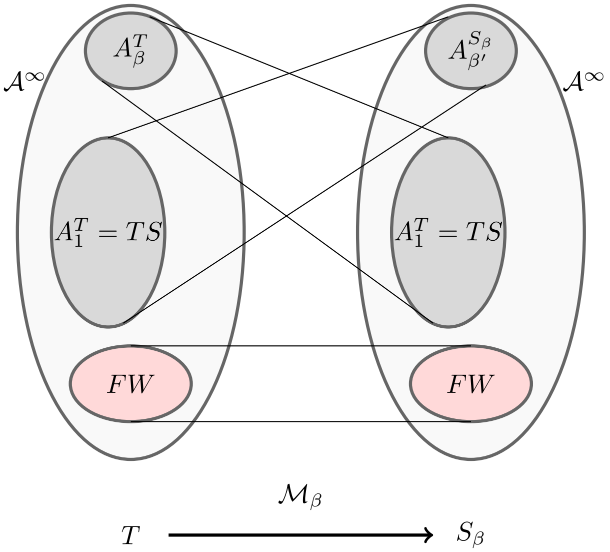

FIG. 39. Infinite partition idea

3. Large Deviation Rate Function

One may be interested in investigating the probability of the class of sequences with the same energy U (same probability). Let's define:

$$I(U) := \lim_{L\to\infty} \left[-\frac{\log_2 \Pr(U^L)}{L} \right] . \tag{66}$$

The probability of classes decay exponentially as a function of L and this is the reason for existence of L in the denominator. This function is called large deviation rate function. It is important to know that $I(U(\beta))$ measures how fast the probability of a whole subset A^L_β decays when we increase L and not how fast the probability of words in A_β decays. When we increase L number of words with the same U increase exponentially, and the probability of every word decreases exponentially. The probability $\Pr(U^L)$ is the multiplication of the number of sequences with the same energy U and the probability of a sequence with the energy U:

$$\Pr(U^L) = N(U^L) \Pr(U^L_w = U) . \tag{67}$$

From Eqs. (15) and (16), one obtains:

$$\Pr(U^L_w = U) = \exp(-LU), \tag{68}$$
$$N(U^L) = \exp(LS(U)) . \tag{69}$$

Then one immediately sees that:

$$I(U) = U - S(U) \ . \tag{70}$$

Thus, the large deviation rate function is closely related to the fluctuation spectrum.

Because S is a convex function of U, we have

$$\left(\frac{\partial^2 I}{\partial U^2}\right) \geq 0 \ , \tag{71}$$

and I would be a concave positive function of U. From eq. 21 it is easy to see that S and U are equal at $\beta = 1$, this means $I(U(1))$ would be zero. That means probability of typical set decays with exponent zero while we increase L. As we mentioned earlier $\beta = 1$ indicate typical set and we expect probability of typical set converges to 1, for large L which is the exact result that I tells us. For other values of β, we expect the probability of A^L_β converges to zero when we increase the length and $I(U(\beta))$ indicate how fast this convergence is.

4. Examples

As an example let's consider biased coin process. Fig.40 shows large deviation rate function versus energy for this process. Green line indicates $\beta = 1$ and purple line indicate $\beta = 0$. From eq. A13 it is clear that maximum of $S(U)$ happens at $\beta = 0$ which is shown with purple line in the figure. As it is shown I is zero at $\beta = 1$ which means we do not have any exponential decay of probability for typical set and probability of typical converges to one when we increase the length of the words. As it is clear from figure, for other subsets we have positive $I(U)$ and that means the probabilities of subsets decrease exponentially with the exponent $I(U)$ when the length of words increase. The most rare class for this process is at U_{max} in the case $\beta \to -\infty$ with exponent around 1.33. From concavity of $I(U)$ it's easy to see the most rare class happens either when β converge to positive infinity or negative infinity. The same calculation for Nemo process is shown in fig. 41. For this case the most rare class happens at U_{min} or in the case $\beta \to +\infty$.

A. Typical Process

Typical processes are those process which do not have non typical set. It means For those processes $\mathcal{A}^\infty$ has only two partitions, which are typical set and forbidden words. Every words that can be generate by a given typical process have the same probability. The partitions for a typical process is shown in fig. 42

Lemma 11. *Typical processes are uniquely determine by their forbidden words, or, for a given set of forbidden words there exists a unique process.*

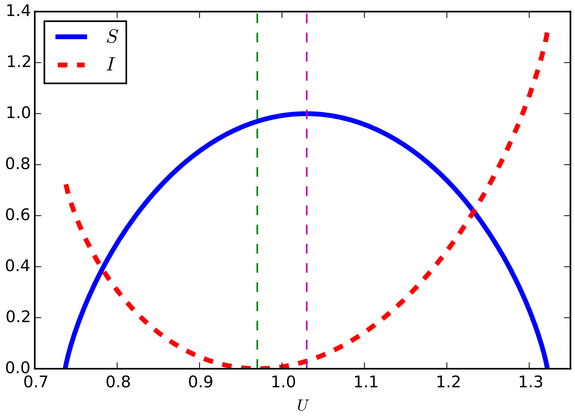

FIG. 40.
Large deviation rate function versus energy for Biased coin process with $p = 0.4$

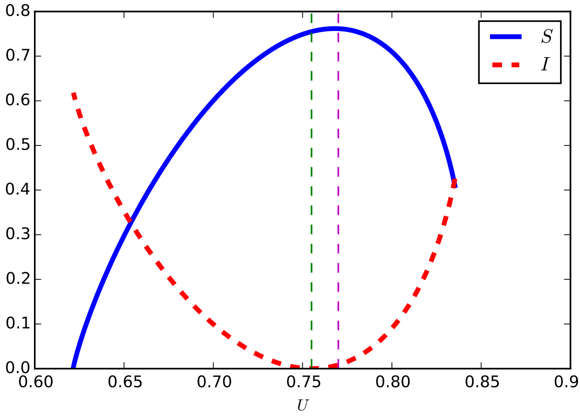

FIG. 41.
Large deviation rate function versus energy Nemo process with $p = 0.65$

Proof. *Having forbidden words, all the other words that are excluded has equal probability for a chosen length and that means we know their probability. Knowing all the probabilities for every words in $\mathcal{A}^\infty$ uniquely determine the process.*

Now we could label all the typical process with their set of forbidden words. We will note a typical process which it's set of forbidden words is $\mathcal{F}$ by $\tau^{\mathcal{F}}$.

We could also define the set of all typical process by

$$\mho = \{\tau^{\mathcal{F}} | \mathcal{F} \subset \mathcal{A}^\infty\}. \tag{72}$$

From the definition it is clear that all the typical pro-

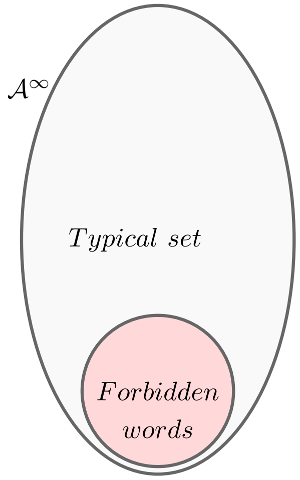

FIG. 42. Partitions for typical process

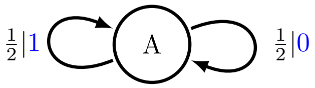

FIG. 43. Fair Coin Process ϵ-machine.

cess has one definite U, that means if we look at their FSD, it would be nonzero only at one definite U. Because that U belongs to a typical set, means $S(U) = U$. So as a result FSD graph for typical processes are just one point with equal $S(U)$ and U.

1. Fair Coin Process

A simplest example for typical process is fair coin process which is defined by the machine in the fig. 43.

Lemma 12. *Fair coin process is a typical process*

Proof. *First, There are no forbidden words for this process and every word can generate. Probabilities of every word with length n are the same and equal to 2^{-n}. That means*

$$\forall w \in \mathcal{A}^n : 2^{-n(H(X)+\epsilon)} \leq \Pr(w) \leq 2^{-n(H(X)-\epsilon)} \tag{73}$$

and that completes the proof.

starting with the biased coin process and changing p we could study how FSD changes. The results are shown in fig. 44. For this process we could calculate U_{max} and U_{min} analytically. From eq.'s 21 and 9 and knowing $S(U)$ at U_{max} and U_{min} for this process is zero, we could rewrite U_{max} and U_{min}

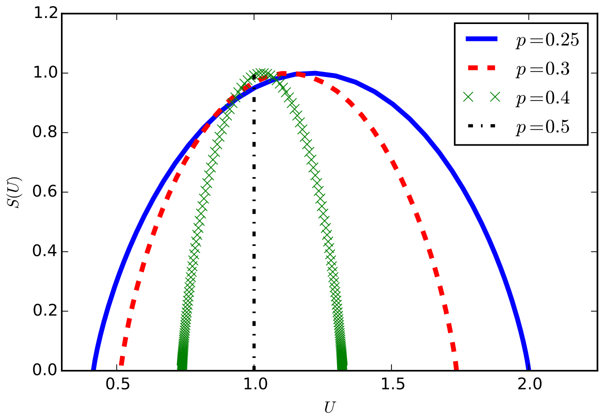

FIG. 44.
FSD for biased and fair coins

$$U_{max} = \lim_{\beta \to -\infty} \frac{-1}{\beta} (\log \widehat{\lambda}_\beta),$$
$$U_{min} = \lim_{\beta \to \infty} \frac{-1}{\beta} (\log \widehat{\lambda}_\beta). \tag{74}$$

we also have

$$\widehat{\lambda}_\beta = P^\beta + (1-p)^\beta, \tag{75}$$

and that means

$$U_{max} = -\log p,$$
$$U_{min} = -\log(1-p). \tag{76}$$

Defining $\Delta := U_{max} - U_{min}$, as a width for $S(U)$, we will have

$$\Delta = \log \frac{1-p}{p}. \tag{77}$$

2. Typical Nemo Process

Another example for typical process is a special case of Nemo process. Here we will find that process analytically. Let's calculate the process at arbitrary temperature.

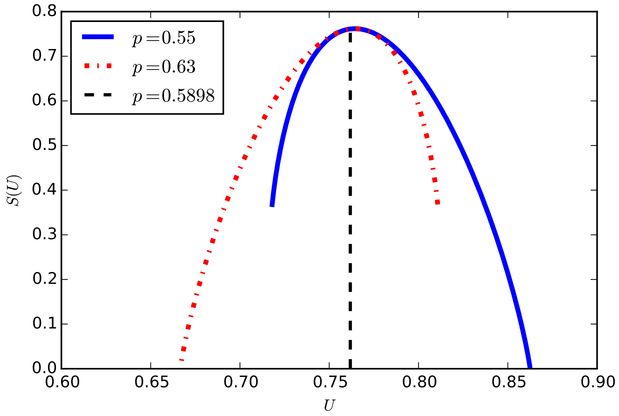

FIG. 45. FSD for Nemo process.

$$S_{0_\beta} := \begin{pmatrix} 0 & (1-\frac{p^\beta}{\widehat{\lambda}_\beta}) & 0 \\ 0 & 0 & 1 \\ \frac{1}{2} & 0 & 0 \end{pmatrix},$$

$$S_{1_\beta} := \begin{pmatrix} \frac{p^\beta}{\widehat{\lambda}_\beta} & 0 & 0 \\ 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 \end{pmatrix}. \qquad (78)$$

which $\widehat{\lambda}_\beta$ satisfies this equation

$$(\frac{\widehat{\lambda}_\beta}{p^\beta})^3 - (\frac{\widehat{\lambda}_\beta}{p^\beta})^2 - 2(\frac{\frac{1}{2}(1-p)}{p^3})^\beta = 0. \qquad (79)$$

For $p = p_0 \approx 0.5898$ which p_0 satisfies

$$2p_0^3 + p_0 - 1 = 0, \qquad (80)$$

equation 79 turns to

$$(\frac{\widehat{\lambda}_\beta}{p^\beta})^3 - (\frac{\widehat{\lambda}_\beta}{p^\beta})^2 - 2 = 0. \qquad (81)$$

This means $\frac{\widehat{\lambda}_\beta}{p^\beta}$ is independent from β and that leads to independence of S_{0_β} and S_{1_β} from temperature. We will prove in the next section that independence of a process from temperature are equivalent to be a typical process. So Nemo process with p_0 is a typical process. FSD for Nemo process for different p is shown in fig.45.

3. *Maximum Entropy versus Typical Process*

Lemma 13. *Maximum of $S(U)$ for a given ϵ-machine only depends on the topology of the machine (in other words the connectivity matrix, $\mathbf{T_0}$) and it is independent of the distribution over the edges.*

Proof. *From eq. A13 the maximum of $S(U)$ happens at $\beta = 0$. Using lemma. 1 and eq. 9, we have*

$$S(U(\beta = 0)) = \mathsf{h}(\beta = 0) = \log_2 \lambda_{\max}. \qquad (82)$$

which $\log_2 \lambda_{\max}$ is independent of distribution over the edges and only depends on topology of the ϵ-machine (connectivity matrix $\mathbf{T_0}$) and this completes the proof.

Theorem 4. *For a given topology (connectivity matrix $\mathbf{T_0}$) there exist a unique distribution over the edges that maximize metric entropy h_μ of ϵ-machines with the same topology.*

Proof. *Assuming an n states machine, for arbitrary state i there is two possibilities which is shown in fig. 46, either there are two exiting edges or one. For the first case in the ith row of the transfer matrix we only have two non zero elements we label the first column $f(i)$ and the second one $g(i)$. In the second case we only have one non zero element in ith row and we label that column $k(i)$.*

Our goal is to for a given topology find $T_{if(i)}$s in a way that maximize the metric entropy. We will use Lagrange multipliers method to do this. The constraints on maximization are

$$\begin{cases} \sum_{i=1}^n \pi_i = 1 \\ \forall i : \pi_i = \sum_{i=1}^n \pi_j T_{ji}. \end{cases} \qquad (83)$$

Hence we maximize

$$\psi = -\sum_{i=1}^n \pi_i (T_{if(i)} \log_2 T_{if(i)} + (1 - T_{if(i)}) \log_2 (1 - T_{if(i)}))$$
$$- a \sum_{i=1}^n \pi_i - \sum_{i=1}^n \mu_i (\pi_i - \sum_{j=1}^n \pi_j T_{ji}). \qquad (84)$$

For the states that we have two exiting edges we have

$$\begin{cases} \partial_{T_{if(i)}} \psi = \pi_i \{\log_2 T_{if(i)} - \log_2 (1 - T_{if(i)}) \\ -\mu_{f(i)} + \mu_{g(i)}\} = 0, \\ \partial_{\pi_i} \psi = \{T_{if(i)} \log_2 T_{if(i)} + (1 - T_{if(i)}) \log_2 (1 - T_{if(i)})\} \\ +a + \mu_i - \mu_{f(i)} T_{if(i)} - \mu_{g(i)} (1 - T_{if(i)}) = 0, \end{cases} \qquad (85)$$

and for the states with one exiting edge we have

$$\partial_{\pi_i} \psi = a + \mu_i - \mu_{k(i)} = 0. \qquad (86)$$

Solving the first set of equation leads to

$$\begin{cases} T_{if(i)} = \exp(-a + \mu_{f(i)} - \mu(i)), \\ T_{ig(i)} = \exp(-a + \mu_{g(i)} - \mu(i)). \end{cases} \qquad (87)$$

If we call the set of states with two exiting edges $\mathfrak{d}$ and the states with one exiting edge $\mathfrak{f}$ then we end up with this

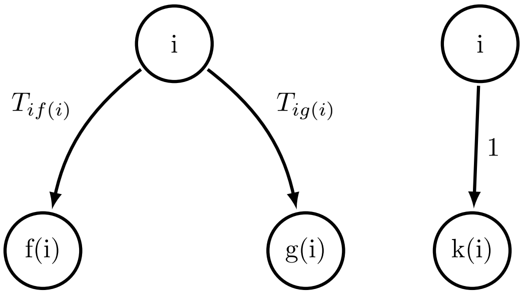

FIG. 46. for arbitrary state i there is two possibilities, either there are two exiting edges or one.

set of equations

$$\begin{cases} \forall i \in \mathfrak{d} \\ \exp(-a + \mu_{f(i)} - \mu(i)) + \exp(-a + \mu_{g(i)} - \mu(i)) = 1, \\ \forall i \in \mathfrak{f} \\ a + \mu_i - \mu_{k(i)} = 0. \end{cases} \tag{88}$$

Now we have n equations and $n+1$ unknown parameters ($\mu_i sanda$) but because in all equations difference of the $\mu_i s$ are appeared we could always set one of them and we end up with n unknown parameters. Solving these equations gives us the $T_{if(i)}$ that maximize metric entropy.

4. Examples

For the first example, we calculate the distribution over the edges for the topology which is shown in fig.10 that maximize metric entropy. Using theorem 4 we could rewrite the transfer matrix as

$$T = \begin{pmatrix} \exp(-a) & \exp(-a + \mu_2 - \mu_1) \\ 1 & 0 \end{pmatrix}, \tag{89}$$

and we have

$$\begin{cases} \exp(-a) + \exp(-a + \mu_2 - \mu_1), \\ a + \mu_2 - \mu_1 = 0. \end{cases} \tag{90}$$

Solving these equations the answer would be

$$T = \begin{pmatrix} \frac{\sqrt{5}-1}{2} & \frac{3-\sqrt{5}}{2} \\ 1 & 0 \end{pmatrix}, \tag{91}$$

which is the same answer that we calculated before.

For the next example consider the topology which is shown in fig. 47. Using theorem 4 the transfer matrix will be

$$T = \begin{pmatrix} 0 & e^{-a+\mu_2-\mu_1} & e^{-a+\mu_3-\mu_1} & 0 \\ 0 & e^{-a} & 0 & e^{-a+\mu_4-\mu_2} \\ 0 & 0 & e^{-a} & e^{-a+\mu_4-\mu_3} \\ 1 & 0 & 0 & 0 \end{pmatrix}. \tag{92}$$

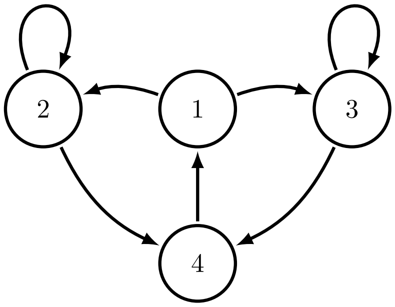

FIG. 47. Given topology.

and we have

$$\begin{cases} \exp(-a + \mu_2 - \mu_1) + \exp(-a + \mu_3 - \mu_1) = 1, \\ \exp(-a) + \exp(-a + \mu_4 - \mu_2) = 1, \\ \exp(-a) + \exp(-a + \mu_4 - \mu_3) = 1, \\ a + \mu_4 - \mu_1 = 0. \end{cases} \tag{93}$$

Solving these equations the answer would be

$$T = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & p_0 & 0 & 1-p_0 \\ 0 & 0 & p_0 & 1-p_0 \\ 1 & 0 & 0 & 0 \end{pmatrix}. \tag{94}$$

which p_0 satisfies

$$2p_0^3 + p_0 - 1 = 0, \tag{95}$$

the same equations that we derived for typical Nemo process and because of the symmetry in the answer the number of states reduce to three and the ϵ-machine looks like fig. 22.

Lemma 14. *The distribution that maximize the metric entropy for any n states topology which every state has two exiting edges is half and half over any pair of exiting edges.*

Proof. *It is easy to see that he answer for set of equations 88 in theorem 4 for this case is*

$$\begin{cases} \forall 1 \leq i \leq n : \mu_i = c \\ \exp(-a) = \frac{1}{2} \end{cases} \tag{96}$$

and that means

$$\forall 1 \leq i \leq n : T_{if(i)} = T_{if(i)} = \frac{1}{2}. \tag{97}$$

For $n > 1$ this answer is not minimal, the minimal version of this answer is a fair coin process. This means for any n states topology which every state has two exiting edges the answer for the process with maximum entropy with the same topology is somehow trivial and it is a fair coin process.

B. Thermodynamic Classes in Process Space

Lemma 15. *For a mapping defined by Eq. (31) we have:*

$$\mathcal{M}_{\beta_1}\mathcal{M}_{\beta_2} = \mathcal{M}_{\beta_1\beta_2} \ . \tag{98}$$

Proof. *Using Lemma 4 the proof is direct.*

Lemma 16. *Duality: Looking at the process T at $\beta = \beta_0$ is equivalent to look at the process $S_{\beta'}$ at $\beta = \beta_0\beta'^{-1}$.*

Proof. *Using Lemma 15:*

$$\mathcal{M}_{\beta_0\beta'^{-1}} S_{\beta'} = \mathcal{M}_{\beta_0\beta'^{-1}}\mathcal{M}_{\beta'} T = \mathcal{M}_{\beta_0} T \ . \tag{99}$$

Lemma 17. *Any process belonging to $\mho$ is invariant under temperature changes.*

Proof. *For any process belonging to $\mho$ all the probabilities for words with same length are equal. When we change the temperature after the twist using Lemma 4 the probabilities are still equal and that means the probabilities are invariant under temperature changes and that complete the proof.*

For a given process T we define a set by:

$$\mathcal{D}_{\mathcal{T}} = \{\mathcal{M}_\beta T | \beta \in \mathbb{R}\} \setminus \mho \ . \tag{100}$$

This set includes all processes that can be generated from process T by changing the temperature, excluding any typical process from it.

Lemma 18. *Any two processes from $\mathcal{D}_{\mathcal{T}}$ convert to each other via temperature change.*

Proof. *From the definition of $\mathcal{D}_{\mathcal{T}}$ we could write two process as:*

$$\begin{aligned} \mathfrak{A} &= \mathcal{M}_{\beta_1} T \\ \mathfrak{B} &= \mathcal{M}_{\beta_2} T \ . \end{aligned} \tag{101}$$

Using Lemma 15:

$$\mathfrak{B} = \mathcal{M}_{\beta_2}\mathcal{M}_{\beta_1^{-1}} T = \mathcal{M}_{\beta_2\beta_1^{-1}} T \ , \tag{102}$$

and this completes the proof.

Lemmas 18 and 17 say that for any arbitrary process either it is typical or belongs to a partition with infinite members that convert into each other via a temperature change. Thus, as Fig. 48 illustrates, the mapping partitions process space to infinite classes, some of which with only one member and others infinite members. Within each all members convert into each other by changing temperature.

Lemma 19. *For a given process T, FSD function for $S_\beta = \mathcal{M}_\beta T$ is related to FSD function of T via:*

$$S_{S_\beta}(U) = S_T\left(\frac{U}{\beta} + \frac{(\beta-1)\mathsf{h}(\beta)}{\beta}\right) \ . \tag{103}$$

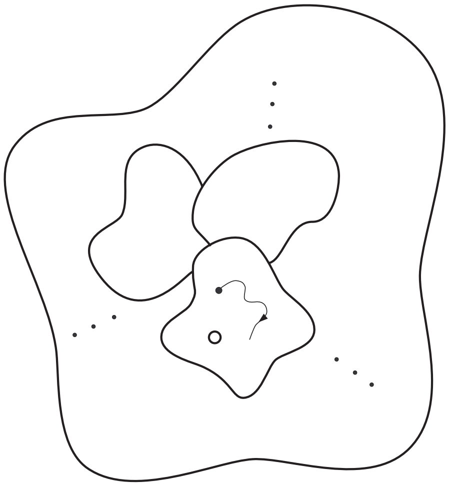

FIG. 48.
Thermodynamic classes in process space.

Proof. *Consider two different energy densities U_1 and U_2, they map to new ones U_1' and U_2' when we map T to S_β. Using Thm. 4 we have:*

$$U_2' - U_1' = \beta(U_2 - U_1) \ , \tag{104}$$

This means we could write the mapping between any U' and U as:

$$U' = \beta U + \mathfrak{C}(\beta) \ .$$

Now, we should find $\mathfrak{C}(\beta)$.

We also know that $U = U(\beta)$ would map to $U' = S(U(\beta))$, because for our new process S_β, $S_{S_\beta}(U)$ should be equal to U at new $\beta = 1$, that gives us:

$$\mathfrak{C}(\beta) = S(U(\beta)) - \beta U(\beta) \ .$$

Using Lemma 2 we have:

$$\mathfrak{C}(\beta) = -(\beta - 1)\mathsf{h}(\beta) \ ,$$

and the mapping between U' and U would be:

$$U' = \beta U - (\beta - 1)\mathsf{h}(\beta) \ .$$

We also have:

$$S_{S_\beta}(U') = S_T(U) \ ,$$

completing the proof.

VII. BEYOND THE LIMITS

Until now we have had several limitation. First have been working on processes with finite number of states and second all of the process were ergodic ones. Here we will go beyond these limitation.

A. Infinite-State Processes

In this section we will look at the processes with infinite number of states. The problem with these processes is that calculating eigenvalues and eigenstates either analytically or computationally is hard for Hilbert space and there are limited number of processes that we could exactly calculate these quantities for them.

We will not do any calculation for any specific process but we will prove a very powerful theorem that gives us an intuition about the difference between processes with finite and infinite number of states.

Consider a sequence of ϵ*-machines* $\Sigma = \{M_n\}$ *which is a list of* ϵ*-machines* $M_1, M_2, ..., M_{n-1}$ *and* M_n *such that each* M_i *has these conditions:*
1. It has i *number of states*
2. State i *has a self loop with probability* c^{-i} *which* $c < 1$ *that generate symbol* s
3. That self loop has the minimum probability between probabilities over all the edges.
4. There are no close path from any of states (except state i*) to that state itself that only generate symbol* s*.*

We define a process M^{Σ} *by*

$$M^{\Sigma} = \lim_{n \to \infty} M_n \tag{105}$$

Theorem 5. *FSD* $S(U)$ *for the process* M^{σ} *has a vertical asymptote when* U *converges to infinity (fig. 49).*

Proof. *Consider* M_n *and let us look at the word* $w = ss...s$ *which has* $2n$ *zeros. Because the condition 4, for the probability of generating this word by* M_n *we will have*

$$\Pr(w) > \Pr(\sigma_i)(c^{-n})^n, \tag{106}$$

and the energy density for this word would be

$$U_w = -\frac{\log_2 \Pr(w)}{n} > n(-\log_2(c)). \tag{107}$$

When we increase the n*, this energy increases linearly with* n *and we also know*

$$U_{max} \geq U_w, \tag{108}$$

that means

$$\lim_{n \to \infty} U_{max} \to \infty. \tag{109}$$

Now using theorem 3 and the result that we just derived $S(U)$ *should has a vertical asymptote when* U *converges to infinity.*

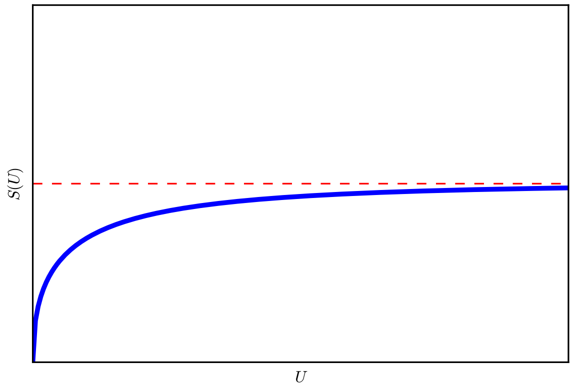

FIG. 49.
S versus U for M^{σ}

Let us look at an example for a sequence Σ*. The nth member of* Σ *is shown in fig. 50. This* ϵ*-machine satisfies all four conditions and that means FSD* $S(U)$ *for the process* M^{σ} *has a horizontal asymptote when* U *converges to infinity.*

B. Nonergodic Processes

In this section we will look at the non ergodic processes. For these process having one realization of data with infinite length is not enough to determine the process. An example for these processes is shown in fig. 52. If we start at state A *and look at the realization only for once we definitely miss one of the branches.*

Consider a non ergodic process like the processes in fig. 51 which each of $M_1, M_2, ..., M_{n-1}$ *are* ϵ*-machine's and* i *is a starting state. At start we are in state* i *then the finite size word* w_k *would generate with probability* p_i *and we will end up at one of the states of* M_k*. That means at first we have a stochastic transient behavior and then one of the machines will be chosen.*

Theorem 6. *For a process which is shown in fig. 51, FSD would be*

$$\forall U > 0 : S(U) = sup\{S_j(U) : 1 \leq j \leq n\}, \tag{110}$$

which $S_j(U)$ *is FSD for* M_j*.*

Proof. *In the limit of* $L \to \infty$ *using eq. 16 and having*

$$N(U^L) = \sum_{j=1}^{n} N_j(U^L), \tag{111}$$

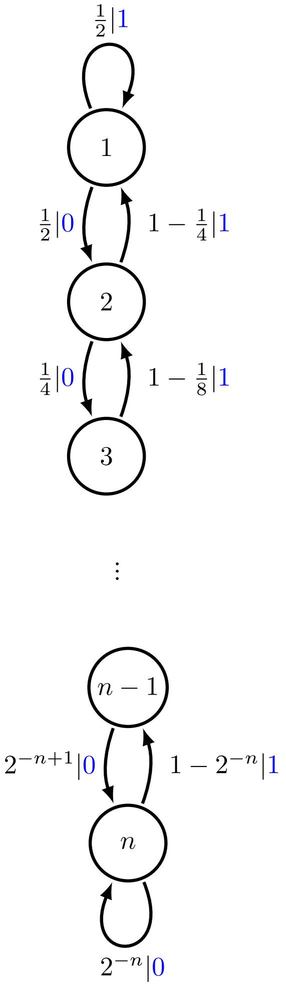

FIG. 50. M_n.

which in $N_j(U^L)$, index j refer to M_j, we will have

$$N(U^L)=\sum_{j=1}^{n}\exp(S_j(U)L). \tag{112}$$

That means we have

$$\begin{aligned}S(U)&=\lim_{L\to\infty}\frac{1}{L}\log_2\sum_{j=1}^{n}\exp(S_j(U)L)\\&=sup\{S_j(U):1\leq j\leq n\}.\end{aligned} \tag{113}$$

As an example for theorem 6 let us look at the process which is shown in fig. 52. Using the theorem FSD for this process would be the graph in fig. 53.

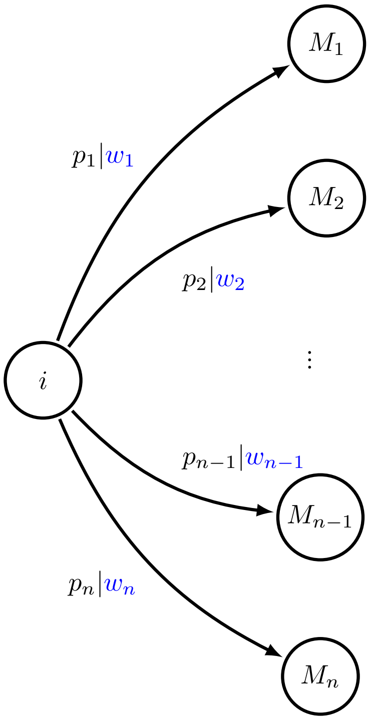

FIG. 51. An example for non ergodic process.

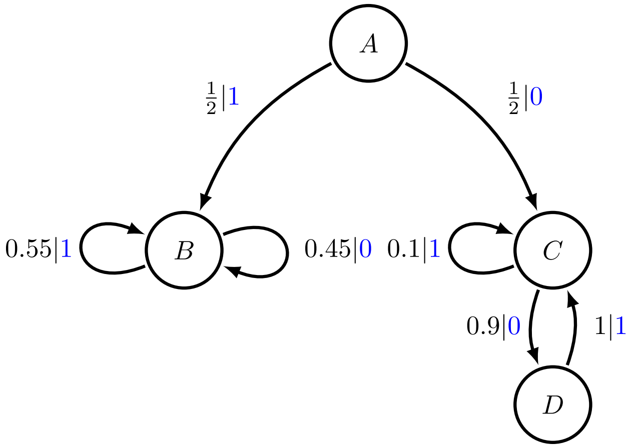

FIG. 52. An example for $n=2$ for the process which is shown in fig. 51.

Appendix A: Proofs

1. Theorem 1 and a Lemma

Theorem 1 A process' Shannon entropy rate can be directly calculated from its ϵ-machine using:

$$h_\mu=-\sum_{i=1}^{N}\Pr(\sigma_i)\sum_{x\in\mathcal{A}}\Pr(x|\sigma_i)\log_2\Pr(x|\sigma_i)\ .$$

Proof. *First, we introduce state-symbol ordered pairs*

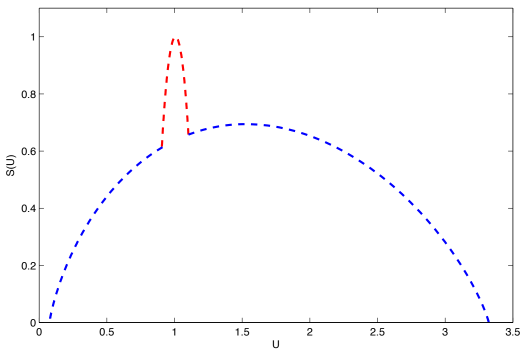

FIG. 53.
FSD versus energy for the process which is shown in fig. 52

$y_i = (\sigma_i, x_i)$ and note that:

$$\Pr(y_{0:\ell}) = \Pr(y_0)\prod_{i=1}^{\ell}\Pr(y_i|y_{i-1}) \ .$$

Defining:

$$\Lambda(y_{0:}) = -\lim_{\ell\to\infty}\frac{1}{\ell}\sum_{\{y_{0:\ell}\}}\Pr(y_{0:\ell})\log_2\Pr(y_{0:\ell}) \ ,$$

we have:

$$\Lambda(y_{0:}) := \lim_{\ell\to\infty}\frac{1}{\ell}\big(-\sum_{\{y_{0:\ell}\}}\Pr(y_0)\prod_{i=1}^{\ell-1}\Pr(y_i|y_{i-1})\log_2\Pr(y_1) \\ -\sum_{j=1}^{\ell}\sum_{y's}\Pr(y_1)\prod_{i=1}^{\ell}\Pr(y_i|y_{i-1})\log_2\Pr(y_j|y_{j-1})\big) \ .$$

To calculate the first term in the righthand side, first sum over y_n, then y_{n-1}, and so on. After summing over y_1 we have:

$$\sum_{\{y_{0:\ell}\}}\Pr(y_0)\prod_{i=1}^{\ell-1}\Pr(y_i|y_{i-1})\log_2\Pr(y_0) \\ = \sum_{y_0}\Pr(y_0)\log_2\Pr(y_0) \ .$$

Now, let's calculate the following summation for arbitrary j:

$$\sum_{\{y_{0:\ell}\}}\Pr(y_0)\prod_{i=1}^{\ell-1}\Pr(y_i|y_{i-1})\log_2\Pr(y_j|y_{j-1}) \ .$$

Summing over y_n, y_{n-1}, and so on to y_{j+1} and then y_1, y_2, and so on till y_{j-2} leads to:

$$\sum_{\{y_{0:\ell}\}}\Pr(y_1)\prod_{i=2}^{n}\Pr(y_i|y_{i-1})\log_2\Pr(y_j|y_{j-1}) \\ = \sum_{y_{j-1},y_j}\Pr(y_j)\Pr(y_j|y_{j-1})\log_2\Pr(y_j|y_{j-1}) \\ = \sum_{y_1,y_2}\Pr(y_1)\Pr(y_2|y_1)\log_2\Pr(y_2|y_1) \ .$$

Now, the entropy rate may be written as:

$$\Lambda(y_{0:}) = -\lim_{\ell\to\infty}\frac{1}{\ell}\big(\sum_{y1}\Pr(y_1)\log_2\Pr(y_1) + \\ (\ell-1)\sum_{y_1,y_2}\Pr(y_1)\Pr(y_2|y_1)\log_2\Pr(y_2|y_1)\big) \\ = -\sum_{y_1,y_2}\Pr(y_1)\Pr(y_2|y_1)\log_2\Pr(y_2|y_1) \\ = -\sum_{\sigma_1,x_1,\sigma_2,x_2}\Pr(\sigma_1,x_1)\Pr(\sigma_2,x_2|\sigma_1,x_1) \\ \log_2\Pr(\sigma_2,x_2|\sigma_1,x_1) \ .$$

Using:

$$\Pr(\sigma_2,x_2|\sigma_1,x_1) = \Pr(\sigma_2|\sigma_1,x_1)\Pr(x_2|\sigma_2,\sigma_1,x_1) \\ = \Pr(\sigma_2|\sigma_1,x_1)\Pr(x_2|\sigma_2) \ ,$$

we have:

$$\Lambda(y_{0:\ell}) = -\sum_{\sigma_1,x_1,\sigma_2,x_2}\Pr(\sigma_1,x_1)\Pr(\sigma_2|\sigma_1,x_1)\Pr(x_2|\sigma_2) \\ \big(\log_2\Pr(\sigma_2|\sigma_1,x_1) + \log_2\Pr(x_2|\sigma_2)\big) \ .$$

$\Pr(\sigma_2|\sigma_1,x_1)$ takes only two values (0 or 1) and so:

$$\Pr(\sigma_2|\sigma_1,x_1)\log_2\Pr(\sigma_2|\sigma_1,x_1) = 0 \ ,$$

which leads to:

$$\Lambda(y_{0:\ell}) = \\ -\sum_{\sigma_1,x_1,\sigma_2,x_2}\Pr(\sigma_1,x_1)\Pr(\sigma_2|\sigma_1,x_1)\Pr(x_2|\sigma_2)\log_2\Pr(x_2|\sigma_2)) \ .$$

Finally, summing over σ_1 and x_1 one arrives at:

$$\Lambda(y_{0:\ell}) = -\sum_{\sigma_2,x_2}\Pr(\sigma_2)\Pr(x_2|\sigma_2)\log_2\Pr(x_2|\sigma_2) \ .$$

Lemma 20. $\lim_{\ell\to\infty}\Lambda(y_{0:\ell}) = h_\mu$.

Proof. *From the definition of an ϵ-machine:*

$$\Lambda(y_{0:\ell}) = \frac{1}{\ell}\mathrm{H}((\sigma,x)_{0:\ell}) \\ = \frac{1}{\ell}\mathrm{H}(\sigma_1,x_{0:\ell}) \ ,$$

that means:

$$\Lambda(y_{0:\ell}) - \frac{1}{\ell}\,\mathrm{H}(x_{0:\ell}) = \frac{1}{\ell}\,\mathrm{H}(\sigma_1, x_{0:\ell}) - \frac{1}{\ell}\,\mathrm{H}(x_{0:\ell}) = \frac{1}{\ell}\,\mathrm{H}(\sigma_1 | x_{0:\ell}) \leq \frac{1}{\ell}\,\mathrm{H}(\sigma_1) ~,$$

where in the infinite limit of ℓ:

$$\lim_{\ell\to\infty} |\Lambda(y_{0:\ell}) - h_\mu| \leq 0 ~.$$

And, this complete the proof.

2. Lemma 1

Lemma 1 The topological and Shannon entropy rates are special cases of the Renyi entropy rate:

$$h = \mathsf{h}(\beta = 0)$$

and

$$h_\mu = \mathsf{h}(\beta \to 1) ~,$$

respectively.

Proof. *From Eq. (20) we have:*

$$\mathsf{h}(\beta = 0) := \lim_{\ell\to\infty} \frac{1}{\ell} \log_2 \sum_{\{w\in\mathcal{A}^\ell, \Pr(w)>0\}} 1 ~.$$

Recalling that:

$$N(\ell) = \sum_{\{w\in\mathcal{A}^L, \Pr(w)>0\}} 1 ~,$$

completes the proof's first part. For the next, we have:

$$\mathsf{h}(\beta \to 1) = \lim_{\beta\to1} \lim_{\ell\to\infty} \frac{1}{\ell(1-\beta)} \log_2 \sum_{\{w\in\mathcal{A}^\ell\}} (\Pr(w))^\beta ~,$$

and using L'Hôpital's rule gives:

$$\begin{aligned}\mathsf{h}(\beta \to 1) &= -\lim_{\beta\to1} \lim_{\ell\to\infty} \frac{1}{\ell} \frac{\sum_{\{w\in\mathcal{A}^\ell\}} (\Pr(w))^\beta \log_2 \Pr(w)}{\sum_{\{w\in\mathcal{A}^\ell\}} (\Pr(w))^\beta} \\ &= -\lim_{\ell\to\infty} \frac{1}{\ell} \sum_{\{w\in\mathcal{A}^\ell\}} \Pr(w) \log_2 \Pr(w) \\ &= h_\mu ~.\end{aligned}$$

3. Lemma 2

Lemma 2 The Renyi entropy rate $\mathsf{h}(\beta)$ and thermodynamic entropy density $S(U)$ are related by:

$$S(U(\beta)) = \beta U(\beta) - (\beta - 1)\mathsf{h}(\beta) ~, \qquad \text{(A1)}$$

where:

$$U(\beta) = \underset{u\in U^\infty}{\operatorname{argmax}} ~(S(u) - \beta u) ~. \qquad \text{(A2)}$$

Proof. *We can translate the word probabilities in Eq. (18) to energies via:*

$$\Pr(w) = \exp(-u\ell) ~.$$

Then, recalling that $e^{S(u)\ell}$ is the size of the class of words with a given probability leads to:

$$(1-\beta)\mathsf{h}(\beta) = \lim_{\ell\to\infty} \frac{1}{\ell} \log \sum_u \exp((S(u) - \beta u)\ell) ~.$$

At fixed β there exists a unique u that maximizes $(S(u) - \beta u)$. So, we define a function $U(\beta)$ that assigns a unique u to each β:

$$U(\beta) = \underset{u\in U^\infty}{\operatorname{argmax}} ~(S(u) - \beta u) ~. \qquad \text{(A3)}$$

We call $U(\beta)$ energy density. *Then, we have:*

$$\begin{aligned}(1-\beta)\mathsf{h}(\beta) &= \lim_{\ell\to\infty} \frac{1}{\ell} \log \big\{ \exp((S(U(\beta)) - \beta U(\beta))\ell)(1 + \sum_{u\neq U(\beta)} \exp(((S(U(\beta)) - \beta U(\beta)) - (S(u) - \beta u))\ell)) \big\} \\ &= S(U(\beta)) - \beta U(\beta)) + \lim_{\ell\to\infty} \frac{1}{\ell} \log \big\{ 1 + \sum_{u\neq U(\beta)} \exp(((S(U(\beta)) - \beta U(\beta)) - (S(u) - \beta u))\ell) \big\} ~.\end{aligned}$$

The exponent in the second term in the sum is always negative and so the second term vanishes, completing the proof.

4. Lemma 3

Lemma 3 The energy density and Renyi entropy are related by:

$$U(\beta) = \frac{\partial}{\partial\beta}((\beta - 1)\mathsf{h}(\beta)) ~. \qquad \text{(A4)}$$

Proof. *Using Eq. (20), we have:*

$$\begin{aligned}
&\frac{\partial}{\partial\beta}(\beta-1)\mathsf{h}(\beta) \\
&\quad = -\lim_{\ell\to\infty}\frac{1}{\ell}\frac{\displaystyle\sum_{\{w\in\mathcal{A}^\ell\}}(\Pr(w))^\beta\log(\Pr(w))}{\mathcal{Z}(\beta)} \\
&\quad = -\lim_{\ell\to\infty}\frac{1}{L}\frac{\displaystyle\sum_u \exp((S(u)-\beta u)L)(-uL)}{\displaystyle\sum_u \exp((S(u)-\beta u)L)}\ ,
\end{aligned} \tag{A5}$$

where we changed from words to their probability classes. Using the energy density of Eq. (A3) and saving the dominant terms in the summation, similar to previous proof, we have:

$$\begin{aligned}
&\frac{\partial}{\partial\beta}(\beta-1)\mathsf{h}(\beta) \\
&\quad = -\lim_{\ell\to\infty}\frac{1}{\ell}\frac{\exp((S(U(\beta))-\beta U(\beta)\ell)(-U(\beta)\ell)}{\exp((S(U(\beta))-\beta U(\beta)\ell)} \\
&\quad = U(\beta)\ ,
\end{aligned}$$

and this complete the proof.

5. Theorem 3

Theorem 3 $S(U)$ is a convex function of U, where the former is given by Eq. (**??**), the latter by Eq. (**??**).

Proof.

$$\begin{aligned}
\frac{\mathrm{d}\widehat{\lambda}_\beta}{\mathrm{d}\beta} &= \sum_{i,j}(\widehat{\mathbf{l}}_\beta)_j\frac{\mathrm{d}(\boldsymbol{T}_\beta)_{ij}}{\mathrm{d}\beta}(\widehat{\mathbf{r}}_\beta)_j \\
&= \frac{1}{\beta}\sum_{i,j}(\widehat{\mathbf{l}}_\beta)_j(\boldsymbol{T}_\beta)_{ij}(\widehat{\mathbf{r}}_\beta)_j\log(\boldsymbol{T}_\beta)_{ij}.
\end{aligned} \tag{A6}$$

From Eqs. (31) and (33) one sees that:

$$(\widehat{\mathbf{l}}_\beta)_j(\boldsymbol{T}_\beta)_{ij}(\widehat{\mathbf{r}}_\beta)_j = \widehat{\lambda}_\beta(\boldsymbol{P}_\beta)_i(\boldsymbol{S}_\beta)_{ij}\ . \tag{A7}$$

Using this in Eq. (A6) gives:

$$\begin{aligned}
\frac{1}{\widehat{\lambda}_\beta}\frac{\mathrm{d}\widehat{\lambda}_\beta}{\mathrm{d}\beta} &= \frac{1}{\beta}\sum_{i,j}(\boldsymbol{P}_\beta)_i(\boldsymbol{S}_\beta)_{ij}\log(\boldsymbol{T}_\beta)_{ij} \\
&= \frac{1}{\beta}\sum_{i,j}(\boldsymbol{P}_\beta)_i(\boldsymbol{S}_\beta)_{ij}\Big[\log(\boldsymbol{S}_\beta)_{ij}+\log\widehat{\lambda}_\beta \\
&\qquad +\log(\widehat{\mathbf{r}}_\beta)_i-(\widehat{\mathbf{r}}_\beta)_j] \\
&= -\frac{S(U(\beta))}{\beta}+\frac{\log\widehat{\lambda}_\beta}{\beta} \\
&\qquad +\frac{1}{\beta}\sum_{i,j}(\boldsymbol{P}_\beta)_i(\boldsymbol{S}_\beta)_{ij}\left[\log(\widehat{\mathbf{r}}_\beta)_i-(\widehat{\mathbf{r}}_\beta)_j\right]\ .
\end{aligned} \tag{A8}$$

To obtain the first term above, the definition of entropy is used and for the second term one makes use of:

$$\begin{aligned}
\sum_{i,j}(\boldsymbol{P}_\beta)_i(\boldsymbol{S}_\beta)_{ij} &= \sum_j(\boldsymbol{P}_\beta)_j \\
&= \sum_j(\widehat{\mathbf{r}}_\beta)_j(\widehat{\mathbf{l}}_\beta)_j \\
&= \boldsymbol{l}_\beta\cdot\boldsymbol{r}_\beta \\
&= 1\ .
\end{aligned} \tag{A9}$$

Now, using:

$$\sum_i(\boldsymbol{P}_\beta)_i(\boldsymbol{S}_\beta)_{ij} = (\boldsymbol{P}_\beta)_j = (\widehat{\mathbf{r}}_\beta)_j(\widehat{\mathbf{l}}_\beta)_j\ , \tag{A10}$$

the third and the fourth terms in Eq. (A8) simply cancel, and one arrives at:

$$\frac{\mathrm{d}}{\mathrm{d}\beta}(\log\widehat{\lambda}_\beta) = -U(\beta)\ . \tag{A11}$$

*Then one may take $S(\cdot)$ as a function of β. Multiplying both sides of Eq. (**??**) by β and differentiating both sides with respect to β, one obtains:*

$$U+\beta\frac{\mathrm{d}U}{\mathrm{d}\beta} = \frac{\mathrm{d}S}{\mathrm{d}\beta}-\frac{\mathrm{d}}{\mathrm{d}\beta}(\log\widehat{\lambda}_\beta), \tag{A12}$$

Using Eq. (A11), one finds:

$$\begin{aligned}
\beta &= \frac{\mathrm{d}S/\mathrm{d}\beta}{\mathrm{d}U/\mathrm{d}\beta} \\
&= \frac{\mathrm{d}S}{\mathrm{d}U}\ .
\end{aligned} \tag{A13}$$

Thus, β indeed plays the same role here as the inverse temperature in statistical physics.

Now, let's consider the convexity of entropy with re-

spect to U. Since:

$$\begin{aligned}\frac{\mathrm{d}^2 S}{\mathrm{d}U^2} &= \frac{\mathrm{d}\beta}{\mathrm{d}U} \\ &= -\left(\frac{\mathrm{d}^2}{\mathrm{d}\beta^2}\log\widehat{\lambda}_\beta\right)^{-1} ~,\end{aligned} \tag{A14}$$

the thermodynamic entropy will be convex with respect to U, as long as $\log\widehat{\lambda}_\beta$ is convex with respect to U. For any two constant matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ [50] and any integer number N:

$$\boldsymbol{B}\boldsymbol{T}_\beta^N\boldsymbol{A} = C\widehat{\lambda}_\beta^N + \cdots ~, \tag{A15}$$

where for large N, the dominant term is $C\widehat{\lambda}_\beta^N$. In any case, $\boldsymbol{B}\boldsymbol{T}_\beta^N\boldsymbol{A}$ contains summation of multiplication of functions of the type $\Pr(x|\sigma_i)^\beta$ or:

$$\sum \exp\left\{\beta\sum\Pr(x|\sigma_i)\right\} ~. \tag{A16}$$

Then, differentiating $\log(\boldsymbol{B}\boldsymbol{T}_\beta^N\boldsymbol{A})$ twice with respect to β results in a positive quantity. However, for large N, this is proportional to two times differentiation of $\log\widehat{\lambda}_\beta$ with respect to β. This completes the proof.

[1] C. H. Bennett. Thermodynamics of computation - a review. *Intl. J. Theo. Phys.*, 21:905, 1982.

[2] S. Still and J. P. Crutchfield. Structure or noise? 2007. Santa Fe Institute Working Paper 2007-08-020; arxiv.org physics.gen-ph/0708.0654.

[3] S. Still, J. P. Crutchfield, and C. J. Ellison. Optimal causal inference: Estimating stored information and approximating causal architecture. *CHAOS*, 20(3):037111, 2010.

[4] S. Marzen and J. P. Crutchfield. Circumventing the curse of dimensionality in prediction:causal rate-distortion for infinite-order markov processes. page in preparation, 2014. SFI Working Paper 14-12-047; arxiv.org:1412.2859 [cond-mat.stat-mech].

[5] R. Bowen and D. Ruelle. *Invent. Math.*, 29:181, 1975.

[6] D. Ruelle. *Thermodynamic Formalism.* Addison-Wesley, Reading, 1978.

[7] Y. Oono and Y. Takahashi. *Prog. Theo. Phys.*, 63:1804, 1980.

[8] H. Touchette. The large deviation approach to statistical mechanics. *Physics Reports*, 478:1–69, 2009.

[9] M. C. Mackey, editor. *Time's Arrow: The Origins of Thermodynamic Behavior.* Springer-Verlag, New York, 1992.

[10] C. Beck and F. Schlögl. *Thermodynamics of Chaotic Systems.* Cambridge University Press, 1993.

[11] J. R. Dorfman. *An Introduction to Chaos in Nonequilibrium Statistical Mechanics.* Cambridge University Press, Cambridge, United Kingdom, 1999.

[12] D. Ruelle. Conversations on nonequilibrium physics with an extraterrestrial. *Physics Today*, 57(5):48–53, May 2004.

[13] J. P. Crutchfield and K. Young. Inferring statistical complexity. *Phys. Rev. Let.*, 63:105–108, 1989.

[14] J. P. Crutchfield. Between order and chaos. *Nature Physics*, 8(January):17–24, 2012.

[15] D. J. Evans, E. G. D. Cohen, and G. P. Morriss. Probability of second law violations in shearing steady flows. *Phys. Rev. Lett.*, 71:2401–2404, 1993.

[16] G. Gallavotti and E. G. D. Cohen. Dynamical ensembles in stationary states. *J. Stat. Phys.*, 80:931–970, 1995.

[17] C. Jarzynski. Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.*, 78(14):2690–2693, 1997.

[18] G. E. Crooks. Nonequilibrium measurements of free energy differences for microscopically reversible markovian systems. *J. Stat. Phys.*, 90(5/6):1481–1487, 1998.

[19] R. Klages, W. Just, and C. Jarzynski, editors. *Nonequilibrium Statistical Physics of Small Systems: Fluctuation Relations and Beyond.* Wiley, New York, 2013.

[20] J. A. Bucklew. *Large Deviation Techniques in Decision, Simulation, and Estimation.* Wiley-Interscience, New York, 1990.

[21] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley-Interscience, New York, second edition, 2006.

[22] O. Penrose. *Foundations of statistical mechanics; a deductive treatment.* Pergamon Press, Oxford, 1970.

[23] H. B. Callen. *Thermodynamics and an Introduction to Thermostatistics.* Wiley, New York, second edition, 1985a.

[24] J. P. Crutchfield and D. P. Feldman. Regularities unseen, randomness observed: Levels of entropy convergence. *CHAOS*, 13(1):25–54, 2003.

[25] K. Young and J. P. Crutchfield. Fluctuation spectroscopy. *Chaos, Solitons, and Fractals*, 4:5 – 39, 1994.

[26] P. Gaspard. Time-reversed dynamical entropy and irreversibility in markovian random processes. *J. Stat. Phys.*, 117(3/4):599–615, 2004.

[27] Unifilar is known as "deterministic" in the finite automata literature [51]. Here, we avoid the latter term so that confusion does not arise due to the stochastic nature of the models being used. It is referred to as "right-resolving" in symbolic dynamics [52].

[28] J. P. Crutchfield, C. J. Ellison, J. R. Mahoney, and R. G. James. Synchronization and control in intrinsic and designed computation: An information-theoretic analysis of competing models of stochastic computation. *CHAOS*, 20(3):037105, 2010. Santa Fe Institute Working Paper 10-08-015; arxiv.org:1007.5354 [cond-mat.stat-mech].

[29] A left (right) stochastic matrix is one all of whose elements are real and nonnegative and the sum of each column (rows) of this matrix is equal to one.

[30] B. D. Johnson, J. P. Crutchfield, C. J. Ellison, and C. S. McTague. Enumerating finitary processes. page submitted, 2012. SFI Working Paper 10-11-027; arxiv.org:1011.0036 [cs.FL].

[31] A. N. Kolmogorov. A new metric invariant of transient dynamical systems and automorphisms in Lebesgue spaces. *Dokl. Akad. Nauk. SSSR*, 119:861, 1958. (Russian) Math. Rev. vol. 21, no. 2035a.

[32] A. N. Kolmogorov. Entropy per unit time as a metric invariant of automorphisms. *Dokl. Akad. Nauk. SSSR*, 124:754, 1959. (Russian) Math. Rev. vol. 21, no. 2035b.

[33] Ja. G. Sinai. On the notion of entropy of a dynamical system. *Dokl. Akad. Nauk. SSSR*, 124:768, 1959.

[34] R. G. James, C. J. Ellison, and J. P. Crutchfield. Anatomy of a bit: Information in a time series observation. *CHAOS*, 21(3):037109, 2011.

[35] D. M. Cvetkovic, M. Doob, and H. Sachs. *Spectra of graphs: Theory and application*, volume 413. Academic press, New York, 1980.

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

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

[38] J. P. Crutchfield and C. J. Ellison. The past and the future in the present. 2014. SFI Working Paper 10-12-034; arxiv.org:1012.0356 [nlin.CD].

[39] C. J. Ellison, J. R. Mahoney, R. G. James, J. P. Crutchfield, and J. Reichardt. Information symmetries in irreversible processes. *CHAOS*, 21(3):037107, 2011.

[40] J. P. Crutchfield, P. Riechers, and C. J. Ellison. Exact complexity: Spectral decomposition of intrinsic computation. submitted. Santa Fe Institute Working Paper 13-09-028; arXiv:1309.3792 [cond-mat.stat-mech].

[41] L. L. Campbell. A coding theorem and Renyi's entropy. *Info. Control*, 8:423, 1965.

[42] Z. Rached, A. Fady, and L. L. Campbell. Renyi's entropy rate for discrete Markov sources. *Proc. CISS*, 99, 1999.

[43] R. S. Ellis. *Entropy, Large Deviations, and Statistical Mechanics*, volume 271 of *Grundlehren der mathematischen Wissenschaften.* Springer-Verlag, New York, 1985.

[44] J. A. Bucklew. A large deviation theory proof of the abstract alphabet source coding theorem. *IEEE Trans. Inf. Theor.*, 34(5):1081–1083, September 2006.

[45] Y. Oono. Large deviation and statistical physics. *Prog. Theo. Phys.*, 99:165, 1989.

[46] D. Mandal and C. Jarzynski. Work and information processing in a solvable model of maxwell's demon. *Proc. Natl. Acad. Sci. USA*, 109(29):11641–11645, 2012.

[47] S. Marzen and J. P. Crutchfield. Information anatomy of stochastic equilibria. *Entropy*, 16:4713–4748, 2014. SFI Working Paper 14-04-005; arxiv.org:1403.3864 [cond-mat].

[48] S. Marzen and J. P. Crutchfield. Informational and causal architecture of discrete-time renewal processes. 2014. SFI Working Paper 14-08-032; arxiv.org:arXiv:1408.6876 [cond-mat.stat-mech].

[49] S. Marzen and J. P. Crutchfield. Informational and causal architecture of continuous-time renewal processes. 2014. SFI Working Paper 14-XX-XXXX; arxiv.org:14XX.XXXX [cond-mat].

[50] The only necessary condition is that these two matrices contain the left and right eigenvectors of $\mathbf{T}_\beta$.

[51] J. E. Hopcroft, R. Motwani, and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation.* Prentice-Hall, New York, third edition, 2006.

[52] D. Lind and B. Marcus. *An Introduction to Symbolic Dynamics and Coding.* Cambridge University Press, New York, 1995.