CHAPTER 1

Computational Mechanics

"Perplexity is the beginning of knowledge. [1]" - Khalil Gibran, trans. Ferris.

§1.1 Philosophy

The first goal of computational mechanics[2] is to provide a common language with which to describe arbitrary dynamical systems. Ptolemy conceived of the Celestial Sphere in terms of planets, epicycles, deferants and equants in a very geometric theory; Maxwell described charges and currents making use of fields and vector calculus; Modern physicists represent particles using constructs ranging from group representations to topological invariants. How can such theories be compared when their components are so diverse? Computational mechanics seeks to describe these various facets of physical phenomena so that we might begin to ask such questions as, 'Which is least predictable, the sun rising, the ticking of a clock, or the exponential decay of ^{87}Rb', and, 'Can we claim that to orchestrate the motion of the planets requires more effort than to know the state the stock market? '.

Even limiting ourselves to the ontology of partial differential equations, we find that there are only a few tools available to us for making comparisons. These include: Lyapunov exponents, fractal dimension of the attractor, etc. These tools have a long and important history that continues to impact many fields. However here we aspire to a more extensive and *principled* accounting of the system, and further, we would like for this accounting to serve as more than a method of comparing, ie. testing likeness, and serve as a satisfactory avatar of the system-in-itself. This incarnation is known as the ϵ-machine.

[1] "$[2^{-\sum_p p \log p} > 1] \Rightarrow [H > 0]$" - Khalil Gibran, trans. anonymous.

[2] For a complementary introduction to computational mechanics, there exist several excellent resources including, but not limited to Refs. **DF thesis, CRS thesis, Hanson?, Upper?**.

The second, and more operative, goal of computational mechanics is to make good predictions. Good predictions can be seen as a consequence of having described a system correctly. Stated as an independent goal, it draws attention to two things. First, prediction will be the practical measure of this work; Funding and interest will continue to fuel this research primarily as it pertains to prediction. Conversely, that prediction is viewed here as a consequence of correct description, as much as it is itself a goal, we hope will help to persuade scientists to consider a different perspective.

§1.1.1 ϵ-machine Prelude

The ϵ-machine is a construct that depends only on two simple ideas. The first is that any description of a dynamical system should be one that is constructed in the language of our interactions with the system. If the way in which we know the system is by sight, then our ϵ-machine ought to be one that is somehow composed of 'sight events'. We might instead, as we will throughout this work, assume to know of our system through digital (wlog binary) measurements. We may wish to think of these as sight events as well since we are likely observing a digital readout display. This is as deeply as we wish to discuss sensory epistemology; we simply wish to contrast our use of sensory data in the construction of an ϵ-machine with a construction that makes use of any preexisting ontology, say, marbles or matrices.

We will use matrices and tensors, but really just as an organizational tool; they will house the manifold probabilities that we must concern ourselves with. We take probability to be something outside of the set of constructive assumptions—fair game for a 'blind' theory. Of course one may argue that this is an assumption with important consequences. For instance, **KW** discusses the implications of an underlying quantum theory. There are surely many interesting things to be said on this topic, but at the risk of becoming overly entangled, we will assume that whatever quantum mechanical operations are at work, are so in a way that is only classically correlated with the observer's data file on their hard drive.

The second idea is an interesting nugget about information—part common sense, part tautology, part kōan. To paraphrase Bateson, *information* is $\{\Delta : \Delta \Rightarrow \Delta'\}$, or 'the differences that make a difference'. What we take from this lesson is that, just as in communication theory, infor-

mation is about deviation from expectation. **say this better** We formalize these concepts in the next section.

§1.2 Our Domain: Processes

The language in the preceding section is purposefully somewhat vague. The idea being that these principles of dynamical system description might be quite broadly applied. Indeed they have been - to: communication channels, cellular automata **CRS, JPC**, spin systems **DF**, continuous space dynamical systems, continuous time dynamical systems **KW**, quantum dynamical systems **KW**, and dynamical networks **Olaf, other?**.

In this work, we focus on discrete-time, finite-alphabet stationary stochastic processes. These will be referred to as just *[stochastic] processes* or *process languages*. Although the results contained are described and proven in this context, we maintain that the spirit of what we do ought to survive translation into many of the other contexts described above, most likely with some degree of reinterpretation or generalization. We feel that the value of this work is as much in the solving of problems for processes as it is in the generic concepts defined and explored.

While generically the event space for each random variable in a set may be different, we will study those sets of random variables for which the event space is identical for each random variable. For this reason, we allow the following definition.

Definition. *An* alphabet, $\mathcal{A}$, *is a set of (possibly continuum) events appropriate to each of a set of random variables.*

Definition. *A* word *is a concatenation of symbols from the alphabet.*

$$w = x_0 x_1 \dots x_k \quad , \quad x_i \in \mathcal{A}$$

We will focus on alphabets that are both discrete and finite. These definitions are as you would expect. In fact, for many purposes it will be sufficient to consider only the alphabet, $\mathcal{A} = \{0, 1\}$.

Definition. *A* discrete-time, finite-alphabet stationary stochastic process *, or just* process, $\mathcal{P}$, *is a bi-infinite string of random variables,*

$$\dots X_{-3}, X_{-2}, X_{-1}, X_0, X_1, X_2, X_3, \dots$$

where the random variables have a finite alphabet, $\mathcal{A}$, and,

$$\Pr(X_t, X_{t+1}, \ldots, X_{t+j}) = \Pr(X_{t+k}, X_{t+1+k}, \ldots, X_{t+j+k})$$

for any $t, j, k \in \mathbb{Z}$.

Since the interpretation of the 'time' index in one of these processes is often *time*, and we, being stuck somewhere in the middle of time, have a notion of *past* and *future*, we will use these words to conveniently describe particular sets of random variables. Since we are interested in stationary processes, we may choose to insert ourselves at $t = 0$ and declare one side the past, and the other the future [3].

Definition. *A* [finite] future, *denoted X_0^k, refers to the k random variables $X_0, X_1, \ldots, X_{k-1}$.*

The short-hand for the finite past is similarly defined.

Definition. *A* [finite] past, *denoted X_{-k}^k, refers to the k random variables $X_{-k}, X_{-k+1}, \ldots, X_{-1}$.*

Naturally, we are interested in the infinite limits of finite futures and pasts.

Definition. *An* infinite future, *or* future, *denoted $\overrightarrow{X}_0$, refers to the infinite set of random variables $X_0, X_1, \ldots$.*

Definition. *An* infinite past, *or* past, *denoted $\overleftarrow{X}_{-1}$, refers to the infinite set of random variables $\ldots, X_{-2}, X_{-1}$.*

When we wish to discuss an instance of a random variable, or future, past, finite or infinite, we will use the lowercase, $x_t, x_t^{t+k}, \overrightarrow{x}_t, \overleftarrow{x}_t$.

Note that there are slight asymmetries in the notation. This is just because when states are introduced'zero' is forced to choose a side. We choose zero to fall on the side of the future because of programmers' predjudice. The more complete view of the indexing scheme will be seen when states are involved (see Fig. 1.6).

To make some use of our shorthand, we can now compactly refer to a process and its variables by the appealing form, $\mathcal{P} = \Pr(\overleftarrow{X}, \overrightarrow{X})$.

[3] Note that the 'time' index may also be used to describe a spatial dimension. Occasionally it seems useful to play with the interpretation of this index as it provides useful alternate perspectives.

§1.3 ϵ-machines

Now that we have laid out the goals and principles of the construction as well as the domain to which it will be applied, we describe the ϵ-machine itself.

Recalling the two simple ideas underlying the ϵ-machine, we see that the first is satisfied by assuming that the process at hand was obtained through digital measurement of some dynamical system. The second is satisfied by utilizing the appropriate equivalence relation.

Definition. *Given a process, $\mathcal{P}$, define an equivalence relation $\sim_\epsilon$ where*

$$\overleftarrow{x} \sim_\epsilon \overleftarrow{x}' \Leftrightarrow \Pr(\overrightarrow{X} \,|\, \overleftarrow{X} = \overleftarrow{x}) = \Pr(\overrightarrow{X} \,|\, \overleftarrow{X} = \overleftarrow{x}')$$

That this relation is reflexive, symmetric and transitive follows straightforwardly from the dependence on the inner equivalence relation, $=$. Throughout this work, we will only use the subscripted symbol, $\sim_\epsilon$, when contrasting with another relation.

Having defined an equivalence relation, several things are immediately forthcoming. First, the equivalence relation defines equivalence classes.

$$[\overleftarrow{x}] = \{\overleftarrow{x}' : \overleftarrow{x}' \sim \overleftarrow{x}\}$$

The set of these equivalence classes is a quotient set.

$$\{\overleftarrow{x}\}/\sim = \{[\overleftarrow{x}'] : \overleftarrow{x}' \in \{\overleftarrow{x}\}\}$$

The relation also induces a surjective map from the original set to the equivalence classes.

$$\epsilon : \{\overleftarrow{x}\} \to \{\overleftarrow{x}\}/\sim \quad , \quad \epsilon(\overleftarrow{x}) = [\overleftarrow{x}]$$

As these equivalence classes are the building blocks of the ϵ-machine, and are at the core of nearly all calculations, we allow ourselves to dub these particular classes, *causal states*. This definition is intended to be suggestive of the ϵ-machine as a probabilistic automaton, and also for notational convenience.

Definition. *The set of* causal states, $\{\mathcal{S}\} = \boldsymbol{\mathcal{S}}$, *is in one-to-one correspondence with the set of equivalence classes,* $\{\overleftarrow{x}\}/\sim$.

The set of causal states [4] can be discrete, fractal, or continuous; see, e.g., Figs. 7, 8, 10, and 17 in Ref. [**?**].

[4] A process's causal states consist of both transient and recurrent states. To simplify the presentation, we henceforth refer *only* to recurrent causal states that are discrete.

Definition. *We say that a causal state, $\mathcal{S}$, is* induced *by a past, $\overleftarrow{x}'$, if $\mathcal{S}$ corresponds to $[\overleftarrow{x}]$ where $\overleftarrow{x}' \in [\overleftarrow{x}]$.*

Let us note that it makes sense to think of causal states are being induced, not only by particular pasts, but also by equivalence classes of pasts.

If two pasts, $\overleftarrow{x}$ and $\overleftarrow{x}'$, are members of the same equivalence class, $[\overleftarrow{x}]$, then by definition, $\Pr(\overrightarrow{X}|\overleftarrow{X} = \overleftarrow{x}) = \Pr(\overrightarrow{X}|\overleftarrow{X} = \overleftarrow{x}')$. We might wish to make a less particular statement such as, $\Pr(\overrightarrow{X}|\overleftarrow{X} = \overleftarrow{x}) = \Pr(\overrightarrow{X}|\overleftarrow{X} = \overleftarrow{x}') \simeq \Pr(\overrightarrow{X}|\overleftarrow{X} = [\overleftarrow{x}])$. This is intuitively correct, but slightly awkward since the random variable $\overleftarrow{X}$ does not have events in the space of equivalence classes. Instead we say $\Pr(\overrightarrow{X}|\overleftarrow{X} = \overleftarrow{x}) = \Pr(\overrightarrow{X}|\overleftarrow{X} = \overleftarrow{x}') = \Pr(\overrightarrow{X}|\mathcal{S} = \sigma)$ where σ is the particular causal state induced by any member of the class $[\overleftarrow{x}]$. This leads to the primary utility of causal states, which is as a short-hand, or stand-in for particular pasts.

So far the causal states have been constructed as sufficient replacement variables for the infinite past. Given that the set of pasts is uncountably infinite and the set of causal states is, for the present, finite, this is a tremendous compactification of knowledge. However, it is not yet the useful tool we desire. What we have provided so far is a 'routing' variable [5]. The real power of the causal state will be its dynamic function in the ϵ-machine—doling out the appropriate future bit by bit. In this way, the causal state is not only a short-hand for the infinite past, but it also shields us from the infinite variety of the future, allowing us to ratchet forward in time one symbol at a time.

In addition to states, we evidently need some notion of dynamic, as it is a dynamical system we are representing. We capture this dynamic with a set of [symbol-] labeled transition matrices. Specifically, the value of an element of the "x'th" matrix is defined,

$$T^{x}_{\sigma,\sigma'} = \Pr(\mathcal{S}_{t+1} = \sigma', X_{t+1} = x|\mathcal{S}_t = \sigma)$$

This is the conditional probability that, given a particular causal state, σ, or equivalently a past that induces σ, the following measurement symbol will be x; this will consequently induce causal state σ'. Since the process is stationary, the value of the variable t is unimportant.

Causal states have a Markovian property that they render the past and future statistically

[5] Imagine a www tool that accepts an infinite past and then provides you with a URL. This URL then leads to a page with an infinite set of infinite futures. This *is* in some sense what we want, but we'd rather not drown in the data just yet. We would hope to control the data flow.

independent; they *shield* the future from the past [**?**]:

$$\Pr(\overleftarrow{X}, \overrightarrow{X} | \mathcal{S}) = \Pr(\overleftarrow{X} | \mathcal{S}) \Pr(\overrightarrow{X} | \mathcal{S}) . \tag{1.1}$$

Moreover, they are optimally predictive [**?**] in the sense that knowing which causal state a process is in is just as good as having the entire past: $\Pr(\overrightarrow{X} | \mathcal{S}) = \Pr(\overrightarrow{X} | \overleftarrow{X})$. In other words, causal shielding is equivalent to the fact [**?**] that the causal states capture all of the information shared between past and future: $I[\mathcal{S}; \overrightarrow{X}] = \mathbf{E}$.

Causal states have a Markovian property that they render the past and future statistically independent; they *shield* the future from the past [**?**]:

$$\Pr(\overleftarrow{X}, \overrightarrow{X} | \mathcal{S}) = \Pr(\overleftarrow{X} | \mathcal{S}) \Pr(\overrightarrow{X} | \mathcal{S}) . \tag{1.2}$$

Moreover, they are optimally predictive [**?**] in the sense that knowing which causal state a process is in is just as good as having the entire past: $\Pr(\overrightarrow{X} | \mathcal{S}) = \Pr(\overrightarrow{X} | \overleftarrow{X})$. In other words, causal shielding is equivalent to the fact [**?**] that the causal states capture all of the information shared between past and future: $I[\mathcal{S}; \overrightarrow{X}] = \mathbf{E}$.

ϵ-Machines have an important structural property called *unifilarity* [**?**, **?**]: From the start state, each symbol sequence corresponds to exactly one sequence of causal states [6]. ϵ-Machine unifiliarity underlies many of the results here. Its importance is reflected in the fact that representations without unifilarity, such as general hidden Markov models, *cannot* be used to directly calculate important system properties—including the most basic, such as, how random a process is. As a practical result, unifilarity is easy to verify: For each state, each measurement symbol appears on at most one outgoing transition [7]. Thus, the signature of unifilarity is that on knowing the current state and measurement, the uncertainty in the next state vanishes: $H[\mathcal{S}_{t+1} | \mathcal{S}_t, X_t] = 0$. In summary, a process's ϵ-machine is its unique, minimal unifilar model.

To summarize, a causal state is a set of pasts that each have the same correlation with, or prediction for, the future, $\Pr(\overrightarrow{X} | \overleftarrow{x})$. The ϵ-machine is obtained by linking neighboring causal states together with the appropriate interstitial observed symbol. As promised, its component states are equivalent to sets of past observations. The probabilistic dynamic induced on the causal states is exactly the one which the data demands as sequential observations induce sequential

[6] Following terminology in computation theory this is referred to as *determinism* [**?**]. However, to reduce confusion, here we adopt the practice in information theory to call it the *unifilarity* of a process's representation [**?**].

[7] Specifically, each transition matrix $T^{(x)}$ has, at most, one nonzero component in each row.

causal states.

§1.4 First Examples

All of this may appear a bit more abstract than necessary, and in some sense this is true. The example ϵ-machines that appear throughout this work are straightforward to draw with a pen and paper in only a minute or two. The beauty is that even these elementary examples will provide us with a means for motivating and uncovering plenty of interesting science.

§1.4.1 IID

We would be remiss if we failed to begin with the an independent, identically distributed (IID) process. By definition, each measurement is independent of the past. Therefore, there is only one conditional distribution: $\Pr(0|\cdot) = p, \Pr(1|\cdot) = 1 - p$. In turn, there is only one equivalence class of histories, and thus only one causal state 1.1.

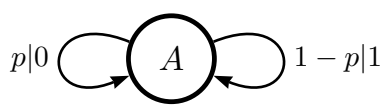

Figure 1.1: IID binary process.

The labeled transition matrices are one-dimensional,

$$T^0_{A,A} = \begin{bmatrix} p \end{bmatrix}, T^1_{A,A} = \begin{bmatrix} 1-p \end{bmatrix}$$

This is a rather trivial, but important process to keep in mind because many kinds of data are taken without consideration for temporal correlation. We will see in how many *independent* ways this process class actually is trivial. Let us use this to motivate our discussion of measures of processes.

§1.4.2 Entropy and Entropy Rate

The first aspect of a process we would like to characterize is its variability; more precisely, we would like to know the degree of uncertainty in particular groups of random variables. Following Shannon **cite shannon**, we characterize the uncertainty by the Shannon information.

Definition. *The* Shannon information, *or here the* entropy [8], *of a random variable is given by,*

$$H[X] = -\sum_{x \in X} \Pr(X = x) \log \Pr(X = x)$$

Analogously, the entropy of a set of random variables, or their joint probability distribution, is defined,

$$H[X, Y] = -\sum_{x \in X, y \in Y} \Pr(X = x, Y = y) \log \Pr(X = x, Y = y)$$

In the trivial case of our first example (Fig. 1.1), the uncertainty in the bi-infinite string of random variables is infinite. Each random variable has some finite uncertainty associated with it, and by definition of being IID, the uncertainty in any given variable is independent of any other variable. Therefore, the uncertainty in the entire string is infinite. Since a categorically infinite entropy leaves little to discuss [9], we will be primarily interested in the functional relation between entropy and length scale as the length grows. Furthermore, we are interested in not only the asymptotic behavior (some kind of exponential envelope), but also the finite length behavior. In fact we are more interested in the finite length behavior as long as we are guaranteed *some* kind of convergence. The parent entropy function that many important process features derive from is the block entropy.

Definition. *The* block entropy function *(or curve [10]) is is the entropy of the distribution of words at length L.*

$$H[X_0^L] = -\sum_{x_0^L \in \mathcal{A}^L} \Pr(X_0^L = x_0^L) \log \Pr(X_0^L = x_0^L)$$

properties: nondecreasing, concave

Definition. *The* entropy rate, h_μ, *of a process is the limit of the conditional entropy,*

$$h_\mu = \lim_{L \to \infty} H[X_0 | X_{-L}, X_{-L+1}, \ldots, X_{-1}]$$

[8] As in most information theory, computer science and symbolic dynamics, the log function in this work will always mean base two.

[9] Actually there is much to discuss. For continuous time or continous valued output, measures such as the differential entropy will generically be infinite. We hope that a talented mathematician will generalize the ideas here to the continuous case.

[10] This is a discrete function and so we use the word curve to be suggestive of the fact that these functions are highly restricted by monotonicity and such and so are, as far as discrete functions go, relatively curve-like.

In our shorthand, $h_\mu = H[X_0|\overleftarrow{X}]$. *We are also interested in finite length approximations to the entropy rate,*

$$\begin{aligned} h_\mu(L) &= H[X_0|X_{-L}, X_{-L+1}, \ldots, X_{-1}] \\ &= H[X_{-L}, X_{-L+1}, \ldots, X_{-1}, X_0] - H[X_{-L}, X_{-L+1}, \ldots, X_{-1}] \end{aligned}$$

properties: nonincreasing, convex The entropy rate estimate at $L = 0$ is defined to be $h_\mu(0) = \log(|\mathcal{A}|)$. This is just stating that if you know nothing about the process other than the size of the alphabet, your uncertainty is maximal,that is, the entropy of a uniform distribution over the alphabet.

$$\begin{aligned} H[\frac{1}{|\mathcal{A}|} \times (1, 1, \ldots, 1)] &= -\sum_{i=1}^{N} \frac{1}{|\mathcal{A}|} \log \frac{1}{|\mathcal{A}|} \\ &= \log |\mathcal{A}| \end{aligned}$$

If we think of this geometrically, the entropy rate is the limit of the discrete slope of the block entropy function. Simple geometric features of these key functions will play a primary role in both motivating classifications of processes and also in understanding features otherwise defined. We can write the probability distribution for an IID process as a product of distributions

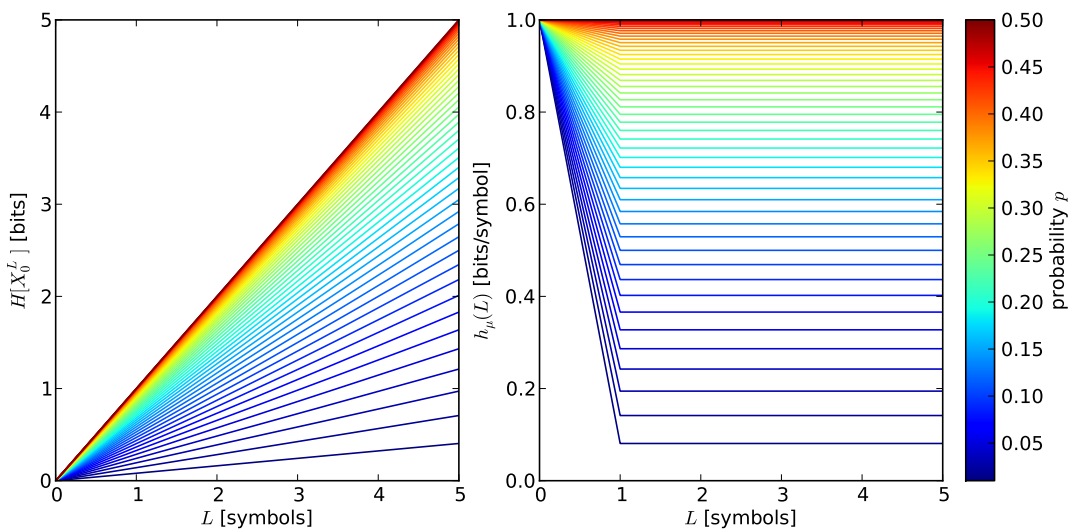

Figure 1.2: (Left) Block entropy curve for all binary IID processes ($\Pr(X = 0) = p$). Each is simply a line through the origin with slope determined by p. Since there is a symmetry in the class about $p = 0.5$, only on half of the p values are illustrated. (Right) Finite length entropy rate approximations. Colorbar indicates the probability, p, that is varied to obtain different members of a process family. The particular probability, p, refers to the variable in Fig. 1.1.

for each random variable.

$$\Pr(\overleftarrow{X}, \overrightarrow{X}) = \ldots \times \Pr(X_{-1}) \times \Pr(X_0) \times \Pr(X_1) \ldots$$

As a consequence, the block entropy is linear in the length, L.

$$\begin{aligned}
H[X_0^L] &= -\sum_{x_0^L \in \mathcal{A}^L} \Pr(X_0^L = x_0^L) \log \Pr(X_0^L = x_0^L) \\
&= -\sum_{x_0, x_1, \ldots, x_{L-1} \in \mathcal{A}} \Pr(X_0 = x_0) \times \ldots \times \Pr(X_{L-1} = x_{L-1}) \log \Pr(X_0 = x_0) \times \ldots \times \Pr(X_{L-1} = x_{L-1}) \\
&= -\sum_{x_0 \in \mathcal{A}} \Pr(X_0 = x_0 \log \Pr(X_0 = x_0)) - \ldots - \sum_{x_{L-1} \in \mathcal{A}} \Pr(X_{L-1} = x_{L-1} \log \Pr(X_{L-1} = x_{L-1})) \\
&= -L \times \sum_{x_0 \in \mathcal{A}} \Pr(X_0 = x_0) \log \Pr(X_0 = x_0) \\
&= -LH[X_0]
\end{aligned}$$

An immediate consequence of this is that the entropy rate is equal to the length-one approximation.

$$\begin{aligned}
h_\mu &= \lim_{L \to \infty} H[X_0^L] - H[X_0^{L-1}] \\
&= \lim_{L \to \infty} LH[X_0] - (L-1)H[X_0] \\
&= H[X_0]
\end{aligned}$$

That is, beyond considering the most trival statistic, there is nothing more to learn about this system. Since this entropy, or entropy rate, will arise rather frequently, it is often referred to simply as the *binary entropy* and denoted $H(p)$.

A more interesting process is one for which the finite length approximations to h_μ are something other than constant. A very simple example that illustrates this [11] is the Golden Mean Process (Fig. 1.3). Notice that in Fig. 1.4, the entropy rate estimates reach the entropy rate at $L = 2$.

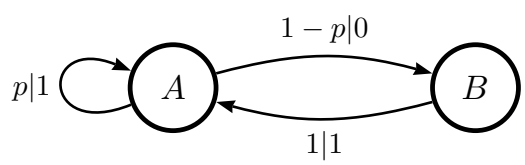

Figure 1.3: The Golden Mean Process is one that disallows consecutive zeros. After a zero [one] is seen, the process is in causal state B [A].

To restate, for the Golden Mean process, the conditional uncertainty in every symbol, including

[11] In fact, *any* process other than IID will illustrate this. I have just chosen a simple one that is convenient because it is simple in other ways.

and beyond the second, is h_μ.

$$H[X_0|X_{-k},\ldots,X_{-1}] = H[X_0|X_{-1}] = h_\mu$$

This leads one to naturally speculate that there is a conditional *probabilistic* independence as well as this conditional *entropic* independence. This is true, but because it is an intuitive result with a somewhat inelegant proof, it can be found in App. **??** We should be sure to point out

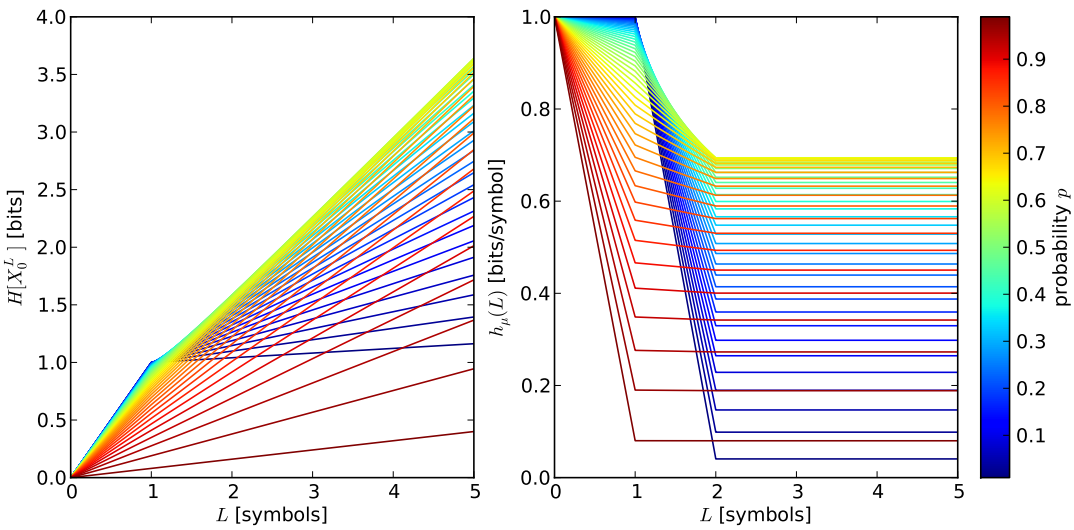

Figure 1.4: (Left) Block entropy curves for all Golden Mean processes. Each is linear for $L \geq 1$. (Right) Each finite length entropy rate estimate reaches its asymptotic value h_μ, at $L = 2$. This indicates that the additional uncertainty in the $L = 2$ blocks, beyond the $L = 1$ blocks, is already h_μ. This implies that the minimum correlation length required for maximal prediction ability is $L = 1$. That is, the Golden Mean is an order-1 Markov process.

an important feature of the ϵ-machine. In contrast with generic hidden Markov models, the ϵ-machine has the property that the entropy rate can be calculated directly from it. We argue that this is a consequence of the ϵ-machine being the natural representation of the process.

Now that the entropy rate has been defined as a limit and is certainly straightforward enough to estimate, we should look for a closed form. We might imagine that, for any hidden Markov model (composed of states $\mathcal{R}$), that the time average surprise is the same as the state average surprise (when weighted by state visitation probabilities). Specifically, that

$$h_\mu \stackrel{?}{=} \sum_\rho \Pr(\mathcal{R}_0 = \rho) H[X_0|\mathcal{R}_0 = \rho])$$

It is easy to see that this is not true. Consider a nonunifilar presentation of the Golden Mean Process. The conditional entropies are $H[X_0|\mathcal{R}_0 = A] = 0$ and $H[X_0|\mathcal{R}_0 = B] = 0$. No weighted

sum of these conditional entropies will yield what we know to be a non-zero entropy rate. The

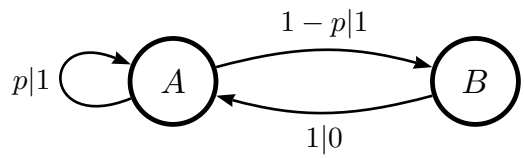

Figure 1.5: Nonunifilar presentation of the Golden Mean Process. The entropy rate of the process is *not* simply the weighted average of the entropies of symbols emitted after visiting each state.

correct form for the entropy rate is in fact this weighted sum of individual state entropies, but *only* when the states in question are *causal* states. So for ϵ-machines, we have the following closed form for the entropy rate in terms of causal state asymptotic probabilities and transition probabilities.

$$\begin{aligned} h_\mu &= H[X_0|\mathcal{S}_0] \\ &= \sum_{\sigma\in\mathcal{A}} \Pr(\mathcal{S}_0=\sigma) H[X_0|\mathcal{S}_0=\sigma]) \\ &= -\sum_{\sigma\in\boldsymbol{\mathcal{S}}} \Pr(\mathcal{S}) \sum_{x\in\mathcal{A}\sigma'\in\boldsymbol{\mathcal{S}}} T^{(x)}_{\mathcal{S}\mathcal{S}'} \log_2 \sum_{\sigma'\in\boldsymbol{\mathcal{S}}} T^{(x)}_{\mathcal{S}\mathcal{S}'} \end{aligned}$$

§1.5 Statistical Complexity

Above, $\Pr(\mathcal{S})$ is the asymptotic probability of the causal states, which is obtained as the normalized principal eigenvector of the transition matrix $T = \sum_{\{x\}} T^{(x)}$. We will use π to denote the distribution over the causal states as a row vector.[12] This distribution over states leads to a second fundamental characterization of processes—the statistical complexity.

Definition. *A process's* statistical complexity, C_μ, *can be directly calculated from its ϵ-machine as it is a property of the dynamic over the causal states:*

$$\begin{aligned} C_\mu &= H[\mathcal{S}] \\ &= -\sum_{\{\mathcal{S}\}} \Pr(\mathcal{S}) \log_2 \Pr(\mathcal{S}) \,. \end{aligned} \tag{1.3}$$

The statistical complexity is a *statistical* complexity as opposed to a deterministic one, such as Kolmogorov complexity,[13] meaning that the measure is intended to capture the complexity

[12] The matrix algebra here follows the 'state on the left, transition matrix on the right' convention.

[13] Kolmogorov complexity is also known as: descriptive complexity, Kolmogorov-Chaitin complexity, stochastic complexity, algorithmic complexity, algorithmic entropy, and program-size complexity.

of a class of data rather than a particular instance. To illustrate, a coin may be flipped to generate a variety of sequences. One such sequence is an alternating sequence of heads and tails, $(HT)^N$. This sequence may be generated by the compact[14] program: i=0,for(i<N){write H, write T, i=i+1}. The increasing majority of sequences will not have this compactness, yet all are possible realizations of the output of this simple dynamical system. The goal of a statistical complexity is to provide a characterization of all of these possiblilities.

Why should we aim to describe the broad class behavior rather than the detailed behavior indicated by a particular data string? If our aim is to describe the physical dynamical system, and if we believe that this system has inherent unpredictability, then describing a particular instance, as the Kolmogorov complexity would do, might actually *overspecify* the physical system. For instance, a flipped coin could certainly produce a binary representation of π; another might code for the name of the next president. Neither one of these things captures the essence of the physical system—that it is IID and uniformly[15] random heads and tails. If we insist on characterizing the coin's behavior by the two instances above, then I argue that we ought to consider *all* possible instances. This is clearly not a productive use of time.

What exactly is the statistical complexity telling us? As C_μ is defined as the entropy of a probability distribution over some event space, it can immediately be understood in the context of communication theory. If Alice wishes Bob to synchronize his ensemble of identical dynamical systems to hers, she must communicate C_μ bits per member of the ensemble.[16] If it is C_μ bits that is passed to Bob to describe the state of the system, then it could be said that each system *carries* that amount of information. This is why the statistical complexity is interpreted as *stored information*.

Why is this a good measure of complexity? As we live in a world where complexity measure abound, it is important to pause and reflect on the particular contribution of a particular measure. We first claim that the above description of C_μ as stored information is strong evidence for its naturalness. Additionally, it has some properties that although somewhat trivial, are not shared by all. In thinking about the range of possible processes, it is hard to argue that IID pro-

[14]We call this compact because as the size of the program goes as $\log(N) + C$. Thus the limit of the ratio of output sequence size to program size is zero. This indicates that this subprocess has and entropy rate of zero.

[15]Some might argue that we have to toss the coin with more vigor (see **diaconis coin**).

[16]Of course she must communicate C_μ bits *on average*, but that is the standard assumption made in information theory. In fact, without it, information does not have the same meaning.

cesses are not on *some* particular extreme. Correspondingly, the statistical complexity of any IID process is zero. This seems to satisfy our intuition about what a complexity measure ought to say about IID processes. In another corner of process space lie completely predictable. These processes, certainly for finite cases, are just the periodic ones. The statistical complexity of these will be the log of the period length. Another good reason is that C_μ has a kind of extensivity. If two uncorrelated processes are 'placed side-by-side', which is the usual thing to do when testing for extensivity, the joint process characterized by the process language over the appropriate tuples of symbols has a statistical complexity which is simply the sum of the individual complexities.[17]

Thus, the ϵ-machine directly gives two important properties: a process's rate (h_μ) of producing information and the amount (C_μ) of historical information it stores in doing so.

§1.6 Excess Entropy

The entropy rate is a property that comes straight out of communication theory. In that context, it is the minimum capacity of an error-free channel; equivalantly, it is the amount of supplementary information required for maintaining perfect decoding (which we think of as prediction). Another concept that arises very naturally in the communication context is the transference of information from input to ouput. Different channels, depending on their capacity, noise present, etc., will have varying abilities to transfer information from one side of the channel to the other. We can cast a dynamical system or time-series as a channel in the following way; The past is considered the input, the future is the output, and the channel itself is the ϵ-machine. Given this picture, the excess entropy is the amount of information about the past that is transmitted via the ϵ-machine channel to the future (See Fig. 1.6). We express this mathematically in terms of a mutual information.

$$I[\overleftarrow{X}; \overrightarrow{X}] = \mathbf{E}$$

Excess entropy has gone by several different names and has been reinvented several times **cite early JPC, Grassberger, etc**. There are several equivalent forms for **E**, see Ref. [**?**, and references

[17] I might argue that this way of thinking about extensivity is a little mundane. The side-by-side test for extensivity is born of thinking about equilibrium systems. As ϵ-machines are definitely not equilibrium systems, it would be most interesting to test for extensivity in different ways. One might sample from the two systems in an alternating manner. We can imagine something more drastic, and maybe harder to motivate, like the graph-join of the two ϵ-machines. Presently, only little is known about about the consequences of these types of actions.

therein]. Here we quote the definition of excess entropy from Ref. **RURO**, where the name is somewhat more intuitive.

Definition. *The* excess entropy *is the sum over word lengths of the degree to which the entropy rate estimate is in excess of the true entropy rate.*

$$\mathbf{E} = \lim_{L'\to\infty} \sum_{L=1}^{L'} (h_\mu(L) - h_\mu)$$

Excess entropy, and related mutual information quantities, are widely used diagnostics for complex systems. They have been applied to detect the presence of organization in dynamical systems [**?**, **?**, **?**, **?**], in spin systems [**?**, **?**, **?**], in neurobiological systems [**?**, **?**], and even in language, to mention only a few applications. For example, in natural language the excess entropy ($\mathbf{E}$) diverges with the number of characters L as $\mathbf{E} \propto L^{1/2}$. The claim is that this reflects the long-range and strongly non-ergodic organization necessary for human communication [**?**, **?**].

It can be demonstrated that this definition is, at least for the types of processes studied in this thesis,[18] equivalent to the mutual information concept.

$$\begin{aligned}
I[\overleftarrow{X}; \overrightarrow{X}] &= \lim_{L\to\infty} I[X_{-L}^L; X_0^L] \\
&= \lim_{L\to\infty} H[X_{-L}^L] + H[X_0^L] - H[X_{-L}^L, X_0^L] \\
&= \lim_{L\to\infty} H[X_0^L] + H[X_0^L] - H[X_0^{2L}] \\
&= \lim_{L\to\infty} 2H[X_0^L] - H[X_0^{2L}] \\
&= \lim_{L\to\infty} 2(H[X_0^L] - Lh_\mu) - H[X_0^{2L}] + 2Lh_\mu \\
&= \lim_{L\to\infty} 2\sum_{L'=1}^{L} (H[X_0^{L'}] - H[X_0^{L'-1}] - h_\mu) - \sum_{L'=1}^{2L} (H[X_0^{L'}] - H[X_0^{L'-1}] + h_\mu) \\
&= \lim_{L\to\infty} 2\sum_{L'=1}^{L} (h_\mu(L') - h_\mu) - \sum_{L'=1}^{2L} (h_\mu(L') + h_\mu) \\
&= 2\lim_{L\to\infty} \sum_{L'=1}^{L} (h_\mu(L') - h_\mu) - \lim_{L\to\infty} \sum_{L'=1}^{2L} (h_\mu(L') + h_\mu) \\
&= 2\mathbf{E} - \mathbf{E} = \mathbf{E}
\end{aligned}$$

The second line follows from the definition of mutual information. The third line is a result of stationarity. We insert some copies of h_μ and rearrange to form the definition of $\mathbf{E}$. To split the limit into two, we assume the existence of the individual limits. This really amounts to assuming

[18] Extensions as benign as the addition of an extra dimension—to a 2D process—necessitate more care with these equivalences **cite feldman 2D**.

the existence of one limit—**E**. In recent work, see Ref. **nick**, it is shown that this limit exists for all ϵ-machines with a finite number of states.

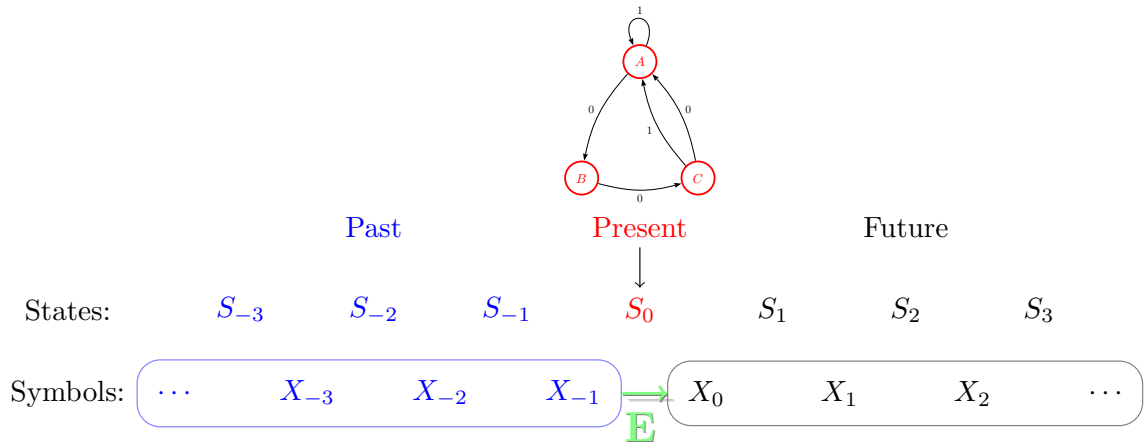

Figure 1.6: A stochastic process can be viewed as a communication channel. The data in the past is the input to the channel. The channel itself is the dynamical system, or ϵ-machine, which transmits information to the future. The total information transmitted from past to future is equal to the excess entropy.

We can begin to collect our understanding of the information theoretic relationships among ϵ-machine variables using an I-diagram (see Fig. 1.7). For a review of I-diagrams, see App. **??**. The other two quantities shown in this diagram are $H[\overleftarrow{X} \mid \overrightarrow{X}]$ and $H[\overrightarrow{X} \mid \overleftarrow{X}]$. Since these will generally be infinite quantities, it can be useful to think of the random variables in their finite forms, X_{-k}^{k} and X_0^k. The rate of growth of these agglomerated variables (with L) is bounded above by $H[X_0]$ and below by h_μ.

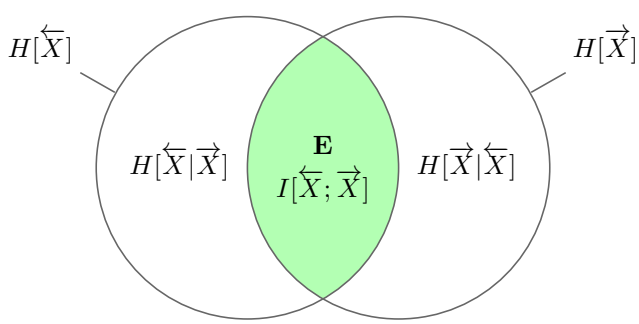

Figure 1.7: This I-diagram highlights the role of excess entropy as the mutual information between past and future data.

To form a complete I-diagram for an ϵ-machine, we must introduce a state variable. Of course this diagram is not a complete description, but it does aid our thinking in several ways.

We start by adding a generic state, actually any random variable at all will do. In Fig. 1.8, all possible information relations among the variables are listed.

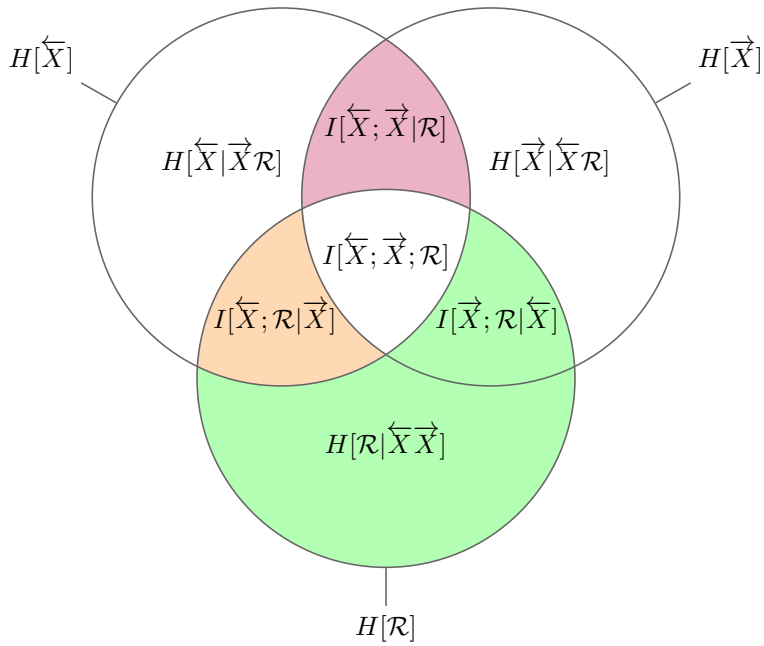

Figure 1.8: The generic relation among the past, future and a state, $\mathcal{R}$ includes 15 nontrivial information quantities. Demanding that the state involved is a causal state effects 4 of these quantities. The green area (which corresponds to two information atoms) is zero because the causal state is a single-valued function of the past. The purple area is zero because causal states are prescient. The orange area is not zero, but is the minimum value possible, given that green and purple are zero.

Note that some of the region are colored. This is to indicate that there is a difference between a generic state variable and a *causal* state in so far as these areas [19] are concerned. Let us explore these individually, substituting a causal state, $\mathcal{S}$ for the generic state, $\mathcal{R}$.

The green area is zero because the causal state is a single-valued function of the infinite past, $\overleftarrow{x}$. Since $H[\mathcal{S}|\overleftarrow{X}\,\overrightarrow{X}]$ is also a conditional entropy, and therefore positive, we have that the green subregions are individually zero.

The purple region is zero because the probability distribution over futures given a past is the same as that given the induced causal state. This implies that a past and the causal state it induces share the same amount of information with the future. Since we already have that

[19] The words 'area', 'region', '[information] quantity' and '[information] atom' are used interchangeably here in light of the correspondence between Venn diagrams and information theory. For more about this relationship refer to App. **??**.

$I[\overrightarrow{X};\mathcal{S}|\overleftarrow{X}] = 0$, this shared information must be the *same* information. Recalling that $I[\overleftarrow{X};\overrightarrow{X}] = \mathbf{E}$, we then arrive at $I[\mathcal{S};\overrightarrow{X}] = \mathbf{E}$.

The orange region is generically not zero, although the ϵ-machine ensures that it is, given that the previously describe regions are zero, the smallest possible value. It is this orange region that will be the subject of much discussion later on. It is a quantity governed by opposing forces; on the one hand, it must be large enough to accomodate capturing all of the information relevant to the future ($\mathbf{E}$), while on the other hand it is asked to be as small as possible, giving the minimal unifilar optimally predictive representation. This quantity is called the crypticity.

§1.7 Estimation of Excess Entropy

The difficulty in obtaining accurate estimates of the excess entropy in even relatively benign systems was the primary (initial) impetus behind our effort to reframe this problem. This section is not intended to provide a comprehensive accounting of the various ways in which estimation can be difficult, nor will it quantify exactly how difficult the estimation is. We will see through a simple example that it is indeed difficult, and argue that this is generic enough to warrant searching for an alternate method. The method having been discovered and detailed in Ref. **PRATISP** obviates the need to revisit and detail the previous study of difficulty in estimation. **a little boring**

§1.7.1 Example of Sharp Convergence : Order-3 Markov

In order to set the stage for the difficult estimation task in the next section, we interrupt to offer an apparently substantial process to contend with (see Fig. 1.9). This process, being an order-3 Markov process, is a fair test case for estimation algorithms as finite order Markov models are used in a wide variety of settings **cite some Markov modeling refs**.

Calculating the standard excess entropy estimates, we see in Fig. 1.10 that for all instances of the class, there is a sharp convergence at $L = 3$. This is a consequence, and additionally an indicator, of the process being order-3 Markov. In fact, excess entropy can be calculated for finite order Markov processes in a finite way. Since we have that the entropy rate estimate becomes exact at $L = R + 1$ for an order-R markov process. This has the effect of truncating the infinite

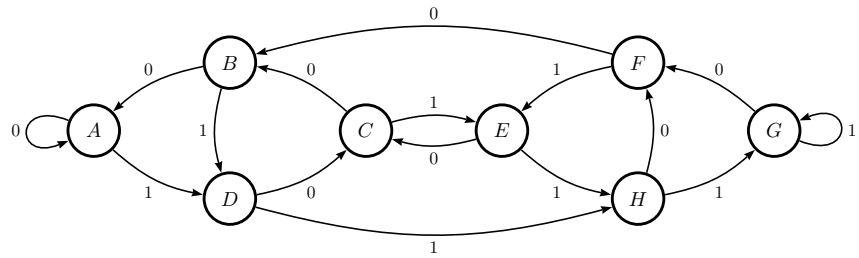

Figure 1.9: A process with 8 causal states. Since each state has two outgoing transitions, each of which has one free parameter, we suppress the probabilities here. The reader may verify that this is the structure of an order-3 Markov process—any 3 symbols will uniquely define a state (the converse happens to also be true in this instance).

sum,

$$\begin{aligned}\mathbf{E} &= \sum_{L=1}^{\infty}(h_\mu(L) - h_\mu) \\ &= \sum_{L=1}^{R}(h_\mu(L) - h_\mu) \\ &= H[X_0^R] - Rh_\mu.\end{aligned}$$

The last step is accomplished by collapsing the telescoping sum of entropy rate estimates. It appears that for finite order Markov processes, certainly for small orders, the excess entropy is easily calculable. We should now ask the questions: "What happens as the Markov order becomes large?", and "What happens when the process is not finite order Markov?"

To answer the first question, we have to calculate the probabilities of roughly $|\mathcal{A}|^R$ different length-R words. The number of words will be smaller than this depending on the process's forbidden words. For large alphabets and large orders, this can quickly become a challenging task. For instance, if we treat the English language as a Markov process over letters $\{a, \ldots, z\}$and allow for a very modest correlation length of 6, we find that there is not even enough space on a modern computer to store the resulting probability distribution.

To illustrate the reponse to the second question, let us investigate a very simple non-Markov process.

§1.7.2 Example of Slow Convergence : Even Process

The example we use to illustrate this complication is the Even Process, as seen in Fig. 1.11. The reason that this process is an appropriate choice for illustrating difficulty with convergence is

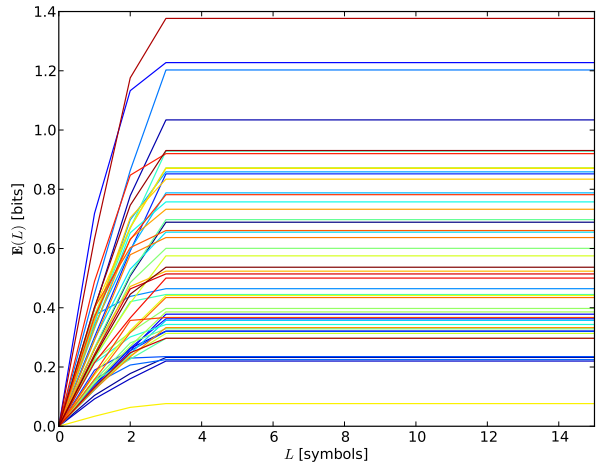

Figure 1.10: Excess entropy estimates for 50 instances of full order-3 Markov chains (see Fig. 1.9 for the topology). Notice that the estimates become extremely good (actually exact) at $L = 3$. This is a consequence of the process being finite order Markov.

that it is not Markovian; we might say that it is infinite-order Markov. This can be intuitively understood in the following way: If we have access to only a finite symbol history, say N symbols, then when we encounter a word of N ones, we cannot provide the proper distribution over futures. At a coarse level, we don't know whether the sequence is currently even in length and therefore has the option of terminating with a zero, or if it is currently odd in length and therefore *must* continue with at least one more one. Therefore no finite history (finite order Markov) model can properly generate the Even Process. This is a fundamental difference[20] between Markov models, or chains; and hidden Markov models, or functions of Markov chains **cite Markov vs hidden Markov**.

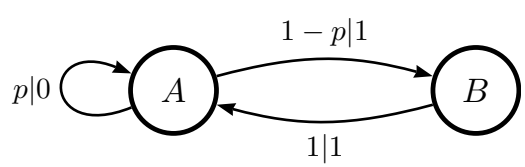

Figure 1.11: The Even Process requires that all blocks of uninterrupted ones, with zeros on either side, be even in length.

[20]This fundamental difference is *not* the difference between node output models and edge output models. Given a unifilar edge output model (this is a 'hidden model') with N nodes and M symbols, there are at most $N \times M$ edges. The corresponding node output model (also a 'hidden' model) then trivially has at most $N \times M$ nodes. This difference is of course important, but never involves transforming a 2 state model to an infinite state model

Prediction	N even	N odd
$\Pr(X_N = 0 \vert X_0 = 1, \ldots, X_{N-1} = 1)$	$1 - \frac{p}{2}$	$\frac{p}{2}$
$\Pr(X_N = 0 \vert X_{-1} = 0, X_0 = 1, \ldots, X_{N-1} = 1)$	p	0
$\Pr(X_N = 0 \vert \mathcal{S}_N = A)$	p	–
$\Pr(X_N = 0 \vert \mathcal{S}_N = B)$	–	0

Table 1.1: The above table illustrates that for the Even Process, for any length N, there exists a word (all ones) such that prediction based on that word alone is different than prediction based on that word knowing that the previous symbol is 0. Notice that the optimal probabilities, those predicted after the block of ones is begun by a zero, are the same as those predicted by the appropriate induced causal state.

It is plain to see in the left pane of Fig. 1.12 that, for a substantial subset of instances, the relative errors in the excess entropy estimates do not fall within acceptable bounds even when considering correlation lengths up to 10. The excess entropy, defined as the infinite sum of the entropy rate overestimates, is continually being fed by new overestimates, as is seen in the right pane of Fig. 1.12. One should probably object at this point saying that with appropriate algorithms and compute power, lengths far beyond 10 must certainly be accessible. This objection is certainly valid, but misses the point of the illustration. First, this is only the 'simplest of the difficult' examples. Depending on the application, model sizes will have dozens or hundreds of nodes. Second, as a matter of theoretical investigation, we would like to be able to calculate the excess entropy for infinite ϵ-machines. Any previous algorithm will fail in this task. Third, the brute force griding out of ever better approximations does not strike us as the proper way to really understand how excess entropy behaves. To illustrate, reverse-type questions about $\mathbf{E}$ are difficult to resolve numerically: The question, "for what value of p does $\mathbf{E} = 1/\pi$?" poses a reasonable computational challenge. Presumably one would have to sample points in the range of p, estimate $\mathbf{E}$ for each, and through interpolation and possibly some manner of successive approximation, hone in on the correct value. This is of course possible, but completely non-generalizable. To examine some other $\mathbf{E}$ value might require resampling a different region of p values. Furthermore, the addition of a single state would require starting the whole procedure from scratch.

We would also like to note that the ability to even represent the relative error in the excess entropy estimates, as in Fig. 1.12, is only made possible by making use of the algorithm we developed to determine the exact value of $\mathbf{E}$. Before access to the limit that our estimates were al-

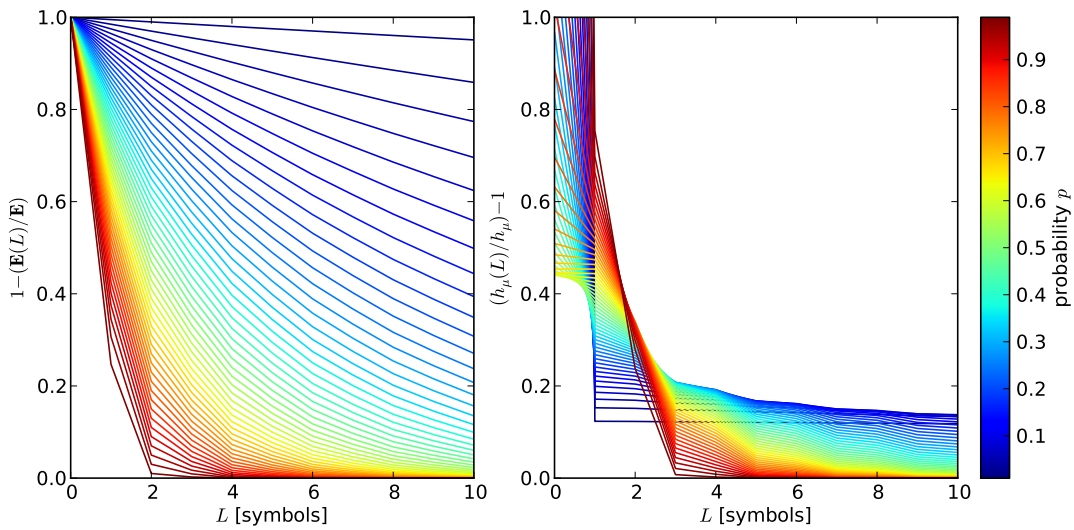

Figure 1.12: The non-Markovianness of the Even Process leads to some members of the family having very slow convergence. (Left) Relative errors in the excess entropy estimates show that even considering correlation lengths up to 10 is grossly inadequate for a large collection of processes. (Right) Relative error of entropy rate estimates are very slow to approach zero for members on the blue end of the spectrum. This process serves as a key motivating example in the search for analytic forms for **E**.

legedly approaching, excess entropy estimate plots were much more undetermined. Figure 1.13 demonstrates the slow, indeterminate growth of the estimates. There is of course the bound from **shalizi** $\mathbf{E} \leq C_\mu$, but from this simulation, there is no clear way to bound any instance away from C_μ at all.

In Ch. **??** we present our method for calculating **E** for an entire parameterized family of ϵ-machines at once. Moreover, this method is finitely terminating even for infinite order Markov processes [21]. To put some concreteness to the technique, it can be rapidly calculated by hand that the solution to the above challenge is the result of this equation,

$$\frac{1}{\pi} = \log(2-p) - \frac{1-p}{2-p}\log(1-p).$$

[21] There are some questions remaining as to what happens in the case of infinite transient states. It appears that when the recurrent states are reachable, that this algorithm will be finite despite the infinite transients. The algorithm needs only to reach all recurrent states (not to reach them via all possible transient paths). When the recurrent states are not reachable, it is known that in some cases, one can define an infinite sequence that converges to the correct result and find the limit analytically. It is hoped that this procedure can be made general. As transient states and infinite states are not discussed here, the reader should look for results in the upcoming Ref. **cite Extension of E algo to infinite**

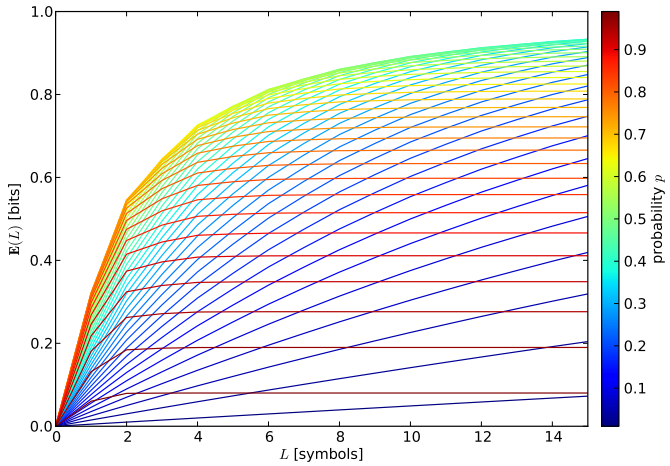

Figure 1.13: Excess entropy estimates for the Even Process without access to the actual limit **E**. Its estimates increase in a very slow manner making claims about convergence, except for very trivial ones, difficult.

§1.8 Crypticity and Cryptic Order

The study of the structure of stochastic processes through their ϵ-machine representations has lead to the the recognition of two new and important quantities: the crypticity, and the cryptic order. Intuitively, these two ideas spring from focusing one's attention not on the information region associated most directly with prediction—the excess entropy or predictive information—but rather with the region that characterizes the information above and beyond **E** necessary for determining the causal state, and thereby for making predictions. This is what we call the crypticity.

§1.8.1 Crypticity

Definition. *The* crypticity *,* χ *, of a process is defined,*

$$\chi = H[\mathcal{S}_0 | \overrightarrow{X}_0]$$

We represent this quantity in our I-diagram as the difference between the statistical complexity and the excess entropy (see Fig. 1.14). At first, it might seem as though the definition of the ϵ-machine ought to obviate any information except for that which is predictive information. The ϵ-machine is, after all, the causal representation of the process. How can we reconcile these intuitions? The essential idea is this: optimal prediction, which is what causal states are built for,

requires not only the ability to match up histories with the appropriate future, or set of futures; it also requires the ability to match up histories with the appropriate distribution over futures, and *these pairings can overlap*.

A positive crypticity means that despite all your hard work in noting the relations between pasts and futures, and determining which class of pasts you are in, there exists a particular future which can follow more than one class (even all classes) of pasts. Supposing that future is realized, you might wish you had been less careful, as the result might[22] have been the same. To say this a little differently, and mathematically,

$$\underbrace{H[X_0^L] - H[X_0^L|\mathcal{S}_0]}_{\text{net earnings}} = \underbrace{H[\mathcal{S}_0]}_{\text{gross earnings}} - \underbrace{H[\mathcal{S}_0|X_0^L]}_{\text{taxes}}.$$

To expand upon this interpretation a little, the 'gross earnings' is the amount that enters consideration. The 'net earnings' is the amount of useful resource. Trivially the, the difference is what is given up in 'taxes'. This analogy is appealing and correct in that only in very rare cases can you get away with paying no taxes [23].

It is the last term—taxes—which is our crypticity. To assure the reader that this information waste is not just a corner-case, note that the Golden Mean, a process we have already introduced, has a crypticity $\chi = 2/3$, a significant fraction of the total stored information, $C_\mu = \log(3) - 2/3 \simeq 0.918$.

§1.8.2 Cryptic Order

As the crypticity is a newly defined quantity, it is natural to attempt to tease it apart in ways similar to quantities we have dealt with in the past. The crypticity can be interpreted as the state-based companion information to the predicive information, **E**. That said, the primary dissection tool used to understand processes by thinking about their predictive information has been the Markov order. The Markov order describes the length scale of the correlations among symbols that give rise to probabilistic conditional independence; this independence is another way of describing an optimal predictor.

[22] There's the rub.

[23] It is not claimed that this is a deep analogy, but the 'economics of information' is an attractive thought; it suggests competition and optimization. It also encourages us to search for off-shore [quantum] information accounts. Other analogies have been explored for the crypticity, most notably in the context of heat engines. There, the statistical complexity is likened to heat transference and excess entropy to the derived work. The ratio then is a measure of the 'thermodynamic efficiency of the machine. There is much work yet to be done to firmly establish an economic or thermodynamic relationship.

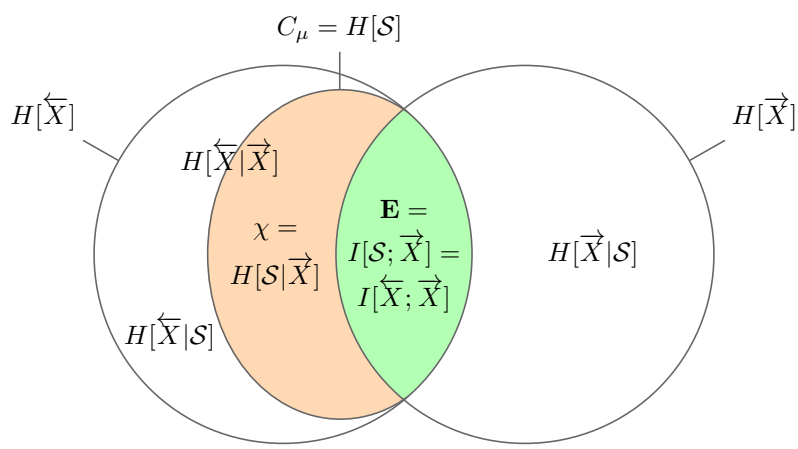

Figure 1.14: This highlights the crypticity χ in orange as the difference between the state information C_μ and the predictive information $\mathbf{E}$. In this sense, crypticity can be thought of as 'modeling overhead'.

To begin, it is useful to have a picture of what Markov order is in terms of our I-diagram. We can stratify the past in terms of the random variables $\{X^1_{-1}, X^2_{-2}, X^3_{-3}, ...\}$. The intersection of this stratification with the future is shown in Fig. 1.15. For the details as to why intersection with a stratification is allowed in this way, and for why it is also non-trivial, see App. **??**. One feature of ϵ-machines is illustrated by the fact that an equivalent way of understanding Markov order is the depth of history required for determining the causal state. It is this which speaks to the fundamental nature of causal states, and which allows us to make the statement that the cryptic order is a 'companion' order.

Cryptic order is a new length scale introduced to characterize the way in which the information associated with crypticity is distributed in the process. We argue that the cryptic order is as fundamental to the nature of processes as the Markov order. It has a slightly different flavor in that it involves causal states, whereas Markov order can be defined without them. For an illustration as to how the cryptic and Markov orders can be different, see Fig. 1.17. The cryptic order is the length scale appropriate for capturing the uncertainty in the causal state *given the future*—it captures the crypticity.

Properties of the crypticity and the cryptic order are the subjects of Ch. **??**. This collection of definitions and proofs marks the beginning of a new and fundamental characterization of stochastic process. This characterization is thought to be fundamental as it is such a close analog to the Markov order. It is deemed impactful because, unlike the Markov order, it makes reference

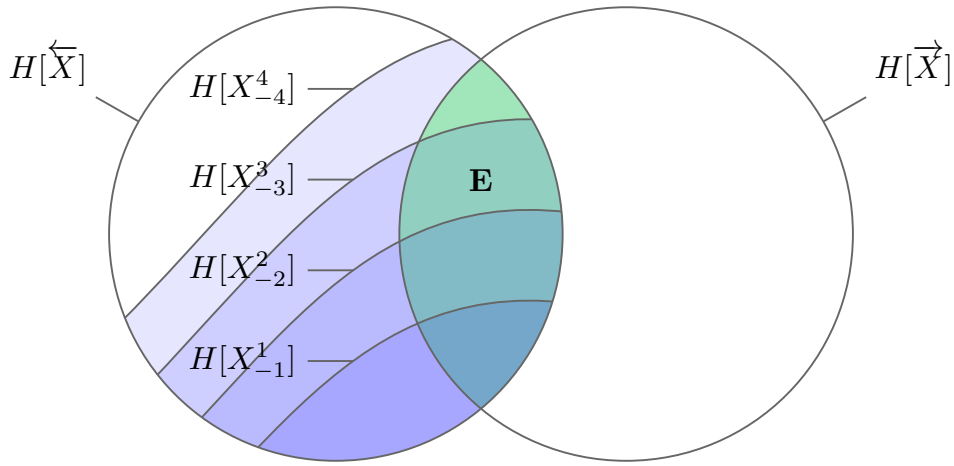

Figure 1.15: An illustration of a process which is order-4 Markov. The past $H[\overleftarrow{X}]$ is shown as being stratified in the standard way. We can see that conditioning on the past 4 variables reduces as much uncertainty in the future as does conditioning on the entire past. Conditioning on only the past 3 variables, however, neglects the upper tip of the mutual information, $\mathbf{E} = I[\overleftarrow{X}; \overrightarrow{X}]$.

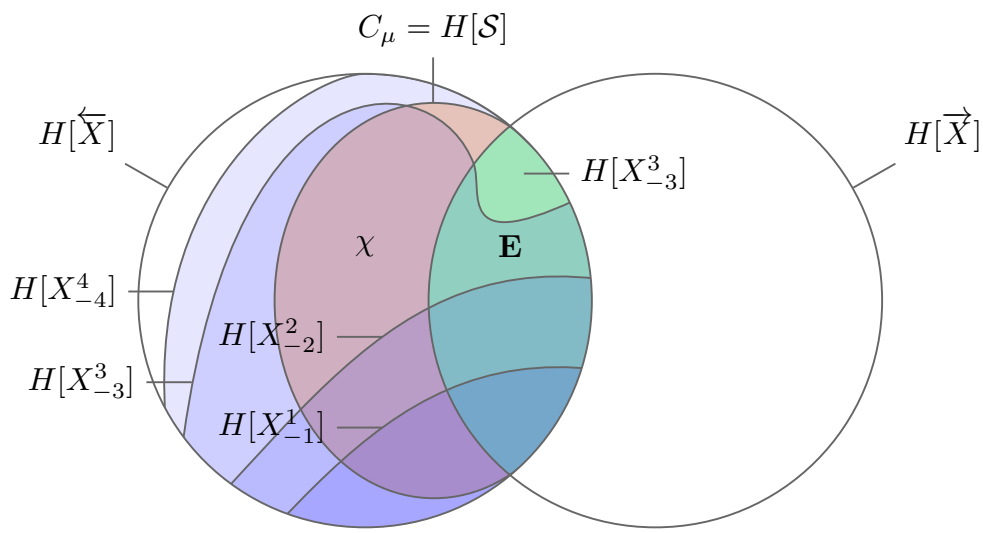

Figure 1.16: This is another illustration of an order-4 Markov process. The causal state has been added to the diagram and the boundaries made a little more curvy to anticipate future I-diagrams. Notice that in addition to the length 4 statistics being sufficient for capturing $\mathbf{E}$, the same is true for capturing χ which is the remainder of C_μ. In contrast, the length 3 statistics are insufficient for both $\mathbf{E}$ and χ. Being insufficient for $\mathbf{E}$ is why the process is order-4 Markov. Being insufficient for χ is why the process is order-4 cryptic.

to states, something that the generically non-zero crypticity strongly suggests we do. Also, the states refered to are not any state, but causal states, and so the naturalness of the ϵ-machine in its ability to deliver quantities such as h_μ and C_μ extends this naturalness to the cryptic order.

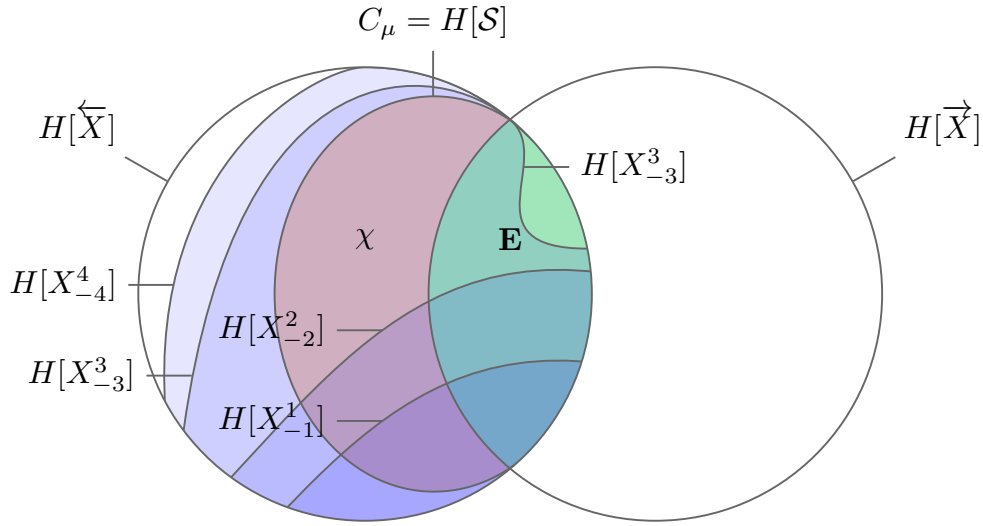

Figure 1.17: An illustration of a process with differing cryptic and Markov orders. The Markov order is 4; this is the first history length which contains all of the predictive information. Notice that the length 3 history curves back again missing a portion of **E**. The cryptic order is 3 because although the length 3 history misses some portion of **E**, it does determine the causal state conditioned on the future. Note that $H[X^3_{-3}]$ is labeled twice for clarity.

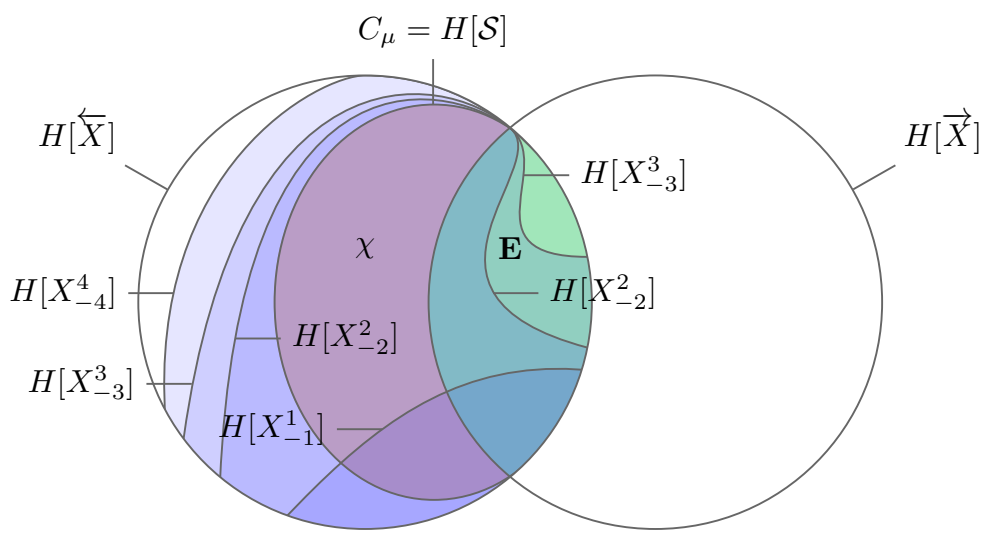

Figure 1.18: The Markov and cryptic orders may differ by more than one. This is an instance where the Markov order is 4, yet the cryptic order is 2. Two entropies are labeled twice for clarity.