Title
Bio-inspired problems in rate-distortion theory

Permalink
https://escholarship.org/uc/item/8234q7f8

Author
Marzen, Sarah

Publication Date
2016

Peer reviewed|Thesis/dissertation
“In neuroscience one must be cautious about using Shannon’s formulation of the role of statistical regularities, because the brain uses information in different ways from those common in communication engineering.”

Barlow, 2001
Acknowledgements

I am intellectually indebted to a great many people, especially James P. Crutchfield, Christopher Hillar, Simon DeDeo, Michael R. DeWeese, Susanne Still, William Bialek, Rob Phillips, Ilya Nemenman, Paul Riechers, Joel Zylberberg, and many others. I was helpfully supported by the U.C. Berkeley Chancellor’s Fellowship, the National Science Foundation Graduate Research Fellowship, and the Ling-Lie Chua Excellence Award.
# Contents

1 Optimal sensors as compressors ............................................. 1
   1.1 What is distortion? ................................................. 2
   1.2 An abridged primer on rate-distortion theory ...................... 4
   1.3 Relating sensory costs to the rate-distortion function ........... 7

2 Perceptual coding of natural image patches ................................. 9
   2.1 Optimal codings are discrete ..................................... 9
   2.2 Validating a new image patch compression scheme ................. 15

3 Sensor tradeoffs in large environments .................................... 19
   3.1 Scaling of requisite codebook size with number of sensory inputs ........................................ 20
   3.2 Universality: when sensor tradeoffs are nearly independent of environment .................................. 23
   3.3 Predictions for phenotypic variation in sensory neuron number ........................................... 27

4 Predictive rate-distortion ...................................................... 31
   4.1 Fundamentally foiled by the curse of dimensionality .............. 33
   4.2 Computing optimal lossy predictive features, regardless ........ 34
   4.3 Examples ................................................................... 38
      4.3.1 Unhidden and Almost Unhidden Processes ................. 40
      4.3.2 First-order Phase Transitions at $\beta = 1$ ............... 43
      4.3.3 Temporal Asymmetry in Lossy Prediction ................ 46
      4.3.4 Predictive Hierarchy in a Dynamical System ............ 50
   4.4 Coarse-grained mixed states as nearly lossless predictive features ........................................... 52

5 Stimuli to test a sensor’s predictive capabilities .......................... 62
   5.1 Renewal processes .................................................... 62
      5.1.1 Discrete-time renewal processes ............................ 63
      5.1.2 Infinitesimal discretization time ......................... 70
      5.1.3 Continuous-time renewal processes ....................... 72
      Continuous-time causal states ................................... 73
      Wave propagation on continuous-time $\epsilon$-machines ...... 76
      Continuous-time excess entropy ................................ 80
   5.2 Restricted hidden semi-Markov processes .......................... 81
   5.3 Not just complicated, but complex ................................. 84
      5.3.1 Complexity from nonergodicity ............................ 86
      5.3.2 Complexity without infinite excess entropy ............ 89
      5.3.3 Complexity in fractal renewal processes ............... 91

Bibliography ......................................................................... 103

A Causal states and minimal maximally predictive models ................ 110
B Rate-distortion theory vs. predictive rate-distortion theory 112
# List of Figures

1.1 Rate-distortion setup .............................................. 5
2.1 Image compression performance of Hopfield network ........ 17
3.1 Scaling of rate-distortion function in low-fidelity regime ... 20
3.2 Scaling of rate-distortion function in high-fidelity regime .. 21
3.3 Signs of weak universality ........................................ 24
3.4 Expected rate-distortion function ................................. 25
3.5 Weak universality .................................................. 25
3.6 Fixed distortion contours .......................................... 28
4.1 Predictive rate-distortion setup ................................... 32
4.2 Curse of dimensionality in estimating predictive rate-distortion functions ............................................. 34
4.3 Simple nonunifilar source and Golden Mean $\epsilon$-machines .. 41
4.4 Predictive rate-distortion functions for the simple nonunifilar source ...................................................... 42
4.5 Predictive rate-distortion functions for the Even Process .... 45
4.6 Random Insertion Process $\epsilon$-machine ......................... 47
4.7 Random Insert Process predictive rate-distortion functions . 48
4.8 Random Insertion Process feature curves .......................... 49
4.9 Tent map iterative dynamics ........................................ 51
4.10 Tent map rate-distortion functions ............................... 53
4.11 Predictive hierarchy revealed by predictive rate-distortion . 60
4.12 Fractal dimensions of mixed state presentations ................ 61
4.13 Fractal dimension of mixed states of the parametrized SNS .. 61
5.1 Hidden Markov models of two renewal process generators .. 63
5.2 The possible $\epsilon$-machine architectures for discrete-time stationary renewal processes ............................ 65
5.3 Smaller time discretization increases apparent length of temporal correlations ............................................ 70
5.4 Generative model of a continuous-time renewal process .... 73
5.5 $\epsilon$-machine for the “generic” renewal process ............... 77
5.6 $\epsilon$-machine of the eventually Poisson process ............... 78
5.7 $\epsilon$-machine of the eventually $\Delta$-Poisson process ........... 79
5.8 Maximally predictive models of certain continuous-time discrete-alphabet processes ...................................... 82
5.9 Prediction hierarchy for stationary ergodic processes ......... 85
5.10 Predictability and complexity of fractal renewal process ..... 101
Dedicated to my sisters, whose work in science has already proven more useful than the entirety of this thesis
Abstract

Bio-inspired problems in rate-distortion theory

by

Sarah E. Marzen

Doctor of Philosophy in Physics

University of California, Berkeley

Professor Michael R. Deweese, Chair

Designed and evolved sensors are faced with a difficult optimization problem: storing more information about the environment and reaping greater “rewards” costs time, energy, and material. Quantifying the benefits and costs associated with storing environmental information will aid design of more efficient sensors and allow other researchers to test whether or not a particular biosensor is, in fact, efficient. One principled approach to this quantification comes from rate-distortion theory. We discuss three biologically inspired problems in rate-distortion theory. First, we validate a new approach to perceptually compressing natural image patches; second, we show that sensing tradeoffs in large fluctuating environments depend only on coarse environmental statistics; and third, we describe a new method for calculating predictive features and memory-prediction trade-offs in stochastic environments. For the last of these problems, we suggest a new class of stimuli that might be used to profitably test whether a particular biosensor is an efficient predictor of its environment.
Chapter 1

Optimal sensors as compressors

Confusing one environmental state for another can be costly due to a subsequent suboptimal choice of action. For example, mistaking a lion for a domesticated cat might lead to a mauled hand, while mistaking a domesticated cat for a lion might lead to unnecessary energy expenditure running from the feline. However, correctly distinguishing between a lion and domesticated cat requires mental effort, e.g. a larger number of neurons devoted to object recognition or a correspondingly larger number of adenosine triphosphate molecules. Can we quantify this tradeoff in some universal way— that is, without specifying the particular biological sensory system under study?

The efficient coding hypothesis (sometimes called the redundancy reduction hypothesis) can lay claim to being such a universal principle. Originally, the efficient coding hypothesis concerned itself with lossless transmission of environmental information (Barlow, 1961). Over the past five and a half decades, the original hypothesis has morphed into related statements about the independence of neural firing, among other things (Simoncelli and Olshausen, 2001, and references therein), as more research made it clear that redundancy of the neural code was not reduced between first (e.g., retina) and later (e.g., primary visual cortex) sensory layers (Barlow, 2001).

We would like to revisit Barlow, 1961 with an eye towards lossy compression and a refocus on material constraints. Although the activity of neurons in primary visual cortex and downstream regions is highly redundant, the activity of earlier sensory regions (e.g., retina) still might be shaped by a focus on “the economy of neuron number” due to the distance between, say, the retina and primary visual cortex and the energetic cost entailed (Attwell and Laughlin, 2001; Chklovskii and Koulakov, 2004). Natural signals are almost always continuous random variables, thereby exceeding the expressive abilities of any finite number of sensory neurons whose spiking patterns are subject to nonzero intrinsic noise (Longtin, 2013). As such, an organism’s sensory systems are constantly operating in a lossy regime, unavoidably losing some of the incoming sensory information en route to higher-order processing stages\(^1\).

We therefore conjecture, following Palmer et al., 2015; Salisbury and Palmer, 2015, that biological sensors are near-optimal lossy compressors of their sensory inputs. The research here on rate-distortion theory revisits

\(^1\)Some of the celebrated confirmations of the efficient coding hypothesis (Atick and Redlich, 1990) assume that neurons communicate via firing rate, for which the above logic would not hold. However, communication via firing rate requires a non-instantaneous response, which we view as undesirable.
two basic questions in various bio-inspired scenarios. First, how little information can we lose while operating within given biophysical constraints? And second, what information should we lose?

Related prior work suggests that biological sensory systems are optimized to communicate sensory information given material, energy, or timing constraints (Laughlin, 2001; Varshney, Sjöström, and Chklovskii, 2006; Balasubramanian, Kimber, and Berry II, 2001; Laughlin, Steveninck, and Anderson, 1998; Levy and Baxter, 1996; Schreiber et al., 2002; Chklovskii and Koulakov, 2004). These previous treatments of such sensory tradeoffs often sought to maximize the mutual information between stimulus and neural response—something that rate-distortion theory penalizes—given resource constraints. This seemingly small difference in objective asks us to specify the behaviorally relevant cost to inaccurate internal representations and treats the mutual information as a universal lower bound on volume, timing, and energy constraints.

Rather than find exact resource costs for a particular biological sensory system, we study fundamental lower bounds on resource constraints for a variety of information sources. Here, rate is a stand-in for material, energy, and timing costs, as described in Sec. 1.3. The distortion measure imbues environmental information with a behaviorally relevant meaning, as described in Sec. 1.1, which suggests modifications to the traditional rate-distortion objective.

There are far more weaknesses in the current setup than one would hope for in that these information-theoretic lower bounds are potentially very weak. Biological organisms necessitate an extension of existing lower bounds to the case of memoryful information sources, memoryful channels, feedback, and instantaneous response. Some recent and not-so-recent work suggests a way forward. The case of memoryful sources and memoryless channels was tackled in (Berger, 1971, Ch. 7), but the rate-distortion function there is difficult to calculate; non-asymptotic results were studied in Kostina and Verdú, 2012. And there are currently no extensions of the rate-distortion theorem to memoryful information sources and memoryful channels, though the recursive information bottleneck suggests an intriguing avenue forward (Still, 2014). Hopefully, some of the methods or results presented here extend to more realistic problem setups.

1.1 What is distortion?

The appropriate choice of information source and distortion measure depends heavily on the particular biological system under study. Some workers have used rate-distortion theory to study everything from chemotaxis (Andrews and Iglesias, 2007) to transcription (Tlusty, 2008) to prediction in the salamander retina (Palmer et al., 2015) to human vision (Sims, 2015). In Andrews and Iglesias, 2007 and Palmer et al., 2015, investigators hypothesized that biological sensory systems under study were optimized for particular distortion measures; in Sims, 2015, a distortion measure was inferred after assuming that the organism cared about rate; and in Tlusty, 2007; Tlusty, 2008, the distortion measure was written in terms of unknown matching constants between different amino acids.
1.1. What is distortion?

Our viewpoint is closest to that espoused in Palmer et al., 2015. Perceptions are only useful insofar as they allow us to reap environmental rewards through our actions. It is instructive to briefly translate between a more common framework in which a sensation $\tilde{x}$ probabilistically leads to an action $a$ according to some $p(a|\tilde{x})$, and in which we reap rewards based on some reward function $r(x, a)$ (Barto, 1998). Our expected reward is

$$\langle r(x, a) \rangle = \sum_{x, \tilde{x}, a} p(x)p(\tilde{x}|x)p(a|\tilde{x})r(x, a) = \sum_{x, \tilde{x}} p(x)p(\tilde{x}|x) \left( \sum_a p(a|\tilde{x})r(x, a) \right).$$

If our action policy correctly utilizes information we obtain about sensory states $x$, then $\sum_a p(a|\tilde{x})r(x, a)$ should be maximized when $\tilde{x} = x$. We assume this to be the case and define a distortion between $x$ and $\tilde{x}$ as

$$d(x, \tilde{x}) := \left( \sum_a P(A = a | \tilde{X} = x)r(x, a) \right) - \left( \sum_a P(A = a | \tilde{X} = \tilde{x})r(x, a) \right).$$

This leads to the relationship

$$\langle r(x, a) \rangle = \left( \sum_{x, \tilde{x}, a} p(x)p(\tilde{x} = x|X = x)P(A = a | \tilde{X} = x)r(x, a) \right) - \langle d(x, \tilde{x}) \rangle$$

and so maximizing expected reward is equivalent to minimizing expected distortion. In other words, the distortion measure can be chosen so as to “integrate out” the reward function, although this may not always be desirable conceptually.

In reality, actions are delayed from perceptions— an inconvenient truth ignored in Chapters 2-3. In other words, in a discrete-time framework, action $a_t$ (dependent on sensory state $s_t$) occurs simultaneously with environmental state $x_{t+1}$.

Thus, in Chapters 4-5, we pivot our focus from estimation to prediction. As before, the organism aims to maximize expected reward, $\langle r(x_{t+1}, a_t) \rangle$. More generally, we might imagine that the reward depends on the entire future trajectory, $r(x_{t+1}, a_t)$. In a more general setting, we imagine that the next actuator state depends on the previous actuator state and current sensor state, $p(a_{t+1}|x_t, s_t, a_t) = p(a_{t+1}|s_{t-1}, a_{t-1})$, and that future environment trajectories depend on past environment trajectories and past actuator states (but not past sensory states), $p(x_{t+1}|x_{t+1}, a_{t-1}, s_{t-1}) = p(x_{t+1}|x_{t+1}, a_t)$. With these relations in hand, we find that expected reward is equivalent to

$$\langle r(x_{t+1}, a_t) \rangle = \tilde{r}((x_{t+1}, a_t), s_t)$$

where

$$\tilde{r}((x_{t+1}, a_t), s_t) = \sum_{x_{t+1}, a_t} p(x_{t+1}|x_{t+1}, a_t)p(a_t|s_{t-1}, a_{t-1})r(x_{t+1}, a_t).$$
Chapter 1. Optimal sensors as compressors

To maximize reward, the organism minimizes a “predictive distortion” \( \langle d(\hat{x}_t, s_t) \rangle \) defined by

\[
d(\hat{x}_t, a_t, s_t) := \max_s \bar{r}(\hat{x}_t, s) - \bar{r}(\hat{x}_t, s_t)
\]

where \( \hat{x}_t \) is shorthand for \( x_{t+1} \), \( a_t \). If the sensorimotor loop isn’t closed—for instance, if an experimentalist completely controls an animal’s surroundings—then we can drop the dependence of the future environment on past actions, so that \( d(\hat{x}_t, s_t) \) can be written as \( d(\hat{x}_t, s_t) \). This framework is easily altered to account for more general dependencies of the current sensor state and current actuator state on past sensor and actuator states, e.g. as in Barnett and Crutchfield, 2015. Depending on the sensor substrate, we can view either \( X_{t+1} \) or \( S_t, X_{t+1} \) (Still, 2014, Recursive Information Bottleneck in) as the information source to be compressed. The former case is discussed in Chapter 4, as it applies directly to the experimental setup in Palmer et al., 2015.

1.2 An abridged primer on rate-distortion theory

In this section, we describe the tradeoff between minimizing expected distortion (see Sec. 1.1) and minimizing material required, heat dissipation, and timing delays. In doing so, we appeal to classic results in rate-distortion theory and newer results in nonequilibrium thermodynamics.

Consider the setup in Fig. 1.1, a minimal version of the usual information theory setup (Yeung, 2008, Ch. 8). An information source sends \( n \) input symbols, \( x_1, \ldots, x_n \), to a channel, which sends one of \( M \) codewords to a receiver. This receiver then attempts to decode the \( n \) input symbols as \( \hat{x}_1, \ldots, \hat{x}_n \). Some distortion measure quantifies the “distance” between the true input and our decoded guess, \( d(x_{1:n+1}, \hat{x}_{1:n+1}) \), and for purposes of this thesis, we assume that this distortion is the sum of single-symbol distortions, \( \frac{1}{n} \sum_{i=1}^n d(x_i, \hat{x}_i) \). We assume that we see a large number of these input strings, so that this distortion tends to \( \langle d(x, \hat{x}) \rangle \) averaged over \( p(x, \hat{x}) \).

Implicit in this setup is the idea that increases in the number of codewords \( M \) is costly. This is true almost no matter the setup. In a more traditional setup, we speak of communicating codewords over a channel, in which at worst each of the \( M \) symbols is represented by a different \( \log_2 M \) bits per input symbol, or rate.

There are some combinations of distortions and rates, \( \langle d(x, \hat{x}) \rangle, \frac{\log M}{n} \), that are achievable, and some that are not. The boundary between achievable and unachievable combinations is given by the rate-distortion function \( R(D) \) shown in Fig. 1.1(bottom). Alternatively, one can speak of its inverse, the distortion-rate function \( D(R) \). The better a sensor, the closer its rate and distortion lie to this boundary, to the extent that high rates are costly for the sensory system.

When the information source and channel are memoryless—meaning
1.2. An abridged primer on rate-distortion theory

A limited number of codewords constitutes an information bottleneck. If the encoder has few enough codewords, the estimated sequence will be different than the input sequence. In applications to biological sensory systems, one can think of the encoder as sensory neurons and the decoder as a homunculus. At bottom, the rate-distortion function $R(D)$ (the blue curve) separates achievable (white) from unachievable (hatched) combinations of rates and distortions. A near-optimal sensor will have a rate and distortion pair which is closer to the rate-distortion function.

Figure 1.1: At top, an illustration of a rate-distortion setup. A limited number of codewords constitutes an information bottleneck. If the encoder has few enough codewords, the estimated sequence will be different than the input sequence. In applications to biological sensory systems, one can think of the encoder as sensory neurons and the decoder as a homunculus. At bottom, the rate-distortion function $R(D)$ (the blue curve) separates achievable (white) from unachievable (hatched) combinations of rates and distortions. A near-optimal sensor will have a rate and distortion pair which is closer to the rate-distortion function.
that neither remembers what they have done previously—then the rate-
distortion theorem (Yeung, 2008, Ch. 8) says that
\[ R(D) = \min_{p(\tilde{x}|x): (d(x, \tilde{x})) \leq D} I[X: \tilde{X}] \]  
(1.7)

where \( \tilde{X} \) is now some random variable that (usually) shares some information with \( X \). If the source is memoryful, then (Berger, 1971)
\[ R(D) = \lim_{n \to \infty} \min_{p(\tilde{x}_0:n|x_0:n): (d(x_0:n, \tilde{x}_0:n)) \leq D} \frac{I[X_0:n; \tilde{X}_0:n]}{n}. \]  
(1.8)

A caveat is in order. First, the “optimal stochastic codebooks”
\[ p^*_D(\tilde{x}|x) := \arg \min_{p(\tilde{x}|x): (d(x, \tilde{x})) \leq D} I[X: \tilde{X}] \]  
(1.9)

are not so optimal from a coding perspective; the entropy rate of the output symbols is not equal to the mutual information between input and output (Berger, 1971). Hence, \( p^*_D(\tilde{x}|x) \) only provides a guide as to the statistics of the codebooks one would use in practice. Optimal deterministic codebooks are a step in the right direction, as their output entropy rate is equal to the mutual information between input and output. However, optimal deterministic codebooks still achieve the lower bound on rate implied by rate-distortion theory only when responses can be arbitrarily delayed from perceptions; see Sec. 1.3.

We can estimate \( R(D) \) by the Blahut-Arimoto algorithm (Blahut, 1987), which essentially amounts to gradient descent with respect to \( p(\tilde{x}|x) \) on the corresponding Lagrangian
\[ L_{\beta} = I[X; \tilde{X}] + \beta(d(x, \tilde{x})). \]  
(1.10)

This is a convex objective in \( p(\tilde{x}|x) \), and so has a unique global minimum in the simplex that satisfies \( \frac{\partial L_{\beta}}{\partial p(\tilde{x}|x)} |_{p^*_D(\tilde{x}|x)} = 0 \). Some algebra shows this condition to be equivalent to
\[ p^*_\beta(\tilde{x}|x) = p^*_\beta(\tilde{x})e^{-\beta d(x, \tilde{x})} / Z_\beta(x) \]  
(1.11)

where \( Z_\beta(x) \) is a normalization factor. (See Chapter 2 for a derivation.) The Blahut-Arimoto algorithm Cover and Thomas, 2006, Ch. 10 of corresponds to iteration of this equation:
\[ p_{t+1}(\tilde{x}|x) = p_t(\tilde{x})e^{-\beta d(x, \tilde{x})} / Z_t(x) \]  
(1.12)

where \( Z_t(x) \) is the appropriate normalization factor. From \( p_\beta(\tilde{x}|x) \), we calculate the resultant rate \( R_\beta \) and expected distortion \( D_\beta \). As \( \beta \) sweeps from 0 (high distortion) to \( \infty \) (low distortion), \( R_\beta \) and \( D_\beta \) parametrically trace out the rate-distortion function \( R(D) \) shown in Fig. 1.1.
1.3 Relating sensory costs to the rate-distortion function

The exact relationship between the rate of a sensor and its material costs, power consumption, or timing delays depends on the physical substrate which lossily communicates environmental information. We attempt to make headway regardless.

For concreteness, consider \( m \) sensory neurons that form an information bottleneck between environment and downstream brain regions. This coding has an expected distortion of the sensory representation, \( \langle d(x, \tilde{x}) \rangle = D \). Though some work suggests that the neural code might be analog (based on spike timing), there is inherent noise in neural circuitry that effectively imposes a minimal discretization time—a few milliseconds (Nemenman et al., 2008)—on which the neural code operates. We therefore think of the sensory neural code as a binary vector of length \( m \) in which \( 1 (0) \) in the \( i \)th position codes for a spike (no spike) from neuron \( i \) in that minimal discretization time.

Material and timing costs can trade off with one another, but both run into fundamental limits quantified by the rate-distortion function. If \( R(D) \leq m \), then one can instantaneously decode each input from a binary vector of length \( m \). Alternatively, we might imagine that \( n \) environmental inputs are encoded and decoded together, in block form. In that case, if \( R(D) \leq m/n \), then one can instantaneously decode each block of \( n \) inputs, \( x_1, \ldots, x_n \), as a binary vector of length \( m \). This timing delay is an unattractive side-effect for biological sensory systems trying to operate in an online fashion. If \( R(D) \geq m \), then we can acquire additional expressiveness by coding each input \( x \) as a string of binary vectors of length \( m \). The expected length of the neuronal output string is no less than \( R(D)/\log_2 2^m = R(D)/m \), which, when multiplied by the number of input symbols sensed thus far, is the timing delay between encoding and decoding. These timing delays are, again, an unattractive side-effect for biological sensory systems trying to operate in an online fashion.

A more generally applicable nonequilibrium thermodynamics viewpoint ties the rate-distortion function to power consumption. Memoryless channels implicitly have a measure-reset cycle: first, the channel senses the environment, and the channel communicates its measurement to some “homunculus”; and afterwards, the channel resets its internal state \(^2\). This cycle is repeated ad nauseum. The total power \( P \) required to maintain such a channel is lower-bounded by \( k_B T I[X; \tilde{X}] \) (Sagawa and Ueda, 2009; Parrowdo, Horowitz, and Sagawa, 2015), which is lower-bounded by

\[
k_B T R_1(D) := k_B T \min_{p(\tilde{x}|x):\langle d \rangle \leq D} I[X; \tilde{X}] \geq k_B T R(D),
\]

where \( R_1(D) = R(D) \) for memoryless source and channel, and \( R_1(D) \geq R(D) \) for memoryful source and memoryless channel (Berger, 1971). This is a different energetic consideration than that mentioned in Tlusty, 2008.

Finally, if downstream regions try to losslessly transmit information provided by the sensory system, then the channel capacity of downstream

\(^2\)Memoryful channels are not so well-conceptualized as measure-reset cycles, and so these bounds may not directly apply to memoryful channels.
brain regions limits the rate at which the sensory system transmits information. There is little evidence to support the claim that brain regions downstream from sensory systems are so “transparent”, so we avoid focusing on this resource constraint.

In short, the rate-distortion function places a lower bound on the size of the physical substrate, on timing delays between encoding and decoding environmental input, and on the power required to maintain sensor functionality.
Chapter 2

Perceptual coding of natural image patches

In this chapter, we use rate-distortion to validate a new bio-inspired approach to coding natural image patches. Strangely, optimal codings are usually discrete. We use this fact, and a new perceptual distortion measure proposed by Wang et al., 2004, to motivate the study of discretized natural image patches.

In Sec. 2.1, we correct minor errors in prior proofs that optimal codings are usually discrete. Readers who are interested primarily in lossy compression of discretized natural image patches are advised to skip directly to Sec. 2.2.

2.1 Optimal codings are discrete

It is well-known that continuous objects are optimally coded by discrete random variables under quite general conditions. Previous proofs of the theorem (Matějka and Sims, 2011; Rose, 1994) in the rate-distortion setting, however, were slightly flawed. The equations used to prove that the reproduction alphabet is usually discrete were derived by assuming the existence of a well-defined probability density function $p_\beta(\tilde{x}|x)$ – which, the proofs then go on to argue, usually does not exist. We use a measure-theoretic formulation to circumvent this difficulty.

For concreteness, we restrict our attention to coding $X$ with realizations $x \in \mathbb{R}$. Let $\mu$ be a probability measure on $\mathbb{R}$ with $\mathbb{P}(X \in S) = \mu(S)$ such that the support of $\mu$ lies in $R = [-A, A]$ for some positive $A$. We define a coding of $\mu$ to be a family of Borel probability measures $\{\nu_x\}_{x \in \mathbb{R}}$, each of which has the property that $x \mapsto \nu_x(E)$ is $\mu$-measurable for each Borel set $E$. Given this data, there is a construction in Johnson, 1984 (usually overlooked as necessary in the literature) of a measure $\lambda$ on the product space $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ satisfying:

$$\int \left( \int f d\nu_x(\tilde{x}) \right) d\mu(x) = \int f d\lambda, \quad \text{for } f \in L^+(\mathbb{R} \times \mathbb{R}).$$

Given such a measure $\lambda$, we have a marginal distribution given by:

$$\nu(E) = \int_R \nu_x(E) d\mu(x) = \lambda(R \times E), \quad \text{for each Borel set } E.$$ 

We now form the standard (independent) product distribution of $\pi = \mu \times \nu$ on $\mathbb{R}^2$. With these general notions in place, we can now define the distortion
and rate of a coding $\{\nu_x\}_{x \in R}$. The expected distortion is given by

$$E[d] = \int d(x, \tilde{x}) \, d\lambda,$$

and the coding rate is

$$I[\lambda] = \sup_{\{G_i\}} \sum_{i=1}^{N} \lambda(G_i) \log \frac{\lambda(G_i)}{\pi(G_i)},$$

where $\{G_i\}_{i=1}^{N}$ is a finite partition of $R \times R$ into rectangles.\(^1\) When $I[\lambda] < \infty$, then the Radon-Nikodym derivative $\frac{d\pi}{d\lambda}$ exists and we have Pinsker, 1964, Thm. 2.4.2:

$$I[\lambda] = \int \log \frac{d\lambda}{d\pi} \, d\lambda.$$

The original optimization problem from Ch. 1 now reads,

$$R(D) = \inf_{\lambda \in \Lambda_D} I[\lambda],$$

where $\Lambda_D$ is the set of measures on $R^2$ such that $\mu$ is the marginal ($\mu(E) = \lambda(E \times R)$ for all Borel sets $E$) and $E[d] = D$.

Finally, we assume that the distortion measure is such that $d(x, \tilde{x})$ is infinite when either $x$ or $\tilde{x}$ are not in $R$. This allows us to assume that the support of $\nu$ is contained in $R$, so that $\lambda$ has bounded support as well. It is straightforward to see that we can then study a compact subset of $\Lambda_D$ under the Lévy-Prokhorov metric (Luenberger, 1997) for which the support of $\nu$ lies in $R$, without loss of generality.

We begin our demonstration of Theorem 3 by sketching a proof of the measure-theoretic analogue of (1.11).

**Lemma 1.** (Optimization lemma) Consider a distortion measure $d(x, \tilde{x})$ which is $\infty$ when $x \notin R$ or $\tilde{x} \notin R$. For all $D > 0$, there is a unique coding and corresponding joint probability measure $\lambda$ which attains the infimum of $I[\lambda]$ on $\Lambda_D$.

**Proof.** Without loss of generality, for finite $D$, we can restrict our attention to codings $\{\nu_x\}_{x \in R}$ for which the support of $\nu_x$ is contained in $R$, since otherwise, our expected distortion is infinite. In an abuse of notation, we now call the space of these restricted codings and corresponding joint probability measures $\Lambda_D$. As $R$ is bounded, $\lambda$ has bounded support, and $\Lambda_D$ is a compact, convex, topological space under the Lévy-Prokhorov metric (Luenberger, 1997). In addition, $I[\lambda]$ is a convex function of $\lambda$ (Smith, 1971). The lemma follows. \(\square\)

To the best of our knowledge, proofs of Lemma 1 or similar results rely on the bounded support of $\lambda$. For instance, in Matějka and Sims, 2011,

---

1In Pinsker, 1964, there are countably many rectangles; but since $R$ is bounded, we can without loss of generality restrict to a finite set.

2We use the convention that $0 \log 0 = 0$ and $0 \log \frac{a}{0} = 0$. We never have to calculate $a \log \frac{a}{0}$ for $a > 0$ by the following logic. Let $G_i = E_i \times F_i$, so $\pi(G_i) = \mu(E_i) \nu(F_i)$. As $\mu$ and $\nu_x$ are positive measures, we have $\lambda(G_i) = \int_{E_i} \int_{F_i} \nu_x \, d\mu \leq \int_{E_i} \int_{F_i} \nu \, d\mu = \nu(F_i)$ and $\lambda(G_i) = \int_{E_i} \int_{F_i} \nu \, d\mu \leq \nu(E_i)$; and as $\lambda$ is also a positive measure, $0 \leq \lambda(G_i) \leq \min(\mu(E_i), \nu(F_i))$. When $\pi(G_i) = 0$, then $\min(\mu(E_i), \nu(F_i)) = 0$ and thus $\lambda(G_i) = 0$. 
2.1. Optimal codings are discrete

boundedness follows from the monotonicity of a suitable distortion measure and the boundedness of $R$. Here, we stipulate a different, somewhat artificial structure to the distortion measure which serves the same purpose. In Lemma 1, we cite a reference (Smith, 1971) that studied a related optimization problem – that of fixing a coding $\{\nu_x\}_{x \in R}$ and asking for the input probability measure $\mu$ which maximizes $I[\lambda]$.

From Lemma 1, there is a unique $\lambda^* \in \Lambda_D$ which minimizes $I[\lambda]$:

$$\lambda^* = \arg \min_{\lambda \in \Lambda_D} I[\lambda].$$

The uniqueness and existence of this $\lambda^*$ is crucial for proving the lemma below.

**Lemma 2.** (First order conditions) Assume that $\mu$ has bounded support $R$ and that $d$ is a bounded, normal distortion measure (meaning that $d(x, x) = 0$ for all $x$ and $d(x, \tilde{x}) < \infty$ for all $x, \tilde{x}$). Define $\lambda^*$ as above, let $\nu^*$ be the corresponding marginal distribution, and let $\pi^* = \mu \times \nu^*$. The Radon-Nikodym derivative $\frac{d\lambda^*}{d\pi^*}$ exists and satisfies, for some $\beta \geq 0$,

$$\frac{d\lambda^*}{d\pi^*} = e^{-\beta d(x, \tilde{x})} Z(x),$$

with $Z(x) = \int e^{-\beta d(x, \tilde{x})} d\nu^*$, for all $(x, \tilde{x})$ in the support of $\lambda^*$. In addition, the following holds for $\tilde{x}$ in the support of $\nu^*$:

$$1 = \int \frac{1}{Z(x)} e^{-\beta d(x, \tilde{x})} d\mu(x).$$

**Proof.** Consider $\lambda \in \Lambda_D$ and define

$$\lambda_\epsilon := (1 - \epsilon)\lambda^* + \epsilon \lambda,$$

with $\lambda_\epsilon \in \Lambda_D$. Denote the weak derivative by

$$f(\lambda^*) := \lim_{\epsilon \to 0} \frac{I[\lambda_\epsilon] - I[\lambda^*]}{\epsilon}.$$

Since $\lambda^*$ is the global minimum of $I[\lambda]$ in $\Lambda_D$, we have $f(\lambda^*) \geq 0$. But from Smith, 1971, a necessary and sufficient condition for $I[\lambda^*] = \inf_{\lambda \in \Lambda_D} I[\lambda]$ is $f(\lambda^*) \leq 0$, so together, we search for $f(\lambda^*) = 0$.

Consider the coding $\nu_x = \frac{1}{m(R)}$ where $m(R)$ is the standard Lebesque measure of $R$ for all $x \in R$. By construction this coding has finite expected distortion and finite rate. The continuity of the rate-distortion function then implies that the rate-distortion function is finite for all nonzero expected distortions, and hence, $I[\lambda^*] = R(D) < \infty$. As stated earlier, this implies that $\frac{d\lambda^*}{d\pi^*}$ exists and that

$$I[\lambda^*] = \int \log \frac{d\lambda^*}{d\pi^*} d\lambda^*.$$
Continuity of $I[\lambda]$ with respect to $\lambda$ (Smith, 1971) implies that $I[\lambda_e]$ is also finite, so that $\frac{d\lambda_e}{d\pi_e}$ exists where $\pi_e = (1 - \epsilon)\pi^* + \epsilon\pi$ and

$$I[\lambda_e] = \int \log \frac{d\lambda_e}{d\pi_e} d\lambda_e.$$ 

We restrict ourselves to perturbations $\lambda$ such that $\lambda_e \ll \lambda^* \ll \pi^* \ll \pi_e$ so that the Chain Rule can be applied to the Radon-Nikodym derivatives. This will still provide us with necessary conditions on $\frac{d\lambda^*}{d\pi^*}$. The Chain Rule gives

$$\frac{d\lambda_e}{d\pi_e} = \frac{d\lambda_e}{d\lambda^*} \frac{d\lambda^*}{d\pi^*} \frac{d\pi^*}{d\pi_e},$$

so that

$$\log \frac{d\lambda_e}{d\pi_e} = \log \frac{d\lambda_e}{d\lambda^*} + \log \frac{d\lambda^*}{d\pi^*} + \log \frac{d\pi^*}{d\pi_e}.$$ 

It is straightforward to show that $\lambda \ll \lambda^*$ under the assumption that $\lambda_e \ll \lambda^*$, since if $\lambda^*(E) = 0$ then $\lambda_e(E) = 0$, giving $\lambda(E) = \frac{\lambda(E) - (1 - \epsilon)\lambda^*(E)}{1 - \epsilon} = 0$. Then $\frac{d\lambda}{d\lambda^*}$ exists and $\frac{d\lambda^*}{d\pi^*} = (1 - \epsilon) + \epsilon \frac{d\pi^*}{d\pi_e}$. Similar logic implies $\frac{d\pi^*}{d\pi_e} = \frac{1}{1 - \epsilon} \left(1 - \epsilon \frac{d\pi}{d\pi_e}\right)$. Together, these allow us to express the difference $I[\lambda_e] - I[\lambda^*] = \int \log \frac{d\lambda_e}{d\lambda^*} d\lambda_e - \int \log \frac{d\lambda^*}{d\pi_e} d\lambda^*$ as

$$I[\lambda_e] - I[\lambda^*] = \int \log \frac{d\lambda_e}{d\lambda^*} d\lambda_e + \int \log \frac{d\lambda^*}{d\pi_e} d\lambda_e + \int \log \frac{d\pi^*}{d\pi_e} d\lambda_e$$

$$= (1 - \epsilon) \int \log \left(1 - \epsilon + \epsilon \frac{d\lambda}{d\lambda^*}\right) d\lambda^*$$

$$+ \epsilon \int \log \left(1 - \epsilon + \epsilon \frac{d\lambda}{d\lambda^*}\right) d\lambda^*$$

$$+ (1 - \epsilon) \int \log \frac{d\lambda^*}{d\pi_e} d\lambda_e$$

$$+ \epsilon \int \log \frac{d\lambda^*}{d\pi_e} d\lambda$$

$$+ \int \log \left(\frac{1}{1 - \epsilon} \left(1 - \epsilon \frac{d\pi}{d\pi_e}\right)\right) d\lambda_e$$

$$- \int \log \frac{d\lambda^*}{d\pi_e} d\lambda^*$$

$$= (1 - \epsilon) \int \log \left(1 + \epsilon \frac{d\lambda}{1 - \epsilon d\lambda^*}\right) d\lambda^*$$

$$+ \epsilon \int \log \left(1 + \epsilon \frac{d\lambda}{1 - \epsilon d\lambda^*}\right) d\lambda^*$$

$$+ \epsilon \int \log \frac{d\lambda^*}{d\pi_e} d\lambda_e$$

$$+ \epsilon \int \log \frac{d\lambda^*}{d\pi_e} d\lambda + \int \log \left(1 - \epsilon \frac{d\pi}{d\pi_e}\right) d\lambda_e.$$
2.1. Optimal codings are discrete

This leads to

\[
\frac{I[\lambda] - I[\lambda^*]}{\epsilon} = \int \frac{(1 - \epsilon) \log \left( 1 + \frac{\epsilon}{1 - \epsilon} \frac{d\lambda}{dx} \right)}{\epsilon} d\lambda^* \\
+ \int \log \left( 1 + \frac{\epsilon}{1 - \epsilon} \frac{d\lambda}{d\lambda^*} \right) d\lambda^* \\
- \int \log \frac{d\lambda^*}{d\pi^*} d\lambda^* + \int \log \frac{d\lambda^*}{d\lambda^*} d\lambda \\
+ \int \log \left( 1 - \frac{\epsilon}{\epsilon} \frac{d\mu}{d\nu^*} \right) d\lambda^*.
\]

Since \( R \) is bounded, each of the Radon-Nikodym derivatives \( \frac{d\lambda}{dx} \) and \( \frac{d\pi}{d\nu^*} \) is bounded; hence, the integrands \( \frac{(1 - \epsilon) \log \left( 1 + \frac{\epsilon}{1 - \epsilon} \frac{d\lambda}{dx} \right)}{\epsilon}, \log \left( 1 + \frac{\epsilon}{1 - \epsilon} \frac{d\lambda}{d\lambda^*} \right), \) and \( \log \left( 1 - \frac{\epsilon}{\epsilon} \frac{d\mu}{d\nu^*} \right) \) are bounded for sufficiently small \( \epsilon \). From an application of L’Hôpital’s Rule, we find that

\[
\lim_{\epsilon \to 0} \frac{(1 - \epsilon) \log \left( 1 + \frac{\epsilon}{1 - \epsilon} \frac{d\lambda}{dx} \right)}{\epsilon} = \frac{d\lambda}{dx^*} \quad \text{and} \quad \lim_{\epsilon \to 0} \frac{\log \left( 1 + \frac{\epsilon}{1 - \epsilon} \frac{d\lambda}{d\lambda^*} \right)}{\epsilon} = -\frac{d\pi}{d\nu^*}.
\]

The dominated convergence theorem then implies that

\[
f(\lambda^*) = \int \frac{d\lambda}{d\lambda^*} d\lambda^* + \int \log \frac{d\lambda^*}{d\pi^*} (\lambda - \lambda^*) - \int \frac{d\pi}{d\pi^*} d\lambda^*.
\]

Immediately from these calculations, we have \( \int \frac{d\lambda}{d\lambda^*} d\lambda^* = 1 \). Note that \( \frac{d\pi}{d\nu^*} = \frac{d\mu}{d\nu^*} \), so that \( \int \frac{d\pi}{d\nu^*} d\lambda^* = 1 \). Hence,

\[
f(\lambda^*) = \int \log \frac{d\lambda^*}{d\pi^*} (\lambda - \lambda^*).
\]

Recall that \( \lambda \in \Lambda_D \), implying that \( \int d(x, \tilde{x}) d(\lambda - \lambda^*) = 0 \) and \( \int \gamma(x) d(\lambda - \lambda^*) = 0 \) for any \( \gamma(x) \in L^+ \), the space of non-negative integrable functions. Thus, a necessary condition for \( \lambda^* \) to have \( I[\lambda^*] = R(D) \) is

\[
0 = \int \left( \log \frac{d\lambda^*}{d\pi^*} + \beta d(x, \tilde{x}) + \gamma(x) \right) d(\lambda - \lambda^*),
\]

for some scalar \( \beta \). Since this holds for any \( \lambda \in \Lambda_D \) with \( \lambda \ll \lambda^* \), we must have \( \log \frac{d\lambda^*}{d\pi^*} + \beta d(x, \tilde{x}) + \gamma(x) = 0 \) or

\[
\frac{d\lambda^*}{d\pi^*} = e^{-\beta d(x, \tilde{x})} Z(x),
\]

for \( (x, \tilde{x}) \) in the support of \( \lambda^* \). Finally, note that for all Borel sets \( E \):

\[
\int_E \left( \int_R \frac{d\lambda^*}{d\pi^*} \, d\nu^* \right) d\mu = \int \frac{d\lambda^*}{d\pi^*} \, d\pi^* = \lambda^*(E \times R) = \mu(E),
\]

which implies that \( \int_R \frac{d\lambda^*}{d\pi^*} \, d\nu^* = 1 \) for all \( x \in R \), giving

\[
Z(x) = \int e^{-\beta d(x, \tilde{x})} \, d\nu^*.
\]
Similarly, for all Borel sets $E$, we have
\[
\int_E \left( \int_R \frac{d\lambda}{d\pi} d\mu \right) d\nu^* = \int_{R \times E} \frac{d\lambda}{d\pi} d\pi^* = \lambda^*(R \times E) = \nu^*(E),
\]
which implies that $\int_R \frac{d\lambda}{d\pi} d\mu = 1$ or
\[
1 = \int e^{-\beta d(x, \tilde{x})} Z(x) d\mu,
\]
for all $\tilde{x}$ in the support of $\nu^*$. The lemma statement follows. \(\square\)

We now argue in Theorem 3 that optimal reproduction alphabets of continuous-valued input sources are usually finite when the distortion measure is analytic.

**Theorem 1.** Assume $\mu$ has bounded support, and that $p(x)$ exists, i.e. $p(x) = \frac{d\mu}{dx}$. Let $d(x, \tilde{x})$ be an analytic difference distortion measure on $R \times R$, so that $d(x, \tilde{x}) = \begin{cases} \rho(\tilde{x} - x) & x, \tilde{x} \in R \\ \infty & \text{else} \end{cases}$. Then $\nu^*$ is discretely supported.

**Proof.** Let $I$ be the support of $\nu^*$. Recall from Lemma 1 and Lemma 2.1 that $\nu^*$ exists, is unique, and that
\[
1 = \int \frac{1}{Z(x)} e^{-\beta d(x, \tilde{x})} d\mu,
\]
for $\tilde{x} \in I$ where $Z(x) = \int e^{-\beta d(x, \tilde{x})} d\nu^*$. For notational ease, we define $F(\tilde{x}) := \int_{-\infty}^{\infty} \frac{p(x)}{Z(x)} e^{-\beta p(x, \tilde{x})} dx - 1$, where we have used $p(x) = \frac{d\mu}{dx}$. Define a related function $G(\tilde{x}) = \int_{-\infty}^{\infty} \frac{p(x)}{Z(x)} e^{-\beta p(x, \tilde{x})} dx - 1$. By construction, $F(\tilde{x}) = G(\tilde{x})$ for $\tilde{x} \in R$, so $G(\tilde{x}) = 0$ for $\tilde{x} \in I$. We work with $G$ rather than $F$ in what follows, as the analyticity of $\rho$ and Lemma III.2 of Matějka and Sims, 2011 implies that $G$ is analytic.

We claim that if $I$ has an accumulation point, then $G(\tilde{x}) = 0$ for all $\tilde{x} \in \mathbb{R}$, not just $\tilde{x} \in I$. As stated above, $G(x)$ is analytic. Suppose that there is an accumulation point $\tilde{x}^* \in I$ of $G(\tilde{x})$; then there exists a sequence $\{x_n\}_{n=1}^{\infty}$ with $\lim_{n \to \infty} x_n = x^*$, which implies that $G(\tilde{x}) = 0$ for all $\tilde{x} \in \mathbb{R}$.\(^3\)

So we have that either $I$ has no accumulation point or $\int_{-\infty}^{\infty} \frac{p(x)}{Z(x)} e^{-\beta p(x, \tilde{x})} dx = 1$ for all $\tilde{x} \in R$. Suppose the latter holds. We claim that the solution $Z(x)$ to this integral equation is unique up to a set of $\mu$-measure 0. If

\(^3\)Because $G$ is analytic, we can write $G(x) = \sum_{k=0}^{\infty} \frac{G^{(k)}(x_\ast)}{k!} (x - x_\ast)^k$ where $G^{(k)}(x_\ast)$ is bounded. For notational ease, define a new sequence $\epsilon_i := x_i - x^*$ (so $\lim_{i \to \infty} \epsilon_i = 0$) and coefficients $a_k := \frac{G^{(k)}(x_\ast)}{k!}$, so that $\sum_{k=0}^{\infty} a_k \epsilon_i^k = 0$ for all $i$. We claim that $a_k = 0$ for all $k$, thereby showing that $G(x) = 0$ for all $x$. To show this, we can use strong induction. First, when $k = 0$, then $a_0 = G(x^*) = 0$ because $x^* \in I$; this is the base case. Next, we assume that $a_k = 0$ for all $k \leq N$ and try to show that $a_{N+1} = 0$. Under these strong induction assumptions, $0 = \sum_{k=N+1}^{\infty} a_k \epsilon_i^k = \epsilon_i^{N+1} \sum_{k=N+1}^{\infty} a_k \epsilon_k^{-N-1}$, and since $\epsilon_i \neq 0$, we have $0 = \sum_{k=N+1}^{\infty} a_k \epsilon_k^{-N-1}$. Rewriting, we have $a_{N+1} = -\sum_{k=N+2}^{\infty} a_k \epsilon_k^{k-N-1}$, or $0 \leq |a_{N+1}| \leq \sum_{k=N+2}^{\infty} |a_k| |\epsilon_k|^{k-N-1}$. Finally, we take the limit as $i \to \infty$ to find $0 \leq |a_{N+1}| \leq \lim_{i \to \infty} \sum_{k=N+2}^{\infty} |a_k| |\epsilon_i|^{k-N-1}$. Without loss of generality, restrict $\{a_k\}_{k=0}^{\infty}$ to a monotone decreasing subsequence; then the monotone convergence theorem implies that $\lim_{i \to \infty} \sum_{k=N+2}^{\infty} |a_k| |\epsilon_i|^{k-N-1} = \sum_{k=N+2}^{\infty} |a_k| \lim_{i \to \infty} |\epsilon_i|^{k-N-1} = 0$. This implies $0 \leq |a_{N+1}| \leq 0$, or $a_{N+1} = 0$, as desired.
there were some other solution \( p(x)/Z'(x) \), then subtraction would yield
\[
\int_{-\infty}^{\infty} p(x) \left( \frac{1}{Z(x)} - \frac{1}{Z'(x)} \right) e^{-\beta \rho(\tilde{x} - x)} dx = 0.
\]
As \( e^{-\beta \rho} \) is everywhere positive, this integral equation would imply that \( Z(x) = Z'(x) \) up to a set of \( \mu \)-measure 0. And when \( x \notin R \), \( Z(x) = \int e^{-\beta d(x,\tilde{x})} dv^* \) is 0 by construction of \( d \), so \( Z(x) \) is unique. It is straightforward to show that this integral equation can be satisfied by
\[
p(x) \int_{-\infty}^{\infty} e^{-\beta \rho(z)} dz = \int e^{-\beta d(x,\tilde{x})} dv^* (\tilde{x}).
\]
Consider integrating both sides over \( x \) from \(-\infty\) to \( \infty \), giving
\[
\int_{-\infty}^{\infty} p(x) dx \int_{-\infty}^{\infty} e^{-\beta \rho(z)} dz = \int_{-\infty}^{\infty} e^{-\beta d(x,\tilde{x})} dx.
\]
First note that
\[
\int_{-\infty}^{\infty} e^{-\beta d(x,\tilde{x})} dx = \int_{-A}^{A} e^{-\beta \rho(\tilde{x} - x)} dx = \int_{\tilde{x} - A}^{\tilde{x} + A} e^{-\beta \rho(z)} dz.
\]
Furthermore, as \( \rho \) is analytic, we have:
\[
\int_{\tilde{x} - A}^{\tilde{x} + A} e^{-\beta \rho(z)} dz < \int_{-\infty}^{\infty} e^{-\beta \rho(z)} dz.
\]
Thus,
\[
\int \int_{-\infty}^{\infty} e^{-\beta d(x,\tilde{x})} dx dv^* (\tilde{x}) < \int_{-\infty}^{\infty} e^{-\beta \rho(z)} dz \int dv^* (\tilde{x}) = \int_{-\infty}^{\infty} e^{-\beta \rho(z)} dz.
\]
However,
\[
\int_{-\infty}^{\infty} p(x) dx \int_{-\infty}^{\infty} e^{-\beta \rho(z)} dz = \int_{-\infty}^{\infty} e^{-\beta \rho(z)} dz
\]
so that the equality given above leads to the following contradiction:
\[
\int_{-\infty}^{\infty} e^{-\beta \rho(z)} < \int_{-\infty}^{\infty} e^{-\beta \rho(z)}.
\]
Hence, \( \int_{-\infty}^{\infty} \frac{p(x)}{Z'(x)} e^{-\beta \rho(\tilde{x} - x)} dx = 1 \) for all \( \tilde{x} \in \mathbb{R} \) cannot hold. In particular, the set \( I \) has no accumulation point. As \( R \) is bounded, we must have that \( I \) is a finite set. The statement of the theorem now follows.

2.2 Validating a new image patch compression scheme

How then, do we optimally code natural images? They are fundamentally continuous objects, but with a limited number of sensory neurons, we must understand them in some discrete way unless the statistics of natural images are matched to a perceptual distortion measure such that the Shannon lower bound (Rose, 1994) is achieved. Such a match is highly unlikely.

Many researchers have tried to develop perceptual distortion measures for natural images, with some success (Wang and Bovik, 2009). Surprisingly, prior research (Hillar, Mehta, and Koepsell, 2014) using the perceptual distortion measure introduced in Wang et al., 2004 has shown that a
reasonable distortion measure between two $L \times L$ natural image patches is actually given by the Hamming distance between two binary vectors of length $2L^2$ constructed from each image patch as follows.

A Hopfield network is a collection of binary nodes ("neurons") with states $x_i \in \{0, 1\}$ connected by weights $J_{ij}$ of possibly varying strength (Hopfield, 1982). When presented with some pattern (corresponding to a binary vector with the length of the number of nodes) the Hopfield network updates the state of its neurons asynchronously according to the following rule:

$$x_i \rightarrow H\left(\sum_j J_{ij} x_j\right).$$

(2.4)

When weights are symmetric, these dynamics cause convergence to a fixed point (Hopfield, 1982). These fixed points are often called "memories". In the context of lossy compression, the Hopfield network implements a many-to-one map that encodes original binary vectors to a smaller number of binary vectors.

In this new image compression scheme, two neurons in a Hopfield network are assigned to each pixel, one “OFF” and one “ON”. If the pixel value is between 0 and 127, then the OFF neuron is turned on and the ON neuron is turned off; if the pixel value is between 127 and 129, then both neurons are turned off; and otherwise, the OFF neuron is turned off and the ON neuron is turned on. Typically, one chooses weights that will store the desired memories. However, in this particular application, the desired memories are unknown, and are instead learned by fitting the probabilistic model implied by the Hopfield network to empirical estimates of probabilities of various binary vectors (Sohl-Dickstein, Battaglino, and DeWeese, 2011; Hillar, Sohl-Dickstein, and Koepsell, 2012).

Prior research (Hillar, Mehta, and Koepsell, 2014) proposed that the memories of these Hopfield networks could be used to code the image patches. We can validate this choice of coding by comparing the rate and distortion of the Hopfield encoding to the rate-distortion function when $L = 2$. These binary vectors of length $2L^2$ inherit a probability distribution from the probability distribution over natural image patches, giving us an estimate of $p(x)$ from the frequency histogram of these binary vectors in van Hateren (Schaaf and Hateren, 1996) natural image patches. We use a Hamming distortion measure, and compute the resulting rate-distortion function using the Blahut-Arimoto algorithm with 1, 5, 50, 500 iterations at each $\beta$. The different choice of iterations showcases the quick convergence of the Blahut-Arimoto algorithm.

Fig. 2.1(top) shows that the Hopfield network is quite close to achievable limits. One benefit to using the memories of a Hopfield network as codewords is that storing the network requires storing $(2L^2)$ real numbers, as opposed to requiring an exponentially large $(2^{2L^2})$ look-up table.

Interestingly, the model implied by these Hopfield networks-- an Ising model with

$$p^{\text{model}}(x) = \frac{1}{Z} \exp \left(\sum_{i,j} J_{ij} x_i x_j\right)$$

(2.5)

is quite bad in that the model consistently under or over estimates the local
2.2. Validating a new image patch compression scheme

Figure 2.1: At top, a comparison between the Hopfield network’s expected distortion and rate relative to the rate-distortion function. The rate-distortion function was calculated using the Blahut-Arimoto algorithm with varying numbers (1, 5, 50, 1000) of iterations. The inset images depict $p_{\beta}(\tilde{x}|x)$ in matrix form for various $\beta$, and the red dot marks the performance of the Hopfield network image compression scheme. At bottom, a comparison between the probabilities $p_{\text{model}}(x)$ implied by the Hopfield network and the frequencies $p_{\text{data}}(x)$. 
maxima of the true probability distribution. But, as shown in Fig. 2.1 (bottom), the Hopfield network correctly estimates the binary vectors which are local maxima of the probability distribution \( p(x) \). An accurate model of the probabilities of these local maxima is not required for good lossy compression.
Chapter 3

Sensor tradeoffs in large environments

In Chapter 2, we analyzed the rate-distortion function for discretized natural image patches, for which there exists a well-defined perceptual distortion measure. However, we have very little idea as to the “distortion measures” that dictate the full behavior of organisms. Can we make predictions about or explain natural phenomena regardless?

Consider a model world in which the reproductive rate of organisms depends only on coding cost and distortion. To specify this world, we need to specify at least two things:

- Environmental statistics: how many sensory states are seen, and with what frequency?
- What does a model organism “care” about? In other words, how are distortion measures chosen?\(^1\)

From knowledge of environmental statistics and the organism-environment interaction, we can calculate the tradeoffs faced by these model organisms in balancing the costs and benefits of sensory perception. This provides a minimal setup for examining the effects of environmental statistics and the perceptual preferences of the organism on the required number of sensory neurons.

We will establish two mathematical results for optimal sensing in large and random environments. First, the minimal cost of confusing one input for another delineates the boundary between resource costs that grow or saturate with the number of environmental symbols. Second, the rate-distortion function is nearly invariant to fluctuations in organism-environment interactions in large environments, which we call “weak universality”.

More concretely, in Sec. 3.1, we assume that there are \( N \) equiprobable environmental inputs, so that \( p(x) = 1/N \) for all \( x \); in Sec. 3.2, we assume that \( p(x) \) is drawn from a Dirichlet distribution with concentration parameter \( \alpha \), i.e. \( p(x_1, ..., x_N) \propto \prod_{i=1}^{N} x_i^{\alpha - 1} \delta(\sum_i x_i - 1) \) where \( x_i \) is the probability of seeing sensory input \( i \). We then take inspiration from Wigner’s treatment of heavy atom energetics and assume that distortions are drawn randomly from some probability density function \( \rho(d) \), except that \( d(x, x) = 0 \) for all \( x \)– there is no cost of confusing \( x \) with itself. A key parameter for \( \rho(d) \) is the “minimal cost of confusion” \( d_{\text{min}} \), defined by

\[
  d_{\text{min}} := \inf \{ d : p(d) > 0 \}. \quad (3.1)
\]

\(^1\)As described in Chapter 1, we are potentially “integrating out” the action policy.
We will examine the effects of varying $\rho(d)$ on both scaling and weak universality in these large, random environments.

Scaling and weak universality have implications for our model organisms. The former implies that some levels of expected distortion are essentially unattainable for any organism. The latter implies that the number of sensory neurons can be chosen without detailed knowledge of environmental statistics, i.e. hard-wired.

### 3.1 Scaling of requisite codebook size with number of sensory inputs

Simulations strongly suggest the presence of two regimes, separated by the minimal confound $d_{\text{min}}$. In the “low-fidelity regime”, where expected distortion is larger than the minimal confound, resource costs asymptote to a finite constant as the number of environmental states grows without bound. See Fig. 3.1. In the “high-fidelity regime”, where expected distortion is smaller than this minimal confound, resource costs grow logarithmically with the number of environmental states. The difference between these two regimes is shown in Figs. 3.2.
3.1. Scaling of requisite codebook size with number of sensory inputs

\[ R(D) \text{ (bits)} \]

\[ N \]

\[
\begin{align*}
D &= 1.5 \\
D &= 1.001 \\
D &= 0.5 \\
\end{align*}
\]

\[ \text{Eq. 15} \]

\[ N \]

\[
\begin{align*}
0 & \leq N < 100 \\
100 & \leq N < 1000 \\
1000 & \leq N < 10000 \\
\end{align*}
\]

\[ R(D) \text{ (bits)} \]

\[ 10^1 \]

\[ 10^2 \]

\[ 10^3 \]

\[ 10^4 \]

\[ D = 0.5 \]

\[ D = 1.001 \]

\[ D = 0.5 \]

\[ \text{Eq. 15} \]

\[ \text{FIGURE 3.2: Scaling of } R(D) \text{ with } N \text{ when } \rho(d) = \begin{cases} 0 & d < 1 \\ e^{-(d-1)} & d \geq 1 \end{cases} \text{, so that } d_{\min} = 1. \text{ At top, distortions are chosen independently. At bottom, rows of independently chosen distortion matrices were randomly scrambled and averaged with the existing row. See text. The lower bound marked as } \text{Eq. 15} \text{ is given by Eq. 3.5. Standard errors show the spread in } R(D) \text{ over 25 different simulations.} \]
Lemma 3. \( R(D) \) is \( O(\log N) \) in the high-fidelity regime \((D < d_{\text{min}})\) and \( O(1) \) in the low-fidelity regime \((D > d_{\text{min}})\).

Proof. In the low-fidelity regime, the suggestive results of the simulations in Figs. 3.1 can be confirmed by a simple analytic argument. For each of the \( N \) environmental states, and expected distortion \( D \), there will be, on average, \( NP(d < D) \) states that are allowable ambiguities—i.e., that the organism can take as synonyms for that percept. While this is not the most efficient coding, it provides an upper bound to the optimal rate of \( 0 \leq R(D) \leq \log_2 \frac{1}{P(d < D)}. \) (3.2)

For the exponential this upper bound means that ambiguities accumulate sufficiently fast that the organism will need at most 6.6 bits (for \( D \) equal to 1%) or 10 bits (for \( D \) equal to 0.1%) even when the set of environmental states becomes arbitrarily large. In short, \( R(D) \) is \( O(1) \) in the low-fidelity regime.

That the bound relies on the conditional distribution function means that our results are robust to heavy-tailed distributions. What matters is the existence of harmless synonyms; for the states we end up being forbidden to confuse, penalties can be arbitrarily large. We can see this in Fig. 3.1, where we consider a log-normal distribution with mean and variance chosen so that \( P(d < 0.1\%) \) is identical to the exponential case. The existence of large penalties in the log-normal case does not affect the asymptotic behavior.

More generally, in structured environments, \( R(D) \) is bounded from above so that \( R(D) \leq \frac{1}{N} \sum_x \log_2 \left( \frac{N}{N_D(x)} \right) \), where \( N_D(x) \) is the number of synonyms for \( x \) with distortion penalty less than \( D \). As long as \( N_D \) is asymptotically proportional to \( N \), this upper bound implies asymptotic insensitivity to the number of states of the environment. This is the case even when off-diagonal entries are partially correlated, as long as correlation scales grow more slowly than environment size so that \( N_D \propto N \).

As with the low-fidelity regime, we can upper-bound coding costs by construction of a sub-optimal codebook; we can also lower-bound coding costs by finding the optimal codebook for a strictly less-stringent environment. Here, the sub-optimal codebook allocates a total probability \( C \) equally to all off-diagonal elements. The rate of this codebook is \( \log_2 N - H_b(C) - C \log_2 (N - 1) \), while its expected distortion is \( C \bar{d} \), where

\[
\bar{d} := \sum_x \sum_{\tilde{x} \neq x} d(x, \tilde{x})/N(N - 1)
\] (3.3)

is the mean off-diagonal distortion. (When \( N \) is large, this tends to \( \langle d \rangle \).) This yields an upper bound on the true rate-distortion function:

\[
R(D) \leq \log_2 N - \frac{D}{\bar{d}} \log_2 (N - 1) - H_b(D/\bar{d}). \tag{3.4}
\]

A less-stringent environment is one in which all distortions are \( d_{\text{min}} \), and since \( d(x, \tilde{x}) \geq d_{\text{min}} \), this environment has strictly lower resource costs for the same distortion. The rate-distortion function of this environment was...
3.2 Universality: when sensor tradeoffs are nearly independent of environment

Given in Shannon, 1959:

$$\log_2 N - \frac{D}{d_{\text{min}}} \log_2(N - 1) - H_b\left(\frac{D}{d_{\text{min}}}\right) \leq R(D). \quad (3.5)$$

Together, these place upper and lower bounds on the true rate-distortion function in the high-fidelity regime. In the large $N$ limit, these bounds approximately simplify to

$$(1 - \frac{D}{d_{\text{min}}}) \log_2 N \leq R(D) \leq (1 - \frac{D}{\langle d \rangle}) \log_2 N, \quad (3.6)$$

plus small $O(1)$ corrections. In short, these bounds show that the rate-distortion function is $O(\log N)$ in the high-fidelity regime, $D < d_{\text{min}}$, as expected.

The bound in Eq. 3.2 is not tight; as can be seen in Fig. 3.1, far better compressions are possible, and the system asymptotes at much lower values. Even so, this bound in Eqn. 3.2 implies that the limit of the expected rate-distortion functions in the low-fidelity regime as environmental complexity increases exists. Simulations suggest that the expected rate-distortion function $E[R(D)]$ is monotone in $N$, and so $\lim_{N \to \infty} E[R(D)]$ exists by the monotone convergence theorem.

3.2 Universality: when sensor tradeoffs are nearly independent of environment

The standard errors in Figs. 3.1-3.6 suggest something quite interesting—that two different distortion measures might yield nearly the same sensor tradeoffs if the environment is sufficiently large. In the low-fidelity regime, we expect these sensor tradeoffs to be roughly independent of environment size as well.

We test these observations more directly through a series of simulations shown in Fig. 3.3 for two different ensembles of distortion measures, $\rho(d)$, and for two different ensembles of input distributions, $p(x)$. See Figs. 3.3-3.5. The former shows 68% confidence intervals for $R(D)$ over the total ensemble at various $N$. Fig. 3.4 also shows the expected $R(D)$ is monotonically increasing with $N$ for a particular ensemble of environments. If this holds, then the monotone convergence theorem and Eq. 3.2 imply that the limit of $E[R(D)]$ as $N$ grows large exists. Finally, Fig. 3.5 shows that the maximal variance in $R(D)$ over an ensemble of environments, $\max_D \text{Var}[R(D)]$, decreases with environment size $N$.

Lemma 4. At sufficiently high $\beta$, when the support set of the reproduction alphabet includes all of the symbols, $R_\beta$ and $D_\beta$ are equal to their expectation values over the ensemble with probability 1 as $N \to \infty$.

Proof. We essentially use the strong law of large numbers.

In the limit that $p(\tilde{x})$ has full support, an exact expression exists for $R_\beta$ and $D_\beta$ from Berger, 1971. Choose an ordering of the sensory inputs. Let $\tilde{p}(x)$ be a vector of input probabilities $p(x)$, let $d$ be the distortion matrix,
FIGURE 3.3: At top, 10 rate-distortion functions calculated for environments with \( p(d) = e^{-d} \), \( \alpha = 2 \), and \( N = 20 \). At bottom, 50 rate-distortion functions calculated for environments with \( p(d) = e^{-d} \), \( \alpha = 2 \), and \( N = 400 \). The 20-fold increase in \( N \) leads to nearly identical rate-distortion functions in different environments.
3.2. Universality: when sensor tradeoffs are nearly independent of environment

Figure 3.4: As $N$ increases, the rate-distortion functions for environment ensembles with $p(d) = e^{-d}$ and concentration parameter $\alpha = 2$ seem to reach a limit; this is confirmed by the monotone convergence theorem, in which the bounds established in Eq. 3.2 imply that the limit of $E[R(D)]$ exists.

Figure 3.5: The maximal variance in $R(D)$ over the ensemble of possible environments, $\max_D \text{Var}(R(D))$ as a function of the number of sensory inputs $N$ for four environment settings, as indicated in the legend: either $p(d) = e^{-d}$ or $p(d) = \frac{1}{\sigma \sqrt{2\pi}} e^{-(\log d - \mu)^2/2\sigma^2}$ with $\mu = 3.014131$ and $\sigma = 2.40717$; and either $\alpha = 2$ or $\alpha = 5$. Variances were estimated from 50 samples of $R(D)$. The different $R(D)$ were calculated at different distortions $D$, and so linear interpolation was used to compare rate-distortion functions at the same distortion.
and let $Q_{x,\tilde{x}} = e^{-\beta d(x,\tilde{x})}$. Then

$$R_\beta = -\beta D_\beta + H[X] + \vec{p}(x)^\top \log(Q^{-1}I)$$  \hspace{1cm} (3.7)

$$D_\beta = \frac{\vec{p}(x)^\top}{Q^{-1}I} Q^{-1} (d \odot Q)(Q^{-1}I).$$  \hspace{1cm} (3.8)

When $\beta$ is sufficiently large, then the entries of $Q - I$ are much smaller than $1$ with high probability, suggesting that

$$Q = I + (Q - I)$$  \hspace{1cm} (3.9)

$$Q^{-1} = (I + (Q - I))^{-1}$$  \hspace{1cm} (3.10)

$$= \sum_{m=0}^{\infty} (-1)^m (Q - I)^m.$$  \hspace{1cm} (3.11)

By the weak law of large numbers, $(Q - I)I$ is highly concentrated around $(N - 1)(e^{-\beta d})I$ as long as the probability density function for $e^{-\beta d}$ has finite variance, so that

$$(Q - I)^m I \approx \left( (N - 1)(e^{-\beta d}) \right)^m I$$  \hspace{1cm} (3.13)

showing that

$$Q^{-1}I = \sum_{m=0}^{\infty} \left( - (N - 1)(e^{-\beta d}) \right)^m I$$  \hspace{1cm} (3.14)

$$\approx (1 + (N - 1)(e^{-\beta d}))^{-1} I.$$  \hspace{1cm} (3.15)

Then we find that

$$\vec{p}(x)^\top \log(Q^{-1}I) \approx \sum_x p(x) \log(1 + (N - 1)(e^{-\beta d}))^{-1}$$  \hspace{1cm} (3.16)

$$= - \log \left( 1 + (N - 1)(e^{-\beta d}) \right)$$  \hspace{1cm} (3.17)

so that

$$R_\beta = -\beta D_\beta + H[X] - \log \left( 1 + (N - 1)(e^{-\beta d}) \right).$$  \hspace{1cm} (3.18)

Similar manipulations, again based on the weak law of large numbers, reveal that

$$D_\beta \approx \frac{(N - 1)(de^{-\beta d})}{1 + (N - 1)(e^{-\beta d})}.$$  \hspace{1cm} (3.19)

When the probability distribution over inputs is uniform, $H[X] = \log N$; when the probability distribution over inputs is drawn from a Dirichlet distribution with concentration parameter $\alpha$, then $H[X]$ is very peaked around $\psi(N\alpha) - \psi(\alpha)$. Thus $D_\beta$ is independent of the particular distortion matrix and dependent only on $\rho(d)$ as long as the pdfs of $e^{-\beta d}$ and $de^{-\beta d}$ have finite variance.
3.3 Predictions for phenotypic variation in sensory neuron number

Because Eqs. 3.18 and 3.19 only refer to expectation values over distributions, the rate-distortion function becomes dependent only on the distribution from which distortions were drawn, and independent on the particular distortion matrix, in the large \( N \) and large \( \beta \) limit. (Extending these arguments to smaller \( \beta \) will be left to future work.) Interestingly, this simple analysis also suggests that—unlike the eigenvalue distribution of random matrices—there is no “strong universality” for these rate-distortion functions. In order for two environment ensembles as specified by \( \rho(d) \) and \( \alpha \) to yield the same \( D_\beta, R_\beta \) in this high \( \beta \) limit, they would need have the same concentration parameter \( \alpha \) and same moment generating functions \( \langle e^{-\beta d} \rangle \), and would therefore have to be the same distribution.

**Remark 1.** Those arguments in Lemmas 6 and 4 are easily extended to the case where different \( d(x, \tilde{x}) \) are correlated.

As a step towards introducing more structure into these environments, we also drew distortion matrices as follows. First, an initial \( N \times N \) distortion matrix is constructed by drawing \( d(x, \tilde{x}) \) i.i.d. from a distribution

\[
p(d) = \begin{cases} 
0 & \text{if } d < 1 \\
\frac{e^{-(d-1)}}{d} & \text{if } d \geq 1 
\end{cases}
\]

Next, each entry \( d(x, \tilde{x}) \) is replaced by \( (d(x, \tilde{x}) + d(x, \tilde{x}') \langle d \rangle) / 2 \), where \( \tilde{x}' \neq x \) is randomly chosen. This final distortion matrix has pairwise-correlated entries. This was exactly the prescription by which distortion matrices in Fig. 3.2(bottom) were drawn.

### 3.3 Predictions for phenotypic variation in sensory neuron number

We now argue that \( \beta \) has a physiological interpretation. An organism’s reproductive rate is fundamentally based on its total energy budget. The distortion \( D \) quantifies how much food energy the organism fails to reap from its environment, while its rate lower-bounds the number of sensory neurons required to reap said energy. Loosely speaking, an organism’s rate of total energy cost is

\[
\langle d(x, \tilde{x}) \rangle + \beta m
\]

where \( m \) is the number of sensory neurons in the simplistic approximation that each sensory neuron has the same power requirement \( \beta \).

If the rate-distortion function is a tight bound on \( m \)—if neurons do implement a dense combinatorial code—then the organism’s rate of total energy cost is approximated by \( D + \beta R(D) \). The distortion \( D \) that minimizes this rate of total energy cost is \( D_\beta \), and simulations in Fig. 3.6 suggest that \( D_\beta > d_{\text{min}} \) unless \( \beta \) grows logarithmically with \( N \).

A simple analytic argument explains why \( \beta \) will at least need to scale as \( \log N \) in order to retain \( D \) below \( d_{\text{min}} \). In the random environments studied below, a non-coding codebook will expected distortion of roughly \( \langle d \rangle \) and a rate of \( 0 \); a codebook in the high-fidelity regime will have some expected distortion \( D < d_{\text{min}} \) and a rate of \( C \log_2 N \) for some constant \( C \) bounded by

\[
1 - \frac{D_{d_{\text{min}}}}{\langle d \rangle} \leq C \leq 1 - \frac{D_{d_{\text{min}}}}{\langle d \rangle}.
\]

The high-fidelity codebook outperforms the non-coding codebook when \( \beta(d) \geq \beta D + C \log_2 N \), or when \( \beta \geq \frac{C}{\langle d \rangle - D} \log_2 N \).

In short, model organisms in large environments will not reliably distinguish between two sensory stimuli that are maximally similar. Doing so would require unreasonably large brains. In the notation of earlier, we expect organisms to exist in the low-fidelity regime.
Figure 3.6: Combinations of \( \beta \) and \( N \) that yield the same \( D \) for \( \rho(d) = \begin{cases} 0 & d < 1 \\ e^{-(d-1)} & d \geq 1 \end{cases} \). Standard errors show the spread in \( \beta \) over 25 different simulations.
3.3. Predictions for phenotypic variation in sensory neuron number

Within this low-fidelity regime, we naively would expect organisms to employ one of a few strategies to cope with fluctuating environments. The first strategy derives from Kelly’s classical analysis of gambling, applied to phenotypic bet-hedging— that a population of organisms should develop into a range of phenotypes to maximize expected log growth rate (Cover and Thomas, 2006, Ch. 6 of). Another strategy would involve delaying development of key sensory regions until the organism has received strong environmental cues— something which does not appear to happen (Williams and Herrup, 1988).

A third strategy would be to essentially ignore environmental fluctuations, and just use hard-wired genomic information. At first, this seems like a suboptimal strategy, in that a population of organisms that employ either of the two strategies listed above would have a higher log growth rate. However, the weak universality results presented here suggest that the necessary size of sensory brain regions, the minimum possible timing delays in sensory perception, and the minimal power required to maintain sensory brain regions all depend only on coarse environmental statistics; and optimal stochastic codebooks— related to optimal neural wiring— fluctuate wildly from environment to environment. In the examples discussed in the main text, these coarse environmental statistics are \( \rho(d) \) and \( \alpha \). More generally, these coarse environmental statistics are the parameters specifying the distribution from which distortion measures are drawn and the distribution from which probability distributions over inputs are drawn.

In apparent agreement with these findings, environmental cues are scarce during development (Williams and Herrup, 1988), there are few reports of neurogenesis in mammalian sensory brain regions (Kaslin, Ganz, and Brand, 2008), and the optimal neuronal wiring in sensory brain regions does require environmental cues. However, if weak universality-type results render generational fluctuations in sensory neuron number unnecessary, we are left with two questions. First, why do investigators find evidence of neurogenesis in non-vertebrate sensory brain regions (Kaslin, Ganz, and Brand, 2008)? And second, why is there high phenotypic variability in sensory neuron number for many animals, including primates (Williams and Herrup, 1988)?

First, Kaslin, Ganz, and Brand, 2008 notes that animals with substantial neurogenesis in sensory areas are also those that grow considerably postnatally, which— in our simple conception of organisms— corresponds to an increase in the possible actions \( a \) taken by the organism. Recall from Chapter 1 that one can connect the distortion measure directly to the reward function \( r(x, a) \) and action policy \( p(a|x) \). Changes in the set of actions will thus change the distortion measure in a (possibly) more structured way than what was considered here. That, in turn, will likely lead to an increase in the requisite sensor size, necessitating adult neurogenesis in sensory areas. We leave a delineation of the induced structure in the distortion measure to future research.

There are at least two possible explanations for phenotypic variability within our minimal model. The first explanation relies on the environment being large but finite: organisms of increasing complexity in large environments will find themselves near \( D = d_{\text{min}} \), a point at which the effect of environmental fluctuations on the rate-distortion function are amplified.
However, we have no evidence that $d_{\min}$ is nonzero. The second explanation is based on the earlier identification of the rate-distortion objective as a fitness function, implying that variability in the Lagrange multiplier $\beta$ (representative of single neuron power usage) is tightly connected to variability in the observed number of sensory neurons. Both of these explanations could be tested in lab experiments in which the distortion measure and probability distribution over inputs are hand-designed.
Chapter 4

Predictive rate-distortion

In Chapter 1, we described a mapping from reward and actions to distortion. In that discussion, we added a key constraint— that actions are delayed from perceptions. Accounting for this changes the nature of the distortion measure drastically, because distortions should now be about one’s ability to predict and not just reconstruct the environment.

How do we predict? Absent another organism providing one with side-information, we must predict based on our memory of the environmental past. If the fluctuations in environmental state have some structure, then we can, in theory, use our memory of past environmental states to predict something about the future. However, we cannot typically store everything we’ve seen, or, rather, doing so is costly, requiring arbitrarily large numbers of neurons. This leads us to two questions:

- What should we remember in order to predict the future? In other words, what are the minimal sufficient statistics of prediction?
- What about the future can be predicted? In other words, what are the minimal sufficient statistics of retrodiction?

Forward-time causal states are minimal sufficient statistics of prediction, while reverse-time causal states are minimal sufficient statistics of retrodiction. See Appendix A.

Unfortunately, in most environments, the number of neurons required to store forward-time causal states is usually infinite. In these cases, as before, limits on the number of sensory neurons provide a bottleneck that forces us to distinguish between more and less useful forward-time causal states.

A slight adaptation of the rate-distortion framework allows us to discuss this more biologically applicable scenario. Consider the setup in Fig. 4.1. It is nearly the same as the original rate-distortion setup in Fig. 1.1 with a few changes:

- Rather than encoding single symbols, the bottleneck (formerly called the encoder) encodes a semi-infinite past ($\vec{x}_1$) with a single codeword $r$.
- Rather than trying to estimate the semi-infinite past, the homunculus (formerly called the decoder) makes predictions about the future based on the codeword $r$, $p(\vec{x}_1|r)$.

As a result of this new emphasis on prediction, we employ distortion measures of the form $d(p(\vec{x}_i|\vec{x}_i), p(\vec{x}_i|r))$; and as a result of trying to code semi-infinite pasts, our coding cost takes the form $I(\vec{X}; R)$. (Such a setup
can be engineered in experiments, e.g. as in Palmer et al., 2015.) Again, some combinations of coding cost and distortion are achievable, and some are not; and the boundary between the two is called a predictive rate-distortion function. Slight modifications to the proof of the rate-distortion theorem (see Appendix B) allow us to compute this boundary as

\[ R(D) = \inf_{p(r|\mathcal{F}): d(\mathcal{F}, r) \leq D} I[\mathcal{X}; R] \]  

(4.1)

where \( d(\mathcal{F}, r) \) is of the form \( d(p(\mathcal{F}|\mathcal{X}), p(\mathcal{F}|r)) \). Again, \( \mathcal{R} \rightarrow \mathcal{X} \rightarrow \mathcal{F} \) forms a Markov chain, so that all information about the future must come from the past.

By far, the most popular predictive distortion measures are informational distortions. The predictive information bottleneck (PIB) (Still and Crutchfield, 2007; Still, Crutchfield, and Ellison, 2010; Creutzig and Sprekeler, 2008; Still, 2014) uses a distortion of the form

\[ d(\mathcal{F}, r) = D_{KL}(P(\mathcal{X}|\mathcal{F}) = \mathcal{F}, P(\mathcal{X}|R = r)) \]  

(4.2)

\[ = \sum_{\mathcal{F}} p(\mathcal{F}|\mathcal{X}) \log \frac{p(\mathcal{F}|\mathcal{X})}{p(\mathcal{F}|r)}, \]  

(4.3)

The expected distortion is more popularly expressed as

\[ \langle d(\mathcal{F}, r) \rangle = I[\mathcal{X}: \mathcal{X}|R] = I[\mathcal{X}: \mathcal{F}] - I[R; \mathcal{X}], \]  

(4.4)

so that penalizing \( \langle d(\mathcal{F}, r) \rangle \) here is equivalent to maximizing \( I[R; \mathcal{X}] \). In many papers, we speak only of maximizing “predictive power” \( I[R; \mathcal{X}] \), and not of minimizing an informational distortion, following Tishby, Pereira, and Bialek, 1999. And rather than plot predictive rate-distortion functions, we can plot predictive information curves (Still and Crutchfield, 2007; Creutzig and Sprekeler, 2008) in which we plot predictive power as a function of coding cost. Other times, we avoid trying to predict semi-infinite futures and focus on prediction of just the next time step, e.g. as in Creutzig, Globerson, and Tishby, 2009, so that expected distortion might take the form \( I[\mathcal{X}_0; X_0|R] \). These kinds of informational distortions can arise when one considers bet-hedging, for instance (Cover and Thomas, 2006, Ch. 6).

That having been said, we try to avoid making too many assumptions as to the form of the predictive distortion measure and even as to the form of the cost.

In what follows, we focus mainly on time series for which we already have an exact or approximate \( \epsilon \)-machine. This may seem counterintuitive, as predictive rate-distortion methods are often used to find approximate
\section*{4.1. Fundamentally foiled by the curse of dimensionality}

\epsilon\text{-}machines. However, we envision a set of experiments similar to those in Palmer et al., 2015, for instance, in which we test whether or not a biological sensory system is efficiently extracting predictive features of its input. By accurately computing predictive rate-distortion functions for these simpler time series, we greatly expand the range of stimuli that can be used to test the ability of such systems to efficiently extract lossy predictive features of this input. We still expect biological sensory systems to operate in the lossy regime, despite the simplicity of some of these artificial stimuli, purely because the organism may not “care” so much about such simple stimuli.

In such applications, accurate computation of predictive rate-distortion functions is paramount. A bad approximation might lead one to conclude that biological sensory systems are excellent predictive feature extractors of artificial stimuli when their use of memory post-adaptation is inefficient.

\subsection*{4.1 Fundamentally foiled by the curse of dimensionality}

The biggest obstacle to accurate computation comes from our desire to compress semi-infinite pasts and predict what we can about semi-infinite futures. Traditionally, we approach this difficulty by looking to pasts of length $M$ and futures of length $N$. When $M$ and $N$ are large enough, the corresponding computed predictive rate-distortion functions and optimal stochastic codebooks are tolerably close to those found in the $M,N \to \infty$ limit. However, the number of trajectories which need be considered increases exponentially with $M$ and $N$, which is computationally taxing. This is a \textit{curse of dimensionality}.

How large must $M$ and $N$ be? It certainly depends on the process under study. Complex time series– those with power law autocorrelation functions, for instance– will likely need very large $M$ and $N$ to achieve tolerable accuracy. Somewhat surprisingly, the curse of dimensionality rears its head for much simpler processes.

Consider the Even Process, whose \epsilon\text{-}machine is shown in Fig. 4.2(top). In the lossless limit, we achieve the maximum predictive power of $I[\hat{X}; \hat{X}]$, often called the excess entropy $E$. By using finite-length pasts of length $M$ and finite-length futures of $N$, we only capture a percentage of the full excess entropy, $I[X_{-N:0}; X_{0:M}]$. The infinity norm between approximated and true predictive information curves is exactly $E - I[X_{-N:0}; X_{0:M}]$, and so $I[X_{-N:0}; X_{0:M}]$ monitors the accuracy of finite-length estimates.

Fig. 4.2(bottom) shows that accurate estimates of the predictive information curve require $M, N \geq 8$. In a way, this is surprising, as the Even Process is generated by a simple two-state \epsilon\text{-}machine. Then again, the Even Process is infinite-order Markov.

Nor is this difficulty confined to the Even Process. When $W$ is diagonalizable with eigenvalues $\{\lambda_i\}$, Ara, Riechers, and Crutchfield, 2014 provides the closed-form expression:

$$E - I[X_{-N:0}; X_{0:M}] = \sum_{i: \lambda_i \neq 1} \frac{\lambda_i^M + \lambda_i^{N+1} - \lambda_i^{M+N+1}}{1 - \lambda_i} \langle \delta_\pi | W_{\lambda_i} | H(W^A) \rangle,$$

(4.5)
where $\langle \delta_\pi | W_{\lambda_i} | H(A) \rangle$ is a dot product between the eigenvector $\langle \delta_\pi | W_{\lambda_i} \rangle$ corresponding to eigenvalue $\lambda_i$ and a vector $H(A)$ of transition uncertainties out of each mixed state.\(^1\) Here, $\pi$ is the stationary state distribution, $\langle \delta_\pi \rangle$ is the probability vector over mixed states with full weight on the mixed state corresponding to the stationary state distribution, and $W_{\lambda_i}$ is the projection operator associated with $\lambda_i$. When $W$’s spectral gap $\gamma = 1 - \max_{i: \lambda_i \neq 1} \lambda_i$ is small, then $I[X_{-N:0}; X_{0:M}]$ necessarily asymptotes more slowly to $E$. When $\gamma$ is small, then (loosely speaking) we need $M, N \sim \log_{1-\gamma}(\epsilon/\gamma)$ in order to achieve a small error $\epsilon \sim E - I[X_{-N:0}; X_{0:M}] \ll 1$ for the predictive information function.

The difficulty can easily become extreme. For instance, altering the Even Process’s lone stochastic transition probability can increase its temporal correlations such that correctly calculating its information function requires massive compute resources. Thus, this curse of dimensionality is a critical concern even for finite-$C_\mu$ processes generated by finite HMMs.

### 4.2 Computing optimal lossy predictive features, regardless

Circumventing the curse of dimensionality in predictive rate-distortion, even given an accurate model of the process, requires an alternative approach to predictive rate distortion that leverages the structural information about

\(^1\)More precisely, each element of $H(A)$ is the entropy in the next observation given that one is currently in the corresponding mixed state.
a process captured by a maximally predictive model. The results in this section describe exactly how this structural information can be utilized. Lemma 5 equates lossy predictive features to lossy forward-time causal states. Theorem 2 shows that, for many predictive distortion measures, reverse-time causal states can replace semi-infinite futures. The predictive information bottleneck—compression of semi-infinite pasts to retain information about semi-infinite futures—can be recast as compression of forward-time causal states to retain information about reverse-time causal states, as a corollary to Theorem 2. The joint probability distribution of forward- and reverse-time causal states may seem somewhat elusive, but previous work has shown that this joint probability distribution can be obtained, or at least approximated, given the process’ model (Ellison, Mahoney, and Crutchfield, 2009; Ellison et al., 2011).

Lemma 5 is built on a simple observation: any predictive codebook can be recast as a codebook over forward-time causal states. Though the old and new codebooks have equivalent predictive distortions, the new codebook is either equivalent to or “smaller” than the old codebook. This observation is summarized in the remark below.

**Remark 2.** Given any codebook \( P(R|X_0) \), construct a new codebook by setting \( P(R|X_0 = x_0) \) to be \( P(R|S^+ = \epsilon^+(x_0)) \). This new codebook has equivalent predictive distortion, since predictive distortion depends only on \( p(r, \sigma^+) \):

\[
\mathbb{E}[d(x_0, r)] = \sum_{x_0, r} p(x_0, r) d(P(X_0|X_0 = x_0), P(X_0|R = r))
\]

\[
= \sum_{x_0, r} p(x_0, r) d(P(X_0|S^+ = \epsilon^+(x_0)), P(X_0|R = r))
\]

\[
= \sum_{\sigma^+, r} p(\sigma^+, r) d(P(X_0|S^+ = \sigma^+), P(X_0|R = r)).
\]

More importantly, this new codebook has equal or smaller rate, since

\[
I[R; X_0] = I[R; S^+] + I[R; X_0|S^+] \geq I[R; S^+] \quad (4.6)
\]

with equality when \( X_0 \rightarrow S^+ \rightarrow R \)– as is true for the new, but not necessarily for the old, codebook.

After the procedure implied by the remark, we could decrease not just the rate, but the number of predictive features by clustering together \( r, r' \) with equivalent future morphs. (In some sense, two predictive features with equivalent future morphs are just copies of the same object.) Then, the number of predictive features would never exceed the number of causal states, and the entropy \( H[R] \) would never exceed the statistical complexity. While potentially useful– some models have rate \( I[R; X_0] \) equivalent to the statistical complexity, despite their non-minimality, effectively by copying one or more causal states– this second operation is unnecessary for the statements below. However, it may prove useful when one considers different quantifications of resource constraints, e.g., constraints on \( H[R] \) with restriction to deterministic codebooks or the number of possible features.

To start, inspired by the previous finding that PIB recovers the forward-time causal states in the lossless limit (Still and Crutchfield, 2007; Still, Crutchfield, and Ellison, 2010), we argue that compressing either the past...
Chapter 4. Predictive rate-distortion

X₀ or forward-time causal states $S^+$ should yield the same lossy predictive features. In other words, lossy predictive features are lossy causal states, and vice versa.

**Lemma 5.** Compressing the past $X_0$ to minimize expected predictive distortion is equivalent to compressing the forward-time causal states $S^+$ to minimize expected predictive distortion.

**Proof.** A codebook which optimally compresses the past to achieve at most a distortion of $\mathbb{E}[d] \leq D$ minimizes rate $I[R; X_0]$, while a codebook which optimally compresses forward-time causal states to achieve at most a distortion of $\mathbb{E}[d] \leq D$ minimizes a rate $I[R; S^+]$. A codebook that optimally compresses forward-time causal states to minimize expected predictive distortion also optimally compresses pasts to minimize expected predictive distortion, as for such a codebook, the objective functions are equivalent: $I[R; X_0] = I[R; S^+]$. In the other direction, suppose that some codebook optimally compresses the past to minimize expected predictive distortion, in that it has the smallest possible rate $I[X_0; R]$ given distortion $\mathbb{E}[d] \leq D$. From Rem. 2, this codebook can be conceptualized as a codebook over forward-time causal states and has rate $I[R; X_0] = I[R; S^+]$. Hence, the corresponding codebook over forward-time causal states also optimally compresses forward-time causal states to minimize expected predictive distortion, as for such a codebook, the objective functions are equivalent. □

This lemma already provides a form of dimensionality reduction; semi-infinite pasts are replaced with the (potentially finite) forward-time causal states. Interestingly, in a nonprediction setting, Banerjee et al., 2004 states Lemma 5 in their Eq. (2.2) without conditions on the distortion measure. However, a distortion measure which is not of the form $d(P(X_0|X_0 = x_0), P(X_0|R = r))$ can still look like a predictive distortion measure, but actually incorporate potentially unnecessary information about the past, e.g. by penalizing the difference between an estimated and true future trajectories. In those situations, Lemma 5 may not apply, depending on the particular future trajectory estimator. Further dimensionality reduction may be possible depending on the predictive distortion, as in Brodú, 2011.

When the distortion measure takes a particular special form, then we can simplify the objective function further. Our inspiration comes from several papers (Crutchfield, Ellison, and Mahoney, 2009; Ellison, Mahoney, and Crutchfield, 2009; Crutchfield and Ellison, 2014) which showed that the mutual information between past and future is identical to the mutual information between forward and reverse-time causal states: $I[X_0; X_0] = I[S^+; S^-]$. In other words, forward-time causal states $S^+$ are the only features needed to predict the future as well as possible, and reverse-time causal states $S^-$ are features one can predict about the future.

**Theorem 2.** Compressing the past $X_0$ to minimize expected distortion of the future $X_0$ is equivalent to compressing the forward-time causal states $S^+$ to minimize expected distortion of reverse-time causal states $S^-$, if the predictive distortion measure is an $f$-divergence.

**Proof.** If $d$ is an $f$-divergence, then it takes the form

$$d(x_0, r) = \sum_{x_0} P(X_0 = x_0) P(X_0 = x_0) f \left( \frac{P(X_0 = x_0; X_0 = x_0)}{P(X_0 = x_0; X_0 = r)} \right)$$
4.2. Computing optimal lossy predictive features, regardless

for some $f$ Csisz et al., 1967. Reverse-time causal states $S^-$ are functions of the future $X_0$ that shield the future from the past and the representation, $R \rightarrow X_0 \rightarrow S^- \rightarrow X_0$, and so

$$P(X_0 = x_0 | X_0 = x_0) = P(X_0 = x_0 | S^- = \epsilon^-(x_0))$$

$x_0 \neq x_0$

and

$$P(X_0 = x_0 | R = r) = P(X_0 = x_0 | S^- = \epsilon^-(x_0))$$

$x_0 \neq x_0$

and so predictive distortions that are $f$-divergences can also be expressed as

$$d(x_0, r) = \sum_{x_0} P(X_0 = x_0 | S^- = \epsilon^-(x_0))$$

$x_0 \neq x_0$

and

$$P(S^- = \sigma^- | X_0 = x_0) f(\frac{P(S^- = \sigma^- | X_0 = x_0)}{P(S^- = \sigma^- | R = r)})$$

Given this fact, and Lemma 5, we recover the theorem statement.

Distortion measures that are not $f$-divergences, such as mean squared-error distortion measures, implicitly emphasize predicting one reverse-time causal state over another. The Kullback-Leibler divergence is an example of an $f$-divergence. It follows from previous discussion that predictive informational distortions treat all reverse-time causal states equally. Corollary 2 then follows, and it recasts the predictive information bottleneck in terms of forward- and reverse-time causal states.

**Corollary 1.** Compressing the past $X_0$ to retain information about the future $X_0$ is equivalent to compressing $S^+$ to retain information about $S^-$. Naturally, there is an equivalent version for the time-reversed setting in which past and future are swapped and the causal state sets are swapped. Also, any forward and reverse-time prescient statistics (Shalizi and Crutchfield, 2001) can be used in place of $S^+$ and $S^-$ in any of the statements above. (Prescient statistics are sufficient, but not necessarily minimal, statistics of prediction.)

These proofs follow almost directly from the definitions of forward- and reverse-time causal states. Variations or portions of Lemma 5, Theorem 2, and Corollary 1 are, hopefully, intuitive. That said, to the best of our knowledge, they are also new.

Throughout, we cavalierly manipulated semi-infinite pasts and futures and their conditional and joint probability distributions—e.g., $P(X_0 | X_0)$. This is mathematically suspect, since then many sums should be measure-theoretic integrals, our codebooks seemingly have an uncountable infinity of codewords, many probabilities vanish, and our distortion measures apparently divide 0 by 0. So, a more formal treatment would instead (i)
consider a series of objective functions that compress finite-length pasts to retain information about finite-length futures for a large number of lengths, giving finite codebooks and finite sequence probabilities at each length, (ii) trivially adapt the proofs of Lemma 5, Theorem 2 and Corollary 1 for these objective functions with finite-time causal states, and (iii) take the limit as those lengths go to infinity; e.g., as in Crutchfield and Ellison, 2014. As long as the finite-time forward- and reverse-time causal states limit to their infinite-length counterparts, which seems to be the case for ergodic stationary processes but not for nonergodic processes, one recovers Lemma 5, Theorem 2 and Corollary 1. We leave the task of an expanded measure-theoretic development to others.

These statements nominally reduce the numerically intractable problem of clustering in the infinite-dimensional sequence space \((X_0, X_0)\) to the potentially tractable one of clustering in \((S^+, S^-)\). This is beneficial when a process’s causal state set is finite. However, many processes have an uncountable infinity of forward-time causal states or reverse-time causal states (Crutchfield, 1994; Upper, 1997). Is Theorem 2 useless in these cases? Not necessarily—predictive rate-distortion functions can be approximated to any desired accuracy by a finite or countable \(\epsilon\)-machine. Furthermore, only a finite \(\epsilon\)-machine can be justified if models are inferred from finite data, a point that we explore in greater detail in Sec. 4.4.

An interesting open problem is to understand how approximations of a process’ minimal maximally predictive model map to approximations of its predictive rate-distortion function. We try to make some headway by studying a few example processes below.

### 4.3 Examples

Theorem 2 suggests a new objective function to define lossy predictive features and predictive rate-distortion functions. In this section, we compare the results of an algorithm suggested by Corollary 1 to results of more commonly used PIB algorithms for some simple stochastic processes to investigate when and why moving to bidirectional model space might be useful.

Traditional PIB algorithms cluster finite-length pasts to retain information about finite-length futures. For simplicity’s sake, we assume that lengths of pasts and futures are both \(L\). These algorithms find \(P(R|\overleftarrow{X}_L)\) which maximizes

\[
L_\beta = I[R; \overleftarrow{X}_L] - \beta^{-1}I[\overleftarrow{X}_L; R],
\]

(4.7)

and vary the Lagrange multiplier \(\beta\) to achieve different distortions. We refer to such algorithms as *optimal causal filtering* (OCF). Using Corollary 1, we can instead search for a codebook \(P(R|S^+)\) which maximizes

\[
L_\beta = I[R; S^-] - \beta^{-1}I[S^+; R],
\]

(4.8)

and again vary the Lagrange multiplier \(\beta\) to achieve different distortions. We refer to algorithms which try to maximize this objective function as *causal information bottleneck* (CIB) algorithms. At large enough \(L\), the approximated predictive features become indistinguishable from the true predictive features. However, some of the examples here give a rather sober
illustration of the substantial errors that can arise when operating at finite $L$ for surprisingly simple processes.

We calculate solutions to both objective functions following Tishby, Pereira, and Bialek, 1999. For example, given $P(S^+, S^-)$, then, one solves for the $P(R|S^+)$ that maximizes the objective function in Eq. 4.8 at each $\beta$ by iterating the dynamical system:

$$P_t(r|\sigma^+) = \frac{P_{t-1}(r)}{Z_t(\sigma^+, \beta)} e^{-\beta D_{KL}[P(\sigma^-|\sigma^+)||P_{t-1}(\sigma^-|r)]}$$  \hfill (4.9)

$$P_t(r) = \sum_{\sigma^+} P_t(r|\sigma^+) P(\sigma^+)$$  \hfill (4.10)

$$P_t(\sigma^-|r) = \sum_{\sigma^+} P(\sigma^-|\sigma^+) P_t(\sigma^+|r),$$  \hfill (4.11)

where $Z_t(\sigma^+, \beta)$ is the normalization constant for $P_t(r|\sigma^+)$. Iterating Eqs. (4.9) and (4.11) at fixed $\beta$ gives (i) one point on the function $(R_\beta, D_\beta)$ and (ii) the explicit optimal lossy predictive features $P(R|S^+)$. We used a similar procedure to calculate finite-$L$ approximations to information functions, but $\sigma^+$ and $\sigma^-$ were replaced by $x_{-L:0}$ and $x_{0:L}$, which were then replaced by finite-time causal states $S^T_{L,0}$ and $S^T_{0,L}$ using a finite-time variant of Corollary 1. The joint probability distribution of these finite-time causal states was calculated exactly by (i) calculating sequence distributions of length $2L$ directly from the $\epsilon$-machine transition matrices and (ii) clustering these into finite-time causal states using the equivalence relation described in App.A. This procedure avoids the complications of finite sequence samples. As a result, differences between the algorithms are entirely due to a difference in the objective function.

We display calculations in two ways. The first is the information function, a rate-distortion function that graphs the code rate $I[X; R]$ versus the distortion $I[X; R] - I[X; R_\epsilon]$.

The second is a feature curve of code rate $I[X; R]$ versus inverse temperature $\beta$. We recall that at zero temperature ($\beta \to \infty$) the code rate $I[X; R] = C^+_\mu$ and the forward-time causal states are recovered: $R \to S^+$. At infinite temperature ($\beta = 0$) there is only a single state that provides no shielding and so the information distortion limits to $I[X; R] = E$. As suggested earlier, these extremes are useful references for monitoring convergence as they are the $L_\infty$ norm between true and approximate information functions.

For each $\beta$, we chose 500 random initial $P_0(r|\sigma^+)$, iterated Eqs. (4.9)-(4.11) 300 times, and recorded the solution with the largest $L_\beta$. This procedure finds local maxima of $L_\beta$, but does not necessarily find global maxima. Thus, if the resulting information function was nonmonotonic, we increased the number of randomly chosen initial $P_0(r|\sigma^+)$ to 5000, increased the number of iterations to 500, and repeated the calculations. This brute force approach to the nonconvexity of the objective function was feasible here only due to analyzing processes with small $\epsilon$-machines. Even so, the estimates might include suboptimal solutions in the lossier regime. A more sophisticated approach would leverage other results, e.g. those of Parker, 2008 and elsewhere, as the informational distortion is the excess entropy less the predictable information captured.
Gedeon, and Dimitrov, 2002; Parker and Gedeon, 2004; Parker, Dimitrov, and Gedeon, 2010, to move carefully from high-\(\beta\) to low-\(\beta\) solutions.

Note that in contrast with deterministic annealing procedures that start at low \(\beta\) (high temperature) and add codewords to expand the codebook as necessary, we can also start at large \(\beta\) with a codebook with codewords \(S^+\) and decrease \(\beta\), allowing the representation to naturally reduce its size. This is usually “naive” (Elidan and Friedman, 2003) due to the large number of local maxima of \(L_\beta\), but here, we know the zero-temperature result beforehand. More importantly, we are usually searching for the lossless predictive features at large \(\beta\), but here, we are asking different questions. Of course, we could also start at low \(\beta\) and increase \(\beta\). The key difference between the algorithm suggested by Corollary 1 and traditional predictive information bottleneck algorithms is not the algorithm itself, but the joint probability distribution of compressed and relevant variables.

Section 4.3.1 gives conditions on a process which guarantee that its information functions can be accurately calculated without first having a maximally-predictive model in hand. Section 4.3.2 describes several processes that have first-order phase transitions in their feature curves at \(\beta = 1\). Section 4.3.3 describes how information functions and feature curves can change nontrivially under time reversal. Finally, Sec. 4.3.4 shows how predictive features describe predictive “macrostates” for the process generated the symbolic dynamics of the chaotic Tent Map.

### 4.3.1 Unhidden and Almost Unhidden Processes

Predictive information bottleneck algorithms that cluster pasts of length \(M \geq 1\) to retain information about futures of length \(N \geq 1\) calculate accurate information functions when \(E(M, N) \approx E\). Such algorithms work exactly on order-\(R\) Markov processes when \(M, N \geq R\), since \(E(R, R) = E\). However, there are many processes that are “almost” order-\(R\) Markov with small \(R\) for which traditional algorithms should work quite well.

The quality of a process’s approximation can be monitored by the convergence error \(E - E(M, N)\), which is controlled by the elusive information (Ara, Riechers, and Crutchfield, 2014)

\[
\sigma_\mu(L) := I[X_0; X_L; |X_0:L].
\]  

To see this, we apply the mutual information chain rule repeatedly:

\[
E = I[X_0; X_0] = I[X_0; X_{0:N-1}] + \sigma_\mu(N) = E(M, N) + I[X_{-M-1}; X_{0:N-1}|X_{-M-1};0] + \sigma_\mu(N).
\]

The last mutual information is difficult to interpret, but easy to bound:

\[
I[X_{-M-1}; X_{0:N-1}|X_{-M-1};0] \leq I[X_{-M-1}; X_0; X_{-M-1};0] = \sigma_\mu(M),
\]
And so, the convergence error is upper-bounded by the elusive information:

$$0 \leq E - E(M, N) \leq \sigma_\mu(N) + \sigma_\mu(M).$$

The inequality of Eq. (4.13) suggests that, as far as accuracy is concerned, if a process has a small $\sigma_\mu(L)$ relative to its $E$ for some reasonably small $L$, then sequences are effective states. This translates into the conclusion that for this class of process calculating information functions by first moving to causal state space is unnecessary.

Let’s test this intuition. The prototypical example with $\sigma_\mu(1) = 0$ is the Golden Mean Process, whose HMM is shown in Fig. 4.3(top). It is order-1 Markov, so OCF with $L = 1$ is provably equivalent to CIB, illustrating one side of the intuition.

A more discerning test is an infinite-order Markov process with small $\sigma_\mu$. One such process is the Simple Nonunifilar Source (SNS) whose (nonunifilar) HMM is shown in Fig. 4.3(bottom). As anticipated, Fig. 4.4(top) shows that OCF with $L = 1$ and CIB yield very similar information functions at low code rate and low $\beta$. In fact, many of SNS’s statistics are well approximated by the Golden Mean HMM.

The feature curve in Fig. 4.4(bottom) reveals a slightly more nuanced story. The SNS is highly cryptic, in that it has a much larger $C_\mu$ than $E$. As a result, OCF with $L = 1$ approximates $E$ quite well but underestimates $C_\mu$, replacing an (infinite) number of feature-discover transitions with a single transition.

This particular type of error—missing predictive features—only matters for predicting the SNS when low distortion is desired. Nonetheless, it is important to remember that the process implied by OCF with $L = 1$—the Golden Mean Process—is not the SNS. The Golden Mean Process is an order-1 Markov process. The SNS HMM is nonunifilar and generates an...
Figure 4.4: Simple Nonunifilar Source: (Top panel) Information function: coding cost versus distortion. (Bottom panel) Feature curve: coding cost as a function of inverse temperature $\beta$. (Blue solid line, circles) CIB with a 10-state approximate $\epsilon$-machine. (Green dashed line, crosses) OCF at $L = 1$. 
infinite-order Markov process and so provides a classic example (Crutchfield, 1994) of how difficult it can be to exactly calculate information measures of stochastic processes.

Be aware that CIB cannot be directly applied to analyze the SNS, since the latter’s causal state space is countably infinite; see Marzen and Crutchfield, 2014b. Instead, we used finite-time causal states with finite past and future lengths and with the state probability distribution given in Marzen and Crutchfield, 2014b. Here, we used $M, N = 10$, effectively approximating the SNS as an order-10 Markov process.

4.3.2 First-order Phase Transitions at $\beta = 1$

Feature curves have discontinuous jumps (“first-order phase transitions”) or are nondifferentiable (“second-order phase transitions”) at critical temperatures when new features or new lossy causal states are discovered. The effective dimension of the codebook changes at these transitions. Symmetry breaking plays a key role in identifying the type and temperature of phase transitions in constrained optimization (Rose, 1994; Parker, Dimitrov, and Gedeon, 2010). Using the infinite-order Markov Even Process described earlier, CIB allows us to explore in greater detail why and when first-order phase transitions occur at $\beta = 1$ in feature curves.

There are important qualitative differences between information functions and feature curves obtained via CIB and via OCF for the Even Process. First, as Fig. 4.5(top) shows, the Even Process CIB information function is a simple straight line, whereas those obtained from OCF are curved and substantially overestimate the coding cost. Second, as Fig. 4.5(bottom) shows, the CIB feature curve is discontinuous at $\beta = 1$, indicating a single first-order phase transition and the discovery of highly predictive states. In contrast, OCF functions miss that key transition and incorrectly suggest several phase transitions at larger $\beta$s.

The first result is notable, as Still, Crutchfield, and Ellison, 2010 proposed that the curvature of OCF information functions define natural scales of predictive coarse-graining. In this interpretation, linear information functions imply that the Even Process has no such intermediate natural scales. And, there are good reasons for this.

So, why does the Even Process exhibit a straight line? Recall that the Even Process’s recurrent forward-time causal states code for whether or not one just saw an even number of 1’s (state A) or an odd number of 1’s (state B) since the last 0. Its recurrent reverse-time causal states (Fig. 2 in Ellison, Mahoney, and Crutchfield, 2009) capture whether or not one will see an even number of 1’s until the next 0 or an odd number of 1’s until the next 0. Since one only sees an even number of 1’s between successive 0’s, knowing the forward-time causal state uniquely determines the reverse-time causal state and vice versa. The Even Process’ forward causal-state distribution is $P(S^+) = (2/3 \ 1/3)$ and the conditional distribution of forward and reverse-time causal states is:

$$P(S^-|S^+) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$
Thus, there is an invertible transformation between $S^+$ and $S^-$, implying:

$$I[R; S^+] = I[R; S^-].$$

(4.14)

And so, we directly calculate the information function as

$$R(I_0) = \min_{I[R; S^-] \geq I_0} I[R; S^+] = \min_{I[R; S^-] \geq I_0} I[R; S^-] = I_0,$$

for all $I_0 \leq E$. Similar arguments hold for periodic process as described in Still and Crutchfield, 2007; Still, Crutchfield, and Ellison, 2010 and for general cyclic (noisy periodic) processes as well. However, periodic processes are finite-order Markov, whereas the infinite Markov-order Even Process hides its deterministic relationship between prediction and retrodiction underneath a layer of stochasticity. This suggests that the bidirectional machine’s switching maps (Ellison, Mahoney, and Crutchfield, 2009) are key to the shape of information functions.

The Even Process’s feature curve in Fig. 4.5(bottom) shows a first-order phase transition at $\beta = 1$. Similar to periodic and cyclic processes, its lossy causal states are all-or-nothing. However, Eq. (4.14) gives:

$$L_\beta = (1 - \beta^2)I[R; S^+] .$$

Recall that $0 \leq I[R; S^+] \leq C_\mu$. For $\beta < 1$, on the one hand, maximizing $L_\beta$ requires minimizing $I[R; S^+]$, so the optimal lossy model is a biased coin approximation of the Even Process—a single-state HMM. For $\beta > 1$, on the other, maximizing $L_\beta$ requires maximizing $I[R; S^+]$, so the optimal lossy features are the causal states $A$ and $B$ themselves. At $\beta = 1$, though, $L_\beta = 0$, and any representation $R$ of the forward-time causal states $S^+$ is optimal. In sum, the discontinuity of coding cost $I[R; S^+]$ as a function of $\beta$ corresponds to a first-order phase transition and the critical inverse temperature is $\beta = 1$.

Both causal states in the Even Process are unusually predictive features: any increase in memory of such causal states is accompanied by a proportionate increase in predictive power. These states are associated with a one-to-one (switching) map between a forward-time and reverse-time causal state. In principle, such states should be the first features extracted by any predictive rate-distortion algorithm. More generally, when the joint probability distribution of forward- and reverse-time causal states can be permuted into diagonal block-matrix form, there should be a first-order phase transition at $\beta = 1$ with one new codeword for each of the blocks.

Many processes do not have probability distributions over causal states that can be permuted, even approximately, into a diagonal block-matrix form; e.g., most of those described in Marzen and Crutchfield, 2014a and Chapter 5. However, we suspect that diagonal block-matrix forms for $P(S^+, S^-)$ might be relatively common in the highly structured processes generated by low entropy-rate deterministic chaos, as such systems often have many irreducible forbidden words. Restrictions on the support of the sequence distribution easily yield blocks in the joint probability distribution of forward-and reverse-time causal states.

For example, the Even Process forbids words with an odd number of 1s, which is expressed by its irreducible forbidden word list $F = \{01^{2k+1}0 : k =$

...
4.3. Examples

Figure 4.5: Even Process analyzed with CIB (solid line, blue circles) and with OCF (dashed lines, colored crosses) at various values of $M = N = L$: (right to left) $L = 2$ (green), $L = 3$ (red), $L = 4$ (light blue), and $L = 5$ (purple). (top) Information functions. (bottom) Feature curves. At $\beta = 1$, CIB functions transition from approximating the Even Process as IID (biased coin flip) to identifying both causal states.
0, 1, 2, . . .}. Its causal states group pasts that end with an even (state $A$) or odd (state $B$) number of 1s since the last 0. Given the Even Process’ forbidden words $F$, sequences following from state $A$ must start with an even number of ones before the next 0 and those from state $B$ must start with an odd number of ones before the next 0. The restricted support of the Even Process’ sequence distribution therefore gives its causal states substantial predictive power.

Moreover, many natural processes are produced by deterministic chaotic maps with added noise (Crutchfield, Farmer, and Huberman, 1982). Such processes may also have $P(S^+, S^-)$ in nearly diagonal block-matrix form. These joint probability distributions might be associated with sharp second-order phase transitions.

However, numerical results for the “four-blob” problem studied in Parker, Dimitrov, and Gedeon, 2010 suggest the contrary. The joint probability distribution of compressed and relevant variables is “a discretization of a mixture of four well-separated Gaussians” and has a nearly diagonal block-matrix form, with each block corresponding to one of the four blobs. If the joint probability distribution were exactly block diagonal—e.g., from a truncated mixture of Gaussians model—then the information function would be linear and the feature curve would exhibit a single first-order phase transition at $\beta = 1$ from the above arguments. The information function for the four-blob problem looks linear; see Fig. 5 of Parker, Dimitrov, and Gedeon, 2010. The feature curve (Fig. 4, there) is entirely different from the feature curves that we expect from our earlier analysis of the Even Process. Differences in the off-diagonal block-matrix structure allowed the annealing algorithm to discriminate between the nearly equivalent matrix blocks, so that there are three phase transitions to identify each of the four blobs. Moreover, none of the phase transitions are sharp. So, perhaps the sharpness of phase transitions in feature curves of noisy chaotic maps might have a singular noiseless limit, as is often true for information measures, e.g. Marzen and Crutchfield, 2014a.

### 4.3.3 Temporal Asymmetry in Lossy Prediction

As described in Crutchfield, Ellison, and Mahoney, 2009; Ellison, Mahoney, and Crutchfield, 2009, the resources required to losslessly predict a process can change markedly under time reversal. The prototype example is the Random Insertion Process (RIP), shown in Fig. 4.6. Its bidirectional machine is known analytically. Therefore, we know the joint $P(S^+, S^-)$ via $P(S^+) = (2/5 1/5 2/5)$ and:

$$
P(S^-|S^+) = \begin{pmatrix} 0 & 1 & 0 & 1/2 \\ 0 & 1/2 & 1/2 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}.
$$

There are three forward-time causal states and four reverse-time causal states, and the forward-time statistical complexity and reverse-time statistical complexity are unequal, making the RIP causally irreversible. For instance, $C_\mu^+ \approx 1.8$ bits and $C_\mu^- \approx 1.5$ bits, even though the excess entropy $E \approx 1.24$ bits is by definition time-reversal invariant.
4.3. Examples

Figure 4.6: Random Insertion Process (RIP): (Top) Forward-time $\epsilon$-machine. (Bottom) Reverse-time $\epsilon$-machine.
Figure 4.7: Random Insertion Process (RIP) Information Functions: RIP is a causally irreversible process: $C_{\mu}^+ < C_{\nu}^-$. There are more causal states in reverse time than forward time, leading to more kinks in the reverse-time process’ information function (bottom) than in the forward-time process’ information function (top). Legend as in previous figure: (solid line, blue circles) CIB function and (dashed lines, colored crosses) OCF at various sequence lengths.
4.3. Examples

Figure 4.8: Random Insertion Process (RIP) Feature Curves: Having more causal states in reverse time than forward time leads to more phase transitions in the reverse-time process’ feature curve (bottom) than in the forward-time process’ feature curve (top). Legend as in previous figure.
However, it could be that the lossy causal states are somehow more robust to time reversal than the (lossless) causal states themselves. We now investigate the difference in RIP’s information and feature curves under time reversal. Figure 4.7 shows information functions for the forward-time and reverse-time processes. Despite RIP’s causal irreversibility, information functions look similar until informational distortions of less than 0.1 bits. RIP’s temporal correlations are sufficiently long-ranged so as to put OCF with $L \leq 5$ at a significant disadvantage relative to CIB, as the differences in the information functions demonstrate. OCF greatly underestimates $E$ by about 30% and both underestimates and overestimates the correct $C_\mu$.

The RIP feature curves in Fig. 4.8 reveal a similar story in that OCF fails to asymptote to the correct $C_\mu$ for any $L \leq 5$ in either forward or reverse time. Unlike the information functions, though, feature curves reveal temporal asymmetry in the RIP even in the lossy (low $\beta$) regime.

Both forward and reverse-time feature curves show a first-order phase transition at $\beta = 1$, at which point the forward-time causal state $C$ and the reverse-time causal state $D$ are added to the codebook, illustrating the argument of Sec. 4.3.2. (Forward-time causal state $C$ and reverse-time causal state $D$ are equivalent to the same bidirectional causal state $C/D$ in RIP’s bidirectional $\epsilon$-machine. See Fig. 2 of Crutchfield, Ellison, and Mahoney, 2009.) This common bidirectional causal state is the main source of similarity in the information functions of Fig. 4.7.

Both feature curves also show phase transitions at $\beta = 2$, but similarities end there. The forward-time feature curve shows a first-order phase transition at $\beta = 2$, at which point both remaining forward-time causal states $A$ and $B$ are added to the codebook. The reverse-time feature curve has what looks to be a sharp second-order phase transition at $\beta = 2$, at which point the reverse-time causal state $F$ is added to the codebook. The remaining two reverse-time causal states, $E$ and $G$, are finally added to the codebook at $\beta = 5$. We leave solving for the critical temperatures and confirming the phase transition order using a bifurcation discriminator (Parker, Gedeon, and Dimitrov, 2002) to the future.

### 4.3.4 Predictive Hierarchy in a Dynamical System

Up to this point, the emphasis was analyzing selected prototype infinite Markov-order processes to illustrate the differences between CIB and OCF. In the following, instead we apply CIB and OCF to gain insight into a nominally more complicated process—a one-dimensional chaotic map of the unit interval—in which we emphasize the predictive features detected. We consider the symbolic dynamics of the Tent Map at the Misiurewicz parameter $a = (\sqrt[3]{9 + \sqrt{57}} + \sqrt[3]{9 - \sqrt{57}}) / \sqrt[3]{9}$, studied in James, Burke, and Crutchfield, 2014. Figure 4.9 gives both the Tent Map and the analytically derived $\epsilon$-machine for its symbolic dynamics, from there. The latter reveals that the symbolic dynamic process is infinite-order Markov. The bidirectional $\epsilon$-machine at this parameter setting is also known. Hence, one can directly calculate information functions as described in Sec. 4.3.

From Fig. 4.10’s information functions, one easily gleams natural coarse-grainings, scales at which there is new structure, from the functions’ steep regions. As is typically true, the steepest part of the predictive information
4.3. Examples

Figure 4.9: Symbolic dynamics of the Tent Map at the Misiurewicz parameter $a$. (top) The map iterates points $x_n$ in the unit interval $[0, 1]$ according to $x_{n+1} = \frac{a}{2}(1 - 2|x_n - \frac{1}{2}|)$, with $x_0 \in [0, 1]$. The symbolic dynamics translates the sequence $x_0, x_1, x_2, \ldots$ of real values to a 0 when $x_n \in [0, \frac{1}{2})$ and to a 1 when $x_n \in [\frac{1}{2}, 1]$. (bottom) Calculations described elsewhere James, Burke, and Crutchfield, 2014 yield the $\epsilon$-machine shown. (Reproduced from Ref. James, Burke, and Crutchfield, 2014 with permission.)
function is found at very low rates and high distortions. Though the information function of Fig. 4.10(top) is fairly smooth, the feature curve (Fig. 4.10(bottom)) reveals phase transitions where the feature space expands a lossier causal state into two distinct representations.

To appreciate the changes in underlying predictive features as a function of inverse temperature, Fig. 4.11 shows the probability distribution $P(S^+ | R)$ over causal states given each compressed variable—the features. What we learn from such phase transitions is that some causal states are more important than others and that the most important ones are not necessarily intuitive. As we move from lossy to lossless ($\beta \to \infty$) predictive features, we add forward-time causal states to the representation in the order $A, B, C$, and finally $D$. The implication is that $A$ is more predictive than $B$, which is more predictive than $C$, which is more predictive than $D$. Note that this predictive hierarchy is not the same as a “stochastic hierarchy” in which one prefers causal states with smaller $H[X_0|S^+ = \sigma^+]$. The latter is equivalent to an ordering based on correctly predicting only one time step into the future. Such a hierarchy privileges causal state $C$ over $B$ based on the transition probabilities shown in Fig. 4.9(bottom), in contrast to how CIB orders them.

### 4.4 Coarse-grained mixed states as nearly lossless predictive features

The previous sections focused on processes with finite or countably infinite $\epsilon$-machines. However, most processes have uncountably infinite $\epsilon$-machines. In this more general scenario, we can place new bounds on memory–prediction tradeoffs in the low distortion limit by coarse-graining mixed states—the probability distribution over states in a generative model given a particular past (Blackwell, 1957). These new bounds give a new operational interpretation to the box-counting and information dimension of the mixed state presentation.

We consider a discrete-time, discrete-state stochastic process $\mathcal{P}$ generated by a Hidden Markov model $\mathcal{G}$, which comes equipped with underlying states $g$ and labeled transition matrices $T_{g,g'} = P(G_{t+1} = g', X_{t+1} = x | G_t = g)$ (Rabiner and Juang, 1986). There are an infinite number of different Hidden Markov models that can generate the same process (Gilbert, 1959), so we specify here that $\mathcal{G}$ is the minimal generative model—that is, the generative model with the minimal number of hidden states. This is not necessarily the same as the generative model with minimal generative complexity (Löhr, 2009).

For reasons that become clear in a later section, we are interested in the block entropy $H(L) = H[X_0: L]$. In particular, the entropy rate $h_{\mu} = \lim_{L \to \infty} \frac{H(L)}{L}$ quantifies the intrinsic “randomness” of the process, while the excess entropy $E = \lim_{L \to \infty} H(L) - h_{\mu} L$ quantifies its intrinsic “predictability”, though see the terminology of Bialek, Nemenman, and Tishby, 2001b. Finite-length entropy rate estimates $h_{\mu}(L) = H[X_0 | X_{-L:0}]$ provide increasingly better approximations to the true entropy rate as $L$ grows large, while
4.4. Coarse-grained mixed states as nearly lossless predictive features

Figure 4.10: Rate distortion analysis for symbolic dynamics of the Tent Map at the Misiurewicz parameter $a$ given in the text. (top) Information functions. (bottom) Feature curves. Comparing CIB (solid line, blue circles) and OCF (dashed lines, colored crosses) at several values of $L$. Legend same as previous.
finite-length excess entropy estimates

\[ \mathbf{E}(L) = H(L) - h_\mu L = \sum_{l=0}^{L-1} h_\mu(l) - h_\mu \] (4.15)

tend to the true excess entropy \( \mathbf{E} \) as \( L \) grows large (Crutchfield and Feldman, 2003). Here, we consider only processes with finite \( \mathbf{E} \).

Predictive features \( \mathcal{R} \) from some alphabet \( \mathcal{F} \) are formed by compressing the past of the process, \( X_{0:L} \), in ways that implicitly retain information about the future of the process, \( X_{0:L} \). A “predictive distortion” quantifies the predictability lost after such a coarse-graining,

\[ d(\mathcal{R}) := I[X_0; X_{0:L} \mid \mathcal{R}] = \mathbf{E} - I[\mathcal{R}; X_{0:L}], \] (4.16)

This distortion achieves a maximum value of \( \mathbf{E} \) when \( \mathcal{R} \) captures no information from the past that could be used for prediction. Distortion can be made 0 trivially by setting \( \mathcal{R} \) to \( X_{0:L} \). In this manuscript, we quantify the cost of a larger feature space in two ways: the number of predictive features \( |\mathcal{F}| \), and their coding cost \( H[\mathcal{R}] \) (Cover and Thomas, 2006).

Ideally, we would identify the minimal number of predictive features \( |\mathcal{F}| \) or the coding cost \( H[\mathcal{R}] \) required to achieve at least a given level of predictive distortion, \( d \). This is almost always a difficult optimization problem to actually solve. However, we can place upper bounds on \( |\mathcal{F}| \) and \( H[\mathcal{R}] \) by constructing two sets of sub-optimal predictive features that achieve a predictive distortion \( d \).

We start by reminding ourselves of the optimal solution in the limit that \( d = 0 \)– the causal states \( \mathcal{S} \) (Crutchfield and Young, 1989a; Shalizi and Crutchfield, 2001; Still and Crutchfield, 2007; Still, Crutchfield, and Ellison, 2010). Causal states can be defined by calling two pasts, \( x_{0:L} \) and \( x'_{0:L} \), “equivalent” when \( P(X_{0:L} \mid X_0 = x_{0:L}) = P(X_{0:L} \mid X_0 = x'_{0:L}) \). One can then form a model from the set of equivalence classes called the \( \epsilon \)-machine, alternatively viewed as the minimal unifilar Hidden Markov model capable of generating the process (Travers and Crutchfield, 2014a). See App. A. Though the mechanics of working with causal states can become rather complicated, the essential idea is that causal states are designed to capture everything about the past relevant to predicting the future, and only that information. In the lossless limit, when \( d = 0 \), one cannot find predictive representations that achieve \( |\mathcal{F}| \) smaller than \( |\mathcal{S}| \) or \( H[\mathcal{R}] \) smaller than \( C_\mu \). Similarly, no optimal lossy predictive representation will ever find \( |\mathcal{F}| > |\mathcal{S}| \) or \( H[\mathcal{R}] \geq C_\mu \). When the number of causal states is infinite or statistical complexity is infinite, as is typical, these bounds are quite useless; but they provide a useful calibration for the proposed feature sets below otherwise.

The first of these sub-optimal predictive feature sets is recall of only the last \( L \) symbols of the past, so that \( \mathcal{F} = A^L \). Here, the features are literally pasts of length \( L \); this feature set can be thought of as constructing an order-\( L \) Markov model of the process.\(^3\) The implied predictive distortion is \( I[X_0; X_{0:L} \mid X_{0:L}] = \mathbf{E} - E(L) \) (Ara, James, and Crutchfield, 2016),

\(^3\)One can construct a better feature set just by applying the causal state equivalence relation to these length-\( L \) pasts. We avoid doing this here because the size of this feature set is more difficult to analyze generically, and because the causal state equivalence relation applied to length-\( L \) pasts often implies no coarse-graining.
while the number of features is the number of length-\(L\) words with nonzero probability, and their entropy is \(H(L)\). When \(L\) is large, \(|F| \sim 2^{h_0 L}\) and \(H(L) \approx h_\mu L\) where \(h_0\) is the topological entropy and \(h_\mu\) the entropy rate. Generally, \(h_\mu(L)\) converges exponentially quickly to the true entropy rate \(h_\mu\) for stochastic processes generated by finite Hidden Markov models; see Travers, 2014 and references therein. Then, at large \(L\), \(h_\mu(L) - h_\mu \sim K e^{-\lambda L}\) in the large \(L\) limit. From Eq. 4.15, we see that this rate of convergence implies an exponential rate of decay for \(E - E(L) \sim K' e^{-\lambda L}\).

In sum, this first set of predictive features—effectively, the construction of order-\(L\) Markov models—yields an algebraic tradeoff between the size of the feature set and the predictive distortion,

\[|F| \sim \left(\frac{1}{d}\right)^{h_0/\lambda},\]  

(4.17)

and a logarithmic tradeoff between the entropy of the features and the predictive distortion,

\[H[R] \sim \frac{h_\mu}{\lambda} \log_2 \left(\frac{1}{d}\right).\]  

(4.18)

This is clearly a sub-optimal feature set when the process in question has only a finite number of causal states; in that case, \(|F|\) optimally saturates at \(|S|\) and \(H[R]\) optimally saturates at \(C_\mu\), achieving \(d = 0\) in that limit. No prior work has indicated whether or not these are good feature sets for the more typical case that \(C_\mu\) is infinite.

The second of these sub-optimal predictive feature sets is based on a coarse-graining of the mixed state simplex. To explain this, we first review the definition of mixed states. As first described in Blackwell, 1957, these mixed states are probability distributions over the internal states of the generative model \(G\) given observations, \(P(G_0|X_{-L:0} = x_{-L:0})\). Transient mixed states are those mixed states which exist at all \(L\), while recurrent mixed states are the sets of mixed states in the limit that \(L\) is infinite. These recurrent mixed states exactly correspond to causal states \(S\) (Ellison, Mahoney, and Crutchfield, 2009). When statistical complexity is infinite, recurrent mixed states often show a fractal pattern on the simplex. See Fig. 4.12.

The simplex is partitioned into cubes of size \(\epsilon\), and each non-empty cube is taken to be a predictive feature in our representation. For clarity, we now use \(R(\epsilon)\) to indicate the predictive features obtained from an \(\epsilon\)-partition of mixed state space.

**Lemma 6.** When \(\epsilon\) is small, then predictive distortion \(d(R(\epsilon))\) scales at most as \(\sim \epsilon\) for processes with uncountably infinite \(\epsilon\)-machines.

**Proof.** For reasons that become clear later, we define

\[d_L(R) := I[X_0; \overrightarrow{X}^L | R] = H[\overrightarrow{X}^L | R] - H[\overrightarrow{X}^L | X_0],\]  

(4.19)

where \(\lim_{L \to \infty} d_L(R) = d(R)\). Let \(\pi(r)\) be the invariant probability distribution over \(\epsilon\)-boxes, and let \(\pi(y|r)\) be the probability measure over mixed states in that \(\epsilon\)-box. Then,

\[d_L(R(\epsilon)) = \sum_r \pi(r) \left(H[\overrightarrow{X}^L | R(\epsilon) = r] - \int_y d\pi(y|r) H[\overrightarrow{X}^L | G \sim y]\right)\]  

(4.20)
where
\[ P(\overrightarrow{X} | \mathcal{R}(\epsilon) = r) = \int_{y} d\pi(y|r)P(\overrightarrow{X} | \mathcal{G} \sim y). \] (4.21)

To place an upper bound on \(d_L(\mathcal{R})\) in terms of \(\epsilon\), we note that for any two mixed states \(y\) and \(y'\) in the same partition,
\[ ||y - y'||_1 \leq \sqrt{|\mathcal{G}|} \epsilon, \] (4.22)
the length of the longest diagonal in a hypercube of dimension \(|\mathcal{G}|\), by construction. Hence,
\[ \|P(\overrightarrow{X} | \mathcal{G} \sim y) - P(\overrightarrow{X} | \mathcal{G} \sim y')\|_{TV} \leq \sum_{i=1}^{|\mathcal{G}|} |y_i - y'_i| \] (4.23)
but, from earlier, this is just
\[ \|P(\overrightarrow{X} | \mathcal{G} \sim y) - P(\overrightarrow{X} | \mathcal{G} \sim y')\|_{TV} \leq ||y - y'||_1 \leq \sqrt{|\mathcal{G}|} \epsilon. \] (4.24)
(Often in the literature, \(\| \cdot \|_{TV}\) is defined as \(\frac{1}{2}\) of the quantity used here.)

From this, we can similarly conclude that
\[ \|P(\overrightarrow{X} | \mathcal{R}(\epsilon) = r) - P(\overrightarrow{X} | \mathcal{G} \sim y')\|_{TV} \leq ||y - y'||_1 \leq \sqrt{|\mathcal{G}|} \epsilon. \] (4.25)
for any mixed state \(y'\) in partition \(r\). Ho and Yeung, 2010 gives us an upper bound on differences in entropy in terms of the total variation, which here implies that
\[ |H[\overrightarrow{X} | \mathcal{R}(\epsilon) = r] - H[\overrightarrow{X} | \mathcal{G} \sim y']| \leq H_b \left( \frac{\sqrt{|\mathcal{G}|} \epsilon}{2} \right) + \frac{\epsilon L \log_2 |\mathcal{A}|}{2}. \] (4.26)

This, in turn, gives
\[ |H[\overrightarrow{X} | \mathcal{R}(\epsilon) = r] - \int_y d\pi(y|r)H[\overrightarrow{X} | \mathcal{G} \sim y]| \leq \int_y d\pi(y|r)|H[\overrightarrow{X} | \mathcal{R}(\epsilon) = r] - H[\overrightarrow{X} | \mathcal{G} \sim y]| \] \[ \leq H_b \left( \frac{\sqrt{|\mathcal{G}|} \epsilon}{2} \right) + \frac{\epsilon L \log_2 |\mathcal{A}|}{2}. \] (4.27)
With that having been said, if there is only one mixed state in some \(\epsilon\)-cube \(r\), the quantity above is zero: \(|H[\overrightarrow{X} | \mathcal{R}(\epsilon) = r] - \int_y d\pi(y|r)H[\overrightarrow{X} | \mathcal{G} \sim y]| = 0\). Denote the probability of a non-empty \(\epsilon\)-cube having only one mixed state
4.4. Coarse-grained mixed states as nearly lossless predictive features

in it as \( \sum_{r:H|Y|\leq r=0} \pi(r) \). Then substitution into Eq. 4.20 gives

\[
d_L(\mathcal{R}(\epsilon)) \leq \left( \sum_{r:H|Y|\leq r=0} \pi(r) \right) \left( H_0 \left( \frac{\sqrt{|\mathcal{G}|\epsilon}}{2} \right) + \frac{\epsilon L \log_2 |A|}{2} \right). \tag{4.28}
\]

This left factor, \( \sum_{r:H|Y|\leq r=0} \pi(r) \) tends to zero as the partitions shrink for a process generated by a countable \( \epsilon \)-machine, and not otherwise. For ease of presentation, we first study the case of uncountably infinite \( \epsilon \)-machines, and then comment on the case of processes produced by countable \( \epsilon \)-machines.

From Eq. 4.28, we can readily conclude that for any finite \( L \),

\[
\lim_{\epsilon \to 0} \frac{d_L(\mathcal{R}(\epsilon))}{\epsilon^\gamma} = 0 \tag{4.29}
\]

when \( \gamma < 1 \). We would like to extend this conclusion to \( d(\mathcal{R}(\epsilon)) \), but inspection of Eq. 4.28 suggests that concluding anything similar for \( d(\mathcal{R}(\epsilon)) \) requires some care. In particular, we would like to show that \( \lim_{\epsilon \to 0} \frac{d(\mathcal{R}(\epsilon))}{\epsilon} = 0 \) for any \( \gamma < 1 \). Recall that \( d(\mathcal{R}) = \lim_{L \to \infty} d_L(\mathcal{R}) \). If we can exchange the limits \( \lim_{\epsilon \to 0} \) and \( \lim_{L \to \infty} \), then

\[
\lim_{\epsilon \to 0} \frac{d(\mathcal{R}(\epsilon))}{\epsilon^\gamma} = \lim_{\epsilon \to 0} \lim_{L \to \infty} \frac{d_L(\mathcal{R}(\epsilon))}{\epsilon^\gamma} = \lim_{L \to \infty} \lim_{\epsilon \to 0} \frac{d_L(\mathcal{R}(\epsilon))}{\epsilon^\gamma} = 0. \tag{4.30}
\]

In order for us to justify exchanging these limits, we appeal to the Moore-Osgood Theorem. Consider any monotone decreasing sequence of \( \epsilon_i \), \( \{\epsilon_i\}_{i=1}^\infty \), such that \( a_{i,L} := \frac{d_L(\mathcal{R}(\epsilon_i))}{\epsilon_i^\gamma} \) is a doubly-infinite sequence. When \( \gamma < 1 \), we have that \( \lim_{i \to \infty} a_{i,L} = 0 \). We provide a plausibility argument for uniform convergence of \( a_{i,L} \) to 0. Recall from Eq. 4.28 that

\[
|a_{i,L} - 0| = a_{i,L} \leq \frac{1}{\epsilon_i} \left( H_0(\sqrt{|\mathcal{G}|\epsilon_i}) + \epsilon_i L \log_2 |A| \right) \tag{4.31}
\]

and since \( H_0(x) \leq -x \log_2 x + x \), this expands to

\[
a_{i,L} \leq \frac{\sqrt{|\mathcal{G}|\epsilon_i^{1-\gamma}}}{2} \log_2 (1/\epsilon_i) + \left( \frac{\sqrt{|\mathcal{G}|}}{2} \log_2 \frac{2}{\sqrt{|\mathcal{G}|}} + \frac{\sqrt{|\mathcal{G}|}}{2} + \frac{L \log_2 |A|}{2} \right) \epsilon_i^{1-\gamma}. \tag{4.32}
\]

At small enough \( \epsilon_i \), the first term will dominate, implying that

\[
a_{i,L} \leq \frac{\sqrt{|\mathcal{G}|}\epsilon_i^{1-\gamma} \log_2 1}{\epsilon_i}. \]

By making \( i \) sufficiently large (i.e., by making \( \epsilon_i \) sufficiently small) we can make this upper bound as small as desired. Finally, Data Processing Inequality implies that \( I[X_0;X^T|\mathcal{R}] \) is monotone increasing in \( L \) and upper bounded by \( E < \log_2 |\mathcal{G}| < \infty \), and so \( a_{i,L} \) is also monotone increasing in \( L \) and upper-bounded by \( E \epsilon_i \); hence, the monotone convergence theorem implies that \( \lim_{L \to \infty} a_{i,L} \) exists for any \( i \). Therefore, the conditions of the

\footnote{This upper bound on \( E \) was noted in Löhr, 2009.}
Moore-Osgood Theorem are satisfied, and \( \lim_{\gamma \to \infty} \lim_{n \to \infty} \frac{d_L(R^{(\epsilon)})}{\epsilon} = 0 \), so \( \lim_{\epsilon \to 0} \frac{d(R^{(\epsilon)})}{\epsilon} = 0 \) for any \( \gamma < 1 \) as desired.

In short, \( d(R^{(\epsilon)}) \) scales no slower than \( \epsilon \) for small \( \epsilon \). Loosely speaking, we say that processes with uncountably infinite \( \epsilon \)-machines have \( d(R^{(\epsilon)}) \sim \epsilon \).

Processes generated by countably infinite \( \epsilon \)-machines are trickier to characterize for two main reasons. First, predictive distortion can decrease faster than \( \epsilon \) for processes generated by countably infinite \( \epsilon \)-machines because of the left-hand factor in Eq. 4.28, \( \sum_v \pi_v \gamma(R^{(\epsilon)}=r) \). We are guaranteed that \( \lim_{\epsilon \to 0} \sum_v \pi_v \gamma(R^{(\epsilon)}=r) = 0 \), but the rate of convergence to 0 is highly process-dependent. Second, the scaling relation between \( N_\epsilon \) and \( 1/\epsilon \) may be (and usually is) sub-power law. Together, though, we can be sure that \( d(R^{(\epsilon)}) \) and scales faster than \( \sim \epsilon \) for processes with countably infinite \( \epsilon \)-machines.

To see these points, consider the parametrized simple nonunifilar source (SNS), represented by Fig. 4.13. It is straightforward to show that the mixed states lie on a line in the simplex: \( \{ (n_1,1-n_1) \}_{n=0}^{\infty} \) with \( v_n = \frac{(1-p)^n(\frac{q}{p^0})}{(1-p)^{n(p-1)}q^{1-q/p}} \). If \( p < q \), then \( \lim_{n \to \infty} v_n = 1 - \frac{p}{q} \); if \( q < p \), then \( \lim_{n \to \infty} v_n = 0 \). It is also straightforward to show that \( v_n \) converges exponentially fast to its limit when \( p \neq q \):

\[
v_n - \lim_{n \to \infty} v_n \approx \begin{cases} 
-\frac{p}{q} \left( 1 - \frac{p}{q} \right)^n & p < q \\
1 - \frac{p}{q} \left( 1 - \frac{p}{q} \right)^n & q < p 
\end{cases} \quad (4.33)
\]

However, when \( p = q \), then \( v_n = \frac{n}{n+\left(\frac{p}{q}-1\right)} \), with \( \lim_{n \to \infty} v_n = 1 \). For any \( p \neq q \), \( v_n - \lim_{n \to \infty} v_n \approx (1/p) - 1/n \) a much slower rate of convergence. This greatly affects the box-counting dimension of the parametrized SNS’s mixed state presentation. In fact, \( \text{dim}_0(Y) \) is nonzero (and roughly \( 1/2 \)) only when \( p = q \); see Fig. 4.13(bottom). Additionally, for the parametrized SNS, causal states act as a counter of the number of 0’s since last 1, so we can roughly think of the predictive distortion at coarse-graining \( \epsilon = 1/n \) as \( E - \text{E}(n) \), which decreases exponentially quickly with \( n \) rather than algebraically under weak conditions, e.g. see Ref. Travers, 2014 and references therein. Alternatively, from Eq. 4.28, we note that \( \sum_{i=0}^{\infty} \pi(i) \) decreases exponentially for the parametrized SNS; see Marzen and Crutchfield, 2014b for details.

Together, we find that the scaling of the number of features with desired distortion \( d \) for countably infinite \( \epsilon \)-machines is very different from either finite or uncountably infinite \( \epsilon \)-machines. For the parametrized SNS with \( p \neq q \), we have \( N_\epsilon \sim \left( \log \log \frac{1}{d(R^{(\epsilon)})} \right)^2 \). When \( p = q \), we have \( N_\epsilon \sim \sqrt{\log \frac{1}{d(R^{(\epsilon)})}} \), which is still markedly faster than the algebraic convergence seen for processes with uncountably infinite \( \epsilon \)-machines.

In summation, unlike the first feature set, this feature set performs quite well on stochastic processes with countably infinite causal states. When there are only a finite number of causal states, then our feature set will, for some nonzero \( \epsilon \), actually be the causal states \( S \). If instead a process has a countable infinity of causal states, there is no such critical \( \epsilon \); but still, as for the finite \(|S|\) case, both the information dimension and sometimes the
4.4. Coarse-grained mixed states as nearly lossless predictive features

The box-counting dimension of the mixed state space will be $0$. More precisely, from Lemma 6 and remarks above, we see that $d(R)$ at coarse-graining $\epsilon$ scales at least as fast as $\epsilon$ in the limit of asymptotically small $\epsilon$. From this, we conclude that

$$|\mathcal{F}| \sim \left(\frac{1}{d}\right)^{\dim_0(Y)}$$

and

$$H[R] \sim \dim_1(Y) \log \frac{1}{d}$$

for processes with uncountably infinite $\epsilon$-machines and faster than these scalings for processes with countably infinite or finite $\epsilon$-machines. This feature set outperforms the feature set of $L$-length pasts when the process has a finite or countably infinite $\epsilon$-machine. It potentially outperforms the feature set of $L$-length pasts when the process has an uncountably infinite $\epsilon$-machine, depending on the ratio of the box-counting dimension of mixed state space $\dim_0(Y)$ to the scaling exponent in Eq. 4.17, $h_0/\lambda$, and on the ratio of $\dim_1(Y)$ to $h_\mu/\lambda$. In general, both $\dim_0(Y)$ and $\dim_1(Y)$ are bounded above by $|G|$, the number of states in the minimal generative model; in principle, $\lambda$ could be arbitrarily small, and so $h_0/\lambda$, $h_\mu/\lambda$ arbitrarily large, even for processes generated by finite-state Hidden Markov models.\(^5\)

The bounds presented here give a sense of how many states might be inferred according to the criteria of Still and Bialek, 2004, i.e. when the uncertainty in estimation of $d(R)$ equals the estimated value. The standard deviation of estimated predictive distortion will tend to decrease as the inverse square root of the amount of data given. As such, we might expect to see the number of states in the inferred minimal maximally predictive model (e.g., see Strelioff and Crutchfield, 2014; Strelioff and Crutchfield, in preparation; Still, 2014; Pfau, Bartlett, and Wood, 2010) scaling algebraically, with a coefficient equal to twice the box-counting dimension of the mixed state space. Data is often processed in real time in many Big Data applications. These types of scaling relations make it possible to estimate when one will require more computer memory.

\(^5\)As an example, consider the spectral gap of the parametrized Even Process.
Figure 4.11: Tent Map predictive features as a function of inverse temperature $\beta$: Each state-transition diagram shows the $\epsilon$-machine in Fig. 4.9(bottom) with nodes gray-scaled by $P(S^n | \mathcal{R} = r)$ for each $r \in \mathcal{R}$. White denotes high probability and black low. Transitions are shown only to guide the eye. The four $\beta$ are chosen to be close to the “critical $\beta$” at which the number of predictive features increases, shown by the $\beta$ at which the feature curve in Fig. 4.10(bottom) appears to jump discontinuously. (a) $\beta = 0.01$: one state that puts unequal weights on states $C$ and $D$. (b) $\beta = 1.9$: two states identified, $A$ and a mixture of $C$ and $D$. (c) $\beta = 3.1$: three states are identified, $A$, $B$, and the mixture of $C$ and $D$. (d) $\beta \rightarrow \infty$: original four states identified, $A$, $B$, $C$, and $D$. 

4.4. Coarse-grained mixed states as nearly lossless predictive features

\[ p(x|A) = \begin{cases} \alpha & x = 0 \\ \frac{1-\alpha}{2} & x = 1, 2' \end{cases} \]

\[ p(x|B) = \begin{cases} \alpha & x = 1 \\ \frac{1-\alpha}{2} & x = 0, 2' \end{cases} \]

\[ p(x|C) = \begin{cases} \alpha & x = 2 \\ \frac{1-\alpha}{2} & x = 0, 1' \end{cases} \]

For Fig. 4.12(middle bottom), \( x = 0.15 \) and \( \alpha = 0.6 \); for Fig. 4.12(right bottom), \( x = 0.05 \) and \( \alpha = 0.6 \). In order to visualize the mixed state simplex in Fig. 4.12, we plot only the first two elements of the mixed state. Box-counting dimension calculated by estimating the slope of \( \log \frac{1}{\epsilon} \) vs. \( \log N_{\epsilon} \).

The box-counting dimensions of these three are as follows: \( \text{dim}_0(Y) \approx 0.3 \) at left, \( \text{dim}_0(Y) \approx 1.8 \) in the middle, and \( \text{dim}_0(Y) \approx 1.9 \) at right.

\[ N_{\epsilon} \approx (\log \frac{1}{\epsilon})^{1/2} \]

\[ N_{\epsilon} \sim (\log \frac{1}{\epsilon})^2 \]

FIGURE 4.12: Mixed state presentations for three different generative models given by the top row. At top left is the first generative model, and at top right is a generative model with \( p(x|A) = \begin{cases} \alpha & x = 0 \\ \frac{1-\alpha}{2} & x = 1, 2' \end{cases} \).

FIGURE 4.13: At top, the generative model of the parametrized simple nonunifilar source (SNS). At bottom, the number of nonempty \( \epsilon \)-cubes as a function of \( 1/\epsilon \) on a log-log plot. When \( p = q \), the scaling relation appears to be power law, \( N_{\epsilon} \approx (1/\epsilon)^{1/2} \); when \( p \neq q \), \( N_{\epsilon} \) scales more slowly than a power law with \( 1/\epsilon \), seemingly at a rate \( N_{\epsilon} \sim (\log \frac{1}{\epsilon})^2 \).
Chapter 5

Stimuli to test a sensor’s predictive capabilities

Chapter 4 described how to test whether or not memory was efficiently utilized for prediction. In this section, we outline types of stimuli that could be used in experiments designed to test predictive capabilities of a sensor (Palmer et al., 2015). Along the way, we are forced to develop new calculational machinery to deal with continuous-time causal states. Recall from Chapter 4 that, in order to calculate memory stored by and predictive power of a transducer, we need the joint probability distribution of forward- and reverse-time causal states (or prescient states), \( p(\sigma^+, \sigma^-) \).

Before we begin in earnest, we must address a “technical difficulty”. Our conception of nature often revolves around a discrete-time version of events, in which the environment provokes a sensory response which provokes an actuator response which then affects the next environment symbol observed. In reality, these dynamical systems evolve continuously, ignoring the discrete-time artifice that we observers may place on the systems. For this reason, Sec. 5.1 is devoted to explicating and understanding the subtle technical difficulties involved in the study of continuous-time ε-machines, and their relationship to corresponding discrete-time ε-machines, for renewal processes—building blocks of the more sophisticated hidden semi-Markov models studied in Sec. 5.2. Not only are renewal processes building blocks of these hidden semi-Markov models, but also they arise commonly in the theoretical neuroscience literature\(^1\). Finally, in Sec. 5.3, we describe how these stimuli can be “complex” using the definition of Bialek, Nemenman, and Tishby, 2001b; Bialek, Nemenman, and Tishby, 2001a.

Readers interested in a description of new ε-machine stimuli should skip to Sec. 5.2.

5.1 Renewal processes

Realizations from a renewal process consist of sequences of events separated by epochs of quiescence, the lengths of which are drawn independently from the same interevent distribution. Throughout, when discussing

\[ \frac{dV}{dt} = g(V) + f(V) \frac{d\eta}{dt} \]  \hspace{1cm} (5.1)

and when \( V \) reaches \( V_{\text{thresh}} \), the neuron “spikes” and resets to \( V_{\text{reset}} \). It is not difficult to see that the corresponding output spike train is a renewal process, in which interspike intervals are drawn independently from some distribution \( \phi(t) \) (Gerstner et al., 2014).

---

\(^1\)A generalized integrate-and-fire neuron which receives a large number of inputs from surrounding neurons has membrane voltage \( V \) which evolves according to

\[ \frac{dV}{dt} = g(V) + f(V) \frac{d\eta}{dt} \]

and when \( V \) reaches \( V_{\text{thresh}} \), the neuron “spikes” and resets to \( V_{\text{reset}} \). It is not difficult to see that the corresponding output spike train is a renewal process, in which interspike intervals are drawn independently from some distribution \( \phi(t) \) (Gerstner et al., 2014).
5.1. Renewal processes

A discrete-time renewal process, we use the following notation: $F(n)$ is the interevent count probability distribution function; $w(n) = \sum_{n'}^\infty F(n')$ is the survival function; and $\mu$ is its mean interevent count. We use the following notation for continuous-time renewal processes: $\phi(t)$ is the waiting time distribution; $\Phi(t) = \int_t^\infty \phi(t')dt'$ is its survival function; and $T$ is the mean interevent interval.

The causal states of a renewal process are surprisingly complicated. We first tackle discrete-time renewal processes and then focus on continuous-time renewal processes. Renewal processes are essentially the simplest processes that require new $\epsilon$-machine calculational machinery.

5.1.1 Discrete-time renewal processes

It will be helpful pedagogically to anchor our theory in the contrast between two different, but still simple, renewal processes. One is the familiar “memoryless” Poisson process with rate $\lambda$. Its HMM generator, a biased coin, is shown at the left of Fig. 5.1. It has an interevent count distribution $F(n) = (1 - \lambda)^n\lambda^n$; a distribution with unbounded support. However, we notice in Fig. 5.1 that it is a unifilar model with a minimal number of states. So, in fact, this one-state machine is the $\epsilon$-machine of a Poisson process. The rate at which it generates information is given by the entropy rate: $h_\mu = H_b(\lambda)$ bits per output symbol. (Here, $H_b(p)$ is the binary entropy function.) It also has a vanishing statistical complexity $C_\mu^+ = 0$ and so stores no historical information.

![Diagram](chart.png)

**Figure 5.1:** (Left) Minimal generative model for the Poisson process with rate $\lambda$. (Right) A generator for the Simple Nonunifilar Source (SNS). Both generate a stationary renewal process. Transition labels $p|s$ denote probability $p$ of taking a transition and emitting symbol $s$. 
The second example is the Simple Nonunifilar Source (SNS) (Crutchfield, 1994), an HMM generator for which is shown on the right of Fig. 5.1. Transitions from state $B$ are unifilar, but transitions from state $A$ are not. In fact, a little reflection shows that the time series produced by the SNS is a discrete-time renewal process. Once we observe the “event” $x_t = 1$, we know the internal model state to be $\sigma_{t+1} = A$, so successive interevent counts are independent.

The SNS generator is not an $\epsilon$-machine and, moreover, it cannot be used to calculate the process’s information per output symbol (entropy rate). If we can only see 0’s and 1’s, we will usually be uncertain as to whether we are in state $A$ or state $B$, so this generative model is not maximally predictive.

We start with a simple Lemma that follows directly from the definitions of a renewal process and the causal states. It allows us to introduce notation that simplifies the development.

Notation 1. Rather than write pasts and futures as semi-infinite sequences, we notate a past as a list of nonnegative integers. The semi-infinite past $X_0:0$ is equivalent to a list of interevent counts $N_0:0$ and the count $N'_0$ since last event. Similarly, the semi-infinite future $X_0:^0$ is equivalent to the count to next event $N_0 - N'_0$ and future interevent counts $N_1:0$.

A prescient state $R$ is a function of the past such that:

$$ H[X_0:|X_0] = H[X_0:|R] . $$

According to Lemma 7, if we remember only the number of counts since the last event and nothing prior, we can predict the future as well as if we had memorized the entire past.

Lemma 7. The count since last event is a prescient statistic of a discrete-time stationary renewal process.

Proof. This follows almost immediately from the definition of stationary renewal process and the definition of causal states, since the random variables $N_i$ are all i.i.d.. Then:

$$ P(X_0:|X_0) = P(N_0 - N'_0|N'_0) \prod_{i=1}^{\infty} P(N_i) . $$

And, therefore, $P(X_0:|X_0 = x_0) = P(X_0:|X_0 = x'_0)$ is equivalent to $P(N_0 - N'_0|N'_0 = n_0) = P(N_0 - N'_0|N'_0 = n'_0)$. Hence, the counts since last event are prescient.

Causal states can be written as unions of prescient states (Shalizi and Crutchfield, 2001). We start with a definition that helps to characterize the converse; i.e., when the prescient states of Lemma 7 are also causal states.

To ground our intuition, recall that Poisson processes are “memoryless”. We therefore expect the prescient states in Lemma 7 to fail to be causal

---

$^2$This may seem counterintuitive from a parameter estimation point of view. After all, one makes better and better estimates of the Poisson rate from observing longer and longer pasts. However, finite data fluctuations in estimating model parameters are irrelevant to the present mathematical setting unless the parameters are themselves random variables. This is not our setting here: see Sec. 5.3.1 for more information.
states precisely when the interevent distribution is similar to that of a Poisson renewal process. This intuition is made precise by Def. 1.

**Definition 1.** A $\Delta$-Poisson process has an interevent distribution

$$F(n) = F(n \mod \Delta) \lambda^\lfloor n/\Delta \rfloor,$$

for all $n$ and some $\lambda > 0$. If this statement holds for multiple $\Delta \geq 1$, then we choose the smallest possible $\Delta$.

A $(\tilde{n}, \Delta)$ eventually $\Delta$-Poisson process has an interevent distribution that is $\Delta$-Poisson for all $n \geq \tilde{n}$:

$$F(\tilde{n} + k\Delta + m) = \lambda^k F(\tilde{n} + m),$$

for all $0 \leq m < \Delta$, for all $k \geq 0$, and for some $\lambda > 0$. If this statement holds for multiple $\Delta \geq 1$ and multiple $\tilde{n}$, then we choose the smallest possible $\Delta$ and the smallest possible $\tilde{n}$.

Thus, a Poisson process is a $\Delta$-Poisson process with $\Delta = 1$ and an eventually $\Delta$-Poisson process with $\Delta = 1$ and $\tilde{n} = 0$. Moreover, we will now show that at some finite $\tilde{n}$, any renewal process is either (i) Poisson, if $\Delta = 1$, or (ii) a combination of several Poisson processes, if $\Delta > 1$.

Why identify new classes of renewal process? In short, renewal processes that are similar to, but not the same as, the Poisson process do not have an infinite number of causal states. The particular condition for when they do not is given by the eventually $\Delta$-Poisson definition. (Their finiteness provides some clue as to how one efficiently approximates predictive rate-distortion functions of renewal processes; see Chapter 4.) Notably, this new class is what emerged, rather unexpectedly, by applying the causal-state equivalence relation $\sim^+$ to renewal processes. The resulting insight
is that general renewal processes, after some number of counts (the “eventually” part) and after some coarse-graining of counts (the $\Delta$ part), behave like a Poisson process.

With these definitions in hand, we can proceed to identify the causal architecture of discrete-time stationary renewal processes.

**Notation 2.** Let $r_n^+ := \{ T : x_{-n-1:0} = 10^n \}$ for $n \in \mathbb{Z}_{\geq 0}$. Recall that $10^n = 100 \cdots 00$, the sequence with $n$ 0s following a 1.

**Remark 3.** Note that $\mathcal{R}^+ = \{ r_n^+ \}_{n=0}^\infty$ is always at least a forward-time prescient rival, if not the forward-time causal states $\mathcal{S}^+$. The probability distribution over $r_n^+$ is straightforward to derive. Saying that $N'_0 = n$ means there were $n$ 0s since the last event, so that the symbol at $X_{-n-1}$ must have been a 1. That is:

$$
\pi(r_n^+) = P(N'_0 = n) = \sum_{x \in A} P(N'_0 = n, X_{-n-1} = x) = P(N'_0 = n, X_{-n-1} = 1) = P(X_{-n-1} = 1)P(X_{-n} = 0^n | X_{-n-1} = 1) .
$$

Since this is a stationary process, $P(X_{-n-1} = 1)$ is independent of $n$, implying:

$$
\pi(r_n^+) \propto P(X_{-n} = 0^n | X_{-n-1} = 1) = \sum_{m=0}^\infty P(X_{-n:m+1} = 0^{n+m+1} | X_{-n-1} = 1) = \sum_{m=n}^\infty F(m) = w(n) .
$$

We see that $\pi(r_n^+) = w(n)/Z$, with $Z$ a normalization constant that makes $\sum_{n=0}^\infty w(n) = T$. And so:

$$
\pi(r_n^+) = \frac{w(n)}{T} .
$$

**Theorem 3.** (a) The forward-time causal states of a discrete-time stationary renewal process that is not eventually $\Delta$-Poisson are groupings of pasts with the same count since last event. (b) The forward-time causal states of a discrete-time eventually $\Delta$-Poisson stationary renewal process are groupings of pasts with the same count since last event up until $\tilde{n}$ and pasts whose count $n$ since last event are in the same equivalence class as $\tilde{n}$ modulo $\Delta$.

**Proof.** From the proof of Lemma 7 in this appendix, we know that two prescient states $r_n^+$ and $r_{n'}^+$ are minimal only when:

$$
P(N_0 - N'_0 | N'_0 = n) = P(N_0 - N'_0 | N'_0 = n') .
$$

Since $P(N_0 - N'_0 = m | N'_0 = n) = P(N_0 = m + n)/P(N'_0 = n)$, $P(N_0 = m + n) = F(m + n)$, and $P(N'_0 = n) = w(n)/T$ from earlier, we find that the equivalence class condition becomes:

$$
\frac{F(m + n)}{w(n)} = \frac{F(m + n')}{w(n')} ,
$$

(5.3)
for all \( m \geq 0 \).

First, note that for these conditional probabilities even to be well-defined, \( w(n) > 0 \) and \( w(n') > 0 \). Hence, if \( F \) has bounded support—\( \max \text{supp} F(n) = N \)—then the causal states do not include any \( r^+_{n'} \) for \( n > N \). Furthermore, Eq. (5.3) cannot be true for all \( m \geq 0 \), unless \( n = n' \) for \( n \) and \( n' \leq N \). To see this, suppose that \( n \neq n' \) but that Eq. (5.3) holds. Then choose \( m = N + 1 - \max(n, n') \) to give \( 0 = F(N + 1 - |n - n'|)/w(n') \), a contradiction unless \( n = n' \).

So, for all remaining cases, we can assume that \( F \) in Eq. (5.3) has unbounded support.

A little rewriting makes the connection between Eq. (5.3) and an eventually \( \Delta \)-Poisson process clearer. First, we choose \( m = 0 \) to find:

\[
\frac{F(n)}{w(n)} = \frac{F(n')}{w(n')},
\]

which we can use to rewrite Eq. (5.3) as:

\[
\frac{F(m + n)}{F(n)} = \frac{F(m + n')}{F(n')},
\]

or more usefully:

\[
F(n' + m) = \frac{F(n')}{F(n)} F(n + m).
\]

A particularly compact way of rewriting this is to define \( \Delta' := n' - n \), which gives \( F(n' + m) = F((n + m) + \Delta') \). In this form, it is clear that the above equation is a recurrence relation on \( F \) in steps of \( \Delta' \), so that we can write:

\[
F((n + m) + k\Delta') = \left( \frac{F(n')}{F(n)} \right)^k F(n + m). \tag{5.4}
\]

This must be true for every \( m \geq 0 \). Importantly, since \( w(n) = \sum_{m=n}^{\infty} F(m) \), satisfying this recurrence relation is equivalent to satisfying Eq. (5.3). But Eq. (5.4) is just the definition of an eventually \( \Delta \)-Poisson process in disguise; relabel with \( \lambda := F(n')/F(n) \), \( \tilde{n} := n \), and \( \Delta = \Delta' \).

Therefore, if Eq. (5.3) does not hold for any pair \( n \neq n' \), the process is not eventually \( \Delta \)-Poisson and the prescient states identified in Lemma 7 are minimal; i.e., they are the causal states.

If Eq. (5.3) does hold for some \( n \neq n' \), choose the minimal such \( n \) and \( n' \) both. The renewal process is eventually \( \Delta \)-Poisson with characterization \( \Delta = n' - n \) and \( \tilde{n} \). And, \( F(\tilde{n} + m)/w(\tilde{n} + m) = F(\tilde{n} + m')/w(\tilde{n} + m') \) implies that \( m \equiv m' \mod \Delta \) since otherwise, the \( n \) and \( n' \) chosen would not be minimal. The theorem follows.

\begin{remark}
For the resulting \( F(n) \) to be a valid interevent distribution, \( \lambda = F(\tilde{n} + \Delta)/F(\tilde{n}) < 1 \) as normalization implies:

\[
\sum_{n=0}^{\tilde{n}-1} F(n) + \sum_{n=\tilde{n}}^{\tilde{n}+\Delta-1} F(n) \frac{1}{1 - \lambda} = 1.
\]

The Poisson process, as an eventually \( \Delta \)-Poisson with \( \tilde{n} = 0 \) and \( \Delta = 1 \), is represented by the one-state \( \epsilon \)-machine despite the unbounded support.
of its interevent count distribution. Unlike most processes, the Poisson process’ $\epsilon$-machine is the same as its generative model shown in Fig. 5.1(left).

The SNS, on the other hand, has an interevent count distribution that is not eventually $\Delta$-Poisson. According to Thm. 3, then, the SNS has a countable infinity of causal states despite its simple two-state generative model in Fig. 5.1(left). Compare the leftmost $\epsilon$-machine in Fig. 5.2. Each causal state corresponds to a different probability distribution over the internal states $A$ and $B$. These internal state distributions are the mixed states of Ellison, Mahoney, and Crutchfield, 2009. Observing more 0’s, one becomes increasingly convinced that the internal state is $B$. For maximal predictive power, however, we must track the probability that the process is still in state $A$. Both Fig. 5.2 and Fig. 5.1(left) are “minimally complex” models of the same process, but with different definitions of model complexity.

The main result is that causal states are sensitive to two features: (i) eventually $\Delta$-Poisson structure in the interevent distribution and (ii) the boundedness of $F(n)$’s support. If the support is bounded, then there are a finite number of causal states rather than a countable infinity of causal states. Similarly, if $F(n)$ has $\Delta$-Poisson tails, then there are a finite number of causal states despite the support of $F(n)$ having no bound. Nonetheless, one can say that the generic discrete-time stationary renewal process has a countable infinity of causal states.

Finding the probability distribution over these causal states is straightforwardly related to the survival-time distribution $w(n)$ and the mean interevent interval $\mu$, since the probability of observing at least $n$ counts since the last event is $w(n)$. Hence, the probability of seeing $n$ counts since the last event is simply the normalized survival function $w(n)/(\mu + 1)$.

We can also endow the causal state space with a transition dynamic to construct the renewal process $\epsilon$-machine. The transition dynamic is sensitive to $F(n)$’s support and not only its boundedness. For instance, the probability of observing an event given that it has been $n$ counts since the last event is $F(n)/w(n)$. For the generic discrete-time renewal process this is exactly the transition probability from causal state $n$ to causal state 0. If $F(n) = 0$, then there is no probability of transition from $\sigma = n$ to $\sigma = 0$.

Figure 5.2 display the causal state architectures for the $\epsilon$-machines in the various cases delineated by Thm. 3. The leftmost is the $\epsilon$-machine of a generic renewal process whose interevent interval can be arbitrarily large and whose interevent distribution never has exponential tails. The second to left is the $\epsilon$-machine of a renewal process whose interevent distribution never has exponential tails but cannot have arbitrarily large interevent counts. The $\epsilon$-machine in second to the right looks quite similar to the $\epsilon$-machine to its left, but it has an additional transition that connects the last state $\tilde{n}$ to itself. This added transition changes our structural interpretation of the process. Interevent counts can be arbitrarily large for this $\epsilon$-machine but past an interevent count of $\tilde{n}$, the interevent distribution is exponential. Finally, the rightmost $\epsilon$-machine in Fig. 5.2 represents an eventually $\Delta$-Poisson process with $\Delta > 1$ whose structure is conceptually most similar to that of the $\epsilon$-machine to its left. (See Def. 1 for the precise version of that statement.) If our renewal process disallows seeing interevent counts of a particular length $L$, then this will be apparent from the $\epsilon$-machine since there will be no transition between the causal state corresponding to an interevent count of $L$ and causal state 0.
We can find an analytic formula for the excess entropy once we characterize its statistical structure in reverse time (Crutchfield, Ellison, and Mahoney, 2009; Ellison, Mahoney, and Crutchfield, 2009).

**Lemma 8.** Groupings of futures with the same counts to next event are reverse-time prescient statistics for discrete-time stationary renewal processes.

**Theorem 4.** (a) The reverse-time causal states of a discrete-time stationary renewal process that is not eventually $\Delta$-Poisson are groupings of futures with the same count to next event. (b) The reverse-time causal states of a discrete-time eventually $\Delta$-Poisson stationary renewal process are groupings of futures with the same count to next event up until $\tilde{n}$ plus groupings of futures whose count since last event $n$ are in the same equivalence class as $\tilde{n}$ modulo $\Delta$.

**Proof.** The proof for both claims relies on a single fact: In reverse-time, a stationary renewal process is still a stationary renewal process with the same interevent count distribution. The lemma and theorem therefore follow from Lemma 7 and Thm. 3.

The excess entropy, being the mutual information between forward and reverse-time prescient states is (Crutchfield, Ellison, and Mahoney, 2009; Ellison, Mahoney, and Crutchfield, 2009):

$$ E = I[R^+; R^-] = H[R^+] - H[R^+ | R^-] . $$

And so, to calculate, we note that:

$$ P(r_n^+, r_m^-) = \frac{F(m + n)}{T} \quad \text{and} \quad P(r_n^+ | r_m^-) = \frac{F(n + m)}{w(m)} . $$

After some algebra, we find that:

$$ H[R^+] = - \sum_{n=0}^{\infty} \frac{w(n)}{T} \log_2 \frac{w(n)}{T} , $$

which is also the statistical complexity $C_\mu$ of the “generic” renewal process, i.e. non-eventually $\Delta$ Poisson. Other calculation gives

$$ H[R^+ | R^-] = - \sum_{m, n=0}^{\infty} \frac{F(n + m)}{T} \log_2 \frac{F(n + m)}{w(m)} \frac{F(m)}{w(m)} \log_2 \frac{F(m)}{T} + \sum_{m=0}^{\infty} \frac{w(m)}{T} \log_2 \frac{w(m)}{T} . $$

The above quantity is the forward crypticity $\chi^+$ (Crutchfield, Ellison, and Mahoney, 2009) when the renewal process is not eventually $\Delta$-Poisson. These together imply:

$$ E = -2 \sum_{n=0}^{\infty} \frac{w(n)}{T} \log_2 \frac{w(n)}{T} + \sum_{m=0}^{\infty} (m + 1) \frac{F(m)}{T} \log_2 \frac{F(m)}{T} . \quad (5.5) $$
5.1.2 Infinitesimal discretization time

One often treats a continuous-time renewal process, such as a spike train from a noisy integrate-and-fire neuron, in a discrete-time setting (Rieke et al., 1999). We investigate how artificial time binning affects estimates of a model neuron’s spike train’s randomness, predictability, and information storage in the limit of infinitesimal time resolution. This is exactly the limit in which analytic formulae for information measures are most useful, as increasing the time resolution artificially increases the apparent range of temporal correlations. See, for instance, Fig. 5.3, in which the ISI distribution $\phi(t)$ is given by

$$\phi(t) = \begin{cases} 0 & t < \tau \\ \frac{\lambda}{2\pi(t-\tau)^3} e^{-\lambda(\mu(t-\tau)-1)^2/2(t-\tau)} & t \geq \tau \end{cases} \quad (5.6)$$

Figure 5.3: An unleaky integrate-and-fire neuron driven by white noise has varying interevent count distributions $F(n)$ that depend on time bin size $\Delta t$. Based on the ISI distribution given in Eq. 5.6 with $\tau = 2$ milliseconds, $1/\mu = 1$ millisecond, and $\lambda = 1$ millisecond. Data points represent exact values of $F(n)$ calculated for integer values of $N$. Dashed lines are interpolations based on straight line segments connecting nearest neighbor points.

Time-binned neural spike trains of noisy integrate-and-fire neurons have been studied for quite some time (Rieke et al., 1999). We do not estimate statistics or reconstruct models from simulated spike train data using non-parametric inference algorithms—e.g., as done in Haslinger, Klinkner, and Shalizi, 2010. Rather, we ask how $\epsilon$-machines ideally extracted from a spike train process and information measures calculated from them vary as a
function of time coarse-graining. Our analytic approach highlights an important lesson about such studies in general: a process’ $\epsilon$-machine and information anatomy are sensitive to time resolution. A secondary and compensating lesson is that the manner in which the $\epsilon$-machine and information anatomy scale with time resolution conveys information about the process’ structure.

Suppose we are given a neural spike train with interspike intervals independently drawn from the same interspike interval (ISI) distribution $\phi(t)$ with mean ISI $1/\mu$. To convert the continuous-time point process into a sequence of binary spike-quiescence symbols, we track the number of spikes emitted in successive time bins of size $\Delta t$. Our goal, however, is to understand how the choice of $\Delta t$ affects reported estimates for various information measures. The way in which each of these vary with $\Delta t$ reveals information about the intrinsic timescales on which a process behaves, e.g. as described for entropy rates (Costa, Goldberger, and Peng, 2002; Costa, Goldberger, and Peng, 2005; Gaspard and Wang, 1993).

In the infinitesimal time-resolution limit, when $\Delta t$ is smaller than any intrinsic timescale, the neural spike train is a renewal process with interevent count distribution:

$$F(n) \approx \phi(n\Delta t) \Delta t \quad (5.7)$$

and survival function:

$$w(n) \approx \int_{n\Delta t}^{\infty} \phi(t)dt \quad (5.8)$$

The interevent distribution $F(n)$ is the probability distribution that the silence separating successive events (bins with spikes) is $n$ counts long, while the survival function $w(n)$ is the probability that the silence separating successive events is at least $n$ counts long. The $\epsilon$-machine transition probabilities therefore change with $\Delta t$. The mean interevent count $T$ is not the mean interspike interval $1/\mu$ since one must convert between counts and spikes:

$$T = \frac{1}{\mu \Delta t} \quad (5.9)$$

In this limit, the $\epsilon$-machines of these model spike trains can take one of the topologies described earlier.

In this subsection, we only focus on two of these $\epsilon$-machine topologies. The first topology corresponds to that of an eventually Poisson process, in which the ISI distribution takes the form $\phi(t) = \phi(T)e^{-\lambda(t-T)}$ for some finite $T$ and $\lambda > 0$. A Poisson neuron with firing rate $\lambda$ and refractory period of time $T$, for instance, generates an eventually Poisson process (hence the name). A Poisson process is a special type of eventually Poisson process with $T = 0$; see the rightmost $\epsilon$-machine in Fig. 5.2. However, the “generic” renewal process has $\epsilon$-machine topology shown leftmost in Fig. 5.2. In other words, the “generic” renewal process has an infinite $\epsilon$-machine, for which the inferred $\epsilon$-machines are only approximations. Almost all current

---

3As the subscript context makes clear, the mean count $\mu$ is not related to that $\mu$ in $C_{\mu}$ and related quantities. In the latter it refers to the measure over bi-infinite sequences generated by a process.
ε-machine inference algorithms can only infer finite ε-machines, so such algorithms applied to renewal processes will often yield an eventually Poisson topology. (Compare Fig. 5.2 to the inferred approximate ε-machine of an integrate-and-fire neuron in Fig. 2 in Haslinger, Klinkner, and Shalizi, 2010.)

We calculated $E$ and $C_\mu$ using the expressions given earlier. Substituting in Eqs. 5.7, 5.8, and 5.9, we find that the excess entropy $E$ tends to:

$$
\lim_{\Delta t \to 0} E(\Delta t) = \int_0^\infty \mu \phi(t) \log_2 (\mu \phi(t)) \, dt - 2 \int_0^\infty \mu \Phi(t) \log_2 (\mu \Phi(t)) \, dt ,
$$

where $\Phi(t) = \int_t^\infty \phi(t') \, dt'$ is the probability that an ISI is longer than $t$. It is easy to see that $E(\Delta t)$ limits to a positive and (usually) finite value as the time resolution vanishes, with some exceptions described below.

Even if $E$ limits to a finite value, the statistical complexity typically diverges due to its dependence on time discretization $\Delta t$. Suppose that the we are observing an eventually Poisson process, such that $\phi(t) = \phi(T) e^{-\lambda (t-T)}$ for $t > T$. Then, from the previous subsection, statistical complexity in the infinitesimal time resolution limit becomes:

$$
C_\mu(\Delta t) \sim \left( \mu \int_0^T \Phi(t) \, dt \right) \log_2 \frac{1}{\Delta t} - \int_0^T (\mu \Phi(t)) \log_2 (\mu \Phi(t)) \, dt - \left( \mu \int_T^\infty \Phi(t) \, dt \right) \log_2 \left( \mu \int_T^\infty \Phi(t) \, dt \right) ,
$$

ignoring terms of $O(\Delta t)$ or higher. The first term diverges, and its rate of divergence is the probability of observing a time since last event less than $T$. This measures the process’ deviation from being Δ-Poisson and so reveals the effective dimension of the underlying causal state space. $C_\mu$’s remaining nondivergent component is equally interesting. In fact, it is the differential entropy of the distribution of the time since last event.

An immediate consequence of the analysis is that this generic continuous-time renewal process is highly cryptic (Crutchfield, Ellison, and Mahoney, 2009). It hides an arbitrarily large amount of its internal state information: $C_\mu$ diverges as $\Delta t \to 0$ but $E$ (usually) asymptotes to a finite value. We have very structured processes that have disproportionately little in the future to predict. Periodic processes constitute an important exception to this general rule of thumb for continuous-time processes. A neuron which fires every $T$ seconds without any jitter has $E = C_\mu$, and both $E$ and $C_\mu$ will diverge logarithmically with $1/\Delta t$.

### 5.1.3 Continuous-time renewal processes

A continuous-time, discrete-alphabet time series is described by a list of symbols $x$ in some alphabet $\mathcal{A}$ and dwell times for those symbols $\tau \in \mathbb{R}_{\geq 0}, \ldots, (x_{-1}, \tau_{-1}), (x_0, \tau_0), (x_1, \tau_1), \ldots$. We focus here on point processes for which $|\mathcal{A}| = 1$, and so we label the time series just by dwell times,
5.1. Renewal processes

\[ t \sim \phi(t) | 0 \]

**Figure 5.4:** Generative model of a continuous-time renewal process. The length of periods of silence are drawn i.i.d. from \( \phi(t) \).

\[ \ldots, \tau_{-1}, \tau_0, \tau_1, \ldots \] We view this time series as a realization of random variables \( \tau = \ldots, \tau_{-1}, \tau_0, \tau_1, \ldots \). When the observed time series is strictly stationary and the process ergodic, then we can (in principle) calculate the probability distribution \( P(\tau) \) from a single realization \( \tau \).

Demarcation of the present splits \( \tau_0 \) into two parts: the time since previous symbol, \( \tau_{0+} \) and the time to next symbol, \( \tau_{0-} \). Thus we define \( \tau_{0+} = \ldots, \tau_{-1}, \tau_{0+} \) as the past and \( \tau_{0-} = \tau_{0-}, \tau_{-1}, \ldots \) as the future. The present itself extends over an infinitesimally small length of time, \( \tau_{0+,0-} \).

Continuous-time renewal processes have a relatively simple generative model. Interevent intervals are drawn from a probability density function \( \phi(t) \), which we assume exists almost everywhere. The survival function \( \Phi(t) = \int_0^t \phi(t')dt' \) is the probability that an interevent interval is greater than or equal to \( t \), and to connect with the neuroscience literature– we define the mean “firing rate” as

\[
\mu = \frac{1}{\int_0^\infty t\phi(t)dt}.
\]

The minimal generative model for a continuous-time renewal process is therefore a single state machine with a continuous alphabet observable, shown in Fig. 5.4.

**Continuous-time causal states**

Renewal processes are temporally symmetric. As such, we will refer to forward-time causal states and the forward-time \( \epsilon \)-machine as simply causal states or the \( \epsilon \)-machine, with the understanding that reverse-time causal states and reverse-time \( \epsilon \)-machines will take the exact same form with slight labelling differences.

We start by describing prescient statistics. The following Lemma exactly parallels that of Lemma 7; the only difference is that the prescient statistic is the time since last event rather than counts since last event.

**Lemma 9.** The time since last event, \( \tau_{0+} \), is a prescient statistic of renewal processes.

**Proof.** From Bayes Rule,

\[
P(\tau_{0-}|\tau_{0+}) = P(\tau_{0-}|\tau_{0+})P(\tau_{1}|\tau_1)
\]

Interevent intervals \( \tau_i \) are independent of one another, so \( P(\tau_{1}|\tau_1) = P(\tau_1) \). The random variables \( \tau_{0+} \) and \( \tau_{0-} \) are functions of \( \tau_0 \) and the location of the present, both of which are independent of other interevent intervals,
and so \( P(\mathcal{T}_0^- | \mathcal{T}_0^+) = P(\mathcal{T}_0^- | \mathcal{T}_0^+) \). Then, this implies,

\[
P(\mathcal{T}_0^- | \mathcal{T}_0^+) = P(\mathcal{T}_1)P(\mathcal{T}_0^- | \mathcal{T}_0^+). \quad (5.14)
\]

The causal state equivalence relation groups two pasts \( \tau_0^+ \) and \( \tau'_0^+ \) together when \( P(\mathcal{T}_0^- | \mathcal{T}_0^+ = \tau_0^+) = P(\mathcal{T}_0^- | \mathcal{T}_0^+ = \tau'_0^+) \). We see that \( \tau_0^+ = \tau'_0^+ \) is a sufficient condition for this from Eq. 5.14. The Lemma follows.

Some renewal processes are quite predictable, while others are purely random. A Poisson process is the latter: interevent intervals are drawn from an exponential distribution, and knowing the time since last event provides no predictive benefit. A fractal renewal process can be the former: there, the interevent interval is so structured that the resultant process can have power law correlations (Lowen and Teich, 1993b), and knowing the time since last event can provide quite a bit of predictive power, as we will discuss in Sec. 5.3.3.

Intermediate between these two extremes is a renewal process whose interevent intervals are structured up to a point, falling off exponentially only after some time \( T^* \). These intermediate cases might be classified as either of the following types of renewal processes, in analogy to the discrete-time classification.

**Definition 2.** An eventually Poisson process has

\[
\phi(t) = \phi(T)e^{-\lambda(t-T)}
\]

for some \( \lambda > 0 \) and \( T > 0 \) almost everywhere. We associate the eventually Poisson process with the minimal such \( T \).

**Definition 3.** An eventually \( \Delta \)-Poisson process with \( \Delta^* > 0 \) has an interevent interval distribution that satisfies

\[
\phi(T^*+s) = \phi(T^*+(s-T^*) \mod \Delta^*)e^{-\lambda s/\Delta^*}
\]

for smallest possible \( T^* \) for which \( \Delta^* \) still exists.

A familiar example of an eventually Poisson process are the spike trains generated by Poisson neurons with refractory periods. There, the neuron is essentially prevented from firing two spikes within some time \( T \) of each other, as its ion channels restore the membrane voltage to equilibrium. After that, the time to next spike is drawn from an exponential distribution. To predict the future of the neural spike train, we would want to know the time since last spike as long as it is less than \( T \); we gain a great deal of predictive power from that piece of information. However, we do not care much about the time since last spike exactly if it is greater than \( T \), because at that point, the neuron acts as a memoryless Poisson neuron. These intuitions are captured by Thm. 5, the continuous time equivalent of Thm. 3.

**Theorem 5.** There are three different types of causal states of a renewal process:

- When the renewal process is not eventually \( \Delta \)-Poisson, the causal states are the time since last event;
- When the renewal process is eventually Poisson, the causal states are the time since last event up until time \( T^* \).
• When the renewal process is eventually $\Delta$-Poisson, the causal states are the time since last event up until time $T^*$, and the time since $T^*$ mod $\Delta$ thereafter.

Proof. Lemma 9 implies that two pasts are causally equivalent if they have the same time since last event, if $\tau_{0+} = \tau'_{0+}$. From the proof of Lemma 9, we further see that two time since last events are causally equivalent when $P(T_0-|T_0+ = \tau_{0+}) = P(T_0-|T_0+ = \tau'_{0+})$. In terms of $\phi(t)$, we find that

$$P(T_0- = \tau_{0-}|T_0+ = \tau_{0+}) = \frac{\phi(\tau_{0-} + \tau_{0+})}{\Phi(\tau_{0+})}$$

(5.17)

using manipulations very similar to those in the proof of Thm. 3. So, to find causal states, we look for $\tau_{0+} \neq \tau'_{0+}$ such that $\frac{\phi(\tau_{0-} + \tau_{0+})}{\Phi(\tau_{0+})} = \frac{\phi(\tau_{0-} + \tau'_{0+})}{\Phi(\tau'_{0+})}$ for all $\tau_{0-} \geq 0$.

To unravel the consequences of $\frac{\phi(\tau_{0-} + \tau_{0+})}{\Phi(\tau_{0+})} = \frac{\phi(\tau_{0-} + \tau'_{0+})}{\Phi(\tau'_{0+})}$, we suppose that $\tau_{0+} < \tau'_{0+}$ without loss of generality. Define $\Delta$ as $\tau'_{0+} - \tau_{0+}$ and $T$ as $\tau_{0+}$, for convenience. The equivalence relation can be rewritten as

$$\phi(T + \Delta + \tau_{0-}) = \lambda \phi(T + \tau_{0-})$$

(5.18)

where $\lambda = \frac{\Phi(T+\Delta)}{\Phi(T)}$, true for any $\tau_{0-} \geq 0$. Iterating this relationship, we find that

$$\phi(T + \tau_{0-}) = \lambda^{\lfloor \frac{\tau_{0-}}{\Delta} \rfloor} \phi(T + (\tau_{0-} \mod \Delta)).$$

(5.19)

This immediately implies the first case of the Theorem statement. If a renewal process is not eventually $\Delta$-Poisson, then $\frac{\phi(\tau_{0-} + \tau_{0+})}{\Phi(\tau_{0+})} = \frac{\phi(\tau_{0-} + \tau'_{0+})}{\Phi(\tau'_{0+})}$ for all $\tau_{0-} \geq 0$ implies $\tau_{0+} = \tau'_{0+}$, so that the prescient statistics of Lemma 9 are also minimal.

To understand the latter two cases, we consider more carefully the set of all pairs $(T, \Delta)$ for which the statement $\frac{\phi(\tau_{0-} + T)}{\Phi(T)} = \frac{\phi(\tau_{0-} + T + \Delta)}{\Phi(T + \Delta)}$ for all $\tau_{0-} \geq 0$ holds true. Define the set

$$S_{T,\Delta} := \{(T, \Delta) : \frac{\phi(\tau_{0-} + T)}{\Phi(T)} = \frac{\phi(\tau_{0-} + T + \Delta)}{\Phi(T + \Delta)} \forall \tau_{0-} \geq 0\}$$

and define the parameters $T^*$ and $\Delta^*$ by

$$T^* := \inf \{T : \exists \Delta | (T, \Delta) \in S_{T,\Delta}\}$$

$$\Delta^* := \inf \{\Delta : (T^*, \Delta) \in S_{T,\Delta}\}.$$
and consider rearranging terms to find
\[ \frac{\phi(T^* + \Delta + \tau_0) - \phi(T^* + \tau_0)}{\phi(T^* + \tau_0)} = \frac{\Phi(T^* + \Delta) - \Phi(T^*)}{\Phi(T^*)}. \]

As \( \Delta^* = 0 \), we can take the limit that \( \Delta \to 0 \) by the definition of \( \inf \) and we find that
\[ \frac{d \log \phi(t)}{dt}|_{t=T^*+\tau_0} = \frac{d \log \Phi(t)}{dt}|_{t=T^*}. \]

The right hand side is some parameter independent of \( \tau_0 \) so, this is a standard ordinary differential equation for \( \phi(t) \) which is solved by \( \phi(t) = \phi(T^*)e^{-\lambda(t-T^*)} \) for \( \lambda := -\frac{d \log \Phi(t)}{dt}|_{t=T^*}. \)

Thm. 5 implies that there is a qualitative change in \( S^+ \) depending on whether or not the renewal process is Poisson, eventually Poisson, eventually \( \Delta \)-Poisson, or not eventually Poisson. In the first case, \( S^+ \) is a discrete random variable; in the second case, \( S^+ \) is a mixed random variable; and in the third and fourth cases, \( S^+ \) is a continuous random variable.

**Wave propagation on continuous-time \( \epsilon \)-machines**

The identification of causal states follows is almost entirely similar to that of discrete-time renewal processes, but the relatively slight differences between the causal states of eventually Poisson, eventually \( \Delta \)-Poisson and not eventually \( \Delta \)-Poisson renewal processes have surprisingly important consequences for continuous-time \( \epsilon \)-machines.

As described by Thm. 5, there are often an uncountable infinity of continuous-time causal states; but as one might anticipate, there is an ordering to this infinity of causal states which makes calculations tractable. There is one major difference between discrete-time \( \epsilon \)-machines and continuous-time \( \epsilon \)-machines—that transition dynamics often amount to specifying the behavior of a probability density function over causal state space.

As such, continuous-time \( \epsilon \)-machines constitute an unusual presentation of a Hidden Markov model. The \( \epsilon \)-machines of continuous-time renewal processes look like conveyer belts, sometimes with a trashbin or second mini-conveyer belt. In a metaphorical sense, they are conveyer belts; they describe the time since the last event.

Unsurprisingly, the exception to this general rule is given by a Poisson process. The \( \epsilon \)-machine of a Poisson process is exactly the minimal generative model shown in Fig. 5.4. At each iteration, an interevent interval is drawn from a probability density function \( \phi(t) = \lambda e^{-\lambda t} \), with some \( \lambda > 0 \). Knowing the time since last event does not aid in predicting the time to next event above and beyond knowing \( \lambda \), and so the \( \epsilon \)-machine of this process has only one state.

Otherwise, the \( \epsilon \)-machine requires describing the evolution of the probability density function over the causal states. For instance, we might search for labeled transition operators \( O^{(x)} \) such that \( \frac{\partial \rho(\sigma,t)}{\partial t} = O^{(x)} \rho(\sigma,t) \). Perhaps unsurprisingly, \( O^{(1)} \) is not particularly illuminating as it is designed to “collapse the wavefunction”, and so we take a different approach and give the partial differential equations governing the labeled transition dynamics.

The \( \epsilon \)-machine of a renewal process which is not eventually-\( \Delta \) Poisson takes a form shown in Fig. 5.5. Let \( \rho(\sigma,t) \) be the probability density
5.1. Renewal processes

Figure 5.5: The $\varepsilon$-machine for the “generic” not eventually Poisson process. Continuous-time causal states are in a one-to-one mapping with the positive real line as they denote time since last event. If no event is seen, probability flows towards increasing time since last event, as described in Eq. 5.26. Otherwise, bent arrows denote allowed transitions back to the “0 node”, which denotes that an event has just occurred.

function over the causal states $\sigma$ at time $t$. Our approach to deriving labeled transition dynamics parallels well-known approaches to determining Fokker-Planck equations using a Kramers-Moyal expansion (Risken, 2007), which is to note that any probability at causal state $\sigma$ at time $t + \Delta t$ could only have come from causal state $\sigma - \Delta t$ at time $t$, if $\sigma \geq \Delta t$. This implies

$$\rho(\sigma; t + \Delta t) = P(S_{t+\Delta t} = \sigma|S_t = \sigma - \Delta t)\rho(\sigma - \Delta t; t).$$

(5.20)

But $P(S_{t+\Delta t} = \sigma|S_t = \sigma - \Delta t)$ is just the probability that the interevent interval is greater than $\sigma$, given that the interevent interval is at least $\sigma - \Delta t$, or

$$P(S_{t+\Delta t} = \sigma|S_t = \sigma - \Delta t) = \frac{\Phi(\sigma)}{\Phi(\sigma - \Delta t)}.$$

(5.21)

Together, this implies that

$$\rho(\sigma; t + \Delta t) = \frac{\Phi(\sigma)}{\Phi(\sigma - \Delta t)}\rho(\sigma - \Delta t; t).$$

(5.22)

From this, we obtain

$$\frac{\partial \rho(\sigma; t)}{\partial t} = \lim_{\Delta t \to 0} \frac{\rho(\sigma; t + \Delta t) - \rho(\sigma; t)}{\Delta t}$$

(5.23)

$$= \lim_{\Delta t \to 0} \frac{\Phi(\sigma)\rho(\sigma - \Delta t; t) - \rho(\sigma; t)}{\Delta t}$$

(5.24)

$$= \lim_{\Delta t \to 0} \frac{(\Phi(\sigma) - 1)\rho(\sigma - \Delta t; t)}{\Delta t}$$

$$+ \lim_{\Delta t \to 0} \frac{\rho(\sigma - \Delta t; t) - \rho(\sigma; t)}{\Delta t}$$

(5.25)

$$= \frac{\partial \log \Phi(\sigma)}{\partial \sigma} \rho(\sigma; t) - \frac{\partial \rho(\sigma; t)}{\partial \sigma}. \quad (5.26)$$

If $O^{(0)}$ exists, it takes the form $O^{(0)} = \frac{\partial \log \Phi(\sigma)}{\partial \sigma} - \frac{\partial}{\partial \sigma}$. The probability density function $\rho(\sigma; t)$ changes discontinuously after an event occurs: all probability mass shifts from $\sigma > 0$ to $\sigma = 0$. In other words, an event **collapses the wavefunction**.

The stationary distribution $\rho(\sigma)$ over causal states is given by setting
$$\partial_t \rho(\sigma; t)$$ to be 0 and solving for the stationary distribution $$\rho(\sigma)$$. (At the risk of confusion, we adopt the convention that $$\rho(\sigma)$$ denotes the stationary distribution, and that $$\rho(\sigma; t)$$ does not.) Some straightforward algebra shows that

$$\rho(\sigma) = \mu \Phi(\sigma).$$  \hfill (5.27)

From this, we can calculate a continuous-time statistical complexity:

$$C_\mu = \int_0^\infty \mu \Phi(\sigma) \log \frac{1}{\mu \Phi(\sigma)} d\sigma. \hfill (5.28)$$

This was the nondivergent component of the infinitesimal time-discretized renewal processes in Eq. 5.11.

Figure 5.6: The $$\epsilon$$-machine for an eventually Poisson process. Continuous-time causal states are in a one-to-one mapping with the real line up until time $$T^*$$, as they again denote time since last event. A leaky absorbing node at $$T^*$$ corresponds to any time since last event $$\geq T^*$$. If no event is seen, probability flows towards increasing time since last event or the leaky absorbing node, as described in Eqs. 5.26 and 5.30. Otherwise, bent arrows denote allowed transitions back to the "0 node", which denotes that an event has just occurred.

As anticipated in Thm. 5, there is a qualitatively different topology to the $$\epsilon$$-machine of an eventually Poisson renewal process, largely because the continuous-time causal states are a mixed random variable. For $$\sigma < T^*$$, there is wave propagation completely analogous to the wave propagation described in Eq. 5.26. However, there is a new kind of continuous-time causal state at $$\sigma = T^*$$, which has a dwell time rather than tiling a dwell time. As such, $$\rho(\sigma; t)$$ (defined for $$\sigma < T^*$$) denotes a probability density function for $$\sigma < T^*$$ and $$\pi(T^*, t)$$ denotes the probability of existing in causal state $$T^*$$. Normalization, then, requires that

$$1 = \int_0^{T^*} \rho(\sigma; t) d\sigma + \pi(T^*, t) \hfill (5.29)$$

The transition dynamics for $$\pi(T^*, t)$$ are obtained similarly to the dynamics for $$\rho(\sigma; t)$$, in that we consider all ways in which probability could flow to $$\pi(T^*, t + \Delta t)$$ in a short time window $$\Delta t$$. Probability could flow from any causal state with $$T^* - \Delta t \leq \sigma < T^*$$ or from $$\sigma = T^*$$ itself. If no event is observed:

$$\pi(T^*, t + \Delta t) = e^{-\lambda \Delta t} \pi(T^*, t) \phi(T^*)$$

+ \int_0^{\Delta t} \rho(T^* - t', t) \frac{\Phi(T^*)}{\Phi(T^* - t')} e^{-\lambda(t' - t)} dt'. \hfill (5.31)$$
5.1. Renewal processes

The term $e^{-\lambda \Delta t} \pi(T^*; t)$ corresponds to probability flow from $\sigma = T^*$ and the term $\rho(T^* - t'; t) \frac{\Phi(T^*)}{\Phi(T^* - t')} e^{-\lambda (\Delta t - t')}$ corresponds to probability influx from state $\sigma = T^* - t'$ for $0 < t' \leq \Delta t$. Assuming differentiability of $\pi(T^*, t)$ with respect to $t$, we find that

$$\frac{\partial}{\partial t} \pi(T^*; t) = -\lambda \pi(T^*; t) + \rho(T^*; t)$$ (5.30)

where $\rho(T^*; t)$ is shorthand for $\lim_{\sigma \to T^*} \rho(\sigma; t)$. This implies that the labeled transition operator $O^{(0)}$ takes a piecewise form which acts as in Eq. 5.26 for $\sigma < T^*$ and as in Eq. 5.30 for $\sigma = T^*$. Similarly to earlier, observation of an event causes us to “collapse the wavefunction” to a delta distribution at $\sigma = 0$.

The stationary distribution over causal states is given again by setting $\partial_t \rho(\sigma; t)$ and $\partial_t \pi(\sigma; t)$ to 0. Equivalently, one can use the prescription suggested by Thm. 5 to calculate $\pi(T^*)$ via integration of the stationary distribution over the prescient machine given earlier:

$$\pi(T^*) = \int_{T^*}^{\infty} \rho(\sigma) d\sigma$$ (5.31)

$$= \mu \int_{T^*}^{\infty} \Phi(\sigma) d\sigma.$$ (5.32)

If we recall that $\Phi(\sigma) = \Phi(T^*) e^{-\lambda (t - T^*)}$, we find that

$$\pi(T^*) = \mu \Phi(T^*) / \lambda.$$ (5.33)

The continuous-time statistical complexity (or the entropy of this mixed random variable) is given by

$$C_\mu = \int_0^{T^*} \mu \Phi(\sigma) \log \frac{1}{\mu \Phi(\sigma)} d\sigma - \frac{\mu \Phi(T^*)}{\lambda} \log \frac{\mu \Phi(T^*)}{\lambda}.$$ (5.34)

This is the sum of the nondivergent component and the rate of divergence of the infinitesimal time-discretized renewal process in Eq. 5.11.

Wave propagation equations, like those in Eq. 5.26, hold true for $\sigma < T^*$.
and for $T^* < \sigma < T^* + \Delta$. At $\sigma = T^*$, probability flows in from both $(T^* + \Delta)^-$ and from $(T^*)^-$, giving rise to the equation

$$\rho(T^*; t + \Delta t) = \rho(T^* - \Delta t; t) + \rho(T^* + \Delta^* - \Delta t; t). \quad (5.35)$$

There is a discontinuous jump in $\rho(\sigma; t)$ at $\sigma = T^*$ coming from $(T^*)^-$ and $(T^* + \Delta^*)^-$, and so we cannot Taylor expand either $\rho(T^* - \Delta t; t)$ or $\rho(T^* + \Delta^* - \Delta t; t)$ about $\Delta t = 0$.

Again, we can use the prescription suggested by Thm. 5 to calculate the probability density function over these causal states, and from that calculate the continuous-time statistical complexity. Below $\sigma < T^*$, the probability density function over causal states is exactly that described in Eq. 5.26–

$$\rho(\sigma) = \sum_{\sigma' : (\sigma - T^*) \mod \Delta^* = \sigma} \mu \Phi(\sigma') \sum_{\sigma'} \Phi(\sigma + i\Delta^*). \quad (5.36)$$

Recalling Def. 3, we see that $\Phi(\sigma + i\Delta^*) = e^{-\lambda i} \Phi(\sigma)$ and so find that for $\sigma > T^*$

$$\rho(\sigma) = \mu \Phi(\sigma) \sum_{i=0}^{\infty} e^{-\lambda i} = \frac{\mu \Phi(\sigma)}{1 - e^{-\lambda}}. \quad (5.37)$$

Altogether, this gives a statistical complexity of

$$C_\mu = \int_0^{T^*} \mu \Phi(\sigma) \log \frac{1}{\mu \Phi(\sigma)} d\sigma + \int_{T^*}^{T^* + \Delta^*} \frac{\mu \Phi(\sigma)}{1 - e^{-\lambda}} \log \frac{1 - e^{-\lambda}}{\mu \Phi(\sigma)} d\sigma. \quad (5.39)$$

**Continuous-time excess entropy**

Prescient machines are adequate for deriving all information measures aside from $C^\pm_\mu$. As such, we focus on the transition dynamics of non-eventually $\Delta$ Poisson $\epsilon$-machines and, implicitly, their bidirectional machines.

To find the joint probability density function of the time to next event $\sigma^-$ and time since last event $\sigma^+$, we note that $\sigma^+ + \sigma^-$ is an inter-event interval; hence,

$$\rho(\sigma^+, \sigma^-) \propto \phi(\sigma^+ + \sigma^-). \quad (5.40)$$

The normalization factor of this distribution is

$$Z = \int_0^{\infty} \int_0^{\infty} \phi(\sigma^+ + \sigma^-) d\sigma^+ d\sigma^- \quad (5.41)$$

$$= \int_0^{\infty} \int_{\sigma^-}^{\infty} \phi(\sigma^+) d\sigma^+ d\sigma^- \quad (5.42)$$

$$= \int_0^{\infty} \sigma^- \phi(\sigma^-) d\sigma^- \quad (5.43)$$

$$= \frac{1}{\mu}. \quad (5.44)$$
So the joint probability distribution is
\[ \rho(\sigma^+, \sigma^-) = \frac{\phi(\sigma^+ + \sigma^-)}{Z} = \mu \phi(\sigma^+ + \sigma^-). \tag{5.45} \]

Equivalently, we could have calculated the conditional probability density function of time to next event given that it has been at least \( \sigma^+ \) since the last event, which (by similar arguments) is \( \phi(\sigma^+ + \sigma^-) \). This would have given the same formula for \( \rho(\sigma^+, \sigma^-) \). To calculate the excess entropy, we merely need calculate (Crutchfield, Ellison, and Mahoney, 2009; Ellison, Mahoney, and Crutchfield, 2009)
\[ E = I[S^+; S^-] \tag{5.46} \]
\[ = \int_0^\infty \int_0^\infty \mu \phi(\sigma^+, \sigma^-) \log \frac{\mu \phi(\sigma^+ + \sigma^-)}{\phi(\sigma^+ + \sigma^-)} d\sigma^+ d\sigma^-, \tag{5.47} \]
and some algebra not shown here gives
\[ E = \int_0^\infty \mu t \phi(t) \log_2 (\mu \phi(t)) dt - 2 \int_0^\infty \mu \Phi(t) \log_2 (\mu \Phi(t)) dt \tag{5.48} \]
Unsurprisingly (Pinsker, 1964), this agrees with the formula derived by considering the limit of infinitesimal time discretization.

### 5.2 Restricted hidden semi-Markov processes

Finite \( \epsilon \)-machines are well-understood, and (from earlier sections) so are maximally predictive models of renewal processes. Here, we introduce a new class of processes which are restricted versions of the hidden semi-Markov models described in Levinson, 1986 but whose dwell time distributions can take any form. (Semi-Markov models are a subset of these restricted hidden semi-Markov models.) Many of the calculations reduce to calculations shown earlier; and when appropriate, we skip those steps.

We start by introducing the restrictions on minimal generative models of the form in Fig. 5.8(top). Let \( \mathcal{G} \) be the set of states in this generative model. Each state \( g \in \mathcal{G} \) emits a symbol, \( x \in A \), and a dwell time for that symbol, \( t \sim \phi_{g,x} \). We assume that the underlying generative model is unifilar and introduce a rather strange restriction on the labeled transition matrices. Define \( \text{supp}(g) := \{ x : p(x|g) > 0 \} \) and \( \text{supp}(g \to g') := \{ x : p(g', x|g) > 0 \} \); then we focus only on generative models for which \( \text{supp}(g) \cap \text{supp}(g \to g') = \emptyset \). This ensures that there is no uncertainty in when one dwell time finishes and another begins. For example, consider Fig. 5.8: if states \( B \) and \( C \) were both to emit a 0 in succession, it would be impossible to tease apart when we had switched from state \( B \) to state \( C \).

A prescient model of this combined process is shown in Fig. 5.8(bottom). Each state \( g \in \mathcal{G} \) comes equipped with one or more renewal process-like tails (“counters”) which, generically, take the form of Fig. 5.5. The leakiness of these (dissipative) counters is given by \( \phi_{g,x} \), the probability density function from which the dwell time is drawn. This is generically the \( \epsilon \)-machine of such a process, but if one or more of the dwell time distributions suggests
Chapter 5. Stimuli to test a sensor’s predictive capabilities

Figure 5.8: At top, a generative model with three hidden states \((A, B, C)\) generates 0’s, 1’s, 2’s, and 3’s for dwell times drawn from probability density functions \(\phi_{B,0}, \phi_{B,1}, \phi_C\) and \(\phi_A\), respectively. At bottom, a prescient machine (and \(\epsilon\)-machine under mild conditions) for the process generated by the left diagram.

The stationary distribution for \(\rho(\tau|g, x)\) exactly follows the treatment for the continuous-time renewal processes earlier, and so

\[
\rho(\tau|g, x) = \mu_{g,x} \Phi_{g,x}(\tau) \quad (5.49)
\]
where $\mu_{g,x} = 1/ \int_0^\infty \Phi_{g,x}(\tau) d\tau$. Then we note that
\[
p(x|g) \propto \frac{T_{g}(x)}{\mu_{g,x}} \rightarrow p(x|g) = \frac{T_{g}(x)}{\sum_{x'} T_{g}(x')/\mu_{g,x'}}.
\] (5.50)
and so
\[
\rho(g, x, \tau) = p(g) \frac{T_g(x)/\mu_{g,x}}{\sum_{x'} T_{g}(x')/\mu_{g,x'}} \mu_{g,x} \Phi_{g,x}(\tau)
\] (5.51)
\[
= p(g) \frac{T_g(x)}{\sum_{x'} T_{g}(x')/\mu_{g,x'}} \Phi_{g,x}(\tau).
\] (5.52)
To find $p(g)$, we again calculate the probability mass dumped at $\tau = 0$ in terms of $p(g)$:
\[
p(g, x, 0) = \sum_{g', x'} \int_0^\infty p(g', x', \tau) \delta_{(g', x'), g} \frac{\phi_{g', x'}(\tau)}{\Phi_{g', x'}(\tau)} T_{g'}(x') d\tau
\] (5.53)
which, after some straightforward substitution of Eq. 5.52 and noting that $\int_0^\infty \phi_{g,x}(\tau) = 1$, we find
\[
\frac{p(g)}{\sum_x T_{g}(x)/\mu_{g,x}} = \sum_{g', x'} \frac{1}{\sum_x T_{g}(x)/\mu_{g,x}} T_{g'}(x') p(g')
\] (5.54)
Let $\pi(g)$ be the stationary distribution for the underlying discrete-state $\epsilon$-machine, $\pi = \text{eig}_1(T_{g', g})$; then
\[
\pi(g) \propto \frac{p(g)}{\sum_x T_{g}(x)/\mu_{g,x}}
\] (5.55)
\[
p(g) = \pi(g) \frac{\sum_x T_{g}(x)/\mu_{g,x}}{\sum_{g', x} \pi(g') T_{g'}(x')/\mu_{g', x}}.
\] (5.56)
Altogether, we find that the steady state distribution is given by
\[
\rho(g, x, \tau) = \left(\pi(g) \frac{\sum_{x'} T_{g}(x')/\mu_{g,x'}}{\sum_{g', x'} \pi(g') T_{g'}(x')/\mu_{g', x'}}\right) \times \left(\frac{T_{g}(x)/\mu_{g,x}}{\sum_{x'} T_{g}(x')/\mu_{g,x'}}\right)
\times (\mu_{g,x} \Phi_{g,x}(\tau))
\] (5.57)
\[
= \frac{\pi(g) T_{g}(x) \Phi_{g,x}(\tau)}{\sum_{g', x'} \pi(g') T_{g'}(x')/\mu_{g', x'}}.
\] (5.58)
The excess entropy of this process can be calculated if we can find the joint probability distribution of forward- and reverse-time causal states, $p(\sigma^+, \sigma^-)$. To this end, we add an additional restriction on the generative model: we focus only on generative models for which $\text{supp}(g') \cap \text{supp}(g \rightarrow g') = \emptyset$. With this restriction on labeled transition matrices, the time-reversed
\(p(\sigma^+, \sigma^-)\) easily:

\[
p(\sigma^+ = (\tau, g, x) | \sigma^- = (g', x', \tau')) = p(\tau | g, x, \tau') \delta_{x, x'} p(g | g', x, x') \phi_{g, x}(\tau + \tau') p(g | g', x) \delta_{x, x'},
\]

where we obtain \(p(g | g', x)\) from standard methods (Ellison, Mahoney, and Crutchfield, 2009; Ellison et al., 2011) applied to (only) the dynamic on \(G\).

Note that \(p(g | g', x', x, \tau')\) reduces to \(p(g | g', x')\) as \(g', x'\) uniquely specify the distribution from which \(\tau'\) is drawn, and since \(x = x'\).

### 5.3 Not just complicated, but complex

In experiments such as Palmer et al., 2015, we seek stimuli which are complex, not just complicated, e.g. natural video. From Bialek, Nemenman, and Tishby, 2001b; Bialek, Nemenman, and Tishby, 2001a, we might say that a process is considered complex when the mutual information between past and future diverges. There is a hierarchy of divergences separating ergodic stationary processes into classes of distinct architecture, depicted in Fig. 5.9. Processes at each level are distinguished by different scalings for their complexity and in how difficult they are to learn and predict.

At the lowest level (Markov) are processes described by finite \(\epsilon\)-machines with finite history dependence (finite Markov order \(R\)); e.g., those described by existing Maximum Caliber models (Pressé et al., 2013) or by measure subshifts of finite type (Lind and Marcus, 1995). Though very commonly posited as models, they inhabit a vanishingly small measure in the space of processes (James et al., 2014). At the next level (Sofic) of structure are processes described by \(\epsilon\)-machines with finite \(C_\mu\). These typically have infinite Markov order; e.g., the measure-sofic processes. Above this level are processes generated by general (that is, nonunifilar) HMMs with uncountable recurrent causal states and divergent statistical complexity that, nonetheless, have finite generative complexity, \(C_{\text{gen}} < \infty\) (Lohr, 2009). Processes at the generative level not only have infinite Markov order and storage, but also require a growing amount of memory for accurate prediction. (See Sec. 4.4.) One consequence is that they are inherently unpredictable by any observer with finite resources. When the smallest generative model is infinite but the process still has short-term memory, we arrive at the class of finitary processes (\(E < \infty\)).

Processes with divergent excess entropy—infinitary processes—inhabit the upper reaches of this hierarchy. Predicting such processes necessarily requires infinite resources, but accurate prediction can also return infinite dividends. We agree, here, with Bialek, Nemenman, and Tishby, 2001b: the asymptotic rate of information divergence is a useful proxy for process complexity.

However, it is as important to know which process mechanism drives the divergence as it is to know the divergence rate. Infinitary Bandit processes store memory entirely in their nonergodic component, as we show
5.3. Not just complicated, but complex

\[ E \rightarrow \infty \]
\[ C_{\text{gen}} \rightarrow \infty \]
\[ C_{\mu} \rightarrow \infty \]
\[ R \rightarrow \infty \]

**Figure 5.9**: Prediction hierarchy for stationary ergodic processes: Each level describes a process class with finite informational quantities. A class above finitely models the processes in the class below. Classes are separated by divergence in the corresponding informational quantity. Moving up the hierarchy corresponds it diverging. Example processes that are finitely presented at each level, but infinitely presented at the preceding lower level. **Sofic**: typical unifilar HMMs, e.g., Even Process (Crutchfield and Feldman, 2003); **Generative**: typical nonunifilar HMMs (Lohr, 2009); **Finitary**: typical infinite nonunifilar HMMs; **Infinitary**: highly atypical infinite HMMs with long memory, e.g. the ergodic construction in Travers and Crutchfield, 2014b.
in Sec. 5.3.1. Generalizing Bandit processes to have structured ergodic components, we now see that even finite $\epsilon$-machines trivially generate infinitary processes when their transition probabilities are continuous random variables. Thus, in this case, we also agree that information divergence is a “necessary but not sufficient” criteria for process complexity Dębowski, 2012b, though see Sec. 5.3.2 for caveats.

Finally, in Sec. 5.3.3, we present an example of a familiar stationary ergodic process—the fractal renewal process—with infinite excess entropy. There are other constructions of stationary ergodic processes with infinite excess entropy, e.g. as in Dębowski, 2012b; Dębowski, 2012a.

### 5.3.1 Complexity from nonergodicity

The simplest construction of a Bandit process is the following. Consider the stochastic process generated by a biased coin whose bias $P$ is itself a random variable. First, a coin bias $p$ is chosen from a user-specified distribution $P(P)$; next, a bi-infinite sequence $x_1 = x_{-1}x_0x_1x_2 \ldots$ is generated from a coin with this particular bias; then, this is repeated for an arbitrarily large number of such trials; generating an ensemble $\{x_1, x_2, x_3, \ldots\}$ of sequences. The process of interest is this sequence ensemble. We denote the random variable block between times $a$ and $b$, but not that at time $b$, as $X_{a:b} = X_aX_{a+1}\ldots X_{b-1}$. We suppress denoting indices that are infinite. To denote the random variable block conditioned on a random variable $Z$ taking realization $z$ we use $X_{a:b}|Z = z$. So here, the subprocess $X|P = p$ is that produced by a coin with bias $p$.

A single one of these bi-infinite sequences comes from an ergodic process that is memoryless in every sense of the word. In particular, since in each trial past and future are independent, the conditional past-future mutual information $I[X_{-M:0}; X_{0:N}|P = p]$ vanishes for any $M$, $N$, and $p$. However, each of these bi-infinite chains is statistically distinct. The mean number of heads, say, in one is very different than the mean number of heads in another. For sufficiently long chains, such differences are almost surely not the consequence of finite-sample fluctuations.

The overall process $X$, does not distinguish between sequences generated by different biased coins. By making the coin bias a random variable, the past and future are no longer independent. Both share information about the underlying coin bias $p$. As we will now rederive (Bialek, Nemenman, and Tishby, 2001b), the shared information or excess entropy $E(M, N) = I[X_{-M:0}; X_{0:N}]$ diverges with $M$ and $N$ when $P$ is a continuous random variable.

To see why, we abstract to a more general case. What follows is an alternative, direct derivation of results in Bialek, Nemenman, and Tishby, 2001b, Sec. 4 that, due to its simplicity, lends additional transparency to the mechanisms driving the divergence.

Let $\Theta$ be a random variable with realizations $\theta$ in a (parameter) space of dimension $K$. $\Theta$ has some as-of-yet unspecified relationship with observations $X_1 = \ldots X_{-2}, X_{-1}, X_0, X_1, \ldots$. We can always perform the following
information-theoretic decomposition of the composite process’s excess entropy:

\[
I[X_{M:0}; X_{0:N}] = I[X_{M:0}; X_{0:N}|\Theta] \\
+ I[X_{M:0}; X_{0:N}; \Theta].
\]  

(5.59)

The first term quantifies the range of temporal correlations of the observed process \( \text{given } \Theta \), and the second term quantifies the dependencies between past and future purely due to \( \Theta \). When the fixed-parameter process \( X.|\Theta = \theta \) is ergodic and the composite process \( X. \) is not, then Eq. (5.59) can be viewed as a decomposition of \( I[X_{M:0}; X_{0:N}] \) into ergodic and nonergodic contributions, respectively.

The second term \( I[X_{M:0}; X_{0:N}; \Theta] \) is a multivariate mutual information (James, Ellison, and Crutchfield, 2011) or co-information (Bell, 2003). It is closely related to parameter estimation, as expected (Bialek, Nemenman, and Tishby, 2001b), since it provides information about the dimension \( K \) of \( \Theta \). Standard information-theoretic identities yield:

\[
I[X_{M:0}; X_{0:N}; \Theta] = H[\Theta] + H[\Theta|X_{M:N}] \\
- H[\Theta|X_{M:0}] - H[\Theta|X_{0:N}].
\]  

(5.60)

The first term \( H[\Theta] \) quantifies our intrinsic uncertainty in the bias. When \( \Theta \) is a continuous random variable, \( H[\Theta] \) is a differential entropy. The subsequent terms describe how our uncertainty in \( \Theta \) decreases after seeing blocks of lengths \( M + N, M, \) or \( N \).

Altogether, Eqs. (5.59) and (5.60) give:

\[
I[X_{M:0}; X_{0:N}] = I[X_{M:0}; X_{0:N}|\Theta] + H[\Theta] \\
+ H[\Theta|X_{M:N}] - H[\Theta|X_{M:0}] \\
- H[\Theta|X_{0:N}].
\]  

(5.61)

Thus, assuming one chose a prior with finite entropy \( H[\Theta] \), divergences in \( I[X_{M:0}; X_{0:N}] \) can come from divergences in \( I[X_{M:0}; X_{0:N}|\Theta] \) or from divergences in \( H[\Theta|X_{M:N}] - H[\Theta|X_{M:0}] - H[\Theta|X_{0:N}] \).

Let’s take the cases covered in Bialek, Nemenman, and Tishby, 2001b, Secs. 4.1-4.4. There, \( \Theta \) consists of the model parameters, \( \theta \) are realizations of \( \Theta \), and \( X.|\Theta = \theta \) consists of (noisy, potentially temporally correlated) sequences generated by the model with parameters \( \theta \). For instance, \( \Theta \) could be the firing rate of a Poisson neuron and \( X.|\Theta = \theta \) could be the time-binned spike trains at firing rate \( \theta \). Or, \( \Theta \) could be transition probabilities in a finite Hidden Markov Model (HMM) and \( X.|\Theta = \theta \) could be the generated process given transition probabilities \( \theta \). The result, in any case, is a nonergodic process \( X. \) constructed from a mixture of ergodic component processes \( X.|\Theta = \theta \).

In these examples, the component-process excess entropy \( I[X_{M:0}; X_{0:N}|\Theta] = \langle I[X_{M:0}; X_{0:N}|\Theta = \theta] \rangle_{\theta} \) does not diverge with \( M \) or \( N \), since finite HMMs have finite excess entropy, which is bounded from above by the internal state entropy (Travers and Crutchfield, 2014b; Lohr, 2009). In fact, the excess entropy for many ergodic stochastic processes is finite, even if generated by infinite-state HMMs. Most of the time, any divergence in the composite process \( I[X_{M:0}; X_{0:N}] \) therefore comes from divergences in \( H[\Theta|X_{M:N}] \).
\[ H[\Theta|X_{-M:0}] - H[\Theta|X_{0:N}] \]

Since the composite process includes sequences \( x^t \) from trials with different \( \theta \), one’s intuition might suggest that \( P(\Theta = \theta|X_{-M:0} = x_{-M:0}) \) is multimodal for most \( x_{-M:0} \). However, existing results (Walker, 1969; Heyde and Johnstone, 1979; Sweeting, 1992; Weng and Tsai, 2008) on asymptotic normality of posteriors carry over to this setting, since they essentially rely on the log-likelihood function \( \log P(X_{-M:0} = x_{-M:0}|\Theta = \theta) \) being sufficiently well-behaved.

For instance, consider the previously described Bandit process construction. A crude derivation of the asymptotic normality of \( P(\Theta = \theta|X_{-M:0} = x_{-M:0}) \) (Hartigan, 1983) starts with Bayes Rule:

\[
P(\Theta = \theta|X_{-M:0} = x_{-M:0}) = \frac{P(|X_{-M:0} = x_{-M:0}|\Theta = \theta)P(\Theta = \theta)}{P(X_{-M:0} = x_{-M:0})}.
\]

The denominator \( P(X_{-M:0} = x_{-M:0}) \) is quite complicated to calculate, but this normalization factor does not affect the \( \theta \)-dependence of \( P(\Theta = \theta|X_{-M:0} = x_{-M:0}) \). More to the point, the prior’s contribution \( P(\Theta = \theta) \) is dwarfed by the likelihood:

\[
P(|X_{-M:0} = x_{-M:0}|\Theta = \theta) = \theta^{\sum_{i=0}^{M-1} x_i} (1 - \theta)^{M - \sum_{i=0}^{M-1} x_i},
\]

in the large-\( M \) limit. Let \( \theta^* \) be the unique maximum of \( P(\Theta = \theta|X_{-M:0} = x_{-M:0}) \): \( \theta^* = \frac{1}{M} \sum_{i=0}^{M-1} x_i + O(1/M) \). Taylor-expanding \( \log P(\Theta = \theta|X_{-M:0} = x_{-M:0}) \) about \( \theta^* \) suggests that \( P(\Theta = \theta|X_{-M:0} = x_{-M:0}) \) is approximately normal in the large-\( M \) limit, with variance decaying as \( \sim 1/M \). Any one of the many sources (Walker, 1969; Heyde and Johnstone, 1979; Sweeting, 1992; Weng and Tsai, 2008) on asymptotic normality of posteriors provides rigorous and generalized statements.

Armed with such asymptotic normality, we now turn our attention to find the asymptotic form of \( H[\Theta|X_{-M:0} = x_{-M:0}], H[\Theta|X_{0:N} = x_{0:N}] \), and \( H[\Theta|X_{-M:N} = x_{-M:N}] \) in the large-\( M \) and \( -N \) limits. The differential entropy of a normal distribution is \( \frac{1}{2} \log |\det \Sigma| \), where \( \Sigma \) is the covariance matrix; here, \( \det \Sigma \sim 1/M \). This captures the error distribution for each of the \( K \) parameters. So, this and asymptotic normality of the posterior imply that:

\[
H[\Theta|X_{-M:0} = x_{-M:0}] \sim -\frac{K}{2} \log M,
\]

plus corrections of \( O(1) \) in \( M \), and thus:

\[
H[\Theta|X_{-M:0}] \sim -\frac{K}{2} \log M,
\]

where \( K \) is the parameter space dimension.

At first blush, the result is counterintuitive. In the limit that \( M \) and \( N \) tend to infinity, and we see longer and longer sequences \( x_{-M:0} \), we become more certain as to \( \Theta \)’s value. This increasing certainty should mean that the conditional entropy \( H[\Theta|X_{-M:0} = x_{-M:0}] \) vanishes. However, if \( \Theta \) is
a continuous random variable (such as a Poisson rate), then $H[\Theta | X_{-M:0} = x_{-M:0}]$ is a differential entropy. As our variance in $\Theta | X_{-M:0} = x_{-M:0}$ decreases to 0, the differential entropy $H[\Theta | X_{-M:0} = x_{-M:0}]$ diverges to negative infinity. It is exactly this well-known divergence that causes a divergence in $I[X_{-M:0}; X_{0:N}]$ for the nonergodic processes we are considering.

From these results and Eq. (5.61), one has:

$$I[X_{-M:0}; X_{0:N}; \Theta] \sim K \frac{1}{2} \log \frac{MN}{M+N}.$$  

And, recalling that the ergodic-component information does not diverge, we immediately recover:

$$I[X_{-M:0}; X_{0:N}] \sim K \frac{1}{2} \log \frac{MN}{M+N}. \quad (5.62)$$

Lower-order terms in $M$ and $N$ include the expected log-determinant of the Fisher information matrix for maximum likelihood estimates of $\Theta$ (Lehman and Casella, 1998).

Logarithmic divergences in excess entropy also occur in stationary ergodic processes, such as exhibited at the onset of chaos through period-doubling (Crutchfield and Young, 1990). Alternative scalings are known, such as power-law divergences (Bialek, Nemenman, and Tishby, 2001b, Sec. 4.5). For natural language texts there is empirical evidence that the excess entropy diverges. One form is referred to as Hilberg’s Law (Ebeling and Nicolis, 1991; Ebeling and Poschel, 1994; Dębowski, 2011):

$$I[X_{-N:0}; X_{0:N}] \propto \sqrt{N}.$$  

In contrast with the previous direct calculation, it is far less straightforward to analyze these power-law divergences:

$$I[X_{0:N}; X_{0:N}] \sim K \frac{1}{2} N^{\gamma}, \quad (5.63)$$

with $\gamma \in [0, 1)$. While there are results on asymptotics of posteriors for nonparametric Bayesian inference, many aim to establish asymptotic normality of the posterior; e.g., as in Wasserman, 1998; Ghosal, 2000. As far as we know, existing asymptotic analyses avoid the essential singularity for the prior utilized in Bialek, Nemenman, and Tishby, 2001b, Sec. 4.5 to obtain power-law divergence.

### 5.3.2 Complexity without infinite excess entropy

A similar information-theoretic decomposition to that of Sec. 5.3.1 can be used to upper-bound the excess entropy of ergodic processes as well. For instance, we can use a similar decomposition to show that excess entropy of an Ising spin on a two-dimensional Ising lattice at criticality is finite.

Dębowski, 2012b pointed out that many infinitary processes do not satisfy intuitive definitions for complexity. It suggested that divergence in $E$ is a “necessary but not sufficient condition” for a process being truly complex. While intuitively compelling, perhaps divergent $E$ is not even a necessary condition.

Spin systems at criticality are one of the most familiar examples of truly complex processes: global correlations emerge from purely local interactions (Binney et al., 1992). Evidence of this complexity appears even if we
are only allowed to observe a single spin’s interaction with another on the lattice. At the critical temperature, the interaction has a power-law autocorrelation function; at all other temperatures, the spin’s autocorrelation function is asymptotically exponential. However, does the temporal excess entropy \( E(M, N) \)—roughly, the interaction a single spin with itself at later times—also diverge at criticality?

Surprisingly, the excess entropy of the dynamics of a single spin on an Ising lattice is finite, even at the critical temperature, unless there are nonlocal spatial interactions between lattice spins. Consider evolving the lattice configurations via Glauber dynamics for concreteness (Binney et al., 1992). That is, spin \( \sigma_j(t) \)'s next state \( \sigma_j(t+1) \) is determined stochastically by its previous state \( \sigma_j(t) \) and its effective magnetic field \( h_j(t) = \sum_i J_{ij} \sigma_i(t) \). In other words, \( h_j(t) \) and \( \sigma_j(t) \) causally shield the past \( \sigma_j(t) \) from the future \( \sigma_j(t+1) \), implying that:

\[
I[\sigma_j(t+1; t); \sigma_j(t) | h_j(t)] = I[\sigma_j(t+1; t); \sigma_j(t) | h_j(t)] 
\leq H[\sigma_j(t)],
\]

Given a finite set of spin values and local interactions, \( h_j(t) \) can only take a finite number of values. Thus, \( H[h_j(t)] < \infty \), and so:

\[
|I[\sigma_j(t-M; t+M); h_j(t)]| \leq H[h_j(t)] < \infty,
\]

as well.

A more familiar example makes this concrete. For the standard two-dimensional Ising lattice \( J_{ij} = J \), if \( i \) and \( j \) are nearest neighbors, and \( J_{ij} = 0 \), otherwise. There, \( h_j(t) \) can only take 5 possible values—\( h_j(t) \in \{0, J, 2J, 3J, and 4J\} \)—giving:

\[
|I[\sigma_j(t-M; t+M); h_j(t)]| \leq H[h_j(t)] \leq \log_2 5 \text{ bits}.
\]

The information-theoretic decomposition in Eq. (5.59) applies in this particular situation. Here, observed variables \( X_t \) are spins \( \sigma_t \), and the parameters \( \Theta \) are replaced by \( h_j(t) \). The bounds above then directly imply that \( E(M, N) < \infty \) for all \( M \) and \( N \). In fact, for the standard two-dimensional Ising lattice, we find that \( E(-\infty, \infty) \leq 1 + \log_2 5 = 3.4 \text{ bits} \). We expect excess entropy to diverge only when \( h_j(t) \) is a continuous random variable. This can happen when \( J_{ij} \) is nonzero for an infinite number of \( i \)'s. However, this necessitates global, not local, spin-spin couplings.

On the one hand, this analysis does not negate \( E \)'s utility as a generalized order parameter. It is still likely maximized at the critical point, even if its temporal version does not diverge (Feldman, 1998). On the other, our analysis shows that phenomena—here, spin lattices with purely local couplings—do not necessarily have divergent \( E \) even when many would consider their dynamics to be truly complex when the system is critical.

At first glance, the analysis contradicts the experiments in Fig. 1 of Bialek, Nemenman, and Tishby, 2001b for the Ising lattice with only local interactions. A more careful look reveals that there is no contradiction.
5.3. Not just complicated, but complex

at all. There, coupling strengths were randomly changed every 4000 iterations, so the resultant time series looked like a concatenation of samples from a Bandit process. The analysis in Sec. 5.3.1 then exactly predicts the observed logarithmic scaling in Fig. 1 for \( N \leq 25 \). However, it also implies that \( \mathbf{E}(-\infty, N) \) will stop increasing logarithmically at or before \( N = 4000 \).

5.3.3 Complexity in fractal renewal processes

Surprisingly, even some renewal processes are also complex in the sense of Bialek, Nemenman, and Tishby, 2001b; Bialek, Nemenman, and Tishby, 2001a, in that they can have infinite \( \mathbf{E} \). Fractal renewal processes—those with power-law interevent interval probability density functions—can have long memory in the sense of Graves et al., 2014, as they can have a Hurst exponent greater than \( H > 1/2 \) (Daley, 1999) and their autocorrelation function can be (asymptotically) a power law (Lowen and Teich, 1993b). Fractal renewal processes have been implicated in a variety of complex natural processes (Lowen and Teich, 1993a; Thurner et al., 1997; Cakir, Grigolini, and Krokhin, 2006; Bianco et al., 2007; Li, Yang, and Komatsuzaki, 2008; Akimoto, Hasumi, and Aizawa, 2010; Kelly et al., 2012; Montero and Villarroel, 2013; Bologna, West, and Grigolini, 2013; Onaga and Shinomoto, 2014). Might these processes also have infinite statistical complexity or infinite excess entropy? To the best of our knowledge, the excess entropy and statistical complexity of fractal renewal processes have yet to be calculated.

Calculating statistical complexity and excess entropy can be challenging when going beyond finite causal-state processes (Crutchfield, Riechers, and Ellison, 2016). To make progress with bounding the excess entropy of fractal renewal processes, we use two tools. The first tool is to coarse grain by time-binning. The Data Processing Inequality (Cover and Thomas, 2006, Ch. 2.8) then implies that the excess entropy of a discrete-time renewal process is always upper-bounded by the excess entropy of the corresponding continuous-time renewal process, and integrals are often easier to evaluate than the corresponding sums. One practical goal, leveraging this below, is to relate the excess entropy of time-binned continuous-time processes to that of corresponding discrete-time renewal processes.

Recall Eq. 5.5 for the excess entropy of a discrete-time renewal process:

\[
\mathbf{E} = \log(\mu + 1) - \frac{2}{\mu + 1} \sum_{n=0}^{\infty} w(n) \log w(n) \\
+ \frac{1}{\mu + 1} \sum_{n=0}^{\infty} (n + 1)F(n) \log F(n) 
\]

and recall Eq. 5.48 for the excess entropy of a continuous-time renewal process:

\[
\hat{\mathbf{E}} = I[X(t)_{t<0}; X(t)_{t\geq0}] \\
= \log T - \frac{2}{T} \int_{0}^{\infty} \Phi(t) \log \Phi(t) dt \\
+ \frac{1}{T} \int_{0}^{\infty} t\phi(t) \log \phi(t) dt . \]

(5.65)
Both are in units of nats and hold only when the mean interevent interval is finite.

Consider time-binning the continuous-time point process \( X(t) \) by asking how many events are observed in an interval \([t, t + \tau)\). If at least one event is observed, then we record a 1; if no events are observed, then we record a 0. This data labeling technique is common, e.g., when studying neural spike trains. The probability of observing at least \( n \) counts between successive 1s is given by:

\[ w_\tau(n) = \Phi(n\tau) . \]

When \( \tau = 1 \), then the survival function of the time-binned process is exactly that of the discrete-time renewal process with excess entropy given in Eq. 5.64.

The excess entropy or estimates thereof for a discrete-time renewal process are upper bounded by the excess entropy of a corresponding continuous-time renewal process, as shown shortly. This is a special case of a more general statement: coarse-graining a time series always reduces its excess entropy, due to the Data Processing Inequality. This statement can be easily generalized to other discrete-alphabet, continuous-time processes. Despite its simplicity, it proves useful for the calculations to come.

In particular, let \( \hat{E} \) denote the excess entropy of a continuous-time renewal process \( X(t) \) with survival function \( \Phi(t) \) and \( E \) the excess entropy of the discrete-time renewal process \( X_t \) with survival function \( w(n) = \Phi(n) \) for all nonnegative integers \( n \). Then, when \( \hat{E} < \infty \):

\[ E \leq \hat{E} . \]

To see this, let \( E_\tau \) denote the excess entropy of the discrete-time process that comes from time-binning the continuous-time renewal process with discretization bin size \( \tau \). To obtain the above inequality, we apply the Data Processing Inequality:

\[ E_{1/n} = I[\ldots, X(-2/n), X(-1/n); X(0), X(1/n), \ldots] \]
\[ \geq I[\ldots, X_{-2}, X_{-1}; X_{0}, X_{1}, \ldots] \]
\[ = E_{1} . \]

If we take the limit of the left-hand side as \( n \to \infty \), we obtain:

\[ E_{\tau=1} \leq \lim_{n \to \infty} E_{1/n} = \lim_{\tau \to 0} E_\tau . \]

Again by the Data Processing Inequality, \( E_{\tau=1} \) is lower-bounded by the mutual information between the counts since last event and counts to next event, as the former is a function of the past and the latter is a function of the future: \( E \leq E_{\tau=1} \). By definition (Pinsker, 1964), \( \lim_{\tau \to 0} E_\tau = \hat{E} \).

The second tool allows us to calculate excess entropy and statistical complexity even when the mean rate of events vanishes by conditioning on the presence of a proxy event. This tool was inspired by previous work (Travers and Crutchfield, 2014b) and is summarized below.
When the mean interevent interval $T$ (or $\mu$) is infinite, the formulae for excess entropy in Eqs. (5.64) and (5.65) no longer apply. Causal states, however, still provide a useful framework for calculating it. Using them we introduce an analysis method for discrete-time renewal processes in this case. Straightforward extensions to continuous-time renewal processes follow when we replace $F(n)$ with $\phi(t)$, $w(n)$ with $\Phi(t)$, and summations with integrals.

We calculate $E(\ell)$ for renewal processes with infinite $\mu$ via an analysis technique inspired by Travers and Crutchfield, 2014b and then calculate $E$ as a limit of $E(\ell)$ as $\ell$ tends to infinity, which (to the best of our knowledge) is valid for ergodic processes. First, we would like to directly calculate $E(\ell)$ in terms of forward and reverse-time causal states (Crutchfield, Ellison, and Mahoney, 2009): $E(\ell) = I[X_0; \overline{X}^L] = I[S^+; S^-_{\ell}]$, where $S^-_{\ell}$ are finite-time reverse-time causal states. Unfortunately, inspecting the corresponding joint probability distribution shows that while we can identify the joint probability distribution up to a normalization constant, this normalization constant is infinite when $\mu$ is infinite.

So, we define a “proxy” binary random variable $U_\ell$ which is 1 if there has been an event sometime in $\overline{X}^L$ and past $X_0$, and 0 otherwise. A little reflection shows that $P(U_\ell = 0) = \lim_{N \to \infty} w(N + \ell) = 0$. Even so, this auxiliary random variable is a surprisingly useful construct. A standard information-theoretic decomposition gives $E(\ell) = I[S^+; S^-_{\ell} | U_\ell = 1] + I[S^+; S^-_{\ell} | U_\ell = 0], but since $P(U_\ell = 0) = 0$, we have that $I[S^+; S^-_{\ell} | U_\ell = 0] = I[S^+; S^-_{\ell} | U_\ell = 1]$. Altogether this yields:

$$E(\ell) = I[S^+; S^-_{\ell} | U_\ell = 1].$$

The conditional probability distribution $P(S^+, S^-_{\ell} | U_\ell = 1)$ is normalizable. From the joint probability distribution over forward- and reverse-time causal states:

$$P(S^+ = \sigma^+, S^-_{\ell} = \sigma^- | U_\ell = 1) = \frac{1}{Z} \begin{cases} F(\sigma^+ + \sigma^-) & \sigma^- \leq \ell \\ 0 & \sigma^- = \ell + 1 \end{cases},$$

where the normalization constant is:

$$Z(\ell) = \sum_{\sigma^- = 0}^{\ell} \sum_{\sigma^+ = 0}^{\infty} F(\sigma^+ + \sigma^-)$$

$$= \sum_{\sigma^- = 0}^{\ell} w(\sigma^-).$$

The marginals are easily calculated:

$$P(S^+ = \sigma^+ | U_\ell = 1) = \frac{1}{Z} (w(\sigma^+) - w(\sigma^+ + \ell + 1))$$
and:

\[ P(S^-_\ell = \sigma^- | U_\ell = 1) = \frac{1}{Z} \begin{cases} w(\sigma^-) & \sigma^- \leq \ell \\ 0 & \sigma^- = \ell + 1 \end{cases} \]

From this, we calculate finite-length excess entropy in nats:

\[
\mathbf{E}(\ell) = H[S^-_\ell | U_\ell = 1] + H[S^+_\ell | U_\ell = 1] - H[S^+_\ell, S^-_\ell | U_\ell = 1]
= \log Z(\ell) - \frac{1}{Z(\ell)} \sum_{n=0}^{\ell} w(n) \log w(n) - \frac{1}{Z(\ell)} \left( \sum_{n=0}^{\infty} (w(n) - w(n + \ell + 1)) \right) 
\times \log(w(n) - w(n + \ell + 1)) + \frac{1}{Z(\ell)} \sum_{n=0}^{\infty} \sum_{m=0}^{\ell} F(n + m) \log F(n + m) .
\]

This simplifies to:

\[
\mathbf{E}(\ell) = \log Z(\ell) - \frac{1}{Z(\ell)} \sum_{n=0}^{\ell} w(n) \log w(n) - \frac{1}{Z(\ell)} \left( \sum_{n=0}^{\infty} (w(n) - w(n + \ell + 1)) \right) 
\times \log(w(n) - w(n + \ell + 1)) + \frac{1}{Z(\ell)} \sum_{n=0}^{\ell} (n + 1) F(n) \log F(n) 
+ \frac{\ell + 1}{Z(\ell)} \sum_{n=\ell+1}^{\infty} F(n) \log F(n) .
\]

Similar manipulations hold for continuous-time processes; as described earlier in Sec. 5.1.3, the time since last event \( t \) and time to next event \( t' \) have a joint probability distribution proportional to \( \phi(t + t') \), since the time since last event plus the time to next event is an interevent interval.

If \( \lim_{t' \to \infty} \mathbf{E}(t') \) diverges, then we look for the asymptotic rate of divergence of \( \mathbf{E}(t') \). Otherwise, the process’ excess entropy can be defined as \( \mathbf{E} = \lim_{t' \to \infty} \mathbf{E}(t') \). We expect \( \mathbf{E} \) will often be finite even when \( \mu \) diverges.

A similar method allows us to calculate \( C_\mu \) when mean interevent count is infinite. This time, we define \( U_\ell \) as a proxy random variable that is 1 if there has been an event in \( X^L \) and 0 otherwise. Since \( U_\ell \) is a function of \( S^+ \), a standard information-theoretic identity implies that:

\[ C_\mu = H[S^+_\ell | U_\ell] + H[U_\ell] \quad (5.66) \]

and, in particular:

\[ C_\mu = \lim_{\ell \to \infty} (H[S^+_\ell | U_\ell] + H[U_\ell]) . \quad (5.67) \]

As before, \( \lim_{\ell \to \infty} P(U_\ell = 0) = \lim_{\ell \to \infty} w(\ell) = 0 \), so \( \lim_{\ell \to \infty} H[U_\ell] = 0 \). Also, \( H[S^+_\ell | U_\ell] = P(U_\ell = 0)H[S^+_\ell | U_\ell = 0] + P(U_\ell = 1)H[S^+_\ell | U_\ell = 1] \)
by definition. Since there is only one semi-infinite past without an event, \( \lim_{\ell \to \infty} H[S^+_\ell | U_\ell = 0] = 0 \). And, \( H[S^+_\ell | U_\ell = 1] = -\sum_{n=0}^{\ell} \frac{w(n)}{Z(\ell)} \log \frac{w(n)}{Z(\ell)} \).

Altogether, this implies:

\[ C_\mu = \lim_{\ell \to \infty} \sum_{n=0}^{\ell} \frac{w(n)}{Z(\ell)} \log \left( \frac{1}{w(n)} \frac{w(n)}{Z(\ell)} \right) . \quad (5.68) \]
5.3. Not just complicated, but complex

One can also study the growth rate of finite-time statistical complexity estimates which are, after a moment’s reflection, the $C_\mu^\ell = -\sum_{n=0}^{\ell} \frac{w(n)}{Z(t)} \log \frac{w(n)}{Z(t)}$ estimates above in Eq. 5.68.

One comment, apparent upon inspection of Eqs. 5.66 and 5.68, is that whether or not $\mathbb{E}$ and $C_\mu$ diverge depends entirely on the asymptotic form of $F(n)$. Another is that the sums in Eq. 5.66 can be quite difficult to evaluate numerically when the renewal process has long-range temporal correlations, since then $F(n)$ decays slowly with $n$.

Fractal renewal processes are typically considered in continuous-time, with interevent intervals generated independently and identically distributed (IID) from the probability density function:

$$\phi(t) = \begin{cases} 0 & t < 1 \\ \alpha t^{-(\alpha+1)} & t \geq 1 \end{cases}. \quad (5.69)$$

The probability of seeing an interevent interval of length $t$ or larger is the survival function:

$$\Phi(t) = \int_t^\infty \phi(t')dt' = \begin{cases} 1 & t < 1 \\ t^{-\alpha} & t \geq 1 \end{cases}. \quad (5.70)$$

Time intervals are given in terms of the shortest possible interevent interval. When $\alpha > 1$, the mean interevent interval $T = \frac{\alpha}{\alpha-1}$ is finite; when $0 < \alpha \leq 1$, the mean interevent interval is infinite, but one will always eventually see an event.

The $\alpha > 1$ case simply requires substituting $\phi(t)$ and $\Phi(t)$ from Eqs. (5.69)-(5.70) into Eq. (5.65) and solving:

$$\hat{\mathbb{E}} = \log T - \frac{2}{\alpha} \int_0^\infty \Phi(t) \log \Phi(t)dt + \frac{1}{\alpha} \int_0^\infty t\phi(t) \log \phi(t)dt. \quad (5.71)$$

After straightforward calculations, we find that:

$$\frac{1}{T} \int_0^\infty \Phi(t) \log \Phi(t)dt = -\frac{1}{\alpha - 1},$$

$$\frac{1}{T} \int_0^\infty t\phi(t) \log \phi(t)dt = \log \alpha - \frac{\alpha + 1}{\alpha - 1}.$$ 

These together yield:

$$\hat{\mathbb{E}} = \log \frac{\alpha^2}{\alpha - 1} - 1.$$ 

Now, we turn our attention to the case of $0 < \alpha \leq 1$. There are two possibilities for $\hat{\mathbb{E}}$ when $0 < \alpha \leq 1$. One is that $\hat{\mathbb{E}}$ diverges, in which case, we only care about the asymptotic rate of divergence of $\hat{\mathbb{E}}(\ell)$. The other possibility is that $\hat{\mathbb{E}}$ does not diverge, in which case, we only care about contributions $Q$ to $\hat{\mathbb{E}}(\ell)$ that are not $o(1)$; i.e., that satisfy $\lim_{\ell \to \infty} Q \neq 0$. Our
strategy in evaluating $\hat{E}(\ell)$ from Eq. 5.71 is to systematically find closed-form expressions for all components that are not $o(1)$.

Direct solution gives:

$$Z = \begin{cases} \frac{\ell^{1-\alpha}}{1-\alpha} & \alpha < 1 \\ \log \ell & \alpha = 1 \end{cases}, \tag{5.72}$$

plus components of $o(1)$:

$$-\frac{1}{Z} \int_0^\ell \Phi(t) \log \Phi(t) dt = \begin{cases} \frac{\alpha}{1-\alpha} + \alpha \log \ell & \alpha < 1 \\ \frac{1}{2} \log \ell & \alpha = 1 \end{cases} \tag{5.73}$$

plus components of $o(1)$; and:

$$\frac{1}{Z} \int_1^\ell t \phi(t) \log \phi(t) dt + \frac{\ell}{Z} \int_\ell^\infty \phi(t) \log \phi(t) dt = \begin{cases} \frac{-1-\alpha-2\alpha^2}{\alpha(1-\alpha)} + \log \alpha - (1+\alpha) \log \ell & \alpha < 1 \\ -2 - \log \ell & \alpha = 1 \end{cases}, \tag{5.74}$$

plus components of $o(1)$.

Finally, we address the only component with no simple closed-form expression:

$$\frac{1}{Z} \int_0^\infty \left(\Phi(t) - \Phi(t + \ell)\right) \log(\Phi(t) - \Phi(t + \ell)) dt$$

$$= \frac{1}{Z} \int_1^\infty (t^{-\alpha} - (t + \ell)^{-\alpha}) \log(t^{-\alpha} - (t + \ell)^{-\alpha}) dt$$

$$+ \frac{1}{Z} \int_0^1 (1 - (t + \ell)^{-\alpha}) \log(1 - (t + \ell)^{-\alpha}) dt.$$

Since:

$$\lim_{\ell \to \infty} \frac{1}{Z} \int_0^1 (1 - (t + \ell)^{-\alpha}) \log(1 - (t + \ell)^{-\alpha}) dt = 0,$$

we ignore that term as a correction of $o(1)$. The case for $\alpha = 1$ can actually be evaluated explicitly since $\frac{1}{t} - \frac{1}{t+\ell} = \frac{\ell}{t(t+\ell)}$:

$$\lim_{\ell \to \infty} \frac{1}{Z} \int_1^\infty \frac{\ell}{t(t+\ell)} \log(\frac{\ell}{t(t+\ell)}) dt = \frac{1}{2} \log \ell.$$
5.3. Not just complicated, but complex

Now, consider the case of $\alpha < 1$. We extract the asymptotic scaling in $\ell$ of the first term by the change of variables $u = \ell t$, giving:

\[
\frac{1}{Z} \int_1^\infty (t^{-\alpha} - (t + \ell)^{-\alpha}) \log(t^{-\alpha} - (t + \ell)^{-\alpha}) dt
= \frac{\ell^{1-\alpha}}{Z} \int_{1/\ell}^\infty (u^{-\alpha} - (1 + u)^{-\alpha}) \log((\ell^{-\alpha}(u^{-\alpha} - (1 + u)^{-\alpha})) du
= -\alpha \frac{\ell^{1-\alpha} \log \ell}{Z} \int_{1/\ell}^\infty u^{-\alpha} - (1 + u)^{-\alpha} du
+ \frac{\ell^{1-\alpha}}{Z} \int_{1/\ell}^\infty (u^{-\alpha} - (1 + u)^{-\alpha}) \log(u^{-\alpha} - (1 + u)^{-\alpha}) du.
\]

The first of the two integrals can be evaluated explicitly as:

\[
\int_{1/\ell}^\infty \frac{u^{-\alpha} - (1 + u)^{-\alpha}}{\ell^{-\alpha}} du = -\frac{\ell^{\alpha-1}}{1-\alpha} + \frac{\ell^{\alpha-1}}{1-\alpha} \ell(1 + 1)^{1-\alpha}
= -\frac{\ell^{-\alpha}}{1-\alpha} \int_{1/\ell}^\infty u^{-\alpha} - (1 + u)^{-\alpha} du \sim -\alpha \log \ell,
\]

We therefore find the first term’s asymptotic behavior to be:

\[-\alpha \frac{\ell^{1-\alpha} \log \ell}{Z} \int_{1/\ell}^\infty u^{-\alpha} - (1 + u)^{-\alpha} du \sim -\alpha \log \ell,
\]

plus corrections of $o(1)$. One of the more notable corrections of $o(1)$ is proportional to $\log \ell$, which is $o(1)$ for $\alpha < 1$ and otherwise has a nonzero limiting value when $\ell \to \infty$.

Surprisingly, the latter of the two integrals limits to a finite value for $\alpha < 1$:

\[
\lim_{\ell \to \infty} \frac{\ell^{1-\alpha}}{Z} \int_{1/\ell}^\infty (u^{-\alpha} - (1 + u)^{-\alpha}) \log(u^{-\alpha} - (1 + u)^{-\alpha}) du
= (1 - \alpha) \int_0^\infty (u^{-\alpha} - (1 + u)^{-\alpha}) \log(u^{-\alpha} - (1 + u)^{-\alpha}) du,
\]

where we used $\lim_{\ell \to \infty} \frac{\ell^{1-\alpha}}{Z} = 1 - \alpha$ for $\alpha < 1$. As a result, we find that:

\[
Q = \frac{1}{Z} \int_0^\infty (\Phi(t) - \Phi(t + \ell)) \log(\Phi(t) - \Phi(t + \ell)) dt
= \begin{cases} 
-\frac{1}{2} \log \ell & \alpha = 1 \\
-\alpha \log \ell + (1 - \alpha) \int_0^\infty (u^{-\alpha} - (1 + u)^{-\alpha}) \log(u^{-\alpha} - (1 + u)^{-\alpha}) du & 0 < \alpha < 1
\end{cases},
\]

(5.75)

plus corrections of $o(1)$. Altogether, combining Eqs. 5.72-5.74 and 5.75 into Eq. 5.71, we find Eq. 5.76.

As discussed there, we still must evaluate $E(\ell)$ at $\alpha = 1$. We focus again on asymptotic expansions in $\ell$ and drop corrections to expressions that do not contribute to $E$. When $\alpha = 1$:

\[
Z(\ell) = 1 + \sum_{n=1}^\ell \frac{1}{n} = \log \ell,
\]
plus corrections of $O(1)$. Next, we evaluate:

$$- \sum_{n=0}^{\ell} w(n) \log w(n) = \sum_{n=1}^{\ell} \log n \frac{\log n}{n} = \sum_{n=2}^{\ell} \log n \frac{\log n}{n}.$$ 

Since $\frac{\log n}{n}$ is a monotone decreasing function with $n$, we lower- and upper-bound this sum using integrals: 

$$\int_{2}^{\ell+1} \frac{\log n}{n} dn \leq \sum_{n=2}^{\ell} \frac{\log n}{n} \leq \log 2 + \int_{2}^{\ell} \frac{\log n}{n} dn.$$ 

These are easily evaluated, giving:

$$- \sum_{n=0}^{\ell} w(n) \log w(n) = - \frac{1}{2} \log^2 \ell,$$

plus corrections of $O(1)$. For other sums, we need an expression for $F(n)$:

$$F(n) = w(n) - w(n+1) = \begin{cases} 0 & n = 0 \\ \frac{1}{n(n+1)} & n \geq 1 \end{cases}.$$ 

Then, we evaluate:

$$\sum_{n=0}^{\ell} (n+1) F(n) \log F(n) = -2 \sum_{n=1}^{\ell} \frac{\log n}{n} + \sum_{n=1}^{\ell} \frac{\log(1 + \frac{1}{n})}{n} = \log^2 \ell,$$

plus corrections of $O(1)$, where we have noted that $\sum_{n=1}^{\infty} \frac{\log(1 + \frac{1}{n})}{n}$ converges since $\int_{1}^{\infty} \frac{\log(1 + x)}{x} dx$ converges. The next term takes the form:

$$(\ell + 1) \sum_{\ell+1}^{\infty} F(n) \log F(n) = -(\ell + 1) \sum_{\ell+1}^{\infty} \frac{\log(n(n+1))}{n(n+1)}.$$ 

We can bound the sum using 

$$\int_{\ell+1}^{\infty} \frac{\log(n(n+1))}{n(n+1)} dn \leq \sum_{\ell+1}^{\infty} \frac{\log(n(n+1))}{n(n+1)} \leq \log((\ell+1)^{\ell+1}) + \int_{\ell+1}^{\infty} \frac{\log(n(n+1))}{n(n+1)} dn.$$ 

These integrals are both easily evaluated, revealing an asymptotic form of:

$$(\ell + 1) \sum_{\ell+1}^{\infty} F(n) \log F(n) = -2 \log \ell,$$

plus corrections of $O(1)$. Finally, to evaluate the last term in the sum, we note that:

$$w(n) - w(n+\ell + 1) = \frac{1}{n(1 + \frac{n}{\ell+1})} = \frac{1/\ell + 1}{n(1 + \frac{n}{\ell+1})},$$

when $n \geq 1$. We define $x_n = \frac{n}{\ell+1}$ with $dx_n = \frac{1}{\ell+1}$ and write:

$$w(n) - w(n+\ell + 1) = \frac{dx_n}{x_n(1 + x_n)}.$$
Then:
\[
\sum_{n=0}^{\infty} (w(n) - w(n + \ell + 1)) \log(w(n) - w(n + \ell + 1)) = (1 - w(\ell + 1)) \log(1 - w(\ell + 1))
\]
\[
+ \log dx_n \sum_{n=1}^{\infty} \frac{dx_n}{x_n(1 + x_n)} + \sum_{n=1}^{\infty} \frac{\log(x_n(1 + x_n))}{x_n(1 + x_n)} dx_n.
\]

The first term is \(o(1)\), since \(\lim_{\ell \to \infty} (1 - w(\ell + 1)) \log(1 - w(\ell + 1)) = 0\). We can view the other two sums as Riemann sums for integrals \(\int_{1/\ell}^{\infty} \frac{dx}{x(1+x)}\) and \(\int_{1/\ell}^{\infty} \frac{\log(x(1+x))}{x(1+x)} dx\) respectively, giving:
\[
\sum_{n=1}^{\infty} \frac{dx_n}{x_n(1 + x_n)} = \log \ell,
\]
plus corrections of \(o(1)\) and:
\[
\sum_{n=1}^{\infty} \frac{\log(x_n(1 + x_n))}{x_n(1 + x_n)} dx_n = -\frac{1}{2} \log^2 \ell.
\]

plus corrections of \(o(1)\). Altogether, substituting the above expressions into Eq. (5.66) yields:
\[
\hat{E}(\ell) = \log \log \ell - 2,
\]
plus corrections of \(o(1)\). The various divergences of order \(\log \ell\) all cancel one another, but the divergence of \(\log \log \ell\) due to the \(\log \ell\) divergence in \(Z(\ell)\) remains, just as for the continuous-time case. When \(F(n)\) is monotone decreasing at some finite \(N\) sufficiently rapidly, manipulations similar to those above imply that divergence in \(\hat{E}\) is a sufficient condition for divergence in \(E\).

In short, we obtain:
\[
\hat{E} = \begin{cases} 
\log \frac{\alpha^2}{\alpha - 1} - 1 & \alpha > 1 \\
\infty & \alpha = 1 \\
\frac{\alpha^2 + \alpha - 1}{\alpha(1-\alpha)} + \log \frac{\alpha}{1-\alpha} - (1-\alpha)K_\alpha & \alpha < 1
\end{cases}
\]

(5.76)

where \(K_\alpha = \int_0^{\infty} (u^{-\alpha} - (1 + u)^{-\alpha}) \log(u^{-\alpha} - (1 + u)^{-\alpha}) du\). Note that at small values of \(\alpha\), \(K_\alpha\) is difficult to evaluate numerically due to the integrand’s long tails, even when \(\hat{E}\) is quite small. For instance, when \(\alpha = 1/4\), \(\hat{E} \approx 0.089\) nats, but \(\int_0^{\infty} (u^{-\alpha} - (1 + u)^{-\alpha}) \log(u^{-\alpha} - (1 + u)^{-\alpha}) du\) does not return positive estimates for the excess entropy until \(N \geq 10^{11}\). A more obvious benefit of Eq. (5.76), then, is that we can study the excess entropy’s asymptotic behavior near \(\alpha = 1\), where \(\hat{E}(\ell) \sim \log \log \ell\). This divergence is slower than any previously reported divergence (Bialek, Nemenman, and Tishby, 2001b; Travers and Crutchfield, 2014b; Debowksi, 2012b), but is a divergence nonetheless.

When \(\alpha > 1\) but close to its critical value, the excess entropy diverges as \(\sim \log \frac{1}{\alpha - 1}\). As \(\alpha \to \infty\), \(\hat{E}\) diverges as \(\log \alpha\).
The discrete-time analog of fractal renewal processes has a survival function:

\[
\begin{aligned}
\quad w(n) &= \begin{cases}
1 & n = 0 \\
\frac{n}{\alpha} & n \geq 1
\end{cases}.
\end{aligned}
\]  

(5.77)

The transient (small \(n\)) behavior of \(w(n)\) may not match that in some applications, but only \(w(n)\)'s asymptotic behavior is relevant to \(E\)'s divergence. Moreover, Eqs. 5.76 guarantees that \(E\) is finite when \(\alpha \neq 1\) and that at \(\alpha = 1\) its divergence is at most \(\log \log \ell\). Additional arguments presented earlier, in turn, show that \(E(\ell)\) diverges at \(\alpha = 1\) as \(\log \log \ell\).

The excess entropy \(E\) captures the amount of predictable randomness of a stochastic process. As a comparison, we are also interested in the statistical complexity \(C_\mu\) of discrete-time and continuous-time fractal renewal processes. The statistical complexity is the number of bits required to losslessly predict (\(E\) nats of) the process' future. Sometimes, \(C_\mu\) is not much larger than \(E\); for discrete-time periodic processes, the two are equivalent and equal to the logarithm of the period. More often than not, \(C_\mu\) is infinite while \(E\) is finite; e.g., for processes generated by most (nonunifilar) Hidden Markov Models.

Cryptic processes have large statistical complexity and small excess entropy (Crutchfield, Ellison, and Mahoney, 2009); the larger the crypticity, the more that a process’ true structure is “hidden” from the observer. An open question is whether or not fractal renewal processes, with their statistical signatures of complexity, are highly cryptic. So, we focus some attention now on evaluating \(C_\mu\) for fractal renewal processes.

We can calculate \(C_\mu\) of time-binned continuous-time renewal processes in the infinitesimal-\(\tau\) limit (Marzen, DeWeese, and Crutchfield, 2015):

\[
C_{\mu, \tau} \sim \log \frac{1}{\tau} - \int_0^\infty \frac{\Phi(t)}{T} \log \frac{\Phi(t)}{T} dt.
\]

The above expression is the differential entropy over continuous-time causal states—the expression given earlier as the “continuous-time statistical complexity” \(\hat{C}_\mu\)—plus the logarithm of our time-bin resolution. Thus, \(C_{\mu, \tau}\)'s \(\log \frac{1}{\tau}\) divergence is an artifact of our failure to use the differential entropy when calculating memory storage requirements of continuous random variables (Cover and Thomas, 2006). As a result, we focus on \(C_{\mu, \tau}\)'s nondivergent component, \(\hat{C}_\mu = \lim_{\tau \to 0} (C_{\mu, \tau} + \log \tau)\), or what was earlier called the continuous-time statistical complexity. Straightforward algebra shows that:

\[
\hat{C}_\mu = \begin{cases}
\frac{1}{\alpha-1} + \log \frac{\alpha}{\alpha-1} & \alpha > 1 \\
\infty & \alpha \leq 1
\end{cases}.
\]  

(5.78)

Again, we can say that the (continuous-time) \(C_\mu\) diverges whenever the mean interevent interval \(T\) diverges. When \(\alpha \leq 1\), finite-length statistical complexity estimates adapted to the continuous-time case from Eq. (5.68) diverge as:

\[
C_{\mu, \tau}^+ \sim \begin{cases}
\log \ell & \alpha < 1 \\
\frac{1}{2} \log \ell & \alpha = 1
\end{cases}.
\]
5.3. Not just complicated, but complex

So, the special nature of \( \alpha = 1 \) is also revealed as a discontinuity in rates of divergence of the finite-length statistical complexity. In particular, the least cryptic fractal renewal process, among fractal renewal processes with divergent statistical complexity, is the process generated when \( \alpha = 1 \).

Equations (5.76) and (5.78) are plotted in Fig. 5.10. The divergences in \( \hat{E} \) and \( \hat{C}_\mu \) at \( \alpha = 1 \) are apparent in the plot. If \( \hat{E} \) and \( \hat{C}_\mu \) are taken to be systems-agnostic order parameters, then a fractal renewal process exhibits a nonequilibrium phase transition exactly when its mean interevent interval diverges.

![Figure 5.10: Excess entropy \( \hat{E} \) and statistical complexity \( \hat{C}_\mu \) of continuous-time fractal renewal processes: Process realizations are generated by drawing interevent intervals IID from the probability density function \( \phi(t) = \alpha t^{-(\alpha+1)} \) for \( t \geq 1 \) and 0 otherwise. \( \hat{E} \) in nats as a function of \( \alpha \), evaluated using Eq. (5.76). The nondivergent component of statistical complexity \( \hat{C}_\mu \) in nats as a function of \( \alpha \), evaluated using Eq. (5.78). Note that \( \hat{C}_\mu \) is a differential entropy and so not necessarily larger than the excess entropy \( \hat{E} \); a subtlety when working with continuous-time processes.

The behavior of \( \hat{E} \) and \( \hat{C}_\mu \) as \( \alpha \) tends to infinity also deserves special mention, as the process appears to become infinitely predictable (\( \hat{E} \to \infty \)) while requiring less memory for prediction (\( \hat{C}_\mu \to 0 \)). As \( \alpha \) tends to \( \infty \), \( \phi(t) \) becomes more and more sharply peaked at \( t = 1 \). In other words, the process moves closer and closer to that of a periodic process with period 1. Periodic processes are random enough, in that the phase of the process could be any real number between 0 and the period. In the language of computational mechanics, the causal state is the phase, and its differential entropy—the continuous-time statistical complexity \( \hat{C}_\mu \)—is the logarithm of the process’ period. As \( \alpha \to \infty \), the mean interevent interval \( \bar{T} = \frac{1}{\alpha-1} \) tends to 1, and the continuous-time statistical complexity correspondingly
tends to $\log 1 = 0$. However, periodic processes are also highly predictable, in that the time to next event is determined by the time since last event; hence, the differential entropy of the time to next event conditioned on the time since last event tends towards negative infinity, resulting in an infinite $\hat{E} = \hat{C}_\mu - H[S^-|S^+] \to \infty$. Similar behavior was seen for model neural spike trains (Marzen, DeWeese, and Crutchfield, 2015) as the variability of interspike intervals tended to zero. The least cryptic fractal renewal process, then, occurs in the limit that $\alpha$ tends to infinity.

A straightforward application of the Data Processing Inequality not shown here reveals that the excess entropy of the processes described in Sec. 5.2 is infinite if and only if one of the $\phi_i(t)$ is asymptotically $\sim 1/t^2$. 
BIBLIOGRAPHY


Appendix A

Causal states and minimal maximally predictive models

We briefly review the field sometimes called “computational mechanics” and sometimes called the “physics of information”. Interested readers can consult Shalizi and Crutchfield, 2001; Crutchfield and Young, 1989b for more detailed expositions.

Causal states are minimal sufficient statistics of prediction for a process. A process is the chain of random variables \( \ldots X_{-3}, X_{-2}, X_{-1}, X_0, \ldots \) generated by any Hidden Markov model. We consider \( \ldots X_{-3}, X_{-2}, X_{-1} \) to be the past and \( X_0, \ldots \) to be the future of the process.

More concretely, let \( x_0 \) and \( x'_0 \) be two different semi-infinite pasts. We consider these two pasts to be equivalent when

\[
P(X_0 | X_0 = x_0) = P(X_0 | X_0 = x'_0)
\]

in that both lead to identical forecasts. This equivalence relation partitions the set of semi-infinite pasts into equivalence classes which we call “forward-time causal states” \( S^+ \) (with realizations \( \sigma^+ \)). When context is clear, we simply call them causal states \( S \) (with realizations \( \sigma \)). The function \( \epsilon^+ \) takes a semi-infinite past to its corresponding forward-time causal state. The forward-time statistical complexity \( C^+_\mu = H[S^+] \) is often denoted as \( C^\mu_\mu \) when context is unambiguous.

Some processes have a finite number of causal states; some have a countable infinity of causal states; and others have an uncountable infinity of causal states. The last of these options seems to be the generic situation. In the first two situations, we can develop a (sometimes) tractable generative Hidden Markov model of the process by calculating labeled transition matrices from the process dynamic itself:

\[
T^{(x)}_{\sigma, \sigma'} = P(S_{t+1} = \sigma', X_t = x | S_t = \sigma).
\]

This Hidden Markov model is also (provably) unifilar.\(^1\)

This Hidden Markov model is the process’ \( \epsilon \)-machine. Hence, the \( \epsilon \)-machine is the minimal maximally predictive model of a process. This leads us to an arguably more useful definition of causal states: they have a one-to-one correspondence with the states in a minimal unifilar Hidden Markov of the

\(^1\)To show this, consider that the addition of a measurement symbol to a semi-infinite past produces another semi-infinite past which belongs to only one equivalence class.
Appendix A. Causal states and minimal maximally predictive models

process. Here, minimal can be defined as either minimal entropy or minimal number of states. The \( \epsilon \)-machine is thus the minimal unifilar Hidden Markov model capable of generating the process.

A third related view of causal states comes from something called the mixed state presentation. Seeing a particular semi-infinite past \( x_{0} \) induces a probability distribution over states in a generative model. These probability distributions live in the “mixed state simplex” and causal states are exactly the accumulation points in this simplex. When there are an uncountable infinity of causal states, as is typical, this viewpoint is the most useful.

One can also construct “reverse-time causal states” \( S^{-} \) (with realizations \( \sigma^{-} \)) by clustering futures with the same conditional probability distribution over pasts. The corresponding reverse-time \( \epsilon \)-machine runs “backwards”, and \( C^{-}_{\mu} = H[S^{-}] \) is the reverse-time statistical complexity. (Usually, \( C^{-}_{\mu} \neq C^{+}_{\mu} \).) The function \( \epsilon^{-} \) takes a semi-infinite future to its corresponding reverse-time causal state.

Both \( S^{+} \) and \( S^{-} \) have special “causal shielding” properties that make them especially useful when calculating various information measures of a process. In the following Markov chain

\[
S^{+} \rightarrow X_{0} \rightarrow X_{0} \rightarrow S^{-},
\]

(A.3)

which clearly holds because \( S^{+} \) and \( S^{-} \) are functions of \( X_{0} \) and \( X_{0} \), \( S^{+} \) and \( X_{0} \) can be interchanged. The same is true for \( X_{0} \) and \( S^{-} \). These properties hold for any sufficient statistics of \( X_{0} \) with respect to \( X_{0} \), or vice versa.

As discussed in Chapter 4, these four Markov chains provide the basis for a new improved algorithm for calculating predictive rate-distortion functions and associated optimal stochastic codebooks.

Recent advances have made it possible to calculate the joint probability distribution of forward- and reverse-time causal states \( P(S^{+}, S^{-}) \) directly from the forward-time (or reverse-time) \( \epsilon \)-machine. These two methods are:

1. Build a bidirectional \( \epsilon \)-machine which can go in forward or reverse time, whose state space is constructed as a direct product of the state spaces of forward- and reverse-time \( \epsilon \)-machines (Ellison et al., 2011). The stationary probability distribution over these bidirectional causal states gives \( P(S^{+}, S^{-}) \).

2. Find a reverse-time generative model from the forward-time \( \epsilon \)-machine by flipping transition arrows and renormalizing, and then constructing the mixed state presentation, looking for accumulation points. By keeping track of the actual distribution over states in the reversed forward-time \( \epsilon \)-machine, we find \( P(S^{+} | S^{-}) \). By building a model from accumulation points, we find \( P(S^{-}) \), and thus also find \( P(S^{+}, S^{-}) \). Details are in Ellison, Mahoney, and Crutchfield, 2009.

No guarantees for convergence for either method are provided, as the reverse-time \( \epsilon \)-machines for finite forward-time \( \epsilon \)-machines are often uncountably infinite.
Appendix B

Rate-distortion theory vs.
predictive rate-distortion
theory

In this appendix, we abuse notation by allowing $x$ to stand for a single
semi-infinite past. Let $\Delta$ be the simplex of semi-infinite futures, a random
variable $Y$ with realizations $y$. (Notation is chosen to better match that of
Tishby, Pereira, and Bialek, 1999.) Distortions take the form $d : \mathcal{A} \times \mathcal{F} \to \mathbb{R}_+$,
so that we use the shorthand

$$d(x, r) = d(P(Y|X = x), P(Y|R = r))$$  \hspace{1cm} (B.1)

where, since $R \rightarrow X \rightarrow Y$,

$$P(Y|R = r) = \sum_x P(Y|X = x)P(X = x|R = r).$$  \hspace{1cm} (B.2)

This type of distortion measure differs from the more typical distortion
measures discussed in Chapters 2 and 3 in that the distortion between $x$ and $r$
depends on the codebook $p(r|x)$.

In light of this difference, we now discuss the slight changes that need
to be made to prove a rate-distortion theorem for distortion measures of
this kind. Following Yeung, 2008, we define $R(D)$ again as the boundary
between achievable and unachievable combinations of rate and predictive
distortion, and define

$$R_I(D) := \min_{p(r|x) : d(x, r) \leq D} I[X; R].$$  \hspace{1cm} (B.3)

Most of the arguments in Yeung, 2008, Ch. 8 carry over exactly, so we comment
mostly on what must be changed in order for the theorem to carry
through unscathed.

In the setup of Fig. 1.1, $n$ input symbols $x_{0:n}$ are sent through a channel
which spits out codeword $r$, which is in turn decoded as $\tilde{x}_{0:n}$, and
differences between $x_{0:n}$ and $\tilde{x}_{0:n}$ are penalized. In the predictive
distortion setup, there is no decoder. Distortion is calculated directly between
the $n$ input symbols $x_{0:n}$ and the $n$ representations $r_{0:n}$. The definitions
of distortion-typical sets are modified slightly in concept, but the asymptotic
equipartition property for jointly typical sets still holds: $d(x_{0:n}, r_{0:n}) = \sum_{i=0}^n d(P(Y|X = x_i), P(Y|R = r_i))$, which (by law of large numbers) tends
to $\mathbb{E}[d(x, r)]$. The rate-distortion theorem ($R(D) = R_I(D)$) then holds for
predictive distortions as well.