On Infinite Complexity: Quantifying the Randomness and Structure of Hidden Markov Processes

By

ALEXANDRA MARIE JURGENS
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

PHYSICS

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

James P. Crutchfield

John Rundle

Richard Scalettar

Committee in Charge

2021

To anyone who is uncertain, and who may derive any amount of knowledge, enjoyment, or encouragement from this work. It is not within my power to resolve your uncertainty, but perhaps we may begin by quantifying it.

Contents

List of Figures vi

Abstract xiii

Acknowledgments xv

Chapter 1. Introduction and Motivation 1

Chapter 2. The Study of Stochastic Processes 9

2.1. Definition of a Process 9

2.2. Hidden Markov Models 11

2.3. The ϵ-Machine 12

2.4. Some Information Theory 13

2.5. Intrinsic Randomness 16

2.6. Intrinsic Structure 17

Chapter 3. The Mixed State Presentation: Building Predictive Models 19

3.1. Introducing Iterated Function Systems 19

3.2. The Mixed-State Presentation 22

3.3. The MSP as an IFS 28

3.4. The Mixed-State Presentation and the ϵ-Machine 31

3.5. Hidden Markov Driven Iterated Function Systems 32

Chapter 4. Intrinsic Randomness, or the Shannon Entropy Rate 34

4.1. Blackwell's Solution 34

4.2. An Updated Solution 35

4.3. Data Requirements 37

4.4. Estimation Errors for Finite-State Autocorrelation 38
4.5. Computational Advantages and Disadvantages 42

Chapter 5. Intrinsic Structure, or the Statistical Complexity Dimension 44
5.1. Defining the Statistical Complexity Dimension 44
5.2. Dimension from Dynamical (In)Stabilities 47
5.3. Calculating Statistical Complexity Dimension 48
5.4. The Overlap Problem 50
5.5. Two-State Hidden Markov Models 51
5.6. Many-state Hidden Markov Models 53

Chapter 6. Introducing the Ambiguity Rate 57
6.1. What is the Ambiguity Rate? 58
6.2. Sources of State Uncertainty Reduction 60
6.3. A Conjecture About the Kaplan-Yorke Conjecture 62
6.4. Interpretations of the Ambiguity Rate 63
6.5. Example: Discrete-Time Renewal Process 66
6.6. Example: Ambiguity Rate in One Dimension 68
6.7. An Algorithm for the Numerical Approximation of Ambiguity Rate 70

Chapter 7. Application: Measurement Induced Randomness and Structure in Qubit Sources 73
7.1. Experimental Motivation 73
7.2. Quantum Formalism 75
7.3. Measured Qubit Processes 76
7.4. Uncountable Predictive Features 77
7.5. Measurement Dependence 80
7.6. Implications and Future Directions 81

Chapter 8. Application: Functional Thermodynamics of Maxwellian Ratchets 84
8.1. Motivation 85
8.2. Information Engines 86
8.3. Mandal-Jarzynski Information Ratchet 90

8.4. Randomizing and Derandomizing Behaviors 93
8.5. Constructing and Deconstructing Patterns 98
8.6. Information Ratchet Mixed-State Attractor Survey 104
8.7. Pattern Deconstruction and Thermodynamic Functionality 105
8.8. Related Efforts 106

Chapter 9. Conclusion and Future Directions 109

Appendix A. Hidden Markov Model Examples 113

Appendix B. Asymptotic Equipartition and Typical Set Contraction 115

Appendix C. Birkhoff's Contraction Coefficient 116

Appendix D. Correspondence with Baker's Map 118

Appendix E. Sierpinski's Triangle 120

Appendix F. Lyapunov Exponents 122

Appendix G. Overlap Estimation 124

Appendix H. Mandal-Jarzynski Ratchet 126

Bibliography 129

List of Figures

1.1 Figures (a) and (b) plot 10^5 states of an uncountably-infinite state ϵ-machine. Each is constructed for a different hidden Markov process, both generated with a 3-state HMM. 6

1.2 text 7

2.1 A hidden Markov model (HMM) with two states, $\{\sigma_1, \sigma_2\}$ and two symbols $\{\square, \triangle\}$. It is unifilar. 11

2.2 A three-variable i-diagram for random variables X, Y and Z, showing all possible information atoms, including conditional entropies, conditional mutual informations, and multivariate mutual information. 15

3.1 How a hidden Markov-driven iterated function system (DIFS) generates a hidden Markov process: An initial state η—a distribution over three states: $(0, 0, 1)$, $(0, 1, 0)$, and $(1, 0, 0)$—in the 2-simplex is associated with a transition probability distribution over the alphabet $\mathcal{A} = \{\square, \triangle\}$. If the emitted symbol selected from this distribution is $\square$, the next state is generated according to the associated mapping function $f^{(\square)}(\eta)$ and the probability distribution is updated accordingly. The same steps are followed if the symbol is $\triangle$ using $f^{(\triangle)}(\eta)$, resulting in an emitted process $\mathcal{P}$ over symbols $\mathcal{A}$. 20

3.2 Determining the mixed-state presentation (MSP) of the 2-state unifilar HMC shown in (A): The invariant state distribution $\pi = (2/3, 1/3)$. It becomes the first mixed state η_0 used in (B) to calculate the next set of mixed states. (C) The full set of mixed states seen from all allowed words. In this case, we recover the unifilar HMC shown in (A) as the MSP's recurrent states. 23

3.3 Determining the mixed-state presentation of the 2-state *nonunifilar* HMC shown in (A). The invariant distribution $\pi = (1/2, 1/2)$. It is the first mixed state η_0 used in (B) to calculate

the next set of mixed states. (B) plots the mixed states along the 1-simplex $\Delta^1 = [0, 1]$. In (C), we translated the points on the simplex to the states of an infinite-state, unifilar HMC. 25

3.4 Figures (a), (b), and (c) each plot 10^5 mixed states of the uncountably-infinite state MSP generated by the parametrized 3-state HMC defined in Eq. (3.7) at various values of x and α. This HMC is capable of generating MSPs with a variety of structures, depending on x and α. However, due to rotational symmetry in the symbol-labeled transition matrices, the attractor is always radially symmetric around the simplex center. 27

3.5 Commuting diagram for probability functions $P = \{p^{(x)}\}$, mixed-state mapping functions $f^{(x)}$, and proposed symbol-distribution mapping functions $g^{(x)}$. 31

4.1 Entropy rate for $250,000$ parametrized HMPs generated according to the definition in Eq. (3.7), over $x \in [0.0, 0.5]$ and $\alpha \in [0.0, 1.0]$. Examples of the MSPs produced are plotted in Fig. 3.4. The entropy rate of each, as estimated by Eq. (4.4), is plotted, showing the gradual change in entropy rate across the parameter space. 36

5.1 Simple HMCs generate MSPs with a wide variety of structures, many fractal in nature. Each subplot displays 10^4 mixed states of a different, highly-nonunifilar 3-state hidden Markov chain. The HMCs themselves are specified in Appendix A. 45

5.2 Overlap problem on the 1-simplex Δ^1: Two distinct IFSs are considered, each with two mapping functions. The images of the mapping functions over the entire simplex are depicted in red and blue. (A) Images of the mapping functions $f^{(0)}$ and $f^{(1)}$ do not overlap—every mixed state $\eta_t \in \boldsymbol{\mathcal{R}}$ has a unique pre-image. (B) Images of the mapping functions overlap (purple Overlapping Region)—there exist $\eta_1, \eta_2 \in \boldsymbol{\mathcal{R}}$ such that $f'^{(0)}(\eta_1) = f'^{(1)}(\eta_2) = \eta_3$. This case is an *overlapping IFS*. 48

5.3 Mixed-state attractors generated by a 3-state HMC parametrized over $\alpha \in [0, 1]$ and $x \in [0, 0.5]$. The HMC itself is given in Eq. (3.7). 10^5 mixed states are plotted for each attractor, with the initial 5×10^4 states thrown away as transients. The ranges of the symbol-labeled maps are color shaded, revealing regions of their image overlap on the attractor. Comparing to Eq. (3.7), the red, blue, and green regions represent the images of the mapping functions defined by $T^{\square}$, $T^{\triangle}$, and T°, respectively. 52

5.4 Attractor area, overlap regions, and Lyapunov dimension of the mixed-state attractors shown in Fig. 5.3, parametrized by $\alpha = [0, 1]$ and $x = [0, 0.5]$. The HMC itself is given in Section 5.6. See Appendix F and Appendix G for a detailed discussion of the production of these plots. 54

6.1 (a) Equivocation: Same input sequence leads to different outputs. (b) Ambiguity: Two different inputs lead to same output. The strategy underlying Shannon's proof of his second coding theorem is to find channel inputs that are least ambiguous given the channel's distortion properties. 58

6.2 Overlap problem on the 2-simplex Δ^2: Two distinct DIFSs (given in Appendix A) are considered, each with three mapping functions. Images of the mapping functions over the entire simplex are depicted as regions in red, blue, and green. (a) Images of the mapping functions $f^{(\triangle)}$,$f^{(\square)}$ and $f^{(\circ)}$ do not overlap—every possible state has a unique pre-image. (b) Images of the mapping functions overlap: there exist $\eta_1, \eta_2 \in \Delta^2$ such that $f'^{(\triangle)}(\eta_1) = f'^{(\square)}(\eta_2) = \eta_3$. This case is an *overlapping DIFS*. 59

6.3 Sources of ambiguity rate depicted in state machines: (A) $H[X_0|\mathcal{R}_0, \mathcal{R}_1] > 0$—Previous state $\mathcal{R}_0$ is mapped to the next state $\mathcal{R}_1$ by two distinct symbols. This occurs when two symbols have identical mapping functions. (B) $I[X_0; \mathcal{R}_0|\mathcal{R}_1] > 0$—Two distinct previous states $\mathcal{R}_0$ map to the same next state by distinct symbols, due to overlapping mapping functions. (C) $H[\mathcal{R}_0|X_0, \mathcal{R}_1] > 0$—Two distinct previous states $\mathcal{R}_0$ map to the same next state by the same symbol. This occurs when a mapping function is noninvertible. 60

6.4 The entropy rate h_μ, which in this case is equivalent to the ambiguity rate h_a, is plotted for the DIFS defined by Eq. (3.9) for $p, q, \in (0, 1)$. 64

6.5 One-dimensional attractor for DIFS given in Eq. (6.5) with $x = 0.25$ and horizontally varying $\alpha \in (0, 1)$. State set $\eta \in \boldsymbol{\mathcal{R}}$ plotted (red, blue green) on top of the images of the mapping functions $\left\{f^{(\triangle)}, f^{(\square)}, f^{(\circ)}\right\}$ applied to $\boldsymbol{\mathcal{R}}$. For $\alpha \in (0.07, 0.78)$, there is overlap in the images of the maps. 66

6.6 Numerical calculation of entropies and dimension quantities for the DIFS given in Eq. (6.5), $\alpha \in (0, 1)$, and $x = 0.25$: (Top) The entropy rate h_μ, the ambiguity rate h_a, and $h_\mu - ha$.

(Bottom) Comparison of $\frac{h_\mu}{\lambda_1}$ to $d_\mu = \frac{h_\mu - h_a}{\lambda_1}$. The latter smoothly departs from the former and is approximately 1 for much of the overlap region, except where it discontinuously jumps to zero at $\alpha = 1/3$. 68

6.7 DIFS for $x = 0.25$ and $\alpha = 0.5$: (Top) Blackwell Measure μ_B approximated by Ulam's method with $k = 400$. The two overlapping regions are overlaid and may be compared with Fig. 6.5. (Middle) Probability of each prior map is plotted. In nonoverlapping regions, only one prior map is possible. In the overlapping regions, there are complicated, fractal-like distributions over multiple prior maps. (Bottom) Shannon entropy over the prior map: Nonzero only in the overlapping regions. 70

7.1 (Top) A general controlled qubit source (CQS) as a discrete-time quantum dynamical system that generates a time series of qubits $\rho_{t-2}\rho_{t-1}\rho_t \ldots$. (Bottom) A cCQS generates a qubit process $|0\rangle\langle 0|, |+\rangle\langle +| \ldots$. Measuring each qubit, an observer sees a classical stochastic process: (Top) $\ldots x_1 x_2 x_3 x_4 x_5$, (Bottom) $\ldots 011100$. 74

7.2 A controlled qubit source emits a discrete-time stochastic process of qubits in pure states ρ_t. Applying measurement E to the qubit emitted at time t in state ρ_t yields outcome x_t. Measuring each qubit at each time step yields a classical stochastic process. 76

7.3 (a) Three-state classically-controlled qubit source (cCQS) that generates a process consisting of qubits in orthogonal states. (i) controller states $\boldsymbol{\mathcal{S}} = \{A, B, C\}$; (ii) orthogonal qubit alphabet $\mathcal{A}_Q = \{\rho_0 = |0\rangle\langle 0|, \rho_1 = |1\rangle\langle 1|\}$; (iii) labeled transition matrices $\mathbb{T}^{\rho_0}$ and $\mathbb{T}^{\rho_1}$, whose state-transition probability components $(\mathbb{T}^{\rho_k})_{ij}$ can be read from the diagram; and (iv) stationary state distribution $\pi = \begin{pmatrix} 1/2 & 1/4 & 1/4 \end{pmatrix}$. If the controller is in state A it has equal probabilities of staying in that state or transitioning to state B. If it transitions to state B, then the system emits a qubit in pure state $|0\rangle\langle 0|$. In the next time step the machine must transition to state C and emits another $|0\rangle\langle 0|$ qubit. Then, in state C it transitions back to state A, emitting either pure state $|0\rangle\langle 0|$ or $|1\rangle\langle 1|$ with equal probability. As the cCQS runs, a time series of orthogonal qubits is generated. (b) Nonorthogonal-qubit cCQS: Three-state HMM that outputs nonorthogonal pure states $|0\rangle\langle 0|$ and $|+\rangle\langle +|$, where $|+\rangle = (|0\rangle + |1\rangle)/\sqrt{2}$. (c) HMM presentation for the classical stochastic process resulting

from measurement ($\theta = \pi/2$) of the quantum process generated by (b). (d) Mixed states for the stochastic process generated by (c) in that HMM's state distribution simplex. Each mixed state is a point of the form (p_A, p_B, p_C) with probabilities of being in state A, B, or C of (c). The color scale shows the logarithm of the probability of the mixed states. 78

7.4 Measurement-induced randomness and structure: (Top) Mixed state sets when measuring the qubit process of Fig. 7.3(b) as a function of measurement angle $\theta \in [0, \pi]$. (Bottom) Entropy rate $\widehat{h_\mu^B}$ (blue curve) as a function of angle. Horizontal line (red) is the entropy rate of the (unmeasured) qubit sequences: $h_\mu^g = 3/4$ bit per output qubit. (Insets) Mixed states three measurement angles: (a) $\theta_a = 0.628$ (purple), (b) $\theta_b = 1.634$ (orange), and (c) $\theta_c = 2.701$ (green). The measured process entropy rates h_μ and statistical complexity dimensions d_μ given there. Both mixed states and complexity measures computed with $\ell = 10^6$ iterates. 79

8.1 Information engine as a finite-state ratchet (controller) connected to a thermal reservoir, a work reservoir, and an information reservoir (depicted as tape whose storage cells may be read or written). 86

8.2 Composing the Mandal-Jarzynski transducer (center, yellow) with a Hidden Markov model (left, green) that describes the process on the input tape gives an output Hidden Markov model (right, purple) that describes the process written to the output tape. Hidden Markov model (HMM) states R are depicted as circles. Directed edges between states represent possible transitions on an observed symbol x. HMM edges are labeled $x : \Pr(x', R_{N+1}|R_N)$. Transducers are similarly depicted by circular states with directed edges representing possible transitions on pairs of input symbols x and output symbols x'. Transducer transitions are specified by $x'|x : \Pr(x', R_{N+1}|x', R_N)$. Left: Input HMMs discussed here, from top to bottom, a (memoryless) Biased Coin, a Period-2 Process, and the Golden Mean Process. Center: The Mandal-Jarzynski ratchet, represented by a three-state transducer. Probabilities are not shown on edge labels for conciseness, but are nonzero for all transitions and all combinations of input-output symbol pairs (x, x'). Each edge probability is a function—denoted by $f(\ldots)$—of the previous state R_N, the next state

R_{N+1}, the input symbol x, and the output symbol x'. See Section 8.3 and Appendix H for further details. Right: Output HMMs resulting in the Mandal-Jarzynski transducer composed with the corresponding input HMM on left. Edge labels are left off for conciseness, but each transition label represents a positive probability of observing a 0 or a 1. 89

8.3 Ratchet schematic adapted from the original Mandal-Jarzynski construction, showing how the dial-and-symbol system is transformed into a three state transducer upon selection of a specific ϵ—determining the energetics of flipping a bit—and τ—determining the interaction interval. For almost every value of ϵ and τ every state-to-state transition has positive probability for every input-output symbol combination. 91

8.4 Mixed states $\eta \in \boldsymbol{\mathcal{R}}$ of the output process generated by the ratchet driven with memoryless input (Fig. 8.2(top row)) plotted on the 2-simplex. Corner labels give the mixed-state probability distributions $\eta = (\Pr(A \otimes D), \Pr(B \otimes D), \Pr(C \otimes D))$. Mixed states at the simplex corners correspond to the HMM being in exactly one of its states, while mixed states in the simplex interior are mixtures of the possible HMM states, with $\eta = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ lying at the center. (Left) Ratchet parameters $\delta = -0.98$, $\epsilon = 0.01$, and $\tau = 0.1$. (Right) Ratchet parameters $\delta = 0.4$, $\epsilon = 0.5$, and $\tau = 0.1$. Insets: Detail of the mixed-state sets, magnified by amount indicated in upper right corner. 92

8.5 Asymptotic work production $\langle W \rangle$, single-symbol ΔH_1 bound, and Kolmogorov-Sinai-Shannon Δh_μ bound when ratchet is driven by period-2 memoryful input. Since the input has no parameters, the parameter sweeps only over ϵ with $\tau = 10$. 95

8.6 Functional thermodynamic regions of the ratchet driven with the Golden Mean Process as a function of parameters $\epsilon \in [-1, 1]$ and $s \in [-1, 1]$ with $\tau = 1$. (Left) Purported functionality identified via by single-symbol entropy bound Eq. (8.2). (Right) Correct functionality identified via the entropy-rate bound IPSL Eq. (8.3). 97

8.7 After passing through the information engine, the input process, which has some initial h_μ and d_μ, is transformed into an output process with a potentially different h'_μ and d'_μ. By carefully selecting input, an information engine can be induced to act as a randomizer ($\Delta h_\mu > 0$) or derandomizer ($\Delta h_\mu < 0$) and a pattern constructor ($\Delta d_\mu > 0$) or deconstructor

($\Delta d_\mu < 0$). We show here that all four regions of the Δh_μ–Δd_μ plane are accessible to the Mandal-Jarzynski ratchet by carefully selecting parameters and input. The two insets on the left show the uncountable set of mixed states of an input process that the ratchet transduces to an IID output. The insets on the right show two uncountably-infinite-state output processes produced by running the Mandal-Jarzynski ratchet on a biased coin. The parameters, clockwise from top right: $\delta_{input} = -0.98, \epsilon = 0.01, \tau = 0.1$; $\delta_{input} = 0.3, \epsilon = 0.5, \tau = 0.1$; $\delta_{output} = 0.8, \epsilon = -0.96, \tau = 0.75$; $\delta_{output} = 0.0, \epsilon = 0.9, \tau = 0.9$. 99

8.8 Mixed-state attractors of the output HMMs of the Mandal-Jarzynski ratchet driven by a Biased Coin as a function of ϵ and input bias δ, given above each square; $(\epsilon, \delta) \in [0, 1] \times [-1, 1]$. Each plot shows $1,000$ mixed states from the attractor at the magnification noted in the lower right corner. The attractors may be compared with Fig. H.1 to determine thermodynamic functionality. 103

8.9 Functional thermodynamic regions of a ratchet pattern deconstructor with an interaction interval of $\tau = 0.75$. The x-axis sweeps over the output bias δ_{out}, while the y-axis sweeps over ϵ. As indicated, there are two parameter regimes where the ratchet is unable to act as a pattern deconstructor. In these regions, the desired output bias is not reachable by the machine at the given ϵ. As $\tau \to \infty$, this region grows, until it encompasses every parameter combination other than $\epsilon = \delta_{out}$, which is always reachable with an input Biased Coin with bias $\delta_{in} = \delta_{out} = \epsilon$. 105

E.1 Attractor of the 3-state, 3-symbol machine specified in Eq. (E.1), with $s = 2$ and $a = \frac{1}{6}$. 120

H.1 MJ ratchet functional thermodynamic regions over $\epsilon \in [1, -1]$ and $b \in [0, 10]$ with $\tau = 1$. (Left) Purported functionality identified via by single-symbol entropy bound Eq. (8.2). (Right) Correct functionality identified via the entropy-rate bound IPSL Eq. (8.3). 128

Abstract

On Infinite Complexity: Quantifying the Randomness and Structure of Hidden Markov Processes

The world around us is awash with structure and pattern. We observe it in the cycles of the seasons, the destructive beauty of coherent large-scale atmospheric events, and the petri-dish bound patterns produced by chemical reactions. In response, we incorporate these patterns into predictive models that allow us to forecast natural phenomena, giving us the ability to say that a plane will fly and a certain medical compound will help, instead of harm. This practice is perhaps intrinsic to the conscious mind, but in the information age these models have become an object of mathematical curiosity, tamed by the axiomatic frameworks of probability and automata theory.

This dissertation focuses on processes generated by a class of models known as finite-state hidden Markov models. Despite the relative simplicity of their generators, the information theoretic properties of these processes are still elusive, eluding our attempts to quantify their structure, randomness, and complexity. This dissertation introduces methods to accurately calculate the Shannon entropy rate (randomness) and statistical complexity dimension (structure) by constructively determining their minimal (though, infinite) set of predictive features.

To do so, we must reframe the stochastic process in the language of random dynamical systems, introducing the set of predictive features as the attractor of a chaotic system. This task accomplished, we turn to the development of a toolset to measure randomness and structure. In the process, we introduce the ambiguity rate, a new intrinsic complexity measure. This quantity is used to measure the rate at which the state space of an infinite model must grow to retain optimal prediction. This quantity is closely related to the fractal dimension of the predictive state set, and we offer a conjectured correction to the Kaplan-Yorke information dimension formula for this class of processes.

To highlight the usefulness of these informational quantities, that otherwise appear rather abstracted from natural systems, we apply these theoretical results to two, rather different, physical domains. The first is to analyze the origin of randomness and structural complexity engendered by quantum measurement. The second is to solve a longstanding problem on exactly determining the thermodynamic functioning of Maxwellian demons, aka information engines. Taken together, we believe the new approach will find even wider use than in these application areas.

Acknowledgments

I must of course first thank and acknowledge my Ph.D. advisor, Professor James P. Crutchfield, without whose vision, guidance, skepticism, encouragement, insight, and above all, support, this dissertation would not exist. Over our years of collaboration, Jim has pushed me to think more deeply, allowed me to find my own ground to stand on, indulged my more creative endeavors, and unwaveringly supported both my work and my person. If I am a better scientist and a deeper thinker today than when I began this journey, it is through his efforts.

The community that Jim has cultivated has been a great source of support, comfort, and comradery during my time at UC Davis. I am thankful for the people that I have met here, not only for the way their brilliance challenges me to go beyond what I ever thought possible, but also due to the deep friendships that I hope and believe will remain through the rest of my life.

In particular, I wish to thank:

Sam Loomis, for his mathematical insight, unwavering friendship, unbounded empathy, and homebrewed mead.

Mihkael Semaan, for his unflinching support, kindness, and for helping keep me sane during quarantine.

David Gier, for his friendship, good humor and always having a word of support.

Adam Rupe, for never turning me away from a late-night desk conversation.

Ryan James, for his knowledge of information theory and his willingness to share it, and for his many words of encouragement.

Alec Boyd, for his thermodynamic insight, and for always pushing me to consider a problem from a new angle.

Many others at the Complexity Sciences Center and beyond, including Ariadna Venegas-Li, Greg Wimsatt, Xincheng Lei, Dany Masante, Anastasiya Salova, Jeffrey Emenheiser, Adam Kunesh,

Korana Burke, Nicolas Brodu, Paul Riechers, and Sarah Marzen, have contributed towards a vibrant and challenging scientific community, and without their presence this work would certainly not exist. I am deeply grateful to know them.

Beyond UC Davis, there are several people and institutions I must acknowledge for their role in my growth as a scientist. Linda Locke, whose tutelage in geometry, while thorough, paled in comparison to her tutelage on how to navigate being a woman in science. The Santa Fe Institute, for introducing me to the broader complexity science community. Ngoc Diep Lai and ENS Paris-Saclay, for hosting me as a young woman and contributing greatly to my understanding of experimental science and the value of international collaboration. SLAC National Accelerator Laboratory and William Schlotter, for providing my first-ever research experience and mentorship that continues to aid me to this day.

Most especially I must thank the late Dr. Cavendish McKay of Marietta College, who brought into my life the beauty and wonder of mathematical physics. His support, enthusiasm, kindness, and humor left a deep impression on me. I cannot overstate the role that his mentorship played in my career: without the impartation of his knowledge of chaos, dynamical systems, and computational physics I would not be here today. His unfortunate death has robbed the world of a brilliant mentor and a better man.

I also thank those in the broader scientific community and adjacent, without whose friendship and late night chats I would not have made it to the finish line. Thank you to Tamara, Sean, Azalee, Matias, Patrick, Emily, among many others.

Above all there is my family, immediate, extended, and found, who have amongst themselves provided support, frustration, shoulders to cry on, contradictory advice, lively game nights, hand-written recipes for layered brownies, road trips with disastrous wrong turns, leftovers for dinner, all this and everything else that a family has to give, foremost amongst these things being love.

My brother, Nicholas, who is in so many ways dear to me, and whose drive, charisma, and strength perpetually inspire me. Growing up alongside him has been one of the greatest gifts a person could receive.

My father, Anthonij, whose encouragement of my interest in science and computers planted deep seeds in my mind. Under his tutoring I first grasped the concept of abstraction, and I offer a

long-overdue apology for regaling him with my frustration on how absurd I found it that “a letter can also be a number.” I can only trust he knows my opinion on the topic has changed.

My mother, Donna, who chipped off a piece of herself and brought me into existence. I do not possess the skill necessary to express my gratitude for her companionship and love, and for her patient and repeated assurances that my path is my own to walk, so I will simply say thank you.

To all those who have come into my life due to this endeavor, I am glad to have met you, and no number of late nights, coffees long gone cold, and outrageously early morning conference sessions is too high a cost to pay for the privilege of having doing so. To those who came before, I am forever indebted to you for staying by my side as I wound my way through this pursuit, despite the path being long and never clearly marked. It is only due to your love and support that I have arrived here.

CHAPTER 1

Introduction and Motivation

In the space of a generation, humanity has found itself in the "Information Age", an epoch heralded by the arrival of "information technology". The most widespread of these technologies is the one you are likely using to read this dissertation: the computer, a machine capable of storing, transforming, and processing information. Despite the rapid proliferation of these machines—and the subsequent economic impact of their widespread adoption—several of the underlying ideas still lack rigorous, physically-motivated explanation. Indeed, our position is not dissimilar to that of the early 19th century, when the first steam locomotive was demonstrated before the term "energy" was first used in its modern scientific sense. The questions that then faced "natural philosophers" of the day—on the creation, transformation, and utilization of this so-called energy—are now recognized as the fundamental questions of physics. Two hundred years later, we find ourselves asking these same questions of *information*: what is it, and how can it be generated, stored, transformed, and utilized? Do natural systems process information—or, *compute*—and how? What role does information and computation have to play in explaining the emergent structure and complexity we so readily observe in the world around us?

In 1948, Claude Shannon [1] made the first stride towards answering these questions by formalizing the mathematical definition of information. He argued that we desire any measure of informational content to meet several axioms: 1) observation of an event X with a certain outcome yields no information; 2) the less probable an event, the more information it yields; and 3) additivity of informational content of independent events. It can be shown that there is a unique function of probability of an outcome x that meets all of these requirements: the logarithmic function $I(X) = -\log(\Pr(X = x))$. When the log is taken in base two, information is measured in *bits*, and the resolution of fair coin flip yields one bit of information. In most cases, we are interested in the *Shannon entropy* $H(X) = E[I(X)]$: the expected value of the information gained by observing X. Properties of $H(X)$ follow naturally from $I(X)$. When there is only one possible outcome, $H(X)$

goes to zero; when all possible outcomes are equally likely, $H(X)$ is maximized. Thus, we see that uncertainty—and the resolution of uncertainty—is inextricably linked to any physical interpretation of information.

Even in Shannon's time, the idea that uncertainty has a role to play in physics was not a novel concept. Long before quantum mechanics ushered us into an inherently probabilistic universe, or even before statistical mechanics formalized thermodynamics via the statistical properties of ensembles, Henri Poincaré's failed 1892 attempt to establish the orderliness of planetary motion showed that both determinism and randomness are essential and unavoidable in the study of physical systems [2, 3, 4, 5]. He showed that a system of three bodies interacting in a gravitational field displayed strange mathematical properties—the existence of infinitely many periodic solutions. In the three-body system, and others like it, intricate structures in the state space allow small differences in trajectories to amplify over time, resulting in divergent futures, despite deterministic evolution of the state space [5]. It was not until the 1960s and 1970s that the rise of dynamical systems theory and the exploration of statistical physics of critical phenomena gave insight into this phenomenon, which we now know as *chaos*.

Chaotic systems—by amplifying small differences over time—generate uncertainty. The rate of this generation can be characterized using information theory: the *Shannon entropy rate* h_μ is how many bits of information a system generates over time, again recalling the relationship between uncertainty and information. It has been shown that this measure is not abstract—perhaps surprisingly, the entropy rate of a system has a deep relationship to its dynamics [6, 7]. For chaotic systems, the entropy rate is positive, and this informs not only their information generation capabilities but also the instability of their orbits. We have in hand, then, a well-motivated description of how much information a system can generate, and a way forward—through studying the system's dynamics—towards calculating it for an arbitrary system. However, this leaves open the most fundamental question we posed above: *how* a system generates information.

Rather than immediately addressing that question, we will consider for a moment an alternative view on the entropy rate. If we were to observe a system for eternity, recording every pattern that we measured, in an attempt to make predictions about future behavior, our efforts would be inescapably limited by the system's h_μ. The entropy rate is a system's *instrinsic randomness*. In other words, it

is not possible to reduce uncertainty in the future of system beyond its entropy rate—this being the mathematical principle that stymied Poincaré's attempt to predict celestial motion. Despite the seemingly grim nature of this reality, the existence of this limit does not frustrate our ability to analyze the natural world, but instead clarifies. The fundamental task of scientific inquiry can be said to be that of optimal prediction, which may now be well defined: for some system, it is the construction of a predictive model that reduces uncertainty in the future to the entropy rate. I wish to argue that the construction of such a model, and the answer to our question of how natural systems generate information, are one and the same.

Let us first approach this argument by metaphor: imagine a machine attached to a row of lights, which causes a single light to flash every second. This machine has been running forever, and will continue running forever. We wish to describe the *complexity* of the machine, only having access to the sequence. A common misconception is that the machine's structure is dictated only by the entropy of the sequence. However, consider that a maximally random sequence is simple to implement: if I need to choose randomly amongst N lights, an N-sided die will suffice. But a completely deterministic sequence—say the first light always flashes, no uncertainty, no variations—may be implemented just as easily. I use the same die as before, replacing the numbers two through N with a one, so each roll returns the same value. In both cases, high and low entropy, the die does not store or process information, because by design, each roll is independent of every previous roll.

On the other hand, say I randomly choose lights one through N except for the 100th time step, when the light is chosen by summing up the last 99 selections and taking modulo N. In this case, I must build a machine capable of storing the previous ninety-nine selections (or at least, keeping a running tally). Due to the interplay between randomness and the *correlations*[1] in the sequence, I require a machine with memory, or, if that term is too loaded, the ability to store information over time. Such a machine is more computationally complex than a die, under any reasonable definition of complexity. However, we must be careful: I could, if I so desired, build for my entirely random sequence a labrynthine Rube Goldberg machine that, at the very end, rolls a die. Such a machine could certainly be made to appear, to the human eye, to be more "complex" than a pocket calculator. To avoid the endless possibility of complexity inflation in this manner, we employ

[1]The correlations we discuss here are temporal, but spatio- and spatio-temporal correlations are of great interest to complexity science as well, see [8, 9, 10, 11, 12].

an Occam's razor argument: if the machine is minimal—its computational resources are exactly sufficient to reproduce all correlations present in the sequence and no more—the complexity of any capable system producing the sequence must be at least as complex as the machine. That this minimal machine—known as the ϵ-*machine*—exists, is unique, and is the optimally predictive model is the fundamental result of the field known as *computational mechanics* [13].

Now answering the question of how systems generate, process, store, and transform information—i.e., how they compute—becomes a task of finding and characterizing the ϵ-machine of the system. However, as we have already implied, consistently measuring the complexity of a machine is not trivial. The definition and meaning of structural complexity is a long standing and much discussed topic [14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25]. Perhaps the most well-known and compelling complexity measure is the Kolmogorov-Chaitin (KC) complexity, length of the minimal program for a given Universal Turing Machine (UTM) required to reconstruct an individual time series [26, 27, 28, 29]. This formalizes the "machine in a box" problem outlined above, and represents both the randomness of the sequence and the complexity of the machine required to construct it by a single value. Unfortunately, due to the halting problem, the KC complexity is uncomputable, even if one has in hand a generative model.

Fortunately, there is a broad class of systems for which the growth of the KC complexity of typical realizations is well characterized: stationary, ergodic processes. They are employed in many fields, from the study of complex systems [30], to coding theory [31], stochastic processes [32], stochastic thermodynamics [33], speech recognition [34], computational biology [35, 36], epidemiology [37], and finance [38]. For these systems, KC complexity grows linearly with time series length, with coefficient equal to h_μ and offset equal to the *statistical complexity* C_μ, a measure of memory resources [14, 30, 39]. Both measures are *intrinsic* to the process, which is to say that they exist independently of the amount of data observed. Furthermore, they are both computable from the ϵ-machine.

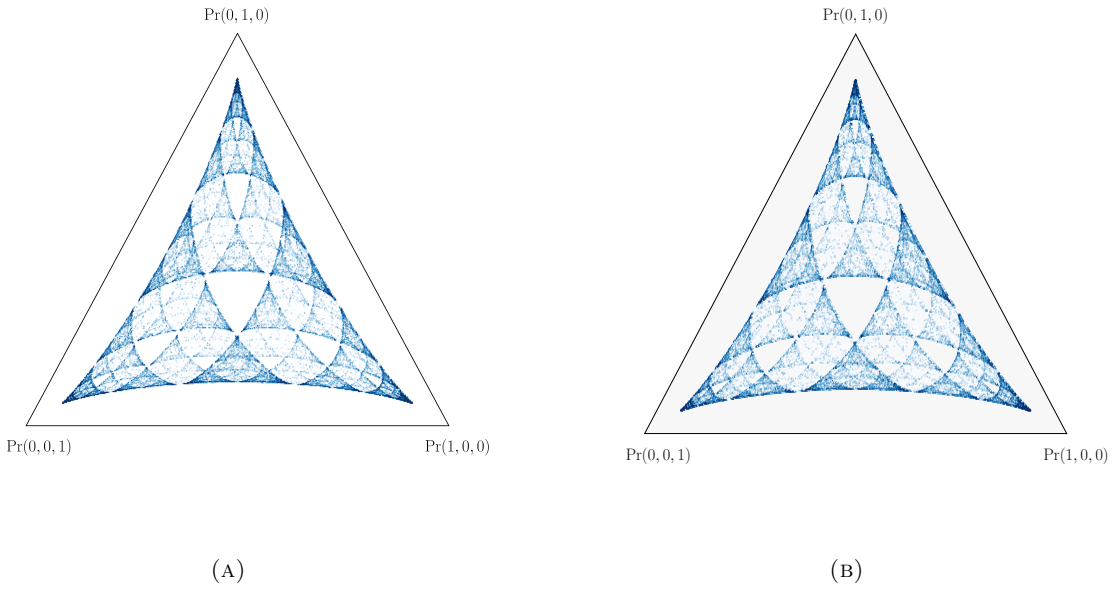
The space of stationary, ergodic processes is well organized by the divergence of their informational quantities into a prediction hierarchy, depicted in figure **??** [40]. At the lowest level are Markov processes, described by finite ϵ-machines with finite history dependence (finite Markov order R); e.g., measure subshifts of finite type [41]. Though very commonly posited as models, they

inhabit a vanishingly small measure in the space of processes [42]. At the next level of structure are Sofic processes described by ϵ-machines with finite C_μ. For processes at this level and below, computational mechanics has shown how to construct [13], characterize, [43, 44, 45] and infer [46] ϵ-machines. These ϵ-machines are represented as finite-state predictive hidden Markov models (HMM), like the two-state example shown in figure 2.1.

Despite this success, the limitation to finite state predictive HMMs is quite restrictive. At the Generative level are processes that may be finitely generated (generative complexity $C_{\text{gen}} < \infty$ [47]) but have ϵ-machines with infinite states [48]. Indeed, the state sets are often fractal, or even continuous. The difficulties of dealing with such systems are well-known; as originally noted by David Blackwell in 1957 [**?**], calculating the entropy rate of such processes is difficult, involving an integral over fractal measures. In addition, the statistical complexity C_μ diverges. Characterizing the computational architecture of such processes was previously limited to the visual identification of the ϵ-machine state sets as fractal-like objects, see figure 3.4 for several examples.

This dissertation addresses this state of affairs by applying dynamical systems theory to generative processes, recasting the state sets of their ϵ-machines as the attractors of chaotic systems. Using this approach, we define the ϵ-machine as a hidden Markov-*driven iterated function system* (DIFSs), easily found given a generative finite state HMM. From this, we may calculate the entropy rate of finitely-generated hidden Markov processes in general, solving Blackwell's long-standing problem [48]. Towards characterizing the computational structure of these models, I introduced the *ambiguity rate* h_a, the rate at which the ϵ-machine *forgets* information [49]. Compare this to the entropy rate, the rate at which the system generates information: when these two rates are equivalent, the ϵ-machine forgets information at the same rate the system generates it, the model size remains constant over time, and the statistical complexity is finite. In this case, a finite-state predictive HMM will suffice.

This gives us intuitive description of systems requiring ϵ-machines with infinite statistical complexity. When the ambiguity rate is less than the entropy rate, the ϵ-machine cannot forget information as quickly as the system generates it, increasing the model size (and therefore, C_μ) over time. Processes at this level require an ever-growing amount of memory for accurate prediction. The



(A) (B)

FIGURE 1.1. Figures (a) and (b) plot 10^5 states of an uncountably-infinite state ϵ-machine. Each is constructed for a different hidden Markov process, both generated with a 3-state HMM.

statistical complexity diverges, and our ϵ-machine has, typically, an uncountable state set[2]. To track the rate of this increase, we introduce the *statistical complexity dimension* $\dim_\mu$, the information dimension of the ϵ-machines' state set, which measures the rate at which the statistical complexity diverges [**51**]. Traditionally, fractal dimensions like the information dimension are difficult to calculate. Again, we make use of dynamical systems: the long standing Kaplan-Yorke conjecture upper bounds the information dimension of an attractor by computing the Lyapunov spectrum of the generating chaotic system. For ϵ-machines with non-zero ambiguity rate, we show that the upper-bound is strict. To solve this problem, the ambiguity rate is proposed as the correction to the conjecture for hidden Markov-DIFSs, allowing direct calculation of $\dim_\mu$.

These results include several important theoretical contributions: solving Blackwell's long-standing problem of the entropy of general HMPs, and improving the Kaplan-Yorke conjecture for a broad class of stochastic processes. In addition, we now have a suite of tools to characterize

[2]It is possible for the statistical complexity to diverge for ϵ-machines with countable state sets, see [**50**]. This occurs when the Minkowski–Bouligand dimension of the state set is strictly an upper bound on the Hausdorff dimension. On the other hand, if the state set is uncountable, the statistical complexity will diverge.

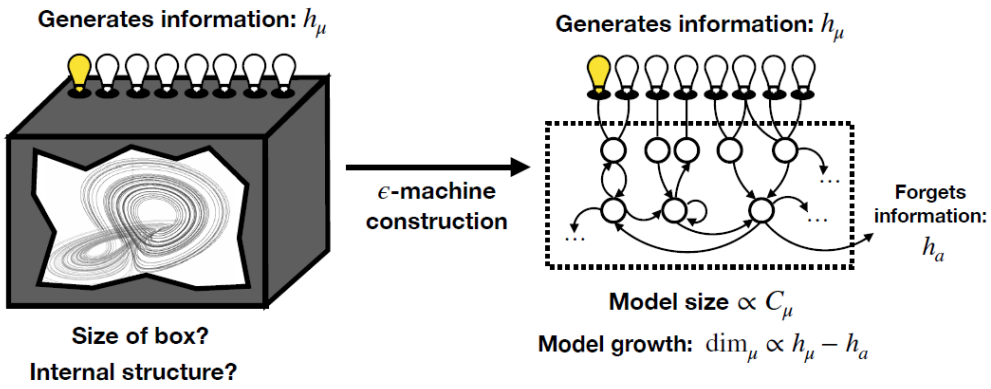


FIGURE 1.2. In a black box, a dynamical system generates h_μ bits per time step, observable via the sequence of lights. We wish to describe the "size" and complexity of the system in the box. The ϵ-machine of the system creates information at the same rate and forgets information h_a bits per time step. To characterize the system, we compute the ϵ-machine's size C_μ, rate of growth $\dim_\mu$, information storage, and information processing.

complex systems with divergent statistical complexity. Figure 1.2 shows the relationship between the quantities introduced. To highlight the usefulness of these informational quantities, that otherwise appear rather abstracted from natural systems, it should be noted that these tools have already found practical application in two, rather different, physical domains. The first analyzed the origin of randomness and structural complexity engendered by quantum measurement [**52**]. We showed that the ϵ-machines of measured quantum processes were highly uncountably-infinite state for almost every angle, but that $\dim_\mu$—the growth of the predictive model—varied continuously as a function of measurement angle. Perceived structural complexity of a measured process is then a complicated product of the measurement mechanism and the underlying "true" process. This has broad implications not only about our ability to communicate through quantum channels, but also how measurement of any system may induce apparent complexity. Our investigation of h_μ and C_μ's dependence on the underlying message versus the measurement channel provides the first steps towards untangling this induced complexity from the underlying system's behavior.

The second application involved information engines, Maxwellian demon-like devices designed to perform useful work by processing information. We solved a longstanding problem on exactly determining the thermodynamic functioning of engines described by general HMMs [**53**]. It was shown that even simple, well-known models require ϵ-machines with uncountably infinite state sets.

In particular, we showed that truncating descriptions of the structural complexity of these engines can result in mischaracterization of their thermodynamic functioning, and even implied violations of the Second Law. This result showed the practical use for consistent and correct measures of structural complexity, and perhaps more importantly, the cost of approximating away divergent complexities. Taken together, we believe these new tools will find even wider use than in these application areas.

CHAPTER 2

The Study of Stochastic Processes

The central object of this work is a class of *stochastic processes*, specifically those that may be finitely generated by a set of states. These processes are variously called *hidden Markov processes* (HMPs), functions of Markov chains, or stochastic finite automata. We will refer to them most frequently as HMPs. Their generating models—the aformentioned set of states and associated dynamic—are called hidden Markov models (HMMs). As described in the previous chapter, our goals are to model, predict, compare, and characterize the computational structure of these processes, and the natural systems that generate them, in the context of complexity theory.

Functions of Markov chains have been studied since the 1950s [**54**], when David Blackwell investigated them as producers of information. Since the 1970s, HMPs have been broadly applied in a number of fields and the simplicity of their description, flexibility of representation, and academic success has contributed to their broad popularity. In this chapter we define processes in general, then turn to HMPs and HMMs, assuming a basic familiarity with probability theory. We discuss the ϵ-machine, the minimal, optimally predictive model of an HMP. We then consider how to best study, characterize, and compare HMPs, introducing several fundamental concepts from *information theory*. We discuss the concepts of *intrinsic randomness* and *intrinsic structure*, the two axes upon which we will measure the computational complexity of HMPs.

2.1. Definition of a Process

Before defining a process exactly, we return to the machine described in Chapter 1: a constantly running mechanism attached to a finite row of lights, such that after certain time interval—say, every second—a single light flashes. Now, imagine that you lower this mechanism into a black box and place it on a laboratory bench, such that the lights are visible but the gears are not. You recall the specifications of the machine, but are no longer able to see how it is oriented. You may choose to place the mechanism in the box, noting its orientation, and begin recording the sequence of

lights immediately. Alternatively, you could walk away from the bench and allow the mechanism to run unobserved for a very long time, only to return and begin recording, having forgotten what orientation the mechanism was placed into the box with.

A process is the bi-infinite sequence of lights, or more technically, a process is a probability measure over a bi-infinite chain of random variables. Let X be a random variable, which may take on values from a finite set $\mathcal{A}$ according to a measure—in the following, a probability measure. In general, we denote random variables with capital letters, and particular *realization*s with the corresponding lower case letter x. Sequences of random variables are written $\ldots X_{t-2} X_{t-1} X_t X_{t+1} X_{t+2} \ldots$, where the subscript denotes a time index, $t \in \mathbb{Z}$. We define a process in terms of the distribution over such sequences.

DEFINITION 1. *Let $\mathcal{A}$ be a countable set. Let $\Omega = \mathcal{A}^{\mathbb{Z}}$ be the set of bi-infinite sequences composed from $\mathcal{A}$, $\mathcal{F}$ the field of cylinder sets of Ω, and P a probability measure. A process $\mathcal{P}$ is the probability space $(\Omega, \mathcal{F}, P)$, with random variable $\overleftrightarrow{X}$.*

We may index into the process: $X_i = x_i$ denotes the ith element of the bi-infinite sequence $\overleftrightarrow{X}$. From this we define the shift map $\tau : \overleftrightarrow{X} \to \overleftrightarrow{X}$ as $(\tau X)_i = X_{i+1}$, which describes symbol-to-symbol dynamics. In practice, we will often work with finite-length *words* $X_{t:t'}$, where the first index is inclusive and the second exclusive: $X_{t:t'} = X_t \ldots X_{t'-1}$. We restrict in the following to stationary processes.

DEFINITION 2. *A proccess $\mathcal{P}$ is stationary if and only if*

$$\Pr(X_{t:t+\ell} = x^\ell) = \Pr(X_{0:\ell} = x^\ell) \tag{2.1}$$

for all $t \in \mathbb{Z}, \ell \in \mathbb{Z}^+$, and all $x^\ell \in X^\ell$.

In such cases, to characterize the process we only need to consider a process's length-ℓ *word distributions* $\Pr(X_{0:\ell})$, $\ell \in \mathbb{Z}^+$. We also consider the infinite sequence following time t, $\overrightarrow{X} = \lim_{\ell \to \infty} X_{t:t+\ell}$. Informally, we refer to this as the *future*. In parallel, we may consider the infinite sequence preceding time t as the *past*, $\overleftarrow{X} = \lim_{\ell \to \infty} X_{t-\ell:t}$. As before, we may refer to a specific future $\overrightarrow{x}$ or a specific past $\overleftarrow{x}$ using the lower case.

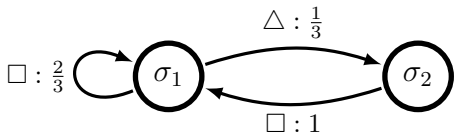


FIGURE 2.1. A hidden Markov model (HMM) with two states, $\{\sigma_1, \sigma_2\}$ and two symbols $\{\square, \triangle\}$. It is unifilar.

A *Markov process*, or Markov chain, is one for which $\Pr(X_t|X_{-\infty:t}) = \Pr(X_t|X_{t-1})$. Consequently, Markov processes are sometimes referred to as "memoryless", since the probability distribution over subsequent realizations depends only on the realization at time t. Functions over Markov processes are called *hidden Markov processes* (HMP), and are discussed at length below.

2.2. Hidden Markov Models

In general, we prefer not to work with processes directly, as dealing with measures over bi-infinite strings is cumbersome. For stochastic processes at the generative level and below (see **??**) we may instead use finite-state generative models called hidden Markov models (HMM).

DEFINITION 3. *A finite-state edge-labeled* hidden Markov model *(HMM) consists of:*

(1) a finite set of states $\boldsymbol{\mathcal{S}} = \{\sigma_1, ..., \sigma_N\}$,

(2) a finite alphabet $\mathcal{A}$ *of* k *symbols* $x \in \mathcal{A}$, *and*

(3) a set of N *by* N *symbol-labeled transition matrices* $T^{(x)}$, $x \in \mathcal{A}$: $T^{(x)}_{ij} = \Pr(\sigma_j, x|\sigma_i)$. *The corresponding overall state-to-state transitions are described by the row-stochastic matrix* $T = \sum_{x\in\mathcal{A}} T^{(x)}$.

The processes generated by HMMs are called hidden Markov processes (HMPs), so named for their internal state sequence $\ldots \mathcal{S}_{t-1}, \mathcal{S}_t, \mathcal{S}_{t+1} \ldots$, which is Markov. The non-Markovian symbol sequence $\ldots X_{t-1}, X_t, X_{t+1} \ldots$ is a function of the underlying state sequence. In general, our assumption is that the state sequence is not observable, i.e., *hidden*, while the symbol sequence is observable.

A given HMP has many possible generative HMMs. To see this, consider that states may be duplicated with identical transitions without changing the statistics of the generative process. Any valid generative HMM may be refered to as a *presentation* of the process. We now introduce a

structural property of HMMs that has important consequences in characterizing process randomness and internal state structure.

DEFINITION 4. *A* unifilar HMM *(uHMM) is an HMM such that for each state* $\sigma_i \in \boldsymbol{\mathcal{S}}$ *and each symbol* $x \in \mathcal{A}$ *there is at most one outgoing edge from state* σ_i *labeled with symbol* x*.*

An important consequence of unifilarity is that a uHMM's states are *predictive* in the sense that each is a function of the prior emitted sequence—the past $x_{-\infty:t} = \ldots x_{t-2}x_{t-1}x_t$. Consider an infinitely-long past that, in the present, has arrived at state σ_t. For uHMMs, it is not required that this infinitely-long past arrive at a unique state, but it is the case that any state arrived at by this past must have the same past-conditioned distribution of future sequences $\Pr(X_{\infty:t}|x_{-\infty:t})$. We call this deterministic relationship between the past and the future a *prediction*. In comparison, a nonpredictive generative (nonunifilar) HMM may return a set of states with varying conditional future distributions upon seeing this infinite past. All that is required for accurate generation is that, if this were to be repeated many times, averaging over these conditional future distributions returns the the unique conditional future distribution $\Pr(X_{\infty:t}|x_{-\infty:t})$ given by the predictive uHMC state.

2.3. The ϵ-Machine

For a given hidden Markov process $\mathcal{P}$, there exist infinitely many generative and predictive presentations. We can see this by considering that if there exists one, it is always possible to split a single state into two without changing the statistics of the generated strings. This frustrates our ability to use HMM representations of HMPs as a gauge of complexity, since it is always possible to construct a machine with additional states. However, there is a canonical presentation that is unique: a process' ϵ-*machine*, the minimal, optimally predictive model of a process. For stationary HMPs the ϵ-machine will be itself an HMM [13], the states of which are a defined by an equivalence class over pasts that lead to the same conditional future.

In other words, the causal states are the range of the ϵ-function

$$\epsilon(\overleftarrow{x} = \overleftarrow{X}) = \{\overleftarrow{x}' \mid \Pr(\overrightarrow{X}|\overleftarrow{X} = \overleftarrow{x}) = \Pr(\overrightarrow{X}|\overleftarrow{X} = \overleftarrow{x}')\} \ . \tag{2.2}$$

If $\epsilon(\overleftarrow{x}) = \epsilon(\overleftarrow{x}')$, both pasts 'belong' to, or induce, the same causal state. The class of causal states is denoted by $\boldsymbol{\mathcal{S}}$, and the random variable $\mathcal{S} = \sigma$.

The causal state dynamic is inherited from the shift map of the process. If $\overleftarrow{x}$ induces the causal state σ_i and $\overleftarrow{x}' = \overleftarrow{x}\,x$ induces the causal state σ_j we say that

$$\Pr(\sigma_j, x|\sigma_i) = \Pr\left((\tau X)_0 = X_1 = x\right) \ .$$

Defining the ϵ-machine via the ϵ-function requires finding a partition over the space of pasts $\overleftarrow{X}$. There is no prescription on the cardinality of $\boldsymbol{\mathcal{S}}$—it may be finite, it may be infinite, it may be fractal, it may be continuous. However, if we have in hand a uHMM of the process, we may easily find the ϵ-machine by enforcing a simple minimality condition.

Definition 5. *An* ϵ-machine *is a uHMM with* probabilistically distinct states*: For each pair of distinct states* $\sigma_i, \sigma_j \in \boldsymbol{\mathcal{S}}$ *there exists a finite word* $w = x_{0:\ell-1}$ *such that:*

$$\Pr(X_{0:\ell} = w|\mathcal{S}_0 = \sigma_k) \neq \Pr(X_{0:\ell} = w|\mathcal{S}_0 = \sigma_j) \ .$$

A process' ϵ-machine is its optimal, minimal presentation, in the sense that the set $\boldsymbol{\mathcal{S}}$ is minimal compared to all its other predictive presentations [13].

2.4. Some Information Theory

In order to meaningfully and effectively characterize stochastic processes and their ϵ-machines, we must turn for a moment to Shannon's information theory [1, 55]. Information theory is a widely-used foundational framework that provides tools to describe how stochastic processes generate, store, and transmit information. We use information theory—as compared to more traditional statistical methods—to study complex systems as it makes minimal assumptions as to the nature of correlations between random variables and handles multi-way, nonlinear correlations that are common in complex processes. Here, we introduce several information theoretic measures that will be required in the following.

The most basic measure of Shannon information theory is the Shannon *entropy*, which measures, intuitively, the amount of information that one learns when observing (or, from the opposite perspective, the amount of uncertainty one faces when predicting) a sample of a random variable. The entropy $H[X]$ of the random variable X is:

$$H[X] = -\sum_{x \in \mathcal{A}} \Pr(X = x) \log_2 \Pr(X = x) \ , \tag{2.3}$$

taking $0 \log_2 0 = 0$. Entropy, when taken with base 2, is measured in *bits* of information.

We can probe the relationship between two jointly-distributed random variables, say, X, drawn from $\mathcal{A}$, and Y, drawn from $\mathcal{B}$. The *joint entropy* $H[X, Y]$ is defined

$$H[X] = -\sum_{x \in \mathcal{A}, y \in \mathcal{B}} \Pr(X = x, Y = y) \log_2 \Pr(X = x, Y = y) \ , \tag{2.4}$$

and is interpreted as the total amount of information learned when observing X and Y simultaneously. Joint entropy may be extended to arbitrarily many random variables, in the obvious way. There is also the *conditional entropy*, which gives the amount of information learned from observation of one random variable given knowledge of another:

$$H[X|Y] = H[X, Y] - H[Y] \ . \tag{2.5}$$

There is a chain rule for multivariable conditional entropies:

$$H[X_1, X_2, \ldots, X_N] = \sum_{i=1}^{N} H[X_i | X_1, \ldots, X_{i-1}] \ .$$

The fundamental measure of correlation between random variables is the *mutual information*. It can be written in terms of Shannon entropies:

$$I[X; Y] = H[X, Y] - H[X|Y] - H[Y|X] \ . \tag{2.6}$$

As should be clear by inspection, the mutual information between two variables is symmetric. When X and Y are independent, the mutual information between them is zero. As with entropy, we may condition the mutual information on another random variable, giving the *conditional mutual*

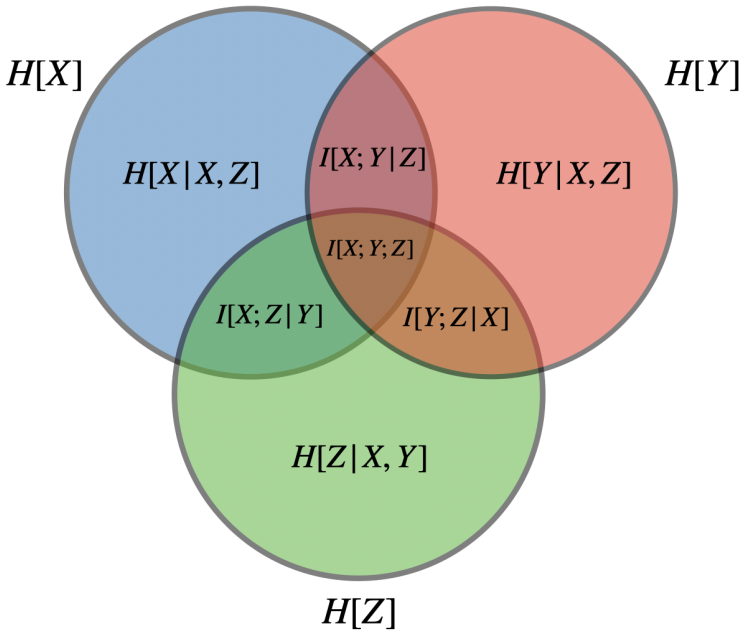


FIGURE 2.2. A three-variable i-diagram for random variables X, Y and Z, showing all possible information atoms, including conditional entropies, conditional mutual informations, and multivariate mutual information.

information:

$$I[X;Y|Z] = H[X|Z] + H[Y|Z] - H[X,Y|Z] \ . \tag{2.7}$$

The conditional mutual information is the amount of information shared by X and Y, given we know a third Z. Note that X and Y can share mutual information, but be conditionally independent. Moreover, conditioning on a third variable Z can either increase or decrease mutual information [**55**]. That is, the two variables can appear more or less dependent, given additional data.

An effective method of representation of information theoretic properties is the *i-diagram*, a type of Venn diagram where the area contained within an atom represents its informational content. Shared information is represented by overlap, and conditioning and mutual informations can be naturally defined. A three variable example is given in Fig. 2.2, showing all possible information atoms.

2.5. Intrinsic Randomness

Now that we have in hand a toolset with which to analyze stochastic processes, we must define the properties of interest. The first is the *instrinic randomness*: the amount of uncertainty remaining in the next observable symbol X_0, given complete knowledge of the infinite past. Taken from the perspective of an observer, this is the average amount of new information learned from every observation of a process. We call this the Shannon *entropy rate* [1] of a process.

DEFINITION 6. *A process'* entropy rate h_μ *is the asymptotic average entropy per symbol [43]:*

$$h_\mu = \lim_{\ell \to \infty} \frac{\mathrm{H}[X_{0:\ell}]}{\ell} \ , \tag{2.8}$$

where $H[X_{0:\ell}]$ *is the Shannon entropy of block* $X_{0:\ell}$*:*

$$H[X_{0:\ell}] = -\sum_{x_{0:\ell} \in \mathcal{A}^\ell} \Pr(x_{0:\ell}) \log_2 \Pr(x_{0:\ell}) \ . \tag{2.9}$$

We can also describe this quantity using the conditional entropy:

$$h_\mu = \lim_{\ell \to \infty} \mathrm{H}[X_0 | X_{-\ell:0}] \ . \tag{2.10}$$

This version of h_μ makes the conditional relationship to the past clear, and converges more rapidly than Eq. (2.8). Still, working with block entropies is difficult, and requires calculation of combinatorially-growing set of probabilities. Therefore, our preferred manner of calculating h_μ is to make use of an uHMM. Given a finite-state unifilar presentation M_u of a process $\mathcal{P}$, we may directly calculate the entropy rate from the transition matrices of the machine [1]:

$$\begin{aligned} h_\mu(\mathcal{P}) &= h_\mu(M_u) \\ &= -\sum_{\sigma \in \boldsymbol{\mathcal{S}}} \Pr(\sigma) \sum_{x \in \mathcal{A}} T^{(x)}_{\sigma\sigma'} \log_2 T^{(x)}_{\sigma\sigma'} \ . \end{aligned} \tag{2.11}$$

Here, $\Pr(\sigma)$ is the internal Markov chain's stationary state distribution, denoted π and determined by T's left eigenvector normalized in probability: $\pi = \pi T$.

However, this equation is only applicable when a finite-state uHMM can be found for the process. This is not always possible: Blackwell showed that for HMPs in general is no closed-form

expression for the entropy rate [54]. For a process generated by a nonunifilar HMC M, applying Eq. (2.11) to M typically overestimates the true entropy rate of the process $h_\mu(\mathcal{P})$:

$$h_\mu(M) \geq h_\mu(\mathcal{P}) \ .$$

Overcoming this limitation is one of our central results.

2.6. Intrinsic Structure

It is common to conflate the intrinsic randomness of a process with its computational complexity, but as discussed in Chapter 1, the two are not equivalent. We must consider not only the randomness of a process but also its *intrinsic structure*. There are many measures of structural complexity, but we will here use measurements of the ϵ-machine, arguing for its suitability as a complexity measure due to its minimality, uniqueness and optimality. Depending on the specific need, we may describe the ϵ-machine either in terms of the number of causal states $|\boldsymbol{\mathcal{S}}|$ or the amount of historical Shannon entropy they store—that is, the *statistical complexity* C_μ.

Definition 7. *A process'* statistical complexity *is the Shannon entropy stored in its ϵ-machine's causal states:*

$$\begin{aligned} C_\mu &= \mathrm{H}[\Pr(\mathcal{S})] \\ &= -\sum_{\sigma \in \mathcal{S}} \pi_\sigma \log_2 \pi_\sigma \ . \end{aligned} \tag{2.12}$$

A process' ϵ-machine is its smallest uHMC presentation, in the sense that both $|\boldsymbol{\mathcal{S}}|$ and C_μ are uniquely minimized by a process' ϵ-machine, compared to all other unifilar presentations. Due to this, the ϵ-machine's state entropy $H[\Pr(\mathcal{S})]$ is a unique measure of a process' structural complexity.

A challenge arises similar to that encountered with determining a process' entropy rate via its nonunifilar HMC presentations: without a finite-state uHMM, there is no closed-form expression for the generated process' C_μ. As we will show, this case is vanishingly small—every finite-state uHMM may be transformed into a nonunfilar HMM with a $\epsilon \to 0$ magnitude perturbation of elements of the transition matrices $T^{(x)}$. In contrast to the entropy rate, which may be at least estimated from

data using Eq. (2.10) or Eq. (2.8), this precludes nearly all processes from a consistent measure of structural complexity.

In the following, we will solve this problem by introducing a method of finding the ϵ-machine for any process that may be generated by a finite state HMM. This allows us to calculate the entropy rate h_μ, however, these ϵ-machines are in general infinite state, and we often will find $C_\mu \to \infty$. Therefore, we develop a new measure of intrinsic structural complexity, the *statistical complexity dimension*, which describes the rate at which C_μ diverges as memory resources are added to the model.

CHAPTER 3

The Mixed State Presentation: Building Predictive Models

The major mathematical conceit of this dissertation is the expression of the state space of ϵ-machines as the attractors of a type of stochastic dynamical system known as iterated function systems (IFSs). By doing so, we are able to make use of the full power of dynamical systems theory to analyze the ϵ-machine. In particular, in Chapter 4 we make use of attractor sampling techniques and ergodic theory to devise a method of calculating the entropy rate, and in Chapter 5 and Chapter 6 we devise new intrinsic complexity measures in this setting.

This chapter describes the relevant properties of our class of IFSs defined by HMMs, which we call hidden Markov-driven iterated function systems (DIFSs). In particular, the uniqueness and invariance of the DIFSs attractor is established, and identified as the states of the ϵ-machine machine for the process generated by the DIFSs. A general algorithm for constructing ϵ-machines from nonunifilar HMMs is given, with several specific examples.

3.1. Introducing Iterated Function Systems

To begin, we will consider iterated function systems in general. Let (Δ^N, d) be a compact metric space with $d(\cdot, \cdot)$ a distance. This notation anticipates our later application, in which Δ^N is N-simplex of discrete-event probability distributions (see Section 3.2.1). However, the results here are general.

Let $f^{(x)} : \Delta^N \to \Delta^N$ for $x = 1, \ldots, k$ be a set of Lipschitz functions with:

$$d\left(f^{(x)}(\eta), f^{(x)}(\zeta)\right) \leq \tau^{(x)} d(\eta, \zeta) \ ,$$

for all $\eta, \zeta \in \Delta^N$ and where $\tau^{(x)}$ is a constant. This notation is chosen to draw an explicit parallel to the stochastic processes discussed in Section 2.2 and to avoid confusion with the lowercase Latin characters used for realizations of stochastic processes. In particular, note that the superscript (x)

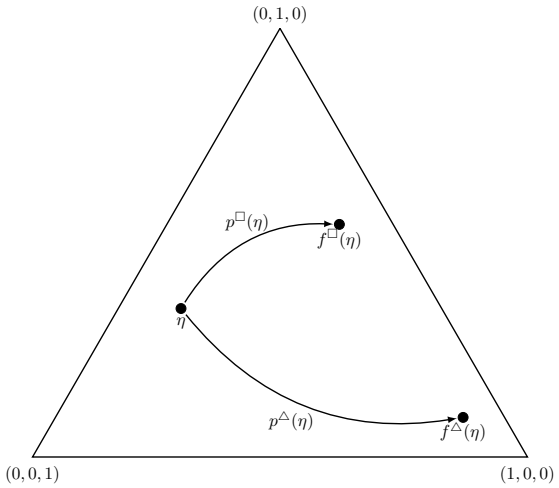


FIGURE 3.1. How a hidden Markov-driven iterated function system (DIFS) generates a hidden Markov process: An initial state η—a distribution over three states: $(0, 0, 1)$, $(0, 1, 0)$, and $(1, 0, 0)$—in the 2-simplex is associated with a transition probability distribution over the alphabet $\mathcal{A} = \{\square, \triangle\}$. If the emitted symbol selected from this distribution is $\square$, the next state is generated according to the associated mapping function $f^{(\square)}(\eta)$ and the probability distribution is updated accordingly. The same steps are followed if the symbol is $\triangle$ using $f^{(\triangle)}(\eta)$, resulting in an emitted process $\mathcal{P}$ over symbols $\mathcal{A}$.

here and elsewhere parallels that of the HMM symbol-labeled transition matrices $T^{(x)}$. The reasons for this will soon become clear.

The Lipschitz constant $\tau^{(x)}$ is the *contractivity* of map $f^{(x)}$. Let $p^{(x)} : \Delta^N \to [0, 1]$ be continuous, with $p^{(x)}(\eta) \geq 0$ and $\sum_{x=1}^{k} p^{(x)}(\eta) = 1$ for all η in M. The triplet $\{\Delta^N, \{p^{(x)}\}, \{f^{(x)}\} : x \in \mathcal{A}\}$ defines a *place-dependent* IFS.

A place-dependent IFS generates a stochastic process over $\eta \in \Delta^N$ as shown in Fig. 3.1. Given an initial position $\eta_0 \in \Delta^N$, the probability distribution $\{p^{(x)}(\eta_0) : x = 1, \ldots, k\}$ is sampled. According to the sample x, apply $f^{(x)}$ to map η_0 to the next position $\eta_1 = f^{(x)}(\eta_0)$. Resample x from the distribution $p^{(x)}(\eta_1)$ and continue, generating $\eta_0, \eta_1, \eta_2, \ldots$.

If each map $f^{(x)}$ is a contraction—i.e., $\tau^{(x)} < 1$ for all $\eta, \zeta \in \Delta^N$—it is well known that there exists a unique nonempty compact set $\Lambda \subset \Delta^N$ that is invariant under the IFS's action:

$$\Lambda = \bigcap_{x=1}^{k} f^{(x)}(\Lambda) \ .$$

Λ is the IFS's *attractor*.

Consider the operator $V : M(\Delta^N) \to M(\Delta^N)$ on the space of Borel measures on the N-simplex:

$$V\mu(B) = \sum_{x=1}^{k} \int_{\left(f^{(x)}\right)^{-1}(B)} p^{(x)}(\eta) d\mu(\eta) \ . \tag{3.1}$$

A Borel probability measure μ is said to be *invariant* or *stationary* if $V\mu = \mu$. It is *attractive* if for any probability measure ν in $M(\Delta^N)$:

$$\int g d(V^n \nu) \to \int g\mu \ ,$$

for all g in the space of bounded continuous functions on Δ^N.

Let us recall here a key result concerning the existence of attractive, invariant measures for place-dependent IFSs.

THEOREM 3.1.1. *[56, Thm. 2.1] Suppose there exists $r < 1$ and $q > 0$ such that:*

$$\sum_{x \in \mathcal{A}} p^{(x)}(\eta) d^q \left(f^{(x)}(\eta), f^{(x)}(\zeta) \right) \leq r^q d^q \left(\eta, \zeta \right) \ ,$$

for all $\eta, \zeta \in \Delta^N$. Assume that the modulus of uniform continuity of each $p^{(x)}$ satisfies Dini's condition and that there exists a $\delta > 0$ such that:

$$\sum_{x : d(f^{(x)}(\eta), f^{(x)}(\zeta)) \leq r d(\eta, \zeta)} p^{(x)}(\eta) p^{(x)}(\zeta) \leq \delta^2 \ , \tag{3.2}$$

for all $\eta, \zeta \in \Delta^N$. Then there is an attractive, unique, invariant probability measure for the Markov process generated by the place-dependent IFS.

In addition, under these same conditions Ref. [57] established an ergodic theorem for IFS *orbits*. That is, for any $\eta \in \Delta^N$ and $g : \Delta^N \to \Delta^N$:

$$\frac{1}{n+1} \sum_{k=0}^{n} g(w_{x_k} \circ \cdots \circ w_{x_1} \eta) \to \int g d\mu \ , \tag{3.3}$$

so that the statistics of any single finite sequence of mixed states $\eta_0, \eta_1, \ldots, \eta_N$ will approach the attractor $\boldsymbol{\mathcal{R}}$ and the measure μ as $N \to \infty$.

3.2. The Mixed-State Presentation

We now return to stochastic processes and their HMM presentations. When discussing the Shannon entropy rate in Section 2.5 we noted that nonunifilar HMC presentations lead to difficulties: (i) the internal Markov-chain $\{\boldsymbol{\mathcal{S}}, T\}$ entropy-rate overestimates the process' entropy rate and (ii) there is no closed-form entropy-rate expression. Furthermore, the states of nonunifilar HMCs are nonpredictive, there is no (known) unique minimal nonunifilar presentation of a given process. This precludes characterizing, in a unique and minimal way, a process' structural complexity directly from a nonunifilar presentation.

To develop the tools needed to resolve these problems, we introduce HMM *mixed states* and their dynamic. To motivate our development, consider the problem of *observer-process synchronization.*

Assume that an observer has a knowledge of a finite HMC M generating a process $\mathcal{P}$. The observer cannot directly observe M's internal states, but wishes to know which internal state M is in at any given time—to *synchronize* to the machine. Since the observer does have knowledge of M's transition dynamic, they can improve on their initial guess $(\Pr(\sigma_1), \Pr(\sigma_2), \ldots, \Pr(\sigma_N))$ by monitoring the output data $x_0\, x_1\, x_2 \ldots$ that M generates.

3.2.1. Mixed States. For a length-ℓ word w generated by M let $\eta(w) = \Pr(\boldsymbol{\mathcal{S}}|w)$ be the observer's guess as to the process' current state after observing w:

$$\eta(w) \equiv \Pr(\mathcal{S}_\ell | X_{0:\ell} = w, \mathcal{S}_0 \sim \pi) \ , \tag{3.4}$$

where the initial guess $\Pr(\mathcal{S}_0|\cdot)$ is π, M's stationary state distribution. When observing a N-state machine, the vector $\langle \eta(w)|$ lives in the *(N-1)-simplex* Δ^{N-1}, the set such that:

$$\{\eta \in \mathbb{R}^N : \langle \eta\, | \mathbf{1} \rangle = 1, \langle \eta\, | \delta_i \rangle \geq 0, i = 1, \ldots, N\} \ ,$$

where $\langle \delta_i | = \begin{pmatrix} 0 & 0 & \ldots & 1 & \ldots & 0 \end{pmatrix}$ and $|\mathbf{1}\rangle = \begin{pmatrix} 1 & 1 & \ldots & 1 \end{pmatrix}$. We use this notation to indicate components of the belief distribution vector η in order to avoid confusion with temporal indexing. When a mixed state appears in probability expressions, the notation refers to the random variable η, not the row vector $|\eta\rangle$, and we drop the bra-ket notation. Bra-ket notation is used in vector-matrix expressions.

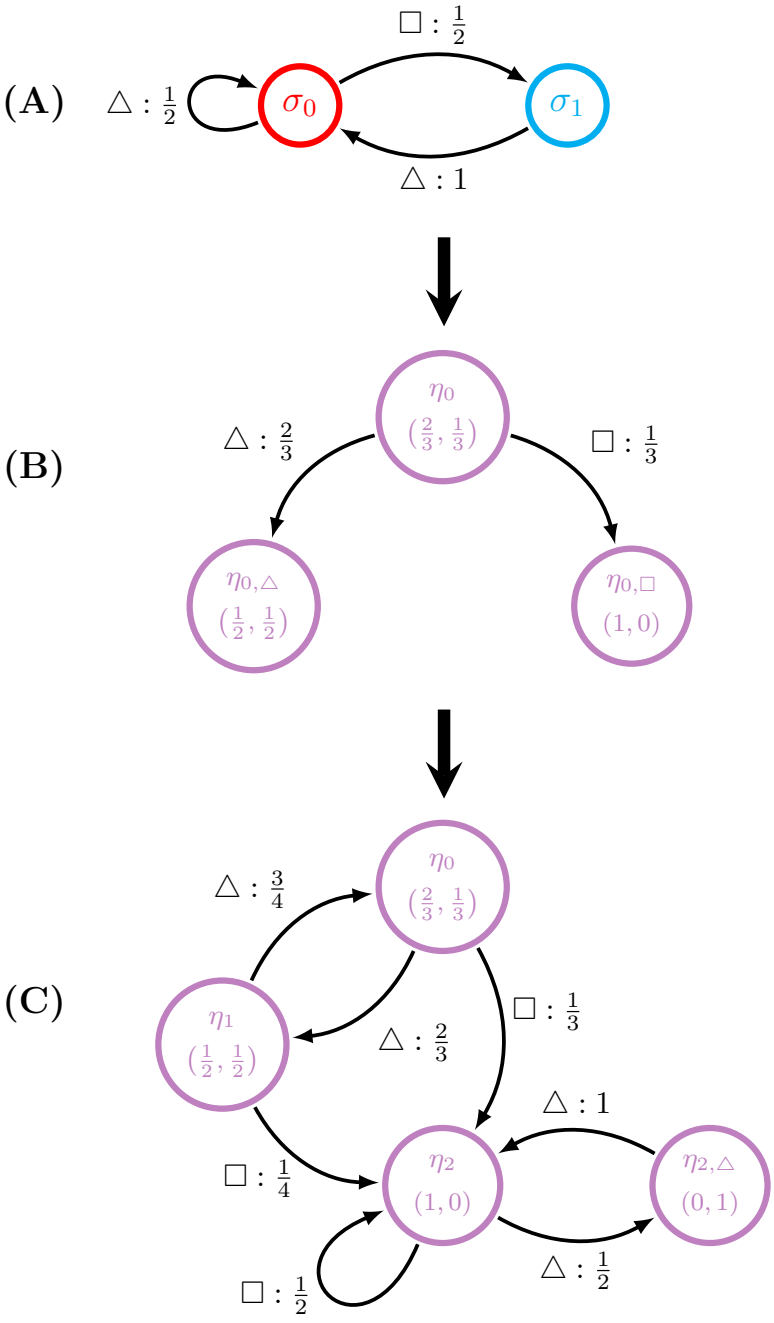


FIGURE 3.2. Determining the mixed-state presentation (MSP) of the 2-state unifilar HMC shown in (A): The invariant state distribution $\pi = (2/3, 1/3)$. It becomes the first mixed state η_0 used in (B) to calculate the next set of mixed states. (C) The full set of mixed states seen from all allowed words. In this case, we recover the unifilar HMC shown in (A) as the MSP's recurrent states.

The 0-simplex Δ^0 is the single point $|\eta\rangle = (1)$, the 1-simplex Δ^1 is the line segment $[0, 1]$ from $|\eta\rangle = (0, 1)$ to $|\eta\rangle = (1, 0)$, and so on.

The set of belief distributions $\eta(w)$ that an HMC can visit defines its set $\boldsymbol{\mathcal{R}}$ of *mixed states*:

$$\boldsymbol{\mathcal{R}} = \{\eta(w) : w \in \mathcal{A}^+, \Pr(w) > 0\} \ .$$

Generically, the mixed-state set $\boldsymbol{\mathcal{R}}$ for an N-state HMC is infinite, even for finite N [54].

3.2.2. Mixed-State Dynamic. The probability of transitioning from $\langle\eta(w)|$ to $\langle\eta(wx)|$ on observing symbol x follows from Eq. (3.4) immediately:

$$\Pr(\eta(wx)|\eta(w)) = \Pr(x|\mathcal{S}_\ell \sim \eta(w)) \ .$$

This defines the mixed-state transition dynamic $\mathcal{W}$. Together the mixed states and their dynamic define an HMC that is unifilar by construction. This is a process' *mixed-state presentation* (MSP) $\mathcal{U}(\mathcal{P}) = \{\boldsymbol{\mathcal{R}}, \mathcal{W}\}$.

We defined a process' $\mathcal{U}$ abstractly. The $\mathcal{U}$ typically has an uncountably infinite set of mixed states, making it challenging to work with in the form laid out in Section 3.2.1. Usefully, however, given any HMC M that generates the process, we can explicitly write down the dynamic $\mathcal{W}$. Assume we have an $N+1$-state HMC presentation M with k symbols $x \in \mathcal{A}$. The initial condition is the invariant probability π over the states of M, so that $\langle\eta_0| = \langle\pi|$. In the context of the mixed-state dynamic, mixed-state subscripts denote time.

The probability of generating symbol x when in mixed state η is:

$$\Pr(x|\eta) = \langle\eta| T^{(x)} |\mathbf{1}\rangle \ , \tag{3.5}$$

where $T^{(x)}$ is M's symbol-labeled transition matrix associated with the symbol x.

From η_0, we calculate the probability $\langle\eta_{1,x}|$ of seeing each $x \in \mathcal{A}$. Upon seeing symbol x, the current mixed state $\langle\eta_t|$ is updated according to:

$$\langle\eta_{t+1,x}| = \frac{\langle\eta_t| T^{(x)}}{\langle\eta_t| T^{(x)} |\mathbf{1}\rangle} \ . \tag{3.6}$$

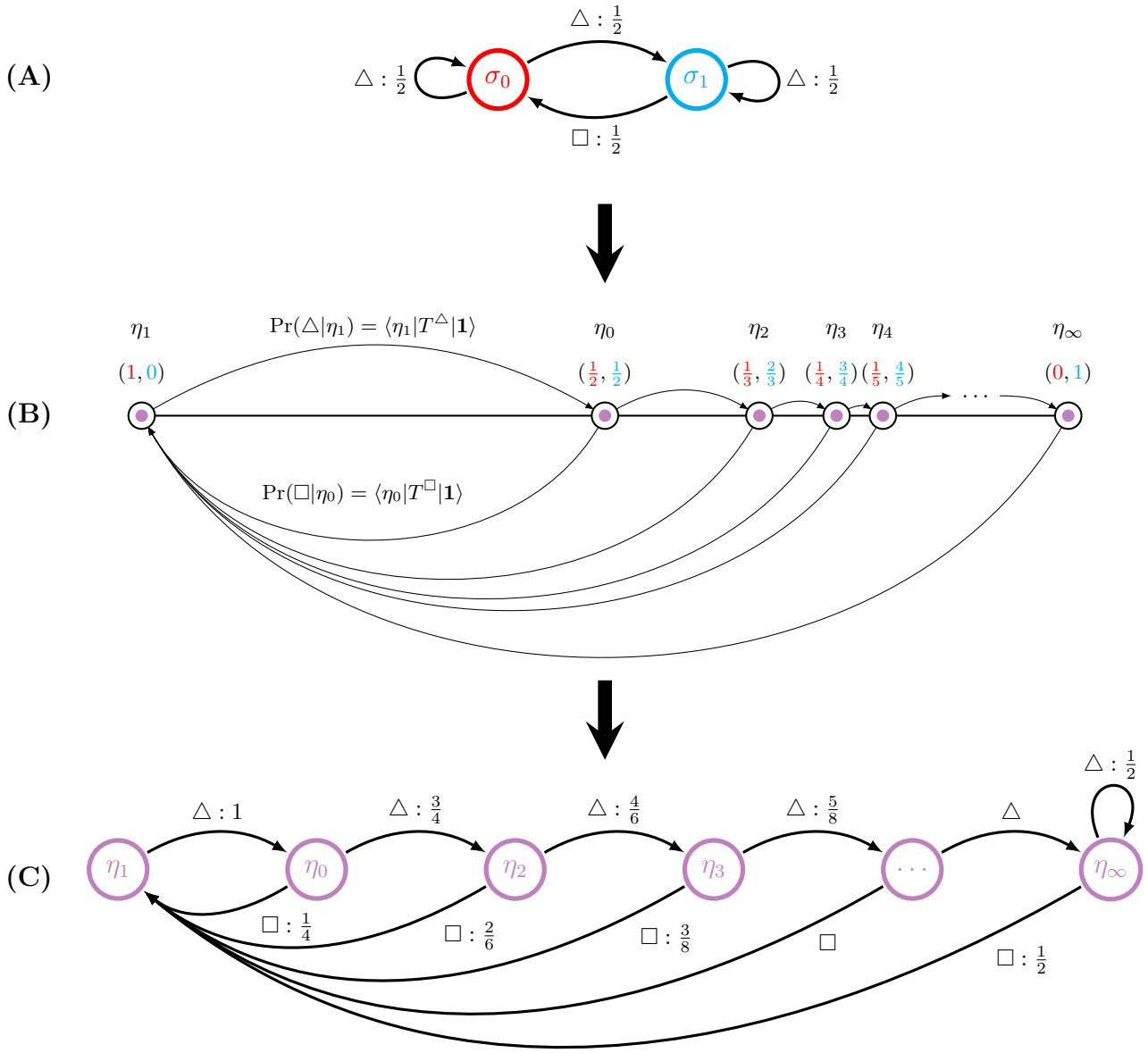


FIGURE 3.3. Determining the mixed-state presentation of the 2-state *nonunifilar* HMC shown in (A). The invariant distribution $\pi = (1/2, 1/2)$. It is the first mixed state η_0 used in (B) to calculate the next set of mixed states. (B) plots the mixed states along the 1-simplex $\Delta^1 = [0, 1]$. In (C), we translated the points on the simplex to the states of an infinite-state, unifilar HMC.

Thus, given an HMC presentation we can restate Eq. (3.4) as:

$$\begin{aligned}\langle \eta(w)| &= \frac{\langle \eta_0| T^{(w)}}{\langle \eta_0| T^{(w)} |\mathbf{1}\rangle} \\ &= \frac{\langle \pi| T^{(w)}}{\langle \pi| T^{(w)} |\mathbf{1}\rangle} \ .\end{aligned}$$

Equation (3.6) tells us that, by construction, the MSP is unifilar, since each possible output symbol uniquely determines the next (mixed) state. Taken together, Eqs. (3.5) and (3.6) define the mixed-state transition dynamic $\mathcal{W}$ as:

$$\begin{aligned}\Pr(\eta_{t+1}, x|\eta_t) &= \Pr(x|\eta_t) \\ &= \langle \eta_t| T^{(x)} |\mathbf{1}\rangle \ ,\end{aligned}$$

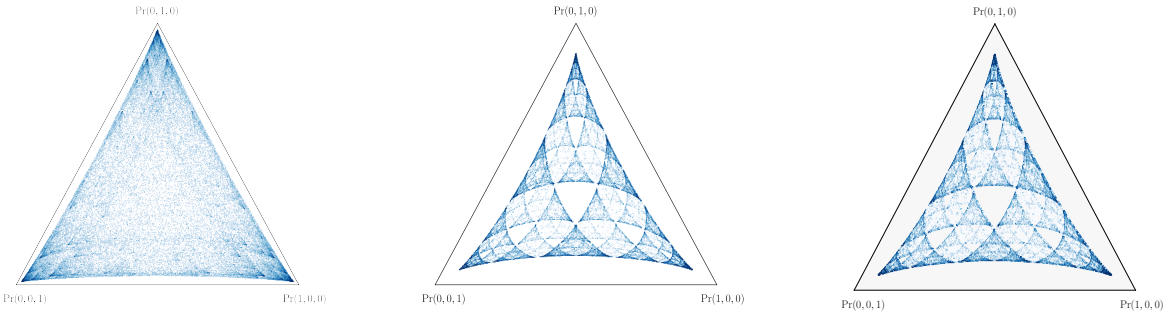
for all $\eta \in \boldsymbol{\mathcal{R}}$, $x \in \mathcal{A}$.

To find the MSP $\mathcal{U} = \{\boldsymbol{\mathcal{R}}, \mathcal{W}\}$ for a given HMC M we apply *mixed-state construction*:

(1) Set $\mathcal{U} = \{\boldsymbol{\mathcal{R}} = \emptyset, \mathcal{W} = \emptyset\}$.
(2) Calculate M's invariant state distribution: $\pi = \pi T$.
(3) Take η_0 to be $\langle \delta_\pi|$ and add it to $\boldsymbol{\mathcal{R}}$.
(4) For each current mixed state $\eta_t \in \boldsymbol{\mathcal{R}}$, use Eq. (3.5) to calculate $\Pr(x|\eta_t)$ for each $x \in \mathcal{A}$.
(5) For $\eta_t \in \boldsymbol{\mathcal{R}}$, use Eq. (3.6) to find the updated mixed state $\eta_{t+1,x}$ for each $x \in \mathcal{A}$.
(6) Add η_t's transitions to $\mathcal{W}$ and each $\eta_{t+1,x}$ to $\boldsymbol{\mathcal{R}}$, merging duplicate states.
(7) For each new η_{t+1}, repeat steps 4-6 until no new mixed states are produced.

Let us walk through these steps with a simple finite-state example. In Fig. 3.2(A) we have a unifilar HMC, which happens to be an ϵ-machine. The invariant state distribution of the machine is $\pi = (2/3, 1/3)$, so in Fig. 3.2(B) this becomes our initial mixed state η_0. Following steps 4 and 5 in the mixed-state construction, we calculate the probabilities of transition for each symbol in the alphabet $\{\triangle, \square\}$ and their resultant mixed states $\{\eta_{0,\triangle}, \eta_{0,\square}\}$. We then relabel these new mixed states $\{\eta_1, \eta_2\}$ and repeat. This process eventually results in Fig. 3.2(C), in which all possible transitions and mixed states have been found.

In Fig. 3.2(C) the recurrent states of the MSP, $\{\eta_2, \eta_3\}$, match exactly with the states of the original machine $\{\sigma_0, \sigma_1\}$. Therefore, the recurrent part of the $\mathcal{U}(M)$ is exactly the ϵ-machine.



(A) $\alpha = 0.6$ and $x = 0.025$

(B) $\alpha = 0.6$ and $x = 0.10$

(C) $\alpha = 0.6$ and $x = 0.49$

FIGURE 3.4. Figures (a), (b), and (c) each plot 10^5 mixed states of the uncountably-infinite state MSP generated by the parametrized 3-state HMC defined in Eq. (3.7) at various values of x and α. This HMC is capable of generating MSPs with a variety of structures, depending on x and α. However, due to rotational symmetry in the symbol-labeled transition matrices, the attractor is always radially symmetric around the simplex center.

When starting with the ϵ-machine, trimming the transient states from the $\mathcal{U}(\epsilon\text{-machine})$ in this way always returns the recurrent-state ϵ-machine, as in the above case. In general, if $\mathcal{U}(M)$ is finite, we find the ϵ-machine by minimizing $\mathcal{U}(M)$ via merging duplicate states: repeat mixed-state construction on $\mathcal{U}(M)$ and trim transient states once more.

Is the MSP always the same as the ϵ-machine? When beginning with a finite, unifilar HMC M generating a process $\mathcal{P}$, the MSP $\mathcal{U}(M)$ is a finite, optimally-predictive rival presentation to $\mathcal{P}$'s ϵ-machine. Trimming the transient states will always return the recurrent-state ϵ-machine, as in the above case. The MSPs of unifilar presentations are interesting and contain additional information beyond the unifilar presentations. For example, containing transient causal states, they are employed in calculating many complexity measures that track convergence statistics [**58**].

However, here we focus on the mixed-state presentations of nonunifilar HMCs, which typically have an infinite mixed-state set $\boldsymbol{\mathcal{R}}$. Figure 3.3 illustrates applying mixed-state construction to a finite, nonunifilar HMC. This produces an infinite sequence of states mixed on the 1-simplex, as depicted in Fig. 3.3(B). In this particular example, the MSP is clearly structured and $\boldsymbol{\mathcal{R}}$ is countably infinite, allowing us to better understand the underlying process $\mathcal{P}$; compared, say, to

the 2-state nonunifilar HMC in Fig. 3.3(A). This is, indeed, the ϵ-machine, as is clear from the fact η_n are probabilistically distinct and predictive. From the causal states, the process' structure—a discrete-time renewal process—becomes manifest and the entropy rate may be directly calculated, adapting Eq. (2.11), as an infinite sum over the states η_n as $n \to \infty$.

That said, MSPs of nonunifilar HMCs typically have an uncountably-infinite mixed-state set $\boldsymbol{\mathcal{R}}$. Consider the following HMC with 3 symbols and 3 states:

$$T^{\square} = \begin{pmatrix} \alpha y & \beta x & \beta x \\ \alpha x & \beta y & \beta x \\ \alpha x & \beta x & \beta y \end{pmatrix}, \; T^{\triangle} = \begin{pmatrix} \beta y & \alpha x & \beta x \\ \beta x & \alpha y & \beta x \\ \beta x & \alpha x & \beta y \end{pmatrix}, \text{ and}$$

$$T^{\circ} = \begin{pmatrix} \beta y & \beta x & \alpha x \\ \beta x & \beta y & \alpha x \\ \beta x & \beta x & \alpha y \end{pmatrix}, \tag{3.7}$$

with $\beta = (1-\alpha)/2$ and $y = 1 - 2x$. By inspection, we see that α takes on any value from 0 to 1 and x may range from 0 to 1/2. Figure 3.4 shows three attractors from the this parameterized 3-state HMC at different points in its parameter space. Our goal then is a set of constructive results for this class of $\mathcal{U}$s: for a given nonunifilar finite-state HMC, determine whether we are guaranteed to have (i) a well-defined, unique, mixed-state set $\boldsymbol{\mathcal{R}}$, (ii) an invariant measure over $\mu(\boldsymbol{\mathcal{R}})$, (iii) an ergodic theorem, and (iv) a notion of minimality. With these established, we can use $\mathcal{U}$ as a candidate for a process' ϵ-machine.

3.3. The MSP as an IFS

With the mixed-state presentation introduced and the goals outlined, our intentions in reviewing iterated function systems (IFSs) become explicit. The MSP exactly defines a place-dependent IFS, where the mapping functions are the set of symbol-labeled mixed-state update functions of Eq. (3.6) and the set of place-dependent probability functions are given by Eq. (3.5). We then have a mapping function and associated probability function for each symbol $x \in \mathcal{A}$ that can be derived from the symbol-labeled transition matrix $T^{(x)}$.

If these probability and mapping functions meet the conditions of Theorem 3.1.1, we identify the attractor Λ as the set of mixed states $\boldsymbol{\mathcal{R}}$ and the invariant measure μ as the invariant distribution π of the potentially infinite-state $\mathcal{U}$—the original HMC's *Blackwell measure*. Since all Lipschitz continuous functions are Dini continuous, the probability functions meet the conditions by inspection.

We now establish that the maps are contractions. For maps defined by nonnegative, aperiodic, and irreducible matrices, we appeal to Birkhoff's 1957 proof that a positive linear map preserving a convex cone is a contraction under the Hilbert projection metric [**59**]. This result, which may be extended to any nonnegative $T^{(x)}$ if there is an $N \in \mathbb{N}^+$ such that $\left(T^{(x)}\right)^N$ is a positive matrix, is summarized in Appendix C.

Although this covers a broad class of nonunifilar HMCs, we are not guaranteed irreducibility and aperiodicity for symbol-labeled transition matrices. Indeed, several of the more interesting examples encountered do not meet this standard. For example, consider the *Simple Nonunifilar Source* (SNS), depicted in Fig. 3.3, defined by the symbol-labeled transition matrices:

$$T^{(\triangle)} = \begin{pmatrix} 1-p & p \\ 0 & 1-q \end{pmatrix} \text{ and } T^{(\square)} = \begin{pmatrix} 0 & 0 \\ q & 0 \end{pmatrix} . \tag{3.8}$$

In this case *both* $T^{(\triangle)}$ and $T^{(\square)}$ are reducible. (A quick check for this property is to examine Fig. 3.3 (A) and ask if there is a length-n sequence consisting of only a single symbol that reaches every state from every other state.) Nonetheless, the HMC has a countable set of mixed states $\boldsymbol{\mathcal{R}}$ and an invariant measure μ.

We can show this with the mapping functions:

$$\begin{aligned} f^{(\triangle)}(\eta) = \Bigg[& \frac{\langle \eta \,|\delta_1\rangle\,(1-p)}{1-(1-\langle \eta \,|\delta_1\rangle)q}, \\ & \frac{\langle \eta \,|\delta_1\rangle\, p + (1-\langle \eta \,|\delta_1\rangle)(1-q)}{1+(1-\langle \eta \,|\delta_1\rangle)q} \Bigg] \text{ and} \end{aligned}$$

$$f^{(\square)}(\eta) = [1, 0] \ . \tag{3.9}$$

Recall here that $\langle \eta \,|\delta_1\rangle$ is simply the first component of η. From any initial state η_0, other than $\eta_0 = \sigma_0 = [1, 0]$, the probability of seeing a $\square$ is positive. Once a $\square$ is emitted, the mixed state is guaranteed to be $\eta = \sigma_0 = [1, 0]$. When the mapping function is constant in this way and the

contractivity is $-\infty$, we call the symbol a *synchronizing* symbol. From σ_0, the set of mixed states is generated by repeated emissions of $\triangle$s, so that $\boldsymbol{\mathcal{R}} = \left\{ \left(f^{(\triangle)} \right)^n (\sigma_0) : n = 0, \ldots, \infty \right\}$. This is visually depicted in Fig. 3.3 for the specific case of $p = q = 1/2$. For all p and q, the measure can be determined analytically; see Ref. [**60**]. This analyticity is due to the HMC's countable-state structure, a consequence of the synchronizing symbol.

This example, including the uniqueness of the IFS attractor Λ, helps establish which HMC class generates ergodic processes: those whose total transition matrix $T = \sum_x T^{(x)}$ is nonnegative, irreducible, and aperiodic. Consider an HMC in this class. Define for any word $w = x_1 \ldots x_\ell \in \mathcal{A}^+$ the associated mapping function $T^{(w)} = T^{(x_1)} \circ \cdots \circ T^{(x_\ell)}$. Consider word w in a process' typical set of realizations (see Appendix B), which approaches measure one as $|w| \to \infty$. Due to ergodicity, it must be the case that $f^{(w)}$ is either (i) a constant mapping—and, therefore, infinitely contracting—or (ii) $T^{(w)}$ is irreducible.

As an example of case (i), any composition of the SNS functions Eq. (3.9) is always a constant function, so long as there is at least one $\square$ in the word, the probability of which approaches one as the word grows in length.

As an example of case (ii), imagine adding to the SNS in Fig. 3.3 (A) a transition on $\square$ from σ_0 to σ_1. For this new machine, both symbol-labeled transition matrices are still reducible, but the composite transition matrices for *any* word including both symbols will be irreducible. By Birkhoff's argument, the map associated with that word is contracting. There are only two sequences for which this does not occur: $w = \square^N$ and $w' = \triangle^N$. However, these sequences are measure zero as $N \to \infty$. Appendix B discusses this argument further.

In short, we extend the result of Theorem 3.1.1 to any HMC with nonnegative substochastic transition matrices, as long as $T = \sum_x T^{(x)}$ is nonnegative, irreducible, and aperiodic, regardless of the properties of the individual maps.

Before moving on, let us highlight the implications of this result. For any process $\mathcal{P}$ that may be generated by a finite-state HMC, we now have a guarantee of a unique, attracting set of mixed states $\boldsymbol{\mathcal{R}}$, with an invariant, attracting measure $\mu(\boldsymbol{\mathcal{R}})$. Furthermore, we appeal to established IFS results for an ergodic theorem over long words [**57**]. So, by introducing a check for minimality, as discussed in Section 3.4, we may identify the MSP $\mathcal{U}(M)$ as the infinite-state ϵ-machine for the

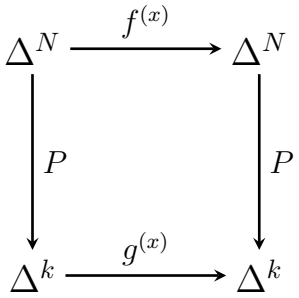


FIGURE 3.5. Commuting diagram for probability functions $P = \{p^{(x)}\}$, mixed-state mapping functions $f^{(x)}$, and proposed symbol-distribution mapping functions $g^{(x)}$.

process $\mathcal{P}$ generated by M. This gives a constructive way to generate the causal states of a broad class of processes and determine their intrinsic randomness and complexity. As Fig. 3.4 makes clear, $\mathcal{U}$s produce highly structured, fractal-like causal-states sets.

3.4. The Mixed-State Presentation and the ϵ-Machine

The minimality of infinite-state mixed-state presentations $\mathcal{U}(M)$ is an open question at present. Mixed-state presentations are not guaranteed to be minimal. In fact, it is possible to construct MSPs with an uncountably-infinite number of states for a process that requires only one state to optimally predict [**61**].

A proposed solution to this problem is a short and simple check on *mergeablility* of mixed states, which here refers to any two distinct mixed states that have the same conditional probability distribution over future strings; i.e., any two mixed states η_0 and ζ_0 for which:

$$\Pr(X_{0:\ell}|\eta_0) = \Pr(X_{0:\ell}|\zeta_0) \ , \tag{3.10}$$

for all $\ell \in \mathbb{N}^+$.

Although minimality does not impact the entropy-rate calculation, one benefit of the IFS formalization of the MSP is the ability to directly check for duplicated states and therefore determine if the MSP is nonminimal. We check this by considering, for an $N+1$ state machine M with alphabet $\mathcal{A} = \{0, 1, \ldots, k\}$, the dynamic not only over mixed states, but probability distributions

over symbols. Let:

$$P(\eta) = \left(p^{(0)}(\eta), \ldots, p^{(k-1)}(\eta)\right) \tag{3.11}$$

and consider Fig. 3.5. For each mixed state $\eta \in \Delta^N$, Eq. (3.11) gives the corresponding probability distribution $\rho(\eta) \in \Delta^k$ over the symbols $x \in \mathcal{A}$. Let M emit symbol x, then the dynamic from one such probability distribution $\rho \in \Delta^k$ to the next is given by:

$$\begin{aligned} g^{(x)}(\rho_t) &= P \circ f^{(x)} \circ P^{-1}(\rho) \\ &= \rho_{t+1,x} \ . \end{aligned} \tag{3.12}$$

From this, we see that if Eq. (3.12) is invertible, $g^{(x)} : \Delta^k \to \Delta^k$ is well defined and has the same functional properties as $f^{(x)}$. In other words, in this case, it is not possible to have two distinct mixed states $\eta, \zeta \in \Delta^N$ with the same probability distribution over symbols. And, the probability distributions can only converge under the action of $g^{(x)}$ if the mixed states also converge under the action of $f^{(x)}$.

3.5. Hidden Markov Driven Iterated Function Systems

With the properties of place dependent IFS defined by an HMM well-established, let us take a moment to coin some terminology and summarize the mathematical object we have developed.

Definition 8. *An N-dimensional hidden Markov-*driven iterated function system *(DIFS)* $\left(\mathcal{A}, \mathcal{V}, \boldsymbol{\mathcal{R}}, \{T^{(x)}\}, \{p^{(x)}\}, \{f^{(x)}\} : x \in \mathcal{A}\right)$ *consists of:*

(1) a finite alphabet $\mathcal{A}$ of k symbols $x \in \mathcal{A}$*,*

(2) a set $\mathcal{V}$ of N presentation states*,*

(3) a set of states $\boldsymbol{\mathcal{R}} \subset \Delta^{(N-1)}$*, over N-dimensional presentation-state distributions $\eta \in \boldsymbol{\mathcal{R}}$,*

(4) a finite set of N by N symbol-labeled substochastic matrices $T^{(x)}$*, $x \in \mathcal{A}$,*

(5) a set of k symbol-labeled probability functions $p^{(x)} = \langle \eta | T^{(x)} \mathbf{1} \rangle$*, and*

(6) a set of k symbol-labeled mapping functions $f^{(x)} = \langle \eta | T^{(x)} \mathbf{1} \rangle / p^{(x)}(\eta)$.

The set of substochastic matrices must sum to the nonnegative, row-stochastic matrix $T = \sum_{x \in \mathcal{A}} T^{(x)}$—the transition matrix for the presentation-state Markov chain. This ensures that $\sum_{x \in \mathcal{A}} p^{(x)}(\eta) = 1$ for all $\eta \in \Delta^{(N-1)}$.

DIFS states are *predictive* in the sense that they are functions of the prior observables. Consider an infinitely-long past that, in the present, has induced some state η. It is not guaranteed that this infinitely-long past induce a unique state, but it is the case that any state induced by this past must have the same conditional future distribution. Indeed, for task of prediction, knowing the previous state is as good as knowing the infinite past: $\Pr(X_{0:\ell} | \mathcal{R}_0 = \eta) = \Pr(X_{0:\ell} | X_{-\infty:0})$ for all $\ell \in \mathbf{N}^+$.

Therefore, the DIFS is a predictive model of the process $\mathcal{P}$ it generates. (This is in contrast to it being merely a generative model.) Borrowing from the language of automata theory, we refer to the set of states $\boldsymbol{\mathcal{R}}$ plus its transition dynamic—$\Pr(x_t | \eta_t)$ and $\Pr(\eta_{t+1} | \eta_t, x_t)$—as a *state machine* or, simply *machine* that optimally predicts $\mathcal{P}$. When we force each infinitely long past to induce a unique state, we produce a canonical predictive model that is unique: a process' *ϵ-machine* [**30**].

A process' ϵ-machine is its optimally-predictive, minimal model, in the sense that the set $\boldsymbol{\mathcal{R}}$ of predictive states is minimal compared to all its other predictive models. By capturing a process' structure and not merely being predictive, an ϵ-machine's states are called *causal states*. Unless otherwise noted, we assume that all DIFS discussed from this point are ϵ-machines.

CHAPTER 4

Intrinsic Randomness, or the Shannon Entropy Rate

In his original study of the Shannon entropy rate, David Blackwell analyzed the entropy of *functions of finite-state Markov chains* [**54**]. With a shift in notation, functions of Markov chains can be identified as general hidden Markov processes. This is to say, both presentation classes generate the same class of stochastic processes. The Shannon entropy rate problem for HMPs that can be predicted with a finite state unifilar hidden Markov model was solved with Shannon's entropy rate expression Eq. (2.11). However, as Blackwell noted, there is no analogous closed-form expression for the entropy rate of an HMP in general. Blackwell proposed an integral solution over a potentially infinite set of features possessing a fractal measure. We show that this set may be identified as the state set of the mixed state presentation introduced in Chapter 3 and that therefore, this set and its associated measure is the attractor of the associated driven iterated function system. This allows us to solve the Shannon entropy rate problem in for HMPs in general by sampling the attractor of their associated DIFS.

4.1. Blackwell's Solution

Blackwell's main result, retaining his original notation, is transcribed here and adapted by us to constructively solve the HMP entropy-rate problem.

THEOREM 4.1.1. *([**54**, Thm. 1].) Let $\{x_n, -\infty < n < \infty\}$ be a stationary ergodic Markov process with states $i = 1, \ldots, I$ and transition matrix $M = \|m(i,j)\|$. Let Φ be a function defined on $1, \ldots, I$ with values $a = 1, \ldots, A$ and let $y_n = \Phi(x_n)$. The entropy of the $\{y_n\}$ process is given by:*

$$H = -\int \sum_a r_a(w) \log r_a(w) dQ(w) \ , \tag{4.1}$$

where Q is a probability distribution on the Borel sets of the set W of vectors $w = (w_1, \ldots, w_I)$ with $w_i \geq 0$, $\sum_i w_i = 1$, and $r_a(w) = \sum_{i=1}^{I} \sum_{j \ni \Phi(j)=a} w_i m(i,j)$. The distribution Q is concentrated on

the sets $W_1, \ldots, W_A$*, where* W_a *consists of all* $w \in W$ *with* $w_i = 0$ *for* $\Phi(i) \neq a$ *and satisfies:*

$$Q(E) = \sum_a \int_{f_a^{-1}E} r_a(w) dQ(w) \ , \tag{4.2}$$

where f_a *maps* W *into* W_a*, with the* j*th coordinate of* $f_a(w)$ *given by* $\sum_i w_i m(i,j)/r_a(w)$ *for* $\Phi(j) = a$*.*

We can identify the w vectors in Theorem 4.1.1 as exactly the mixed states of Section 3.2. Furthermore, it is clear by inspection that $r_a(w)$ and $f_a(w)$ are the probability and mapping functions of Eqs. (3.5) and (3.6), respectively, with a playing the role of our observed symbol x.

Therefore, Blackwell's expression Eq. (4.1) for the HMP entropy rate, in effect, replaces the average over a finite set $\boldsymbol{\mathcal{S}}$ of unifilar states in Shannon's entropy rate formula Eq. (2.11) with (i) the mixed states $\boldsymbol{\mathcal{R}}$ and (ii) an integral over the Blackwell measure μ. In our notation, we write Blackwell's entropy formula as:

$$h_\mu^B = -\int_{\boldsymbol{\mathcal{R}}} d\mu(\eta) \sum_{x \in \mathcal{A}} p^{(x)}(\eta) \log_2 p^{(x)}(\eta) \ . \tag{4.3}$$

Thus, as with Shannon's original expression, this too uses unifilar states—now, though, states from the mixed-state presentation $\mathcal{U}$. This, in turn, maintains the finite-to-one internal (mixed-) state sequence to observed-sequence mapping. Therefore, one can identify the mixed-state entropy rate itself as the process' entropy rate.

4.2. An Updated Solution

As noted in Section 3.1, the contractivity of our substochastic transition matrix mappings guarantees ergodicity over the words generated by the mixed-state presentation [**57**]. With this established in Section 3.3, we can replace Eq. (4.3)'s integral over $\boldsymbol{\mathcal{R}}$ with a time average over a mixed-state trajectory $\eta_0, \eta_1, \ldots$ determined by a long allowed word, using Eqs. (3.5) and (3.6). This gives a new limit expression for the HMP entropy rate:

$$\widehat{h_\mu}^B = -\lim_{\ell \to \infty} \frac{1}{\ell} \sum_{t=0}^{\ell} \sum_{x \in \mathcal{A}} \Pr(x|\eta_\ell) \log_2 \Pr(x|\eta_\ell) \ , \tag{4.4}$$

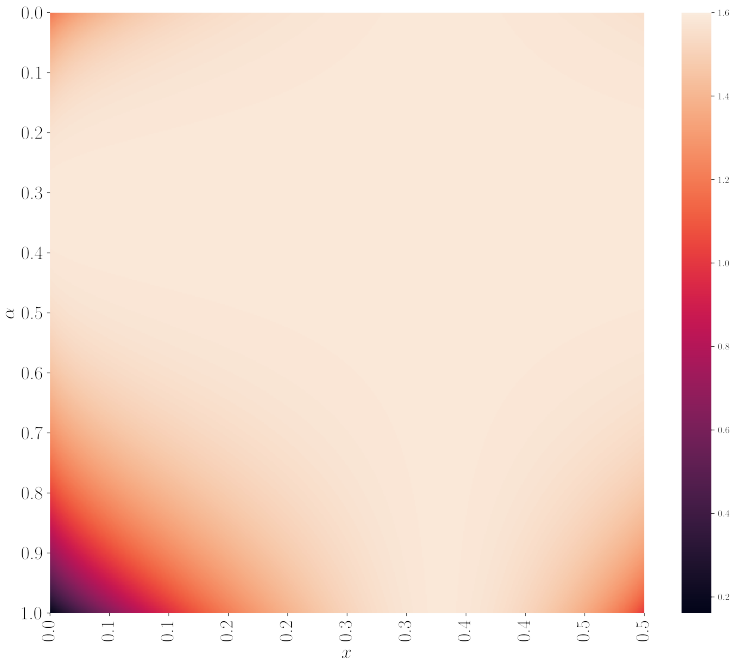


FIGURE 4.1. Entropy rate for 250,000 parametrized HMPs generated according to the definition in Eq. (3.7), over $x \in [0.0, 0.5]$ and $\alpha \in [0.0, 1.0]$. Examples of the MSPs produced are plotted in Fig. 3.4. The entropy rate of each, as estimated by Eq. (4.4), is plotted, showing the gradual change in entropy rate across the parameter space.

where $\eta_\ell = \eta(w_{0:\ell})$ and $w_{0:\ell}$ is the first ℓ symbols of an arbitrarily long sequence $w_{0:\infty}$ generated by the process. Note that $w_{0:\ell}$ will be a typical trajectory, if ℓ is sufficiently long. To remove convergence-slowing contributions from transient mixed states, one can ignore some number of the initial mixed states. The exact number of transient states that should be ignored is unknown in general and discussed in Section 4.3. We can say that it depends on the initial mixed state η_0, which is generally taken to be $\langle \delta_\pi |$, and the diameter of the attractor.

Figure 4.1 plots the entropy rate for 250,000 HMPs. The HMP definition is given in Eq. (3.7) and parametrized by two variables, x and α. Three examples of the MSPs from this parameter space displayed in Fig. 3.4. Despite the visual distinction of the MSPs, the entropy rate smoothly varies across parameters and is often close to $\log_2(3)$. This is partially due to the radial symmetry of the mixed-state attractors. The connection between the MSP structure and information measures will be more fully addressed in Chapter 5.

4.3. Data Requirements

Although we developed our HMP entropy-rate expression in terms of IFSs, determining a process' entropy rate can be recast as Markov chain Monte Carlo (MCMC) estimation. In MCMC, the mean of a function $f(x)$ of interest over a desired probability distribution $\pi(x)$ is estimated by designing a Markov chain with a stationary distribution π. For HMPs the desired distribution is the Blackwell measure μ, which is the stationary distribution μ over the MSP states $\boldsymbol{\mathcal{R}}$. Then, the Markov chain is simply the transition dynamic $\mathcal{W}$ over $\boldsymbol{\mathcal{R}}$.

With this setting, we estimate the entropy rate $\widehat{h_\mu}^B$ as the mean of the stochastic process defined by taking the entropy $H[X_\eta]$ over symbols emitted from state η for a sequence of mixed states generated by $\mathcal{W}$. In effect, we estimate the entropy rate as the mean of this stochastic process:

$$\begin{aligned} \widehat{h_\mu}^B &= \mu_H \\ &= \langle H[X_\eta] \rangle_\mu \ . \end{aligned} \tag{4.5}$$

Mathematically, little has changed. The advantage, though, of this alternative description is that it invokes the extensive body of results on MCMC estimation. In this, it is well known that there are two fundamental sources of error in the estimation. First, there is that due to *initialization bias* or undesired statistical trends introduced by the initial transient data produced by the Markov chain before it reaches the desired stationary distribution. Second, there are errors induced by *autocorrelation in equilibrium*. That is, the samples produced by the Markov chain are correlated. And, the consequence is that statistical error cannot be estimated by $1/\sqrt{N}$, as done for N independent samples.

To address these two sources of error, we follow common MCMC practice, considering two "time scales" that arise during estimation. Consider the autocorrelation of the stationary stochastic process:

$$C_f(t) = \langle f_s f_{s+t} \rangle - \mu_f^2 \ ,$$

where μ_f is f's mean. Also, consider the *normalized* autocorrelation, defined:

$$\rho_f(t) = \frac{C_f(t)}{C_f(0)} \ .$$

If the autocorrelation decays exponentially with time, we define the *exponential autocorrelation time*:

$$\tau_{exp,f} = \lim_{t\to\infty} \sup \frac{1}{-\log|\rho_f(t)|}$$

and

$$\tau_{exp} = \sup_f \tau_{exp,f} \ .$$

So, τ_{exp} upper bounds the rate of convergence from an initial nonequilibrium distribution to the equilibrium distribution.

For a given observable, we also define the *integrated autocorrelation time* $\tau_{init,f}$ as:

$$\tau_{int,f} = \frac{1}{2}\sum_{-\infty}^{\infty} \rho_f(t) \ . \tag{4.6}$$

This relates the correlated samples selected by the chain to the variance of independent samples for the particular function f of interest. The variance of $f(x)$'s sample mean in MCMC is higher by a factor of $2\tau_{int,f}$. In other words, the errors for a sample of length ℓ are of order $\sqrt{\tau_{int,f}/N}$. Thus, targeting 1% accuracy requires $\approx 10^4\tau_{int,f}$ samples.

In practice, it is difficult to find τ_{exp} and τ_{int} for a generic Markov chain. There are two options. The first is to use numerical approximations that estimate the autocorrelation function, and therefore τ, from data. If we have the nonunifilar model in hand, it is a simple matter of sweeping through increasingly long strings of generated data until we observe convergence of the autocorrelation function.

4.4. Estimation Errors for Finite-State Autocorrelation

Alternatively, taking inspiration from previous treatments of nonunifilar models, we make a finite-state approximation to the MSP by coarse-graining the simplex into boxes of length ϵ and

employ a suitable method, such as Ulam's, to approximate the transition operator. Using methods previously discussed in Ref. [45], this allows calculating the autocorrelation function directly.

Coarse-graining the mixed-state simplex into a set $\mathcal{C}$ of boxes of width ϵ, we may construct a finite-state approximation of the infinite-state MSP. It has been shown that given such an approximation, for any given box c, the bound on the difference in the entropy rate over the symbol distribution between the coarse-grained approximation and a mixed state within that box is bounded by [60]:

$$\left|H[X_0|\mathcal{C}=c] - H[X_0|\eta \in f]\right| \leq H_b\left(\frac{\sqrt{G}\epsilon}{2}\right), \tag{4.7}$$

where $H_b(\cdot)$ is the binary entropy function. Our task here is to consider the error in the autocorrelation in the sequence of mixed states since, if we can show that this is bounded, the error in the autocorrelation of the branching entropy must also be bounded.

At time zero, the autocorrelation is equal to $A(\ell = 0) = \langle X_0 \overline{X_0} \rangle$, so for the finite-state approximation, we have:

$$A_{\mathcal{C}}(\ell = 0) = \sum_i \pi_{\mathcal{C}}(i) c_i \overline{c_i} \ ,$$

where $\pi_{\mathcal{C}}$ is the stationary distribution over the coarse-grained mixed states, $\pi_{\mathcal{C}}(i)$ is the stationary probability of cell i, and c_i is the center of cell i. For the true process, we have:

$$\begin{aligned} A(\ell = 0) &= \int_{\boldsymbol{\mathcal{R}}} d\mu(\eta) \eta \overline{\eta} \\ &= \sum_i \pi_{\mathcal{C}}(i) \int_{\eta \in \mathcal{C}_i} d\mu(\eta|i) \eta \overline{\eta} \ , \end{aligned}$$

where $d\mu(\eta|i)$ is the distribution over mixed states within cell i. The maximum distance between any two mixed states in a cell i is bounded by:

$$\|\eta - \zeta\|_1 \leq \sqrt{G}\epsilon \ ,$$

the length of the longest diagonal in a hypercube of dimension $|G|$, by construction. Since the gradient of the L_2 norm is simply $\nabla \|\mathbf{x}\|_2 = \mathbf{x}/\|\mathbf{x}\|_2$, we have a bound on the difference in the

autocorrelation at time zero:

$$|A_{\mathcal{C}}(\ell = 0) - A(\ell = 0)| \leq N\sqrt{|G|}\epsilon \ .$$

With increasing length we have:

$$A_{\mathcal{C}}(\ell) = \sum_i \pi_{\mathcal{C}}(i) c_i \sum_{w \in \mathcal{A}^\ell} \overline{f^{(w)}(c_i)} p^{(w)}(c_i)$$

and:

$$\begin{aligned} A(\ell) &= \int_{\boldsymbol{\mathcal{R}}} d\mu(\eta)\eta \sum_{w \in \mathcal{A}^\ell} \overline{f^{(w)}(\eta)} p^{(w)}(\eta) \\ &= \sum_i \pi_{\mathcal{C}}(i) \int_{\eta \in \mathcal{C}_i} d\mu(\eta|i)\eta \sum_{w \in \mathcal{A}^\ell} \overline{f^{(w)}(\eta)} p^{(w)}(\eta) \ . \end{aligned}$$

Let $\eta = c_i + \delta$ for some mixed state in cell i. Then we can write:

$$|A_{\mathcal{C}}(\ell) - A(\ell)| \leq \sum_i \pi_{\mathcal{C}}(i) \left[c_i \sum_{w \in \mathcal{A}^\ell} \overline{f^{(w)}(c_i)} p^{(w)}(c_i) - (c_i + \delta) \sum_{w \in \mathcal{A}^\ell} \overline{f^{(w)}(c_i + \delta)} p^{(w)}(c_i + \delta) \right] \ .$$

Now, note that:

$$p^{(w)}(c_i + \delta) \approx p^{(w)}(c_i) + \nabla p^{(w)}(c_i) \cdot \delta$$

and:

$$f^{(w)}(c_i + \delta) \approx f^{(w)}(c_i) + e^{\lambda^x} \delta$$

where λ^w is the leading Lyapunov exponent of the mapping function. Substituting this and eliminating terms of order δ^2 gives us:

$$|A_{\mathcal{C}}(\ell) - A(\ell)| \leq \sum_i \pi_{\mathcal{C}}(i) \left[c_i \sum_{w \in \mathcal{A}^\ell} \left(\overline{f^{(w)}(c_i)} \nabla p^{(w)} \cdot \delta + e^{\lambda^w} \overline{\delta} p^{(w)}(c_i) \right) + \delta \sum_{w \in \mathcal{A}^\ell} \overline{f^{(w)}(c_i)} p^{(w)}(c_i) \right] \ .$$

These terms identify three sources of approximation error: (i) that due to a difference in the probability distribution over symbols, (ii) that in the mapping functions, and (iii) that from approximating the points at the center of their cells.

For the first, we note that total variation in the probability distribution over symbols is bounded by the distance between the mixed states at which the distributions are computed. So, for any two mixed states in the same cell, $\| \Pr(X = x|\eta) - \Pr(X = x|\zeta)\|_{TV} \leq \sqrt{G}\epsilon$. Then, the first term is the error due to the difference in the expectation value of the next state, given that we have calculated the probability distribution at $c_i + \delta$, rather than c_i. Using Hölder's inequality, for two distributions over $P(X)$ and $Q(X)$, we may say:

$$\begin{aligned} E[f]_Q - E[f]_P &= \sum_x f(x)(P(x) - Q(x)) \\ \sum_x f(x)(P(x) - Q(x)) &\leq \sum_x f(x)|P(x) - Q(x)| \\ \sum_x f(x)|P(x) - Q(x)| &\leq \|f\|_p \|P - Q\|_q \ , \end{aligned}$$

where $1/p + 1/q = 1$. Setting $q = 1$:

$$E[f]_Q - E[f]_P \leq \|f\|_\infty \|P - Q\|_{TV} \ .$$

So, after taking the product with the cell centers c_i, we have that the first error is bounded by $N\sqrt{G}\epsilon$ at all lengths.

For the second, we note that since the maps are contractions, $\lambda < 0$, and the distance between $f^x(\eta)$ and $f^x(\zeta)$, where η and ζ are in the same cell i, is bounded by $\sqrt{G}\epsilon$. As the length of a word w grows, $\lambda^w \to -\infty$ and the distance $f^w(\eta) - f^w(\zeta) \to 0$. At large ℓ, this term vanishes, at a rate equal to the average maximal Lyapunov exponent of the IFS.

The final error is that in the autocorrelation in the cell approximation which is, likewise, bounded by the cell size—this is the same error from $A(0)$, viz. $N\sqrt{G}\epsilon$.

And so, in combination with the bound on the entropy, we may say, loosely speaking, that the error in the autocorrelation vanishes as $\epsilon \to 0$. Therefore, to find τ and estimate the error in Eq. (4.5) as a function of sample size, we take finer coarse-grained approximations until convergence in the autocorrelation curve is observed, and then calculate τ directly.

4.5. Computational Advantages and Disadvantages

This completes our development of the procedure to determine the HMP entropy rate. We now consider the computational advantages and disadvantages of using the MSP and Eq. (4.4) to find the entropy rate, in comparison to existing methods, as well as the practical issues of the resources needed for accurate estimation.

One impetus driving our development is a recurring need for an entropy estimation method that is both fast and general—one that applies to extensive surveys of HMPs that may have varying structural elements and transition probabilities. This challenge is broadly encountered. In point of fact, these structural tools and the entropy-rate method introduced here have already been put to practical use in two prior works. One diagnosed the origin of randomness and structural complexity in quantum measurement [**52**]. The other exactly determined the thermodynamic functioning of Maxwellian information engines [**53**], when there had been no previous method for this. Both applications relied critically on analyzing parametrized HMPs and required reliable and fast calculation of entropy rates and other information measures across a variety of HMP topologies.

HMP Shannon entropy rate has been studied in terms of upper and lower bounds [**55**, **62**], with exact expressions [**1**, **63**, **64**], and as the solution to an integral equation [**54**]. Naturally, exact expressions are preferable where applicable and needed. Unfortunately, though, they are available only for restricted subsets of HMP topologies, such as unifilar or countable HMPs.

A well-known result is that the upper and lower finite-length conditional entropy estimates [**55**]:

$$H(X_0|X_1, \ldots, X_\ell, \sigma_{\ell+1}) \leq h_\mu \leq H(X_0|X_1, \ldots, X_\ell)$$

converge exponentially in word length ℓ to the entropy rate for *path mergeable* HMPs, a property that is straightforwardly checked [**62**]. This being said, while testing for the path mergeability condition is feasible, the algorithm to do so runs in polynomial time in the number of states and symbols, making testing impractical for a large-scale survey of HMPs with many states and/or symbols. Furthermore, while the conditional entropy estimates of h_μ converge in exponentially in ℓ, calculating conditional entropies is nontrivial. This is particularly so when the exponential convergence rate α is arbitrarily close to 0. Thus, for accuracy within a desired bound the required

ℓ may become arbitrarily large. Finally, while $H(X_\ell|X_1,\ldots,X_{\ell-1})$ may be calculated exactly for all ℓ, given knowledge of the HMP, at large ℓ this requires calculation of the full distribution over $|\mathcal{A}|^\ell$ ℓ-length words. Needless to say, for applications involving many states, large alphabets, and/or many HMPs, bounding the entropy rate in this way becomes computationally impractical.

The new method largely obviates these problems. When mixed-state construction returns a countable-state HMP, we directly apply Shannon's entropy rate formula Eq. (2.11) and find the entropy rate exactly. When the MSP is uncountable we apply Eq. (4.4). It runs in $\mathcal{O}(N)$ where N is the number of mixed states generated, with no direct dependence on number of HMP states or alphabet size. Section 4.3 and Section 4.4 give a full discussion of the data-length requirements and error of the mixed-state method, respectively.

The net result is that, being cognizant of the data requirements, entropy rate estimation is well behaved, convergent, and accurate. One concludes that the most effective manner of calculating of entropy rate for large-scale surveys will likely employ a combination of our methodology and the other techniques just mentioned, with deployment of exact expressions where possible. Furthermore, we note that the development of the MSP as the ϵ-machine, and the constructive method of producing the causal state set $\boldsymbol{\mathcal{R}}$, is of interest beyond computational advantages in characterizing the complexity of the underlying system beyond the entropy rate.

CHAPTER 5

Intrinsic Structure, or the Statistical Complexity Dimension

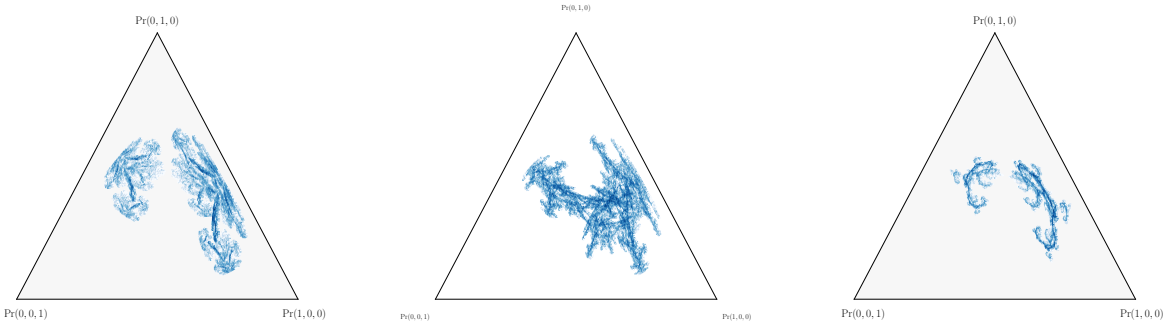
We have now thoroughly explored how to use the mixed-state presentation to determine a process' intrinsic randomness. Now we must complement the measure of randomness with a measure of structure. The naive approach of measuring the structure with statistical complexity C_μ introduces a problem: the statistical complexity diverges for an HMC with an uncountably-infinite state set $\boldsymbol{\mathcal{R}}$. In general, MSPs of HMCs are uncountably-infinite state, precluding distinguishing them via C_μ. This being said, it is clear visually from figures like Fig. 3.4 and Fig. 5.1 that HMCs with uncountably-infinite state spaces still have distinct structures. We wish to find a way to measure and distinguish such structure. For this, we take inspiration from Shannon's dimension rate [1] and call on a familiar tool.

Fractal dimension measures the rate at which a chosen size metric of a set diverges with the scale at which the set is observed [6, 65, 66, 67, 68]. Fractal dimension is also useful to probe the "size" of objects when cardinality is not informative. For example, the mixed-state presentation, generically, has an uncountable infinity of causal states. That observation is far too coarse, though, to distinguish the clearly distinct mixed-state sets $\boldsymbol{\mathcal{R}}$ in Fig. 5.1. Each is uncountably infinite, but the $\boldsymbol{\mathcal{R}}$'s geometries differ. Determining their fractal and other dimensions will allow us to distinguish them and allow us to introduce additional insights into the original process' intrinsic information processing.

5.1. Defining the Statistical Complexity Dimension

Consider the mixed-state set $\boldsymbol{\mathcal{R}}$ on the simplex for an N-state HMC M that generates a process $\mathcal{P}$. We consider two types of dimension for $\boldsymbol{\mathcal{R}}$: the Minkowski-Bouligand or box-counting dimension, often simply called the fractal dimension, and the information dimension.

To calculate the first, coarse-grain the N-simplex with evenly spaced subsimplex cells of side length ϵ. Let $\mathcal{F}(\epsilon)$ be the set of cells that encompass at least one mixed state. Then $\boldsymbol{\mathcal{R}}$'s *box-counting*



(A) Mixed state attractor for 3-state "alpha" machine.

(B) Mixed state attractor for 3-state machine from Eq. (3.7), with $\alpha = 0.6$ and $x = 0.1$.

(C) Mixed state attractor for 3-state "beta" machine.

FIGURE 5.1. Simple HMCs generate MSPs with a wide variety of structures, many fractal in nature. Each subplot displays 10^4 mixed states of a different, highly-nonunifilar 3-state hidden Markov chain. The HMCs themselves are specified in Appendix A.

dimension is:

$$d_0(\boldsymbol{\mathcal{R}}) = -\lim_{\epsilon \to 0} \frac{\log |\mathcal{F}(\epsilon)|}{\log \epsilon} \ , \tag{5.1}$$

where $|C|$ is the size of set C.

The information dimension tracks how the Blackwell measure $\mu_B(\boldsymbol{\mathcal{R}})$ scales with ϵ. Let each cell in $\mathcal{F}(\epsilon)$ be a state and approximate the dynamic over $\mathcal{U}(M)$ by grouping all transitions to and from states encompassed by the same cell. This results in a finite-state Markov chain that generates an approximation of the original mixed-state process and has a stationary distribution $\mu(\mathcal{F}(\epsilon))$. Then $\mu_B(\boldsymbol{\mathcal{R}})$'s *information dimension* is:

$$d_1(\mu_B(\boldsymbol{\mathcal{R}})) = \lim_{\epsilon \to 0} \frac{H_\mu[\mathcal{F}(\epsilon)]}{\log \epsilon} \ , \tag{5.2}$$

where $H_\mu[\mathcal{F}(\epsilon)] = -\sum_{C_i \in \mathcal{F}(\epsilon)} \mu(C_i) \log \mu(C_i)$ is the Shannon entropy over the set $\mathcal{F}(\epsilon)$ of cells that cover attractor $\boldsymbol{\mathcal{R}}$ with respect to μ.

These dimensions give two complementary resource-scaling laws for HMC-generated processes. Rearranging Eq. (5.1), we see that the number of mixed states in our finite-state approximation to

$\mathcal{U}(M)$ scales algebraically with $\boldsymbol{\mathcal{R}}$'s box-counting dimension:

$$|\mathcal{F}(\epsilon)| \sim \epsilon^{-d_0(\boldsymbol{\mathcal{R}})} \ . \tag{5.3}$$

In other words, for an uncountably infinite MSP, the exponential growth rate of mixed states is $d_0(\boldsymbol{\mathcal{R}})$.

Similarly, the entropy of the Blackwell measure scales with the information dimension. Rearranging Eq. (5.2) shows that the state entropy of the finite-state approximation to $\mathcal{U}(M)$ scales logarithmically with $\boldsymbol{\mathcal{R}}$'s information dimension with respect to the Blackwell measure:

$$H_\mu[\mathcal{F}] \sim d_1(\mu_B) \cdot \log \epsilon \ . \tag{5.4}$$

As $\epsilon \to 0$, $|\mathcal{F}|$ and $H_\mu(\mathcal{F})$ diverge and d_0 and d_1 are the divergence rates, respectively. The remainder focuses on d_1 as applied to the ϵ-machine, for which it describes the rate of divergence of statistical complexity C_μ.

We refer to the information dimension $d_1(\mu)$ of the ϵ-machine the *statistical complexity dimension* d_μ. Applying d_1 to the Blackwell measure $\mu_B(\boldsymbol{\mathcal{R}})$ gives the rate of divergence of C_μ as one constructs increasingly better finite-state approximations to the infinite-state ϵ-machine. In this way, d_μ describes the divergence of memory resources when attempting to optimally predict a process that requires an uncountably-infinite number of predictive features. This is a unique, minimal description of the process' structural complexity. This solves the challenge posed in the introduction: quantifying structure for a broad class of truly complex systems.

When a process may be optimally predicted with a finite number of predictive features, the statistical complexity dimension vanishes. In this case, the more relevant complexity measure is the original ϵ-machine statistical complexity C_μ, which is is finite. When a process requires uncountably-infinite causal states, but may be generated with a minimal N-state (nonunifilar) HMC, the statistical complexity is less than or equal to $N - 1$. This is the associated IFS's *embedding dimension*, since the mixed states lie in a space of dimension $N - 1$.

Unfortunately, directly calculating the information dimension using Eq. (5.2)—and therefore calculating the statistical complexity dimension d_μ—is nontrivial, as it requires estimating a fractal

measure. Fortunately, to calculate the d_μ of the mixed state attractor $\boldsymbol{\mathcal{R}}$, we can leverage the associated generating dynamical system (see Section 3.3).

5.2. Dimension from Dynamical (In)Stabilities

We can link the information dimension of an MSP's mixed state set $\boldsymbol{\mathcal{R}}$ to the stability properties of the associated IFS. This starts with determining the local time-average stability and instability of orbits within an attractor via the *spectrum of Lyapunov characteristic exponents* $\Gamma = \{\lambda_1, \ldots, \lambda_N : \lambda_i \geq \lambda_{i+1}\}$ [69, 70]. Individual LCEs λ_i measure the average local growth or decay rate of orbit perturbations. The net result is a list of quantities that indicate long-term orbit instability ($\lambda_i > 0$) and orbit stability ($\lambda_i < 0$) in complementary directions. Usefully, their sum gives the net state-space divergence—volume loss for dissipative systems. The sum of the positive LCEs is the dynamical system's entropy rate [6, 69, 71, 72, 73]—the net information generation.

To motivate our present use of the Lyapunov spectrum, it will help to develop a simple intuition for how the LCEs relate to dimension. Both of our previously defined dimensional quantities—Eq. (5.2) and Eq. (5.1)—depended on the growth rate of cells needed to cover the attractor as we take the side length of the cells to zero. First, imagine the attractor of a two-dimensional map with LCEs $\lambda_1 > 0 > \lambda_2$ and whose state space is covered with equally spaced squares of side length ϵ. After iterating the map q times, for ϵ small enough, the local action of the map is approximately linear. From the LCE definition, this means it takes the initial square cells to rectangles of average length $(e^{\lambda_1 q})\epsilon$ and average width $(e^{\lambda_2 q})\epsilon$. Now, consider covering the attractor with a new set of squares of side length $(e^{\lambda_2 q})\epsilon$. We can see by inspection that this requires roughly $e^{(\lambda_1/\lambda_2)q}$ squares per rectangle. In this way, the Lyapunov exponents relate to the scalings measured by dimension quantities [74].

Indeed, this relationship has resulted in the definition of a *Lyapunov dimension* d_Γ in terms of the spectrum Γ [75]:

$$d_\Gamma = \begin{cases} k + \dfrac{\Lambda(k)}{|\lambda_{k+1}|} \ , & \Lambda(N) < 0 \\ N \ , & \Lambda(N) \geq 0 \end{cases} , \tag{5.5}$$

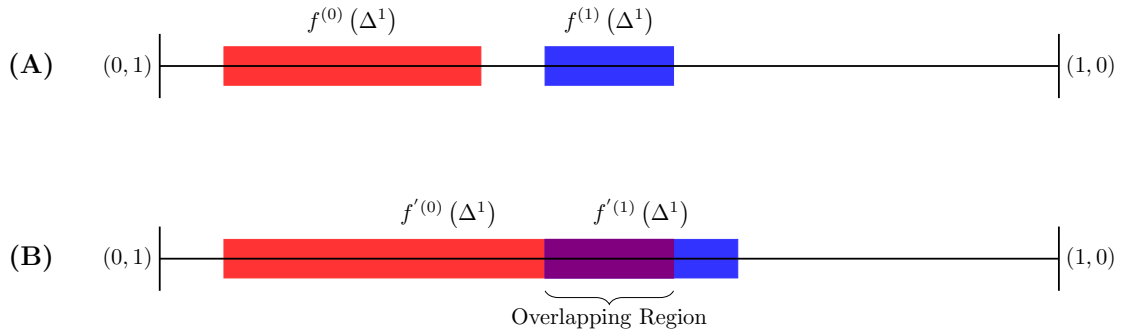


FIGURE 5.2. Overlap problem on the 1-simplex Δ^1: Two distinct IFSs are considered, each with two mapping functions. The images of the mapping functions over the entire simplex are depicted in red and blue. (A) Images of the mapping functions $f^{(0)}$ and $f^{(1)}$ do not overlap—every mixed state $\eta_t \in \boldsymbol{\mathcal{R}}$ has a unique pre-image. (B) Images of the mapping functions overlap (purple Overlapping Region)—there exist $\eta_1, \eta_2 \in \boldsymbol{\mathcal{R}}$ such that $f'^{(0)}(\eta_1) = f'^{(1)}(\eta_2) = \eta_3$. This case is an *overlapping IFS*.

where we introduce the Lyapunov spectrum partial sum $\Lambda(m) = \sum_{i=1}^{m} \lambda_i$ and $k = 0, 1, 2, \ldots, N$ is the largest index for which $\Lambda(k) > 0$. If $\lambda_1 < 0$, we take $\Lambda(0) = 0$ and $d_\Gamma = 0$. Its name helps distinguish the conditions under which the relationships between the various dimensions actually hold. (There are many conditions and system classes that this summary necessarily leaves out.)

It has been shown that for any ergodic, invariant probability measure μ:

$$d_1(\mu) \leq d_\Gamma \ ,$$

with equality when μ is a Sinai-Bowen-Ruelle (SBR) measure [7]. Furthermore, it was conjectured that for "typical dynamical systems", the Lyapunov dimension d_Γ equals the information dimension d_1 [75]. This remarkable relationship directly relates a system's dynamics to the geometry and natural measure of its attractor. Additionally, and usefully, Eq. (5.5) gives us a tractable method to find the information dimension of an attractor generated by a dynamical system when the equality holds and an upper bound when it does not.

5.3. Calculating Statistical Complexity Dimension

We now have in hand two important pieces. First, the definition of statistical complexity dimension d_μ as the information dimension of an ϵ-machine. Second, we have a bound on the

information dimension of an attractor, given knowledge of the generating system's dynamics. To complete our picture, we now address the final puzzle piece.

As Section 5.2 discussed, there is a direct relationship between the information dimension of a chaotic attractor and the dynamics of the system to which the attractor belongs. Furthermore, as discussed in Section 3.3, every HMC has an associated random dynamical system—the iterated function system (IFS)—which has the HMC's set of mixed states $\boldsymbol{\mathcal{R}}$ as its unique attractor. Combining these two facts allows us to exactly calculate d_μ in many cases of interest.

An advantage of working with IFSs defined by HMCs is a clean division exhibited by their Lyapunov spectrum. It has been shown that the entropy rate of the generated process is equivalent to the largest Lyapunov exponent [**76**, **77**]; conceptually, this appends h_μ to the Lyapunov spectrum as λ_0. Calculating h_μ was the main topic of the prequel to the present work [**48**]. Furthermore, due to the contractivity of the IFS mapping functions, all other Lyapunov exponents $(\lambda_i, i = 1, 2, \ldots, N)$ will necessarily be negative. For a review of calculating the Lyapunov exponents for IFSs, see Appendix F.

Therefore, the Lyapunov dimension of an IFS is:

$$\widetilde{d_\Gamma} = \begin{cases} k + \dfrac{\Lambda(k) + h_\mu}{|\lambda_{k+1}|}, & -\Lambda(N) > h_\mu \\ N \ , & -\Lambda(N) \leq h_\mu \end{cases} , \tag{5.6}$$

where $k = 0, 1, 2, \ldots, N$ is now the largest index for which $-\Lambda(k) < h_\mu$. In contrast to Eq. (5.5), in this case $\Lambda(m) < 0$ for all $m = 1, 2, \ldots, N$. We still take $\Lambda(0) = 0$, but the dimension does not necessarily go to zero: $\widetilde{d_\Gamma} > 0$ when $h_\mu > 0$. Readers familiar with the Lyapunov dimension should take care as this expression effectively re-indexes the traditional presentation of d_Γ as given in Eq. (5.5).

Under specific technical conditions to be discussed shortly, the IFS d_Γ is exactly equal to the information dimension of the IFS's attractor and, therefore, is d_μ [**78**]. In general, relaxing those conditions, $\widetilde{d_\Gamma}$ upper bounds the statistical complexity dimension:

$$\widetilde{d_\Gamma} \geq d_\mu \ . \tag{5.7}$$

Assembling these pieces together determines the basic algorithm to calculate (or bound) the statistical complexity dimension:

(1) For an N-state HMC M with $|\mathcal{A}| = k$, write down the associated IFS with k symbol-labeled mapping functions $f^{(x)}$ and probability functions $p^{(x)}$.
(2) Calculate the entropy rate h_μ using the Blackwell limit (see [**48**]).
(3) Calculate the negative Lyapunov exponents $\{\lambda_1, \ldots, \lambda_{N-1}\}$ (see Appendix F).
(4) Compute the Lyapunov dimension d_Γ using Eq. (5.6).

As mentioned, in specific cases, the Lyapunov dimension is exactly equal to the statistical complexity dimension, and our task is complete. However, there are major technical concerns with when we have only the bound in Eq. (5.7) and with its tightness then.

5.4. The Overlap Problem

A subtle disadvantage of working with IFSs is a direct result of the stochastic nature of them as random dynamical systems. We must consider the *overlap problem*, which concerns the ranges of the symbol-labeled mapping functions $f^{(x)}$, illustrated in Fig. 5.2. Specifically, the problem means that we must distinguish between IFSs that meet the open set condition [**79**, **80**] and those that do not.

Definition 9. *An iterated function system with mapping functions $f^{(x)} : \Delta \to \Delta$ satisfies the* open set condition *(OSC) if there exists an nonempty open set $U \in \Delta$ such that for all $x, x' \in \mathcal{A}$:*

$$f^{(x)}(U) \cap f^{(x')}(U) = \emptyset, \quad x \neq x' \ .$$

IFSs that meet the OSC are nonoverlapping IFSs.

When the images of the symbol-labeled mappings overlap the inequality in Eq. (5.7) is strict. To briefly outline the consequences, for an overlapping IFS the entropy rate h_μ does not accurately capture state-space expansion. And, this causes the IFS d_Γ (Eq. (5.6)) to overestimate the information dimension. As a rule of thumb, the degree to which the mappings overlap determines the magnitude of the bound's error. The impact of overlaps is significant. It is explored both in Section 5.6, where

we calculate the statistical complexity dimension for HMCs with and without overlap, as well as in the sequel, which diagnoses the problem's origins and provides an algorithmic solution.

For now, to give a workable approach, we simply introduce two extra steps to the d_μ algorithm from the previous section:

(5) Determine if the Open Set Condition is met using the mapping functions $f^{(x)}$.
(6) If the OSC is not met, estimate the degree of overlap to determine the closeness of the bound on d_μ.

5.5. Two-State Hidden Markov Models

Finally, we analyze two-state HMCs, for which Eq. (5.6) simplifies significantly. When there is no overlap in the maps, we have exact results for d_μ.

For two-state nonunifilar HMCs, the mixed-state set lives on the 1-simplex Δ^1—the unit interval from $\eta = (0, 1)$ to $\eta = (1, 0)$. Mixed states $\eta \in \boldsymbol{\mathcal{R}}$ and the dynamic on them exist in a one-dimensional space and, thus, there is a single negative Lyapunov exponent $\lambda_1 < 0$.

In this case, the calculation of the negative Lyapunov exponent is particularly direct, since the maps are all one-dimensional. The negative Lyapunov exponent for a one-dimensional map $\eta_{n+1} = f(\eta_n)$ is:

$$\lambda(\eta_0) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N-1} \log \left| \frac{df(\eta_i)}{d\eta} \right|$$

for an orbit starting at η_0. For an IFS with a set of of mapping functions $\{f^{(x)}\}$, we find λ as the weighted average of the Lyapunov exponents of each map:

$$\lambda_\mu = \int \sum_x p^{(x)}(\eta) \log \left| \frac{df^{(x)}(\eta)}{d\eta} \right| d\mu \ ,$$

where μ is the IFS's Blackwell measure. We can apply ergodicity to transform this into a summation over time for ease of calculation.

If a two-state HMC has an MSP with an uncountable infinity of mixed states and its corresponding IFS satisfies the OSC, there is a simple relationship between the entropy rate, the Lyapunov

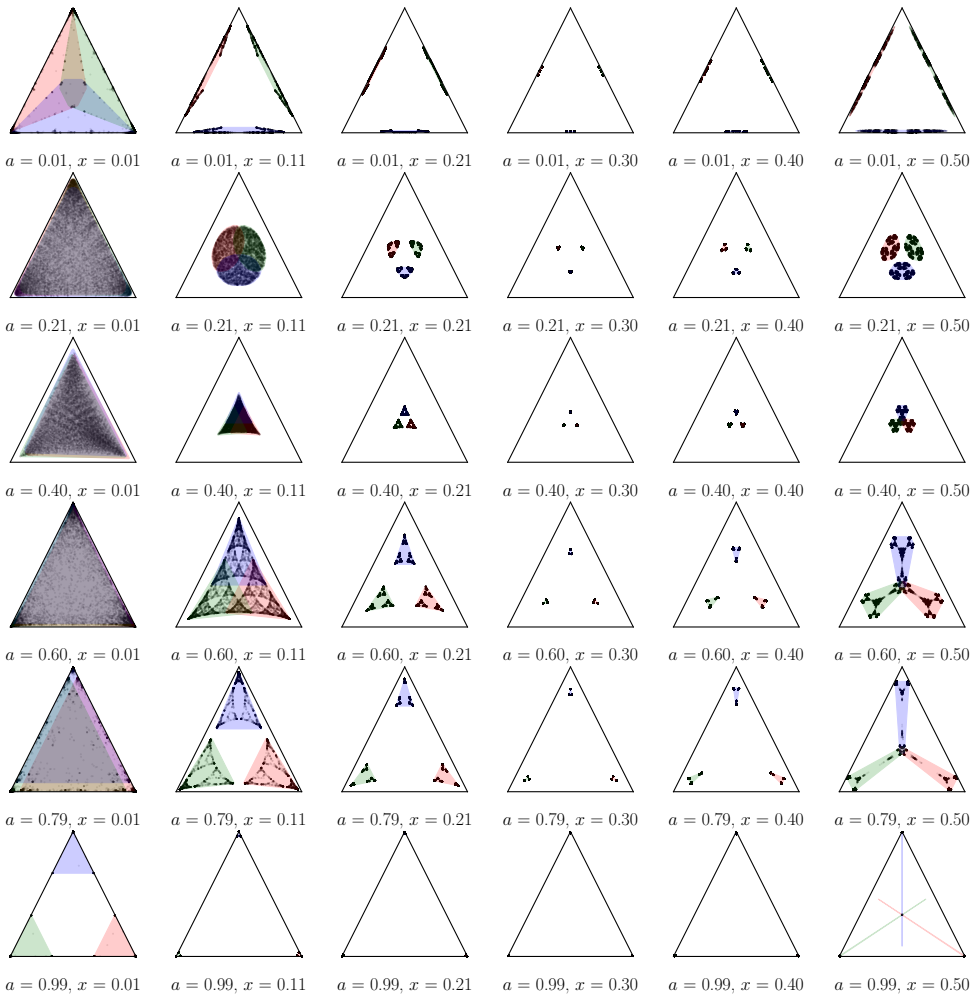


FIGURE 5.3. Mixed-state attractors generated by a 3-state HMC parametrized over $\alpha \in [0, 1]$ and $x \in [0, 0.5]$. The HMC itself is given in Eq. (3.7). 10^5 mixed states are plotted for each attractor, with the initial 5×10^4 states thrown away as transients. The ranges of the symbol-labeled maps are color shaded, revealing regions of their image overlap on the attractor. Comparing to Eq. (3.7), the red, blue, and green regions represent the images of the mapping functions defined by $T^{\square}$, $T^{\triangle}$, and T°, respectively.

exponent, and the statistical complexity dimension. This is given by:

$$d_\mu(\mu) = -\frac{h_\mu}{\lambda_\mu} \ , \tag{5.8}$$

recalling that $h_\mu > 0$, so that the dimension is always positive. Failing OSC, this ratio is an upper bound on the dimension of the measure μ [**81**]. For a discussion of the intuition behind this formula, see Appendix D.

5.6. Many-state Hidden Markov Models

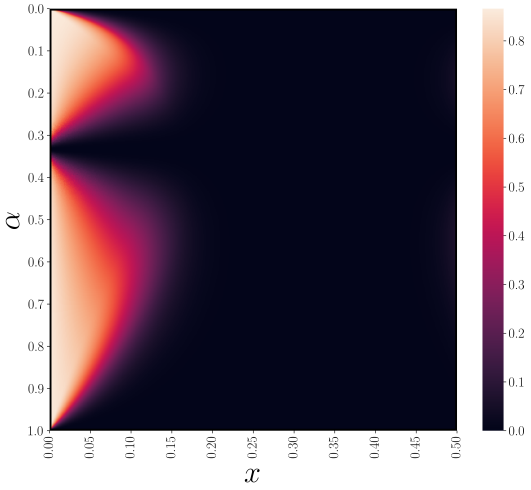
Notably, the Lyapunov dimension Eq. (5.8) for more-than-two-state HMCs is easily shown to be correct when the maps are similitudes and the probability functions are constant. The latter is seen, for example, with the Sierpinski triangle, as discussed in Appendix E.

However, we are generally interested in multi-state HMCs that do not produce perfectly self-similar fractals. Furthermore, we are often interested in considering physical systems described by parametrized HMCs, such as those that arose in the two prequels on quantum measurement processes and information engine functionality [**52**, **53**]. In such cases, an HMC determined by an application may meet the OSC in some regions of parameter space and fail to do so in others. Let's consider an HMC that spans the breadth of these possible behaviors, from zero overlap to complete overlap. This will demonstrate the range of applicability of our statistical complexity dimension algorithm.

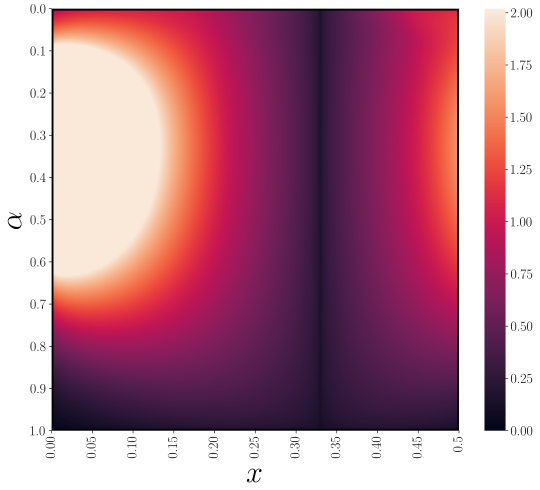
Consider the HMC with 3 symbols and 3 states defined in Eq. (3.7). Figure 5.3 shows how the MSP attractors change across the (α, x) parameter space. Each black dot is a generated mixed state, while the colored regions show the range of each symbol-labeled map.

For example, on one hand, in the top left corner with $\alpha = 0.01$ and $x = 0.01$, we find an attractor that fills the simplex, with moderate amounts of overlap. On the other, $\alpha = 0.79$ and $x = 0.11$ produces an attractor with no overlap, and clearly defined regions.

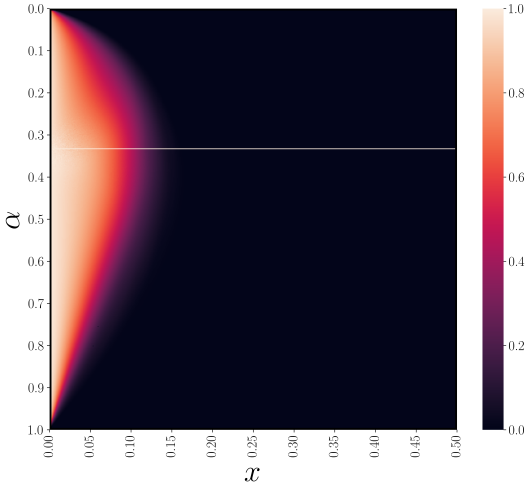
Moreover, for any α, choosing $x = 1/3$ leads the MSP attractor to collapse to a finite 3-state HMC, since the symbol-labeled mapping functions become constant functions. In this case, there is no overlap, as each symbol-labeled map takes on a different constant value.



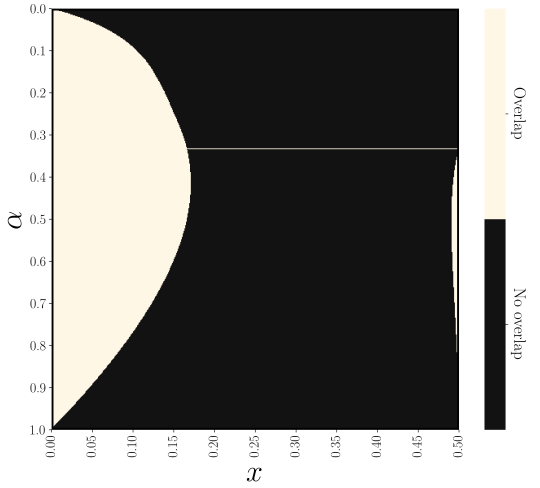
(A) Mixed-state attractor area estimated across α and x. The area is minimized along lines of $\alpha = 1/3$ and $x = 1/3$, where the attractor becomes point-like.



(B) Mixed-state attractor Lyapunov dimension does not detect the collapse to zero dimension along $\alpha = 1/3$, which is due to overlaps.



(C) Percentage of attractor area in which there is overlap.



(D) Overlap in the attractor area. Comparing with (c), for much of this area, the overlap is very small.

FIGURE 5.4. Attractor area, overlap regions, and Lyapunov dimension of the mixed-state attractors shown in Fig. 5.3, parametrized by $\alpha = [0, 1]$ and $x = [0, 0.5]$. The HMC itself is given in Section 5.6. See Appendix F and Appendix G for a detailed discussion of the production of these plots.

However, when $\alpha = \beta = 1/3$, all symbol-labeled mapping functions are identical. Therefore, the attractor is the single fixed-point shared by all three maps—a single-state HMC. This is a case of maximal possible overlap. Along both lines in parameter space the MSP collapses to a finite-state

HMC, so $d_\mu = 0$, by definition. However, these different mechanisms of state-collapse are relevant in calculation of d_Γ via Eq. (5.6).

First, consider Fig. 5.4a, which illustrates the estimated area on the simplex taken up by the attractor across parameter space. For a discussion of how attractor area was estimated, see Appendix G. This figure matches the grid in Fig. 5.3: lower values of x produce larger attractors, excepting the region near $\alpha = 1/3$, where the area drops to zero.

Now, the area taken up by an attractor is not a good proxy for dimension—we may have very small-in-size attractors with large dimension. Indeed, this is the case for most values of α near $1/3$, where the attractors are very small but whose $d_\mu \to 2$. However, we know from analyzing the attractor grid that for HMCs lying exactly on this line, the attractor is instantaneously finite state. And so, the statistical complexity dimension d_μ must discontinuously drop to zero. However, this is not accurately reflected by d_Γ, as seen in Fig. 5.4b.

That said, d_Γ clearly smoothly approaches zero as $x \to 1/3$. Along the vertical line, d_Γ is correctly exactly zero. This is the other line in parameter space where the MSP is finite state and d_μ is analytically known to be zero. The disparity, in both correctness and continuity, is due to the different mechanisms driving the collapse noted above.

As $x \to 1/3$, the slopes of the symbol-labeled mapping functions approach zero. When $x = 1/3$, the symbol-labeled mapping functions become constant values, as reflected in the Lyapunov exponents and consequently in d_Γ. The constant functions have negative Lyapunov exponents of negative infinity, sending Eq. (5.6) to zero.

In contrast, along the $\alpha = 1/3$ line, the contraction in the state space is a result of the maps instantaneously sharing a fixed point. For $\alpha = 1/3 \pm \epsilon$, the attractor is not finite. In this region of parameter space, symbol-labeled maps are not infinitely contracting, so d_Γ badly overestimates d_μ along the $\alpha = 1/3$ line. This illustrates the importance of the OSC on the bound.

This poses the question, which regions in HMC parameter space exhibit overlap IFS maps? Figure 5.4d depicts parameter space regions in terms of overlap or no overlap. Figure 5.4c shows this as a percentage of the total attractor area. For a discussion of how overlap was determined, please see Appendix G. Comparing the two, we see that there is a significant region over which overlap does exist for $x < 0.15$ and a smaller region where $x > 0.48$. However, for much of that region the

attractor's overlap area is relatively small. As a rule of thumb, the gap between d_Γ and d_μ for the mixed-state set is determined by the percentage of the attractor that is affected by overlap. If the overlap region is relatively small, in comparison to the size of the attractor, d_Γ may be very close to d_μ.

However, if the overlap is very large, d_Γ may be a dramatic overestimation of d_μ. This occurs when $\alpha = 1/3$ and $x < 0.15$. The statistical complexity dimension d_μ vanishes, yet the Lyapunov dimension saturates at $d_\Gamma = 2.0$.

We also note that the mechanism driving the collapse of the MSP attractor at $\alpha = 1/3$ is a discontinuity in the parameter space, as compared to $x = 1/3$. This is because state space collapse due to overlap requires the maps to be identical, and even minute differences in the symbol-labeled transition matrices will produce an uncountably-infinite MSP, potentially with $d_\mu = 2.0$. This encourages us to consider not just the statistical complexity dimension, but also the area of the attractor and the nearby regions in parameter space for a clearer understanding of the underlying HMC. In this, the tools developed here, by allowing (computationally-efficient) surveys of large regions of parameter space, are particularly useful.

CHAPTER 6

Introducing the Ambiguity Rate

The ϵ-machine performs all relevant aspects of scientific modeling—prediction, generation, and pattern recognition. It captures a system's generation, storage, and transmission of information. We may, at various times, imagine it as distinct from the system, a predictor of its behavior, or as representative of the system itself, generating data that is statistically indistinguishable from observation. We may also consider the ϵ-machine as a memoryful channel, mapping pasts to futures. Consider a channel interacting with a system via a time-series of discrete observations, at each time step updating its configuration represent the current optimal prediction of the system. Then the possible configurations of the channel are the causal states, and channel has C_μ bits of memory.

The student of information theory will recall that Shannon, in his analysis of information transmission through channels, introduced two mechanisms: *equivocation*, in which the same input may lead to distinct outputs, and *ambiguity*, in which two different inputs may lead to the same output; see Fig. 6.1. When our channel is taken to be the ϵ-machine, the equivocation rate of the channel is the *entropy rate* h_μ of the underlying system—the rate at which the system generates future information. This is guaranteed by the predictive optimality of the ϵ-machine—the only noise in the channel is due to the intrinsic randomness of our complex system.

In this chapter, we introduce the parallel quantity, the *ambiguity rate* h_a. The ambiguity rate tracks the rate at which the system discards past information by introducing uncertainty over the infinite past. Explicitly, if a process can be optimally modeled with a finite set of predictive features, its ϵ-machine must forget information at the same rate at which the system generates it, so as to not grow the size of the model over time: $h_\mu = h_a$. However, this is atypical, as we have already shown.

In Chapter 5, accurate calculation of d_μ was contingent on the DIFS meeting restrictive technical conditions, outlined in Section 5.4. Introducing ambiguity rate h_a reframes these constraints information-theoretically, effectively lifting them. The result is a new method to accurately calculate

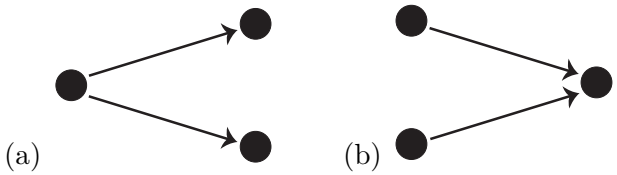


FIGURE 6.1. (a) Equivocation: Same input sequence leads to different outputs. (b) Ambiguity: Two different inputs lead to same output. The strategy underlying Shannon's proof of his second coding theorem is to find channel inputs that are least ambiguous given the channel's distortion properties.

d_μ for a broad class of complex processes. More abstractly, we propose h_a as a new intrinsic complexity measure of a stochastic process—the *growth rate* of the information stored in a process' optimally predictive features. When $h_\mu = h_a$, this growth rate vanishes and the associated ϵ-machine's internal causal-state process is stationary. However, when $h_\mu > h_a$, the latter process is nonstationary and any optimal predictor must accumulate new information over time to sustain accurate predictions.

6.1. What is the Ambiguity Rate?

As discussed in Chapter 5, the *overlap problem* is a long-standing concern for iterated function systems that arises from the overlap of the ranges of the symbol-labeled mapping functions $f^{(x)}$. Figure 6.2 illustrates the issue. Specifically, to quantitatively count system orbits we must properly monitor orbit divergence and convergence. This then requires distinguishing between iterated function systems that meet the open set condition (OSC) and those that do not.

When the OSC is not met, the inequality in the d_μ bound Eq. (5.7) becomes strict. This is a consequence of using h_μ as our measure of state space expansion in Eq. (5.6). The Shannon entropy rate tracks the uncertainty in the next symbol x given our current causal state η_t, averaged over the Blackwell measure. From a dynamical systems point of view, we identify this as the typical growth rate of orbits (words) in symbol space.

When the OSC is met, the Shannon entropy rate also measures the typical growth rate of orbits in the $(N-1)$-simplex. Observing x, current state η_t transitions to the next state η_{t+1} via application of the mapping function $\eta_{t+1} = f^{(x)}(\eta_t)$. Because the images of the map do not overlap,

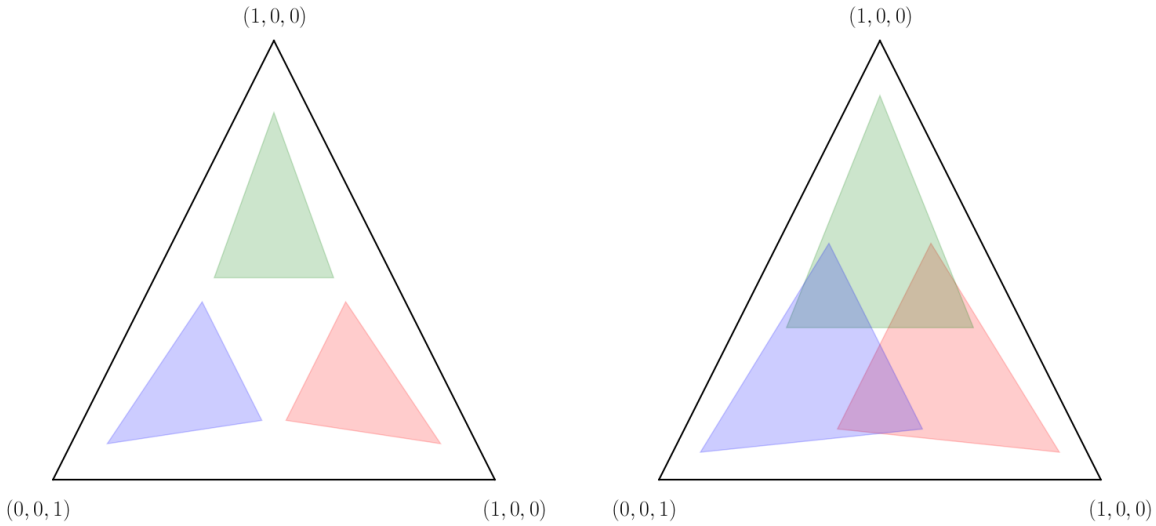
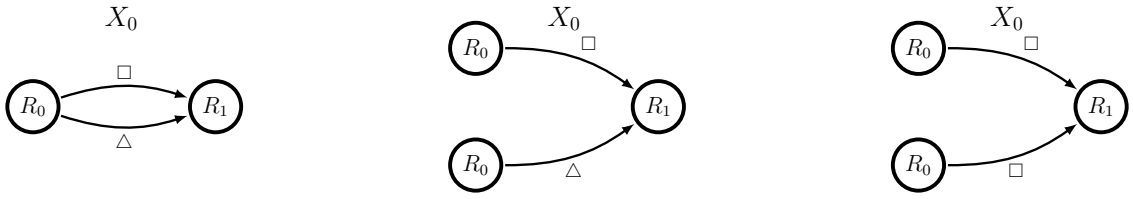


FIGURE 6.2. Overlap problem on the 2-simplex Δ^2: Two distinct DIFSs (given in Appendix A) are considered, each with three mapping functions. Images of the mapping functions over the entire simplex are depicted as regions in red, blue, and green. (a) Images of the mapping functions $f^{(\triangle)}$,$f^{(\square)}$ and $f^{(\circ)}$ do not overlap—every possible state has a unique pre-image. (b) Images of the mapping functions overlap: there exist $\eta_1, \eta_2 \in \Delta^2$ such that $f'^{(\triangle)}(\eta_1) = f'^{(\square)}(\eta_2) = \eta_3$. This case is an *overlapping DIFS*.

this is guaranteed to be a distinct new state—thus state sequences grow at the same rate as words do. Then, we may use h_μ to measure expansion of the state space.

However, when the OSC is not met, it is possible for two distinct states $\eta_t, \zeta_t \in \Delta$ to map to the same next state on different symbols, by occupying the "overlapping region", as depicted in Fig. 5.2. In this case, $\eta_{t+1} = \zeta_{t+1}$ has no unique pre-image. This introduces ambiguity about the past, given knowledge of the current state. As a consequence, using the Shannon entropy rate as a proxy for the state expansion rate implies a more rapid expansion in state space than is actually occurring. This indicates the need adjust our use of h_μ when determining d_μ.

We propose the *ambiguity rate* as the correction to $\widetilde{d_\Gamma}$'s overcounting of orbits. Since the problem at hand is an overestimation in uncertainty in our state space, we must identify and quantify mechanisms of state uncertainty reduction when the OSC is not met. Consider that when the OSC is met, every state η_t has a unique pre-image η_{t-1} that can only be reached via a single, specific



(A) $H[X_0|\mathcal{R}_0, \mathcal{R}_1]$ (B) $I[X_0; \mathcal{R}_0|\mathcal{R}_1]$ (C) $H[\mathcal{R}_0|X_0, \mathcal{R}_1]$

FIGURE 6.3. Sources of ambiguity rate depicted in state machines: (A) $H[X_0|\mathcal{R}_0, \mathcal{R}_1] > 0$—Previous state $\mathcal{R}_0$ is mapped to the next state $\mathcal{R}_1$ by two distinct symbols. This occurs when two symbols have identical mapping functions. (B) $I[X_0; \mathcal{R}_0|\mathcal{R}_1] > 0$—Two distinct previous states $\mathcal{R}_0$ map to the same next state by distinct symbols, due to overlapping mapping functions. (C) $H[\mathcal{R}_0|X_0, \mathcal{R}_1] > 0$—Two distinct previous states $\mathcal{R}_0$ map to the same next state by the same symbol. This occurs when a mapping function is noninvertible.

observed symbol. When the OSC is not met, for a subset of $\eta \in \boldsymbol{\mathcal{R}}$ there is uncertainty about the previous state, the previous symbol, or both. Quantifying this ambiguity about the past is the goal in constructing the ambiguity rate h_a.

Intuitively, it would seem that generating uncertainty in reverse time is equivalent to reduction of uncertainty in forward time. The following shows that this is the case and that the ambiguity rate is the necessary correction to the DIFS dimension formula Eq. (5.6).

6.2. Sources of State Uncertainty Reduction

For ϵ-machines represented as DIFSs, there are three distinct mechanisms that contribute to the ambiguity rate, as depicted in Fig. 6.3.

The first is *identical mapping functions*, depicted in Fig. 6.3 (a). When for $x, x' \in \mathcal{A}$, $f^{(x)}(\eta) = f^{(x')}(\eta)$ for all $\eta \in \boldsymbol{\mathcal{R}}$, we say that x and x' have identical mapping functions. In this case, the distinction between x and x' is not reflected in state sequences and produces ambiguity in the symbol sequence. We quantify this as the Shannon entropy in our current symbol, conditioned on the previous state and the next state: $H[X_t|\mathcal{R}_t, \mathcal{R}_{t+1}]$.

The second is overlapping mapping functions, which motivated this investigation and already have been defined. Their impact on the state machine is shown in Fig. 6.3 (b). In this case, two distinct symbols $x, x' \in \mathcal{A}$ map two distinct states $\eta, \zeta \in \boldsymbol{\mathcal{R}}$ to the same next state. Although the

previous state affects the probability distribution over the observed symbol, the next state "forgets" that distinction. This is quantified by the mutual information shared by the current symbol and the previous state, conditioned on the next state: $I[X_t; \mathcal{R}_t | \mathcal{R}_{t+1}]$.

Finally, there is noninvertibility in the mapping functions. If a single mapping function maps distinct states $\eta, \zeta \in \boldsymbol{\mathcal{R}}$ to the same next state, the pasts that led to η and ζ can no longer be distinguished. Figure 6.3 (c) shows this in general. However, it may also be observed in $f^{(\square)}$ from Fig. 3.3, which maps every state to $\eta_0 = (1, 0)$. Numerically, the reduction via this mechanism is measured by the Shannon entropy in the previous state, given our next state and current symbol: $H[\mathcal{R}_t | X_t, \mathcal{R}_{t+1}]$.

Combining these three sources of uncertainty reduction defines the *ambiguity rate*:

$$\begin{aligned} h_a &= H[X_t | \mathcal{R}_t, \mathcal{R}_{t+1}] + I[X_t; \mathcal{R}_t | \mathcal{R}_{t+1}] \\ &\quad + H[\mathcal{R}_t | X_t, \mathcal{R}_{t+1}] \\ &= H[X_t, \mathcal{R}_t | \mathcal{R}_{t+1}] \ . \end{aligned} \tag{6.1}$$

This can be rewritten as a integral over $\boldsymbol{\mathcal{R}}$:

$$h_a = -\int_{\eta \in \boldsymbol{\mathcal{R}}} \mathrm{d}\mu_B(\eta) \sum_{\substack{x \in \mathcal{A}, \\ \zeta \in (f^{(x)})^{-1}(\eta)}} \Pr(x, \zeta | \eta) \log_2 \Pr(x, \zeta | \eta). \tag{6.2}$$

In this, we must be careful about the pre-images of η, due to the possibility of noninvertible mapping functions. The probability distribution inside the summation is given by the relationship:

$$\begin{aligned} &\Pr(X_0 = x, \mathcal{R}_0 = \zeta | \mathcal{R}_1 = \eta) = \\ &\qquad \frac{\mu_B(\mathcal{R}_0 = \zeta)}{\mu_B(\mathcal{R}_1 = \eta)} \times \Pr(X_0 = x | \mathcal{R}_0 = \zeta) \ . \end{aligned} \tag{6.3}$$

Calculating this distribution requires calculating or estimating the Blackwell measure, which may be nontrivial. Section 6.5 and Section 6.6 discuss this in greater depth.

6.3. A Conjecture About the Kaplan-Yorke Conjecture

The information-theoretic decomposition of ambiguity rate facilitates combining h_a and h_μ. Recall that for prediction, the states of a predictive model are equivalent to knowledge of the infinite past. Due to this, the Shannon entropy rate may be written $H[X_t|\mathcal{R}_t]$. Combining this with the ambiguity rate gives:

$$\begin{aligned} h_\mu - h_a &= H[X_t|\mathcal{R}_t] - H[X_t, \mathcal{R}_t|\mathcal{R}_{t+1}] \\ &= H[\mathcal{R}_{t+1}|\mathcal{R}_t, X_t] + H[\mathcal{R}_{t+1}] - H[\mathcal{R}_t] \\ &= \Delta H[\mathcal{R}_t] \ . \end{aligned}$$

Moving to the third line called on the fact that the symbol and state transitions are defined by functions. So, the difference between the Shannon entropy rate and the ambiguity rate gives the rate of growth of our causal state set $\mathcal{R}$.

Recall that the information dimension, as defined in Eq. (5.2), compares the average growth of occupied cells $\mathcal{F}$—taking into account the measure over those cells—as the cell size ϵ shrinks. To adhere to the main development, here we will not walk through the heuristic for how a dimensional quantity is determined from the Γ. Nonetheless, we will show how the relationship between d_1, $h_\mu - h_a$, and Γ is intuitive for DIFSs in one dimension.

When the DIFS states lie in the 1-simplex, Γ consists of only one exponent $\lambda_1 < 0$, which is the weighted average of the Lyapunov exponents of each map:

$$\lambda_1 = \int \sum_x p^{(x)}(\eta) \log \left| \frac{df^{(x)}(\eta)}{d\eta} \right| d\mu \ ,$$

where μ is the Blackwell measure.

Now, consider a line segment in Δ^1 of length ϵ. Mapping this line forward k times by the DIFS produces, averaging over several iterations of this action, $2^{(h_\mu - h_a)k}$ new lines of length $\epsilon e^{\lambda_1 k} < \epsilon$. (Note that the use of base-two for Shannon entropy rather than base e follows convention; retained here for familiarity. In numerical calculation of d_μ, we recommend a consistent base be chosen for h_μ, h_a, and the Γ.) The logarithmic ratio of the growth rate of lines (as averaged over the Blackwell

measure) compared to the shrinking of these lines is the simple ratio:

$$d_\mu = -\frac{h_\mu - h_a}{\lambda_1} \ .$$

This, of course, is exactly the definition of the information dimension Eq. (5.2) and is, assuming the DIFS is an ϵ-machine, the statistical complexity dimension d_μ.

For higher-dimensional DIFSs, we conjecture that the ambiguity rate is the adjustment to the IFS Lyapunov dimension formula that gives the information dimension:

$$\widetilde{d_\mu} = \begin{cases} k + \dfrac{\Lambda(k) + h_\mu - h_a}{|\lambda_{k+1}|}, & -\Lambda(N) > h_\mu - h_a \\ N \ , & -\Lambda(N) \leq h_\mu - h_a \end{cases} , \tag{6.4}$$

where as in Eq. (5.6), $\Lambda(m)$ is the Lyapunov spectrum partial sum $\Lambda(m) = \sum_{i=1}^{m} \lambda_i$ and $k = 0, 1, 2, \ldots, N-1$ is the largest index for which $-\Lambda(k) < h_\mu - h_a$.

6.4. Interpretations of the Ambiguity Rate

Up to this point, we motivated ambiguity rate as correcting over counting in the DIFS statistical complexity dimension d_μ. It is worth discussing the quantity in more depth.

On the one hand, note that when $h_\mu - h_a = 0$, the causal-state process is stationary and C_μ time-independent: $\Delta H[\mathcal{R}_t] = 0$. This occurs for finite-state DIFSs, as well as many with countably-infinite states; see Section 6.5. When this occurs, applying Eq. (6.4) returns a vanishing statistical complexity dimension $d_\mu = 0$, as expected.

On the other hand, when ambiguity rate vanishes, C_μ grows at the Shannon entropy rate: $\Delta H[\mathcal{R}_t] = h_\mu$. This occurs when there are no identical maps, no overlap, and no noninvertibility in the mapping functions. In short, $h_a = 0$ when the process is "perfectly self-similar" and every new observed symbol produces a new, distinct state.

With this in mind, we can use the ambiguity rate, and specifically $h_\mu - h_a$, to describe the stationarity of the model's internal state process. The state set is time independent. When $h_a > 0$, however, to optimally predict the process $\mathcal{P}$ requires a nonstationary model (temporally-growing state set $\boldsymbol{\mathcal{R}}$), even though $\mathcal{P}$ is itself stationary. This is a consequence of modeling "out of class".

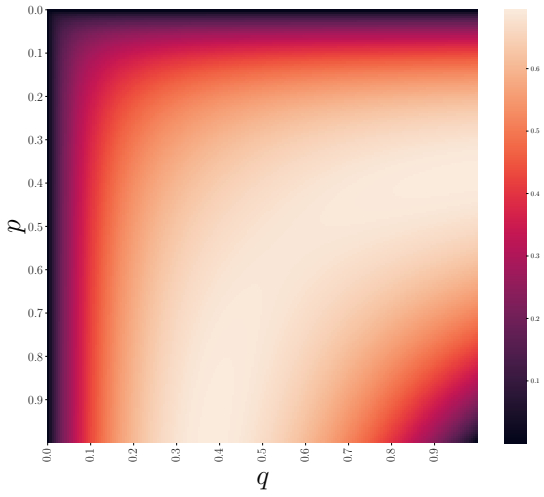


FIGURE 6.4. The entropy rate h_μ, which in this case is equivalent to the ambiguity rate h_a, is plotted for the DIFS defined by Eq. (3.9) for $p, q, \in (0, 1)$.

That is, predicting a perfectly self-similar $\mathcal{P}$ requires differentiating every possible infinite past. This is only possible with a DIFS by storing new states at the rate new pasts are being created. (Moving to a more powerful model class by, say, imbuing our states with counters or stacks, may make it possible to model $\mathcal{P}$ with a stationary model.)

This perspective naturally leads to another that probes the efficacy of the causal-state mapping. Considering the space of all possible infinite pasts $\overleftarrow{X}$, the causal-state mapping $f_\epsilon(\overleftarrow{X}) \rightarrowtail \boldsymbol{\mathcal{R}}$ is defined such that:

$$f_\epsilon(\overleftarrow{X} = \overleftarrow{x}) = f_\epsilon(\overleftarrow{X} = \overleftarrow{x'}) = \eta_i \ ,$$

if $\Pr(X_{0:\ell}|\overleftarrow{X} = \overleftarrow{x}) = \Pr(X_{0:\ell}|\overleftarrow{X} = \overleftarrow{x'})$ for all $\ell \in \mathbf{N}^+$. When the process is perfectly self-similar, the causal-state mapping is one-to-one and $h_a = 0$. In this case, storing the causal states is no better for prediction than simply tracking the space of all pasts. (Although the causal-state set $\boldsymbol{\mathcal{R}}$ is still informative in characterizing how we might approximate the process with a finite state machine [60].) The number of pasts each state "contains" is stationary and given by $2^{h_a} = 1$.

In general, for a stationary process $\mathcal{P}$, the average number of pasts contained by a given causal state grows at the rate 2^{h_a}. When the process has a stationary state set, the number of pasts each state contains must necessarily grow at the rate new pasts are being generated, and so $2^{h_\mu} = 2^{h_a}$.

Finally, let's close with a short historical perspective. The development of d_μ was partially inspired by Shannon's definition in the 1940s of *dimension rate* [1]:

$$\lambda = \lim_{\delta \to 0} \lim_{\epsilon \to 0} \lim_{T \to \infty} \frac{N(\epsilon, \delta, T)}{T \log \epsilon} \ ,$$

where $N(\epsilon, \delta, T)$ is the smallest number of elements that may be chosen such that all elements of a trajectory ensemble generated over time T, apart from a set of measure δ, are within the distance ϵ of at least one chosen trajectory. This is the minimal "number of dimensions" required to specify a member of a trajectory (or message) ensemble. Unfortunately, Shannon devotes barely a paragraph to the concept, leaving it largely unmotivated and uninterpreted.

Therefore, it appears the first modern discussion of a dimensional quantity of this nature for stochastic processes motivated the development using resource theory [60], noting that the d_1 of the causal-state set Eq. (5.2) characterizes the *distortion rate* when coarse-graining an uncountably-infinite state set. Starting from the dimensional quantity, the relationship to statistical complexity was then forged.

In this light, developing ambiguity rate and calling out its easy mathematical connection to $\Delta H[\boldsymbol{\mathcal{R}}_t]$ flips this motivation. The quantity $h_\mu - h_a$ can be defined purely in terms of $\mathcal{P}$ and has an intuitive relationship to the causal-state mapping. The dimensional quantity d_μ naturally falls out when we compare this rate of model-state growth to the dynamics of the causal states in the mixed-state simplex. Therefore, we may motivate d_μ not as only a resource-theoretic tool for finitizing infinitely-complex state machines, but also as an intrinsic measurement of a process' structural complexity.

We now consider two examples. The first is a parametrized discrete-time renewal process that has a countably-infinite state space for all parameters. This allows us to explicitly write down the Blackwell measure and calculate ambiguity rate exactly using Eq. (6.2). The second is a parametrized machine with three maps, which has an uncountably-infinite state space for nearly all parameters. Calculating the ambiguity rate in this case requires us to approximate the Blackwell measure using Ulam's method.

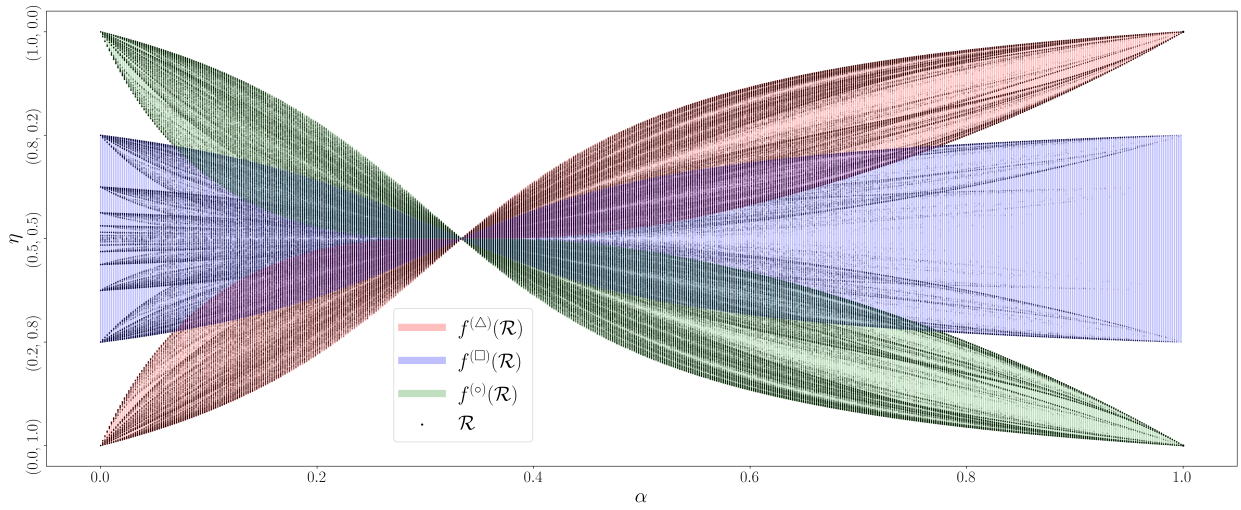


FIGURE 6.5. One-dimensional attractor for DIFS given in Eq. (6.5) with $x = 0.25$ and horizontally varying $\alpha \in (0, 1)$. State set $\eta \in \boldsymbol{\mathcal{R}}$ plotted (red, blue green) on top of the images of the mapping functions $\left\{f^{(\triangle)}, f^{(\square)}, f^{(\circ)}\right\}$ applied to $\boldsymbol{\mathcal{R}}$. For $\alpha \in (0.07, 0.78)$, there is overlap in the images of the maps.

6.5. Example: Discrete-Time Renewal Process

Recall the previously-discussed DIFS depicted in Fig. 3.3, which depicts the stages of producing the MSP for the substochastic matrices given in Eq. (3.9) when where $p = q = \frac{1}{2}$. For the general case where $p, q \in (0, 1)$ are left unspecified, we have the probability function set:

$$p^{(\triangle)}(\eta) = 1 - \langle \eta | \delta_2 \rangle p \ ,$$

$$p^{(\square)}(\eta) = \langle \eta | \delta_2 \rangle p \ ,$$

and the mapping function set:

$$f^{(\triangle)}(\eta) = \left(\frac{\langle \eta | \delta_1 \rangle (1-q)}{1 - \langle \eta | \delta_2 \rangle p}, \frac{\langle \eta | \delta_1 \rangle q + \langle \eta | \delta_2 \rangle (1-p)}{1 - \langle \eta | \delta_2 \rangle p} \right) \ ,$$

$$f^{(\square)}(\eta) = (1, 0) \ .$$

Since $\boldsymbol{\mathcal{R}}$ is countable, we may write down the state set $\boldsymbol{\mathcal{R}}$ as a sequence:

$$\eta_n = \left[\frac{(p-q)(1-q)^n}{p(1-q)^n - q(1-p)^n}, \frac{q(1-q)^n - q(1-p)^n}{p(1-q)^n - q(1-p)^n} \right] \ ,$$

where n is the number of $\triangle$s seen since the last $\square$ and $p \neq q$. This simple structure allows us to give the Blackwell measure explicitly:

$$\mu_B(n) = \frac{p(1-q)^n - q(1-p)^n}{p-q} \times \frac{pq}{p+q} \ ,$$

where $\mu_B(n)$ is the asymptotic invariant measure over the state induced after seeing n $\triangle$s since the last $\square$.

With the Blackwell measure in hand, the entropy rate can be explicitly calculated as the infinite sum:

$$\begin{aligned} h_\mu &= \sum_{n=1}^{\infty} \mu_n H[X_n | \mathcal{R}_n = \eta_n] \\ &= -\sum_{n=1}^{\infty} \mu_n \Big(p^{(\triangle)}(\eta_n) \log_2 p^{(\triangle)}(\eta_n) \\ &\qquad + p^{(\square)}(\eta_n) \log_2 p^{(\square)}(\eta_n) \Big) \ . \end{aligned}$$

Figure 6.4 plots h_μ for $p, q \in (0, 1)$. In calculating h_μ, there is a contribution from every state except the first—η_0—since the first state transitions to the second with probability one and there is no branching uncertainty. Every other state transitions on a coin flip of a determined bias between $(\triangle, \square)$, generating uncertainty with each transition.

In contrast to how h_μ averages over all mixed states, ambiguity rate accumulates in only one state—η_0. From Fig. 3.3, we see that $H[x_n, \eta_n | \eta_{n+1}] = 0$ for all n other than $n = 0$. That is, each state η_n is only accessed via the prior state η_{n-1}, except for η_0, which may be accessed from every other state. So, ambiguity in the past can only be introduced by visiting η_0. Since these transitions only occur on a $\square$, we must find the probability distribution $\Pr(X_0 = \square, \mathcal{R}_0 = \eta_n | \mathcal{R}_1 = \eta_0)$.

Applying Eq. (6.2) and Eq. (6.3), we explicitly write down the ambiguity rate as:

$$h_a = \mu_0 \sum_{n=1}^{\infty} \left(\frac{\mu_n}{\mu_0} p^{(\square)}(\eta_n) \right) \log_2 \left(\frac{\mu_n}{\mu_0} p^{(\square)}(\eta_n) \right) \ .$$

Both h_μ and h_a are infinite summations, but when calculating the ambiguity rate, the sum refers to calculating a single Shannon entropy over the infinite, discrete distribution representing the probability distribution over prior states when arriving in η_0.

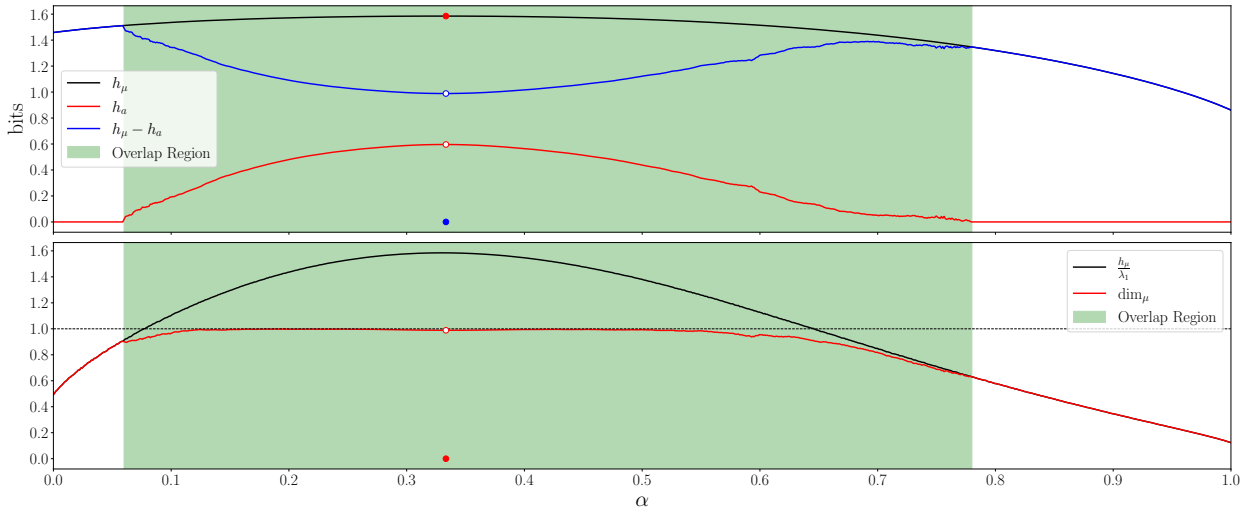


FIGURE 6.6. Numerical calculation of entropies and dimension quantities for the DIFS given in Eq. (6.5), $\alpha \in (0,1)$, and $x = 0.25$: (Top) The entropy rate h_μ, the ambiguity rate h_a, and $h_\mu - ha$. (Bottom) Comparison of $\frac{h_\mu}{\lambda_1}$ to $d_\mu = \frac{h_\mu - h_a}{\lambda_1}$. The latter smoothly departs from the former and is approximately 1 for much of the overlap region, except where it discontinuously jumps to zero at $\alpha = 1/3$.

Since the state space does not grow—$\Delta H[\boldsymbol{\mathcal{R}}_t] = 0$—the entropy rate $h_\mu = h_a$ as $n \to \infty$. Therefore, d_μ vanishes for all values of p and q. This will always be the case for finite-state DIFSs and, in general, for those with countable state spaces.

6.6. Example: Ambiguity Rate in One Dimension

Now, let's turn to the more general case, those DIFSs with uncountably-infinite state spaces. For the moment we restrict to one-dimensional DIFSs, so that the states lie in the 1-simplex. Consider a DIFS with the alphabet $\mathcal{A} = \{\triangle, \square, \circ\}$ and the associated substochastic matrices:

$$T^{\triangle} = \begin{pmatrix} \alpha y & \beta x \\ \alpha x & \beta y \end{pmatrix}, \; T^{\square} = \begin{pmatrix} \beta y & \beta x \\ \beta x & \beta y \end{pmatrix}, \text{ and}$$

$$T^{\circ} = \begin{pmatrix} \beta y & \alpha x \\ \beta x & \alpha y \end{pmatrix}, \tag{6.5}$$

with $\alpha = 1 - 2\beta, \alpha \in (0,1)$ and $x = 1 - y, x \in (0,1)$.

Figure 6.5 depicts all of the DIFSs for the slice of the parameter space where $x = 0.25$. The vertical axis is the 1-simplex and each vertical slice plots the state space $\boldsymbol{\mathcal{R}}(\alpha)$ at the appropriate value of α, given on the horizontal axis. Additionally, the images of the functions $\left\{f^{(\triangle)}(\boldsymbol{\mathcal{R}}), f^{(\square)}(\boldsymbol{\mathcal{R}}), f^{(\circ)}(\boldsymbol{\mathcal{R}})\right\}$ are shaded in red, blue, and green, respectively.

At $\alpha = 1/3$, the mixed state set $\boldsymbol{\mathcal{R}}$ contracts to a finite set, and h_μ must equal to h_a, making $d_\mu = 0$. At this point in parameter space, the state set consists of only one state; $\boldsymbol{\mathcal{R}}(\alpha = 1/3) = \{(1/2, 1/2)\}$. At every other value of α, $h_a < h_\mu$. There is overlap in the images of the maps for, approximately, $\alpha \in (0.07, 0.78)$. In this regime, $h_a > 0$.

To calculate the ambiguity rate and therefore the statistical complexity dimension d_μ we use a modified Ulam's method to approximate the Blackwell measure and then approximate the integral equation Eq. (6.2). This method is not the only way to find the ambiguity rate, but does have several advantages, including speed and the ability to control the accuracy of our approximation. This method is discussed in depth in Section 6.7.

The top plot in Fig. 6.6 gives the entropy rate, ambiguity rate, and $h_\mu - h_a$ for the DIFSs pictured in Fig. 6.5. As α is increased, h_a smoothly increases from zero as overlap begins to occur. It approaches ≈ 0.6 around $\alpha = 1/3$, but is discontinuously equal to h_μ at this point. The reason for this is an instantaneous equality in the fixed points of the mapping functions, causing the state space to collapse. As α increases to 1, h_a smoothly decreases back to zero. The roughness seen in the plot is due to numerical precision.

For a large portion of the overlap region, d_Γ saturates at 1.0. The bottom plot in Fig. 6.6 instead depicts $\frac{h_\mu}{\lambda_1}$ to show how this quantity smoothly changes across parameter space, maxing out at around 1.6. In comparison, $d_\mu = \frac{h_\mu - h_a}{\lambda_1}$ smoothly departs from the Lyapunov dimension when overlap begins and instead hugs the underside of the dim = 1.0 line for much of the overlap region. Again, at $\alpha = 1/3$ there is the discontinuous drop to $d_\mu = 0$, followed by d_μ smoothly rejoining with the Lyapunov dimension as the overlap region ends.

Unsurprisingly, calculating the ambiguity rate in higher dimensions is more challenging. Although, in principle, Ulam's method still applies and we may in principle follow the algorithm laid out in Section 6.7, higher-dimensional mapping functions introduce additional error sources in the

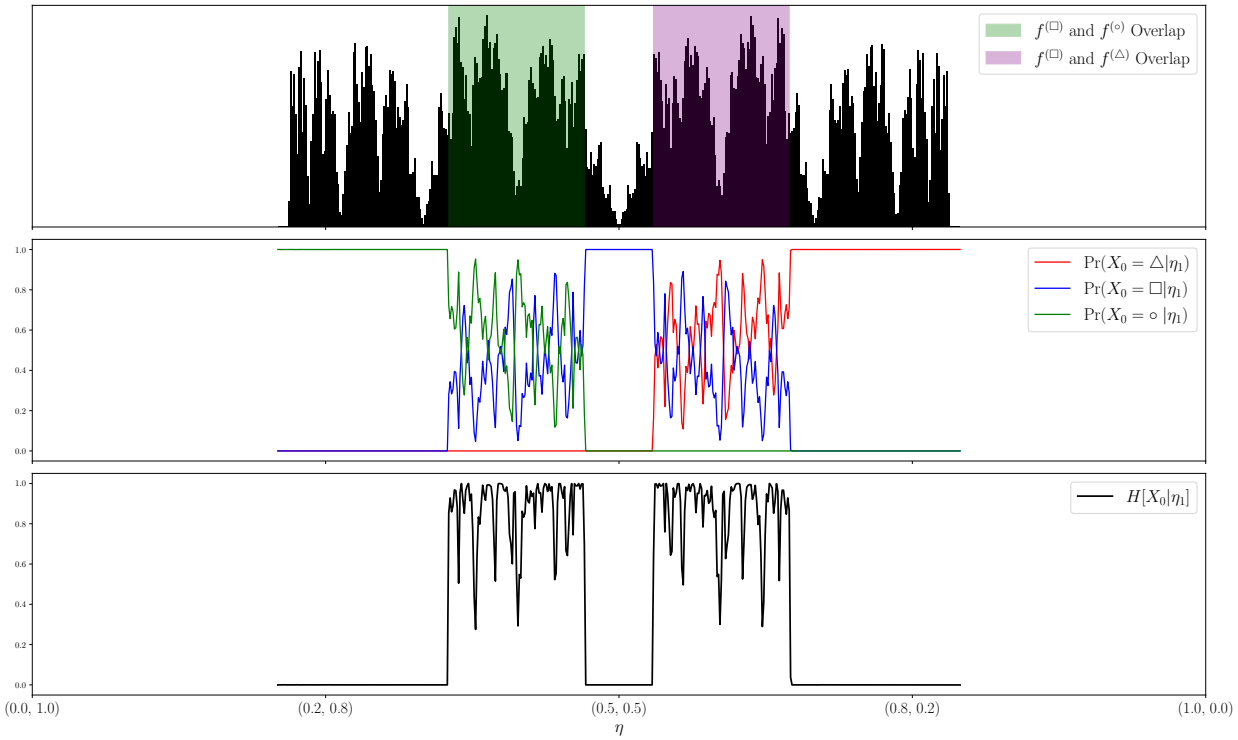


FIGURE 6.7. DIFS for $x = 0.25$ and $\alpha = 0.5$: (Top) Blackwell Measure μ_B approximated by Ulam's method with $k = 400$. The two overlapping regions are overlaid and may be compared with Fig. 6.5. (Middle) Probability of each prior map is plotted. In nonoverlapping regions, only one prior map is possible. In the overlapping regions, there are complicated, fractal-like distributions over multiple prior maps. (Bottom) Shannon entropy over the prior map: Nonzero only in the overlapping regions.

approximation. Developing an algorithm to efficiently and accurately calculate the ambiguity rate in higher dimensions is of great interest.

6.7. An Algorithm for the Numerical Approximation of Ambiguity Rate

To numerically approximate the ambiguity rate for a DIFSlying in the 1-simplex, we may use Ulam's method to approximate the Blackwell measure, then compute Eq. (6.2). Given a partition $\{A_1, \ldots A_k\}$ of the simplex, define:

$$P_{ij}^{(x)} = \frac{m(f^{(x)}(A_i) \cap A_j)}{m(f^{(x)}(A_i))} \times p^{(x)}\left(\overline{A_i}\right) \ ,$$

where m is the Lebesgue measure over Δ and $\overline{A_i}$ is the center of a partition element. Let $P = \sum P^{(x)}$ and find the left eigenvalue $p = pP$. Then, the invariant-measure approximation is:

$$\mu_n(A) = \sum_i p_i \frac{m(A \cap A_i)}{m(A_i)} \ .$$

For this example, let's walk through estimating the ambiguity rate for one DIFS–setting $x = 0.25$ and $\alpha = 0.5$. The partition $\{A_1, \ldots A_k\}$ is created by dividing the 1-simplex into k boxes of equal length. The approximated Blackwell measure $\widehat{\mu_B}$ for the DIFS, using $k = 400$, is shown in the top plot of Fig. 6.7. The overlay indicates the region (green) of the state space that exhibits overlap. Compare the two regions depicted in Fig. 6.7 to the overlap shown in Fig. 6.5, for the vertical slice at $\alpha = 0.5$.

Note that the partition may be defined as desired. We have found that defining the partition by calculating the set of fixed points of the mapping functions $\left\{p^x : f^{(x)}(p^x) = p^x\right\}$. Then, as many times as is desired, find all possible iterates of each fixed point, constructing a new set $\left\{f^{(w)}(p^x) : x \in \mathcal{A}, w \in \bigcup_{n=0}^{N} \mathcal{A}^n\right\}$, where $N \in \mathbb{Z}^+$. Removing duplicates and ordering the set gives a list of endpoints for a partition of the 1-simplex. Increasing N produces increasingly fine partitions. This method of defining partitions has advantages when calculating h_a across parameter space as we have in Section 6.6, since the position of the fixed point iterates in the simplex are smooth functions of α.

Regardless, once the partition is selected and $\widehat{\mu_B}$ is determined, we again use the partition. For each cell A_i, we find the probability distribution over the maps that could have transitioned into A_i: by applying Eq. (6.3) and assuming invertibility of the mapping functions:

$$\Pr(X_0 = x | \mathcal{R} \in A_i) = \frac{\widehat{\mu_B}\left(\left(f^{(x)}\right)^{-1}(A_i)\right)}{\widehat{\mu_B}(A_i)} \times p^x\left(\overline{\left(f^{(x)}\right)^{-1}(A_i)}\right) \ .$$

For our example DIFS, the probability of the previous map given current location in the simplex is plotted in the middle figure of Fig. 6.7. For parts of the simplex outside the overlapping regions, only one prior map is possible and it has probability one. Within the overlapping regions, the distribution over the possible prior maps may be very complicated. The Shannon entropy over

the prior map distribution $H\left[X_0 = x|\mathcal{R} \in A_i\right]$ is shown in the third plot of Fig. 6.7. Once these entropies are calculated, the final step is to approximate the integral equation Eq. (6.2) with a summation over cells in the partition:

$$h_a = \sum_i \widehat{\mu_B}(A_i) \sum_{x \in \mathcal{A}} H\left[X_0 = x|\mathcal{R} \in A_i\right] \ .$$

In our example, the ambiguity rate is found to be $h_a = 0.4499$. Since the DIFS entropy rate is $h_\mu = 1.5596$, this gives an adjusted state space expansion rate of $h_\mu - h_a = 1.1098$. Calculating the DIFS's Γ and applying Eq. (6.4) results in a statistical complexity dimension of $d_\mu = 0.9815$.

The advantage of Ulam's method is its relative ease and speed. Additionally, it is deterministic given the partition. We may may increase the accuracy of our approximation simply by tuning our partition, although increasingly fine partitions increase computation time. Additionally, when the set becomes highly rarified, noisiness will be observed in the calculation of h_a. This can be seenin our example DIFS on either end of the overlap region, although it is worst when $\alpha \in (0.6, 0.78)$. This may be understood when comparing Fig. 6.5 to Fig. 6.6— from $\alpha \in (0.6, 0.78)$ are bands of high density in the overlapping region that increas in probability as the overlapping region itself is shrinking. Calculating the h_a accurately in this region requires increasingly fine partitioning. An immediate improvement may be made by adapting the method to use adaptive partitioning as it sweeps parameter space, taking into account the structure of the state set. The method may be applied to any DIFS in the 1-simplex with overlaps.

CHAPTER 7

Application: Measurement Induced Randomness and Structure in Qubit Sources

We now will turn from our theoretical development to a specific application of our newly developed tools. The first topic we tackle is the impact of measurement on apparent randomness and structure: in particular, classical measurement of qubit sources.

7.1. Experimental Motivation

Temporal sequences of controlled quantum states are key to fundamental physics and its engineering deployment. Quantum entanglement [82] between emitted qubits, such as photons [83], is central to Bell probes [84] of inconsistencies between quantum mechanics and local hidden variable theories [85]. Complementing their scientific role, entangled qubits are now recognized as basic resources for quantum technologies—quantum key distribution [86], teleportation [87], metrology [88], and computing [89]. The quest there is for qubit sources that allow *on-demand* generation: at a certain time a source should emit one and only one pair of entangled photons. Qubit sources should also be *efficient*: qubits emitted and collected with a high success rate. And, individual qubits should have *specified properties*. In EPR experiments photons in emitted pairs should be identical from trial to trial. And, in communication systems polarization states should manipulable at the highest possible rates [90].

Much experimental effort has been invested to develop qubit sources that, for example, extract entangled photons from trapped atoms [91, 92], spontaneous parametric down-conversion [93], quantum dots [94], and related CQED systems [95]. To date, though, there is still no single-qubit source that exactly meets the performance desiderata. The on-demand criterion has been particularly vexing [96]. Addressing these challenges leads rather directly to a common question, one that touches on both fundamental physics and quantum engineering: *How to characterize the statistical and*

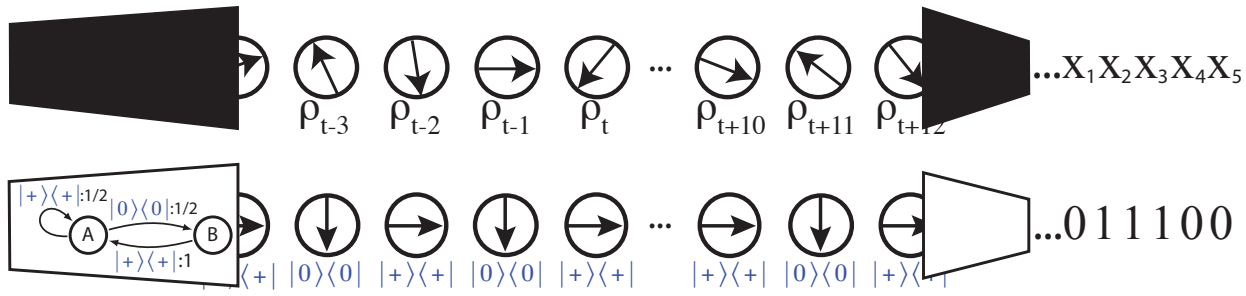


FIGURE 7.1. (Top) A general controlled qubit source (CQS) as a discrete-time quantum dynamical system that generates a time series of qubits $\rho_{t-2}\rho_{t-1}\rho_t \ldots$. (Bottom) A cCQS generates a qubit process $|0\rangle\langle 0|, |+\rangle\langle +| \ldots$. Measuring each qubit, an observer sees a classical stochastic process: (Top) $\ldots x_1 x_2 x_3 x_4 x_5$, (Bottom) $\ldots 011100$.

structural properties of qubit time series? The underlying challenge is that a systematic description of quantum processes with memory in terms of experimental measurements has yet to be given [97].

Complementing this, techniques developed within quantum process tomography and process reconstruction [98, 99, 100, 101, 102, 103, 104] achieve partial or total reconstruction of channels and system evolution. This task is of particular importance with the advent of physically realizing processing in multi-qubit systems; notably, some are now available as open cloud services [105, 106]. This progress only heightens the need to fully characterize quantum process information and statistics. In the domain of classical stochastic processes, these properties determine simulation capabilities, memory requirements, and predictability, and they quantify a process' randomness and structure [17, 30].

To address these challenges we concern ourselves with a source that generates a single qubit at a time. We imagine that the on-demand source is used repeatedly, producing an arbitrarily-long time series, which we call a *qubit process*. A simple example arises when monitoring sequential emissions from a blinking quantum dot [107, 108]. We refer to their generators as *controlled qubit sources* (CQSs). We wish to determine how random and structured they appear to be to an experimentalist.

We can now state our main result: Even with a finite-state HMM control, generically a CQS produces a measured qubit process whose minimal optimal predictor requires an infinite number of causal states. [1] Prediction resources (C_μ) diverge; though at a quantifiable rate. We establish the

[1] Genericity here refers to any HMM that is ergodic, aperiodic, and nonnegative.

result constructively, by determining h_μ and exploring C_μ's divergence for qubit processes and by identifying the driving mechanism as *measurement-induced nonunifilarity*.

7.2. Quantum Formalism

We introduce qubit information sources observed through quantum measurement, as depicted in Fig. 7.1. Noting that their outputs—what an experimentalist sees—are classical stochastic processes, we note that the classical measures of complexity discussed at length in this dissertation are appropriate to this kind of qubit process, and for identifying the mechanism in quantum measurement that generates the observed complexity. We concern ourselves with qubit sources that generate single qubits at a time, putting aside entangled pairs for now, and we ask the source to determine which qubit property, out of a finite set, is generated at each time. We imagine that the on-demand source is used repeatedly, producing an arbitrarily-long time series, which we call a *qubit process*.

The quantum evolution of a qubit is typically considered to occur in the Hilbert space H_2 which contains the one-parameter (time) family of states $\rho(t)$. While this representation is appropriate for many problems, it does not accurately describe the time series of qubits that concern us. In our time series, a different qubit is emitted at each time step t. The time parameter t is discrete and labels the qubit state ρ_t emitted at time t. A different Hilbert space H_2^t contains the state of each distinct qubit—the state of the qubit emitted at time t belongs to the Hilbert space: $\rho_t \in H_2^t$. The state of the entire bi-infinite time series lies in the Hilbert space is $\mathcal{H} = \lim_{\ell\to\infty} \bigotimes_{t=-\ell}^{+\ell} H_2^t$. And, when considering a finite part of the time series of length ℓ, the Hilbert space of interest is a truncation of $\mathcal{H}$ denoted $\mathcal{H}_{t:t+l} = \bigotimes_{k=t}^{t+l-1} H_2^k$.

In our setting, a *classically controlled qubit source* (cCQS) emits a qubit at each time step. The cCQS also determines state ρ_t of each output qubit. The latter is taken to be a pure state and it remains constant until measured. The qubit chain's state then is $\ldots \rho_{t-1} \otimes \rho_t \otimes \rho_{t+1} \otimes \ldots$. These restrictions guarantee that there is no temporal entanglement and that one can apply a single-qubit projective measurement E to each output qubit without affecting the states of the other qubits in the time series.

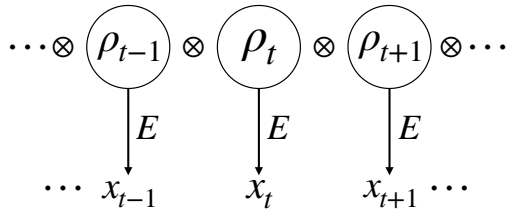


FIGURE 7.2. A controlled qubit source emits a discrete-time stochastic process of qubits in pure states ρ_t. Applying measurement E to the qubit emitted at time t in state ρ_t yields outcome x_t. Measuring each qubit at each time step yields a classical stochastic process.

Generating the qubits in a time series is governed by a cCQS that, without loss of generality, we take to be an ϵ-machine for which the symbols emitted during state-to-state transitions consist of qubit-states. This choice ensures that the source's internal complexity used in generating the qubit process can be quantified. Both the entropy rate h_μ^g and the statistical complexity C_μ^g of the controller can be exactly computed from the cCQS. We restrict the emitted qubits to be pure-state density matrices; that is, $\rho^2 = \rho$. This limits the type of correlations that can be present across the qubit time series to classical correlations and leaves time series with temporal entanglement for future exploration. Since each qubit ρ_t is in a pure state, the quantum state of the random variable chain that forms the time series can be regarded as the tensor product of the individual qubits: $\ldots \rho_{t-2} \otimes \rho_{t-1} \otimes \rho_t \ldots$.

The states of the qubits output by a cCQS form a stochastic process; two examples of the latter are shown in Figs. 7.3 (a) and (b). The cCQS in (a) generates a qubit time series of orthogonal pure states $|0\rangle\langle 0|$ and $|1\rangle\langle 1|$. The cCQS in (b) generates a qubit time series of nonorthogonal pure states $|0\rangle\langle 0|$ and $|+\rangle\langle +|$, where $|+\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$.

7.3. Measured Qubit Processes

The observer interacts with such processes by applying to each qubit a projective measurement as depicted in Fig. 7.2, consisting of the set of orthonormal measurement operators $\{E_0, E_1\}$ with measurement basis $E_0 = |\psi_0\rangle\langle\psi_0|$ and $E_1 = |\psi_1\rangle\langle\psi_1|$ parametrized by the Bloch angles θ and ϕ via:

$$|\psi_0\rangle = \cos\frac{\theta}{2}|0\rangle + e^{i\phi}\sin\frac{\theta}{2}|1\rangle \text{ and} \tag{7.1a}$$

$$|\psi_1\rangle = \sin\frac{\theta}{2}|0\rangle - e^{i\phi}\cos\frac{\theta}{2}|1\rangle \ . \tag{7.1b}$$

The outcome of each measurement can then be labeled 0 or 1, respectively, resulting in a binary classical stochastic process.

Knowledge of the controller and the measurement basis allows us to directly construct an HMM that generates the measured qubit process itself. We call this the *measured* cCQS, it has the same states and stationary distribution π as the original cCQS. Its labeled transition matrices $\{T^{(x)}\}$ with $x \in \mathcal{A}$ are:

$$T^{(x)} = \sum_{\rho_j} \mathbb{T}^{\rho_j} \Pr(x|\rho_j) \ , \tag{7.2}$$

where $\Pr(x|\rho_j) = tr(E_x \rho_j E_x^\dagger)$ and the cCQS labeled transition matrices $\mathbb{T}^{\rho_j}$ are defined in Sec. 2. A key step is that one can determine the HMM of the measured process by composing the measurement operator with the the qubit controller HMM that governs the cCQS.

See the HMM in Fig. 7.3 (c). It generates the classical process resulting from measuring the qubit process generated by Fig. 7.3 (b) with angles $\phi = 0$ and $\theta = \pi/2$. Note that the measured cCQS is particularly convenient as we no longer have to simulate the original CQS to generate qubit sequences and then apply measurement operators some, presumably large, number of times to each qubit to produce a large ensemble of output sequences. That is, the measured cCQS directly generates the classical output process.

7.4. Uncountable Predictive Features

This simple setup raises several natural questions about characterizing qubit processes generated by CQSs. How random is the qubit process? How much memory does the source use to generate the qubit series? Can we identify the internal control mechanism from the qubit time series alone?

One would hope that, since here we know the measured cCQS, as shown in Fig.7.3(c) for the example there, and it generates the measured qubit process, we could apply Eqs. (2.12)-(2.11) to

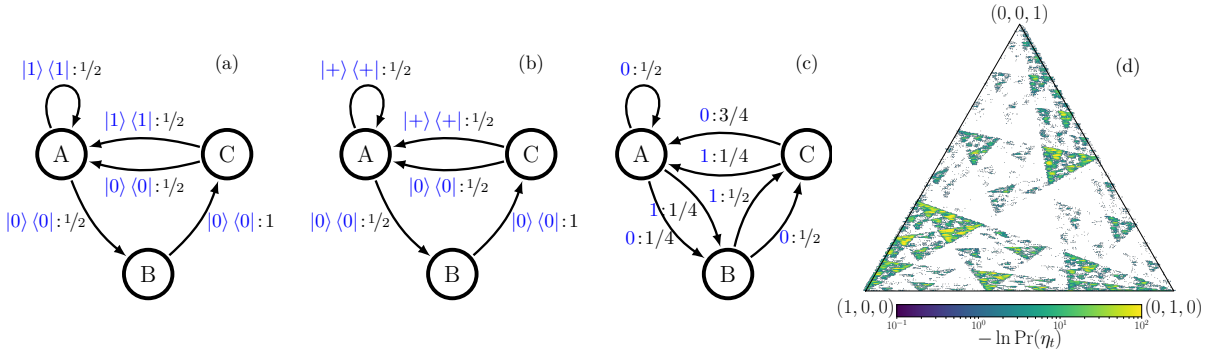


FIGURE 7.3. (a) Three-state classically-controlled qubit source (cCQS) that generates a process consisting of qubits in orthogonal states. (i) controller states $\boldsymbol{\mathcal{S}} = \{A, B, C\}$; (ii) orthogonal qubit alphabet $\mathcal{A}_Q = \{\rho_0 = |0\rangle\langle 0|, \rho_1 = |1\rangle\langle 1|\}$; (iii) labeled transition matrices $\mathbb{T}^{\rho_0}$ and $\mathbb{T}^{\rho_1}$, whose state-transition probability components $(\mathbb{T}^{\rho_k})_{ij}$ can be read from the diagram; and (iv) stationary state distribution $\pi = (1/2 \quad 1/4 \quad 1/4)$. If the controller is in state A it has equal probabilities of staying in that state or transitioning to state B. If it transitions to state B, then the system emits a qubit in pure state $|0\rangle\langle 0|$. In the next time step the machine must transition to state C and emits another $|0\rangle\langle 0|$ qubit. Then, in state C it transitions back to state A, emitting either pure state $|0\rangle\langle 0|$ or $|1\rangle\langle 1|$ with equal probability. As the cCQS runs, a time series of orthogonal qubits is generated. (b) Nonorthogonal-qubit cCQS: Three-state HMM that outputs nonorthogonal pure states $|0\rangle\langle 0|$ and $|+\rangle\langle +|$, where $|+\rangle = (|0\rangle + |1\rangle)/\sqrt{2}$. (c) HMM presentation for the classical stochastic process resulting from measurement ($\theta = \pi/2$) of the quantum process generated by (b). (d) Mixed states for the stochastic process generated by (c) in that HMM's state distribution simplex. Each mixed state is a point of the form (p_A, p_B, p_C) with probabilities of being in state A, B, or C of (c). The color scale shows the logarithm of the probability of the mixed states.

calculate our measures of randomness and memory directly from that model. Unfortunately, a problem arises. The measured cCQS is not an ϵ-machine since the generated measurement sequences are not in one-to-one correspondence with the internal state sequences. This is the problem of measured-cCQS nonunifilarity and it stymies any attempt to directly calculate the randomness and memory of the measured quantum process. In fact, and this is the first part of our result, nonunifilarity is generic to measured cCQSs, since randomly sampled HMMs are nonunifilar.

Though Fig. 7.3(c)'s HMM generates the observed qubit process, we cannot use it to directly determine even the most basic process properties. Fortunately, this measured cCQS can be converted to an ϵ-machine by calculating the cCQS's MSP, as discussed in Chapter 3.

Measurement-induced nonunifilarity results in the number of causal states diverging. That is, despite the controller having only a finite number of states and the controlling cCQS being unifilar, measurement means that predicting the observed process requires an uncountable number of causal states. In turn, to calculate the entropy rate h_μ we must make use of the tools developed in Chapter 4. To characterize the structural complexity of the measured process, we calculate the statistical complexity dimension.

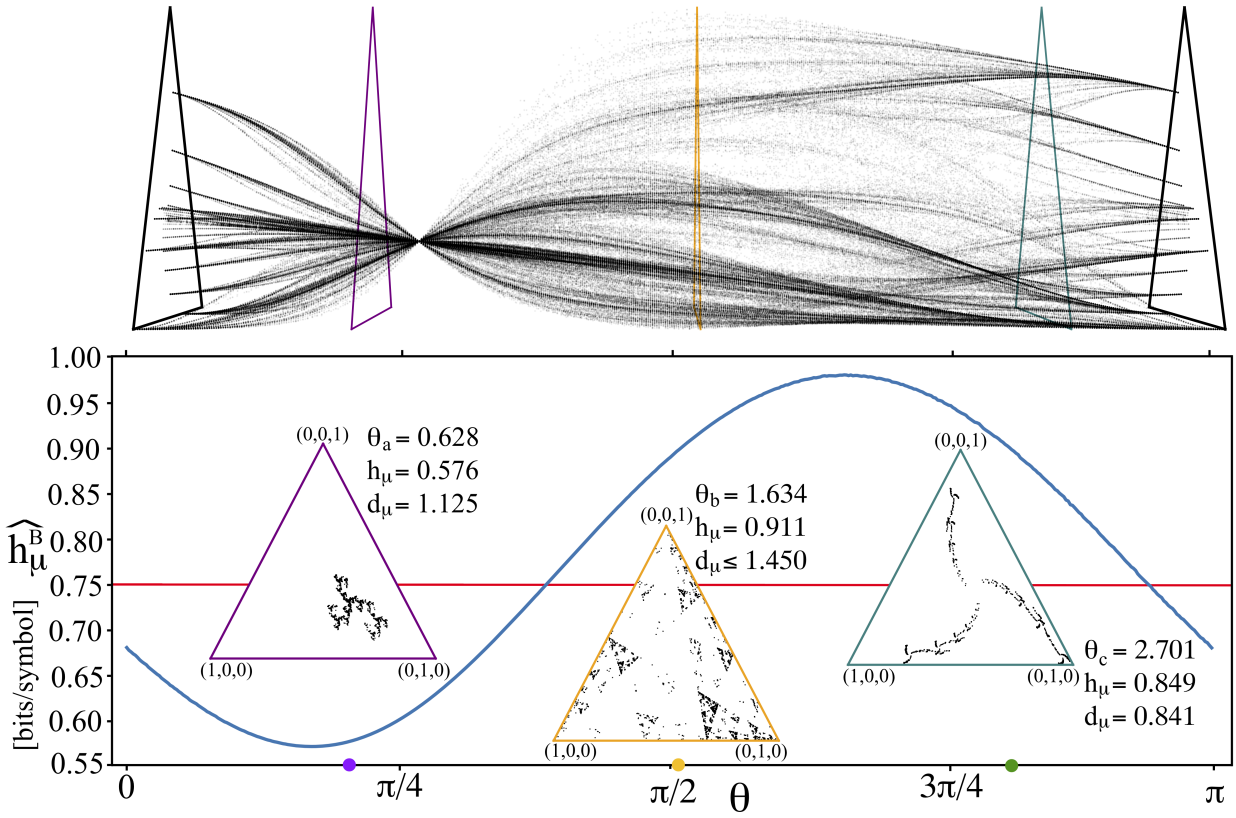


FIGURE 7.4. Measurement-induced randomness and structure: (Top) Mixed state sets when measuring the qubit process of Fig. 7.3(b) as a function of measurement angle $\theta \in [0, \pi]$. (Bottom) Entropy rate $\widehat{h_\mu^B}$ (blue curve) as a function of angle. Horizontal line (red) is the entropy rate of the (unmeasured) qubit sequences: $h_\mu^q = 3/4$ bit per output qubit. (Insets) Mixed states three measurement angles: (a) $\theta_a = 0.628$ (purple), (b) $\theta_b = 1.634$ (orange), and (c) $\theta_c = 2.701$ (green). The measured process entropy rates h_μ and statistical complexity dimensions d_μ given there. Both mixed states and complexity measures computed with $\ell = 10^6$ iterates.

7.5. Measurement Dependence

Equations (7.1) and (7.2) indicate that the choice of measurement basis alters the observed process. This, in turn, implies that the process entropy rate and statistical complexity dimension also depend on measurement. To explore this with an example, we calculate the dependence of the above complexity measures as a function of measurement angle θ, with fixed ϕ for Fig. 7.3(b)'s cCQS, determining the measured cCQS at each measurement setting.

Figure 7.4 (top) shows the results. The cCQS is measured in 500 different bases, holding $\phi = 0$ fixed and varying $\theta \in [0, \pi]$ uniformly. For each measured cCQS the MSP is computed and the resulting series of mixed-state sets is plotted. Figure 7.4 (bottom) plots $\widehat{h_\mu^B}(\theta)$ and highlights three particular measurement angles $\{\theta_a, \theta_b, \theta_c\}$, showing the unique attractor found in latter's mixed-state simplices. MSP entropy rate and the statistical complexity dimension are estimated using Eqs. (4.4) and (5.6), respectively.

Common characteristics are apparent, such as a smooth behavior of $h_\mu(\theta)$ with well-defined maxima and minima and the systematic change in the MSP structure as a function of θ which is consistent with the quoted dimensions d_μ. Angles $\theta = 0$ and $\theta = \pi$ give particularly simple behaviors with finite statistical complexity and $d_\mu = 0$, in accord with the countable MSPs there. The measured machines at these two values of θ are identical, aside from a symbol swap—all 0's become 1's and vice versa. They both have $C_\mu = 0.6813$ bits.

Figure 7.4 (top) exhibits a case of interest at $\theta = \pi/4$. The mixed states converge to a single point: a single-state machine that represents a biased coin. This occurs since the underlying cCQS has a binary quantum alphabet $\mathcal{A}_Q = \{\rho_0, \rho_+\}$ and the measurement basis corresponding to $\phi = 0$ and $\theta = \frac{\pi}{4}$ with basis vectors $|\psi_0\rangle$ and $|\psi_1\rangle$ is such that $\Pr(0|\rho_0) = \Pr(0|\rho_+)$ and $\Pr(1|\rho_0) = Pr(1|\rho_+)$. This basis is equidistant from both quantum states in $\mathcal{A}_Q$. Therefore, applying the measurement to one state or the other yields the same probability distribution over outcomes. One loses all information about the underlying structure and the measured cCQS generates an *independent identically distributed* process.

To compare the randomness and organization of the underlying generator process, the horizontal line in the $\widehat{h_\mu^B}(\theta)$ plot gives the entropy rate of the (unmeasured) qubit process: $h_\mu^g = 3/4$ bit per

output qubit. Its statistical complexity is $C_\mu^g = \mathrm{H}[\pi] = 1.5$ bits. The differences between these constant values and those of the measured cCQS values makes it clear that quantum measurement can both add or remove randomness and structure. A measurement close to $\theta = \frac{\pi}{4}$, on the one hand, rarely distinguishes between the measured quantum states ($|0\rangle\langle 0|$ or $|+\rangle\langle +|$) and has a greater number of measured 0s, making the measured process less random than the generating process. On the other hand, a measurement close to $\theta = 3\pi/4$ results in a more even distribution of measurement outcomes, introducing randomness to the measured process.

7.6. Implications and Future Directions

That randomness and complexity arise when observing qubit processes can be too facilely appreciated. Indeed, quantum measurement often comes steeped in mystery. We dispelled some of that mystery by showing that (i) an infinite number of predictive features are required to describe measured qubit time series and (ii) measurement can both introduce and subtract information and correlation. These characters of measurement greatly complicate learning about the informational and dynamical organization of quantum systems. However, at least now, we can appreciate more fully what the task is, what mechanism drives it (nonunifilarity), and why it is challenging.

Unexpectedly, analyzing the quantum physics necessitated novel theory and efficient algorithms for quantifying the randomness and complexity of ergodic, stationary processes generated by nonunifilar hidden Markov models. Mathematically, these gave a constructive answer to the longstanding information-theoretic problem of characterizing functions of Markov chains—a problem that until now had only been formally, not constructively, solved [**54**].

Solving the problem of measurement-induced complexity is a first step to fully describing quantum systems in terms of measurements. The results shed light on the fundamental ways in which the measurement act influences the observed complexity of quantum systems. With the newly-introduced tools, on the one hand, the next steps to reduce observed complexity easily come to mind: using POVMs, implementing multi-qubit measurements, and developing adaptive measurement schemes. On the other hand, introducing quantum controllers will bring results that bear directly on contemporary experimental systems, such as single-photon sources.

The developments here complement recent progress on representing classical processes via quantum channels, which showed that quantum representations can be markedly smaller than classical [109, 110, 111] and that classical and quantum physics are at odds when it comes to measures of organization [112, 113]. In a similar, complementary way, we hope that our results aid in developing a systematic description of memoryful quantum processes built on experimentally accessible quantities [97].

Future directions are best appreciated in light of the work's intentions. Though ostensibly about the complexity of quantum dynamics, the development here actually served three purposes. First, most directly it answered questions about measured quantum-state time series, demonstrating that measurement renders them more or less random and more or less structured than the internal controller dynamics. It identified the nonunifilarity mechanism that gives rise to those behaviors. Second, it announced innovations from dynamical systems and ergodic theory and introduced algorithms for processes generated by general HMMs. Specifically, it showed how to employ mixed state presentations to define entropy rate and statistical complexity dimension for processes admitting nonunifilar presentations. Third, it introduced a relatively simple setting that forms the basis for analyzing a family of increasingly more sophisticated and physically-realistic qubit-process generators.

The contributions immediately suggest several relatively straightforward pathways to future explorations—explorations that will have even broader impact. The first relates back to the task of reconstructing the hidden qubit controller from the observed process. Clearly, the results on randomness and structure say this is a challenging problem. A second avenue is to employ the full analytic power afforded by spectral decompositions arising from the meromorphic functional calculus [114]. Likely, this will be necessary for quantally-controlled CQSs. These tools were simply not engaged here. Nor, despite noting that statistical complexity is a memory resource, we did not engage resource theory fully. This is a third avenue for further exploration. A fourth is to explore the novel correlations and structure in temporally entangled qubit processes. For example, can the internal controller remember enough of its past and operate on it so that the generated qubit time series is entangled?

In addition, there is a need to shift the setting towards more physical realism and experimental relevance. The first task along these lines, just alluded to, would extend the current setting to quantally controlled qubit sources (qCQSs)—a fully quantum mechanical qubit generator—and to the quantum communication channel setting in which qubits are input, quantally processed, then output. Advances here will more directly impact information processing and computing performed by quantum dynamical systems. A second extension is to address qubit source timing issues, moving away from the admittedly simple use of discrete time here. Fortunately, the cCQS model can easily be extended to discrete-event continuous-time hidden semi-Markov models [**115**]. This immediately will give measures of randomness and structure, paralleling the development here. Finally, the projective measurements used here should be replaced with positive-operator valued measurements (POVM). Results on continuous-time, quantum-generated process measured via POVMs will have broad applications. And, these results will naturally complement recent progress on representing classical processes via quantum channels, which showed that quantum representations can be markedly smaller than classical [**109**, **110**, **111**] and that classical and quantum physics are at odds when it comes to measures of organization [**112**].,

CHAPTER 8

Application: Functional Thermodynamics of Maxwellian Ratchets

In 1867, James Clerk Maxwell introduced a thought experiment designed to challenge the Second Law of Thermodynamics [116,117]; what Lord Kelvin later came to call "Maxwell's Demon". Exploiting the fact that the Second Law holds only on average—i.e., the thermodynamic entropy $\langle S \rangle$ cannot decrease over repeated transformations—the experiment conjured an imaginary, intelligent being capable of detecting and then harvesting negative entropy fluctuations to do work. The paradox that Maxwell put forward is that by using its "intelligence" this being apparently violates the Second Law of Thermodynamics. Maxwell's challenge was the first indication that the Second Law must take into account information processing.

The puzzle's solution came from recognizing that the "very observant" and "neat-fingered" Demon must manipulate memory to perform its detection and control task and, critically, that such information processing comes at a cost [118, 119]. To operate, the Demon's intelligence has thermodynamic consequences. This is summarized by Landauer's Principle: "any logically irreversible manipulation of information ... must be accompanied by a corresponding entropy increase in non-information-bearing degrees of freedom of the information-processing apparatus or its environment" [120]. This recasts the Demon as a type of engine—an *information engine* that uses correlations in an *information reservoir* to leverage thermodynamic fluctuations in a *heat reservoir* to do useful work.This class of information engines—Maxwellian demons and their generalized ratchets—has been subject to extensive study [121, 122, 123, 124]. However, previous determinations of their thermodynamic functionality were stymied by the difficulty of accurately calculating the entropic change in what Landauer identified as the system's "information-bearing degrees of freedom".

8.1. Motivation

Consider a Maxwellian ratchet designed to read an infinite input tape, perform a computation and thermodynamic transformation, and write to an infinite output tape, as depicted in Fig. 8.1. The relevant entropic change then is quantified by the difference in the Kolmogorov-Sinai entropies of the inputs to the ratchet (h_μ) and of the outputs to the information reservoir (h'_μ) [**122**]. However, this calculation ranges from very difficult to intractable when the processes generating the input and output information have temporal correlations. And, more troubling, this problem is generic—when driving a ratchet with uncorrelated input, even simple finite-state memoryful ratchets produce output processes with temporal correlations. Fundamental progress was halted since determining thermodynamic functionality in the most general case—temporally correlated input driving a memoryful ratchet—was intractable. Attempts to circumvent these problems either heavily restricted thermodynamic-controller architecture [**122**], invoked approximations that misclassified thermodynamic functioning, or flatly violated the Second Law [**121**]. It appears that—and this is one practical consequence of the results reported in the following—a number of recent analyses of information-engine efficiency and functioning must be revisited and corrected. Our contribution is that the latter is now possible.

Re-examining a well-known information ratchet, we apply newly discovered techniques to accurately measure the Kolmogorov-Sinai-Shannon entropy of temporally correlated processes in general [**48**, **51**]. We show that, via the Information Processing Second Law [**122**], this allows accurate determination of the functional thermodynamics of arbitrary finite-state ratchets. The net result is a shift in perspective. To guarantee that the output information could be studied analytically, previous successful efforts designed ratchet *structure*—the states and transitions—in accord with a given input's correlational structure [**123**]. One consequence is that follow-on efforts adopted a fixed input-output-centric view of information engines. Here, following the example of the earliest discussions of information ratchets [**121**], the new methods shift the focus back to the engine itself, setting its design and then exploring all possible input-dependent thermodynamic functionalities.

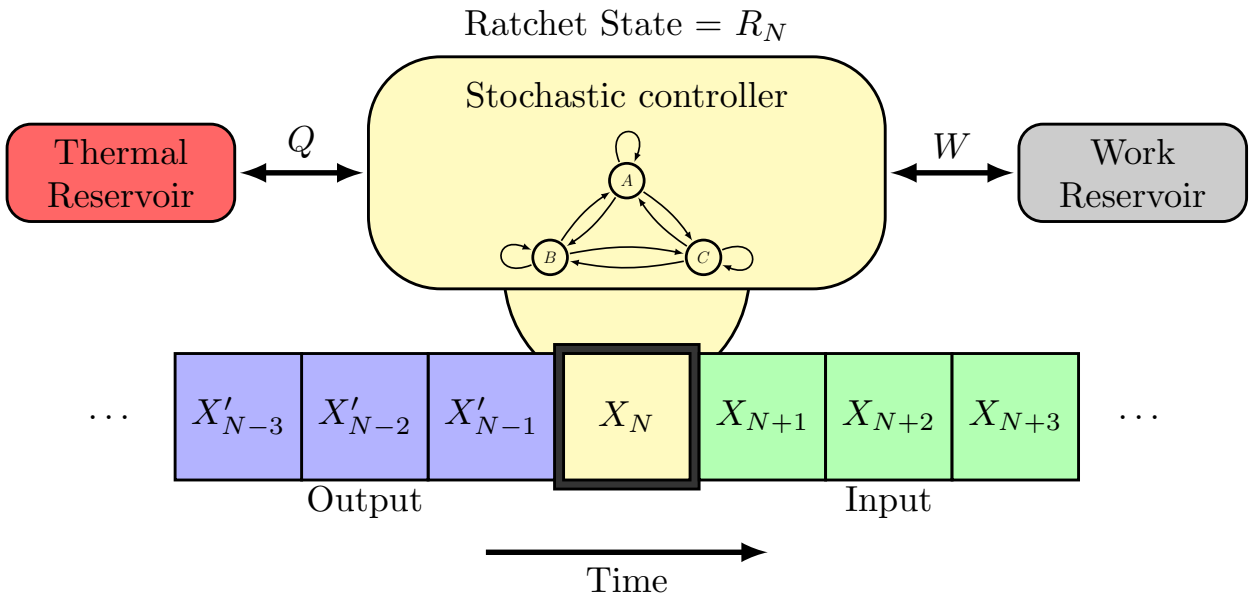


FIGURE 8.1. Information engine as a finite-state ratchet (controller) connected to a thermal reservoir, a work reservoir, and an information reservoir (depicted as tape whose storage cells may be read or written).

8.2. Information Engines

The information engines of interest consist of a finite-state stochastic controller or *ratchet* that interacts with a *thermal reservoir*, a *work reservoir*, and an *information reservoir*. These are connected as shown in Fig. 8.1 and are embedded in a thermal environment at constant temperature T. The information reservoir takes the form of an *input tape*, which stores a binary-symbol string. Its state is described by the random variable $X_{0:\infty} = X_0 X_1 \ldots$. We restrict to binary input and output alphabets, so that each X_N realizes an element $x_n \in \mathcal{X} = \{0, 1\}$. The ratchet operates in continuous time; the controller state at time $t = N\tau$ is represented by the random variable R_N, which realizes an element $r \in \mathcal{R}$—the ratchet's discrete, finite state space.

At each step, X_N couples to the ratchet controller for an interaction of duration τ. During this time, thermal fluctuations continuously drive transitions in the coupled state space $\mathcal{R} \times \mathcal{X}$ of the ratchet and the current tape symbol. After the interaction interval, the ratchet is in a potentially different state R_{N+1}, and the symbol X_N has been transduced into an output symbol $X'_N = x'_N \in \mathcal{X}$, which is written to the tape. The strings of possible output symbols are expressed by the random variable $X'_{0:\infty} = X'_0 X'_1 \ldots$. The tape moves forward, and the next input symbol X_{N+1} begins its

interaction with the ratchet, which now starts in state R_{N+1}. The joint transitions between states of the ratchet and symbol have energetic consequences, capturing energy flows between the thermal and work reservoirs.

8.2.1. Energetics. These information engines are autonomous and transitions in the coupled ratchet-symbol system are driven by fluctuations in the thermal reservoir. Recently, Ref. [**123**] introduced a general formalism for determining the energetics of such information engines. Under detailed balance, transitions over the joint ratchet-symbol state space $\mathcal{R} \times \mathcal{X}$ are described by a Markov chain M, where every transition with positive probability—denoted:

$$M_{r_N \otimes x_N \to r_{N+1} \otimes x'_N} = \Pr(R_{N+1} = r_{N+1}, X'_N = x'_N | R_N = r_N, X_N = x_N)$$

—must have a reverse transition with positive probability. Energy changes associated with an internal-state transition are then determined by the forward-reverse transition probability ratio:

$$\Delta E_{r_N \otimes x_N \to r_{N+1} \otimes x'_N} = k_{\mathrm{B}} T \ln \frac{M_{r_{N+1} \otimes x'_N \to r_N \otimes x_N}}{M_{r_N \otimes x_N \to r_{N+1} \otimes x'_N}} \ .$$

Assuming that all energy exchanges with the heat reservoir occur during the ratchet-symbol interaction interval τ and that all energy exchanges with the work reservoir occur between interaction intervals, the average asymptotic work is:

$$\langle W \rangle = \sum_{r,r' \in \mathcal{R}, x, x' \in \mathcal{X}} \pi_{r \otimes x} M_{r \otimes x \to r' \otimes x'} \Delta E_{r \otimes x \to r' \otimes x'} \ , \tag{8.1}$$

where $\pi_{r \otimes x}$ is the asymptotic distribution over the joint state of the ratchet-symbol system at the beginning of an interaction interval.

8.2.2. Structure. To discuss the computational *structure* of information engines, we first cast the input and output strings in terms of the hidden Markov Models (HMMs) that read and generate them. This representation allows us to consider the internal states $\mathcal{S}$ of the *input machine* as well as the internal states $\mathcal{S}'$ of the *output machine*. The latter are the joint states of the input process and the ratchet: $\mathcal{S}' = \mathcal{S} \times \mathcal{R}$.

When the string of inputs or outputs can be generated by an HMM with only a single internal state, they are *memoryless*, since they can store no information from the past. The random variables generated by the associated HMM are *independent and identically distributed* (IID). When there is more than a single state, in contrast, the associated process is *memoryful* and the random variables generated may be correlated in time.

Similarly, we cast the ratchet controller as a *transducer* that maps from input sequences to distributions over output sequences.

DEFINITION 10. *A finite-state edge-labeled* transducer *consists of:*

(1) A finite set of states $\mathcal{R} = \{R_1, ..., R_{N'}\}$,

(2) A finite input alphabet $\mathcal{A}$ *of* k *symbols* $x \in \mathcal{A}$,

(3) A finite output alphabet $\mathcal{A}'$ *of* k' *symbols* $x' \in \mathcal{A}'$, *and,*

(4) A set of N' *by* N' *input-output symbol-labeled transition matrices* $T^{(x,x')}$, $x, x' \in \mathcal{A} \times \mathcal{A}'$*:*

$$T_{ij}^{(x,x')} = \Pr(r_j, x' | r_i, x) \ .$$

The transducer formulation allows us to calculate the output HMM in terms of the ratchet and the input machine. The exact method is given in Appendix H.0.1. As with the input machine, a ratchet is memoryless when it possesses only one internal state, and memoryful otherwise. A key feature of a ratchet we focus on here is its ability to alter temporal correlations by altering the structure of an input process. If memoryless (IID) input is fed to a memoryful ratchet, generally the output will be memoryful, since this guarantees the state space dimension of the output $|\mathcal{R}'| = |\mathcal{S} \times \mathcal{R}| > 1$. Figure 8.2 graphically illustrates the composition of various input process HMMs with the ratchet transducer we analyze in detail shortly.

8.2.3. Informatics. Following Landauer, extensions of the Second Law of Thermodynamics were proposed to bound the thermodynamic costs of information processing by an information engine. Reference [**121**] employed a bound that compares the *Shannon entropy* of single input and single output symbols. Comparing the single-symbol Shannon entropy in the input string to that in the output string quantifies how the ratchet transforms randomness in individual symbols. This

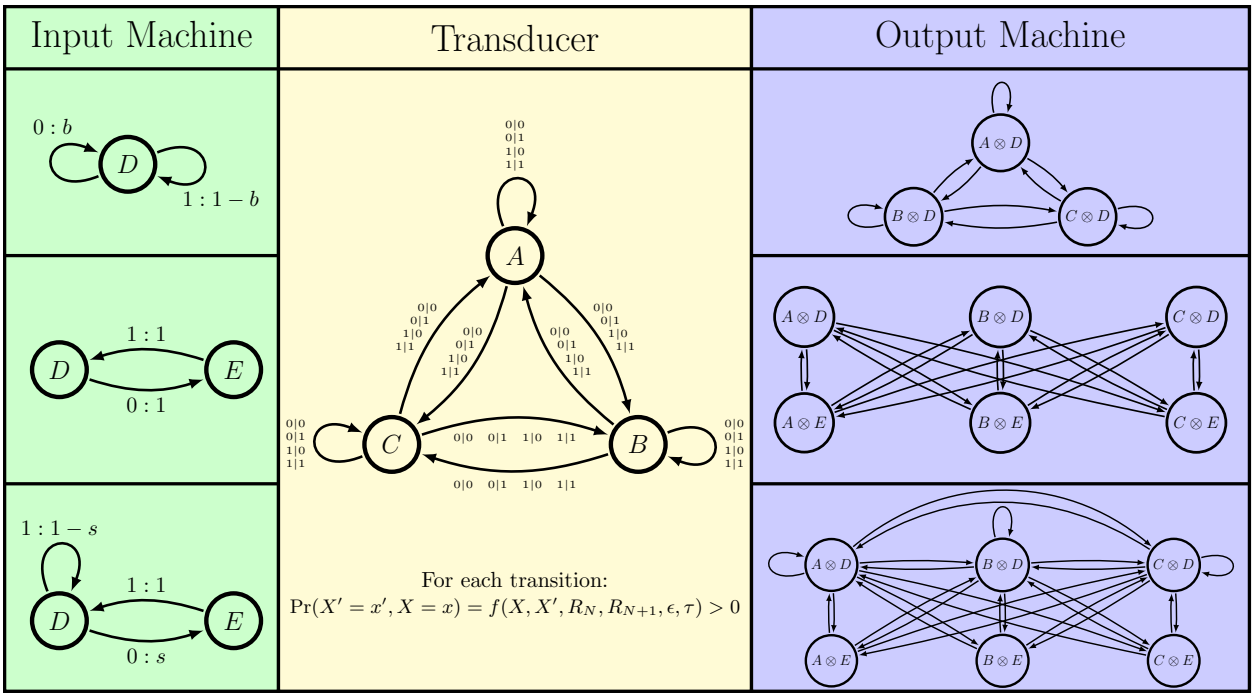


FIGURE 8.2. Composing the Mandal-Jarzynski transducer (center, yellow) with a Hidden Markov model (left, green) that describes the process on the input tape gives an output Hidden Markov model (right, purple) that describes the process written to the output tape. Hidden Markov model (HMM) states R are depicted as circles. Directed edges between states represent possible transitions on an observed symbol x. HMM edges are labeled $x : \Pr(x', R_{N+1}|R_N)$. Transducers are similarly depicted by circular states with directed edges representing possible transitions on pairs of input symbols x and output symbols x'. Transducer transitions are specified by $x'|x : \Pr(x', R_{N+1}|x', R_N)$. Left: Input HMMs discussed here, from top to bottom, a (memoryless) Biased Coin, a Period-2 Process, and the Golden Mean Process. Center: The Mandal-Jarzynski ratchet, represented by a three-state transducer. Probabilities are not shown on edge labels for conciseness, but are nonzero for all transitions and all combinations of input-output symbol pairs (x, x'). Each edge probability is a function—denoted by $f(\ldots)$—of the previous state R_N, the next state R_{N+1}, the input symbol x, and the output symbol x'. See Section 8.3 and Appendix H for further details. Right: Output HMMs resulting in the Mandal-Jarzynski transducer composed with the corresponding input HMM on left. Edge labels are left off for conciseness, but each transition label represents a positive probability of observing a 0 or a 1.

difference captures one aspect of the ratchet's information processing. And, it was proposed as an upper bound on the asymptotic work done $\langle W \rangle$ [**121**]:

$$\langle W \rangle \stackrel{?}{\leq} k_{\mathrm{B}} T \Delta H_1 \ , \tag{8.2}$$

where $H_1 = H_1[X]$ is the entropy averaged over the input tape, $H'_1 = H_1[X']$ is that averaged over the output, and $\Delta H_1 = H'_1 - H_1$ is the change in the single-symbol statistics produced by the ratchet's operation.

Note, however, that while the H_1s track the average information in any single instance of X_t or X'_t, they do not account for temporal correlations within input sequences or within output sequences. This is key, as information ratchets change more than the statistical bias in an individual symbol, they alter temporal correlations in symbol strings. These altered correlations are related to the fact that the ratchet induces structural change in its input. Recognizing this is central to bounding the thermodynamic costs of the ratchet's interaction with the input process.

To properly address how correlations affect costs, we calculate a process' intrinsic randomness when all temporal correlations are taken into account, as measured by the entropy rate. Replacing the input and output Shannon entropies in Eq. (8.2) with their respective entropy rates gives the *Information Processing Second Law* (IPSL) [**122**]:

$$\begin{aligned} \langle W \rangle &\leq k_{\mathrm{B}} T \ln 2 \left(h'_\mu - h_\mu \right) \\ &= k_{\mathrm{B}} T \ln 2 \ \Delta h_\mu \ . \end{aligned} \tag{8.3}$$

The IPSL correctly expresses the upper bound on work, taking into account the presence of temporal correlations in input and output processes.

The importance of Eq. (8.3) cannot be overstated—any memoryful ratchet induces temporal correlations in its output, even for IID input. Using Eq. (8.2) in the IID case typically overestimates the upper limit on available work. Additionally, temporal correlations in the input are known to be a thermodynamic resource [**125**]. In fact, suitably designed ratchets can leverage such correlations to do useful work. Thus, inappropriately applying Eq. (8.2) in these cases often results in claims that violate the Second Law. In short, Eq. (8.3) generalizes Landauer's Principle to the case of correlated environments and finite-state memoryful ratchets that generate correlated outputs.

8.3. Mandal-Jarzynski Information Ratchet

To demonstrate the descriptive power of these dynamical-thermodynamic results on ratchet entropy, dimension, mixed states, and function, we apply them to a well-known example of an

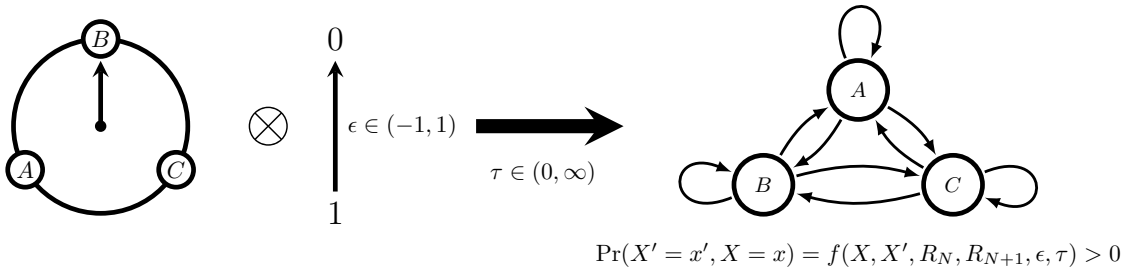


FIGURE 8.3. Ratchet schematic adapted from the original Mandal-Jarzynski construction, showing how the dial-and-symbol system is transformed into a three state transducer upon selection of a specific ϵ—determining the energetics of flipping a bit—and τ—determining the interaction interval. For almost every value of ϵ and τ every state-to-state transition has positive probability for every input-output symbol combination.

information engine—the Mandal-Jarzynski ratchet [**121**]; hereafter, *the ratchet*. Although initially introduced without reference to HMMs and transducers, following Ref. [**122**] we translate the original ratchet model into the HMM-transducer formalism outlined in Section 8.2. In these terms, the ratchet is a three-state, fully connected transducer, designed such that only transitions that flip an incoming symbol are energetically consequential. As shown in Fig. 8.3, the ratchet's transition probabilities are parametrized by $\tau \in [0, \infty)$—duration of the ratchet-symbol interaction—and $\epsilon \in (-1, 1)$—the *weight parameter*. For a given τ and ϵ, the Mandal-Jarzynski model may be written down as the three-state transducer shown in the center column Fig. 8.2. See Appendix H for how to calculate the transducer, which is based on a rate-transition matrix, and Appendix H.0.1 for the input-transducer composition method.

Any interaction interval in which the input symbol is unchanged is energetically neutral. Therefore, we measure the average work done by the ratchet by the difference in the probability of reading a 1 on the input tape cell versus writing a 1 to the output tape cell:

$$\langle W \rangle = k_{\mathrm{B}} T \; w \left(\Pr(X' = 1) - \Pr(X = 1) \right) \; ,$$

where $w(\epsilon) = \log((1 + \epsilon)/(1 - \epsilon))$. When $\epsilon = 0$, flips $0 \to 1$ and $1 \to 0$ are both energetically neutral; when $\epsilon \to \pm 1$, symbol flips in one direction are energetically favored over the other. Note that this computation finds the same asymptotic work production as Eq. (8.1); recalled here as an aid to intuition.

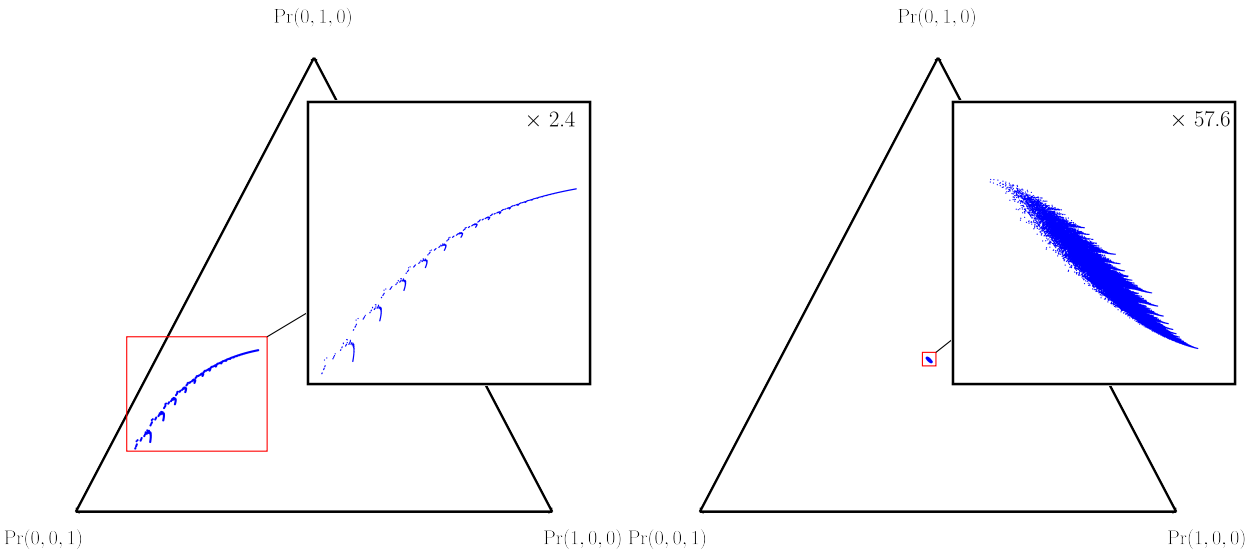


FIGURE 8.4. Mixed states $\eta \in \boldsymbol{\mathcal{R}}$ of the output process generated by the ratchet driven with memoryless input (Fig. 8.2(top row)) plotted on the 2-simplex. Corner labels give the mixed-state probability distributions $\eta = (\Pr(A \otimes D), \Pr(B \otimes D), \Pr(C \otimes D))$. Mixed states at the simplex corners correspond to the HMM being in exactly one of its states, while mixed states in the simplex interior are mixtures of the possible HMM states, with $\eta = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ lying at the center. (Left) Ratchet parameters $\delta = -0.98$, $\epsilon = 0.01$, and $\tau = 0.1$. (Right) Ratchet parameters $\delta = 0.4$, $\epsilon = 0.5$, and $\tau = 0.1$. Insets: Detail of the mixed-state sets, magnified by amount indicated in upper right corner.

Reference [**121**]'s initial analysis considered only uncorrelated inputs. That is, their input machine was a single-state HMM—a biased coin, with bias $\delta = \Pr(0) - \Pr(1)$. To identify their ratchet's thermodynamic functionality, the work bound was approximated via Eq. (8.2)—that is, assuming tape symbols were statistically independent. However, the ratchet is memoryful (due to its three internal states) and, therefore, in general induces correlations in its output, even for uncorrelated inputs. Since the single-symbol entropy only upper-bounds the true Shannon entropy rate—$h_\mu \leq H_1$—Eq. (8.2) is suspect when used to identify actual thermodynamic functioning. Using new results here, the following shows that, while approximately correct for uncorrelated input, the single-symbol entropy bound is violated for correlated input. Its incorrect use mischaracterizes thermodynamic functioning and can lead to violations of the Second Law.

In addition, our new methods give insight into how the ratchet processes structural information. Due to its inherent nonunifilarity, even when driven by a finite-state ϵ-machine, the ratchet produces

nonunifilar output machines that generate processes with an uncountably-infinite set of mixed states, as Fig. 8.4 shows. Moreover, the figure also demonstrates that as ratchet parameters vary the mixed-state sets have strikingly different structure.

Previous interpretations of ratchet thermodynamic functioning were limited to considering only transformations of randomness; i.e., for given ratchet parameters and input, what is the sign and magnitude of $k_B T \ln 2 \ \Delta h_\mu$ and how does this affect $\langle W \rangle$? Such questions ignore the key second dimension of information processing illustrated so vividly by Fig. 8.4. That is, given the same ratchet, parameters, and input, what is the sign and magnitude of ΔC_μ and Δd_μ? Does the ratchet construct new patterns in its output ($\Delta C_\mu > 0$ or $\Delta d_\mu > 0$) or deconstruct patterns passed to it from the input ($\Delta C_\mu < 0$ or $\Delta d_\mu < 0$)? How do these then affect $\langle W \rangle$? Answering structural questions requires a more thorough taxonomy of thermodynamic functionality than the original engine/dud/eraser categories.

8.4. Randomizing and Derandomizing Behaviors

The ratchet's previously-identified thermodynamic functions *engine*, *eraser*, and *dud* were identified by comparing the sign and magnitude of $k_B T \ln 2 \ \Delta h_\mu$ to the asymptotic work production. As such, there are three physically possible orderings:

- Engine: $0 < \langle W \rangle \leq k_B T \ln 2 \ \Delta h_\mu$;
- Eraser: $\langle W \rangle \leq k_B T \ln 2 \ \Delta h_\mu < 0$; and
- Dud: $\langle W \rangle \leq 0 \leq k_B T \ln 2 \ \Delta h_\mu$.

A ratchet randomizing inputs ($\Delta h_\mu > 0$) can operate as an engine, if it is leveraging the change in entropy rate to do useful work. It may also act as a dud, if the randomization produces no useful work or, worse, if the ratchet is using work. A ratchet derandomizing inputs ($\Delta h_\mu < 0$) is termed an "eraser" and can only derandomize up to $\langle W \rangle / k_B T \ln 2$ bits using $\langle W \rangle$ joules of work. The ordering $k_B T \ln 2 \ \Delta h_\mu < \langle W \rangle < 0$ would imply that the ratchet is derandomizing beyond the physical limitations of Landauer's principle.

As noted already, Ref. [**121**] originally identified these functionalities using the entropy-change approximation ΔH_1 rather than the exact change Δh_μ introduced by Ref. [**122**]. As previously shown, driving a memoryful ratchet with a memoryful input violates Eq. (8.2) [**125**]. In all other

cases, Eq. (8.2) is valid, but may mischaracterize the functional thermodynamic regimes. A natural question, therefore, is how much difference does using the correct entropy rate make in identifying function? To see this, we now compare Δh_μ and ΔH_1.

There are three possibilities. First, $\Delta H(1) = \Delta h_\mu$. In this case, a ratchet does not change the presence of temporal correlations. This occurs when a memoryless ratchet is driven by memoryless input.

Second, $\Delta H(1) > \Delta h_\mu$. Here, a ratchet reduces the presence of temporal correlations, which occurs when a memoryless ratchet has been driven by memoryful input. In this regime, the difference in single-symbol entropy is a tighter bound on the correlation change than the difference in entropy rate. Critical to this case, though, recall that our goal is not a tight bound, but rather an accurate measurement of the gap between information processing and asymptotic work. The upshot is that using Eq. (8.2) in this case may mischaracterize thermodynamic functionality.

Finally, $\Delta H(1) < \Delta h_\mu$, which occurs when a memoryful ratchet is driven by memoryless input. In this case, the ratchet increases temporal correlations in the output, so that the difference in entropy rates is a tighter bound on the asymptotic work production. This is the scenario in the first treatment of the Mandel-Jarzynski ratchet [**121**]. Note that when a memoryful ratchet is driven with memoryful input, the most generic case, all orderings of Δh_μ and $\Delta H(1)$ are possible.

Let's now turn to consider in detail how the ratchet operates in three distinct environments: Memoryless, periodic, and memoryful inputs. This gives more direct insight into the ratchet's transformational capabilities.

8.4.1. Memoryless Input. When the ratchet is driven with a memoryless input, as in the original analysis, Eq. (8.2) is valid, but IPSL always offers a tighter or equal bound on work production than the single-symbol entropy approximation. This holds since the input is memoryless, while the three-state output machine is memoryful and nonunifilar for almost every parameter setting. As such, one cannot calculate the entropy rate h'_μ in closed form. However, the new techniques above can determine the mixed-state presentations of the output HMMs and this gives accurate numerical calculation of both the single-symbol and the IPSL work bounds.

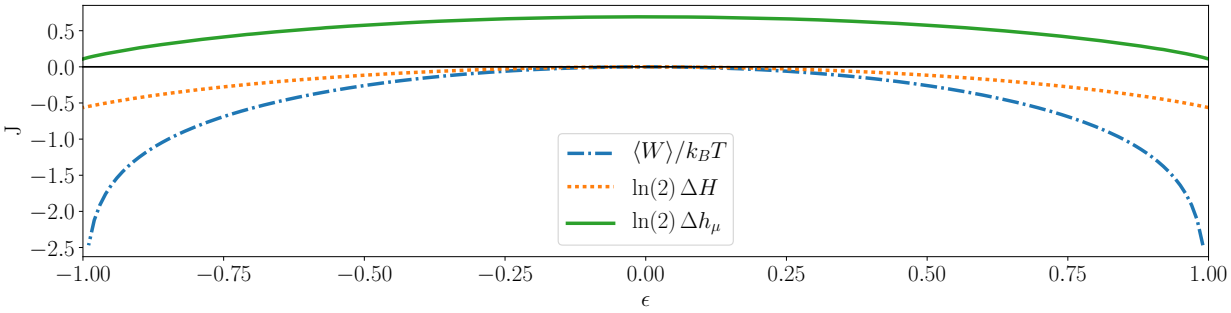


FIGURE 8.5. Asymptotic work production $\langle W \rangle$, single-symbol ΔH_1 bound, and Kolmogorov-Sinai-Shannon Δh_μ bound when ratchet is driven by period-2 memoryful input. Since the input has no parameters, the parameter sweeps only over ϵ with $\tau = 10$.

This all being said, for most parameter values of the Mandal-Jarzynski ratchet, in practice we find that $\Delta h_\mu \approx \Delta H_1$. In other words, when driven with a memoryless input, the ratchet's functional thermodynamic regions are not significantly changed when identified via the single-symbol entropy—a minor quantitative difference without a functional distinction. (See Fig. H.1 for a comparison of the functional thermodynamic regions found by each bound.) Exploring output-machine MSPs shows this arises from the ratchet's transition topology. As shown in the middle column of Fig. 8.2, the ratchet's transducer is fully connected, and all transitions to any other state on any combination of symbols are possible. Therefore, it is impossible to be certain about which state the ratchet is in; or, indeed, to even be sure which states the ratchet is *not* in. Graphically, this is represented by the fact that the output-machine mixed states η always lie deep in the simplex $\boldsymbol{\mathcal{R}}$'s interior, as illustrated in Fig. 8.4 (Right).

Mixed states lying at $\boldsymbol{\mathcal{R}}$'s center correspond to an equal belief in each of the output HMM's three states: $A \otimes D$, $B \otimes D$, and $C \otimes D$. While those on $\boldsymbol{\mathcal{R}}$'s border indicate certainty of *not* being in at least one state. Since the probability distribution over the next symbol is a continuous function over the mixed states, the diameter of the mixed-state set is a rough measure of the presence of temporal correlations in the ratchet's behavior. To explicitly illustrate this, the mixed states for two example output processes generated by the ratchet are shown in Fig. 8.4. On the left, the mixed states are spread out, indicating that at the selected parameters, the ratchet induces stronger temporal correlations than in the next example (right). There, all mixed states lie very

close together and very near the simplex center. The mixed-state set has very small diameter. For most parameter values, one finds that the mixed states of the memoryless-driven ratchet's output process cluster closely in the middle of the simplex. (See Fig. 8.8 for a broader survey of ratchet MSPs for memoryless input.) So, by giving insight into the mixed states of the output process, our new techniques rather directly explain why $\Delta h_\mu \approx \Delta H_1$ for this particular ratchet.

8.4.2. Periodic Input. Now, consider driving the ratchet with a periodic input. The *Period*-2 *Process*, shown in the middle row of Fig. 8.2, is memoryful, with two internal states. So, it now is possible that Eq. (8.2) is violated. Since $H_1 = 1$ and $h_\mu = 0$, the presence of temporal correlations in the input is maximized. Noting this, and the near-memoryless behavior of the ratchet as discussed in Section 8.4.1, we can see that for almost all parameters, the ratchet decreases the presence of temporal correlations in transforming the input process to the output. The periodically-driven ratchet output HMMs have six states and are nonunifilar for nearly all parameter values; see Fig. 8.2 (middle, last column). And so, we must calculate these machine's mixed-state presentations to estimate h'_μ. Comparing Eq. (8.2) and Eq. (8.3) in Fig. 8.5 to the asymptotic work production shows that Eq. (8.2) is not violated. As predicted above, it is a tighter bound on $\langle W \rangle$ than Eq. (8.3).

Although it may seem desirable to use the tighter bound, the single-symbol and entropy rate bounds identify the ratchet's thermodynamic functioning differently: Since $\langle W \rangle \leq k_B T \ln 2 \, \Delta H_1 \leq 0$ for all values of ϵ, the single-symbol entropy bound classifies the ratchet as an eraser, dissipating work to reduce the randomness in the input. However, when considering temporal correlations, we see that the ratchet is in plain fact a dud—$\Delta h_\mu > 0$. That is, the ratchet dissipates work while *increasing* the tape's intrinsic randomness. This marked mischaracterization of thermodynamic function by the single-symbol entropy highlights an important lesson: *Bounding the asymptotic work production as tightly as possible is not the same as correctly identifying the functional thermodynamics.* As Ref. [**126**] recently showed, rather than merely a bound, Eq. (8.3) is meaningful only when comparing $k_B T \ln 2 \, \Delta h_\mu$ to $\langle W \rangle$. The difference in the two quantifies the amount of work the ratchet can do, if it were an optimal, globally-integrated information processor. This shows that even when it may appear to outperform Eq. (8.3), in general Eq. (8.2) cannot serve as a reliable bound on asymptotic work production. We return to this in our final example, where applying Eq. (8.2) implies a violation the Second Law.

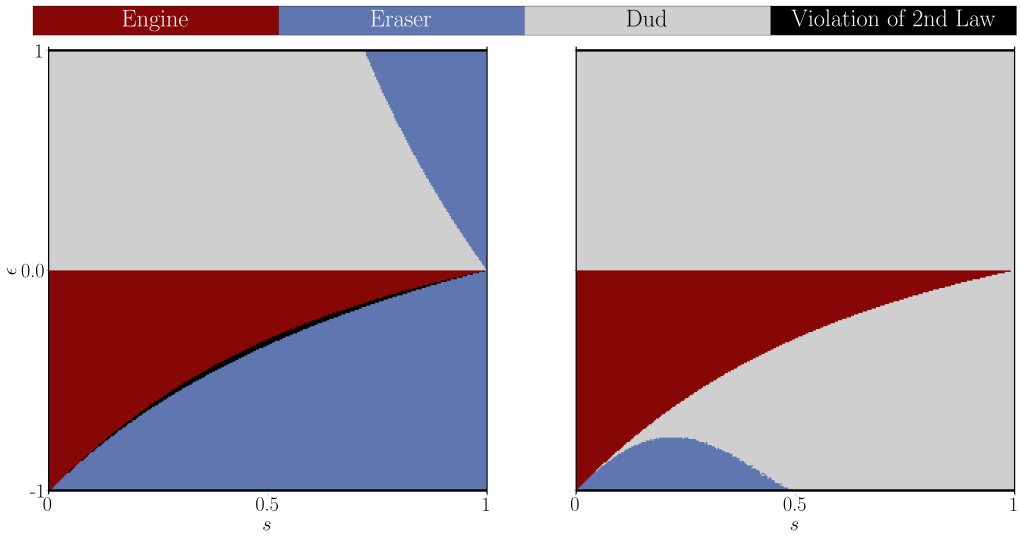


FIGURE 8.6. Functional thermodynamic regions of the ratchet driven with the Golden Mean Process as a function of parameters $\epsilon \in [-1, 1]$ and $s \in [-1, 1]$ with $\tau = 1$. (Left) Purported functionality identified via by single-symbol entropy bound Eq. (8.2). (Right) Correct functionality identified via the entropy-rate bound IPSL Eq. (8.3).

8.4.3. Memoryful Input. Finally, let's drive the ratchet with a mixed-complexity memoryful process—partly regular, partly stochastic—the *Golden Mean Process*. As depicted in Fig. 8.2 (bottom row), this two-state HMM generates a family of processes parametrized by $s \in [0, 1]$. When $s = 1$, the process is period-2. Decreasing s lets the process emit multiple 1s in a row. This increases in probability until at $s = 0$, where the process emits only 1s. The driven ratchet's output HMMs have six states and are nonunifilar for nearly all parameter values; see Fig. 8.2(bottom, last column). So, again, we must calculate mixed-state presentations to get h'_μ and identify functionality.

In Fig. 8.6, we apply both Eq. (8.2) and Eq. (8.3) for the same set of ratchet parameters. For both, we find asymmetry in the functional thermodynamic regions with respect to ϵ, in contrast to the highly symmetric regions found for memoryless input, shown in Fig. H.1. This is due to the asymmetry in input. In fact, it is not possible for the Golden Mean Process to produce strings biased towards 0. Thermodynamically, for $\epsilon > 0$, the ratchet is not able to extract work. When applying the single-symbol bound, as shown on the left in Fig. 8.6, the bound reports large regions of eraser behavior. And, most importantly, between the engine and lower eraser region lies a region

where we see that Eq. (8.2) implies

$$\langle W \rangle > k_B T \ln 2 \ \Delta H_1 \ , \tag{8.4}$$

a violation of the Second Law!

Of course, when we apply Eq. (8.3) in Fig. 8.6 (Right), the violation region disappears, to be correctly identified as duds. Additionally, the large region of eraser functionality in Fig. 8.6 (Left) shrinks significantly in Fig. 8.6 (Right). Figure 8.6 (Left)'s regions have been mischaracterized similar to the case discussed in Section 8.4.2. It is more subtle here, though, since $h_\mu > 0$. However, the fundamental problem is the same—by considering only the single-symbol entropy, it appears that the ratchet performs work to make the input less random, since $\Delta H_1 < 0$. In fact, the output is more intrinsically random than the input, and the ratchet dissipates work uselessly. In the violation region on the left, the ratchet is identified as not dissipating sufficient work to reduce the randomness as much as ΔH_1 implies it must be. This leads to the Second Law violation. This contradiction is resolved when we take into account that the input's intrinsic randomness was actually much lower than its single-symbol entropy. And so, the apparent decrease in randomness was in fact an increase.

It is already known that Eq. (8.2) may be violated in cases of a memoryful ratchet driven by memoryful input. However, the Mandal-Jarzynski ratchet was not designed to find such a violation, as has been done previously [**123**]. Rather, we find that driving a simple transition-rate based ratchet with a mixed-complexity process creates regions of violation when applying Eq. (8.2). Since such ratchets are common in application, and any such ratchet will be highly stochastic by nature, for reasons further discussed in Appendix H, we conclude that Eq. (8.2) is not suitable to be broadly applied. On the positive side, we see that the dynamical-systems techniques introduced here apply broadly, giving consistent and accurate characterizations stochastic-control information engines.

8.5. Constructing and Deconstructing Patterns

Up to this point, we monitored how the ratchet changed the *amount* of intrinsic randomness present in a symbol sequence and leveraged this to do useful work. When information ratchets are memoryful, they can alter not only the statistical bias of a symbol sequence, but also the presence of temporal correlations. This has thermodynamic consequences, as discussed above. Now, we turn to

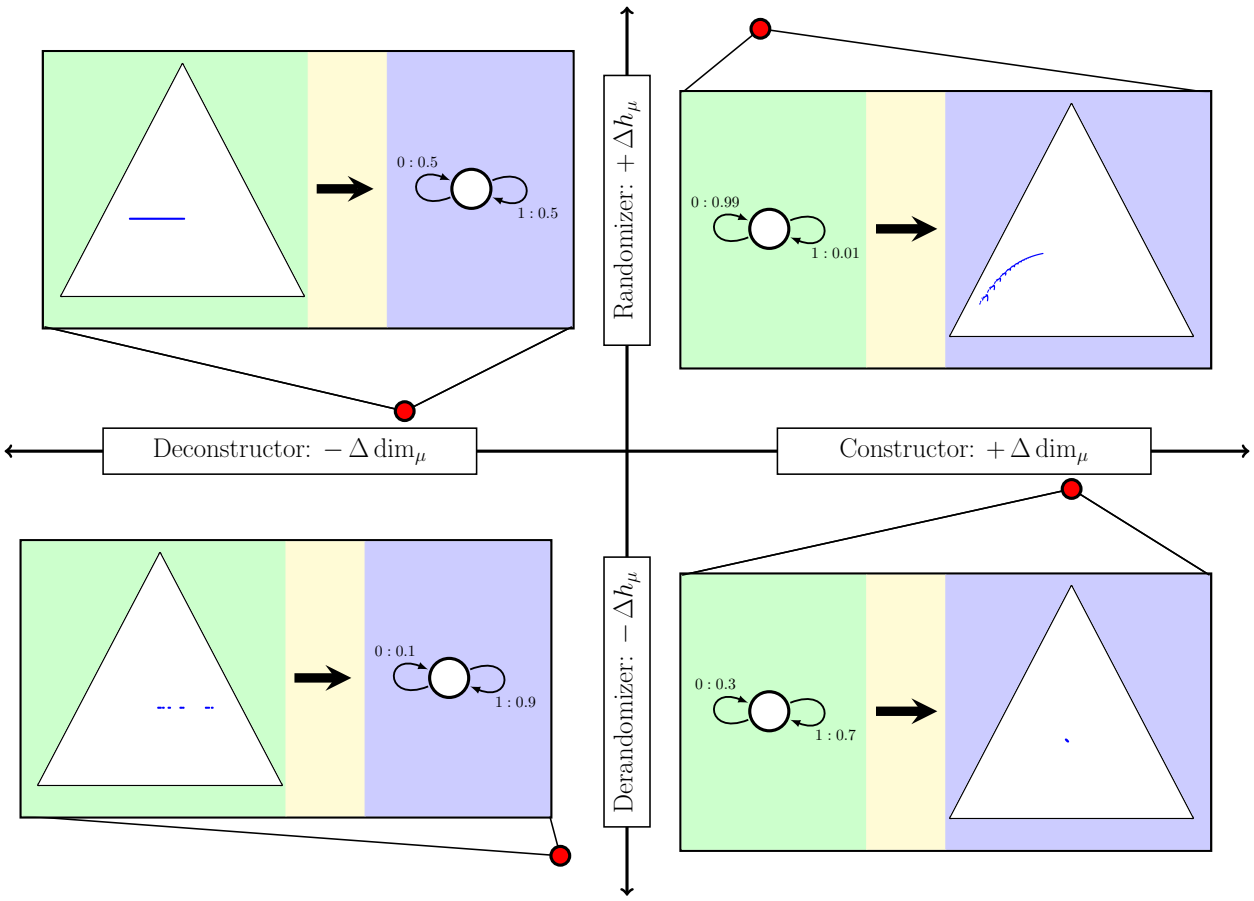


FIGURE 8.7. After passing through the information engine, the input process, which has some initial h_μ and d_μ, is transformed into an output process with a potentially different h'_μ and d'_μ. By carefully selecting input, an information engine can be induced to act as a randomizer ($\Delta h_\mu > 0$) or derandomizer ($\Delta h_\mu < 0$) and a pattern constructor ($\Delta d_\mu > 0$) or deconstructor ($\Delta d_\mu < 0$). We show here that all four regions of the Δh_μ–Δd_μ plane are accessible to the Mandal-Jarzynski ratchet by carefully selecting parameters and input. The two insets on the left show the uncountable set of mixed states of an input process that the ratchet transduces to an IID output. The insets on the right show two uncountably-infinite-state output processes produced by running the Mandal-Jarzynski ratchet on a biased coin. The parameters, clockwise from top right: $\delta_{input} = -0.98, \epsilon = 0.01, \tau = 0.1$; $\delta_{input} = 0.3, \epsilon = 0.5, \tau = 0.1$; $\delta_{output} = 0.8, \epsilon = -0.96, \tau = 0.75$; $\delta_{output} = 0.0, \epsilon = 0.9, \tau = 0.9$.

consider by what *mechanisms* an information ratchet changes the presence of temporal correlations, which manifests in changes in sequence structure and organization.

By structure and organization, we refer to the internal states of the HMM that generates the input symbol sequence, the ratchet states and transitions, and the output sequence. As depicted

in Fig. 8.2, the input and the ratchet each have their own set of internal states. Since the output machine is the composition of the ratchet transducer and input HMM, its states are the Cartesian product of the set of input states and set of output states.

In the simplest case, when a memoryless ratchet is driven by memoryless input, there is only ever one state, and no temporal correlations are present at any stage. The only possible action of the ratchet then is to change the statistical bias of individual input symbols and transform this change in Shannon entropy to a change in thermodynamic entropy.

When one or both of the input and ratchet are memoryful, the internal structure of the output will be, in general, memoryful. That is, the ratchet has induced a structural change in processing the input to generate the output. Consider two basic structural-change operating modes [**126**]: *pattern construction*, where the output is more structured than the input, and *pattern deconstruction*, the output is less structured. As before, these modalities are input-dependent—the same ratchet may exhibit either. Note that structural change to the symbol sequence does not uniquely determine the thermodynamic functionality associated with changes in randomness. It is possible for an information engine to act as an engine, eraser, or dud while constructing patterns. The same is true of deconstruction. Rather, transformations of randomness and structure are orthogonal, and a ratchet's information processing capabilities may lie anywhere in the Δh_μ—Δd_μ plane sketched in Fig. 8.7.

8.5.1. Pattern Construction. Ideal pattern construction occurs when a ratchet takes structureless input—an IID process—to structured output. Therefore, when the ratchet is driven with a biased coin input, it is operating as an ideal pattern constructor. As discussed in Section 8.4.1, driving the ratchet with memoryless input results in an uncountably-infinite set of states in the output HMM for most parameter values. The exception occurs along the line $\delta = \epsilon$ in parameter space, where the ratchet returns the input unchanged, implying $\Delta C_\mu = 0$. At every other point in ratchet parameter space $\Delta C_\mu = +\infty$ and the ratchet acts as a pattern constructor. As can be seen from Appendix H.0.2's Fig. H.1, this type of structural change can be associated with any thermodynamic behavior.

The resulting divergence of C_μ is a direct consequence of the nonunifilarity induced by the ratchet. The structure generated by any ratchet driven by an IID process is the set of mixed

states of the ratchet, given knowledge of the outputs. Due to the ratchet's topology, there is an uncountable infinity of such mixed states. In this circumstance one uses the statistical complexity dimension of the set of output mixed states to monitor the rate of the memory-resource divergence. Δd_μ distinguishes between output machines with an uncountable infinity of states, and so is able to compare the structural information processing of the ratchet across parameter space.

Figure 8.7 places two examples of ideal pattern construction on the right side of the Δh_μ–Δd_μ plane with the associated input and output machines, the latter plotted on the 2-simplex. Although it may appear that the more entropic ratchet in the upper half of the plane constructs a more "complex" pattern, this is not so. Refer back to Fig. 8.4 and compare the dimension of the two sets of mixed states to see the opposite is true. The ratchet operating in the $-\Delta h_\mu$ half of the plane produces a much denser set of states, resulting in a larger Δd_μ. In addition to the structural transformation, the ratchet in the $+\Delta h_\mu$ plane randomizes inputs as a dud, while the other derandomizes inputs as an eraser.

8.5.2. Pattern Deconstruction. In a complementary fashion, the ratchet can deconstruct patterns. In ideal pattern deconstruction, a ratchet transforms a memoryful input sequence, with $C_\mu > 0$, to memoryless, IID output, with $C_\mu = 0$. When taking a ratchet-focused view, as we do here, ideal pattern deconstruction is a more involved task than ideal pattern construction, since we must carefully design inputs that a ratchet will transform into a biased coin. Any correlations in the input must be recognizable by the ratchet so that the ratchet can map them to randomness. Similar to the previous discussion, we consider the induced ratchet mixed states, but now we have knowledge of the inputs. The algorithm to design the required input process, given knowledge of the ratchet, is discussed in Section 8.7.

Critically, pattern deconstruction is not possible for all ratchet parameters and desired output. That said, the Mandal-Jarzynski ratchet can perform as an engine, eraser, or dud while deconstructing patterns, as can be seen in Section 8.7's Fig. 8.9. As τ increases, the parameter-space region in which the ratchet can extract patterns shrinks. At $\tau \to \infty$ pattern extraction may only occur along the line $\delta = \epsilon$. In a mirror of pattern construction, generating the input processes requires reference to the uncountably infinite set of mixed states of the ratchet. In general, this implies that an input process which maps to a memoryless output process also has an uncountably-infinite set of states and

$\Delta C_\mu \to -\infty$. In other words, to properly ensure the output symbols are temporally uncorrelated, the input process must remember its *infinite past*. Once again, the associated statistical complexity dimension d_μ—now of the set of input mixed states—quantifies the rate of the memory-resource divergence.

Two examples of ideal pattern deconstruction are placed on the left side of the Δh_μ–Δd_μ plane in Fig. 8.7, with the ratchet mixed states—on the 2-simplex—and the output machine. Δd_μ is approximate, based on the dimension of the ratchet mixed states, which are conjectured to have the same dimension as the input mixed states.

8.5.3. Thermodynamic Taxonomy of Construction and Deconstruction. From its highly stochastic nature and from parameter sweeps like the one shown in Section 8.6's Fig. 8.8, we conclude that for almost all parameters, the Mandal-Jarzynski ratchet is only able to construct patterns with infinite sets of predictive features (mixed states). We conjecture that likewise, it is only able to perfectly deconstruct infinite patterns. An interesting note is that the input and output mixed-state sets and their dimensions are asymmetric. We can visually see the asymmetry in Fig. 8.7, which sketches the Δh_μ–Δd_μ plane and shows an example of the Mandal-Jarzynski ratchet operating in all four quadrants. Infinite state output constructed by the ratchet may span the simplex, but the mixed states of the ratchet, while acting as a deconstructor, always lie along a line in the simplex. This implies that while the ratchet may construct patterns up to $\Delta d_\mu = 2.0$, it is only able to deconstruct patterns up to $\Delta d_\mu = -1.0$. The difference in Δd_μ in these two modalities points to a difference in memory-resource divergence for pattern construction versus pattern deconstruction.

This asymmetry is not necessarily surprising. Recall the asymmetry in the ratchet's ability to randomize and derandomize behavior. The combined area of dud and engine regions in Fig. 8.6 comprise the ratchet's randomizing regime, while the derandomizing regime is the comparatively small eraser region. One interpretation of this asymmetry comes from the thermodynamic limitations on the ordering of Δh_μ and $\langle W \rangle$: While an increase in Δh_μ is thermodynamically unbounded, Δh_μ is constrained by the Second Law to only drop as low as the minimum asymptotic work. This strongly suggests that there is a thermodynamic taxonomy of structural transformation—one that parallels our existing thermodynamic taxonomy of randomness transformation. We must leave

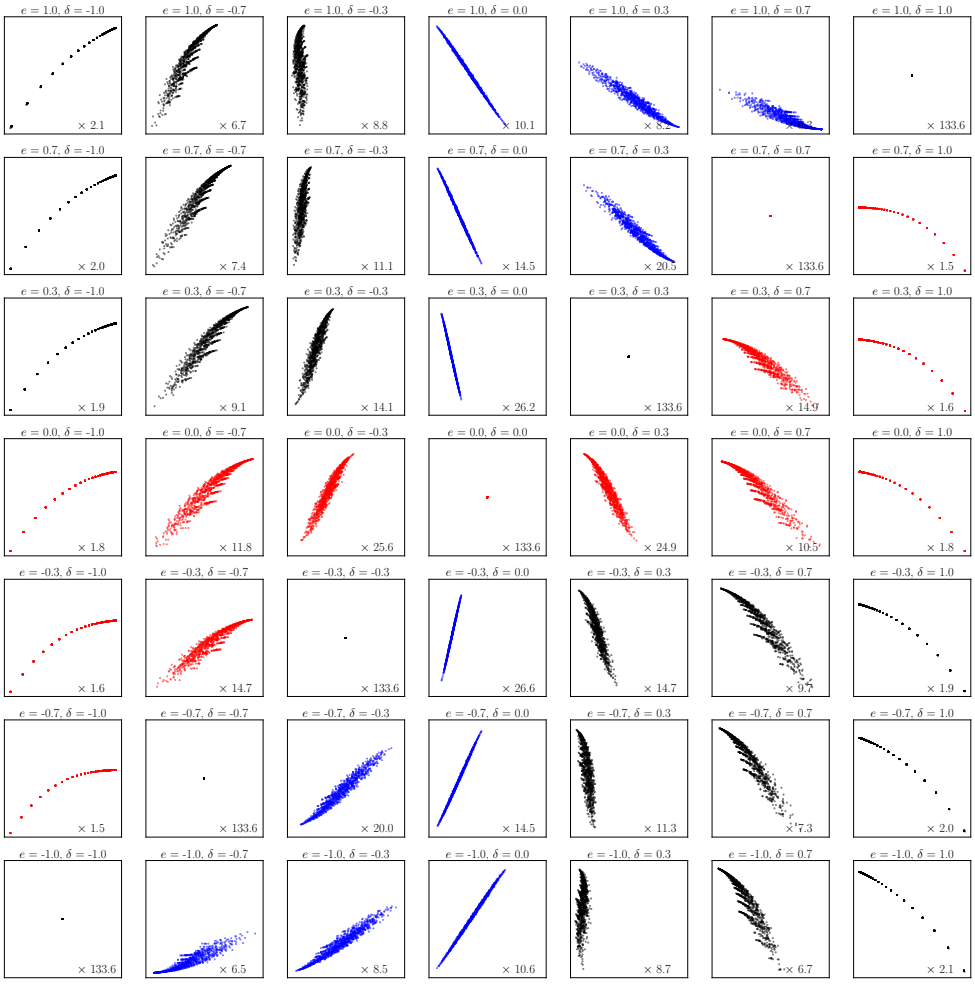


FIGURE 8.8. Mixed-state attractors of the output HMMs of the Mandal-Jarzynski ratchet driven by a Biased Coin as a function of ϵ and input bias δ, given above each square; $(\epsilon, \delta) \in [0, 1] \times [-1, 1]$. Each plot shows $1,000$ mixed states from the attractor at the magnification noted in the lower right corner. The attractors may be compared with Fig. H.1 to determine thermodynamic functionality.

finding such a taxonomy and the analysis of more general ratchets with input-dependent structural behavior to the future.

8.6. Information Ratchet Mixed-State Attractor Survey

To emphasize how exploring a ratchet's mixed states elucidates the underlying physics, Fig. 8.8 presents the attractors of the Mandal-Jarzynski ratchet driven by the Biased Coin, as a function of ϵ and b, in analogy with Fig. H.1. Each square in the grid shows the mixed-state attractor for the output HMM produced by the composition of the Mandal-Jarzynski ratchet at the given ϵ with a Biased Coin at the given bias δ. The grid is laid out identically to the functional thermodynamic plots above, with ϵ varying on the y-axis and the input bias b varying on the x-axis. Note that the squares are not at the same scale: each is magnified to show the structure of the attractor; the magnification factor is given in the lower right corner. Compare to Fig. 8.4 to see the mixed-state attractors in further detail. Additionally, the attractors are color coded to show thermodynamic functionality: red for engines, blue for erasers, and black for duds.

The symmetry of the Mandal-Jarzynski ratchet around $\epsilon = 0$ is revealed by how structure of the output HMM attractors is reflected and reversed over the $\epsilon = \delta$ line. Along this diagonal, we see that the mixed-state attractor collapses to a single state—a single point. This reflects the fact that at any $\epsilon = \delta$ the output HMM is the input Biased Coin, so $\langle W \rangle = \Delta H = \Delta h_\mu = 0$. Furthermore, we see that the structure of the mixed-state attractor does not have a strong effect on the thermodynamic functionality—very similar attractors act as duds and as erasers on each side of the $\epsilon = 0$ line. This is as expected since, although thermodynamic functionality appears to change suddenly, the grids in Figs. 8.6, 8.8 and H.1 actually sweep over output machines with smoothly changing transition probabilities. And, changes in functionality represented by the boundaries of thermodynamic regions are actually due to small, smooth changes in the comparative magnitude of $\langle W \rangle$ and Δh_μ. Figure 8.8 illustrates this clearly, as the mixed-state attractor changes smoothly under the parameter sweep.

Note that the construction of Fig. 8.8 was only possible due to the new dynamical-systems techniques outlined in this thesis. The recently developed guarantee of ergodicity and quick generation of mixed states allows us to easily plot and investigate the mixed-state attractors of arbitrary HMMs. And, this allows for parameter sweeps of attractors of HMM families and rapid calculation of their entropy rates. The latter was required to determine the thermodynamic functionality color coding in Fig. 8.8.

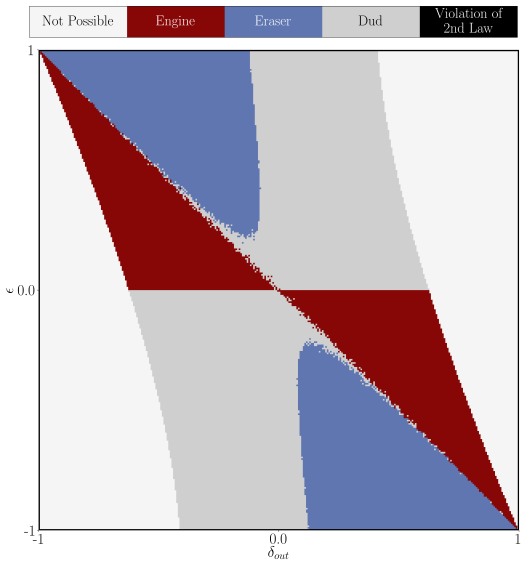


FIGURE 8.9. Functional thermodynamic regions of a ratchet pattern deconstructor with an interaction interval of $\tau = 0.75$. The x-axis sweeps over the output bias δ_{out}, while the y-axis sweeps over ϵ. As indicated, there are two parameter regimes where the ratchet is unable to act as a pattern deconstructor. In these regions, the desired output bias is not reachable by the machine at the given ϵ. As $\tau \to \infty$, this region grows, until it encompasses every parameter combination other than $\epsilon = \delta_{out}$, which is always reachable with an input Biased Coin with bias $\delta_{in} = \delta_{out} = \epsilon$.

8.7. Pattern Deconstruction and Thermodynamic Functionality

While it is relatively simple to run the Mandal-Jarzynski ratchet as an ideal pattern constructor, forcing the ratchet to perfectly deconstruct patterns is a more challenging task. As previously discussed in Ref. [126], to deconstruct patterns, the input and ratchet must remain synchronized. However, the Mandel-Jarzynski ratchet, being highly stochastic, resists synchronization. Since the input process cannot stay synced to the states of the Mandal-Jarzynski ratchet, it must synchronize to the mixed states instead. To design an input sequence that the ratchet transduces to an IID output process with bias δ we calculate as follows:

(1) Pick a ratchet mixed state;
(2) Determine the input-output probability distribution;
(3) Calculate the input probability distribution such that $\Pr(X' = 0) - \Pr(X' = 1) = \delta$.
(4) Step forward, record the input.
(5) Use the input to update the ratchet mixed state; and

(6) Repeat the procedure starting at Step (1), using the new mixed state.

Note that one must ensure that the output probability distribution remains constant at each time step.

As might be suspected from the algorithm, this is not possible at all parameters. For example, the ratchet may be so heavily biased to flip $0 \to 1$ that emitting a sequence of mainly 1s is mathematically impossible. This is expressed in the algorithm by finding a required input probability distribution with a negative component. This is illustrated in Fig. 8.9, where the functional thermodynamic regions associated with the Mandal-Jarzynski ratchet acting as a pattern deconstructor are shown. There are two inaccessible regions, where the desired output is not possible for the ratchet at the given value of ϵ. As $\tau \to \infty$ these regions grow in size, until at large τ the only parameter region capable of pattern deconstruction is $\delta = \epsilon$. This is where the ratchet becomes memoryless, so it is trivially a pattern deconstructor along this line.

8.8. Related Efforts

We can now place the preceding methods and new results in the context of prior efforts to identify the thermodynamic functioning of information engines. In short, though, having revealed the challenge of exact entropy calculations and the inherent divergence in structural complexity, the new methods appear to call for a substantial re-evaluation of previous claims. We start noting a definitional difference and then turn to more consequential comparisons.

The framework of information reservoirs discussed here differs from alternative approaches to the thermodynamics of information processing, which include: (i) active feedback control by external means, where the thermodynamic account of the Demon's activities tracks the mutual information between measurement outcomes and system state [**127**, **128**, **129**, **130**, **131**, **132**, **133**, **134**, **135**, **136**, **137**, **138**, **139**]; (ii) the multipartite framework where, for a set of interacting, stochastic subsystems, the Second Law is expressed via their intrinsic entropy production, correlations among them, and transfer entropy [**140**, **141**, **142**, **143**]; and (iii) steady-state models that invoke time-scale separation to identify a portion of the overall entropy production as an information current [**144**, **145**]. A unified approach to these perspectives was attempted in Refs. [**146**, **147**, **148**].

These differences being called out, Maxwellian demon-like models designed to explore plausible automated mechanisms that do useful work by decreasing the physical entropy, at the expense of positive change in reservoir Shannon information, have been broadly discussed elsewhere [**121**, **144**, **149**, **150**, **151**, **152**, **153**]. However, these too neglect correlations in the information-bearing components and, in particular, the mechanisms by which those correlations develop over time. In effect, they account for thermodynamic information-processing by replacing the Shannon information of the components as a whole by the sum of the components' individual Shannon informations. Since the latter is larger than the former [**55**], using it can lead to either stricter or looser bounds than the correct bound derived from differences in total configurational entropies. Of more concern, though, bounds that ignore correlations can simply be violated. Finally, and just as critically, the bounds refer to configurational entropies, not the intrinsic dynamical entropy over system trajectories—the Kolmogorov-Sinai entropy. A more realistic model was suggested in Ref. [**154**]. Issues aside, these designs have been extended to enzymatic dynamics [**155**], stochastic feedback control [**156**], and quantum information processing [**157**, **158**].

In comparison, our approach expands on Ref. [**122**] that considers a Demon in which all correlations among the system components are addressed and accounted for. As shown above, this has significant impact on the analysis of Demon thermodynamic functionality. To properly account for correlations, we developed a new suite of tools that allow quickly and efficiently analyzing nonunifilar HMMs and related stochastic controllers, which removes the mathematical intractability of analyzing correlations for arbitrary demons. We note that our approach and results are consistent with the analyses that consider the entropy of the system as a whole, therefore treating correlations in the system implicitly, an approach epitomized by Ref. [**159**]. Since correlations are not ignored, this approach is fully consistent with our treatment. This being said, insofar as that work does not address specific partitioning of the system, it does not offer an explicit accounting of the system's internal correlations, as is done here. As previously discussed, one may derive information ratchet-type results from that approach by considering an explicit partitioning [**125**, **126**, **160**]. While the results are consistent, leaving the role of correlations implicit does not allow for investigating how to best leverage them. It also does not give a way to analyze internal computational structure. These remarks highlight the importance of explicitly considering information engine-style partitioning.

The dynamical-systems methods additionally allowed us to consider a Demon's internal structure, which had only previously been investigated for unifilar ratchets in Ref. [126]. From engineering and cybernetics to biology and now physics, questions of structure and how an agent, here understood as the ratchet, interacts with and leverages its environment—i.e., input—is a topic of broad interest [161,162]. General principles for how an agent's structure must match that of its environment will become essential tools for understanding how to take thermodynamic advantage of correlations in structured environments, whether the correlations are temporal or spatial. Ashby's Law of Requisite Variety—a controller must have at least the same variety as its input so that the whole system can adapt to and compensate that variety and achieve homeostasis [161]—was an early attempt at such a general principle of regulation and control. For information engines, a controller's variety should match that of its environment [125]. Above, paralleling this, but somewhat surprisingly, we showed that for the Mandal-Jarzynski ratchet to extract patterns from its environment, the input must have an uncountably infinite set of memory states synchronized to the ratchet's current mixed state. One cannot but wonder how such requirements manifest physically in adaptive thermodynamic nanoscale devices and biological agents.

CHAPTER 9

Conclusion and Future Directions

This dissertation opened by considering the role that *information* has to play in our understanding of physical systems. We argued that like energy before it, information is a measurable, fundamental quantity of physical importance. We further argued that simply quantifying information content is not enough to fully understand the role that it has to play in the natural world—we must understand how information can be generated, stored, transformed, and utilized. To make a project of this scope tractable, we narrowed our focus to a specific class of stochastic processes that has been used broadly to model natural systems. This class is known as hidden Markov processes, which may be generated by hidden Markov models.

A central challenge in studying this class of processes has been quantifying randomness, patterns, and structure and doing so in a mathematically-consistent but calculable manner. For well over a half a century Shannon entropy rate has stood as the standard by which to quantify randomness in a time series. Until now, however, calculating it for processes generated by nonunifilar HMMs has been difficult and inaccurate, at best.

We began our analysis of this problem by recalling that, in general, hidden Markov models that are not unifilar have no closed-form expression for the Shannon entropy rate of the processes they generate. Despite this, these HMMs can be *unifilarized* by calculating the mixed states. The resulting mixed-state presentations are themselves HMMs that generate the process. However, adopting a unifilar presentation comes at a heavy cost: Generically, these presentations have uncountably-infinite state sets and so Shannon's expression cannot be used. Nonetheless, we showed how to work constructively with these mixed-state presentations.

In particular, we showed that they fall into a common class of dynamical system known as place-dependent iterated function systems. Analyzing the IFS dynamics associated with a finite-state nonunfilar HMM allows one to extract useful properties of the original process. For instance, we

can easily find the entropy rate of the generated process from long orbits of the IFS. That is, one may select any arbitrary starting point in the mixed-state simplex and calculate the entropy over the IFS's place-dependent probability distribution. We evolve the mixed state according to the IFS and sequentially sample the entropy of the place-dependent probability distribution at each step. Using an arbitrarily long word and taking the mean of these entropies, the method converges on the process' entropy rate.

Although the IFS-HMC connection has been considered previously [**163**, **164**], our development complements this by expanding it to address the role of mixed-state presentations in calculating the entropy rate and to connect it to existing approaches to randomness and structure in complex processes. In particular, while our results focused on quantifying and calculating a process' randomness, we left open questions of pattern and structure. Towards this, we showed how the attractor of the IFS defined by an HMC is, assuming uniqueness of the mixed states as discussed in Section 3.4, the set of causal states $\boldsymbol{\mathcal{R}}$ of the process generated by that HMC. Practically, this gives a method to construct the causal states of a process $\mathcal{P}$, so long as it can be finitely generated. For instance, Fig. 3.2 demonstrated how the highly structured nature of the Simple Nonunifilar Source is made topologically explicit through calculating its mixed-state presentation—which is also its ϵ-machine.

In point of fact, many information-theoretic properties of the underlying process may be directly extracted from its mixed-state presentation. These sets are often fractal in nature and quite visually striking. See Fig. 3.4 for several examples. We established that the information dimension of the mixed-state attractor is exactly the divergence rate of the *statistical complexity* [**17**]—a measure of a process' structural complexity that tracks memory. Thus, processes in this class effectively increase their use of memory, "creating" mixed or causal states, on the fly. Furthermore, we introduced a method to calculate the information dimension of the mixed-state attractor from the Lyapunov spectrum of the mixed-state IFS. In this way, we demonstrated that coarse-graining the mixed-state simplex—the previous method for studying the structure of infinite-state processes [**60**]—can be avoided altogether. This greatly improves accuracy and computational speed and deepens our understanding of the origins of complexity in stochastic processes.

That was not the end of the story, since calculating d_μ is difficult due to long-standing problems in the field of IFS dimension theory. In particular, the overlapping problem posed a significant

hurdle—restricting the preceding results to only nonoverlapping IFSs. This restricted us to the class of stochastic processes with a one-to-one past-to-causal state mapping. In a sense, these processes are the most complex but with structure that is the least interesting. That is, for DIFSs we simply store every past to build an optimally-predictive model.

This state of affairs led directly to the present development and to introducing the ambiguity rate. The latter allows smoothly varying between ϵ-machines with countable state spaces ($h_a = h_\mu$ and $\Delta H[\boldsymbol{\mathcal{R}}] = 0$) and those with perfectly self-similar state spaces ($h_a = 0$ and $\Delta H[\boldsymbol{\mathcal{R}}] = h_\mu$), including all those lying in between, with $h_\mu > h_a > 0$ and $\Delta H[\boldsymbol{\mathcal{R}}] = h_\mu - h_a$. This model class is much more general, generating an exponentially larger family of stochastic processes. As such, we anticipate that this class will be of great interest and likely to lead to significant further progress in analyzing the randomness and structure generated by hidden Markov processes.

In the introduction, we noted the ubiquity of multi-scale systems with several levels of simultaneous mechanism. We noted that our ultimate goal is to describe behavior at all levels of these system, and measure information flow up and down levels of structure. To briefly expand on what is meant by "information flow", we note here that the ambiguity rate h_a is the first step towards an information anatomy of models. Taking inspiration from decompositions of information and information flow in processes [**165**, **166**, **167**, **168**], I wish to better characterize the role of information in the ϵ-machine itself. As an example, C_μ, while vital to characterizing computational complexity, is static in the sense that it gives the quantity of memory resources available to the model at one moment in time. On the other hand, $h_\mu - h_a$ describes the *rate* at which the memory resources in the model is increasing (see again figure 1.2) and can described as the information flux of the model. By decomposing this flux further based on how information retains or loses temporal correlations as it flows through the model, we can build out the informational "anatomy of a machine". This brings to mind classic work in computation theory aimed at uncovering mechanism through computational structure [**169**] and will be necessary to fully understand how multi-scale and hierarchical systems transform, store, and process information.

Despite the run away successes of the Information Age, it is well known that our ability to predict complex systems have been balanced out by our inability to "see under the hood", so to speak, of the black-box techniques so widely used today. As physicists, we must be unsatisfied

with this partial success, desiring models of natural system that combine predictive power with mechanistic explication. This goal, by no means to be underestimated, underlies my long-term academic goals and research program. As has become inescapably clear in the last century, deepening our understanding of the natural world points us towards ever more complex, structured, hierarchical systems, replete with interacting subsystems and complicated modes of information flow. Our ability to model and predict these systems is tantamount to the goals of science itself.

APPENDIX A

Hidden Markov Model Examples

We reproduce here the HMCs used to create the various examples of MSPs present in the text.

First, the "alpha" HMC, from Fig. 5.1a, is given by:

$$T^{\square} = \begin{pmatrix} 2.734 \times 10^{-2} & 0.392 & 1.924 \times 10^{-2} \\ 0.475 & 2.176 \times 10^{-2} & 2.766 \times 10^{-4} \\ 0.224 & 2.711 \times 10^{-3} & 0.236 \end{pmatrix},$$

$$T^{\triangle} = \begin{pmatrix} 1.845 \times 10^{-3} & 0.133 & 0.259 \\ 3.913 \times 10^{-2} & 0.315 & 2.789 \times 10^{-2} \\ 0.467 & 1.015 \times 10^{-2} & 4.699 \times 10^{-3} \end{pmatrix},$$

$$T^{\circ} = \begin{pmatrix} 9.782 \times 10^{-2} & 3.374 \times 10^{-2} & 3.644 \times 10^{-2} \\ 5.422 \times 10^{-2} & 6.503 \times 10^{-2} & 2.090 \times 10^{-3} \\ 5.328 \times 10^{-2} & 1.278 \times 10^{-3} & 8.778 \times 10^{-4} \end{pmatrix}.$$

Fig. 5.1b is given by Eq. (3.7), at $\alpha = 0.6$ and $x = 0.1$. The "beta" HMC, in Fig. 5.1c, is given by:

$$T^{\square} = \begin{pmatrix} 5.001 \times 10^{-2} & 0.388 & 4.251 \times 10^{-2} \\ 0.464 & 4.484 \times 10^{-2} & 2.495 \times 10^{-2} \\ 0.232 & 2.720 \times 10^{-2} & 0.243 \end{pmatrix},$$

$$T^{\triangle} = \begin{pmatrix} 1.708 \times 10^{-3} & 0.123 & 0.240 \\ 3.623 \times 10^{-2} & 0.292 & 2.583 \times 10^{-2} \\ 0.432 & 9.397 \times 10^{-3} & 4.351 \times 10^{-3} \end{pmatrix},$$

$$T^{\circ} = \begin{pmatrix} 9.0576 \times 10^{-2} & 3.124 \times 10^{-2} & 3.374 \times 10^{-2} \\ 5.020 \times 10^{-2} & 6.021 \times 10^{-2} & 1.935 \times 10^{-3} \\ 4.933 \times 10^{-2} & 1.183 \times 10^{-3} & 8.127 \times 10^{-4} \end{pmatrix}.$$

The mapping images shown in Fig. 5.2 are produced by the three symbol DIFSdefined in Eq. (3.7), with $\alpha = 0.63, x = 0.2$ for the overlapping example in Fig. 6.2a and $\alpha = 0.6, x = 0.15$ for the nonoverlapping example in Fig. 6.2b.

Due to finite numerical accuracy, reproduction of the attractors using these specifications may differ slightly from Fig. 5.1.

APPENDIX B

Asymptotic Equipartition and Typical Set Contraction

The *asymptotic equipartition property* (AEP) states that for a discrete-time, ergodic, stationary process X:

$$-\frac{1}{\ell}\log_2 \Pr(X_1, X_2, \ldots, X_\ell) \to h_\mu(X) \ , \tag{B.1}$$

as $\ell \to \infty$ [55]. This effectively partitions the set of sequences in two: the *typical set*—sequences for which the AEP holds—and the *atypical set*, for which it does not. As a consequence of the AEP, it must be the case that the typical set is measure one in the space of all allowed realizations and all sequences in the atypical set approach measure zero as $\ell \to \infty$.

We argue that while our IFS class includes reducible maps, any composition of maps corresponding to a word in the typical set will be irreducible. This can be seen intuitively by considering the SNS, shown in Fig. 3.2, and adding an additional transition on a $\square$ from σ_0 to σ_1. This produces an HMC with two reducible symbol-labeled transition matrices, but an irreducible total transition matrix. However, as $|w| \to \infty$, the only words such that $T^{(w)}$ remains reducible are $\square^\ell$ and $\triangle^\ell$. We can see that these words cannot possibly be in the typical set, since $-\frac{1}{\ell}\log_2 \Pr(\square^\ell) = -\log_2 \Pr(\square) \neq h_\mu(X)$. The entropy rate h_μ is by definition the branching entropy averaged over the mixed states. And so, any word that visits only a restricted subset of the mixed states—i.e., a word with a reducible transition matrix—cannot approach h_μ, regardless of length. Therefore, only words with an irreducible mapping will be in the typical set, implying that there exists an integer word length $|w| > 0$ for which words without a contractive mapping are measure zero.

APPENDIX C

Birkhoff's Contraction Coefficient

First, we define the relevant metric.

DEFINITION C.0.1. Given an integer $N \geq 2$, let C^N be the nonnegative cone in $\mathbb{R}^N$, so that C^N consists of all vectors $z = (z_1, z_2, \ldots, z_N)$ satisfying $z \neq 0$ and $z_i \geq 0$ for all i. The *projective distance* $d : C^N \times C^N \to [0, \infty)$ is defined:

$$d(z, y) := \max \left\{ \left| \log \left(\frac{z_r}{z_s} \frac{y_s}{y_r} \right) \right| : r, s = 1, \ldots, N; r \neq s \right\} \tag{C.1}$$

for $z, y \in C^N$, where $d(z, z) = 0$. If one of the points is on the cone boundary, the distance is taken to be $+\infty$.

If $T^{(x)}$ is an $N \times N$ positive matrix, we have $d(zT^{(x)}, yT^{(x)}) < d_N(z, y)$ for every $z, y \in C^N$ such that $d(y, z) > 0$. We define the *projective contractivity* $\tau^{(x)}$ associated with $T^{(x)}$ as:

$$\tau_x := \sup_{\{z,y \in C_N : d(z,y) > 0\}} \frac{d(zT^{(x)}, yT^{(x)})}{d(z, y)} \ ,$$

so that $\tau^{(x)}$ satisfies $\tau^{(x)} \leq 1$. As the theorem below indicates, this inequality is strict.

THEOREM C.0.1. *([170, Thm. 1].) Let the integers $m, n \geq 2$ be arbitrary. For each matrix $T^{(x)} = \left[t_{ij}^{(x)} \right]$ of order $m \times n$ with positive components, $\tau^{(x)}$ is given by the following Birkhoff formula:*

$$\tau^{(x)} = \frac{1 - \left(\phi^{(x)} \right)^{1/2}}{1 + \left(\phi^{(x)} \right)^{1/2}} \ ,$$

where:

$$\phi(H) := \min_{r,s,j,k} \frac{t_{rj}^{(x)} t_{sk}^{(x)}}{t_{sj}^{(x)} t_{rk}^{(x)}} \ .$$

By inspection we see that $\phi(H) > 0$ and $\tau < 1$. As Ref. [171] notes, not only does the projective metric turn all positive linear transformations into contraction mappings, it is the only metric that does so.

The Birkhoff result extends to any nonnegative matrix $T^{(x)}$ for which there exists an $\ell \in \mathbb{N}^+$ such that $\left(T^{(x)}\right)^\ell$ is a positive matrix. Then there will be a $\tau^{(x)} < 1$ such that $d\left(\eta\left(T^{(x)}\right)^\ell, \zeta\left(T^{(x)}\right)^\ell\right) < d(\eta, \zeta)$.

APPENDIX D

Correspondence with Baker's Map

The simple dimension formula in Eq. (5.8) may not seem easily motivated. Especially, considering that, in general, both positive and negative Lyapunov exponents are required to have a nontrivial attractor. However, for iterated function systems, all Lyapunov exponents are negative and the expansive role played by positive Lyapunov exponents is instead played by an IFS's stochastic map selection, as measured by the entropy rate h_μ.

This is more intuitively appreciated by comparing the two-state IFS with the Baker's map. Consider the Baker's map:

$$x_{n+1} = \begin{cases} \dfrac{x_n}{s_0} \ , & y < p \\ \dfrac{x_n + s_1 - 1}{s_1} \ , & y \geq p \end{cases} \quad \text{and}$$

$$y_{n+1} = \begin{cases} \dfrac{y_n}{p} \ , & y < p \\ \dfrac{y_n - p}{p - 1} \ , & y \geq p \end{cases}$$

It has LCE spectrum $\Lambda = \{\lambda_1, \lambda_2\}$, where:

$$\lambda_1 = p \log(p) + (1 - p) \log(1 - p)$$

$$\lambda_2 = p \log(1/s_0) + (1 - p) \log(1/s_1) \ .$$

Note that $\lambda_1 > 0$ and $\lambda_2 < 0$. Then, the Lyapunov dimension is:

$$d_\Gamma = 1 - \frac{\lambda_1}{\lambda_2} \ .$$

To compare this to an IFS, take:

$$\{f(x)\} = \left\{ \frac{x_n}{s_0}, \frac{x_n + s_1 - 1}{s_1} \right\} \text{ and}$$
$$\{p(x)\} = \{p, 1-p\} \ .$$

Thus, we identify the y coordinate as controlling the stochastic map choice. The dynamic over position in the y direction exactly determines the IFS entropy rate. Since the Baker's map is volume preserving in y, the extra dimension always contributes a plus one in the dimension formula. In other words, the dimension along a slice of constant y equals the IFS dimension.

APPENDIX E

Sierpinski's Triangle

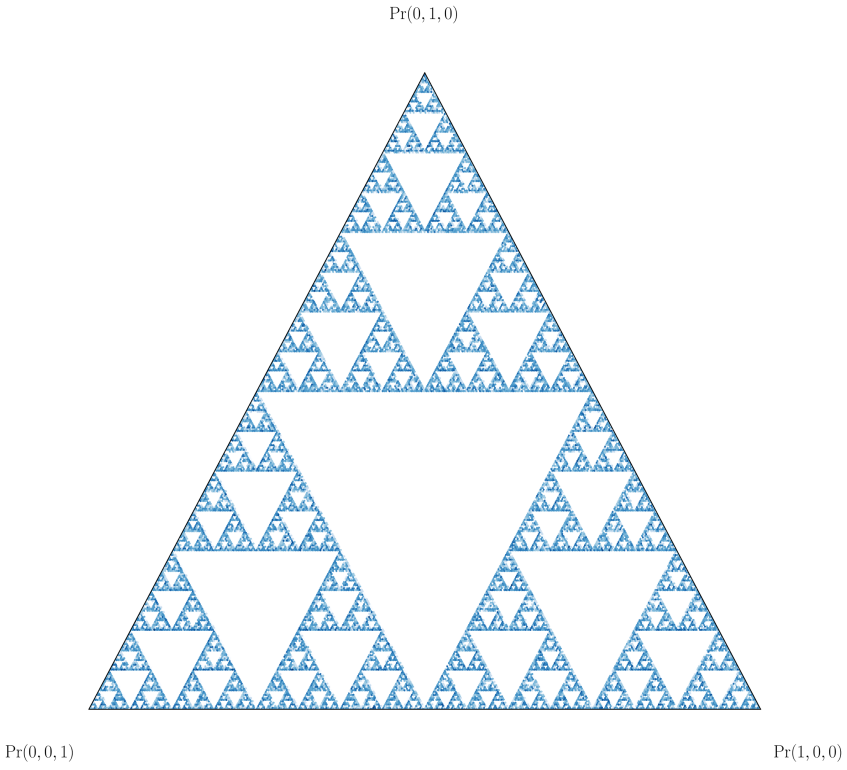


FIGURE E.1. Attractor of the 3-state, 3-symbol machine specified in Eq. (E.1), with $s = 2$ and $a = \frac{1}{6}$.

The Sierpinski triangle is a canonical Cantor set in two dimensions. An HMC that generates a MSP attractor that is the Sierpinski triangle is:

$$T^{(0)} = \begin{pmatrix} a & 0 & a(s-1) \\ 0 & a & a(s-1) \\ 0 & 0 & as \end{pmatrix}, \; T^{(1)} = \begin{pmatrix} \frac{1-as}{2} & 0 & 0 \\ \frac{(1-as)(s-1)}{2s} & \frac{1-as}{2s} & 0 \\ \frac{(1-as)(s-1)}{2s} & 0 & \frac{1-as}{2s} \end{pmatrix}, \text{ and} \tag{E.1}$$

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

where s controls the contraction coefficient and a controls the probability of selecting the maps. This HMC produces constant probability functions:

$$p^{(0)} = as \; , \; p^{(1)} = \frac{1-as}{2} \; , \text{ and } p^{(2)} = \frac{1-as}{2} \; .$$

and, therefore, linear mappings, since $f^{(0)} = \langle \eta | \, T^{(i)}/p^{(i)}(\eta)$. The constant probability functions make the entropy rate trivial to calculate. And, the linearity of the mappings does the same for the Lyapunov exponents.

Setting $s = 2$ and $a = 1/6$, results in equal probability for all maps and gives the standard Sierpinski triangle shown in Fig. E.1. In this case, the entropy rate is $h_\mu = \log_2 3$ and the Lyapunov exponents are both $-\log_2 2$. Plugging this into Eq. (5.6) returns the well-known fractal dimension of the Sierpinski triangle: $\log_2 3 / \log_2 2 \approx 1.585$.

APPENDIX F

Lyapunov Exponents

A *Lyapunov characteristic exponent* for a dynamical system measures the exponential rate of separation of trajectories that begin infinitesimally close. Since, typically, the separation rate depends on the direction of the initial separation, we use a spectrum of Lyapunov exponents, with one exponent for each state-space dimension. In a chaotic dynamical system, at least one Lyapunov exponent is positive. In general, the Lyapunov exponent spectrum for an N-dimensional dynamical system with mapping $x_{n+1} = F(x_n)$ depends on the initial condition x_0. However, here we consider ergodic systems, for which the spectrum does not.

Consider the map's *Jacobian* matrix:

$$J = \frac{\partial F}{\partial x}$$

and the evolution of vectors in the tangent space, controlled by:

$$\dot{Y} = YJ \ ,$$

where $Y(0) = \mathbb{I}_N$ and $Y(t)$ describes how an infinitesimal change in $x(0)$ has propagated to $x(t)$. Let $\{y_1, \ldots, y_N\}$ be the eigenvalues of the matrix $Y(t)Y(t)^\intercal$. Then, the Lyapunov exponents are:

$$\lambda_i = \lim_{t\to\infty} \frac{1}{2t} \log y_i \ .$$

The *Lyapunov numbers* were introduced and proven to exist by Oseledets [**172**]. The Lyapunov exponents are merely the logarithms of Lyapunov numbers.

The most common way of calculating an IFS's Lyapunov spectrum, employed to produce the results in Fig. 5.4, is the *pull-back* method. The basic idea is that for an IFS in the $N-1$ simplex, defined by an N-state HMC, there will be $N-1$ independent directions of contraction. These directions are represented by a coordinate frame of $N-1$ vectors that are kept orthogonal and

normalized. This coordinate frame is carried along a long orbit of the IFS. At each time step, the Jacobian is used to evolve the frame. We track the contraction rate of each vector, and this becomes the estimation of the Lyapunov exponents.

In contrast to applying this approach to deterministic dynamical systems, as is more familiar, an IFS's stochastic nature introduces additional error. Since the Jacobian varies not just across the simplex, but also for the selected maps, the orbit must be long enough to sample from the IFS attractor's distribution accurately for each possible mapping function. This being said, it has been established that the pull-back method works for IFS spectra, given a sufficiently long orbit [**57**].

The prequel, on estimating the entropy rate of HMC processes, made use of error-bounding techniques from Markov chain Monte Carlo (MCMC) [**48**]. Since here we are estimating Lyapunov exponents by sampling from the Blackwell distribution, similar error-bounding techniques apply. In this analysis, there are two fundamental sources of estimation error. First, that due to *initialization bias* or undesired statistical trends introduced by the initial transient data produced by the Markov chain before it reaches the desired stationary distribution. Second, there are errors induced by *autocorrelation in equilibrium*. That is, the samples produced by the Markov chain are correlated. And, the consequence is that statistical error cannot be estimated by $1/\sqrt{N}$, as done for N independent samples.

Bounding these error sources requires estimating the autocorrelation function, which can be done from long sequences of samples. If we have the nonunifilar model in hand, it is a simple matter of sweeping through increasingly long sequences of generated samples until we observe convergence of the autocorrelation function. An alternative method of approximating the infinite-state HMC with a finite-state approximation is discussed in detail in our previous work [**48**]. The upshot is that the method here generally efficiently leads to accurate estimates of the LCE spectrum.

For completeness, we note that there are alternative methods to calculate Lyapunov exponents; see, e.g., Refs. [**173**, **174**]. These methods may be more appropriate in specific applications. That said, the accuracy and applicability of Lyapunov exponent estimation is not the focus here.

APPENDIX G

Overlap Estimation

To estimate the size of a mixed-state attractor and overlap of mapping functions in Fig. 5.4, a combination of techniques were used. We will briefly summarize the method here.

First, $250,000$ different HMCs were generated using a 500×500 parameter grid over $\alpha = [0, 1]$ and $x = [0, 0.5]$. Each HMC was defined by plugging the appropriate parameter values into the symbol-labeled transition matrices in Eq. (3.7). From this HMC, the mapping and probability functions were defined (see Eqs. (3.5) and (3.6)), producing a place-dependent IFS.

For each IFS and associated HMC, $10,000$ mixed states were generated from an initial randomized state, throwing away the first $5,000$ as transients. Using a spatial algorithm from the SciPy Python package, a convex hull was drawn around this set of points, with a small buffer. This convex hull (the attractor "outline") was converted into a polygon. This polygon was then evolved independently by each symbol-labeled mapping function, producing three polygons, each associated with a symbol. This may be visualized by referencing Fig. 5.3, where the evolved polygons are depicted on top of the mixed-state attractor, each with a different color. We can see that the combination of these polygons must necessarily cover the attractor.

These three symbol-labeled polygons were then combined into a single polygon or multipolygon (a polygon with "holes" that are themselves polygons) using the geometry-processing module Shapely [**175**]. This produces a more accurate outline of the attractor than the convex hull. This process may then be repeated with the new outline for as many iterations as desired, until a polygon or multipolygon that covers the mixed state attractor with the desired level of accuracy is produced. The same result could be achieved by beginning with the entire simplex as the initial outline, without any production of mixed-states. However, the step of estimating the convex hull sharply reduces the number of required iterations and, more importantly, makes the required number more equal across parameter space. To see this, consider that attractors taking up less of the simplex require

several iterations to converge to the small size of the attractor. By initializing with the convex hull, the process of converging to the attractor's basic shape is skipped, and the iterations are merely refinements.

In producing Fig. 5.4, we found that we could determine good outlines across parameter space by evolving the convex hull three times. To produce Fig. 5.4a, the area of the resultant polygon or multipolygon was found using Shapely. To produce Fig. 5.4d and Fig. 5.4d, the outline was evolved one more time by each map, and the resultant polygons and/or multipolygons were checked for intersection. For the binary overlap/no-overlap plot in Fig. 5.4d, only the existence of overlap somewhere on the attractor was considered. For the percentage overlap in Fig. 5.4c, the area of the total outline that was comprised of overlapping polygons—whether only two or all three—was compared to the total area. The subtlety of whether an overlap region included two or three maps was largely ignored here, but will be analyzed in future explorations of the overlap problem.

APPENDIX H

Mandal-Jarzynski Ratchet

To work with the Mandal-Jarzynski ratchet, we reformulated it in computational mechanics terms, which is explained in Section 8.2. In its original conception, the model was imagined as a single symbol ("bit") interacting with a dial that may smoothly transition between three positions, as shown on the left in Fig. 8.3. This results in six possible states of the joint dial-symbol system, $\{A \otimes 0, A \otimes 1, B \otimes 0, B \otimes 1, C \otimes 0, C \otimes 1\}$. The transitions among these six states are modeled as a Poisson process, where $\mathcal{R}_{ij}$ is the infinitesimal transition probability from state j to state i, with $i; j \in \{A \times 0, \ldots, C \times 1\}$ [**121**]. The *weight parameter* ϵ, so named because it is intended to model the effect of attaching a mass to the side of the dial, impacts the probability of transitions among the six states by making $0 \to 1$ transitions energetically distinct from $1 \to 0$ transitions. This creates a preferred "rotational direction", since bit flips in one direction will be more energetically beneficial than the other. This is what allows the ratchet to do useful work.

Explicitly, the transition rate matrix $\mathcal{R}$ is:

$$\mathcal{R} = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -2+\epsilon & 1+\epsilon & 0 & 0 \\ 0 & 0 & 1-\epsilon & -2-\epsilon & 1 & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix} . \tag{H.1}$$

To express the ratchet's evolution over a single interaction interval of length τ, we calculate $\mathcal{T}(\tau, \epsilon) = \left(e^{\mathcal{R}(\epsilon)\tau}\right)^{\mathsf{T}}$, the transition matrix of the six-state Markov model representing the Mandal-Jarzynski model. In turn, this six-state model with the states $\{A\otimes 0, A\otimes 1, B\otimes 0, B\otimes 1, C\otimes 0, C\otimes 1\}$ may be transformed into a three-state transducer, with states $\{A, B, C\}$ and input and output

symbols in $\{0,1\}$. To do this, we define the projection matrices:

$$\mathbb{P}_0 = \begin{pmatrix} \mathbb{I}_3 \\ \mathbf{0}_3 \end{pmatrix} \text{ and } \mathbb{P}_1 = \begin{pmatrix} \mathbf{0}_3 \\ \mathbb{I}_3 \end{pmatrix} .$$

Then, the transducer input-output matrices $K^{\text{in,out}}(\tau,\epsilon)$ for a given $\mathcal{T}(\tau,\epsilon)$ are given by:

$$K^{\text{in,out}}(\tau,\epsilon) = (\mathbb{P}_{\text{in}})^{\intercal}\, \mathcal{T}(\tau,\epsilon)\mathbb{P}_{\text{out}} .$$

For all $\epsilon \in (0,1)$ and $\tau \in (0,\infty]$, all four $K^{\text{in,out}}(\tau,\epsilon)$ are positive definite matrices. This is what is meant by "fully-connected, highly stochastic" controller—all transitions on all combinations of symbols have positive probability. Explicitly, the probability function $f(\ldots)$ referenced in both Fig. 8.2 and Fig. 8.3 is given by:

$$\begin{aligned} \Pr(X' = x', X = x, R_N = r, R_{N+1} = r') & \\ &= f(x, x', r, r', \epsilon, \tau) \\ &= (\mathbb{P}_x)^{\intercal}\, \mathcal{T}_{R,R'}(\tau,\epsilon)\mathbb{P}_{x'} . \end{aligned} \tag{H.2}$$

H.0.1. Composing a Ratchet with an Input Process' Machine. Given an input process generated by an HMM with transition matrices $T^{(x)}$, such that $x \in \{0,1\}$, we may exactly calculate the transition matrices $T'^{(x')}$ of the output process' HMM:

$$T'^{(x')}_{R_N \times S_N, R_{N+1} \times S_{N+1}} = \sum_x K^{x,x'}_{R_N, R_{N+1}} T^{(x)}_{S_N, S_{N+1}} ,$$

noting that the state space of the output HMM is the Cartesian product of the state space of the transducer $\mathcal{R}$ and the state space of the input machine $\mathcal{S}$. Although presenting this in the setting of the Mandal-Jarzynski ratchet specifically, this method applies for any input machine and transducer, given that the transducer is able to recognize the input [**176**]

That said, there are several interesting points specific to the Mandal-Jarzynski ratchet we should highlight. As noted in the previous section, the Mandal-Jarzynski transducer matrices are positive definite, guaranteeing that the output machine will be nonunifilar, although disallowed state transitions in input machines are preserved in the output. (Composing the Mandal-Jarzynski

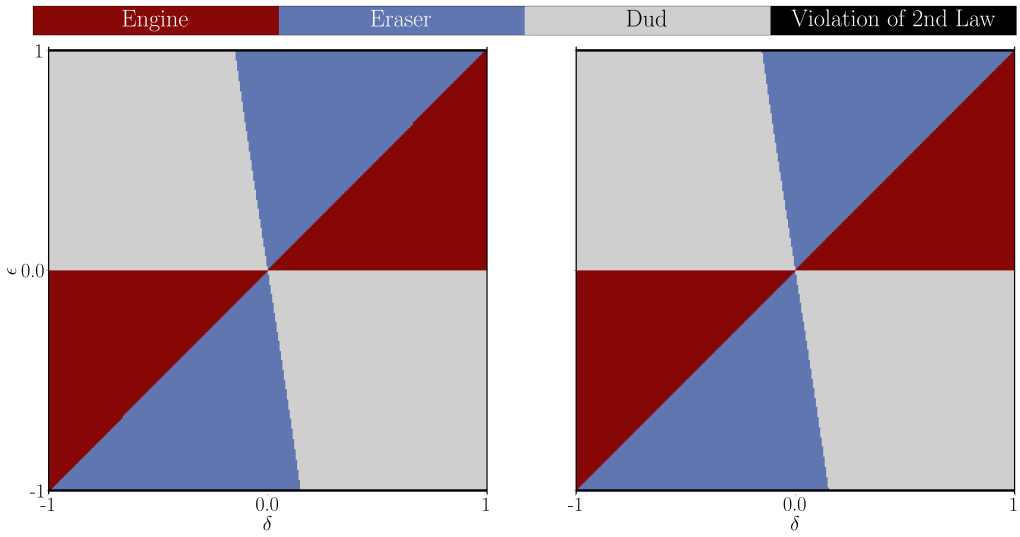


FIGURE H.1. MJ ratchet functional thermodynamic regions over $\epsilon \in [1, -1]$ and $b \in [0, 10]$ with $\tau = 1$. (Left) Purported functionality identified via by single-symbol entropy bound Eq. (8.2). (Right) Correct functionality identified via the entropy-rate bound IPSL Eq. (8.3).

ratchet with the Golden Mean Process in Fig. 8.3 illustrates this effect.) This is characteristic of any transducer defined via the rate transition matrix method outlined above. The conclusion is that the techniques required to analyze nonunifilar HMMs are required in general.

H.0.2. Biased Coin Parameter Sweep. As Section 8.4.1 discusses, we recreated the results from Mandal and Jarzynski's original ratchet [**121**] using the techniques outlined in this section and in Chapter 8. There, the ratchet is driven by a memoryless Biased Coin and the functional thermodynamic regions are identified via Eq. (8.2) [**121**]. These results are shown in Fig. H.1, on the left, and demonstrate close agreement with the original results. As previously noted, calculating the thermodynamic regions via Eq. (8.3) did not significantly change the identified regions, as can be seen by comparison to the figure on the right. Although not shown here, we also recreated the results at $\tau = 10$, which again show strong agreement with results reported in Ref. [**121**].

Bibliography

[1] C. E. Shannon. A mathematical theory of communication. *Bell Sys. Tech. J.*, 27:379–423, 623–656, 1948.

[2] D. Goroff, editor. *H. Poincaré, New Methods Of Celestial Mechanics, 1: Periodic And Asymptotic Solutions*. American Institute of Physics, New York, 1991.

[3] D. Goroff, editor. *H. Poincaré, New Methods Of Celestial Mechanics, 2: Approximations by Series*. American Institute of Physics, New York, 1993.

[4] D. Goroff, editor. *H. Poincaré, New Methods Of Celestial Mechanics, 3: Integral Invariants and Asymptotic Properties of Certain Solutions*. American Institute of Physics, New York, 1993.

[5] J. P. Crutchfield, N. H. Packard, J. D. Farmer, and R. S. Shaw. Chaos. *Sci. Am.*, 255:46 – 57, 1986.

[6] Ya. B. Pesin. *Dimension Theory in Dynamical Systems: Contemporary Views and Applications*. University of Chicago Press, 1997.

[7] F. Ledrappier and L. S. Young. The metric entropy of diffeomorphisms: Part ii: Relations between entropy, exponents and dimension. *Annals of Mathematics*, 122:540–574, 1985.

[8] D. P. Feldman and J. P. Crutchfield. Structural information in two-dimensional patterns: Entropy convergence and excess entropy. *Phys. Rev. E*, 67(5):051103, 2003.

[9] A. Rupe, J. P. Crutchfield, K. Kashinath, and Prabhat. A physics-based approach to unsupervised discovery of coherent structures in spatiotemporal systems. Santa Fe Institute Working Paper 2017-09-033; arXiv.org:1709.03184 [physics.flu-dyn].

[10] J. E. Hanson. *Computational Mechanics of Cellular Automata*. PhD thesis, University of California, Berkeley, 1993. Published by University Microfilms Intl, Ann Arbor, Michigan.

[11] A. Rupe and J. P. Crutchfield. Local causal states and discrete coherent structures. arXiv.org:1709.1801.00515 [cond-mat.stat-mech].

[12] A. Rupe and J. P. Crutchfield. Spacetime computational mechanics. 2017. in preparation.

[13] C. R. Shalizi and J. P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *J. Stat. Phys.*, 104:817–879, 2001.

[14] D. P. Feldman and J. P. Crutchfield. Measures of statistical complexity: Why? *Phys. Lett. A*, 238:244–252, 1998.

[15] T. Hogg and B. A. Huberman. Order, complexity, and disorder. 1985.

[16] P. Grassberger. Toward a quantitative theory of self-generated complexity. *Intl. J. Theo. Phys.*, 25:907, 1986.

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

[18] J. P. Crutchfield and N. H. Packard. Symbolic dynamics of noisy chaos. *Physica*, 7D(1-3):201 – 223, 1983.

[19] P. Szépfalusy and G. Györgyi. Entropy decay as a measure of stochasticity in chaotic systems. *Phys. Rev. A*, 33:2852–2855, 1986.

[20] S. Wolfram. Universality and complexity in cellular automata. *Physica*, 10D:1, 1984.

[21] R. Shaw. *The Dripping Faucet as a Model Chaotic System*. Aerial Press, Santa Cruz, California, 1984.

[22] C. H. Bennett. Dissipation, information, computational complexity, and the definition of organization. In D. Pines, editor, *Emerging Syntheses in the Sciences 1985*, pages 297–313. Addison-Wesley, Redwood City, 1988.

[23] K. Lindgren and M. G. Nordahl. Complexity measures and cellular automata. *Complex Systems*, 2:409, 1988.

[24] W. Li. On the relationship between complexity and entropy for Markov chains and regular languages. *Complex Systems*, 5(4):381–399, 1991.

[25] James P. Crutchfield. The calculi of emergence: Computation, dynamics, and induction. *Physica D*, 75:11–54, 1994.

[26] A. N. Kolmogorov. Three approaches to the concept of the amount of information. *Prob. Info. Trans.*, 1:1, 1965.

[27] G. Chaitin. On the length of programs for computing finite binary sequences. *J. ACM*, 13:145, 1966.

[28] P. Martin-Lof. The definition of random sequences. *Info. Control*, 9:602–619, 1966.

[29] A. N. Kolmogorov. Combinatorial foundations of information theory and the calculus of probabilities. *Russ. Math. Surveys*, 38:29–40, 1983.

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

[31] B. Marcus, K. Petersen, and T. Weissman, editors. *Entropy of Hidden Markov Process and Connections to Dynamical Systems*, volume 385 of *Lecture Notes Series*. London Mathematical Society, 2011.

[32] Y. Ephraim and N. Merhav. Hidden markov processes. *IEEE Trans. Info. Th.*, 48(6):1518–1569, 2002.

[33] J. Bechhoefer. Hidden markov models for stochastic thermodynamics. *New. J. Phys.*, 17:075003, 2015.

[34] L. R. Rabiner and B. H. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, January:4–16, 1986.

[35] E. Birney. Hidden Markov models in biological sequence analysis. *IBM J. Res. Dev.,*, 45(3.4):449–454, 2001.

[36] S. Eddy. What is a hidden Markov model? *Nature Biotech.*, 22:1315–1316, Oct 2004.

[37] C. Bretó, D. He, E. L. Ionides, and A. A. King. Time series analysis via mechanistic models. *Ann. App. Statistics*, 3(1):319–348, Mar 2009.

[38] T. Rydén, T. Teräsvirta, and S. Åsbrink. Stylized facts of daily return series and the hidden Markov model. *J. App. Econometrics*, 13:217–244, 1998.

[39] J. P. Crutchfield and E. van Nimwegen. The evolutionary unfolding of complexity. In L. F. Landweber and E. Winfree, editors, *Evolution as Computation: DIMACS Workshop, Princeton, 1999*, Natural Computing Series, pages 67–94, New York, 2002. Springer-Verlag. Santa Fe Institute Working Paper 99-02-015; adap-org/9903001.

[40] J. P. Crutchfield. The calculi of emergence: Computation, dynamics, and induction. *Physica D*, 75:11–54, 1994.

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

[42] R. G. James, J. R. Mahoney, C. J. Ellison, and J. P. Crutchfield. Many roads to synchrony: Natural time scales and their algorithms. *Physical Review E*, 89:042135, 2014. Santa Fe Institute Working Paper 10-11-025; arxiv.org:1010.5545 [nlin.CD].

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

[44] P. Riechers and J. P. Crutchfield. Spectral simplicity of apparent complexity, Part I: The nondiagonalizable metadynamics of prediction. *Chaos*, 28:033115, 2018. Santa Fe Institute Working Paper 2017-05-018; arxiv.org:1705.08042.

[45] P. Riechers and J. P. Crutchfield. Spectral simplicity of apparent complexity, Part II: Exact complexities and complexity spectra. *Chaos*, 28:033116, 2018. Santa Fe Institute Working Paper 17-06-019; arxiv.org:1706.00883.

[46] C. C. Strelioff, J. P. Crutchfield, and Alfred Hübler. Inferring markov chains: Bayesian estimation, model comparison, entropy rate, and out-of-class modeling. *Phys. Rev. E*, 76(1):011106, 2007.

[47] J. Ruebeck, R. G. James, John R. Mahoney, and J. P. Crutchfield. Prediction and generation of binary Markov processes: Can a finite-state fox catch a markov mouse? *CHAOS*, 28:013109, 2018. SFI Working Paper 2017-08-027; arxiv.org:1708.00113 [cond-mat.stat-mech].

[48] A. Jurgens and J. P. Crutchfield. Shannon entropy rate of hidden Markov processes. *J. Statistical Physics*, to appear, 2020. arXiv.org:2008.12886.

[49] A. Jurgens and J. P. Crutchfield. Ambiguity rate of hidden markov processes. *arXiv:2105.07294*, 2021.

[50] S. Marzen and J. P. Crutchfield. Statistical signatures of structural organization: The case of long memory in renewal processes. *Phys. Lett. A*, 380(17):1517–1525, 2016.

[51] A. Jurgens and J. P. Crutchfield. Divergent predictive states: The statistical complexity dimension of stationary, ergodic hidden Markov processes. *Chaos*, 2021.

[52] A. Venegas-Li, A. Jurgens, and J. P. Crutchfield. Measurement-induced randomness and structure in quantum dynamics. *Phys Rev E.*, 2020.

[53] A. Jurgens and J. P. Crutchfield. Functional thermodynamics of Maxwellian ratchets: Constructing and deconstructing patterns, randomizing and derandomizing behaviors. *Phys. Rev. Research*, 2(3):033334, 2020.

[54] D. Blackwell. The entropy of functions of finite-state Markov chains. volume 28, pages 13–20, Publishing House of the Czechoslovak Academy of Sciences, Prague, 1957. Held at Liblice near Prague from November 28 to 30, 1956.

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

[56] M.F. Barnsley, S.G. Demko, J.H. Elton, and J.S. Geronimo. Invariant measures arising from iterated function systems with place dependent probabilities. *Ann. Inst. H. Poincare*, 24:367–394, 1988.

[57] J. H. Elton. An ergodic theorem for iterated maps. *Ergod Th & Dynam Sys*, 7:481–488, 1987.

[58] J. P. Crutchfield, P. Riechers, and C. J. Ellison. Exact complexity: Spectral decomposition of intrinsic computation. *Phys. Lett. A*, 380(9-10):998–1002, 2016. Santa Fe Institute Working Paper 13-09-028; arXiv:1309.3792 [cond-mat.stat-mech].

[59] Garrett Birkhoff. Extensions of jentzsch's theorem. *Transations of the American Mathematical Society*, 85(1):219–227, 1957.

[60] S. E. Marzen and J. P. Crutchfield. Nearly maximally predictive features and their dimensions. *Phys. Rev. E*, 95(5):051301(R), 2017. SFI Working Paper 17-02-007; arxiv.org:1702.08565 [cond-mat.stat-mech].

[61] A. Jurgens and J. P. Crutchfield. Minimal embedding dimension of minimally infinite hidden markov processes. *in preparation*, 2020.

[62] N. F. Travers. Exponential bounds for convergence of entropy rate approximations in hidden markov models satisfying a path-mergeability condition. *Stochastic Proc. Appln.*, 124(12):4149–4170, 2014.

[63] N. Travers and J. P. Crutchfield. Infinite excess entropy processes with countable-state generators. *Entropy*, 16:1396–1413, 2014. SFI Working Paper 11-11-052; arxiv.org:1111.3393 [math.IT].

[64] A. Allahverdyan. Entropy of hidden Markov processes via cycle expansion. *J. Stat Phys.*, 133:535–564, 2008.

[65] A. Renyi. On the dimension and entropy of probability distributions. *Acta Math. Hung.*, 10:193, 1959.

[66] B. B. Mandelbrot. *The Fractal Geometry of Nature*. W. H. Freeman and Company, San Francisco, California, 1982.

[67] G. A. Edgar. *Measure, Topology, and Fractal Geometry*. Springer-Verlag, New York, 1990.

[68] K. Falconer. *Fractal geometry: Mathematical foundations and applications*. John Wiley, Chichester, 1990.

[69] I. Shimada and T. Nagashima. A numerical approach to ergodic problem of dissipative dynamical systems. *Prog. Theo. Phys.*, 61:1605, 1979.

[70] G. Benettin, L. Galgani, A. Giorgilli, and J.-M. Strelcyn. Lyapunov characteristic exponents for smooth dynamical systems and for hamiltonian systems; a method for computing all of them. *Meccanica*, 15:9, 1980.

[71] Ja. B. Pesin. Ljapunov characteristic exponents and ergodic properties of smooth dynamical systems with an invariant measure. *Doklady Akademii Nauk*, 226:774–777, 1976.

[72] D. Ruelle. Sensitive dependence on initial conditions and turbulent behavior of dynamical systems. *Annals of the New York Academy of Sciences*, 316(1):408–416, 1977.

[73] G. Benettin, L. Galgani, and J.-M. Strelcyn. Kolmogorov entropy and numerical experiments. *Phys. Rev. A*, 14(6):2338–2345, 1976.

[74] J. D. Farmer, E. Ott, and J. A. Yorke. The dimension of chaotic attractors. *Physica*, 7D:153, 1983.

[75] J. Kaplan and J. Yorke. *Springer Lecture Notes in Math.*, 730:204, 1979.

[76] P. Jacquet, G. Seroussi, and W. Szpankowski. On the entropy of a hidden Markov process. *Theo. Comp. Sci.*, 395:203–219, 2008.

[77] T. Holliday, A. Goldsmith, and P. Glynn. Capacity of finite state channels based on Lyapunov exponents of random matrices. *IEEE Trans. Info. Th.*, 52:3509–3532, 2006.

[78] B. Barany. On the ledrappier-young formula for self-affine measures. *Mathematical Proceedings of the Cambridge Philosophical Society*, 159(3):405–432, 2015.

[79] P. A. P. Moran. Additive functions of intervals and hausdorff measure. *Mathematical Proceedings of the Cambridge Philosophical Society*, 42(1):15–23, 1946.

[80] C. Bandt, N. Viet Hung, and H. Rao. On the open set condition for self-similar fractals. *Proceedings of the American Mathematical Society*, 134(5):1369–1374, 2006.

[81] J. Jaroszewska and M. Rams. On the hausdorff dimension of invariant measures of weakly contracting on average measurable ifs. *Journal of Statistical Physics*, 2008.

[82] E. Schrödinger. The present situation in quantum mechanics. *Proc. Am. Philos. Soc.*, 124:323, 1935.

[83] C. S. Wu and I. Shaknov. The angular correlation of scattered annihilation radiation. *Phys. Rev.*, 77:136, 1950.

[84] J. S. Bell. On the Einstein Podolsky Rosen paradox. *Physics*, 1:195, 1964.

[85] A. Einstein, B. Podolsky, and N. Rosen. Can quantum-mechanical description of physical reality be considered complete? *Phys. Rev.*, 47:777, 1935.

[86] A. K. Ekert. Quantum cryptography based on Bell's theorem. *Phys. Rev. Lett.*, 67:661, 1991.

[87] C. H. Bennett, G. Brassard, C. Crepeau, R. Jozsa, A. Peres, and W. K. Wootters. Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels. *Phys. Rev. Lett.*, 70:1895, 1993.

[88] V. Giovannetti, S. Lloyd, and L. Maccone. Quantum metrology. *Phys. Rev. Lett.*, 96:010401, 2006.

[89] R. Raussendorf and H. J. Briegel. A one-way quantum computer. *Phys. Rev. Lett.*, 86:5188, 2001.

[90] M. Lindemann, G. Xu, T. Pusch, R. Michalzik, M. R. Hofmann, I. Zutic, and N. C. Gerhardt. Ultrafast spin-lasers. *Nature*, 5678:212, 2019.

[91] S. J. Freedman and J. F. Clauser. Experimental test of local hidden-variable theories. *Phys. Rev. Lett.*, 28:938, 1972.

[92] R. Miller, T. E. Northup, K. M. Birnbaum, A. Boca, A. D. Boozer, and H. J. Kimble. Trapped atoms in cavity qed: coupling quantized light and matter. *J. Phys. B: At. Mol. Opt. Phys.*, 38(9):S551–S565, 2005.

[93] P. G. Kwiat, K. Mattle, H. Weinfurter, A. Zeilinger, A. V. Sergienko, and Y. Shih. New high-intensity source of polarization-entangled photon pairs. *Phys. Rev. Lett.*, 75:4337, 1995.

[94] O. Benson, C. Santori, M. Pelton, and Y. Yamamoto. Regulated and entangled photons from a single quantum dot. *Phys. Rev. Lett.*, 84:2513, 2000.

[95] H. Walther, B. T. H. Varcoe, B.-G. Englert, and T. Becker. Cavity quantum electrodynamics. *Rep. Prog. Phys.*, 69(5):1325–1382, 2006.

[96] M. D. Eisaman, J. Fan, A. Migdall, and S. V. Polyakov. Single-photon sources and detectors. *Rev. Sci. Instr.*, 82(071101), 2011.

[97] F. A. Pollock, C. Rodriguez-Rosario, T. Frauenheim, M. Paternostro, and K. Modi. Non-Markovian quantum processes: Complete framework and efficient characterization. *Phys. Rev. A*, 97:012127, 2018.

[98] M. A. Nielsen and I. L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, Cambridge, United Kingdom, Tenth anniversary edition, 2011.

[99] R. Rey de Castro, R. Cabrera, D. I. Bondar, and H. Rabitz. Time-resolved quantum process tomography using Hamiltonian-ecoding and observable-decoding. *New J. Phys.*, 15:025032, 2013.

[100] M. Holzapfel, T. Baumgratz, M. Cramer, and M.B. Plenio. Scalable reconstuction of unitary processes and Hamiltonians. *Phys. Rev. A*, 91, 2015.

[101] C. Granade, J. Combes, and D.G. Cory. Practical Bayesian tomography. *New J. Phys.*, 18, 2016.

[102] M. Kliesch, R. Kueng, J. Eisert, and D. Gross. Guaranteed recovery of quantum process from few measurements. *Quantum*, 3:171, 2019.

[103] A.M. Palmieri, E. Kovlakov, and F. Bianchi. Experimental neural network enhanced quantum tomography. *npj Quantum Inf*, 6:20, 2020.

[104] L.C.G. Govia, G.J. Ribeill, D. Riste, M. Ware, and H. Krovi. Bootrstrapping quantum process tomography via perturbative ansatz. *Nature Comm.*, 11:1084, 2020.

[105] IBM quantum experience. https://quantum-computing.ibm.com.

[106] Rigetti cloud services. https://rigetti.com.

[107] C. Galland, Y. Ghosh, A. Steinbru, M. Sykora, J. A. Hollingsworth, V. I. Klimov, and H. Htoon. Two types of luminescence blinking revealed by spectroelectrochemistry of single quantum dots. *Nature*, 479:203–207, 2011.

[108] A. L. Efros and D. J. Nesbitt. Origin and control of blinking in quantum dots. *Nature Nanotech.*, 11:661–671, 2016.

[109] M. Gu, K. Wiesner, E. Rieper, and V. Vedral. Quantum mechanics can reduce the complexity of classical models. *Nature Comm.*, 3(762):1–5, 2012.

[110] J. R. Mahoney, C. Aghamohammadi, and J. P. Crutchfield. Occam's quantum strop: Synchronizing and compressing classical cryptic processes via a quantum channel. *Scientific Reports*, 6:20495, 2016.

[111] P. M. Riechers, J. R. Mahoney, C. Aghamohammadi, and J. P. Crutchfield. Minimized state-complexity of quantum-encoded cryptic processes. *Phys. Rev. A*, 93(5):052317, 2016.

[112] C. Aghamohammadi, J. R. Mahoney, and J. P. Crutchfield. The ambiguity of simplicity. *Physics Letters A*, 381(14):1223–1227, 2017.

[113] S. Loomis and J. P. Crutchfield. Strong and weak optimizations in classical and quantum models of stochastic processes. *J. Stat. Phys.*, pages 1–26, 2019. arXiv:1808.08639.

[114] P. M. Riechers and J. P. Crutchfield. Beyond the spectral theorem: Decomposing arbitrary functions of nondiagonalizable operators. *AIP Advances*, 8:065305, 2018.

[115] S. Marzen and J. P. Crutchfield. Structure and randomness of continuous-time discrete-event processes. *J. Stat. Physics*, 169(2):303–315, 2017.

[116] J. C. Maxwell. *Theory of Heat*. Longmans, Green and Co., London, United Kingdom, ninth edition, 1888.

[117] Koji Maruyama, Franco Nori, and Vlatko Vedral. Colloquium: The physics of maxwell's demon and information. *Rev. Mod. Phys.*, 81:1–23, Jan 2009.

[118] O. Penrose. *Foundations of Statistical Mechanics; A Deductive Treatment*. Oxford: Pergamon, 1970.

[119] C. H. Bennet. Thermodynamics of computation—-a review. *Int. J. Theor. Phys.*, 21:905–940, 1982.

[120] R. Landauer. Irreversibility and heat generation in the computing process. *IBM J. Res. Develop.*, 5(3):183–191, 1961.

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

[122] A. B. Boyd, D. Mandal, and J. P. Crutchfield. Identifying functional thermodynamics in autonomous Maxwellian ratchets. *New J. Physics*, 18:023049, 2016. SFI Working Paper 15-07-025; arxiv.org:1507.01537 [cond-mat.stat-mech].

[123] A. B. Boyd, D. Mandal, and J. P. Crutchfield. Correlation-powered information engines and the thermodynamics of self-correction. *Phys. Rev. E*, 95(1):012152, 2017.

[124] A. B. Boyd, D. Mandal, P. M. Riechers, and J. P. Crutchfield. Transient dissipation and structural costs of physical information transduction. *Phys. Rev. Lett.*, 118:220602, 2017. arXiv.org:1612.08616 [cond-mat.stat-mech].

[125] A. B. Boyd, D. Mandal, and J. P. Crutchfield. Leveraging environmental correlations: The thermodynamics of requisite variety. *J. Stat. Phys.*, 167(6):1555–1585, 2016. SFI Working Paper 16-09-021; arxiv.org:1609.05353 [cond-mat.stat-mech], http://rdcu.be/qJj6.

[126] A. B. Boyd, D. Mandal, and J. P. Crutchfield. Above and beyond the landauer bound: Thermodynamics of modularity. 2017. SFI Working Paper 2017-08-030; arxiv.org:1708.03030 [cond-mat.stat-mech].

[127] L. B. Kish and C. G. Granqvist. Energy requirement of control: Comments on szilard's engine and maxwell's demon. *EPL (Europhysics Letters)*, 98(6):68001, Jun 2012.

[128] Takahiro Sagawa and Masahito Ueda. Fluctuation theorem with information exchange: Role of correlations in stochastic thermodynamics. *Phys. Rev. Lett.*, 109:180602, Nov 2012.

[129] Anupam Kundu. Nonequilibrium fluctuation theorem for systems under discrete and continuous feedback control. *Phys. Rev. E*, 86:021107, Aug 2012.

[130] David Abreu and Udo Seifert. Thermodynamics of genuine nonequilibrium states under feedback control. *Phys. Rev. Lett.*, 108:030601, Jan 2012.

[131] Suriyanarayanan Vaikuntanathan and Christopher Jarzynski. Modeling maxwell's demon with a microcanonical szilard engine. *Phys. Rev. E*, 83:061120, Jun 2011.

[132] D. Abreu and U. Seifert. Extracting work from a single heat bath through feedback. *EPL (Europhysics Letters)*, 94(1):10001, Mar 2011.

[133] Léo Granger and Holger Kantz. Thermodynamic cost of measurements. *Phys. Rev. E*, 84:061110, Dec 2011.

[134] Jordan M. Horowitz and Juan M. R. Parrondo. Thermodynamic reversibility in feedback processes. *EPL (Europhysics Letters)*, 95(1):10005, Jun 2011.

[135] M. Ponmurugan. Generalized detailed fluctuation theorem under nonequilibrium feedback control. *Phys. Rev. E*, 82:031129, Sep 2010.

[136] S. Toyabe, T. Sagawa, M. Ueda, E. Muneyuki, and M. Sano. Experimental demonstration of information-to-energy conversion and validation of the generalized Jarzynski equality. *Nature Physics*, 6:988–992, 2010.

[137] Takahiro Sagawa and Masahito Ueda. Generalized jarzynski equality under nonequilibrium feedback control. *Phys. Rev. Lett.*, 104:090602, Mar 2010.

[138] Francisco J. Cao, Luis Dinis, and Juan M. R. Parrondo. Feedback control in a collective flashing ratchet. *Phys. Rev. Lett.*, 93:040603, Jul 2004.

[139] Hugo Touchette and Seth Lloyd. Information-theoretic limits of control. *Physical Review Letters*, 84(6):1156–1159, Feb 2000.

[140] Sosuke Ito and Takahiro Sagawa. Information thermodynamics on causal networks. *Phys. Rev. Lett.*, 111:180603, Oct 2013.

[141] D Hartich, A C Barato, and U Seifert. Stochastic thermodynamics of bipartite systems: transfer entropy inequalities and a maxwell's demon interpretation. *Journal of Statistical Mechanics: Theory and Experiment*, 2014(2):P02016, Feb 2014.

[142] Jordan M. Horowitz and Massimiliano Esposito. Thermodynamics with continuous information flow. *Phys. Rev. X*, 4:031015, Jul 2014.

[143] Jordan M Horowitz. Multipartite information flow for multiple maxwell demons. *Journal of Statistical Mechanics: Theory and Experiment*, 2015(3):P03006, Mar 2015.

[144] Philipp Strasberg, Gernot Schaller, Tobias Brandes, and Massimiliano Esposito. Thermodynamics of a physical model implementing a maxwell demon. *Phys. Rev. Lett.*, 110:040601, Jan 2013.

[145] Massimiliano Esposito and Gernot Schaller. Stochastic thermodynamics for "maxwell demon" feedbacks. *EPL (Europhysics Letters)*, 99(3):30003, Aug 2012.

[146] Jordan M Horowitz and Henrik Sandberg. Second-law-like inequalities with information and their interpretations. *New Journal of Physics*, 16(12):125007, Dec 2014.

[147] A. C. Barato and U. Seifert. Unifying three perspectives on information processing in stochastic thermodynamics. *Phys. Rev. Lett.*, 112:090601, Mar 2014.

[148] Jordan M. Horowitz, Takahiro Sagawa, and Juan M. R. Parrondo. Imitating chemical motors with optimal information motors. *Phys. Rev. Lett.*, 111:010602, Jul 2013.

[149] Dibyendu Mandal, H. T. Quan, and Christopher Jarzynski. Maxwell's refrigerator: An exactly solvable model. *Physical Review Letters*, 111(3), Jul 2013.

[150] A. C. Barato and U. Seifert. An autonomous and reversible maxwell's demon. *EPL (Europhysics Letters)*, 101(6):60001, Mar 2013.

[151] J. Hoppenau and A. Engel. On the energetics of information exchange. *EPL (Europhysics Letters)*, 105(5):50002, Mar 2014.

[152] Z. Lu, D. Mandal, and C. Jarzynski. Engineering Maxwell's demon. *Phys. Today*, 67:60, August 2014.

[153] Jaegon Um, Haye Hinrichsen, Chulan Kwon, and Hyunggyu Park. Total cost of operating an information engine. *New Journal of Physics*, 17(8):085001, Aug 2015.

[154] Z. Lu, D. Mandal, and C. Jarzynski. Engineering Maxwell's demon. *Physics Today*, 67(8):60–61, January 2014.

[155] Yuansheng Cao, Zongping Gong, and H. T. Quan. Thermodynamics of information processing based on enzyme kinetics: An exactly solvable model of an information pump. *Phys. Rev. E*, 91:062117, Jun 2015.

[156] Naoto Shiraishi, Sosuke Ito, Kyogo Kawaguchi, and Takahiro Sagawa. Role of measurement-feedback separation in autonomous maxwell's demons. *New Journal of Physics*, 17(4):045012, Apr 2015.

[157] Giovanni Diana, G. Baris Bagci, and Massimiliano Esposito. Finite-time erasing of information stored in fermionic bits. *Phys. Rev. E*, 87:012111, Jan 2013.

[158] A. Chapman and A. Miyake. How can an autonomous quantum Maxwell demon harness correlated information? arXiv:1506.09207, 2015.

[159] J. M. R. Parrondo, J. M. Horowitz, and T. Sagawa. Thermodynamics of information. *Nature Physics*, 11:131–139, 2015.

[160] P. M. Riechers. *Transforming Metastable Memories: The Nonequilibrium Thermodynamics of Computation*. SFI Press, 2019. arXiv:1808.03429.

[161] W. R. Ashby. *An Introduction to Cybernetics, 2nd Edn.* Wiley, New York, 1960.

[162] M. Ehrenberg and C. Blomberg. Thermodynamic constraints on kinetic proofreading in biosynthetic pathways. *Biophys. J.*, 31:333–358, 1980.

[163] M. Rezaeian. Hidden Markov process: A new representation, entropy rate and estimation entropy. *arXiv:0606114*.

[164] W. Słomczyński, J. Kwapień, and K. Życzkowski. Entropy computing via integration over fractal measures. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 10(1):180–188, Mar 2000.

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

[166] P. M. Ara, R. G. James, and J. P. Crutchfield. The elusive present: Hidden past and future dependence and why we build models. *Phys. Rev. E*, 93(2):022143, 2016. SFI Working Paper 15-07-024; arxiv.org:1507.00672 [cond-mat.stat-mech].

[167] R. G. James, N. Barnett, and J. P. Crutchfield. Information flows? A critique of transfer entropies. *Phys. Rev. Lett.*, 116(23):238701, 2016. SFI Working Paper 16-01-001; arxiv.org:1512.06479 [cond-mat.stat-mech].

[168] R. G. James and J. P. Crutchfield. Multivariate dependence beyond shannon information. *Entropy*, 19:531, 2017. Santa Fe Institute Working Paper 16-09-017; arxiv.org:1609.01233 [math.IT].

[169] M. Minsky. *Computation: Finite and Infinite Machines*. Prentice-Hall, Englewood Cliffs, New Jersey, 1967.

[170] Rolando Cavazos-Cadena. An alternative derivation of birkhoff's formula for the contraction coefficient of a positive matrix. *Linear Algebra and its Applications*, 375:291–297, 2003.

[171] E. Kohlberg and J. W. Pratt. The contraction mapping approach to the Perron-Frobenius theory: Why Hilbert's metric? *Math. Oper. Res.*, 7(2), 1982.

[172] V. I. Oseledets. A multiplicative ergodic theorem. lyapunov characteristic numbers for dynamical systems. *Trans. Moscow Math. Soc.*, 19:197–231, 1968.

[173] G. Froyland and K. Aihara. Rigorous numerical estimation of lyapunov exponents and invariant measures of iterated function systems and random matrix products. *Intl. J. Bifn. Chaos*, 10(1):103–122, 2000.

[174] A. Boyarsky and Y. S. Lou. A matrix method for approximating fractal measures. *Intl. J. Bifn. Chaos*, 02:167–175, 1992.

[175] Shapely: Geometric objects, predicates, and operations. version 1.7.1. https://doi.org/10.21105/joss.00738, https://pypi.org/project/Shapely/.

[176] N. Barnett and J. P. Crutchfield. Computational mechanics of input-output processes: Structured transformations and the ϵ-transducer. *J. Stat. Phys.*, 161(2):404–451, 2015.