Title

Invariant Properties of Ergodic Processes, with Applications to Quantum Computing, Data Science and Emissions Modeling

Author

Loomis, Samuel

Publication Date

2022

Peer reviewed|Thesis/dissertation

INVARIANT PROPERTIES OF ERGODIC PROCESSES, WITH APPLICATIONS TO QUANTUM COMPUTING, DATA SCIENCE AND EMISSIONS MODELING

By

SAMUEL P. LOOMIS
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

Daniel Cox

Mark Cooper

Committee in Charge

2022

"Sometimes I go about in pity for myself,
and all the while
a great wind carries me across the sky."

To my parents

Contents

INVARIANT PROPERTIES OF ERGODIC PROCESSES, WITH APPLICATIONS TO QUANTUM COMPUTING, DATA SCIENCE AND EMISSIONS MODELING

Abstract

The mathematics which underly the intrinsic structures of stochastic processes and dynamics of probability are further developed, and broad applications are considered. I provide a general and rigorous definition of predictive states in stochastic processes, and demonstrate how they may be reliably and convergently estimated from time-series data. I connect this to new developments in the machine learning of dynamical systems. I further demonstrate that the dynamics of predictive states for a given stochastic process generates an algebraic structure, the observable semigroup, and show that this constrains the structure of physical systems which can generate said process. I apply this result to studying quantum machines which generate stochastic processes. By combining the algebra of the semigroup with that of majorization theory, I show that the constraints of the semigroup induce minimal costs in memory and energy required for these machines, and I compare these costs with classical machines, finding overall quantum advantage in memory but more ambiguous results in energy. I close by returning to questions of data science, and show how the mathematics of stochastic processes and majorization can help separate genuine structure from artifact in models of carbon footprints derived from global trade data.

Acknowledgments

I hated math—to me it was frustrating tables I couldn't memorize and quizzes I couldn't complete in time. My father, seeking to remedy this, bought me a book: Theoni Pappas's "The Joy of Mathematics." Within a year or two I was telling my middle-school counselor that my career path was "theoretical physicist." I have been told I am prone to over-correction.

At the time we had just moved to the Research Triangle in North Carolina, not for the burgeoning business scene but because of my need for special therapy due to my autism. I won't claim sole credit for the move, but it was just one of the many ways that my parents sacrificed and risked much for their children. I know there were many around that time who had doubted that I might ever live independently—much less that I might one day be writing these acknowledgements while flying back to North Carolina next to my wife and dog, after spending six years in California pursuing the degree I first started dreaming of in sixth grade. I don't know where I would be without the stalwart love and support my parents have always shown, and all that they have given. No "thank you" can ever cover that, but I may as well try. Thank you.

To my sister, Chelsea—I guess some of your passion for teaching rubbed off on me; even in some of the toughest quarters these past years, I've always found joy in the job. I credit you for that, so thank you.

To my brother, Jack—your imagination has always inspired mine. Even when we've been on our own paths, I'm looking for new adventures. I know Tolkein says "not all who wander are lost"—I think our wandering is what *keeps* us from getting lost. I know it did that for me these past years. Thank you.

The integrity of my sanity is entirely attributable to my wife, Catie. I hope I've been as stabilizing a force for her as well. I think there was no accident in the way that we finally took notice of each other at the very end of our undergraduate years, before each embarking on our own graduate journey. Those separate journeys eventually brought us back together in Davis. Every day she inspires me to be a better version of myself. Writing this as we finally fly back to the place we both call home, I am so grateful for the journeys we've shared and excited for the journeys to come. Thank you for keeping me alive and sane these last six years.

To Ruby, our chiweenie—thanks for being my buddy for the last year, and for rigorously enforcing social distancing by borking at anyone who gets within six feet of me. Even though, just so you know, that's wasn't strictly necessary anymore even when we got you. Thanks anyways.

I've been fortunate to work with a number of excellent colleagues and mentors during my time as a student. They say it takes a village to raise a child. I've learned that in scholarship, it takes a village to make much more than an abstract. There are therefore a lot of people responsible, behind the scenes, for this dissertation.

Dr. Brown—thanks for giving me a place to nerd out about general relativity in undergrad, and for teaching me how to find the fun in physics even at its most frustrating.

Dr. Carlip—thanks for your guidance and mentorship in early graduate school, which helped me get a clearer sense of where I was going and what I was doing.

Dr. Fushing—thanks for fun discussions about data science and encouraging me to get my head out of theory once in a while and go get some real data.

Mark—thanks for all the great discussions about geography, data epistemology, and philosophy of science, and encouraging me to pursue the work that became Chapter 7 of this dissertation.

Jim—thanks for the mentorship and support, for helping me navigate the Kafkaesque bureaucracies of academia, and for teaching me a whole different way of thinking about physics.

John and Cina—thanks for taking me under your wing when I first came to the group, getting me set up with the projects that would become much of this thesis (chapters 4,5 and 6).

My post-doc mentors Ryan and Fabio—thanks for being my gurus in information theory and quantum information, who I could bug at any time with my oddball questions and harebrained schemes, and for being full of sage advice on graduate research to boot.

David, Alex, Mikhael and Tamara—thanks for all the great discussions, movie nights, and meetups, and for being a solid "quarantine crew" through thick and thin.

To the rest of the Complexity Sciences Center (and affiliates)—thanks for the group meetings, the Complexi-teas, and providing an encouraging and stimulating research environment.

CHAPTER 1

Gliding o'er all: Ergodicity and structure in nonlinear systems

Gliding o'er all, through all,
Through Nature, Time, and Space,
As a ship on the waters advancing,
The voyage of the soul—not life alone,
Death, many deaths I'll sing.

Walt Whitman, *Leaves of Grass*

1.1. Overture

It would be pointless to pretend that the work which comprises this dissertation was undertaken with a single unified intent. There are, however, common themes which underly the disparate subjects covered herein. Three, in particular, are worth drawing out clearly for the reader.

The first major theme of this work centers on the fact that, and methods for how, we can fundamentally understand nonlinear systems with linear algebraic tools. This is less of a contradiction than it seems at first glance. For better or worse, the methods and models we so often use to comprehend our complex, nonlinear world are awash in uncertainty. This uncertainty arises typically from a lack of sufficient data, and is amplified by nonlinear structural processes such as chaos and interconnectedness. Uncertainty, in turn, is quantified by probabilities.

It is at this level of perspective that nonlinear systems *become* linear: probabilities are inherently linear objects, which follow linear laws, even when describing the most intricately intertwined phenomena. Linear structures are also often the most easily generalizable, gliding easily between different representations of the same system, finding relevance in all. Like a "ship on the waters" they can smoothly transition between ports of perspective or advance forwards in time, even as the systems beneath them roil with turbulence.

The second major theme of this work is, in many ways, a corollary of the first: we can use the algebra of probability theory to better understand the structure and behavior of *models*. Models are often taken in science as fundamental tools, which must be taken for granted as we use them to study a given system; the only question is whether the model is appropriate or not. Less frequently do we undertake the "voyage of the soul" necessary to examine the nature of models themselves. The contrary perspective of this thesis is that models can in fact be objects of rich study.

The models we construct of our complex world so often go uninterrogated in any dimension before their application to "real data," but even in the midst of sophisticated statistics there is a pronounced lack of probabilistic perspective in analyzing the consequences and constraints of models. Linear constraints often have much to say about a system's dynamics, and when we tune into the probabilistic stratum, we will find that they positively sing: telling us what it may cost to implement a model, or what we can gain, and most importantly what types of systems the model can even accurately describe without a loss of information.

These questions are, of course, ambitious, and so to make any progress I have constrained this work to a particular kind of process, typically called "stationary and ergodic." *Stationary* means that the system's behavior is invariant with respect to translations in time. *Ergodic* means that the range of possible system behaviors will, ultimately, be fully explored over time. These are necessary assumptions if we are to construct accurate models of systems from data collected over time. We can think of these processes as those for which a definitive mathematical link can be constructed between the superficial, empirical presentation of a process (the collected data) and the possible underlying models.

Thus, the third theme of this thesis applies specifically to ergodic processes, and is the *self-similarity of time*. Generally, the term self-similar is used to describe fractal structures, often spatial in nature, which contain as components smaller copies of themselves. However, the recurrent nature of stationary and ergodic processes means that the temporal stream of data they produce has its own self-similar structure; strip off the first N observations, and the statistics of the remaining data is unchanged. This self-similarity has crucial implications both for our mathematical understanding of these datasets, but also for the nature of memory and persistence in the models that produce said data. In these processes, memory may be long, but it is *always* transient: all information about

past goings-on is erased to make way for repetition. The "many deaths" of ergodic processes will be recurrent throughout this work, being central to the results each chapter.

When these three themes come together, they provide a powerful framework for thinking about many sorts of processes in both mathematics and the physical world. This approach is primarily phenomenological, not concerned so much with uncovering hidden models as describing the behaviors we find directly in the data we handle. Despite this, our approach will also engage a wide range of technical mathematics. I have tried to keep the most specific technicalities quarantined to their respective chapters. However, in the remainder of this introduction I will review some concepts that will prove useful in multiple chapters.

1.2. Basic notations and concepts

In the wide range between quantum computers to carbon footprints there is a lot of conceptual ground to cover. The advantage of a mathematical perspective is that we can recycle concepts and notation like old grocery bags, whenever they are suitable to carry the relevant subjects on the page. I will have mercy on the reader and avoid carrying this sort of abstraction too far (for that way lies category theory)—but it will suffice to say that we shall always be concerned with systems that may have a variety of *states*, circumscribed by some *state space*, and we will be interested in transformations on these systems, characterized by *mappings* between state spaces.

To that end we will strive for some consistent notation. States will be denoted by lowercase letters, either Greek ($\alpha, \beta, \gamma,$ *etc.*) or Latin ($a, b, c,$ *etc.*); state spaces will be denoted by capital calligraphic letters ($\mathcal{A}$,$\mathcal{B}$,$\mathcal{C}$, *etc.*); mappings will be denoted by capital italic letters ($F, G, H,$ *etc.*).

Functions in general have a more relaxed treatment. Certain functions which appear frequently and with a standard definition will be indicated by roman letters, such as the entropy of a random variable $\mathrm{H}[X]$ or the probability of an event $\Pr(x)$. In other cases, when dealing with functions as abstract objects of interest, we will just adopt the usual f, g, h notation.

Sometimes it is useful to bundle a bunch of objects together in a tuple, such as in the case of hidden Markov models, which are a triple $\left(\mathcal{S}, \mathcal{X}, \left\{ T^{(x)} \right\}\right)$ of a state set, a symbol set, and a set of stochastic transformations, respectively. We will often label such tuples with a Fraktur letter ($\mathfrak{A}, \mathfrak{B}, \mathfrak{C},$ *etc.*) to emphasize the composite nature of the object.

A special conceptual space must be afforded to the spaces $\mathbb{R}^n$; that is, vector spaces over reals with n dimensions. These spaces are particularly useful for representing probability distributions over finite sets. Elements of these spaces will be represented by lowercase bold letters ($\mathbf{a}, \mathbf{b}, \mathbf{c}$, *etc.*); however, it will also be useful to refer to their components, which we will do with the notation $\mathbf{a} = (a_i)$, which indicates that a_i is the ith component of $\mathbf{a}$ in $\mathbb{R}^n$. Sometimes we will write $\mathbf{a} = (a_i)_{i \in \mathcal{S}}$ for some finite set $\mathcal{S}$; this still denotes a finite vector, but means that the index i instead refers directly to the elements of $\mathcal{S}$ instead of being numerically indexed.

Since linear maps between $\mathbb{R}^n$ and $\mathbb{R}^m$ are representable by $m \times n$ matrices, we will denote them similarly to vectors in the form $\mathbf{T} = (T_{i,j})$, with capital letters and multiple indices instead.

A similar notation will be used to describe *sequences*. Given a set $\mathcal{S}$, a sequence taking values in $\mathcal{S}$ will be denoted with an arrow as $\overrightarrow{x} = (x_i)_{i \in \mathbb{N}}$. This indicates that the ith element of the sequence $\mathbf{x}$ is given by x_i. The $\mathbb{N}$ indicates that $i = 1, 2, 3, \ldots$. Sometimes we will write $\overleftrightarrow{x} = (x_i)_{i \in \mathbb{Z}}$ to denote a bi-infinite sequence, which stretches infinitely in both directions: $i = \ldots, -1, 0, 1, \ldots$. The space of sequences over $\mathcal{S}$ is denoted $\mathcal{S}^{\mathbb{N}}$ and the space of bi-infinite sequences is $\mathcal{S}^{\mathbb{Z}}$. We can select subsequences of a sequence $\mathbf{x}$ using a slicing notation similar to that found in `python` and `MATLAB`: $\overrightarrow{x}_{k:\ell}$ denotes the sub-sequence $(x_k, \ldots, x_{\ell-1})$. We will often just write $x_k \ldots x_{\ell-1}$ without the commas or parentheses. $\overrightarrow{x}_k$ just denotes the infinite subsequence $(x_{k+i-1})_{i \in \mathbb{N}}$.

In addition to the three core themes addressed in the previous section, I will now overview here three threads of technical thought which will be utilized throughout this dissertation. The first is the formalism of *stochastic processes* and, related, the idea of the *predictive state*. The second thread is the multi-faceted concept of *ergodicity*, and a very important theorem which relates all these facets, called the *Perron-Frobenius theorem*. Lastly I will cover some of the basic principles of *resource theory* and *information theory*, which jointly describe how probabilistic systems balance abstract informational resources against physical constraints. This chapter is not a literature review, and the reader will be pointed to relevant technical literature in the relevant introductions of each chapter.

1.3. Stochastic processes

The field of stochastic processes [90, 175] can become easily fraught with mathematical formalism. There will be a time for some of that formalism in the next chapter, but we need not worry about it in order to get the idea of a stochastic process.

Let's just define a process as a data source. The source produces a steady sequence of observations. For instance, when I walk my dog every morning, I observe the weather. In the little town of Davis, this creates a process whose sequence looks like

$$(\ldots, \text{SUNNY}, \text{SUNNY}, \text{SUNNY}, \text{SUNNY}, \text{SUNNY}, \text{SUNNY}, \text{SUNNY}, \ldots)$$

in the summer,

$$(\ldots, \text{RAINY}, \text{RAINY}, \text{SUNNY}, \text{RAINY}, \text{SUNNY}, \text{RAINY}, \text{RAINY}, \ldots)$$

in the winter, and something more like

$$(\ldots, \text{RAINY}, \text{SUNNY}, \text{SUNNY}, \text{SUNNY}, \text{RAINY}, \text{SUNNY}, \text{SUNNY}, \ldots)$$

around spring and autumn.

Notice that a process necessarily involves both a system (the local weather) and a measurement (my ambulatory observations). In what follows, we will be considerably agnostic as to the nature of this underlying system and the measurement used to observe it. Our approach to studying stochastic processes makes little effort to disentangle the system and the measurement. Rather, we will seek to describe the structures in the data observed—whatever they are, whatever their origin. This model-agnostic approach will guide our developments in Chapters 2 and 3. In Chapters 4, 5 and 6 we will pay a fair bit more attention to the underlying system-and-measurement assemblage which is actually generating our process, but we will use our understanding gleaned from Chapter 2 to learn about the informational and physical costs imposed on any system which is capable of producing the observed data. This reverses the typical direction of things, as we use the data to study the possible models, rather than using the model to analyze the data.

For now, though, we will just take a birds-eye view of things. Generally we will suppose that the observations which comprise our data are drawn from a particular set, which we may call the *alphabet*, and denote $\mathcal{X}$. A process is then just a thing which, when we encounter it, begins to generate a sequence $\overrightarrow{x} = (x_t)$ from $\mathcal{X}^{\mathbb{N}}$. A finite subsequence $x_1 \dots x_L$ will be called a *word*.

Now, given a sequence (which we will conveniently suppose is infinite but could just be very long), we can ask about its *statistics*. Generally speaking, given any quantitative measurement of my observations, we can average the value of that measurement over time to get its *mean* which is the most fundamental statistic. A quantitative measurement can be thought of as any function $f : \mathcal{X}^L \to \mathbb{R}$ which takes a length-L subsequence of observations and returns a number. Then the mean of f is

$$\mathbb{E}[f] = \lim_{N\to\infty} \frac{1}{N} \sum_{t=0}^{N} f(x_t \dots x_{t+L-1})$$

(The E means *expectation*, which is a silly notation since we don't actually expect the mean most times, but it is traditional.)

For instance, sunny days can be around 85 degrees and rainy days around 65 degrees. We could built a few types of statistics from this. If we let $T(\text{SUNNY}) = 85$ and $T(\text{RAINY}) = 65$ then $\mathbb{E}[T]$ just gives us the average temperature over time in Davis. We could also define a rolling average

$$T_L(x_1 \dots x_L) = \frac{1}{L} \sum_{k=1}^{L} T(x_k)$$

but then $\mathbb{E}[T_L]$ would just be the same as $\mathbb{E}[T]$, which is not very interesting. More interesting would be the rolling *variance*, given by

$$V_L(x_1 \dots x_L) = \frac{1}{L} \sum_{k=1}^{L} (T(x_k) - T_L(x_1 \dots x_L))^2$$

Then $\mathbb{E}[V_L]$ would tell you something about the weekly weather volatility in Davis. (It is not very high.)

A very commonly used statistic is the probability. If the set $\mathcal{U} \subseteq \mathcal{X}^L$ is some "bag of words" then we can ask how frequently the observed word (that is, the L most recent observations) is in that bag. Using the *indicator function* $\mathbf{1}_{\mathcal{U}}(x_1 \dots x_L)$, which equals 1 when $x_1 \dots x_L$ is in $\mathcal{U}$ and 0 otherwise, we can compute the mean $\mathbb{E}[\mathbf{1}_{\mathcal{U}}]$. This just means counting the number of times that $x_1 \dots x_L$ is in

$\mathcal{U}$ and dividing by the total number of observations. Rather than actually using the cumbersome indicator function every time we want to talk about this proportion, we denote it as:

$$\Pr\left(x_t \dots x_{t+L-1} \in \mathcal{U} \right) = \frac{\#\text{ of times } x_t \dots x_{t+L-1} \in \mathcal{U}}{\#\text{ of observations}}$$

If $\mathcal{X}$ is *discrete*, like in the case of our weather example, then we can simplify this notation greatly by simply asking how many times we observed a particular word. In that case we just write

$$\Pr\left(x_1 \dots x_L \right) = \frac{\#\text{ of times } x_1 \dots x_L \text{ observed}}{\#\text{ of observations}}$$

and dispense with any set notation whatsoever. These statistics have some key properties; namely, they are always non-negative, and if we sum over all words of a given length, the probabilities add to one:

$$\sum_{w \in \mathcal{X}^L} \Pr\left(w \right) = 1$$

(This also implies that $\Pr\left(w \right) \leq 1$ for all words w.)

Now, I have made much ado of the words "stationary and ergodic" in reference to processes. The concepts defined above allow us to give proper meaning to these words. Generally one uses the word "stationary" to describe any process for which the computed statistics do not depend on when I start computing them. Now, we must be careful not to be too strict about this. I take a rather loose definition of stationary: if for every measurement $f : \mathcal{X}^L \to \mathbb{R}$ (over any word length), the time-dependent expectation

$$\mathbb{E}_{\mathrm{t:N}}[f] = \frac{1}{N-t} \sum_{k=t}^{N} f(x_k \dots x_{k+L-1})$$

(note that the sum starts at t!) has a well-defined limit as $N \to \infty$ (that is, as the sample size becomes longer), and that limit is *independent* of t, then the process is stationary.

For example, there was a time when the weather in the region now called Davis was probably more-or-less stationary. Certainly, the short-term averages depended greatly on when you measured them, but the *long-run* averages (which is exactly what $\mathbb{E}[f]$ denotes) did not. Of course, that has changed in recent years; one who starts measuring the long-run temperatures in Davis, or the long-run proportion of rainy days, in the 21st century will arrive at distinctly warmer (or dryer)

results than in previous years. The global climate has never been *fully* stationary, but recent years have very much seen an accelerated change in its underlying behavior.

To discuss ergodicity, we must define a new kind of statistic, which can be built up from the standard probabilities. Let $w_1 = x_1 \dots x_L$ and $w_2 = x_{L+1} \dots x_{L+K}$ be two words. Then we denote

$$\Pr(\, w_2 \mid w_1 \,) = \frac{\Pr(\, w_1 w_2 \,)}{\Pr(\, w_1 \,)}$$

The pair $w_1 w_2$ denotes the concatenated word $x_1 \dots x_L x_{L+1} \dots x_{L+K}$. Notice that the number of times we observe $w_1 w_2$ cannot be more than the number of times that we observe w_1; further, if we sum $\Pr(\, w_2 \mid w_1 \,)$ over all w_2 of length K, these quantities must add up to 1, because then we are just counting how many times the word w_1 appears followed by *anything* (and it is always followed by something). We can call $\Pr(\, w_2 \mid w_1 \,)$ a *conditional probability*, and interpret it as telling us the proportion of times we see w_2 following w_1, out of all the times we have seen w_1.

Now, let us consider the set of all words of length L which have positive probability, which we denote by $\mathcal{X}_L$. Consider the conditional probabilities $\Pr(\, w_2 \mid w_1 \,)$ for every pair of such words. We can construct a graph (that is, a network) where each node is a length L word, and there is a directed edge pointing from w_1 to w_2 if $\Pr(\, w_2 \mid w_1 \,) > 0$ (that is, if we *ever* see w_2 following w_1). Now examine this graph. If you can draw a path from any one word to any other word by following the directed edges, then we say the graph is *strongly connected*. In this case, this would tell us that starting from any observed word of length L, we are likely to eventually see every other word in $\mathcal{X}_L$. If the graph of $\Pr(\, w_2 \mid w_1 \,)$ is strongly connected, then we say the process is *ergodic*. What that means is that over time we will observe every possible behavior, and then we will eventually see that behavior again. In the following section, we'll discuss some of the more specific consequences of this strong temporal connectivity and recurrence, but before we do so it will be helpful to give one further note on stochastic processes.

Suppose we try to use the length-R block probabilities $\Pr(\, w_2 \mid w_1 \,)$ to *recreate* the process. Will the result have the same statistics as the original? What I mean by this is: suppose that we

(1) generate a length-R word, w_1, using a random number generator and the distribution $\Pr(\, w \,)$,

(2) then generate another word w_2 using the distribution $\Pr(\, w_2 \mid w_1 \,)$,

(3) concatenate that with the existing sequence,

(4) repeat the last two steps indefinitely with the most recent word as w_1.

To assess the feasibility of reproducing the process in this way, it is helpful to note the following telescopic identity for any word of length N:

$$\Pr(x_1 \dots x_N) = \Pr(x_1) \Pr(x_2 \mid x_2) \Pr(x_3 \mid x_1 x_2) \dots \Pr(x_N \mid x_1 \dots x_{N-1})$$

Now, the process generated by the outlined steps will have the following simplifying rule: for all $N \geq R$,

$$\Pr(x_t \mid x_{t-N} \dots x_{t-1}) = \Pr(x_t \mid x_{t-R} \dots x_{t-1})$$

That is, the newest observation cannot depend on any more than the last R symbols. There are some processes for which this is true; we say they are Markov with order R. Most processes, however, are not Markov at any order R.

A core concept underlying much of the work in this thesis is the *predictive state* [**39**]. Simply put, this is just the limit of how the next block of observations depends on an infinite amount of past information:

$$\Pr(x_1 \dots x_\ell \mid \overleftarrow{x}) = \lim_{N \to \infty} \Pr(x_1 \dots x_\ell \mid x_{-N} \dots x_0)$$

where $\overleftarrow{x}$ is some infinite sequence of past observations, $\overleftarrow{x} = (x_0, x_{-1}, \dots)$, with x_{-k} being the $k+1$th most recent observation. While the length-R conditional probabilities are not typically sufficient to give us the full range of behavior for a process, if we understand the structure of the predictive state function, then we *do* have a full characterization of the process's behavior.

It is also useful to note that the full predictive state is characterized by *all* the distributions $\Pr(x_1 \dots x_\ell \mid \overleftarrow{x})$ for *each* $\ell = 1, 2, \dots$. This is an infinite number of probability distributions, which together encompass our understanding of the full set of possible futures. If we want to bundle all of these distributions into a nice, compact mathematical object, we will need to use the formalism of *measure theory*. This will be undertaken in Chapter 2. In that chapter we will also discuss how these predictive states can be represented by vector space embeddings which can be useful in machine learning. Subsequent chapters (namely 4 and 6) will examine how predictive states can be used to understand models of processes and their physical constraints.

1.4. Ergodicity and connectivity

Let's return now to the concept of ergodicity. It is a very ubiquitous concept in the theories of stochastic processes, dynamical systems, thermodynamics and elsewhere. Wherever it appears it often comes with a specialized definition for the setting at hand. There are typically three basic sorts of characterizations of ergodic processes [**96**, **175**]:

(1) An ergodic process is one where "spatial averages"—that is, averages taken at a moment in time over all possibilities—are equal to temporal averages—the long-range averages we have been considering. Put in formulaic terms, for any function $f(x_1 \dots x_L)$ andd any particular sample of the process $\overrightarrow{x} = (x_t)$,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} f(x_{t:t+N}) = \sum_{w \in \mathcal{X}^L} f(w) \Pr(w)$$

The reader can check that this is a consequence of the definitions we have chosen for the probabilities, stationarity and ergodicity.

(2) An ergodic process is one where the "word dynamics" (characterized by the graph of $\Pr(w_2 \mid w_1)$ which shows the probability of each word being followed by another) has no proper invariant subsets. To unpack this, the only invariant sets of the dynamics (that is, a set of words where all outward arrows point back to the same set) are the empty set, or the whole set $\mathcal{X}_L$. In a moment, we will see how this is a consequence of the definition we have already chosen.

(3) An ergodic process is one where the graph of $\Pr(w_2 \mid w_1)$ is strongly connected. This is the definition we gave in the previous section.

This is just my physicist's opinion, which doesn't really have a formal justification, but I consider the strong connectivity definition to be in a certain sense the most *causal* definition of ergodicity, in the sense that the other two definitions follow as rather natural consequences. For this reason I consider ergodicity to be a very *graphical* concept, and it will therefore behoove us to take a moment to discuss a little bit more graph theory. There is a theorem (or really, a cluster of theorems) we will then be able to discuss, called the *Perron-Frobenius theorem.* This theorem provides insight to how the graphical structure of the word dynamics $\Pr(w_2 \mid w_1)$ determines the long-run behavior of the

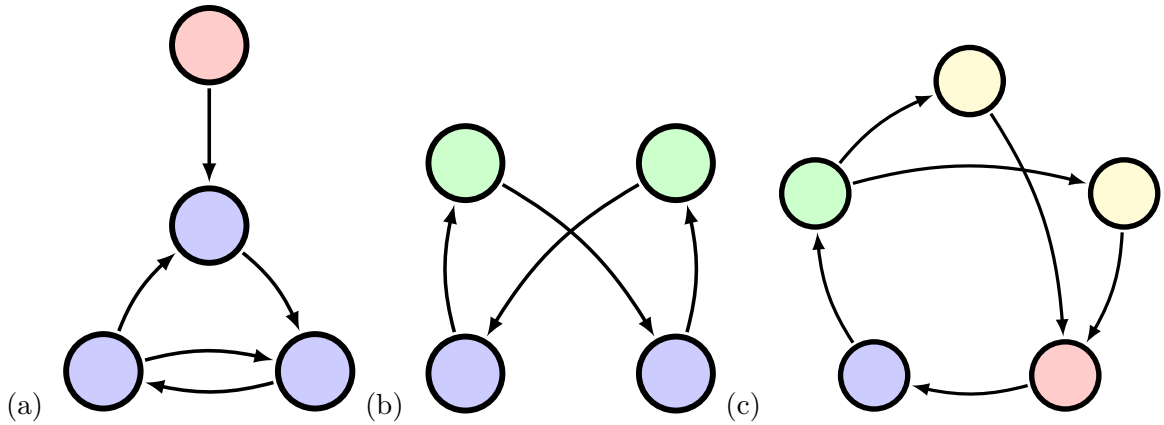

FIGURE 1.1. Perron-Frobenius theory: *(a)* Example of an *aperiodic* graph with a transient node (red) and a strongly connected part (blue). *(b)* An example of a period-2 graph, with the blocks colored separately. *(c)* Example of a period-4 graph, also with separately colored blocks.

stochastic process. These insights will guide our understanding of ergodic processes in Chapters 4 through 6, and there is also an extent to which the graph-theoretic picture will inform our intuition in 7, when we analyze trade networks.

The two main properties of graphs that we will want to understand are *connectedness* and *period* (Fig. 1.1). We have already defined the concept of strong connectivity, but it will be good to consider the different ways that this can break down. There are a few distinct possibilities which can arise. In imagining these possibilities, it will be helpful to the reader to imagine that the directed edges of a graph are indicating the "flow" of some conserved fluid, such as water, and to think of the long-term behavior of a process as all the ways the water can turn and pool throughout the graph.

(1) In the case of strong connectivity, every ounce of the water will at some point flow through each node of the graph, as no matter where it starts, there are paths (at least one, and typically more) it can take to get to any destination.

(2) In the case of completely separated subgraphs, there is no exchange of flow between the subsets of the nodes, and so each subset has its own private water supply which is never accessed by outside nodes.

(3) It may be the case that some subset of nodes has *outflow* to another subset, but no returning inflow. In this case, though the graph is fully connected, it is not *strongly* connected. We call the part which only has net outflow the *transient part* and the subset which only has

net inflow the *recurrent part*. Any water starting in the transient part will eventually flow out into the recurrent part, and the transient part will be left empty.

Notice that we have said little about any sort of asymptotic or stationary distribution of water; to talk about that aspect, we need to think about period. What connectivity tells us about, though, is the the *temporal average* of water flow: how much water will flow through a node over a long enough range of time? In a strongly-connected graph, every node will have some consistent rate of flow over the long-time average. In disconnected graphs, this flow will be restricted to water originating from the nodes in the same connected part. In weakly connected graphs, the transient part receives no flow at all in the long run.

Period in graphs is a somewhat more subtle concept with a fairly clunky formal definition. A *loop* in a directed graph is any sequence of edges, such that the "tail node" of each edge is the "head node" of the preceding edge, and the head node of the final edge is the tail node of the initial edge. For a given graph, we can consider the set of all possible loops; the lengths of these loops may take on quite a variety of values. However, if these loop lengths all have a greatest common divisor p, then we say that the graph has period p.

Now that we have given the number-theoretic definition, let's give the intuitive definition. If I can divide the nodes of my graph into p blocks, and label each block with some number in $1, \ldots, p$ such that all the outward edges from nodes of the kth block point into the $k+1$th block, or the 1st block in the case of the final pth block, then the graph has period p. Note that when we say *all* the outward edges from nodes in a block point to the next, we really mean *all*. There are no internal connections between nodes in the same block, and no "self-loop" edges connecting a node to itself. This obviously disqualifies many graphs from having any period at all. We say these graphs are *aperiodic*.

As with connectivity, it will be helpful to consider for the moment the implications of periodicity for flows:

(1) In the case of periodicity, the there is an extent to which the distribution of water flow through the nodes never settles into any asymptotic behavior. If all the water starts in a given block, it will all be back in that block p steps later. Thus, if you know which block a given ounce of water is in at time t, then you know which block it will be in at any

other time. Despite this, there may still be a well-defined unique stationary distribution of water (the uniform distribution where water is contained in each block in equal amounts), and there may still be well-defined long-time averages of water levels in each node (due to connectivity properties).

(2) In the case of aperiodicity (and strong connectivity), the distribution of water across nodes will, over time, settle into a unique stationary distribution. This is due to a combination of the stable long-time averages arising form connectivity with the phenomenon of *mixing*, by which aperiodicity introduces temporal "stutters" causing the juxtaposition of temporally-shifted possibilities. The consequence is that the instantaneous distribution of the water begins to reseemble the long-time average. A simple example which makes this phenomenon evident is the case of a mostly circular graph of n nodes, where one of the nodes has a single self-loop. Despite every other node only pointing to the next, water in any initial distribution will eventually settle into the uniform distribution due to the action of the "eddy" in the self-looped node catching and stalling water from completing the circle in exactly n steps.

All of the considerations we have just given for connected and periodic graphs can be summarized by a theorem called the Perron-Frobenius theorem.

The Perron-Frobenius theorem is stated in terms of *positive matrices*, and concerns their spectral properties; that is, it tells us about the eigenvectors and eigenvalues of positive matrices, and relates their spectra to the structural properties of the corresponding graph [**32**]. Thus, when we talk about stationary distributions, Perron-Frobenius talks about eigenvectors; when we talk about asymptotic behavior and mixing, Perron-Frobenius talks about the relative magnitudes of eigenvalues. We will state the theorem and then relate its assertions to the graph concepts we have already discussed.

Recall that the graph of a positive matrix $\mathbf{T} = (T_{ij} \geq 0)$ is the directed graph such that the edge from node j to node i has weight T_{ij} if $T_{ij} > 0$ (or does not exist otherwise). Further recall that the spectral radius ρ of a matrix $\mathbf{T}$ is the maximum magnitude taken over all its eigenvalues.

THEOREM 1 (Perron-Frobenius Theorem (abridged)). *Let* $\mathbf{T} : \mathbb{R}^n \to \mathbb{R}^n$ *be a matrix with non-negative values,* $\mathbf{T} = (T_{ij} \geq 0)$. *Suppose the graph of* $\mathbf{T}$ *is strongly connected. Let* ρ *be the spectral radius of* $\mathbf{T}$.

(1) $\rho > 0$ *is a simple eigenvalue of* $\mathbf{T}$*; that is, it has only one left- and right-eigenvector.*

(2) If $\mathbf{T}$*'s graph has period* p*, then there are exactly* p *simple complex eigenvalues of magnitude* ρ*, and they take the form* $\rho e^{i2\pi k/p}$ *for* $k = 0, \ldots, p-1$*. If* $\mathbf{T}$*'s graph is aperiodic, then there is only one eigenvalue of magnitude* ρ*, which is* ρ *itself.*

(3) The right- and left-eigenvectors corresponding to ρ*, denoted by* $\boldsymbol{\pi}$ *and* $\boldsymbol{\phi}$ *respectively, each have all positive components:* $\pi_i > 0$ *and* $\phi_i > 0$ *for all* $i = 1, \ldots, n$.

(4) No other eigenvectors of $\mathbf{T}$ *have all positive components.*

Let's discuss the meaning of each point successively.

(1) If we consider the scaled matrix $\rho^{-1}\mathbf{T}$, which has a spectral radius of 1, then the first point of Perron-Frobenius tells us that $\rho^{-1}\mathbf{T}$ has a single unique stationary vector $\boldsymbol{\pi}$, so that $\mathbf{T}\boldsymbol{\pi} = \boldsymbol{\pi}$. In particular, for conditional probability matrices, $\rho = 1$ automatically, and so Perron-Frobenius theorem says that an ergodic process has a unique stationary distribution over words, given by the vector $\boldsymbol{\pi} = (\pi_w)$ where $\pi_w = \Pr(w)$. (Also, for a conditional probability, it can be easily checked that $\boldsymbol{\phi}$ is just a vector of all 1's.)

(2) The second point tells us that, if the graph structure of $\mathbf{T}$ has period p, then the dynamic will "rotate" the distribution through p distinct states (complex eigenvalues are always an indicator of rotation). Alternatively, if $\mathbf{T}$ is aperiodic, then there is a "spectral gap" between the eigenvalue ρ and all other eigenvalues. The consequence of a spectral gap is that, as the dynamic system progresses, the stationary vector drowns out all other vectors. We will examine this more closely in a moment.

(3) The positivity of the eigenvectors means that the stationary distribution can be interpreted, when scaled, as a probability distribution, and has positive (*i.e.* nonzero) weight on every node of the graph.

(4) The last point implies that any other eigenvector of $\mathbf{T}$ cannot describe a probability distribution itself; however, they can describe *differences* between distributions. We can

consider the non-maximal eigenvectors to describe all the possible ways that a distribution can differ from the stationary distribution $\boldsymbol{\pi}$, and their corresponding eigenvectors describe how quickly that difference decays (or, in the periodic case, how that difference is rotated over time).

Points 2 and 4 bear further elaboration. Let us start by considering the aperiodic case. It is helpful to break the matrix $\mathbf{T}$ into two parts: the unital eigenvector part, $\rho\boldsymbol{\pi}\boldsymbol{\phi}^\top$ (where $(\cdot)^\top$ is the transpose), and the remainder $\tilde{\mathbf{T}} = \mathbf{T} - \rho\boldsymbol{\pi}\boldsymbol{\phi}^\top$. Now, because ρ is the spectral radius, and we are assuming aperiodicity, the Perron-Frobenius theorem tells us that the matrix $\tilde{\mathbf{T}}$ has a spectral radius less than ρ (after all, all it has left are smaller eigenvalues). Then for any vector $\mathbf{v}$ it will be the case that $\left\|\tilde{\mathbf{T}}\mathbf{v}\right\| \leq r \left\|\mathbf{v}\right\|$ for some $r < \rho$. It must then be the case that

$$\lim_{n\to\infty} \frac{1}{\rho^n}\mathbf{T}^n\mathbf{v} = (\boldsymbol{\phi}\cdot\mathbf{v})\,\boldsymbol{\pi} + \lim_{n\to\infty} \frac{1}{\rho^n}\tilde{\mathbf{T}}^n\mathbf{v} = (\boldsymbol{\phi}\cdot\mathbf{v})\,\boldsymbol{\pi}$$

Here $\cdot$ is the dot product between two vectors. The last limit drops off to zero because its norm scales as $(r/\rho)^n$. So, asymptotically, the action of $\rho^{-1}\mathbf{T}$ on a vector $\mathbf{v}$ is the same as just multiplying it by $\boldsymbol{\pi}\boldsymbol{\phi}^\top$. If we drop the $\mathbf{v}$ (since it is arbitrary), we have the formula

$$\lim_{n\to\infty} \frac{1}{\rho^n}\mathbf{T}^n = \boldsymbol{\pi}\boldsymbol{\phi}^\top$$

The periodic case is not that much worse. We will skip the details, but the gist is that the rotational eigenvectors simply pick up powers of their corresponding complex phase $e^{i2\pi k/p}$. When we multiply by $\rho^{-1}\mathbf{T}$ a number of p times, and sum over the results, these complex phases cancel out and eliminate the rotational eigenvectors. As a consequence,

$$\lim_{n\to\infty} \frac{1}{p}\sum_{k=1}^{p} \frac{1}{\rho^{n+k}}\mathbf{T}^{n+k} = \boldsymbol{\pi}\boldsymbol{\phi}^\top$$

Here, rather than looking at the repeated action of $\rho^{-1}\mathbf{T}$ alone, we are looking at the average over a length-p block of large powers of $\rho^{-1}\mathbf{T}$. This "seasonal" average removes the lingering effects of periodicity to get an asymptotic result.

For both periodic and aperiodic matrices, the Cesàro average formula holds:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \frac{1}{\rho^k} \mathbf{T}^k = \boldsymbol{\pi}\boldsymbol{\phi}^\top \tag{1.1}$$

This simply generalizes the asymptotic formulae above to a long-run temporal average.

The Perron-Frobenius formula covers matrices whose graphs are strongly connected. The summary of what it tells us is that there is a unique stationary state of the matrix's dynamics, which in the aperiodic case is the asymptotic endpoint of the dynamical process, and in the periodic case is arrived at via seasonal or long-time averages.

The reader may wonder just what Perron-Frobenius has to say on matrices whose graphs are *not* strongly connected. In this case, our earlier delineation of the possible cases will be helpful. Earlier we noted that every graph can be decomposed into *transient* parts, which have only a net outflow and no net inflow, and *recurrent* parts, which have net inflow (and potentially net outflow). Further, a graph can have multiple disconnected recurrent parts. We will suppose we have partitioned the nodes of the graph of $\mathbf{T}$ into the sets $\mathcal{T}$ and $\mathcal{R}_s$, where $\mathcal{T}$ is the set of transient nodes and $\mathcal{R}_s$ are the disconnected recurrent components, also called *ergodic components*, indexed by s. If we denote $\mathbf{T}|_s = (T_{ij})_{i,j\in\mathcal{R}_s}$, $\mathbf{T}|_{\mathcal{T}} = (T_{ij})_{i,j\in\mathcal{T}}$, and $\mathbf{T}|_{s,\mathcal{T}} = (T_{ij})_{i\in\mathcal{R}_s,j\in\mathcal{T}}$ as the restrictions of $\mathbf{T}$ to its constituent blocks, then the matrix takes the overall block-matrix shape

$$\mathbf{T} = \begin{bmatrix} \mathbf{T}|_{\mathcal{T}} & 0 & 0 & \cdots & 0 \\ \mathbf{T}|_{1,\mathcal{T}} & \mathbf{T}|_1 & 0 & \cdots & 0 \\ \mathbf{T}|_{2,\mathcal{T}} & 0 & \mathbf{T}|_{\mathcal{T},2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{T}|_{S,\mathcal{T}} & 0 & 0 & \cdots & \mathbf{T}|_S \end{bmatrix}$$

where S is the number of ergodic components.

By definition, the restricted matrices $\mathbf{T}|_s$ satisfy the requirements of the Perron-Frobenius theorem. In particular, each has a spectral radius ρ_s and a stationary state $\boldsymbol{\pi}|_s$.

The spectral radius ρ of the entire matrix $\mathbf{T}$ would just be the largest spectral radius of each of the recurrent parts, $\max_s \rho_s$. If we naïvely consider the behavior of $\rho^{-1}\mathbf{T}$, we would find that the recurrent component with the largest spectral radius drowns out all the others, so that the only

stationary state of $\rho^{-1}\mathbf{T}$ is the vector which is zero on all components except for the maximal one, where it takes the values of the vector $\boldsymbol{\pi}|_s$.

The more interesting, and thankfully far more common case for stochastic matrices is the case where all ergodic components have the same spectral radius, which we will for simplicity assume to be 1. In that case, the dynamical process described by $\mathbf{T}$ has an infinite number of possible stationary distributions. If we let $\mathbf{p} = (p_s)$ be a probability vector over the ergodic components, then any vector which is zero on the transient component and takes the values $p_s\boldsymbol{\pi}|_s$ on the ergodic components is itself a stationary distribution. That is, any vector which takes the block form:

$$\mathbf{v} = \begin{bmatrix} 0 \\ p_1\boldsymbol{\pi}|_1 \\ p_2\boldsymbol{\pi}|_2 \\ \cdots \\ p_S\boldsymbol{\pi}|_S \end{bmatrix}$$

is an invariant distribution under the action of $\mathbf{T}$. Consequently, there may be a considerable degree of mixing of possibilities going on within each recurrent part, but each recurrent part itself remains insulated from the others, and discrepancies between them remain stable over time.

1.5. Resources and information

Knowledge is power—so they say. It is more apt to say that knowledge is a resource. There is, of course, little novel insight in pointing out that having knowledge about the current state and behavior of a system gives an observer the ability to extract value from it. Quantifying this ability, on the other hand, remains a highly active area of research.

The modern science of *information theory* developed from precisely these sorts of questions during the heyday of operations research, in the height and aftermath of the second World War. The most familiar form of information theory, Shannon's theory of communication [**177**], has become foundational to signal processing and electrical and computer engineering, and has also drawn considerable interest from social and biological sciences [**36**, **119**].

Efforts by quantum physicists to adapt Shannon's concept of information to the realm of quantum probabilities gradually revealed the existence of numerous extensions describing a variety of distinct forms of quantum informational resources. This frontier of research, which spans quantum physics, nonequilibrium thermodynamics and nanoscale computing, has become known as *resource theory* [**35**]. In addition to Shannon's theory of communication, with various entropic measures, resource theory also draws heavily upon Blackwell's theory of statistical decisions [**18**, **19**], utilizing additional tools such as majorization [**123**]. We will give a brief overview of Shannon's and Blackwell's perspectives here, presenting them as dual parts of a unified theory of data production and data processing. Shannon's and Blackwell's definitions of information will together be central to Chapters 4 through 6, as well as in Chapter 7.

To discuss data processing, we have to make something of a frame shift from the preceding sections. There we discussed the dynamics of a system with the use of conditional probabilities. These told us, if we observed a block of observations as the word w, that we would observe the next block as w' with a frequency characterized by the probability $\Pr(\, w' \mid w \,)$. We can also use conditional probabilities to describe a method of signal processing. If a stimulus signal a is received, then we produce the response signal b with frequency $\Pr(\, b \mid a \,)$. Whereas before the dynamical matrix was always square, signal-processing matrices can be rectangular, as the set of responses may not be the same as the set of stimuli.

When discussing information, the two dual perspectives on its quantification involve, on the one hand, describing the *variety* of possible signals which can be observed from a source, and on the other hand, describing the capacity of a data-processing mechanism for *distinguishing* between different signal sources. These are dual precisely because the easiest way for two systems to distinguish themselves from one another is to limit their signals to have minimal overlap—reducing their individual variety. Variety and distinction are, in a sense, the two primary resources of interest in information and resource theory. We shall discuss each in turn.

1.5.1. Variety as resource. The stochastic processes which we have previously discussed can be characterized by their word probabilities $\Pr(\, w \,)$. These describe for us the range of possible behaviors of the process, and how frequently different parts of this range are utilized.

Our ablity to predict the behavior of the process—and thus, if it is linked to any physical resources, to exploit its behavior—is dependent on our ability to comprehend, transcribe and interpret the range of its behavior [13]. We will refer to the scope of this range of possiblities as the *variety*.

All resource theories are theories of data processing. Certain kinds of data processing will impact corresponding kinds of informational resources. Thus every theory of a particular resource has a corresponding set of *free operations*, which are taken to be those data processing operations which cannot increase the resource. If we want to study variety—we must start by ask ourselves what kind of conditional probabilities $\Pr(\, b \mid a \,)$ cannot increase the variety of the signal.

After considering this question for a moment, the reader might settle on *deterministic operations*. These are conditional probabilities $\Pr(\, b \mid a \,)$ where for each stimulus signal a there is only *one* response b with nonzero probability. Deterministic operations are more commonly known as functions: essentially, they map each stimulus a to a specific response $f(a)$. The resource theory of variety is essentially just the theory of processing data with functions. Variety, in essence, is *defined* as that quality of data which cannot be increased by applying functions to it.

This abstraction is very nice but not, on its own, particularly useful. Resource theories also provide several means of quantifying the resource under investigation. Quantifications of variety are called *entropies*. If we have a random variable X—that is, a variable whose possible values are described by a probability distribution $\Pr_X(\, x \,)$—then an entropy of X is any function $\mathrm{F}[\cdot]$ with the property that

$$\mathrm{F}[f(X)] \leq \mathrm{F}[X]$$

for all random variables X and functions f.

Many functions have this property, but it will be useful to consider ones which have practical interpretations. Variety can have many origins in nature. Our challenge, as those who would make use of nature, is to *represent* variety. Useful entropies are those which tell us how difficult or costly it is to represent the variety of a process. Good representations will be reversible—we can use them as a reference to reconstruct the original behavior that was observed. This means every useful representation process consists of a function f, the *encoding*, which maps a stimulus to its symbol,

and another function g, the *decoding*, which maps a symbol back to a stimulus of the original type:

$$a \xrightarrow{\text{encoding } f} b \xrightarrow{\text{decoding } g} a'$$

When $a' = a$, we have successfully described the stimulus. Otherwise, we have an error. The goal of representation is to encode as much variety as possible with minimal error.

Variety as a *resource* comes into play because it should be immediately evident to the reader that in order to construct an error-free representation of a stimulus, the set of symbols $\mathcal{B}$ must be at least as large as the set of possible stimuli $\mathcal{A}$. Thus, the system being used as a representation must have at least as much potential variety as the system being observed. This leads to our first entropy, the *max entropy*:

$$\mathrm{H}_{\max}[A] = \log_2 | \{ a \in \mathcal{A} \mid \Pr_A (a) > 0 \} |$$

The log is for historical reasons that will become evident in a minute—but the purpose of the max-entropy is to quantify the minimal number of distinct symbols required to represent the variety of a process with zero error. This is just the number of observed stimuli.

On the other hand, we might simply ask ourselves what would be the chance of error if we simply chose to *always* decode the symbol as the *most likely* stimuli. This way, we only need one symbol, and have a success rate of $\max_{a \in \mathcal{A}} \Pr_A (a)$. The corresponding entropy is called the *min-entropy*:

$$\mathrm{H}_{\min}[A] = -\log_2 \left(\max_{a \in \mathcal{A}} \Pr_A (a) \right)$$

The logarithm is, again, for historical reasons. The negative is so that $\mathrm{H}_{\min}[\cdot]$ has the proper behavior of an entropy, and decreases when a function is applied to the random variable.

Between these two quantities is the *Shannon* entropy, which we will be the entropy we are referring to if we just say "entropy" and do not otherwise specify. The Shannon entropy, unlike the previous two we have considered, represents an ideal *rate* which can be achieved when we utilize the economy of scales, and use the same symbol set to encode a large number of signals at once. The Shannon entropy has a simple formula:

$$\mathrm{H}[A] = -\sum_{a \in \mathcal{A}} \Pr_A (a) \log_2 \Pr_A (a)$$

The Shannon entropy has a dual interpretation. On the one hand, if we wish to encode a large number N of copies of a signal X with a controlled error rate, the minimal amount of symbols required is $2^{N\,\mathrm{H}[X]}$. On the other hand, to achieve this economy of scales, we admit an asymptotically vanishing error rate of $2^{-N\,\mathrm{H}[X]}$. Thus the Shannon entropy can be seen simultaneously as a generalization of both the max- and min-entropies to the economy of scales.

The Shannon entropy has many remarkable properties. For one, it satisfies a formula called the *chain rule of entropies*, which allows it to scale over multiple variables. If we have a joint variable AB with probability $\Pr_{AB}(\,a,b\,)$ and conditional probabilities $\Pr_{B|A}(\,b \mid a\,)$,

$$\begin{aligned}\mathrm{H}[AB] &= -\sum_{a\in\mathcal{A}} \Pr_A(\,a\,)\log_2 \Pr_A(\,a\,) - \sum_{\substack{a\in\mathcal{A}\\ b\in\mathcal{B}}} \Pr_{AB}(\,a,b\,)\log_2 \Pr_{B|A}(\,b \mid a\,)\\ &= \mathrm{H}[A] + \mathbb{E}_{\mathrm{A}}[\mathrm{H}[B|A=a]]\end{aligned}$$

The second term is the mean of the entropy for each conditional distribution of B, averaged over the conditioning variable A. It is common to write more compactly

$$H[B|A] = \mathbb{E}_{\mathrm{A}}[\mathrm{H}[B|A=a]] = -\sum_{\substack{a\in\mathcal{A}\\ b\in\mathcal{B}}} \Pr_{AB}(\,a,b\,)\log_2 \Pr_{B|A}(\,b \mid a\,)$$

This is called the conditional entropy, and quantifies how much variety is left in B once A is known.

Another property of entropy is its *subadditivity*. This means that the joint entropy is always less than the sum of the two individual entropies. The difference is called the *mutual information*:

$$\mathrm{I}[A:B] = \mathrm{H}[A] + \mathrm{H}[B] - \mathrm{H}[AB] = \sum_{\substack{a\in\mathcal{A}\\ b\in\mathcal{B}}} \Pr_{AB}(\,a,b\,)\log_2 \frac{\Pr_{AB}(\,ab\,)}{\Pr_A(\,a\,)\Pr_B(\,b\,)}$$

Due to the chain rule, it is also the case that $\mathrm{I}[A:B] = \mathrm{H}[B] - \mathrm{H}[B|A]$. The mutual information quantifies the correlation between A and B: it tells us how much of their overall variety is in fact shared betweeen the two sources.

1.5.2. Distinguishability as resource. We have examined how to quantify the variety observed in stochastic processes, with an aim towards the describing the memory costs of describing that variety. This is one approach to thinking about information as a resource. Another approach

begins when we ask ourselves about the *hidden* variety which may be driving the behavior of a process.

By hidden variety we mean potential *latent* states which determine the sorts of behavior a process may display. Since latent states are unobservable, we can only hypothesize their value based on the observable variables which they influence. If we wish to efficiently make use of the information stored in processes, then it is important that we be able to infer knowledge about the latent states, which may tell us about available energy and other physical resources.

We can only know latent states through the effects they induce. The important thing about latent states, then, is how they affect observable quantities. This can be characterized by conditional probabilities: if a is the observable signal and s is the latent state, then this means the probabilities $\Pr(\, a \mid s \,)$. Telling the difference between two latent states s_1 and s_2 means telling the difference between $\Pr(\, a \mid s_1 \,)$ and $\Pr(\, a \mid s_2 \,)$. The problem of distinguishing between different latent states can then be reduced to the problem of distinguishing between different probability distributions.

For simplicity, let us consider pairs of probability vectors $(\mathbf{p}, \mathbf{q})$. Our ability to tell the difference between these distributions (based on observing samples) is called their *distinguishability* and is a kind of resource [209]. As with variety, our investigation of it must begin with a definition of what kind of data processing operation does not increase distinguishability. In this case, the answer is that *no* data processing operation can increase distinguishability, so long as you perform the *same* operation on each distribution in the pair.

That is, given any conditional probability matrix $\mathbf{T} = (T_{ij})$ (with $\sum_i T_{ij} = 1$, $T_{ij} \geq 0$), the pair $(\mathbf{Tp}, \mathbf{Tq})$ cannot be considered more distinguishable than $(\mathbf{p}, \mathbf{q})$. It is common, given two distributions $(\mathbf{p}, \mathbf{q})$ and $(\mathbf{p}', \mathbf{q}')$, to say that $(\mathbf{p}, \mathbf{q})$ *relatively majorizes* $(\mathbf{p}', \mathbf{q}')$, written $(\mathbf{p}, \mathbf{q}) \succsim (\mathbf{p}', \mathbf{q}')$, if there is a conditional probability matrix $\mathbf{T}$ such that $\mathbf{p}' = \mathbf{Tp}$ and $\mathbf{q}' = \mathbf{Tq}$. The partial ordering of relative majorization ranks pairs of distributions by their overall distinguishability.

The principle is simple: doing the same thing to two distributions can't make them any more different than they already are. The consequences of this observation are profound. This perspective on information processing was first expounded by Blackwell [18, 19], not long after Shannon expressed his own theory of communication. Blackwell's approach takes us down a somewhat different road,

but ultimately leads to a very similar place, and we will even see how entropies arise naturally as a consequence of this principle.

As with the resource theory of variety, there are a number of ways to *quantify* distinguishability. Before we get to those, however, it is important to discuss an elegant and powerful tool unique to the theory of distinguishability: we can *visualize* the difference between two distributions, using a concept called *relative majorization* (Fig. 1.2).

The visualizations of pairs $(\mathbf{p}, \mathbf{q})$, called the Lorenz curves $\Lambda(\mathbf{p}, \mathbf{q})$, have a partial ordering on them which is isomorphic to the partial ordering of relative majorization. To be specific, given two pairs $(\mathbf{p}, \mathbf{q})$ and $(\mathbf{p}', \mathbf{q}')$, the Lorenz curves satisfy $\Lambda(\mathbf{p}, \mathbf{q}) \geq \Lambda(\mathbf{p}', \mathbf{q}')$ if and only if $(\mathbf{p}, \mathbf{q}) \succsim (\mathbf{p}', \mathbf{q}')$.

The value in having a visual isomorphism of the theory of relative majorization is obvious. To construct the Lorenz curve of a pair, we first sort the indices of the vectors $i_1, \ldots, i_n$ so that the ratio p_{i_k}/q_{i_k} is decreasing (n is assumed to be the number of possible outcomes to the distributions $\mathbf{p}$ and $\mathbf{q}$). We then draw a piecewise linear loop connecting the $n+1$ vertices $(x_0, y_0), \ldots, (x_n, y_n)$ defined by

$$x_k = \sum_{\ell=1}^{k} q_{i_k}$$

$$y_k = \sum_{\ell=1}^{k} p_{i_k}$$

Note that $(x_0, y_0) = (0, 0)$ and $(x_n, y_n) = (1, 1)$ always. This loop is the Lorenz curve $\Lambda(\mathbf{p}, \mathbf{q})$.

The Lorenz curve outlines a convex set, which is the area between the diagonal connecting $(0, 0)$ to $(1, 1)$ and a convex curve which "bows out" from this diagonal. We say that $\Lambda(\mathbf{p}, \mathbf{q}) \geq \Lambda(\mathbf{p}', \mathbf{q}')$ if the convex seet outlined by $\Lambda(\mathbf{p}, \mathbf{q})$ contains the curve $\Lambda(\mathbf{p}', \mathbf{q}')$. Alternatively, this means that the bowed-out portion of $\Lambda(\mathbf{p}, \mathbf{q})$ is always vertically higher than that of $\Lambda(\mathbf{p}', \mathbf{q}')$.

The Blackwell-Sherman-Stein theorem states that $\Lambda(\mathbf{p}, \mathbf{q}) \geq \Lambda(\mathbf{p}', \mathbf{q}')$ *if and only if* $(\mathbf{p}, \mathbf{q}) \succsim (\mathbf{p}', \mathbf{q}')$ [**19**]. (It actually says quite a bit more than this, but this is the consquence on pairs of probability distributions.)

The visualization afforded by the Lorenz curve, and the concept of relative majorization, are a powerful tools which will very important to our results in Chapters 4, 5 and 7.

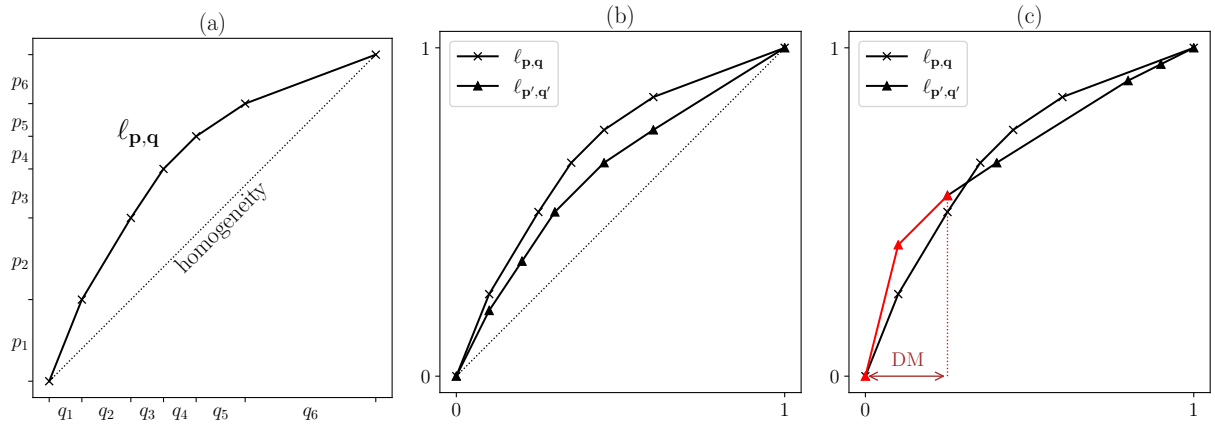

FIGURE 1.2. Majorization primer: also seen in Chapter 7. *(a)* Example a Lorenz curve for a pair of distributions $(\mathbf{p}, \mathbf{q})$ over 6 elements. We assume the elements are indexed so that p_i/q_i is monotonically decreasing. $\mathbf{p}$ and $\mathbf{q}$ are not homogeneous with respect to one another, and so the Lorenz curve bows out above the diagonal. *(b)* An example of two pairs, $(\mathbf{p}, \mathbf{q})$ and $(\mathbf{p}', \mathbf{q}')$, such that $(\mathbf{p}, \mathbf{q}) \succeq (\mathbf{p}', \mathbf{q}')$. Visually, this means that the second Lorenz curve is fully beneath the first and, therefore, closer to the line of homogeneity. *(c)* Example where majorization does not hold. "DM" stands for dismajorization, which describes the extent to which majorization fails (covered in more detail in Chap. 7).

It will also be important to understand the ways of quantifying distinguishability. A *divergence* is any function of pairs of distributions $\mathrm{D}\left(\, \mathbf{p} \,\|\, \mathbf{q} \,\right)$ (the middle double-bar is traditional) which satisfies

$$\mathrm{D}\left(\, \mathbf{T}\mathbf{p} \,\|\, \mathbf{T}\mathbf{q} \,\right) \leq \mathrm{D}\left(\, \mathbf{p} \,\|\, \mathbf{q} \,\right)$$

As with entropy, we will primarily be interested with divergences that are interpretable; in fact, here we will just discuss one: the Kullback-Liebler divergence, which is extremely versatile and appears in a multitude of settings.

The Kullback-Liebler divergence, also sometimes called the *relative entropy* (for reasons that will be soon apparent), is defined as

$$\mathrm{D}_{\mathrm{KL}}\left(\, \mathbf{p} \,\|\, \mathbf{q} \,\right) = \sum_{j=1}^{n} p_j \log_2 \left(\frac{p_j}{q_j} \right)$$

A direct interpretation of the relative entropy is provided by *Sanov's theorem*. If we collect N independent samples from distribution $\mathbf{q}$ without knowing the distribution, and we attempt to determine if the statistics of the samples match distribution $\mathbf{p}$, then we will erroneously conclude

that the sample matches $\mathbf{p}$ with a probability which scales as $2^{-N\mathrm{D}_{\mathrm{KL}}(\,\mathbf{p}\,\|\,\mathbf{q}\,)}$. Just as with the Shannon entropy, the relative entropy characterizes an exponential rate of error decay, in this case relating to the classification error between two distributions.

The relative entropy has a far wider reach than hypothesis testing, however. To consider its range, we will consider some additional resource theories which are derived as embedded resource theories within the theory of distinguishability.

One such embedded resource theory is the theory of *nonuniformity* [64]. The nonuniformity of a distribution $\mathbf{p}$ is the degree to which it is not the uniform distribution, which we will denote by $\mathbf{u}$. To build a resource theory for this property, we would first have to characterize the set of operations which do not increase nonuniformity. We can leverage the concept of majorization to do so. Let $\mathbf{B}$ be a conditional probability matrix which has the uniform distribution as a fixed point: $\mathbf{Bu} = \mathbf{u}$. Such a matrix is called *bistochastic*. By the definition of majorization, it is then the case that $(\mathbf{Bp}, \mathbf{u}) \succsim (\mathbf{p}, \mathbf{u})$ for all $\mathbf{p}$. In other words, $\mathbf{B}$ can never make a distribution more different from the uniform distribution than it already is. Bistochastic matrices are therefore a suitable "free operation" for a theory of nonuniformity.

Notice what we have done: we have just taken a subset of the free operations of the theory of distinguishability, and defined a new resource theory which is embedded in the old one. It stands to reason that we can then import all the other properties of the theory of distinguishability. For instance, the Lorenz curve $\Lambda(\mathbf{p}, \mathbf{u})$ can be used to visualize the nonuniformity of a distribution. We can define a partial ordering on distributions where $\mathbf{p} \succsim \mathbf{q}$ if and only if $(\mathbf{p}, \mathbf{u}) \succsim (\mathbf{q}, \mathbf{u})$. (This ordering is traditionally just called *majorization*.) Last but certainly not least, any divergence used in the theory of distinguishability can be used as a quantification of nonuniformity.

When we use the relative entropy on the pair $(\mathbf{p}, \mathbf{u})$, a familiar friend arrives:

$$\begin{aligned}\mathrm{D}_{\mathrm{KL}}\,(\,\mathbf{p}\,\|\,\mathbf{u}\,) &= \sum_{j=1}^{n} p_j \log_2 \left(\frac{p_j}{1/n}\right) = \log_2 n - \sum_{j=1}^{n} p_j \log_2 p_j \\ &= \log_2 n - \mathrm{H}[\mathbf{p}]\end{aligned}$$

(We are abusing the standard notation a bit here—we have not defined a random variable to go with $\mathbf{p}$, but we hope the reader understands our meaning.) The divergence of any distribution from the

uniform is just the negative of the entropy of that distribution. This is sensible, since the uniform distribution maximizes entropy. This quantity is often called the *negentropy*.

Another example of an embedded resource theory is the theory of nonequilibrium thermodynamics (sometimes called the theory of *thermal operations*) [77]. Suppose to each state j of our system, there is a corresponding energy level E_j; then the system is in thermal equilibrium at temperature T if its distribution matches the Gibbs distribution $\boldsymbol{\gamma}_T = (\gamma_{T,j})$,

$$\gamma_{T,j} = \frac{1}{Z(T)} e^{-E_j/\mathrm{k_B}T}, \quad Z(T) = \sum_j e^{-E_j/\mathrm{k_B}T}$$

A system in nonequilibrium is just any system whose distribution differs from the Gibbs distribution. Such a system can have work extracted from it as it relaxes to equilibrium. As with the theory of nonuniformity, we can define free operations at temperature T as those conditional probability matrices $\mathbf{G}$ which satisfy $\mathbf{G}\boldsymbol{\gamma}_T = \boldsymbol{\gamma}_T$. These are called *thermal operations*, or sometimes *Gibbs-stochastic*. Remarkably, thermal operations are precisely those which result from only doing energy-conserving operations with the aid of a thermal bath at temperature T. Thus, the constraints of nonequilibrium thermoynamics can be directly embedded into the resource theory of distinguishability.

The analogue of majorization here is called *thermo-majorization* [71], and when we apply the relative entropy to the pair $(\mathbf{p}, \boldsymbol{\gamma}_T)$, we find another old friend, this time for the physicist:

$$\begin{aligned} \mathrm{D_{KL}}\left(\, \mathbf{p} \parallel \boldsymbol{\gamma}_T \,\right) &= \log_2 Z - \sum_{j=1}^{n} p_j \log_2 \left(p_j e^{E_j/\mathrm{k_B}T} \right) \\ &= \frac{\mathbb{E}_{\mathrm{j}\sim\mathbf{p}}[E_j]}{\mathrm{k_B}T \ln 2} - \mathrm{H}[\mathbf{p}] + \log_2 Z \end{aligned}$$

The logarithm of the partition function is, up to a scaling factor of $\mathrm{k_B}T\ln 2$, the negative of the free energy of the system when it is in thermal equilibrium. Preceding this term is the average energy minus the *Shannon* entropy—an information-theoretic modification to the traditional Helmholtz formula of internal energy minus thermodynamic entropy, $F = U - \mathrm{k_B}TS$. If we denote

$$F_{\mathrm{eq}} = -\mathrm{k_B}T \ln Z$$

$$F_{\mathrm{neq}} = \mathbb{E}_{\mathrm{j}\sim\mathbf{p}}[E_j] - \mathrm{k_B}T \ln 2\, \mathrm{H}[\mathbf{p}]$$

as the equilibrium and nonequilibrium free energies, respectively, then

$$\mathrm{D_{KL}}\left(\,\mathbf{p}\parallel\boldsymbol{\gamma}_T\,\right)=\frac{F_{\mathrm{neq}}-F_{\mathrm{eq}}}{\mathrm{k_B}T\ln 2}$$

So the relative entropy in the theory of thermal nonequilibrium is just the difference between the nonequlibrium free energy and equilibrium free energy; this indicates the energy *over and above* the equilibrium free energy which can be extracted from the system as it relaxes to equilibrium.

The theory of thermal nonequilibrium thus provides one of the clearest examples of the intuition of resource theory: that is, quantifying the informational resources of a system, such as our ability to distinguish different distributions, can provide deep insights to our ability to extract value (in this case, energy) from those systems.

Before we close, it would be helpful to discuss the implications of the relation

$$\mathrm{D_{KL}}\left(\,\mathbf{Tp}\parallel\mathbf{Tq}\,\right)\leq\mathrm{D_{KL}}\left(\,\mathbf{p}\parallel\mathbf{q}\,\right)$$

This is often called the *data-processing inequality* [36, 119]. It is perhaps one of the most important equations in all of information theory. Partly this is because of how many kinds of quantities can be expressed as relative entropies. We have already seen this in the case of nonequilibrium free energy and negentropy. Another example is the mutual information defined in the previous section. If two random variables X and Y are distributed according to a joint distribution $\mathbf{p}^{(XY)}=(p_{xy})$ with marginalizations $\mathbf{p}^{(X)}=\left(\sum_y p_{xy}\right)$ and $\mathbf{p}^{(Y)}=(\sum_x p_{xy})$, then the mutual information can be expressed as

$$\mathrm{I}[X:Y]=\mathrm{D_{KL}}\left(\,\mathbf{p}^{(XY)}\,\middle\|\,\mathbf{p}^{(X)}\otimes\mathbf{p}^{(Y)}\,\right)$$

where $\mathbf{p}^{(X)}\otimes\mathbf{p}^{(Y)}$ denotes the distribution $p_x^{(X)}p_y^{(Y)}$.

This identity results in another common form of the data processing inequality. Suppose we transform the variable Y into another variable, Z, by use of conditional probabilities $\Pr_{Z|Y}\left(\,z\mid y\,\right)$ which have no explicit dependence on X. This situation, where three variables are related only by an intermediary, is called a Markov chain and denoted $X-Y-Z$. Let $\mathbf{T}$ be the operation on XY which leaves X unchanged and transforms Y to Z; then the data processing inequality implies

$$\mathrm{I}[X:Z]\leq\mathrm{I}[X:Y]$$

In other words, the correlation between X and Z can be no more than the correlation of X with the intermediate variable Y.

The data processing inequality, in both of the common forms presented here, will be examined more closely in chapter 6.

1.6. Itinerary of results

Since this introduction is full of triples, one more should suffice to provide the reader with an overview of the contents and results of this dissertation. The chapters can be thought of as organized into three distinct categories.

In Chapters 2 and 3, we will examine the concept of a *predictive state*, which we defined at the end of Section 1.3. Chapter 2 will lay down the mathematical foundation for a study of predictive states, proving that they are well-defined, convergent objects, and determining which sorts of geometric tools are most appropriate for their analysis. To accomplish this, we rely heavily on the theory of measures [**90**] and ergodic dynamical systems [**96**, **175**].

The foundations laid by this chapter will be built on in Chapter 3, where we will examine how predictive states may be studied in the context of machine learning. The geometric insights gleaned from Chapter 2 will be used to understand why one commonly used method, the *reproducing kernel Hilbert space*, has been so successful, and we will also show how previously unapplied methods such as the Wasserstein distance can also be useful.

The foundations laid in Chapter 2 will also prove the bedrock of the next major thrust of this thesis. Chapters 4, 5 and 6 will look at how stochastic processes can be modeled and generated by other physical and mathematical systems. Chapter 4 will combine our insights from Chapter 2 with tools from representation theory and resource theory, allowing us not only to begin mapping out the space of all possible models for a given stochastic process, but also to understand what physical constraints we may face when implementing these models. Our discussions on ergodicity and the Perron-Frobenius theorem (Section 1.4) as well as our discussion of Shannon information theory (Section 1.5.1) will be of use here. Chapter 5 will build on its predecessor by extending the analysis to *quantum* models of stochastic processes. In both Chapters 4 and 5 we will focus on *memory* costs

of modeling and generating processes; 5 will show how quantum models routinely offer significant improvements in memory cost in comparison to classical models.

Chapter 6 will expand the resoure theoretic discussion to include *thermodynamics*, following the resource theory of thermal operations we discussed in Section 1.5.2. Here we will consider how both classical and quantum generators fare. One of the primary results of this chapter is that it is possible to generate a process with zero dissipation, but this involves a strict minimal memory requirement. The methods we use to prove this are very reminiscent of our discussion in Section 1.4 on systems with multiple ergodic components: essentially, dissipation-free information processing requires only operating on these ergodic components of a system in an invertible manner. Additionally, we find that in *any* case where a quantum model is used over a classical one for the purpose of memory compression, this will result in nonzero dissipation.

Our Chapter 7 stands alone—but it is not lonely. We will switch into a very different domain—the calculation of carbon footprints from trade data via input-output models [**99**, **100**]—but in this domain we will find all the familiar algebra of stochastic processes, majorization, and ergodicity. We will examine how the assumptions of input-output analysis essentially model "embodied carbon" as travelling through the trade network via a Markovian process; importantly, it is driven by the *same* Markov process that is assumed to drive the flow of money through the same network. Our knowledge of ergodicity from the Perron-Frobenius theorem (Section 1.4), and our understanding of the theory of distinguishability via majorization (Section 1.5.2) both indicate the inevitable consequence: the distribution of carbon footprints and the income distribution rapidly converge to one another. Using majorization, we demonstrate how this *statistical* phenomenon is in fact responsible for a number of quantitative results in the field of input-output analysis, which have previously been presented as *empirical* results derived from economic data. We show how this phenomenon can be easily identified using a null model network. In essence, the strong ergodicity of input-output models drowns out information from the actual data being analyzed, leaving us with artifacts being presented as policy impacts. It is our hope that this work will promote more careful use of input-output analysis, and a greater motivation for policy analysts to interrogate the structures of their models.

Thus while this dissertation spans a great number of subjects and fields, ultimately the ideological space we explore is quite small, and we hope will be increasingly familiar to the intrepid reader. Stochastic processes, ergodic models, and information theory are extremely powerful tools which are relevant in settings from the esoteric to the ubiquitous. Through all of these tools, we find that linear algebra continues to be a first-class approach to understanding even the most complex and nonlinear of systems, though we must make significant shifts in perspective to make effective use of it. Once we make this shift, we can see more clearly how stochastic processes imprint their invariant patterns across a wide variety of natural and manmade phenomena—including our own models for those phenomena.

CHAPTER 2

Unfolding time: Prediction in stochastic processes

Life can only be understood backwards; but it must be lived forwards.

Søren Kierkegaard

2.1. Introduction

The best place to start off with this chapter is precisely where Section 1.3 left off: the predictive states of processes.

In that section, we discussed how a (stationary, ergodic) stochastic process can be understood in terms of its word probabilities $\Pr(x_1 \ldots x_\ell)$, where the observations x_k are drawn from an "alphabet set" $\mathcal{X}$. We will start here by noting that this approach, while intuitive, only works well for the case that $\mathcal{X}$ is a discrete set, so that these probabilities can be nonzero. In the continuous case, we would be more likely to deal with a probability *density*, and more specifically, a probability *measure*. These details will be worked out later in this chapter; for now, if continuous states are your concern, you may mentally imagine that $\Pr(\cdot)$ denotes a density.

We also discussed in section 1.3 how conditional probabilities can be defined, allowing us to condition the likelihood of subsequent observations on prior observations:

$$\Pr(w_2 \mid w_1) = \frac{\Pr(w_1 w_2)}{\Pr(w_1)} \tag{2.1}$$

But wait!—you might be wondering—would this construction be well-defined in the continuous case, where the probability of a specific observation is zero? It would be nice to suppose that, in an appropriate limit, we could just take the ratio of densities. Generally, this *is* the case, but we will also have to address this in the main body of the chapter.

If we want to have as complete as possible a prediction of the next ℓ observations, it is generally necessary (unless the process is Markov at some finite order R) that condition on as much information

about the past as possible. The logical conclusion of this approach is to eventually condition on the infinite past:

$$\Pr\left(x_1 \ldots x_\ell \mid \overleftarrow{x} \right) = \lim_{N \to \infty} \Pr\left(x_1 \ldots x_\ell \mid x_{-N} \ldots x_0 \right) \tag{2.2}$$

For a specified ℓ, we may call this limit a *future morph*. The collection of all future morphs (at every ℓ) we will term the *predictive state*.

How should we think about this object? It is first helpful to draw an analogy between working with infinite sequences and working with continuous data. The set of infinite sequences has the cardinality of the continuum, and more importantly has close topological similarities to the real line. We will be discussing these mathematical aspects of temporal data in section 2.3. For now we will point out that the sequence of future morphs $\Pr\left(x_1 \ldots x_\ell \mid \overleftarrow{x} \right)$ for increasing ℓ specifies the probability of an increasingly *specific* future event, much as the sequence of probabilities for a continuous variable being in increasingly smaller intervals. Thus, the approach of measure theory will be useful for thinking about temporal limits just as it is for thinking about continuous data.

Additionally, the conditional limit in Eq. (2.2) is very similar to the "appropriate limit" that we earlier speculated would make Eq. (2.1) sensible for continuous variables. This is also no accident. A deeper understanding of the theory of conditional measures, and under which circumstances these conditional limits are valid, will be developed in Section 2.4.

Before we take our deep dive into the esotericities of conditional measure theory and infinite sequences, it would be prudent for me to justify to the reader why predictive states are worth the trouble, and in particular why the level of mathematical care we will take in this chapter is worth our while. We have established the concept of the predictive state; let us now review its history.

The predictive (or sometimes *causal* state) appeared in the study of dynamical systems, and the attempt to reconstruct the geometry of chaotic attractors from a data stream of measurements limited in number of variables (typically one) and precision (discretization at scale ϵ). In essence, it was discovered that sufficient information for characterizing the hidden attractor could be recovered in the form of the ϵ-*machine*: the term given for the dynamical model reconstructed from the data at precision scale ϵ [**44**]. The state space for the ϵ-machine was precisely the set of predictive states of the measured data.

Though we will postpone in-depth discussion of *models* of stochastic proceesses to Chapter 4, it will be helpful to briefly explain how the ϵ-machine functions. It is a kind of stochastic automaton known as a hidden Markov model. The states of this model are given by the unique predictive states of the process; that is, if we define an equivalence relation on pasts such that $\overleftarrow{x} \sim \overleftarrow{x}'$ if and only if

$$\Pr\left(w \mid \overleftarrow{x} \right) = \Pr\left(w \mid \overleftarrow{x}' \right)$$

for all words w, then the set of ϵ-machine states is the set of all equivalence classes inscribed by $\sim$.

Let η be one such equivalence class; it corresponds to a unique prediction $\Pr_\eta\left(w \right)$ of future words w. Let us focus on just the very next observation x. The epsilon machine, starting in state η, generates the observation x with probability $\Pr_\eta\left(x \right)$ and then transitions to the new state η' given by

$$\Pr_{\eta'}\left(w \right) = \frac{\Pr_\eta\left(xw \right)}{\Pr_\eta\left(x \right)}$$

(Equivalently, if $\overleftarrow{x}$ is in the equivalence class of η, then the new state η' is the equivalence class of $\overleftarrow{x}\, x$.)

From a data science standpoint, the broad goals of predictive-state analysis are threefold [**173**]. The first is to understand the overall structure of how the predictive states relate to one another geometrically. As mathematical objects, predictive states can be compared and distances between them can be quantified. In the inference setting, we may use this geometry to classify pasts based on equivalence of their predictive states. The second goal is to actually reproduce the prediction to a specified accuracy. That is, once we have embedded the predictive states in an abstract geometry, we must actually recover the useful information about a predictive state: what it predicts. The third is to understand the dynamics of how predictions evolve under a stream of new observations. This involves simulating the ϵ-machine of the reconstructed state.

Following the initial introduction of the predictive state approach and the ϵ-machine, both concepts have been employed in numerous settings, such as classical and quantum thermodynamics [**22**, **23**, **24**, **25**, **26**, **27**, **28**, **86**, **107**, **108**], quantum information and computing [**4**, **6**, **7**, **17**, **66**, **106**, **109**, **120**, **204**], condensed matter [**41**, **57**, **130**, **200**, **201**, **202**, **205**], renewal processes and spike-trains [**126**, **127**], dynamical systems [**51**, **87**, **88**, **89**, **167**], cellular automata [**162**], and model inference [**31**, **124**, **128**,

129, **131**, **163**, **164**, **165**, **174**, **176**, **184**, **185**, **186**]. While the original emphasis on the measurement-dependent aspect of the ϵ-machine has declined, the *name* "ϵ-machine" has stuck.

Predictive states were originally considered in the case where the process could be generated by a finite-state model, analagous to the regular languages of computational linguistics, and the natural "next step" appeared to be extending the concept to processes with more complex patterns, such as those that might be generated by a stack automaton or an indexed grammar [**44**]. In the time since, however, the finite-state case has proven to be an incredibly rich direction of study on its own, and a sophisticated mathematical theory has developed around the predictive states of hidden Markov models and their generalizations [**76**, **173**, **180**, **193**, **195**, **196**, **197**, **199**].

The content of the chapter is primarily drawn from the manuscript *Topology, Convergence, and Reconstruction of Predictive States* [**111**] in which we built new foundations for predictive state theory on the bedrock of measure theory. Our discussion of Cantor set geometry is from *Predictive State Geometry via Cantor Embeddings and Wasserstein Distance* [**112**]. After the writing of both these manuscripts I became aware that the specific result of the convergence of predictive states had been previously in Ref. [**199**], using Lévy's upwards theorem. Despite this, our independently derived approach is covered in full in this chapter, as the process of deriving the result offers key insights beyond the measure-theoretic convergence into the topology and geometry of predictive states, as well as the practical means by which the convergence is achieved.

From these new foundations, we are able to extend the applicability of predictive states beyond finite-state processes to *any* stationary, ergodic process, no matter how complex. Additionally, our extensions allow the definition of predictive states on continuous-valued data (how far from the days of ϵ we have come!). In Chapter 3 we will discuss how our broadened perspective offers insight to the utility of machine learning tools for reconstructing predictive state geometry.

The concepts handled in this chapter also lay the foundations for the use of predictive states in understanding models and their physical implementations in Chapters 4 through 6. In Chapters 4 and 5 we will show how predictive states form the bedrock of an entire theory of models of stochastic processes which encompasses hidden Markov models and several of their generalizations. We will learn to see predictive states as *dynamical* objects, and their dynamics have invariant symmetries which we will prove to be present in every dynamical model of the same stochastic process. Indeed,

we will in fact tie each dynamical model to an algebraic representation of the symmetry group of the predictive states. In Chapter 6, this representation theory of models will be used in conjunction with the resource theory of Section 1.5 to study the memory and energetic costs of physically implementing models of stochastic processes. Thus, predictive states allow us to concretely explore model space and understand its physical manifestations.

We hope this will justify to the reader why predictive states deserve such a close and scrutinizing examination such as we will provide in this chapter. We will now proceed with an overview of the mathematical tools and questions which will be the core of this chapter.

2.2. A brief introduction to measure theory and conditioning

The theory of measures and stochastic processes is deep and intricate, and even an introduction to these concepts will be fraught with rabbit holes which, while interesting, are irrelevant to our current considerations. Here we will briefly review the general concepts of measures and Lebesgue integration, but we will resist formal definitions and refer the reader to more appropriate venues for a properly rigorous introduction [**90**, **96**, **175**]. Once the general foundations have been laid we will also give a brief review of the problem of conditioning on measures, as this is the central formal roadblock which the current chapter seeks to divert. In the main body of this chapter we will be much more rigorous in establishing how stochastic processes may be handled by measure-theoretic approaches.

The *measure* is the mathematical formalization of the concept of "size," particularly as applied to sets. Given a space $\mathcal{X}$, a collection of its subsets Σ (frequently called the Σ-algebra) is selected, and the subsets within Σ are considered to be *measurable*. A (signed) measure is a function $\mu : \Sigma \to \mathbb{R} \cup \{\pm\infty\}$, which is subject to the conditions that $\mu(\emptyset) = 0$ and

$$\mu\left(\bigcup_j E_j\right) = \sum_j \mu(E_j)$$

where $\{E_j\}$ is an *at most countably large* collection of disjoint measurable sets. Thus a measure μ is considered to provide a sense of size to measurable sets which scales additively as sets are combined. It may take infinite values, though most of the measures that we will consider are bounded. Note

that implicitly, the Σ-algebra is required to be closed under countable unions. It is also supposed to be closed under countable intersection.

A standard example of a measure is the Lebesgue measure on $\mathbb{R}$, typically denoted by λ. The Σ-algebra is generally taken to be generated by all intervals $[a, b] \subset \mathbb{R}$, and the measure is given by $\lambda([a, b]) = b - a$.

Measures are a natural way to formulate the ideas of probability theory. The long-time frequency with which a variable is observed to take a value within the set U forms a measure μ with the property that $\mu(U) > 0$ always and $\mu(\mathcal{X}) = 1$ (where $\mathcal{X}$ is the full domain of possible values). Any measure satisfying these properties is called a probability measure.

Certain kinds of functions, known as *measurable* functions, can be integrated with respect to measures. Specifically, the definition of measurable for a function is relative to the measure μ in question; however, examples of non-measurable functions are typically non-constructive, relying on the axiom of choice to prove their existence, and so we will go forwards with the assumption that we are always dealing with measurable functions, since all the functions we will be concerned with have unambiguously constructive definitions.

The Lebesgue integral of a function f with respect to a measure μ over a measurable set U is written in the form

$$\int_U f(x) d\mu(x)$$

We will not give the precise definition here but will describe the basic intuition. Suppose we subdivide U into smaller, measurable sets U_i, and take the maximum (minimum) of f on each of these sets, denoted f_i, and compute the quantity $\sum_i f_i \mu(U_i)$; then the integral is the maximal (minimal) value attainable by this procedure over all subdivisions of U. If f is measurable, the max-min approach to this limit should be the same as the min-max approach. This approach is very similar to the Riemann integral, where the subdivisions are restricted to be intervals; when f is Riemann-integrable (that is, continuous everywhere except for a measure-zero set), the Lebesgue integral coincides with the Riemann integral.

It is common to think of probability distributions over continuous variables in terms of a probability *density*. If the probability density function of a distribution is given by a positive function f, then

the probability mass of a set $U \subseteq \mathbb{R}$ is given by

$$\Pr(U) = \int_U f_\mu(x)dx$$

In measure-theoretic terms, we are describing a measure defined by an integral over the Lebesgue measure:

$$\mu(U) = \int_U f_\mu(x)d\lambda(x)$$

Some features of this definition should be noted. For any set U with Lebesgue measure zero ($\lambda(U) = 0$), we must also have $\mu(U) = 0$ from the definition of the Lebesgue integral and the positivity of f.

Generally speaking, we say for two measures μ and ν that $\nu \ll \mu$, said "ν is absolutely continuous with respect to μ," if $\nu(U) = 0$ whenever $\mu(U) = 0$. Whenever this is true, the *Radon-Nikodym theorem* stipulates that there exists a function, denoted $d\nu/d\mu(x)$, such that

$$\nu(U) = \int_U \frac{d\nu}{d\mu}(x)d\mu(x)$$

The function $d\nu/d\mu(x)$ is called the *Radon-Nikodym derivative*, and is a generalization of the notion of a probability density function.

The Radon-Nikodym derivative is the pivot point for us to turn towards discussing the problem with conditioning on measures. There are two facets of conditioning which are extremely important to this chapter. First, we will discuss how conditional measures can be defined as a special application of the Radon-Nikodym derivative. This part is (deceptively) straightforward. The second facet is, as they say, the "catch": the Radon-Nikodym theorem is *non-constructive*, only proving the existence of the derivative, but giving us no method with which to compute it. Thus, pinning the definition of a conditional probability to the Radon-Nikodym derivative means conditional probabilities are also inherently non-constructive. Fortunately, there is a body of work that has provided numerous tools to construct the Radon-Nikodym derivative, and therefore conditional probabilities, in certain special cases [**156**]. Our task in this chapter will be to firmly establish stochastic processes as one of those cases.

When we talk about conditional probabilities, we will specifically assume that we are dealing with a joint measure μ_{XY} over a product space $\mathcal{X} \times \mathcal{Y}$. In this setting a conditional measure for Y conditioned on X is typically defined by a *regular conditional probability*, which is a function $\kappa_{Y|X} : \Sigma_Y \times \mathcal{X} \to \mathbb{R}$. Note that $\kappa_{Y|X}$ is a function of a point in $\mathcal{X}$ and a measurable set in $\mathcal{Y}$; effectively, it can also be seen as a function from $\mathcal{X}$ to measures on $\mathcal{Y}$. The conditional measure must satisfy the formula

$$\mu_{XY}(U \times V) = \int_U \kappa_{Y|X}(V, x) d\mu_X(x)$$

for all measurable $U \subset \mathcal{X}$ and $V \subset \mathcal{Y}$. The idea is that the probability of the joint event $x \in U$, $y \in V$ is given by the integral (over U) of the conditional probability of event $y \in V$.

This equation can be satisfied by use of the Radon-Nikodym theorem. Specifically, for a given set V, let $\mu_{X,V}(U) = \mu_{XY}(U \times V)$ be a measure on X. Then the conditional probability defined by

$$\kappa_{Y|X}(V, x) = \frac{d\mu_{X,V}}{d\mu_X}(x) \tag{2.3}$$

satisfies the conditional measure equation. Thus, the existence of regular conditional probabilities is just a special application of the Radon-Nikodym theorem; further, if we *have* a constructive definition of the derivative, then the conditional probability may be computed using that definition. Let us now turn our attention to this task.

The name "derivative" evokes the fundamental theorem of calculus: ν is the integral of $d\nu/d\mu$ because $d\nu/d\mu$ is the derivative of ν. One might intuitively then suppose that the Radon-Nikodym derivative is also a *differentiation*—that is, it is the limit of variations, just as the standard derivative df/dx is defined. Specifically, it would be *quite* desirable if the Radon-Nikodym derivative were equal to the limit

$$\frac{d\nu}{d\mu}(x) = \lim_{U \to x} \frac{\nu(U)}{\mu(U)} \tag{2.4}$$

where the limit is taken over sets U which contain x and "approach" it by contracting until they *only* contain x.

This is, indeed, possible, but only when certain conditions are met. When the measure's support is countable, the limit is trivial as points x have nonzero probability mass to themselves, and the

Radon-Nikodym derivative at x is *just* the ratio of the measures at x. The uncountable case is more subtle [**156**]. The crux of the issue is whether a certain property—called the *Vitali property*—holds to some degree of strength, in which case there exists a collection of sets— called the *differentiation basis*—which can be used to take the limit in Eq. (2.4).

Skipping several steps which will be elaborated upon in Section 2.3, the predictive state definition as

$$\Pr(w \mid \overleftarrow{x}) = \lim_{k \to \infty} \frac{\Pr(x_{-k} \ldots x_0 w)}{\Pr(x_{-k} \ldots x_0)} \tag{2.5}$$

is essentially the application of Eq. (2.3) and Eq. (2.4) in sequence: namely, we are identifying the conditional probability as a Radon-Nikodym derivative, and then identifying that derivative as a limit of the ratio of increasingly specific probabilities. As we have explained, Eq. (2.3) is a universally valid step, but Eq. (2.4) may not be. It is certainly not immediately clear that the second step ought to hold for *all* stationary, ergodic processes. Demonstrating that this is in fact the case is the primary goal of sections 2.3 and 2.4 of this chapter.

One of our first tasks in this chapter will be to demonstrate that the Vitali property holds in its strongest form on the set of sequences $\mathcal{X}^{\mathbb{N}}$ when $\mathcal{X}$ is a finite set. This proves Eq. (2.5) will converge to a valid conditional probability for any discretely-valued stochastic processes.

When $\mathcal{X}$ is a subset of $\mathbb{R}^d$, the situation is considerably more complicated. Even in $\mathbb{R}$ itself, the existence of a Vitali property is not trivial. For the Lebesgue measure, only a weak Vitali property holds, though this is still sufficient for the equivalence between Radon-Nikodym derivatives and likelihood ratios. The differentiation basis in this setting can be taken to be comprised of all intervals (a, b) on the real line. Going from $\mathbb{R}$ to $\mathbb{R}^d$, constraints must be placed on the differentiation basis. An "interval" here is really the Cartesian product of intervals, but for a Vitali property to hold we must only consider products of intervals whose edges are held in a fixed ratio to one another, so that the edges converge uniformly to zero. Likelihood ratios for fixed-aspect boxes of this kind can converge to the Radon-Nikodym derivative.

This requirement poses a challenge for generalizing the Vitali property to infinite dimensions, as we must to study sequences of real numbers. A fixed-aspect "box" around a sequence of real numbers is not a practical construction. In the empirical setting, we can only observe information about a finite number of past outputs. We therefore cannot obtain any "uniform" knowledge of the entire

past. A direct generalization of the case for $\mathbb{R}^d$ does not suffice. Nor, however, does a more relaxed generalization, where only a finite number of axes are constrained to be fixed-aspect at a time: the Vitali condition can be proven to be violated in this case [**80**]. This fact poses seemingly dire consequences for the validity of Eq. (2.5).

The tools for surmounting this obstacle actually come from the same source as the identification of the obstacle itself: the early work on integration of sequences by Jessen [**79**] and later Enomoto [**53**]. Their results focused on generalizing Lebesgue measure to $(S^1)^{\mathbb{N}}$, where S^1 is the circle. In Section 2.4 we will show that their results can be significantly extended. Though the Vitali property does not hold on $(S^1)^{\mathbb{N}}$ (or more generally $\mathcal{X}^{\mathbb{N}}$ for $\mathcal{X} \subset \mathbb{R}$), our generalization of Enomoto's Theorem will provide a differentiation basis for $\mathcal{X}^{\mathbb{N}}$ under which Eq. (2.5) provides a well-defined conditional probability.

An aspect of formulae like Eq. (2.4) and Eq. (2.5) which we have so far brushed over is the matter of *convergence of measures*. Both the aforementioned formulae describe the convergence of an *evaluation* of the measure: in Eq. (2.4), we are concerned with its evaluation on a set V, and in Eq. (2.5) with its evaluation on a word w.

What do these limits over evaluations say about the convergence of the *measure* in question? There are a variety of distinctly different manners in which a sequence of measures can converge. The kind of convergence which applies in our case is *convergence in distribution* [**90**]. A sequence of measures (μ_n) over $\mathcal{X}$ is defined to converge in distribution if

$$\lim_{n\to\infty} \int f(x) d\mu_n(x) = \int f(x) d\mu(x)$$

for every continuous function $f : \mathcal{X} \to \mathbb{R}$. We will try to avoid abstract topology here, but for our purposes it will suffice to assume that $\mathcal{X}$ is either discrete (in which case any function is continuous), a subset of $\mathcal{R}$ (in which case we simply inherent the definition of continuity from $\mathbb{R}$), or a space of sequences (in which case we will address continuity in Section 2.3). It turns out that convergence in distribution of measures is equivalent to the convergence-on-words that we establish holds in Eq. (2.5) (we will prove this equivalence in Section 2.3).

The purpose of this section has been to give the reader a brief introduction to the key concepts of measure theory, and why our lengthy examination of conditioning on stochastic processes in the

main body is necessary. In Sections 2.3, 2.4 and 2.3.4 we will elaborate further on how to address these issues for stochastic processes.

2.3. The structure of temporal data

In this section we will cover the basic mathematics of stochastic processes and symbolic dynamics, introducing some example processes which will follow us throughout the text as well as some useful theorems. The eexamples will be presented in Section 2.3.1.

The methods here are intended to be applied to stationary and ergodic stochastic processes that generate a discrete-time sequence of data. In the case of categorical data, we can consider a stochastic process to be a collection of probability distributions $\Pr_\mu(x_1 \ldots x_L)$ over any finite, contiguous sequence, taking values in a finite set $\mathcal{X}$. Formally, this describes a measure μ over the set of all bi-infinite sequences $(\ldots, x_{-1}, x_0, x_1, \ldots) \in \mathcal{X}^{\mathbb{Z}}$. For real-valued data, we can similarly consider the process to be a collection of *measures* over $\mathcal{X}^L$, which altogether constitute a single measure over $\mathcal{X}^{\mathbb{Z}}$. In Sections 2.3.2 and 2.3.3 we will cover the formal nuances of defining these measures and the geometry of sequences. In Section 2.3.4 we demonstrate that the convergence in distribution for measures on $\mathcal{X}^{\mathbb{N}}$ is equivalent to the convergence of the measure on all finite-word probabilities.

2.3.1. Example processes. Stochastic processes are generated by a number of systems with widely varying complexity. Most popularly studied are those often characterized as having a degree of "finite memory": *Markov chains*, *hidden Markov models*, and *observable operator models* (also termed *generalized hidden Markov models*) [**76**, **199**]. Beyond these, one can also generate processes using probabilistic grammars, such as *probabilistic context-free* and *indexed grammars* [**63**]. Additionally, coarse-grained data from chaotic dynamical systems—such as the *logistic map*—display behavior varying widely in complexity [**38**].

We refer back frequently to the following example processes:

(1) The *even process* can be generated by repeatedly tossing a coin and writing down a 0 for every tail and 11 for every head; thus a sample might look like "011001101111." The process is essentially random except that 1s only appear in contiguous blocks of even size bounded by 0s. The even process has infinite Markov order but can be generated by a two-state hidden Markov chain [**42**].

(2) A *renewal process*, usually defined over continuous-time, can be defined for discrete time as follows. A renewal process emits 0s for a randomly selected duration before emitting a single 1 and then randomly selecting a new duration to fill with 0s [**125**]. A *renewal process* is specified by the survival probability $\Phi(n)$ that a contiguous block of 0s has length at least n. The exact probability of a given length is $F(n) := \Phi(n) - \Phi(n+1)$. It is always assumed that $\Phi(1) = 1$. Further, stationarity requires that $m := \sum_{n=1}^{\infty} \Phi(n)$ be finite, as this gives the mean length of a block of 0s.

(3) The $a^n b^n$ *process* can be generated by choosing a random integer $n \geq 1$ (we suppose via a Poisson process) and writing n `a`'s followed by an equal number of `b`'s, and then repeating this procedure indefinitely. This results in sequences where any contiguous block of `a`s is followed by a block of `b`s of equal size. The $\texttt{a}^n\texttt{b}^n$ process cannot be generated by any finite hidden Markov chain, though it is a simple example of a probabilistic context-free language [**69**].

(4) The $x+f(x)$ *process* is a probabilistic context-free language modeling the syntactic structure of simple mathematical expressions. It has terminal symbols {`(` , `)` , `;` , `+` , `f` , `x`} and nonterminals $\{A, B, C\}$, and starts with a sequence of As. Sequences are generated by applying the production rules:

$$A \mapsto B + C \ ; \mid C \ ;$$

$$B \mapsto B + C \mid C$$

$$C \mapsto \texttt{f}(B) \mid \texttt{x} \ .$$

(5) The $a^n b^n c^n$ *process* is a probabilistic indexed language [**69**] that is analogous to $\texttt{a}^n\texttt{b}^n$ except after writing the blocks of `a`'s and `b`'s, we also write a block of `c`'s of length n.

(6) The *Morse-Thue process* is generated by sampling from the time series of the logistic map at critical "onset of chaos" parameter $r_c \approx 3.56995$:

$$y_{t+1} = r y_t (1 - y_t)$$

and then coarse-graining the data by taking $x_t = 0$ if $0 < y_t \leq \frac{1}{2}$ and $x_t = 1$ if $\frac{1}{2} < y_t < 1$ [96]. Alternatively, we can generate this process by starting with a single 0 and executing the replacements $0 \mapsto 11$ and $1 \mapsto 01$ consecutively. The resulting process is an indexed-context free language [38].

2.3.2. Processes via measures. A *stochastic process* is typically defined as a function-valued random variable $X : \Omega \to \mathcal{X}^{\mathcal{T}}$, where (Ω, Σ, μ) is a measure space, $\mathcal{T}$ is a set of temporal indices (perhaps the real line, perhaps a discrete set), and $\mathcal{X}$ is a set of possible observations (also potentially real or discrete in nature). We take the sample space Ω to be the set $\mathcal{X}^{\mathcal{T}}$ and X to be the identity. In this way, a stochastic process is identified solely with the measure μ over $\Omega = \mathcal{X}^{\mathcal{T}}$.

When $\mathcal{T}$ is $\mathbb{Z}$, we say the process is *discrete-time*; when it is $\mathbb{R}$ we say *continuous-time*. Unless specified otherwise we assume discrete-time, later treating continuous-time as an extension of the discrete case. In discrete time, it is convenient to write $X(t)$ as an indexed sequence (x_t), where each x_t is an element of $\mathcal{X}$. When $\mathcal{X}$ is a discrete finite set, we say that the process is *discrete-observation*; by *continuous-observation* we typically mean the case where $\mathcal{X}$ is an interval in $\mathbb{R}$ or a Cartesian product of intervals in $\mathbb{R}^d$. These are the only cases we consider rigorously. That said, we believe they are sufficient for many practical purposes or, at least, not too cumbersome to extend if necessary.

The temporal *shift operator* $\tau : \mathcal{X}^{\mathbb{Z}} \to \mathcal{X}^{\mathbb{Z}}$ simply translates $t \mapsto t + 1$: $(\tau X)(t) = X(t + 1)$. It also acts on measures of $\mathcal{X}^{\mathbb{Z}}$: $(\tau\mu)(A) = \mu(\tau^{-1}A)$. A stochastic process paired with the shift operator—$(\mathcal{X}^{\mathbb{Z}}, \Sigma, \mu, \tau)$—becomes a dynamical system and is *stationary* if $\tau\mu = \mu$. It is further considered *ergodic* if, for all shift-invariant sets $\mathcal{I} \subseteq \mathcal{X}^{\mathbb{Z}}$, either $\mu(\mathcal{I}) = 1$ or $\mu(\mathcal{I}) = 0$. Here, we assume all processes are both stationary and ergodic.

We will note two implications of this ergodicity. The first is the ergodic theorem, which states that for any function $f : \mathcal{X}^{\mathcal{Z}} \to \mathbb{R}$,

$$\lim_{T \to \infty} \sum_{t=0}^{\infty} f(\tau^t X) = \int f(X) d\mu(X) \tag{2.6}$$

In other words, temporal averages are equal to instantaneous averages. The other implication of ergodicity (which follows from the above) is that the "path" taken from any starting point eventually fills the system. If we let $\mathbf{1}_U$ denote the indicator function of any set $U \subseteq \mathcal{X}^{\mathcal{Z}}$, so that $\mathbf{1}_U(X) = 1$

whenever $X \in U$ and $\mathbf{1}_U(X) = 0$ otherwise, we have

$$\lim_{T\to\infty} \sum_{t=0}^{\infty} \mathbf{1}_U(\tau^t X) = \mu(U) \tag{2.7}$$

In other words, if U has positive measure, we will revisit it recurrently over time, no matter our starting point. The reader will hopefully note that these three characterizations of ergodicity match those we provided in the context of Section 1.4.

If $\mathcal{X}$ is discrete, then the measurable sets of $\mathcal{X}^{\mathcal{Z}}$ are generated by the *cylinder sets*:

$$U_{t,w} := \{\, X : x_{t+1} \dots x_{t+\ell} = w \,\} \ ,$$

where $w \in \mathcal{X}^\ell$ is a *word* of length ℓ. For a stationary process, the *word probabilities*:

$$\Pr_\mu (\, x_1 \dots x_\ell \,) := \mu \left(U_{0,x_1 \dots x_\ell} \right)$$

are sufficient to uniquely define the measure μ.

In the continuous-observation case, the issue is more subtle. A cylinder set instead takes the form:

$$U_{t,I_1 \dots I_\ell} := \{\, X : x_{t+1} \in I_1, \dots, x_{t+\ell} \in I_\ell \,\} \ ,$$

where each I_t is an interval in $\mathcal{X}$. This does not lend itself well to expressing simple word probabilities. However, we can define the *word measures* μ_ℓ by restricting μ to the set $\mathcal{X}^\ell$ describing the first ℓ values. We use the notation

$$\Pr_\mu (\, I_1 \dots I_\ell \,) := \mu \left(U_{0,I_1 \dots I_\ell} \right)$$

similarly to the discrete case when we want to specify that a word is constrained by a sequence of intervals.

The set of all measures over $\mathcal{X}^{\mathbb{N}}$ will be denoted $\mathbb{M}(\mathcal{X}^{\mathbb{N}})$, and the set of all *probability* measures will be denoted $\mathbb{P}(\mathcal{X}^{\mathbb{N}})$.

2.3.3. The self-similar geometry of time. The geometry of sequences is inherently self-similar. Given an infinite sequence $\overrightarrow{x} = (x_1, x_2, \dots)$, we can split it into its leading word $x_1 x_2 \dots x_L$

and a following sequence $\overrightarrow{x}_L = (x_L, x_{L+1}, \dots)$. That is, the space of sequences $\mathcal{X}^{\mathbb{N}}$ can be factored into $\mathcal{X}^L \times \mathcal{X}^{\mathbb{N}}$ for any L. The fractal nature of sequence-space is encoded in the structure of its *product topology*.

Topology is, in many ways, just the study of similarity: defining when things are similar to one another, and how different kinds of similarity are linked. In the case of sequences, we say that two sequences $\overrightarrow{x}$ and $\overrightarrow{y}$ are similar if they share matching symbols, and they are more similar the more matching symbols they share. This intuition is encoded in the product topology, which defines the basic "neighborhoods" of a sequence $\overrightarrow{x}$ to be generated by its cylinder sets: $U_{0,x_1 \dots x_\ell}$ in the discrete case and in the continuous case, $U_{0,I_1 \dots I_\ell}$ where the intervals I_k are open and $x_k \in I_k$. The product topology endows the space of sequences with several useful properties; most importantly, if $\mathcal{X}$ is a compact set (automatically true for discrete $\mathcal{X}$, and true for closed and bounded $\mathcal{X} \subset \mathbb{R}$), then $\mathcal{X}^{\mathbb{Z}}$ and $\mathcal{X}^{\mathbb{N}}$ are also compact in the product topology.

Two useful families of distance metrics, equivalent to the product topology on $\mathcal{X}^{\mathbb{N}}$, are the Euclidean metrics, one for the discrete and real case each:

$$D_{\mathrm{E},\gamma}(X,Y)^2 := \begin{cases} \sum_{t=1}^{\infty} (1 - \delta_{x_t y_t}) \gamma^{2t} & \mathcal{X} \text{ discrete} \\ \sum_{t=1}^{\infty} \|x_t - y_t\|^2 \gamma^{2t} & \mathcal{X} \subset \mathbb{R}^d \end{cases},$$

for some $0 < \gamma < 1$. These distance metrics arise from embedding $\mathcal{X}^{\mathbb{N}}$ in a Hilbert space. Given an orthogonal basis (e_i), the components of this embedding for the discrete case are given by:

$$c_i(X) = \begin{cases} \gamma^{\lfloor i/|\mathcal{X}| \rfloor} & x_{\lfloor i/|\mathcal{X}| \rfloor} = i \bmod |\mathcal{X}| \\ 0 & \text{otherwise} \end{cases}$$

and in the continuous case ($\mathcal{X} \subset \mathbb{R}^d$) by:

$$c_i(X) = \gamma^t x_{k,t}, \quad i = k \mod d \, .$$

The Euclidean distance has an interesting "Pythagorean theorem." Define the restricted distance on $\mathcal{X}^\ell$:

$$D^{(\ell)}_{\mathrm{E},\gamma}(X,Y)^2 := \begin{cases} \sum_{t=1}^{\ell}(1-\delta_{x_t y_t})\gamma^{2t} & \mathcal{X} \text{ discrete} \\ \sum_{t=1}^{\ell} \|x_t - y_t\|^2 \gamma^{2t} & \mathcal{X} \subset \mathbb{R}^d \end{cases},$$

for $X, Y \in \mathcal{X}^\ell$. Then:

$$\begin{aligned} D_{\mathrm{E},\gamma}(X,Y)^2 =& D^{(\ell)}_{\mathrm{E},\gamma}(x_1 \dots x_\ell, y_1 \dots y_\ell)^2 \\ &+ \gamma^{2\ell} D_{\mathrm{E},\gamma}(x_{\ell+1} \dots, y_{\ell+1} \dots)^2 \end{aligned} . \tag{2.8}$$

This theorem provides an algebraic expression of the self-similarity of the product topology.

For discrete sequences, we can exploit the self-similar geometry in an interesting way by constructing another distance metric which also generates the product topology. This is done by constructing an embedding between sequence space and the celebrated *Cantor set* (or one of its generalizations). Suppose a symbolic sequence $(x_1, x_2, \dots)$ takes values in an alphabet $\mathcal{X}$ of size $|\mathcal{X}|$. To each $x \in \mathcal{X}$ we associate a unique integer between 0 and $|\mathcal{X}| - 1$ inclusive; call this $J(x)$. Then, there is a function $C : \mathcal{X}^{\mathbb{N}} \to [0,1]$ that maps every sequence to a positive real number:

$$C(x_1, x_2, \dots) = \sum_{k=1}^{\infty} \frac{2J(x_k)}{(2|\mathcal{X}|-1)^k} .$$

For instance, suppose that $|\mathcal{X}| = 2$ has two elements; then C maps the sequence to a point in the traditional Cantor set fractal. For a finite sequence of length L, truncate the sum at $k = L$.

Remarkably, the embedding C has the property that for any continuous function f on $[0,1]$, the function $F(\overrightarrow{x}) = f(C(\overrightarrow{x}))$ is continuous on $\mathcal{X}^{\mathbb{N}}$. Further, if F is continuous on $\mathcal{X}^{\mathbb{N}}$, then $f(y) = F(C^{-1}(y))$ is continuous on the image. Thus, the embedding C respects the basic structure of the product topology [96].

Stationary processes, due to their time-translation invariance, inherit the fractal temporality of sequence space. This can be easily visualized: Given a length-L sample $x_1 \dots x_L$, and $n, k > 0$, take a sliding window of pasts and futures, $(x_{t-k+1} \dots x_t, x_{t+1} \dots x_n)$ for $t = k, \dots, L-n$. For each past-future pair, compute the truncated Cantor embeddings on the *reversed* past and (unreversed) future: $(C(x_t \dots x_{t-k+1}), C(x_{t+1} \dots x_n))$. The resulting pairs of real numbers can be plotted as

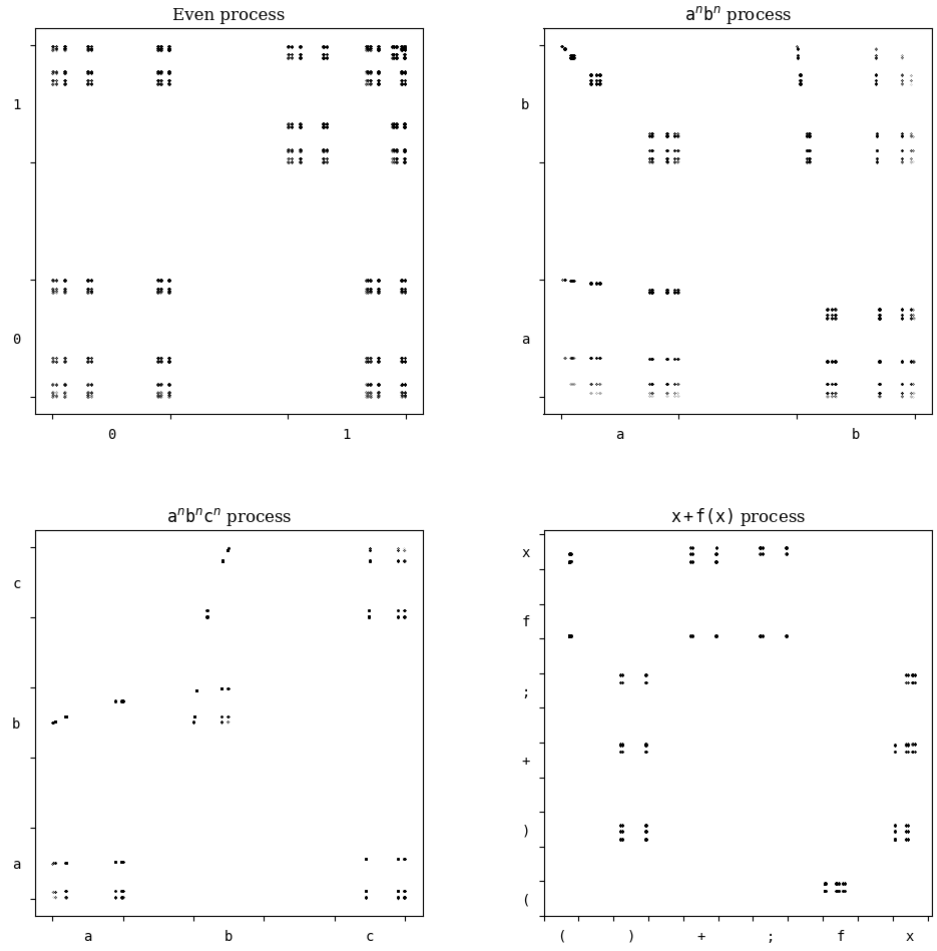

FIGURE 2.1. Cantor plots for the even, $\mathtt{a}^n\mathtt{b}^n$, $\mathtt{a}^n\mathtt{b}^n\mathtt{c}^n$, and $\mathtt{x}+\mathtt{f}(\mathtt{x})$ processes. Each point (x, y) corresponds to a pair of sequences corresponding to the past and future, respectively. The symbol on the x (y) axis indicates that all points above (to the right of) that symbol have a past (future) whose most recent observation is that symbol. Though not marked, further proportional subdivisions of each segment of the axes indicate the value of the second, third, *etc.* symbols. For instance, one can read from the $\mathtt{x}+\mathtt{f}(\mathtt{x})$ fractal that any past ending in `f` must be paired with a future beginning in (`f` or (`x`.

(x, y)-values on a scatter plot. The fractal that emerges contains, in essence, all information necessary to understand a process' temporal structures. See Fig. 2.1 for examples and guidance on how to interpret the visualization.

Note that for $|\mathcal{X}| > 2$ the embedding C introduces additional structure that may or may not be desired. Associating each symbol x with an integer j_x endows an ordinal structure on the set $\mathcal{X}$. This ordinality is present in the macroscopic geometry of $C\left(\mathcal{X}^{\mathbb{N}}\right)$.

2.3.4. Continuity and convergence of measures. A central feature of our result on predictive states is that they converge in distribution as more information from the past is provided. Convergenec in distribution is defined in terms of continuous functions. Namely, a sequence of measures μ_k over $\mathcal{X}^{\mathbb{N}}$ converges to a measure μ in distribution if, for every continuous function $f : \mathcal{X}^{\mathbb{N}} \to \mathbb{R}$,

$$\lim_{k\to\infty} \int_{\mathcal{X}^{\mathbb{N}}} f(\overrightarrow{x}) d\mu_k(\overrightarrow{x}) = \int_{\mathcal{X}^{\mathbb{N}}} f(\overrightarrow{x}) d\mu(\overrightarrow{x})$$

To relate this definition of convergence to our own intuitions of stochastic processes, we must have a better understanding of continuity in sequence-space.

The definition of continuity depends on the product topology, whose neighborhoods are cylinder sets. For the discrete case, a simple rendering of the definition of continuity is this: a function f is continuous if, for every $\overrightarrow{x} \in \mathcal{X}$ and some small number $\epsilon > 0$, there is a sufficiently large time t such that $|f(\overrightarrow{y}) - f(\overrightarrow{x})| < \epsilon$ whenever $y_1 \dots y_t = x_1 \dots x_t$. In other words, if two sequences match sufficiently far into the future, then their function values will be arbitrarily close.

Another feature of continuity on $\mathcal{X}^{\mathbb{N}}$ comes to us by virtue of the compactness of the space. The Heine-Cantor theorem asserts that any continuous function on a compact space is *uniformly* continuous. This means that we can in fact amend our definition of continuity: for any small $\epsilon > 0$, there is a single time $t > 0$ after which it is guaranteed that $|f(\overrightarrow{y}) - f(\overrightarrow{x})| < \epsilon$ for any two $\overrightarrow{x}, \overrightarrow{y}$ who match on the first t symbols: $y_1 \dots y_t = x_1 \dots x_t$. In other words, convergence occurs at (at most) a uniform rate at every point; there are no "straggler points" who take an arbitrarily long time to converge compared to other points.

For the following theorem, we call a measure μ "full" if it assigns positive measure to every cylinder set.

PROPOSITION 1 (Continuity via word averages). *A function $f : \mathcal{X}^{\mathbb{N}} \to \mathbb{R}$ is continuous if and only if the functions*

$$f_{\mu,\ell}(x_1 \dots x_\ell) = \frac{\int_{U_{0,x_1\dots x_\ell}} f(\overrightarrow{x}) d\mu(\overrightarrow{x})}{\mu(U_{0,x_1\dots x_\ell})}$$

are continuous on $\mathcal{X}^\ell$ and converge to $f(\overrightarrow{x})$ uniformly over $\overrightarrow{x}$ and every full measure μ, as $\ell \to \infty$.

PROOF. *Suppose f is continuous; then it is uniformly continuous, and so for every $\epsilon > 0$ there is a t so that, for every $\overrightarrow{x}$ and μ, $|f_{\mu,\ell}(x_1 \dots x_\ell) - f(\overrightarrow{x})| < \epsilon$. This follows because f will be close to $f(\overrightarrow{x})$ on the cylinder set being averaged over, and so the average will be close. Further, continuity of $f_{\mu,\ell}$ will be inherited from the continuity of f. Thus the forward implication is true.*
For the converse, consider the fact that $f_{\mu,\ell}$ can be extended to a function on $\mathcal{X}^{\mathbb{N}}$ as $f_{\mu,\ell}(\overrightarrow{x}) = f_{\mu,\ell}(x_1 \dots x_\ell)$. Each of these extensions is necessarily continuous on $\mathcal{X}^{\mathbb{N}}$. Because they converge uniformly to f, f must be continuous by virtue of the Uniform Limit theorem (which states that a uniform convergence of continuous functions results in a continuous function).

Let us keep in mind that if $\mathcal{X}$ is discrete, then the requirement that $f_{\mu,\ell}$ is continuous is trivial.
We can now demonstrate that convergence-in-distribution will be equivalent to convergence over word distributions. We will state the full result and then explain the implications afterward.

THEOREM 2 (Equivalence of convergence-over-words and convergence-in-distribution). *Let μ_k be a sequence of measures on $\mathcal{X}^{\mathbb{N}}$ and let μ be a measure over the same. Then $\mu_k \to \mu$ in distribution if and only if $\mu_k|_\ell \to \mu|_\ell$ in distribution for every ℓ, where $\nu|_\ell$ is the marginalization of the measure ν to $\mathcal{X}^\ell$.*

PROOF. *By virtue of Prop. 1, for any continuous function f we will have for all measures ν:*

$$\lim_{\ell\to\infty} \int f_{\nu,\ell}(x_{1:\ell}) d\left(\nu|_\ell\right)(x_{1:\ell}) = \int f(\overrightarrow{x}) d\nu(\overrightarrow{x})$$

If we replace ν with μ_k and μ respectively, then convergence in distribution has the form

$$\lim_{k\to\infty} \lim_{\ell\to\infty} \int f_{\mu_k,\ell}(x_{1:\ell}) d\left(\mu_k|_\ell\right)(x_{1:\ell}) \lim_{k\to\infty} \int f(\overrightarrow{x}) d\mu_k(\overrightarrow{x}) = \int f(\overrightarrow{x}) d\mu(\overrightarrow{x})$$

for all continuous f.
On the other hand, convergence of $\mu_k|_\ell \to \mu|_\ell$ for all ℓ takes the form of the requirement

$$\int f(\overrightarrow{x}) d\mu(\overrightarrow{x}) = \lim_{\ell\to\infty} \int f_{\mu,\ell}(x_{1:\ell}) d\left(\mu|_\ell\right)(x_{1:\ell}) = \lim_{\ell\to\infty} \lim_{k\to\infty} \int f_{\mu_k,\ell}(x_{1:\ell}) d\left(\mu_k|_\ell\right)(x_{1:\ell})$$

for all continuous f.

Equivalence of these two convergences then just boils down to the interchange of limits $\lim_{\ell\to\infty}\lim_{k\to\infty} \leftrightarrow \lim_{k\to\infty}\lim_{\ell\to\infty}$. *By the Moore-Osgoode theorem, this interchange is in fact valid whenever the limit*

$$\lim_{\ell\to\infty}\int f_{\mu_k,\ell}(x_{1:\ell})d\left(\mu_k|_\ell\right)(x_{1:\ell}) = \int f(\overrightarrow{x})d\mu_k(\overrightarrow{x})$$

is uniformly convergent over all k. *But this is guaranteed by Prop. 1, and so the two forms of convergence are equivalent.*

Theorem 2 guarantees that in order to demonstrate convergence in distribution, it is sufficient that the measures converge on ther marginalizations to finite words. For discrete $\mathcal{X}$, this means that

$$\lim_{k\to\infty}\Pr_{\mu_k}(\,w\,) = \Pr_{\mu}(\,w\,)$$

for all w is equivalent to convergence in distribution. This is extremelly convenient, as word probabilities are perhaps the most intuitive way to interact with the measure.

For the case of $\mathcal{X}\subset\mathbb{R}$, the situation is more subtle. The Portmanteau theorem [**90**] states that convergence in distribution is equivalent to a very weak *bounded* convergence over open sets. In our case, this means that

$$\liminf_{k\to\infty}\Pr_{\mu_k}(\,I_1\ldots I_\ell\,) \geq \Pr_{\mu}(\,I_1\ldots I_\ell\,)$$

for all open neighborhoods $I_1\times\cdots\times I_\ell$ of any length is equivalent to convergence in distribution. In fact, what we will prove for predictive states is a somewhat stronger form of convergence than this, where we maintain the equality at each ℓ, though this is still not nearly as strong as other forms of convergence over $\mathcal{X}^{\mathbb{N}}$, and in most practical cases it is equivalent to convergence in distribution.

2.4. Conditioning on sequences: discrete intuitions

Each element $x\in\mathcal{X}^{\mathbb{Z}}$ can be decomposed from a bidirectional infinite sequence to a pair of unidirectional infinite sequences in $\mathcal{X}^{\mathbb{N}}\times\mathcal{X}^{\mathbb{N}}$, by the transformation $\ldots x_{-1}x_0x_1\cdots\mapsto(x_0x_{-1}\ldots,x_1x_2\ldots)$. The first sequence in this pair we call the *past* $\overleftarrow{x}$ and the second we call the *future* $\overrightarrow{x}$. In this perspective, a stochastic process is a bipartite measure over pasts and futures. The intuitive definition of a *predictive state* is as a measure over future sequences that arises from conditioning

on past sequences. Heuristically, $\Pr_\mu\left(\overrightarrow{X} \,\middle|\, \overleftarrow{x} \right)$ represents the "predictive state" associated with past $\ldots x_{-1}x_0$.

Formally, the predictive state can be defined in terms of a Radon-Nikodym derivative. Consider the discrete case. Let $\overleftarrow{\mu}$ denote the restriction of μ to pasts, and let $\overleftarrow{\mu}_{x_\ell \ldots x_1}$ be the measure on pasts which precede the word $w := x_1 \ldots x_\ell$. These are given by:

$$\Pr_{\overleftarrow{\mu}}(x_0 \ldots x_{-k}) := \Pr_\mu(x_{-k} \ldots x_0)$$

$$\Pr_{\overleftarrow{\mu}_{x_\ell \ldots x_1}}(x_0 \ldots x_{-k}) := \Pr_\mu(x_{-k} \ldots x_0 x_1 \ldots x_\ell) \ .$$

Then the predictive state can be defined as a Radon-Nikodym derivative

$$\Pr_\mu(x_1 \ldots x_\ell \mid \overleftarrow{x}) = \frac{d\overleftarrow{\mu}_{x_\ell \ldots x_1}}{d\overleftarrow{\mu}}(\overleftarrow{x}) \ . \tag{2.9}$$

This definition has the benefit that, for any word $w = x_{-n} \ldots x_0$,

$$\Pr_\mu(x_{-n} \ldots x_0 x_1 \ldots x_\ell) = \int_{U_{0,x_0 \ldots x_{-n}}} \Pr_\mu(x_1 \ldots x_\ell \mid \overleftarrow{x}) \, d\overleftarrow{\mu}(\overleftarrow{x}) \ . \tag{2.10}$$

Thus the predictive state acts as a proper conditional probability. The continuous case is equivalent, with the replacement of symbols x_k with intervals I_k.

As discussed in Section 2.2, there is a downside to the Radon-Nikodym formulation, which is that it is not generally constructive. We therefore have to provide a definition of predictive state which is grounded directly in ratios of probabilities which can be empirically measured. In this section we will show the following:

(1) The predictive state, defined in the form of a Radon-Nikodym derivative, has a computable formulation as a limit of ratios of word probabilities. In the discrete case, this looks like:

$$\Pr_\mu(x_1 \ldots x_\ell \mid \overleftarrow{x}) = \lim_{k \to \infty} \frac{\Pr_\mu(x_{-k} \ldots x_0 x_1 \ldots x_\ell)}{\Pr_\mu(x_{-k} \ldots x_0)}$$

This limit works for *every* stationary and ergodic process, regardless of complexity.

(2) Because the above limit exists for every word $w = x_1 \ldots x_\ell$, the convergence of the predictive state *as a measure* is convergence in distribution, as per Thm. 2. The measure associated with the predictive state $\Pr_\mu(\cdot \mid \overleftarrow{x})$ will be denoted $\epsilon[\overleftarrow{x}]$.

(3) An example will be provided (namely, the Feigenbaum process) of a distribution for which stronger forms of convergence (namely, convergence on sets and convergence in total variation) do not hold, thus making convergence in distribution the strongest form of convergence guaranteed for predictive states.

This section will focus on the case of discrete $\mathcal{X}$ in section 2.4.1. This will introduce us to the basic issues at hand which, while not trivial for the discrete case, may be handled fairly straightforwardly. In Section 2.3.4 we will examine the case of continuous observations, reviewing the previous literature on the nuances of this domain and extending its results for our present purposes, in sections 2.3.4 through 2.5.2.

Because the mathematics in this section is far from either elegant or intuitive, we will close by reviewing our results in Section 2.4.2 with example cases.

2.4.1. Discrete observations. Let $\overleftarrow{\mu}$ denote the restriction of μ to pasts. We will establish the following:

THEOREM 3. *For all measures μ on $\mathcal{X}^{\mathbb{Z}}$, all $\ell \in \mathbb{N}$, all $w = x_1 \ldots x_\ell \in \mathcal{X}^\ell$, and $\overleftarrow{\mu}$-almost all pasts $\overleftarrow{x}$, where $\mathcal{X}$ is a finite set, the following limit is convergent:*

$$\Pr_\mu (w \mid \overleftarrow{x}) = \lim_{k \to \infty} \frac{\Pr_\mu (x_{-k} \ldots x_0 x_1 \ldots x_\ell)}{\Pr_\mu (x_{-k} \ldots x_0)} \tag{2.11}$$

Note that by Thm. 2, this implies that the predictive state converges in distribution.

For all $\overleftarrow{x}$ where Eq. (2.11) converges, we can define a measure $\epsilon[\overleftarrow{x}] \in \mathbb{P}(\mathcal{X}^{\mathbb{N}})$ over future sequences, uniquely determined by the requirement $\epsilon[\overleftarrow{x}](U_{0,w}) = \Pr_\mu (w \mid \overleftarrow{x})$. This $\epsilon[\overleftarrow{x}]$ is the *predictive state* of $\overleftarrow{x}$ and the function $\epsilon : \mathcal{X}^{\mathbb{N}} \to \mathbb{M}(\mathcal{X}^{\mathbb{N}})$, the *prediction mapping*.

The proof strategy will consist in redefining the problem. The limit Eq. (2.11) can be recast as what is called a *likelihood ratio*. The convergence of likelihood ratios is itself closely related to the theory of Radon-Nikodym derivatives between measures. Specifically, the Radon-Nikodym derivative can be computed as a convergence of likelihood ratios of that convergence is taken over a particular class of neighborhoods, called a *differentiation basis*, and that basis has a property called the *Vitali property*. We will define these concepts for the reader below and use them to prove Theorem 1.

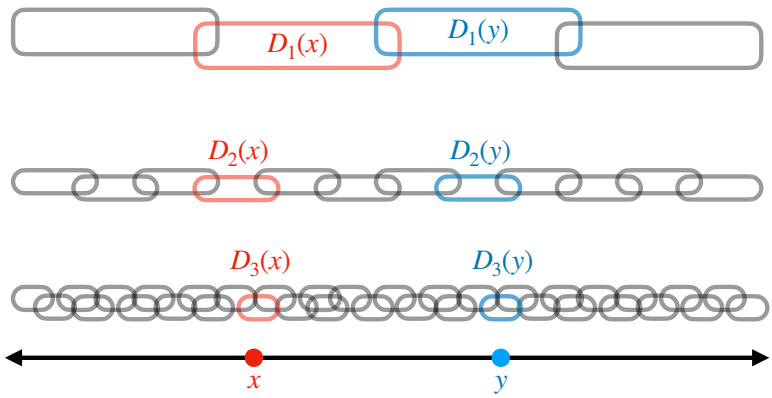

FIGURE 2.2. *Snapshot of a differentiation basis.* A differentiation basis is a collection of neighborhoods in $\mathcal{X}^{\mathbb{N}}$, which have hierarchical structure. For every point $x \in \mathcal{X}^{\mathbb{N}}$, there must be a sequence of neighborhoods converging on that point. Pictured above, a line is shown with a partial representation of its differentiation basis above it in the form of a hierarchical collection of rounded rectangles. For two points x and y we show the corresponding sequence of sets $(D_j(x))$, $(D_j(y))$ converging on each.

Let $\overleftarrow{\mu}$ and $\overleftarrow{\mu}_{x_\ell \dots x_1}$ be defined as in the start of this section. Then Eq. (2.11) can be recast in the form of a convergence of likelihood ratios, taken over a sequence of cylinder sets $U_k := U_{0,x_0 \dots x_{-k}}$ converging on $\overleftarrow{x}$:

$$\Pr_\mu (x_1 \dots x_\ell \mid \overleftarrow{x}) = \lim_{k \to \infty} \frac{\overleftarrow{\mu}_{x_\ell \dots x_1}(U_k)}{\overleftarrow{\mu}(U_k)} \ . \tag{2.12}$$

This reformulation, though somewhat conceptually cumbersome, is useful because of existing theorems relating the convergence of likelihood ratios to the Radon-Nikodym derivative. Indeed, wherever Eq. (2.12) converges, it will be equal to the Radon-Nikodym derivative $d\overleftarrow{\mu}_{x_\ell \dots x_1} / d\overleftarrow{\mu}(\overleftarrow{x})$.

To use these theorems we must define a *differentiation basis*. Any collection of neighborhoods $\mathcal{D}$ in $\mathcal{X}^{\mathbb{N}}$ may be considered a differentiation basis if for every $\overleftarrow{x} \in \mathcal{X}^{\mathbb{N}}$, there exists a sequence of neighborhoods (D_k) such that $\lim_{k \to \infty} D_k = \{ \overleftarrow{x} \}$. See Fig. 2.2.

The Vitali theorem states that whenever the differentiation basis $\mathcal{D}$ possesses the *Vitali property* with respect to two measures ν and μ, then for μ-almost all $\overleftarrow{x}$, the limit of likelihood ratios exists for any sequence $(V_k) \subset \mathcal{D}$ converging on $\overleftarrow{X}$ and is equal to the Radon-Nikodym derivative $d\mu / d\nu(\overleftarrow{x})$

at that point [156]. This sort of very flexible limit is denoted by

$$\lim_{\substack{V \in \mathcal{D} \\ V \ni \overleftarrow{x}}} \frac{\mu(V)}{\nu(V)} = \frac{d\mu}{d\nu}(\overleftarrow{x})$$

The Vitali property has strong and weak forms, but we will be able to prove the strong form. The differentiation basis $\mathcal{D}$ has the strong Vitali property with respect to μ if for every measurable set A and for every a sub-differentiation basis $\mathcal{D}' \subseteq \mathcal{D}$ covering A, there is an *at most countable* subset $\{ D_j \} \subseteq \mathcal{D}'$ such that $D_j \cap D_{j'}$ is empty for all $j \neq j'$ and

$$\mu\left(A - \left(\bigcup_j D_j\right)\right) = 0 \tag{2.13}$$

In other words, we must be able to cover "almost all" of A with a countable number of nonoverlapping sets from the differentiation basis [156].

We now demonstrate that the differentiation basis $\mathcal{D}$ generated by cylinder sets on $\mathcal{X}^{\mathbb{N}}$ has the Vitali property for *any measure* μ.

Proposition 2 (Vitali property for stochastic processes.). *For any stochastic process* $(\mathcal{X}^{\mathbb{N}}, \Sigma, \mu)$*, let* $\mathcal{D}$ *be the differentiation basis of allowed cylinder sets. Then* $\mathcal{D}$ *has the strong Vitali property.*

Proof. Let $\mathcal{D}' \subseteq \mathcal{D}$ be any sub-differentiation basis covering $\mathcal{X}^{\mathbb{N}}$. (Our proof trivially generalizes to any $A \subseteq \mathcal{X}^{\mathbb{N}}$.) Because $\mathcal{D}'$ is a differentiation basis, for all $\overleftarrow{x} \in \mathcal{X}^{\mathbb{N}}$ there must be a sequence $(D_j(\overleftarrow{x}))$ of cylinder sets converging on $\overleftarrow{X}$. We can without loss of generality suppose that $D_j(\overleftarrow{x}) = U_{-\ell_j, x_{-\ell_j+1} \ldots x_0}$ with ℓ_j monotonically increasing. (If this is not the case, we take a subsequence of $D_j(\overleftarrow{x})$ for which it is the case.)

Now consider the combination of all such sequences:

$$\mathcal{D}'' := \bigcup_{\overleftarrow{x} \in \mathcal{X}^{\mathbb{N}}} \left\{ D_j(\overleftarrow{X}) \,\middle|\, j \in \mathbb{N} \right\} .$$

We note that $\mathcal{D}''$, though it is a union of an uncountable number of sets, can itself cannot be larger than a countable set, as the elements of the sets from which it is composed are characterized by finite words, and finite words themselves only form a countable set. That is, there is significant redundancy in $\mathcal{D}''$ which keeps it countable. Furthermore, $\mathcal{D}''$ has a lattice structure given by the

set inclusion relation $\subseteq$ with the particular property that for $U, V \in \mathcal{D}''$, $U \cap V$ is nonempty only if $U \subseteq V$ or vice-versa.

We then choose the set $\mathcal{C}$ of all maximal elements of this lattice: that is, those $U \in \mathcal{D}''$ such that there is no $V \in \mathcal{D}''$ containing U. These maximal elements must exist since for each $U \in \mathcal{D}''$ there is only a finite number of sets in $\mathcal{D}''$ that can contain it.

It must be the case that all sets in $\mathcal{C}$ are nonoverlapping. Furthermore, for any $V \in \mathcal{D}''$, not in $\mathcal{C}$, there can only be a finite number of such sets containing V. One of them must be maximal and therefore in $\mathcal{C}$. In particular, for every $\overleftarrow{x} \in \mathcal{X}^{\mathbb{N}}$, each of its neighborhoods in $\mathcal{D}''$ is contained by the union of $\mathcal{C}$.

This implies $\mathcal{C}$ is a complete covering of $\mathcal{X}^{\mathbb{N}}$. Since it is also nonoverlapping and countable, the strong Vitali property is proven. $\square$

As a consequence, the likelihood ratios in Eq. (2.12) must converge for $\overleftarrow{\mu}$-almost every past $\overleftarrow{X}$ and every finite-length word w—proving Theorem 3.

We note that this result follows as a relatively straightforward application of the Vitali property, which holds for any measure μ on $\mathcal{X}^{\mathbb{Z}}$ and $\mathcal{X}^{\mathbb{N}}$. Our good fortune is due to the particularly well-behaved topology of sequences of discrete observations. For continuous observations, a less direct path to predictive states must be taken.

2.4.2. The space of predictive states. This section has defined predictive states in terms of Radon-Nikodym derivatives, and then proven that thay may be computed using the method of likelihood ratios. Providing concete examples of these ideas may be useful to the reader.

It will also help to lay down some new terminology surrounding predictive states. Let the (closure of the) set of a process's predictive states be denoted by:

$$\mathcal{K}(\mu) := \overline{\left\{ \epsilon[\overleftarrow{X}] \;\middle|\; \overleftarrow{X} \in \mathcal{X}^{\mathbb{N}} \right\}} .$$

The closure is taken under convergence in distribution.

The relation between pasts and predictive states may be highly redundant. For instance, in the process generated by the results of a random coin-toss, since the future observations do not depend on past observations, $\mathcal{K}(\mu)$ is trivial. Meanwhile, for a periodic process of period k, $\mathcal{K}(\mu)$ has k

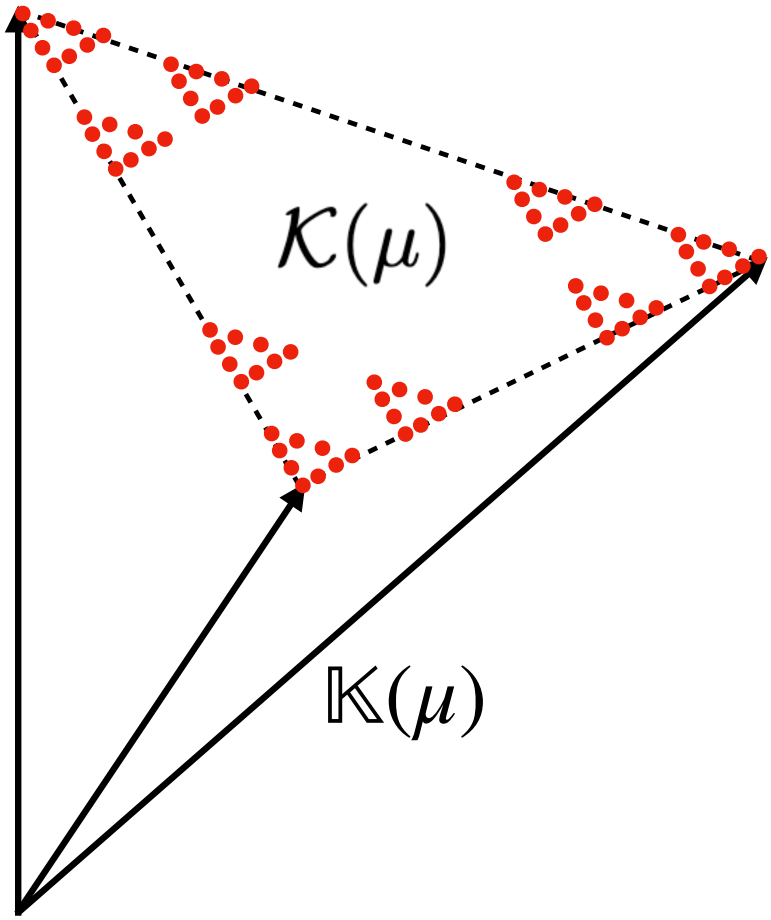

FIGURE 2.3. *Predictive states and their closed span.* The red dots are a hypothetical set of predictive states (shaped like the Sierpinski set), $\mathcal{K}(\mu)$, which is uncountably infinite but which has a finite-dimensional closed span $\mathbb{K}(\mu)$.

elements, corresponding to the k distinct states—the process' phases. In more complex cases, $\mathcal{K}(\mu)$ may have countable and uncountable cardinality.

We may also consider the vector space of signed measures generated by the closed span of $\mathcal{K}(\mu)$, denoted $\mathbb{K}(\mu)$. This is the smallest closed vector space which contains the predictive states. This, too, may demonstrate redundancy in the form of linear dependence, regardless of the cardinality of $\mathcal{K}(\mu)$. For instance, it is possible to have an uncountably infinite set of predictive states $\mathcal{K}(\mu)$

whose dimension, $\dim \mathbb{K}(\mu)$, is finite. In fact it is the general case that any process generated by an HMM or GHMM will have finite dimensional $\mathbb{K}(\mu)$, but is not guaranteed to have finite $\mathcal{K}(\mu)$ [89]. See Fig. 2.3.

We will start with some specific examples where, given a process, we construct the set of predictive states and, where applicable, also directly compute their Radon-Nikodym derivatives with respect to the stochastic process measure.

First we will consider the case of the Markov process:

EXAMPLE 1 (Markov process). *Consider a Markov process process where* $\Pr_\mu(x_1 \ldots x_\ell) = \pi_{x_1} P_{x_2|x_1} \ldots P_{x_\ell|x_{\ell-1}}$ *for some stochastic map* $\mathbf{P}$ *from* $\mathcal{X}$ *to itself and stationary distribution* $\boldsymbol{\pi}$ *on* $\mathcal{X}$*. Because the future only depends on the most recent symbol, the set of causal states is given by* $\mathcal{K}(\mu) = \{ \epsilon_x \mid x \in \mathcal{X} \}$*, where* $\epsilon_x(U_{1,w}) = \Pr_\mu(w \mid x)$*.*

For each $x \in \mathcal{X}$*, the predictive state's Radon-Nikodym derivative with* μ *is given by*

$$\frac{d\epsilon_x}{d\mu}(\overrightarrow{x}) = \frac{P_{x_1|x}}{\pi_{x_1}}$$

To summarize, a Markov process has a finite number of predictive states, each associated with a possible observation, and their Radon-Nikodym derivative with the process measure is just the ratio of the probability of the next symbol with the stationary probability of that symbol.

Markovian processes, as one would probably guess, have fairly simple causal structure. It is perhaps less obvious that many non-Markovian processes have equivalently simple causal structures.

EXAMPLE 2 (Even process). *Consider the Even process, which is generated by a hidden Markov model. A hidden Markov model (HMM) is a hidden state space* $\mathcal{S} = \{\mathrm{A}, \mathrm{B}\}$ *and a set of symbol-labeled transition matrices* $\left\{ \mathbf{T}^{(x)} \;\middle|\; x \in \mathcal{X} \right\}$*, where each matrix* $\mathbf{T}^{(x)} = (T^{(x)}_{rs})$ *has its rows and columns indexed by* $r, s \in \mathcal{S}$*. The generated process is defined by*

$$\Pr_\mu(x_1 \ldots x_\ell) = \sum_{s_1 \ldots s_{\ell+1}} T^{(x_\ell)}_{s_{\ell+1} s_\ell} \ldots T^{(x_1)}_{s_2 s_1} \pi(s_1)$$

where $\pi(s)$ *is the stationary distribution* $\pi(r) = \sum_{x,s} T^{(x)}_{rs} \pi(s)$. *For the Even process, we have* $\mathcal{S} = \{\mathrm{A}, \mathrm{B}\}$, $\mathcal{X} = \{0, 1\}$ *and matrices*

$$\mathbf{T}^{(0)} = \begin{pmatrix} 1/2 & 0 \\ 0 & 0 \end{pmatrix}, \ \mathbf{T}^{(1)} = \begin{pmatrix} 0 & 1 \\ 1/2 & 0 \end{pmatrix}$$

with $\pi(A) = 2/3$, $\pi(B) = 1/3$. *The name of the process derives from the fact that it only assigns nonzero probability to sequences symbols where the* 1*'s appear in blocks of even length. It has the useful property that* 0 *is a synchronizing symbol, in that if* $x_t = 0$ *at any* t, *then it must be the case that* $s_t = s_{t+1} = \mathrm{A}$.

Based on this property, it can be computed that

$$\Pr_\mu (w \mid \overleftarrow{x}) = \begin{cases} \Pr_\mathrm{A} (w) & \overleftarrow{x} \in \overleftarrow{\mathcal{X}}_0 \\ \Pr_\mathrm{B} (w) & \overleftarrow{x} \in \overleftarrow{\mathcal{X}}_1 \\ \text{undefined} & \overleftarrow{x} = \overleftarrow{1} \end{cases}$$

where $\overleftarrow{\mathcal{X}}_0$ *is the set of all pasts where the smallest* ℓ *such that* $x_{-\ell} = 0$ *is even,* $\overleftarrow{\mathcal{X}}_1$ *that where it is odd, and* $\overleftarrow{1}$ *is the past of all* 1*'s where a* 0 *never appears. Furthermore,*

$$\Pr_\mathrm{A} (x_1 \ldots x_\ell) := \sum_{s_1 \ldots s_{\ell+1}} T^{(x_\ell)}_{s_{\ell+1} s_\ell} \ldots T^{(x_1)}_{s_2 \mathrm{A}}$$

$$\Pr_\mathrm{B} (x_1 \ldots x_\ell) := \sum_{s_1 \ldots s_{\ell+1}} T^{(x_\ell)}_{s_{\ell+1} s_\ell} \ldots T^{(x_1)}_{s_2 \mathrm{B}}$$

Thus, we see that every past except $\overleftarrow{1}$ *has a well-defined causal prediction, and the causal state set has only two elements,* $\mathbb{K}(\mu) = \{ \epsilon_\mathrm{A}, \epsilon_\mathrm{B} \}$ *where* ϵ_A *and* ϵ_B *are given by the predictions* $\Pr_\mathrm{A} (\cdot)$ *and* $\Pr_\mathrm{B} (\cdot)$, *respectively. In other words, the predictive states are just isomorphic to the HMM states* $\mathcal{S}$. *It has been proven elsewhere that for any HMM which is* recurrent *(the full transition matrix is ergodic),* unifilar *(*$T^{(x)}_{ss'}$ *is positive for only one* s' *for fixed* s, x*), and* minimal *(each state has a unique prediction of the future), the predictive states of the produced process will be isomorphic to the HMM states [**197**].*

In addition, we can use the synchronizing property of 0 *to calculate*

$$\frac{d\epsilon_{\mathrm{A}}}{d\mu}\left(\overrightarrow{x}\right) = \begin{cases} 0 & \overrightarrow{x} \in \overrightarrow{\mathcal{X}}_0 \\ 3/2 & \overrightarrow{x} \in \overrightarrow{\mathcal{X}}_1 \\ \text{undefined} & \overrightarrow{x} = \overrightarrow{1} \end{cases}$$

$$\frac{d\epsilon_{\mathrm{B}}}{d\mu}\left(\overrightarrow{x}\right) = \begin{cases} 3 & \overrightarrow{x} \in \overrightarrow{\mathcal{X}}_0 \\ 0 & \overrightarrow{x} \in \overrightarrow{\mathcal{X}}_1 \\ \text{undefined} & \overrightarrow{x} = \overrightarrow{1} \end{cases}$$

where $\overrightarrow{\mathcal{X}}_0$ *is the set of all futures where the smallest* ℓ *such that* $x_\ell = 0$ *is even,* $\overrightarrow{\mathcal{X}}_1$ *that where it is odd, and* $\overrightarrow{1}$ *is the future of all* 1*'s where a* 0 *never appears.*

We'll now turn our attention to two examples of somewhat more complicated processes. The first is a type of renewal process; our example will first consider the structure of the causal states for general renewal processes, and then focus on the specific structure of a dual Poisson process.

Example 3 (Renewal process). *A renewal process is characterized by output strings with individual 1's (called* events*) with a certain number of 0's between them. The probability of a contiguous block of 0's, bounded on either side by 1's, having length at least* n *is given by the survival probability* $\Phi(n)$*. This quantity starts at* $\Phi(0) = 1$ *and monotonically decreases such that* $\lim_{n\to\infty} \Phi(n) = 0$*. It is assumed that the length of each block is independent of other blocks, so that the "closure" of a block with the symbol 1 resets the process.*

The probability that a contiguous block of 0's has actual length k *is given by* $F(k) := \Phi(k) - \Phi(k+1)$*. The mean inter-event length is given by*

$$m := \sum_{k=0}^{\infty} kF(k) = \sum_{k=1}^{\infty} \Phi(k)$$

and is assumed to be finite.

We will partition pasts and futures into the sets

$$\overleftarrow{\mathcal{X}}_k = \left\{ \overleftarrow{x} \;\middle|\; x_{-k} \dots x_0 = 10^k \right\}$$
$$\overrightarrow{\mathcal{X}}_k = \left\{ \overrightarrow{x} \;\middle|\; x_{-k} \dots x_0 = 0^k 1 \right\}$$

where $10^k = 10 \dots 0$ *is a* 1 *followed by* k *0's, and similarly for* $0^k 1$*. These sets include every past except* $\overleftarrow{0}$ *and every future except* $\overrightarrow{0}$*.*

Each value of $k \in \mathbb{N}$ *generally represents a distinct predictive state* $\epsilon_k := \epsilon_{\overleftarrow{x}}$ *for* $\overleftarrow{x} \in \overleftarrow{X}_k$*. We can evaluate the predictive states in the Radon-Nikodym form as*

$$\frac{d\epsilon_k}{d\mu}(\overrightarrow{x}) = \begin{cases} \frac{(m+1)F(k+\ell)}{\Phi(k)\Phi(\ell)} & \overrightarrow{x} \in \overrightarrow{\mathcal{X}}_\ell \\ \lim_{\ell\to\infty} \frac{(m+1)F(k+\ell)}{\Phi(k)\Phi(\ell)} & \overrightarrow{x} = \overrightarrow{0}, \text{ limit exists} \\ \text{undefined} & \overrightarrow{x} = \overrightarrow{0}, \text{ otherwise} \end{cases}$$

We consider a specific case now: the dual Poisson process. This models a continuous process which is observed at discrete time intervals Δt*; after each event, a decay rate is chosen from the distinct set* $\{\gamma_\mathrm{A}, \gamma_\mathrm{B}\}$ *with probabilities* p_A*,* p_B *and the time until the next event is determined by a Poisson process with the chosen decay rate. We have*

$$\Phi(n) = p_\mathrm{A}\Gamma_\mathrm{A}^n + p_\mathrm{B}\Gamma_\mathrm{B}^n$$
$$F(n) = p_\mathrm{A}\Gamma_\mathrm{A}^n(1-\Gamma_\mathrm{A}) + p_\mathrm{B}\Gamma_\mathrm{B}^n(1-\Gamma_\mathrm{B})$$
$$m = \frac{p_\mathrm{A}\Gamma_\mathrm{A}}{1-\Gamma_\mathrm{A}} + \frac{p_\mathrm{B}\Gamma_\mathrm{B}}{1-\Gamma_\mathrm{B}}$$

where $\Gamma_s := \exp(-\gamma_s \Delta t)$*. If we assume WLOG that* $\gamma_\mathrm{A} < \gamma_\mathrm{B}$*, then*

$$\frac{d\epsilon_k}{d\mu} = \frac{p_\mathrm{A}\Gamma_\mathrm{A}^k}{p_\mathrm{A}\Gamma_\mathrm{A}^k + p_\mathrm{B}\Gamma_\mathrm{B}^k}\frac{d\eta_\mathrm{A}}{d\mu} + \frac{p_\mathrm{B}\Gamma_\mathrm{B}^k}{p_\mathrm{A}\Gamma_\mathrm{A}^k + p_\mathrm{B}\Gamma_\mathrm{B}^k}\frac{d\eta_\mathrm{B}}{d\mu}$$

where

$$\frac{d\eta_{\mathrm{A}}}{d\mu}(\overrightarrow{x}) = \begin{cases} \frac{(m+1)\Gamma_{\mathrm{A}}^{\ell}(1-\Gamma_{\mathrm{A}})}{\Phi(\ell)} & \overrightarrow{x} \in \overrightarrow{\mathcal{X}}_{\ell} \\ \frac{(m+1)(1-\Gamma_{\mathrm{A}})}{p_{\mathrm{A}}} & \overrightarrow{x} = \overrightarrow{0} \end{cases}$$

$$\frac{d\eta_{\mathrm{B}}}{d\mu}(\overrightarrow{x}) = \begin{cases} \frac{(m+1)\Gamma_{\mathrm{B}}^{\ell}(1-\Gamma_{\mathrm{B}})}{\Phi(\ell)} & \overrightarrow{x} \in \overrightarrow{\mathcal{X}}_{\ell} \\ 0 & \overrightarrow{x} = \overrightarrow{0} \end{cases}$$

We can think of the predictions η_{A} *and* η_{B} *as representing what we would expect if we were given the secret knowledge of which decay rate was chosen. That all our predictive states can be represented as a linear combination of just two predictions is significant; it tells us that* $\dim \mathbb{K}(\mu)$ *is finite. We can take the closure of the causal states: this just involves adding the state* ϵ_∞*, whose Radon-Nikodym form is given by*

$$\frac{d\epsilon_\infty}{d\mu} := \lim_{k\to\infty} \frac{d\epsilon_k}{d\mu} = \frac{d\eta_{\mathrm{A}}}{d\mu}$$

So we see that η_A *is in fact the causal state* ϵ_∞*. The interpretation of this is that the larger the block observed, the more asymptotically certain we are that the decay rate is in fact the slowest of the two.*

These facts also imply the existence of an HMM with only two states that generates the dual Poisson process. This HMM has states $\mathcal{S} := \{\mathrm{A}, \mathrm{B}\}$*, and is given by the transition matrices*

$$\mathbf{T}^{(0)} = \begin{pmatrix} \Gamma_{\mathrm{A}} & 0 \\ 0 & \Gamma_{\mathrm{B}} \end{pmatrix}$$

$$\mathbf{T}^{(1)} = \begin{pmatrix} (1-\Gamma_{\mathrm{A}})p_{\mathrm{A}} & (1-\Gamma_{\mathrm{B}})p_{\mathrm{A}} \\ (1-\Gamma_{\mathrm{A}})p_{\mathrm{B}} & (1-\Gamma_{\mathrm{B}})p_{\mathrm{B}} \end{pmatrix}$$

Each HMM state represents the separate Poisson processes being mixed, and each event (an observation of 1) results in a scrambling of these two states.

The dual Poisson process example is significant: it tells us that by observing the properties of the predictive states, we can in fact infer a simple hidden Markov model underlying the process—this in spite of the infinite number of distinct predictive states. Understanding the geometry of predictive states informs us about the potential models which can produce a process.

Our last example looks at a process generated by pushdown automaton, whose allowed language is context-free but not regular. This example marks a strict divergence from much of the previous literature on predictive states, which has focused on renewal processes and hidden Markov models.

EXAMPLE 4. *Consider the 0^n1^n process, which will generate a block of 0's—we will denote the survival probability that this block has length at least n as $\Phi(n)$, which is monotonically decreasing—and then generate a block of 1's of equal length, before starting over with another 0-block. Such a process can be generated by a stochastic pushdown automaton.*

To discuss the causal states, we will partition pasts and futures into the sets

$$\begin{aligned}
\overleftarrow{\mathcal{X}}_0 &= \left\{ \overleftarrow{x} \;\middle|\; x_{-2k} \ldots x_0 = 10^k 1^k,\ k > 0 \right\} \\
\overleftarrow{\mathcal{X}}_k &= \left\{ \overleftarrow{x} \;\middle|\; x_{-k} \ldots x_0 = 10^k,\ k > 0 \right\} \\
\overleftarrow{\mathcal{X}}_{-k} &= \left\{ \overleftarrow{x} \;\middle|\; x_{-2\ell+k} \ldots x_0 = 10^\ell 1^{\ell-k},\ \ell > k \right\} \\
\overrightarrow{\mathcal{X}}_0 &= \left\{ \overrightarrow{x} \;\middle|\; x_1 \ldots x_{2k+1} = 0^k 1^k 0,\ k > 0 \right\} \\
\overrightarrow{\mathcal{X}}_k &= \left\{ \overrightarrow{x} \;\middle|\; x_1 \ldots x_{k+1} = 1^k 0,\ k > 0 \right\} \\
\overrightarrow{\mathcal{X}}_{-k} &= \left\{ \overrightarrow{x} \;\middle|\; x_1 \ldots x_{2\ell-k+1} = 0^{\ell-k} 1^\ell 0,\ \ell > k \right\}
\end{aligned}$$

for $k > 0$. These sets include every past and future except $\overleftarrow{0}$, $\overleftarrow{1}$, $\overrightarrow{0}$ and $\overrightarrow{1}$.

Using these we can define the causal states ϵ_0, ϵ_k and ϵ_{-k} for $k > 0$, where each causal state corresponds to the similarly labeled partition of pasts. Using temporal symmetry and the fact that

$F(k) := \Phi(k) - \Phi(k+1)$, *we can evaluate the causal states in Radon-Nikodym form as*

$$\frac{d\epsilon_0}{d\mu}(\overrightarrow{x}) = \begin{cases} 2m & \overrightarrow{x} \in \overrightarrow{\mathcal{X}}_0 \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{d\epsilon_k}{d\mu}(\overrightarrow{x}) = \begin{cases} \frac{2m}{\Phi(k)} & \overrightarrow{x} \in \overrightarrow{\mathcal{X}}_{-k} \\ \frac{2mF(k)}{\Phi(k)^2} & \overrightarrow{x} \in \overrightarrow{\mathcal{X}}_k \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{d\epsilon_{-k}}{d\mu}(\overrightarrow{x}) = \begin{cases} \frac{2m}{\Phi(k)} & \overrightarrow{x} \in \overrightarrow{\mathcal{X}}_k \\ 0 & \text{otherwise} \end{cases}$$

2.5. Jessen-Enomoto theory: continuous conditioning

2.3.4 In the last section we introduced the concept of a differentiation basis and the Vitali property, and explained how these concepts provide us with the necessary foundation to define predictive states based on a computable limit.

In Section 2.2 we discussed why the extension of the Vitali property to $\mathcal{X}^{\mathbb{N}}$ is far less straightforward in the case where $\mathcal{X} \subset \mathbb{R}$. To summarize, there is not really a useful form of the Vitali property in this domain. Instead, we must rely on an alternative approach to computing Radon-Nikodym derivatives, which relies on a series of theorems derived by Jessen [**79**] and Enomoto [**53**]. The capstone of this body of work is Enomoto's theorem, provided a differentiation basis for $(S^1)^{\mathbb{N}}$ which is both practical and useful.

In sections 2.5.1 and 2.5.2, we will revisit the work of Jessen and Enomoto, proving a generalized version of Enomoto's theorem:

THEOREM 4 (Generalized Enomoto's Theorem). *Let $\mathcal{X}$ be an interval of $\mathbb{R}$, and let μ be any probability measure over $\mathcal{X}^{\mathbb{N}}$. Let $f : \mathcal{X}^{\mathbb{N}} \to \mathbb{R}^+$ and let F be its indefinite integral under μ. Let $\mathcal{V}$ denote the differentiation basis consisting of sets of the form*

$$V_{n,\delta}(\overleftarrow{x}) = \{\ \overleftarrow{y}\ \mid |y_j - x_j| < \delta,\ j = 1, \ldots, n\ \}\ .$$

Then:

$$\lim_{\substack{V \in \mathcal{V} \\ V \ni \overleftarrow{x}}} \frac{F(V)}{\mu(V)} = f(\overleftarrow{x}) \tag{2.14}$$

for $\overleftarrow{\mu}$-almost all $\overleftarrow{x}$.

Note that the differentiation basis is very similar to the basis of cylinder sets, but is more restrictive. Each $V_{n,\delta}$ is evidently a cylinder set, but of a very particular kind. As we take $\delta \to 0$ and $n \to \infty$, we extend the "window" of the cylinder set to the entire past while simultaneously narrowing its width *uniformly*. This turns out to be sufficient to replicate the same effect as the fixed-aspect boxes in the finite-dimensional case.

As a direct corollary of Theorem 4, we will have the following result for predictive states:

COROLLARY 1. *For all measures μ on $\mathcal{X}^{\mathbb{Z}}$, where $\mathcal{X} \subset \mathbb{R}^d$ is a compact set, all neighborhoods $U_{0,I_1 \ldots I_\ell} \subset \mathcal{X}^\ell$, and all $\ell \in \mathbb{N}$, and for $\overleftarrow{\mu}$-almost all pasts $\overleftarrow{x} = \ldots x_{-1}x_0$, the limit:*

$$\Pr_\mu (I_1 \ldots I_\ell \mid \overleftarrow{x}) := \lim_{n \to \infty} \frac{\mu(V_{n,\delta(n)}(\overleftarrow{x}) \times U_{0,I_1 \ldots I_\ell})}{\mu(V_{n,\delta(n)}(\overleftarrow{x}))} \tag{2.15}$$

converges as long as $\delta(n) > 0$ for all n and $\delta(n) \to 0$.

Note here that we allowed $\mathcal{X} \subset \mathbb{R}^d$. This can be obtained from Enomoto's theorem by simply reorganizing a sequence of d-dimensional coordinates from $(\mathbf{x}_1, \mathbf{x}_2, \ldots)$ to $(x_{11}, \ldots, x_{d1}, x_{12}, \ldots, x_{d2}, \ldots)$. Enomoto's theorem then requires uniformity of the intervals across past instances as well as within each copy of $\mathbb{R}^d$.

To phrase this corollary in the η-notation of the previous section, we may define $\eta_{\ell,\delta}[\overleftarrow{x}]$ to correspond to the measure given by

$$\Pr_{\eta_{\ell,\delta}[\overleftarrow{x}]} (\cdot) := \frac{\mu(V_{n,\delta(n)}(\overleftarrow{x}) \times \cdot)}{\mu(V_{n,\delta(n)}(\overleftarrow{x}))}$$

Then Corollary 1 tells us that $\eta_{n,\delta(n)}[\overleftarrow{x}]$ converges to $\epsilon[\overleftarrow{x}]$ as $n \to \infty$. This is an immediate consequence of Thm. 4, but we get what we paid for: on its own, it is not as elegant as the result of for discrete processes. What we would like, and will find useful later, is a result on the convergence

of $\eta_\ell[\overleftarrow{x}]$, which would be defined directly as the conditional measures at each past length ℓ:

$$\Pr_{\eta_{\ell,\delta}[\overleftarrow{x}]}(\,\cdot\,) := \Pr_\mu(\,\cdot \mid x_{-\ell+1}\ldots x_0\,)$$

These conditional measures are themselves defined as Radon-Nikodym derivatives over the space $\mathcal{X}^\ell$. As discussed in Section 2.2, these can be computed via likelihood ratios of intervals with fixed-aspect ratios. In our case this means that $\eta_{\ell,\delta}[\overleftarrow{x}] \to \eta_\ell[\overleftarrow{x}]$ as $\delta \to 0$ with fixed ℓ and $\overleftarrow{x}$.

It would be useful and elegant if $\eta_\ell[\overleftarrow{x}] \to \epsilon[\overleftarrow{x}]$. As it happens, we can prove just this.

As before, the quantities $\Pr_\mu(\,U \mid \overleftarrow{x}\,)$ define a unique measure $\epsilon[\overleftarrow{x}]$ on $\mathcal{X}^{\mathbb{N}}$. It is determined by:

$$\epsilon[\overleftarrow{x}](U_{0,I_1\ldots I_\ell}) = \Pr_\mu(\,I_1\ldots I_\ell \mid \overleftarrow{x}\,)$$

for all $U_{0,I_1\ldots I_\ell}$.

THEOREM 5. *For all measures μ on $\mathcal{X}^{\mathbb{Z}}$, all $\ell \in \mathbb{N}$, all intervals $I_1 \times \cdots \times I_\ell \subset \mathcal{X}^\ell$, and $\overleftarrow{\mu}$-almost all pasts $\overleftarrow{x}$, where $\mathcal{X}$ is a finite set, the following limit is convergent:*

$$\Pr_\mu(\,w \mid \overleftarrow{x}\,) = \lim_{k\to\infty} \Pr_\mu(\,x_1\ldots x_\ell \mid x_{-k}\ldots x_0\,) \tag{2.16}$$

PROOF. *For each $\theta > 0$, let $\Delta(\ell,\theta)$ be chosen such that $\eta_{\ell,\Delta(\ell,\theta)}[\overleftarrow{x}]$ is θ-close to $\epsilon[\overleftarrow{x}]$, in the sense that*

$$\left|\Pr_{\eta_{\ell,\Delta(\ell,\theta)}[\overleftarrow{x}]}(\,w\,) - \Pr_{\epsilon[\overleftarrow{x}]}(\,w\,)\right| < \theta$$

Now let $\bar{\Delta}(L,\theta)$ be given by

$$\bar{\Delta}(L,\theta) = \min_{0\le\ell\le L} \Delta(\ell,\theta)$$

Now let $\delta(\ell,\zeta) = \ell^{-1}\bar{\Delta}(\ell,\theta)$. It is clear that $\delta(\ell,\zeta) \to 0$ as $\ell \to \infty$, so it must be case that $\eta_{\ell,\delta(\ell,\zeta)}[\overleftarrow{x}] \to \epsilon[\overleftarrow{x}]$ by Cor. 1. Trivially, then, it is also true that

$$\Pr_\mu(\,w \mid \overleftarrow{x}\,) = \lim_{\zeta\to 0}\lim_{\ell\to\infty} \Pr_{\eta_{\ell,\delta(\ell,\zeta)}[\overleftarrow{x}]}(\,w\,)$$

But it is also the case that $\Pr_{\eta_{\ell,\delta(\ell,\zeta)}[\overleftarrow{x}]}(\,w\,) \to \Pr_\mu(\,w \mid x_{-\ell+1\ldots x_0}\,)$ as $\zeta \to 0$. In fact, this limit is uniformly convergent in ℓ and $\overleftarrow{x}$, because ζ is defined to be the residual error, and so the error is ζ

regardless of all other values. Due to the uniform convergence, we can exchange limits:

$$\begin{aligned}\Pr_\mu(\, w \mid \overleftarrow{x}\,) &= \lim_{\ell\to\infty}\lim_{\zeta\to 0}\Pr_{\eta_{\ell,\delta(\ell,\zeta)}[\overleftarrow{x}]}(\, w\,)\\ &= \lim_{\ell\to\infty}\Pr_\mu(\, w \mid x_{-\ell+1\ldots x_0}\,)\end{aligned}$$

Thus proving the theorem.

This is more properly the continuous analogue of Thm. 3.

Both Corollary 1 and Theorem 5 are reliant on Enomoto's theorem. Enomoto's theorem itself is the capstone result in a sequence of theorems initiated by Jessen [**79**]. To prove Theorem 4, we must start from the beginning, generalizing Jessen's results. Fortunately, the bulk of the work to be done is in generalizing the first of these results—Jessen's correspondence principle. After this, the generalization follows quite trivially to the subsequent theorems. The next section provides the full proof for a generalized correspondence principle and explains how this result impacts the proofs of the subsequent theorems. For completeness, we also give the full proof of the generalized Enomoto's theorem, though it does not differ much from Enomoto's—published in French—once the preceding theorems are secured.

2.5.1. Jessen's correspondence principle. The Jessen and Enomoto theory rests on a profound correspondence between cylinder sets on $\mathcal{X}^{\mathbb{N}}$ and intervals on $\mathbb{R}$. To state it, we must define the concept of a net.

A net is similar to but formally separate from a differentiation basis, but like the latter allows for a notion of differentiation, called *differentiation-by-nets*. This is weaker than the Vitali property on a differentiation basis, but following on Jessen's work, Enomoto showed that differentiation-by-nets can be extended to describe a particular differentiation basis with the Vitali property.

Let $\mathcal{X}$ be a finite interval on $\mathbb{R}$. A *dissection* $D = (b_1, \ldots, b_N)$ of $\mathcal{X}$ is simply a sequence of cut points, that generate a sequence of adjacent intervals (b_k, b_{k+1}) spanning $\mathcal{X}$, covering all but a finite set of points (the edges of the intervals). See Fig. 2.4. Denote the intervals $\mathcal{I}(D) = \{\, (b_k, b_{k+1}) \mid k = 1, \ldots, N-1 \,\}$. The length of the largest interval in $\mathcal{I}(D)$ is denoted $|D|$. (Not to be confused with D's cardinality, that we have no need to reference.) A *net* $\mathcal{N} = (D_n)$ is a sequence

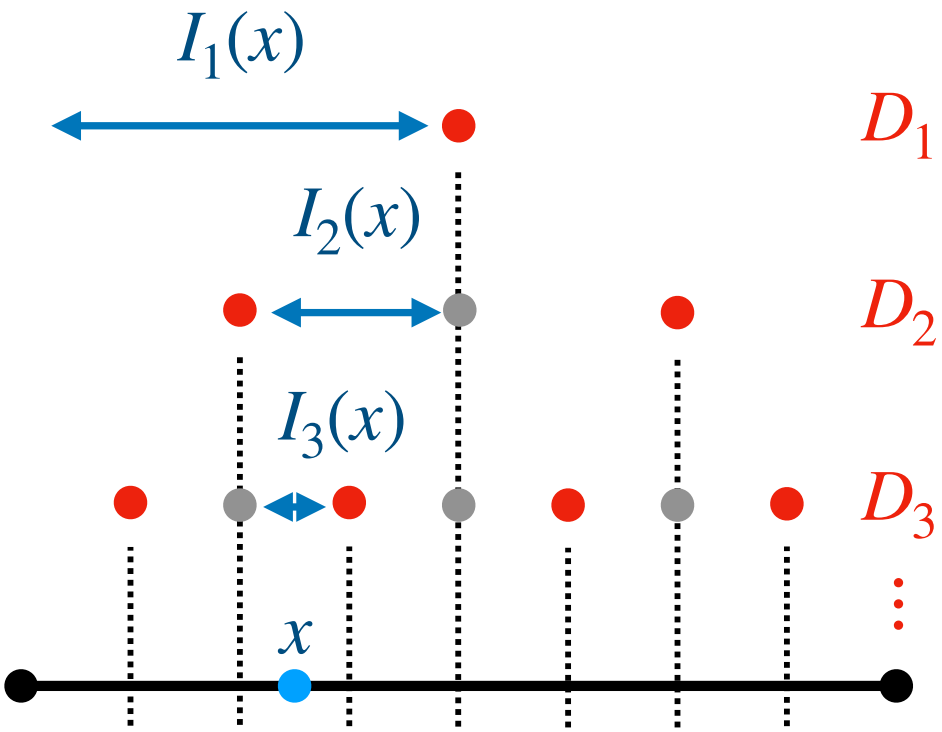

FIGURE 2.4. *Snapshot of a differentiation net.* A differentiation net defined on a line segment. $D_1, D_2, D_3, \ldots$ represents the dissections which comprise the net. Each dissection contains the last; new points are indicated in red and old points in grey. These points define intervals; a sequence of these intervals is shown, $(I_k(x))$, converging on the point x.

of dissections so that $D_n \subset D_{n+1}$ (that is, each new dissection only adds further cuts) and $|D_n| \to 0$ (the largest interval length goes to zero). The boundary $\partial\mathcal{N} = \bigcup_n D_n$ denotes all the boundary points from the sequence and is always a countable set.

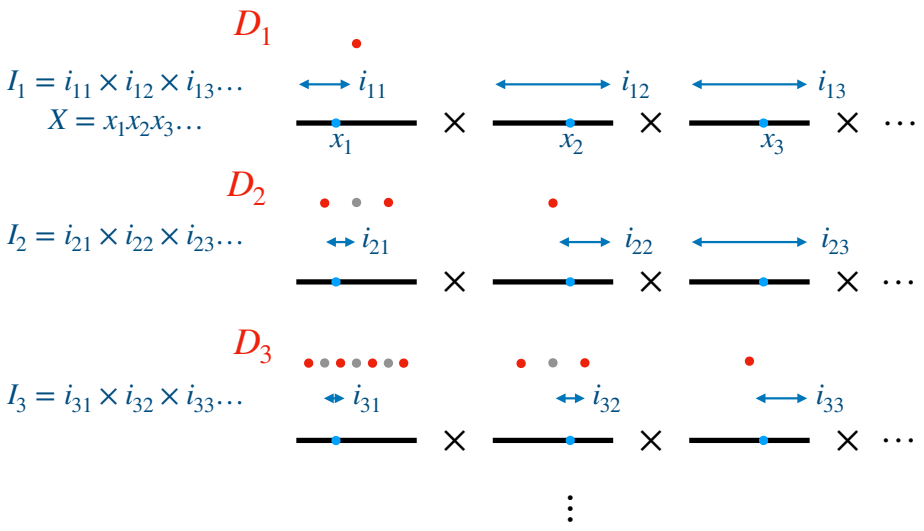

FIGURE 2.5. *Snapshot of a differentiation net on a product space.* A differentiation net defined on a product space $\mathcal{X}^{\mathbb{N}}$. This is comprised of an increasingly detailed dissection on each factor space. Also shown is a sequence of product intervals converging on a point $X = x_1 x_2 x_3 \ldots$.

We can similarly define a dissection $D = (d_1, \ldots, d_\ell)$ on $\mathcal{X}^{\mathbb{N}}$ as a set of ℓ dissections, one for each of the first ℓ copies of $\mathcal{X}$. D intervals $\mathcal{I}(D) = \left\{ i_1 \times \ldots i_\ell \times \mathcal{X}^{\mathbb{N}} \;\middle|\; i_k \in \mathcal{I}(d_k) \right\}$ are the cylinder sets generated by the intervals of each individual dissection. See Fig. 2.5. The boundary of a dissection is the set of all points that do not belong to these intervals: $\partial D = \left\{ \overrightarrow{x} \in \mathcal{X}^{\mathbb{N}} \;\middle|\; \exists k : x_k \in d_k \right\}$. The size of the dissection is $|D| := \max_k |d_k|$. For a finite measure μ, there are always dissections with $\mu(\partial D) = 0$ of any given $|D| = \max_k |d_k|$, because $\mu|_{\mathcal{X}^\ell}$ can only have at most countably many singular points. A net $\mathcal{N} = (D_n = (d_{1,n}, \ldots, d_{\ell_n,n}))$ of $\mathcal{X}^{\mathbb{N}}$ is a sequence of dissections of increasing depth ℓ_n so that each sequence $(d_{k,n})$ for fixed k is a net for the kth copy of $\mathcal{X}$. $\partial\mathcal{N} = \bigcup_n \partial D_n$ denotes all the accumulated boundary points of this sequence. Again, for finite measure μ, nets always exist that have $\mu(\partial\mathcal{N}) = 0$ for all n; nets with this property are called *μ-continuous nets*.

Note that for any net, every sequence of intervals (I_n), $I_n \in \mathcal{I}(D_n)$ and $I_{n+1} \subset I_n$, uniquely determines a point $\overrightarrow{x} \in \mathcal{X}^{\mathbb{N}}$. If $\overrightarrow{x} \notin \partial D$, then X uniquely determines a sequence of intervals.

The following result can be proven (generalized from Ref. [79]):

THEOREM 6 (Generalized correspondence principle). *Let $\mathcal{X} \subset \mathbb{R}$ be an interval and let λ be the Lebesgue measure on $\mathcal{X}$, normalized so $\lambda(\mathcal{X}) = 1$. Let μ be a finite measure on $\mathcal{X}^{\mathbb{N}}$ that has no singular points. Let $\mathcal{N} = (D_n)$ be any μ-continuous net of $\mathcal{X}^{\mathbb{N}}$. Then there exists a net $\mathcal{M} = (d_n)$ of $\mathcal{X}$ so that:*

(1) There exists a function Φ_n that maps each interval in $\mathcal{I}(D_n)$ of positive measure to one and only one interval in $\mathcal{I}(d_n)$, and vice-versa for Φ_n^{-1};

(2) $\lambda(\Phi_n(I)) = \mu(I)$ for all $I \in \mathcal{I}(D_n)$ with $\mu(I) > 0$; and

(3) The mapping $\phi : \mathcal{X}^{\mathbb{N}} - \partial\mathcal{N} \to \mathcal{X} - \partial\mathcal{M}$, generated by $\overrightarrow{x} \mapsto (I_n) \mapsto (\Phi_n(I_n)) \mapsto x$, is measure-preserving.

To summarize this technical statement: For any method of indefinitely dissecting the set $\mathcal{X}^{\mathbb{N}}$ into smaller and smaller intervals, there is in fact an "equivalent" such method for dissecting the much simpler set $\mathcal{X}$. It is equivalent in the sense that all the resulting intervals are in one-to-one correspondence with one another, a correspondence that preserves measure. Since interval sequences uniquely determine points (and vice-versa for a set of full measure), this induces a one-to-one correspondence between points that is also measure-preserving.

The proof consists of two parts. The first proves the first two claims about $\mathcal{M}$. Namely, there is an interval correspondence and it is measure-preserving. The second shows this extends to a correspondence between $\mathcal{X}^{\mathbb{N}}$ and $\mathcal{X}$ that is also measure-preserving.

PROOF (Interval correspondence). *The proof proceeds by induction. For a given μ-continuous net $\mathcal{N} = (D_n)$, suppose we already constructed dissections $d_1, \ldots, d_N$ of $\mathcal{X}$ so that a function Φ_n between positive-measure intervals in D_n and d_n exists with the desired properties (1) and (2), for all $n = 1, \ldots, N$. Now, for D_{n+1}, a certain set of the intervals in $\mathcal{I}(D_n)$ will be divided. Suppose $I \in \mathcal{I}_n$ is divided into I' and I''. If either of these, say I'', has measure zero then we discard it and set $\Phi_{n+1}(I') = \Phi_n(I)$. Otherwise, suppose that $\Phi_n(I) = (a, b)$. Then we will divide $\Phi_n(I)$ into the intervals:*

$$\Phi_{n+1}(I') := \left(a, \frac{a\mu(I) + (b-a)\mu(I')}{\mu(I)}\right)$$
$$\Phi_{n+1}(I'') := \left(\frac{a\mu(I) + (b-a)\mu(I')}{\mu(I)}, b\right) ,$$

which clearly have Lebesgue measures $\lambda(\Phi_{n+1}(I')) = \mu(I')$ and $\lambda(\Phi_{n+1}(I'')) = \mu(I'')$, respectively. Generalizing this to more complicated divisions of I is straightforward.

Now, we can always suppose for a given net $\mathcal{N}$ that D_0 is just the trivial dissection that makes no cuts and only one interval. However, this has a trivial correspondence with $\mathcal{X}$; namely, $\Phi_0(\mathcal{X}^{\mathbb{N}}) = \mathcal{X}$. By induction, then, the desired $\mathcal{M}$ can always be constructed.

With the existence of the interval correspondence established, we further demonstrate the existence of a point correspondence between μ-almost-all of $\mathcal{X}^{\mathbb{N}}$ and λ-almost-all of $\mathcal{X}$.

Proof (Point correspondence). *For every $\overrightarrow{x} \in \mathcal{X}^{\mathbb{N}} - \partial\mathcal{N}$, there is a unique sequence (I_n) of concentric intervals, $I_n \in \mathcal{I}(D_n)$ and $I_{n+1} \subset I_n$, such that $\bigcap_n I_n = \{\overrightarrow{x}\}$. If $\overrightarrow{x}$ is in the support of μ, then we define:*

$$\phi(\overrightarrow{x}) := \bigcap_n \Phi_n(I_n)$$

as the corresponding point in $\mathcal{X} - \partial\mathcal{M}$. Due to the interval correspondence, this mapping is invertible. By measure-preserving we mean that for all $A \subseteq \mathcal{X}^{\mathbb{N}} - \partial\mathcal{N}$, $\lambda(\phi(A)) = \mu(A)$ and vice-versa for ϕ^{-1}. Both the Lebesgue measure and μ must be outer regular, due to being finite measures. Outer regular means that the measure of a set A is the infimum of the measure of all open sets containing A, a property we use to our advantage.

Consider for each n the minimal covering $\mathcal{C}_n$ of A by intervals in $\mathcal{I}(D_n)$. The measure of this covering is denoted $m_n := \mu(\bigcup \mathcal{C}_n)$. Clearly, $m_n \geq \mu(A)$ and $m_n \to \mu(A)$. The corresponding covering $\Phi_n(\mathcal{C}_n)$ in $\mathcal{I}(d_n)$ is a covering of $\phi(A)$ and has the same measure m_n. By outer regularity, then, $m_n \geq \lambda(\phi(A))$ for all n. And so, $\mu(A) \geq \lambda(\phi(A))$.

Now, by the exact reverse argument of the previous paragraph, going from $\mathcal{X}$ to $\mathcal{X}^{\mathbb{N}}$ via ϕ^{-1}, we can also deduce that $\mu(A) \leq \lambda(\phi(A))$. Therefore $\mu(A) = \lambda(\phi(A))$, and the function ϕ is measure-preserving.

2.5.2. Corollaries and Enomoto's Theorem. Jessen's correspondence principle is an extremely powerful device. Among its consequences are the following theorems regarding functions on

$\mathcal{X}^{\mathbb{N}}$. We state their generalized forms here and for the proofs refer to Jessen [79], as each is a direct application of Theorem 6 without making any further assumptions on the measure μ.

The first offers a much weaker (and on its own, insufficient for our purposes) concept of differentiation of measures that we refer to as *differentiation-by-nets*.

Corollary 2 (Differentiation-by-nets). *Let $f : \mathcal{X}^{\mathbb{N}} \to \mathbb{R}^+$ and let F be the measure defined by its indefinite integral: $F(A) := \int_A f(\overrightarrow{x})d\mu(\overrightarrow{x})$. Further let $\mathcal{N} = (D_n)$ be a net on $\mathcal{X}^{\mathbb{N}}$ and denote by $\hat{f}_n$ a piecewise function such that $\hat{f}_n(\overrightarrow{x}) = F(I_n)/\mu(I_n)$ for all $X \in I_n$ and each $I_n \in D_n$. Then $\hat{f}_n(\overrightarrow{x}) \to f(\overrightarrow{x})$ as $n \to \infty$ for μ-almost all $\overrightarrow{x}$.*

Though the full proof is found in Ref. [79], we summarize the key point of the proof: Using the correspondence of intervals, we write $F(I_n)/\mu(I_n) = \tilde{F}(\Phi(I_n))/\lambda(\Phi(I_n))$, where $\tilde{F}$ is the indefinite integral of $f \circ \phi^{-1}$ with respect to λ. The limit then holds due to the Vitali property of λ on $\mathcal{X}$. However, we also note that Corollary 2 is *not* an extension of the Vitali property to cylinder sets on $\mathcal{X}^{\mathbb{N}}$. Jessen himself offers a counterexample to this effect in a later publication [80].

Jessen's second corollary is key to demonstrating that $\mathcal{V}$, the differentiation basis defined in Theorem 4, *will* have the Vitali property we are after.

Corollary 3 (Functions as limits of integrals). *Let $f : \mathcal{X}^{\mathbb{N}} \to \mathbb{R}^+$, and let $f_n(\overrightarrow{x})$ be a sequence of functions given by:*

$$f_n(x_1 x_2 \dots) := \int_{Y \in \mathcal{X}^{\mathbb{N}}} f(x_1 \dots x_n Y) d\mu(Y) \ .$$

That is, we integrated over all observations after the first n. Thus, f_n only depends on the first n observations. Then $f_n(\overrightarrow{x}) \to f(\overrightarrow{x})$ as $n \to \infty$ for μ-almost all $\overrightarrow{x}$.

This proof we also skip, again referring the reader to Jessen [79], as no step is directly dependent on the measure μ itself and only on properties already proven by the previous theorems.

We now have sufficient knowledge to prove the generalized Enomoto's theorem (generalized from Ref. [53]).

Proof (Generalized Enomoto's Theorem). *First, we must demonstrate, for almost every $\overrightarrow{x}$, that there exists a sequence $V_j(\overrightarrow{x})$ converging on $\overrightarrow{x}$ such that the limit holds. By Corollary 3, there*

must be, for μ-almost all $\overrightarrow{x}$ and any $\epsilon > 0$, a $K(\overrightarrow{x}, \epsilon)$ such that $|f_n(\overrightarrow{x}) - f(\overrightarrow{x})| < \epsilon/2$ for all $n > K(\overrightarrow{x}, \epsilon)$. Now, from the Vitali property on μ_n and the fact that f_n only depends on the first n observations, it must be true that for any $\epsilon > 0$ and almost all $\overrightarrow{x}$, there is a $0 < \Delta(\overrightarrow{x}, n, \epsilon) < 1$ so that:

$$\left| f_n(\overrightarrow{x}) - \frac{F(V_{n,\delta}(\overrightarrow{x}))}{\mu(V_{n,\delta}(\overrightarrow{x}))} \right| < \epsilon/2 \ ,$$

whenever $\delta < \Delta(\overrightarrow{x}, n, \epsilon)$. For a given ϵ, there is a countable number of conditions (one for each n). As such, the set of points $\overrightarrow{x}$ for which all conditions hold is still measure one. Then, taking for each $\overrightarrow{x}$ the integer $K := K(\overrightarrow{x}, \epsilon)$ and subsequently the number $\Delta := \Delta(\overrightarrow{x}, k(\overrightarrow{x}, \epsilon), \epsilon)$, we can choose $V_{K,\Delta}(\overrightarrow{x})$ and by the triangle inequality we must have:

$$\left| f(\overrightarrow{x}) - \frac{F(V_{K,\Delta}(\overrightarrow{x}))}{\mu(V_{K,\Delta}(\overrightarrow{x}))} \right| < \epsilon \ . \tag{2.17}$$

This completes the proof's first part.

However, the second part—that all *sequences $V_{n_j,\delta_j}(\overrightarrow{x})$ of neighborhoods give converging likelihood ratios—further follows from the above statements, as:*

$$\left| f(\overrightarrow{x}) - \frac{F(V_{n_j,\delta_j}(\overrightarrow{x}))}{\mu(V_{n_j,\delta_j}(\overrightarrow{x}))} \right| < \epsilon$$

must hold for any $n_j > K(\overrightarrow{x}, \epsilon)$ and any $\delta_j < \Delta(\overrightarrow{x}, K(\overrightarrow{x}, \epsilon), \epsilon)$, which must eventually be true for any converging sequence to $\overrightarrow{x}$.

Now, the previous theorem does not directly prove the Vitali property but rather bypasses it. Demonstrating that the differentiation basis $\mathcal{V}$ may be used to recover Radon-Nikodym derivatives. This, then, is sufficient for Corollary 5 to hold, guaranteeing the existence of predictive states $\epsilon[\overleftarrow{X}]$ for μ-almost all $\overleftarrow{X}$.

2.6. Discussion

In this chapter we have taken a *very* close look at predictive states. By this point the reader may be experiencing a mathematical form of semantic satiation: the phenomenon in which repeating a word many times causes it to temporarily lose meaning to the listener. What *is* a predictive state, even?

The resolute answer given by this chapter is that a predictive state is a conditional measure. It is conditional, in that it makes predictions of future behavior contingent on an infinite number of past observations; it is a measure, in that these predictions take the form of an entire family of probabilities assigned to finite future observations.

The infinities entailed in this conception of predictive states required us to ground ourselves in the formalism of mathematics. While this kind of mathematics can ground the truth or falsehood of our claims, it rarely grounds our understanding. It encourages us to pick at the details, to the point that even our own musings may take on unending depth. Let us therefore take the bird's eye view of what we have accomplished in this chapter.

(1) We gave the concept of a predictive state $\Pr_\mu (\cdot \mid \overleftarrow{x})$ a firm mathematical definition via conditional measures.

(2) We proved, using this definition and the mathematical properties of sequences, that

$$\Pr_\mu (\cdot \mid \overleftarrow{x}) = \lim_{k \to \infty} \Pr_\mu (\cdot \mid x_{-k} \ldots x_0)$$

where the right-hand side is the predictive state based on a finite amount of past information. That is, the predictive state is *stable*; the more we know about the past, the more stable it becomes, and this makes it well-suited for use in empirical data analysis. The convergence theorems we have proven will also tell us much about the dynamics of models which generate stochastic processes in Chapters 4 and 5.

(3) We showed that the stability of the predictive state is of a particular kind, which in Chapter 3 we will demonstrate has significant consequences for machine learning.

To the extent that these results look more straightforward in hindsight than the seemed in this chapter, it is because of how foundational they are to the core reasons for the utility of predictive states.

Taken altogether, the results fill-in important gaps in the foundations of predictive states, while strengthening those foundations for further development, extension, and application. Previously, predictive states were only examined in the context of hidden Markov models, their generalizations, and hidden semi-Markov models. We provided a definition applicable to any stationary and ergodic

process with discrete and real-valued observations. Further, our results indicate that predictive states for these processes are learnable from empirical data.

One important extension is to continuous-time processes. By exploiting the full generality of Jessen's and Enomoto's theorems we believe this extension is quite feasible. As long as the set of possible pasts and futures constitutes a separable space, they should be expressible in the form of a countable basis, to which these theorems may then be applied. The issue will lie in constructing an appropriate and useful basis. We leave this for future work.

CHAPTER 3

Model the noise: Inference with predictive states

Hollowed out, clay makes a pot.
Where the pot's not is where it's useful.

Lao Tzu, *Tao Te Ching*, transl. Ursula Le Guin

3.1. Introduction

As we have seen, predictive states can offer a characterization of the dynamics of a stochastic process. One aspect of this offer which we have not sufficiently harped on at this point is that this characterization is *model-free*; that is, we have assumed nothing and will assume nothing about the underlying system which produces the stochastic process. We have only assumed the minimal necessary assumptions on the qualitative nature of the data (that is, its temporal invariance via stationarity and ergodicity) that allow us to quantify the statistics of the process in the first place.

This sort of approach entails a shift in perspective for both the typical modeler and data analyst. The standard paradigm, though challenged on many fronts, remains to construct a model as a "signal" process, with intricate and detailed features, shrouded in uninteresting and unstructured "noise." The predictive state paradigm, however, requires a fundamental reversal in how data is perceived, in the same style as a Gestalt figure-ground inversion: *model the noise*. From this angle, data becomes immensely richer, providing us with information about the entire family of models which can produce the process, and the invariant qualities all these models share. Focusing on separating signal and noise is like focusing solely on the clay of the pot in this chapter's leading epigraph: to do so blinds us to all that the pot can contain.

When we start modeling the noise, we see that noise can itself have structure, with intricate dependencies on the present system state determining how deviations from the mean are generated. What we perceive as noise, or randomness, is frequently just the lack of all relevant information [42]. Where the useful information for predicting future behaviors is not available in the present, we must

avail ourselves of historical data, which may contain information that is not immediately at hand. Modeling the noise, then, if it is to be done properly, requires conditioning the present and future behavior of the data on as much of the past data as we can reliably utilize. This is precisely what the predictive state accomplishes.

But how can we utilize the predictive state? Because predictive states are characterized by a potentially *infinite* number of probabilities, they cannot be directly represented but must be represented by some kind of embedded vector. Several works have sought to reconstruct representations of predictive states using a formalism known as *reproducing kernel Hilbert spaces* (RKHS) [**20**, **31**, **181**, **182**]. This has been achieved to great effect for a specific reason, and to understand this, we must discuss what an RKHS is.

A kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ generates a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ if $k(\cdot, \cdot)$ is positive semi-definite and symmetric [**11**]. This is a shorthand for saying that, for any collection of points $\{x_i\} \subset \mathcal{X}$, the matrix $\mathbf{K} = (k(x_i, x_j))$ is positive semi-definite and symmetric. $\mathcal{H}$ is typically defined as a space of *functions* (from $\mathcal{X} \to \mathbb{R}$), which is generated as the span of all functions $\phi_x(y) = k(x, y)$ for each $x \in \mathcal{X}$, and has an inner product defined by $\langle \phi_x | f \rangle = f(x)$. This is called the *reproducing property*.

Each RKHS also allows the embedding of measures on $\mathcal{X}$ into the function space through $\mu \mapsto f_\mu$ where

$$f_\mu(x) = \int k(x, y) d\mu(y)$$

Incidentally, the resulting inner product over measures has a natural expression in terms of the original kernel:

$$\langle f_\mu | f_\nu \rangle_k := \int \int k(x, y) d\mu(x) d\nu(y) \ .$$

Thus, an RKHS can also be viewed as endowing the space of measures with an inner product, and therefore a geometry. The embedding of measures into this space is non-degenerate if the kernel is *characteristic*. Further—and importantly—convergence in the norm of the Hilbert space is equivalent to convergence in distribution whenever the kernel is *universal* [**183**].

This is the key reason that RKHS embeddings have proven to be successful in reproducing the predictive states from empirical data. In this chapter we prove that predictive states are convergent in distribution, and this is exactly the kind of convergence which is consistent with the geometry of RKHS embeddings.

The naturalness of the RKHS geometry for the specific kind of convergence demonstrated by predictive states is a novel insight provided by the work we present in this chapter. Crucially, understanding the source of the power of RKHS methods in predictive-state analysis frees us to consider other alternatives.

The content of the chapter is primarily drawn from the publication *Topology, Convergence, and Reconstruction of Predictive States* [**110**] and *Predictive State Geometry via Cantor Embeddings and Wasserstein Distance* [**112**]. In Section 3.2 we consider a novel approach to reconstructing the geometry of predictive states, inspired by the fractal Cantor set [**96**] and the Wasserstein distance [**145**]. The "Cantor embedding" will rely on an isomorphism between the Cantor set and the space of sequences, and the Wasserstein distance provides a geometry between measures which, like RKHS geometry, respects convergence in distribution. The resulting distance matrix is then used to find low-dimensional embeddings [**21**] of the geometry or hierarchical clusterings [**139**] of the predictive states. When combined with the fractal embedding, the latter, in particular, provides a highly interpretable visualization of the predictive-state space.

Following this, in Section 3.3 we revisit the previous literature on RKHS embeddings, and expound on how previous successes can be both explained and generalized by our measure-theoretic formalism for predictive states. In doing so we are able to add new terms to the traditional asymptotic convergence bounds.

It is in this chapter, primarily, that our mathematical efforts from Chapter 2 shall pay off; understanding the structure of predictive states will demonstrate why certain embedding strategies work far better than others and offer directions for new techniques. The predictive-state-as-embedding provides a quantitative and computationally concrete manifestation of the abstract predictive-state-as-measure.

3.2. Cantor-Wasserstein embeddings

In this section and the following (Sec. 3.3), we will turn our attention towards the problem of learning the geometry of the predictive states from empirical data.

As discussed in the previous section (Sec. 2.4), predictive states can be calculated as the convergent limit of likelihood ratios. However, this limit has a caveat: it is generally convergent in *distribution*, but further convergence cannot be assumed. Therefore, if we are to compute the geometry of predictive states, then the distances which define said geometry must replicate the topology of convergence in distribution. In the existing literature, the most popular method for predictive state embeddings has been via reproducing kernel Hilbert spaces (RKHS). In Sec. 3.3, we will examine the formal reasons for why these are appropriate, and use our results to provide more specific convergence bounds for predictive state embeddings. Before this, though, we will consider a more visually intuitive approach, which also utilizes the topology of convergence in distribution.

One method for accomplishing this can be inferred by analyzing the diagrams from Fig. [] back in Sec. 2.3. The Cantor fractals represent probability distributions: We interpret a vertical slice of the fractal, located at horizontal position $C(\overleftarrow{x})$, as visualizing the predictive state $P_{\overleftarrow{x}}$ as a distribution over Cantor-embedded futures $C(\overrightarrow{x})$.

For example, by examining the even process' Cantor fractal, one notices that there are effectively only 2 distinct predictive states—every vertical column is just one of two types. This corresponds with the 2 states of the hidden Markov model that generates the even process.

We can compare predictive states not only on how much their supports overlap, but on how geometrically close their supports are to one another. For the $\mathtt{a}^n\mathtt{b}^n$ process, for example, we see that the first few columns (corresponding to pasts of the form $\ldots\mathtt{ba}^n$ for some n) are inherently similar to one another, though they are shifted upwards the closer to the y-axis they are. (This corresponds to the increasing number of $\mathtt{b}$s in the predicted future as n increases.)

The intuitive distance metric between probability measures for capturing this "weight-shifting" geometry is the Wasserstein metric [**145**]. Given two measures μ and ν defined on a metric space

$\mathcal{M}$ with metric d, the Wasserstein distance between μ and ν is given by:

$$W(\mu, \nu) = \min_{\pi \in \Gamma(\mu,\nu)} \int_{\mathcal{M}\times\mathcal{M}} d(x, y) d\pi(x, y) \ ,$$

where $\Gamma(\mu, \nu)$ is the set of all measures on $\mathcal{M} \times \mathcal{M}$ whose left and right marginals are μ and ν, respectively. It is the minimal cost to "shift" the probability mass from one distribution to match shape of the other.

$W(\mu, \nu)$ is the solution to a constrained linear optimization. As a function of distributions, $W(\mu, \nu)$ is continuous with respect to convergence in distribution. In fact, convergence under the Wasserstein distance is equivalent to convergence in distribution on compact spaces [**145**]. This makes $W(\mu, \nu)$ ideal for measuring geometry between predictive states, since empirical estimates of these are known to converge in distribution.

When $\mathcal{M} \subseteq \mathbb{R}$, there is in fact a closed-form solution to the Wasserstein optimization problem [**190**]. Let F and G respectively be the cumulative distributions functions of μ and ν. Then:

$$W(\mu, \nu) = \int_{-\infty}^{\infty} |F(t) - G(t)| \, dt \ .$$

This closed-form solution is considerably faster to compute than the linear optimization required for arbitrary metric spaces. Since the Cantor embedding embeds the space of sequences directly into $[0, 1]$, we can directly employ this formula.

One can therefore achieve useful visualizations of predictive state geometry from using empirical data and calculating the Wasserstein distance between reconstructed predictive states. To elucidate the relative geometry of the predictive states, we can use the Wasserstein distance matrix to perform additional methods of geometric data analysis, such as hierarchical clustering [**139**] and multidimensional scaling [**21**]. This will be performed in the following sections.

3.2.1. Visualization with hierarchical clustering. Figure 3.1 displays the result of collecting the Cantor-embedded empirical predictions for all pasts of a given length for four processes—even, $\mathtt{a}^n\mathtt{b}^n$, $\mathtt{a}^n\mathtt{b}^n\mathtt{c}^n$, and $\mathtt{x} + \mathtt{f}(\mathtt{x})$. For each, the Wasserstein distance between every pair of predictions was computed and used to hierarchically cluster the pasts with others that produced similar predictions, using the Ward method [**139**].

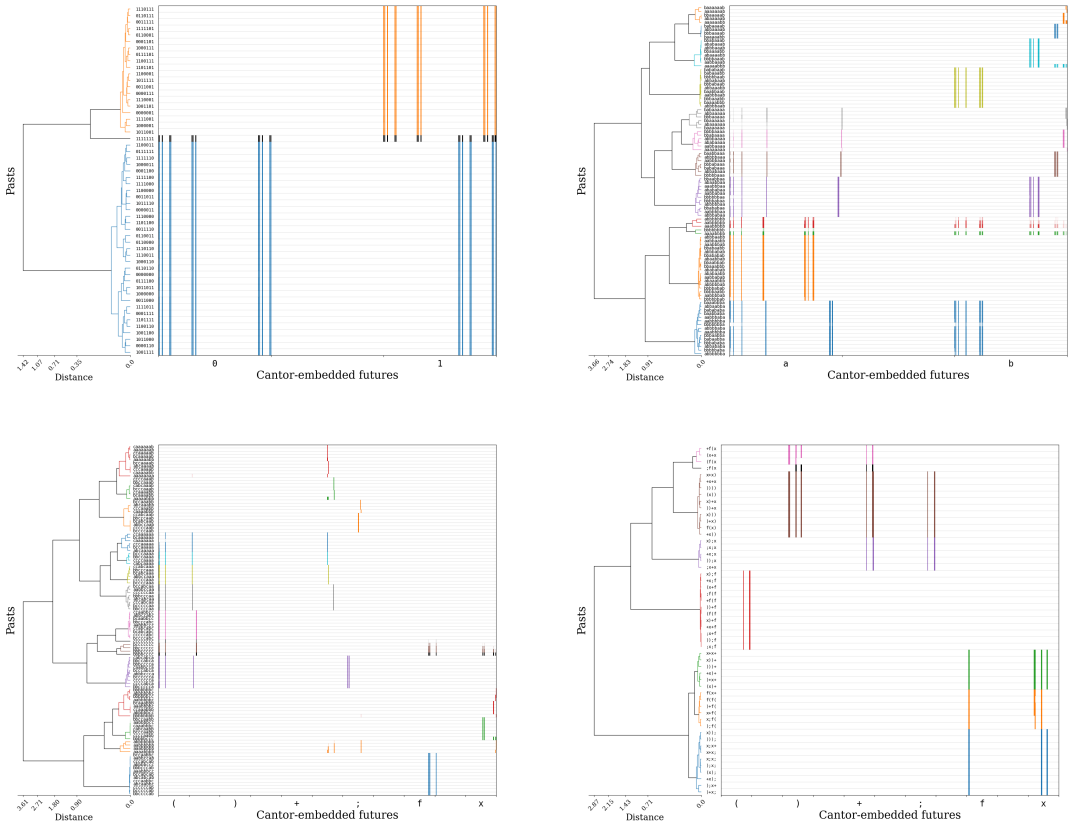

FIGURE 3.1. (Upper left to lower right) Clustered Cantor diagrams of the even, $\mathtt{a}^n\mathtt{b}^n$, $\mathtt{a}^n\mathtt{b}^n\mathtt{c}^n$, and $\mathtt{x} + \mathtt{f}(\mathtt{x})$ processes. Zoom for detail. For each, the vertical axis shows all pasts of a given length k along with their hierarchically clustered dendrogram. $k = 8$ for the even, $\mathtt{a}^n\mathtt{b}^n$, and $\mathtt{a}^n\mathtt{b}^n\mathtt{c}^n$ processes and $k = 4$ for the $\mathtt{x} + \mathtt{f}(\mathtt{x})$ process. For present purposes, the coloring threshold was chosen to aid visual interpretation. The lines in each row show the empirical distribution of Cantor-embedded futures observed following each past. As such, the horizontal axis corresponds exactly to the vertical axis of Fig. 2.1.

The resulting clustered Cantor plots offer a highly interpretable visualization of the relationship between pasts and futures, and of the predictive states' geometry. Each plot, in a certain sense, sorts the columns in the Cantor fractals of Fig. 2.1 with the white space between columns removed. For instance, the even process's clustered Cantor plot clearly contains the two major states, with a third "transient" state visible. (The latter corresponds to the increasingly unlikely event of never seeing a 0 in a block of length n.) This third state was previously hidden mostly out of view on the far-right side of the 2-dimensional Cantor plot of the even process in Fig. 2.1.

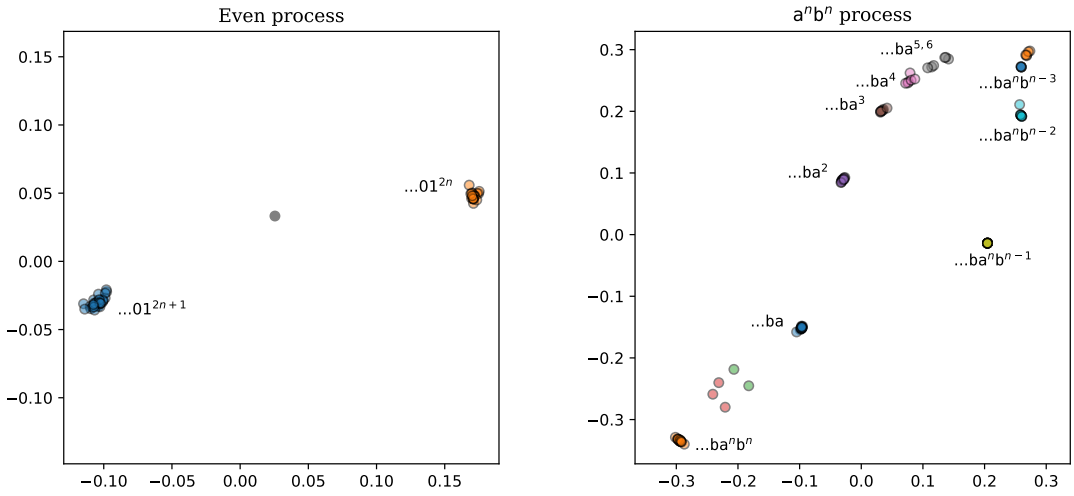

FIGURE 3.2. Scatterplots of the first two MDS coordinates of the reconstructed predictive states: (Left) Even process. (Right) $\mathtt{a}^n\mathtt{b}^n$ process. Clusters colored according to the scheme determined by the dendrogram in Fig. 3.1 and the label on each cluster describes the pattern that uniquely characterizes the pasts in that cluster.

Other features are worth calling out. Close observation shows that hierarchical clustering reveals the (mostly) scale-free distinctions between pasts with subtle differences. For the $\mathtt{a}^n\mathtt{b}^n$ process, pasts of the form $\ldots\mathtt{ba}^n$ are distinguished for different n, as each involves a distinct number of $\mathtt{b}$'s appearing in the near future. Meanwhile, the clustering scheme carefully distinguishes pasts of the form $\ldots\mathtt{ba}^n\mathtt{b}^{n-k}$ for different k but *not* for different n, as k is the essential variable for predicting the remaining number of $\mathtt{b}$'s. (The scale-free discernment of the algorithm breaks down past $n = 5$—the scale at which sampling error becomes relevant for our chosen sample size.)

Similar discernment is seen for the $\mathtt{a}^n\mathtt{b}^n\mathtt{c}^n$ and $\mathtt{x} + \mathtt{f}(\mathtt{x})$ processes as well. We draw attention to the manner in which the presence of a semicolon in pasts from $\mathtt{x} + \mathtt{f}(\mathtt{x})$ affects the comparison of predictions.

By analyzing clustered Cantor plots, one gains insight into the properties of pasts that make them similar in terms of future predictions, even if they are superficially quite distinct. Furthermore, the horizontal axis allows for continued use of the Cantor set's natural geometry for visualizing the future forecasts associated with each cluster of predictions.

3.2.2. Visualization with multidimensional scaling. Sacrificing direct visualization of future predictions leads to a more intuitive picture of predictive-state space geometry. Applying any desired dimension reduction algorithm to the matrix of Wasserstein distances between predictions yields a coordinate representation of the similarities between predictive states.

Figure 3.2 plots the first two dimensions of a multidimensional scaling (MDS) decomposition [**21**] for the even and $\mathtt{a}^n\mathtt{b}^n$ processes. Clusters are colored in the same manner as in Fig. 3.1 and labeled by the specific pattern that distinguishes the pasts in some of the clusters. Note that the clusters and labels are directly drawn from Fig. 3.1 for reference. They are not the result of the MDS algorithm itself. However, interactive plotting approaches may allow for similar exploration from these decompositions without the need for prior clustering.

The even process, as in all other cases seen thus far, has two dominant prediction clusters. These correspond to the predictive states that result from seeing an even-sized block of 1s (or, equivalently, no 1s), and that result from seeing an odd-sized block of 1s.

The $\mathtt{a}^n\mathtt{b}^n$ plot is much more sophisticated. Intriguingly, its geometry not only clearly distinguishes predictively distinct states, but organizes them in a manner highly suggestive of an *pushdown stack*. The latter is particularly appropriate given that stack automata are the natural analog of hidden Markov chains but for context-free languages. Observing more `a`s pushes more symbols onto the stack, with the predictive states moving further up towards the plot's upper-right corner. And, as more `b`'s are observed the top symbol is popped off the stack, and the predictive states move back towards the lower left. The latter represents equality between `a`s and `b`s.

The geometric approach is particularly insightful when computing the Wasserstein matrix between predictions estimated from Morse-Thue process data. Recall that the Morse-Thue process is a coarse-graining of the iterated logistic map $y_{t+1} = ry_t(1 - y_t)$ at the critical chaos parameter $r_c \approx 3.56995$. The resulting stream of 0s and 1s is a well-known instance of high complexity at the "order-disorder border". Specifically, setting parameter r on either side of r_c results in sequences that can be generated by finite hidden Markov chains. However, at r_c itself the resulting Morse-Thue process is context-sensitive and therefore requires infinite predictive states. That is, when it comes to capturing its behavior, the process is several orders higher in model complexity. It is further up the Chomsky language hierarchy.

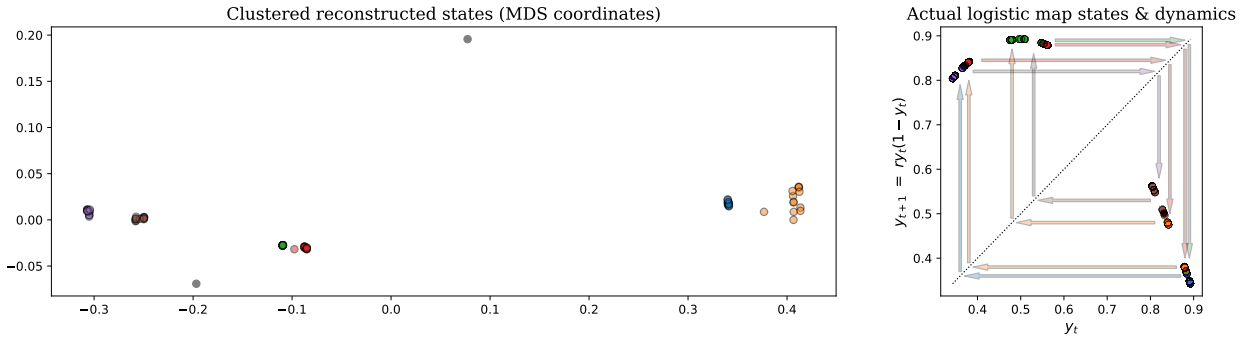

FIGURE 3.3. (Left) Scatterplot of the first two MDS coordinates of the reconstructed predictive states for the Morse-Thue process, color-coded by cluster. (Right) Scatterplot of the corresponding points in the domain of the logistic map, plotting for each point both the present value y_t and the next value y_{t+1}, with the $x = y$ line for reference. Each pair of color-coded arrows shows where each cluster maps to under the action of the logistic map. The predictively reconstructed clusters thus correspond to dynamically similar neighborhoods of the logistic map domain.

Despite this high order of structural complexity, the predictive state geometry reconstructed from a sufficiently large sample of the Morse-Thue process recovers the neighborhoods of $[0, 1]$ that are relevant to the dynamics of the original logistic map. Said differently, there is a correspondence between each past $x_{-k+1} \dots x_0$ and a subset $V_{x_{-k+1} \dots x_0}$, such that $V_{x_{-k+1} \dots x_0}$ is the set of all points y for which $x(f^{-t}(y)) = x_{-t}$ for $0 \leq t < n$. (Here, $f(y) = ry(1-y)$ and $x(y)$ is the encoding $y \mapsto \{0, 1\}$.) As it happens, pasts $x_{-k+1} \dots x_0$ whose predictive states are close under the Wasserstein distance are also pasts for which the sets $f(V_{x_{-k+1} \dots x_0})$ are close. That is, they correspond to predictively similar ranges of the logistic map variable.

Figure 3.3 directly visualizes the relationship between the reconstructed predictive states of the Morse-Thue process, neighborhoods of the logistic variable y, and the logistic map dynamics. In short, despite the fact that the Morse-Thue process is a highly coarse-grained form of the logistic map, the essential geometry of that map can be recovered by reconstructing predictive state geometry with the Wasserstein metric and the Cantor embedding.

Note that, due to the deterministic nature of the Morse-Thue process, the combination of the Wasserstein metric and the Cantor embedding is particularly important to achieving this result. Asymptotically, each past corresponds to a unique future. And so, there is asymptotically no overlap between predictions. The choice of the Cantor map facilitates placing together forecasts

that match up to a certain time in the future. And, the Wasserstein distance allows directly comparing predictions whose supports are geometrically close. In this way, the combination of the two approaches enables the straightforward recovery of the underlying dynamical system's (logistic map's) geometry.

3.3. Embedding predictive states in reproducing kernel Hilbert spaces

Thus far, we demonstrated that for discrete and real $\mathcal{X}$, measures over $\mathcal{X}^{\mathbb{N}}$ possess a well-defined feature called *predictive states* that relate how past observations constrain future possibilities. These states are defined by convergent limits that can be approximated from empirical time series in the case of stationary, ergodic processes.

We turn our attention now to the topological and geometric structure of these states, the spaces they live in, and how the structure of these spaces may be leveraged in the inference process. The results make contact between predictive states *as elements of a Hilbert space* and the well-developed arena of reproducing kernel Hilbert spaces. We discussed the basic concepts of reproducing kernel Hilbert spaces in Sec. 3.1; now, we will revisit these concepts by applying them to distributions over sequences. This section will address the Hilbert space embedding of measures over sequences in general. In Section 3.3.1 we will examine the truncation error which occurs in embedding when we only have a finite length of a sequence to work with. Then, in Section 3.3.2 we will take the bull by the horns, combining our results thus far with the theory of conditional RKHS embeddings to establish the convergence of RKHS representations of predictive states.

The space $\mathbb{K}(\mu)$ of predictive states is a subspace of $\mathbb{M}(\mathcal{X}^{\mathbb{N}})$, the space of measures over $\mathcal{X}^{\mathbb{N}}$. On $\mathbb{M}(\mathcal{X}^{\mathbb{N}})$, given any symmetric positive-definite kernel $k : \mathcal{X}^{\mathbb{N}} \times \mathcal{X}^{\mathbb{N}} \to \mathbb{R}$, we can define an inner product over measures:

$$\langle \mu, \nu \rangle_k := \int \int k(\overrightarrow{x}, \overrightarrow{y}) d\mu(\overrightarrow{x}) d\nu(\overrightarrow{y}) \ . \tag{3.1}$$

Positive-definite means that for any finite set $\{\overrightarrow{x}_i\}$ of $\overrightarrow{x}_i \in \mathcal{X}^{\mathbb{N}}$ and any set $\{c_i\}$ of values $c_i \in \mathbb{R}$, both sets having the same cardinality:

$$\sum_{i,j} k(\overrightarrow{x}_i, \overrightarrow{x}_j) c_i c_j \geq 0 \ ,$$

with equality only when $c_i = 0$ for all i. If this is true, then the inner product Eq. (3.1) is positive-definite for all *measures*. That is, $\langle \mu, \mu \rangle_k \geq 0$ with equality only when $\mu = 0$ [**183**].

Since $\mathcal{X}^{\mathbb{N}}$ is compact, if the kernel k satisfies a property of being *universal* then norm convergence under the inner product defined by k is equivalent to convergence in distribution of measures [**183**]. A simple example of a universal kernel is the Gaussian radial basis function, when paired with an appropriate distance—namely, one defined from embedding $\mathcal{X}^{\mathbb{N}}$ in a Hilbert space [**34**]. The reader will hopefully recall that the Euclidean metrics on sequences $D_{\mathrm{E},\gamma}$, which we introduced in section 2.3.3 are of precisely this class. The kernel takes the form:

$$k_{\beta,\gamma}(\overrightarrow{x}, \overrightarrow{y}) := \exp\left(-\frac{D_{\mathrm{E},\gamma}(\overrightarrow{x}, \overrightarrow{y})^2}{\beta^2}\right) .$$

We denote the associated inner products by $\langle \cdot, \cdot \rangle_{\beta,\gamma}$. $\mathcal{H}_{\beta,\gamma} := \left(\mathbb{M}(\mathcal{X}^{\mathbb{N}}), \langle \cdot, \cdot \rangle_{\beta,\gamma}\right)$ defines a Hilbert space, since it has the topology of convergence in distribution and $\mathbb{M}(\mathcal{X}^{\mathbb{N}})$ is complete in this topology.

When referring to a measure μ as an element of $\mathcal{H}_{\beta,\gamma}$ we denote it $|\mu\rangle_{\beta,\gamma}$ and inner products in the bra-ket are $\langle \mu | \nu \rangle_{\beta,\gamma}$. Now, it should be noted that to every ket $|\mu\rangle_{\beta,\gamma}$ there is a bra $\langle \mu |_{\beta,\gamma}$ that denotes a dual element. However, the dual elements of $\mathbb{M}(\mathcal{X}^{\mathbb{N}})$ correspond to continuous functions. The function f_μ corresponding to $\langle \mu |_{\beta,\gamma}$ is given by:

$$f_\mu(\overrightarrow{x}) := \int k_{\beta,\gamma}(\overrightarrow{x}, \overrightarrow{y}) d\mu(\overrightarrow{y}) , \tag{3.2}$$

so that:

$$\langle \mu | \nu \rangle_{\beta,\gamma} = \int f_\mu(\overrightarrow{x}) d\nu(\overrightarrow{x}) .$$

Let $\mathcal{F}_{\beta,\gamma}$ denote the space of all f_μ that can be constructed from Eq. (3.2). This function space, when paired with the inner product $\langle f_\mu, f_\nu \rangle := \langle \nu | \mu \rangle$, is isomorphic to $\mathcal{H}_{\beta,\gamma}$. $\mathcal{F}_{\beta,\gamma}$ is then a reproducing kernel Hilbert space with kernel $k_{\beta,\gamma}$.

It is necessarily the case that $\mathcal{F}_{\beta,\gamma}$ is a proper subset of the space of continuous functions. Furthermore, the $\mathcal{F}_{\beta,\gamma}$ are not identical to one another, obeying the relationship $\mathcal{F}_{\beta,\gamma} \subset \mathcal{F}_{\beta',\gamma}$ when $\beta > \beta'$ [**216**].

However, it is also the case that each $\mathcal{F}_{\beta,\gamma}$ is dense in the space of continuous functions, so their representative capacity is still quite strong [183].

We note an important rule regarding the scaling of our inner products, as constructed. The distances $D_{\mathrm{E},\gamma}(\overrightarrow{x}, \overrightarrow{y})$ have finite diameter on our spaces. Let Δ denote the diameter of $\mathcal{X}$. For discrete $\mathcal{X}$ we simply have $\Delta = 1$; for $\mathcal{X} \subset \mathbb{R}^d$ it is determined by the Euclidean distance. Then $\mathcal{X}^{\mathbb{N}}$'s diameter is given by $\Delta/\sqrt{1-\gamma^2}$. Since the Gaussian is bounded below by $1 - D_\gamma^2/\beta^2$, for arbitrarily large β:

$$\|\mu - \nu\|_{\beta,\gamma}^2 \leq \frac{\|\mu - \nu\|_{\mathrm{TV}} \Delta^2}{(1-\gamma^2)\beta^2} + O(\beta^{-3}) \ , \tag{3.3}$$

where $\|\cdot\|_{\beta,\gamma}$ is simply the norm of $\mathcal{H}_{\beta,\gamma}$ and $\|\cdot\|_{\mathrm{TV}}$ is the total variation norm. This tells us that the norm is less discriminating between measures as $\beta \to \infty$. Naturally, this can be remedied by rescaling the kernel with a β^2 factor. As it happens, Eq. (3.3) will be useful later.

3.3.1. Finite-length embeddings. Our goal is to study how reproducing kernel Hilbert spaces may be used to encode information about predictive states gleaned from empirical observations. Given that such observations are always finite in length, we must determine whether and in what manner the Hilbert space representations of measures over finite-length observations converges to the Hilbert space representation of a measure over infinite sequences. We call the residual the "truncation error."

In Section 2.3.3 we also defined the restricted distance $D_{\mathrm{E},\gamma}^{(\ell)}$, and showed a "Pythagorean theorem" relation of the form $D_{\mathrm{E},\gamma}(\overrightarrow{x}, \overrightarrow{y})^2 = D_{\mathrm{E},\gamma}^{(\ell)}(x_{1:\ell+1}, y_{1:\ell+1}) + \gamma^{2\ell} D_{\mathrm{E},\gamma}(x_{\ell+1:}, y_{\ell+1:})^2$. Now, using $D_{\mathrm{E},\gamma}^{(\ell)}$, define kernels $k_{\beta,\gamma}^{(\ell)}$ in the same style as for $\mathcal{X}^{\mathbb{N}}$. These generate inner products on $\mathbb{P}(\mathcal{X}^\ell)$. Denote by $\mathcal{H}_{\beta,\gamma}^{(\ell)}$ the resulting Hilbert spaces. These are related to the original $\mathcal{H}_{\beta,\gamma}$ by the following factorization theorem:

Proposition 3. *The predictive Hilbert space $\mathcal{H}_{\beta,\gamma}$ factors into $\mathcal{H}_{\beta,\gamma}^{(\ell)} \otimes \mathcal{H}_{\beta\gamma^{-\ell},\gamma}$.*

Before stating the proof, we should explain the above. The factorization $\mathcal{H}_{\beta,\gamma} = \mathcal{H}_{\beta,\gamma}^{(\ell)} \otimes \mathcal{H}_{\beta\gamma^{-\ell},\gamma}$ denotes a separation of the infinite-dimensional $\mathcal{H}_{\beta,\gamma}$ into two pieces—one of which is finite-dimensional, but retains the same kernel parameters, and another *reparameterized* infinite-dimensional Hilbert space. The reparameterization is $\beta \to \beta\gamma^{-\ell}$. This constitutes, essentially, a renormalization technique, in which the the topology of words starting at depth ℓ is equivalent to a reparameterization

of the usual topology. This reparameterization works precisely because of the Pythagorean theorem for sequences Eq. (2.8), and is a reflection of the self-similar geometry of sequences discussed in Section 2.3.

PROOF. *We are demonstrating an isomorphism—a particularly natural one. Let* $\delta_{\overrightarrow{x}}$ *be the Dirac delta measure concentrated on* $\overrightarrow{x}$*. We note that for any measure* μ*:*

$$|\mu\rangle_{\beta,\gamma} = \int |\delta_{\overrightarrow{x}}\rangle_{\beta,\gamma} \, d\mu(\overrightarrow{x}) \ .$$

Now, consider the linear function from $\mathcal{H}_{\beta,\gamma}$ *to* $\mathcal{H}^{(\ell)}_{\beta,\gamma} \otimes \mathcal{H}_{\beta\gamma^{-\ell},\gamma}$ *that maps:*

$$|\delta_{\overrightarrow{x}}\rangle_{\beta,\gamma} \mapsto |\delta_{x_1 \ldots x_\ell}\rangle^{(\ell)}_{\beta,\gamma} \otimes |\delta_{x_{\ell+1}\ldots}\rangle_{\beta\gamma^{-\ell},\gamma} \ ,$$

for every $\overrightarrow{x}$*. Then by Eq. (2.8) we can see that this preserves the inner product and so is an isomorphism.*

Note that for any of these Hilbert spaces there exists an element corresponding to the constant function $\mathbf{1}(\overrightarrow{x}) = 1$ for all $\overrightarrow{x}$. This function always exists in $\mathcal{F}_{\beta,\gamma}$. We denote its corresponding element in $\mathcal{H}_{\beta,\gamma}$ as $\langle \mathbf{1}|_{\beta,\gamma}$, so that $\langle \mathbf{1}|\mu\rangle_{\beta,\gamma} = 1$ for all μ. Then the operator $\Pi^{(\ell)}_{\beta,\gamma} : \mathcal{H}_{\beta,\gamma} \to \mathcal{H}^{(\ell)}_{\beta,\gamma}$ is given by:

$$\Pi^{(\ell)}_{\beta,\gamma} := I^{(\ell)} \otimes \langle \mathbf{1}|_{\beta,\gamma} \ ,$$

where $I^{(\ell)}$ is the identity on $\mathcal{H}^{(\ell)}_{\beta,\gamma}$. It provides the canonical mapping from a measure μ to its restriction μ_ℓ: That is, $\Pi^{(\ell)}_{\beta,\gamma} |\mu\rangle_{\beta,\gamma} = |\mu_\ell\rangle^{(\ell)}_{\beta,\gamma}$.

Now we will consider the "truncation error"—that is, the residual error remaining when representing a measure by its truncated form μ_ℓ rather than by its full form μ. We quantify this in terms of an embedding. That is, there exists an embedding of truncated measures $\mathbb{P}(\mathcal{X}^\ell)$ into the space of full measures $\mathbb{P}(\mathcal{X}^{\mathbb{N}})$ such that the distance between any full measure and its truncated embedding is small:

THEOREM 7. *There exist isometric embeddings* $\mathcal{H}^{(\ell)}_{\beta,\gamma} \mapsto \mathcal{H}^{(\ell')}_{\beta,\gamma}$ *and* $\mathcal{H}^{(\ell)}_{\beta,\gamma} \mapsto \mathcal{H}_{\beta,\gamma}$ *for any* $\ell \leq \ell'$*. Furthermore, let* μ *be any measure and* μ_ℓ *be its restriction to the first* ℓ *observations, and let* $|\hat{\mu}_\ell\rangle_{\beta,\gamma}$ *be the embedding of* μ_ℓ *into* $\mathcal{H}_{\beta,\gamma}$*. Then* $|\hat{\mu}_\ell\rangle_{\beta,\gamma} \to |\mu\rangle_{\beta,\gamma}$ *as* $\ell \to \infty$*, with* $\|\mu - \hat{\mu}_\ell\|_{\beta,\gamma} \sim O(\beta^{-1}\gamma^\ell)$*.*

PROOF. *Let $\lambda_{\beta,\gamma} \in \mathbb{P}(\mathcal{X}^{\mathbb{N}})$ denote the measure such that $\langle \lambda_{\beta,\gamma}|_{\beta,\gamma} = \langle \mathbf{1}|_{\beta,\gamma}$ for a given β, γ. For a measure μ with restriction μ_ℓ let $\hat{\mu}_\ell$ denote the measure on $\mathcal{X}^{\mathbb{N}}$ with the property:*

$$\hat{\mu}_\ell(A \times B) = \mu_\ell(A)\lambda_{\beta\gamma^{-\ell},\gamma}(B) \ ,$$

for $A \in \mathcal{X}^\ell$ and $B \in \mathcal{X}^{\mathbb{N}}$. Then the mapping $\mu_\ell \mapsto \hat{\mu}_\ell$ is an isomorphism, since:

$$\begin{aligned}
\langle \hat{\mu}_\ell | \hat{\nu}_\ell \rangle_{\beta,\gamma} &= \int\int k_{\beta,\gamma}(\overrightarrow{x}, \overrightarrow{y}) d\hat{\mu}_\ell(\overrightarrow{x}) d\hat{\nu}_\ell(\overrightarrow{y}) \\
&= \int\int k^{(\ell)}_{\beta,\gamma}(x_1 \dots x_\ell, y_1 \dots y_\ell) d\mu_\ell d\nu_\ell \ \times \int\int k_{\beta\gamma^{-\ell},\gamma}(x_\ell \dots, y_\ell \dots) d\lambda_{\beta\gamma^{-\ell},\gamma} d\lambda_{\beta\gamma^{-\ell},\gamma} \\
&= \langle \mu_\ell | \nu_\ell \rangle^{(\ell)}_{\beta,\gamma} \int d\lambda_{\beta\gamma^{-\ell},\gamma} = \langle \mu_\ell | \nu_\ell \rangle^{(\ell)}_{\beta,\gamma} \ .
\end{aligned}$$

Now, as a result of Eq. (2.8), note that for any two measures μ and ν:

$$\begin{aligned}
&\langle \mu, \nu \rangle = \\
&\int d\mu_\ell(x_1 \dots x_\ell) \int d\nu_\ell(y_1 \dots y_\ell) \exp\left(-\beta^2 D^{(\ell)}_\gamma(x_1 \dots x_\ell, y_1 \dots y_\ell)\right) \langle \mu(\cdot|x_1 \dots x_\ell), \nu(\cdot|y_1 \dots y_\ell) \rangle_{\beta\gamma^\ell,\gamma}
\end{aligned}$$

If we combine this fact with the bound Eq. (3.3), we have the result:

$$\begin{aligned}
&\|\mu - \hat{\mu}_\ell\|^2_{\beta,\gamma} = \\
&\quad \int d\mu_\ell(x_1 \dots x_\ell) \int d\mu_\ell(y_1 \dots y_\ell) e^{-\beta^2 D^{(\ell)}_\gamma(x_1 \dots x_\ell, y_1 \dots y_\ell)} \|\mu(\cdot|x_1 \dots x_\ell) - \hat{\mu}_\ell(\cdot|y_1 \dots y_\ell)\|^2_{\beta\gamma^{-\ell}} \\
&\quad \leq \int d\mu_\ell(x_1 \dots x_\ell) \int d\mu_\ell(y_1 \dots y_\ell) \frac{\|\mu - \hat{\mu}_\ell\|_{\mathrm{TV}} \Delta^2 \gamma^{2\ell}}{(1-\gamma^2)\beta^2} = \frac{\|\mu - \hat{\mu}_\ell\|_{\mathrm{TV}} \Delta^2 \gamma^{2\ell}}{(1-\gamma^2)\beta^2} \ .
\end{aligned}$$

Thus, $\|\mu - \hat{\mu}_\ell\|_{\beta,\gamma} \sim O(\beta^{-1}\gamma^\ell)$.

In summary, representing measures μ over $\mathcal{X}^{\mathbb{N}}$ by their truncated forms μ_ℓ leads to a Hilbert space representation that admits an approximate isomorphism to the space of full measures. The resulting truncation error is of order $O(\beta^{-1}\gamma^\ell)$.

We close this part with a minor note about a lower bound on the distance between measures. Given a word w, the function on $\mathcal{X}^\ell$ that equals 1 when $X = w$ and zero otherwise has a representation in $\mathcal{H}^{(\ell)}_{\beta,\gamma}$, $|w\rangle^{(\ell)}_{\beta,\gamma}$. (This follows since for finite $\mathcal{X}$, all functions on $\mathcal{X}^\ell$ belong to $\mathcal{F}^{(\ell)}_{\beta,\gamma}$.) The extension of this to $\mathcal{H}_{\beta,\gamma}$ is $|w\rangle_{\beta,\gamma} := |w\rangle^{(\ell)}_{\beta,\gamma} \otimes |\lambda_{\beta,\gamma}\rangle_{\beta\gamma^{-\ell},\gamma}$. This has the convenient property that $\langle w|\mu\rangle_{\beta,\gamma} = \Pr_\mu(\, w \,)$.

Then, by the Cauchy-Schwarz inequality, for any measures μ and ν and any word w:

$$\begin{aligned} \|\mu - \nu\|_{\beta,\gamma} &\geq \frac{|\langle w|\mu - \nu\rangle|}{\sqrt{\langle w|w\rangle_{\beta,\gamma}}} \\ &= \frac{|\Pr_\mu(w) - \Pr_\nu(w)|}{\sqrt{\langle w|w\rangle_{\beta,\gamma}}} \ . \end{aligned} \tag{3.4}$$

So, word probabilities function as lower bounds on the Hilbert space norm.

3.3.2. Predictive states from kernel conditional measures. A prominent use of reproducing kernel Hilbert spaces is to approximate empirical measures [138]. Given a measure μ over a space $\mathcal{X}$ and N samples X_k drawn from this space, one constructs an approximate representation of μ via:

$$|\hat{\mu}\rangle := \frac{1}{N}\sum_{k=1}^{N} |\delta_{X_k}\rangle \ .$$

In other words, μ is approximated as a sum of delta functions centered on the observations. Convergence of this approximation to $|\mu\rangle$ is (almost surely) $O(N^{-1/2})$ [138].

This fact, combined with our Theorem 7, immediately gives the following result for $\mathcal{H}_{\beta,\gamma}$:

Proposition 4. *Suppose for some $\mu \in \mathbb{P}(\mathcal{X}^{\mathbb{N}})$ we take N samples of length ℓ, denoted $\{X_k \in \mathcal{X}^\ell\}$ $(k = 1 \ldots N)$, and construct the state:*

$$|\hat{\mu}_{\ell,N}\rangle_{\beta,\gamma} = \frac{1}{N}\sum_{k=1}^{N} |\delta_{X_k}\rangle^{(\ell)}_{\beta,\ell} \otimes |\lambda_{\beta\gamma^{-\ell},\gamma}\rangle_{\beta\gamma^{-\ell},\gamma} \ .$$

Then $|\hat{\mu}_{\ell,N}\rangle_{\beta,\gamma} \to |\mu\rangle$ converges almost surely as $N, \ell \to \infty$ with error $O(N^{-1/2} + \beta^{-1}\gamma^{\ell})$.

A more nuanced application of RKHS for measures lies in reconstructing conditional distributions [20, 58, 138, 181, 182]. Let μ be a joint measure on some $\mathcal{X} \times \mathcal{Y}$, and let $\mu|_{\mathcal{X}}$ and $\mu|_{\mathcal{Y}}$ be its marginalizations. Given N samples (X_k, Y_k), construct the covariance operators:

$$\hat{C}_{XX} := \frac{1}{N}\sum_k |\delta_{X_k}\rangle\langle\delta_{X_k}| \text{ and}$$

$$\hat{C}_{YX} := \frac{1}{N}\sum_k |\delta_{Y_k}\rangle\langle\delta_{X_k}| \ .$$

Let $\mu_{\mathcal{Y}|X}$ be the conditional measure for $X \in \mathcal{X}$. For some $g \in \mathcal{H}_{\mathcal{Y}}$—the RKHS constructed on $\mathcal{Y}$—let $F_g(X) := \langle g|\mu_{\mathcal{Y}|X}\rangle$ be a function on $\mathcal{X}$. If $F_g \in \mathcal{H}_{\mathcal{X}}$ for all $g \in \mathcal{H}_{\mathcal{Y}}$, then $\hat{C}_{YX}\left(\hat{C}_{XX} - \zeta I\right)^{-1}|\delta_X\rangle$ converges to $|\mu_{\mathcal{Y}|X}\rangle$ as $N \to \infty$, $\zeta \to 0$, with convergence rate $O\left((N\zeta)^{-1/2} + \zeta^{1/2}\right)$. (The requirement essentially tells us that the structure of the conditional measure is compatible with the structures represented by the RKHS.) This approach is called *conditional kernel densities* [**58**].

In section 2.4, we have defined predictive states as a conditional measure, obtained by taking the convergent limit of finite-length probabilities. Our results for predictive states (Thm. 3 and Cor. 5) and for truncation error (Thm. 7 and Prop. 4) can be combined with the technique of conditional kernel densities to obtain the following theorem, which is the capstone of our work in this chapter. To state it we will remind the reader of the notation $\eta_\ell[\overleftarrow{x}]$ to denote the measure given by

$$\Pr_{\eta_\ell[\overleftarrow{x}]}(\,\cdot\,) = \Pr_\mu(\,\cdot \mid x_{-\ell+1}\ldots x_0\,)$$

In Theorems 3 and 5, we demonstrated that $\eta_\ell[\overleftarrow{x}] \to \epsilon[\overleftarrow{x}]$ in distribution as $\ell \to \infty$ for (μ-almost) all $\overleftarrow{x}$. However, the convergence rate was not fixed; we will explicitly include the unknown rate in the following theorem.

THEOREM 8. *Let $\mu \in \mathbb{P}(\mathcal{X}^{\mathbb{Z}})$ be a stationary and ergodic process. Suppose we take a long sample $w \in \mathcal{X}^L$ and from this sample subwords of length 2ℓ, $w_t = x_{t-\ell+1}\ldots w_{t+\ell}$ for $t = \ell, \ldots, L-\ell$. (There are $L - 2\ell + 1$ such words.) Split each word into a past $\overleftarrow{w}_t = x_{t-\ell+1}\ldots w_t$ and a future $\overrightarrow{w}_t = x_{t+1}\ldots x_{t+\ell}$, each of length ℓ. Define the operators:*

$$\hat{C}^{(\overleftarrow{x}\,\overleftarrow{x})}_{\beta,\gamma} = \frac{1}{L-2\ell+1}\sum_{t=\ell}^{L-\ell} |\hat{\delta}_{\overleftarrow{w}_t}\rangle_{\beta,\gamma} \otimes |\hat{\delta}_{\overleftarrow{w}_t}\rangle_{\beta,\gamma} \quad \textit{and}$$

$$\hat{C}^{(\overleftarrow{x}\,\overrightarrow{x})}_{\beta,\gamma} = \frac{1}{L-2\ell+1}\sum_{t=\ell}^{L-\ell} |\hat{\delta}_{\overleftarrow{w}_t}\rangle_{\beta,\gamma} \otimes |\hat{\delta}_{\overrightarrow{w}_t}\rangle_{\beta,\gamma} \ .$$

Now, suppose for every $g \in \mathcal{F}_{\beta,\gamma}$ that $\langle \epsilon[\overleftarrow{x}], g\rangle \in \mathcal{F}_{\beta,\gamma}$, and $\langle \eta_\ell[\overleftarrow{x}], g\rangle \to \langle \epsilon[\overleftarrow{x}], g\rangle$ at a rate of $O(h_{\overleftarrow{x}}(\ell))$. Then for all $\overleftarrow{x}$:

$$\hat{C}^{(\overleftarrow{x}\,\overrightarrow{x})}_{\beta,\gamma}\left(\hat{C}^{(\overleftarrow{x}\,\overleftarrow{x})}_{\beta,\gamma} - \zeta \cdot I_{\beta,\gamma}\right)^{-1} |\delta_{\overleftarrow{x}}\rangle_{\beta,\gamma} \to |\epsilon[\overleftarrow{x}]\rangle_{\beta,\gamma}$$

μ-almost surely in $\overleftarrow{x}$, as $L, \ell \to \infty$ and $\zeta \to 0$, at the rate $O\left((L\zeta)^{-1/2} + \zeta^{1/2} + \gamma^{-\ell} + h_{\overleftarrow{x}}(\ell)\right)$.

This integrates all our results thus far with the usual kernel Bayes' rule. Since $\epsilon[\overleftarrow{x}]$ is not generally continuous, the theorem's strict requirements on $\epsilon[\overleftarrow{x}]$ are not satisfied. That said, weaker versions hold. If $\langle \epsilon[\overleftarrow{x}], g \rangle$ as a function of $\overleftarrow{x}$ does not belong to $\mathcal{F}_{\beta,\gamma}$ as a function of $\overleftarrow{x}$, then the representational error scaling depends on the precise form of $\epsilon[\overleftarrow{x}]$. The latter can be obtained by choosing the ζ-parameter through cross-validation analysis [58, 138].

3.4. Convergence rates: case studies

The next natural question is how rapidly convergence occurs for each past, in a given process. So far, we only guaranteed that convergence exists, but said nothing on its rate. This is process-dependent. Here we will give several examples of processes and process types with their convergence rate. The most useful way to think of the rate is in the form of "probably-almost-correct"-type statements, as exemplified in the following result:

PROPOSITION 5. *Let μ be a probability measure on $\mathcal{X}^\ell$. For every $\Delta_1, \Delta_2 > 0$, we have for sufficiently large ℓ:*

$$\Pr_{\overleftarrow{\mu}} \left(\|\eta_\ell[\overleftarrow{x}] - \epsilon[\overleftarrow{x}]\| > \Delta_1 \right) < \Delta_2 \ .$$

That is, the probability of an error beyond Δ_1 is less than Δ_2.

This is a consequence of the fact that all $\overleftarrow{x}$ must eventually converge. The possible relationships between Δ_1, Δ_2, and ℓ in particular is explored in our examples.

It will help at this point to have a few examples. The simplest case, as usual, is the Markov process:

EXAMPLE 5. *Recall that a Markov process is a stochastic process where each observation x_t statistically depends only on the previous observation x_{t-1}. An order-R Markov process is one where each observation x_t depends only on the previous R observations $x_{t-R} \ldots x_{t-1}$. As such, the predictive states are simply given by:*

$$\Pr_\mu \left(x \mid \overleftarrow{x} \right) = \frac{\Pr_\mu \left(x_{-R+1} \ldots x_0 x \right)}{\Pr_\mu \left(x_{-R+1} \ldots x_0 \right)} \ ,$$

for each $\overleftarrow{x} = x_0 x_{-1} \ldots$. *Since the predictive state is entirely defined after a finite number of observations, and this number is bounded by* R*, there is no conditioning error when* R *is taken as the observation length.*

Going beyond Markov processes, processes generated by HMMs offer multiple layers of subtlety. Recall that an HMM $(\mathcal{S}, \mathcal{X}, \left\{ \mathbf{T}^{(x)} \right\})$ is defined as a finite set $\mathcal{S}$ of states, a set $\mathcal{X}$ of observations, and a set $\mathbf{T}^{(x)} = (T^{(x)}_{ss'})$ of transition matrices, labeled by elements of $\mathcal{X}$ and whose components are indexed by $\mathcal{S}$ [199]. The elements are constrained so that $T^{(x)}_{ss'} > 0$ and $\sum_{x,s'} T^{(x)}_{ss'} = 1$ for all $s \in \mathcal{S}$. Let $\mathbf{T} = \sum_x \mathbf{T}^{(x)}$ and $\boldsymbol{\pi}$ be its left-eigenvector such that $\boldsymbol{\pi}\mathbf{T} = \boldsymbol{\pi}$. HMMs generate a stochastic process μ defined by the word probabilities:

$$\Pr_\mu (x_1 \ldots x_\ell) := \sum_{s'} \left[\boldsymbol{\pi} \mathbf{T}^{(x_1)} \ldots \mathbf{T}^{(x_\ell)} \right]_{s'} .$$

An extension of HMMs, called *generalized hidden Markov models* (GHMMs) [199] (or elsewhere *observable operator models* [76]), is defined as $(\mathbf{V}, \mathcal{X}, \left\{ \mathbf{T}^{(x)} \right\})$ where $\mathbf{V}$ is a finite-dimensional vector space. The only constraint on the transition matrices $\mathbf{T}^{(x)}$ is that $\mathbf{T}^{(x)}$ have a simple eigenvector of eigenvalue 1, the left-eigenvector is still denoted $\boldsymbol{\pi}$, the right-eigenvector denoted $\boldsymbol{\phi}$, and the word probabilities:

$$\Pr_\mu (x_1 \ldots x_\ell) := \boldsymbol{\pi} \mathbf{T}^{(x_1)} \ldots \mathbf{T}^{(x_\ell)} \boldsymbol{\phi}$$

are positive [199]. GHMMs generate a strictly broader class of processes than finite hidden Markov models can [75, 76, 199], though their basic structure is very similar.

We will now consider first a special class of HMMs—called *sofic processes*—with a very well-defined convergence law. Then we will consider the general case.

EXAMPLE 6. *consider sofic processes. A* sofic process *is one that is not Markov at any finite order, but that is still expressible in a certain finite way. Namely, a sofic process is any that can be generated by a finite-state hidden Markov model with the* unifilar property. *An HMM has the unifilar property if* $T^{(x)}_{s's} > 0$ *only when* $s' = f(x, s)$ *for some deterministic function* $f : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$. *Unifilar HMMs are the stochastic generalization of deterministic finite automata in computation theory [69].*

*The most useful property of sofic processes is that the states of their minimal unifilar HMM correspond exactly to the predictive states, of which there is always a finite number. Unlike with order-R Markov processes, there is no upper bound to how many observations it may take to δ-synchronize the predictive states. However, closed-form results on the synchronization to predictive states for unifilar HMMs is already known: at L past observations, with $L \to \infty$, the conditioning error is exponentially likely (in L) to be exponentially small (in L) [**197**, **198**]. In terms of our Hilbert space norm, there are constants α and C such that*

$$\Pr_{\overleftarrow{\mu}}\left(\|\eta_\ell[\overleftarrow{x}] - \epsilon[\overleftarrow{x}]\|_{\beta,\gamma} > \alpha^\ell \right) < C\alpha^\ell \ .$$

As such, for $\overleftarrow{\mu}$-almost-all pasts, the corresponding convergence rate for the kernel Bayes' rule applied to a sofic process is $O\left((L\zeta)^{-1/2} + \zeta^{1/2} + \min(\alpha,\gamma)^{-\ell}\right)$.

EXAMPLE 7. *Not all discrete-observation stochastic processes can be generated with a finite-state unifilar hidden Markov model; though still encompassing only a small slice of processes, general hidden Markov models have a considerably larger scope of representation than finite unifilar models, as noted above. The primary challenge in this setting is to relate the structure of a given HMM to the predictive states of its process. This is achieved through the notion of mixed states. A* mixed state *ρ is a distribution over the states of a finite HMM. A given HMM, with the stochastic dynamics between its own states, induces a higher-order dynamic on its mixed states and, critically for analysis, this is an iterated function system (IFS). Under suitable conditions the IFS has a unique invariant measure, and the support of this measure maps surjectively onto the process' set of predictive states. See Refs. [**86**, **87**, **88**, **89**] for details on this construction.*

If $\rho = (\rho)$ is a mixed state, then the updated mixed state after observing symbol x is:

$$f_s^{(x)}(\rho) := \frac{1}{\sum_{s'} \left[\mathbf{T}^{(x)}\boldsymbol{\rho}\right]_{s'}} \left[\mathbf{T}^{(x)}\boldsymbol{\rho}\right]_s$$

Let the matrix $\left[\mathbf{D}f^{(x)}\right]_{s's}(\rho)$ be given by the Jacobian $\partial f_{s'}^{(x)}/\partial\rho_s$ at a given value of ρ. There is a statistic, called the Lyapunov characteristic exponent *$\lambda < 0$, such that:*

$$\lambda = \lim_{\ell\to\infty} \frac{1}{\ell} \log \frac{\left\|\mathbf{D}f^{(x_\ell)}(\rho_\ell)\cdots\mathbf{D}f^{(x_1)}(\rho_1)\mathbf{v}\right\|}{\|\mathbf{v}\|} \ ,$$

where $\rho_t := f^{(x_{t-1})} \circ \cdots \circ f^{(x_1)}(\rho)$*, for any vector* $\mathbf{v}$ *tangent to the simplex, almost any* ρ *(in the invariant measure), and almost any* $\overrightarrow{x} = x_1 x_2 \ldots$ *(in the measure of the prediction induced by* ρ*). The exponent* λ *then determines the rate at which conditioning error for predictive states converges to zero: for all* ϵ *and sufficiently large* ℓ*:*

$$\Pr_{\overleftarrow{\mu}} \left(\|\eta_\ell[\overleftarrow{x}] - \epsilon[\overleftarrow{x}]\|_{\beta,\gamma} < Ce^{\lambda \ell} \right) > 1 - \epsilon \ .$$

This is somewhat less strict—depending on how rapidly the Lyapunov exponent converges in probability. In any case, for $\overleftarrow{\mu}$*-almost all pasts, the convergence of the conditional kernel density is* $O\left((L\delta)^{-1/2} + \delta^{1/2} + \min(e^{\lambda}, \gamma)^{\ell}\right)$*, very similar to the sofic process rate.*

We anticipate that these rules still broadly apply to generalized hidden Markov models, though we recommend more detailed analysis on this question.

Lastly we will consider renewal processes, as they offer an important addendum to Theorem 8.

Example 8. *Recall that a* renewal process *is specified by the survival probability* $\Phi(n)$ *that a contiguous block of* 0*s has length at least* n*. The exact probability of a given length is* $F(n) := \Phi(n) - \Phi(n+1)$*. It is always assumed that* $\Phi(1) = 1$*. Further, stationarity requires that* $m := \sum_{n=1}^{\infty} \Phi(n)$ *be finite, as this gives the mean length of a block of* 0*s. In the most general case the predictive states are given by:*

$$\epsilon[\overleftarrow{x}] = \begin{cases} \epsilon_k & \overleftarrow{x} = 0^k 1 \ldots \\ \text{undefined} & \overleftarrow{x} = 0^\infty \end{cases} ,$$

where the measures ϵ_k *are recursively defined by the word probabilities:*

$$\Pr_{\epsilon_k}\left(0^\ell 1 w\right) = \frac{F(k+\ell)}{\Phi(k)} \Pr_{\epsilon_0}(w) \ .$$

Now, it can be easily seen that each past $\overleftarrow{x}$ *converges to zero conditioning error at a finite length since (almost) all pasts have the structure* $\ldots 10^k$*, and so only the most recent* $k+1$ *values need be observed to know the predictive state. Therefore the kernel Bayes' rule has an asymptotic convergence rate for each past* $\overleftarrow{x}$ *of* $O\left((L\delta)^{-1/2} + \delta^{1/2} + \gamma^{-\ell}\right)$*. However, this does not tell the entire story, as*

obviously not all pasts converge uniformly. A probabilistic expression of the conditioning error gives more information:

PROPOSITION 6. *Suppose μ is a renewal process with $\Phi(n) \propto n^{-\alpha}$, $\alpha > 1$. Then there exist constants C and K such that:*

$$\Pr_{\overleftarrow{\mu}}\left(\|\eta_{x_1 \ldots x_\ell} - \epsilon[\overleftarrow{x}]\|_{\beta,\gamma} > C\ell^{-1} \right) > K\ell^{-\alpha} \ .$$

That is, the probability the conditioning error decays as $1/\ell$ is itself at least power-law decaying in ℓ.

PROOF. *Recall from Eq. (3.4):*

$$\begin{aligned} &\|\eta_{x_1 \ldots x_\ell} - \epsilon[\overleftarrow{x}]\|_{\beta,\gamma} \\ &\quad > \frac{|\Pr_\mu(\, w \mid x_1 \ldots x_\ell\,) - \Pr_\mu(\, w \mid \overleftarrow{x}\,)|}{\sqrt{\langle w|w\rangle_{\beta,\gamma}}} \ , \end{aligned}$$

for every word w, so we can choose any w and obtain a lower bound on the conditioning error. If our past $\overleftarrow{x}$ has the form $0^k 1 \ldots$ for $k < \ell$, then we are already synchronized to the predictive state and the conditioning error is zero. Thus, we are specifically interested in the case $k \geq \ell$ and we will further consider the large-ℓ limit.

Now, under our assumptions, $\Phi(n) = n^{-\alpha}$ for some constant Z. For large n, $F(n) \sim \alpha n^{-\alpha-1}$. Then for any j:

$$\Pr_\mu\left(0^j 1 \,\middle|\, \overleftarrow{x} \right) = \frac{F(k+j)}{\Phi(k)} \sim \frac{\alpha}{k}\left(\frac{k+j}{k}\right)^{-\alpha-1} \ .$$

Meanwhile, so long as $k \geq \ell$, the truncated prediction has the form:

$$\begin{aligned} \Pr_\mu\left(0^j 1 \,\middle|\, 0^\ell \right) &= \sum_{n=1}^{\infty} \frac{\Phi(n+\ell)}{\sum_p \Phi(p+\ell)} \frac{F(n+\ell+j)}{\Phi(n+\ell)} \\ &= \frac{\Phi(\ell+j)}{\sum_p \Phi(p+\ell)} \sim \frac{\alpha-1}{\ell}\left(\frac{\ell+j}{\ell}\right)^{-\alpha} \ . \end{aligned}$$

Now, choose $0 < C < \alpha - 1$ and define:

$$B = \left(1 - \frac{C+1}{\alpha}\right)^{-1} \ .$$

Then it can be checked straightforwardly that whenever $k > B\ell$, we have:

$$\begin{aligned}\Pr_\mu\left(1 \,\middle|\, 0^\ell\right) - \Pr_\mu\left(1 \mid \overleftarrow{x}\right) &\sim \frac{1}{\ell}\left[\alpha\left(1 - \frac{\ell}{k}\right) - 1\right] \\ &> \frac{C}{\ell} \ .\end{aligned}$$

The probability that $k > B\ell$ is given by $\Phi(B\ell) = B^{-\alpha}\ell^{-\alpha}$. Setting $K = B^{-\alpha}/\sqrt{\langle 1|1\rangle_{\beta,\gamma}}$ proves the theorem.

Therefore, while every sequence $\overleftarrow{x}$ converges to zero conditioning error at finite length, this convergence is not uniform, to such a degree that the proportion of pasts that retain conditioning error of $1/\ell$ has a fat tail in ℓ. This is a matter of practical importance that is not cleanly expressed in the big-O expression of the conditioning error from Thm. 8.

3.5. Discussion

Any use of predictive states in machine learning will require embedding techniques to represent the information contained in the state. We demonstrated that the popular RKHS embedding is a valid embedding of predictive states, in that it inherits their stability. We also showed how this stability extends to other methods, such as an embedding based on combining the Wasserstein metric with the Cantor fractal. This means that our results provide *general* insight into efficient and stable machine learning algorithms for time series: one path to viability for an algorithm is that it emulates the predictive state formalism.

Compared to using reproducing kernel Hilbert spaces—a dominant approach to predictive states at present—our Cantor-Wasserstein embedding may appear a mere toy model. However, as the results demonstrated, there are strong benefits to the approach, which synergizes the benefits of both the Wasserstein distance and the Cantor embedding. The topology of convergence in distribution can be replicated with both the Wasserstein distance and the RKHS inner product. However, the Wasserstein distance depends on far fewer parameters—such as, the choice of the eponymous kernel in RKHS approaches. Moreover, its value is directly interpretable in terms of the shapes of the distributions it compares.

Similarly, there are many ways to metrize the product topology on sequences, but the Cantor embedding offers a direct way to connect the product topology with a visualizable geometry. And, embedding in a single dimension enables efficient computation of the Wasserstein metric. The benefits of the Cantor and Wasserstein approaches adds interpretability to the resulting predictive-state geometry along two distinct axes, most clearly seen in Fig. 3.1's clustered Cantor diagrams. We hope that the success of this approach in providing clear insights will complement existing thrusts in the direction of abstract embeddings and mathematical formalism by motivating further development on interpretable approaches to predictive state analysis.

We close this chapter by noting that predictive states are not merely static objects. They predict the probabilities of future observations. And, once those observations are made, the predictive state may be updated to account for new information. Thus, predictive states provide the stochastic rules for their own transformation into future predictive states. This dynamical process has been explored in great detail in the cases where the process is generated by a finite hidden Markov model—this is found in former work on the ϵ-machine (*e.g.* [**172**]) and the mixed states of HMMs (*e.g.* [**87**]). In the following Chapters 4 through 6, we will show how the dynamics of predictive states must be emulated in the internal states of any model or physical system which generates the stochastic process. This constraint tells us an enormous amount about the space of possible models for a given process, and the resource constraints those models must face if they are to be implemented in the physical world.

CHAPTER 4

There you are: Generating processes with memory

Wherever you go, there you are.

Old proverb

4.1. Introduction

When studying classical stochastic processes, we often seek models and representations of the underlying system that allow us to simulate and predict future dynamics. If the process is memoryful, then models that generate it or predict its future behaviors must also have memory. Memory, however, comes at some resource cost; both in a practical sense—consider, for instance, the substantial resources required to generate predictions of weather and climate [**113**, **114**]—and in a theoretical sense—seen in analyzing resource use in thermodynamic systems such as information engines [**24**]. It is therefore beneficial to seek out a process' minimally resource-intensive implementation. Notably, this challenge remains an open problem with regards to both classical and quantum processes.

In Chapter 1 we defined a process phenomenologically, as a "source" which produces a data stream. When that stream is stationary and ergodic, a process can be described by its "word probabilities" $\Pr_\mu(x_1 \ldots x_\ell)$, which describe how frequently certain observations appear consecutively. In Chapter 2 we showed that these word probabilities characterize a *measure* μ, which is the mathematical formalization of a process. We also defined the notion of a *predictive state*. The arena of *computational mechanics* relates predictive states to the information processing capacities required to produce and predict a process [**38**, **39**, **43**, **173**]. The minimal information processing required to predict the sequence is represented by a type of hidden Markov model called the *ϵ-machine*. The statistical complexity C_μ—the memory rate for ϵ-machines to simultaneously generate many copies of a process—is a key measure of a process' memory resources. Where finite, C_μ is known to be the minimal memory rate over all classical implementations.

Computational mechanics has largely been developed in the domain of traditional information theory. A generalization of information theory, termed *resource theory* has recently emerged within quantum information theory as a toolkit for addressing resource consumption in the contexts of entanglement, thermodynamics, and numerous other quantum and even classical resources [**35**]. Its fundamental challenge is to determine when one system (a *resource*) can be converted to another using a predetermined set of *free* or *allowed* operations.

Resource theory is closely related to the theory of *majorization*, which the reader will recall from Section 1.5. On the one hand, majorization is a preorder relation $\succsim$ on positive vectors (typically probability distributions) computed by evaluating a set of inequalities [**123**]. If the majorization relations hold between two vectors, then one can be converted to the other using a certain class of operations. Majorization is used in several resource theories to numerically test for convertibility between two resources [**64**, **71**, **142**].

In this chapter we explore the concept of a model as the algebraic representation of the predictive state dynamics. Specifically, we formulate the manner in which predictive states transform into one another over the passage of time as a *semigroup*, and define models as a certain kind of representation of this semigroup. Essentially, the invariant geometry of a process's predictive states is again and again repeated in the state space of *any* model of that process; as the old adage says, "Wherever you go—there you are."

Representation theory for stochastic processes was first considered in Refs. [**81**, **82**, **83**], though these works did not receive much attention at the time. Later, though not formulated in the language of representation theory, equivalent results were independently reached by both Ref. [**199**] and Refs. [**76**, **193**, **217**]. We synthesize the perspectives from these approaches and combine them with our results on conditional measures from Chapter 2 to build up a general theory of models and generators of stochastic processes which will carry us through Chapters 5 and 6. This is accomplished in Section 4.2.

In Sections 4.3 and 4.4, we will shift our attention to focusing on hidden Markov models of stochastic processes. As a special case of representation models, we can interpret HMMs physically as describing system dynamics. We use majorization theory to develop a resource-theoretic interpretation of model memory, and in this new formalism we generalize existing results on the minimality of the

ϵ-machine as a predictive model of a stochastic process. These sections are mostly drawn from the publication *Strong and Weak Optimizations in Classical and Quantum Models of Stochastic Processes* [**106**].

4.2. Models and representations

Because systems and measurements are so intertwined in this discussion, we will not define them as separate entities, but rather as one object: a *generator*. A generator is a triple $\mathfrak{G} = (\mathcal{S}, \mathcal{X}, T)$, where $\mathcal{S}$ is a set of states which we will identify as the system, $\mathcal{X}$ is an alphabet of measurement outcomes, and $T : \mathcal{S} \to \mathbb{P}(\mathcal{X} \times \mathcal{S})$ is a conditional measure which, given a current system state as input, provides a measure over pairs of observations and updated system states.

Why the name "generator"? This is because because generators sequentially emit observations, and therefore produce a process. Suppose, for instance, that $\mathcal{X}$ is discrete, in which case we can write T instead as a set of functions $T^{(x)} : \mathcal{S} \to \mathbb{M}(\mathcal{S})$ which map each state to a positive measure, so that $\sum_x T^{(x)}$ always results in a probability measure. Then we can define the process generated by $\mathfrak{G}$ and initial state s_0 using the word probabilities

$$\Pr_{\mathfrak{G}}(\, x_1 \ldots x_\ell \mid s_0 \,) = \int_{\mathcal{S}^\ell} dT^{(x_\ell)}(s_\ell | s_{\ell-1}) \times \cdots \times dT^{(x_1)}(s_1 | s_0)$$

If we also assume that $\mathcal{S}$ is discrete, then we can think of each $T^{(x)}$ as a matrix $\mathbf{T}^{(x)} = (T^{(x)}_{s's})$, and write the word probabilities as

$$\Pr_{\mathfrak{G}}(\, x_1 \ldots x_\ell \mid s_0 \,) = \sum_{s_\ell} \left[\mathbf{T}^{(x_\ell)} \ldots \mathbf{T}^{(x_1)} \right]_{s_\ell s_0}$$

using matrix multiplication.

Generators encapsulate together the system dynamics and measurement process. As it happens, this encapsulated approach will be most useful when it comes to relating generators to predictive states. The relation arises from two fundamental facts. The first we have essentially already stated: the previous paragraph demonstrates that each generator $\mathfrak{G}$ induces a mapping from its state space $\mathcal{S}$ to the space $\mathbb{P}(\mathcal{X}^{\mathbb{N}})$ of probability measures over sequences. We will call this mapping $P_{\mathfrak{G}}(s)$, so that

$$\Pr_{P_{\mathfrak{G}}(s)}(\, x_1 \ldots x_\ell \,) = \Pr_{\mathfrak{G}}(\, x_1 \ldots x_\ell \mid s \,)$$

The second fact has to do with predictive states themselves. The reader may recall that in Sec. 2.1 we described the existence of a mathematical object, the ϵ-*machine*, which characterizes how the predictive state is altered upon making new observations. We now have more precise language to express the nature of the ϵ-machine: it is a generator of the process whose set of states is just the predictive states. Specifically, the ϵ-machine of a process μ is the triple $\mathfrak{E}(\mu) = (\mathcal{K}(\mu), \mathcal{X}, T)$ with the property that $T^{(x)}(\epsilon[\overleftarrow{x}])$ produces a scaled Dirac delta measure $\Pr_\mu(\, x \mid \overleftarrow{x}\,)\, \delta_{\epsilon[\overleftarrow{x}\, x]}$.

Let us break this down. Given a starting predictive state $\epsilon[\overleftarrow{x}]$ conditional measure T assigns a probability to the next symbol of $\Pr_\mu(\, x \mid \overleftarrow{x}\,)$, and noiselessly transitions to the updated state $\epsilon[\overleftarrow{x}\, x]$. This just means we are observing x with the probability determined by conditioning on all previously observed data, and then reconfiguring our prediction of the future based on our observation of x.

One of the most remarkable facts about the ϵ-machine is that it has a parallel interpretation, not in terms of generators but in terms of vector spaces. Predictive states are, after all, measures, and can be linearly combined; the vector space they span we have named $\mathbb{K}(\mu)$. Consider, now, the mapping for any probability measure η over $\mathcal{X}^{\mathbb{N}}$ given by $\eta \mapsto \tau^{(x)}\eta$, defined as

$$\Pr_{\tau^{(x)}\eta}(\, w\,) = \Pr_\eta(\, xw\,)$$

(You saw this earlier in Sec. 2.1.) The reader may quickly convince themselves that $\tau^{(x)}$ is a *linear map*. It is also related to the ϵ-machine: for each past $\overleftarrow{x}$, we have $\tau^{(x)}\epsilon[\overleftarrow{x}] = \Pr_\mu(\, x \mid \overleftarrow{x}\,)\, \epsilon[\overleftarrow{x}\, x]$. The set $\{\tau^{(x)}\}$ of these operators (spanning over x) generates an entire semigroup $\mathcal{T}(\mu) = \{\tau^{(x_1 \ldots x_\ell)}\}$ of operators, each corresponding to a sequence of observations as $\tau^{(x_1 \ldots x_\ell)} = \tau^{(x_\ell)} \ldots \tau^{(x_1)}$. We call this semigroup the *observable semigroup*.

We will return to the observable semigroup in a moment. Let us now combine what we know about generators with what we know about predictive states and the ϵ-machine. From the definition of the map $P_{\mathfrak{G}}$, which maps generator states $\mathcal{S}$ into the vector space $\mathbb{P}(\mathcal{X}^{\mathbb{N}})$, and that of the observable operators $\tau^{(x)}$, we have the notable fact:

$$\tau^{(x)} \circ P_{\mathfrak{G}}(s) = \int_{\mathcal{S}} P_{\mathfrak{G}}(s')\, dT^{(x)}(s'|s) \tag{4.1}$$

At this point, the reader may notice a multitude of consequences.

(1) Practically, this equation shows that the operator $\tau^{(x)}$ can "pass through" the mapping $P_{\mathfrak{G}}$ between the prediction space $\mathbb{K}(\mu)$ and the system state space $\mathcal{S}$, transitioning from a linear map to a conditional measure.

(2) This "passing-through" (or should we say "gliding over"?) guarantees that several linear and spectral properties must be shared between $\tau^{(x)}$ and $T^{(x)}$ (when the latter is viewed as a linear map on the space $\mathbb{M}(\mathcal{S})$).

(3) The reader with background in mathematical physics may also notice that this equation guarantees, when the conditional maps $T^{(x)}$ are viewed as linear maps on the space $\mathbb{M}(\mathcal{S})$, that they form a *representation* of the observable semigroup $\mathcal{T}(\mu)$. Further, the map $P_{\mathfrak{G}}$ forms essentially a kind of homomorphism between representations, called an intertwining operator.

The representation-theoretic language allows us to encapsulate these facts in a single statement. For any generator of a process, the conditional maps $T^{(x)}$ over its state space form a homomorphic representation of the predictive state dynamics for the same process.

It is in the representation-theoretic context that we will now, finally, introduce a definition of an *observable operator model*. Our definition essentially follows that used by [76, 193, 217], and is also equivalent to the generalized hidden Markov model used in [199]. An OOM of a process μ is any quartet $\mathfrak{M} = (\mathcal{V}, \mathcal{X}, \{T^{(x)}\}, P)$, where $\mathcal{V}$ is a vector space, $\mathcal{X}$ is the alphabet, each $T^{(x)} : \mathcal{V} \to \mathcal{V}$ is a linear operator on $\mathcal{V}$, and $P : \mathcal{V} \to \mathbb{M}(\mathcal{X}^{\mathbb{N}})$ is a *not necessarily linear* mapping from vectors to measures over sequences, satisfying the relation

$$\tau^{(x)} \circ P = P \circ T^{(x)} \tag{4.2}$$

As with generators, this definition guarantees that a model of μ always contains a representation of the observable semigroup.

The primary distinction between an OOM and a generator is semantic but, we feel, important. A generator is defined simply as some dynamics paired with an observation mechanic, which happens to produce a process as a byproduct. An OOM, on the other hand, is always defined relative to a process's observable semigroup. This is also the main difference between our definition of an OOM and that used in [76] (in which the OOM is not bound to a process, and indeed may not even be

guaranteed to generate a positive measure over sequences). This distinction will be helpful to keep in mind in the current chapter. In any case, it should be clear from what we have discussed that every generator contains an OOM implicitly (in the vector space of distributions over its states); similarly, every OOM contains a generator.

4.2.1. The observable semigroup. Let us discuss the observable semigroup in greater detail, and discuss its properties. The main point which we would like to impart to the reader is that the observable semigroup for a given process is just as fundamental an object as the predictive states.

Let us illustrate what we mean. Any process μ (defined as a measure on $\mathcal{X}^{\mathbb{Z}}$) induces a measure over "future" sequences, $\overrightarrow{\mu}$, which is a measure over $\mathcal{X}^{\mathbb{N}}$. The stationarity of the process means that the measure $\overrightarrow{\mu}$ is invariant to the action of the shift operator $\tau = \sum_x \tau^{(x)}$:

$$\Pr_{\tau\overrightarrow{\mu}}(\, w \,) = \sum_x \Pr_{\overrightarrow{\mu}}(\, xw \,) = \Pr_{\overrightarrow{\mu}}(\, w \,)$$

(Recall that *ignoring* the first observation of a stationary process should have no impact on our predictions, if we know nothing else about the current state.)

Now, given a set of operators $\mathcal{T}$ and a vector v, the *cyclic subspace* generated by that pair is the subspace spanned by all vectors of the form $T_1 \dots T_n v$, where $T_1, \dots, T_n \in \mathcal{T}$. In other words, it is the space "reachable" by applying the set of operators to the vector. (More specifically, it is really the topological *closure* of that set.) The cyclic subspace of the observable semigroup $\mathcal{T}(\mu)$ acting on $\overrightarrow{\mu}$ would be the closure of all vectors of the form

$$\tau^{(x_0)} \dots \tau^{(x_{-k})} \overrightarrow{\mu} = \Pr_{\mu}(\, x_{-k} \dots x_0 \,)\, \eta_\ell[\overleftarrow{x}\,]$$

for each past $\overleftarrow{x}$ and $k \geq 0$. Then, due to the convergence theorem of predictive states Thm. 3, the cyclic subspace of $\mathcal{T}(\mu)$ acting on $\overrightarrow{\mu}$ must just be the space $\mathbb{K}(\mu)$.

To rephrase this fact, the just as the observable semigroup can be defined in terms of the natural dynamics on the predictive state space $\mathbb{K}(\mu)$, it is also true that $\mathbb{K}(\mu)$ is recoverable from the observable semigroup as a cyclic subspace. We can therefore see the observable semigroup as being just as fundamental to the characterization of a stochastic process as predictive states are. What is

additionally remarkable, with considerable significance, is that $\mathbb{K}(\mu)$ can be generated as the cyclic subspace of $\mathcal{T}(\mu)$ starting from *any* probability measure $\eta \in \mathbb{K}(\mu)$.

THEOREM 9. *For any process μ and any almost any $\eta \in \mathcal{K}(\mu)$, the cyclic subspace generated by the action of $\mathcal{T}(\mu)$ is just $\mathbb{K}(\mu)$.*

Proving this requires making an interesting observation, which is an extension o fthe earlier result Prop. 5. There we noted that for every $\Delta_1, \Delta_2 > 0$, for sufficiently large ℓ

$$\Pr_{\overleftarrow{\mu}} (\|\eta_\ell[\overleftarrow{x}] - \epsilon[\overleftarrow{x}]\| > \Delta_1) < \Delta_2 \ .$$

The norm here stands in for any distance which generates the topology of convergence in distribution. A much stronger form is the following:

PROPOSITION 7. *For almost every $\overleftarrow{x}$, and any $\Delta_1, \Delta_2 > 0$, there is a sufficiently large ℓ such that*

$$\frac{\overleftarrow{\mu} (\{ \overleftarrow{y} \mid \|\epsilon[\overleftarrow{y}] - \epsilon\overleftarrow{x}\| < \Delta_2, \ y_{-\ell} \ldots y_0 = x_{-\ell} \ldots x_0 \})}{\overleftarrow{\mu}(\{ \overleftarrow{y} \mid y_{-\ell} \ldots y_0 = x_{-\ell} \ldots x_0 \})} > 1 - \Delta_1 \ .$$

In other words, there is an arbitrarily high probability that "close" pasts (in the cylinder set sense) will also have close predictive states.

PROOF. *This is a consequence of applying the Chebyshev inequality. We already know, for most pasts, that $\Pr_\mu (w \mid x_{-\ell} \ldots x_0)$ converges to $\Pr_\mu (w \mid \overleftarrow{x})$. Now consider instead the square $\Pr_\mu (w \mid \overleftarrow{x})^2$. By the Vitali property (Prop. 2),*

$$\lim_{\ell \to \infty} \frac{\int_{U_{-\ell-1, x_{-\ell} \ldots x_0}} \Pr_\mu (w \mid \overleftarrow{y})^2 \, d\overleftarrow{\mu}(\overleftarrow{y})}{\overleftarrow{\mu}(U_{-\ell-1, x_{-\ell} \ldots x_0})} = \Pr_\mu (w \mid \overleftarrow{x})^2$$

Let us consider this in terms of expectations and moments. The left-hand side corresponds to the second moment $\mathbb{E}_\mu \Big[\Pr_\mu (w \mid \overleftarrow{y})^2 \Big| \overleftarrow{y} \in U_{-\ell-1, x_{-\ell} \ldots x_0} \Big]$, while the right-hand side corresponds to the square of the first moment $\mathbb{E}_\mu \Big[\Pr_\mu (w \mid \overleftarrow{y}) \Big| \overleftarrow{y} \in U_{-\ell-1, x_{-\ell} \ldots x_0} \Big]^2$. Their equality in the limit implies that the variance $\mathrm{Var}_\mu \Big[\Pr_\mu (w \mid \overleftarrow{y}) \Big| \overleftarrow{y} \in U_{-\ell-1, x_{-\ell} \ldots x_0} \Big]$ converges to zero as $\ell \to \infty$. Then by the Chebyshev inequality (or any other similar inequality), the theorem is proven.

This proposition proves an important property of the function $\Pr_\mu (w \mid \overleftarrow{x})$, as a function of $\overleftarrow{x}$. We say it is *essentially continuous*. What this means is that, for almost all $\overleftarrow{x}$, there is a sufficiently

small neighborhood so that almost everywhere in that neighborhood the function is close to its value on $\overleftarrow{x}$. It is a very weak form of continuity. However, it will prove a very useful concept going forwards.

Now we can prove Thm. 9.

PROOF. *Due to the Prop. 7, starting from any predictive state $\epsilon[\overleftarrow{x}]$, there is a sufficiently long sequence of symbols $y_{-\ell} \dots y_0$ such that $\epsilon[\overleftarrow{x}\, y_{-\ell} \dots y_0]$ is arbitrarily close to $\epsilon[\overleftarrow{y}]$, for a given $\overleftarrow{y}$. Since $\epsilon[\overleftarrow{x}\, y_{-\ell} \dots y_0]$ is proportional to $\tau^{(y_0)} \dots \tau^{(y_{-\ell})}$, this means that $\epsilon[\overleftarrow{y}]$ is reachable. Therefore all of $\mathbb{K}(\mu)$ is in the cyclic subspace of $\mathcal{T}(\mu)$ starting from $\epsilon[\overleftarrow{x}]$.*

Theorem 9 is significant because we can interpret it as requiring a kind of *ergodicity* over the dynamics of predictive states as driven by the observable semigroup. Remember that one interpretation of ergodicity is that "everything can be reached from everything else," and this is exactly what we have shown here.

What we will demonstrate next is how the relationship between $\mathcal{T}(\mu)$ and $\mathbb{K}(\mu)$ is mirrored in representations of $\mathcal{T}(\mu)$, and therefore is mirrored in any model of the process μ. In particular, just as ergodicity has implications on finite graphs in the form of the Perron-Frobenius theorem, we will soon see how this allows us to implement some form of Perron-Frobenius theory for general models.

4.2.2. Irreducibility and indecomposibility. We have defined an OOM $\mathfrak{M} = (\mathcal{V}, \mathcal{X}, \{T^{(x)}\}, P)$ as a representation of the observable semigroup, where each element $\tau^{(x)}$ has a corresponding $T^{(x)} \in \mathrm{GL}(\mathcal{V})$, paired with a correspondence relation $P : \mathcal{V} \to \mathbb{M}(\mathcal{X}^{\mathbb{N}})$ mapping vectors in $\mathcal{V}$ to measures over future sequences, satisfying the relationship

$$\tau^{(x)} \circ P = P \circ T^{(x)} \tag{4.3}$$

The map P is not necessarily a linear map, so it cannot be characterized as a homomorphism between representations (which are traditionally linear). If it is linear, however, we will call the model a *linear OOM*.

An example of a nonlinear OOM would be the class of *quadratic OOMs*, which we will return to in Chapter 5. We will for now focus on linear models, as there is much of interest we can say about them.

In mathematical physics, as soon as the subject of representation comes up, it is almost always accompanied by the classification of all irreducible representations for the algebraic object. Typically this object is a *group*. This is significant, because it means that the word "irreducible" takes on two equivalent meanings, which are *not* necessarily equivalent in the semigroup setting. The standard meaning of an irreducible representation is one which contains no proper subspaces which are invariant under the action of the group representation. Since group objects are invertible, this also implies *indecomposibility*: the representation space cannot be the direct sum of two invariant sub-representations.

The difference which arises in the case of semigroups like $\mathcal{T}(\mu)$ is the possibility of *transient* subspaces. This concept should be familiar from our discussion of the Perron-Frobenius theorem in Sec. 1.4. Transient subspaces "sink" into invariant subspaces irreversibly, meaning that one can have a proper invariant subspace without decomposibility: the transient subspace is neither invariant nor a representation in its own right. This is not possible with a group action because of its inherent invertibility, meaning there is no such thing as irreversible transience.

What we shall do here is point out a handful of useful observations about irreducibility and indecomposibility for linear OOMs. These observations will be useful in a variety of settings later, but our present interest will simply be to provide their straightforward derivation.

First, we will start by considering the nature of the linear map P. Let $\mathbf{1} : \mathbb{M}(\mathcal{X}^{\mathbb{N}})$ be the linear map on measures which corresponds to evaluation on the entire space: $\mathbf{1}\nu = \nu(\mathcal{X}^{\mathbb{N}})$. This mapping always sends probability measures to 1, hence the name. Define the composition $\mathbf{1} \circ P$, which we will name $\mathbf{1}_{\mathfrak{M}}$. A vector $v \in \mathcal{V}$ is called *unital* if $\mathbf{1}_{\mathfrak{M}}(v) = 1$.

It is clear, then, that certain unital vectors in $\mathcal{V}$ should map to probability measures over $\mathcal{X}^{\mathbb{N}}$. We can, in fact, directly access the word probabilities of these measures by using the homomorphism rule Eq. (4.3):

$$\Pr_{\mathfrak{M}}(x_1 \dots x_\ell \mid v) = \mathbf{1}_{\mathfrak{M}} T^{(x_\ell)} \dots T^{(x_1)} v$$

Now, it should be noted that the map $T = \sum_x T^{(x)}$ always maps unital vectors to unital vectors; this is just a consequence of the fact that summing over any of the symbols in the above equation still results in a probability measure. It is therefore the case that $\mathbf{1}_{\mathfrak{M}} \circ T = \mathbf{1}_{\mathfrak{M}}$.

This fact is extremely important: it tells us that 1 is an eigenvalue of T, because $\mathbf{1}_{\mathfrak{M}}$ is a fixed point. It must therefore be the case that there also exists *at least one* vector, say π, in $\mathcal{V}$ which is a fixed point of T: $T\pi = \pi$.

The following result concerns the relationship between decomposibility of a model $\mathfrak{M}$ and the dimension of T's stationary subspace.

THEOREM 10. *Let $\mathfrak{M} = (\mathcal{V}, \mathcal{X}, \{T^{(x)}\}, P)$ be a linear model. Let $\{\pi_j\}$ be a linearly independent set of stationary states of the matrix $T = \sum_x T^{(x)}$, such that $T\pi_j = \pi_j$ for each j. Then $\mathfrak{M}$ in fact contains j sub-models, each inhabiting an invariant subspace of $\mathcal{V}$. The converse (that the number of stationary states is equal to the number of sub-models) is also true.*

To prove this theorem we must do some legwork which will also have further value down the line. We will introduce here the concept of *embedding pasts in models*, and prove an important lemma about embeddings of pasts. This lemma is a much more specific form of Theorem 3.4.3 from Ref. [**199**].

LEMMA 1. *Let $\mathfrak{M} = (\mathcal{V}, \mathcal{X}, \{T^{(x)}\}, P)$ be a linear model of the process μ such that $\mathcal{V}$ is finite-dimensional, and let π be any one of the stationary states of T. Then there is a mapping $E_{\mathfrak{M},\pi} : \mathcal{X}^{\mathbb{N}} \to \mathcal{V}$, called a* past embedding*, with the properties*

(1) $E_{\mathfrak{M},\pi}(\overleftarrow{x})$ is a homomorphism of the shift space:

$$T^{(x)} E_{\mathfrak{M},\pi}(\overleftarrow{x}) = \Pr_{\mu}(\, x \mid \overleftarrow{x}\,)\, E_{\mathfrak{M},\pi}(\overleftarrow{x}\, x)\ ; \tag{4.4}$$

(2) $E_{\mathfrak{M},\pi}(\overleftarrow{x})$ is predictively consistent:

$$\Pr_{PE_{\mathfrak{M},\pi}(\overleftarrow{x})}(\, w\,) = \Pr_{\mu}(\, w \mid \overleftarrow{x}\,)\ ; \tag{4.5}$$

(3) $E_{\mathfrak{M},\pi}$ is essentially continuous (in the same sense as used in Prop. 7);

(4) $E_{\mathfrak{M},\pi}(\mathcal{X}^{\mathbb{N}})$ spans the cyclic subspace generated by $\{T^{(x)} | x \in \mathcal{X}\}$ starting from π or $E_{\mathfrak{M},\pi}(\overleftarrow{x})$ for any $\overleftarrow{x}$.

PROOF. *We will define $E_{\mathfrak{M},\pi}$ as a Radon-Nikodym derivative first. If $\mathcal{V}$ is finite dimensional, we can give each vector $v \in \mathcal{V}$ coorinates $\mathbf{v} = (v_j)$. For each j, define the measure ν_j over pasts as*

$$\nu_j\left(U_{-\ell, x_{-\ell} \ldots x_0}\right) = \left[T^{(x_0)} \ldots T^{(x_{-\ell})} \pi\right]_j$$

Then $E_{\mathfrak{M},\pi}(\overleftarrow{x})$ *is the unique vector for which*

$$[E_{\mathfrak{M},\pi}(\overleftarrow{x})]_j = \frac{d\nu_j}{d\overleftarrow{\mu}}(\overleftarrow{x})$$

Property (1) follows from the fact that

$$\sum_j T^{(x)}_{j'j}[E_{\mathfrak{M},\pi}(\overleftarrow{x})]_j = \frac{d}{d\overleftarrow{\mu}}\left(\sum_j T^{(x)}_{j'j}\nu_j\right)(\overleftarrow{x}) = \Pr_\mu(\, x \mid \overleftarrow{x}\,)\,\frac{d\nu_{j'}}{d\overleftarrow{\mu}}(\overleftarrow{x}\,x)$$

which is a consequence of the definition of ν_j*. Property (2) is itself a direct consequence of property (1) and the homomorphism law for models, Eq.* (4.3)*.*

Property (3), on the other hand, follows from the Vitali property of $\mathcal{X}^{\mathbb{N}}$ *for exactly the same reasons given in Prop. 7, which we will not repeat here. Similarly, property (4) follows just as Thm. 9 followed from Prop. 7.*

Now it should be evident then that to each stationary π_j associated with a model $\mathfrak{M}$, there corresponds an invariant ergodic subspace spanned by the image of $E_{\mathfrak{M},\pi_j}$. Due to this ergodicity, in each of these subspaces there can only be one stationary state: π_j itself. So, none of these subspaces can intersect except at the origin. This proves Theorem 10.

What we have shown, then, is that every model either decomposes into a collection of sub-models (plus some transient subspaces), or is itself indecomposible. In the latter case, due to Theorem 10, there can only be one stationary state. Since it goes without saying that any sub-model must contain at least one stationary state, we find that *the existence of a single stationary state* π and *being indecomposible* are equivalent properties for OOMs. Further, due to Lemma 1, this implies the existence of a unique embedding of past states into any indecomposible model.

4.2.3. Time-reversal of models and processes. Thus far we have thought of processes, their dynamics, and the predictive states and models which describe processes, all in terms of a forward-pointing arrow of time. One symbol is generated after another, and each model state transforms to the next. The future is conditioned on the past.

There is nothing stopping us, however, from turning the clock backwards. In fact, everything we have proven remains true. As it happens, however, the structures we define in the reverse direction

for a process are rarely mirror images of the same structures done forwards. Many times, they provide us with additional crucial information about the process under study. Here we will just make some simple definitions of reverse-time objects and operations which will be useful later in this chapter.

We will start be defining the reverse of a process. Given a stochastic process μ, the reverse process, denoted μ^R, is defined by the word probabilities

$$\Pr_{\mu^R}(x_1 \ldots x_\ell) = \Pr_\mu (x_\ell \ldots x_1)$$

That is, it assigns to every word the probability which μ assigned the reversed word. In this sense, a reverse process "turns back time" by reversing the order of our observations.

Some processes are symmetric under reversal. An example of this is the Even process, which is defined only by the fact that 1's only appear in even-sized contiguous blocks. This property of a word is unchanged under reversal and so the reverse of the Even process is still the Even process. On the other hand, consider the $\mathtt{a}^n\mathtt{b}^n$ process, where every contiguous block of `a`'s is followed by a block of `b`'s of the same size. Obviously, under time reversal, the role of `a` and `b` are reversed, and so the process is not symmetric under time reversal. The can be visualized in Fig. 2.1: the Cantor embedding of a process being symmetric under the swapping of the axes is a visualization of time-reversal symmetry.

Now, just as μ has predictive states, so does μ^R. Consider a future $\overrightarrow{x} = x_1 x_2 \ldots$ and its reversal $\overleftarrow{x} = \ldots x_2 x_1$. Then we define the *retrodictive state* of $\overrightarrow{x}$, denoted by $\epsilon^R[\overrightarrow{x}]$, as the measure over *pasts* defined by the probabilities

$$\Pr_{\epsilon^R[\overrightarrow{x}]}(y_{-\ell} \ldots y_0) = \Pr_{\mu^R}(y_0 \ldots y_{-\ell} \mid \overleftarrow{x})$$

While predictive states are measures over the future (conditioned on the past), retrodictive states are the opposite: they are measures over the past, conditioned on the future. They can be mathematically treated as the predictive states of the reverse process μ^R and inherit all of its properties (when all temporalities are appropriately reversed).

The last thing we will define here is the time reversals of generators and models. For clarity we will do so under the assumption that the underlying spaces are finite (that is, that $\mathcal{S}$ is finite or $\mathcal{V}$ is finite-dimensional).

Consider a finite generator $\mathfrak{G} = (\mathcal{S}, \mathcal{X}, \{\mathbf{T}^{(x)}\})$. We will suppose it is ergodic, in the sense that every state can be reached by every other via the stochastic dynamic of the $\mathbf{T}^{(x)}$'s; then there is a unique stationary distribution $\boldsymbol{\pi} = (\pi_s)$ which satisfies $\mathbf{T}\boldsymbol{\pi} = \boldsymbol{\pi}$. The reverse generator is then defined as $\mathfrak{G}^R = (\mathcal{S}, \mathcal{X}, \{\tilde{\mathbf{T}}^{(x)}\})$, where

$$\tilde{T}^{(x)}_{s's} = \frac{\pi_{s'} T^{(x)}_{ss'}}{\pi_s}$$

It is fairly straightforward to check that $\mathfrak{G}^R$ generates the reverse of the process that is generated by $\mathfrak{G}$ [50]. When the generator has multiple ergodic components, we can apply the above formula to each ergodic block of the transition matrices. Further, when the state space is infinite, particularly continuous, the reversal in the above equation can be accomplished with a Radon-Nikodym derivative.

Now consider a finite-dimensional linear OOM $\mathfrak{M} = (\mathcal{V}, \mathcal{X}, \{T^{(x)}\}, P)$. We shall define reversal here in a very different way. Let $\mathcal{V}^*$ be the *dual space* of $\mathcal{V}$. It is isomorphic to $\mathcal{V}$, and represents the space of all linear functions $\phi : \mathcal{V} \to \mathbb{R}$. Since it is finite dimensional, like $\mathcal{V}$, it can be expressed in coordinates; the element ϕ has coordinates $\boldsymbol{\phi} = (\phi_j)$ so that $\phi(v) = \sum_j \phi_j v_j$. For every operator S on $\mathcal{V}$, define the adjoint operator $S^\dagger : \mathcal{V}^* \to \mathcal{V}^*$ be defined by the formula $(S^\dagger \phi)_j = \sum_{j'} \phi_{j'} S_{j'j}$. (This way, $(S\phi)(v) = \phi(Sv)$.) Lastly, let π be the stationary state of T; then define the mapping $\tilde{P} : \mathcal{V}^* \to \mathbb{M}(\mathcal{X}^{\mathbb{N}})$ as

$$\Pr_{\tilde{P}\phi}(\, x_1 \dots x_\ell \,) = \left(T^{(x_\ell)\dagger} \dots T^{(x_1)\dagger} \phi \right)(\pi)$$

Then the OOM $\mathfrak{M}^R = (\mathcal{V}^*, \mathcal{X}, \{T^{(x)\dagger}\}, \tilde{P})$ is the reverse model of $\mathfrak{M}$, and generates the reverse process.

4.3. Memory and variety in physical generators

We have thus far dealt with models and generators in the very abstract; let us now make matters more concrete. Our primary interest is in using real, physical systems to simulate stochastic processes. The concept of *generator* which we have defined is obviously the best object for this sort of analysis; the states, described by the set $\mathcal{S}$, can be interpreted as a set of possible microstates

which a physical system may occupy at a given time, and the dynamics described by $T^{(x)}$ describe how the system's state is altered by the measurement of some macrostate as well as the normal passage of time—all, of course, under the limitations of classical statistical mechanics.

The question of *implementing* a generator, then, comes to the forefront. In order to produce a particular stochastic process in nature, we need a physical system in hand whose states and dynamics correspond to one of the many generators that can produce the desired process! This is always easier said than done. In particular, if the process is significantly complex—which we will characterize here as meaning that the process displays a wide variety of behaviors over the course of time—then it must draw its complexity, or rather variety, from somewhere. Specifically, variety in the process's behavior over time is accomplished by a generator which has a variety of states and is capable of exploring all of them via its dynamics.

The ability of a generator to capture the variety in a process's behavior is doubly important if we are using the generator with the intention of tracking its state in order to more accurately predict future outcomes. Performing this task optimally requires that the generator states correspond to the predictive states of the process. This is precisely the function of the ϵ-machine. Therefore, the memory required to implement the ϵ-machine is an important benchmark for the task of predicting process behaviors.

In Section 1.5.1, we discussed the idea of *variety* as a resource and a form of information. In the context of stochastic processes, we call this resource *memory*. Just as variety is defined for distributions over random variables, memory is defined in terms of distributions over state spaces of generators. For each (indecomposible) generator, the stationary distribution over states, $\boldsymbol{\pi}$, is the primary object which characterizes the potential information costs of implementing said generator physically. These costs can be quantified using entropies.

In this section we will forget, for the moment, the language of generators and focus entirely on distributions and their entropies. We will also extensively discuss *majorization*, which, as we discussed in Section 1.5.2, is an important tool in the resource theory of nonuniformity, whose costs are also formulated using entropies. For this reason majorization is also very useful in the discussion of memory.

4.3.1. Majorization as a tool for memory. First off, an overview of important relevant concepts from majorization and information theory is in order.

The majorization of positive vectors provides a qualitative description of how concentrated the quantity of a vector is over its components. For ease of comparison, consider vectors $\mathbf{p} = (p_i)$, $i \in \{1, \ldots, n\}$, whose components all sum to a constant value, which we take to be unity ($\sum_{i=1}^{n} p_i = 1$), and are nonnegative: $p_i \geq 0$. For our purposes, we interpret these vectors as probability distributions.

In Section 1.5.2 we first introduced relative majorization, and then majorization as a special case. We introduce majorization here on its own terms, following Ref. [**123**]. The historical definition of majorization is also the most intuitive, starting with the concept of a *transfer operation*.

A *transfer operation* $\mathbf{T}$ on a vector $\mathbf{p} = (p_i)$ selects two indices $i, j \in \{1, \ldots, n\}$, such that $p_i > p_j$, and transforms the components in the following way:

$$(Tp)_i := p_i - \epsilon$$

$$(Tp)_j := p_j + \epsilon \ ,$$

where $0 < \epsilon < p_i - p_j$, while leaving all other components equal; $(Tp)_k := p_k$ for $k \neq i, j$. Intuitively, these operations reduce concentration, since they act to equalize the disparity between two components, in such a way as to not create greater disparity in the opposite direction. This is the *principle of transfers*.

Suppose now that we have two vectors $\mathbf{p} = (p_i)$ and $\mathbf{q} = (q_i)$ and that there exists a sequence of transfer operations $\mathbf{T}_1, \ldots, \mathbf{T}_m$ such that $\mathbf{T}_m \circ \cdots \circ \mathbf{T}_1 \mathbf{p} = \mathbf{q}$. We will say that $\mathbf{p}$ *majorizes* $\mathbf{q}$; denoted $\mathbf{p} \succsim \mathbf{q}$. The relation $\succsim$ defines a *preorder* on the set of distributions, as it is reflexive and transitive but not necessarily antisymmetric.

There are, in fact, a number of equivalent criteria for majorization. We list three relevant to our development in the following composite theorem.

THEOREM 11 (Majorization Criteria). *Given two vectors* $\mathbf{p} := (p_i)$ *and* $\mathbf{q} := (q_i)$ *with the same total sum, let their orderings be given by the permuted vectors* $\mathbf{p}^{\downarrow} := (p_i^{\downarrow})$ *and* $\mathbf{q}^{\downarrow} := (q_i^{\downarrow})$ *such that* $p_1^{\downarrow} > p_2^{\downarrow} > \cdots > p_n^{\downarrow}$ *and the same for* $\mathbf{q}^{\downarrow}$. *Then the following statements are equivalent:*

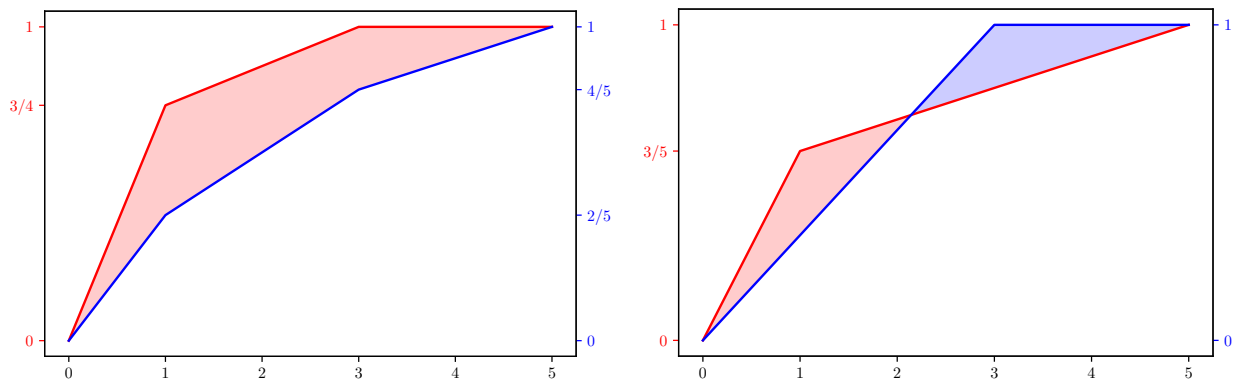

FIGURE 4.1. (*Left*) Lorenz curves when **p** and **q** are comparable and the first majorizes the second: $\mathbf{p} \succsim \mathbf{q}$. Here, we chose $\mathbf{p} = (3/4, 1/8, 1/8, 0, 0)$ and $\mathbf{q} = (2/5, 1/5, 1/5, 1/10, 1/10)$. Tick marks indicate kinks in the Lorenz curve. (*Right*)Lorenz curves when **p** and **q** are incomparable. Here, we chose $\mathbf{p} = (3/5, 1/10, 1/10, 1/10, 1/10)$ and $\mathbf{q} = (1/3, 1/3, 1/3, 0, 0)$.

(1) Hardy-Littlewood-Pólya: *For every* $1 \leq k \leq n$,

$$\sum_{i=1}^{k} p_i^{\downarrow} \geq \sum_{i=1}^{k} q_i^{\downarrow} \ ;$$

(2) Principle of transfers: $\mathbf{p}$ *can be transformed to* $\mathbf{q}$ *via a sequence of transfer operations;*

(3) Schur-Horn: *There exists a unitary matrix* $\mathbf{U} := (U_{ij})$ *such that* $\mathbf{q} = \mathbf{D}\mathbf{p}$, *where* $\mathbf{D} := \left(|U_{ij}|^2\right)$, *a* uni-stochastic *matrix.*

The Hardly-Littlewood-Pólya criterion provides a visual representation of majorization in the form of the *Lorenz curve*. For a distribution $\mathbf{p} := (p_i)$, the Lorenz curve is simply the function $\beta_{\mathbf{p}}(k) := \sum_{i=1}^{k} p_i^{\downarrow}$. See Fig. 4.1. We can see that $\mathbf{p} \succsim \mathbf{q}$ so long as the area under $\beta_{\mathbf{q}}$ is completely contained in the area under $\beta_{\mathbf{p}}$.

The Lorenz curve can be understood via a social analogy, by examining rhetoric of the form "The top x% of the population owns y% of the wealth". Let y be a function of x in this statement, and we have the Lorenz curve of a wealth distribution. (Majorization, in fact, has its origins in the study of income inequality.)

If neither **p** nor **q** majorizes the other, they are *incomparable*. (See Fig. 4.1.)

It is worthwhile to note an ambiguity when comparing distributions defined over different numbers of elements. There are generally two standards for such comparisons that depend on application.

In the resource theory of nonuniformity [64], one compares distributions over different numbers of events by "squashing" their Lorenz curves so that the x-axis ranges from 0 to 1. Under this comparison, the distribution $\mathbf{p}_3 = (1, 0, 0)$ has more informational nonequilibrium than $\mathbf{p}_2 = (1, 0)$. In the following, however, we adopt the standard of simply extending the smaller distribution by adding events of zero probability. In this case, $\mathbf{p}_3$ and $\mathbf{p}_2$ are considered equivalent. This choice is driven by our interest in the Rényi entropy costs and not in the overall nonequilibrium. (The latter is more naturally measured by Rényi *negentropies* $\bar{H}_\alpha(\mathbf{p}) = \log n - H_\alpha(\mathbf{p})$, where n is the number of events.)

Now, as noted, majorization is a preorder, since there may exist distinct $\mathbf{p}$ and $\mathbf{q}$ such that $\mathbf{p} \succsim \mathbf{q}$ and $\mathbf{q} \succsim \mathbf{p}$. This defines an equivalence relation $\sim$ between distributions. It can be checked that $\mathbf{q} \succsim \mathbf{p}$ if and only if the two vectors are related by a permutation matrix $\mathbf{P}$. Every preorder can be converted into a partial order by considering equivalence classes $[\mathbf{p}]_\sim$.

If majorization, in fact, captures important physical properties of the distributions, we should expect that these properties may be quantified. The class of monotones that quantify the preorder of majorization are called *Schur-convex* and *Schur-concave* functions. A function $f : \mathbb{R}^n \to \mathbb{R}$ is called Schur-convex (-concave) if $\mathbf{p} \succsim \mathbf{q}$ implies $f(\mathbf{p}) \geq f(\mathbf{q})$ ($f(\mathbf{p}) \leq f(\mathbf{q})$). f is strictly Schur-convex (concave) if $\mathbf{p} \succsim \mathbf{q}$ and $f(\mathbf{p}) = f(\mathbf{q})$ implies $\mathbf{p} \sim \mathbf{q}$.

An important class of Schur-concave functions consists of the Rényi entropies:

$$\mathrm{H}_\alpha[\mathbf{p}] := \frac{1}{1-\alpha} \log_2 \left(\sum_{i=1}^{n} p_i^\alpha \right) .$$

In particular, the three limits:

$$\mathrm{H}[\mathbf{p}] := \lim_{\alpha \to 1} \mathrm{H}_\alpha[\mathbf{p}] = -\sum_{i=1}^{n} p_i \log_2 p_i \ ,$$

$$\mathrm{H}_{\max}[\mathbf{p}] := \lim_{\alpha \to 0} \mathrm{H}_\alpha[\mathbf{p}] = \log_2 |\{1 \leq i \leq n : p_i > 0\}| \ , \text{ and}$$

$$\mathrm{H}_{\min}[\mathbf{p}] := \lim_{\alpha \to \infty} \mathrm{H}_\alpha[\mathbf{p}] = -\log_2 \max_{1 \leq i \leq n} p_i$$

—*Shannon* entropy, *topological* entropy, and *min*-entropy, respectively—describe important practical features of a distribution. In order, they describe (i) the asymptotic rate at which the outcomes can be accurately conveyed, (ii) the single-shot resource requirements for the same task, and (iii) the

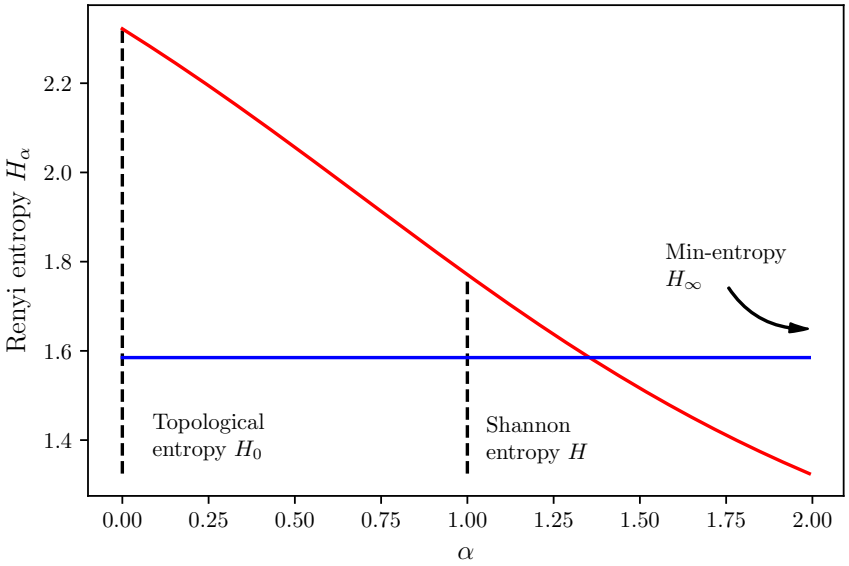

FIGURE 4.2. Rényi entropies of the two incomparable distributions **p** and **q** from Fig. 4.1.

probability of error in guessing the outcome if no information is conveyed at all (or, alternatively, the single-shot rate at which randomness can be extracted from the distribution) [**158**, **194**]. As such, they play a significant role in communication and memory storage.

We note that the Rényi entropies for $0 < \alpha < \infty$ are *strictly* concave.

The example of two incomparable distributions $\mathbf{p}$ and $\mathbf{q}$ can be analyzed in terms of the Rényi entropies if we plot $\mathrm{H}_\alpha[\mathbf{p}]$ and $\mathrm{H}_\alpha[\mathbf{q}]$ as a function of α, as in Fig. 4.2.

4.3.2. Strong and weak optimization. The central idea explored in the following is how majorization may be used to determine when it is possible to simultaneously optimize all entropy monotones—or, alternatively, to determine if each monotone has a unique extremum. Obviously, this distinction is a highly practical one to make when possible. This leads to defining *strong maxima* and *strong minima*. Let S be a set of probability distributions. If a distribution $\mathbf{p} \in S$ satisfies $\mathbf{p} \precsim \mathbf{q}$ ($\mathbf{p} \succsim \mathbf{q}$), for all $\mathbf{q} \in S$, then $\mathbf{p}$ is a *strong maximum* (*minimum*) of the set S.

The extrema names derive from the fact that the strong maximum maximizes the Rényi entropies and the strong minimum minimizes them. One can extend the definitions to the case where $\mathbf{p} \notin S$, but is the least-upper-bound such that any other $\mathbf{p}'$ satisfying $\mathbf{p}' \precsim \mathbf{q}$ must obey $\mathbf{p}' \precsim \mathbf{p}$. This case

would be called a *strong supremum* (or in the other direction a *strong infimum*). However, these constructions may not be unique as $\succsim$ is a preorder and not a partial order. However, if we sort by equivalence class, then the strongly maximal (minimal) class is unique if it exists.

One example of strong minimization is found in quantum mechanics. Let ρ be a density matrix and X be a maximal diagonalizing measurement. For a given measurement Y, let $\rho|_Y$ be the corresponding probability distribution that comes from measuring ρ with Y. Then $\rho|_X \succsim \rho|_Y$ for all maximal projective measurements Y. (This follows from the unitary matrices that transform from the basis of X to that of Y and the Schur-Horn lemma.)

Another, recent example is found in Ref. [72], where the set $B_\epsilon(\mathbf{p})$ of all distributions ϵ-close to $\mathbf{p}$ under the total variation distance δ is considered:

$$B_\epsilon(\mathbf{p}) := \{\mathbf{q} : \delta(\mathbf{p}, \mathbf{q}) \leq \epsilon\} \ .$$

This set has a strong minimum, called the *steepest distribution* $\overline{\mathbf{p}}_\epsilon$, and a strong maximum, called the *flattest distribution* $\underline{\mathbf{p}}_\epsilon$.

When a strong minimum or maximum does not exist, we refer to the individual extrema of the various monotones as *weak* extrema.

4.4. Classical generators and hidden Markov models

Recall that a (finite) generator of a process μ is defined as a triple $\mathfrak{G} = (\mathcal{S}, \mathcal{X}, \left\{ \mathbf{T}^{(x)} \right\})$ consisting of a finite set of hidden states $\mathcal{S}$, an alphabet $\mathcal{X}$, and a set of transition maps $\mathbf{T}^{(x)}$ which jointly encode the probability of producing a given symbol and then transitioning to another state. When $\mathcal{S}$ is finite, this is, incidentally, also the definition of a hidden Markov model (HMM).

If $\mathcal{S}$ has size $|\mathcal{S}|$ then the generator $\mathfrak{G}$ also corresponds to the linear OOM $(\mathbb{R}^{|\mathcal{S}|}, \mathcal{X}, \left\{ T^{(x)} \right\}, P)$ where $T^{(x)}$ are the linear operators corresponding to the conditional matrices $\mathbf{T}^{(x)}$ and P is given by

$$\Pr_{Pv}(\, x_1 \ldots x_\ell \,) = \mathbf{1}^\top \mathbf{T}^{(x_\ell)} \ldots \mathbf{T}^{(x_1)} \mathbf{v}$$

In Section 4.2.2 we discussed the conditions for the existence of unique stationary states in linear OOMs. These results were a sort of extension of the Perron-Frobenius theorem, utilizing the ergodic nature of the predictive states. If $\mathbf{T} = \sum_x \mathbf{T}^{(x)}$ is an irreducible matrix, we can directly apply the

Perron-Frobenius theorem, which also guarantees a unique stationary state, $\mathbf{T}\boldsymbol{\pi} = \boldsymbol{\pi}$. The *process generated by* $\mathfrak{G}$ is defined by the word probabilities

$$\Pr_{\mathfrak{G}}(x_1 \dots x_\ell) = \mathbf{1}^\top \mathbf{T}^{(x_\ell)} \dots \mathbf{T}^{(x_1)} \boldsymbol{\pi}$$

Note that this both defines a measure over $\mathcal{X}^{\mathbb{N}}$ but also a (stationary, ergodic) measure over $\mathcal{X}^{\mathbb{Z}}$, due to the stationarity of $\boldsymbol{\pi}$.

The Embedding Lemma (Lemma 1) implies that there is a unique mapping, $E_{\mathfrak{G}} : \mathcal{X}^{\mathbb{N}} \to \mathbb{R}^{|\mathcal{S}|}$, which maps each past to a "mixed state," that is, a probability distribution over $\mathcal{S}$. This mixed state $E_{\mathfrak{G}}(\overleftarrow{x})$ represents the asymptotic result of starting from any generator state and observing a long sequence of symbols $x_{-\ell} \dots x_0$ which matches the past $\overleftarrow{x}$. Since $E_{\mathfrak{G}}(\overleftarrow{x})$ is defined as a vector, we will denote it as a probability using the notation $\Pr_{\mathfrak{G}}(s \mid \overleftarrow{x}) = [E_{\mathfrak{G}}(\overleftarrow{x})]_s$. We have, to some extent, discussed mixed states before, in Section 7. Even for finite measures, the set of mixed states which can result may be uncountably infinite and have complex fractal structures. Mixed states over generators were first considered in [**199**], and interest in them has recently been re-ignited due to their usefulness in studying the behavior of hidden Markov models [**86**, **87**, **88**, **89**].

4.4.1. Moving through the space of generators. Let us now generalize a concept which has appeared more than once so far, and will appear again later. We have noted that OOMs are "homomorphisms" of the predictive state dynamics, in the sense of Eq. (4.3). We have also encountered a similar notion in the relationship between the embedding dynamics and the natural dynamics as pasts transform into other pasts: Eq. (4.4). There is a general relation that can exist between generators which matches this intuition. Let $\mathfrak{F} = (\mathcal{R}, \mathcal{X}, \left\{ \mathbf{M}^{(x)} \right\})$ and $\mathfrak{G} = (\mathcal{S}, \mathcal{X}, \left\{ \mathbf{T}^{(x)} \right\})$ be two generators of the same process. We say that $\mathfrak{G}$ embeds $\mathfrak{F}$, written $\mathfrak{G} \succsim \mathfrak{F}$, if there is a stochastic matrix $\mathbf{P} = (P_{s|r})$ such that

$$\mathbf{P}\mathbf{M}^{(x)} = \mathbf{T}^{(x)}\mathbf{P} \tag{4.6}$$

for all $x \in \mathcal{X}$. The idea of the relation $\succsim$ is that each state in $\mathfrak{F}$ corresponds directly to a mixed state in $\mathfrak{G}$, in a manner which commutes with the dynamics. The embedding relation $\succsim$ between

generators should not be confused with majorization between distributions. There is, unfortunately, such a lack of good ordering symbols.

Let us define two particular "landmark" generators (which are *not* finite) which will help us illustrate this point. The *pasts generator* is the generator $\mathfrak{P}_\mu = (\mathcal{X}^{\mathbb{N}}, \mathcal{X}, \sigma_F)$ where $\sigma_F : \mathcal{X}^{\mathbb{N}} \to \mathbb{M}(\mathcal{X} \times \mathcal{X}^{\mathbb{N}})$ is the pushforward on pasts, which for each past $\overleftarrow{x}$ generates x and maps to $\overleftarrow{x}x$ with probability $\Pr_\mu(\, x \mid \overleftarrow{x}\,)$. The *futures generator* is the generator $\mathfrak{R}_\mu = (\mathcal{X}^{\mathbb{N}}, \mathcal{X}, \sigma_R)$ where $\sigma_R : \mathcal{X}^{\mathbb{N}} \to \mathbb{M}(\mathcal{X} \times \mathcal{X}^{\mathbb{N}})$ is the pullback on futures, which for each future $\overrightarrow{x} = x_1 x_2 \ldots$ generates x_1 with with certainty and then transitions to $\overrightarrow{x}' = x_2 x_3 \ldots$.

The reason these are "landmark" generators is that, due to the Embedding Lemma, every generator $\mathfrak{G}$ of a process μ embeds its corresponding pasts generator $\mathfrak{P}_\mu$; further, due to Eq. (4.3), the futures generator $\mathfrak{R}_\mu$ embeds every generator $\mathfrak{G}$. In other words,

$$\mathfrak{P}_\mu \succsim \mathfrak{G} \succsim \mathfrak{R}_\mu \tag{4.7}$$

for every generator $\mathfrak{G}$ of μ.

Two other "landmarks," whose importance will be seen shortly, are the (forward, or predictive) ϵ-machine and the reverse (or retrodictive) ϵ-machine. (When we refer to the ϵ-machine alone, we will mean the forward variety.) The forward ϵ-machine we have already defined as the generator $\mathfrak{E}_\mu = (\mathcal{K}(\mu), \mathcal{X}, \{T^{(x)}\})$ with the property that $T^{(x)}(\epsilon[\overleftarrow{x}])$ produces a scaled Dirac delta measure $\Pr_\mu(\, x \mid \overleftarrow{x}\,)\, \delta_{\epsilon[\overleftarrow{x}x]}$. The ϵ-machine is, essentially, the generator which directly arises from the predictive states [**176**].

Conversely, the reverse ϵ-machine arises from the retrodictive states; it is the generator $\mathfrak{E}^R_\mu = (\mathcal{K}^R(\mu), \mathcal{X}, \{\tilde{T}^{(x)}\})$ with the property that $\tilde{T}^{(x)}(\epsilon^R[\overrightarrow{x}])$ works by randomly selecting a future $\overrightarrow{y} = y_1 y_2 \ldots$ with $\epsilon^R[\overrightarrow{y}] = \epsilon^R[\overrightarrow{x}]$ and then generating the symbol $x = y_1$ and transitioning to $\epsilon^R[y_2 y_3 \ldots]$.

The complexity of a finite generator $\mathfrak{G}$ can be characterized by the entropies of its stationary distribution: $\mathrm{H}_{\min}[\pi]$, $\mathrm{H}_{1/2}[\pi]$, $\mathrm{H}[\pi]$, and all the Rényi entropies $\mathrm{H}_\alpha[\pi]$ in between. As mentioned previously, these characterize the memory costs imposed on any physical system which implements the generator.

4.4.2. Unifilarity and prediction. We will now focus our attention on generators which can be used to track the behavior of a process and predict its future behavior. We will define some closely related properties of process which help to this end.

First, let us reintroduce the concept of a random variable, which is represented by a capital letter and which takes values whose probabilities are described by some distribution. For a generator $\mathfrak{G} = (\mathcal{S}, \mathcal{X}, \left\{ \mathbf{T}^{(x)} \right\})$, for instance, its state at any given time is represented by the random variable S, which is distributed according to the stationary state, $S \sim \pi$. For the pasts and futures generators, $\mathfrak{P}_\mu$ and $\mathfrak{G}_\mu$, the state random variables are denoted $\overleftarrow{X}$ and $\overrightarrow{X}$ respectively, each distributed as $\overleftarrow{X} \sim \overleftarrow{\mu}$ and $\overrightarrow{X} \sim \overrightarrow{\mu}$.

Given three joint random variables XYZ distributed as $\Pr_{XYZ}(\, xyz \,)$, we say that Y is Markov for X and Z, denoted $X - Y - Z$, whenever

$$\Pr_{XYZ}(\, x, y, z \,) = \Pr_X(\, x \,) \Pr_{Y|X}(\, y \mid x \,) \Pr_{Z|Y}(\, z \mid y \,) \tag{4.8}$$

The Markov chain relation indicates that Y is "sufficient" for knowing Z, in the sense that also knowing X provides no additional information on the distribution of Z.

The embedding rule introduced for generators Eq. (4.7) tells us that, given a value of the past $\overleftarrow{X} = \overleftarrow{x}$, the distribution for the generator state and the subsequent future must be given by the rule $\Pr_{S\overrightarrow{X}}\left(S, \overrightarrow{X} \,\middle|\, \overleftarrow{x} \right) = \Pr_{\mathfrak{G}}(\, S \mid \overleftarrow{x} \,) \Pr_{\mathfrak{G}}\left(\overrightarrow{X} \,\middle|\, S \right)$. (Here we are playing fast and loose with notations—obviously $\Pr\left(\overrightarrow{X} \right)$ is not a well defined function and must be understood as a measure—but we hope by now we have earned the reader's trust and they are willing to play along.) It is therefore a consequence of the rule Eq. (4.7) that $\overleftarrow{X} - S - \overrightarrow{X}$ for any generator state S. Every generator state is Markov between the past and the future.

Now we shall define a *predictive generator*. These are any generator which satisfies the rule $S - \overleftarrow{X} - \overrightarrow{X}$. To understand this rule we must think about the generator state as being an *encoding* of the past, which contains the information in the past which is useful for understanding the future. This is a consequence of the rule $\overleftarrow{X} - S - \overrightarrow{X}$. However, that rule also allows that the state S also represents *additional* information, which constrains the outcomes in the future beyond what is required by knowledge of the past [**161**].

This "oracular" knowledge may appear useful at first glance—but it is a mirage. If we are using the generator as a kind of model, tracking the behavior of an existing process and attempting to guess its internal state, the oracular dimensions of the state can never be known to us because, by definition, they are not dependent on the past $\overleftarrow{X}$. These dimensions of the generator state are, from a prediction standpoint, merely wasteful. For this reason a predictive generator is defined as one where $S - \overleftarrow{X} - \overrightarrow{X}$, for this means that S cannot hold any information about the future which is not already contained in the past.

This definition, though now ideologically justified, is too abstract to be constructive. In the next section we'll introduce an equivalent characterization of predictive models, which is much more constructive, but we will need intermediate definitions.

To that end, let us define another concept, which is much more concrete and essentially graph-theoretic. A generator $\mathfrak{G} = (\mathcal{S}, \mathcal{X}, \left\{ \mathbf{T}^{(x)} \right\})$ is said to be *unifilar* (read: Latin for "single-threaded") if the transition matrices $\mathbf{T}^{(x)} = (T^{(x)}_{s's})$ have the form $T^{(x)}_{s's} = \Pr(\, x \mid s \,)\, \delta_{s',f(x,s)}$ where $f : \mathcal{X} \times \mathcal{S} \to \mathcal{S}$ is a deterministic function. A unifilar generator, in random variable terms, is one where the second state is simply a function of the previous state and generated symbol: $S' = f(X, S)$. If we know the initial state of a unifilar generator, then by tracking its output over time we can know with certainty its state at *any future time* (as opposed to non-unifilar generators, whose states may follow "many threads").

A key example of a unifilar generator is the ϵ-machine. The ϵ-machine is also predictive—because its state is defined as a function of the past, $S = f(\overleftarrow{X})$, it must be the case that $S - \overleftarrow{X} - \overrightarrow{X}$.

Unifilar generators provide a concretization of the intuition of predictive generators. After all, the state at any time is just a function of all the emitted symbols (and some distant past state). It is indeed the case that every unifilar generator is predictive (which will be demonstrated in the next section). It is, however, not the case that every predictive generator is unifilar.

Before moving on, let us define the time-reversed equivalents of these concepts. A generator is *retrodictive* if it obeys $\overleftarrow{X} - \overrightarrow{X} - S$, and it is called *co-unifilar* if the *previous* state is a function of the observed symbol and the future state: $S = f(X, S')$. As with the forward ϵ-machine, the reverse ϵ-machine is both co-unifilar and retrodictive.

It may seem futile to define such generators—we have just discussed why they are not optimal in the prediction setting—but prediction is not the only reason to utilize generators. We will see in Chapter 6 that retrodictive generators are particularly thermodynamically advantageous.

4.4.3. State-merging generators. Since memory is a cost associated with generators, it is potentially desireable to reduce those costs with minimal loss of useful information. We recall from 1.5.1 that the entropies of a random variable X are always lower for any function $F(X)$. Let us now leverage this fact to define a universally memory-reducing operation on generators.

Given a generator $\mathfrak{G} = (\mathcal{S}, \mathcal{X}, \left\{ \mathbf{T}^{(x)} \right\})$ and a function $F : \mathcal{S} \to \mathcal{R}$ for some function F to a new state space $\mathcal{R}$, we define the *merged generator* as $\mathfrak{G}|_F = (\mathcal{R}, \mathcal{X}, \left\{ \hat{\mathbf{T}}^{(x)} \right\})$, where

$$\hat{T}^{(x)}_{r'r} = \sum_{\substack{s' \in F^{-1}(r) \\ s \in F^{-1}(r)}} T^{(x)}_{s's} \frac{\pi_s}{\hat{\pi}_r}$$

and $\hat{\pi}_r = \sum_{s \in F^{-1}(r)} \pi_s$. We say that $\mathfrak{G}$ is *mergeable under* F if $\mathfrak{G}|_F$ generates the same process as $\mathfrak{G}$.

An important form of merging is via the *predictive equivalence class*. Two states $s, s' \in \mathcal{S}$ of a generator $\mathfrak{G}$ are *predictively equivalent*, written $s \sim_P s'$, if $\Pr_{\mathfrak{G}}(\cdot \mid s) = \Pr_{\mathfrak{G}}(\cdot \mid s')$. The predictive equivalence map $[\cdot]_P$ maps states s to their equivalence class $[s]_P$. The predictively merged generator $\mathfrak{G}|_P$ is the result of merging states under the predictive equivalence relation.

Proposition 8. *Any generator $\mathfrak{G} = (\mathcal{S}, \mathcal{X}, \{T^{(x)}_{s's}\})$ is mergeable under the predictive equivalence relation $[\cdot]_\sim$.*

Proof. *We will prove the forward (predictive) case; the retrodictive case follows from time reversal. The proof proceeds by induction. Let us suppose that, for words w of length ℓ, $\Pr_{\mathfrak{G}|_P}(w \mid r) = \Pr_{\mathfrak{G}}(w \mid s)$ for all $s \in r$ (here r represents an equivalence class). Then*

$$\begin{aligned}
\Pr_{\mathfrak{G}}(xw \mid s) &= \sum_{s'} \Pr_{\mathfrak{G}}(w \mid s') T^{(x_0)}_{s's} \\
&= \sum_{r'} \Pr_{\mathfrak{G}|_P}(w \mid r') \left(\sum_{s' \in r'} T^{(x_0)}_{s's} \right)
\end{aligned}$$

We can average over all $\hat{s}$ in the given equivalence class $[s]_P$, which leaves the left-hand side unchanged (by the definition of $[s]_P$):

$$\begin{aligned}
\Pr_{\mathfrak{G}}(xw \mid s) &= \sum_{r'} \Pr_{\mathfrak{G}|_P}(w \mid r') \left(\sum_{\substack{s' \in r' \\ \hat{s} \sim_P s}} T^{(x)}_{s'\hat{s}} \frac{\pi_{\hat{s}}}{\hat{\pi}_{[s]_P}} \right) \\
&= \sum_{r'} \Pr_{\mathfrak{G}|_P}(w \mid r') \hat{T}^{(x)}_{r'[s]_P} \\
&= \Pr_{\mathfrak{G}|_P}(w \mid [s]_P)
\end{aligned}$$

So, if "$\Pr_{\mathfrak{G}|_P}(w \mid r) = \Pr_{\mathfrak{G}}(w \mid s)$ for all $s \in r$" is true for all words of length ℓ, it is also true for words of length $\ell + 1$.

But this is easily seen to be true for $\ell = 1$: this is evident simply from the definition of $\hat{T}^{(x)}_{r'r}$, as summing over r' gives the length-1 conditional probabilities.

So, by induction it is true for all words w that $\Pr_{\mathfrak{G}|_P}(w \mid r) = \Pr_{\mathfrak{G}}(w \mid s)$ for all $s \in r$. When we average over the stationary state on both sides, this gives $\Pr_{\mathfrak{G}|_P}(w) = \Pr_{\mathfrak{G}}(w)$.

Before going on, let us fix intuitions, considering several example generators which can be predictively merged. We will use majorization as a visualization to see how this process reduces memory.

First, consider the Biased Coin Process, a memoryless process (in the sense that it has only one state) in which, at each time step, a coin is flipped with probability p of generating a 1 and probability $1 - p$ of generating a 0. Figure 4.3 displays three models for it. Model (a) is the ϵ-machine of the process, and models (b) and (c) are each 2-state uniflar generators. Notice that in both models (b) and (c), the two states are predictively equivalent.

Continuing, Fig. 4.4 displays two alternative models of the Even-Odd Process. This process is uniformly random save for the constraint that 1s appear only in blocks of even number and 0s only in blocks of odd number. We see in Fig. 4.4(a) the process' ϵ-machine. In Fig. 4.4(b), we see an alternative unifilar generator. Notice that its states E and F predict the same futures and so are not probabilistically distinct. They both play the role of state C in the ϵ-machine, in terms of the futures they predict.

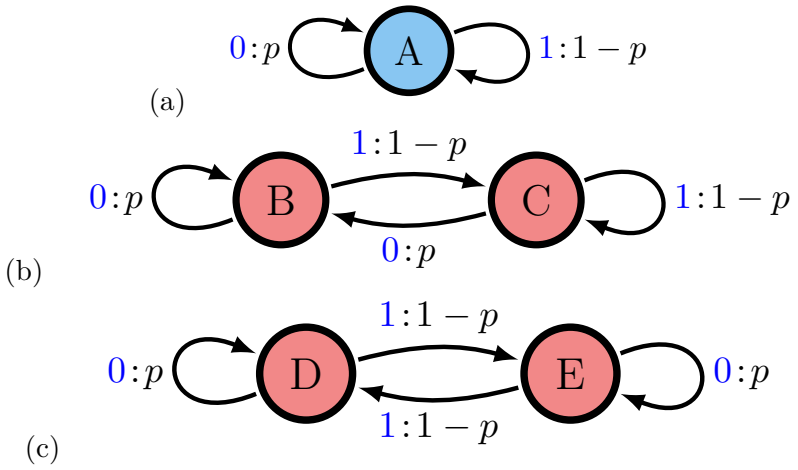

FIGURE 4.3. The diagrammatic form of a FSM is read as follows. The colored circles represent hidden states from the finite set $\mathcal{R}$. The edges are labeled by a blue number, the symbol x, and a probability p. The edges with symbol x represent the transition matrix $\mathbf{T}^{(x)} := (T^{(x)}_{r'|r})$, where the tail of the arrow is the starting state r, the head is the final state r', and $p = T^{(x)}_{r'|r}$. (a) ϵ-Machine for a coin flipped with bias p. (b) Alternate representation with bias p to be in state B and $1-p$ to be in state C. (c) Alternate representation with biases p to stay in current state and $1-p$ to switch states.

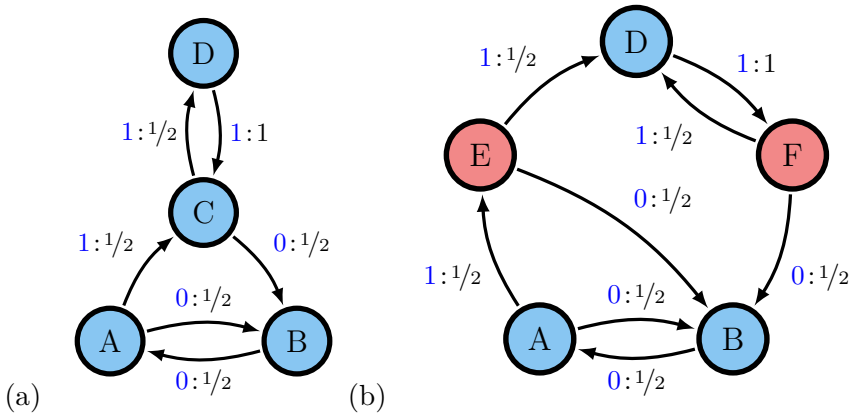

FIGURE 4.4. (a) ϵ-Machine for Even-Odd Process. (b) Refinement of the Even-Odd Process ϵ-machine, where the ϵ-machine's state C has been split into states E and F.

Majorization, and Lorenz curves in particular, allow us to compare the various models for each of these processes—see Fig. 4.5. We notice that the ϵ-machine state distribution always majorizes the state distribution of the alternative machines.

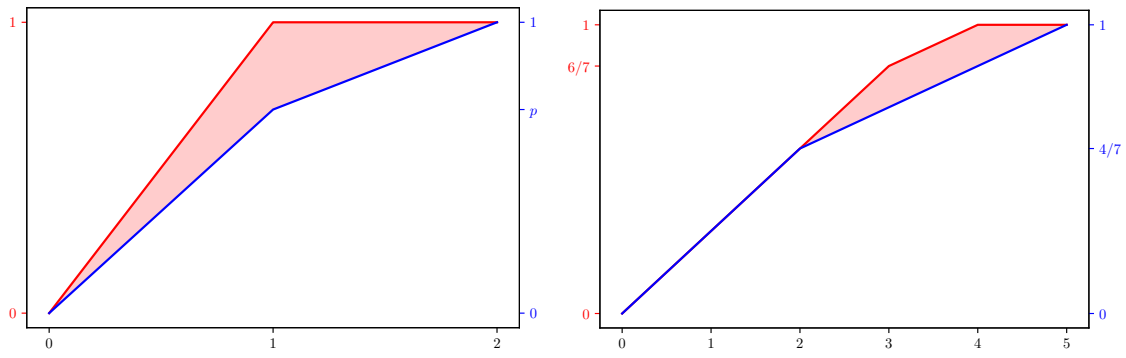

FIGURE 4.5. (*Left*) Lorenz curves for Fig. 4.3**(a)**'s ϵ-machine and Fig. 4.3**(b)**'s alternative predictor of the Biased Coin Process. (*Right*) Same comparison for the Even-Odd Process ϵ-machine Fig. 4.4**(a)** and alternative predictor Fig. 4.4**(b)**.

4.4.4. Minimality of the ϵ-machine. We will close this section on classical generators by demonstrating that predictively merging any predictive generator always results in the ϵ-machine. This provides a useful characterization of predictive generators—the are, in essence, "state-splittings" of the ϵ-machine. Significantly, this means that the stationary state of the ϵ-machine majorizes the state distributions of all other predictive generators: put otherwise, the ϵ-machine *strongly minimizes* the set of all predictive generators, attaining the minimum for every possible entropy.

We start by stating a core theorem regarding finite-state generators and the central role the ϵ-machine plays among them [**197**].

THEOREM 12 (Uniqueness of the ϵ-machine). *Any generator* $\mathfrak{G} = (\mathcal{S}, \mathcal{X}, \left\{ \mathbf{T}^{(x)} \right\})$ *for which*

(1) $\mathbf{T}$ *is irreducible,*

(2) the transitions are unifilar*, and*

(3) the states are predictively distinct ($s \sim_P s'$ *implies* $s = s'$)

is isomorphic to the ϵ-machine.

Keeping this result in mind we will now turn to giving a new characterization of what it means for a process to be predictive.

PROPOSITION 9. *Any generator* $\mathfrak{G} = (\mathcal{S}, \mathcal{X}, \left\{ \mathbf{T}^{(x)} \right\})$ *whose predictively state-merged generator* $\mathfrak{G}|_P$ *is unifilar is also a predictive generator; the converse also holds.*

PROOF. *If $\mathfrak{G}|_P$ is unifilar, then it must be the ϵ-machine, because by definition it will have predictively distinct states. Now, let us denote the predictive state by the random variable Σ and the generator state by S. Since the ϵ-machine is predictive, $\Sigma - \overleftarrow{X} - \overrightarrow{X}$ is Markov, but also we have $S - \Sigma - \overrightarrow{X}$ because of the state-merging. These two together mean that $S - \overleftarrow{X} - \overrightarrow{X}$.*

Now, consider the converse. If $S - \overleftarrow{X} - \overrightarrow{X}$ then it must be the case that the embedded mixed state $\Pr_{\mathfrak{G}}\left(S \,\middle|\, \overleftarrow{X}\right)$ can only take non-zero values over predictively equivalent states. Since embedded pasts must map to other embedded pasts under the generator dynamics (a consequence of $\mathfrak{P}_\mu \precsim \mathfrak{G}$), generator states in a given equivalent block can only map into another equivalent block after a symbol x is observed. This means that, after the predictively equivalent blocks are merged, the dynamics will be unifilar.

Thus, the set of all predictive generators is actually the set of all generators which predictively merge to unifilar ones, and specifically, which merge to the ϵ-machine. (The equivalent statement for retrodictors is obtained by time-reversal: the reverse ϵ-machine is the unique recurrent, retrodictively minimal counifilar generator, and all retrodictors merge to it.)

We therefore have the result:

COROLLARY 4. *The ϵ-machine $\mathfrak{E}(\mu)$ strongly minimizes the set of all predictive generators for μ; if Σ represents the ϵ-machine state, $\mathrm{H}_\alpha[\Sigma] \leq \mathrm{H}_\alpha[S]$ where S is the state of any other predictive generator of μ.*

This is a direct consequence of the fact that the ϵ-machine arises from a state-merging of every predictive generator of μ.

Because of the universal optimality of the ϵ-machine among predictive machines, it is considered a benchmark for the memory costs of predicting the process μ. To that end, the *statistical complexity* of the process μ is defined as

$$C_\mu = \mathrm{H}[\Sigma] \tag{4.9}$$

and similarly for the Rényi statistical complexities $C_\mu^{(\alpha)} = \mathrm{H}_\alpha[\Sigma]$.

4.5. Discussion

In this chapter we synthesized much of the past literature on models of stochastic processes together with new results in a manner which allow the application of resource theoretic methods. In his seminal work *Reflections on the Motive Power of Fire*, Sadi Carnot expressed the importance of discovering laws which apply "not only to steam engines but to *all imaginable* heat-engines." In this chapter we established rules, such as the Embedding Lemma 1, which constrain the internal mechanics of all imaginable models and classical generators of stochastic processes. We demonstrated one particular use of these principles, strengthening previous results on the fundamental memory constraints of the tasks of prediction and retrodiction of processes, and identified in each case a single generator (the ϵ-machine and the reverse ϵ-machine) which achieves universal memory minimality. But all imaginable means *all imaginable*, and in the day and age of quantum computing it is not sufficient to stop at the boundary of classical possibility. In Chapter 5 we will extend our results on models to describe quantum models and generators, and we will assess the extent to which our results on memory for classical generators can be extended to the quantum setting. Classical and quantum generators will both feature heavily in Chapter 6 as we examine the *thermodynamic* aspects of their physical implementations, and determine the nature of the tradeoff between energy savings and memory compression. In both of the following chapters, results such as the Embedding Lemma 1 and our definition of the embedding relation $\succsim$ between generators will provide crucial insights to the nature of resource costs in models.

The Embedding Lemma, and more importantly the definition of a model as a homomorphism, provided by Eq. (4.3), was in some sense anticipated by the field of cybernetics, for instance in the work *Every good regulator of a system must be a model of that system* by Conant and Ashby. There they defined a model of a process in terms of group homomorphisms while also appealing to the intuitive picture of a model as any system which contains a "scale model" of another system; in comparison, we have defined models in the chapter as semigroup homomorphisms, and used this fact in conjunction with the topological and measure-theoretic results of Chapter 2 to demonstrate that such models do indeed contain an embedded "scale model" of the predictive states.

The dynamics of this scale model place strict algebraic constraints on the spectra and fixed points of the overarching model, and it is on this basis that we are able to discuss *memory* in terms of

the entropies of stationary distributions over generator states. Physical constraints on the channels which drive the evolution of the scale model will be the subject of Chapter 6 and will be the key to determining which *kinds* of models are able to achieve thermodynamic optimality.

CHAPTER 5

All imaginable: Quantum generators of processes

There are no answers, only cross-references.

Norbert Wiener's "Law of Libraries"

5.1. Introduction

Recently, Google AI announced a breakthrough in quantum supremacy, using a 54-qubit processor ("Sycamore") to complete a target computation in 200 seconds, claiming the world's fastest supercomputer would take more than 10,000 years to perform a similar computation [12]. Shortly afterward, IBM announced that they had proven the Sycamore circuit could be successfully simulated on the Summit supercomputer, leveraging its 250 PB storage and 200 petaFLOPS speed to complete the target computation in a matter of days [146]. This episode highlights two important aspects of quantum computing: first, the importance of memory and, second, the subtle relationship between computation and simulation.

When simulating classical processes, quantum implementations can be constructed that have smaller memory requirements than the ϵ-machine [17, 66, 121, 192]. The study of such implementations is the task of *quantum computational mechanics*. Over a wide range of processes, a particular implementation of quantum simulation—the *q-machine*—has shown advantage in reduced memory rate; often the advantage over classical implementations is unbounded [4, 6, 61, 187]. For quantum machines, the minimal memory rate C_q has been determined in cases such as the Ising model [187] and the Perturbed Coin Process [191], where the q-machine attains the minimum rate. Though a given q-machine's memory can be readily calculated [160], in many cases the absolutely minimal C_q is not known.

In this chapter we will follow on the discussion of Chapter 4, with much the same goal. We will develop the representation theory of stochastic processes to include quantum models, and we will apply majorization and information theory to examine the memory costs involved.

The idea of using quantum-mechanical tools and Hilbert spaces as an alternative setting for representations of stochastic processes has been around at least as long as the representation theory of stochastic processes. In particular, representations involving the generation of a process via sequential measurements of a qubit had been proposed independently at least twice before the advent of the q-machine (see, for instance, [**83**] and [**217**]). The primary difference with the q-machine [**66**] is that it contains a direct embedding of the predictive states from the ϵ-machine, a feature not considered in prior approaches. Additionally, the q-machine formalism allows for fully general complex phase parameters [**102**].

We synthesize the theory of quantum representations of stochastic processes with a novel result that extends our Embedding Lemma 1 to quantum models. We expand on the role of the q-machine within this larger framework, and also describe a concrete alternative construction, the *reverse q-machine*. We additionally explore the structure of the space of quantum models for a given process by examining the effects of gauge invariants which arise from the choice of phase parameters in the model definition.

In Chapter 4, we showed that the ϵ-machine occupies an important information-theoretic role in the class of all classical predictive models of processes, in that it strongly minimizes that class in memory; this means that it achieves the minimal value for a wide spectrum of choices of entropy measure. In this chapter we demonstrate that, on the hand, every q-machine of any phase choice has a strong memory advantage over the ϵ-machine (and similarly every reverse q-machine has strong advantage over the reverse ϵ-machine)—but, unlike the case of classical predictors, there is generally no single *strongly* minimal q-machine; different entropies may be minimized by different choices of phase parameters. Additionally, this chapter sets the stage for Chapter 6, where we will examine classical and quantum models in an equivalent thermodynamic context, and compare their tradeoff of memory and thermodynamic resources.

This chapter is a synthesis of material on quantum generators from the publications *Strong and Weak Optimizations in Classical and Quantum Models of Stochastic Processes* [**106**], *Thermal Efficiency of Quantum Memory Compression* [**107**], and *Thermodynamically Efficient Local Computation* [**108**].

5.2. Quadratic models and quantum generators

In Section 4.2 we focused almost exclusively on linear process models—though we alluded to the existence of *quadratic* observable operator models (OOMs).

As with any OOM, a quadratic OOM of a process μ is a quartet $\mathfrak{K} = (\mathcal{H}, \mathcal{X}, \{K^{(x)}\}, P)$ composed of a vector space $\mathcal{H}$, an alphabet $\mathcal{X}$, a set of $\mathcal{X}$-labeled linear maps $K^{(x)}$ over $\mathcal{H}$, and a mapping $P : \mathcal{H} \to \mathbb{M}(\mathcal{X}^{\mathbb{N}})$ satisfying the equation:

$$P\left(K^{(x)}\psi\right) = \tau^{(x)}P(\psi)$$

for each $\psi \in \mathcal{H}$.

(The reader will notice we have changed quite a few letters, and even written the homomorphism law of models (originally Eq. (4.3)) quite differently. Some of these changes are just intended to start subtly shifting the reader's intuitions for what is going on—others are foreshadowing for the particularly *quantum* interpretation of these models which we will soon utilize.)

The difference between a quadratic OOM and its linear brethren is that the vector space $\mathcal{H}$ is assumed to be a Hilbert space, and the mapping P, rather than being a linear map, is defined as:

$$\Pr_{P\psi}(\, x_1 \ldots x_\ell \,) = \langle\psi|K^{(x_1)\dagger} \ldots K^{(x_\ell)\dagger} K^{(x_\ell)} \ldots K^{(x_1)}|\psi\rangle$$

The moniker "quadratic" arises from the quadratic norm-like nature of the mapping P. One consequence which should immediately stand out to the reader from this form is that the resulting measures are *always* positive. (This was not necessarily the case for linear OOMs, though it is of course necessarily true on the cone spanned by the past-embeddings.)

Let us now take some time to consider what we can learn about our model, and in particular the operators $K^{(x)}$. To start with, just as we defined unital vectors in linear OOMs as vectors v for which Pv is a normalized probability distribution, we will do the same for quadratic OOMs. Note

that *unital vector* does not necessarily mean *unit vector* (but just wait!). A unital vector ψ will be one which satisfies

$$\sum_{w \in \mathcal{X}^\ell} \langle \psi | K^{(w)\dagger} \dots K^{(w)\dagger} K^{(w)} \dots K^{(w)} | \psi \rangle = 1$$

for words of every length ℓ (here we have abbreviated $K^{(w)} = K^{(x_\ell)} \dots K^{(x_1)}$). Let us define the object

$$K_2^{(\ell)} = \sum_{w \in \mathcal{X}^\ell} K^{(w)\dagger} K^{(w)}$$

This is by construction a Hermitian and at least positive-semidefinite matrix. Consider its eigenvectors, v_i. It should be clear enough that whether a vector is unital is just a matter of scaling; so, let us assume that each v_i is unital. Then it must be the case that

$$\lambda_i \langle v_i | v_i \rangle = 1$$

Immediately evident is that no eigenvalue λ_i can equal zero; $K_2^{(\ell)}$ is actually positive definite. Further, we can now characterize all unital vectors as having the form

$$|\psi\rangle = \sum_i \frac{c_i}{\lambda_i} |v_i\rangle$$

where $\mathbf{c} = (c_i)$ *is* a unit vector.

Well, at this point, there hardly seems any point in not just rescaling the direction of each eigenvector by a factor of λ_i, which necessarily modifies the $K^{(x)}$ matrices but results in a quantum OOM where the unital vectors are *actually* unit vectors. We shall therefore, without *any* loss of generality, make it an additional requirement of quantum OOMs that unit vectors generate probability distributions under P.

A by-product of this rescaling is that now it is simply the case that

$$\sum_{w \in \mathcal{X}^\ell} K^{(w)\dagger} K^{(w)} = I$$

where I is the identity map on $\mathcal{H}$. The reader can check that it is only necessary that the above equation be true for $\ell = 1$, and then it will be true for all ℓ. The above constraint on a set $\{K^{(x)}\}$ of Hilbert space operators is called the *completeness* relation.

Operators satisfying such a relation are typically called *Kraus operators* and, incidentally, correspond to a quantum *positive operator-valued measure*, or POVM. POVMs are, in essence, the most general sort of measurement one can perform on a quantum system. They come about from allowing system described by $\mathcal{H}$ to entangle with an auxiliary system before performing a projective measurement on the auxiliary system and discarding it:

$$U\left|\psi\right\rangle \otimes \left|\emptyset\right\rangle = \sum_x K^{(x)}\left|\psi\right\rangle \otimes \left|x\right\rangle$$

The symbols x correspond to the outcome of the auxiliary system measurement and the operator $K^{(x)}$ describes both the probability of outcome x given ψ and the altered quantum state after measurement:

$$\Pr\left(\, x \mid \psi \,\right) = \left\langle\psi\right|K^{(x)\dagger}K^{(x)}\left|\psi\right\rangle$$

$$\left|\psi\right\rangle\left\langle\psi\right| \mapsto \frac{K^{(x)}\left|\psi\right\rangle\left\langle\psi\right|K^{(x)\dagger}}{\Pr\left(\, x \mid \psi \,\right)}$$

We can therefore consider every quadratic OOM to have an exact correspondence with a quantum measurement system. This motivates us, now, to expand our interest to *quantum generators*.

Quantum generators are both an extension of quadratic OOMs and also a useful framework with which to study them. We define a quantum generator as a triple $\mathfrak{Q} = (\mathcal{H}, \mathcal{X}, \{\mathcal{E}^{(x)}\})$, where $\mathcal{H}$ is a Hilbert space, $\mathcal{X}$ is the alphabet, and each $\mathcal{E}^{(x)} : \mathcal{B}(\mathcal{H}) \to \mathcal{B}(\mathcal{H})$ is a completely positive map on the space of bounded operators on $\mathcal{H}$, $\mathcal{B}(\mathcal{H})$, such that $\mathcal{E} = \sum_x \mathcal{E}^{(x)}$ is a completely positive and trace-preserving (CPTP) map.

Recall that a positive map is any which maps all positive-definite operators to positive-definite operators, and a completely positive map $\mathcal{E} : A \to B$ is one where $\mathrm{Id}\otimes\mathcal{E} : \mathbb{C}^{k\times k}\otimes\mathcal{B}(\mathcal{H}) \to \mathbb{C}^{k\times k}\otimes\mathcal{B}(\mathcal{H})$ is positive for all k, where Id is the identity. A map is trace-preserving if $\mathrm{Tr}\left[\mathcal{E}(A)\right] = \mathrm{Tr}\left[A\right]$ for all $A \in \mathcal{B}(\mathcal{H})$.

Choi's theorem for completely positive maps tells us that each completely positive map $\mathcal{E}$ can be decomposed into Kraus operators:

$$\mathcal{E}(\rho) = \sum_\alpha K^{(\alpha)}\rho K^{(\alpha)\dagger}$$

If $\mathcal{E}$ is a CPTP, then these operators must be complete.

It should be evident that any quadratic OOM corresponds to a quantum generator where $\mathcal{E}^{(x)}(\rho) = K^{(x)}\rho K^{(x)\dagger}$. Thus, we can see that quadratic models are a kind of extremal form of quantum generator: those where the $\mathcal{E}^{(x)}$ are pure Kraus operators.

One use of the quantum generator approach over that of the quadratic OOM is that each quantum generator is, also, a linear OOM. By this we mean there is a model $\mathfrak{M} = (\mathcal{B}(\mathcal{H}), \mathcal{X}, \{\mathcal{E}^{(x)}\}, \mathcal{P})$ where $\mathcal{P}$ is given by

$$\Pr_{\mathcal{P}(\rho)}(\, x_1 \dots x_\ell \,) = \mathrm{Tr}\left[\mathcal{E}^{(x_\ell)} \circ \cdots \circ \mathcal{E}^{(x_1)}(\rho)\right]$$

Note that the trace Tr plays the role here that **1** played before; formally, they accomplish the same goal.

Because a quantum generator is also a linear OOM, we can apply our results from that domain. Either it is decomposible into other linear OOMs, or it has a unique stationary state. We will generally here suppose that we are only dealing with indecomposible quantum generators. There must therefore exist a stationary density matrix ρ_π such that $\mathcal{E}(\rho_\pi) = \rho_\pi$.

Our results about embedding of pasts also holds. This means that there is a mapping $E_{\mathfrak{Q}} : \mathcal{X}^{\mathbb{N}} \to \mathcal{B}(\mathcal{H})$, sending each past $\overleftarrow{x}$ to a density matrix, which we will denote as $\rho_{\overleftarrow{x}}$.

An interesting result then follows for quadratic OOMs. This is an extension of the embedding theorem which ensures that for quadratic OOMs, the embedding is always a *pure* state. Proving this requires some appeals to quantum Perron-Frobenius theory; we will direct the reader to appropriate theorems in [**213**], but we will also encourage the reader to refresh their memory on classical Perron-Frobenius theory in Section 1.4.

THEOREM 13. *Suppose $\mathfrak{K} = (\mathcal{H}, \mathcal{X}, \{K^{(x)}\}, P)$ is a quadratic OOM, and further that it is indecomposible. Then the embedding map $E_{\mathfrak{K}}$ associated with the quantum generator in fact maps each past $\overleftarrow{x}$ to a* pure *quantum state $\psi_{\overleftarrow{x}}$, satisfying*

(1) $E_{\mathfrak{K}}(\overleftarrow{x})$ is a homomorphism of the shift space:

$$K^{(x)}\,|\psi_{\overleftarrow{x}}\rangle = e^{i\Phi(x,\overleftarrow{x})}\sqrt{\Pr_\mu(\, x \mid \overleftarrow{x} \,)}\,|\psi_{\overleftarrow{x}\,x}\rangle \tag{5.1}$$

for some $\Phi : \mathcal{X} \times \mathcal{X}^{\mathbb{N}} \to \mathbb{R}$;

(2) $E_{\mathfrak{K}}(\overleftarrow{x})$ *is predictively consistent:*

$$\Pr_{P\psi_{\overleftarrow{x}}}(w) = \Pr_{\mu}(w \mid \overleftarrow{x}) \ ; \tag{5.2}$$

(3) $E_{\mathfrak{K}}$ *and* Φ *are essentially continuous;*

(4) $E_{\mathfrak{K}}(\mathcal{X}^{\mathbb{N}})$ *spans the cyclic subspace generated by* $\{K^{(x)} | x \in \mathcal{X}\}$ *starting from any* $\psi_{\overleftarrow{x}}$.

PROOF. *Most of this is just a direct application of Lem. 1 to the linear OOM associated with the quantum generator. The key novel feature here is the fact that we are asserting the embedded states must be pure.*

Let us suppose for a minute we did not assume this. Then the cyclic subspace requirement of embeddings would still require that, starting from any linear combination of embedded states $\rho_{\overleftarrow{x}}$*, we end up "arbitrarily close" to a (ray of) any other embedded state. However, it is also the case that the space of pure states is invariant under the action of Kraus operators; thus, if any single linear combination of embedded states is pure, then all embedded states are pure.*

The question is whether it is possible that no *linear combination of embedded states is pure. For this we must invoke a result from the literature on completely positive quantum maps; we will essentially draw from a result of the Perron-Frobenius theory of completely positive maps.*

Suppose first that $\mathcal{E} = \sum_x K^{(x)} \cdot K^{(x)\dagger}$*, as a linear map, has only one eigenvalue of magnitude 1;* i.e. *it is a* primitive *map. Then there is a result (see Theorem 6.8 of [213]) that the cyclic subspace of Kraus operators acting on any item in* $\mathcal{B}(\mathcal{H})$ *is simply the entire space* $\mathcal{B}(\mathcal{H})$ *(recall we are assuming finite-dimensionality!). Applying this cyclic subspace to* ρ_π *directly means, however, that there must be some embedding states whose linear combination gives rise to a pure state* $|\psi\rangle\langle\psi|$.

Now let us suppose that $\mathcal{E}$ *is not primitive. This is of little matter. Consider instead* $\bar{\mathcal{E}} = \frac{1}{p}\sum_p \mathcal{E}^p$ *where* p *is the period of* $\mathcal{E}$*; the Kraus operators of this map are just the set of all composite Kraus operators* $K^{(x_1 \ldots x_\ell)} = K^{(x_\ell)} \ldots K^{(x_1)}$ *of length* $\ell \leq p$. $\bar{\mathcal{E}}$ *is primitive, and its cyclic subspaces are therefore still composed from the application of Kraus operators. We can therefore reach the same conclusion as the previous paragraph, proving the theorem.*

Thus, the linear OOM perspective of quantum generators provides a useful leverage point for proving interesting results about quadratic OOMs.

Let us now turn our attention back to the quadratic OOMs, and see what we can determine about their structure and memory.

5.3. The q-machine

We have essentially just demonstrated that every quadratic OOM $\mathfrak{K}$ actually embeds pasts as pure states in its model, in a homomorphic manner. To stretch the meaning of the embedding relation $\succsim$ from before, we have shown that $\mathfrak{P}_\mu \precsim \mathfrak{K}$, where $\mathfrak{P}_\mu$ is the past machine.

There is no particular reason why we should take special interest in $\mathfrak{P}_\mu$, however. Let us tentatively say that $\mathfrak{K}$ embeds a classical generator $\mathfrak{G} = (\mathcal{S}, \mathcal{X}, \{\mathbf{T}^{(x)}\})$ whenever there is a set of pure states $\{\psi_s | s \in \mathcal{S}\}$ such that

$$K^{(x)} |\psi_s\rangle = \sum_{s'} e^{i\phi_{x,s,s'}} \sqrt{T^{(x)}_{s's}} \, |\psi_{s'}\rangle \tag{5.3}$$

where $\{\phi_{x,s,s'} | x \in \mathcal{X}, s, s' \in \mathcal{S}\}$ is some set of phases. (Watch those phases! They are physically significant and, as we will see, can have a solenoid-like effect on the geometry of the states.)

What does this equation tell us? Essentially that—up to some phases—the amplitudes in the superpositions of states induced by $K^{(x)}$ mirror the probabilities in the mixed states induced by $T^{(x)}$.

However, we must not take Eq. (5.3) at face value. All quadratic OOMs must satisfy a completeness relation. Let us consider its implications for the overlap between two states $r, s \in \mathcal{S}$:

$$\langle \psi_r | \psi_s \rangle = \sum_{\substack{x \in \mathcal{X} \\ r', s' \in \mathcal{S}}} e^{i(\phi_{x,s,s'} - \phi_{x,r,r'})} \sqrt{T^{(x)}_{r'r} T^{(x)}_{s's}} \, \langle \psi_{r'} | \psi_{s'} \rangle \ , \tag{5.4}$$

This equation provides a recursive constraint on the geometry of the embedded states; specifically, what their state overlaps must be. We can solve this equation, *if it has a solution*, by finding the matrix $\Omega_{rs} := \langle \psi_r | \psi_s \rangle$ which is the "eigenvector" for the superoperator described by the above equation.

There is a quick way to check if a solution exists at all: let $r = s$, in which case we must have $\Omega_{rr} = 1$. The reader will quickly notice that this does not generally work for Eq. (5.4). In fact, there appear to be only two ways to guarantee that the solution can have $\Omega_r r = 1$ for all $r \in \mathcal{S}$.

The first is to require that $T^{(x)}_{r'r}$ be either unifilar. For in that case, then for each x in the sum $r' = s' = f(x, r)$, and so a constant diagonal is a viable solution.

Another way to guarantee a solution exists is to suppose that $T^{(x)}_{rs}$ is co-unifilar and $\Omega_{rs} = \delta_{rs}$. In particular $\Omega_{rs} = \delta_{rs}$ constrains that two states $r \neq s$ each map to different final states after observing a specific symbol x, which is the condition of co-unifilarity.

Now we will pause and remember that the ϵ-machine and reverse ϵ-machine are the result of state-merging unifilar and co-unifilar models, respectively; therefore, if we are constrained to embed only unifilar and co-unifilar models but are concerned about memory efficiency, we can save ourselves time by focusing our examination directly on embeddings of the ϵ-machine and the reverse ϵ-machine. These are called the q-machine and reverse q-machine, respectively.

5.3.1. The forward q-machine. Let us suppose that there does exist a solution Ω_{rs} which is physically consistent, for a unifilar dynamic described by $T^{(x)}_{r'r}$. It has in fact been proven that under these conditions a unique solution exists for any choice of phases $\{\phi_{xs}\}$ [102]. The physical properties of each quantum generator are entirely determined by its overlap matrix $\Omega_{rs} = \langle\psi_r|\psi_s\rangle$. However, this in itself contains nonphysical degrees of freedom [106]. None of the invariant geometry of our generators is modified when under the transformation $|\psi_s\rangle \mapsto e^{i\Psi_s}|\psi_s\rangle$ on the signal states. Thus, these represent a *gauge transformation*. In terms of the overlap matrix, this means that our generators are invariant under the transformations $\Omega_{rs} \mapsto e^{i(\Psi_s - \Psi_r)}\Omega_{rs}$. We will discuss the implications of this in a later section.

For now, it is enough to note that once Ω_{rs} is determined, the encoding states and Kraus operators can be explicitly constructed. Let $\sqrt{\pi_r\pi_s}\Omega_{rs} = \sum_\alpha U_{r\alpha}U^*_{s\alpha}\lambda_\alpha$ be the singular value decomposition of $\sqrt{\pi_r\pi_s}\Omega_{rs}$ into a unitary $U_{i\alpha}$ and singular values $\lambda_\alpha > 0$. Suppose $\alpha = 1, \ldots, d$. Then given any computational basis $\{|\alpha\rangle : \alpha = 1, \ldots, d\}$, we can construct:

$$|\psi_s\rangle = \sum_\alpha \sqrt{\frac{\lambda_\alpha}{\pi_s}} U^*_{s\alpha}|\alpha\rangle \text{ and} \tag{5.5}$$

$$K^{(x)} = \sum_{\alpha,\beta,s} e^{i\phi_{xs}} U^*_{s'\beta}U_{s\alpha}\sqrt{\frac{\lambda_\beta\pi_s}{\lambda_\alpha\pi_{s'}}T^{(x)}_{s's}}\,|\beta\rangle\langle\alpha| \ . \tag{5.6}$$

It is easy to check that $\langle\psi_r|\psi_s\rangle = \Omega_{rs}$ and that Eq. (5.3) is satisfied by this construction. Notice that:

$$\rho_\pi = \sum_s \pi_s \left|\psi_s\right\rangle \left\langle\psi_s\right| = \sum_\alpha \lambda_\alpha \left|\alpha\right\rangle \left\langle\alpha\right| \ . \tag{5.7}$$

So, the computational basis α is the diagonal basis of the stationary state ρ_π.

When the process being generated is a Markov process, so that predictive states correspond to symbols ($\mathcal{S} = \mathcal{X}$), the Kraus operator $K^{(x)}$ always ends maps to the state $|\psi_x\rangle$. In this case, the singular value decomposition approach may be overkill, and a simpler approach is justified: we are seeking a set of states $|\psi_x\rangle$ and *dual* states $|\phi_x\rangle$ such that

$$\langle\phi_x|\psi_y\rangle = e^{i\phi_{xy}}\sqrt{T_{xy}}$$

Here $T_{xy} = T_{xy}^{(x)}$ is a shorthand for Markov processes. Given these states the Kraus operators may be written simply as $K^{(x)} = |\psi_x\rangle\langle\phi_x|$, and completeness corresponds to

$$\sum_x \left|\phi_x\right\rangle \left\langle\phi_x\right| = I$$

In this case we would not be so interested in the overlap matrix Ω_{rs} as we are in the amplitude matrix $A_{xy} = e^{i\phi_{xy}}\sqrt{T_{xy}}$, whose singular value decomposition $A_{xy} = \sum_\alpha U_{x\alpha} s_\alpha V^*_{y\alpha}$ provides the state representations as

$$|\psi_x\rangle = \sum_\alpha s_\alpha V^*_{x\alpha} \left|\alpha\right\rangle \text{ and} \tag{5.8}$$

$$|\phi_x\rangle = \sum_\alpha U^*_{x\alpha} \left|\alpha\right\rangle \ . \tag{5.9}$$

The reader can check that this satisfies the completeness relations. The dual-basis approach first appeared in [4] for all q-machines, and remains particularly useful for the Markov chain case.

We will note lastly that the the quantum generator corresponding to the q-machine is a *proper* embedding of the ϵ-machine, in the sense defined by our $\succsim$ relation in the previous chapter. Let $\mathcal{E}^{(x)} = K^{(x)} \cdot K^{(x)\dagger}$ be the positive maps of the q-machine, and let $\mathcal{P} : \mathcal{S} \to \mathcal{B}(\mathcal{H})$ be the mapping from states to density matrices given by $\mathcal{P}(s) = |\psi_s\rangle\langle\psi_s|$. One can straightforwardly check, as a

result of Eq. (5.3) and the unifilarity of the ϵ-machine, that

$$\mathcal{E}^{(x)} \circ \mathcal{P}(s) = \sum_{s'} \mathcal{P}(s') T^{(x)}_{s's}$$

So, for each q-machine $\mathfrak{K}$, we have $\mathfrak{K} \succsim \mathfrak{E}(\mu)$.

5.3.2. The reverse q-machine. Let us now consider the case where the matrices are co-unifilar, and $\Omega_{rs} = \delta_{rs}$. Similarly to the unifilar case, we will focus exclusively on the reverse ϵ-machine. Before we dive into this, let us remind the reader that every co-unifilar generator is just the time-reverse of a unifilar generator, and in particular the time-reverse of the reverse ϵ-machine is the ϵ-machine of the reverse process.

It stands to reason that for time-reversing the q-machine. Let $\mathfrak{E}_R(\mu)$ be the reverse ϵ-machine of a process μ. Its time reverse is the forward ϵ-machine of the reverse process, $\mathfrak{E}(\mu^R)$. From it, we can construct a q-machine $\mathfrak{Q}(\mu^R)$, with Kraus operators expressed by $\left\{\widetilde{K}^{(x)}\right\}$. Recall that the generated process of the q-machine is given by

$$\Pr_{\mathfrak{Q}(\mu^R)}(\, x_1 \ldots x_t \,) := \mathrm{Tr}\left[\widetilde{K}^{(x_t \ldots x_1)} \rho_\pi \widetilde{K}^{(x_t \ldots x_1)\dagger}\right] \ . \tag{5.10}$$

This is the time-reverse of the process generated by $\mathfrak{E}_R(\mu)$, expressed in the equation $\Pr_{\mathfrak{Q}(\mu^R)}(\, x_1 \ldots x_t \,) = \Pr_\mu(\, x_t \ldots x_1 \,)$. In terms of the q-machine, we can write:

$$\begin{aligned} \Pr_\mu(\, x_t \ldots x_1 \,) &:= \mathrm{Tr}\left[\widetilde{K}^{(x_t)} \ldots \widetilde{K}^{(x_1)} \rho_\pi \widetilde{K}^{(x_1)\dagger} \widetilde{K}^{(x_t)\dagger}\right] \\ &= \mathrm{Tr}\left[K^{(x_t)} \ldots K^{(x_1)} \rho_\pi K^{(x_1)\dagger} \ldots K^{(x_t)\dagger}\right] \ , \end{aligned} \tag{5.11}$$

where $K^{(x)} = \rho_\pi^{1/2} \widetilde{K}^{(x)\dagger} \rho_\pi^{-1/2}$. This is, essentially, the Petz reversal of the POVM $\{\widetilde{K}^{(x)}\}$, and it constitutes a formal time-reversal of the quantum process [37].

Computing $K^{(x)}$ is straightforward using Eq. (5.5), as this gives the Kraus operators in the diagonal basis of ρ_π, where it is easiest to compute $\rho_\pi^{1/2}$ and its inverse. We have:

$$K^{(x)} = \sum_{\alpha,\beta,s} e^{-i\phi_{xs}} U_{s'\beta} U^*_{s\alpha} \sqrt{\frac{\pi_s}{\pi_{s'}} \widetilde{T}^{(x)}_{s's}} \, |\alpha\rangle \langle\beta| \ . \tag{5.12}$$

Now, take the basis $|\psi_s\rangle = \sum_\alpha U^*_{s\alpha} |\alpha\rangle$. In this basis:

$$K^{(x)} = \sum_{s'} e^{-i\phi_{xs'}} \sqrt{T^{(x)}_{s's}} \, |\psi_{s'}\rangle \langle\psi_s| \ . \tag{5.13}$$

Now we can check and see that the basis $\{|\psi_s\rangle\}$ and Kraus operators $\{K^{(x)}\}$ satisfy the equations Eq. (5.1). Furthermore, it is evident that $\langle\psi_r|\psi_s\rangle = \delta_{rs}$ when the number of nonzero singular values λ_α is at least equal to the number of states; when it is less, the matrix $\langle\psi_r|\psi_s\rangle$ is still a projector. This means that by following our intuition regarding time-reversal we have actually discovered an *additional* way of solving equation Eq. (5.4) for co-unifilar $T^{(x)}_{r'r}$ without assuming $\Omega_{rs} = \delta_{rs}$. The possibility of "dimension-reducing" solutions to quantum models will play a key role in our investigation of memory in quantum generators.

Note that the stationary state of a time-reversed q-machine is just the stationary state of the original q-machine—this is not altered under time reversal. However, we find a new expression for the stationary state, in terms of the basis $\{|\psi_s\rangle\}$:

$$\rho_\pi = \sum_{r,s,\alpha} \lambda_\alpha U^*_{r\alpha} U_{s\alpha} \, |\psi_s\rangle \langle\psi_r| \tag{5.14}$$

$$= \sum_{r,s} \sqrt{\pi_r \pi_s} \Omega_{sr} \, |\psi_s\rangle \langle\psi_r| \ . \tag{5.15}$$

So, ρ_π is generally not diagonal in the basis $\{|\psi_s\rangle\}$. The extent to which ρ_π commutes with $\{|\psi_s\rangle\}$ is the extent to which Ω_{rs} is block-diagonal.

For Markov chains, we can see clearly that every q-machine of a process is also a reverse q-machine. This is evident in the dual-basis form: the $|\phi_x\rangle$ states correspond to the retrodictive state embeddings while the $|\psi_s\rangle$ states correspond to the predictive state embeddings.

Unlike the q-machine, the reverse q-machine is not actually a proper embedding of the reverse ϵ-machine in the homomorphism sense; in fact, the opposite is the case. If we let $\mathcal{M} : \mathcal{B}(\mathcal{H}) \to \mathbb{M}(\mathcal{S})$ be the mapping from states to distributions given by the projective measurement $\mathcal{M}(\rho)_r = \langle\psi_r|\rho|\psi_r\rangle$, then we can check that Eq. (5.3) and the co-unifilarity of the reverse ϵ-machine imply that

$$\mathbf{T}^{(x)} \circ \mathcal{M} = \mathcal{M} \circ \mathcal{E}^{(x)}$$

So, for any reverse q-machine $\mathfrak{K}$ we have $\mathfrak{E}_R(\mu) \succsim \mathfrak{K}$.

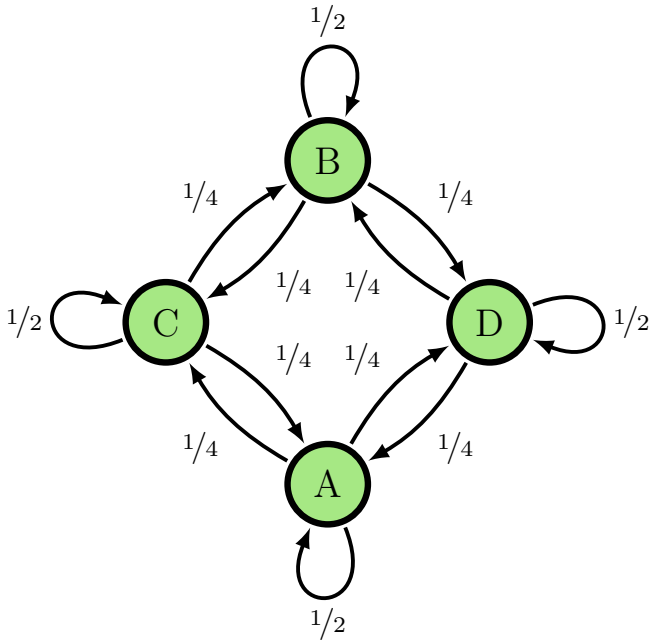

FIGURE 5.1. The 4-state MBW Process as a Markov chain (which is the ϵ-machine).

5.3.3. Examples of q-machines. In this section we will provide some examples of q-machine construction; first, we will provide two examples of Markov chains and apply the dual construction to demonstrate two distinct q-machines for each example. Then we will consider two more general sorts of process, for which we will examine the gauge freedoms of the overlap matrix and parameterize all (gauge-invariant) q-machines for those processes.

EXAMPLE 9. *First we will consider the "MBW Process" introduced in Ref. [**133**], where they provided the first example demonstrated a machine whose q-machine with nonzero phases. Consider the process generated by the* 4-state MBW *machine shown in Fig. 5.1.*

This process' HMM is simply a Markov chain, and its representation in Fig. 5.1 is its ϵ-machine. Denote this classical representation by $\mathfrak{M}_4$. If we take $\{|A\rangle, |B\rangle, |C\rangle, |D\rangle\}$ as an orthonormal basis

of a Hilbert space, we can construct the q-machine with the states:

$$
\begin{aligned}
|\psi_A\rangle &:= \frac{1}{\sqrt{2}}|A\rangle + \frac{1}{2}\left(|C\rangle + |D\rangle\right) \ , \\
|\psi_B\rangle &:= \frac{1}{\sqrt{2}}|B\rangle + \frac{1}{2}\left(|C\rangle + |D\rangle\right) \ , \\
|\psi_C\rangle &:= \frac{1}{\sqrt{2}}|C\rangle + \frac{1}{2}\left(|A\rangle + |B\rangle\right) \ , \ \textit{and} \\
|\psi_D\rangle &:= \frac{1}{\sqrt{2}}|D\rangle + \frac{1}{2}\left(|A\rangle + |B\rangle\right) \ .
\end{aligned}
$$

Since it is a Markov chain, we can write the Kraus operators as $K_x := |\psi_x\rangle\langle\phi_x|$*, where* $\langle\phi_x|\psi_{x'}\rangle \propto \sqrt{P_{x'|x}}$*. For q-machines of Markov chains, then, the dual basis is just* $\langle\phi_x| = \langle x|$*. We denote the q-machine model of the 4-state MBW Process as* $\mathfrak{Q}_4$*.*

It turns out that there is a smaller quantum model embedded in two dimensions, with states:

$$
\begin{aligned}
|\psi'_A\rangle &:= |0\rangle \ , \\
|\psi'_B\rangle &:= |1\rangle \ , \\
|\psi'_C\rangle &:= \frac{1}{\sqrt{2}}\left(|0\rangle + |1\rangle\right) \ , \ \textit{and} \\
|\psi'_D\rangle &:= \frac{1}{\sqrt{2}}\left(|0\rangle - |1\rangle\right) \ .
\end{aligned}
$$

In this case, $\langle\phi'_x| = \frac{1}{\sqrt{2}}\langle\psi'_x|$ *derives the q-machine. This gives the proper transition probabilities for the 4-state MBW model. We denote this dimensionally-smaller model* $\mathfrak{D}_4$*.*

EXAMPLE 10. *Now consider the 3-state MBW model, denoted* $\mathfrak{M}_3$ *and displayed in Fig. 5.2. This is a generalization of the previous example to three states instead of four. We will compute the corresponding q-machine* $\mathfrak{Q}_3$ *and show that there also exists a dimensionally-smaller representation* $\mathfrak{D}_3$*. In this case, however,* $\mathfrak{D}_3$ *is not smaller in its statistical memory.*

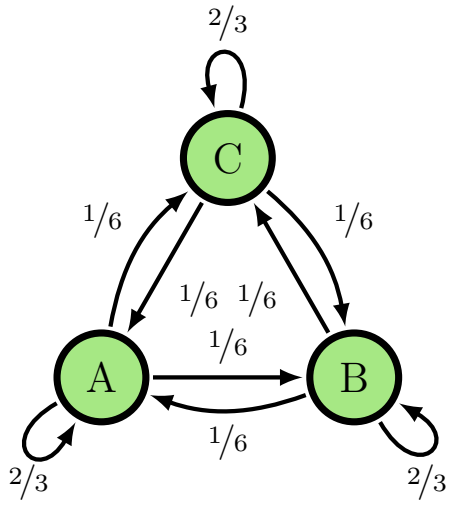

FIGURE 5.2. 3-state MBW Process as a Markov chain (which is the process' ϵ-machine).

The q-machine $\mathfrak{Q}_3$ of this Markov chain is given by the states:

$$|\psi_A\rangle := \sqrt{\frac{2}{3}}\,|A\rangle + \frac{1}{\sqrt{6}}\,(|B\rangle + |C\rangle) \ ,$$
$$|\psi_B\rangle := \sqrt{\frac{2}{3}}\,|B\rangle + \frac{1}{\sqrt{6}}\,(|A\rangle + |C\rangle) \ , \ \textit{and}$$
$$|\psi_C\rangle := \sqrt{\frac{2}{3}}\,|C\rangle + \frac{1}{\sqrt{6}}\,(|A\rangle + |B\rangle) \ ,$$

and Kraus operators defined similarly to before.

The lower-dimensional model $\mathfrak{D}_3$ is given by the states:

$$|\psi_A\rangle := |0\rangle \ ,$$
$$|\psi_B\rangle := \frac{1}{2}\,|0\rangle + \frac{\sqrt{3}}{2}\,|1\rangle \ , \ \textit{and}$$
$$|\psi_C\rangle := \frac{1}{2}\,|0\rangle - \frac{\sqrt{3}}{2}\,|1\rangle \ ,$$

with $\langle\phi'_x| = \sqrt{\frac{2}{3}}\,\langle\psi'_x|$. This gives the proper transition probabilities for the 3-state MBW model.

In this case, we can provide a simple proof that $\mathfrak{D}_3$ is geometrically (that is, up to gauge invariance) the only *2-dimensional model of the 3-MBW process. Let us set up the general equations for the*

states and dual states:

$$\begin{aligned}
\langle\phi_A|\psi_A\rangle &= e^{i\phi_{AA}}\sqrt{\tfrac{2}{3}}\ , \quad \langle\phi_A|\psi_B\rangle = e^{i\phi_{AB}}\tfrac{1}{\sqrt{6}}\ , \quad \langle\phi_A|\psi_C\rangle = e^{i\phi_{AC}}\tfrac{1}{\sqrt{6}}\ , \\
\langle\phi_B|\psi_A\rangle &= e^{i\phi_{BA}}\tfrac{1}{\sqrt{6}}\ , \quad \langle\phi_B|\psi_B\rangle = e^{i\phi_{BB}}\sqrt{\tfrac{2}{3}}\ , \quad \langle\phi_B|\psi_C\rangle = e^{i\phi_{BC}}\tfrac{1}{\sqrt{6}}\ , \\
\langle\phi_C|\psi_A\rangle &= e^{i\phi_{CA}}\tfrac{1}{\sqrt{6}}\ , \quad \langle\phi_C|\psi_B\rangle = e^{i\phi_{CB}}\tfrac{1}{\sqrt{6}}\ , \quad \langle\phi_C|\psi_C\rangle = e^{i\phi_{CC}}\sqrt{\tfrac{2}{3}}
\end{aligned}$$

Now keep in mind the possible gauge transformation $|\psi_x\rangle \mapsto e^{i\Psi_x}|\psi_x\rangle$ *(accompanied by* $|\phi_x\rangle \mapsto e^{i\Psi_x}|\phi_x\rangle$*). This means that the amplitude matrix* $A_{xy} = \langle\phi_x|\psi_y\rangle$ *can undergo the transformation* $A_{xy} \mapsto e^{i(\Psi_y - \Psi_x)}A_{xy}$ *without changing the invariant geometry.*

That these states are embedded in a 2D Hilbert space requires that A_{xy} *be degenerate. One way to enforce this to check that the characteristic polynomial* $\det(\mathbf{A} - \lambda\mathbf{I}_3)$ *has an overall factor of* λ*. For simplicity, we compute the characteristic polynomial of* $A\sqrt{6}$*:*

$$\begin{aligned}
\det(\sqrt{6}\mathbf{A} - \lambda\mathbf{I}_3) = (2-\lambda)^3 + & \\
& \left(e^{i(\phi_{AB}+\phi_{BC}+\phi_{CA})} + e^{i(\phi_{BA}+\phi_{CB}+\phi_{AC})}\right) - \\
& (2-\lambda)\left(e^{i(\phi_{AB}+\phi_{BA})} + e^{i(\phi_{AC}+\phi_{CA})} + e^{i(\phi_{BC}+\phi_{CB})}\right)\ .
\end{aligned}$$

To have an overall factor of λ*, we need:*

$$\begin{aligned}
0 = 8 &+ \left(e^{i(\phi_{AB}+\phi_{BC}+\phi_{CA})} + e^{i(\phi_{BA}+\phi_{CB}+\phi_{AC})}\right) \\
&- 2\left(e^{i(\phi_{AB}+\phi_{BA})} + e^{i(\phi_{AC}+\phi_{CA})} + e^{i(\phi_{BC}+\phi_{CB})}\right)\ .
\end{aligned}$$

Typically, there will be several ways to choose phases to cancel out vectors, but in this case since the sum of the magnitudes of the complex terms is 8, the only way to cancel is at the extreme point where $\phi_{AB} = -\phi_{BA} = \phi_1$*,* $\phi_{BC} = -\phi_{CB} = \phi_2$*, and* $\phi_{CA} = -\phi_{AC} = \phi_3$ *and:*

$$\phi_1 + \phi_2 + \phi_3 = \pi\ .$$

To recapitulate the results so far, $\mathbf{A}$ *has the form:*

$$\mathbf{A} = \frac{1}{\sqrt{6}} \begin{pmatrix} 2 & e^{i\phi_1} & -e^{i(\phi_1+\phi_2)} \\ e^{-i\phi_1} & 2 & e^{i\phi_2} \\ -e^{-i(\phi_1+\phi_2)} & e^{-i\phi_2} & 2 \end{pmatrix} .$$

But we can see that this is exactly equal to $A_{xy} = e^{i(\Psi_y - \Psi_x)} \hat{A}_{xy}$*, where* $\Psi_x = (-\phi_1, 0, \phi_2)$ *and*

$$\hat{\mathbf{A}} = \frac{1}{\sqrt{6}} \begin{pmatrix} 2 & 1 & -1 \\ 1 & 2 & 1 \\ -1 & 1 & 2 \end{pmatrix} .$$

So, there is geometrically only one 2-dimensional solution, and its amplitude matrix corresponds to the one we already discovered in $\mathfrak{D}_3$*.*

The following examples will explore in more depth the nature of gauge invariance with respect to phase. We have mentioned that the overlap matrix may undergo the transformation $\Omega_{rs} \mapsto e^{i(\Psi_s - \Psi_r)}\Omega_{rs}$ for some vector Ψ_s without changing the geometry between states.

It is helpful to consider these gauge properties in terms of how they act on the phases $\{\phi_{xs}\}$ that determine the quantum generator. Applying the gauge transformation to the consistency formula gives:

$$\Omega_{rs} = \sum_x \sqrt{\Pr(x|r)\Pr(x|s)} e^{i(\tilde{\phi}_{xs} - \tilde{\phi}_{xr})} \Omega_{f(x,r)f(x,s)} \ ,$$

where:

$$\tilde{\phi}_{xs} = \phi_{xs} - \Psi_s + \Psi_{f(x,s)} \tag{5.16}$$

is the induced transformation on the generator's phases. Eq. (5.16) can be taken as a fundamental description of the gauge transformation.

Using Eq. (5.16) allows us to determine the *gauge invariants*—that is, combinations of the phases $\{\phi_{xs}\}$ that do not change under a gauge transformation. In this case, the gauge invariants are best understood graphically, in terms of the hidden Markov models from before. Each phase $\{\phi_{xs}\}$ can

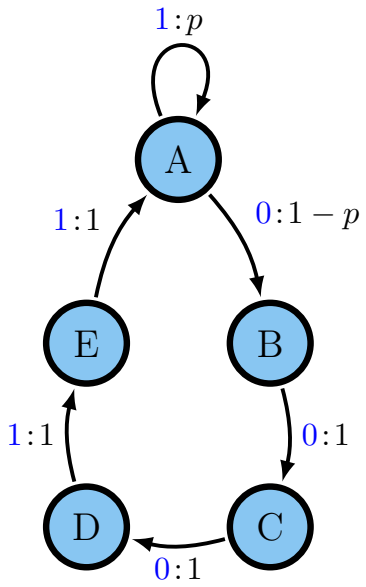

FIGURE 5.3. ϵ-machine of the $3, 2$-Golden Mean process.

be understood as being assigned to an edge, while each phase in the gauge transformation $\{\Psi_s\}$ can be seen as being assigned to a state.

For each loop of edges, we can take a linear combination of the constituent edges' phases ϕ_{xs}, adding positive and negative signs based on the direction of the edges. These loop sums are the gauge invariants. For instance, the Nemo process has $\Phi_0 = \phi_{0A}$, $\Phi_1 = \phi_{1C} - \phi_{0C}$, and $\Phi_2 = \phi_{1A} + \phi_{1B} + \phi_{1C}$ as gauge invariants.

EXAMPLE 11 ((R, k)-Golden Mean Generators). *An R, k-Golden Mean Generator is one with $R + k$ memory states. These states can be considered to belong to two groups: the A state, which is the only nondeterministic state and the B-states $\mathcal{B} \equiv \{B_1, \ldots, B_{R+k-1}\}$. The B-states are further broken down into a Markov part $\mathcal{R} \equiv \{B_1, \ldots, B_{R-1}\}$ and a cryptic part $\mathcal{K} \equiv \{B_R, \ldots, B_{R+k-1}\}$. The*

dynamic on the generator is given by:

$$\Pr\left(s', 0|s\right) = \begin{cases} 1-p & s = A,\ s' = B_1 \\ 1 & s = B_r,\ s' = B_{r+1},\ 0 \le r < R \\ 0 & \text{otherwise} \end{cases}$$

and

$$\Pr\left(s', 1|s\right) = \begin{cases} p & s', s = A \\ 1 & s = B_r,\ s' = B_{r+1},\ R \le r \le R+k-2 \\ 1 & s = B_{R+k-1},\ s' = A \\ 0 & \text{otherwise} \end{cases}.$$

We can check that:

$$\Pr{}_0(s) = \begin{cases} \frac{1}{1+(R+k-1)(1-p)} & s = A \\ \frac{1-p}{1+(R+k-1)(1-p)} & s = A \end{cases}$$

is the stationary distribution. Letting $Z = 1 + (R+k-1)(1-p)$*, we have:*

$$\Pr\left(s', 0, s\right) = \begin{cases} \frac{1-p}{Z} & s = A,\ s' = B_1 \\ \frac{1-p}{Z} & s = B_r,\ s' = B_{r+1},\ 0 \le r < R \\ 0 & \text{otherwise} \end{cases}$$

and

$$\Pr\left(s', 1, s\right) = \begin{cases} \frac{p}{Z} & s', s = A \\ \frac{1-p}{Z} & s = B_r,\ s' = B_{r+1},\ R \le r \le R+k-2 \\ \frac{1-p}{Z} & s = B_{R+k-1},\ s' = A \\ 0 & \text{otherwise} \end{cases}$$

It is helpful to also have:

$$\Pr(X=0) = \frac{R(1-p)}{Z} \ ,$$
$$\Pr(X=1) = \frac{(k-1)(1-p)-1}{Z} \ ,$$
$$\Pr\left(s'|0\right) = \begin{cases} \frac{1}{R} & s=A,\ s'=B_1 \\ \frac{1}{R} & s=B_r,\ s'=B_{r+1},\ 0\leq r<R \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\Pr\left(s'|1\right) = \begin{cases} \frac{p}{(k-1)(1-p)-1} & s',s=A \\ \frac{1-p}{(k-1)(1-p)-1} & s=B_r,\ s'=B_{r+1},\ R\leq r\leq R+k-2 \\ \frac{1-p}{(k-1)(1-p)-1} & s=B_{R+k-1},\ s'=A \\ 0 & \text{otherwise.} \end{cases}$$

First, we wish to show that regardless of the chosen phases $\{\phi_{xs}\}$ we get the equivalent quantum model. Recall that the formula defining the overlaps is given by:

$$\Omega_{rs} = \sum_x \sqrt{\Pr(x|r)\Pr(x|s)}e^{i(\phi_{xs}-\phi_{xr})}\Omega_{f(r,s)f(x,s)} \ .$$

In this case, we have:

$$\Omega_{AB_{R+k-1}} = \sqrt{p}e^{i(\phi_{1B_{R+k-1}}-\phi_{1A})}$$
$$\Omega_{B_rB_s} = e^{i(\phi_{1B_r}-\phi_{1B_s})}\Omega_{B_{r+1}B_{s+1}}$$
$$\Omega_{AB_r} = \sqrt{p}e^{i(\phi_{1B_r}-\phi_{1A})}\Omega_{AB_{r+1}} \ ,$$

which has the solution:

$$\frac{\Omega_{AB_{R+m}}}{\sqrt{p^{k-m}}} = e^{i\left(\sum_{j=m}^{k-1}\phi_{1B_{R+j}}-(k-m)\phi_{1A}\right)}$$
$$\frac{\Omega_{B_{R+m}B_{R+n}}}{\sqrt{p^{m-n}}} = e^{i\left(\sum_{j=n}^{k-1}\phi_{1B_{R+j}}-\sum_{j=m}^{k-1}\phi_{1B_{R+j}}-(m-n)\phi_{1A}\right)} \ .$$

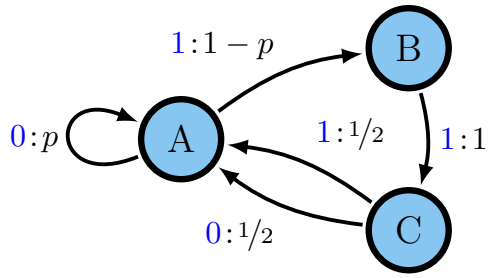

FIGURE 5.4. ϵ-machine of the Nemo process.

Note that under the gauge transformation $\Psi_A = k\phi_{1A}$ *and* $\Psi_{B_m} = \sum_{j=m}^{k-1} \phi_{1B_{R+j}} + m\phi_{1A}$*, we can eliminate phases and end up simply with:*

$$\begin{aligned} \Omega_{AB_{R+m}} &= \sqrt{p^{k-m}} \\ \Omega_{B_{R+m}B_{R+n}} &= \sqrt{p^{m-n}} \end{aligned} \tag{5.17}$$

We note that this matrix only explicitly depends upon k *and not* R*. This extends a result from Ref. [160] to all* R *and* k*, as well as to all choices of phase* $\{\phi_{xs}\}$*.*

EXAMPLE 12 (Nemo Process). *For the Nemo Process, not all phases* $\{\phi_{xs}\}$ *give the equivalent implementation. To analyze the situation in more detail, we make use of the gauge invariants.*

The gauge invariants of the Nemo q-machines are:

$$\begin{aligned} \Phi_0 &= \phi_{0A} \\ \Phi_1 &= \phi_{1C} - \phi_{0C} \\ \Phi_2 &= \phi_{1A} + \phi_{1B} + \phi_{1C} \end{aligned} \tag{5.18}$$

We work to express the overlap matrix in terms of these invariants.

The formula defining the overlaps, for the Nemo process, this gives the system of equations:

$$\Omega_{AB} = \sqrt{1-p}e^{i(\phi_{1C}-\phi_{1A})}\Omega_{BC}$$

$$\Omega_{BC} = \frac{1}{\sqrt{2}}e^{i(\phi_{1C}-\phi_{1B})}\Omega_{CA}$$

$$\Omega_{CA} = \sqrt{\frac{p}{2}}e^{i(\phi_{0A}-\phi_{0C})} + \sqrt{\frac{1-p}{2}}e^{i(\phi_{1A}-\phi_{1C})}\Omega_{AB}$$

which has the solution:

$$\Omega_{AB} = \frac{\sqrt{p(1-p)}}{1+p} e^{i(\phi_{1C}-\phi_{1A}+\phi_{0A}-\phi_{0C})}$$
$$\Omega_{BC} = \frac{\sqrt{p}}{1+p} e^{i(\phi_{1C}-\phi_{1A}+\phi_{0A}-\phi_{0C})}$$
$$\Omega_{CA} = \frac{\sqrt{2p}}{1+p} e^{i(\phi_{0A}-\phi_{0C})}$$

Now, we gauge fix ϕ_{1A} and ϕ_{1B} so that Ω_{AB} and Ω_{BC} are phaseless. The result is:

$$\Omega_{AB} = \frac{\sqrt{p(1-p)}}{1+p}$$
$$\Omega_{BC} = \frac{\sqrt{p}}{1+p} \tag{5.19}$$
$$\Omega_{CA} = \frac{\sqrt{2p}}{1+p} e^{i(2\Phi_0+2\Phi_1-\Phi_2)}$$

We see that the overlap matrix then only depends on the gauge invariants in the single phase $\Phi = 2\Phi_0 + 2\Phi_1 - \Phi_2$. This generalizes a result from Ref. [121] to all input phases $\{\phi_{xs}\}$.

5.4. Quantum information and memory compression

Almost every concept and relation in Shannon information theory either has a direct analogue or—in the most interesting cases—some subversion in the quantum realm. We will mostly be interested here in quantum analogues of the Rényi entropies. Given a density matrix ρ with eigenvalue distribution $\boldsymbol{\lambda}$, the quantum Rényi entropy can be defined simply as

$$\mathrm{H}_{\alpha,\mathrm{q}}[\rho] = \mathrm{H}_{\alpha}[\boldsymbol{\lambda}]$$

Generally these can be expressed with eigenvalues in the form $\mathrm{H}_{\alpha,\mathrm{q}}[\rho] = \frac{1}{1-\alpha}\mathrm{Tr}\left[\rho^{\alpha}\right]$. Some special cases of interest are:

$$\begin{aligned}
\mathrm{H}_{\mathrm{q}}[\rho] &= \lim_{\alpha\to 1} \mathrm{H}_{\alpha,\mathrm{q}}[\rho] = \mathrm{H}[\boldsymbol{\lambda}] = -\mathrm{Tr}\left[\rho \log_2 \rho\right] \\
\mathrm{H}_{\max,\mathrm{q}}[\rho] &= \lim_{\alpha\to 0} \mathrm{H}_{\alpha,\mathrm{q}}[\rho] = \mathrm{H}_{\max}[\boldsymbol{\lambda}] = \log_2 \mathrm{rank}(\rho) \\
\mathrm{H}_{\min,\mathrm{q}}[\rho] &= \lim_{\alpha\to \infty} \mathrm{H}_{\alpha,\mathrm{q}}[\rho] = \mathrm{H}_{\min}[\boldsymbol{\lambda}] = -\log_2 \inf\{\, \lambda \mid \rho \leq \lambda I \,\} \\
\mathrm{H}_{1/2,\mathrm{q}}[\rho] &= \mathrm{H}_{1/2}[\boldsymbol{\lambda}] = 2\log_2 \mathrm{Tr}\left[\sqrt{\rho}\right]
\end{aligned}$$

The quantity $\mathrm{H}_{\mathrm{q}}[\cdot]$ is often called the *von Neumann* entropy. It is also necessary at this point to address an unfortunate conflict of notation which occurs between Shannon and quantum information. In quantum information, it is common to refer to $\mathrm{H}_{1/2,\mathrm{q}}[\cdot]$ as the max-entropy and write it as $\mathrm{H}_{\max,\mathrm{q}}[\cdot]$. In classical information, this is reserved for the limit $\mathrm{H}_{\alpha}[\cdot]$ as $\alpha \to 0$. In the quantum setting, this deviation arises from the fact that $\mathrm{H}_{1/2,\mathrm{q}}[\cdot]$ and $\mathrm{H}_{\min,\mathrm{q}}[\cdot]$ have very interesting dual relationships which we will in fact observe in the next chapter. However, due to the fact that we are stepping so frequently between classical and quantum information theory, and since $\mathrm{H}_{0,\mathrm{q}}[\cdot]$ proves so important in this section, I will retain the Shannon notation throughout this thesis.

Now, each of these quantities has several operational interpretations in the setting of pure quantum information theory, in the form of various quantum coding theorems, and these theorems often mirror those of the corresponding quantities in Shannon information theory. We will not be terribly concerned with those results here. Instead we will be more interested in how quantum systems can be used as an intermediate step by encoding some classical information (say, a generator state) and producing a classical output (say, a generated process). The question, then, is how quantum information quantities relate to their Shannon information counterparts when we shift between classical and quantum settings.

The question, then, is how quantum information quantities relate to their Shannon information counterparts when we shift between classical and quantum settings. In making this comparison, majorization will prove very useful.

Let us start with the task of embedding a classical random variable in a quantum state.

PROPOSITION 10. *Let X be a random variable with values in $\mathcal{X}$ and distribution $\mathbf{p}$, and let $\{|\psi_x\rangle | x \in \mathcal{X}\}$ be an ensemble of pure states. Consider the eigenvalues $\boldsymbol{\lambda}$ of the density matrix*

$$\rho = \sum_{x \in \mathcal{X}} p_x \, |\psi_x\rangle \, \langle \psi_x|$$

Then $\boldsymbol{\lambda} \succsim \mathbf{p}$, with similarity $\boldsymbol{\lambda} \sim \mathbf{p}$ only when the ensemble $\{|\psi_x\rangle\}$ is orthogonal.

PROOF. *We know that:*

$$\begin{aligned} \rho &= \sum_{x \in \mathcal{X}} p_x \, |\psi_x\rangle \, \langle \psi_x| \\ &= \sum_{x \in \mathcal{X}} |\eta_x\rangle \, \langle \eta_x| \ , \end{aligned}$$

where $|\eta_x\rangle := \sqrt{p_x} \, |\psi_x\rangle$. However, we can also write ρ in the eigenbasis:

$$\begin{aligned} \rho &= \sum_{i=1}^{d} \lambda_i \, |i\rangle \, \langle i| \\ &= \sum_{i=1}^{d} |\theta_i\rangle \, \langle \theta_i| \ , \end{aligned}$$

where $|\theta_i\rangle := \sqrt{\lambda_i} \, |i\rangle$. Then the two sets of vectors can be related via:

$$|\eta_x\rangle = \sum_{i=1}^{d} U_{xi} \, |\theta_i\rangle \ ,$$

where U_{xi} is a $|\mathcal{X}| \times d$ matrix comprised of d rows of orthonormal $|\mathcal{X}|$-dimensional vectors [74]. Now, we have:

$$\begin{aligned} p_x &= \langle \eta_x | \eta_x \rangle \\ &= \sum_{i=1}^{d} |U_{xi}|^2 \lambda_i \ . \end{aligned}$$

Note that U_{xi} is not necessarily square, but we can take $\lambda_i = 0$ for $i > d$, and simply extend U_{xi} into a square unitary matrix by filling out the bottom $|\mathcal{X}| - d$ rows with more orthonormal vectors.

This leaves the equation unchanged. We can then write:

$$p_x = \sum_{i=1}^{n} |U_{xi}|^2 \lambda_i \ .$$

Then by the Schur-Horn criterion from Thm. 11, $\boldsymbol{\lambda} \succsim \mathbf{p}$*. Similarity (*$\mathbf{p} \succsim \boldsymbol{\lambda}$*) requires that* $|U_{xi}|^2$ *be a permutation matrix, and this only occurs if there is a one-to-one mapping between* $\mathcal{X}$ *and the eigenvectors of* ρ*, so that the* $|\psi_x\rangle$ *are orthogonal.*

The implication of this fact is that, for any entropy, we have $\mathrm{H}_{\alpha,\mathrm{q}}[\rho] \leq \mathrm{H}_\alpha[X]$. In other words, we can only lose entropy by encoding a variable in a quantum state. Additionally, the reason for this loss is evidently the non-orthogonality of the embedding states. There is an intuitive logic to this: non-orthogonal states cannot be fully distinguished by quantum measurement, and so information encoded in this way is fundamentally unrecoverable. The reduction in entropy reflects this loss.

A direct result of this is the following:

Corollary 5. *Let* $\mathfrak{E}(\mu)$ *be the* ϵ*-machine of a process* μ*, and let* $\mathfrak{Q}$ *be any q-machine which embeds* $\mathfrak{E}(\mu)$*, with stationary state* ρ_π*. Then for all* α*,* $\mathrm{H}_{\alpha,\mathrm{q}}[\rho_\pi] \leq C_\mu^{(\alpha)}$*, with equality only if the embedding states* $|\psi_s\rangle$ *are orthogonal.*

Proof. *This simply follows from Prop. 10 and the fact that* $\rho_\pi = \sum_s \pi_s |\psi_s\rangle \langle\psi_s|$*, where* π_s *is the stationary distribution of the* ϵ*-machine.*

The implication of this theorem is that every q-machine has a kind of memory advantage over the ϵ-machine: they can generate the same process with less memory. Further, since state entropy is unchanged under time-reversal, this also implies that every reverse q-machine has a lower memory than the reverse ϵ-machine.

This memory compression does not come freely, however. The consequence is the non-orthogonality of the embedded states, which means they cannot be recovered by measurement without disrupting the system. Specifically, q-machines cannot be used as predictors, despite the fact that they are a deterministic embedding of the past (which in the classical case is sufficient to guarantee unifilarity and thus predictivity).

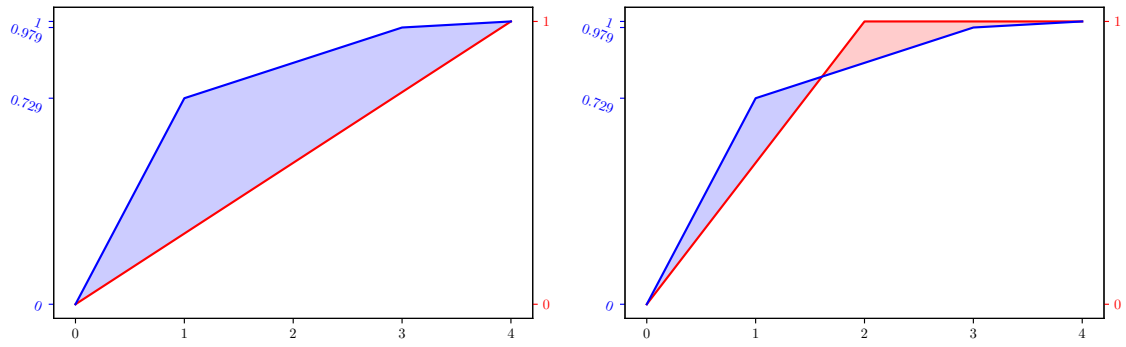

FIGURE 5.5. (*Left*) Lorenz curves for the 4-state MBW ϵ-machine $\mathfrak{M}_4$ and the associated q-machine $\mathfrak{Q}_4$. (*Right*) Lorenz curves for the 4-state MBW q-machine $\mathfrak{Q}_4$ and a dimensionally-smaller model $\mathfrak{D}_4$.

Generally speaking, however, the reduction in memory is quite significant, and thus far it appears that q-machines frequently outperform *non*-predictive classical generators in memory reduction as well.

One may reasonably wonder if there exists, among q-machines (since there are many, depending on phase choices), a single *strongly minimal* q-machine, in the same way that the ϵ-machine strongly minimized all predictors. Let us reconsider two examples—the 4-MBW process and the 3-MBW process—to get an intuition for this. For clarity, when referring to the entropy of the stationary state of two different models of the same process, we will use the model variable as the entropy argument.

First, let's examine the majorization between $\mathfrak{Q}_4$ and the Markov model via the Lorenz curves of $\boldsymbol{\lambda}$, the eigenvalues of ρ_π, and the stationary state of the Markov chain. See Fig. 5.5. This demonstrates the "strong advantage" indicated by Cor. 5.

Figure 5.5 compares the Lorenz curve of its stationary eigenvalues $\boldsymbol{\lambda}'$ to those of $\mathfrak{Q}_4$. One sees that it does not majorize the q-machine, but it does have a lower statistical memory: $\mathrm{H_q}[\mathfrak{D}_4] = 1.0$ and $\mathrm{H_q}[\mathfrak{Q}_4] \approx 1.2$ bit. (On the other hand, the q-machine has a smaller min-memory, with $\mathrm{H_{min}}[\mathfrak{D}_4] = 1.0$ and $\mathrm{H_{min}}[\mathfrak{Q}_4] \approx 0.46$.)

We can also examine the majorization between the q-machine and ϵ-machine of the 3-MBW process by plotting the Lorenz curves of $\boldsymbol{\lambda}$, the eigenvalues of ρ_π, and the stationary state of the Markov

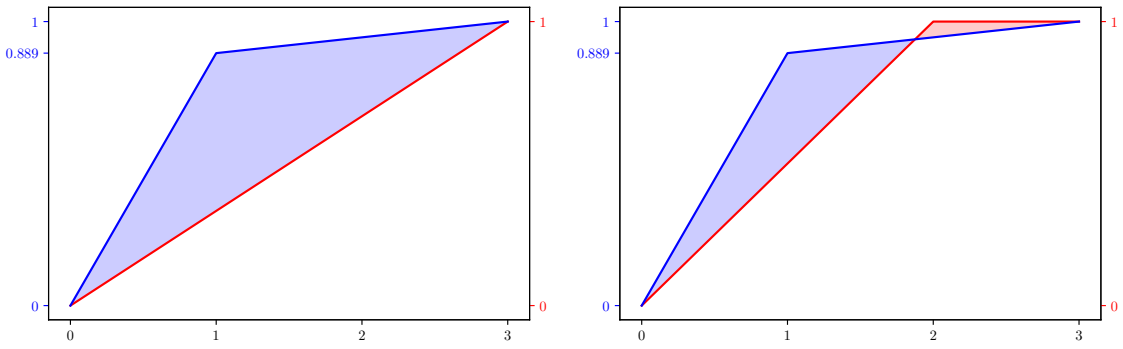

FIGURE 5.6. (*Left*) Lorenz curves for the 3-state MBW ϵ-machine $\mathfrak{M}_3$ and the associated q-machine $\mathbf{Q}_3$. (*Right*) Lorenz curves for the 3-state MBW q-machine, $\mathfrak{Q}_3$ and a dimensionally-smaller model $\mathfrak{D}_3$.

chain, shown in Fig. 5.6. Again, we see a majorization curve indicating strong memory advantage over the ϵ-machine.

Figure 5.6 compares the Lorenz curve of its stationary eigenvalues $\boldsymbol{\lambda}'$ to that of $\mathfrak{Q}_3$. We see that it does not majorize $\mathfrak{Q}_3$. And, this time, this is directly manifested by the fact that the smaller-dimension model has a larger entropy: $\mathrm{H_q}[\mathfrak{D}_3] = 1.0$ and $\mathrm{H_q}[\mathfrak{Q}_3] \approx 0.61$ bit.

The reader should note by now that none of the examples covered above are strong minima among q-machines. One way to prove that no strong minimum exists for, say, the 3-state MBW process requires showing that there does not exist *any other* quantum model in 2 dimensions that generates the process. This would imply that no other model can majorize $\mathfrak{D}_3$. But we already did this in Example 10!

COUNTEREXAMPLE (Weak Minimality of $\mathfrak{D}_3$). *The quantum model $\mathfrak{D}_3$ weakly minimizes topological complexity for all quantum generators of the 3-state MBW Process; consequently, the 3-state MBW Process has no strongly minimal quantum model.*

We have therefore demonstrated that while the ϵ-machine is a strong minimum among its own class of predictive models, q-machines do not appear to have any strong minima among themselves; nevertheless, they all provide strong advantage over the ϵ-machine in terms of memory costs for generation of the process.

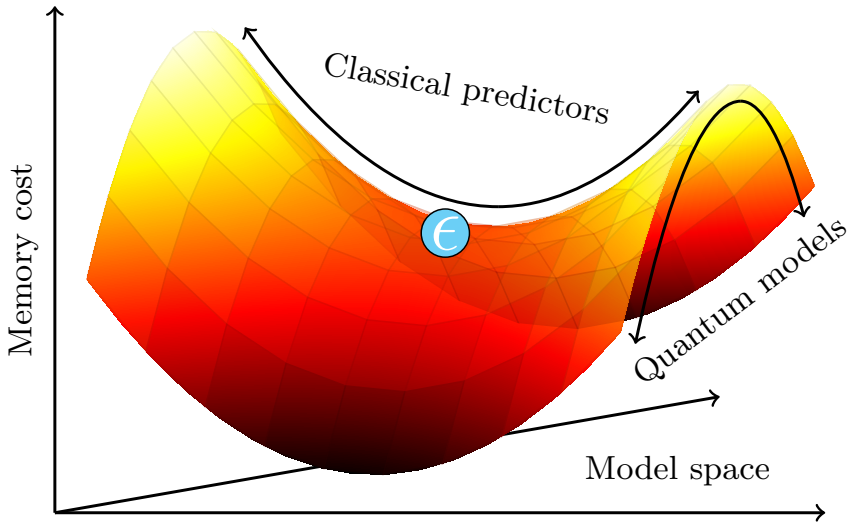

FIGURE 5.7. Proposed majorization saddle structure of model-space: The ϵ-machine (labeled ϵ) is located at a saddle-point with respect to majorization, where classical deviations (state-splitting) move up the lattice and quantum deviations (utilizing state overlap) move down the lattice.

5.5. Discussion

Our results on quantum models of processes complement our previous chapter's results on the memory costs of predictive models; namely:

(1) The ϵ-machine majorizes all classical predictive models of the same process and so simultaneously minimizes many different measures of memory cost.

(2) The q-machine, and indeed any quantum realization of the ϵ-machine, always majorizes the ϵ-machine, and so simultaneously improves on all the measures of memory cost.

(3) For at least one process, there does not exist any quantum pure-state model that majorizes all quantum pure-state models of that process. Thus, while an ϵ-machine may be improved upon by different possible quantum models, there is not a unique one quantum model that is unambiguously the "best" choice.

Imagining the ϵ-machine as an invariant "saddle-point" in the majorization structure of model-space, Fig. 5.7 depicts the implied geometry. That is, we see that despite its nonminimality among all models, the ϵ-machine still occupies a topologically important position in model-space—one that is

invariant to one's choice of memory measure. However, no similar model plays the topologically minimal role for quantum pure-state models.

The quantum statistical complexity C_q has been offered up as an alternative quantum measure of structural complexity—a rival of the statistical complexity C_μ [189]. One implication of our results here is that the nature of this quantum minimum C_q is fundamentally different than that of C_μ. This observation should help further explorations into techniques required to compute C_q and the physical circumstances in which it is most relevant.

That the physical meaning of C_q—as a von Neumann entropy—involves generating an asymptotically large number of realizations of a process may imply that it cannot be accurately computed by only considering machines that generate a single realization. This is in contrast to C_μ which, being strongly minimized, must be attainable in the single-shot regime along with measures like $C_\mu^{(0)}$ and $C_\mu^{(\infty)}$.

In this way, the quantum realm again appears ambiguous. Ambiguity in structural complexity has been previously observed in the sense that there exist pairs of processes such that one process may have a larger C_μ than the other while having a smaller C_q [5]. The classical and quantum paradigms for modeling can disagree on simplicity—there is no universal Ockham's Razor. How this result relates to strong versus weak optimization deserves further investigation.

In general, what all these considerations tell us is that quantum models and q-machines occupy an important but critically distinct role in the space of all models of a classical process. They are in many ways incomparable to other classes of models, such as predictors and retrodictors. A q-machine, like a unifilar predictor, operates "deterministically" (pure state to pure state), but due to its non-orthogonality cannot practically function as one.

Further, while the class of predictors is organized hierarchically by state-merging relations, q-machines are far more egalitarian, organized by the choice of phase parameters ϕ_{xs}, which imbues the space of q-machines for a given process with the topology of a multi-dimensional torus. The dimensionality of this torus is determined by gauge invariants, which are induced by a kind of solenoidal action on the dynamical state space of the embedded ϵ-machine.

There is a fascinating structure to q-machine space which is as elegant as it is frustrating, for it does not offer any clear direction to a "universally" advantageous q-machine, but instead offers

us potentially different models for different practical considerations. A proper understanding of quantum models of stochastic processes will, I think, require continued patience and thorough analysis. Here especially Wiener's maxim that there are "no answers, only cross-references" should be kept at the forefront: we can only understand the full potential of q-machines by continuing to relate them to, and distinguish them from, each other and our classical intuitions.

The methods and results here should also be extended to analyze more general (*i.e.* non-predictive or non-retrodictive) classical generative models which, in many ways, bear resemblances in their functionality to the quantum models [**104**, **105**, **161**]. These drop the requirement of unifilarity, similar to how the quantum models relax the notion of orthogonality. Important questions to pursue in this vein are whether generative models are strongly maximized by the ϵ-machine and whether they have their own strong minimum or, like the quantum models, only weak minima in different contexts.

CHAPTER 6

Forgetful demons: Heat extraction with quantum simulation

Heat may be generated and destroyed by certain processes,
and this shows that heat is not a substance.
James Clerk Maxwell, *Theory of Heat*

6.1. Introduction

Richard Feynman [85] broached the notion that quantum computers would be singularly useful for the simulation of quantum processes, without supposing that this would also make them advantageous at simulating classical processes. Quantum information and quantum advantage have recently benefited from the study of, on the one hand, quantum *memory compression* [17, 66, 102, 106, 121, 160, 191, 192], particularly for simulating stochastic processes, and, on the other, quantum *thermodynamics* [30, 45, 48, 55, 71, 116, 117, 206]. As a complement to their independent contributions, here we explore the thermal efficiency of quantum memory compression in physical implementations, illustrating a fruitful new cross over that elucidates how physical systems generate and process information.

Quantum computational mechanics recently explored how to simulate and transform *classical* stochastic processes using *quantum* systems [17, 66, 121, 192]. Generally, quantum simulators of complex processes require less memory (measured by the quantum-state von Neumann entropy) than classical (measured by the statistical complexity—the classical-state Shannon entropy) [160, 191]. While this quantum advantage holds for all memory metrics, from the single-shot to the asymptotic [106], here an important contrast with the classical case arises: There is no quantum equivalent to the ϵ-machine that simultaneously minimizes all metrics. Rather each process has a family of quantum simulators that may each be relevant in different settings—some favorable in the asymptotic regime, with others favorable in the single-shot [102, 106].

Quantum thermodynamics [206], though recently advancing via thermal resource theories [30, 71, 116, 117] and single-shot thermodynamics [29, 45, 48, 55], has not yet been applied to examine quantum simulators. However, it is known that Landauer's lower bound, as given in the form of Shannon and von Neumann entropies, is not generally attainable—a more nuanced view is necessary [29, 45]. As in quantum computational mechanics, transitioning from classical to quantum regimes leads to a sharp separation between single-shot and asymptotic settings.

Using the laws of nonequilibrium thermodynamics, it is possible for physical systems to extract useful work from heat baths while generating samples of some stochastic process. This is a consequence of the *information processing Second Law* (IPSL) [24], which describes the minimal cost of transforming a given stochastic process into another, and itself relies on Landauer's principle for memory erasure [97]. Previous work focusing on classical generators of stochastic processes has extended the IPSL to the *thermodynamics of modularity*, which describes the thermodynamic costs of operating locally on separate parts of a system, and provides more realistic bounds for the work extraction of specific physical generators [27].

Here, we explore issues raised by the recent developments in quantum computing, focusing on the problem of simulating classical stochastic processes via stochastic and quantum computers. To link quantum memory compression with its associated thermodynamics, we calculate upper bounds on the work cost of quantum implementations of classical simulators. Functionally, these become lower bounds on the work that can be extracted while generating a process. These bounds—achievable in the asymptotic limit of parallel generation—mirror classical results and show a direct relationship between memory compression achieved by a quantum implementation and the change in extractable work via the same. We also demonstrate the existence of a natural partial ordering on generators which is monotonic in work extraction rate; that is, this mathematical ordering is also an ordering of thermodynamic efficiency.

The physical setting of our work is in the realm of *information reservoirs*—systems all of whose states have the same energy level. Landauer's Principle for quantum systems says that to change an information reservoir A from state ρ_A to state ρ'_A requires a work cost satisfying the lower bound:

$$W \geq k_{\mathrm{B}} T \ln 2 \left(\mathrm{H}_{\mathrm{q}}[\rho_A] - \mathrm{H}_{\mathrm{q}}[\rho'_A] \right) \ . \tag{6.1}$$

where $\mathrm{H_q}[\rho_A]$ is the von Neumann entropy [141]. Note that the lower bound

$$W_{\min} := k_{\mathrm{B}} T \ln 2 \left(\mathrm{H_q}[\rho_A] - \mathrm{H_q}[\rho'_A] \right)$$

is simply the change in free energy for an information reservoir. Further, due to an information reservoir's trivial Hamiltonian, all of the work W becomes heat Q. Then the total entropy production—of system and environment—is:

$$\begin{aligned} \Delta S &:= Q + k_{\mathrm{B}} T \ln 2 \Delta \, \mathrm{H}[A] \\ &= W - W_{\min} \ . \end{aligned} \tag{6.2}$$

Thus, not only does Landauer's Principle provide the lower bound, but reveals that any work exceeding $W_{\min}$ represents dissipation.

The efficiency ordering among generators points directly to a particular class of generators as the optimally efficient, achieving the Landauer rate: these are classical retrodictive states. We prove that any generator not in this class, including quantum compressions of retrodictive generators, has a nonzero modularity dissipation over and above the Landauer cost. In other words, using quantum computers to simulate classical processes typically requires nonzero thermodynamic cost, while stochastic computers can theoretically achieve zero cost in simulating classical processes. This supports the viewpoint originally put forth by Feynman—that certain types of computers would each be advantageous at simulating certain physical processes—which challenges the current claims of quantum supremacy. Furthermore, we show that in both classical and quantum simulations, thermodynamic efficiency places a lower bound on the required memory of the simulator.

To accomplish these results, we must prove a new theorem on the thermodynamic efficiency of local operations. Correlation is a resource: it has been investigated as such, in the formalism of *resource theories* [35] such as that of local operations with classical communication [73], with public communication [132], and many others, as well as the theory of local operations alone, under the umbrella term of *common information* [59, 95, 214]. Correlations have long been recognized as a thermal resource [25, 26, 97, 103], enabling efficient computation to be performed when taken properly into account. Local operations that act only on part of a larger system are known to

never increase the correlation between the part and the whole; most often, they are destructive to correlations and therefore resource-expensive.

Thermodynamic dissipation induced by a local operation—say on system A of a bipartite system AB to make a new joint system CB—is classically proportional to the difference in mutual informations [27]:

$$\Delta S_{\text{loc}} = k_{\text{B}}T \left(\text{I}[A : B] - \text{I}[C : B] \right) \ .$$

This can be asymptotically achieved for quantum systems [107]. By the data processing inequality [36, 141], it is always nonnegative: $\Delta S_{\text{loc}} \geq 0$. Optimal thermodynamic efficiency is achieved when $\Delta S_{\text{loc}} = 0$.

To identify the conditions, in both classical and quantum computation, when this is so, we draw from prior results on saturated information-theoretic inequalities [68, 136, 137, 149, 150, 151, 166]. Specifically, using a generalized notion of quantum sufficient statistic [78, 98, 118, 151], we show that a local operation on part of a system is efficient if and only if it unitarily preserves the minimal sufficient statistic of the part for the whole. Our geometric interpretation of this also draws on recent progress on fixed points of quantum channels [8, 14, 33, 67].

Paralleling previous results on ΔS_{loc} [27], our particular interest in locality arises from applying it to thermal transformations that generate and manipulate stochastic processes. This is the study of *information engines* [24, 25, 26, 60, 62, 122]. Previous work explored optimal conditions for a classical information engine to generate a process. Working from the hidden Markov model (HMM) [199] that determines an engine's memory dynamics, it was conjectured that the HMM must be *retrodictive* to be optimal. For this to hold, the current memory state must be a sufficient statistic of the *future* data for predicting the *past* data [27].

Employing a general result on conditions for reversible local computation, the following confirms this conjecture, in the form of an equivalent condition on the HMM's structure. We then extend this, showing that it holds for quantum generators of stochastic processes. Notably, quantum generators are known to provide potentially unbounded advantage in memory storage when compared to classical generators of the same process. Surprisingly, the advantage is contingent: optimally-efficient generators—those with $\Delta S_{\text{loc}} = 0$—must not benefit from any memory compression. We

show this to be true not only for previously published quantum generators, but for a new family of quantum generators as well, derived from time reversal [37, 40, 50, 191].

Combining our two major results in this chapter—the positive relationship between *embedding* and efficiency, and the negative relationship between *compression* and efficiency—one concludes that a quantum-compressed generator is efficient with respect to the generator it compresses but, to the extent that it is compressed, it cannot be optimally efficient. In short, only classical retrodictive generators achieve the lower bound dictated by the IPSL. Practically, this highlights a pressing need to experimentally explore the thermodynamics of quantum computing.

This chapter is a synthesis of material on the thermodynamics of quantum generators from the publications *Thermal Efficiency of Quantum Memory Compression* [107], and *Thermodynamically Efficient Local Computation* [108]. Our result on embedding and efficiency, Theorem 14, is a generalization of the result we originally derived in Ref. [107], and similarly we have greatly generalized our results from Ref. [108] in Theorem 19, which applies to all quantum generators, though our original publication had only proven the same result for q-machines and reverse q-machines.

6.2. Energy and information

To analyze the thermodynamics of physical generators, we must establish rules that circumscribe what we consider physically allowed and the correspondence to thermodynamic quantities such as work and heat.

Here, we consider the *resource theory of thermal operations* [30, 71]. Generally, on a quantum system S we allow operations of the form:

$$\mathcal{E}\left(\rho_S\right) := \mathrm{Tr}_B\left(U\rho_S \otimes \frac{e^{-\beta H_B}}{Z_B} U^\dagger\right) , \tag{6.3}$$

where A and B are auxiliary systems with Hamiltonians H_A and H_B, B a thermal bath, and U acts on the joint Hilbert space of $\mathcal{H}_A$ and $\mathcal{H}_B$. The unitary operator U satisfies the rule of *microscopic* conservation of energy, where we constrain $[U, H_S + H_B] = 0$.

In the special case of classical operations, thermal operations reduce to stochastic operations $\mathbf{T}$ which satisfy $\mathbf{T}\boldsymbol{\gamma} = \boldsymbol{\gamma}$, where $\boldsymbol{\gamma}$ is the Gibbs distribution. This property is called *Gibbs-stochasticity*.

As discussed in Section 1.5.2, this approach leads us into theory of relative majorization, which can be used to determine the work costs for transforming one distribution into another [157]. Using Hamiltonian control (which is a special case of relative majorization methods), Ref. [27] showed that *any* stochastic channel can be implemented in a way which achieves the Landauer bound. That is, applying a channel $\Pr(y|x)$ to a random variable X, resulting in Y, can be performed with the work cost $W = k_{\mathrm{B}}T\left(\mathrm{H}[X] - \mathrm{H}[Y]\right)$.

The quantum regime is unfortunately not quite so simple; here the single-shot costs for applying a stochastic channel rarely have simple expressions in terms of Shannon or von Neumann entropies, and we must use entropies more suited to the single-shot domain in order to understand the complex constraints of quantum nonequilibrium thermodynamics [29].

6.2.1. A single-shot form of Landauer's bound. We have previously (Sec. 4.3) discussed the min-entropy $\mathrm{H}_{\min,\mathrm{q}}[\cdot]$ and Rényi 1/2-entropy $\mathrm{H}_{1/2,\mathrm{q}}[\cdot]$, which in the literature of quantum information and thermodynamics is frequently called the *max-entropy*. There is an important generalization of both these quantities to the bipartite domain. For two systems A and B with joint state ρ_{AB}, the min- and max-entropies are given by:

$$\begin{aligned}
\mathrm{H}_{\min}\left[A|B\right]_{\rho} &\equiv \min_{\sigma_B} \sup\{\lambda : \rho_{AB} \leq 2^{-\lambda} 1_A \otimes \sigma_B\} \\
\mathrm{H}_{1/2}\left[A|B\right]_{\rho} &\equiv \max_{\sigma_B} 2\log_2 F\left(\rho_{AB}, 1_A \otimes \sigma_B\right) \ ,
\end{aligned}$$

where $F(\rho,\sigma) = \mathrm{Tr}\left(\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}}\right)$ is the *fidelity*. The smooth entropies are optimizations of these quantities over all $\tilde{\rho}_{AB}$ within the ϵ-ball $B_\epsilon(\rho_{AB})$; that is, all states such that $\sqrt{1 - F\left(\tilde{\rho}_{AB}, \rho_{AB}\right)} < \epsilon$:

$$\begin{aligned}
\mathrm{H}^{\epsilon}_{\min}\left[A|B\right] &\equiv \max_{\tilde{\rho}_{AB}} \mathrm{H}_{\min}\left[A|B\right]_{\tilde{\rho}} \\
\mathrm{H}^{\epsilon}_{1/2}\left[A|B\right] &\equiv \min_{\tilde{\rho}_{AB}} \mathrm{H}_{1/2}\left[A|B\right]_{\tilde{\rho}} \ .
\end{aligned}$$

When B is decoupled from A, $\rho_{AB} = \rho_A \otimes \rho_B$, the resulting quantities are independent of B and so we have the marginal smooth entropies $\mathrm{H}^{\epsilon}_{\min}\left[A\right]$ and $\mathrm{H}^{\epsilon}_{\min}\left[A\right]$.

We import the following result from Ref. [48]: Given a system S correlated with an auxiliary A, and any $\epsilon > 0$, there is a procedure for erasing A while preserving S, with probability of failure ϵ,

which has a work cost of no more than:

$$\frac{W}{k_{\mathrm{B}}T\ln 2} \leq \mathrm{H}_{1/2}^{\epsilon^2/16}\left[A|S\right] + O\left(\log\frac{1}{\epsilon}\right) . \tag{6.4}$$

We use this to prove a generalization of the "detailed" Landauer cost. Suppose we have a quantum channel $\mathcal{E}$ we wish to implement and we do so on a system S with average state ρ_S. The target state is $\rho'_S = \mathcal{E}\left(\rho_S\right)$. We perform the map in the following way. Using the Stinespring dilation of $\mathcal{E}$, we couple S to an auxiliary system A in state $|0\rangle\langle 0|_A$ and perform a unitary operation on both systems:

$$\rho'_{SA} = U_{AB}\rho_S \otimes |0\rangle\langle 0|_A U_{AB}^{\dagger} ,$$

such that $\mathcal{E}\left(\rho_S\right) = \mathrm{Tr}_A\left(\rho'_{SA}\right)$. At the end of the procedure we must erase A. This can be done with cost Eq. (6.4). This form of the cost for implementing a channel is given in Ref. [**55**].

Now, we utilize a result on smooth entropies that generalizes the chain rule on von Neumann entropy [**207**]. We state two somewhat streamlined versions of the theorem here. For any $\delta > 0$ and systems AB:

$$\mathrm{H}_{1/2}^{\delta}\left[B|A\right] \leq \mathrm{H}_{1/2}^{4\delta}\left[AB\right] - \mathrm{H}_{\min}^{\delta}\left[A\right] + O\left(\log\frac{1}{\delta}\right) \tag{6.5}$$

$$\mathrm{H}_{\min}^{\delta}\left[B|A\right] \leq \mathrm{H}_{\min}^{4\delta}\left[AB\right] - \mathrm{H}_{\min}^{\delta}\left[A\right] + O\left(\log\frac{1}{\delta}\right) \tag{6.6}$$

Applying (6.5) to (6.4), we have:

$$\frac{W}{k_{\mathrm{B}}T\ln 2} \leq \mathrm{H}_{1/2}^{\epsilon^2/4}\left[S'A'\right] - \mathrm{H}_{\min}^{\epsilon^2/16}\left[S'\right] + O\left(\log\frac{1}{\epsilon}\right)$$

However, $\mathrm{H}_{1/2}^{\epsilon^2/4}\left[S'A'\right] = \mathrm{H}_{1/2}^{\epsilon^2/4}\left[S\right]$ by unitary equivalence, so we have the erasure cost:

$$\frac{W}{k_{\mathrm{B}}T\ln 2} \leq \mathrm{H}_{1/2}^{\epsilon^2/4}\left[S\right] - \mathrm{H}_{\min}^{\epsilon^2/16}\left[S'\right] + O\left(\log\frac{1}{\epsilon}\right) . \tag{6.7}$$

Since we can perform the initial unitary with no work, this is the only work cost involved in implementing the channel. To summarize: The channel $\mathcal{E}$ can be performed on the system S with a work cost not exceeding Eq. (6.7).

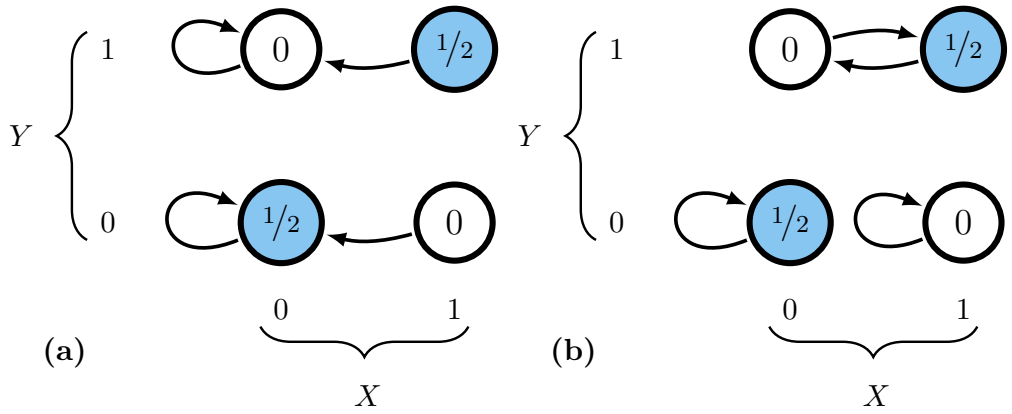

FIGURE 6.1. Thermodynamics of locality: Suppose we have two bits XY in a correlated state where $1/2$ probability is in $XY = 00$ and $1/2$ probability is in $XY = 11$. **(a)** A thermodynamically irreversible operation can be performed to erase only X (that is, set $X = 0$ without changing Y) if we are not allowed to use knowledge about the state of Y. **(b)** A reversible operation can be performed to erase X if we are allowed to use knowledge about Y. Both operations have the same outcome given our initial condition, but the nonlocal operation **(a)** is more thermodynamically costly because it is irreversible. According to Thm. 17, operation **(a)** is costly since it erases information in X that is correlated with Y.

Now, suppose we choose instead to implement parallel generation of our process. That is, we have N independent systems on which we want to implement N independent copies of the channel $\mathcal{E}$ with probability of error less than $\epsilon > 0$. Naturally, the work cost becomes:

$$\frac{W}{k_{\mathrm{B}}T\ln 2} \leq \mathrm{H}_{1/2}^{\epsilon^2/4}\left[S^{\otimes N}\right] - \mathrm{H}_{\min}^{\epsilon^2/16}\left[S'^{\otimes N}\right] + O\left(\log\frac{1}{\epsilon}\right) .$$

Significantly, the error term does not depend on N. When we further account for the Asymptotic Equipartition Theorem of smooth entropies, we have the remarkable result for the work rate:

$$\frac{W}{Nk_{\mathrm{B}}T\ln 2} \leq \mathrm{H}_{\mathrm{q}}[S] - \mathrm{H}_{\mathrm{q}}[S'] + O\left(\sqrt{\frac{1}{N}\log\frac{1}{\epsilon}}\right) . \tag{6.8}$$

With Landauer's bound sandwiching the work from below, we find a tight result on the achievable work cost. By scaling error with N, for instance $\epsilon \sim 2^{-\sqrt{N}}$, Landauer's bound can, in the limit $N \to \infty$, be achieved for quantum channels. In the single-shot regime, the bound of Eq. (6.7) gives us a somewhat less certain range of achievability.

6.2.2. Thermodynamics of modularity. To derive useful results, we must place further constraints on the system dynamics to see how Landauer's bound is affected. Reference [27] introduced the following perspective. Consider a bipartite information reservoir AB, on which we wish to apply the local channel $\mathcal{E} \otimes I_B$, where $\mathcal{E} : \mathcal{B}(\mathcal{H}_A) \to \mathcal{B}(\mathcal{H}_C)$ maps the states of system A into those of system C, transforming the initial joint state ρ_{AB} to the final state ρ_{CB}. The Landauer bound for $AB \to CB$ is given by $W_{\text{min}} = k_{\text{B}}T \ln 2 \left(\text{H}[\rho_{AB}] - \text{H}[\rho_{CB}]\right)$. However, since we constrained ourselves to use local manipulations, the lowest achievable bound is actually $W_{\text{loc}} := k_{\text{B}}T \ln 2 \left(\text{H}[\rho_A] - \text{H}[\rho_C]\right)$. Thus, we must have dissipation of at least:

$$\begin{aligned} \Delta S &\geq W_{\text{loc}} - W_{\text{min}} \\ &= k_{\text{B}}T \ln 2 \left(\text{H}[\rho_A] - \text{H}[\rho_{AB}] - \text{H}[\rho_C] + \text{H}[\rho_{CB}]\right) \\ &= k_{\text{B}}T \ln 2 \left(\text{I}[A:B] - \text{I}[C:B]\right) \ . \end{aligned} \tag{6.9}$$

where $\text{I}[A:B] = \text{H}[\rho_A] + \text{H}[\rho_B] - \text{H}[\rho_{AB}]$ is the quantum mutual information. And so, we have a minimal *locality dissipation*:

$$\Delta S_{\text{loc}} := k_{\text{B}}T \ln 2 \left(\text{I}[A:B] - \text{I}[C:B]\right) \ , \tag{6.10}$$

which arises because we did not use the correlations to facilitate our erasure. See Fig. 6.1 for a simple example of this phenomenon.

This local form of Landauer's Principle is still highly general, but the following shows how to examine it for specific classical and quantum computational architectures The key question we ask is: For which architectures can ΔS_{loc} be made to exactly vanish? We first we consider this problem generally and then provide a solution.

6.3. Thermodynamics of generators

Generators, as we have defined them, are typically a triplet $\mathfrak{G} = (\mathcal{S}, \mathcal{X}, \{\mathbf{T}^{(x)}\})$ of a state space $\mathcal{S}$, an alphabet $\mathcal{X}$, and a stochastic matrix $\mathbf{T}^{(x)} = (T^{(x)}_{s's})$ which describes the probability, when starting in state $s \in \mathcal{S}$, of transitioning to the state s' and emitting a symbol $x \in \mathcal{X}$. To think of generators as thermodynamic objects, we need a more concrete picture. One such approach is to visualize generators as writing on a tape while continually erasing and rewriting their internal

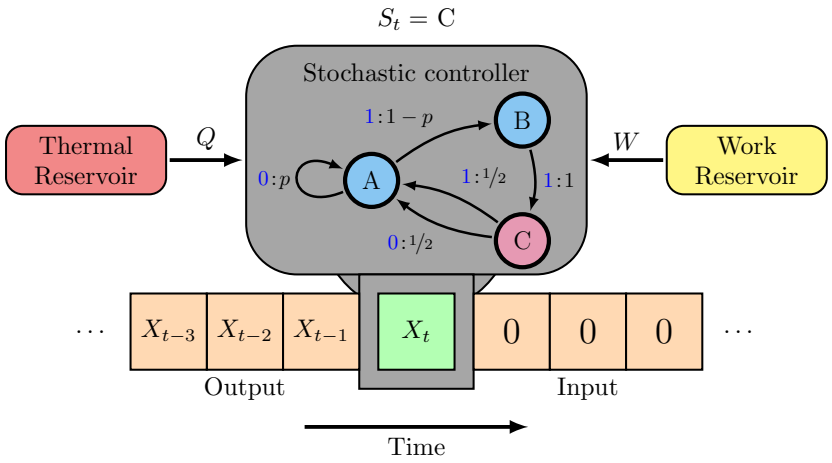

FIGURE 6.2. Information ratchet sequentially generates a symbol string on an empty tape: At time step t, S_t is the random variable for the ratchet state. The generated symbols in the output process are denoted by $X_{t-1}, X_{t-2}, X_{t-3}, \ldots$. The most recently generated symbol X_t (green) is determined by the internal dynamics of the ratchet's memory, using heat Q from the thermal reservoir as well as work W from the work reservoir. *(Ratchet interior.)* The memory dynamics and symbol production are governed by the conditional probabilities $\Pr(s_{t+1}, x_t|s_t)$, where s_t is the current state at time t, x_t is the generated symbol and s_{t+1} is the new state. Diagrammatically, this is a hidden Markov model—a labeled, directed graph in which nodes are states s and edges represent transitions $s \to s'$ labeled by the emitted symbol and associated probability: $x : \Pr(s', x|s)$.

memory Fig. 6.2. To accomplish these tasks, they must exchange energy with a work reservoir and heat bath.

Erasure generally requires work, drawn from the work reservoir, while the creation of noise often allows the extraction of work, which is represented in our sign convention by drawing negative work from the reservoir. Producing a process $X_1 \ldots X_t \sim \Pr(x_1 \ldots x_t)$ of length t has an associated work cost $W \geq -k_{\mathrm{B}} T \ln 2 H(X_1 \ldots X_t)$. The negative sign, as discussed, indicates work $k_{\mathrm{B}} T \ln 2 \; H(X_1 \ldots X_t)$ may be transferred from the thermal reservoir to the work reservoir. For large t, this can be asymptotically expressed by the work rate $W/t \geq -k_{\mathrm{B}} T \ln 2 \; h_\mu$, where:

$$h_\mu := \lim_{t \to \infty} \frac{1}{t} \mathrm{H}[X_1 \ldots X_t] \tag{6.11}$$

is the process' *Kolmogorov-Sinai entropy rate* [24]. This is a reasonable description of the average entropy rate of a process that is *stationary*—that is, $\Pr(X_t \ldots X_{t+\ell} = x_1 \ldots x_{\ell-1})$ is independent of

t—and *ergodic*. Said differently, for large t a typical realization $x_1 \dots x_t$ contains the word $\widehat{x}_1 \dots \widehat{x}_\ell$ approximately $t \times \Pr(\widehat{x}_1 \dots \widehat{x}_\ell)$ times. Recurrent generators produce exactly these sorts of processes.

6.3.1. Thermodynamic implementations of quantum generators. For quantum generators $\mathfrak{Q} = (\mathcal{H}, \mathcal{X}, \left\{ \mathcal{E}^{(x)} \right\})$, we can offer quite a specific strategy for their implementation which is well-suited to the resource theory of thermal operations. This approach involves a memory space $\mathcal{H}_S$, symbol space $\mathcal{H}_X$, auxiliary space $\mathcal{H}_A$, and bath space $\mathcal{H}_B$; and a unitary acting on $\mathcal{H}_S \otimes \mathcal{H}_X \otimes \mathcal{H}_A \otimes \mathcal{H}_B$, such that the channel:

$$\mathcal{T}_{SX}\left(\rho_{SX}\right) := \mathrm{Tr}_{AB}\left(U \rho_{SX} \otimes |0\rangle\langle 0|_A \otimes \rho_B U^\dagger\right)$$

satisfies:

$$\mathcal{T}_{SX}\left(\rho_S \otimes |0\rangle\langle 0|_X\right) = \sum_x \mathcal{E}^{(x)}(\rho_S) \otimes |x\rangle\langle x|_X \quad . \tag{6.12}$$

Suppose that there are Hamiltonians H_S, H_X, H_A, H_B for each system such that $\rho_B = Z_B^{-1} \exp\left(-\beta H_B\right)$ is the Gibbs distribution of its Hamiltonian. If $[U, H_S + H_X + H_A + H_B] = 0$, then we say that the implementation is *thermal*, as these implementations are those allowed by the resource theory of thermal operations.

In quantum mechanics, the rule of microscopic conservation $[U, H_S + H_X + H_A + H_B] = 0$ brings coherence with respect to the Hamiltonian into play as a resource [**116**, **117**]. The type of systems we consider here are what are often, in the literature of information engines, called *information reservoirs*: systems whose Hamiltonian is trivially flat, so that energetics does not play a direct role in their dynamics. On such systems, tracking coherence is no longer at issue, as all operators commute with a flat Hamiltonian.

The most well-developed class of quantum generators are the q-machine and reverse q-machine [**17**, **102**, **108**], which belong themselves to the broader class of quadratic OOMs. We can implement a quadratic OOM $\mathfrak{K} = (\mathcal{H}, \mathcal{X}, \{K^{(x)}\}, P)$ as follows.

In the first step, the *evolution step*, we act only on the memory and the output SX with the unitary U_{SX} defined by the action:

$$U_{SX} |\psi\rangle \otimes |\emptyset\rangle = \sum_x K^{(x)} |\psi\rangle \otimes |x\rangle \ . \tag{6.13}$$

In the second step—the *measurement step*—the symbol is observed, sending the pure state $U_{SX} |\psi\rangle \langle\psi| \otimes |\emptyset\rangle \langle\emptyset| U_{SX}^\dagger$ to the mixed state in Eq. (6.12). This is done by coupling the system X to the auxiliary system A and applying a unitary so that:

$$U_{XA} |x\rangle_X \otimes |\emptyset\rangle_A \propto |x\rangle_X \otimes |x\rangle_A \ .$$

When the auxiliary is discarded (or, more realistically, reset) we are left with the state on SX, as desired.

6.3.2. Dissipation in generators. Now, a given generator cannot necessarily be implemented as efficiently as the minimal work rate $W_{\min} := -k_B T \ln 2 \ h_\mu$ indicates. This is because a generator acts temporally locally, only being able to use its current memory state to generate the next memory state and symbol. That is, the generator only acts directly on S at a given time to produce $S'X$. The true cost at time t must be bounded below by $W_{\text{loc}} := W_{\min} + \Delta S_{\text{loc}}$, where in this case the asymptotic locality dissipation is [**27**]:

$$\Delta S_{\text{loc}} = k_B T \ln 2 \ (\mathrm{I}\left[S : \overleftarrow{X}\right] - \mathrm{I}\left[S'X : \overleftarrow{X}\right]) \ . \tag{6.14}$$

In this case, the dissipation does not represent work lost to heat but rather the increase in tape entropy which did not facilitate converting heat into work. This minimal dissipation is also achievable for quantum processes, due to Section 6.2.1:

$$\Delta S_{\text{loc}} = k_B T \ln 2 \ (\mathrm{I}_q\left[S : \overleftarrow{X}\right] - \mathrm{I}_q\left[S'X : \overleftarrow{X}\right]) \ . \tag{6.15}$$

Henceforth we will drop the q from our entropies, as the formulae remain equivalent between the quantum and classical domain.

Another form for the dissipation, starting not from the locality dissipation but directly from Landauer's bound, is

$$\begin{aligned} \Delta S_{\text{loc}} &= k_{\text{B}} T \ln 2 \ \left[\left(\text{H}[S] - \text{H}\left[S'X\right] \right) - h_\mu \right] \\ &= k_{\text{B}} T \ln 2 \ \left[\text{I}\left[S' : X\right] - \left(\text{H}[X] - h_\mu \right) \right] \end{aligned} \tag{6.16}$$

This form offers us an immediately useful insight into the structure of dissipation costs for classical and quantum generators. Recall that we defined the partial order on generators, $\succsim$, which defines when one generator may be embedded into another. In particular, for $\mathfrak{F} = (\mathcal{R}, \mathcal{X}, \left\{ \mathbf{M}^{(x)} \right\})$ and $\mathfrak{G} = (\mathcal{S}, \mathcal{X}, \left\{ \mathbf{T}^{(x)} \right\})$ be two generators of the same process. We say that $\mathfrak{G}$ embeds $\mathfrak{F}$, written $\mathfrak{G} \succsim \mathfrak{F}$, if there is a stochastic matrix $\mathbf{P} = (P_{s|r})$ such that

$$\mathbf{P}\mathbf{M}^{(x)} = \mathbf{T}^{(x)}\mathbf{P}$$

for all $x \in \mathcal{X}$. We also extended this for quantum models; in particular, all q-machines embed the ϵ-machine of a process. One implication to this rule is that it renders the chain $S' - R' - X$ Markov, where S' and R' are the updated states of $\mathfrak{G}$ and $\mathfrak{F}$, respectively. This leads to the following theorem.

THEOREM 14. *Let $\Delta S_{\text{loc}}(\mathfrak{G})$ indicate the locality dissipation of a generator $\mathfrak{G}$. Suppose for two generators that that $\mathfrak{G} \succsim \mathfrak{F}$; then*

$$\Delta S_{\text{loc}}(\mathfrak{G}) \leq \Delta S_{\text{loc}}(\mathfrak{F}) \tag{6.17}$$

PROOF. *The only generator-dependent part of ΔS_{loc} is the mutual information $\text{I}[S' : X]$. By the Markov chain $S' - R' - X$ and the data processing inequality, whenever $\mathfrak{G} \succsim \mathfrak{F}$, we have $\text{I}[S' : X] \leq \text{I}[R' : X]$.*

In particular, this means that the q-machine produces less dissipation than the ϵ-machine when implemented thermodynamically.

Obviously, the minimal amount of dissipation that can be produced is ΔS_{loc}. It has been observed that this is achieved by retrodictive generators [27]. These are generators for whose states the Markov chain $\overleftarrow{X} - \overrightarrow{X} - S$ holds. But it is also the case that $S - S'X - \overrightarrow{X}$ (since the future depends

only on the next symbol probabilistically); this means that $\overleftarrow{X} - S'X - S$. Therefore we must have $\mathrm{I}\left[S : \overleftarrow{X}\right] - \mathrm{I}\left[S'X : \overleftarrow{X}\right] \leq 0$; since we cannot have $\Delta S_{\mathrm{loc}} < 0$, we must have $\Delta S_{\mathrm{loc}} = 0$.

In Ref. [27] it was also conjectured that retrodictive generators are the *only* generator which can achieve zero dissipation. This conjecture was made in the classical context. Regardless, we will not only prove that this statement is true among all classical generators, but that it also applies to quantum generators. To accomplish this we will have to prove a new theorem regarding the information- and resource-theoretic properties of local operations. Specifically, we will determine necessary and sufficient conditions for a locally performed operation $\mathcal{E} \otimes I$ mapping a quantum system AB to CB to saturate the data-processing inequality: $\mathrm{I}[A : B] = \mathrm{I}[A : C]$.

6.4. Efficient local computation

One of the most fundamental information-theoretic inequalities is the monotonicity of the relative entropy under transformations. This is the *data processing inequality* [36]. For a quantum channel $\mathcal{E}$ it says:

$$\mathrm{D}\left(\mathcal{E}\left(\rho_A\right) \| \mathcal{E}\left(\sigma_A\right)\right) \leq \mathrm{D}\left(\rho_A \| \sigma_A\right) \tag{6.18}$$

where $\mathrm{D}_{\rho\|\sigma}(=) \operatorname{Tr}\left[\rho \log_2 \rho - \rho \log_2 \sigma\right]$ is the quantum version of the Kullback-Liebler divergence from Section 1.5.2. The condition for equality requires constructing the *Petz recovery channel*:

$$\mathcal{R}_\sigma(\cdot) = \sigma^{1/2} \mathcal{E}^\dagger \left(\mathcal{E}(\sigma_A)^{-1/2} \cdot \mathcal{E}(\sigma_A)^{-1/2}\right) \sigma^{1/2} \ . \tag{6.19}$$

It is easy to check that $\mathcal{R}_\sigma \circ \mathcal{E}\left(\sigma_A\right) = \sigma_A$. A markedly useful result is that $\mathrm{D}\left(\mathcal{E}\left(\rho_A\right) \| \mathcal{E}\left(\sigma_A\right)\right) = \mathrm{D}\left(\rho_A \| \sigma_A\right)$ if and only if $\mathcal{R}_\sigma \circ \mathcal{E}\left(\rho_A\right) = \rho_A$ as well [149, 150].

Two other forms of the data processing inequality are useful to note here. The first uses another quantity for measuring distance between states called the *fidelity*:

$$F\left(\rho, \sigma\right) := \left(\operatorname{Tr}\left[\sqrt{\sqrt{\rho}\sigma\sqrt{\rho}}\right]\right)^2 \ . \tag{6.20}$$

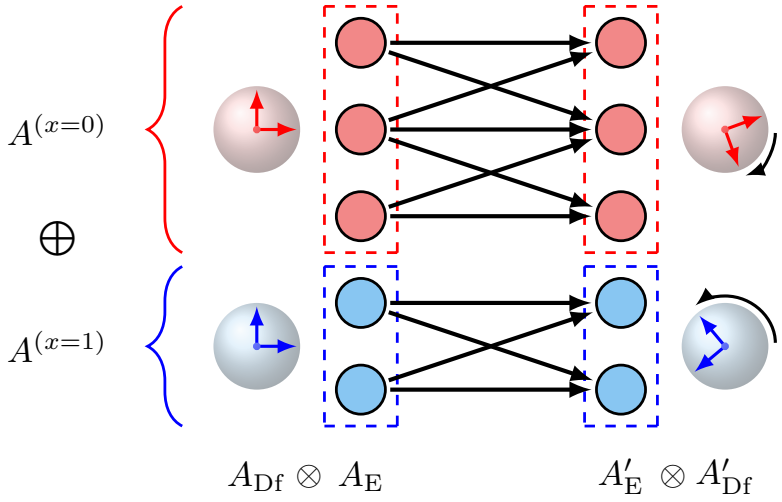

FIGURE 6.3. **Quantum channel decompositions**: Conserved measurement X divides the Hilbert space "vertically" via an orthogonal decomposition, $\mathcal{H}_A^{(x=0)} \oplus \mathcal{H}_A^{(x=1)}$, represented above by labels $A^{(x=0)}$ and $A^{(x=1)}$. For each value of x, there is a "horizontal" decomposition into the tensor product of an ergodic subspace and a decoherence-free subspace: $\mathcal{H}_A^{(x)} = \mathcal{H}_{A_{\mathrm{E}}}^{(x)} \otimes \mathcal{H}_{A_{\mathrm{Df}}}^{(x)}$, represented respectively by the labels A_{E} and A_{Df}. According to Theorem 16, information-theoretic reversibility requires storing data in the conserved measurement and the decoherence-free subspace. Any information stored coherently with respect to the conserved measurement (stored in the ergodic subspace) will be irreversibly modified under the channel's action.

It takes value $F = 0$ when the states ρ and σ are completely orthogonal and value $F = 1$ if and only if $\rho = \sigma$. The data processing inequality for fidelity states that for any quantum channel $\mathcal{E}$:

$$F\left(\mathcal{E}(\rho), \mathcal{E}(\sigma)\right) \geq F\left(\rho, \sigma\right) \ . \tag{6.21}$$

This is yet another way of saying that states map closer together under a quantum channel $\mathcal{E}$.

The second form arises from applying Eq. (6.18) to the mutual information. Let $\mathcal{E}_A : \mathcal{B}\left(\mathcal{H}_A\right) \to \mathcal{B}\left(\mathcal{H}_C\right)$ be a quantum channel. The local operation $\mathcal{E}_A \otimes I_B$ maps a bipartite system AB to CB. The data processing inequality Eq. (6.18) implies that we have $\mathrm{I}[C : B] \leq \mathrm{I}[A : B]$.

In these terms, our thermodynamic efficiency goal—$\Delta S_{\mathrm{loc}} = 0$—translates into determining conditions for equality—$\mathrm{I}[C : B] = \mathrm{I}[A : B]$—using the Petz recovery channel and channel fixed points.

6.4.1. Reversible information processing. To understand our result on local channels, an illustrative starting point is a key result on fixed points of quantum channels [14, 33] that leads to a natural decomposition, as illustrated in Fig. 6.3.

THEOREM 15 (Channel and Stationary State Decomposition). *Suppose* $\mathcal{E} : \mathcal{B}(\mathcal{H}_A) \to \mathcal{B}(\mathcal{H}_A)$ *is a quantum channel, Hilbert space* $\mathcal{H}_A$ *has a transient subspace* $\mathcal{H}_{\mathrm{T}}$*, and there is a projective measurement* $X = \left\{\Pi^{(x)}\right\}$ *on* $\mathcal{H}_{\mathrm{T}}^{\perp}$ *with countable outcomes* $\mathcal{X}$*, such that* $\mathcal{H}_A = \mathcal{H}_{\mathrm{T}} \oplus \left(\bigoplus_x \mathcal{H}_A^{(x)}\right)$*, where* $\mathcal{H}_A^{(x)}$ *is the support of* $\Pi^{(x)}$*. Then:*

(1) Subspace $\mathcal{H}_A^{(x)}$ *is preserved by* $\mathcal{E}$*, in that for all* $\rho \in \mathcal{B}\left(\mathcal{H}_A^{(x)}\right)$*, we have* $\mathcal{E}(\rho) \in \mathcal{B}\left(\mathcal{H}_A^{(x)}\right)$*.*

(2) Subspace $\mathcal{H}_A^{(x)}$ *further decomposes into an ergodic subspace* $\mathcal{H}_{A_{\mathrm{E}}}^{(x)}$ *and decoherence-free subspace* $\mathcal{H}_{A_{\mathrm{Df}}}^{(x)}$*:*

$$\mathcal{H}_A^{(x)} = \mathcal{H}_{A_{\mathrm{E}}}^{(x)} \otimes \mathcal{H}_{A_{\mathrm{Df}}}^{(x)} \ , \tag{6.22}$$

such that the Kraus operators of $\mathcal{E}|_{\mathcal{H}_{\mathrm{T}}^{\perp}}$ *decompose as [67]:*

$$K^{(\alpha)} = \bigoplus_{x \in \mathcal{X}} K_{A_{\mathrm{E}}}^{(\alpha,x)} \otimes U_{A_{\mathrm{Df}}}^{(x)} \tag{6.23}$$

and the map $\mathcal{E}_{A_{\mathrm{E}}}^{(x)}(\cdot) := \sum_\alpha K_{A_{\mathrm{E}}}^{(\alpha,x)} \cdot K_{A_{\mathrm{E}}}^{(\alpha,x)\dagger}$ *has a unique invariant state* $\pi_{A_{\mathrm{E}}}^{(x)}$*.*

(3) Any subspace of $\mathcal{H}$ *satisfying the above two properties is, in fact,* $\mathcal{H}_A^{(x)}$ *for some* $x \in \mathcal{X}$*.*

Furthermore, if ρ_A *is any invariant state—that is,* $\rho_A = \mathcal{E}_A(\rho_A)$*—then it decomposes as:*

$$\rho_A = \bigoplus_{x \in \mathcal{X}} \Pr(x)\, \pi_{A_{\mathrm{E}}}^{(x)} \otimes \rho_{A_{\mathrm{Df}}}^{(x)} \ , \tag{6.24}$$

for any distribution $\Pr(x)$ *and state* $\rho_{A_{\mathrm{Df}}}^{(x)}$ *satisfying* $U_{A_{\mathrm{Df}}}^{(x)} \rho_{A_{\mathrm{Df}}}^{(x)} U_{A_{\mathrm{Df}}}^{(x)\dagger} = \rho_{A_{\mathrm{Df}}}^{(x)}$*.*

Figure 6.3 gives the geometric structure implied by the theorem. The ergodic subspace of quantum channel $\mathcal{E}$ has two complementary decompositions. First, there is an orthogonal decomposition $\mathcal{H}_{\mathrm{T}}^{\perp} = \bigoplus_x \mathcal{H}_A^{(x)}$ induced by a projective measurement X whose values are conserved by $\mathcal{E}$'s action. This conservation is decoherent: only states compatible with X are stationary under the action of $\mathcal{E}$. X is called the *conserved measurement* of $\mathcal{E}$ [8]. Then, each $\mathcal{H}_A^{(x)}$ has a tensor decomposition $\mathcal{H}_A^{(x)} = \mathcal{H}_{A_{\mathrm{E}}}^{(x)} \otimes \mathcal{H}_{A_{\mathrm{Df}}}^{(x)}$ into an *ergodic* (E) and a *decoherence-free* (Df) part. The decoherence-free

subspace $\mathcal{H}_{A_{\mathrm{Df}}}^{(x)}$ undergoes only a unitary transformation [67]. The ergodic part $\mathcal{H}_{A_{\mathrm{E}}}^{(x)}$ is irreducibly mixed such that there is a single stationary state.

This result's contribution here is to aid in identifying when the data-processing inequality saturates. That is, using Thm. 15 and the Petz recovery channel, we derive necessary and sufficient constraints on the structures of ρ_A, σ_A, and $\mathcal{E}_A$ for determining when $\mathrm{D}\left(\mathcal{E}\left(\rho_A\right) \| \mathcal{E}\left(\sigma_A\right)\right) = \mathrm{D}\left(\rho_A \| \sigma_A\right)$.

To achieve this, we recall a previously known result [136, 137], showing that it can be derived using only the Petz recovery map and Thm. 15. The immediate consequence is a novel proof.

THEOREM 16 (Reversible Information Processing). *Suppose for two states ρ_A and σ_A on a Hilbert space $\mathcal{H}_A$ and a quantum channel $\mathcal{E} : \mathcal{B}\left(\mathcal{H}_A\right) \to \mathcal{B}\left(\mathcal{H}_A\right)$, we have:*

$$\mathrm{D}\left(\rho_A \| \sigma_A\right) = \mathrm{D}\left(\rho_C \| \sigma_C\right) \ , \tag{6.25}$$

where $\rho_C = \mathcal{E}_A\left(\rho_A\right)$ and $\sigma_C = \mathcal{E}_A\left(\sigma_A\right)$. Then there is a measurement X on A with countable outcomes $\mathcal{X}$ and orthogonal decompositions $\mathcal{H}_{\mathcal{E}} = \bigoplus_x \mathcal{H}_A^{(x)}$ and $\mathcal{H}_C^{(x)}$ such that:

(1) For all $\rho \in \mathcal{B}\left(\mathcal{H}_A^{(x)}\right)$ and $\mathcal{E}\left(\rho\right) \in \mathcal{B}\left(\mathcal{H}_C^{(x)}\right)$, the mapping $\mathcal{E}|_{\mathcal{H}_A^{(x)}}$ onto $\mathcal{B}\left(\mathcal{H}_C^{(x)}\right)$ is surjective.
(2) Subspaces $\mathcal{H}_A^{(x)}$ further decompose into:

$$\mathcal{H}_A^{(x)} = \mathcal{H}_{A_{\mathrm{E}}}^{(x)} \otimes \mathcal{H}_{A_{\mathrm{Df}}}^{(x)} \text{ and} \tag{6.26}$$

$$\mathcal{H}_C^{(x)} = \mathcal{H}_{C_{\mathrm{E}}}^{(x)} \otimes \mathcal{H}_{C_{\mathrm{Df}}}^{(x)} \ , \tag{6.27}$$

so that $\mathcal{H}_{C_{\mathrm{E}}}^{(x)}$ and $\mathcal{H}_{A_{\mathrm{Df}}}^{(x)}$ are unitarily equivalent and the Kraus operators decompose as:

$$L^{(\alpha)} = \bigoplus_{x \in \mathcal{X}} L_{A_{\mathrm{E}}}^{(\alpha,x)} \otimes U_{A_{\mathrm{Df}}}^{(x)} \ . \tag{6.28}$$

Furthermore, states ρ_A and σ_A satisfy:

$$\rho_A = \sum_{x \in \mathcal{X}} \Pr\left(x; \rho\right) \pi_{A_{\mathrm{E}}}^{(x)} \otimes \rho_{A_{\mathrm{Df}}}^{(x)} \tag{6.29a}$$

$$\sigma_A = \sum_{x \in \mathcal{X}} \Pr\left(x; \sigma\right) \pi_{A_{\mathrm{E}}}^{(x)} \otimes \sigma_{A_{\mathrm{Df}}}^{(x)} \tag{6.29b}$$

for some $\pi^{(x)}_{A_{\mathrm{E}}}$. *And, their images are:*

$$\rho_C = \sum_{x\in\mathcal{X}} \Pr(x;\rho)\, \pi^{(x)}_{C_{\mathrm{E}}} \otimes \rho^{(x)}_{C_{\mathrm{Df}}} \tag{6.30a}$$

$$\sigma_C = \sum_{x\in\mathcal{X}} \Pr(x;\sigma)\, \pi^{(x)}_{C_{\mathrm{E}}} \otimes \sigma^{(x)}_{C_{\mathrm{Df}}} \ , \tag{6.30b}$$

where $\pi^{(x)}_{C_{\mathrm{E}}} = \mathcal{E}^{(x)}_{A_{\mathrm{E}}}\left(\pi^{(x)}_{A_{\mathrm{E}}}\right)$, $\rho^{(x)}_{C_{\mathrm{Df}}} = U^{(x)}_{A_{\mathrm{Df}}} \rho^{(x)}_{A_{\mathrm{Df}}} U^{(x)\dagger}_{A_{\mathrm{Df}}}$, *and* $\sigma^{(x)}_{C_{\mathrm{Df}}} = U^{(x)}_{A_{\mathrm{Df}}} \sigma^{(x)}_{A_{\mathrm{Df}}} U^{(x)\dagger}_{A_{\mathrm{Df}}}$.

Proof. *We know that* $\mathcal{N}_\sigma := \mathcal{R}_\sigma \circ \mathcal{E}$ *must have both* ρ_A *and* σ_A *as stationary distributions. Let* X *be the conserved measurement of* $\mathcal{N}_\sigma$. *It induces the decompositions* $\mathcal{H}_A = \mathcal{H}_{\mathrm{T}} \oplus \left(\bigoplus_x \mathcal{H}^{(x)}_A\right)$ *and* $\mathcal{H}^{(x)}_A = \mathcal{H}^{(x)}_{A_{\mathrm{E}}} \otimes \mathcal{H}^{(x)}_{A_{\mathrm{Df}}}$, *as well as the state decompositions Eq.* (6.29).

Now, we leverage the fact that $\mathcal{N}_\sigma|_{\mathcal{H}^\perp_{\mathrm{T}}}$ *has Kraus decomposition of the form Eq.* (6.23). *The net effect is summed up by two constraints:*

(1) For each α, $K^{(\alpha)}$ *maps each* $\mathcal{H}^{(x)}_A$ *to itself.*

(2) Let $K^{(\alpha,x)}$ *be the block of* $K^{(\alpha)}$ *restricted to* $\mathcal{H}^{(x)}_A$. *Let* M *be a complete projective measurement on* $\mathcal{H}^{(x)}_{A_{\mathrm{Df}}}$ *with basis* $\{|m\rangle\}$ *and define the transformed basis* $\{|\widetilde{m}\rangle = U^{(x)}|m\rangle\}$. *Now, let* $\mathcal{H}^{(x,m)}_A = \left\{|\psi\rangle \otimes |m\rangle : |\psi\rangle \in \mathcal{H}^{(x)}_{A_{\mathrm{E}}}\right\}$ *and similarly for* $\mathcal{H}^{(x,\widetilde{m})}_A$. *Then,* $K^{(z,x)}$ *maps* $\mathcal{H}^{(x,m)}_A$ *to* $\mathcal{H}^{(x,\widetilde{m})}_A$. This holds for any measurement M.

Proving Eq. (6.28) *requires these two constraints. Each is a form of distinguishability criterion for the total channel* $\mathcal{N}_\sigma|_{\mathcal{H}^\perp_{\mathrm{T}}}$. *Since* $\mathcal{N}_\sigma$ *can tell certain orthogonal outcomes apart, so too must* $\mathcal{E}$. *Or else,* $\mathcal{R}_\sigma$ *would "pull apart" nonorthogonal states. However, this is impossible for a quantum channel. By formally applying this notion to constraints 1 and 2 above, we recover Eq.* (6.28).

Let $L^{(\alpha)}$ *be the Kraus operators of* $\mathcal{E}|_{\mathcal{H}^\perp_{\mathrm{T}}}$. *Then* $K^{(\alpha)} = \sigma^{1/2} L^{(\alpha)\dagger} \mathcal{E}(\sigma)^{1/2} L^{(z)}$. *Now, if* $L^{(\alpha)}$ *did not map each* $\mathcal{H}^{(x)}_A$ *to some orthogonal subspace* $\mathcal{H}^{(x)}_C$, *then for some distinct* x *and* x' *there would be* $|\psi\rangle \in \mathcal{H}^{(x)}_A$ *and* $|\phi\rangle \in \mathcal{H}^{(x')}_C$ *such that* $F\left(\mathcal{E}(|\psi\rangle\langle\psi|), \mathcal{E}(|\phi\rangle\langle\phi|)\right) > 0$; *recall Eq.* (6.20) *defines fidelity. However, we must also have* $F\left(\mathcal{N}_\sigma(|\psi\rangle\langle\psi|), \mathcal{N}_\sigma(|\phi\rangle\langle\phi|)\right) = 0$, *which is impossible by Eq.* (6.21) *since as applying* $\mathcal{R}_\sigma$ *cannot reduce fidelity. So, it must be the case that* $L^{(\alpha)}$ *maps each* $\mathcal{H}^{(x)}_A$ *to some orthogonal subspace* $\mathcal{H}^{(x)}_C$. *This proves Claim 1 in Thm. 16.*

Let $L^{(\alpha,x)}$ *be the block of* $L^{(z)}$ *restricted to* $\mathcal{H}^{(x)}_A$. *Further, let* M *be a complete measurement on* $\mathcal{H}^{(x)}_{A_{\mathrm{Df}}}$. *Then* $\mathcal{E}|_{\mathcal{H}^{(x)}_A}$ *must map each* $\mathcal{H}^{(x,m)}_A$ *onto orthogonal subspaces* $\mathcal{H}^{(x,m)}_C$ *of* $\mathcal{H}^{(x)}_C$, *lest* $\mathcal{N}_\sigma|_{\mathcal{H}^{(x)}_A}$

could not map each $\mathcal{H}_A^{(x,m)}$ *to orthogonal spaces* $\mathcal{H}_A^{(x,\widetilde{m})}$*. This follows from fidelity, as in the previous paragraph.*

Now, let $L^{(\alpha,x,m)} : \mathcal{H}_{A_{\mathrm{E}}}^{(x)} \to \mathcal{H}_C^{(x,m)}$*, so that:*

$$L^{(\alpha,x)} = \bigoplus_m L^{(\alpha,x,m)} \otimes \langle m| \ . \tag{6.31}$$

Let N *be another complete measurement on* $\mathcal{H}_{A_{\mathrm{Df}}}^{(x)}$ *such that* $|n\rangle = \sum_m w_{m,n} |m\rangle$*, with* $w_{m,n}$ *a unitary matrix. And, let* $n, n' \in \mathcal{N}$ *be distinct. We have for any* $|\psi\rangle$ *that:*

$$\begin{aligned} \mathcal{E}^{(x)}\left(|\psi\rangle\langle\psi| \otimes |n\rangle\langle n|\right) &= \sum_\alpha \bigoplus_{m,m'} w_{m',n} w^*_{m,n} L^{(\alpha,x,m)} |\psi\rangle\langle\psi| L^{(\alpha,x,m')\dagger} \text{ and} \\ \mathcal{E}^{(x)}\left(|\psi\rangle\langle\psi| \otimes |n'\rangle\langle n'|\right) &= \sum_\alpha \bigoplus_{m,m'} w_{m',n'} w^*_{m,n'} L^{(\alpha,x,m)} |\psi\rangle\langle\psi| L^{(\alpha,x,m')\dagger} \ . \end{aligned} \tag{6.32}$$

Now, it must be that $\mathcal{E}^{(x)}\left(|\psi\rangle\langle\psi| \otimes |n\rangle\langle n|\right)$ *and* $\mathcal{E}^{(x)}\left(|\psi\rangle\langle\psi| \otimes |n'\rangle\langle n'|\right)$ *are orthogonal. However:*

$$\mathrm{Tr}\left[\mathcal{E}^{(x)}\left(|\psi\rangle\langle\psi| \otimes |n\rangle\langle n|\right)\mathcal{E}^{(x)}\left(|\psi\rangle\langle\psi| \otimes |n'\rangle\langle n'|\right)\right] = \sum_{\alpha,\alpha'}\sum_m w_{m,n} w^*_{m,n'} \langle\psi| L^{(\alpha,x,m)\dagger} L^{(\alpha,x,m)} |\psi\rangle \ . \tag{6.33}$$

This vanishes for arbitrary N *and* $|\psi\rangle$ *only if* $L^{(\alpha,x,m)\dagger} L^{(\alpha',x,m)}$ *is independent of* m *for each* α *and* α'*. This implies that* $L^{(\alpha,x,m)} = W^{(m)} L_{\mathrm{E}}^{(\alpha,x)}$ *for some unitary* $W^{(m)}$*.*

All of which leads one to conclude that the $\mathcal{H}_C^{(x,m)}$ *for each* m *must be unitarily equivalent. And so, the decomposition* $\mathcal{H}_C^{(x)} = \bigoplus_m \mathcal{H}_C^{(x,m)}$ *instead becomes a tensor product decomposition* $\mathcal{H}_C^{(x)} = \mathcal{H}_{C_{\mathrm{E}}}^{(x)} \otimes \mathcal{H}_{C_{\mathrm{Df}}}^{(x)}$ *and further that* $L^{(\alpha,x)} = L_{A_{\mathrm{E}}}^{(\alpha,x)} \otimes V_{A_{\mathrm{Df}}}^{(x)}$*.*

Finally, the constraints Eq. (6.30) follow from the form of $\mathcal{E}$ *and Eq. (6.29).*

Theorem 16's main implication is that, when a channel $\mathcal{E}$ acts, information stored in the conserved measurement and in the decoherence-free subspaces is recoverable. Two states that differ in terms of the conserved measurement and the decoherence-free subspaces remain different and do not grow more similar under $\mathcal{E}$'s action. Conversely, information stored in measurements not compatible with the conserved measurement or stored in the ergodic subspaces is irreversibly garbled by $\mathcal{E}$.

The next section uses this decomposition to study how locally acting channels impact correlations between subsystems. This directly drives the thermodynamic efficiency of local operations. Namely,

for thermodynamic efficiency correlations must be stored specifically in the conserved measurement and decoherence-free subspaces of the local channel.

6.4.2. Quantum sufficient statistics. Stating our result requires first defining the quantum notion of a sufficient statistic. Previously, quantum sufficient statistics of A for B were defined when AB is a classical-quantum state [98]. That is, when ρ_{AB} commutes with a local measurement on A. They were also introduced in the setting of sufficient statistics for a family of states [78, 151]. This corresponds to the case where AB is quantum-classical—ρ_{AB} commutes with a local measurement on B. Our definition generalizes these cases to fully-quantal correlations between A and B.

We start, as an example, by giving the following definition of a minimum sufficient statistic of a classical joint random variable $XY \sim \Pr(x, y)$ in terms of an equivalence relation. We define the predictive equivalence relation $\sim$ for which $x \sim x'$ if and only if $\Pr(y|x) = \Pr(y|x')$ for all y. The *minimum sufficient statistic* (MSS) $[X]_Y$ is given by the equivalence classes $[x]_Y := \{x' : x \sim x'\}$. Let us denote $\Sigma := [X]_Y$ and let $\Pr(y|\sigma) := \Pr(y|x)$ for any $x \in \sigma$.

This cannot be *directly* generalized to the quantum setting since correlations between A and B cannot always be described in the form of states conditioned on the outcome of a local measurement on A. If the latter were the case, the state would be classical-quantum, but general quantum correlations can be much more complicated than these. However, we can take the most informative local measurement that does not disturb ρ_{AB} and then consider the "atomic" quantum correlations it leaves behind.

Let ρ_{AB} be a bipartite quantum state. A *maximal local commuting measurement* (MLCM) of A for B is any local measurement X with projectors $\{\Pi^{(x)}\}$ on system A such that:

$$\rho_{AB} = \bigoplus_x \Pr(X = x)\rho_{AB}^{(x)} , \tag{6.34}$$

where:

$$\Pr(X = x) = \mathrm{Tr}\left((\Pi_X^{(x)} \otimes I_B)\rho_{AB}\right) \tag{6.35}$$

and:

$$\Pr(X = x)\rho_{AB}^{(x)} = (\Pi_X^{(x)} \otimes I_B)\rho_{AB}(\Pi_X^{(x)} \otimes I_B) \ , \tag{6.36}$$

and any further local measurement Y on $\rho_{AB}^{(x)}$ disturbs the state:

$$\rho_{AB}^{(x)} \neq \sum_y (\Pi_Y^{(y)} \otimes I_B)\rho_{AB}^{(x)}(\Pi_Y^{(y)} \otimes I_B) \ . \tag{6.37}$$

We call the states $\left\{\rho_{AB}^{(x)}\right\}$ *quantum correlation atoms.*

Proposition 11 (MLCM uniqueness). *Given a state ρ_{AB}, there is a unique MLCM of A for B.*

Proof. *Suppose there were two distinct MLCMs, X and Y. Then:*

$$\rho_{AB} = \sum_y (\Pi_Y^{(y)} \otimes I_B)\rho_{AB}(\Pi_Y^{(y)} \otimes I_B) \ . \tag{6.38}$$

This can be written as:

$$\rho_{AB} = \bigoplus_x \sum_y \Pr(X = x)(\Pi_Y^{(y)} \otimes I_B)\rho_{AB}^{(x)}(\Pi_Y^{(y)} \otimes I_B) \ . \tag{6.39}$$

However, this means for each x:

$$(\Pi_Y^{(y)} \otimes I_B)\rho_{AB}^{(x)}(\Pi_Y^{(y)} \otimes I_B) = \rho_{AB}^{(x)} \ . \tag{6.40}$$

So, X is not a MLCM, giving a contradiction.

It will be helpful in our study of quantum generators to have the following fact as well:

Proposition 12 (MLCM for a classical-quantum state). *Given a classical-quantum state:*

$$\rho_{AB} := \sum_x \Pr(x) \left|x\right\rangle \left\langle x\right| \otimes \rho_B^{(x)} \ , \tag{6.41}$$

the MLCM is the most refined measurement Θ such that:

$$\rho_B^{(x)} = \sum_\theta \Pi^{(\theta)} \rho_B^{(x)} \Pi^{(\theta)} \tag{6.42}$$

for all x.

PROOF. *Given that Θ is a commuting local measurement, the question is whether it is maximal. If it is not maximal, though, there is a refinement Y that is also a commuting local measurement. By Θ's definition, there is an x such that $\rho_B^{(x)} \neq \sum_y \Pi^{(y)} \rho_B^{(x)} \Pi^{(y)}$. This implies $\rho_{AB} \neq \sum_y (I \otimes \Pi^{(y)}) \rho_{AB} (I \otimes \Pi^{(y)})$, contradicting the assumption that Y is commuting local.*

Now, as in the classical setting, we define an equivalence class over the values of the MLCM via the equivalence between their quantum correlation atoms. Classically, these atoms are simply the conditional probability distributions $\Pr(\cdot|x)$; in the classical-quantum setting, they are the conditional quantum states $\rho_B^{(x)}$. Note that each is defined as a distribution on the variable Y or system B. In contrast, the general quantum correlation atoms $\rho_{AB}^{(x)}$ depend on both systems A and B.

The resulting challenge is resolved in the following way. Let ρ_{AB} be a bipartite quantum state and let X be the MLCM of A for B. We define the *correlation equivalence* relation $x \sim x'$ over values of X where $x \sim x'$ if and only if $\rho_{AB}^{(x)} = (U \otimes I_B) \rho_{AB}^{(x')} (U^\dagger \otimes I_B)$ for a local unitary U.

Finally, we define the *Minimal Local Sufficient Statistic* (MLSS) $[X]_\sim$ as the equivalence class $[x]_\sim := \{x' : x' \sim x\}$ generated by the relation $\sim$ between correlation atoms. Thus, our notion of sufficiency of A for B is to find the most informative local measurement and then coarse-grain its outcomes by unitary equivalence over their correlation atoms. The correlation atoms and the MLSS $[X]_\sim$ together describe the correlation structure of the system AB.

6.4.3. Reversible local operations. Finally, with the aid of the MLSS and Theorem 16 we can prove our result on reversible local operations.

THEOREM 17 (Reversible local operations). *Let ρ_{AB} be a bipartite quantum state and let $\mathcal{E}_A \otimes I_B$ be a local operation with $\mathcal{E}_A : \mathcal{B}(\mathcal{H}_A) \to \mathcal{B}(\mathcal{H}_C)$. Suppose X is the MLCM of ρ_{AB} and Y, that of $\rho_{CB} = \mathcal{E}_A \otimes I_B (\rho_{AB})$. The decomposition into correlation atoms is:*

$$\rho_{AB} = \bigoplus_x \Pr{}_A(x) \, \rho_{AB}^{(x)} \text{ and} \tag{6.43}$$

$$\rho_{CB} = \bigoplus_y \Pr{}_C(y) \, \rho_{CB}^{(y)} \, . \tag{6.44}$$

Then, $\mathrm{I}[A:B] = \mathrm{I}[C:B]$ *if and only if* $\mathcal{E}_A$ *can be expressed by Kraus operators of the form:*

$$K^{(\alpha)} = \bigoplus_{x,y} e^{i\phi_{xy\alpha}} \sqrt{\Pr(y,\alpha|x)} U^{(y|x)} \ , \tag{6.45}$$

where $\phi_{xy\alpha}$ *is any arbitrary phase and* $\Pr(y,\alpha|x)$ *is a stochastic channel that is nonzero only when* $\rho^{(x)}_{AB}$ *and* $\rho^{(y)}_{CB}$ *are equivalent up to a local unitary operation* $U^{(y|x)}$ *that maps* $\mathcal{H}^{(x)}_A$ *to* $\mathcal{H}^{(y)}_C$.

Proof. *We can apply the Reversible Information Processing Theorem (Thm. 16) from the previous section here. This demands that there be a measurement* X *and a decomposition of the Hilbert space* $\mathcal{H}_A = \mathcal{H}_{A_{\mathrm{E}}} \otimes \mathcal{H}_{A_{\mathrm{Df}}}$ *such that:*

$$\rho_{AB} = \sum_x \Pr{}_A(x) \, \rho^{(x)}_{(AB)_{\mathrm{E}}} \otimes \rho^{(x)}_{(AB)_{\mathrm{Df}}} \tag{6.46}$$

$$\rho_A \otimes \rho_B = \sum_x \Pr{}_A(x) \, \rho^{(x)}_{(AB)_{\mathrm{E}}} \otimes \left(\rho^{(x)}_{A_{\mathrm{Df}}} \otimes \rho^{(x)}_{B_{\mathrm{Df}}} \right) \ , \tag{6.47}$$

such that $\mathcal{E}_A \otimes I_B$ *conserves measurement* X *and acts decoherently on* $(AB)_{\mathrm{E}}$ *and coherently on* $(AB)_{\mathrm{Df}}$. *However, the local nature of* $\mathcal{E}_A \otimes I_B$ *makes it clear we can simplify this decomposition to:*

$$\rho_{AB} = \sum_x \Pr{}_A(x) \, \rho^{(x)}_{A_{\mathrm{E}}} \otimes \left(\rho^{(x)}_{A_{\mathrm{Df}}B} \right) \tag{6.48}$$

$$\rho_A \otimes \rho_B = \sum_x \Pr{}_A(x) \, \rho^{(x)}_{A_{\mathrm{E}}} \otimes \left(\rho^{(x)}_{A_{\mathrm{Df}}} \otimes \rho_B \right) \ , \tag{6.49}$$

where $\mathcal{E}_A$ *conserves the* local *measurement* X *on* A *and acts decoherently on* A_{E} *and acts as a local unitary* $U_{A_{\mathrm{Df}}} \otimes I_B$ *on* $A_{\mathrm{Df}}B$.

Suppose now, however, that given X *the variable* Y_x *is the diagonalizing measurement of* $\rho^{(x)}_{A_{\mathrm{E}}}$ *and* Z_x *is the MLCM of* $\rho^{(x)}_{A_{\mathrm{Df}}B}$. *The joint measurement* XY_XZ_X *—where* X *is measured first and then the other two measurements are determined with knowledge of its outcome—is the MLCM of* ρ_{AB}. *Note that for any* x *and* z, *the outcomes* (x,y,z) *and* (x,y',z) *are correlation equivalent: measurement* Y_X *is completely decoupled from system* B. *Then, the MLSS* $\Sigma := [XY_XZ_X]_B$ *is simply a function of* X *and* Z_X.

Since XZ_X *is conserved by the action of* $\mathcal{E}_A \otimes I$ *—where* X *is the conserved measurement, while* Z_X *is preserved through the unitary evolution—the MLSS* Σ *must be preserved and each correlation atom is transformed only by a local unitary. This results in the form Eq. (6.45).*

This proves that $\mathrm{I}[A:B] = \mathrm{I}[C:B]$ *implies Eq.* (6.45)*. The converse is straightforward to check. Let* $\Sigma = [X]_B$ *and let* $I\left(A:B|\Sigma=\sigma\right)$ *be the mutual information of* $\rho_{AB}^{(x)}$ *for any* $x \in \sigma$*. (This is the same for all such* x *by local unitary equivalence.) Then:*

$$\mathrm{I}[A:B] = \sum_{\sigma} \Pr\left(\Sigma=\sigma\right) \mathrm{I}[A:B|\Sigma=\sigma] \ . \tag{6.50}$$

Similarly, let $\Sigma' = [Y]_B$ *and let* $I\left(C:B|\Sigma'=\sigma'\right)$ *be the mutual information of* $\rho_{CB}^{(y)}$ *for any* $y \in \sigma'$*; then:*

$$\mathrm{I}[A:B] = \sum_{\sigma'} \Pr\left(\Sigma'=\sigma'\right) \mathrm{I}\left[C:B|\Sigma'=\sigma'\right] \ . \tag{6.51}$$

Since $\mathcal{E}$ *isomorphically maps each* σ *to a unique* σ'*, such that* $\mathrm{I}[A:B|\Sigma=\sigma] = \mathrm{I}[C:B|\Sigma'=\sigma']$ *by unitary equivalence, we must have* $\mathrm{I}[A:B] = \mathrm{I}[C:B]$*.*

The theorem's classical form follows as a corollary.

Corollary 6 (Reversible local operations, classical). *Let* XY *be a joint random variable and let* $\Pr(Z=z|X=x)$ *be a channel from* $\mathcal{X}$ *to some set* $\mathcal{Z}$*, resulting in the joint random variable* ZY*. Then* $\mathrm{I}[X:Y] = \mathrm{I}[Z:Y]$ *if and only if* $\Pr(Z=z|X=x) > 0$ *only when* $\Pr(Y=y|Z=z) = \Pr(Y=y|X=x)$ *for all* y*.*

6.5. Dissipation-free generators

With Theorem 17 in hand, we can proceed to demonstrate the conditions for a generator to achieve $\Delta S_{\mathrm{loc}} = 0$. We will begin by considering the case of classical-only generators, as this will provide us the necessary intuition.

6.5.1. Classical generators. In the classical case, we will demonstrate that $\Delta S_{\mathrm{loc}}(\mathfrak{G}) = 0$ implies that the $\mathfrak{G}$ is a retrodictor by utilizing the operational characterization of retrodictors which we proved in Prop. 9 in Chapter 4. Specifically, we will use the fact that $\mathfrak{G} = (\mathcal{S}, \mathcal{X}, \{\mathbf{T}^{(x)}\})$ is a retrodictor if and only if its retrodictive state-merging $\mathfrak{G}|_R$ is co-unifilar; that is, if $T^{(x)}_{s's} > 0$ only when $s = f(x, s')$ for some function f.

THEOREM 18. *A generator* $\mathfrak{G} = (\mathcal{S}, \mathcal{X}, \{\mathbf{T}^{(x)}\})$ *satisfies* $\mathrm{I}\left[S : \overleftarrow{X}\right] = \mathrm{I}\left[S'X : \overleftarrow{X}\right]$ *if and only if it is a retrodictor.*

PROOF. *We already know that retrodiction implies* $\mathrm{I}\left[S : \overleftarrow{X}\right] = \mathrm{I}\left[S'X : \overleftarrow{X}\right]$*; we must prove the forward direction.*

To see this, recall from Cor. 6 that $\mathrm{I}\left[S : \overleftarrow{X}\right] = \mathrm{I}\left[S'X : \overleftarrow{X}\right]$ *only if* $\Pr_{\mathfrak{G}}(s_1, x_1|s_0) > 0$ *implies that* s_0 *and* s_1, x_1 *generate the same conditional distribution over pasts. This means for every* $t \geq 0$*,*

$$\Pr_{\mathfrak{G}}(x_{-t} \ldots x_0|s_1, x_1) = \Pr_{\mathfrak{G}}(x_{-t} \ldots x_0|s_0) \ . \tag{6.52}$$

We use this to write:

$$\begin{aligned} \Pr_{\mathfrak{G}}(x_{-t} \ldots x_1|s_1) &= \Pr_{\mathfrak{G}}(x_{-t} \ldots x_0|s_1 x_1) \Pr(x_1|s_1) \\ &= \Pr_{\mathfrak{G}}(x_{-t} \ldots x_0|s_0) \Pr(x_1|s_1) \ . \end{aligned} \tag{6.53}$$

Rearranging and using the retrodictive equivalence partitions $\sigma_t := [s_t]_\sim$*, we have:*

$$\Pr_{\mathfrak{G}|_R}(x_{-t} \ldots x_0|\sigma_0) = \frac{\Pr_{\mathfrak{G}|_R}(x_{-t} \ldots x_1|\sigma_1)}{\Pr_{\mathfrak{G}|_R}(x_1|\sigma_1)} \ . \tag{6.54}$$

Define a function $f : \mathcal{S} \times \mathcal{X} \to \mathcal{S}$ *as follows. For a given* σ' *and* x*, let* $f(\sigma', x)$ *be the state of* $\mathfrak{G}|_R$ *class such that:*

$$\Pr_{\mathfrak{G}|_R}\left(\overleftarrow{X}|f(\sigma', x)\right) = \frac{\Pr_{\mathfrak{G}|_R}\left(\overleftarrow{X}x|\sigma'\right)}{\Pr_{\mathfrak{G}|_R}(x|\sigma')} \ . \tag{6.55}$$

Such an state $f(\sigma', x)$ *must exist by Eq. (6.54). It is unique since, by definition, equivalence classes* σ *have unique distributions* $\Pr\left(\overleftarrow{X}|\sigma\right)$*. Then* $\sigma = f(\sigma', x)$ *is a requirement for* $\Pr_{\mathfrak{G}|_R}(\sigma', x|\sigma) > 0$*. This means that* $\mathfrak{G}|_R$ *is co-unifilar, and therefore* $\mathfrak{G}$ *is a retrodictor.*

Our theorem confirms the conjecture from [27] that the necessary and sufficient condition for $\Delta S_{\mathrm{loc}} = 0$ is that the generator in question is a retrodictor. A similar result, for classical generators, was presented in [62] where a lower bound on ΔS_{loc} was derived for predictive generators (Eq. (A23) in [62]). A consequence of this bound is that $\Delta S_{\mathrm{loc}} = 0$ only when the predictor is also a retrodictor. However, this bound does not extend to nonpredictive generators. In contrast, Thm. 18 applies to all generators.

Our result is complemented by another recent result [**60**], which demonstrated how from a predictive generator one can construct a sequence of generators that asymptotically approach a retrodictor and whose dissipation ΔS_{loc} asymptotically approaches zero. Helpfully, this result points to possible perturbative extensions of Thm. 18.

These results bear on the trade-off between dissipation and *memory* for classical generators. As we saw in Chapter 4, the reverse ϵ-machine, being a state-merging of any retrodictive generator, is minimal with respect to the retrodictive generators via all quantifications of the memory, such as the number of memory states $|\mathcal{S}|$ and the entropy $\mathrm{H}[S]$ [**106**]. We now see that the above showed that any *thermodynamically efficient* generator must be retrodictive; consequently, thermodynamic efficiency comes with a memory constraint, defined by the reverse ϵ-machine. And, when the memory falls below that of the reverse ϵ-machine, *dissipation must be present.*

6.5.2. Quantum generators. The result for classical generators, Thm. 18, will form the foundation for the equivalent result for quantum generators. However, in order to apply it, we will need to gain additional insight into the quantum sufficient statistics at play in quantum generators. In Section 6.3.1, we described how a quantum generator $\mathfrak{Q} = (\mathcal{H}, \mathcal{X}, \{\mathcal{E}^{(x)}\})$ could be physically implemented using thermodynamically allowed operations. The implementation updates its state system S while writing its output $X_1 \ldots X_t$ onto a sequence $A_1 \ldots A_t$ of auxiliary systems.

The state of the entire implementation at time t is given by

$$\rho_{\mathfrak{G}}(t) = \sum_{x_1 \ldots x_t} \Pr_{\mathfrak{Q}}(\, x_1 \ldots x_t \,)\, \rho_{x_1 \ldots x_t} \otimes |x_1 \ldots x_t\rangle \langle x_1 \ldots x_t|$$

where $\rho_{x_1 \ldots x_t} = \Pr_{\mathfrak{Q}}(\, x_1 \ldots x_t \,)^{-1}\, \mathfrak{E}^{(x_1 \ldots x_t)}(\rho_\pi)$. Now, by the Embedding Lemma (Lem. 1) applied to quantum generators, for sufficiently large t this converges to an embedding of a past $\rho_{\overleftarrow{x}} = E_{\mathfrak{Q}}(\overleftarrow{x})$. Consider the full set of past embeddings $\{\rho_{\overleftarrow{x}}\}$. Let the most refined measurement which commutes with *all* of these states be called Θ. By Prop. 12, we see that Θ is the MCLM of S in for the asymptotic system $S\overleftarrow{X}$, and further that ΘX is the MCLM of $S'X$ for the system $S'\overleftarrow{X}X$.

THEOREM 19 (Maximally-efficient quantum generator). *Let $\mathfrak{Q} = (\mathcal{S}, \mathcal{X}, \{\mathcal{E}^{(x)}\})$ quantum generator. Then $\Delta S_{\text{loc}}(\mathfrak{Q}) = 0$ if and only if $\mathfrak{Q}$ is unitarily equivalent to a classical retrodictor paired with an irrelevant quantum system.*

PROOF. *The quantum channel:*

$$\mathcal{E}(\rho) = \sum_x \mathcal{E}^{(x)}(\rho) \otimes |x\rangle\langle x| \tag{6.56}$$

can be expanded as

$$\mathcal{E}(\rho) = \sum_{x,\alpha} K^{(x,\alpha)} \rho K^{(x,\alpha)\dagger} \otimes |x\rangle\langle x| \tag{6.57}$$

where $K^{(x,\alpha)}$, indexed by α, are the Kraus operators for each $\mathcal{E}^{(x)}$. This means that the whole channel $\mathcal{E}$ has Kraus operators $L^{(x,\alpha)} := K^{(x,\alpha)} \otimes |x\rangle$.

Let Θ denote the MCLM of S (so that ΘX is the MCLM of $S'X$). Then, if we require zero dissipation, then Thm. 17 demands that the Kraus operators $L^{(x,\alpha)}$ have the form:

$$\begin{aligned} L^{(x,\alpha)} &= \bigoplus_{\theta,\theta'} \sqrt{\Pr(\theta', x, \alpha|\theta)} U^{(x,\theta'|\theta)} \\ &= \bigoplus_{\theta,\theta'} \sqrt{\Pr(\theta', x, \alpha|\theta)} U^{(x)}_{\theta \mapsto \theta'} \otimes |x\rangle \ . \end{aligned} \tag{6.58}$$

This implies that the quantum-generator Kraus operators have the form:

$$K^{(x,\alpha)} = \bigoplus_{\theta,\theta'} \sqrt{\Pr(\theta', x, \alpha|\theta)} U^{(x)}_{\theta \mapsto \theta'} \ . \tag{6.59}$$

The values $\Pr(\theta', x|\theta)$ must be positive only when $\theta' x \sim \theta$.

Now consider the generator $\mathfrak{R} = (\mathcal{R}, \mathcal{X}, \{\hat{T}^{(x)}_{\theta'\theta}\})$ with states $\theta \in \mathcal{R}$ and transition probabilities

$$\hat{T}^{(x)}_{\theta'\theta} = \Pr(\theta', x|\theta) = \sum_\alpha \Pr(\, \theta', x, \alpha \mid \theta\,) \tag{6.60}$$

The equation (6.59) implies the following:

(1) The subspaces $\mathcal{H}^{(\theta)}_S$ associated with each θ must all be unitarily equivalent to one another;

(2) For any $\rho_\theta \leq \Pi^{(\theta)}$ associated with a given subspace,

$$\mathcal{E}^{(x_t)} \dots \mathcal{E}^{(x_1)}(\rho_\theta) = \Pr_{\mathfrak{R}}(\, x_1 \dots x_t \mid \theta\,)\, U \rho_\theta U^\dagger$$

for some unitary U;

(3) Consequently, $\mathfrak{R}$ generates the same process as $\mathfrak{Q}$, and any information contained in ρ_θ beyond its Θ-subspace is superfluous.

Thus we can decompose $\mathcal{H}_S = \mathcal{H}_\Theta \otimes \mathcal{H}_U$, where $\mathcal{H}_\Theta$ is the space corresponding to the classical states of $\mathfrak{R}$ and $\mathcal{H}_U$ is a superfluous system undergoing unitary rotations.

The Kraus operators $K^{(x,\alpha)}$ in this format decompose as

$$K^{(x,\alpha)} = \sum_{\theta,\theta'} \sqrt{\Pr(\theta', x, \alpha|\theta)}\, |\theta'\rangle \langle\theta| \otimes V^{(x)}_{\theta\theta'}$$

for some unitary $V^{(x)}_{\theta\theta'}$. Let

$$\bar{K}^{(x,\alpha)} = \sum_{\theta,\theta'} \sqrt{\Pr(\theta', x, \alpha|\theta)}\, |\theta'\rangle \langle\theta|$$

Now consider the quantum generator $\overline{\mathfrak{Q}} = (\mathcal{H}_\Theta, \mathcal{X}, \{\bar{\mathcal{E}}^{(x)}\})$ where

$$\bar{\mathcal{E}}^{(x)}(\rho) = \sum_\alpha \bar{K}^{(x,\alpha)} \rho \bar{K}^{(x,\alpha)\dagger}$$

The embedded mixed states $\hat{\rho}_{\overleftarrow{x}}$ of this generator must, by the definition of Θ, commute with Θ. But now Θ is a maximal projective measurement on $\mathcal{H}_\Theta$; its projectors form the whole basis of the space. Thus all the $\hat{\rho}_{\overleftarrow{x}}$ must commute with one another. It must therefore be the case that the $\bar{\mathcal{E}}^{(x)}$ are decohering and have the form

$$\bar{\mathcal{E}}^{(x)}(\rho) = \sum_{\theta,\theta'} \Pr\left(\, \theta', x \mid \theta \,\right) \langle\theta|\rho|\theta\rangle\, |\theta'\rangle \langle\theta'|$$

which tells us that $\overline{\mathfrak{Q}}$ is just a direct implementation of the classical (and retrodictive) generator $\mathfrak{R}$. This completes the proof.

This result applies to all quantum generators, even those of classes which we have not yet considered. However, it can be interpreted in light of our two key examples of quantum generators: the forward and reverse q-machines.

We earlier found that, in the limit of asymptotically parallel generation, a q-machine is always more thermodynamically efficient than its corresponding ϵ-machine, in that it has a lower dissipation; this is a special case of Thm. 14. Yet this does not imply dissipation can be made to vanish for quantum generators of a process. In fact, only for processes whose forward ϵ-machine is also a retrodictor

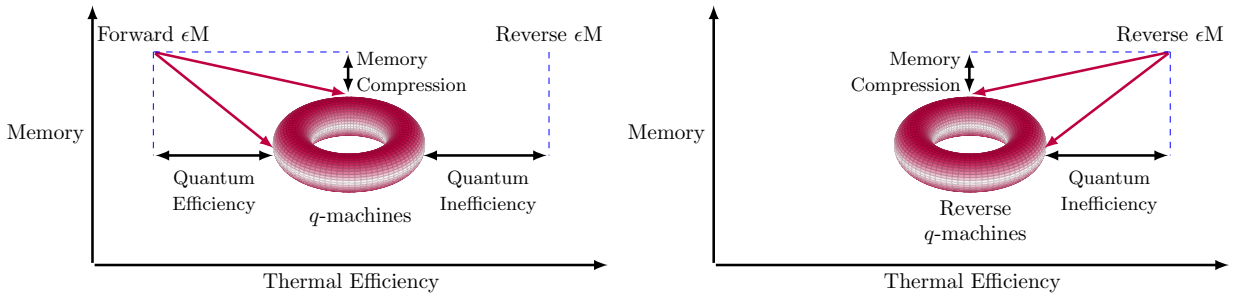

FIGURE 6.4. (*Left*) Performance trade-offs for q-machines. Under all ways of quantifying memory, the q-machines constructed from a predictor achieve nonnegative memory compression, and they also have a smaller dissipation ΔS_{loc}, rendering them more thermodynamically efficient. However, to achieve positive compression, they must also have a nonzero ΔS_{loc}, rendering them less efficient than a classical retrodictor. (*Right*) Performance trade-offs for reverse q-machines: Under all quantifications of memory, the reverse q-machines constructed from a retrodictor achieve nonnegative memory compression. However, to achieve positive compression, they must also have a nonzero dissipation ΔS_{loc}. The latter renders them less thermodynamically efficient than their classical counterparts. In both diagrams, the dependence of q-machine properties on phases $\{\phi_{xs}\}$ is represented by a torus.

can dissipation be made to vanish: this follows from the fact that the proof of Thm. 19 requires all embedding states commute, which we previously (in Chapter 4) demonstrated is only possible when the ϵ-machine is also co-unifilar. A further consequence is that positive memory compression is intrinsically tied to positive dissipation. The situation is heuristically represented in Fig. 6.4.

Similarly, reverse ϵ-machines which are memory-compressed by their reverse q-machine can no longer be thermodynamically efficient; the only reverse ϵ-machines which can be quantally compressed while remaining efficient are those which are also predictive generators. This also implies that memory compression is tied to dissipation; see also Fig. 6.4.

Overall, this is a profound result on the efficiency of quantum memory compression. Distinct from the classical case, where Thm. 18 established that *every* process has certain generators that do achieve zero dissipation, Thm. 19 implies that only *certain* processes have q-machines which are thermodynamically efficient; namely, those whose ϵ-machines are co-unifilar. Moreover, those particular processes achieve no memory compression; the q-machine is isomorphic to the ϵ-machine. The memory states, being orthogonally encoded, take no advantage of the quantum setting to reduce their memory cost.

6.6. Discussion

In single-shot quantum thermodynamics, von Neumann entropy is no longer the only meaningful quantifier of cost [29, 45]. In Chapter 5 we showed that using memory as a resource when quantally generating stochastic processes (see also Refs. [102, 106]). This makes comparing classical and quantum resource costs for generating a process challenging. Despite this, we derived bounds on the single-shot work cost of quantum generator implementations and showed that von Neumann entropies can be recovered as physically attainable work costs in the asymptotic limit of parallel generation. The first of these results opens the pathway for single-shot comparisons between classical and quantum resources in process generation, while the second allows direct comparison in terms of asymptotic quantities.

We identified the conditions under which local operations circumvent the thermodynamic dissipation ΔS_{loc} that arises from destroying correlation. We started by showing how a useful theorem can be derived using recent results on the fixed points of quantum channels. We applied it to the setting of local operations to determine the necessary and sufficient conditions for vanishing ΔS_{loc} in classical and quantum settings, with the aid of a generalized notion of quantum sufficient statistic. We employed this fundamental result to review and extend previous results on the thermodynamic efficiency of generators of stochastic processes. We confirmed a recent conjecture regarding the conditions for vanishing ΔS_{loc} in a classical generator. And, then, we showed the exact same conditions hold for quantum generators, even to the point of requiring orthogonal encoding of memory states. This implies the profound result that quantum memory compression and perfect efficiency ($\Delta S_{\text{loc}} = 0$) are incompatible.

It is appropriate here to recall the lecture by Feynman in the early days of thinking about quantum computing, in which he observed that quantum systems can only be simulated on classical (even probabilistic) computers with great difficulty, but on a fundamentally-quantum computer they could be more realistically simulated [85]. Here, we considered the task of simulating a classical stochastic process by two means: one by using fundamentally-classical but probabilistic machines and the other by using a fundamentally-quantum machine. Previous results generally indicated quantum machines are advantageous in memory for this task, in comparison to their classical counterparts.

Historically, this led to a much stronger notion of "quantum supremacy" than Feynman proposed: quantum computers may be advantageous in *all* tasks [**153**].

We have demonstrated that any quantum generator of a classical process, though potentially advantageous in memory, requires nonzero dissipation in order to cash in on that advantage. Furthermore, not every process necessarily has a q-machine that achieves zero dissipation. This is in sharp contrast to the classical outcome. And so, this returns us to the spirit of Feynman's vision for simulating physics, in which it may sometimes be the case that the best machine to simulate a classical stochastic process is a classical stochastic computer—at least, thermodynamically speaking.

CHAPTER 7

Where the light is: Statistical physics in carbon footprinting

The ages that are past
Are now a book with seven seals protected:
What you "the Spirit of the Ages" call
Is nothing but the spirit of you all,
Wherein the Ages are reflected.

Goethe, *Faust*, transl. Bayard Taylor

7.1. Introduction

As our planet faces environmental catastrophe of unprecedented scope, it has become necessary to address human impacts on the climate and global ecosystem through multilateral action by the world's governments. Warnings have been raised about the pitfalls of too short-sighted a response. For instance, treaties that only address pollution at the point of production may effectively outsource carbon-intensive activities from signatory nations to nonsignatories, a process known as *carbon leakage* [47, 148]. Instead, a holistic response that accounts for the multifaceted social relations driving environmental impacts is required [15, 70, 84, 159, 179]. Acquiring the data needed to make such holistic assessments, however, is a challenge in its own right.

Multiregional input-output (MRIO) tables provide data on the monetary transactions between national-level industries, both within a nation's borders and across them [92, 99, 100, 169, 211]. These can be used to construct models of the interconnected global economy. MRIO tables can be ecologically extended (called EE-MRIO tables) by adding local data on the environmental impacts that arise as byproducts of production.

Leontief analysis is a method frequently paired with EE-MRIO tables to attribute production-level impacts to the (potentially distant) activities that they support—typically consumption [9, 15, 16, 46, 47, 49, 54, 94, 101, 134, 135, 143, 144, 147, 148, 152, 178, 212, 215]. These attributed

impacts are said to be *embodied* in the consumed product. The flows of embodied impacts computed from EE-MRIO tables have been utilized in policy analysis by global and national government institutions [**1**, **2**, **3**, **65**]. Leontief analysis makes key assumptions, however, whose accuracy has been brought into question, particularly when applied to existing MRIO tables [**92**, **169**, **170**].

The following explores the *consequences* of these assumptions, through the lens of statistical mechanics. Given the considerable importance of the questions being asked in Leontief analysis and the potential for policy impact, it is important that we ensure the methods of data analysis used provide us with actual insights into the hidden structures at play. Otherwise, we run the risk of misidentifying the "spirit of our data" with the "spirit of our assumptions." We identify in this chapter that the assumptions of Leontief analysis are indeed strong drivers of the quantitative metrics observed.

In particular, we find that when certain reasonable conditions hold on an EE-MRIO dataset—namely, relatively small trade deficits among nations and geographically heterogeneous impact intensities—the directionality of embodied impact flows to and from extremal regions is heavily influenced by the impact intensities. Notably, the MRIO tables themselves have only a secondary effect. We call the phenomenon mediating this *eco-majorization*.

Our analysis of eco-majorization relies on the general theory of majorization and Lorenz curves [**123**] which have found wide application in economic and social analysis [**115**, **203**], statistical decision theory [**18**, **19**], and statistical physics [**29**, **30**, **71**, **73**, **77**, **116**]. We have thus far encountered majorization and Lorenz curves in Section 1.5 and Chapters 4 through 6. They are a means of characterizing the differences between two distributions without reducing those differences to a single parameter. The intuition of majorization has a natural foothold in the assumptions of Leontief analysis and EE-MRIO tables. Due to this, heterogeneities in impact intensities drive embodied flows in a manner directly analogous to physical diffusion.

Disentangling the results of a mode of analysis from mathematical artifacts that arise from the assumptions entailed is a difficult and often overlooked practice when working with complex data. For this reason, the use of specialized null models in network science has become increasingly popular [**10**, **56**, **154**, **155**, **171**, **210**]. The null models are used to randomly generate networks with special constraints designed to replicate the structural assumptions of a dataset while otherwise

reducing structural biases via random connections. Following this, we use null models specifically constructed to address the structures of EE-MRIO datasets, providing numerical confirmation of how majorization mediates the relationship between the assumptions of Leontief analysis and the embodied flows it detects.

This chapter is largely based on the manuscript *Nonequilibrium Thermodynamics in Measuring Carbon Footprints: Disentangling Structure and Artifact in Input-Output Accounting* [**110**]. In Section 7.2 we will provide an overview of the relevant features of input-output tables and Leontief analysis. Following this, in Section 7.3 we will construct a null network model of an MRIO and demonstrate that certain quantitative results of Leontief analysis are largely independed of the generated network data. In Section 7.4 we will provide a mathematical explanation of this phenomenon rooted using majorization theory, and define a particular form of majorization which in this context may be called *eco-majorization*.

7.2. Input-output models and ecological footprints

The following describes the basic components of EE-MRIO analysis, focusing on aspects relevant to the developments in Sections 7.3 and 7.4. Useful reviews of input-output methods are found in Refs. [**92**, **144**, **169**].

7.2.1. Basic input-output analysis. An *economy* $(\mathcal{I}, \mathcal{V}, \mathcal{D})$ is composed of finite sets of industrial sectors $\mathcal{I}$, value-added sectors $\mathcal{V}$, and final demand sectors $\mathcal{D}$. Value-added sectors usually include factors of production, such as labor, capital, land, and natural resources. Final demand sectors indicate the various forms of consumption: traditionally, private consumption, government spending, and business investment.

An economy's operation is cast as various kinds of *flow* between the sectors. To capture these for an economy, an input-output table is defined as the triple $(\mathbf{Z}, \mathbf{V}, \mathbf{D})$ of matrices with interindustry flows Z_{ij} $(i, j \in \mathcal{I})$; value-added flows V_{ui} $(i \in \mathcal{I},\ u \in \mathcal{V})$; and final-demand flows D_{ia}, $(i \in \mathcal{I},\ a \in \mathcal{D})$. Interindustry flows describe transactions between industrial firms; value-added flows describe factor returns such as wages, profits, and rent; final-demand flows describe the direct spending by individuals, governments, and businesses on consumable commodities, services, and fixed capital. By focusing on monetary flows, we are implicitly assuming that the necessary material

requirements for each industry and demand are met, and that supply equals demand overall. The effects of relaxing this assumption are beyond the scope of the present work.

Each matrix component describes the flow of money from the column sector to the row sector over a given time period (typically a year). For instance, Z_{ij} describes the total flow of money from sector j to sector i. Each industrial sector is assumed to be balanced, so that the total outlays equal the total output:

$$\underbrace{\sum_{u\in\mathcal{V}} V_{ui} + \sum_{j\in\mathcal{I}} Z_{ji}}_{\text{Outlays}} = \underbrace{\sum_{j\in\mathcal{I}} Z_{ij} + \sum_{a\in\mathcal{D}} D_{ia}}_{\text{Outputs}} \tag{7.1}$$

We define, respectively, the total outlays z_i of industry i, total value-added v_i by industry i, and total demand d_i of industry i as follows:

$$\begin{aligned} z_i &:= \sum_{u\in\mathcal{V}} V_{ui} + \sum_{j\in\mathcal{I}} Z_{ji} = \sum_{j\in\mathcal{I}} Z_{ij} + \sum_{a\in\mathcal{D}} D_{ia}, \\ v_i &:= \sum_{u\in\mathcal{V}} V_{ui}, \text{ and } d_i := \sum_{a\in\mathcal{D}} D_{ia} \ . \end{aligned} \tag{7.2}$$

We additionally define the total income as $Y := \sum_i v_i$, which is necessarily equal to the total spending $\sum_i d_i$ by Eq. (7.1).

A primary use of input-output analysis is to attribute the impacts of various activities to each of the final demand sectors. This provides a useful way to conceptualize the complex economy's interconnected causal relationships. To do this, we first define the technical coefficients C_{ij} as $C_{ij} := Z_{ij}/z_j$. These specify the outlays on activity i required to produce a single monetary unit of output in sector j. Equation (7.1)'s balance condition can then be written in matrix form as $\mathbf{z} = \mathbf{C}\mathbf{z} + \mathbf{d}$, a linear algebra problem whose solution (for $\mathbf{z}$) is:

$$\mathbf{z} = (\mathbf{I} - \mathbf{C})^{-1}\,\mathbf{d} \ , \tag{7.3}$$

where $\mathbf{I}$ is the identity matrix and one uses the matrix inverse. Written more explicitly we have:

$$z_i = \sum_{a\in\mathcal{D}} \left[(\mathbf{I} - \mathbf{C})^{-1}\,\mathbf{D}\right]_{ia} \ . \tag{7.4}$$

This expresses the total output as a column-sum of the matrix $(\mathbf{I} - \mathbf{C})^{-1}\mathbf{D}$. The sum allows us to break sector i's total output into parts, each attributed to a particular final demand $a \in \mathcal{D}$. The attribution matrix $\mathbf{A}$:

$$A_{ia} := \frac{\left[(\mathbf{I} - \mathbf{C})^{-1}\mathbf{D}\right]_{ia}}{z_i} , \tag{7.5}$$

describes, for each dollar of output in sector i, how much of that dollar is attributed to final demand a. Determining these attributions is called *Leontief analysis* after its originator [99, 100].

(Note how the value-added terms v_i have been whisked away in this analysis; the approach may be reversed to attribute outputs to factors $\mathcal{Y}$ rather than to final demands $\mathcal{D}$. While not explicitly considered here, the results derived for demand-based accounting apply symmetrically to factor-based accounting.)

It will be important, later, to note that $\sum_a A_{ia} = 1$. When this property holds for a matrix, in addition to the condition of nonnegative components, we say the matrix is *stochastic*. It has considerable importance for majorization.

The utility of Leontief analysis rests on two main assumptions [169]:

(L-1) **Sectors produce homogeneous products**: Due to this, we do not reweight the technical coefficients to reflect differences between the purchasing sectors—they purchase the same item.

(L-2) **Sectoral products are homogeneously priced**: Every buyer pays the same unit price. This again allows using the technical coefficients without modification to reflect differences in the inputs required per dollar for different purchasers.

These have been particularly singled out because of their crucial role in justifying the use of the single matrix $A_{ri,s}$ to compute all attribution vectors. We will return to the assumptions later to discuss how this constraint impacts the results of Leontief analysis.

For now, we mention two important reflections of these assumptions in the results above. First is the fact that $\sum_s A_{ri,s} = 1$ for all regions r and industries i. Additionally, because $(\mathbf{I} - \mathbf{C})^{-1}$ is a positive matrix [188], $A_{ri,s}$ is also positive. These two facts give the matrix $A_{ri,s}$ the property of *stochasticity*. That the direct impacts and attributed impacts can be related by a single stochastic

matrix is a reflection of the homogeneity in each of these assumptions, and is the central factor at play in our results.

Second, income itself can be treated as an impact. The value-added vector, when passed through the attribution matrix, returns the regional spending vector:

$$\hat{x}_s = \sum_{ri} v_{ri} A_{ri,s} \ , \tag{7.6}$$

While Sec. 7.2.2 explains the identity's mathematical origin, conceptually it arises from assumption (L-2): Impacts are attributed to demand in the exact same manner that incomes are attributed to spending. Both the stochasticity of $A_{ri,s}$ and Eq. (7.6) will be critically important when applying majorization.

7.2.2. Multiregional models. Multiregional input-output (MRIO) tables deepen the structure of input-output tables by dividing sectors into regions [**99**]. Specifically, we suppose there is a finite set $\mathcal{R}$ of regions and each set of sectors is organized as $\mathcal{I} = \mathcal{R} \times \mathcal{I}_0$, $\mathcal{V} = \mathcal{R} \times \mathcal{V}_0$, and $\mathcal{D} = \mathcal{R} \times \mathcal{D}_0$, where $\mathcal{I}_0$, $\mathcal{V}_0$, and $\mathcal{D}_0$ are the regional-level industrial, value-added, and final-demand sectors, respectively. An economy with this structure is said to be a *multiregional economy*.

The matrices and vectors described above can be adapted to this geographic picture by replacing each individual index $i \in \mathcal{I}$ (or others) with the pair (r, i), $r \in \mathcal{R}$ and $i \in \mathcal{I}_0$. This corresponds to re-envisioning the matrices and vectors as block-matrices and block-vectors, with rows and columns organized by regional blocks. In a block matrix each row and column indicates a pair (r, i), so that the first $|\mathcal{I}|$ indices correspond to all the industries in one region, and so on; Fig. 7.1 gives a visual aid. Our notation indicates this through use of commas: $Z_{ri,sj}$ is a compact way of denoting the matrix element $Z_{r\times|\mathcal{I}|+i,s\times|\mathcal{I}|+j}$, and similarly for v_{ri} and d_{ri}.

Additionally, simplifying constructs are often used in collecting and apportioning the data. For instance, in the dataset GTAP 8 (discussed shortly), $\mathcal{I}$ contains a duplicate of each industry, one for managing imports and the other for domestic industries. Cross-regional trade from r to s can only go from the domestic copy of industry i in r to the import copy of industry i in s. That is, cross-regional trade is not treated as cross-industry. This is represented visually by the diagonal structure of off-diagonal blocks in Fig. 7.1.

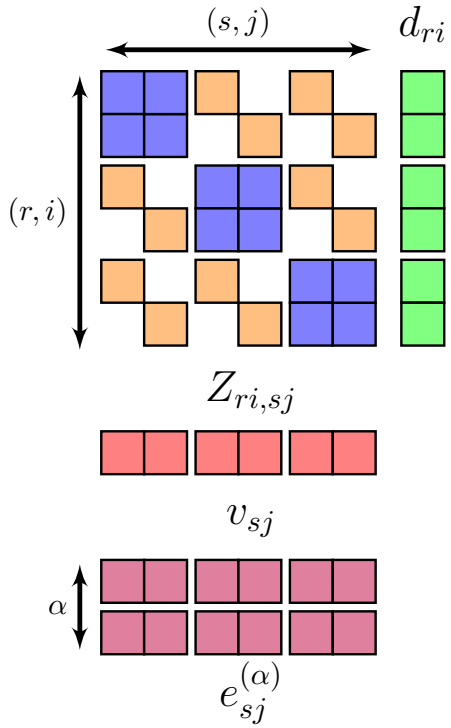

FIGURE 7.1. Block-matrix structure of a typical EE-MRIO table with $|\mathcal{I}| = 2$ and $|\mathcal{R}| = 3$. Blue and orange blocks are arranged into the single block-matrix $Z_{ri,sj}$, representing the value of all inter-industry transactions. Red blocks correspond to the value-added vector v_{ri}, consisting of all returns to wages, profits, and rent. And, the green blocks correspond to the final demand vector d_{ri}, consisting of all expenditures by consumers, governments, and investors. Pink blocks represent two separate ecological impact distributions $e_{ri}^{(\alpha)}$.

For another common simplification in trade datasets, while the inter-industry flows $Z_{ri,sj}$ may involve considerable inter-regional interaction, it is typically supposed that $V_{ru,si} = 0$ and $D_{ri,sa} = 0$ whenever $r \neq s$. In other words, regional factors are paid directly by a same-region industrial sector and regional consumption purchases directly from a same-region sector. (The direct consumption of imports is addressed by introducing intra-regional importing sectors to mediate the inter-regional interaction, usually doubling the size of $\mathcal{I}_0$.) This makes $\mathbf{V}$ and $\mathbf{D}$ block-diagonal.

Three concepts are important when identifying majorization. First, while global income and global spending equal one another, it is not necessarily the case that regional income and regional spending are equal. In fact, this difference is directly related to the trade deficit, by Eq. (7.1). Denoting

the regional income $\hat{y}_r := \sum_i v_{ri}$, the regional spending $\hat{x}_r := \sum_i d_{ri}$, and the inter-regional trade $\hat{Z}_{rs} := \sum_{ij} Z_{ri,sj}$, we have:

$$\underbrace{\hat{y}_r - \hat{x}_r}_{\text{Income – Spending}} = \underbrace{\sum_{s\in\mathcal{R}} \left(\hat{Z}_{rs} - \hat{Z}_{sr}\right)}_{\text{Exports – Imports}} .$$

We will refer to $\hat{\mathbf{y}} - \hat{\mathbf{x}}$ as simply the *regional deficit vector*. Its properties will be important in our study of majorization in Leontief analysis.

Second, when using MRIO tables to attribute economic activities to their corresponding demands, considerably more focus is given to the region of demand than the actual sector. Presently, this is our entire concern. We therefore define the regional attribution matrix $\hat{\mathbf{A}}$ as:

$$\hat{A}_{ri,s} := \sum_{a\in\mathcal{D}_0} A_{ri,sa} ,$$

where $\mathbf{A}$ is the block-matrix form of the attribution matrix defined in Eq. (7.5). In $\hat{\mathbf{A}}$, the rows are block-structured by regions, while the each column directly corresponds to a unique region. $\hat{\mathbf{A}}$ retains the stochastic property.

Third, a profoundly important identity emerges when applying the regional attributions matrix to the value-added vector from Eq. (7.2): we arrive at the regional spending vector $\hat{x}_r$:

$$\mathbf{v}\hat{\mathbf{A}} = \hat{\mathbf{x}} . \tag{7.7}$$

This follows from the relation:

$$\frac{v_{ri}}{z_{ri}} = 1 - \sum_{\substack{s\in\mathcal{R}\\ j\in\mathcal{I}_0}} C_{sj,ri} ,$$

that, in turn, is a consequence of Eq. (7.1). What Eq. (7.7) tells us is that the total value of the income for which each region's consumption is responsible is just that region's spending. And, this is the only consistent attribution if global income is to equal global spending. Leontiefian assumptions (L-1) and (L-2) suppose that environmental impacts may be attributed along the same lines as monetary flows. The fact that $\hat{A}_{ri,s}$ accurately attributes income to spending is a reflection of these assumptions and plays a central role in the appearance of majorization.

7.2.3. Environmentally extended tables. As mentioned already, we wish to explore the use of MRIO tables to attribute the impacts of economic activities to the demand sectors that stimulate them. Impacts themselves are often accounted for by environmentally extending the input-output table. A MRIO table with environmental extension is an *environmentally-extended* MRIO (EE-MRIO) table.

In the multiregional setting, an environmental extension is a family of block vectors $\{\mathbf{e}^{(\alpha)}\}$, indexed by impact α, with the form $e_{ri}^{(\alpha)}$, $r \in \mathcal{R}$ and $i \in \mathcal{I}_0$. The quantity $e_{ri}^{(\alpha)}$ gives the total impact of the activity in sector i and region r during the same time period as the other input-output matrices. For instance, if α corresponds to greenhouse gas emissions, then $e_{ri}^{(\alpha)}$ is the quantity of greenhouse gases emitted measured in carbon dioxide equivalents. The regional impacts are then given by $\hat{e}_r^{(\alpha)} = \sum_i e_{ri}^{(\alpha)}$ and the total impacts are denoted $E^{(\alpha)} = \sum_r \hat{e}_r^{(\alpha)}$.

Given impact α, we define the *α-intensity* $f_{ri}^{(\alpha)}$ of sector (r, i) as the ratio of environmental impact to economic impact. While there are many different definitions, for our needs we define it as:

$$f_{ri}^{(\alpha)} := \frac{e_{ri}^{(\alpha)}/v_{ri}}{E^{(\alpha)}/Y} \ .$$

That is, we take the ratio of the impact $e_{ri}^{(\alpha)}$ to the *value-added* v_{ri} by sector (r, i). This directly relates the emission at a given stage of production to the value added to the region by that production. (Not counted in this is the economic input from other industries, as this value has already been counted as value-added in another industry.) To remove dependence on the units used, we normalize the ratio by comparing it to the total ratio between emissions and income.

We can similarly define the *regional intensity* by

$$\begin{aligned} \hat{f}_r^{(\alpha)} &:= \frac{\hat{e}_r^{(\alpha)}/\hat{y}_r}{E^{(\alpha)}/Y} \\ &= \sum_{i \in \mathcal{I}_0} \frac{v_{ri}}{\hat{y}_r} f_{ri}^{(\alpha)} \ . \end{aligned} \tag{7.8}$$

As Eq. (7.8) indicates, this can be conceptualized either as a ratio of totals or the regional average of sectoral intensities.

Impacts happen at the point of production, but this activity meets a demand somewhere potentially geographically distant. Impacts may be thought of as becoming *embodied* in their product, which

travels from production to demand [211]. EE-MRIO tables have been frequently used to compute the flow of embodied impacts from production to final demand. While impact at the point of production is described the impacts vector $\hat{\mathbf{e}}^{(\alpha)}$, the embodied impacts attributed to each region r are given by the attribution vector:

$$\hat{\mathbf{a}}^{(\alpha)} := \mathbf{e}^{(\alpha)}\hat{\mathbf{A}} \ . \tag{7.9}$$

In theory, the quantity $\hat{a}_r^{(\alpha)}$ describes the total impact, originating anywhere in the global economy, required to meet the demands of region r. We again emphasize that this depends on the assumptions (L-1) and (L-2). For now, it is sufficient to appreciate that these are the standard calculations employed in Leontief analysis of EE-MRIO tables.

Embodied flows resulting from Leontief analysis are frequently quantified by one or both of the following proxy measures [9, 15, 16, 46, 49, 54, 101, 134, 143, 178, 215]:

(1) The net export $\boldsymbol{\xi}^{(\alpha)} = (\xi_r^{(\alpha)})$ of attributed impact, as a share of total global impact:

$$\xi_r^{(\alpha)} = \frac{\hat{e}_r^{(\alpha)} - \hat{a}_r^{(\alpha)}}{\sum_r \hat{e}_r^{(\alpha)}} \ .$$

(2) Or, the ratio $\boldsymbol{\rho}^{(\alpha)} = (\rho_r^{(\alpha)})$ of attributed impact to direct impact:

$$\rho_r^{(\alpha)} = \frac{\hat{a}_r^{(\alpha)}}{\hat{e}_r^{(\alpha)}} \ .$$

Naturally, these are closely related, as they ultimately express the relationship between the relative sizes of produced impacts and consumed impacts.

7.3. Data or artifact?: Consulting the null model

Null models, also *configuration models*, are a popular tool in the study of complex networks. They are widely applied to social [93, 210], animal [56, 155], and biological [154] networks to separate-out structures detected in empirical networks from those engendered by methodological assumptions [10, 171]. Often particular aspects of the networks—such as degree distribution—are held constant while all remaining aspects are randomized.

The following constructs a null model of global trade that maintains similar technical coefficients between industries, but "social coefficients"—those determining the relations between nations, dependency on imports, proportion paid to factors, and so on—are entirely randomized. This way, structures that consistently arise from Leontief analysis of networks drawn from this model cannot be attributed to social relations. They are, rather, artifacts of the assumptions of Leontief analysis.

7.3.1. Null models and GTAP 8. The Global Trade Aggregation Project (GTAP) [**140**] offers an extensive collection of transaction tables over a large number of regions and sectors. GTAP has been used as the EE-MRIO source data in many studies of embodied carbon emissions and other impacts [**15**, **16**, **46**, **152**, **215**]. We used GTAP 8 covering 134 regions (114 of which are countries), with 57 industrial sectors within each region, as well as 5 factor sectors (skilled and unskilled labor, capital, land, and natural resources), the standard 3 final demand sectors, and further sectors that fall outside the scope of our analysis. The data on these sectors was used to construct a multiregional input-output table. GTAP 8 comes directly with environmental extensions for carbon and energy use, and a number of satellite datasets exist providing further extensions. As an additional point of comparison, we used the satellite GMig2 [**208**] that provides information on labor inputs in human-years.

The carbon and labor data provided by GTAP 8 and GMig2 allowed us to compute the carbon and labor intensities, $\hat{\mathbf{f}}^{(\mathrm{CO_2})}$ and $\hat{\mathbf{f}}^{(\mathrm{L})}$, respectively, over 17 *megaregions* formed by aggregating the 134 GTAP standard regions.

We then constructed a null model for generating trade datasets over a reduced MRIO system with 4 factors, 16 industrial sectors, and 17 regions (aggregated from the GTAP sectors and regions). The null model has two parameters: a scalar ζ_X and a vector $\zeta_{C,i}$ taking values over industrial sectors. It makes liberal use of Dirichlet distributions as the source for drawing randomly-generated stochastic matrices from which the input-output tables are procedurally constructed.

Recall that a Dirichlet distribution $\mathrm{Dir}(\alpha_1, \ldots, \alpha_K)$ is defined on the simplex of probability vectors over K elements. The density function of $\mathrm{Dir}(\alpha_1, \ldots, \alpha_K)$ is given by:

$$d(p_1, \ldots, p_K) \propto \prod_{i=1}^{K} p_i^{\alpha_i - 1} \ .$$

Notably, when $\alpha_1, \ldots, \alpha_K$ all equal 1, the Dirichlet distribution draws from all probability vectors with equal weight (a uniform Dirichlet). Otherwise, it tends to draw probability vectors whose weight distribution is similar to that of the vector $(\alpha_1 - 1, \ldots, \alpha_K - 1)$, varying to a degree that is inversely proportional to $\alpha_0 := \sum_{i=1}^{K} \alpha_i$ (a nonuniform Dirichlet).

The null model uses nonuniform Dirichlet distributions to draw the values of the regional technical coefficients so that they are similar to the global coefficients $\tilde{C}_{ij}^{(\mathrm{GTAP})}$ from GTAP 8 with a degree of variation controlled by the parameter $\zeta_{C,i}$ for each industrial sector i. All other coefficients—which we call *social coefficients*—are constructed from combining uniform Dirichlet distributions. Social coefficients are those that depend on exogenous social parameters, such as acceptable wage levels, trade agreements, and consumer ethics. Technical constraints, as calculated by the GTAP 8 dataset, are emulated in the null model; social constraints are randomized entirely.

Using the technical and social coefficients, we can use another nonuniform Dirichlet to generate the regional spending and income distributions so that the regional deficit is small, controlled by parameter ζ_X. Baseline values $\bar{\zeta}_{C,i}$ and $\bar{\zeta}_X$ are derived to replicate the variation among regional technical coefficients and deficits in the GTAP 8 data.

To begin, we define the regional technical coefficients $\hat{C}_{i,sj}$ and global technical coefficients $\tilde{C}_{i,j}$ to be:

$$\hat{C}_{i,sj} = \frac{\sum_r Z_{ri,sj}}{\sum_{r,i} Z_{ri,sj}} \ , \quad \tilde{C}_{i,j} = \frac{\sum_{r,s} Z_{ri,sj}}{\sum_{r,i,s} Z_{ri,sj}} \ .$$

The first characterizes the input requirements of regional industries relative to other industries. Note that the social choice of from which regions to acquire inputs is not described by this matrix. The second $\tilde{C}_{i,j}$ averages these industrial requirements over the entire globe.

Our null model is then constructed as follows:

(1) The global technical coefficients $\tilde{C}_{ij}^{(\mathrm{GTAP})}$ from GTAP 8 are used as a baseline to generate regional technical coefficients. Namely, for each region r and industry j the technical coefficients $\hat{C}_{i,rj}$ are drawn as:

$$\hat{C}_{i,rj} \sim \mathrm{Dir}\left(\alpha_i = \zeta_{C,i} \hat{C}_{i,rj}\right) \ .$$

This ensures that the technical composition of local industries is similar to that found in the GTAP 8 database. The vector parameter $\zeta_{C,i}$ controls the degree of similarity.

(2) The regional spending distribution $s_{a,r}$, consumption coefficients $c_{i,ra}$, import coefficients $M_{i,sj}$, regional supply coefficients $R_{r,si}$, value-added coefficients U_{ri}, and factor coefficients $F_{u,ri}$ are all drawn from uniform Dirichlets:

$$\begin{aligned} s_{a,r} &\sim \text{Dir}\,(\alpha_a = 1) \\ c_{i,ra} &\sim \text{Dir}\,(\alpha_i = 1) \\ (M_{i,sj}, 1 - M_{i,sj}) &\sim \text{Dir}\,(\alpha_1 = 1, \alpha_2 = 1) \\ R_{r,si} &\sim \text{Dir}\,(\alpha_r = 1) \\ (U_{ri}, 1 - U_{ri}) &\sim \text{Dir}\,(\alpha_1 = 1, \alpha_2 = 1) \\ F_{u,ri} &\sim \text{Dir}\,(\alpha_u = 1) \end{aligned} \quad . \tag{7.10}$$

In order, these describe: the distribution of regional spending among the final demand sectors; the distribution of spending by each final demand sector among its products of consumption; the proportion by which a given industrial sector will import a particular input instead of source it domestically; the regional probability of importing an input from another specific region; the proportion of capital outlay towards factors by a given industrial sector; and the distribution of that capital among the factors. These describe socially-determined relations between the regions and sectors. These are the relationships we seek to randomize in the null model.

(3) The full technical coefficients are constructed as:

$$C_{ri,sj} = \begin{cases} (1 - U_{ri})(1 - M_{i,sj})\hat{C}_{i,sj} & r = s \\ (1 - U_{ri})R_{r,si}M_{i,sj}\hat{C}_{i,sj} & r \neq s \end{cases} .$$

From these we define the matrix:

$$K_{r,s} := \sum_{i,j,a} U_{ri} \left(\mathbf{I} - \mathbf{C}\right)^{-1}_{ri,sj} c_{j,sa} s_{a,s} \ .$$

This matrix describes the likelihood that money originating in a final demand sector in region s ends up in a value-added sector in region r. Now, given any particular global spending distribution $\hat{\mathbf{x}}$, the regional incomes are determined by $\hat{\mathbf{y}} = \mathbf{K}\hat{\mathbf{x}}$. We calculate the eigenvector $\boldsymbol{\pi}$ such that $\mathbf{K}\boldsymbol{\pi} = \boldsymbol{\pi}$. (Its existence is guaranteed by the fact that $\mathbf{L}$ is stochastic in its left index.) If spending matched this eigenvector, then no region would hold a trade deficit or surplus. All trade would be equally balanced. The spending in our model is drawn from a Dirichlet as:

$$\hat{x}_r \sim \mathrm{Dir}(\alpha_r = \zeta_X \pi_r) \ .$$

Thus, the parameter ζ_X controls the scale of trade imbalance.

(4) The multiregional input-output table can now be constructed as:

$$\begin{aligned} D_{ri,ra} &= c_{i,ra} s_{a,r} \hat{x}_r \\ Z_{ri,sj} &= \sum_{\substack{t\in\mathcal{R} \\ k\in\mathcal{I}_0 \\ a\in\mathcal{D}_0}} C_{ri,sj} \left(\mathbf{I} - \mathbf{C}\right)^{-1}_{sj,tk} D_{tk,ta} \\ V_{ru,ri} &= \sum_{\substack{s\in\mathcal{R} \\ j\in\mathcal{I}_0 \\ a\in\mathcal{D}_0}} F_{u,ri} U_{ri} \left(\mathbf{I} - \mathbf{C}\right)^{-1}_{ri,sj} D_{sj,sa} \end{aligned} \quad . \tag{7.11}$$

From $Z_{ri,sj}$ we compute the sector activities $z_{ri} = \sum_{s,j} Z_{ri,sj}$ and finally the attribution matrix:

$$A_{ri,sa} = \frac{\left[\left(\mathbf{I} - \mathbf{C}\right)^{-1} \mathbf{D}\right]_{ri,sa}}{z_{ri}} \ .$$

The parameters will be generally set at baseline values $\bar{\zeta}_{C,i}$ and $\bar{\zeta}_X$, determined by:

$$\begin{aligned} \bar{\zeta}^{-1}_{C,i} &= \sum_{\substack{r\in\mathcal{R} \\ j\in\mathcal{I}_0}} \frac{v^{(\mathrm{GTAP})}_{ri}}{\sum_s v^{(\mathrm{GTAP})}_{si}} \hat{C}^{(\mathrm{GTAP})}_{j,ri} \log\left(\frac{\hat{C}^{(\mathrm{GTAP})}_{j,ri}}{\tilde{C}^{(\mathrm{GTAP})}_{j,i}}\right) \\ \bar{\zeta}^{-1}_{X} &= \sum_{r\in\mathcal{R}} \hat{x}^{(\mathrm{GTAP})}_r \log\left(\frac{\hat{x}^{(\mathrm{GTAP})}_r}{\pi^{(\mathrm{GTAP})}_r}\right) \end{aligned} \quad . \tag{7.12}$$

This uses the Kullback-Liebler divergence [36] as a proxy for the degree of difference between various empirical distributions. This sets the baseline for similar variations within the null model.

Additionally, we built a second, simpler null model for producing environmental extensions. It generates impact intensities, with control over the heterogeneity of intensity across regions and sectors. The imaginary resource that this impact represents is termed *unobtainium* or simply U. The null model involves constructing a new parameter ζ_U. We sample *unnormalized* intensities, denoted $\phi_{ri}^{(\mathrm{U})}$, as:

$$\phi_{ri}^{(\mathrm{U})} \sim \mathrm{Dir}(\alpha_{ri} = \zeta_U) \ .$$

ζ_U is set to a low value—our baseline is $\bar{\zeta}_U = 0.05$—resulting in the Dirichlet sampling distributions with high peakedness around randomly selected sectors and so assuring heterogeneity. By increasing ζ_U, we can reduce heterogeneity. The normalized intensities, for a given MRIO, are then determined as:

$$f_{ri}^{(\mathrm{U})} = \frac{\phi_{ri}^{(\mathrm{U})}}{\sum_{s,j} v_{sj} \phi_{sj}^{(\mathrm{U})}} \ .$$

Combining both null models allows generating a wide sampling of EE-MRIO tables with random social coefficients, while controlling for technical coefficients, regional deficits, and impact heterogeneity. These last two attributes, in particular, are important for majorization, as we will show.

7.3.2. Flows in the null model. Leontief analysis of the null-model MRIO tables paired with the empirical labor and CO_2 distributions exposed a strong bias for high-intensity regions to be net exporters and low-intensity regions to be net importers of embodied CO_2 emissions. This finding mirrors previous results [**9**, **15**, **16**, **46**, **47**, **49**, **54**, **94**, **101**, **135**, **143**, **147**, **148**, **178**, **212**, **215**]. The key difference is that our trade networks were entirely randomized. Importantly, this suggests that our findings are not a consequence of global trade relations, but an *artifact resulting from pairing unequal intensities with Leontief analysis*. Given that this is a natural research strategy for the field, the conclusion is a cautionary lesson.

We drew 1000 samples from the null model with $\zeta_{C,i}$ and ζ_X set at the baseline values $\bar{\zeta}_{C,i}$ and $\bar{\zeta}_X$. For each sample from the null model, we juxtaposed the resulting attribution matrix with the predetermined CO_2 intensity $\hat{\mathbf{f}}^{(\mathrm{CO_2})}$ and labor intensity $\hat{\mathbf{f}}^{(\mathrm{L})}$ to calculate the embodied flows of these impacts. For each region and impact, we calculated the proportion of samples for which

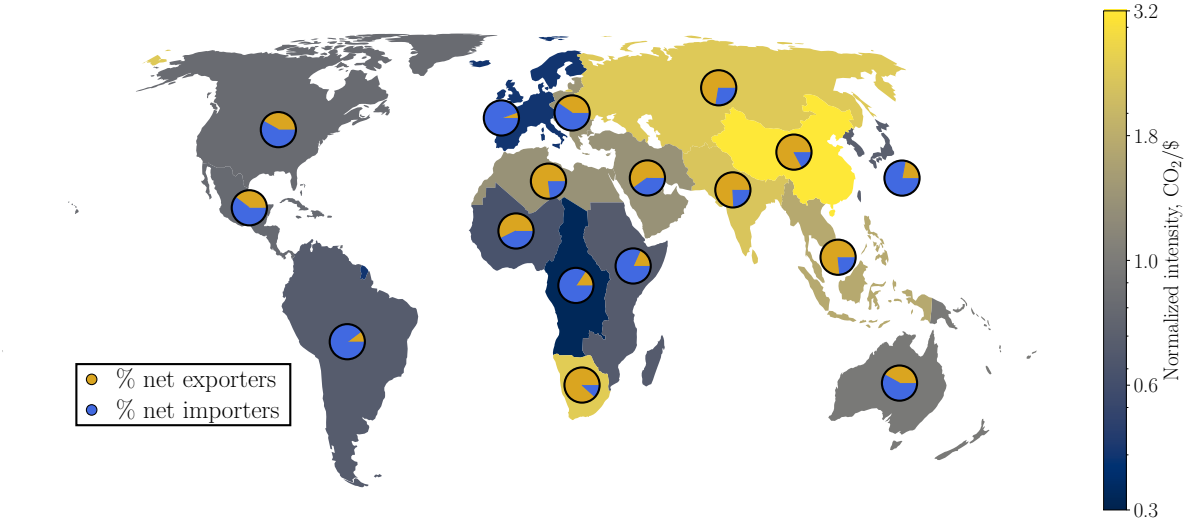

FIGURE 7.2. Embodied CO_2 flows: Two sets of information for each of 17 megaregions. *Regional color*: Normalized intensity $\hat{f}_r^{(CO_2)}$ for each region r with bright yellow indicating higher intensities and dark blue indicating lower. *Regional pie charts*: Proportion of null models for which the given region was a net exporter (yellow) or a net importer (blue). A chart's amount of yellow corresponds to the quantity $\Xi_r^{(\alpha)}$—sample proportion of net positive exports; see text. This map uses the Equal Earth projection [**168**].

the net exports $\xi_r^{(\alpha)}$ were positive, denoting the proportion $\Xi_r^{(\alpha)}$. In addition to being a function of the region r, $\Xi_r^{(\alpha)}$ is dependent on the null model parameters as well as the impact distribution $\hat{\mathbf{e}}^{(\alpha)}$, and may be termed the *null likelihood of net exports* for region r with respect to resource α. Figure 7.2 displays the impact intensities in color and the null export likelihood as a pie chart for each region with respect to CO_2.

Despite the null model having no preferred directionality between regions or, for that matter, even any preferred tendency between imported and domestic sources, one sees that high-intensity regions have a strong tendency to export embodied CO_2 while low-intensity regions have a strong tendency to import embodied CO_2. Moderately-intense regions do not exhibit bias.

The relationship between intensity $\hat{f}_r^{(\alpha)}$ and export likelihood $\Xi_r^{(\alpha)}$ can be expressed using a nonlinear measure of correlation, such as Kendall's τ [**91**]. We found that the correlation between the two quantities for carbon was $\tau^{(CO_2)} = 0.68$. While, for labor, the correlation was a remarkable

$\tau^{(\mathrm{L})} = 0.96$. Likely, this is due to the fact that labor intensities are determined on the factor level—consequently, there is no intra-regional variation in labor intensity that confounds the relationship between intensity and global trade.

In this way, the null models demonstrate that complete randomization of social factors in MRIO tables has little effect on the directionality of embodied flows: they are still directed from high- to low-intensity regions. The following section offers an explanation, via majorization, for how the assumptions underlying Leontief analysis itself drive the correlation between impact intensities and embodied flows.

7.4. Eco-majorization: Visualizing the effects of Leontiefian assumptions

Section 7.3 discusses the primary results from our null-model MRIO table. In this section we will provide a definition of *eco-majorization*. We show how, in the framework of Leontief analysis, it drives global flows of trade, in a manner that explains the results of Section 7.3. This leads us to analyze the conditions under which Leontief analysis is biased towards eco-majorized results. Lastly, we alter the parameters of the null model to explore the consequences of relaxing these conditions. This reveals a strong relationship between eco-majorization and the directionality of embodied CO_2 flows.

7.4.1. Majorization and statistical mechanics. To fully appreciate the results coming from the null models, we must compare various distributions. To this end, we introduce a tool with a long tradition that recently gained traction and found significant development in information theory and statistical physics. The tool in question is *majorization*—more specifically, relative majorization [**123**,**157**,**203**]. We describe the necessary background below and also provide a primer in Fig. 7.3.

Given two probability distributions $\mathbf{p} = (p_i)$ and $\mathbf{q} = (q_i)$ defined over a finite set $\mathcal{S}$, we construct their *Lorenz curve* $\ell_{\mathbf{p},\mathbf{q}} : [0,1] \to [0,1]$ as the piecewise convex function connecting the points (x_n, y_n):

$$x_n = \sum_{m=1}^{n} p_{i_m} \ , \quad y_n = \sum_{m=1}^{n} q_{i_m} \ ,$$

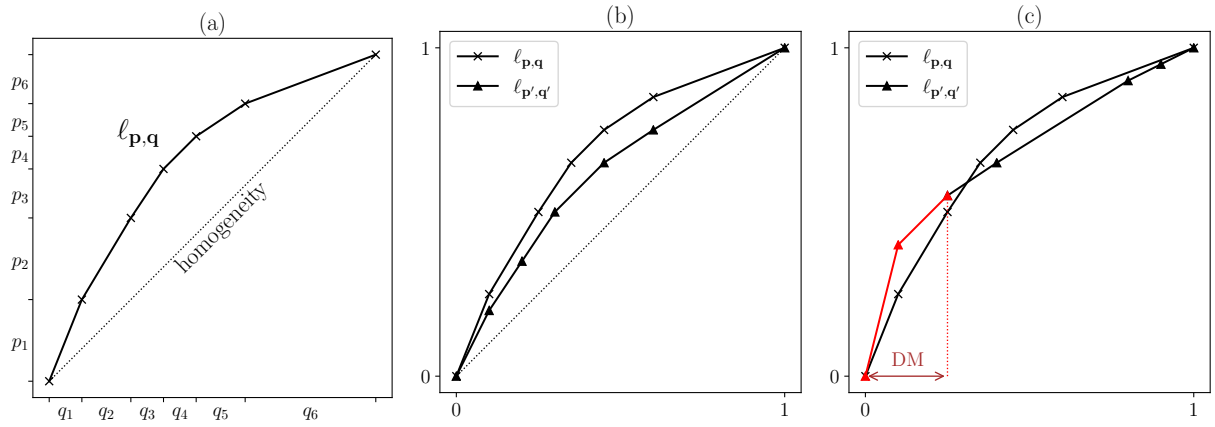

FIGURE 7.3. Majorization Primer: *(a)* Example a Lorenz curve for a pair of distributions $(\mathbf{p}, \mathbf{q})$ over 6 elements. We assume the elements are indexed so that p_i/q_i is monotonically decreasing. $\mathbf{p}$ and $\mathbf{q}$ are not homogeneous with respect to one another, and so the Lorenz curve bows out above the diagonal. *(b)* An example of two pairs, $(\mathbf{p}, \mathbf{q})$ and $(\mathbf{p}', \mathbf{q}')$, such that $(\mathbf{p}, \mathbf{q}) \succeq (\mathbf{p}', \mathbf{q}')$. Visually, this means that the second Lorenz curve is fully beneath the first and, therefore, closer to the line of homogeneity. *(c)* Example where majorization does not hold. The extent of failure can be described by the *dismajorization*, defined as the total $\mathbf{q}'$-probability associated with the majorization-breaching vertices. "DM" stands for dismajorization.

where $(i_m) = (i_1, i_2, \dots)$ orders the set $\mathcal{S}$ so that p_{i_m}/q_{i_m} is monotonically decreasing in m.

Given two pairs of distributions $(\mathbf{p}, \mathbf{q})$ and $(\mathbf{p}', \mathbf{q}')$, if $\ell_{\mathbf{p},\mathbf{q}}(x) \geq \ell_{\mathbf{p}',\mathbf{q}'}(x)$ for all $x \in [0, 1]$, then we say that $(\mathbf{p}, \mathbf{q})$ *majorizes* $(\mathbf{p}', \mathbf{q}')$ [203]:

$$(\mathbf{p}, \mathbf{q}) \succeq (\mathbf{p}', \mathbf{q}') \ .$$

An intuitive application of majorization in fact arose in its first use as an indicator of economic inequality [115]. In this context, if $\mathbf{p}$ describes the population distribution and $\mathbf{q}$ the wealth distribution, the Lorenz curve completes statements of the type "The richest x% of the population holds $\ell_{\mathbf{p},\mathbf{q}}(x)$% of the wealth." One country can be said to be definitively more unequal than another if its Lorenz curve is always higher. Majorization generalizes this relationship.

The connection between majorization and nonequilibrium statistical mechanics, as well as its connection to our work, arises as a consequence of the *Blackwell-Sherman-Stein* (BSS) theorem [18, 19]: $(\mathbf{p}, \mathbf{q}) \succeq (\mathbf{p}', \mathbf{q}')$ if and only if there exists a stochastic matrix $\mathbf{T}$ such that $\mathbf{Tp} = \mathbf{p}'$ and $\mathbf{Tq} = \mathbf{q}'$. This connection is profound, given the frequent appearance of stochastic matrices in

statistical mechanics, information processing, stochastic processes, game theory, and decision theory, and far more.

It has quite recently found significant application in the intersection between information theory and nonequilbrium statistical mechanics. Actions taken upon a thermodynamic system can be described as stochastic matrices over a system's microstates. In this setting, it can be shown that any action (described by stochastic matrix $\mathbf{t}$) that satisfies (i) energy conservation, (ii) Liouville's theorem or information conservation, and (iii) access to a thermal reservoir of temperature T must obey the constraint:

$$\sum_j t_{i|j}\gamma_j(T) = \gamma_i(T) \ , \tag{7.13}$$

where γ_i is the Boltzmann-Gibbs distribution:

$$\gamma_i(T) = \frac{e^{-E_i/kT}}{Z(T)}, \quad Z(T) = \sum_i e^{-E_i/kT}$$

and $\mathbf{E} = (E_i)$ defines the energies of each microstate i [77].

The significance of this observation is that even when operating on a distribution $\mathbf{p}$ that is far from equilibrium—that is, *not* equal to $\boldsymbol{\gamma}(T)$—the actions we take must still satisfy Eq. (7.13). Using the BSS theorem, we learn that a distribution $\mathbf{p}$ can be physically transformed into another $\mathbf{p}'$ using a bath at temperature T only if $(\mathbf{p}, \boldsymbol{\gamma}(T)) \succeq (\mathbf{p}', \boldsymbol{\gamma}(T))$. It is said in this case that $\mathbf{p}$ *thermo-majorizes* $\mathbf{p}'$ [71]. This connection between majorization and thermodynamics has been used to derive a number of relations that leverage thermodynamic fluctuations to extract work in the nanoscale, single-shot regime [29, 30, 71, 116, 157].

When majorization does not hold, it may hold approximately—a fact that can still result in many of the same consequences of majorization. It will be useful to have a quantification of dismajorization that allows us to distinguish between small and large violations of majorization.

We define the *dismajorization* $\mathrm{DM}\left[(\mathbf{p}, \mathbf{q}); (\mathbf{p}', \mathbf{q}')\right]$ of two pairs of curves in the following way. Let (i_m) be the same ordering of indices used above and let x'_n and y'_n be defined in the same manner as x_n and y_n but for the pair $(\mathbf{p}', \mathbf{q}')$. Finally, let $\mathcal{N}$ be the set of n such that $y'_n > \ell_{\mathbf{p},\mathbf{q}}(x'_n))$. Then

we define:

$$\mathrm{DM}\left[(\mathbf{p},\mathbf{q});(\mathbf{p}',\mathbf{q}')\right] = \sum_{n\in\mathcal{N}} q'_{i_n}$$

as the total probability associated with the points where the second Lorenz curve exceeds the first.

7.4.2. Eco-majorization and Leontief bias. Majorization's key benefit is that it readily explains the internal mechanics of input-output analysis. In this way, the present use is yet another example of generalizing thermodynamic logic to new settings. Such applications have, for instance, already been powerfully applied to develop quantum resource theories, which make frequent use of majorization to study entanglement and other quantum properties as a nonfungible resources [**29**, **30**, **71**, **73**, **77**].

The following, using it, shows that Leontief analyses tends to detect flows of embodied impacts from high-intensity regions to low-intensity regions. The effect is physically analogous to particles diffusing from high-density to low-density regions. In this way, majorization connects these two settings.

Majorization is defined on probability vectors, whose total sum is normalized to 1, but we will for simplicity write unnormalized vectors in the majorization pairs as a shorthand for the majorization of their normalized forms. We will demonstrate that if the following conditions hold for an EE-MRIO with impact α:

(MRIO-1) The regional impact intensities $\hat{\mathbf{f}}^{(\alpha)}$ are highly heterogeneous and

(MRIO-2) The regional deficit $\hat{\mathbf{x}} - \hat{\mathbf{y}}$ is small as a proportion of regional income across regions,

then with high probability Leontief analysis results in (or approximately results in) the majorization:

$$\left(\hat{\mathbf{e}}^{(\alpha)}, \hat{\mathbf{y}}\right) \succeq \left(\hat{\mathbf{a}}^{(\alpha)}, \hat{\mathbf{y}}\right) \ . \tag{7.14}$$

As this phenomenon links both ecological impacts and economic activity levels, we call this relation *eco-majorization*, where the prefix may refer to either. We note that both assumptions hold for the GTAP 8 dataset.

Since the regional income distribution $\hat{\mathbf{y}}$ plays a role similar to the thermodynamic Gibbs distribution, the stated relation tells us that the embodied impacts $\hat{\mathbf{a}}^{(\alpha)}$ are more similarly distributed to the

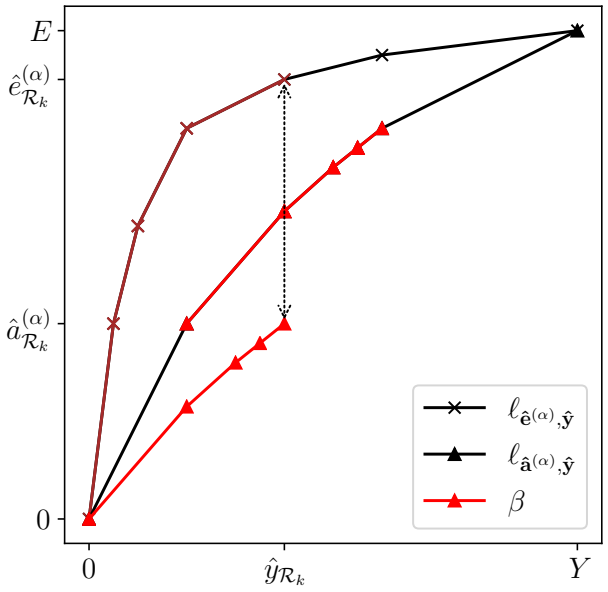

FIGURE 7.4. Most intense regions must be net exporters: Graphical proof that, as long as eco-majorization holds, the k most intense regions $\mathcal{R}_k$ will have $\hat{\mathbf{a}}^{(\alpha)}_{\mathcal{R}_k} \leq \hat{\mathbf{e}}^{(\alpha)}_{\mathcal{R}_k}$. This rests on the fact that the partial Lorenz curve β corresponding only to the regions $\mathcal{R}_k$ must be below the Lorenz curve $\ell_{\hat{\mathbf{a}}^{(\alpha)},\hat{\mathbf{y}}}$. And that, in turn, falls below $\ell_{\hat{\mathbf{e}}^{(\alpha)},\hat{\mathbf{y}}}$.

regional incomes than to the local impacts $\hat{\mathbf{e}}^{(\alpha)}$. This necessitates transferring embodied impacts from high-intensity regions, where distributional imbalance is most positive, to low-intensity regions, where it is most negative.

We quantify the previous statement using net exports $\boldsymbol{\xi}^{(\alpha)}$ and impact ratios $\boldsymbol{\rho}^{(\alpha)}_r$. Let $\mathcal{R}_k$ be that subset of regions containing the k regions with the highest values of $\hat{f}^{(\alpha)}_r$. Then it can be shown that, if eco-majorization holds:

$$\sum_{r\in\mathcal{R}_k} \xi^{(\alpha)}_r \geq 0 \text{ and } \sum_{r\in\mathcal{R}_k} \frac{\hat{e}^{(\alpha)}_r}{E}\rho^{(\alpha)}_r \leq 1\ . \tag{7.15}$$

Both indicators then tell us that regions in $\mathcal{R}_k$ must be net exporting or, at least, never net importing. Figure 7.4 presents a visual proof using Lorenz curves. Since a Lorenz curve is monotonically decreasing in slope, the curve $\beta(x)$ formed by taking only a subset of segments must always be lower in height than the original curve. From this it can be seen that when $\left(\hat{\mathbf{e}}^{(\alpha)},\hat{\mathbf{y}}\right) \succeq \left(\hat{\mathbf{a}}^{(\alpha)},\hat{\mathbf{y}}\right)$, we must have $\sum_{r\in\mathcal{R}_k}\hat{\mathbf{a}}_r \leq \sum_{r\in\mathcal{R}_k}\hat{\mathbf{e}}_r$, from which Eqs. (7.15) hold.

Eco-majorization, then, places rigid constraints on the directionality of trade flows. Our null model simulation decisively demonstrates it is at play in the observed relationship between intensity and exports: embodied labor flows were eco-majorized for 100% of the simulated networks and embodied CO_2 flows were eco-majorized for 72% of the simulated networks.

These results suggest that while eco-majorization is not guaranteed, it is a highly ubiquitous phenomenon among randomly generated MRIO tables. The issue, we argue, rests in the presence of the two conditions (MRIO-1) and (MRIO-2).

Due to Eqs. (7.6) and (7.9), the BSS theorem automatically entails:

$$\left(\mathbf{e}^{(\alpha)}, \mathbf{v}\right) \succeq \left(\hat{\mathbf{a}}^{(\alpha)}, \hat{\mathbf{x}}\right) \ . \tag{7.16}$$

This differs from Eq. (7.14) in two key respects. First, the lefthand side refers to the sectoral impact distribution $\mathbf{e}^{(\alpha)}$ and the factor income distribution $\mathbf{v}$ rather than the regionalized distributions $\hat{\mathbf{y}}^{(\alpha)}$ and $\hat{\mathbf{y}}$, respectively. We call this difference *regionalization*. Second, the righthand side uses the spending distribution $\hat{\mathbf{x}}$ instead of the income distribution $\hat{\mathbf{y}}$. This is of little concern issue when the regional deficit is small, as assumed in (MRIO-2). By the nature of Lorenz curves, small changes in the underlying distributions result in correspondingly small changes in curve's shape.

The subtlety, and the only reason why 100% of simulated networks do not display majorization for all impacts, arises from regionalization: Eq. (7.16) shows that majorization holds for the full sectoral distributions, but does not say anything about regional distributions. It is entirely *possible* that Eq. (7.16) may be true and Eq. (7.14) may be false. The crux of this issue is, in fact, the same as a major point of thermo-majorization theory: namely, that when the majorizing pair of distributions are coarse-grained, majorization might no longer hold. Notably, many recent results on the work cost of driving systems away from thermodynamic equilibrium exploit this phenomenon [**30**, **157**].

We are not interested here in how to induce this phenomenon. Rather, we are interested in why it does not appear to naturally arise in either the existing trade data or the null model. Our argument is that condition (MRIO-1) significantly constrains the possible configurations that may result in a violation of Eq. (7.14).

The argument is as follows. For a set $\mathcal{S}$ of regions, define:

$$\hat{y}_{\mathcal{S}} := \sum_{s \in \mathcal{S}} \hat{y}_s \ , \quad \hat{f}_{\mathcal{S}}^{(\alpha)} := \frac{\sum_{s \in \mathcal{S}} \hat{e}_s^{(\alpha)}/E}{\hat{y}_{\mathcal{S}}/Y} \ , \text{ and } \hat{f}_{\mathcal{S}}^{(\alpha)'} := \frac{\sum_{s \in \mathcal{S}} \hat{a}_s^{(\alpha)}/E}{\hat{y}_{\mathcal{S}}/Y} \ .$$

Further, let $\hat{y}_k := \hat{y}_{\mathcal{R}_k}$ for simplicity. To violate majorization there must be a set $\mathcal{S}$ of regions such that:

$$\hat{f}_{\mathcal{S}}^{(\alpha)'} \geq \left(\frac{\hat{y}_{\mathcal{S}} - \hat{y}_k}{\hat{y}_{k+1} - \hat{y}_k} \right) \hat{f}_{\mathcal{R}_k}^{(\alpha)} + \left(\frac{\hat{y}_{k+1} - \hat{y}_{\mathcal{S}}}{\hat{y}_{k+1} - \hat{y}_k} \right) \hat{f}_{\mathcal{R}_{k+1}}^{(\alpha)} \ , \tag{7.17}$$

where k is the unique integer such that $\hat{y}_k \leq \hat{y}_{\mathcal{S}} < \hat{y}_{k+1}$. (This simply restates the definition of majorization via Lorenz curves.) We can actually suppose without loss of generality that $\hat{y}_k = \hat{y}_{\mathcal{S}}$. This can be achieved by splitting regions into smaller but structurally identical subregions. So, Eq. (7.17) can be expressed more simply as $\hat{f}_{\mathcal{S}}^{(\alpha)'} \geq \hat{f}_{\mathcal{R}_k}^{(\alpha)}$.

Now, we may rewrite $\hat{f}_{\mathcal{S}}^{(\alpha)'}$ as:

$$\hat{f}_{\mathcal{S}}^{(\alpha)'} = \sum_{\substack{r \in \mathcal{R} \\ i \in \mathcal{I}_0}} \frac{\hat{A}_{ri,s} y_{ri}}{\hat{y}_{\mathcal{S}}} f_{ri}^{(\alpha)} \ . \tag{7.18}$$

This considerably constrains the structure of matrices $\hat{\mathbf{A}}$ that yield a large value for $\hat{f}_{\mathcal{S}}^{(\alpha)'}$. Specifically, if $\hat{f}_{\mathcal{S}}^{(\alpha)'} \geq \hat{f}_{\mathcal{R}_k}^{(\alpha)}$, then $\hat{A}_{ri,s}$ either must put great weight on the most intense sectors within the regions of $\mathcal{R}_k$ or it must draw from similarly intense sectors that may, with small probability, have arisen in less intense regions. In either case, $\hat{A}_{ri,s}$ must give high weight to sectors (r, i) with intensities $f_{ri}^{(\alpha)}$ that exceed the average intensity of the highest k regions: $f_{ri} > \hat{f}_{\mathcal{R}_k}^{(\alpha)}$.

The Markov inequality states that for any positive random variable X with mean value $\hat{x}$, the probability that an instance exceeds the mean by a proportion β is constrained [**36**]:

$$\Pr_{(X \geq \beta \hat{x})} (\leq) \, \frac{1}{\beta} \ .$$

Then for each region r, the weight (under $\mathbf{v}$) that a given sectoral intensity exceeds $\hat{f}_k^{(\alpha)}$ by a proportion β is:

$$\sum_{i : f_{ri} > \beta \hat{f}_{\mathcal{R}_k}^{(\alpha)}} v_{ri} \leq \frac{q_r \hat{f}_r^{(\alpha)}}{\beta \hat{f}_{\mathcal{R}_k}^{(\alpha)}} \ .$$

Thus, due to the Markov inequality, high-intensity sectors are suppressed in weight, in a manner determined by the relative proportions of intensities between regions. When condition (MRIO-2) holds—that is, the regional intensities are highly heterogeneous—this suppression is strengthened.

To counter this suppression in Eq. (7.18), $\hat{A}_{ri,s}$ must place extremely high relative weight on high-intensity sectors. In this case, however, very little weight remains to distribute among other sectors. Combinatorially, then, matrices $\hat{\mathbf{A}}$ violating regional majorization occupy a relatively small niche in the space of all configurations.

To summarize, as a consequence of the BSS theorem and fundamental facts of Leontief analysis, Eq. (7.16) must hold for any EE-MRIO table. When assumptions (MRIO-1) and (MRIO-2) hold, implying heterogeneity of regional intensities and small regional deficits, Eq. (7.16) further supports eco-majorization Eq. (7.14) by constraining the possible configurations which are not eco-majorized.

In the analogous thermodynamic setting, the experimenter (a Maxwell's "demon") may intentionally configure matrices that violate majorization after coarse-graining. In the setting of global trade, however, such a matrix must come about as the result of a strict bias among some regions to only consume high-intensity products. This is hardly realistic: Even at a national level, imports are a function of the variegated needs of multiple consumers and corporations. And, they necessarily draw their consumption from high-intensity and low-intensity industries. Regions, in short, do not operate as Maxwellian demons—at least, not in regards to their bulk imports.

Furthermore, as the null model is symmetrically generated without knowledge of local intensities, the null model is not be likely to draw models from the small niche required significantly violate majorization. Indeed, it is worth noting that our null model effectively acts as a Monte-Carlo model, calculating the total probability mass of the configuration space where majorization is violated.

7.4.3. Relaxing assumptions. To verify that conditions (MRIO-1) and (MRIO-2) are indeed responsible for the appearance of majorization and, consequently, the coupling between intensities and embodied flows, we made two modifications to the null model. First, we introduced null-impact intensities. Second, we varied parameters of both the MRIO null model and the impact intensity null model to determine their effects. We performed these in two separate trials.

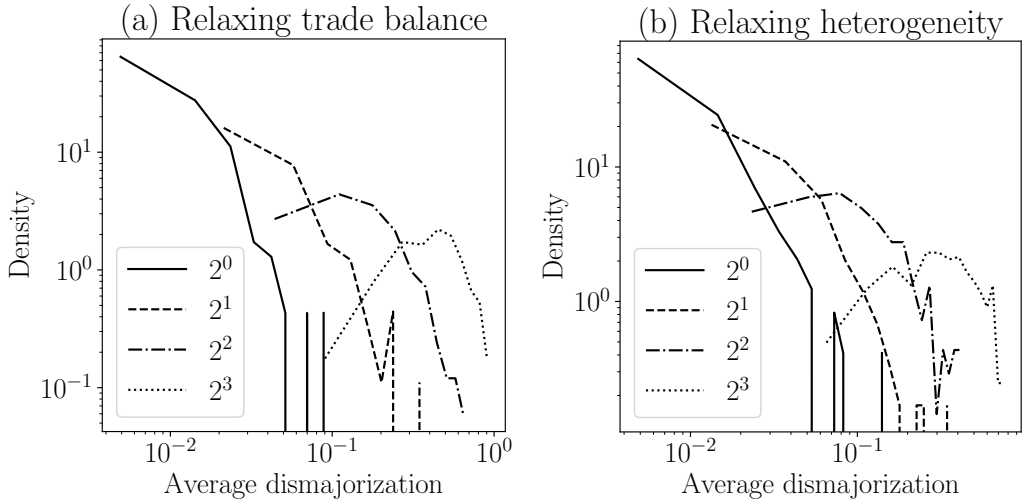

FIGURE 7.5. Emergence of eco-majorization: The results of varying the conditions that, when paired with Leontief analysis, result in eco-majorization. *(a)* Relaxing trade balance: Four suites of ensembles where the trade balance parameter was taken as $\zeta_X = \sigma^{-1}\zeta_X$, $\sigma = 2^0, 2^1, 2^2, 2^3$. As σ grows, so does the ensemble's overall trade deficit. A Gaussian kernel density was computed for the dismajorization and the null likelihood of net export over the null model samples. The left plot demonstrates that trade deficit's increase has the effect of increasing the average dismajorization. The right plot demonstrates that this also has the impact of shifting the net export density from being bimodal to unimodal. *(b)* Relaxing heterogeneity: Four suites of ensembles where the impact heterogeneity parameter was taken as $\zeta_U = \sigma\zeta_U$, $\sigma = 2^0, 2^1, 2^2, 2^3$. As σ grows, heterogeneity decreases. This has similar effects on the dismajorization density plot and the net export density plot.

For the first, we generated 1000 unobtainium intensity functions with parameter $\zeta_U = \bar{\zeta}_U$. For each scaling factor $\sigma = 1, 2, 4, 8$, we generated 250 MRIO tables from the null model with $\boldsymbol{\zeta_C} = \bar{\boldsymbol{\zeta}}_{\boldsymbol{C}}$ and $\zeta_X = \sigma^{-1}\bar{\zeta}_X$. This results in 4 ensembles of $250,000$ EE-MRIO tables each, with increasing trade imbalances from one ensemble to the next.

For the second, we flipped the structure of this approach, generating 1000 MRIO tables from the null model with $\boldsymbol{\zeta_C} = \bar{\boldsymbol{\zeta}}_{\boldsymbol{C}}$ and $\zeta_X = \bar{\zeta}_X$. Then, for each scaling factor $\sigma = 1, 2, 4, 8$, we generated 250 unobtainium intensities with parameter $\zeta_U = \sigma\bar{\zeta}_U$. This results in the same number of ensembles, but with decreasing heterogeneity of intensities.

To quantify the effect of the changes in parameter on majorization, we used dismajorization $\mathrm{DM}\left[(\hat{\mathbf{e}}^{(\mathrm{U})}, \hat{\mathbf{y}}); (\hat{\mathbf{a}}^{(\mathrm{U})}, \hat{\mathbf{y}})\right]$, as defined in Section 7.4.1. For the first suite of ensembles, for every sampled MRIO table we calculated the average dismajorization over all unobtainum distributions.

Then, a density plot of average dismajorizations over all samples with a given scaling factor σ was computed. The resulting density functions are shown in Fig. 7.5. Similarly, for the second suite, for every sampled unobtainium distribution we calculated the average dismajorization over all MRIO tables. Density plots were similarly taken for each scaling factor σ. The resulting density functions are also shown in Fig. 7.5. We find in each case that the likely dismajorization increases dramatically as the assumptions (MRIO-1) and (MRIO-2) are relaxed.

For both suites, we also sought to examine the impact of the scaling factors on trade-flow directionality. To this end, we calculated the null likelihoods of net exports Ξ_r for each region and each unobtainium distribution $\hat{\mathbf{e}}^{(U)}$. We made a density plot of observed values of Ξ_r for each value of $\sigma = 1, 2, 4, 8$. At the baseline parameter values ($\sigma = 1$), this density plot is bimodal, with one mode close to zero (countries that tend to be net importers) and one mode close to one (those that tend to be net exporters). However, as the baseline values are altered to relax assumptions, the density of Ξ_r becomes unimodal—the sharp distinction between exporters and importers vanishes. This is true as either assumption is relaxed.

These results complement those of the previous section. In Section 7.4.2 we showed that that the assumptions (MRIO-1) and (MRIO-2) are together sufficient for majorization to be predominant; Fig. 7.5 shows that they are each necessary. Additionally, relaxing either assumption (and the consequent irrelevancy of majorization) diminishes an initially strong dichotomy between exporting and importing nations, as the density of Ξ_r goes from bimodal to unimodal.

7.5. Discussion

A common adage in the social sciences involves a man looking for his keys under a streetlamp. When asked if that is where he lost them, he "No—but it's where the light is!"

Despite the increasing abundance of input-output data, and the convenient applicability of Leontief analysis, we should not be too hasty in drawing conclusions from its results. The two assumptions of Leontief analysis, (L-1) and (L-2), are critically important for understanding (i) the emergence of majorization in this setting and (ii) how exploring the validity of Leontiefian assumptions has great merit in assessing its results' usefulness. The "foothold" that majorization makes in our analysis

begins with Eqs. (7.6) and (7.9):

$$\mathbf{v}\hat{\mathbf{A}} = \hat{\mathbf{x}} \text{ and } \mathbf{e}^{(\alpha)}\hat{\mathbf{A}} = \hat{\mathbf{a}}^{(\alpha)} \ .$$

These equations relate factor incomes to the regional spending and production impacts to attributed impacts. Since both use the same attribution matrix $\hat{\mathbf{A}}$, the initial majorization relation Eq. (7.16) holds. Our mathematical analysis rests on this fact.

In information theory, stochastic matrices act as lossy channels that transmit information in a degraded condition. This leads distinct channel inputs to become more similar. While we sorted out the subtleties here, it is primarily for this reason that embodied emissions become more similar to global income: they are both channeled through the same lines of flow, determined by the single matrix $\hat{\mathbf{A}}$.

However, employing the same attribution matrix—whose primary determinant is *monetary* flows—to also drive embodied impact flows is a possibility only allowed by (L-1) and (L-2). The homogeneity of products and prices allows us to assume that monetary flows are entirely sufficient to reconstruct the commodity chains in which embodied flows are materialized. Without these assumptions, one would have a unique attribution matrix $\hat{\mathbf{A}}^{(\alpha)}$ for each impact α, distinct from the monetary attribution matrix. Majorization would no longer necessarily hold. And so, it would no longer drive the relationship between local intensities and embodied flows. In this way, our results suggest that the outcomes of Leontief analysis are highly dependent on the assumptions made. Rather than a neutral tool of analysis, Leontiefian methods embody significant ideological consequences.

Consider in this light the carbon leakage hypothesis. Essentially, firms from high-income countries foist the direct carbon costs of their production (which feeds local consumption) onto lower-income countries. This maintains consumption patterns while lowering compliance costs by superficially adhering to climate treaties to which they are signatories [**47**, **148**].

When this extends beyond carbon to other environmental impacts, the flow of embodied impacts from low- to high-income countries is *ecologically unequal exchange*—a major topic in modern geography and ecological economics [**70**, **84**, **159**, **179**]. One can even consider impacts such as labor-time. This leads to the more traditional hypothesis of unequal exchange [**52**] of exploited labor

from low- to high-income countries. Each of these hypotheses rests on fundamental assumptions about the social and economic power relations between nations and regions.

Multiregional input-output tables and Leontief analysis have been put to use evaluating these hypotheses previously, frequently in tandem with quantities such as the net exports $\boldsymbol{\xi}^{(\alpha)}$ and consumption-to-production ratio $\boldsymbol{\rho}^{(\alpha)}$ which we have analyzed here [**9**, **15**, **16**, **46**, **49**, **54**, **101**, **134**, **143**, **178**, **215**].

Our results directly bear on these previous studies. We showed that these quantities are, in fact, strongly driven by the assumptions of Leontief analysis and are largely independent of the relational data within the MRIO tables used. This calls into question the validity of Leontief analysis as a tool for empirically verifying hypotheses of unequal exchange, contributing a new perspective to previous critiques of this application [**49**]. Employing MRIO tables and Leontief analysis for empirical purposes must be done with caution; it would be unwise to place undue faith in its results merely because this is "where the light is."

To this end, we recommend an approach based on that taken here. Specifically, for this we make two contributions. First, use modern tools from information theory and statistical physics to better understand the consequences of methods like Leontief analysis, embedded as they are with numerous stochastic matrices and distributional relationships. Second, frequently consult with null models. This will aid in disentangling data structures, mathematical artifacts, and hypotheses, as otherwise these can be quite difficult to tease apart when using sophisticated modeling assumptions. This recommendation, of course, extends beyond MRIO tables and Leontief analysis. However, applying this approach to other studies of carbon accounting, environmental impacts, and ecologically unequal exchange will remain for future work.

Bibliography

[1] *Environmental Accounts of the Netherlands, 2012*, Statistics Netherlands, The Hague, 2012.

[2] *Industrial Development Report 2016: The Role of Technology and Innovation in Inclusive and Sustainable Industrial Development*, United Nations Industrial Development Organization, Vienna, 2015.

[3] *Renewable Energy Jobs: Future Growth in Australia*, Climate Council of Australia Ltd., 2016.

[4] C. Aghamohammadi, S. P. Loomis, J. R. Mahoney, and J. P. Crutchfield, *Extreme quantum memory advantage for rare-event sampling*, Phys. Rev. X, 8 (2018), p. 011025.

[5] C. Aghamohammadi, J. R. Mahoney, and J. P. Crutchfield, *The ambiguity of simplicity in quantum and classical simulation*, Phys. Lett. A, 381 (2017), pp. 1223–1227.

[6] ———, *Extreme quantum advantage when simulating classical systems with long-range interaction*, Scientific Reports, 7 (2017).

[7] C. Aghamohammdi, J. R. Mahoney, and J. P. Crutchfield, *Extreme quantum advantage when simulating strongly coupled classical systems*, Sci. Reports, 7 (2017), pp. 1–11.

[8] V. V. Albert, *Asymptotics of quantum channels: conserved quantities, an adiabatic limit, and matrix product states*, Quantum, 3 (2019), p. 151.

[9] A. Alsamawi, J. Murray, and M. Lenzen, *The Employment Footprints of Nations*, Journal of Industrial Ecology, 18 (2014), pp. 59–70. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jiec.12104.

[10] G. Ansmann and K. Lehnertz, *Constrained randomization of weighted networks*, Phys. Rev. E, 84 (2011), p. 026103. Publisher: American Physical Society.

[11] N. Aronszajn, *Theory of reproducing kernels*, Trans. Amer. Math. Soc., 68 (1950), pp. 337–404.

[12] F. Arute et al., *Quantum supremacy using a programmable superconducting processor*, Nature, 574 (2019), pp. 505–510.

[13] W. R. Ashby, *An Introduction to Cybernetics*, John Wiley and Sons, New York, second ed., 1960.

[14] B. Baumgartner and H. Narnhoffer, *The structures of state space concerning quantum dynamical semigroups*, Rev. Math. Phys., 24 (2012), p. 1250001.

[15] L. Bergmann, *Bound by Chains of Carbon: Ecological–Economic Geographies of Globalization*, Annals of the Association of American Geographers, 103 (2013), pp. 1348–1370.

[16] L. Bergmann and M. Holmberg, *Land in Motion*, Annals of the American Association of Geographers, 106 (2016), pp. 932–956.

[17] F. C. Binder, J. Thompson, and M. Gu, *A practical, unitary simulator for non-Markovian complex processes*, Phys. Rev. Lett., 120 (2017), p. 240502.

[18] D. Blackwell, *Comparison of Experiments*, Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, (1951), pp. 93–102. Publisher: University of California Press.

[19] ——, *Equivalent Comparisons of Experiments*, The Annals of Mathematical Statistics, 24 (1953), pp. 265–272. Publisher: Institute of Mathematical Statistics.

[20] B. Boots, A. Gretton, and G. Gordon, *Hilbert space embeddings of predictive state representations*, in Proceedings of the 29th International Conference on Uncertainty in Artificial Intelligence, 2013, pp. 92–101.

[21] I. Borg and P. J. F. Groenen, *Modern Multidimensional Scaling: Theory and Applications*, Springer Series in Statistics, Springer, New York, NY, second ed., 2005.

[22] A. B. Boyd and J. P. Crutchfield, *Maxwell demon dynamics: Deterministic chaos, the Szilard map, and the intelligence of thermodynamic systems*, Physical Review Letters, 116 (2016), p. 190601.

[23] A. B. Boyd, J. P. Crutchfield, and M. Gu, *Thermodynamic machine learning through maximum work production*, arXiv:2006.15416.

[24] A. B. Boyd, D. Mandal, and J. P. Crutchfield, *Identifying functional thermodynamics in autonomous Maxwellian ratchets*, New J. Physics, 18 (2016), p. 023049.

[25] ——, *Leveraging environmental correlations: The thermodynamics of requisite variety*, J. Stat. Phys., 167 (2016), pp. 1555–1585.

[26] ——, *Correlation-powered information engines and the thermodynamics of self-correction*, Phys. Rev. E, 95 (2017), p. 012152.

[27] ——, *Thermodynamics of modularity: Structural costs beyond the Landauer bound*, Physical Review X, 8 (2018), p. 031036.

[28] 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 (2017), p. 220602.

[29] F. G. S. L. Brandão, M. Horodecki, N. Ng, J. Oppenheim, and S. Wehner, *The second laws of quantum thermodynamics*, Proc. Natl. Acad. Sci. USA, 112 (11) (2015), pp. 3275–3279.

[30] F. G. S. L. Brandão, M. Horodecki, J. Oppenheim, J. M. Renes, and R. W. Spekkens, *Resource theory of quantum states out of thermal equilibrium*, Phys. Rev. Lett., 111 (2013), p. 250404.

[31] N. Brodu and J. P. Crutchfield, *Discovering causal structure with reproducing-kernel hilbert space ϵ-machines*, arXiv:2011.14821, (2020).

[32] R. A. Brualdi and D. Cvetković, *A Combinatorial Approach to Matrix Theory and its Applications*, Taylor & Francis Group, Boca Raton, FL, 2009.

[33] R. Carbone and Y. Pautrat, *Irreducible decompositions and stationary states of quantum channels*, Reports on Math. Phys., 77 (2016), p. 293.

[34] A. Christman and I. Steinwart, *Universal kernels on non-standard input spaces*, in Proceedings of the 23rd International Conference on Neural Information Processing Systems, vol. 1, ACM, 2010, pp. 406–414.

[35] B. Coecke, T. Fritz, and R. W. Spekkens, *A mathematical theory of resources*, Info. Comp., 250 (2016), p. 59.

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

[37] G. Crooks, *Quantum operation time reversal*, Phys. Rev. A, 77 (2008), p. 034101.

[38] J. P. Crutchfield, *The calculi of emergence: Computation, dynamics, and induction*, Physica D, 75 (1994), pp. 11–54.

[39] ———, *Between order and chaos*, Nature Physics, 8 (2012), pp. 17–24.

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

[41] J. P. Crutchfield and D. P. Feldman, *Statistical complexity of simple one-dimensional spin systems*, Phys. Rev. E, 55 (1997), pp. R1239–R1243.

[42] ———, *Regularities unseen, randomness observed: Levels of entropy convergence*, CHAOS, 13 (2003), pp. 25–54.

[43] J. P. Crutchfield and K. Young, *Inferring statistical complexity*, Phys. Rev. Let., 63 (1989), pp. 105–108.

[44] ———, *Thermodynamics of minimal reconstructed machines*. in preparation, 1989.

[45] O. C. O. Dahlsten, R. Renner, E. Rieper, and V. Vedral, *Inadequacy of von Neumann entropy for characterizing extractable work*, New J. Phys., 13 (2011).

[46] S. J. Davis and K. Caldeira, *Consumption-based accounting of CO2 emissions*, PNAS, 107 (2010), pp. 5687–5692. Publisher: National Academy of Sciences Section: Biological Sciences.

[47] S. J. Davis, G. P. Peters, and K. Caldeira, *The supply chain of CO2 emissions*, PNAS, 108 (2011), pp. 18554–18559. Publisher: National Academy of Sciences Section: Biological Sciences.

[48] L. del Rio, J. Åberg, R. Renner, O. Dahlsten, and V. Vedral, *The thermodynamic meaning of negative entropy*, Nature, 474 (2011), pp. 61–63.

[49] C. Dorninger and A. Hornborg, *Can EEMRIO analyses establish the occurrence of ecologically unequal exchange?*, Ecological Economics, 119 (2015), pp. 414–418.

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

[51] J. Emenheiser, A. Chapman, M. Posfai, J. P. Crutchfield, M. Mesbahi, and R. M. D'Souza, *Patterns of patterns of synchronization: Noise induced attractor switching in rings of coupled nonlinear oscillators*, Chaos, 26 (2016), p. 094816.

[52] A. Emmanuel, *Unequal Exchange: A Study of the Imperialism of Trade*, Monthly Review Press, 1972.

[53] P. S. Enomoto, *Dérivation par rapport à un système de voisinages dans l'espace de tore*, Proc. Japan Acad., 30 (1954), pp. 721–725.

[54] K.-H. Erb, F. Krausmann, W. Lucht, and H. Haberl, *Embodied HANPP: Mapping the spatial disconnect between global biomass production and consumption*, Ecological Economics, 69 (2009), pp. 328–334.

[55] P. Faist, F. Dupuis, J. Oppenheim, and R. Renner, *The minimal work cost of information processing*, Nature Comm., 6 (2015), p. 7669.

[56] D. R. Farine, *A guide to null models for animal social network analysis*, Methods in Ecology and Evolution, 8 (2017), pp. 1309–1320. _eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/ 10.1111/2041-210X.12772.

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

[58] K. Fukumizu, L. Song, and A. Gretton, *Kernel bayes' rule: Bayesian inference with positive definite kernels*, J. Mach. Learn. Res., 14 (2013), pp. 3753–3783.

[59] P. Gacs and J. Korner, *Common information is much less than mutual information*, Problems Contr. Inform. Th., 2 (1973), pp. 149–162.

[60] A. J. P. Garner, *Oracular information and the second law of thermodynamics*, (2019). arXiv:1912.03217 [quant-ph].

[61] A. J. P. Garner, Q. Liu, J. Thompson, V. Vedral, and M. Gu, *Provably unbounded memory advantage in stochastic simulation using quantum mechanics*, New J. Physics, 19 (2017), p. 103009.

[62] A. J. P. Garner, J. Thompson, V. Vedral, and M. Gu, *Thermodynamics of complexity and pattern manipulation*, Phys. Rev. E, 95 (2017), p. 042140.

[63] S. Geman and M. Johnson, *Probabilistic grammars and their applications*, in In International Encyclopedia of the Social & Behavioral Sciences. N.J. Smelser and P.B, 2000, pp. 12075–12082.

[64] G. Gour, M. P. MuHller, V. Narasimhachar, R. W. Spekkens, and N. Y. Halpern, *The resource theory of informational nonequilibrium in thermodynamics*, Physics Reports, 583 (2015), pp. 1–58.

[65] W. B. Group, *World Development Report 2020: Trading for Development in the Age of Global Value Chains*, International Bank for Reconstruction and Development / The World Bank, Washington, D. C., 2020.

[66] M. Gu, K. Wiesner, E. Rieper, and V. Vedral, *Quantum mechanics can reduce the complexity of classical models*, Nature Comm., 3 (2012).

[67] J. Guan, Y. Feng, and M. Ying, *The structure of decoherence-free subsystems*, (2018). arXiv:1802.04904 [quant-ph].

[68] P. Hayden, R. Josza, D. Petz, and A. Winter, *Structure of states which satisfy strong subadditivity of quantum entropy with equality*, Comm. Math. Phys., 246 (2004), p. 359.

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

[70] A. Hornborg, *The Unequal Exchange of Time and Space: Toward a Non-Normative Ecological Theory of Exploitation*, Journal of Ecological Anthropology, 7 (2003), pp. 4–10.

[71] M. Horodecki and J. Oppenheim, *Fundamental limitations for quantum and nanoscale thermodynamics*, Nature Comm., 4 (2013), p. 2059.

[72] M. Horodecki, J. Oppenheim, and C. Sparaciari, *Approximate majorization*, J. Phys. A: Math. Theor., 51 (2018), p. 305301.

[73] R. Horodecki, P. Horodecki, M. Horodecki, and K. Horodecki, *Quantum entanglement*, Rev. Mod. Phys., 81 (2009), p. 865.

[74] L. P. Hughston, R. Jozsa, and W. K. Wootters, *A complete classification of quantum ensembles having a given density matrix*, Phys. Lett. A, 183 (1993), pp. 12–18.

[75] H. Ito, S.-I. Amari, and K. Kobayashi, *Identifiability of hidden Markov information sources and their minimum degrees of freedom*, IEEE Info. Th., 38 (1992), p. 324.

[76] H. Jaeger, *Observable operator models for discrete stochastic time series*, Neural Computation, 12 (2000), pp. 1371–1398.

[77] D. Janzing, P. Wocjan, R. Zeier, R. Geiss, and T. Beth, *Thermodynamic Cost of Reliability and Low Temperatures: Tightening Landauer's Principle and the Second Law*, International Journal of Theoretical Physics, 39 (2000), pp. 2717–2753.

[78] A. Jenčová and D. Petz, *Structure of sufficient quantum coarse-grainings*, Comm. Math. Phys., 263 (2006), p. 259.

[79] B. Jessen, *The theory of integration in a space of an infinite number of dimensions*, Acta Math., 63 (1934), pp. 249–323.

[80] ———, *A remark on strong differentiation in a space of infinitely many dimensions*, Mat. Tidsskr. B., (1952), pp. 54–57.

[81] D. P. Johnson, *Representations of classifications of stochastic processes*, Trans. Amer. Math. Soc., 188 (1974).

[82] ———, *Representations of general stochastic processes*, Journal of Multivariate Analysis, 9 (1979), pp. 16–59.

[83] ———, *Hilbert space representations of general discrete time stochastic processes*, Stochastic Processes and their Applications, 19 (1985), pp. 183–187.

[84] A. K. Jorgenson and J. Rice, *The sociology of ecologically unequal exchange in comparative perspective*, in Routledge Handbook of World-Systems Analysis, Routledge, May 2012.

[85] M. Junge, R. Renner, D. Sutter, M. M. Wilde, and A. Winter, *Recoverability in quantum information theory*, Int. J. Theor. Phys., 21 (1982).

[86] A. Jurgens and J. P. Crutchfield, *Functional thermodynamics of maxwellian ratchets: Constructing and deconstructing patterns, randomizing and derandomizing behaviors*, Phys. Rev. Res., 2 (2020), p. 033334. arXiv.org:2003.00139.

[87] ———, *Shannon entropy rate of hidden Markov processes*, J. Statistical Physics, 183 (2020), pp. 1–18.

[88] ———, *Ambiguity rate of hidden Markov processes*, Phys. Rev. E, (2021).

[89] ———, *Divergent predictive states: The statistical complexity dimension of stationary, ergodic hidden Markov processes*, Chaos, (2021).

[90] O. Kallenberg, *Foundations of Modern Probability*, Springer, New York, 2 ed., 2001.

[91] M. G. Kendall, *A new measure of rank correlation*, Biometrika, 30 (1938), pp. 81–93.

[92] J. Kitzes, *An Introduction to Environmentally-Extended Input-Output Analysis*, Resources, 2 (2013), pp. 489–503. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

[93] S. Kojaku and N. Masuda, *Core-periphery structure requires something else in the network*, New J. Phys., 20 (2018), p. 043012. Publisher: IOP Publishing.

[94] F. Krausmann, K.-H. Erb, S. Gingrich, C. Lauk, and H. Haberl, *Global patterns of socioeconomic biomass flows in the year 2000: A comprehensive assessment of supply, consumption and constraints*, Ecological Economics, 65 (2008), pp. 471–487.

[95] G. R. Kumar, C. T. Li, and A. E. Gamal, *Exact common information*, (2014). arXiv:1402.0062 [cs.IT].

[96] P. Kůrka, *Topological and Symbolic Dynamics*, Société Mathématique de France, Paris, 2003.

[97] R. Landauer, *Irreversibility and heat generation in the computing process*, IBM J. Res. Develop., 5 (1961), pp. 183–191.

[98] M. S. Leifer and R. W. Spekkens, *A Bayesian approach to compatibility, improvement, and pooling of quantum states*, J. Phys. A: Math. Theor., 47 (2014), p. 275301.

[99] W. Leontief, *Essays in Economics: Theories, Theorizing, Facts, and Policies*, Transaction Publishers, 1966.

[100] ———, *The economy as a circular flow*, Structural Change and Economic Dynamics, 2 (1991), pp. 181–212.

[101] H. Liu, W. Liu, X. Fan, and Z. Liu, *Carbon emissions embodied in value added chains in China*, Journal of Cleaner Production, 103 (2015), pp. 362–370.

[102] Q. Liu, T. J. Elliot, F. C. Binder, C. D. Franco, and M. Gu, *Optimal stochastic modeling with unitary quantum dynamics*, Phys. Rev. A, 99 (2019), p. 062110.

[103] S. Lloyd, *Use of mutual information to decrease entropy: Implications for the second law of thermodynamics*, Phys. Rev. A, 39 (1989), p. 5378.

[104] W. Löhr and N. Ay, *Non-sufficient memories that are sufficient for prediction*, in Complex Sciences 2009, J. Zhou, ed., vol. 4 of Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Springer, New York, 2009, pp. 265–276.

[105] ———, *On the generative nature of prediction*, Adv. Complex Sys., 12 (2009), pp. 169–194.

[106] S. Loomis and J. P. Crutchfield, *Strong and weak optimizations in classical and quantum models of stochastic processes*, J. Stat. Phys, 176 (2019), pp. 1317–1342.

[107] ———, *Thermal efficiency of quantum memory compression*, Phys. Rev. Lett., 125 (2020), p. 020601.

[108] ———, *Thermodynamically-efficient local computation and the inefficiency of quantum memory compression*, Phys. Rev. Res., 2 (2020), p. 023039. arXiv:2001.02258.

[109] S. Loomis, J. R. Mahoney, C. Aghamohammdi, and J. P. Crutchfield, *Optimizing quantum models of classical channels: The reverse Holevo problem*, (2020).

[110] S. P. Loomis, M. Cooper, and J. P. Crutchfield, *Nonequilibrium thermodynamics in measuring carbon footprints: Disentangling structure and artifact in input-output accounting*, (2021). arXiv:2106.03948.

[111] S. P. Loomis and J. P. Crutchfield, *Topology, convergence, and reconstruction of predictive states*, (2021). Submitted to Physica D. arXiv:2109.09203.

[112] ———, *Predictive state geometry via cantor embeddings and wasserstein distance*, (2022). Submitted to KDD 2022.

[113] E. N. Lorenz, *Deterministic nonperiodic flow*, J. Atmos. Sci., 20 (1963), p. 130.

[114] ———, *The problem of deducing the climate from the governing equations*, Tellus, XVI (1964), p. 1.

[115] M. O. Lorenz, *Methods of Measuring the Concentration of Wealth*, Publications of the American Statistical Association, 9 (1905), pp. 209–219. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

[116] M. Lostaglio, *Thermodynamic laws for populations and quantum coherence: A self-contained introduction to the resource theory approach to thermodynamics*. arXiv:1807.11549.

[117] M. Lostaglio, D. Jennings, and T. Rudolph, *Description of quantum coherence in thermodynamic processes requires constraints beyond free energy*, Nature Comm., 6 (2015), p. 6383.

[118] A. Łuczak, *Quantum sufficiency in the operator algebra framework*, Int. J. Theor. Phys., 53 (2013), p. 3423.

[119] D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms*, Cambridge, Cambridge, United Kingdom, 2003.

[120] 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 (2016), p. 20495.

[121] ———, *Occam's quantum strop: Synchronizing and compressing classical cryptic processes via a quantum channel*, Scientific Reports, 6 (2016).

[122] D. Mandal and C. Jarzynski, *Work and information processing in a solvable model of Maxwell's demon*, PNAS, 109 (2012), p. 11641.

[123] A. W. Marshall, I. Olkin, and B. C. Arnold, *Inequalities: Theory of Majorization and Its Applications*, Springer, New York, NY, 3 ed., 2011.

[124] S. Marzen and J. P. Crutchfield, *Predictive rate-distortion for infinite-order markov processes*, J. Stat. Phys., 163 (2014), pp. 1312–1338.

[125] ———, *Informational and causal architecture of discrete-time renewal processes*, Entropy, 17 (2015), pp. 4891–4917.

[126] ———, *Statistical signatures of structural organization: The case of long memory in renewal processes*, Phys. Lett. A, 380 (2016), pp. 1517–1525.

[127] S. Marzen, M. R. DeWeese, and J. P. Crutchfield, *Time resolution dependence of information measures for spiking neurons: Scaling and universality*, Front. Comput. Neurosci., 9 (2015), p. 109.

[128] S. E. Marzen and J. P. Crutchfield, *Inference, prediction, and entropy-rate estimation of continuous-time, discrete-event processes*, arxiv:2005.03750.

[129] ———, *Nearly maximally predictive features and their dimensions*, Phys. Rev. E, 95 (2017), p. 051301(R). SFI Working Paper 17-02-007; arxiv.org:1702.08565 [cond-mat.stat-mech].

[130] ———, *Optimized bacteria are environmental prediction engines*, Physical Review E, 98 (2018), p. 012408.

[131] ———, *Probabilistic deterministic finite automata and recurrent networks, revisited*, (2019). arxiv.org:1910.07663 [cs.LG].

[132] U. M. Maurer, *Secret key agreement by public discussion from common information*, IEEE Trans. Info. Th., 39 (1993), p. 3.

[133] A. Monras, A. Beige, and K. Wiesner, *Hidden quantum Markov models and non-adaptive read-out of many-body states*, Appl. Math. Comput. Sci., 3 (2011), p. 93.

[134] D. D. Moran, M. Lenzen, K. Kanemoto, and A. Geschke, *Does ecologically unequal exchange occur?*, Ecological Economics, 89 (2013), pp. 177–186.

[135] D. D. Moran, M. C. Wackernagel, J. A. Kitzes, B. W. Heumann, D. Phan, and S. H. Goldfinger, *Trading spaces: Calculating embodied Ecological Footprints in international trade using a Product Land Use Matrix (PLUM)*, Ecological Economics, 68 (2009), pp. 1938–1951.

[136] M. Mosonyi, *Entropy, Information and Structure of Composite Quantum States*, PhD thesis, KU Leuven, 2005.

[137] M. Mosonyi and D. Petz, *Structure of sufficient quantum coarse-grainings*, Lett. Math. Phys., 68 (2004), p. 19.

[138] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf, *Kernel mean embedding of distributions: A review and beyond*, Found. Trends in Machine Learn., 10 (2017), pp. 1–141.

[139] D. Müllner, *Modern hierarchical, agglomerative clustering algorithms*, (2011). 1109.2378v1.

[140] B. Narayanan G., A. Aguiar, and R. McDougall, *Global Trade, Assistance, and Production: The GTAP 8 Data Base*, Center for Global Trade Analysis, Purdue University, 2012.

[141] M. Nielsen and I. Chuang, *Quantum Computation and Quantum Information*, Cambridge University Press, New York, 2010.

[142] M. A. Nielsen, *Conditions for a class of entanglement transformations*, Phys. Rev. Lett., 83 (1999).

[143] A. Oita, A. Malik, K. Kanemoto, A. Geschke, S. Nishijima, and M. Lenzen, *Substantial nitrogen pollution embedded in international trade*, Nature Geoscience, 9 (2016), pp. 111–115. Number: 2 Publisher: Nature Publishing Group.

[144] J. Oosterhaven, *Basic, Demand-Driven IO Quantity Models*, in Rethinking Input-Output Analysis: A Spatial Perspective, J. Oosterhaven, ed., SpringerBriefs in Regional Science, Springer International Publishing, Cham, 2019, pp. 5–18.

[145] V. M. Panaretos and Y. Zemel, *Statistical aspects of wasserstein distances*, Annu. Rev. Stat. Appl., 6 (2019), pp. 405–431.

[146] E. Pednault, J. A. Gunnels, G. Nannicini, L. Horesh, and R. Wisnieff, *Leveraging Secondary Storage to Simulate Deep 54-qubit Sycamore Circuits.*, (2019). arXiv:1910.09534 [quant-ph].

[147] G. P. Peters, *From production-based to consumption-based national emission inventories*, Ecological Economics, 65 (2008), pp. 13–23.

[148] G. P. Peters, J. C. Minx, C. L. Weber, and O. Edenhofer, *Growth in emission transfers via international trade from 1990 to 2008*, PNAS, 108 (2011), pp. 8903–8908. Publisher: National Academy of Sciences Section: Biological Sciences.

[149] D. Petz, *Sufficient subalgebras and the relative entropy of states of a von neumann algebra*, Comm. Math. Phys., 105 (1986), p. 123.

[150] ——, *Sufficiency of channels over von neumann algebras*, Quart. J. Math. Oxford, 2 (1988), p. 97.

[151] ——, *Quantum Information Theory and Quantum Statistics*, Springer, Berlin, Heidelberg, Springer-Verlag Berlin Heidelberg, 2008.

[152] C. Prell, K. Feng, L. Sun, M. Geores, and K. Hubacek, *The Economic Gains and Environmental Losses of US Consumption: A World-Systems and Input-Output Approach*, Social Forces, 93 (2014), pp. 405–428.

[153] J. Preskill, *Quantum computing and the entanglement frontier*, arxiv:1203.5813.

[154] B. J. Prettejohn, M. J. Berryman, and M. D. McDonnell, *Methods for Generating Complex Networks with Selected Structural Properties for Simulations: A Review and Tutorial for Neuroscientists*, Front. Comput. Neurosci., 5 (2011). Publisher: Frontiers.

[155] R. W. Rankin, J. Mann, L. Singh, E. M. Patterson, E. Krzyszczyk, and L. Bejder, *The role of weighted and topological network information to understand animal social networks: a null model approach*, Animal Behaviour, 113 (2016), pp. 215–228.

[156] M. M. Rao, *Conditional Measures and Applications*, CRC Press, Boca Raton, FL, second ed., 2005.

[157] J. M. Renes, *Relative submajorization and its use in quantum resource theories*, J. Math. Phys., 57 (2016), p. 122202.

[158] R. Renner and S. Wolf, *Smooth Rényi entropy and applications*, in 2004 IEEE Intl. Symp. Info. Th.: Proceedings, I. I. T. Society, ed., Piscataway, N.J., 2004, IEEE, p. 232.

[159] J. Rice, *Ecological Unequal Exchange: Consumption, Equity, and Unsustainable Structural Relationships within the Global Economy*, International Journal of Comparative Sociology, 48 (2007), pp. 43–72. Publisher: SAGE Publications Ltd.

[160] P. M. Riechers, J. R. Mahoney, C. Aghamohammadi, and J. P. Crutchfield, *Minimized state-complexity of quantum-encoded cryptic processes*, Phys. Rev. A, 93 (2016), p. 052317.

[161] J. B. Ruebeck, R. G. James, J. R. Mahoney, and J. P. Crutchfield, *Prediction and generation of binary Markov processes: Can a finite-state fox catch a Markov mouse?*, Chaos, 28 (2018).

[162] A. Rupe and J. P. Crutchfield, *Local causal states and discrete coherent structures*, Chaos, 28 (2018), pp. 1–22. arXiv.org:1801.00515 [cond-mat.stat-mech].

[163] ———, *Spacetime autoencoders using local causal states*, AAAI Fall Series 2020 Symposium on "Physics-guided AI for Accelerating Scientific Discovery", (2020).

[164] A. Rupe, J. P. Crutchfield, K. Kashinath, and Prabhat, *A physics-based approach to unsupervised discovery of coherent structures in spatiotemporal systems*, in Proceedings of the 7th International Workshop on Climate Informatics: CI 2017, V. Lyubchich, N. C. Oza, A. Rhines, and E. Szekely, eds. NCAR Technical Note NCAR/TN-536+PROC.

[165] A. Rupe, N. Kumar, V. Epifanov, K. Kashinath, O. Pavlyk, F. Schimbach, M. Patwary, S. Maidanov, V. Lee, Prabhat, and J. P. Crutchfield, *Disco: Physics-based unsupervised discovery of coherent structures in spatiotemporal systems*, arxiv:1909.XXXXX.

[166] M. B. Ruskai, *Inequalities for quantum entropy: A review with conditions for equality*, J. Math. Phys., 43 (2002), p. 4358.

[167] A. Salova, J. Emenheiser, J. P. Crutchfield, and R. M. D'Souza, *Koopman operator and its approximations for systems with symmetries*, Chaos, 29 (2019), p. 093128.

[168] B. Šavrič, T. Patterson, and B. Jenny, *The Equal Earth map projection*, International Journal of Geographical Information Science, 33 (2019), pp. 454–465. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/13658816.2018.1504949.

[169] A. Schaffartzik, M. Sachs, D. Wiedenhofer, and N. Eisenmenger, *Market Interaction and Efficient Cooperation*, Tech. Rep. 154, Institute of Social Ecology, IFF, Vienna, 2014.

[170] A. Schaffartzik, D. Wiedenhofer, and N. Eisenmenger, *Raw Material Equivalents: The Challenges of Accounting for Sustainability in a Globalized World*, Sustainability, 7 (2015), pp. 5345–5370. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

[171] M. Á. Serrano, M. Boguñá, and R. Pastor-Satorras, *Correlations in weighted networks*, Phys. Rev. E, 74 (2006), p. 055101. Publisher: American Physical Society.

[172] C. R. Shalizi, *Causal Architecture, Complexity and Self-Organization in Time Series and Cellular Automata*, PhD thesis, University of Wisconsin, Madison, Wisconsin, 2001.

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

[174] ———, *Information bottlenecks, causal states, and statistical relevance bases: How to represent relevant information in memoryless transduction*, Adv. Compl. Sys., 5 (2002), pp. 91–95.

[175] C. R. Shalizi and A. Kontorovich, *Almost none of the theory of stochastic processes*, February 2013. Lecture notes.

[176] C. R. Shalizi, K. L. Shalizi, and J. P. Crutchfield, *Pattern discovery in time series, Part I: Theory, algorithm, analysis, and convergence*, Journal of Machine Learning Research, (2002). Santa Fe Institute Working Paper 02-10-060; arXiv.org/abs/cs.LG/0210025.

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

[178] M. S. Simas, L. Golsteijn, M. A. J. Huijbregts, R. Wood, and E. G. Hertwich, *The "Bad Labor" Footprint: Quantifying the Social Impacts of Globalization*, Sustainability, 6 (2014), pp. 7514–7540. Number: 11 Publisher: Multidisciplinary Digital Publishing Institute.

[179] D. A. Smith, *Trade, unequal exchange, global commodity chains*, in Routledge Handbook of World-Systems Analysis, Routledge, May 2012.

[180] L. Song, B. Boots, S. Siddiqi, G. Gordon, and A. Smola, *Learning and discovery of predictive state representations in dynamical systems with reset*, in Proceedings of the Twenty-first International Conference on Machine Learning, ACM, 2004, p. 53.

[181] ———, *Hilbert space embeddings of hidden markov models*, in Proceedings of the 27th International Conference on Machine Learning, Omnipress, 2010, pp. 991–998.

[182] L. Song, J. Huang, A. Smola, and K. Fukumizu, *Hilbert space embeddings of conditional distributions with applications to dynamical systems*, in Proceedings of the 26th International Conference on Machine Learning, ACM, 2009, p. 961–968.

[183] B. Sriperembudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. G. Lanckriet, *Hilbert space embeddings and metrics on probability measures*, J. Machine Learn. Res., 11 (2010), pp. 1517–1561.

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

[185] C. C. Strelioff and J. P. Crutchfield, *Bayesian structural inference for hidden processes*, Physical Review E, 89 (2014), p. 042119.

[186] C. C. Strelioff, J. P. Crutchfield, and A. Hübler, *Inferring markov chains: Bayesian estimation, model comparison, entropy rate, and out-of-class modeling*, Phys. Rev. E, 76 (2007), p. 011106.

[187] W. Y. Suen, J. Thompson, A. J. P. Garner, V. Vedral, and M. Gu, *The classical-quantum divergence of complexity in modelling spin chains*, Quantum, 1 (2017), p. 25.

[188] S. Suh and R. Heijungs, *Power series expansion and structural analysis for life cycle assessment*, The International Journal of Life Cycle Assessment, 12 (2007), p. 381.

[189] R. Tan, D. R. Terno, J. Thompson, V. Vedral, and M. Gu, *Towards quantifying complexity with quantum mechanics*, Eur. J. Phys. Plus, 129 (2014), p. 191.

[190] O. Thas, *Comparing Distributions*, Springer Series in Statistics, Springer, New York, NY, 2010.

[191] J. Thompson, A. J. P. Garner, J. R. Mahoney, J. P. Crutchfield, V. Vedral, and M. Gu, *Causal asymmetry in a quantum world*, Phys. Rev. X, 8 (2018), p. 031013.

[192] J. Thompson, A. J. P. Garner, V. Vedral, and M. Gu, *Using quantum theory to simplify input-output processes*, npj Quantum Information, 3 (2017), p. 6.

[193] M. Thon and H. Jaeger, *Links between multiplicity automata, observable operator models and predictive state representations – a unified learning framework*, Journal of Machine Learning Research, 16 (2015), pp. 103–147.

[194] M. Tomamichel, *A Framework for Non-Asymptotic Quantum Information Theory*, PhD thesis, ETH Zurich, Zurich, 2012.

[195] N. Travers and J. P. Crutchfield, *Asymptotic synchronization for finite-state sources*, J. Stat. Phys., 145 (2011), pp. 1202–1223.

[196] ———, *Exact synchronization for finite-state sources*, J. Stat. Phys., 145 (2011), pp. 1181–1201.

[197] ———, *Equivalence of history and generator ϵ-machines*, (2014). arxiv.org:1111.4500 [math.PR].

[198] ———, *Infinite excess entropy processes with countable-state generators*, Entropy, 16 (2014), pp. 1396–1413.

[199] D. R. Upper, *Theory and Algorithms for Hidden Markov Models and Generalized Hidden Markov Models*, PhD thesis, University of California, Berkeley, 1997. Published by University Microfilms Intl, Ann Arbor, Michigan.

[200] D. P. Varn, G. S. Canright, and J. P. Crutchfield, *Discovering planar disorder in close-packed structures from X-Ray diffraction: Beyond the fault model*, Phys. Rev. B, 66 (2002), pp. 174110–3.

[201] ———, *ϵ-machine spectral reconstruction theory: A direct method for inferring planar disorder and structure from X-ray diffraction studies*, Acta. Cryst. Sec. A, 69 (2013), pp. 197–206.

[202] D. P. Varn and J. P. Crutchfield, *Chaotic crystallography: How the physics of information reveals structural order in materials*, Curr. Opin. Chem. Eng., 7 (2015), pp. 47–56.

[203] A. F. Veinott, *Least d-Majorized Network Flows with Inventory and Statistical Applications*, Management Science, 17 (1971), pp. 547–567. Publisher: INFORMS.

[204] A. Venegas-Li, A. Jurgens, and J. P. Crutchfield, *Measurement-induced randomness and structure in controlled qubit processes*, Physical Review E, 102 (2020), p. 040102(R).

[205] V. S. Vijayaraghavan, R. G. James, and J. P. Crutchfield, *Anatomy of a spin: The information-theoretic structure of classical spin systems*, Entropy, 19 (2017), p. 214. Santa Fe Institute Working Paper 15-10-042; arxiv.org:1510.08954 [cond-mat.stat-mech].

[206] S. Vinjanampathy and J. Anders, *Quantum thermodynamics*, Contemporary Physics, 57:4 (2016), pp. 545–579.

[207] A. Vitanov, F. Dupuis, M. Tomamichel, and R. Renner, *Chain rules for smooth min- and max-entropies*, IEEE Trans. Info. Th.,, 59 (2013), pp. 2603–2612.

[208] T. L. Walmsley and S. A. Ahmed, *A Global Bilateral Migration Data Base: Skilled Labor, Wages and Remittances*, (2007), p. 32.

[209] X. Wang and M. Wilde, *Resource theory of asymmetric distinguishability*, (2019). arXiv:1905.11629 [quant-ph].

[210] H. Whitehead, *Investigating structure and temporal scale in social organizations using identified individuals*, Behavioral Ecology, 6 (1995), pp. 199–208.

[211] T. Wiedmann, *Editorial: Carbon Footprint and Input–Output Analysis – an Introduction*, Economic Systems Research, 21 (2009), pp. 175–186. Publisher: Routledge _eprint: https://doi.org/10.1080/09535310903541256.

[212] T. Wiedmann, R. Wood, J. C. Minx, M. Lenzen, D. Guan, and R. Harris, *A Carbon Footprint Time Series of the Uk – Results from a Multi-Region Input–Output Model*, Economic Systems Research, 22 (2010), pp. 19–42. Publisher: Routledge _eprint: https://doi.org/10.1080/09535311003612591.

[213] M. M. Wolf, *Quantum channels & operations: Guided tour*, (2012). Lecture notes.

[214] A. D. Wyner, *The common information of two dependent random variables*, IEEE Trans. Info. Theory,, 21 (1975).

[215] Y. Yu, K. Feng, and K. Hubacek, *Tele-connecting local consumption to global land use*, Global Environmental Change, 23 (2013), pp. 1178–1186.

[216] H. Zhang and L. Zhao, *On the inclusion relation of reproducing kernel Hilbert spaces*, Analysis and Applications, 11 (2013), p. 1350015.

[217] M. Zhao and H. Jaeger, *Norm-observable operator models*, Neural Comp., 22 (2010), pp. 1927–59.