Journal of Statistical Physics (2019) 176:1317–1342
https://doi.org/10.1007/s10955-019-02344-x

Strong and Weak Optimizations in Classical and Quantum Models of Stochastic Processes

Samuel P. Loomis[1] · **James P. Crutchfield**[1]

Received: 29 December 2018 / Accepted: 18 June 2019 / Published online: 26 June 2019

Abstract
Among the predictive hidden Markov models that describe a given stochastic process, the ϵ-machine is strongly minimal in that it minimizes every Rényi-based memory measure. Quantum models can be smaller still. In contrast with the ϵ-machine's unique role in the classical setting, however, among the class of processes described by pure-state hidden quantum Markov models, there are those for which there does not exist any strongly minimal model. Quantum memory optimization then depends on which memory measure best matches a given problem's circumstance.

Keywords Stochastic process · Hidden Markov model · ϵ-machine · Causal states · Quantum information

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 [1,2]—and in a theoretical sense—seen in analyzing resource use in thermodynamic systems such as information engines [3]. 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.

The mathematical idealization of a system's behaviors is its *stochastic process*, and the study of the resource costs for predicting and simulating processes is known as *computational*

Communicated by Hal Tasaki.

✉ James P. Crutchfield
chaos@ucdavis.edu

Samuel P. Loomis
sloomis@ucdavis.edu

[1] Complexity Sciences Center and Physics Department, University of California at Davis, One Shields Avenue, Davis, CA 95616, USA

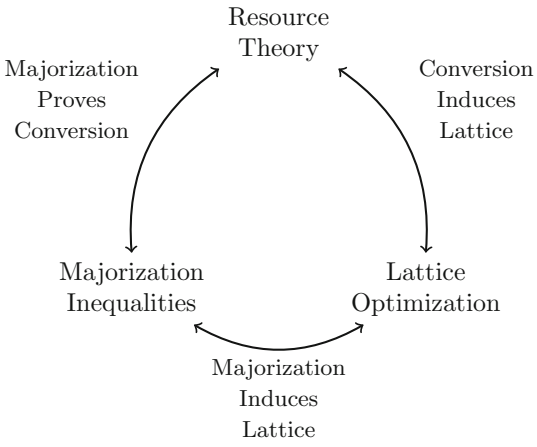


Fig. 1 Triumvirate of resource theory, majorization, and lattice theory

mechanics (CM) [4–7]. To date CM has largely focused on discrete-time, discrete-state stochastic processes. These are probability measures $\mathbb{P}(\ldots x_{-1}x_0x_1\ldots)$ over sequences of symbols that take values in a finite alphabet $\mathcal{X}$. 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.

When simulating classical processes, quantum implementations can be constructed that have smaller memory requirements than the ϵ-machine [8,9]. The study of such implementations is the task of *quantum computational mechanics* (QCM). 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 [10–13]. For quantum machines, the minimal memory rate C_q has been determined in cases such as the Ising model [11] and the Perturbed Coin Process [14], where the q-machine attains the minimum rate. Though a given q-machine's memory can be readily calculated [15], in many cases the absolutely minimal C_q is not known.

Another structural formalism, developed parallel to CM, provides a calculus of quantum informational resources. This field, termed *quantum resource theory* (QRT) recently emerged in quantum information theory as a toolkit for addressing resource consumption in the contexts of entanglement, thermodynamics, and numerous other quantum and even classical resources [16]. Its fundamental challenge is to determine when one system (a QRT *resource*) can be converted to another using a predetermined set of *free* or *allowed* operations.

QRT is closely allied with two other areas of mathematics, namely *majorization* and *lattice theory*. Figure 1 depicts their relationships.

On the one hand, majorization is a preorder relation $\succsim$ on positive vectors (typically probability distributions) computed by evaluating a set of inequalities [17]. 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 [18–20].

Lattice theory, on the other hand, concerns partially ordered sets and their suprema and infima, if they exist [21]. Functions that quantify the practical uses of a resource are monotonic with respect to the partial orders induced by convertibility and majorization. Optimization

of monotones of memory is then related to the problem of finding the extrema of the lattice. Majorization and resource convertibility are both relations that generate lattice-like structures on the set of systems.

The following brings the tools of CM and QRT together for the first time. Section 3 starts with a review of majorization theory for the unfamiliar and introduces *strong* and *weak* optimization which, as we show, have eminently practical implications for process predictors and simulators. Section 4 briefly reviews CM and demonstrates how strong/weak optimizations shed new light on the fundamental role of the ϵ-machine in the hierarchy of implementations for a given process. In particular, among classical predictive models the ϵ-machine is strongly minimal in that it simultaneously minimizes all measures of memory. Sections 5 and 6 then take these notions into the quantum setting, demonstrating the universally advantageous nature of quantum modeling when it comes to memory resources, but showing that no analog of (strong minimal) ϵ-machines exists in the hierarchy of quantum machines.

2 Processes, Probabilities, and Measures

The objects whose probabilities we study span both finite and infinite spaces, each of which entails its own notation.

Most of the objects of study in the following can be described with finite probability distributions. Finite here refers to random variables (e.g., X) that take values in a finite set (e.g., $\mathcal{X}$). Distribution refers to the probability of outcomes $x \in \mathcal{X}$ given by a vector $\mathbf{p} := (p_x)$ with components indexed by $\mathcal{X}$ that sum to unity: $\sum_{x \in \mathcal{X}} p_x = 1$.

Probability vectors may be transformed into one another by stochastic matrices. Here, we write such matrices as $\mathbf{T} := (T_{y|x})$ to represent a stochastic mapping from $\mathcal{X}$ to $\mathcal{Y}$. The matrix components are indexed by elements $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ and the stochasticity constraint is $\sum_{y \in Y} T_{y|x} = 1$ and $T_{y|x} \geq 0$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

The following development works with one object that is not finite—a stochastic process. Starting with a finite set $\mathcal{X}$ of symbols x (the "alphabet"), a length-ℓ word $w := x_1 \ldots x_\ell$ is a concatenation of ℓ symbols and the set of these is denoted $\mathcal{X}^\ell$. A bi-infinite word $\overleftrightarrow{x} := \ldots x_{-1} x_0 x_1 \ldots$ is a concatenation of symbols that extends infinitely in both directions and the set of these is denoted $\mathcal{X}^\infty$.

A stochastic process is a probability distribution over bi-infinite words. This implies a random variable $\overleftrightarrow{X}$ taking values in the set $\mathcal{X}^\infty$. However, this set is uncountably infinite, and the notation of measure theory is required to appropriately work with it [22]. In this case, probability values are taken over sets rather than distinct elements. We distinguish probabilities of sets from those of elements using the symbol $\mathbb{P}$. Often, we ask for the probability of seeing a given length-ℓ word w. This asks for the probability of the *cylinder set* $c_w := \left\{ \overleftrightarrow{x} : \overleftrightarrow{x} = \ldots x_t w x_{t+\ell+1} \ldots \text{ for some } t \in \mathbb{Z} \right\}$ of bi-infinite words containing word w. The measure then induces a finite distribution $\mathbf{p} := (p_w)$ over $\mathcal{X}^\ell$ describing a random variable W:

$$p_w := \mathbb{P}(c_w) .$$

When discussing the process as a whole, we refer to it by its random variable $\overleftrightarrow{X}$.

Following these conventions, lowercase boldface letters such as $\mathbf{p}$ and $\mathbf{q}$ denote probability vectors; uppercase boldface letters such as $\mathbf{T}$ denote linear transformations on the probability vectors; and uppercase cursive letters such as $\mathcal{X}$ denote finite sets (and almost always come

with an associated random variable X). Lowercase italic letters generally refer to elements of a finite set, though p and q are reserved for components of probability vectors.

Notation for quantum systems follows standard practice. Cursive letters do double-duty, as $\mathcal{H}$ is exclusively reserved for a Hilbert space, and quantum states are given by lowercase Greek letters. Linear operators are upper-case but not boldface.

3 Majorization and Optimization

First off, an overview of important relevant concepts from majorization and information theory is in order. Those familiar with these may skip to strong/weak optimization (Definition 3), though the intervening notational definitions might be useful.

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, as just discussed in Sect. 2.

Our introduction to majorization here follows Ref. [17]. The historical definition of majorization is also the most intuitive, starting with the concept of a *transfer operation*.

Definition 1 (*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:

$$\begin{aligned}(Tp)_i &:= p_i - \epsilon \\ (Tp)_j &:= p_j + \epsilon,\end{aligned}$$

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 1 (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:*

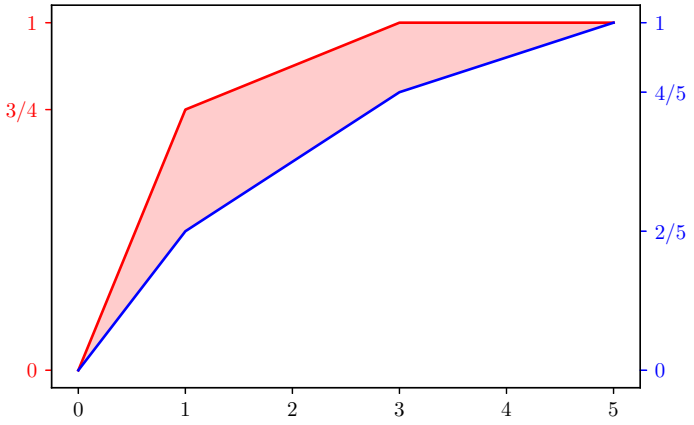


Fig. 2 Lorenz curves when $\mathbf{p}$ and $\mathbf{q}$ are comparable and the first majorizes the second: $\mathbf{p} \succsim \mathbf{q}$. Here, we chose $\mathbf{p} = (3/4, 1/8, 1/8, 0, 0)$ and $\mathbf{q} = (2/5, 1/5, 1/5, 1/10, 1/10)$. Tick marks indicate kinks in the Lorenz curve

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* $\mathbf{q} = \mathbf{Dp}$, *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. 2. 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 $\mathbf{p}$ nor $\mathbf{q}$ majorizes the other, they are *incomparable*.[1] (See Fig. 3.)

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 $q \sim p$ 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 $[\mathbf{p}]_{\sim}$.

[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 informational nonequilibrium [20], 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.)

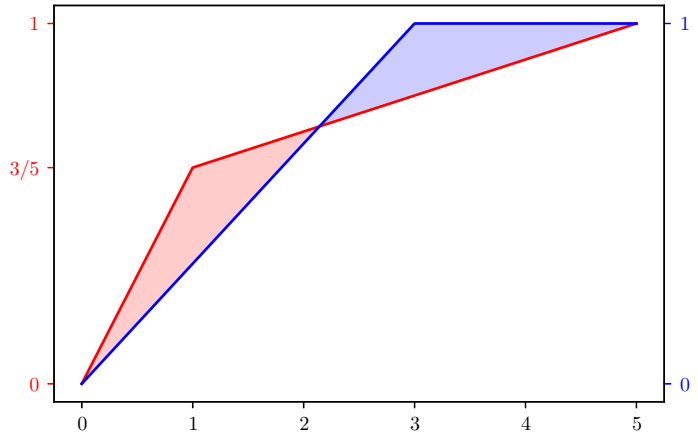


Fig. 3 Lorenz curves when $\mathbf{p}$ and $\mathbf{q}$ are incomparable. Here, we chose $\mathbf{p} = (3/5, 1/10, 1/10, 1/10, 1/10)$ and $\mathbf{q} = (1/3, 1/3, 1/3, 0, 0)$

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.

Definition 2 (*Schur-convex* (*-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:

$$H_\alpha(\mathbf{p}) := \frac{1}{1-\alpha} \log_2 \left(\sum_{i=1}^{n} p_i^\alpha \right).$$

In particular, the three limits:

$$\begin{aligned} H(\mathbf{p}) &:= \lim_{\alpha \to 1} H_\alpha(\mathbf{p}) = -\sum_{i=1}^{n} p_i \log_2 p_i, \\ H_0(\mathbf{p}) &:= \lim_{\alpha \to 0} H_\alpha(\mathbf{p}) = \log_2 |\{1 \leq i \leq n : p_i > 0\}|, \text{ and} \\ H_\infty(\mathbf{p}) &:= \lim_{\alpha \to \infty} H_\alpha(\mathbf{p}) = -\log_2 \max_{1 \leq i \leq n} p_i \end{aligned}$$

—*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 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) [23,24]. As such, they play a significant role in communication and memory storage.

We note that the Rényi entropies for $0 < \alpha < \infty$ are *strictly* Schur-concave.

The example of two incomparable distributions $\mathbf{p}$ and $\mathbf{q}$ can be analyzed in terms of the Rényi entropies if we plot $H_\alpha(\mathbf{p})$ and $H_\alpha(\mathbf{q})$ as a function of α, as in Fig. 4.

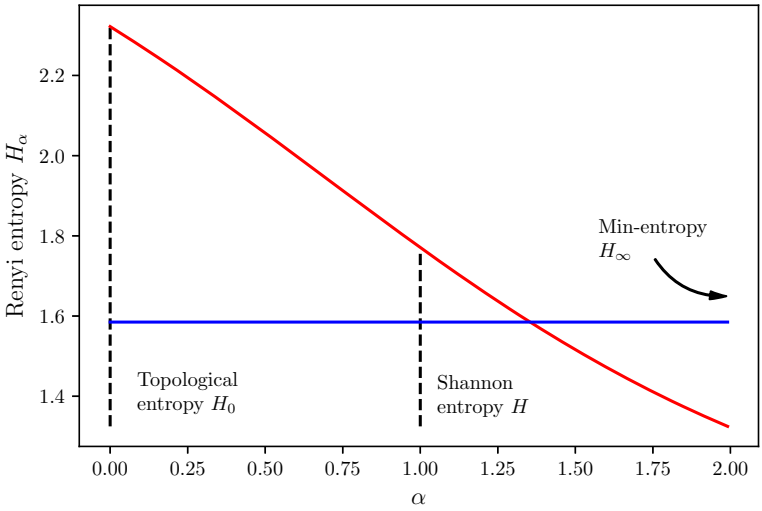


Fig. 4 Rényi entropies of the two incomparable distributions **p** and **q** from Fig. 3

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*.

Definition 3 (*Strong maximum* (*minimum*)) 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*). These constructions may not be unique as $\precsim$ 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.

In lattice-theoretic terms, the strong maximum is essentially the lattice-theoretic notion of a *meet* and the strong minimum is a *join* [21].

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 $\rho|_X \precsim \rho|_Y$ 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. [25], 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 Strong Minimality of the ϵ-Machine

We spoke in the introduction of simulating and predicting processes; this task is accomplished by *hidden Markov models* (HMMs) [26]. Here, we study a particular class of HMMs which we term *finite predictive models* (FPM).[2]

Definition 4 (*Finite predictive model*) A *finite predictive model* is a triplet $\mathfrak{M} := (\mathcal{R}, \mathcal{X}, \{\mathbf{T}^{(x)} : x \in \mathcal{X}\})$ containing:

1. A finite set of *hidden states* $\mathcal{R}$,
2. A finite *alphabet* $\mathcal{X}$,
3. Nonnegative transition matrices $\mathbf{T}^{(x)} := \left(T^{(x)}_{r'|r}\right)$, labeled by symbols $x \in \mathcal{X}$ with components indexed by $r, r' \in \mathcal{R}$,

satisfying the properties:

1. *Irreducibility*: $\mathbf{T} := \sum_{x\in\mathcal{X}} \mathbf{T}^{(x)}$ is stochastic and irreducible.
2. *Unifilarity*: $T^{(x)}_{r'|r} = P_{x|r}\delta_{r', f(r,x)}$ for some stochastic matrix $P_{x|r}$ and deterministic function f.

A finite predictive model is thought of as a dynamical object; the model transitions between states $r, r' \in \mathcal{R}$ at each timestep while emitting a symbol $x \in \mathcal{X}$ with probabilities determined by the transition matrices $\mathbf{T}^{(x)} := \left(T^{(x)}_{r'|r}\right)$. Unifilarity ensures that, given the model state $r \in \mathcal{R}$ and symbol $x \in \mathcal{X}$, the next state $r' \in \mathcal{R}$ is unique.

What makes this model *predictive*? Here, it is the *unifilarity* property that grants predictivity: In a unifilar model, the hidden state provides the most information possible about the future behavior as compared to other nonunifilar models [6].

Given a FPM $\mathfrak{M}$, the state transition matrix $\mathbf{T}$ has a single right-eigenvector $\boldsymbol{\pi}$ of eigenvalue 1, by the Perron-Frobenius theorem, satisfying $\mathbf{T}\boldsymbol{\pi} = \boldsymbol{\pi}$. We call this state distribution the *stationary distribution*. The finite set $\mathcal{R}$ and distribution $\boldsymbol{\pi}$ form a random variable R describing the asymptotic distribution over hidden states.

A stationary[3] stochastic process $\overleftrightarrow{X}$ is entirely determined by specifying its probability vectors $\mathbf{p}^{(\ell)} := (p^{(\ell)}_w)$ over words $w = x_1 \ldots x_\ell$ of length ℓ, for all $\ell \in \mathbb{Z}^+$. Using the stationary distribution $\boldsymbol{\pi}$ we define the process $\overleftrightarrow{X}_{\mathfrak{M}}$ generated by $\mathfrak{M}$ using the word distributions $p^{(\ell)}_w := \mathbf{1}^\top \mathbf{T}^{(x_\ell)} \ldots \mathbf{T}^{(x_1)}\boldsymbol{\pi}$, where $w := x_1 \ldots x_\ell$ and $\mathbf{1}$ is the vector with all 1's for its components. If we let $\boldsymbol{\delta}_r$ be a distribution on $\mathcal{R}$ that assigns the state $r \in \mathcal{R}$ probability 1, then the vector $\mathbf{p}^{(\ell)}_r := (p^{(\ell)}_{w|r})$ with components $p^{(\ell)}_{w|r} := \mathbf{1}^\top \mathbf{T}^{(x_\ell)} \ldots \mathbf{T}^{(x_1)}\boldsymbol{\delta}_r$ is the probability of seeing word w after starting in state r.[4]

Given a model with stationary distribution $\boldsymbol{\pi}$, we define the model's Rényi memory as $H_\alpha(\mathfrak{M}) := H_\alpha(\boldsymbol{\pi})$. This includes the topological memory $H_0(\mathfrak{M})$, the statistical memory

[2] The following uses the words *machine* and *model* interchangeably. *Machine* emphasizes the simulative nature of the implementation; *model* emphasizes the predictive nature.

[3] A process is stationary if it is time-invariant.

[4] This portrait of a process, in terms of stochastic matrices, is introduced in Refs. [27–29] and has important parallels to the matrix product state formalism. Reference [30] explores these parallels in the quantum setting.

$H(\mathfrak{M}) = H_1(\mathfrak{M})$, and the min-memory $H_\infty(\mathfrak{M})$. Given a process $\overleftrightarrow{X}$, we define the Rényi *complexity* as [4]

$$C_\mu^{(\alpha)}\left(\overleftrightarrow{X}\right) := \min_{\mathfrak{M}:\, \overleftrightarrow{X} = \overleftrightarrow{X}_{\mathfrak{M}}} H_\alpha(\mathfrak{M})$$

These include the topological complexity $C_\mu^{(0)}$, the statistical complexity $C_\mu := C_\mu^{(1)}$, and the min-complexity $C_\mu^{(\infty)}$.

The question, then, of strong or weak optimization with regards to memory in prediction and simulation is really the question of whether, for a given process $\overleftrightarrow{X}$, a particular model achieves all $C_\mu^{(\alpha)}$ (strong optimization), or whether a separate model is required for different values of α (weak optimization). As each α may have practical meaning in a particular scenario, this question is highly relevant for problems of optimal modeling.

Among the class of FPMs, a particularly distinguished member is the ϵ-*machine*, first considered in Ref. [4]. We use the definition given in Ref. [31].

Definition 5 (*Generator* ϵ-*machine*) A generator ϵ-machine is a finite predictive model $\mathfrak{M} := (\mathcal{S}, \mathcal{X}, \{\mathbf{T}^{(x)} : x \in \mathcal{X}\})$ such that $\mathbf{p}_s^{(\ell)} = \mathbf{p}_{s'}^{(\ell)}$ for all $\ell \in \mathbb{Z}^+$ implies $s = s'$ for $s, s' \in \mathcal{S}$.

In other words, a generator ϵ-machine must be irreducible, unifilar, and its states must be *probabilistically distinct*, so that no pair of distinct states predict the same future.

An important result of computational mechanics is that the generator ϵ-machine is unique with respect to the process it generates [31].

Theorem 2 (Model-Process Uniqueness Theorem) *Given an* ϵ-*machine* $\mathfrak{M}$*, there is no other* ϵ-*machine that generates* $\overleftrightarrow{X}_{\mathfrak{M}}$.

This is a consequence of the equivalence of the generator definition with another, called the *history* ϵ-machine, which is itself provably unique (up to isomorphism) [6]. A further important result is that the ϵ-machine minimizes both the statistical complexity C_μ and the topological complexity $C_\mu^{(0)}$ [6].

To fix intuitions, and to begin introducing majorization concepts into CM, we will now consider several example processes and their models.

First, consider the Biased Coin Process, a memoryless process 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 5 displays three models for it. Model (a) is the process' ϵ-machine, and models (b) and (c) are each 2-state alternative finite predictive models. Notice that in both models (b) and (c), the two states generate equivalent futures.

Continuing, Fig. 6 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. 6a the process' ϵ-machine. In Fig. 6b, we see an alternative finite predictive model. 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.

Majorization and Lorenz curves, in particular, allow us to compare the various models for each of these processes—see Fig. 7. We notice that the ϵ-machine state distribution always majorizes the state distribution of the alternative machines.

The key to formalizing this observation is the following corollary of Model-Process Uniqueness Theorem:

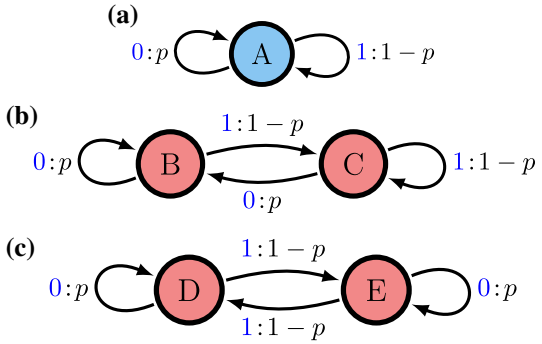


Fig. 5 The diagrammatic form of a FSM is read as follows. The colored circles represent hidden states from the finite set $\mathcal{R}$. The edges are labeled by a blue number, the symbol x, and a probability p. The edges with symbol x represent the transition matrix $\mathbf{T}^{(x)} := \left(T^{(x)}_{r'|r}\right)$, 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

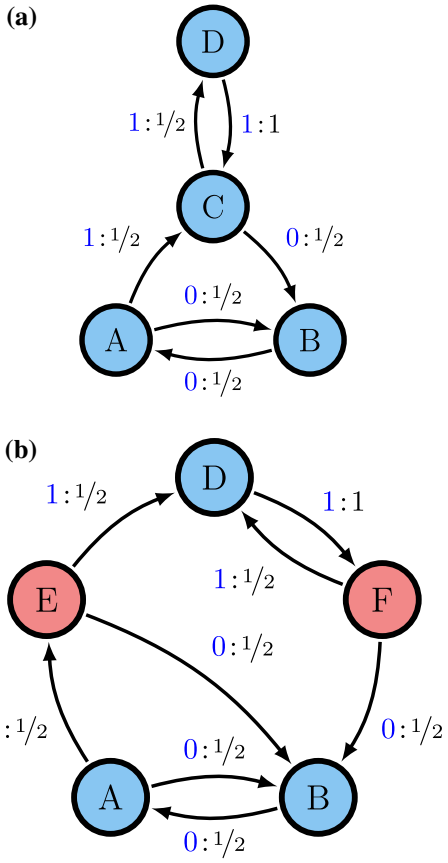


Fig. 6 **a** ϵ-Machine for even–odd process. **b** Refinement of the even–odd process ϵ-machine, where the ϵ-machine's state C has been split into states E and F

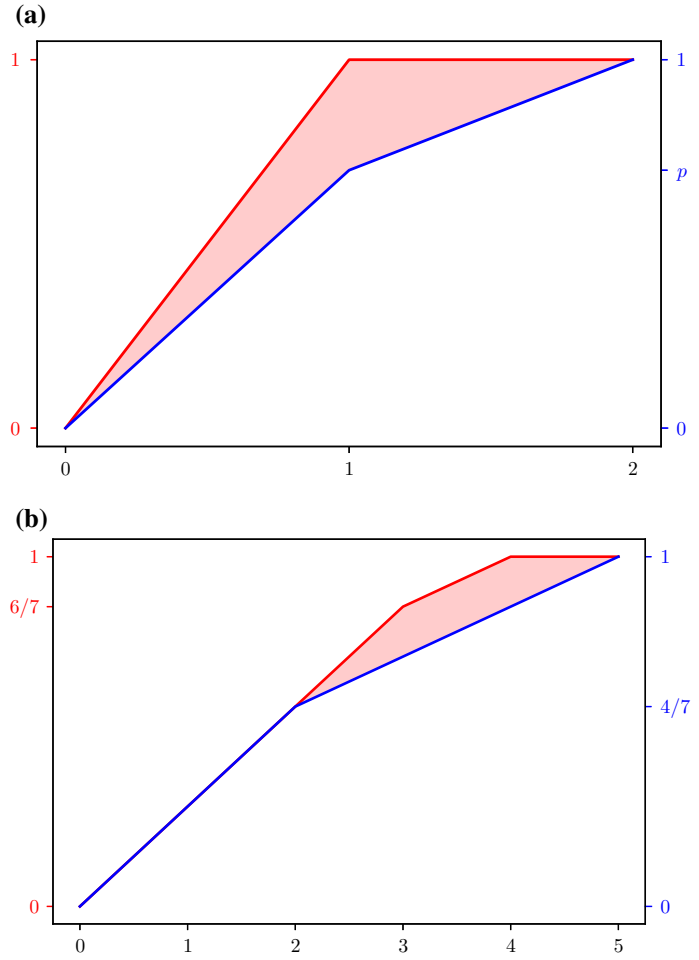


Fig. 7 **a** Lorenz curves for Fig. 5a's ϵ-machine and Fig. 5b's alternative predictor of the Biased Coin Process. **b** Same comparison for the even–odd process ϵ-machine Fig. 6a and alternative predictor Fig. 6b

Corollary 1 (State Merging) *Let* $\mathfrak{M} := (\mathcal{R}, \mathcal{X}, \{\mathbf{T}^{(x)} : x \in \mathcal{X}\})$ *be a finite predictive model that is not an* ϵ*-machine. Then the machine created by merging its probabilistically equivalent states is the* ϵ*-machine of the process* $\overleftrightarrow{X}_{\mathfrak{M}}$ *generated by* $\mathfrak{M}$.

Proof Let $\sim$ be the equivalence relation where $r \sim r'$ if $\mathbf{p}_r^{(\ell)} = \mathbf{p}_{r'}^{(\ell)}$ for all $\ell \in \mathbb{Z}^+$. Let $\mathcal{S}$ consist of the set of equivalence classes $[r]_\sim$ generated by this relation. For a given class $s \in \mathcal{S}$, consider the transition probabilities associated with each $r \in s$. For each $x \in \mathcal{X}$ such that $P_{x|r} > 0$, there is a outcome state $r_x := f(x, r)$. Comparing with another state in the same class $r' \in s$, we have the outcome state $r'_x := f(x, r')$.

For the future predictions of both states r and r' to be equivalent, they must also be equivalent after seeing the symbol x. That is, $p_{w|r}^{(\ell)} = p_{w|r'}^{(\ell)}$ for all w and ℓ also implies $p_{xw|r}^{(\ell+1)} = p_{xw|r'}^{(\ell+1)}$ for all x, w and ℓ. But $p_{xw|r}^{(\ell+1)} = p_{w|r_x}^{(\ell)}$, and so we have $r_x \sim r'_x$ for all $x \in \mathcal{X}$.

The upshot of these considerations is that we can define a consistent and unifilar transition dynamic $\{\widetilde{\mathbf{T}}^{(x)} : x \in \mathcal{X}\}$ on $\mathcal{S}$ given by the matrices $\widetilde{T}^{(x)}_{s'|s} := \widetilde{T}^{(x)}_{r'|r}$ for any $r \in s$ and $r' \in s'$. It inherits unifilarity from the original model $\mathfrak{M}$ as well as irreducibility. It has probabilistically distinct states since we already merged all of the probabilistically equivalent states. Therefore, the resulting machine $\mathfrak{M}_{\mathcal{S}} := (\mathcal{S}, \mathcal{X}, \{\widetilde{\mathbf{T}}^{(x)} : x \in \mathcal{X}\})$ is the ϵ-machine of the process $\overleftrightarrow{X}_{\mathfrak{M}}$ generated by $\mathfrak{M}$; its uniqueness follows from Model-Process Uniqueness Theorem. □

The state-merging procedure here is an adaptation of the Hopcroft algorithm for minimization of deterministic finite (nonstochastic) automate, which is itself an implementation of the Nerode equivalence relation [32]. The Hopcroft algorithm has been applied previously to analyze synchronization in ϵ-machines [33].

Using Corollary 1, we prove this section's main result.

Theorem 3 (Strong Minimality of ϵ-Machine) *Let $\mathfrak{M}_{\mathcal{S}} := (\mathcal{S}, \mathcal{X}, \{\widetilde{\mathbf{T}}^{(x)} : x \in \mathcal{X}\})$ be the ϵ-machine of process $\overleftrightarrow{X}$ and $\mathfrak{M}_{\mathcal{R}} := (\mathcal{R}, \mathcal{S}, \{\mathbf{T}^{(x)} : x \in \mathcal{X}\})$ be any other finite generating machine. Let the stationary distributions be $\boldsymbol{\pi}_{\mathcal{S}} := \left(\pi_{s|\mathcal{S}}\right)$ and $\boldsymbol{\pi}_{\mathcal{R}} := \left(\pi_{r|\mathcal{R}}\right)$, respectively. Then $\boldsymbol{\pi}_{\mathcal{S}} \gtrsim \boldsymbol{\pi}_{\mathcal{R}}$, with equivalence $\sim$ only when $\mathfrak{M}_{\mathcal{S}}$ and $\mathfrak{M}_{\mathcal{R}}$ are isomorphic.*

Proof By Corollary 1, the states of the ϵ-machine $\mathfrak{M}_{\mathcal{S}}$ are formed by merging equivalence classes $s = [r]$ on the finite predictive model $\mathfrak{M}_{\mathcal{R}}$. Since the machines are otherwise equivalent, the stationary probability $\pi_{s|\mathcal{S}}$ is simply the sum of the stationary probabilities for each $r \subseteq s$, given by $\pi_{r|\mathcal{R}}$. That is:

$$\pi_{s|\mathcal{S}} = \sum_{r \in s} \pi_{r|\mathcal{R}}.$$

One can then construct $\boldsymbol{\pi}_{\mathcal{R}}$ from $\boldsymbol{\pi}_{\mathcal{S}}$ by a series of transfer operations in which probability is shifted out of the state s into new states r. Since the two states are related by a series of transfer operations, $\boldsymbol{\pi}_{\mathcal{S}} \gtrsim \boldsymbol{\pi}_{\mathcal{R}}$. □

It immediately follows from this that not only does the ϵ-machine minimize the statistical complexity C_μ and the topological complexity $C_\mu^{(0)}$, but it also minimizes every other Rényi complexity $C_\mu^{(\alpha)}$ as well. That this was so for C_μ and $C_\mu^{(0)}$ has previously been proven; the extension to all α is a new result here.

The uniqueness of the ϵ-machine is extremely important in formulating this result. This property of ϵ-machines follows from the understanding of predictive models as partitions of the past and of the ϵ-machines as corresponding to the coarsest graining of these predictive partitions [6]. Other paradigms for modeling will not necessarily have this underlying structure and so may not have strongly minimal solutions. Indeed, in the following we will see that this result does not generalize to quantum machines.

5 Strong Quantum Advantage

A pure-state quantum machine can be generalized from the classical case by replacing the classical states s with quantum state vectors $|\eta_s\rangle$ and the symbol-labeled transition matrices $\mathbf{T}^{(x)}$ with symbol-labeled Kraus operators $K^{(x)}$.[5] The generalization is called a *pure-state quantum model* (PSQM).

[5] The definition here using Kraus operators can be equivalently formulated in terms of a unitary quantum system [34]. While that alternate definition is more obviously physical, our formulation makes the classical parallels explicit.

Definition 6 (*Pure-state quantum model*) A pure-state quantum model is a quintuplet $\mathfrak{M} := (\mathcal{H}, \mathcal{X}, \mathcal{S}, \{|\eta_s\rangle : s \in \mathcal{S}\}, \{K^{(x)} : x \in \mathcal{X}\})$ consisting of:

1. A finite-dimensional Hilbert space $\mathcal{H}$,
2. A finite alphabet $\mathcal{X}$,
3. Pure states $|\eta_s\rangle$ indexed by elements $s \in \mathcal{S}$ in a finite set $\mathcal{S}$,
4. Nonnegative Kraus operators $K^{(x)}$ indexed by symbols $x \in \mathcal{X}$,

satisfying the properties:

1. *Completeness*: The Kraus operators satisfy $\sum_x K^{(x)\dagger} K^{(x)} = I$.
2. *Unifilarity*: $K^{(x)} |\eta_s\rangle \propto \left|\eta_{f(s,x)}\right\rangle$ for some deterministic function $f(s, x)$.

This is a particular kind of hidden *quantum* Markov model (HQMM) [35] in which we assume the dynamics can be described by the evolution of pure states. This is practically analogous to the assumption of unifilarity in the classical predictive setting.

It is not necessarily the case that the states $\{|\eta_s\rangle\}$ form an orthonormal basis; rather, nonorthonormality is the intended advantage [8,9]. Overlap between the states allows for a smaller von Neumann entropy for the process' stationary state distribution. We formalize this notion shortly.

It is assumed that the Kraus operators have a unique stationary density matrix ρ_π analogous to a classical model's stationary state $\boldsymbol{\pi}$. One way to compute it is to note the matrix $P_{x|s} = \langle\eta_s| K^{(x)\dagger} K^{(x)} |\eta_s\rangle$ and the function $s \mapsto f(s, x)$ together determine a finite predictive model as defined above. The model's stationary state $\boldsymbol{\pi} := (\pi_s)$ is related to the stationary density matrix of the quantum model via:

$$\rho_\pi = \sum_s \pi_s |\eta_s\rangle \langle\eta_s|.$$

The process generated by a pure-state quantum model has the length-ℓ word distribution, for words $w = x_1 \ldots x_\ell$:

$$p_w^{(\ell)} := \mathrm{Tr}\left[K^{(x_\ell)} \ldots K^{(x_1)} \rho_\pi K^{(x_1)\dagger} \ldots K^{(x_\ell)\dagger}\right].$$

The eigenvalues $\{\lambda_i\}$ of the stationary state ρ_π form a distribution $\boldsymbol{\lambda} = (\lambda_i)$. The Rényi entropies of these distributions form the *von Neumann-Rényi* entropies of the states:

$$S_\alpha(\rho_\pi) = H_\alpha(\boldsymbol{\lambda}).$$

We noted previously that for a given density matrix, these entropies are strongly minimal over the entropies of all projective, maximal measurements on the state. Given a model $\mathfrak{M}$ with stationary state ρ_π, we may simply write $S_\alpha(\mathfrak{M}) := S_\alpha(\rho_\pi)$ as the Rényi memory of the model. Important limits, as before, are the *topological memory* $S_0(\mathfrak{M})$, the *statistical memory* $S(\mathfrak{M}) = S_1(\mathfrak{M})$, and the *min-memory* $S_\infty(\mathfrak{M})$, which represent physical limitations on memory storage for the generator.

To properly compare PSQMs and FPMs, we define the *classical equivalent model* of a PSQM.

Definition 7 (*Classical equivalent model*) Let $\mathfrak{M} := (\mathcal{H}, \mathcal{X}, \mathcal{S}, \{|\eta_s\rangle : s \in \mathcal{S}\}, \{K^{(x)} : x \in \mathcal{X}\})$ be a pure-state quantum model, with probabilities $P_{x|s} := \langle\eta_s| K^{(x)\dagger} K^{(x)} |\eta_s\rangle$ and deterministic function $f(s, x)$ such that $K^{(x)} |\eta_s\rangle \propto \left|\eta_{f(s,x)}\right\rangle$. Its classical equivalent $\mathfrak{M}_{\mathrm{cl}} = (\mathcal{S}, \mathcal{X}, \{\mathbf{T}^{(x)} : x \in \mathcal{X}\})$ is the classical finite predictive model with state set $\mathcal{S}$, alphabet $\mathcal{X}$, and symbol-based transition matrices $\mathbf{T}^{(x)} = \left(T^{(x)}_{s'|s}\right)$ given by $T^{(x)}_{s'|s} = P_{x|r}\delta_{r', f(r,x)}$.

Each PSQM $\mathfrak{M}$ generates a process $\overleftrightarrow{X}_{\mathfrak{M}}$, which is the same process that is generated by the classical equivalent model: $\overleftrightarrow{X}_{\mathfrak{M}_{\text{cl}}} = \overleftrightarrow{X}_{\mathfrak{M}}$.

We now prove that a classical equivalent model for a PSQM is always strongly improved in memory by said PSQM.

Theorem 4 (Strong quantum advantage) *Let $\mathfrak{M} := (\mathcal{H}, \mathcal{X}, \mathcal{S}, \{|\eta_s\rangle : s \in \mathcal{S}\}, \{K^{(x)} : x \in \mathcal{X}\})$ be a pure-state quantum model with stationary state ρ_π, and let $\mathfrak{M}_{\text{cl}}$ be the classical equivalent model with stationary state $\boldsymbol{\pi} := (\pi_s)$ (with $s = 1, \ldots, n$). Let $D := \dim \mathcal{H}$ and $N := |\mathcal{S}|$. (We have $N \geq D$: if not, then we can take a smaller Hilbert space that spans the states.) Let $\boldsymbol{\lambda} = (\lambda_i)$ be an N-dimensional vector where the first D components are the eigenvalues of ρ_π and the remaining elements are 0. Then $\boldsymbol{\lambda} \succsim \boldsymbol{\pi}$.*

Proof We know that:

$$\begin{aligned}\rho_\pi &= \sum_{s \in \mathcal{S}} \pi_s \, |\eta_s\rangle \, \langle \eta_s| \\ &= \sum_{s \in \mathcal{S}} |\phi_s\rangle \, \langle \phi_s| \, ,\end{aligned}$$

where $|\phi_s\rangle := \sqrt{\pi_s} \, |\eta_s\rangle$. However, we can also write ρ_π in the eigenbasis:

$$\begin{aligned}\rho_\pi &= \sum_{i=1}^{D} \lambda_i \, |i\rangle \, \langle i| \\ &= \sum_{i=1}^{D} |\psi_i\rangle \, \langle \psi_i| \, ,\end{aligned}$$

where $|\psi_i\rangle := \sqrt{\lambda_i} \, |i\rangle$. Then the two sets of vectors can be related via:

$$|\phi_s\rangle = \sum_{i=1}^{D} U_{si} \, |\psi_i\rangle \, ,$$

where U_{si} is a $N \times D$ matrix comprised of d rows of orthonormal N-dimensional vectors [36]. Now, we have:

$$\begin{aligned}\pi_s &= \langle \phi_s | \phi_s \rangle \\ &= \sum_{i=1}^{D} |U_{si}|^2 \lambda_i \, .\end{aligned}$$

Note that U_{si} is not square, but since we have taken $\lambda_i = 0$ for $i > D$, we can simply extend U_{si} into a square unitary matrix by filling out the bottom $n - D$ rows with more orthonormal vectors. This leaves the equation unchanged. We can then write:

$$\pi_s = \sum_{i=1}^{n} |U_{si}|^2 \lambda_i \, .$$

Then by Theorem 1, $\boldsymbol{\lambda} \succsim \boldsymbol{\pi}$. □

It helps to recall now that majorization is a preorder, which means we could have $\boldsymbol{\pi} \sim \boldsymbol{\lambda}$, in which case there would be no advantage *per se*. This happens when $|U_{si}|^2$ is a permutation matrix. However, one quickly sees that this is true if and only if $\{|\eta_s\rangle\}$ are orthogonal. Thus, any nonorthogonality in the quantum states automatically induces advantage.

Corollary 2 *$S_\alpha(\mathfrak{M}) \leq H_\alpha(\mathfrak{M}_{\text{cl}})$ for all $\alpha \geq 0$, with equality for $0 < \alpha < \infty$ if and only if the states $\{|\eta_s\rangle\}$ of $\mathfrak{M}$ are orthogonal.*

As in the classical case, it immediately follows from this that not only is the classical equivalent model improved upon by its corresponding PSQM in terms of $S_0(\mathfrak{M})$ and $S(\mathfrak{M})$ (as was previously known in certain cases), but it is improved in *all* Rényi memories $S_\alpha(\mathfrak{M})$.

Many alternative pure-state quantum models may describe the same process. The "first mark", so to speak, for quantum models is the q-machine, which directly embeds the dynamics of the ϵ-machine into a quantum system while already leveraging the memory advantage due to quantum state overlap. The notion of the q-machine originates in [8], and its definition was further refined in Refs. [9,15]. We use an equivalent definition first introduced in Ref. [34]; however, there an equivalent unitary formalism is used instead of Kraus operators.

Definition 8 (*q-Machine*) Given an ϵ-machine $\mathfrak{M} := \left(\mathcal{S}, \mathcal{X}, \{\mathbf{T}^{(x)} : x \in \mathcal{X}\}\right)$, where $T^{(x)}_{ss'} := P_{x|s}\delta_{s', f(s,x)}$ for some deterministic function $f(s, x)$, construct the corresponding q-machine in the following way:

1. The states $|\eta_s\rangle$ are built to satisfy the recursive relation:
$$\langle\eta_s|\eta_{s'}\rangle = \sum_{x\in\mathcal{X}} \sqrt{P_{x|s}P_{x|s'}}\left\langle\eta_{f(s,x)}|\eta_{f(s',x)}\right\rangle.$$
2. $\mathcal{H}$ is the space spanned by the states $|\eta_s\rangle$.
3. The Kraus operators $K^{(x)}$ are determined by the relations:
$$K^{(x)}|\eta_s\rangle = \sqrt{P_{x|s}}\left|\eta_{f(s,x)}\right\rangle.$$

Then Corollary 2 can be applied here. The q-machine is matched in memory by the ϵ-machine when and only when the states $|\eta_s\rangle$ are orthogonal, $\langle\eta_s|\eta_{s'}\rangle = \delta_{ss'}$. The recursive relation becomes:

$$\delta_{ss'} = \sum_{x\in\mathcal{X}} \sqrt{P_{x|s}P_{x|s'}}\delta_{f(s,x)f(s',x)}.$$

This holds if and only if $\delta_{f(s,x)f(s',x)} = \delta_{ss'}$ for all x satisfying $P_{x|s}, P_{x|s'} > 0$. This constrains the structure of the ϵ-machine: two distinct states s and s' cannot map to the same state on the same symbol. In other words, given a state and an incoming symbol, the previous state must be determined. Such a structure is called *co-unifilar* [37]. Examples of co-unifilar machines are shown in Fig. 5a and c.

To be clear, then, the q-machine offers strict advantage over any ϵ-machine which is not co-unifilar and matches the ϵ-machine when it is co-unifilar. That the q-machine offers statistical memory advantage with respect to the ϵ-machine was previously shown in Ref. [9] and with respect to topological memory in Ref. [14]. Theorem 4 implies those results as well as advantage with respect to all Rényi measures of memory.

One can check that the q-machine satisfies the completeness relations and has the correct probability dynamics for the process generated by the ϵ-machine.

6 Weak Quantum Minimality

An open problem is to determine the minimal quantum pure-state representation of a given classical process. This problem is solved in some specific instances such as the Ising model

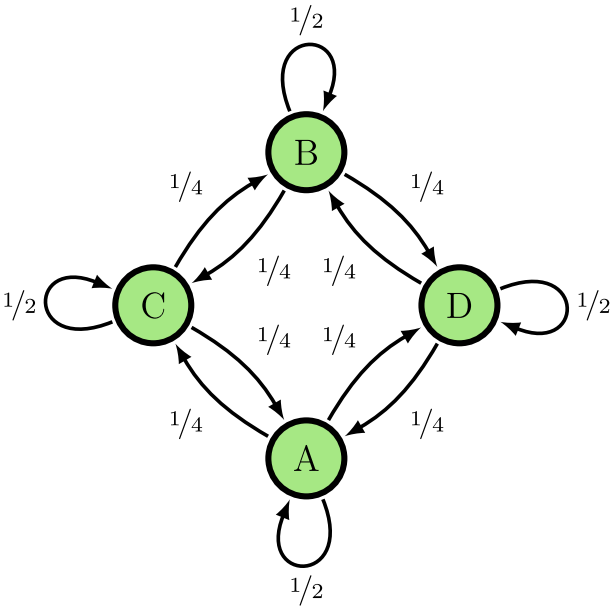


Fig. 8 The 4-state MBW process as a Markov chain (which is the ϵ-machine)

[11] and the Perturbed Coin Process [14]. In these cases it is known to be the q-machine. We denote the smallest value of the Rényi entropy of the stationary state as:

$$C_q^{(\alpha)}\left(\overleftrightarrow{X}\right) := \min_{\mathfrak{M}:\, \overleftrightarrow{X}_{\mathfrak{M}} = \overleftrightarrow{X}} S_\alpha\left(\mathfrak{M}\right),$$

called the *quantum Rényi complexities*, including the limits, the *quantum topological complexity* $C_q^{(0)}$, the *quantum min-complexity* $C_q^{(\infty)}$, and the *quantum statistical complexity* $C_q := C_q^{(1)}$.

If a strongly minimal quantum pure-state model exists, these complexities are all attained by the same pure-state model. Our primary result here is that there are processes for which this does not occur.

We start by examining two examples. The first, the MBW process introduced in Ref. [35], demonstrates a machine whose q-machine is not minimal in the von Neumann complexity. Consider the process generated by the *4-state MBW* machine shown in Fig. 8.

This process' HMM is simply a Markov chain, and its representation in Fig. 8 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:

$$\begin{aligned}
|\eta_A\rangle &:= \frac{1}{\sqrt{2}}|A\rangle + \frac{1}{2}\left(|C\rangle + |D\rangle\right), \\
|\eta_B\rangle &:= \frac{1}{\sqrt{2}}|B\rangle + \frac{1}{2}\left(|C\rangle + |D\rangle\right), \\
|\eta_C\rangle &:= \frac{1}{\sqrt{2}}|C\rangle + \frac{1}{2}\left(|A\rangle + |B\rangle\right), \text{ and}
\end{aligned}$$

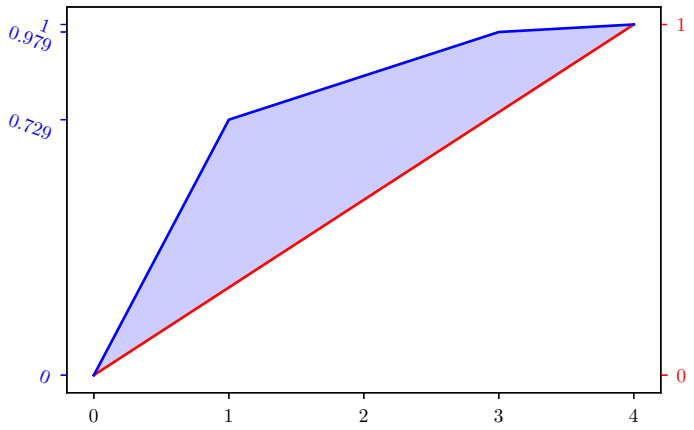


Fig. 9 Lorenz curves for the 4-state MBW ϵ-machine $\mathfrak{M}_4$ and the associated q-machine $\mathfrak{Q}_4$

$$|\eta_D\rangle := \frac{1}{\sqrt{2}}|D\rangle + \frac{1}{2}\left(|A\rangle + |B\rangle\right).$$

Since it is a Markov chain, we can write the Kraus operators as $K_x := |\eta_x\rangle\langle\epsilon_x|$, where $\langle\epsilon_x|\eta_{x'}\rangle \propto \sqrt{P_{x'|x}}$. This is a special case of the construction used in Ref. [13]. For q-machines of Markov chains, then, the dual basis is just $\langle\epsilon_x| = \langle x|$. We denote the q-machine model of the 4-state MBW process as $\mathfrak{Q}_4$.

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. 9.

It turns out that there is a smaller quantum model embedded in two dimensions, with states:

$$\begin{aligned}
\left|\eta'_A\right\rangle &:= |0\rangle\,, \\
\left|\eta'_B\right\rangle &:= |1\rangle\,, \\
\left|\eta'_C\right\rangle &:= \frac{1}{\sqrt{2}}\left(|0\rangle + |1\rangle\right), \text{ and} \\
\left|\eta'_D\right\rangle &:= \frac{1}{\sqrt{2}}\left(|0\rangle - |1\rangle\right).
\end{aligned}$$

In this case, $\left\langle\epsilon'_x\right| = \frac{1}{\sqrt{2}}\left\langle\eta'_x\right|$ derives the q-machine. This gives the proper transition probabilities for the 4-state MBW model. We denote this dimensionally-smaller model $\mathfrak{D}_4$. Figure 10 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: $S(\mathfrak{D}_4) = 1.0$ and $S(\mathfrak{Q}_4) \approx 1.2$ bit. (On the other hand, the q-machine has a smaller min-memory, with $S_\infty(\mathfrak{D}_4) = 1.0$ and $S_\infty(\mathfrak{Q}_4) \approx 0.46$.)

Now consider something in the opposite direction. Consider the 3-state MBW model, denoted $\mathfrak{M}_3$ and displayed in Fig. 11. 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.

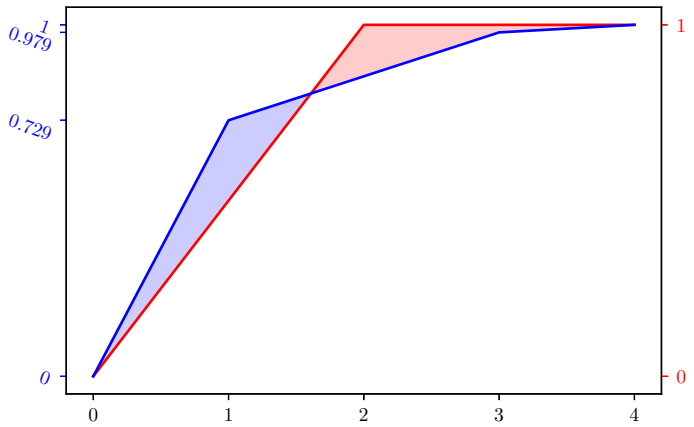
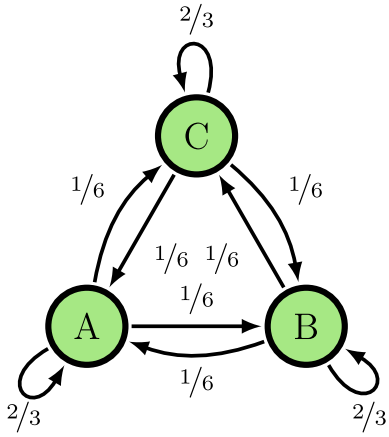


Fig. 10 Lorenz curves for the 4-state MBW q-machine $\mathfrak{Q}_4$ and a dimensionally-smaller model $\mathfrak{D}_4$

Fig. 11 The 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:

$$|\eta_A\rangle := \sqrt{\frac{2}{3}}\,|A\rangle + \frac{1}{\sqrt{6}}\,(|B\rangle + |C\rangle)\,,$$

$$|\eta_B\rangle := \sqrt{\frac{2}{3}}\,|B\rangle + \frac{1}{\sqrt{6}}\,(|A\rangle + |C\rangle)\,, \text{ and}$$

$$|\eta_C\rangle := \sqrt{\frac{2}{3}}\,|C\rangle + \frac{1}{\sqrt{6}}\,(|A\rangle + |B\rangle)\,,$$

and Kraus operators defined similarly to before. We can examine the majorization between the q-machine and the Markov model by plotting the Lorenz curves of $\boldsymbol{\lambda}$, the eigenvalues of ρ_π, and the stationary state of the Markov chain, shown in Fig. 12.

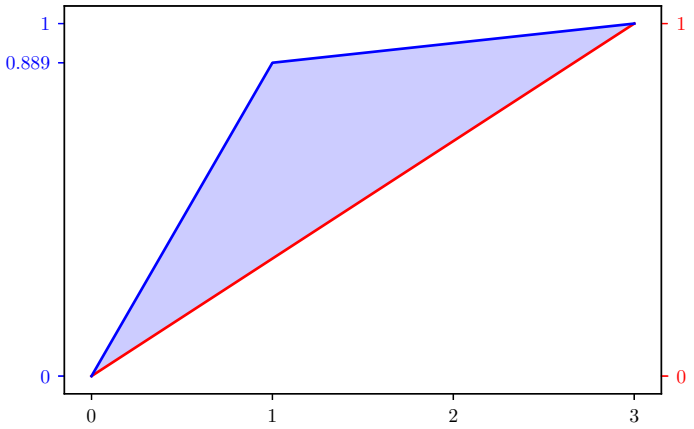


Fig. 12 Lorenz curves for the 3-state MBW ϵ-machine $\mathfrak{M}_3$ and the associated q-machine $\mathbf{Q}_3$

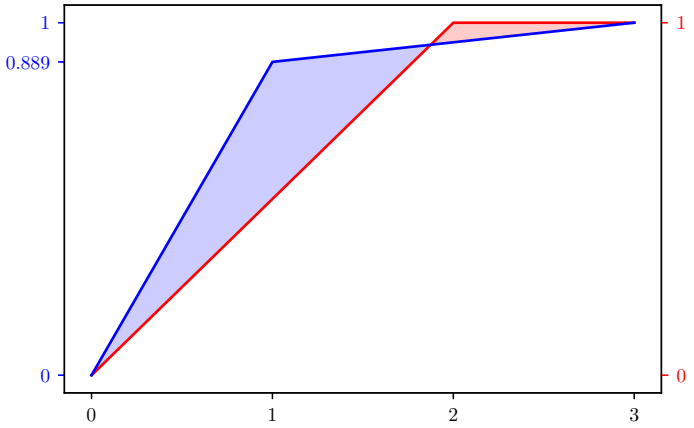


Fig. 13 Lorenz curves for the 3-state MBW q-machine, $\mathfrak{Q}_3$ and a dimensionally-smaller model $\mathfrak{D}_3$

The lower-dimensional model $\mathfrak{D}_3$ is given by the states:

$$
\begin{aligned}
|\eta_A\rangle &:= |0\rangle\,, \\
|\eta_B\rangle &:= \frac{1}{2}\,|0\rangle + \frac{\sqrt{3}}{2}\,|1\rangle\,, \text{ and} \\
|\eta_C\rangle &:= \frac{1}{2}\,|0\rangle - \frac{\sqrt{3}}{2}\,|1\rangle\,,
\end{aligned}
$$

with $\langle \epsilon'_x | = \sqrt{\frac{2}{3}}\,\langle \eta'_x |$. This gives the proper transition probabilities for the 3-state MBW model. Figure 13 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: $S(\mathfrak{D}_3) = 1.0$ and $S(\mathfrak{Q}_3) \approx 0.61$ bit.

After seeing the ϵ-machine's strong minimality with respect to other classical models and its strong maximality with respect to quantum models, it is certainly tempting to conjecture

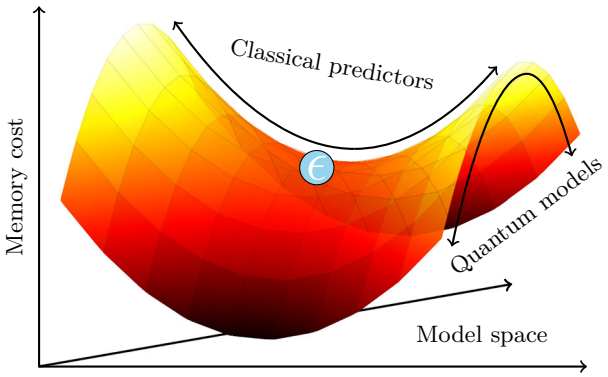


Fig. 14 Proposed majorization saddle structure of model-space: the ϵ-machine (labeled ϵ) is located at a saddle-point with respect to majorization, where classical deviations (state-splitting) move up the lattice and quantum deviations (utilizing state overlap) move down the lattice

that a strongly minimal quantum model exists. However, the examples we just explored cast serious doubt. None of the examples covered above are strong minima.

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 $\mathfrak{D}_3$. Since this model is not strongly minimal, no strongly minimal solution can exist.

Appendix 1 proves exactly this—thus, demonstrating a counterexample to the strong minimality of quantum models.

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.*

7 Concluding Remarks

Majorization provides a means to compare a process' alternative models in both the classical and quantum regimes. When it holds, majorization implies the simultaneous minimization of a large host of functions. As a result we showed that:

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. 14 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 [38]. 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 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, $\overleftrightarrow{X}$ and $\overleftrightarrow{Y}$, such that $C_\mu\left(\overleftrightarrow{X}\right) > C_\mu\left(\overleftrightarrow{Y}\right)$ but $C_q\left(\overleftrightarrow{X}\right) < C_q\left(\overleftrightarrow{Y}\right)$ [39]. 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.

The methods and results here should also be extended to analyze classical generative models which, in many ways, bear resemblances in their functionality to the quantum models [40–42]. 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.

We also only explored finite-state, discrete-time processes. Processes with infinite memory [43] and continuous generation [44,45] are also common in nature. Applying our results to understand these requires further mathematical development.

We close by noting that we have committed a sleight of hand here, using the tools of resource theory to study CM and, particularly, memory in stochastic processes. This is still far from formulating a *resource theory of memory*. It is also far from applying the memoryful logic of CM to extend resource theory, which often studies memoryless collections of resources, in which there is no temporal structure. Both of these directions will be developed elsewhere and, in doing so, will likely shed significant light on the above questions.

Acknowledgements The authors thank Fabio Anza, John Mahoney, Cina Aghamohammadi, and Ryan James for helpful discussions, as well as Felix Binder for clarifying suggestions. As a faculty member, JPC thanks the Santa Fe Institute and the Telluride Science Research Center for their hospitality during visits. This material is based upon work supported by, or in part by, John Templeton Foundation Grant 52095, Foundational Questions Institute Grant FQXi-RFP-1609, the U.S. Army Research Laboratory and the U. S. Army Research Office under Contract W911NF-13-1-0390 and Grant W911NF-18-1-0028, and via Intel Corporation support of CSC as an Intel Parallel Computing Center.

Author contributions SPL and JPC conceived of the project, developed the theory, and wrote the manuscript. SPL performed the calculations. JPC supervised the project.

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no competing financial interests.

Appendix: Weak Minimality of $\mathbf{D_3}$

Here, we prove that $\mathfrak{D}_3$ is the unique 2D representation of the 3-state MBW process. We show this by considering the entire class of 2D models and applying the completeness constraint.

We note that a pure-state quantum model of the 3-state MBW process must have three states $|\eta_A\rangle$, $|\eta_B\rangle$, and $|\eta_C\rangle$, along with three dual states $\langle\epsilon_A|$, $\langle\epsilon_B|$, and $\langle\epsilon_C|$ such that:

$$\begin{aligned}
\langle\epsilon_A|\eta_A\rangle &= e^{i\phi_{AA}}\sqrt{\frac{2}{3}},\\
\langle\epsilon_A|\eta_B\rangle &= e^{i\phi_{AB}}\frac{1}{\sqrt{6}}, \text{ and}\\
\langle\epsilon_A|\eta_C\rangle &= e^{i\phi_{AC}}\frac{1}{\sqrt{6}},\\
\langle\epsilon_B|\eta_A\rangle &= e^{i\phi_{BA}}\frac{1}{\sqrt{6}},\\
\langle\epsilon_B|\eta_B\rangle &= e^{i\phi_{BB}}\sqrt{\frac{2}{3}}, \text{ and}\\
\langle\epsilon_B|\eta_C\rangle &= e^{i\phi_{BC}}\frac{1}{\sqrt{6}},
\end{aligned}$$

and

$$\begin{aligned}
\langle\epsilon_C|\eta_A\rangle &= e^{i\phi_{CA}}\frac{1}{\sqrt{6}},\\
\langle\epsilon_C|\eta_B\rangle &= e^{i\phi_{CB}}\frac{1}{\sqrt{6}},\\
\langle\epsilon_C|\eta_C\rangle &= e^{i\phi_{CC}}\sqrt{\frac{2}{3}}.
\end{aligned}$$

We list the available geometric symmetries that leave the final stationary state unchanged:

1. Phase transformation on each state, $|\eta_x\rangle \mapsto e^{i\phi_x}|\eta_x\rangle$;
2. Phase transformation on each dual state, $|\epsilon_x\rangle \mapsto e^{i\phi_x}|\epsilon_x\rangle$; and
3. Unitary transformation $|\eta_x\rangle \mapsto U|\eta_x\rangle$ and $\langle\epsilon_x| \mapsto \langle\epsilon_x|U^\dagger$.

From these symmetries we can fix gauge in the following ways:

1. Set $\langle 0|\eta_x\rangle$ to be real and positive for all x.
2. Set $\phi_{AA} = \phi_{BB} = \phi_{CC} = 0$.
3. Set $\langle 0|\eta_A\rangle = 0$ and set $\langle 1|\eta_B\rangle$ to be real and positive.

These gauge fixings allow us to write:

$$\begin{aligned}
|\eta_A\rangle &= |0\rangle\,,\\
|\eta_B\rangle &= \alpha_B|0\rangle + \beta_B|1\rangle\,, \text{ and}\\
|\eta_C\rangle &= \alpha_C|0\rangle + e^{i\theta}\beta_C|1\rangle\,,
\end{aligned}$$

for $\alpha_B, \alpha_C \geq 0$, $\beta_B = \sqrt{1-\alpha_B^2}$ and $\beta_C = \sqrt{1-\alpha_C^2}$ and a phase θ.

That these states are embedded in a 2D Hilbert space means there must exist some linear consistency conditions. For some triple of numbers $\mathbf{c} = (c_A, c_B, c_C)$ we can write:

$$c_A|\eta_A\rangle + c_B|\eta_B\rangle + c_C|\eta_C\rangle = 0.$$

Up to a constant, this triplet has the form:

$$(c_A, c_B, c_C) = \left(e^{i\theta}\alpha_B \frac{\beta_C}{\beta B} - \alpha_C,\ -e^{i\theta}\frac{\beta_C}{\beta B},\ 1 \right).$$

Consistency requires that this relationship between vectors is preserved by the Kraus operator dynamic. Consider the matrix $\mathbf{A} := (A_{xy}) = \left(\langle \epsilon_x | \eta_y \rangle\right)$. The vector $\mathbf{c}$ must be a null vector of $\mathbf{A}$; i.e. $\sum_y A_{xy} c_y = 0$. This first requires that A_{xy} be degenerate. One way to enforce this is 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 $\mathbf{A}\sqrt{6}$:

$$\begin{aligned}\det(\sqrt{6}\mathbf{A} - \lambda \mathbf{I}_3) = \quad & (2-\lambda)^3 + \left(e^{i(\phi_{AB}+\phi_{BC}+\phi_{CA})} + e^{i(\phi_{BA}+\phi_{CB}+\phi_{AC})} \right) \\ & - (2-\lambda)\left(e^{i(\phi_{AB}+\phi_{BA})} + e^{i(\phi_{AC}+\phi_{CA})} + e^{i(\phi_{BC}+\phi_{CB})} \right).\end{aligned}$$

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:

$$\mathbf{A} = \frac{1}{\sqrt{6}} \begin{pmatrix} 2 & e^{i\phi_1} & -e^{i(\phi_1+\phi_2)} \\ e^{-i\phi_1} & 2 & e^{i\phi_2} \\ -e^{-i(\phi_1+\phi_2)} & e^{-i\phi_2} & 2 \end{pmatrix}.$$

We now need to enforce that $\sum_y A_{xy} c_y = 0$. We have the three equations:

$$\begin{aligned} 2c_A + e^{i\phi_1} c_B - e^{i(\phi_1+\phi_2)} c_C &= 0, \\ 2c_B + e^{-i\phi_1} c_A + e^{i\phi_2} c_C &= 0, \text{ and} \\ 2c_C + e^{-i\phi_2} c_B - e^{-i(\phi_1+\phi_2)} c_A &= 0. \end{aligned}$$

It can be checked that these are solved by:

$$\begin{aligned} c_A &= e^{i(\phi_1+\phi_2)} c_C \text{ and} \\ c_B &= -e^{i\phi_2} c_C. \end{aligned}$$

Taking our formulation of the $\mathbf{c}$ vector, we immediately have $\beta_B = \beta_C = \beta$ (implying $\alpha_B = \alpha_C = \alpha$), $\phi_2 = \theta$, and:

$$\begin{aligned} e^{-i\phi_3} &= \alpha(1 - e^{i\theta}) \\ &= -2i\alpha \sin(\theta) e^{i\theta/2} \\ &= \alpha \sin(\theta) e^{i(\theta-\pi)/2}. \end{aligned}$$

This means:

$$\alpha = \frac{1}{2}\left|\csc\left(\frac{\theta}{2}\right)\right| \text{ and}$$
$$\phi_3 = \frac{-\theta + \text{sgn}(\theta)\pi}{2},$$

where we take $-\pi \leq \theta \leq \pi$ and $\text{sgn}(\theta)$ is the sign of θ.

Note, however, that for $-\frac{\pi}{3} < \theta < \frac{\pi}{3}$, we have $|\csc(\theta)| > 1$, so these values are unphysical.

We see that all parameters in our possible states $|\eta_x\rangle$, as well as all the possible transition phases, are dependent on the single parameter θ. To construct the dual basis, we start with the new forms of the states:

$$|\eta_A\rangle = |0\rangle\,,$$
$$|\eta_B\rangle = \alpha\,|0\rangle + \beta\,|1\rangle\,, \text{ and}$$
$$|\eta_C\rangle = \alpha\,|0\rangle + e^{i\theta}\beta\,|1\rangle\,.$$

We note directly that we must have:

$$\langle\epsilon_A|0\rangle = \sqrt{\frac{2}{3}},$$
$$\langle\epsilon_B|0\rangle = \frac{1}{\sqrt{6}}e^{-i\phi_1}, \text{ and}$$
$$\langle\epsilon_C|0\rangle = \frac{1}{\sqrt{6}}e^{i\phi_3},$$

from how the dual states contract with $|\eta_A\rangle$. These can be used with the contractions with $|\eta_B\rangle$ to get:

$$\langle\epsilon_A|1\rangle = \frac{1}{\beta}\sqrt{\frac{2}{3}}\left(\frac{1}{2}e^{i\phi_1} - \alpha\right),$$
$$\langle\epsilon_B|1\rangle = \frac{1}{\beta}\sqrt{\frac{2}{3}}\left(1 - \frac{1}{2}\alpha e^{-i\phi_1}\right), \text{ and}$$
$$\langle\epsilon_C|1\rangle = \frac{1}{2\beta}\sqrt{\frac{2}{3}}\left(e^{-i\phi_2} - \alpha e^{i\phi_3}\right).$$

It is quickly checked that these coefficients are consistent with the action on on $|\eta_C\rangle$ by making liberal use of $e^{-i\phi_3} = \alpha(1 - e^{i\theta})$.

Recall that with the correct dual states, the Kraus operators take the form:

$$K_A = |\eta_A\rangle\langle\epsilon_A|,$$
$$K_B = |\eta_B\rangle\langle\epsilon_B|, \text{ and}$$
$$K_C = |\eta_C\rangle\langle\epsilon_C|.$$

Completeness requires:

$$|\epsilon_A\rangle\langle\epsilon_A| + |\epsilon_B\rangle\langle\epsilon_B| + |\epsilon_C\rangle\langle\epsilon_C| = I.$$

Define the vectors $u_x = \langle\epsilon_x|0\rangle$ and $v_x = \langle\epsilon_x|1\rangle$. One can check that the above relationship implies $\sum_x u_x^* u_x = \sum_x v_x^* v_x = 1$ and $\sum_x u_x v_x^* = 0$. However, for our model, it is straightforward (though a bit tedious) to check that:

$$\sum_x u_x^* u_x = \frac{2}{3} + \frac{1}{6} + \frac{1}{6} = 1 \text{ and}$$

$$\sum_x v_x^* v_x = \frac{1}{\beta^2}\left(1 + \alpha^2 - \alpha\cos\phi_1\right).$$

Using the definitions of α, β, and ϕ_1, the second equation can be simplified to:

$$\sum_x v_x^* v_x = \frac{2 + \csc^2\frac{\theta}{2}}{4 - \csc^2\frac{\theta}{2}}.$$

This is unity only when $\csc^2\frac{\theta}{2} = 1$, which requires that $\theta = \pi$. This is, indeed, the model $\mathfrak{D}_3$ that we have already seen.

This establishes that the only two-dimensional pure-state quantum model which reproduces the 3-state MBW process is the one with a nonminimal statistical memory $S(\rho_\pi)$. This means there cannot exist a quantum representation of the 3-state MBW process that majorizes all other representations of the same. For, if it existed, it must be a two-dimensional model and also minimize $S(\rho_\pi)$.

References

1. Lorenz, E.N.: Deterministic nonperiodic flow. J. Atmos. Sci. **20**, 130 (1963)
2. Lorenz, E.N.: The problem of deducing the climate from the governing equations. Tellus **XVI**, 1 (1964)
3. Boyd, A.B., Mandal, D., Crutchfield, J.P.: Identifying functional thermodynamics in autonomous Maxwellian ratchets. New J. Phys. **18**, 023049 (2016)
4. Crutchfield, J.P., Young, K.: Inferring statistical complexity. Phys. Rev. Lett. **63**, 105–108 (1989)
5. Crutchfield, J.P.: The calculi of emergence: computation, dynamics, and induction. Physica D **75**, 11–54 (1994)
6. Shalizi, C.R., Crutchfield, J.P.: Computational mechanics: pattern and prediction, structure and simplicity. J. Stat. Phys. **104**, 817–879 (2001)
7. Crutchfield, J.P.: Between order and chaos. Nat. Phys. **8**, 17–24 (2012)
8. Gu, M., Wiesner, K., Rieper, E., Vedral, V.: Quantum mechanics can reduce the complexity of classical models. Nat. Commun. **3**, 762 (2012)
9. Mahoney, J.R., Aghamohammadi, C., Crutchfield, J.P.: Occam's quantum strop: synchronizing and compressing classical cryptic processes via a quantum channel. Sci. Rep. **6**, 20495 (2016)
10. Aghamohammadi, C., Mahoney, J.R., Crutchfield, J.P.: Extreme quantum advantage when simulating classical systems with long-range interaction. Sci. Rep. **7**, 6735 (2017)
11. Suen, W.Y., Thompson, J., Garner, A.J.P., Vedral, V., Gu, M.: The classical-quantum divergence of complexity in modelling spin chains. Quantum **1**, 25 (2017)
12. Garner, A.J.P., Liu, Q., Thompson, J., Vedral, V., Gu, M.: Provably unbounded memory advantage in stochastic simulation using quantum mechanics. New J. Phys. **19**, 103009 (2017)
13. Aghamohammadi, C., Loomis, S.P., Mahoney, J.R., Crutchfield, J.P.: Extreme quantum memory advantage for rare-event sampling. Phys. Rev. X **8**, 011025 (2018)
14. Thompson, J., Garner, A.J.P., Mahoney, J.R., Crutchfield, J.P., Vedral, V., Gu, M.: Causal asymmetry in a quantum world. Phys. Rev. X **8**, 031013 (2018)
15. Riechers, P.M., Mahoney, J.R., Aghamohammadi, C., Crutchfield, J.P.: Minimized state-complexity of quantum-encoded cryptic processes. Phys. Rev. A **93**(5), 052317 (2016)
16. Coecke, B., Fritz, T., Spekkens, R.W.: A mathematical theory of resources. Inf. Comput. **250**, 59–86 (2016)
17. Marshall, A.W., Olkin, I., Arnold, B.C.: Inequalities: Theory of Majorization and Its Applications, 3rd edn. Springer, New York (2011)
18. Nielsen, M.A.: Conditions for a class of entanglement transformations. Phys. Rev. Lett. **83**, 436 (1999)
19. Horodecki, M., Oppenheim, J.: Fundamental limitations for quantum and nanoscale thermodynamics. Nat. Commun. **4**, 2059 (2013)
20. Gour, G., Müller, M.P., Narasimhachar, V., Spekkens, R.W., Halpern, N.Y.: The resource theory of informational nonequilibrium in thermodynamics. Phys. Rep. **583**, 1–58 (2015)

21. Grätzer, G.: Lattice Theory: Foundation. Springer, Basel (2010)
22. Lind, D., Marcus, B.: An Introduction to Symbolic Dynamics and Coding. Cambridge University Press, New York (1995)
23. Renner, R., Wolf, S.: Smooth Rényi entropy and applications. In: Proceedings 2004 IEEE International Symposium on Information Theory, IEEE Information Theory Society, Piscataway, p. 232 (2004)
24. Tomamichel, M.: A Framework for Non-Asymptotic Quantum Information Theory. PhD thesis, ETH Zurich, Zurich(2012)
25. Horodecki, M., Oppenheim, J., Sparaciari, C.: Extremal distributions under approximate majorization. J. Phys. A **51**, 305301 (2018)
26. Upper, D.R.: Theory and Algorithms for Hidden Markov Models and Generalized Hidden Markov Models. PhD thesis, University of California, Berkeley. Published by University Microfilms Intl, Ann Arbor (1997)
27. Crutchfield, J.P., Riechers, P., Ellison, C.J.: Exact complexity: spectral decomposition of intrinsic computation. Phys. Lett. A **380**(9–10), 998–1002 (2016)
28. Riechers, P.M., Crutchfield, J.P.: Spectral simplicity of apparent complexity. I. The nondiagonalizable metadynamics of prediction. Chaos **28**, 033115 (2018)
29. Riechers, P.M., Crutchfield, J.P.: Spectral simplicity of apparent complexity. II. Exact complexities and complexity spectra. Chaos **28**, 033116 (2018)
30. Yang, C., Binder, F.C., Narasimhachar, V., Gu, M.: Matrix product states for quantum stochastic modelling. arXiv:1803.08220 [quant-ph] (2018)
31. Travers, N.F., Crutchfield, J.P.: Equivalence of history and generator ϵ-machines. arxiv.org:1111.4500 [math.PR]
32. Hopcroft, J.: An $n \log n$ algorithm for minimizing states in a finite automaton. In: Paz, A., Kohavi, Z. (eds.) Theory of Machines and Computations, pp. 189–196. Academic Press, New York (1971)
33. Travers, N.F., Crutchfield, J.P.: Exact synchronization for finite-state sources. J. Stat. Phys. **145**, 1181–1201 (2011)
34. Binder, F.C., Thompson, J., Gu, M.: A practical, unitary simulator for non-Markovian complex processes. Phys. Rev. Lett. **120**, 240502 (2017)
35. Monras, A., Beige, A., Wiesner, K.: Hidden quantum Markov models and non-adaptive read-out of many-body states. Appl. Math. Comput. Sci. **3**, 93 (2011)
36. Hughston, L.P., Jozsa, R., Wootters, W.K.: A complete classification of quantum ensembles having a given density matrix. Phys. Lett. A **183**, 12–18 (1993)
37. Crutchfield, J.P., Ellison, C.J., Mahoney, J.R., James, R.G.: Synchronization and control in intrinsic and designed computation: an information-theoretic analysis of competing models of stochastic computation. CHAOS **20**(3), 037105 (2010)
38. Tan, R., Terno, D.R., Thompson, J., Vedral, V., Gu, M.: Towards quantifying complexity with quantum mechanics. Eur. J. Phys. Plus **129**, 191 (2014)
39. Aghamohammadi, C., Mahoney, J.R., Crutchfield, J.P.: The ambiguity of simplicity in quantum and classical simulation. Phys. Lett. A **381**(14), 1223–1227 (2017)
40. Löhr, W., Ay, N.: Non-sufficient memories that are sufficient for prediction. In: Zhou, J. (ed.) Complex Sciences 2009. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 4, pp. 265–276. Springer, New York (2009)
41. Löhr, W., Ay, N.: On the generative nature of prediction. Adv. Complex Syst. **12**(02), 169–194 (2009)
42. Ruebeck, J.B., James, R.G., Mahoney, J.R., Crutchfield, J.P.: Prediction and generation of binary Markov processes: Can a finite-state fox catch a Markov mouse? Chaos **28**, 013109 (2018)
43. Crutchfield, J.P., Marzen, S.: Signatures of infinity: nonergodicity and resource scaling in prediction, complexity and learning. Phys. Rev. E **91**, 050106 (2015)
44. Crutchfield, J.P., Marzen, S.: Structure and randomness of continuous-time, discrete-event processes. J. Stat. Phys. **169**(2), 303–315 (2017)
45. Elliot, T.J., Garner, A.J.P., Gu, M.: Quantum self-assembly of causal architecture for memory-efficient tracking of complex temporal and symbolic dynamics. arxiv.org:1803.05426 (2018)