Provably superior accuracy in quantum stochastic modeling

Chengran Yang,[1,2,*] Andrew J. P. Garner,[2,3] Feiyang Liu,[4] Nora Tischler,[5,6] Jayne Thompson,[1]
Man-Hong Yung,[4] Mile Gu,[2,1,7,†] and Oscar Dahlsten[8,4,‡]

[1]*Centre for Quantum Technologies, National University of Singapore, 3 Science Drive 2, Singapore 117543, Singapore*
[2]*Nanyang Quantum Hub, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371, Singapore*
[3]*Institute for Quantum Optics and Quantum Information, Austrian Academy of Sciences, Boltzmanngasse 3, Vienna 1090, Austria*
[4]*Shenzhen Institute for Quantum Science and Engineering and Department of Physics,*
Southern University of Science and Technology, Shenzhen 518055, China
[5]*Centre for Quantum Computation and Communication Technology,*
Centre for Quantum Dynamics, Griffith University, Yuggera Country, Brisbane, Queensland 4111, Australia
[6]*Dahlem Center for Complex Quantum Systems, Freie Universität Berlin, 14195 Berlin, Germany*
[7]*MajuLab, CNRS-UNS-NUS-NTU International Joint Research Unit, UMI 3654, Singapore 117543, Singapore*
[8]*Department of Physics, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong*

(Received 13 November 2022; revised 30 May 2023; accepted 14 July 2023; published 10 August 2023)

In the design of stochastic models, there is a constant trade-off between model complexity and accuracy. Here we prove that quantum models enable a more favorable trade-off. We present a technique for identifying fundamental upper bounds on the predictive accuracy of dimensionality-constrained classical models. We identify quantum models that surpass this bound by creating an algorithm that learns quantum models given time-series data. We demonstrate that this quantum accuracy advantage is attainable in a present-day noisy quantum device. These results illustrate the immediate relevance of quantum technologies to time-series analysis and offer an instance where their resulting accuracy advantage can be provably established.

DOI: 10.1103/PhysRevA.108.022411

I. INTRODUCTION

In data analytics, the curse of dimensionality is a well-acquainted adversary [1]. As we seek to predict the future behavior of ever-more complex processes, tracking all potentially relevant past observations becomes quickly intractable. Even when the process emits only binary outputs at each point in time, the cost of accounting for temporal correlations in the last n time steps grows as 2^n—making the exact simulation of highly non-Markovian processes computationally infeasible. These considerations motivate the need for modeling with dimensional constraints. Tractability is restored by constraining the number of possible pasts we consider, at a potential cost to the model accuracy. A big question is then: Given certain constraints on model complexity, what is the most accurate predictive model? Indeed, much of machine learning, from dimensional reduction, feature detection to auto-encoders, concerns variants of this question [2–4].

In traditional approaches, predictive models are classical. The possible observed pasts are classified through classical bits of data. For example, consider a discrete-time clock, which "ticks" by emitting "1" every N time steps and "0" at all other times. A model using a single bit of memory cannot count to N. Should $N > 2$, future distortions become unavoidable. Assigning more bits to our model would, of course, mitigate this distortion. However, this comes at a cost of increased memory dimension. Forcing a bounded memory dimension can thus result in unavoidable model distortion.

Here we ask: *Can dimensionally constrained quantum models—quantum machines that record relevant past information using qubits—reduce such distortions?*

To answer this, we first develop a means to bound the predictive accuracy that can be achieved by the classical models with constrained memory dimensions. We then demonstrate that these bounds can be surpassed by quantum-mechanical models with the same memory dimension. These quantum models were identified by designing a systematic algorithm that takes raw time-series data as input. We implement a model found by our algorithm on the IBM quantum computer ("ibmq athens"). Despite the noise, a statistically significant accuracy advantage remains, demonstrating a provable accuracy advantage from executing quantum models. Our quantum models, capable of generating possible futures in quantum superposition, can be a key resource for contemporary quantum algorithms in risk and stochastic time-series analysis [5–7] in the noisy intermediate-scale quantum (NISQ) era.

II. CLASSICAL AND QUANTUM MODELS

Consider a physical system observed at discrete times t with corresponding outcomes $x_t \in \mathcal{X}$ over some finite alphabet $\mathcal{X}$. We assume that the system is stationary and stochastic, such that each x_t is drawn from random variable X_t, all of

[*] yangchengran92@gmail.com
[†] mgu@quantumcomplexity.org
[‡] oscar.dahlsten@cityu.edu.cn

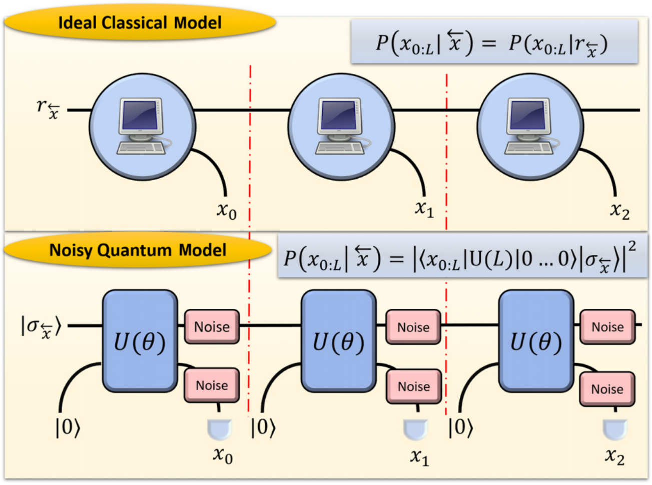

Ideal Classical Model
$P(x_{0:L}|\overleftarrow{x}) = P(x_{0:L}|r_{\overleftarrow{x}})$
$r_{\overleftarrow{x}}$
x_0
x_1
x_2
Noisy Quantum Model
$P(x_{0:L}|\overleftarrow{x}) = |\langle x_{0:L}|U(L)|0\ldots0\rangle|\sigma_{\overleftarrow{x}}\rangle|^2$
$|\sigma_{\overleftarrow{x}}\rangle$
$U(\theta)$
Noise
$|0\rangle$
x_0
x_1
x_2

FIG. 1. Comparison of classical and quantum models that produce statistical samples of a stochastic process's future behavior. A classical model stores classical states in its memory. In quantum models, the quantum memory and the output register—initialized in state $|0\rangle$—interact through the unitary transformation U, and a measurement of the output register in the computational basis yields x_i. This can be repeated to sample $x_{0:L}$. $U(L)$ denotes the unitary coupling the memory with L consecutive output registers. We train the parameters θ of U on a classical computer. We execute the resulting model on a noisy quantum device and prove that the accuracy outperforms any same-dimension classical model.

which are governed by a time-translation invariant distribution $P(\overleftarrow{X}, \overrightarrow{X})$. Here, $\overleftarrow{X} := \cdots X_{-2}X_{-1}$ and $\overrightarrow{X} := X_0X_1X_2\cdots$ denote random variables governing stochastic outcomes in the past and future.

An (exact) *predictive model* is a systematic algorithm that takes each possible past $\overleftarrow{x}$, encodes it via the deterministic map $\mathcal{E}$ to some suitable state $\mathcal{E}(\overleftarrow{x})$ within a physical memory M, while at each subsequent time-step sequentially outputting $x_0, x_1, x_2\cdots$ with probability $P(\overrightarrow{X} = \overrightarrow{x}|\overleftarrow{x})$, as shown in Fig. 1. Knowing the state of the memory M is thus as useful as $\overleftarrow{x}$ for purposes of inferring future statistics.

Regardless of how many steps L of the process are produced, it is always possible to work out what state we should update M to, such that the model can continue the simulation process. If a predictive model in state $r_{\overleftarrow{x}} = \mathcal{E}(\overleftarrow{x})$ emits some output x_0, we know with certainty that the model memory will transition to state $r_{\overleftarrow{x},x_0} = \mathcal{E}(\overleftarrow{x}, x_0)$ [8]. That requirement of a unique new memory state for a given output and memory state is termed *unifilarity* (see also Appendix F).

The memory size is naturally quantified by the number of distinguishable memory states: d. There has been significant interest in identifying the memory-minimal models for replicating the statistics of a process. There are many motivations, from inferring causal structure, quantifying complexity, energy harvesting, to simulation of rare events [9–13]. This research has resulted in the definition of the *topological complexity* d_c of a process, quantifying the minimal memory dimension (the number of mutually distinguishable internal states) a model needs to predict its future. For Markovian processes, knowing the very last output, x_{-1} is sufficient for distortion-free future prediction. Thus the process's topological complexity needs never exceed the size of the output alphabet, $|\mathcal{X}|$. For non-Markovian processes though, d_c can scale without bound [14]. For example, the discrete-time clock that ticks every N seconds aforementioned would have a topological complexity of N. Using a model with a lower memory dimension would result in a necessary *distortion* in predicted future statistics, highlighting a trade-off between memory cost and accuracy.

In conventional models, the memory M is assumed to be classical. Quantum mechanics, however, enables each past to be encoded in a suitable quantum state $|\sigma_{\overleftarrow{x}}\rangle$ [15–20]. Here we are interested in (1) whether quantum models can lead to a favorable trade-off of memory dimension and accuracy, and if so (2) how such models can be designed. One motivation is as a proof-of-principle demonstration of quantum advantage in noisy intermediate scale quantum (NISQ) devices, where our quantum memory is naturally constrained. The second is that dimensional constraints are useful even when there are no clear memory constraints. This is because many applications require models to be learned directly from data. In such scenarios, statistical fluctuations imply that aiming for zero distortion is likely meaningless—resulting in high-dimensional models that are prone to overfitting. Indeed, recent work has shown that models that remember less past information may be more effective at generalizing [5].

III. QUANTIFYING DISTORTION

We begin by introducing a formal quantifier of model distortion. We employ the Kullback–Leibler (KL) divergence between probability distributions $P(x)$ and $Q(x)$,

$$\mathcal{D}(P, Q) = \sum_x P(x)\log_2\left(\frac{P(x)}{Q(x)}\right). \tag{1}$$

The KL divergence has operational significance in diverse contexts including machine learning, hypothesis testing and thermodynamics [21,22]. Consider a candidate model that, when fed $\overleftarrow{x}$, generates conditional future statistics $\hat{P}_{\overrightarrow{X_{0:L}}|\overleftarrow{x}}$ instead of desired statistics $P_{\overrightarrow{X_{0:L}}|\overleftarrow{x}}$, where $X_{0:L} = X_0X_1\ldots X_{L-1}$ denotes the first L future outputs. One *possible* definition of the distortion associated with the candidate model is then

$$D_L(P, \hat{P}|\overleftarrow{x}) = \frac{1}{L}\mathcal{D}(P_{X_{0:L}|\overleftarrow{x}}, \hat{P}_{X_{0:L}|\overleftarrow{x}}). \tag{2}$$

However, (i) we want a distortion measure that does not depend on L, the number of future time points. Moreover, (ii) different pasts $\overleftarrow{x}$ are associated with different distortions. To take (i) and (ii) into account we define the distortion associated with $\hat{P}$ as

$$D_e(P, \hat{P}) = \sum_{\overleftarrow{x}} P(\overleftarrow{X} = \overleftarrow{x}) \lim_{L\to\infty} D_L(P, \hat{P}|\overleftarrow{x}), \tag{3}$$

where the e subscript reminds us it is the expected (meaning average) distortion rate over pasts. Operationally, $D_e(P, \hat{P})$ represents how the likelihood of mistaking n samples from a sequence of length-L conditional futures from $\hat{P}$ for one generated by the true distribution P scales with L. Specifically, this likelihood scales as 2^{-nLD_e} in the limit of large L and n [22,23].

IV. CLASSICAL DISTORTION BOUNDS

Our goal here is then to first bound the minimal distortion achievable for a classical model of given memory dimension $\hat{d}_c$, and find means to surpass this bound using quantum models under identical memory constraints. To derive such bounds, we first make use of *computational mechanics*, which provides a means to construct memory-minimal classical models in the zero-distortion limit. Such constructions—ε *machines* [8,10,24,25]—rationalize that if two $\overleftarrow{x}$ and $\overleftarrow{x}'$ have coinciding future statistics, then mapping them into the same memory state s_i will not increase distortion. Thus, the ε machine then allocates one memory state for each equivalence class of pasts with the same conditional future [26]. The number of equivalence classes provably coincides with the process's classical topological complexity, d_c. Thus, there exists a systematic method to find an *exact* model whenever $\hat{d}_c \geqslant d_c$.

The situation when $\hat{d}_c < d_c$ is significantly less trivial. There exists no known means to prove a candidate predictive model achieves minimal distortion. Indeed, most research has been focused on inferring approximate classical models from data—owing to its practical usefulness [27]. Such models constitute upper bounds on achievable distortion. However, we need the lower bound on achievable distortion in order to be able to claim a provable quantum advantage for distortion.

An intuition behind our derivation of the lower bound is that it is natural to create approximate classical models by "merging" causal states, a process which can also be termed a coarse-graining of the memory states. For example, a predictive model that has no distortion if one uses three causal states (s_1, s_2, s_3) may be approximated by a model with just two causal states, such as (s_1, s_2) where s_2 can be the encoded state for the case of the output associated with s_3 in the exact model. We prove no other method of creating approximate models has lower distortion.

The proof involves introducing a more general type of "models" which can be used to bound the achievable distortion of predictive models. We call this wider family *premodels*. Such premodels still include an encoding function $\mathcal{E}$ that encodes each possible past within a memory M. However, they are not required to output future predictions in temporal order: instead of generating predictions of X_0, X_1, $X_2, \ldots$ by repeated application of the same process, they may generate a prediction of all future outcomes simultaneously (see Appendix F for a precise definition). Since these premodels constitute a wider family, if no premodel with memory dimension d can achieve a distortion of D_e, then no classical model can either. We consider a class of premodels that correspond to "coarse-grained" ε machines—where additional memory is saved by merging two or more causal states. We can then prove (see Appendix F) that the distortion achievable through coarse-graining lower bounds that of all classical predictive models:

Theorem 1. Let P be any stationary process. Then, for every model dimension $\hat{d}_c$, the algorithm in Box 1 produces a bound on the minimum distortion D_e of a unifilar model by considering all possible mergers of causal states and searching over the next K output statistics.

Thus, minimizing the distortion over the finite set of coarse-grained classical models, yields a lower bound on the distortion of *all* classical models for the given memory dimension.

Theorem 1 accordingly enables a set procedure (see Box 1) that allows us to identify a fundamental bound on the minimal distortion (as characterized by the normalized Kullback–Leibler divergence) of any classical model under set memory constraints. Furthermore, we show in Theorem 2 of the Appendix that this bound is tight for Markovian processes (for which an approximate model is needed if the memory dimension is smaller than the size of the output alphabet) in that we can find a minimum distortion model through the procedure of Box 1. Thus we rigorously bound the accuracy of classical models of a given dimension.

Box 1: Bounding the distortion of classical models

Inputs:

P – the process P with Markov order κ
$\hat{d}_c$ – the dimension of the classical model.

Outputs:

Lower bound on distortion D_e of dimension $\hat{d}_c$ models for P.

Algorithm overview:

(a) Make use of Theorem 1 that optimal approximate models have causal states that are coarse-grainings (mergings) of the exact model's causal states.

(b) For all encodings $\mathcal{E}$ merging causal states of process P: Vary the associated conditional probabilities $\hat{P}(X_{0:\kappa}|\mathcal{E}(\overleftarrow{x}))$ to minimize the distortion relative to P:

$$D_\kappa(P, \hat{P}) = \sum_{\overleftarrow{x}} P(\overleftarrow{x}) D_\kappa(P, \hat{P}|\overleftarrow{x}).$$

This minimization is done with gradient descent (note $D_\kappa(P, \hat{P})$ is convex in $\hat{P}$).

(c) Return the minimal $D_\kappa(P, \hat{P})$ among all encodings $\mathcal{E}$.

V. INFERRING QUANTUM MODELS

A. Overview

We now describe our algorithm for learning memory-constrained quantum models for classical time-series data, summarized in Box 2. Recall that a $\hat{d}_q$-dimensional quantum model is one where each past $\overleftarrow{x}$ is encoded into a suitable quantum state $|\sigma_{\overleftarrow{x}}\rangle$ within a $\hat{d}_q$- dimensional quantum memory M. The action that generates conditional future outcomes $x_1, x_2, \ldots$ at each time-step can thus be a quantum process. Figure 1 illustrates one potential circuit realization of such models. There is (i) the coupling of M to an output register initially in $|0\rangle$ via a time-step-independent unitary U (ii) the measurement of the output register, and the subsequent emission of the measurement outcome.

Just as in the classical case, identifying distortion-minimizing quantum models of fixed memory dimension is also highly nontrivial. Our goal here then is not to necessarily find the provably optimal quantum model, but to find one that achieves a distortion smaller than our fundamental classical bound. To do this, we construct an algorithm that infers quantum models from time-series data. The algorithm

Box 2: Quantum model discovery algorithm.

```
Inputs:
  $\hat{d}_q$ – the desired model memory dimension.
  $x_{0:L}$ – a length L sequence from the process.
Outputs:
  $U$ – the unitary dynamics of the quantum model.
  $\mathcal{E}$ – the encoding map from observed histories
      to $\hat{d}_q$–dimensional quantum memory states.
Algorithm:
  1. To learn the unitary operator $U$:
     (a) Randomly initialize parameter set $A$ corresponding up to completeness to the set of Kraus operator matrices $\{A_x := \langle x|U|0\rangle\}$.
     (b) Evaluate the cost function $C := -\log P(x_{0:L}; A)$ and its gradient $\nabla C$.
     (c) Update the Kraus operators $\{A_x\}$ using gradient-descent based method (such as Adam).
     (d) Repeat (b)-(c) until the cost function decrease is sufficiently small.
     (e) Save final $A$ as candidate parameters.
     (f) Repeat (a)-(e) as desired, to minimize the impact of initial choice of $A$. Choose the final $A$ with the lowest C.
     (g) For that $A$, recover the completeness relation for $\{A_x\}$.
     (h) Construct a unitary operator $U$ from these $\{A_x\}$.
  2. To compute the encoding $\mathcal{E}$:
     (a) Compute the leading eigenvector $|\sigma_0\rangle$ of one of the Kraus operators (e.g., $A_0$).
     (b) $\mathcal{E}$ is then defined as
         $\mathcal{E}(\overleftarrow{x}) := A_{\overleftarrow{x}}|\sigma_0\rangle/\|A_{\overleftarrow{x}}|\sigma_0\rangle\|_2$.
```

takes two inputs: (i) a data sequence $x_{0:L}$, assumed to be drawn from some stationary stochastic process P and (ii) our desired memory dimension $\hat{d}_q$. From this data, the algorithm learns a quantum model of the process, including (i) an encoding map $\mathcal{E}$ which specifies how to encode each past into a quantum memory of dimension $\hat{d}_q$, and (ii) the relevant physical process—described by a unitary U as in Fig. 1—whose repeated application produces entangled output states. Measurements on such output states emit outputs that approximate the process's conditional future behavior.

To operate the learning algorithm, we cast the problem as minimizing some cost function over a parameter set B. This involves finding (a) an effective parametrization of all such models to optimize over and (b) a computable cost function that proxies the expected KL-divergence D_e.

To tackle (a), first note that our model's output behavior is entirely defined by a family of Kraus operators $A = \{A_x\}$, where $A_x = \langle x|U|0\rangle$ captures the action of the memory when it interacts with an output register that is subsequently observed to have output x. Observe also that A_x informs the *encoding map* from pasts onto quantum memory states. To see this, note that a memory initially in ρ is (up to normalization) mapped to $A_x \rho A_x^\dagger$ after observation of x. Thus, observation of a past $\overleftarrow{x}$ implies applying a sequence of Kraus operators $A_{\overleftarrow{x}} := A_{x_{-1}} A_{x_{-2}} \cdots$. We can therefore encode the $\overleftarrow{x}$ as

$$\mathcal{E}(\overleftarrow{x}) := |\sigma_{\overleftarrow{x}}\rangle := cA_{\overleftarrow{x}}|\sigma_0\rangle, \tag{4}$$

where c is the constant of normalization. Here, the initial quantum state $|\sigma_0\rangle$ can be the leading eigenvector of any Kraus operator A_x (see Appendix C).

To tackle (b), the obvious candidate cost function is D_e itself. However, evaluating D_e requires writing out probabilities of conditional futures whereas in a practical scenario only a sample $x_{0:L}$ of the desired process P would be given as input. We therefore employ the negative log-likelihood of producing $x_{0:L}$,

$$C(A) = -\log_2 P_A(x_{0:L}), \tag{5}$$

as a proxy, where the subscript A reminds us that the model being trained depends on the free parameters A. In the limit of $L \to \infty$, a model minimizing the negative log-likelihood of Eq. (5) exactly replicates the desired stochastic behavior [28,29]. With both the cost function and parameter spaces established, the use of gradient-descent optimization algorithms such as Adam is enabled.

B. Model parametrization

The candidate quantum models of dimension $\hat{d}_q$ for a process with output alphabet $\mathcal{X}$ can be parametrized by a complete set of $|\mathcal{X}|$ Kraus operators $A = \{A_x\}_{x\in\mathcal{X}}$, where A_x denotes the Kraus operator describing how the model updates upon emission of output x. However, this parametrization is nonideal for optimization due to the completeness constraint.

Here, we demonstrate an alternative parametrization using $B = \{B_x\}_{x\in\mathcal{X}}$, where each B_x is a general $\hat{d}_q \times \hat{d}_q$ complex matrix. Notably, we show that given any such B, it is possible to recover a corresponding set A via the following process: First, consider the linear map $\mathcal{E}^{B^*}(\cdot) := \sum_x B_x^\dagger \cdot B_x$. By construction, this map is completely positive, so its leading eigenvalue λ is real and positive, and the associated eigenmatrix V is positive semidefinite. Thus, V admits a decomposition $V = W^\dagger W$ (where W is invertible). We can then set A_x such that

$$A_x := W B_x W^{-1}/\sqrt{\lambda}. \tag{6}$$

Every set $\{A_x\}$ formed in this way satisfies the completeness relation (see Appendix A). Thus, each B_x can be used directly to construct a valid set of corresponding Kraus operators A_x. This further results in valid unitary operator U (see Appendix B). Likewise, via Eq. (6) one can also infer the encoding map $\mathcal{E}$ from B (see Appendix C).

C. Computing the cost function

The learning algorithm relies on optimizing the parameters to minimize the cost function. A direct way is optimizing over the Kraus operators A, but this is generally cumbersome as A is constrained by the completeness relation $\sum_x (A_x)^\dagger A_x = \mathbb{I}$. Thus, we instead devise a way to optimize over a set of unconstrained $\hat{d}_q \times \hat{d}_q$ complex matrices $B = \{B^x\}$, whose value enables the deduction of A via tensor network techniques [30]. We show that C can be computed directly from B without first deducing A, thereby boosting optimization efficiency. Once optimization concludes, the corresponding U for generating predictions is retrieved from B.

To see how to compute the cost function $C(B) = -\log_2 P_B(x_{0:L})$ directly from B, first consider a model with Kraus operators $A = \{A_x\}$ initialized in state ρ_0 at $t = 0$. The probability it outputs x at $t = 1$ is then given by $P(x|\rho_0)|_A = \mathrm{Tr}(A_x\rho_0(A_x)^\dagger)$, whereby the state transitions to $\rho_1 = A_x\rho_0(A_x)^\dagger$. As such, repeated iterations of the model will output $x_{0:L}$ with probability

$$P(x_{0:L}|\rho_0)|_A = \mathrm{Tr}\big(A_{x_{0:L}}\rho_0 A^\dagger_{x_{0:L}}\big), \tag{7}$$

where $A_{x_{0:L}} = A_{x_{L-1}}\cdots A_{x_0}$. In addition, if ρ_0 is the stationary memory state averaged over all histories of the process, we obtain the probability of output sequence $x_{0:L}$ when averaged over all pasts. We can then write this likelihood directly in terms of B^x by applying Eq. (6),

$$P_B(x_{0:L}) = \mathrm{Tr}\big(B_{x_{0:L}}\tilde{\rho}B_{x_{0:L}}{}^\dagger V\big)/\lambda^L, \tag{8}$$

where $\tilde{\rho} = W^{-1}\bar{\rho}W^{-1\dagger}$ is the leading eigenmatrix of $\mathcal{E}^B$.

We finally remark on the costs of computing the cost function. Consider the scaling in the length of the time-series data L and the memory dimension $\hat{d}_q$. The computational complexity of computing the cost function with a classical computer, as undertaken in this paper via the Kraus operators and matrix product state methods, is $\Theta(L\hat{d}_q^3)$. This is because there are L steps of matrix multiplication, which is known to have a worst-case complexity going as N^3 for N-dimensional matrices. In present contexts, since we are interested in finding quantum models with small fixed d_q which can simulate a target process with reduced distortion, this computational cost is not a significant constraint.

D. Training process

Here we specify the details of the training process to minimize the cost function $C(B) = -\log_2 P_B(x_{0:L})$. At this stage, any number of different optimization techniques could be employed. Specifically, we used Adam optimization [31]. The Adam method is a sophisticated form of gradient descent. Recall that standard gradient descent involves computing ∇C, the partial derivatives with respect to each degree of freedom in B (referred to as a free parameter), and updating B according to $B' = B - \eta\nabla C$, where $\eta > 0$ is an adjustable learning rate. The Adam method fine tunes this by using individual adaptive learning rates for each free parameter, computed using estimates of the first and second moments of the gradients ∇C [31]. Due to the capacity to relate the cost function directly to B [see Eq. (8)], this can be done without ever needing to write out the quantum circuit itself.

In each run of the training process, we begin with a random B. The update process is then repeated until the decrease of updated cost function is below a chosen threshold. Like nearly all gradient descent variants, Adam can converge to local minima. However, in our training we found roughly three quarters of all runs give the same minimal value for the negative log-likelihood C of Eq. (5), and no further gains were made after taking the optimal of three runs (see Appendixes D and E). The result of this algorithm for the examples discussed in this article is given in Appendix E.

VI. PERFORMANCE OF QUANTUM MODELS

We are now in a position to explore the question: *when can memory-constrained quantum models exhibit improved accuracy*? First, we illustrate provable quantum accuracy advantage in two settings: (1) cases where classical models cannot avoid distortion, while quantum models can avoid distortion entirely, and (2), cases where both quantum and classical models have nonzero distortion, but quantum models can achieve less distortion than any classical counterpart. We then conclude via a case study of modeling a family of discrete-time renewal processes, a stochastic generalization of discrete-time clocks aforementioned. We only consider the cases where the dimensionality of the approximate quantum models $\hat{d}_q > 1$, such that there is a nontrivial memory. For a trivial memory ($\hat{d}_q = 1$) the outputs are sampled from the same distribution at each time and any given distribution can be realized by both a quantum and classical model, so there can be no quantum accuracy advantage.

A. Quantum models with zero distortion

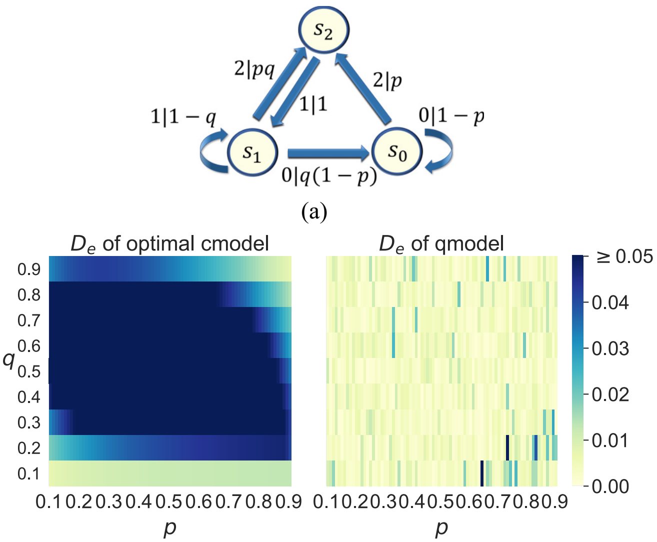
There are stochastic processes with quantum models whose memory dimension $d_q < d_c$. Consider any process with classical topological complexity d_c. We know then that if we constrain memory dimension to some $\hat{d}_c < d_c$, then no classical model can achieve zero distortion. Thus, provided we can find a quantum model of dimension $\hat{d}_q$ with zero distortion, we can establish a setting in which quantum models have a provable accuracy advantage. Such settings have been discovered [15–20,32].

We illustrate the quantum advantage in memory size using simple asymmetric processes. Each process in this family has three causal states [see its optimal predictor in Fig. 2(a)], which means $d_c = 3$. On the other hand, a quantum model can generate statistically identical predictions using a single qubit as memory [33]. Thus, if we limit our models to have a memory dimension of two, quantum models can maintain zero distortion—an impossible feat for any classical counterpart. In Fig. 2(b), we show minimal classical distortion (computed using our causal coarse-graining algorithm). The bound is exact since the process is Markovian. We observe that the normalized KL-divergence exceeds 0.05 for much of the parameter space. Therefore, quantum models have a definite accuracy advantage.

To determine if we can infer models with an accuracy advantage from finite training data, we apply our quantum inference algorithm to a data string containing a sample of the process over several time-steps (e.g., 1000). We indeed see in Fig. 2(b) an order of magnitude reduction in the distortion relative to the best possible approximate classical model. This is remarkable given that the quantum model was learned from finite data. In fact, there are families of processes—such as dual-Poisson processes (see Appendix I)—whose exact classical models require a memory dimension d_c that grows without bound, while $\hat{d}_q$ remains finite.

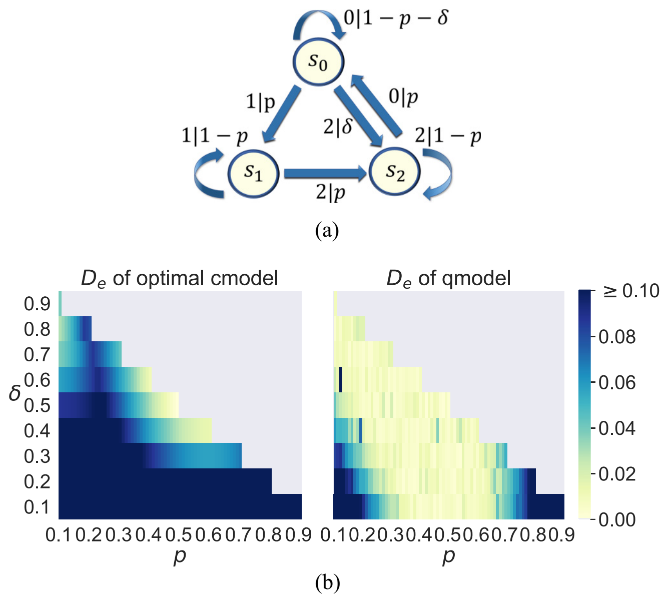
B. Finite quantum distortions

Previous discussion focused on analytically tractable cases where quantum models had *zero* distortion. However, in many



(b)

FIG. 2. Example: The asymmetric process. (a) ε machine for the asymmetric process, parametrized by $p, q \in (0, 1)$. The nodes are the causal states s_0, s_1, and s_2. The arrows indicate transitions between states, where each label $x|P(x)$ indicates output x is produced with probability $P(x)$. (b) Classical vs quantum distortion. Distortion in the two-dimensional models for the asymmetric process across a range of process parameters p, q. The left-hand-side shows the optimal approximate classical model's distortion. The right-hand-side shows the distortion of the most accurate quantum model found by our algorithm.



(b)

FIG. 3. Example: Quasicycle process. (a) ε machine for the three-state quasicycle parametrized by $p \in (0, 1)$, $\delta \in (0, 1-p]$. The nodes are the causal states s_0, s_1, and s_2. The arrows indicate transitions between states, where each label $x|P(x)$ indicates output x is produced with probability $P(x)$. (b) Classical vs quantum distortion comparison. Distortion in the two-dimensional models for the quasicycle, across a range of process parameters p, δ. The left-hand-side shows the optimal approximate classical model's error. The right-hand-side shows the distortion of the most accurate quantum model found by our algorithm.

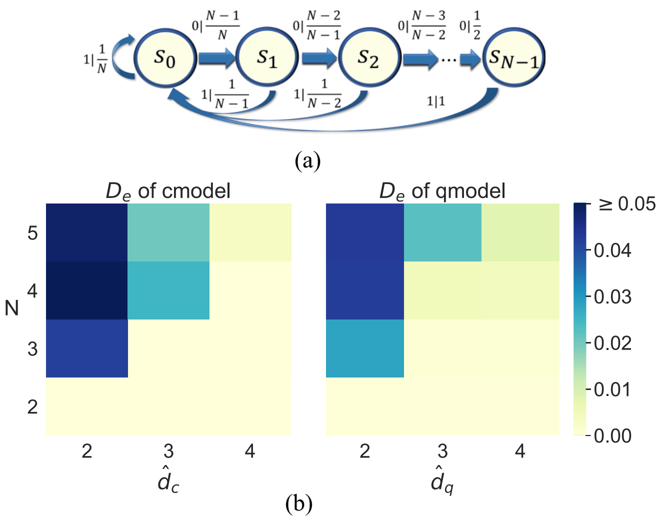
realistic scenarios, both quantum and classical models inevitably introduce distortions (e.g., when the model memory is constrained to some dimension that is less than the minimal dimension for exact simulation, or when models are learned from finite data). Indeed, dimensional quantum advantage for exact modeling is rare. For instance, for most of the parameters p and δ in the model described below in Fig. 3(a), the quantum models found do not have strictly zero distortion, which would in fact require memory dimension three. Thus, for quantum models to be useful in practical settings, we need to go beyond exact quantum models and determine situations where quantum models have an advantage in distortion-memory trade-of, and demonstrate that this advantage is much more generic.

To do this, we consider the minimal distortion caused when we constrain both classical and quantum models to a fixed memory dimension d. We achieve this by introducing coarse-graining, a technique that allows us to bound the minimal distortion we must suffer when constrained to d-dimensional classical models (Box 1). We then use our quantum inference protocol (Box 2) to design quantum models that surpass this bound and establish a provable quantum accuracy advantage.

We illustrate this technique on a family of generalized three-state quasicycles [see Fig. 3(a)]. Quasicycles are for example of interest in thermodynamics—certain quasicycles are valid thermal operations in that they preserve the Gibbs' thermal state, but interestingly without respecting detailed balance [34]. They constitute a two-parameter family of processes where zero-distortion models generally require both quantum and classical models to have three-dimensional memories—save for a one-dimensional subfamily that can be exactly modeled by quantum models of memory dimension two [35]. The resulting error of our inferred quantum model (see Appendix E) is compared with that of the classical limit in Fig. 3(b). In contrast, there is a broad area of process parameters (p, δ) which can be accurately approximated by a compressed quantum model. Thus, our results unveil that quantum models exhibit an accuracy advantage in a much broader class of processes once we drop requirements of exact simulation.

C. Modelling renewal processes

We now apply our inference algorithm to the family of discrete-time renewal processes—a family of progressively more non-Markovian processes that generalize the aforementioned ticking clock. This family of models allows us to investigate the quantum accuracy advantage for varying numbers of causal states. At each time step, the process may emit one of two outputs: 0, representing no tick, and 1, representing a tick. Their probability of ticking at each time step depends only on k, the number of time steps since their previous tick [36]. In the event that the system is guaranteed to tick for a particular value of k, we have a period clock. In the other extreme case where the probability is independent of k, we have a memoryless Poissonian process whose topological (classical or quantum) complexity is zero. Most physical clocks interpolate between these extremes [37].



(a)

(b)

FIG. 4. Example: The uniform renewal process. (a) The memory-minimal classical model of N-step uniform renewal process with zero distortion. The resulting model has N memory states (denoted $s_0, \ldots, s_{N-1}$). Arrows indicate the transition probability between states, where each label $x|P(x)$ indicates that output x is produced with probability $P(x)$. Any model with a memory dimension of less than N would necessarily feature distortion. Panel (b) plots a lower bound on this distortion for all possible classical models for various max-periodicity N and memory dimension $\hat{d}$. This is compared with the distortion achieved by quantum models delivered by our inference algorithm. We see that the quantum models achieve distortions significantly below classical limits in certain parameter regimes (e.g., when $N = 4$ and $\hat{d} = 3$).

Here we consider a class of period-N uniform renewal processes $\mathcal{U}_N$. These represent processes where the number of 0s between each two neighboring ticks is uniformly distributed between 0 and $N-1$. As N scales, the process becomes progressively more non-Markovian. Indeed, its topological complexity is N, following similar reasoning to the case of the periodic clock [see corresponding ε machine in Fig. 4(a)]. Reducing the number of distinct memory configurations to $\hat{d}_c < N$ results in unavoidable distortion.

Using causal state coarse-graining, we can establish a provable lower bound to this distortion for classical models for various $\hat{d}_c$ and N up to 5 [Fig. 4(b)]. As expected, classical models achieve zero distortion only when $\hat{d}_c \geqslant N$. Simultaneously, we plot the distortions of quantum models delivered by our discovery algorithm [Fig. 4(b)] using training data with 40 000 data points. Even though the learned quantum models have the additional handicap of being trained on finite-length data, we see that quantum models outperform optimal classical counterparts away from the zero-error regime.

To examine the resilience of the quantum accuracy advantage on present-day noisy devices, we implement the quantum model on one of IBM's quantum systems ("ibmq athens") for the case of $\hat{d}_q = 2$, $N = 3$ (see Appendix H). Error bars for computing distortion were estimated by running the model inference algorithm over 50 independent batches of data. We see that our quantum models can still exhibit a statistically significant accuracy advantage (of 2.12 standard deviations) over the classical bound, as shown in Fig. 5.

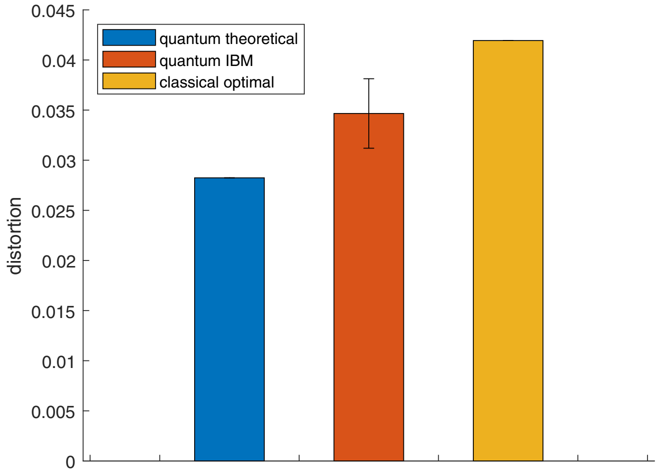


FIG. 5. Accuracy advantage on present-day quantum hardware. We executed a learned quantum model on ibmq-athens. To demonstrate a statistically significant accuracy advantage, 40 000 runs were divided into 50 batches and the distortion evaluated for each batch. Despite inherent noise in ibmq-athens, our quantum model achieves a distortion 2.12 standard deviations below our proven classical limit. For comparison, the blue bar shows the expected distortion of the learned quantum model without experimental noise. See Appendix H for details.

The intuition behind this accuracy advantage is that extra degrees of freedom are afforded by coherence in the quantum memory. Specifically, recall that ϵ machines, the classical optimal predictive models of zero distortion, operate by assigning a distinct memory to each equivalence class of pasts with statistically identical futures. Quantum models can replicate their statistical predictions exactly, while storing these memory states in nonmutually orthogonal states within the quantum memory [15]. While these memory states are typically independent and thus there are no immediate savings in memory dimension, this nonorthogonality does generally imply that the memory costs of exact modeling are reduced according to entropic measures. That is, certain subspaces within the memory of such quantum models can have significantly reduced occupation probability compared with classical counterparts. When we impose a dimensional memory constraint, removing such subspaces could result in less distortion compared with classical counterparts. Furthermore, since differences in entropic memory advantage of quantum models can scale without bound for generic classes of problems [38], we anticipate that accuracy advantages can persist in certain situations even when the quantum models have much more stringent dimensional constraints than classical counterparts.

VII. DISCUSSION

Modeling complex processes requires memory, allowing us to generate correct future behavior. As processes become more complex, so does the amount of memory they require—necessitating a trade-off between model accuracy and available memory. Here, we derived a minimal bound on this unavoidable distortion that applies to all classical models of a given stochastic process with a mixed memory dimension (i.e., the number of distinct past sequences it can reliably store). We then demonstrated that quantum models with the same memory dimension can surpass this bound and furthermore, such models can be learned from finite data. Thus, quantum information processing enables a more advantageous trade-off between accuracy and memory cost. We witnessed this advantage using an IBM quantum computer, where it generated predictions featuring a two sigma reduction in distortion beyond classical limits.

Taken together, these results significantly advanced our present understanding of quantum-enhanced predictive modeling. Previously, all works in minimizing model memory concerned the task of exact simulation—where the dimensional memory requirements of quantum and classical models are compared in the context of requiring zero distortion [33,35,38]. In this context, a quantum advantage was established. However, the stochastic processes involved needed to be finely tuned. The only general memory advantage occurred in the independent and identically distributed (i.i.d.) scenario, where $N \gg 1$ processes were simulated in unison. Here, we see that by considering the general cost of memory-distortion trade-offs, the advantage of quantum models is much more general.

We note also that this article was concerned with quantum advantage in the *execution* of a model. The algorithm for *learning* such models is entirely classical and as such, it scales (1) inefficiently with the number of qubits in quantum memory, (2) does not produce an explicit gate sequence for the discovered quantum models. This limits their applicability to scenarios where very low-dimensional quantum models can outperform very high-dimensional classical counterparts (e.g., renewal processes [38]). To go beyond such constraints, hybrid learning approaches will be necessary. One approach is to parametrize the unitary U using a shallow circuit ansatz, such that a suitable cost function is evaluated quantum-mechanically and optimized through variational means [39]. Such an algorithm can then leverage quantum computers to ensure that any models inferred can be directly implemented.

Doing so could have exciting algorithmic benefits. Observe that our quantum models can generate a superposition of all possible futures over L time steps weighted by their likelihood of occurrence, with circuit complexity scaling linearly with L. The resulting circuit is then a fundamental primitive for various data analytics protocols (e.g., amplitude amplification, Grover's search, value at risk, and importance sampling) [7,40,41]. Indeed, exact quantum models have already been deployed in quantum-enhanced stochastic analysis [6]. More fundamentally, recent results show that the less information a candidate model stores about past data, the greater its capacity to generalize [5]. This yields an exciting possibility: that quantum models may be fundamentally better for learning in regimes where training data are sparse.

ACKNOWLEDGMENTS

We are grateful for discussions with Thomas Elliott, Jim Crutchfield, and Kelvin Onggadinata. This research is supported by the National Research Foundation, Singapore, and Agency for Science, Technology and Research (A*STAR) under its QEP2.0 program (NRF2021-QEP2-02-P06), The Singapore Ministry of Education Tier 1 Grants RG77/22 and RG146/20, the QEP1.0 Grant QEP-SF3, Grant No. FQXI R-710-000-146-720 (Are quantum agents more energetically efficient at making predictions?) from the Foundational Questions Institute and Fetzer Franklin Fund (a donor-advised fund of Silicon Valley Community Foundation), the Quantum Engineering Program QEP-SP3. N.T. acknowledges support by the Griffith University Postdoctoral Fellowship Scheme, the Australian Research Council (ARC) Centre of Excellence CE170100012, and by the Alexander von Humboldt Foundation. O.D. acknowledges support from the National Natural Science Foundation of China (Grants No. 12050410246, No. 1200509, and No. 12050410245), a collaborative HiSilicon project, and City University of Hong Kong (Project No. 9610623).

Codes associated with this work are available online [42].

APPENDIX A: PROOF OF COMPLETENESS

Here we prove that for each model parametrized by $B = \{B_x\}_x$, the set $A = \{A_x\}_x$ is complete. This follows from

$$\begin{aligned}\sum_x A_x^\dagger A_x &= \sum_x \frac{1}{\lambda}(W^{-1})^\dagger B_x{}^\dagger W^\dagger W B_x W^{-1} \\ &= \frac{1}{\lambda}(W^\dagger)^{-1}\left(\sum_x B_x{}^\dagger W^\dagger W B_x\right) W^{-1} \\ &= (W^\dagger)^{-1} W^\dagger W W^{-1} = \mathbb{I}, \qquad \text{(A1)}\end{aligned}$$

where we use $(W^\dagger)^{-1} = (W^{-1})^\dagger$.

APPENDIX B: BUILDING UNITARY CIRCUITS

Once we have the Kraus operators A_x that describe a model, standard techniques enable construction of a full unitary circuit. Specifically, we need to find a $|\mathcal{X}|\hat{d}_q \times |\mathcal{X}|\hat{d}_q$ unitary operator U such that

$$\langle x|U|0\rangle = A_x. \qquad \text{(B1)}$$

As a result, some of the elements of the unitary operators are predefined, i.e.,

$$\langle j|\langle x|U|k\rangle|0\rangle = A_{x,jk}. \qquad \text{(B2)}$$

These predefined elements form $\hat{d}_q$ columns of the unitary operators, i.e., $|\phi_{k0}\rangle := U|k\rangle|0\rangle$. Such columns $|\phi_{k0}\rangle$ are mutually orthogonal quantum states, due to the completeness relation of A_x. Since an operator U is unitary if and only if its columns form an orthogonal basis, the remaining task is to find the complementary quantum states $|\phi_{kx}\rangle$ so that they form orthogonal states. This can be done by the Gram-Schmidt process [43].

APPENDIX C: INITIAL CONFIGURATION OF THE QUANTUM MODEL

We now describe how the initial state is picked when executing the learned model. The initial configuration of the quantum model is a pure quantum state $|\sigma_{\overleftarrow{x}}\rangle$, associated with the past, $\overleftarrow{x}$. This is required because the model is trained to work with the memory being in $|\sigma_{\overleftarrow{x}}\rangle$ states. We now describe how we extract such a state $|\sigma_{\overleftarrow{x}}\rangle$ from the Kraus operators A_i resulting from the training.

Consider first models with finite Markov order κ where only the latest κ steps of a given past $\overleftarrow{x}$ affect the future statistics, i.e., $P(\overrightarrow{X}|\overleftarrow{x}) \equiv P(\overrightarrow{X}|x_{-\kappa:0})$.

Then,

$$P(\overrightarrow{x}|\overleftarrow{x}) = P(\overrightarrow{x}|0^{\infty}, x_{-\kappa:0}), \tag{C1}$$

where $0^{\infty} := \ldots, 0, 0, 0$. The equality would also hold if 0^{∞} were replaced with other well-defined sequences of outcomes x. Recall we denote $A_0 = \langle 0|U|0\rangle$ as the Kraus operator associated with outcome 0. Thus A_0^{∞} is associated with the outcome 0^{∞}. In our simulations we used the leading eigenvector of A_0, $|\sigma_0\rangle$, as $|\sigma_{0^{\infty}}\rangle$. When there is a unique leading eigenvector $|\sigma_0\rangle$, which we generally expect when A_0 is the result of a learning process, then for any far distant past state $|\psi\rangle$, $A_0^{\infty}|\psi\rangle = |\sigma_0\rangle$ up to normalization.

A subtlety is that, although in our simulations the leading eigenvector of A_0, a somewhat messy learned matrix, has always been observed to be unique, one could manufacture degenerate cases where this is not the case. The uniqueness could then be recovered by adding a small perturbation to A_0. Moreover, degenerate models are unlikely to be the result of training because any of the degenerate initial configurations would exhibit the same outcome probabilities for all sequences of outcomes [44] and in that sense make trivial use of the memory. There are in fact two cases of quantum processes: (1) most of sequences of length L of Kraus operators $A_{x_{-L-\kappa:-\kappa}}$ have a unique leading eigenvector for sufficiently large L and the memory is used, or (2) the process makes trivial use of the memory [44].

When the process has *infinite* Markov order, perfect initialization of the machine would require an observation of the entire process history. However, one can approximate the perfect initialization by taking a long but finite string, which will have exponentially diminishing error with the length of the string [45]. As such, the above choice of $|\sigma_0\rangle$ remains valid.

APPENDIX D: HYPERPARAMETERS

The length of the sample in the first two examples was 1000, while that of the third example was 5000. The third example—the renewal process—has more causal states, and training quantum models with higher $\hat{d}_q$ also requires more data. For all of the three examples, the learning rate η is 0.1 and the optimizer is Adam [31]. We repeat the training three times and choose the quantum model of minimal cost.

The evaluation of the error $D_e(P, \hat{P})$ relies on computing the KL-divergence $\mathcal{D}_{\mathrm{KL}}(P(X_{0:L}|\overleftarrow{x})||\hat{P}(X_{0:L}|\epsilon(\overleftarrow{x})))$. Here, the length of the past $\overleftarrow{x}$ is five and the length of the future L is one (see Appendix G). This amounts to assuming that the Markov order of the process is less than or equal to five, which is true by construction for our examples. The update process is then repeated until ΔC is less than 0.1.

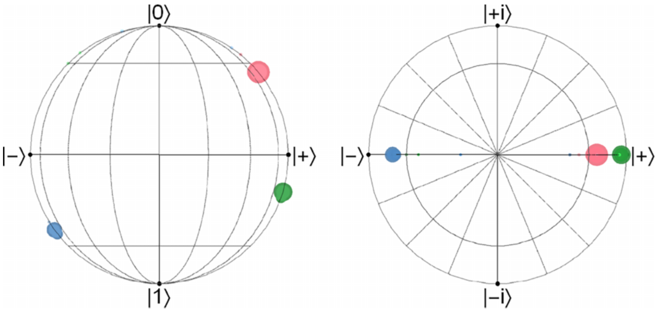


FIG. 6. Example: Asymmetric process—encoded states. Quantum states produced by the learned $\hat{d}_q = 2$ encoding map for the asymmetric process with $p = 0.3$, $q = 0.8$, represented on a Bloch sphere (viewed sideways and top-down). Each point represents a pure state mapped to by one history (of length five), and the color represents the history's associated causal state. For larger points (representing the most probable 99% of histories), the area is proportional to the probability of that history. For smaller points (representing the next 0.999% of histories), the opacity is proportional to the probability.

APPENDIX E: LEARNED QUANTUM MODELS

Here we present details of the quantum models produced by our model inference algorithm. We show the Kraus operators together with the corresponding encoding quantum states. As $\hat{d}_q$ is set to 2, each encoding quantum state is represented by a point in a Bloch sphere.

Case 1: The asymmetric process. For a $\hat{d}_q = 2$ model of the asymmetric process [recall Fig. 2(a)] with $p = 0.3$, $q = 0.8$, we found the Kraus operators:

$$\begin{aligned} A_0 &= \begin{bmatrix} 0.676 & 0.317 \\ 0.316 & 0.150 \end{bmatrix}, \\ A_1 &= \begin{bmatrix} -0.264 & 0.534 \\ -0.336 & 0.729 \end{bmatrix}, \\ A_2 &= \begin{bmatrix} -0.241 & -0.095 \\ 0.449 & 0.227 \end{bmatrix}. \end{aligned} \tag{E1}$$

For the asymmetric processes, real amplitudes are sufficient to train a quantum model with $\hat{d}_q = 2$ to good accuracy [32].

The quantum states $|\sigma_{\overleftarrow{x}}\rangle$ associated with encodings of pasts of length five are plotted in Fig. 6. The more likely quantum states lie in three clusters. This coincides with our expectation from the classical ε machine, which has three causal states—but here, this has been discovered directly from the sample during training. Moreover, the location of these states aligns with our expectation from the theoretical optimal quantum states ascertained in Ref. [32]: two clusters (pink and blue) approximately correspond to orthogonal quantum states, while the third (green) lies biased between the two.

Case 2: The quasicycle. For the quasicycle, we show the case where $p = 0.5$, $\delta = 0.1$. The Kraus operators found for $\hat{d}_q = 2$ are

$$A_0 = \begin{bmatrix} 0.021 - 0.077i & -0.104 + 0.084i \\ -0.277 + 0.236i & 0.656 + 0.098i \end{bmatrix},$$

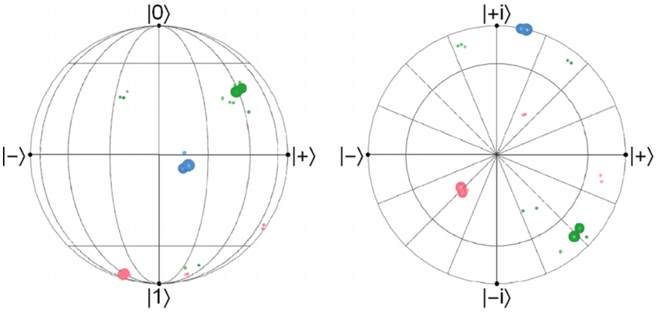


FIG. 7. Example: Quasicycle—encoded states. Quantum states produced by the learned $\hat{d}_q = 2$ encoding map for a quasicycle with $p = 0.5,\ \delta = 0.1$, represented on a Bloch sphere. The diagram is interpreted as in Fig. 6.

$$A_1 = \begin{bmatrix} 0.508 + 0.135i & -0.233 + 0.465i \\ 0.279 - 0.171i & 0.073 + 0.277i \end{bmatrix},$$

$$A_2 = \begin{bmatrix} 0.371 + 0.271i & 0.304 + 0.010i \\ -0.209 + 0.473i & 0.055 + 0.308i \end{bmatrix}. \tag{E2}$$

The quantum states $|\sigma_{\overleftarrow{x}}\rangle$ are plotted in Fig. 7. As shown in Fig. 7, most of the states (weighted by probability of occurrence) lie in three distinct clusters—but here, no pair of clusters is orthogonal.

Case 3: The uniform renewal process. For the discrete renewal process [Fig. 4(a)] we show the results for $N = 3$, $\hat{d}_q = 2$. The Kraus operators found are

$$A_0 = \begin{bmatrix} 0.064 + 0.170i & 0.043 + 0.246i \\ -0.196 + 0.825i & 0.499 - 0.079i \end{bmatrix},$$

$$A_1 = \begin{bmatrix} 0.490 - 0.053i & 0.304 + 0.753i \\ 0.005 - 0.068i & 0.048 - 0.142i \end{bmatrix}. \tag{E3}$$

The associated encoded states are plotted in Fig. 8.

APPENDIX F: BOUNDING THE MINIMAL DISTORTION OF CLASSICAL MODELS

Here we derive Theorem 1, a lower bound on the minimal distortion of classical models with a given memory dimension $\hat{d}_c$. In principle, such a bound could be found by exhaustive search over all permitted classical models, but this is not tractable since the freedom in the transition probabilities grows as $O(\hat{d}_c^{|\mathcal{X}|+1})$, where $|\mathcal{X}|$ is the alphabet size. We solve this problem by proving that an intuitively natural method for creating a tractable set of approximate models is optimal in the sense of realizing minimal distortion: clump together some of d_c original states to create a model of memory size $\hat{d}_c < d_c$. There is a subtlety in that this clumping together, a type of coarse-graining, may remove the unifilarity of the model, such that it may be in a wider set of models we shall call *premodels*. The task then becomes to prove that the minimal distortion, over all premodels, of the original model, can be realized by a coarse-grained model, as depicted in Fig. 9.

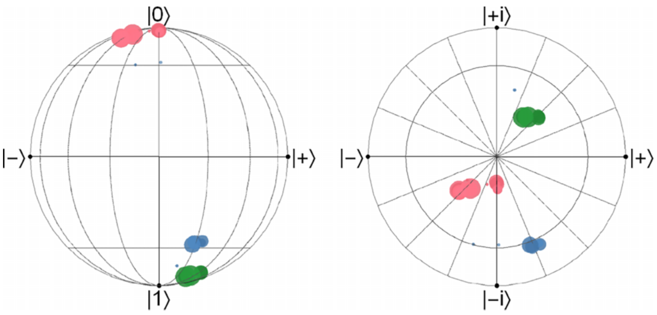


FIG. 8. Example: Renewal process—encoded states. Quantum states produced by the learned $\hat{d}_q = 2$ encoding map for an $N = 3$ discretization of the uniform renewal process, represented on a Bloch sphere. The diagram is interpreted as in Fig. 6.

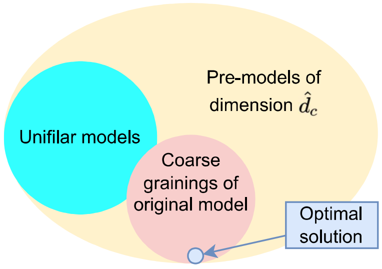

Pre-models of dimension $\hat{d}_c$
Unifilar models
Coarse grainings of original model
Optimal solution

FIG. 9. Proof overview. Sets of models of a given memory dimension $\hat{d}_c$ are depicted. In a *premodel* the internal state update is not generally defined. If the memory state is updated entirely as a function of the output and the memory state, the model is termed *unifilar*. An exact original model with dimension $d_c > \hat{d}_c$ can be coarse-grained by clumping together memory states, to create a lower-dimensional model. We prove that the set of coarse-grained models contains models realizing minimal distortion relative to the original model (for the given $\hat{d}_c$).

The derivation of Theorem 1 is structured into three parts: (1) Model definitions, (2) Calculating the distortion, and (3) Bounding the minimum distortion. Finally, we make additional stronger statements for a special case, showing in particular that for Markov processes the coarse-grained processes are unifilar.

1. Model definitions

We now give relevant definitions of models which shall be used in the proof.

Let $\overleftarrow{X}$ be the set of all possible observable histories of some process P, and $\mathcal{R}$ be a countable set of possible model memory states. A *partition of histories* or *encoding* $\mathcal{E}$ is then a function $\mathcal{E} : \overleftarrow{X} \to \mathcal{R}$. We refer to the number of distinct values in $\mathcal{R}$ taken by $\mathcal{E}$ [excluding those that occur with measure zero with respect to $P(\overleftarrow{X})$] as the model *dimension* of $\mathcal{E}$, which we write as $\hat{d}_c$ for approximate classical models and d_c for exact classical models. That is, each $\mathcal{E}$ effectively divides the set of histories into $\hat{d}_c$ mutually exclusive partitions. To use this encoding as a model, we must also supply a second map $\mathcal{F} : \mathcal{R} \to \overrightarrow{X}$ from memory states to a set of statistics over the future of the process known. The $\mathcal{F}$s are termed the *future morphs* of the memory states because they represent the “shape” (morph in Greek) of the future [8]. Together, the composition of $\mathcal{E}$ and $\mathcal{F}$ defines a conditional probability distribution $\hat{P}(\overrightarrow{X}|\mathcal{E}(\overleftarrow{x}))$, of the model which emulates the process's conditional probability distribution $P(\overrightarrow{X}|\overleftarrow{x})$ for each history $\overleftarrow{x}$. Finally, a model may also have a rule enabling running a simulation over several time steps: an action $\mathcal{P}$ that updates the memory state conditional on the output and the last memory state. It is often demanded that the action $\mathcal{P}$ is time-independent such that the simulation corresponds to a stationary process.

We shall find it convenient to also consider "underdefined" models, such that brute force searches over models become more tractable. In particular we term models defined via the couple $(\mathcal{E}, \mathcal{F})$ alone (without necessarily specifying the action $\mathcal{P}$) *premodels* $\mathcal{M}$. Such premodels arise naturally when considering coarse-grainings of exact models. As an example, a predictive model that has no distortion if one uses three causal states (s_1, s_2, s_3), may be approximated by a model with just two causal states, such as (s_1, s_2) where s_2 can be the encoded state for the case of the output associated with s_3 in the exact model. Suppose the exact model has a well-defined memory-updating action $\mathcal{P}$; this still leaves many choices for the memory-updating action of the coarse-grained model and one may consider leaving that undefined. We shall be particularly interested in premodels that are coarse-grainings of exact models and output some finite number K of time series outputs in one go, without a need to specify the update of the memory state. We will use the distortion of such premodels on the K outputs as a lower bound to the distortion of any approximate admissible model with the tightness of the bound in general changing with K. Checking the distortion of these coarse-grained premodels involves checking possible coarse-grainings which is much more tractable than considering all memory-updating actions.

For technical reasons we introduce a class of models that are intermediate between the K-output coarse-grained premodels and the unifilar models we are seeking to approximate (recall unifilar models have a unique memory state update for a given output and previous memory state). We term a premodel *K-unifilar* if it has a memory-updating action $\mathcal{P}$ that deterministically decides the updated memory state based on the last memory state and an output of K time steps:

Definition 1. (K-unifilar premodel). For $K \in \mathbb{Z}^+$, a K-unifilar premodel is the triple $(\mathcal{E}, \mathcal{F}, \mathcal{P}_K)$ consisting of a premodel $(\mathcal{E}, \mathcal{F})$ with a deterministic map satisfying the consistency condition $\mathcal{P}_K : \mathcal{R} \times \mathcal{X}^{\otimes K} \to \mathcal{R}$ such that $\mathcal{P}_K(\mathcal{E}(\overleftarrow{x}), x_{0:K}) = \mathcal{E}(\overleftarrow{x}\, x_{0:K})$ for all $x_{0:K}$ and all $\overleftarrow{x}$.

We write the set of all K-unifilar premodels of maximum dimension d as $\mathcal{M}_K^d$. We write the set of premodels with maximum dimension d and infinite future lengths as $\mathcal{M}_\infty^d$.

Crucially, the valid morphs of K-unifilar premodels are highly constrained:

Lemma 1. A K-unifilar premodel $(\mathcal{E}, \mathcal{F}, \mathcal{P}_K)$ can always be specified by the triple $(\mathcal{E}, \mathcal{F}_K, \mathcal{P}_K)$ where $\mathcal{F}_K$ is a map from $\mathcal{R}$ to probability distributions over words of length K.

Proof. First, let $r := \mathcal{E}(\overleftarrow{x})$ and note that the definitions of $\mathcal{P}_K$ and $\mathcal{F}$ imply that $\mathcal{F}(\mathcal{E}(\overleftarrow{x}\, x_{0:K})) = \mathcal{F}(\mathcal{P}_K(\mathcal{E}(\overleftarrow{x}), x_{0:K}))$ allowing for any substitutions of the form (for any $L > K$):

$$P(X_{K:L} = x_{K:L}|r x_{0:K}) = P(X_{0:L-K} = x_{K:L}|\mathcal{P}_K(r, x_{0:K})), \quad \text{(F1)}$$

since if this was not the case then $\mathcal{F}$ would not consistently assign the correct future morph for some memory states.

For notational brevity, we recursively define the set of functions $\{r_n : \mathcal{R} \times \mathcal{X}^{nK} \to \mathcal{R}\}_n$ as

$$r_n(r, x_{0:nK}) = \mathcal{P}_K(r_{n-1}(r, x_{0:(n-1)K}), x_{(n-1)K:nK}) \qquad \text{(F2)}$$

for $n \geqslant 1$, and $r_0 = r$. For any $M \in \mathbb{Z}^+$, we can thus expand the probability distribution

$$\begin{aligned} &P(X_{0:MK} = x_{0:MK}|r)\\ &\quad = P(X_{0:K} = x_{0:K}|r)P(X_{K:KM} = x_{K:MK}|r x_{0:K})\\ &\quad = P(X_{0:K} = x_{0:K}|r_0)P(X_{0:(M-1)K} = x_{K:MK}|r_1(r, x_{0:K}))\\ &\quad = P(X_{0:K} = x_{0:K}|r_0)P(X_{0:K} = x_{K:2K}|r_1(r, x_{0:K}))\\ &\qquad \times P(X_{K:(M-1)K} = x_{2K:MK}|r_1(r, x_{0:K})x_{K:2K})\\ &\quad = P(X_{0:K} = x_{0:K}|r_0)P(X_{0:K} = x_{K:2K}|r_1(r, x_{0:K}))\\ &\qquad \times P(X_{0:(M-2)K} = x_{2K:MK}|r_2(r, x_{0:2K}))\\ &\quad = \cdots\\ &\quad = \prod_{i=0}^{M-1} P(X_{0:K} = x_{iK:(i+1)K}|r_i(r, x_{0:iK})). \qquad \text{(F3)} \end{aligned}$$

Here, Bayesian expansion allows us to make the first and third equalities, and K unifilarity [via Eqs. (F1) and (F2)] allows us to make the substitutions for the second and fourth equalities. Thus, we can use this to generate a future morph for words of any length KM as a function of the probability distribution over the next K symbols. (Probabilities over words of length that are not multiples of K can always be found by taking marginals of a longer word that is a multiple of K.)

The contrapositive implies that if $\mathcal{F}$ does not assign probabilities in this way, then it cannot satisfy Eq. (F1), and hence is not a K-unifilar premodel. ■

An immediate corollary of Lemma 1 is that a one-unifilar premodel is exactly a unifilar hidden Markov model. It also immediately follows that any K-unifilar premodel for $K \in \mathbb{Z}^+$ is also an NK-unifilar premodel for $N \in \mathbb{Z}^+$—and, particularly, unifilar models are K-unifilar premodels for all $K \in \mathbb{Z}^+$. For any maximum dimension d, this sets up the hierarchy $\mathcal{M}_1^d \subseteq \mathcal{M}_K^d \subseteq \mathcal{M}_\infty^d$.

2. Calculating the distortion D_e

In this section we derive certain lemmas relating different quantifications of distortion that we shall use in the proof of the bound.

Consider two processes P and $\hat{P}$, and for calculational convenience, define the arguments within the limit of Eqs. (2) and (3) as

$$D_L(P, \hat{P}) = \sum_{\overleftarrow{x}} P(\overleftarrow{x}) D_L(P, \hat{P}|\overleftarrow{x}), \qquad \text{(F4)}$$

such that Eq. (3) may be written as

$$D_e(P, \hat{P}) = \lim_{L\to\infty} D_L(P, \hat{P}). \qquad \text{(F5)}$$

Lemma 2. For stationary processes P and $\hat{P}$:

$$D_e(P, \hat{P}) = D_K(P, \hat{P}) = D_1(P, \hat{P})\ \forall\ K \in \mathbb{Z}^+. \qquad \text{(F6)}$$

Proof. According to Bayesian rules, for $L > 1$:

$$P(x_{0:L}|\overleftarrow{x}) = P(x_0|\overleftarrow{x})P(x_{1:L}|\overleftarrow{x}x_0). \tag{F7}$$

Then (expanding the definition of KL divergence),

$$\begin{aligned} LD_L(P,\hat{P}) &= \sum_{\overleftarrow{x}} P(\overleftarrow{x}) \sum_{x_{0:L}} P(x_{0:L}|\overleftarrow{x}) \log_2 \frac{P(x_{0:L}|\overleftarrow{x})}{\hat{P}(x_{0:L}|\overleftarrow{x})} \\ &= \sum_{\overleftarrow{x}} P(\overleftarrow{x}) \sum_{x_0} P(x_0|\overleftarrow{x}) \log_2 \frac{P(x_0|\overleftarrow{x})}{\hat{P}(x_0|\overleftarrow{x})} \\ &\quad + \sum_{\overleftarrow{x}} P(\overleftarrow{x}) \sum_{x_{0:L}} P(x_{0:L}|\overleftarrow{x}) \log_2 \frac{P(x_{1:L}|\overleftarrow{x}x_0)}{\hat{P}(x_{1:L}|\overleftarrow{x}x_0)}. \end{aligned} \tag{F8}$$

Since P and $\hat{P}$ are stationary, we can substitute $P(x_{1:L}|\overleftarrow{x}x_0) = P(x_{0:L-1}|\overleftarrow{x})$, and hence for $L > 1$:

$$LD_L(P,\hat{P}) = D_1(P,\hat{P}) + (L-1)D_{L-1}(P,\hat{P}). \tag{F9}$$

Then, inductively

$$D_L(P,\hat{P}) = D_1(P,\hat{P})\ \forall\ L \geqslant 1, \tag{F10}$$

and hence also the limit

$$D_e(P,\hat{P}) = \lim_{L\to\infty} D_L(P,\hat{P}) = D_1(P,\hat{P}). \tag{F11}$$

■

Similarly, this enables a shortcut to calculate the distortion between a process and a K-unifilar premodel:

Lemma 3. Let P be some stationary process, and $(\mathcal{E}, \mathcal{F}_K, \mathcal{P}_K)$ be a K-unifilar premodel whose future morph is generated by $\hat{P}(X_{0:K}|\mathcal{E}(\overleftarrow{x}))$. Then,

$$D_e(P,\hat{P}) = D_K(P,\hat{P}). \tag{F12}$$

Proof. We can group each word of K contiguous symbols from process P over alphabet $\mathcal{X}$ into a single symbol of process Q with alphabet $\mathcal{Y} := \mathcal{X}^K$, such that the distributions over the two processes are related by

$$Q(\ldots, y_0, y_1, \ldots) = P(\ldots, x_{0:R}, x_{R:2R}, \ldots). \tag{F13}$$

We similarly group $\hat{P}$ into $\hat{Q}$. Since $(\mathcal{E}, \mathcal{F}, \mathcal{P}_K)$ is a K-unifilar premodel for $\hat{P}$, it immediately defines a unifilar model for $\hat{Q}$ with encoding $\mathcal{E}_Q : \overleftarrow{Y} \to \mathcal{R}$ satisfying $\mathcal{E}_Q(\overleftarrow{y}) = \mathcal{E}(\overleftarrow{x})$ where $\overleftarrow{y}$ is the grouping of $\overleftarrow{x}$. Such combination has no effect on the KL divergence and hence the distortion satisfies

$$D_e(Q,\hat{Q}) = KD_e(P,\hat{P}). \tag{F14}$$

Meanwhile, using Lemma 2,

$$\begin{aligned} D_e(Q,\hat{Q}) &= D_1(Q,\hat{Q}) \\ &= \sum_{\overleftarrow{y}} Q(\overleftarrow{y}) D_1(Q,\hat{Q}|\overleftarrow{y}) \\ &= \sum_{\overleftarrow{x}} P(\overleftarrow{x}) KD_K(P,\hat{P}|\overleftarrow{x}) \\ &= KD_K(P,\hat{P}). \end{aligned} \tag{F15}$$

And hence $D_e(P,\hat{P}) = D_K(P,\hat{P})$. ■

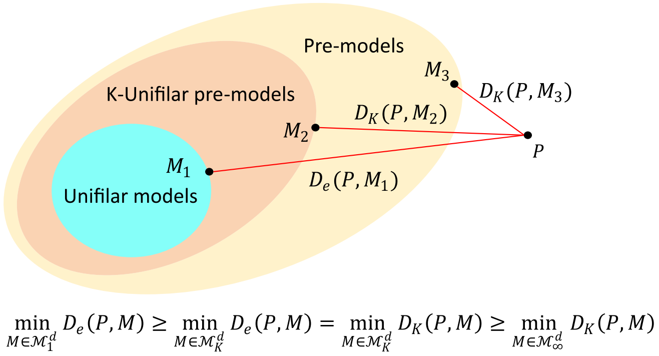

Pre-models
K-Unifilar pre-models
Unifilar models
M_1
M_2
M_3
P
$D_K(P, M_3)$
$D_K(P, M_2)$
$D_e(P, M_1)$
$\min_{M\in\mathcal{M}_1^d} D_e(P,M) \geq \min_{M\in\mathcal{M}_K^d} D_e(P,M) = \min_{M\in\mathcal{M}_K^d} D_K(P,M) \geq \min_{M\in\mathcal{M}_\infty^d} D_K(P,M)$

FIG. 10. P represents the given stochastic process. For a given model dimension d, $\mathcal{M}_\infty^d$ represent all premodels, $\mathcal{M}_K^d$ represents all K-unifilar premodels, and $\mathcal{M}_1^d$ represents all unifilar models.

3. Bounding the minimum distortion

We now build further on the definitions and lemmas of the two previous sections to show that coarse-grained (pre-) models can indeed achieve a distortion that lower-bounds that of any other permissible classical model of the given memory dimension. The proof is illustrated in Fig. 10.

A consequence of the above lemmas (Lemmas 2 and 3) is to make it easier to calculate bounds that minimize K-unifilar premodels and unifilar models:

Lemma 4. Let P be a stationary process and let $K, N \in \mathbb{Z}^+$. Then, for every dimension d,

$$\min_{M\in\mathcal{M}_K^d} D_e(P,M) \geqslant \min_{M\in\mathcal{M}_{KN}^d} D_K(P,M). \tag{F16}$$

Proof. First, Lemma 3 tells us that for any K-unifilar premodel $D_e(P,M) = D_K(P,M)$ and hence $\min_{M\in\mathcal{M}_K} D_e(P,M) = \min_{M\in\mathcal{M}_K} D_K(P,M)$. Next, the hierarchy of models $\mathcal{M}_K^d \subseteq \mathcal{M}_{KN}^d$, and hence a minimization over $\mathcal{M}_{KN}^d$ lower-bounds a minimization over $\mathcal{M}_K^d$. This proves Eq. (F16). ■

Why might we choose some finite number of outputs K rather than minimizing the distortion D_e over generic ($K = \infty$) premodels, i.e., the set $\mathcal{M}_\infty$? $\mathcal{M}_\infty$ is so permissive that it is possible to achieve no distortion. Consider any (stationary, aperiodic) process, and let us assign a one-dimensional model, but whose single future morph defines probabilities arbitrarily far into the future, with statistics that increasingly match those of P (e.g., by assigning a weighted average of P's causal state's future morphs). For such a (highly nonunfilar) model, $D_e \to 0$, saturating the arithmetically obvious bound. Second, the parameter space of $P(\overrightarrow{X}|\mathcal{E}(\overleftarrow{x}))$ is *a priori* infinite, and the above lemmas allow us to greatly reduce the relevant parameter space of future morphs in our calculations to probability distributions over words of finite length K.

Although the above lemmas (Lemma 4 in particular) allow us to reduce the dimensionality of $\mathcal{F}$ assigned to each memory state, there still remains the problem that there are infinitely many possible partitions of the past $\mathcal{E}$. We will argue, with the next set of lemmas, that for exploring minimum bounds, we can restrict ourselves to the finite set of coarse-grainings over the causal states.

First, we note that the problem of finding a minimum distortion model is trivial if we allow models with high enough dimension:

Lemma 5. For a process P with topological complexity d_c, then for all model dimensions $\hat{d}_c \geqslant d_c$, there is a zero-distortion unifilar model M such that

$$D_K(P, M) = 0 \,\forall\, K \in \mathbb{Z}^+. \tag{F17}$$

Proof. Such a minimum distortion model can be realized by the ε machine of P, which has zero distortion. By definition of causal states, $P(X_{0:K}|\overleftarrow{x}) = P(X_{0:K}|\varepsilon(\overleftarrow{x}))$ for all K and $\overleftarrow{x}$, and hence $\mathcal{D}_{\mathrm{KL}}(P(X_{0:K}|\overleftarrow{x})||P(X_{0:K}|\varepsilon(\overleftarrow{x}))) = 0$ for all L and $\overleftarrow{x}$. This implies that all $D_K(P, \hat{P}) = 0$. ■

Lemma 6. For any process P, for every dimension d, and for every $K \in \mathbb{Z}^+$, if two pasts $\overleftarrow{y}$ and $\overleftarrow{z}$ have identical future morphs in P, then there exists a premodel $(\mathcal{E}^d_{\min}, \mathcal{F})$ that minimizes D_K and satisfies

$$\mathcal{E}^d_{\min}(\overleftarrow{y}) = \mathcal{E}^d_{\min}(\overleftarrow{z}). \tag{F18}$$

Proof. If $d \geqslant d_c$, the ε machine has the minimum distortion (see Lemma 5) and the causal state partition ε satisfies Eq. (F18) by definition. Likewise, if $d_c = 1$, there is only one encoding onto a single state and so Eq. (F18) is trivially satisfied. Thus, it remains only to prove the cases where $1 < d < d_c$.

Suppose $\overleftarrow{y}$ and $\overleftarrow{z}$ have identical future morphs in P, and we have a premodel $(\mathcal{E}, \mathcal{F})$ where encoding $\mathcal{E}(\overleftarrow{y}) \neq \mathcal{E}(\overleftarrow{z})$. Then without loss of generality (by switching the labels of $\overleftarrow{y}$ and $\overleftarrow{z}$ if necessary) the implied statistics $\hat{P}$ satisfy

$$D_K(P, \hat{P}|\overleftarrow{y}) \leqslant D_K(P, \hat{P}|\overleftarrow{z}), \tag{F19}$$

where $D_K(P, \hat{P}|\overleftarrow{x})$ is defined in Eq. (F4).

We can then construct a new premodel with identical $\mathcal{F}$, but with a new mapping $\mathcal{E}'$ identical to $\mathcal{E}$ in every way, except it now maps $\overleftarrow{z}$ to $\mathcal{E}(\overleftarrow{y})$ instead of $\mathcal{E}(\overleftarrow{z})$. This has implied statistics $\hat{P}'$, and

$$D_K(P, \hat{P}'|\overleftarrow{z}) = D_K(P, \hat{P}|\overleftarrow{y}) \leqslant D_K(P, \hat{P}|\overleftarrow{z}). \tag{F20}$$

Hence D_K of $(\mathcal{E}', \mathcal{F})$ satisfies

$$\begin{aligned}
&D_K(P, \hat{P}') \\
&\quad= P(\overleftarrow{z})D_K(P, \hat{P}'|\overleftarrow{z}) + \sum_{\overleftarrow{x}\neq\overleftarrow{z}} P(\overleftarrow{x})D_K(P, \hat{P}'|\overleftarrow{x}) \\
&\quad= P(\overleftarrow{z})D_K(P, \hat{P}'|\overleftarrow{z}) + \sum_{\overleftarrow{x}\neq\overleftarrow{z}} P(\overleftarrow{x})D_K(P, \hat{P}|\overleftarrow{x}) \\
&\quad\leqslant P(\overleftarrow{z})D_K(P, \hat{P}|\overleftarrow{z}) + \sum_{\overleftarrow{x}\neq\overleftarrow{z}} P(\overleftarrow{x})D_K(P, \hat{P}|\overleftarrow{x}) \\
&\quad= D_K(P, \hat{P}).
\end{aligned} \tag{F21}$$

It then follows that for any premodel where $\overleftarrow{y}$ and $\overleftarrow{z}$ with identical future morphs map to different memory states, there is another premodel of the same (or lower) dimension such that $\overleftarrow{y}$ and $\overleftarrow{z}$ map to the same memory state, and this premodel has the same or lower distortion. Hence, among the minimum distortion premodels of a given process, there will always be a premodel where $\overleftarrow{y}$ and $\overleftarrow{z}$ map to the same memory state. ■

The above lemma implies the following:

Lemma 7. For a process P with topological complexity d_c, for every dimension $d < d_c$, and every $K \in \mathbb{Z}^+$ there is a premodel M that minimizes $D_K(P, M)$ whose encoding $\mathcal{E}^d_{\min}$ is a map onto a coarse graining of the causal states $\mathcal{S}$.

Proof. This follows from Lemma 6 by noting that pasts in the same causal state have the same future morph by definition. ■

Such coarse-grained models admit a computational shortcut for calculating their distortion: we can define $\pi_i := \sum_{\overleftarrow{x}\in s_i} P(\overleftarrow{x})$ for each causal state $s_i \in \mathcal{S}$, such that the distortion is

$$D_K(P, \hat{P}) = \sum_{i=1}^{d_c} \pi_i D_K(P, \hat{P}|s_i), \tag{F22}$$

where for each s_i, $D_K(P, \hat{P}|s_i) := D_K(P, \hat{P}|\overleftarrow{x})$ for one arbitrary choice of $\overleftarrow{x} \in s_i$ (since by definition the value is the same for all such choices).

Crucially, the above lemma enables an exhaustive search for D_K-optimal premodels for any process with finite topological complexity—instead of having to consider an infinite number of possible partitions, we can iterate through the various combinations of causal states.

Lemma 8. For a process P, for every dimension d, and every $K \in \mathbb{Z}^+$ there is premodel $M = (\mathcal{E}^d_{\min}, \mathcal{F}^d_{\min}) \in M^d_\infty$ that minimizes $D_K(P, M)$ such that every memory state r_i whose entire pre-image [except for a measure zero subset with respect to $P(\overleftarrow{x})$] is mapped to the causal state $s_{i'}$, has the future morph $\mathcal{F}^d_{\min}(r_i) = P(\overrightarrow{X}|s_{i'})$.

Proof. Again, we consider the case where model dimension $d < d_c$ since the case $d \geqslant d_c$ is trivially satisfied by the ε machine (Lemma 5). From Lemma 7 we may restrict ourselves to coarse-grainings of ε machines.

Let s_i be the causal states of the process, indexed such that $i = 1, \dots, k$ correspond to states that are not merged, and $i = k+1, \dots, d_c$ are merged into states $r_1, \dots, r_{d-k}$ (i.e., the states of the model are $\mathcal{R} = \{s_1, \dots, s_k, r_1, \dots, r_{d-k}\}$). As the lemma's claim is vacuously true if $k = 0$, we consider the cases when $k \geqslant 1$. Then, using the coarse-graining structure, and choosing an arbitrary past $\overleftarrow{x}_i \in s_i$ for each causal state,

$$D_K(P, \hat{P}) = \sum_{i=1}^{k} \pi_i D_K(P, \hat{P}|\overleftarrow{x}_i) + \sum_{i=k+1}^{d} \pi_i D_K(P, \hat{P}|\overleftarrow{x}_i). \tag{F23}$$

If for any $i \in [1, k]$ we have $D_K(P, \hat{P}|\overleftarrow{x}_i) > 0$, we can instead form a new premodel that assigns the future morph $\mathcal{F}(s_i) = P(\overrightarrow{X}|s_i)$. Then, in this new premodel $D_k(P, \hat{P}'|\overleftarrow{x}_i) = 0$ for all such s_i. Thus,

$$D_K(P, \hat{P}') = \sum_{i=k+1}^{d} \pi_i D_K(P, \hat{P}|s_i) \leqslant D_K(P, \hat{P}). \tag{F24}$$

■

In simpler words: when we look for a D_K-optimal premodel by merging states of an ε machine, we should not alter the future morphs of any "unmerged" states from how they are in the ε machine.

Theorem 1. Let P be any stationary process. Then, for every model dimension $\hat{d}_c$, the algorithm in Box 1 produces a bound on the minimum distortion D_e of a unifilar model by considering all possible mergers of causal states, and searching over the next K output statistics.

Proof. Lemma 7 tells us that a D_K-optimal premodel exists that is formed by merging causal states but says nothing as to whether such a premodels is unifilar. However, Lemma 4 implies that such a D_K-optimized premodel bounds the D_e–optimized unifilar models. Moreover, D_K, as a function of $\hat{P}(X_{0:K}|\mathcal{E}(\overleftarrow{x}))$, depends only on the next K output statistics. As such, to find this minimum bound, we can iterate through the various coarse-grainings of causal states of a given dimension and search through finite parameter space of $\hat{P}$ to minimize D_K. The minimum found here will lower bound the lowest value of D_e achievable by a unifilar model. ■

4. Minimum-distortion approximations of Markov processes

For Markov processes, the bound of Theorem 1 is tight. First:

Lemma 9. Let P be a process with Markov order κ. Then, every encoding map $\mathcal{E}$ formed by coarse-graining the causal states of P admits a κ-unifilar premodel. At least one such κ-unifilar premodel minimizes $D_\kappa(P, \hat{P})$.

Proof. Let ε be the encoding map onto causal states, and $\mathcal{C} : \mathcal{S} \to \mathcal{R}$ be the coarse–graining map from causal states $\mathcal{S}$ to memory states $\mathcal{R}$, such that $\mathcal{E} = \mathcal{C} \circ \varepsilon$. We must demonstrate that there exists a $\mathcal{P}_\kappa$ such that $\mathcal{P}_\kappa(\mathcal{E}(\overleftarrow{x}), x_{0:\kappa}) = \mathcal{E}(\overleftarrow{x}x_{0:\kappa})$ for all $\overleftarrow{x}$ and all $x_{0:\kappa}$. Since P has Markov order κ, there exists an $\tilde{\varepsilon} : \mathcal{X}^{\otimes\kappa} \to \mathcal{S}$ that acts on only the final κ symbols of the history to identify the causal states, and for every $\overleftarrow{x}$ that ends in the same $x_{-\kappa:0}$, $\varepsilon(\overleftarrow{x}) = \tilde{\varepsilon}(x_{-\kappa:0})$. We thus have an $\tilde{\mathcal{E}} = \mathcal{C} \circ \tilde{\varepsilon}$ that maps to the same memory state as $\mathcal{E}$ for every $x_{-\kappa:0}$ and every $\overleftarrow{x}$ that ends in $x_{0:\kappa}$.

Thus, now consider $\mathcal{P}_\kappa(\mathcal{E}(\overleftarrow{x}), x_{0:\kappa}) = \mathcal{E}(\overleftarrow{x}x_{0:\kappa}) = \tilde{\mathcal{E}}(x_{0:\kappa})$. A candidate $\mathcal{P}_\kappa$ can be defined such that it is completely independent of its first argument (the current machine state $r \in \mathcal{R}$) and instead takes exactly the same value as $\tilde{\mathcal{E}}$ applied to its second input (the recently output word $x_{0:\kappa} \in \mathcal{X}^\kappa$). Thus, there exists a κ-unifilar premodel for any such $\mathcal{E}$.

Now, from Lemma 7, we know that the D_κ optimal (∞-unifilar) premodel $(\mathcal{E}, \mathcal{F})$ is formed by merging causal states, but *a priori* we have not demonstrated that such a model is κ-unifilar. To calculate D_κ, we need only evaluate the probabilities associated with the first κ outputs. Thus, we can define $\mathcal{F}_\kappa$ as per Lemma 1 such that $\mathcal{F}_\kappa$ and $\mathcal{F}$ perfectly agree on the statistics of the first word of length κ in the morph. Then, taking the encoding map of $\mathcal{E}$ from this premodel, we form a κ-unifilar model $(\mathcal{E}, \mathcal{F}_\kappa, \mathcal{P}_\kappa)$ where $\mathcal{P}_\kappa$ is defined as above. Since D_κ only depends on the first κ steps of the future, minimizing the value of $D_\kappa(P, \hat{P})$ provides the optimal κ-unifilar premodel. As such, we have formed a minimum-distortion κ-unifilar premodel. ■

When the process is Markovian, these lemmas then give us a systematic method for finding the minimum distortion unifilar hidden Markov model.

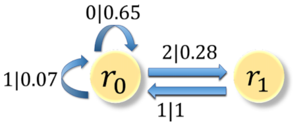


FIG. 11. Approximate model for asymmetric process. The approximate classical model for the asymmetric example with $p = 0.3$ and $q = 0.8$. s_0 and s_1 of the original asymmetric process are merged into r_0, with transition probabilities identified by a minimization. s_2 is mapped to r_1 with an unchanged future morph.

Theorem 2. Let P be a Markov process. Then, for every model dimension d, we can find a minimum distortion unifilar model by merging causal states, and searching over the next–output statistics.

Proof. If $d \geqslant d_c$, this encoding is the ε machine (Lemma 5). In the case $d < d_c$, we use Lemma 7, to find a $\mathcal{E}^d_{\min}$ that is a coarse–graining of causal states. Then Lemma 9 specialized to $\kappa = 1$ implies that this model can be made unifilar. ■

APPENDIX G: CLASSICAL MINIMUM DISTORTION MODELS: EXAMPLES

As illustrative examples, we present the minimum–distortion approximate classical model for the asymmetric process with $p = 0.3$, $q = 0.8$ in Fig. 11 and for quasicycle with $p = 0.5$, $\delta = 0.1$ in Fig. 12.

For the third example, we evaluate D_N as a lower bound on the error of the optimal classical model for any discrete renewal process with N states, as shown in Fig. 4(b).

APPENDIX H: IMPLEMENTATION OF QUANTUM MODELS FOR THE DISCRETE RENEWAL PORCESS ON IBM's DEVICES

To investigate the real-world performance of the models found by the algorithm, we implement one of the discovered quantum models for the discrete renewal process on IBM's superconducting cloud quantum computer. We consider the case of $\hat{d}_q = 2$, $N = 3$ and decompose the quantum model into a quantum circuit consisting of single qubit gates and CNOT gates, as shown in Fig. 13. The IBM device used was "ibmq athens." We ran the circuit 40 000 times for each input quantum state and obtained the output probabilities by measuring the output register.

Figure 5 compares the distortion realized on the IBM device with the optimal classical bound for $\hat{d}_c = 2$ and the ideal realization of the learned quantum model. The 40 000 runs

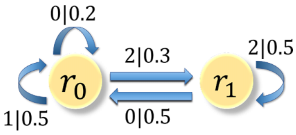


FIG. 12. Approximate model for quasicycle. The approximate classical model for the quasicycle example with $p = 0.5$ and $\delta = 0.1$. s_0 and s_1 of the original quasicycle process are merged into r_0. The transition probability is minimized by going through the probability vector space. s_2 is mapped to r_1 with an unchanged future morph.

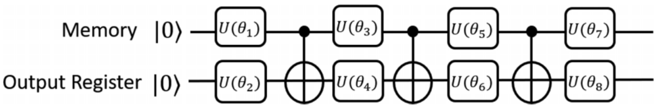


FIG. 13. Circuit diagram for a single step output. The upper line represents the memory qubit while the lower line represents the output register. The circuit involves eight single qubit gates $[U(\theta_i)]$ and three CNOT gates. Each single qubit gate is specified by three real parameters. $U(\theta_1)$ encodes the input memory state.

were divided into 50 batches, and the distortion was calculated for each batch. The error bar depicts the standard deviation in distortion of those 50 batches. Our quantum model maintains a statistically significant accuracy advantage even in a noisy environment: there is a 2.12 standard deviations gap between the experimental quantum distortion and our lower bound on the best classical distortion.

Figure 14 compares the conditional probability distribution realized on "ibmq athens" with the corresponding noiseless distribution of the learned quantum model.

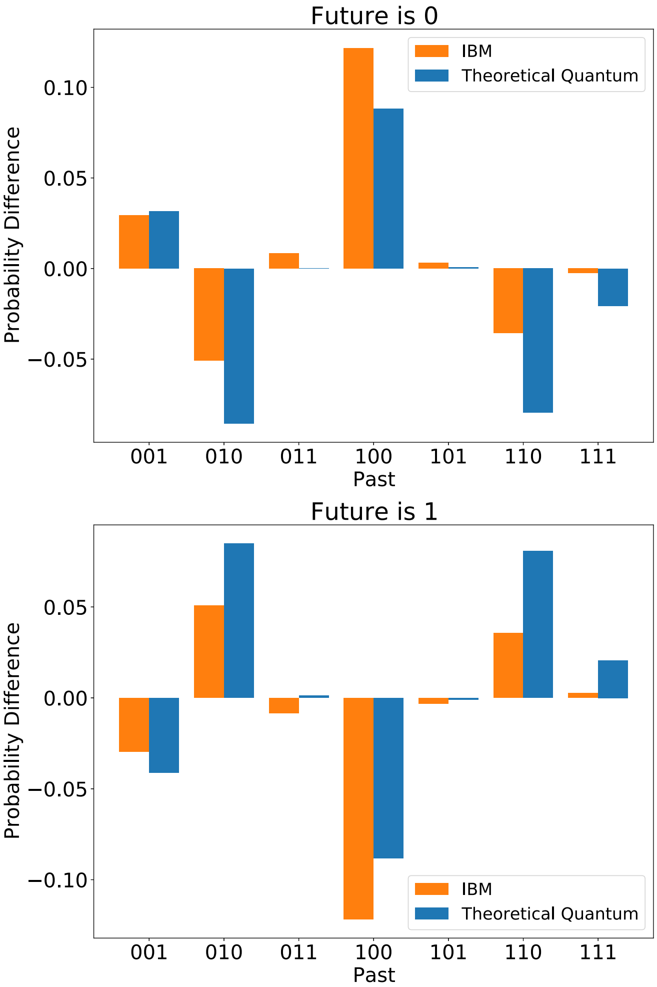


FIG. 14. Comparison of experimental (orange) and theoretical (blue) realizations of the learned quantum model to the true process. Each datum represents the difference between the model's conditional probability for a given past, and the true process's conditional probability.

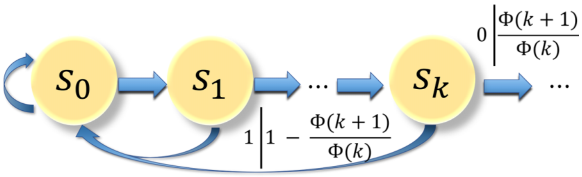

$0 \left| \frac{\Phi(k+1)}{\Phi(k)} \right.$
S_0
S_1
…
S_k
…
$1 \left| 1 - \frac{\Phi(k+1)}{\Phi(k)} \right.$

FIG. 15. ε machine for dual Poisson processes.

APPENDIX I: DUAL POISSON PROCESSES

Here we consider a process whose quantum models outperform classical models. Consider a system that consists of two Poissonian channels with rates γ_1 and γ_2. At any time, only one of the two channels is turned on, undergoing Poissonian decay events. Once an event occurs, both of the two channels are turned off. Then the system randomly turns on one of the two channels with probability p and $1-p$, respectively. An observer measures the system at discrete time intervals Δt, recording a 1 when an event occurs, and 0 otherwise. The series of zeros and ones form a stochastic process, called a dual Poisson process. Many important features of dual Poisson processes can be characterized by the so-called survival probability $\Phi(k)$, which is the probability of observing a contiguous string of zeros (book-ended by ones) of at least length k:

$$\Phi(k) = pq_1^k + (1-p)q_2^k, \tag{I1}$$

where $q_i = \exp(-\gamma_i \Delta t)$.

The ε machine of the dual Poisson process consists of infinitely many causal states $\{S_k\}$, as shown in Fig. 15. Each causal state S_k represents the class of the pasts $\overleftarrow{x}$, which contains k contiguous zeros since the last one when tracking back the past $\overleftarrow{x}$ (e.g., $S_2 = \{\cdots 100\}$). The transition probability between causal states S_k can then be induced from the survival probability,

$$\begin{aligned} P(S_{k+1}, 0 \mid S_k) &= \frac{\Phi(k+1)}{\Phi(k)}, \\ P(S_0, 1 \mid S_k) &= 1 - \frac{\Phi(k+1)}{\Phi(k)}. \end{aligned} \tag{I2}$$

Quantum models exhibit unbounded memory advantages for exact simulation of the dual Poisson processes [17]. Specifically, quantum models achieve such a significant memory compression by encoding the causal states S_i into quantum states $|\sigma_i\rangle$, which lie in a Hilbert space with dimension two:

$$|\sigma_k\rangle = \frac{\sqrt{pq_1^k} + ig\sqrt{\bar{p}q_2^k}}{\sqrt{\Phi(k)}}|0\rangle + \frac{i\sqrt{(1-g^2)\bar{p}q_2^k}}{\sqrt{\Phi(k)}}|1\rangle, \tag{I3}$$

where

$$\bar{p} = 1-p, \quad g = \frac{\sqrt{(1-q_1)(1-q_2)}}{1-\sqrt{q_1 q_2}}. \tag{I4}$$

There exists unitary operator U that carries out the transition dynamics between quantum states $|\sigma_i\rangle$

$$U|\sigma_k\rangle|0\rangle = \sqrt{\frac{\Phi(k+1)}{\Phi(k)}}|\sigma_{k+1}\rangle|0\rangle + \sqrt{1-\frac{\Phi(k+1)}{\Phi(k)}}|\sigma_0\rangle|1\rangle. \tag{I5}$$

When the memory resource is limited, any classical models will inevitably introduce distortion whereas the simulations of quantum models are exact.

[1] M. Verleysen and D. François, The curse of dimensionality in data mining and time series prediction, in *International Work-Conference on Artificial Neural Networks* (Springer, 2005), pp. 758–770.

[2] L. V. D. Maaten, E. Postma, and J. V. den Herik, Dimensionality reduction: A comparative, J. Mach. Learn. Res. **10**, 13 (2009).

[3] E. Alpaydin, *Introduction to Machine Learning* (MIT Press, Cambridge, Massachusetts, 2020).

[4] A. Ng *et al.*, Sparse autoencoder, Stanford University, CS294A Lecture notes **72**, 1 (2011).

[5] L. Banchi, J. Pereira, and S. Pirandola, Generalization in quantum machine learning: A quantum information standpoint, PRX Quantum **2**, 040321 (2021).

[6] C. Blank, D. K. Park, and F. Petruccione, Quantum-enhanced analysis of discrete stochastic processes, npj Quantum Inf. **7**, 126 (2021).

[7] S. Woerner and D. J. Egger, Quantum risk analysis, npj Quantum Inf. **5**, 15 (2019).

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

[9] R. N. Muñoz, A. Leung, A. Zecevik, F. A. Pollock, D. Cohen, B. V. Swinderen, N. Tsuchiya, and K. Modi, General anesthesia reduces complexity and temporal asymmetry of the informational structures derived from neural recordings in drosophila, Phys. Rev. Res. **2**, 023219 (2020).

[10] J. B. Park, J. W. Lee, J. S. Yang, H. H. Jo, and H. T. Moon, Complexity analysis of the stock market, Phys. A (Amsterdam, Neth.) **379**, 179 (2007).

[11] A. J. Palmer, C. W. Fairall, and W. A. Brewer, Complexity in the atmosphere, IEEE Trans. Geosci. Remote Sens. **38**, 2056 (2000).

[12] C. Aghamohammadi and J. P. Crutchfield, Minimum memory for generating rare events, Phys. Rev. E **95**, 032101 (2017).

[13] A. B. Boyd, D. Mandal, and J. P. Crutchfield, Thermodynamics of Modularity: Structural Costs Beyond the Landauer Bound, Phys. Rev. X **8**, 031036 (2018).

[14] S. E. Marzen and J. P. Crutchfield, Informational and causal architecture of discrete-time renewal processes, Entropy **17**, 4891 (2015).

[15] M. Gu, K. Wiesner, E. Rieper, and V. Vedral, Quantum mechanics can reduce the complexity of classical models, Nat. Commun. **3**, 762 (2012).

[16] C. Yang, F. C. Binder, V. Narasimhachar, and M. Gu, Matrix Product States for Quantum Stochastic Modeling, Phys. Rev. Lett. **121**, 260602 (2018).

[17] T. J. Elliott, C. Yang, F. C. Binder, A. J. P. Garner, J. Thompson, and M. Gu, Extreme Dimensionality Reduction with Quantum Modeling, Phys. Rev. Lett. **125**, 260501 (2020).

[18] F. C. Binder, J. Thompson, and M. Gu, Practical Unitary Simulator for Non-Markovian Complex Processes, Phys. Rev. Lett. **120**, 240502 (2018).

[19] J. R. Mahoney, C. Aghamohammadi, and J. P. Crutchfield, Occam's quantum strop: Synchronizing and compressing classical cryptic processes via a quantum channel, Sci. Rep. **6**, 20495 (2016).

[20] C. Aghamohammadi, S. P. Loomis, J. R. Mahoney, and J. P. Crutchfield, Extreme Quantum Memory Advantage for Rare-Event Sampling, Phys. Rev. X **8**, 011025 (2018).

[21] A. Hobson, *Concepts in Statistical Mechanics* (CRC Press, Boca Raton, Florida, 1987).

[22] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (John Wiley & Sons, 2012).

[23] V. Vedral, The role of relative entropy in quantum information theory, Rev. Mod. Phys. **74**, 197 (2002).

[24] J. P. Crutchfield and K. Young, Inferring Statistical Complexity, Phys. Rev. Lett. **63**, 105 (1989).

[25] R. Haslinger, K. L. Klinkner, and C. R. Shalizi, The computational structure of spike trains, Neural Comput. **22**, 121 (2010).

[26] Mathematically such equivalence classes are governed according to the relation $\overleftarrow{x} \sim \overleftarrow{x}'$ if and only if $P(\overrightarrow{X}|\overleftarrow{x}) = P(\overrightarrow{X}|\overleftarrow{x}')$.

[27] C. R. Shalizi, K. L. Shalizi, and J. P. Crutchfield, An algorithm for pattern discovery in time series, arXiv:cs/0210025.

[28] T. Petrie, Probabilistic functions of finite state Markov chains, Ann. Math. Stat. **40**, 97 (1969).

[29] B. Juang and L. R. Rabiner, A probabilistic distance measure for hidden Markov models, AT&T Tech. J. **64**, 391 (1985).

[30] G. Vidal, Entanglement Renormalization, Phys. Rev. Lett. **99**, 220405 (2007).

[31] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv:1412.6980.

[32] 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**, 031013 (2018).

[33] F. Ghafari, N. Tischler, J. Thompson, M. Gu, L. K. Shalm, V. B. Verma, S. W. Nam, R. B. Patel, H. M. Wiseman, and G. J. Pryde, Dimensional Quantum Memory Advantage in the Simulation of Stochastic Processes, Phys. Rev. X **9**, 041013 (2019).

[34] M. Horodecki and J. Oppenheim, Fundamental limitations for quantum and nanoscale thermodynamics, Nat. Commun. **4**, 2059 (2013).

[35] Q. Liu, T. J. Elliott, F. C. Binder, C. Di Franco, and M. Gu, Optimal stochastic modeling with unitary quantum dynamics, Phys. Rev. A **99**, 062110 (2019).

[36] S. Marzen and J. P. Crutchfield, Informational and causal architecture of continuous-time renewal processes, J. Stat. Phys. **168**, 109 (2017).

[37] M. P. Woods, R. Silva, G. Pütz, S. Stupar, and R. Renner, Quantum clocks are more accurate than classical ones, PRX Quantum **3**, 010319 (2022).

[38] T. J. Elliott, A. J. P. Garner, and M. Gu, Memory-efficient tracking of complex temporal and symbolic dynamics with quantum simulators, New J. Phys. **21**, 013021 (2019).

[39] M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio *et al.*, Variational quantum algorithms, Nat. Rev. Phys. **3**, 625 (2021).

[40] L. K. Grover, Quantum Mechanics Helps in Searching for a Needle in a Haystack, Phys. Rev. Lett. **79**, 325 (1997).

[41] G. Brassard, P. Hoyer, M. Mosca, and A. Tapp, Quantum amplitude amplification and estimation, Contemp. Math. **305**, 53 (2002).

[42] github.com/Yangchengran/LearningQuantumStochasticModellingCode.

[43] R. A. Horn and C. R. Johnson, *Matrix Analysis* (Cambridge University Press, Cambridge, 2012).

[44] H. Maassen and B. Kümmerer, Purification of quantum trajectories, *Lecture Notes-Monograph Series* (2006), pp. 252–261.

[45] S. E. Marzen and J. P. Crutchfield, Nearly maximally predictive features and their dimensions, Phys. Rev. E **95**, 051301(R) (2017).