Measures of distinguishability between stochastic processes

Chengran Yang [1,2,*] Felix C. Binder [3,†] Mile Gu [1,2,4,‡] and Thomas J. Elliott [2,1,§]

[1]*School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371*

[2]*Complexity Institute, Nanyang Technological University, Singapore 637335*

[3]*Institute for Quantum Optics and Quantum Information (IQOQI) Vienna, Austrian Academy of Sciences, Boltzmanngasse 3, 1090 Vienna, Austria*

[4]*Centre for Quantum Technologies, National University of Singapore, 3 Science Drive 2, Singapore 117543*

(Received 19 September 2019; accepted 4 May 2020; published 23 June 2020)

Quantifying how distinguishable two stochastic processes are is at the heart of many fields, such as machine learning and quantitative finance. While several measures have been proposed for this task, none have universal applicability and ease of use. In this article, we suggest a set of requirements for a well-behaved measure of process distinguishability. Moreover, we propose a family of measures, called divergence rates, that satisfy all of these requirements. Focusing on a particular member of this family—the coemission divergence rate—we show that it can be computed efficiently, behaves qualitatively similar to other commonly used measures in their regimes of applicability, and remains well behaved in scenarios where other measures break down.

DOI: 10.1103/PhysRevE.101.062137

I. INTRODUCTION

How alike are the behaviors of two systems? How similar are the trajectories of two stock prices? Much of the physical world can be described as a collection of interacting stochastic processes; understanding how distinguishable two processes are allows us to answer such questions. These problems are of universal relevance—for example, quantifying how closely a model replicates its target has applications in fields such as protein homology [1] and speech recognition [2,3]. Meanwhile, understanding how much external noise or perturbations impact the behavior of a system is a central task in studies of open systems [4], quantum computation [5], and machine learning [6,7].

Many measures have been proposed for this task [8–16]. However, a number of these are founded on measures tailored for quantifying distances between distributions. Though they work well for quantifying distances between finite strings, they typically do not behave well in the context of processes where infinite strings of observational data arise as a process continues to run. Particularly, they fail to quantify *how* different processes are [8,10,12–14]. Others, such as the Kullback-Leibler (KL) divergence [9,11], require intensive computational resources to evaluate and can behave pathologically in seemingly innocuous situations such as processes with different output alphabets. Finally, measures that possess an intrinsic dependence on a representation of a process rather than the process itself will often fail by misidentifying different models of the same process with identical observable behavior as being distinguishable [16,17].

In this article, we suggest several requirements that a measure of process distinguishability should satisfy. We then propose a family of measures that satisfy all of these properties. We focus on a specific member of this family and develop an efficient method to compute the exact value of this measure. Furthermore, we illustrate our proposal by applying it to a set of example scenarios designed to highlight where other measures either cannot be applied or behave pathologically.

II. STOCHASTIC PROCESSES AND DISTANCES

Consider a bi-infinite, discrete-time, discrete-alphabet stochastic process $\mathcal{P}$, which generates an output x drawn from an alphabet $\mathcal{A}$ at each time step. A contiguous output sequence $x_{t:t+L} := x_t x_{t+1} \ldots x_{t+L-1}$ occurs with probability $P(x_{t:t+L})$, where t denotes the initial time and L is the length of sequence. Many naturally occurring processes can be described within this formalism, such as biological processes [18,19] and speech recognition [20,21]. We shall here consider processes that are both stationary and ergodic: A process is stationary if the distribution of its output sequences are invariant with respect to time, i.e., $P(x_{t:t+L}) = P(x_{0:L}) \forall t, L \in \mathbb{Z}, x_{0:L} \in \mathcal{A}^L$; a process is ergodic if its time-average behavior is identical to its ensemble-average and its statistical properties can be deduced from a single sufficiently long sample of an output sequence.

A. Criteria for a measure

With the above questions as motivation, we suggest that a good measure of distinguishability between stochastic processes $R(\mathcal{P}, \mathcal{Q})$ should satisfy the following criteria:

(1) *Non-negativity*: $R(\mathcal{P}, \mathcal{Q}) \geqslant 0$.

(2) *Symmetry*: $R(\mathcal{P}, \mathcal{Q}) = R(\mathcal{Q}, \mathcal{P})$.

(3) *Identity of indiscernibles*: $R(\mathcal{P}, \mathcal{Q}) = 0 \Leftrightarrow \mathcal{P} = \mathcal{Q}$, i.e., $R(\mathcal{P}, \mathcal{Q}) = 0$ if and only if (iff) the processes are identical.

*Yangchengran92@gmail.com

†quantum@felix-binder.net

‡mgu@quantumcomplexity.org

§physics@tjelliott.net

Together these three conditions define a semimetric distance. Note that as with many proposed measures of distance between processes, we do not demand the triangle inequality $R(\mathcal{P}, \mathcal{Q}) \leqslant R(\mathcal{P}, \mathcal{G}) + R(\mathcal{Q}, \mathcal{G}) \forall \mathcal{G}$ be satisfied, and so the measure will not necessarily be a metric.

(4) *Model independence*: $R(\mathcal{P}, \mathcal{Q})$ should depend only on observable properties (i.e., the outputs) of the processes and not any underlying models.

This condition enforces the idea that the measure should be identical when calculated relative to any representation of the processes, alleviating the issues discussed above for model-dependent measures.

(5) *Continuity*: Suppose stochastic process $\mathcal{Q}$ depends on a continuous parameter δ. Continuity mandates that $\lim_{\delta \to \delta_0} R(\mathcal{P}, \mathcal{Q}(\delta)) = R(\mathcal{P}, \mathcal{Q}(\delta_0))$.

Smooth deformations in the parameters defining a process will smoothly change the distinguishability between it and other processes—this condition enforces that this is reflected in the measure.

B. Existing measures

With these requirements at hand, we are able to assess the behavior of existing measures that have been proposed for quantifying how distinct two stochastic processes are. One such widely used measure, often serving as a cost function for many machine learning works, is the aforementioned KL divergence,

$$D_{\mathrm{KL}}(P||Q) = \sum_{x_{t:t+L}} P(x_{t:t+L}) \log_2 \left[\frac{P(x_{t:t+L})}{Q(x_{t:t+L})} \right], \tag{1}$$

It can be seen however, that the KL divergence does not satisfy the continuity criterion. It becomes singular when two stochastic processes contain different sets of possible output sequences, no matter how small the probability is of these unique sequences occurring. Moreover, while not a violation of any of the above criteria, a further drawback of the KL divergence is its high computational cost, as it requires a calculation over all output sequences, the number of which grows exponentially with their length L.

Other measures, such as the Jensen-Shannon divergence [14], though free from singularities, can still fail the continuity criterion in the context of processes where the output sequence lengths are infinite. This problem is endemic to measures based on distances between distributions, such as trace norms and the Bures distance [10], and highlights a crucial key difference between distributions over finite sequences and processes. Specifically, these measures asymptotically saturate to their maximal value as the length of the output sequences increases, because any two different processes can be asymptotically distinguished for sufficiently long output sequences. As a result, these measures are either 0 or maximal—identifying whether $\mathcal{P}$ and $\mathcal{Q}$ are different processes, but not how different they are. A good measure of process distinguishability should be equipped to handle this distinction.

Finally, the model independence criterion rules out other measures [16,17] that are explicitly based on the structure of a particular model. That is, for such measures two models of the same stochastic process may be identified as having nonzero distance between them despite exhibiting identical observable behavior. Furthermore, such model-dependent measures may also be impossible to evaluate for some pairs of models with sufficiently distinct structures.

C. Divergence rates

In light of such issues with commonly used measures, we seek a measure that satisfies all of our criteria. We propose a family of measures of process distinguishability, called divergence rates, which measure how quickly the observed behavior of two processes becomes distinguishable. That is, they quantify how much the distance between output sequence probability distributions grows with their length.

Consider a metric distance measure between distributions $D(P, Q)$ that is normalized such that $0 \leqslant D(P, Q) \leqslant 1$. We introduce the notion of *similarity* S_D, that can be thought of as the complement to the distance, satisfying

$$S_D(P, Q) := \sqrt{1 - D(P, Q)^2}. \tag{2}$$

We then define the D *divergence rate* as

$$R_D(\mathcal{P}, \mathcal{Q}) := - \lim_{L \to \infty} \frac{1}{L} \log_2 [S_D(P, Q; L)], \tag{3}$$

where $S_D(P, Q; L)$ and similarly $D(P, Q; L)$ are used to denote these quantities evaluated for distributions formed from sequences of length L output by the processes. The D divergence rate can be seen to parameterize the rate at which the similarity (according to the distance D) of the two processes decays once many symbols have been observed. The limit of $L \to \infty$ accounts for strings with infinite length and suppresses the distance induced by different initial states. This parallels the notion of entropy rates [22], which similarly capture long-term behavior by averaging over long sequences.

Theorem 1. Suppose a continuous, normalized metric distance $D(P, Q; L)$ that scales with L as $D(P, Q; L) \sim 1 - \alpha \exp(-\eta L)$ for non-negative $\alpha \leqslant 1$ and non-negative real η with continuous dependence on the stochastic processes. The $R_D(\mathcal{P}, \mathcal{Q})$ induced by the distance $D(P, Q; L)$ fulfils all above requirements for a measure of process distinguishability. For such a distance, we have that $R_D(\mathcal{P}, \mathcal{Q}) = \eta/2$.

Proof 1. Conditions 1, 2, and 4 immediately follow from the definition of $R_D(\mathcal{P}, \mathcal{Q})$ and the properties of $D(P, Q)$ as a metric. By directly inserting the scaling $D(P, Q; L) = 1 - \alpha \exp(-\eta L)$ into Eq. (3), it can be seen that $R_D(\mathcal{P}, \mathcal{Q}) = \eta/2$. As $D(P, Q)$ is a metric it follows that if two processes are identical we must have $\eta = 0$, and conversely $\eta = 0$ indicates that two processes have identical long-term behavior—and hence condition 3 is satisfied. Note that α is irrelevant and depends only on transient behavior resulting from the initial configuration of the two processes. Finally, condition 5 follows from the equality $R_D = \eta/2$ and the continuity of the decay rate η.

The above theorem holds only for specific metric distances, which depend continuously on the stochastic processes and scale as desired. Furthermore, whenever $D(P, Q)$ exhibits the required scaling, we see that $R_D(\mathcal{P}, \mathcal{Q})$ is infinite iff $D(P, Q; L)$ becomes exactly 1 within finite L, rather than just asymptotically approaching it. Qualitatively, this can be understood as the measure being infinite iff the two processes

can be discriminated with certainty by observing a sufficiently long yet finite sequence of outputs.

D. Coemission divergence rate

We now consider the case where the distance used is the L_2 norm, given by

$$D_{L_2}(P, Q) := \frac{1}{\sqrt{2}}\|\hat{P} - \hat{Q}\|_2, \tag{4}$$

where $\hat{P} = P/\|P\|_2$ and $\hat{Q} = Q/\|Q\|_2$. Noting that $\|\hat{P} - \hat{Q}\|_2^2 = 2 - 2\langle\hat{P}, \hat{Q}\rangle$, where $\langle\cdot,\cdot\rangle$ is the inner product, this distance can be expressed in terms of so-called coemission probabilities [13] $C(P, Q; L) = \sum_{x_{0:L}} P(x_{0:L})Q(x_{0:L})$. Using this, we obtain

$$R_C(\mathcal{P}, \mathcal{Q}) = -\lim_{L\to\infty} \frac{1}{2L} \log\left[\frac{C(P, Q; L)}{\sqrt{C(P, P; L)C(Q, Q; L)}}\right], \tag{5}$$

which we call the *coemission divergence rate* (CDR).

Theorem 2. The CDR satisfies all of the requirements specified for a good measure of process distinguishability.

As D_{L_2} is a continuous metric distance, we need only to show that the measure obeys the specified scaling.

Lemma 1. $D_{L_2}(P, Q; L)$ scales as $1 - \alpha\exp(-\eta L)$ with η depending continuously on the processes.

The proof of this employs a recently developed correspondence between tensor networks and stochastic processes [23], and is given in detail in Appendix B.

Any bi-infinite, stationary stochastic process can be represented in terms of a hidden Markov model (HMM) [24]. Such models consist of a set of hidden internal states s_i. At each time step, based on the current state s_i the model generates output x and transitions to state s_j with probability $P(s_j, x|s_i)$. In proving Lemma 1, we obtain an efficient way to compute the CDR between any two processes for which is known a HMM representation.

Corollary 1. Given a HMM representation of process $\mathcal{P}$ with transition probabilities $P(s_j, x|s_i)$ and of process $\mathcal{Q}$ with $Q(\tilde{s}_n, x|\tilde{s}_m)$, the CDR between them is given by

$$R_C(\mathcal{P}, \mathcal{Q}) = -\frac{1}{2}\log_2\left[\frac{\mu_{PQ}}{\sqrt{\mu_P}\sqrt{\mu_Q}}\right], \tag{6}$$

where μ_P, μ_Q, and μ_{PQ} are the leading eigenvalues of the transfer matrices $\mathbb{E}_{PP}$, $\mathbb{E}_{QQ}$, and $\mathbb{E}_{PQ}$, defined as

$$(\mathbb{E}_{PQ})_{im,jn} := \sum_x P(s_j, x|s_i)Q(\tilde{s}_n, x|\tilde{s}_m). \tag{7}$$

The computational complexity of calculating these eigenvalues (and hence the CDR) depends only on the number of hidden states $|S|$ in our HMM representations of $\mathcal{P}$ and $\mathcal{Q}$, scaling polynomially with both. We need only calculate the leading eigenvalue of a $|S_\mathcal{P}||S_\mathcal{Q}| \times |S_\mathcal{P}||S_\mathcal{Q}|$ matrix for μ_{PQ}, and similarly for μ_P and μ_Q. Moreover, as we only need the leading eigenvalues, we can use tools such as the power method [25] rather than full spectral decomposition. Crucially, there is no scaling of complexity with the length of sequences considered (the $L \to \infty$ limit is implicitly accounted for). And unlike Monte Carlo methods used to estimate, e.g., KL divergences, the result is exact.

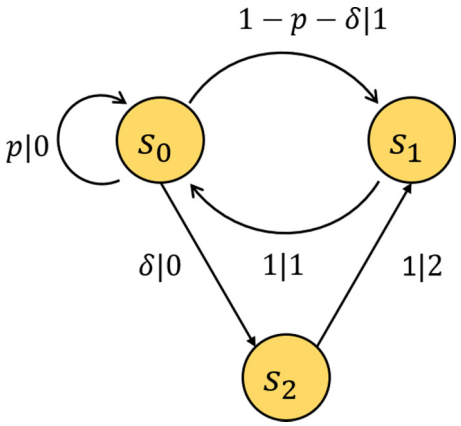

FIG. 1. The example process we consider can be represented by a three-state HMM with two variables p and δ. The edge label $P|x$ between states s_i and s_j signifies that if the model is in state s_i it will transition to s_j while emitting symbol x with probability P.

Corollary 1 can be applied to any stationary, ergodic stochastic processes, since any such process can be represented as by a HMM [24]. Although for some processes the number of hidden states in the HMM must be infinite for an exact representation, the CDR can be approximated by using approximate finite-sized HMMs.

III. EXAMPLES

In Appendix C, we work through a pedagogical example that demonstrates how our efficient method for calculating the CDR as described in Corollary 1 may be used. Here in the main text we present an illustrative example using a highly tunable process that highlights several scenarios in which our measure can be employed, where other previously proposed measures of process distinguishability break down. The most general form of this example process can be represented by a HMM with three hidden states, as illustrated in Fig. 1. The model has two variable parameters p and δ; we use $\mathcal{G}(p, \delta)$ to represent the process generated by the model for a particular set of parameters

First, we show that the CDR exhibits qualitatively similar behavior to the (symmetric) KL divergence (per symbol) in a scenario where the latter can be applied. Note that we must utilize Monte Carlo methods [26] to estimate the KL divergence, due to its computationally intensive nature. Let process $\mathcal{P} = \mathcal{G}(p, 0)$ for $p \in [0.1, 0.9]$, and similarly process $\mathcal{Q} = \mathcal{G}(q, 0)$ for $q \in [0.1, 0.9]$. We calculate the CDR using the method described in Corollary 1, while the symmetric KL divergence per symbol is estimated using the Monte Carlo method for sequences of length $L = 1000$ and a sampling set size $M = 50$.

From Fig. 2, we see that the CDR and KL divergence exhibit similar behavior. We emphasize that we are able to efficiently compute the exact CDR, while we are only able to estimate the KL divergence as it requires exponentially growing resources with sequence length. Moreover, as the

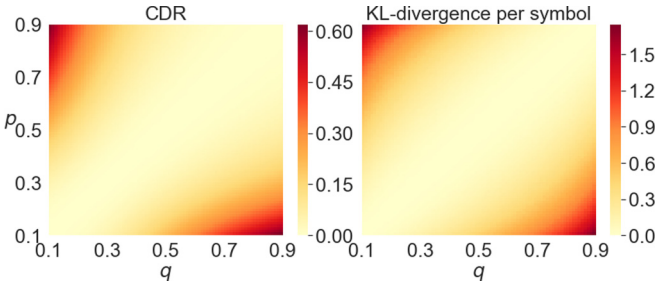

FIG. 2. Comparison of CDR and KL divergence per symbol for distinguishing between different parameter values of the process at $\delta = 0$. We see that the qualitative behavior of the two is very similar. As would be expected, both are zero along the line $p = q$ where the processes are equal, and grow as the difference $|p - q|$ increases.

processes considered have infinite Markov order (i.e., their behavior is conditioned on outputs from infinitely far back into the past), no measure based on sequences with finite L can capture the full behavior of the processes exactly.

Second, we consider a scenario where the KL divergence cannot be suitably used. Consider the case where again process $\mathcal{P} = \mathcal{G}(p, 0)$ for $p \in [0.1, 0.9]$, but now $\mathcal{Q} = \mathcal{G}(q_1, q_2)$ for $q_1 \in [0.1, 0.9]$ and $q_2 \in [0, 1 - q_1]$. When $\delta = 0$ the hidden state s_2 cannot be reached, and so the symbol 2 is never emitted—thus for any $q_2 \neq 0$, $\mathcal{P}$ and $\mathcal{Q}$ have different output alphabets and so exhibit infinite KL divergence (per symbol). Nevertheless, the CDR varies smoothly with the parameters, and we are still able to efficiently calculate it, as shown in Fig. 3 for the plane defined by $p = q_1$. Furthermore, since the HMM representations of the processes have different numbers of accessible states for $q_2 \neq 0$, a number of other measures

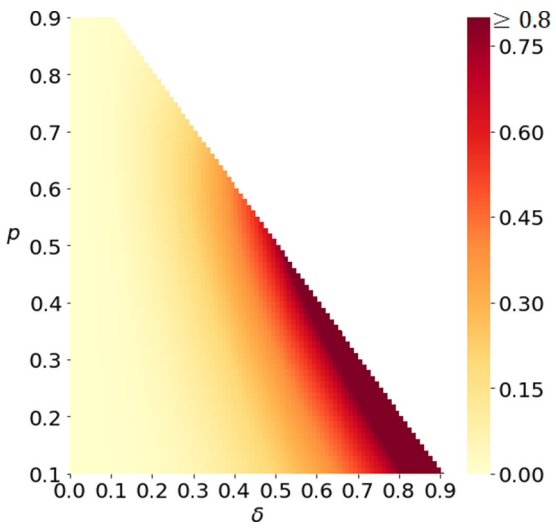

FIG. 3. The CDR is able to be calculated to ascertain the distinguishability between processes with different output alphabets and representations with different numbers of states, unlike other measures such as the KL divergence. As expected, we see that the CDR increases as one process becomes increasingly likely to emit a symbol the other cannot. The upper right white region represents an unphysical parameter regime.

based on the model topology cannot be properly applied [17].

We also show how CDR can be applied to quantify the distance between continuous-time processes that generate continuous outputs through discretization. We consider Markov processes with transition probabilities $P(x, t|x', t')$, describing the probability of finding process in state x at time t given a previous state x' at t'. For instance, the Ornstein-Uhlenbeck (OU) process [27], a model of Brownian motion, can be described by the Fokker-Planck equation

$$\frac{\partial P}{\partial t} = \theta \frac{\partial}{\partial x}(xP) + D\frac{\partial^2 P}{\partial x^2}. \tag{8}$$

After discretizing, the transition probabilities are

$$P(x, \Delta t|x_0, 0) = \mathcal{N}\left(x_0 e^{-\theta \Delta t}, \sqrt{\frac{D}{\theta}\left(1 - e^{-2\theta \Delta t}\right)}\right), \tag{9}$$

where $\mathcal{N}(\mu, \sigma)$ represents a Gaussian distribution with mean $\mu = x_0 e^{-\theta \Delta t}$ and standard deviation $\sigma = \sqrt{D/\theta[1 - \exp(-2\theta \Delta t)]}$. We use the CDR to compare an OU process to a totally random Gaussian process with null correlation and transition probabilities

$$P(x, \Delta t|x_0, 0) = \mathcal{N}(0, \sigma). \tag{10}$$

To show the convergence of the CDR for strings of increasing length, we introduce the approximation $R_C(\mathcal{P}, \mathcal{Q}) \approx g(L + 1) - g(L)$, where

$$g(L) := \frac{1}{2} \log \left[\frac{C(P, Q; L)}{\sqrt{C(P, P; L)C(Q, Q; L)}}\right], \tag{11}$$

noting that this approximation becomes exact as $L \to \infty$. In Fig. 4(a), we show how this approximation converges to the exact CDR as calculated from Corollary 1, for two different initial seed states of the OU process. In Fig. 4(b), we display how the exact CDR varies with the size of the time step used in the discretization. We see that for large Δt the distance vanishes (as the large time step wipes out the dependence on the current position) and appears to converge toward a particular value as we tend toward the continuous limit.

IV. DISCUSSION

Though our efficient method for computing the CDR relies on having HMM representations of the processes considered, the measure itself does not rely on this. In lieu of HMM representations, Monte Carlo methods can be employed on the sequence probabilities to calculate the CDR, as with the KL divergence. We also note that while we have here considered stationary processes, a modified form of the CDR can be applied to nonstationary processes, where instead of taking $L \to \infty$ we take the longest sequence possible; the measure will then yield the average decay of process similarity per symbol.

A further generalization that can be considered is the effect of relabeling the alphabet, such that each output is ascribed to a different symbol. In general, the observed statistics after such a relabeling will be different, and as such the CDR between the original and relabeled processes will generally be nonzero. However, from another perspective,

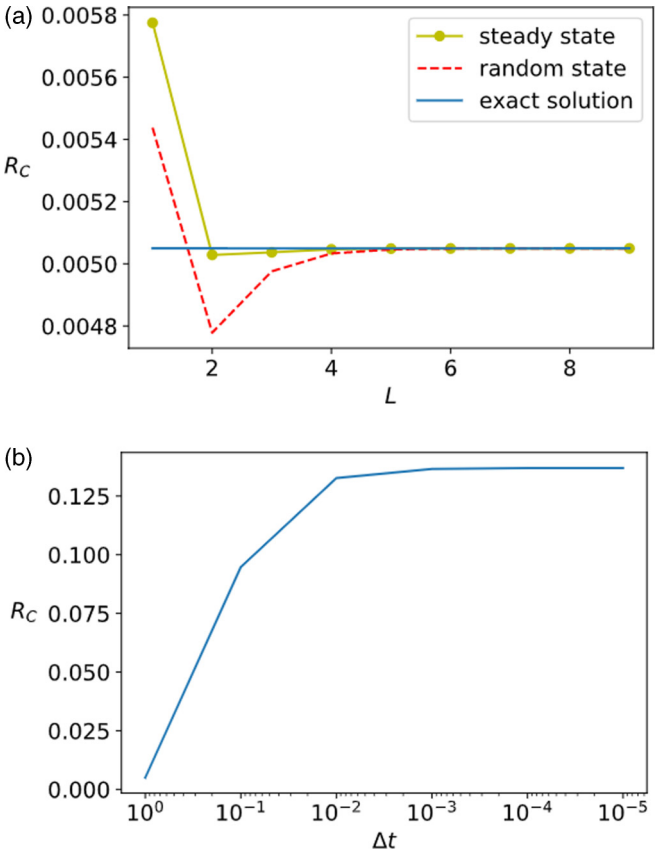

FIG. 4. CDR between an OU process and a totally random Gaussian process showing (a) convergence with L for different initial states, and (b) convergence with decreasing time-step size Δt. For (a) $\Delta t = 1$, and for both plots we take x in the range $(-3, 3)$ with step size $\Delta x = 2$, $\theta = 1$, $\mu = 0$, $\sigma = 1$. The qualitative behavior of the red curve appears typical in order of magnitude and convergence length.

one can argue that the two processes still exhibit statistically identical behavior—with merely a different nomenclature for the events. To remedy this, one can consider an alphabet-symmetrized form of the divergence rate, where it is minimized over all permutations of the alphabet for one of the processes. This would then identify a process and its relabeled version as having zero CDR.

Finally, we remark that there exist other members of the divergence rate family which satisfy all of the requirements for a process distinguishability measure. Consider the Bures [10] or Hellinger distance [8] $D_B(P, Q; L) = \sqrt{1 - F(P, Q; L)}$, where the fidelity $F(P, Q; L) := \sum_{x_{0:L}} \sqrt{P(x_{0:L})Q(x_{0:L})}$. Taking this as our distance measure, we obtain the *fidelity divergence rate* (FDR):

$$R_F(\mathcal{P}, \mathcal{Q}) = -\lim_{L\to\infty} \frac{1}{2L} \log_2[F(P, Q; L)]. \tag{12}$$

In Appendix D, we show that the FDR satisfies all the requirements and provide an efficient way to calculate it from deterministic HMM representations of the processes.

V. CONCLUSION

To summarize, we have discussed a set of conditions we believe a good measure of process distinguishability should satisfy and proposed a family of divergence rates that satisfy them. We focused on a particular example of this family, the CDR, and developed an efficient method for its computation. We illustrate the advantages of our measure relative to previously proposed measures by applying it to example scenarios where other measures behave pathologically. Finally, we discussed a number of possible generalizations of the measure.

Our measure can be applied to a broad range of areas, particularly those dealing with stochastic processes such as HMMs [17], computational mechanics [24,28] and quantum stochastic modeling [29–35]. Other areas of application include assessment of the accuracy of machine learning models, benchmarking the performance of time-series inference protocols, finding the optimal approximate representations of a process, and quantifying the robustness of processes to noise. Our method for efficiently computing the CDR uses tools from tensor networks [36–39], adding to the growing list of applications of these methods for stochastic processes [23,40–45] and machine learning [46–54].

ACKNOWLEDGMENTS

This research is supported by the National Research Foundation (NRF). Singapore, under its NRFF Fellow programme (Award No. NRF-NRFF2016-02), the Lee Kuan Yew Endowment Fund (Postdoctoral Fellowship), Singapore Ministry of Education Tier 1 Grants No. MOE2017-T1-002-043 and No FQXi-RFP-1809 from the Foundational Questions Institute and Fetzer Franklin Fund (a donor-advised fund of Silicon Valley Community Foundation). F.C.B. acknowledges funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska Curie Grant Agreement No. 801110 and the Austrian Federal Ministry of Education, Science, and Research (BMBWF). T.J.E., C.Y., and F.C.B. thank the Centre for Quantum Technologies for their hospitality. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

APPENDIX A: TENSOR NETWORKS AND THEIR RELATION TO HMMS

A tensor network decomposes a large tensor into several smaller tensors connected by a network structure. These techniques have many promising applications, a key one being in simplifying the numerical simulation of quantum many-body systems. They possess a comprehensive pictorial representation in which each tensor is represented by a node with several legs, as shown in Fig. 5(a).

To represent a HMM we can use a special type of tensor network, called matrix product states [23]. Transitions between states in the HMM are described by the transition matrix $T^x_{ij} := P(s_j, x|s_i)$; this is a rank 3 tensor and is thus represented by a node with three legs, as shown in Fig. 5(b). The stationary distribution of the HMM states π_i is represented

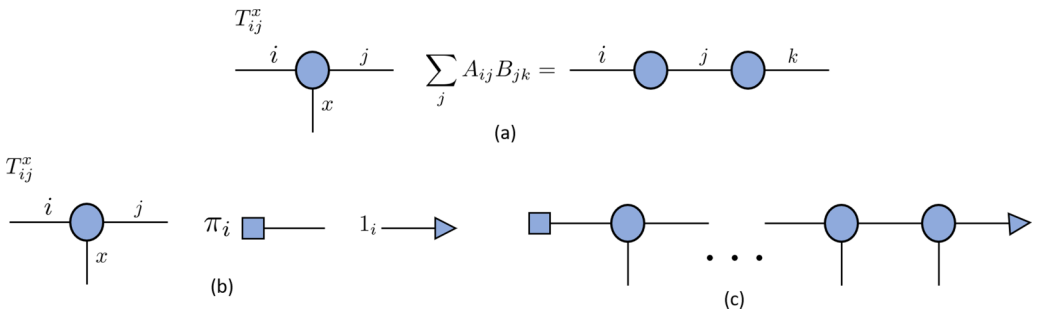

FIG. 5. Pictorial representation of tensor networks. (a) Each leg represents an index of a tensor, with linking between legs representing summation over the corresponding index. (b) Tensor network representation of a HMM. (c) Multi-index tensor representing the probability of an output sequence.

by the square, and the triangle represents 1_i, a column vector filled with 1s.

The probability of a particular sequence $x_{0:L}$ being generated by a HMM is given by

$$P(x_{0:L}) = \sum_{s_i^0, s_i^1 \cdots s_i^{L-1}} \pi_i^0 P\left(s_i^1, x_0 | s_i^0\right) \ldots P\left(s_i^L, x_{L-1} | s_i^{L-1}\right). \tag{A1}$$

This can be represented by a tensor network, as shown in Fig. 5(c). The nomenclature "matrix product state" becomes clear: The sequence tensor is obtained by multiplying by a matrix T^{x_l} at each step.

A HMM, and its tensor network representation, decompose the large tensor $P(x_{0:L})$ into products of small tensors T^x_{ij}. As a result, HMMs exponentially reduce the memory requirement of representing a stochastic process to $O(LN^2)$ from $O(|\mathcal{A}|^L)$ where L is the length of the sequence, N is the number of states of the HMM, and $|\mathcal{A}|$ is the size of output alphabet.

APPENDIX B: PROOF OF THEOREM 2, LEMMA 1, AND COROLLARY 1

Here, we present our efficient method of computing the coemission divergence rate (CDR), in the process proving Theorem 2 and Lemma 1. Every stationary stochastic process has a HMM representation [24]; we consider two stationary stochastic processes $\mathcal{P}$ and $\mathcal{Q}$ with HMMs $T^x_{ij} := P(s_j, x|s_i)$ and $\tilde{T}^x_{mn} := Q(\tilde{s}_n, x|\tilde{s}_m)$, where s_i and $\tilde{s}_m$ are the corresponding hidden states. The corresponding pictorial representations are shown in Fig. 6(a).

The coemission probability is

$$C(P, Q; L) = \sum_{x_{t:t+L}} P(x_{t:t+L}) Q(x_{t:t+L}), \tag{B1}$$

obtained by contracting the output indices $x_{t:t+L}$ over tensors T and $\tilde{T}$, as shown in Fig. 6(b). The tensor structure in the dashed square, which repeatedly appears in the network, has four legs, i.e., is a rank 4 tensor. Combining the left two legs together as a row index, and the right two legs as a column index, this becomes the transfer matrix

$$(\mathbb{E}_{PQ})_{im,jn} := \sum_x P(s_j, x|s_i) Q(\tilde{s}_n, x|\tilde{s}_m). \tag{B2}$$

The leftmost and rightmost tensors represent the left and right boundaries, respectively. The left boundary $\langle bl|$ is a row vector with elements $v_{ij} = \pi_i \tilde{\pi}_j$. The right boundary $|br\rangle$ is a column vector filled with 1s, such that the hidden states at the last step are equally weighted, i.e., $P(x_{t:t+L})Q(x_{t:t+L}) = \sum_{s_j, s_n} P(x_{t:t+L}, s_j) Q(x_{t:t+L}, s_n)$.

If $\mathbb{E}_{PQ}$ is diagonalizable, it has an eigenvalue decomposition

$$\mathbb{E}_{PQ} = \sum_i \mu_i |r_i\rangle\langle l_i|, \tag{B3}$$

where μ_i are eigenvalues of $\mathbb{E}_{PQ}$, sorted in order of decreasing magnitude, and $|r_i\rangle$ and $\langle l_i|$ are the associated right and left eigenvectors. Consequently, we have

$$\mathbb{E}^L_{PQ} = \mu_1^L (|r_1\rangle\langle l_1| + \sum_{i \neq 1} \left(\frac{\mu_i}{\mu_1}\right)^L |r_i\rangle\langle l_i|). \tag{B4}$$

As $\mathbb{E}_{PQ}$ is constructed from probabilities, it is non-negative. Its left- and right-leading eigenvectors are then non-negative according to the Perron-Frobenius theorem [55,56]. Thus, the left- and right-boundary vectors have nonzero overlap with the associated leading left and right eigenvectors of the matrix $\mathbb{E}_{PQ}$, and therefore the coemission probability has the following scaling:

$$\begin{aligned} C(P, Q; L) &= \langle bl|\mathbb{E}^L_{PQ}|br\rangle \\ &= \mu^L_{PQ}\left\{\alpha_{PQ} + O\left[\left(\frac{\mu_2}{\mu_{PQ}}\right)^L\right]\right\} \sim \alpha_{PQ}\mu^L_{PQ}, \end{aligned} \tag{B5}$$

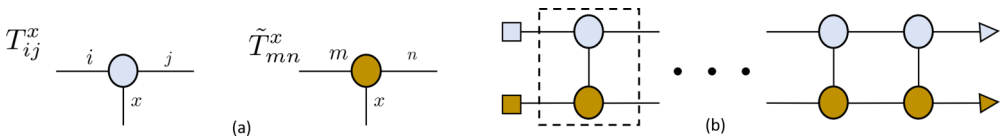

FIG. 6. (a) Tensor network representation of the transition matrices of $\mathcal{P}$ and $\mathcal{Q}$. (b) Tensor network representation of the coemission probability.

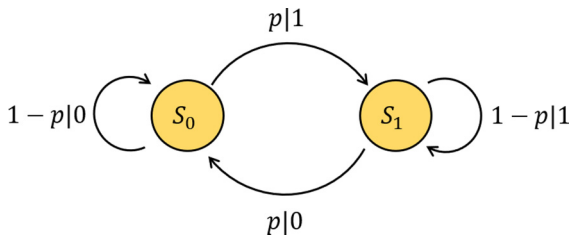

FIG. 7. The perturbed coin process has two hidden states, s_0 and s_1. The system occupies s_0 when the last output was 0, and similarly, s_1 after output 1.

where $\mu_{PQ} := \mu_1$ and $\alpha_{PQ} = \langle bl|r_1\rangle\langle l_1|br\rangle$ is positive. This scaling holds even if $\mathbb{E}_{PQ}$ is not diagonalizable; this can be proved using the Jordan form of the matrix.

Using the same argument, we also have

$$C(P,P;L) \sim \alpha_P \mu_P^L \quad \text{and} \quad C(Q,Q;L) \sim \alpha_Q \mu_Q^L, \tag{B6}$$

where μ_P is the leading eigenvalue of the transfer matrix $\mathbb{E}_{PP}$ and μ_Q is the leading eigenvalue of the transfer matrix $\mathbb{E}_{QQ}$. Then, we have

$$D_{L_2}(P,Q;L)^2 = 1 - \frac{C(P,Q;L)}{\sqrt{C(P,P;L)C(Q,Q;L)}} \sim 1 - \alpha\left(\frac{\mu_{PQ}}{\sqrt{\mu_P\mu_Q}}\right)^L, \tag{B7}$$

where $\alpha = \alpha_{PQ}/\sqrt{\alpha_P\alpha_Q}$. Thus, the distance has the desired scaling. The continuity of the decay rate η follows from the continuity of the leading eigenvalues, which depend continuously on the coefficients of the characteristic equations. This proves Lemma 1, and in turn Theorem 2. Taking $L \to \infty$ leads to

$$\lim_{L\to\infty} \frac{1}{L} \log_2 \left[\sum_{x_{t:t+L}} P(x_{t:t+L})Q(x_{t:t+L})\right] = \log_2 \mu_{PQ}, \tag{B8}$$

and analogously,

$$\lim_{L\to\infty} \frac{1}{L} \log_2 \left[\sum_{x_{t:t+L}} P(x_{t:t+L})P(x_{t:t+L})\right] = \log \mu_P, \tag{B9}$$

$$\lim_{L\to\infty} \frac{1}{L} \log_2 \left[\sum_{x_{t:t+L}} Q(x_{t:t+L})Q(x_{t:t+L})\right] = \log \mu_Q. \tag{B10}$$

Therefore,

$$R_C(\mathcal{P},\mathcal{Q}) = -\frac{1}{2}\log_2 \frac{\mu_{PQ}}{\sqrt{\mu_P\mu_Q}}. \tag{B11}$$

This proves Corollary 1.

APPENDIX C: PEDAGOGICAL EXAMPLE OF CALCULATING THE CDR

As a pedagogical example of how our efficient method for computing the CDR works, we study the distinguishability between two versions of the perturbed coin process [29] (representable by the HMM in Fig. 7) with different parameters. This is a Markov process, as output 0 indicates the hidden state is s_0 and output 1 indicates the hidden state s_1. Consider two perturbed coin processes $\mathcal{P}$ and $\mathcal{Q}$ with parameters p and q respectively. Then the transfer matrices are

$$\mathbb{E}_{PP} = \begin{bmatrix} (1-p)^2 & 0 & 0 & p^2 \\ (1-p)p & 0 & 0 & (1-p)p \\ (1-p)p & 0 & 0 & (1-p)p \\ p^2 & 0 & 0 & (1-p)^2 \end{bmatrix},$$

$$\mathbb{E}_{QQ} = \begin{bmatrix} (1-q)^2 & 0 & 0 & q^2 \\ (1-q)q & 0 & 0 & (1-q)q \\ (1-q)q & 0 & 0 & (1-q)q \\ q^2 & 0 & 0 & (1-q)^2 \end{bmatrix},$$

$$\mathbb{E}_{PQ} = \begin{bmatrix} (1-p)(1-q) & 0 & 0 & pq \\ (1-p)q & 0 & 0 & p(1-q) \\ p(1-q) & 0 & 0 & (1-p)q \\ pq & 0 & 0 & (1-p)(1-q) \end{bmatrix}. \tag{C1}$$

Evaluating the leading eigenvalues of these matrices, we obtain

$$\begin{aligned} \mu_P &= p^2 + (1-p)^2, \\ \mu_Q &= q^2 + (1-q)^2, \\ \mu_{PQ} &= pq + (1-p)(1-q). \end{aligned} \tag{C2}$$

Therefore, the CDR is

$$R_C(\mathcal{P},\mathcal{Q}) = -\frac{1}{2}\log_2 \frac{pq + (1-p)(1-q)}{\sqrt{[p^2+(1-p)^2][q^2+(1-q)^2]}}. \tag{C3}$$

Clearly, $R(\mathcal{P},\mathcal{Q}) = 0$ iff the two processes are identical, i.e., $p = q$.

APPENDIX D: FIDELITY DIVERGENCE RATE

Here, we present another member of the divergence rate family that also satisfies the desired properties of a process distinguishability measure. This divergence rate is called fidelity divergence rate (FDR), as the associated distance is the Bures-Hellinger distance $D_B(P,Q;L) = \sqrt{1 - F(P,Q;L)}$, which is expressed in terms of the fidelity $F(P,Q;L) := \sum_{x_{t:t+L}} \sqrt{P(x_{t:t+L})Q(x_{t:t+L})}$. Then $S_{D_B}(P,Q;L) = \sqrt{1 - D_B(P,Q;L)^2} = \sqrt{F(P,Q;L)}$, and the FDR is

$$R_F(\mathcal{P},\mathcal{Q}) = -\frac{1}{2}\lim_{L\to\infty}\frac{1}{L}\log_2[F(P,Q;L)]. \tag{D1}$$

Similar to the CDR, we will demonstrate that D_B exhibits the required scaling for the FDR to satisfy our requirements and provide an efficient method for its evaluation given deterministic HMM representations of the processes. A deterministic (or unifilar) HMM is one for which the current hidden state can always be deduced with certainty given the previous state and output. This means that each state only has at most one outgoing edge for each symbol. Every stationary stochastic process has a deterministic HMM representation [24].

Theorem 3. The FDR satisfies all the proposed requirements for a measure of process distinguishability.

Consider two stationary stochastic processes, $\mathcal{P}$ and $\mathcal{Q}$, with deterministic HMM representations $P(s_j, x|s_i)$ and $Q(\tilde{s}_n, x|\tilde{s}_m)$, respectively, and associated transfer matrix $(\mathbb{E}^F_{PQ})_{im,jn} = \sum_x \sqrt{P(s_j|x,s_i)Q(\tilde{s}_n,x|\tilde{s}_m)}$.

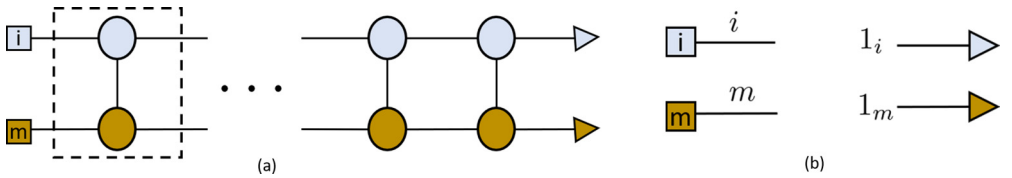

FIG. 8. Tensor network representation of (a) the fidelity and (b) boundary vectors.

Lemma 2. $D_B(P, Q; L)$ scales as $1 - \alpha \exp(-\eta L)$.

The fidelity $\sum_{x_{0:L}} \sqrt{P(x_{0:L}|s_i)Q(x_{0:L}|\tilde{s}_m)}$, conditioned on starting in hidden states $(s_i, \tilde{s}_m)$, has a pictorial representation as shown in Fig. 8(a). The left boundary represents the ith and mth standard basis vectors $\langle i|$ and $\langle m|$ in the corresponding space, while the right boundary, denoted by $|br\rangle$, is the column vector filled with 1s, as shown in Fig. 8(b).

The tensor structure in the dashed square is the transfer matrix $\mathbb{E}^F_{PQ}$, which acts repeatedly on the left boundary $\langle i, m|$. Similar to the proof for the coemission, if $\mathbb{E}^F_{PQ}$ is diagonalizable we have the eigenvalue decomposition

$$\mathbb{E}^F_{PQ} = \sum_i \mu_i |r_i\rangle\langle l_i|, \tag{D2}$$

where μ_i are the eigenvalues of $\mathbb{E}^F_{PQ}$ sorted in order of decreasing magnitude, and $|r_i\rangle$ and $\langle l_i|$ are the associated right and left eigenvectors. Consequently,

$$\left(\mathbb{E}^F_{PQ}\right)^L = \mu_1^L(|r_1\rangle\langle l_1| + \sum_{i\neq 1}\left(\frac{\mu_i}{\mu_1}\right)^L |r_i\rangle\langle l_i|), \tag{D3}$$

where $\alpha = \langle bl|r_1\rangle\langle l_1|br\rangle$ is the overlap between the left vector and the leading eigenvector of $\mathbb{E}^F_{PQ}$. Because $\mathbb{E}^F_{PQ}$ is non-negative matrix, its left- and right-leading eigenvectors are non-negative according to the Perron-Frobenius theorem [55,56]. Since $\langle i, m|$ spans the whole space, there always exists a vector $\langle i, m|$ such that it has nonzero overlap with leading left eigenvector of transfer operator $\mathbb{E}^F_{PQ}$, i.e., $\langle i, m|r_1\rangle > 0$. As with the coemission, the above scaling still holds when $\mathbb{E}^F_{PQ}$ is not diagonalizable, as can be shown using the Jordan form. Thus, we see that the fidelity decays exponentially with the length of the sequence, and thus D_B exhibits the required scaling. The continuity of decay rate also follows from the continuity of leading eigenvalues, hence proving Lemma 2 and Theorem 3.

Corollary 2. Given deterministic HMM representations $P(s_j, x|s_i)$ and $Q(\tilde{s}_n, x|\tilde{s}_m)$ of processes $\mathcal{P}$ and $\mathcal{Q}$, the FDR is given by

$$R_F(\mathcal{P}, \mathcal{Q}) = -\tfrac{1}{2}\log_2 \mu_{PQ}, \tag{D4}$$

where μ_{PQ} is the leading eigenvalue of operator $\mathbb{E}^F_{PQ} = \sum_x \sqrt{P(s_j|x, s_i)Q(\tilde{s}_n, x|\tilde{s}_m)}$.

For certain boundary vectors $\langle i, m|$, we have

$$\lim_{L\to\infty}\frac{1}{L}\log_2 \sum_{x_{0:L}} \sqrt{P(x_{0:L}|s_i)Q(x_{0:L}|\tilde{s}_m)} = \log_2 \mu_{PQ}, \tag{D5}$$

where $\mu_{PQ} = \mu_1$ The above quantity is the fidelity conditional on certain past $(s_i, \tilde{s}_m)$. We now bound the nonconditioned fidelity in the following:

$$\begin{aligned} F(P, Q; L) &:= \sum_{x_{0:L}} \sqrt{P(x_{0:L})Q(x_{0:L})} \\ &= \sum_{x_{0:L}} \sqrt{\sum_i \pi_i P(x_{0:L}|s_i)} \times \sqrt{\sum_m \tilde{\pi}_m Q(x_{0:L}|\tilde{s}_m)}. \end{aligned} \tag{D6}$$

The inequality $\sqrt{x+y} \leqslant \sqrt{x} + \sqrt{y}$ implies

$$\begin{aligned} F(P, Q; L) &\leqslant \sum_{x_{0:L}, i, m} \sqrt{\pi_i \tilde{\pi}_m} \times \sqrt{P(x_{0:L}|s_i)Q(x_{0:L}|\tilde{s}_m)} \\ &\leqslant \max_{i,m}\left(\sum_{x_{0:L}} \sqrt{P(x_{0:L}|s_i)Q(x_{0:L}|\tilde{s}_m)}\right). \end{aligned} \tag{D7}$$

Therefore, we have

$$\begin{aligned} R_F(\mathcal{P}, \mathcal{Q}) &= \frac{1}{2}\lim_{L\to\infty} -\frac{1}{L}\log_2 F(P, Q; L) \\ &\geqslant \frac{1}{2}\lim_{L\to\infty} -\frac{1}{L}\log_2 \max_{i,m} \\ &\quad \times \left(\sum_{x_{0:L}} \sqrt{P(x_{0:L}|s_i)Q(x_{0:L}|\tilde{s}_m)}\right) \\ &= -\frac{1}{2}\log_2 \mu_{PQ}. \end{aligned} \tag{D8}$$

On the other hand, using inequality $\sqrt{\sum_i p_i x_i} \geqslant \sum_i p_i \sqrt{x_i}$, we have

$$\begin{aligned} F(P, Q; L) &\geqslant \sum_{i,m} \pi_i \tilde{\pi}_m \sum_{x_{0:L}} \sqrt{P(x_{0:L}|s_i)Q(x_{0:L}|\tilde{s}_m)} \\ &\geqslant \max_{i,m} \pi_i \tilde{\pi}_m \sum_{x_{0:L}} \sqrt{P(x_{0:L}|s_i)Q(x_{0:L}|\tilde{s}_m)}. \end{aligned} \tag{D9}$$

Similarly, we also obtain

$$R_F(P, Q) \leqslant -\tfrac{1}{2}\log_2 \mu_{PQ}.$$

Thus, the proof is completed.

The fidelity divergence rate can thus be obtained by evaluating the leading eigenvalue of the transfer matrix $\mathbb{E}^F_{PQ}$. The computational complexity of this method depends only polynomially on the number of hidden states in the deterministic HMM representations of each process, and thus can be efficiently computed.

Finally, we provide upper and lower bounds for the FDR that can be calculated even when we do not have deterministic representations of the processes, from their statistics alone.

Theorem 4. Suppose two stochastic processes $\mathcal{P}$, $\mathcal{Q}$ have finite Markov order and the larger one is κ. Then the fidelity divergence rate has the following upper and lower bounds:

$$\begin{aligned} R^{\downarrow} &:= \min_{x_{-\kappa:0}} -\log F[P(x|x_{-\kappa:0}), Q(x|x_{-\kappa:0})] \leqslant 2R_F(\mathcal{P}, \mathcal{Q}), \\ R^{\uparrow} &:= \max_{x_{-\kappa:0}} -\log F[P(x|x_{-\kappa:0}), Q(x|x_{-\kappa:0})] \geqslant 2R_F(\mathcal{P}, \mathcal{Q}). \end{aligned} \tag{D10}$$

First, having Markov order κ implies that

$$P(x_{0:L+\kappa+1}) = P(x_{L+\kappa}|x_{0:L+\kappa})P(x_{0:L+\kappa}) = P(x_{L+\kappa}|x_{L:L+\kappa})P(x_{0:L+\kappa}). \tag{D11}$$

From this, we find that

$$P(x_{0:L+\kappa+1}) \leqslant P(x_{0:L+\kappa}) \max_{x_{L:L+\kappa}} P(x_{L+\kappa}|x_{L:L+\kappa}). \tag{D12}$$

Thus, we have

$$\begin{aligned} F(P, Q; L+\kappa+1) &= \sum_{x_{0:L+\kappa+1}} \sqrt{P(x_{0:L+\kappa+1})Q(x_{0:L+\kappa+1})} \\ &= \sum_{x_{0:L+\kappa}} \sqrt{P(x_{0:L+\kappa})Q(x_{0:L+\kappa})} \times \sum_{x_{L+\kappa}} \sqrt{P(x_{L+\kappa}|x_{L:L+\kappa})Q(x_{L+\kappa}|x_{L:L+\kappa})} \\ &\leqslant \sum_{x_{0:L+\kappa}} \sqrt{P(x_{0:L+\kappa})Q(x_{0:L+\kappa})} \times \max_{x_{L:L+\kappa}} \sum_{x_{L+\kappa}} \sqrt{P(x_{L+\kappa}|x_{L:L+\kappa})Q(x_{L+\kappa}|x_{L:L+\kappa})} \\ &= F(P, Q; L+\kappa) \times \max_{x_{L:L+\kappa}} \sum_{x_{L+\kappa}} \sqrt{P(x_{L+\kappa}|x_{L:L+\kappa})Q(x_{L+\kappa}|x_{L:L+\kappa})}. \end{aligned} \tag{D13}$$

Substituting the above into the definition of FDR leads to

$$\begin{aligned} 2R_F(\mathcal{P}, \mathcal{Q}) &= \lim_{L\to\infty} -\frac{1}{L}\log_2[F(P, Q; L)] \\ &\leqslant \lim_{L\to\infty} -\frac{1}{L}[\log_2 F(P, Q; \kappa) + (L-\kappa)R^{\uparrow}] = R^{\uparrow}. \end{aligned} \tag{D14}$$

The lower bound $2R_F(\mathcal{P}, \mathcal{Q}) \leqslant R^{\downarrow}$ can similarly be obtained by replacing maximizations with minimizations and reversing the directions of the inequalities.

[1] J. Söding, Bioinformatics **21**, 951 (2004).
[2] L. R. Rabiner, Proc. IEEE **77**, 257 (1989).
[3] J. Silva and S. Narayanan, IEEE Trans. Signal Process. **56**, 4176 (2008).
[4] W. Horsthemke, in *Non-equilibrium Dynamics in Chemical Systems* (Springer, Berlin, 1984), pp. 150–160.
[5] J. Preskill, Quantum **2**, 79 (2018).
[6] R. A. Kennewick, D. Locke, M. R. Kennewick, Sr., M. R. Kennewick, Jr., and T. Freeman, System and Method for Filtering and Eliminating Noise from Natural Language Utterances to Improve Speech Recognition and Parsing, U.S. Patent No. 8,140,327, 20 March 2012.
[7] J. Huang and B. Kingsbury, in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (IEEE, Piscataway, 2013), pp. 7596–7599.
[8] E. Hellinger, J. Angew. Math. **136**, 210 (1909).
[9] S. Kullback and R. A. Leibler, Ann. Math. Stat. **22**, 79 (1951).
[10] D. Bures, Trans. Am. Math. Soc. **135**, 199 (1969).
[11] B.-H. Juang and L. R. Rabiner, AT&T Tech. J. **64**, 391 (1985).
[12] E. F. Krause, *Taxicab Geometry: An Adventure in Non-Euclidean Geometry*, Dover Books on Mathematics Series (Dover, New York, 1986).
[13] R. B. Lyngso, C. N. Pedersen, and H. Nielsen, in *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology* (AAAI Press, Heidelberg, Germany, 1999), Vol. 99, pp. 178–186.
[14] B. Fuglede and F. Topsoe, in *Proceedings of the International Symposium on Information Theory, 2004 ISIT* (IEEE, Piscataway, 2004), p. 31.
[15] S.-H. Cha, Int. J. Math. Model. Meth. Appl. Sci. **1**, 300 (2007).
[16] S. M. E. Sahraeian and B.-J. Yoon, IEEE Signal Process. Lett. **18**, 87 (2011).
[17] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, Bell Syst. Tech. J. **62**, 1035 (1983).
[18] P. Baldi, Y. Chauvin, T. Hunkapiller, and M. A. McClure, Proc. Natl. Acad. Sci. USA **91**, 1059 (1994).
[19] C. Barrett, K. Karplus, and R. Hughey, Bioinformatics **14**, 846 (1998).
[20] B. H. Juang and L. R. Rabiner, Technometrics **33**, 251 (1991).
[21] B. Schuller, G. Rigoll, and M. Lang, in *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003 (ICASSP '03)* (IEEE, Piscataway, 2003), Vol. 2, p. II-1.
[22] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (John Wiley & Sons, New York, 2012).

[23] C. Yang, F. C. Binder, V. Narasimhachar, and M. Gu, Phys. Rev. Lett. **121**, 260602 (2018).
[24] C. R. Shalizi and J. P. Crutchfield, J. Stat. Phys. **104**, 817 (2001).
[25] R. Mises and H. Pollaczek-Geiringer, Z. Angew. Math. Mech. **9**, 152 (1929).
[26] J. R. Hershey and P. A. Olsen, in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07* (IEEE, Piscataway, 2007), Vol. 4, p. IV-317.
[27] J. L. Doob, Ann. Math. **43**, 351 (1942).
[28] J. P. Crutchfield and K. Young, Phys. Rev. Lett. **63**, 105 (1989).
[29] M. Gu, K. Wiesner, E. Rieper, and V. Vedral, Nat. Commun. **3**, 762 (2012).
[30] J. R. Mahoney, C. Aghamohammadi, and J. P. Crutchfield, Sci. Rep. **6**, 20495 (2016).
[31] C. Aghamohammadi, S. P. Loomis, J. R. Mahoney, and J. P. Crutchfield, Phys. Rev. X **8**, 011025 (2018).
[32] T. J. Elliott and M. Gu, npj Quantum Inf. **4**, 18 (2018).
[33] F. C. Binder, J. Thompson, and M. Gu, Phys. Rev. Lett. **120**, 240502 (2018).
[34] T. J. Elliott, A. J. P. Garner, and M. Gu, New J. Phys. **21**, 013021 (2019).
[35] Q. Liu, T. J. Elliott, F. C. Binder, C. Di Franco, and M. Gu, Phys. Rev. A **99**, 062110 (2019).
[36] D. Perez-Garcia, F. Verstraete, M. M. Wolf, and J. I. Cirac, arXiv:quant-ph/0608197).
[37] R. Orús and G. Vidal, Phys. Rev. B **78**, 155117 (2008).
[38] K. Temme and F. Verstraete, Phys. Rev. Lett. **104**, 210502 (2010).
[39] R. Orús, Ann. Phys. **349**, 117 (2014).
[40] Y. Hieida, J. Phys. Soc. Jpn. **67**, 369 (1998).
[41] E. Carlon, M. Henkel, and U. Schollwöck, Eur. Phys. J. B **12**, 99 (1999).
[42] E. Carlon, M. Henkel, and U. Schollwöck, Phys. Rev. E **63**, 036101 (2001).
[43] A. Critch and J. Morton, SIGMA **10**, 95 (2014).
[44] M. Kliesch, D. Gross, and J. Eisert, Phys. Rev. Lett. **113**, 160503 (2014).
[45] T. H. Johnson, T. J. Elliott, S. R. Clark, and D. Jaksch, Phys. Rev. Lett. **114**, 090602 (2015).
[46] I. V. Oseledets, SIAM J. Sci. Comput. **33**, 2295 (2011).
[47] E. Stoudenmire and D. J. Schwab, in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016* (Curran Associates, Inc., Red Hook, NY, 2016), pp. 4799–4807.
[48] Z.-Y. Han, J. Wang, H. Fan, L. Wang, and P. Zhang, Phys. Rev. X **8**, 031012 (2018).
[49] E. M. Stoudenmire, Quantum Sci. Technol. **3**, 034003 (2018).
[50] C. Guo, Z. Jie, W. Lu, and D. Poletti, Phys. Rev. E **98**, 042114 (2018).
[51] J. Chen, S. Cheng, H. Xie, L. Wang, and T. Xiang, Phys. Rev. B **97**, 085104 (2018).
[52] S. R. Clark, J. Phys. A: Math. Th. **51**, 135301 (2018).
[53] I. Glasser, N. Pancotti, and J. I. Cirac, arXiv:1806.05964.
[54] Y. Levine, O. Sharir, N. Cohen, and A. Shashua, Phys. Rev. Lett. **122**, 065301 (2019).
[55] O. Perron, Math. Ann. **64**, 248 (1907).
[56] F. G. Frobenius, *Über Matrizen aus Nicht Negativen Elementen* (Königliche Akademie der Wissenschaften, Berlin, Germany, 1912).