Quantum predictive filtering

Arne L. Grimsmo[1,*] and Susanne Still[2,†]

[1]*Institut quantique and Département de Physique, Université de Sherbrooke, Sherbrooke, Québec, Canada J1K 2R1*
[2]*Department of Information and Computer Sciences and Department of Physics and Astronomy, University of Hawai'i at Mānoa,*
Honolulu, Hawaii 96822, USA

How can relevant information be extracted from a quantum information source? In many situations, only some part of the total information content produced by an information source is useful. Can one then find an efficient encoding, in the sense of retaining the largest fraction of relevant information? This paper offers one possible solution by giving a generalization of a classical method designed to retain as much relevant information as possible in a lossy data compression. A key feature of the method is to introduce a *second* information source to define relevance. We quantify the advantage a quantum encoding has over the best classical encoding in general, and we demonstrate using examples that a substantial quantum advantage is possible. We show analytically, however, that if the relevant information is purely classical, then a classical encoding is optimal.

I. INTRODUCTION

Predicting future outcomes based on past observations is a fundamental problem, not only for science and technology, but also for living organisms. A central question is how much of the available data is *useful* for prediction. This question is closely related to the challenge of finding quantitative measures of complexity. One such measure is the fraction of information about a dynamical process's past states that is needed to describe its future states [1–5]. Efficient use of information then boils down to storing only information that is relevant for prediction: If two representations of past events yield equally good predictions, the simpler one is typically preferable [6–10].

We have shown recently that this type of information efficiency is directly related to efficient thermodynamic operation [11,12], implying that considerations regarding predictive filtering might be of relevance to the design of power efficient small scale devices, man-made or natural. Since real environments are ultimately quantum, this immediately raises a question: Is there an advantage gained by encoding the information in a quantum memory rather than a classical memory? In this paper we investigate under what conditions a quantum memory can be more predictive than a classical memory and whether it can do so with higher efficiency.

This basic information-theoretic question is particularly enticing in a biological context. Living systems are masters at adapting to their environment, and predicting future events is key to their survival. Efficient information use is found throughout the nervous system and may constitute a building principle of biological computing machines, such as neurons and brains [13]. Filtering useful bits from "nonpredictive clutter" [13] furthermore allows for thermodynamic efficiency [11,12]. Recent work has found that biophysical devices indeed evolved to use energy in a highly efficient manner [14–16].

There is mounting evidence that quantum effects play an important role in the efficient operation of some microscopic

biological systems [17–19]. For example, the extremely high energetic efficiency of excitation transport in light-harvesting complexes in bacteria and plants may be using some form of quantum random walk [20]. It has been shown that a delicate balance of noisy and coherent quantum processes is necessary to reach optimal efficiency [20–22], indicating that quantum effects might be exploited for an evolutionary advantage. Although quantifying quantum effects in biological systems and understanding advantages they might result in remain a great challenge, we hope that the work presented in this paper will provide a fruitful framework for addressing questions regarding the utility of storing *quantum* information for systems that operate in a quantum environment.

A. Extraction of relevant information from classical information sources

Any continuous information source contains an infinite amount of information, but not all of this information is useful to the receiver at the other end of a communication channel. One would therefore like to delineate relevant from irrelevant information. Shannon [23] addressed this problem by pointing out that the rate of an information source for a given quality of signal reproduction should be taken as the smallest amount of information required to specify the source, subject to a given constraint on average distortion. The distortion measure, which has to be chosen *ad hoc* by the practitioner, implicitly contains a notion of relevance [24].

Relevant information is treated explicitly in a similar framework, called the "information bottleneck" method [26]. Given a data source X and a relevant variable Y that depends on X, the method finds an optimal encoding $X \to M$ of the data into a representation M, such that information about Y is kept while irrelevant bits are filtered out. The method achieves this by solving a constrained optimization problem, minimizing mutual information $I[X,M]$, subject to a constraint on $I[M,Y]$. The quantity $I[X,M]$ is taken to measure the coding cost, in line with Shannon's work, and the quantity $I[M,Y]$ measures the amount of relevant information kept in the "memory," M. It cannot exceed the total relevant information the raw data contains, i.e., $I[M,Y] \leqslant I[X,Y]$, due to the data processing inequality. The trade-off between

———————
*arne.loehre.grimsmo@usherbrooke.ca
†sstill@hawaii.edu

coding cost and relevant bits retained is controlled by a Lagrange multiplier which parametrizes a family of solutions. The resulting compression scheme finds sufficient statistics in the limit where all of the relevant information is kept, i.e., $I[M,Y] = I[X,Y]$. When applied to time series prediction, this limiting model can be used to quantify the complexity of a dynamical system, as mentioned above [27].

B. Organization of the paper

In this paper we introduce a quantum generalization of the information bottleneck method in Sec. II. The data source, X, is now quantum, and relevance is defined with respect to a second quantum data source, Y. The coding protocol we introduce here can be seen as a generalization of the protocol used in quantum rate-distortion coding [28].

Following the classical method, we find an encoding of the information in X into a memory M, maximizing the information about Y while discarding irrelevant information. To that end, we derive self-consistent equations that any optimal encoding must obey (Sec. II C and Appendix A). These equations form the basis for an iterative algorithm (Appendix C) which allows us to illustrate the behavior of optimal quantum encodings using a series of numerical examples (Sec. V).

We show analytically that a quantum advantage is possible only for nonclassical relevant information, i.e., only when the co-occurrence statistics between X and Y cannot be described by a classical probability distribution (Sec. III B and Appendix B 2). A system operating in an environment where relevant features can be fully approximated by a classical model therefore does not gain from encoding quantum information.

In general, however, using a quantum memory allows for storing more relevant information without having to increase the size of the memory to do so. Moreover, we demonstrate that it is possible to find encodings with more information about Y than what is *maximally* achievable for any classical memory. We specify the quantum advantage in Sec. IV and analyze examples in Sec. V to verify that there can be a significant quantum advantage. For the examples, we compute entanglement and quantum discord present in the optimal encodings to shed light on the role played by quantum correlations.

We find that a quantum advantage is possible even when the map from X to Y breaks any entanglement with the memory (Sec. V B). We furthermore demonstrate in Sec. V C that in a quantum process with redundant information, irrelevant features get filtered out: The numerical algorithm is able to pick out and purify only the relevant information.

II. RELEVANT QUANTUM INFORMATION ENCODING

Given a information source X, how can one introduce the notion that some information is more relevant than the rest? In conventional rate-distortion coding, the goal is to encode the information in X with minimal distortion [23,28–31]. However, in practice, it may be difficult to define an appropriate distortion measure. In general, we may want to encode only those aspects of the data that we deem important and filter out the rest. This task is closely related to predictive inference, where we want to extract from X exactly those features that
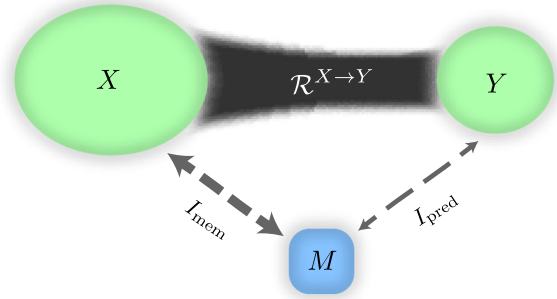

FIG. 1. We consider a data source X and define "relevance" by introducing a quantum channel, $\mathcal{R}^{X \to Y}$, to a second information source, Y. Information from X is encoded in a memory M, and only information about Y is deemed important. Relevant information is quantified by an information measure, I_{pred}, quantifying the correlations between M and Y. We refer to I_{pred} as "predictive power." The correlations between M and X, quantified by an information measure I_{mem}, we refer to as "memory." The latter quantifies the total encoded information. An encoding is considered more efficient if it has greater I_{pred} at the same I_{mem}.

are useful for making inferences about the outcome of some dynamical process that takes X as its input.

The information bottleneck method [26] gives a simple solution to this problem: Introduce a second information source Y, a "relevant variable," which depends on X, and let this dependence *define* relevance. Classically, the co-occurrence statistics of X and Y determine the available relevant information, which can be filtered out when the data are represented by a memory M. In a quantum mechanical generalization, two scenarios are possible: (i) Think of X and Y as two quantum systems in a joint state ρ_{XY}, or (ii) think of Y as output after sending X through a quantum channel, $\rho_Y = \mathcal{R}(\rho_X)$. We here focus on the latter scenario, illustrated in Fig. 1. Due to the causal relationship between X and Y in this scenario, we refer to information about X as "memory" and information about Y extracted from X as "predictive information." The total amount of encoded predictive information we call the encoding's "predictive power". The map $\mathcal{R}^{X \to Y}$ defines relevance, and we refer to it as "the relevance channel".

One cannot simultaneously send X through two independent quantum channels, one with Y as output and the other with M as output. That is, we cannot send X through both the relevance channel *and* an independent encoding channel. Classically, this problem does not arise, and one can always make a copy of X, but quantum mechanically, no such physical process exists in general [32]. We therefore consider a protocol similar to that used in quantum rate-distortion coding [28–31]: We take the input to the problem to be a purification of the state ρ_X by introducing a second quantum system that acts as a reference, R. Information is then encoded in the memory by mapping XR to MR. Subsequently, MR is mapped to MY via the relevance channel.

In the spirit of the classical information bottleneck method [26], we measure encoding cost and encoding quality using mutual information. Here we depart from the approach used in quantum rate-distortion coding, as we do not use the usual

qubit encoding rate to quantify cost and the usual entanglement fidelity to quantify quality [28–31]. Instead, we measure quality by predictive power, i.e., the relevant information quantifying the correlations between M and Y. Memory cost is measured by mutual information between M and X.

A. Notation and definitions

Hilbert spaces associated to quantum systems are denoted by $\mathcal{H}_A$, where the subscript is used to differentiate between systems. We assume that all Hilbert spaces have finite dimension and let d_A denote the dimension of system A, etc. The set of linear operators on $\mathcal{H}_A$ we denote $L(\mathcal{H}_A)$. We reserve the symbol $I_A \in L(\mathcal{H}_A)$ for the identity operator on A. A quantum state is a positive semidefinite operator $\rho_A \in L(\mathcal{H}_A)$, with unit trace. For a given state ρ_{AB} defined on a composite system, AB, we use the convention that ρ_A and ρ_B denote the reduced states, $\rho_A = \mathrm{tr}_B \rho_{AB}$ and $\rho_B = \mathrm{tr}_A \rho_{AB}$, respectively.

Quantum channels, i.e., completely positive and trace-preserving maps, are denoted by uppercase calligraphic letters, such as $\mathcal{E}^{A \to B} : L(\mathcal{H}_A) \to L(\mathcal{H}_B)$. The superscript is used to emphasize the input and output systems, but is left out when a channel is applied to a state, e.g., $\mathcal{E}(\rho_A)$, to simplify notation. We use $\mathcal{I}_A$ to denote the identity channel on system A. A given quantum channel, $\mathcal{E}^{A \to B}$, can be extended to an ancilla system, C, by tensoring with the identity channel. Such an extension we denote by putting a hat on the same symbol as that used for the original channel: $\hat{\mathcal{E}}^{AC \to BC} = \mathcal{E}^{A \to B} \otimes \mathcal{I}_C$.

A purification of a state, ρ_A, is a pure state $\psi_{AR} = |\psi_{AR}\rangle\langle\psi_{AR}|$, satisfying $\mathrm{tr}_R \psi_{AR} = \rho_A$. Here R is a second quantum system with dimension at least as large as A. There are many possible purifications, but we here fix a choice with R isomorphic to A,

$$|\psi_{AR}\rangle = \sum_i \sqrt{p_i} |i\rangle |i'\rangle, \tag{1}$$

where $\rho_A = \sum_i p_i |i\rangle\langle i|$ is the spectral decomposition of ρ_A and $|i\rangle$ and $|i'\rangle$ are orthonormal bases for the two systems A and R, respectively.

The von Neumann entropy of system A in a state ρ_A is defined as $S[A] = S(\rho_A) = -\mathrm{tr}[\rho_A \log_2 \rho_A]$. The mutual information of two systems, A and B, in a bipartite state, ρ_{AB}, is given in terms of von Neumann entropies,

$$I[A : B] = I(\rho_{AB}) = S(\rho_A) + S(\rho_B) - S(\rho_{AB}). \tag{2}$$

Mutual information was given an operational meaning in [33], where it was shown that it quantifies the smallest rate at which one must inject noise into the system for A and B to become uncorrelated in the usual asymptotic limit. It can therefore be taken as a measure of the total correlations between the two systems.

We make use of the Choi-Jamiołkowski isomorphism to represent channels in terms of positive operators. More specifically, we use the framework developed in [34] and introduce *conditional quantum operators* to represent quantum channels. That is, any channel $\mathcal{E}^{A \to B}$ we can write uniquely in terms of a positive operator $\mathcal{E}_{B|A} \in \mathcal{L}(H_A \otimes H_B)$, satisfying $\mathrm{tr}_B \mathcal{E}_{B|A} = I_A$ and [34]

$$\mathcal{E}(\rho_A) = \mathrm{tr}_A\big[\mathcal{E}_{B|A}^{T_A} \rho_A\big], \tag{3}$$

for all states ρ_A on A. T_A here denotes the partial transpose on A.

For *given* initial state ρ_A and channel $\mathcal{E}^{A \to B}$, we can define a bipartite quantum state on AB [34]:

$$\rho_{AB} = \rho_A^{1/2} \mathcal{E}_{B|A} \rho_A^{1/2}. \tag{4}$$

Vice versa, for any *given* bipartite state ρ_{AB} we can define a channel through a conditional quantum operator

$$\mathcal{E}_{B|A} = \rho_A^{-1/2} \rho_{AB} \rho_A^{-1/2}. \tag{5}$$

This framework allows us to switch back and forth between channels and states for representing the evolution of a given initial state.

Classical probability distributions we denote by $p(x)$ and $q(x)$, and conditional probability distributions we denote by $p(y|x)$ and $q(y|x)$. A conditional probability distribution satisfies $\sum_y p(y|x) = 1$ for all x. A state on a bipartite system XY that can be written

$$\rho_{XY} = \sum_{x,y} p(x,y) |x\rangle\langle x| |y\rangle\langle y| \tag{6}$$

for a joint probability distribution $p(x,y)$ and orthonormal bases $|x\rangle$ for $\mathcal{H}_X$ and $|y\rangle$ for $\mathcal{H}_Y$, we call a *classical state*, reflecting that X and Y are only classically correlated. A state that can be written

$$\rho_{XY} = \sum_y p(y) \rho_{X|y} |y\rangle\langle y|, \tag{7}$$

where $\rho_{X|y}$ are arbitrary states of system X, we call *quantum classical*, reflecting that X can be quantum while Y is classical. Similarly, a quantum channel, $\mathcal{E}^{X \to Y}$, is said to be *classical* if it can be written in terms of a conditional quantum operator of the form

$$\mathcal{E}_{Y|X} = \sum_{x,y} p(y|x) |y\rangle\langle y| |x\rangle\langle x|, \tag{8}$$

where $p(y|x)$ is a conditional probability distribution. If this classical channel is applied to a state of X that is diagonal in the same basis $|x\rangle$, $\rho_X = \sum_x p(x) |x\rangle\langle x|$, we call this a *classical process*. The output is in this case just $\rho_Y = \sum_y p(y) |y\rangle\langle y|$, where $p(y) = \sum_x p(x,y) = \sum_x p(x) p(y|x)$.

B. Relevant information optimization problem

Denote by ρ_X a state of a quantum system X, with purification ψ_{XR}. This state represents the data from which we wish to extract information. The reference system R is sent through the *relevance channel*, $\mathcal{R}^{R \to Y}$, with output system Y. We seek an optimal encoding of the information in X to a memory, M, in terms of a quantum channel $\mathcal{E}^{X \to M}$ in the sense that we wish to retain as much information about Y as possible, without storing any unnecessary data.

The protocol we consider is similar to that used in quantum rate-distortion coding [28–31]. Starting from the purification, ψ_{XR}, information is encoded in the memory by a map,

$$\hat{\mathcal{E}}^{XR \to MR} = \mathcal{E}^{X \to M} \otimes \mathcal{I}_R. \tag{9}$$

This map is an extension of the local encoding map, $\mathcal{E}^{X \to M}$, to the reference system R. Information is now stored in the state

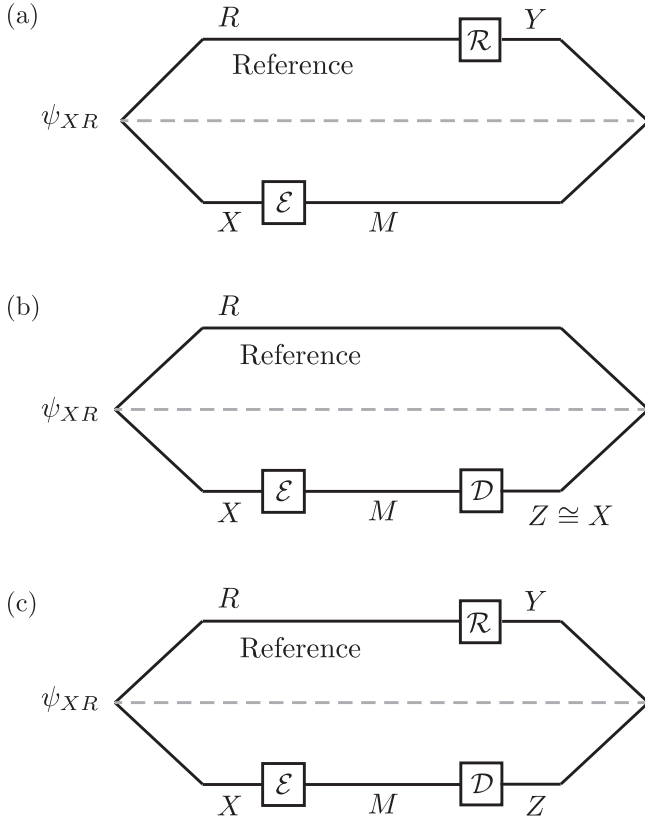$$\rho_{MR} = \hat{\mathcal{E}}(\psi_{XR}). \tag{10}$$

FIG. 2. An illustration of the causal structure of three related coding protocols. In each case the input is a purification ψ_{XR} of the initial data ρ_X. The protocols differ only by local maps applied to either the data, X, or the purification reference, R. (a) The "relevance coding protocol" considered in this paper. The data are encoded in a memory by an encoding map, $\mathcal{E}^{X \to M}$, while a "relevance channel" $\mathcal{R}^{R \to Y}$ is applied to the reference to define relevant information. (b) The conventional protocol for quantum rate-distortion coding. No "relevance channel" is used, and all of the information in the initial data is considered equally important. A decoding map, $\mathcal{D}^{M \to Z}$, needs to be introduced to measure the fidelity of the encoded and original data. (c) A generalized rate-distortion coding protocol, of which the two protocols in (a) and (b) are special cases.

Subsequently, the reference is sent through the relevance channel $\mathcal{R}^{R \to Y}$, which we extend to the memory system,

$$\hat{\mathcal{R}}^{MR \to MY} = \mathcal{I}_M \otimes \mathcal{R}^{R \to Y}, \tag{11}$$

so that information about Y is finally encoded in the state

$$\rho_{MY} = \hat{\mathcal{R}}(\rho_{MR}). \tag{12}$$

The protocol is depicted in Fig. 2(a).

We use mutual information to quantify both the information encoded about the initial data, which we refer to as *memory*,

$$I_{\text{mem}} = I_{\text{mem}}[\rho_{MR}] = I[M : R], \tag{13}$$

and the information available about the output on Y, which we refer to as *predictive power*,

$$I_{\text{pred}} = I_{\text{pred}}[\rho_{MR}] = I[M : Y]. \tag{14}$$

Following the classical information bottleneck method, we wish to filter out the relevant information about Y.

We do so by maximizing I_{pred}. However, keeping more relevant information comes at the cost of increased memory, I_{mem}. We define an *optimal* encoding as one that maximizes I_{pred}, while simultaneously obeying a constraint on I_{mem}. Optimal encodings therefore must be solutions to the following optimization problem:

$$\max_{\mathcal{E}^{X \to M}} I_{\text{pred}} \quad \text{such that} \quad I_{\text{mem}} \leqslant \text{const.} \tag{15}$$

It should be noted that the dual problem of minimizing I_{mem} subject to a constraint on I_{pred} is essentially equivalent, in the sense that for a given value of I_{mem} and I_{pred} the same solution, ρ_{MR}, is found.

Before we discuss solutions to this optimization problem, let us pause to consider the relationship to quantum rate-distortion coding in more detail [28–31]. There, one also considers an information source X in a state ρ_X, with purification ψ_{XR}, and seeks to find a minimal encoding of X into a memory, M, given by a map $\mathcal{E}^{X \to M}$. The encoding is subject to a constraint quantifying the allowed degree of distortion. To measure the distortion, one introduces a decoding map, $\mathcal{D}^{M \to Z}$, where the system Z is isomorphic to X. A distortion function measuring the fidelity of the decoded signal with the original, can then be introduced on the final output space of Z and the reference system R. The protocol is depicted in Fig. 2(b). The goal in rate-distortion coding is to find the most compact encoding by minimizing the dimension of the memory Hilbert space, d_M [35]. We here have a different goal, by maximizing a mutual information measure subject to a constraint. We do, however, note that we will find optimal encodings for which d_M is the smallest possible value that allows for reaching the maximum of the constrained optimization problem.

Another key difference between the protocol considered in this paper and quantum rate-distortion coding is that in the latter, distortion is measured relative to the original data source, X, whereas we here measure it relative to a *second* information source, Y. We seek to extract only the information from X that is relevant for making an inference about Y. Conceptually, one could say that we measure average distortion by the negative mutual information, $-I_{\text{pred}} = -I[M : Y]$ (large I_{pred} meaning low distortion) [36]. Therefore, we do not need to introduce a decoding map to define distortion. More generally though, a decoding scheme can be introduced by introducing a map, $\mathcal{D}^{M \to Z}$, following $\mathcal{E}^{X \to M}$. This generalized protocol is depicted in Fig. 2(c).

C. Optimal encodings

Finding the solutions to the optimization problem in Eq. (15) is equivalent to solving

$$\rho_{MR}^{\text{opt}} = \arg \max_{\sigma_{MR}} \{ I[\hat{\mathcal{R}}(\sigma_{MR})] - \alpha I(\sigma_{MR}) - \text{tr}_R[\Lambda_R \sigma_R] \} \tag{16}$$

(this is shown in Appendix A). Here $\alpha \geqslant 0$ and Λ_R are Lagrange multipliers, where Λ_R is a Hermitian operator on R. α is for the constraint on $I_{\text{mem}} = I[M : R]$, and Λ_R is for the constraint that $\sigma_R = \text{tr}_M \sigma_{MR}$ must equal the given initial state ρ_R (denoting the reduced state $\rho_R = \text{tr}_X \psi_{XR}$, which is

identical to the given initial state ρ_X, but defined on system R). Recall that $\hat{\mathcal{R}}^{R\to Y}$ is the relevance channel; see Eq. (12).

An optimal encoding $\mathcal{E}_{\mathrm{opt}}^{X\to M}$ can be constructed from an optimal state found in Eq. (16), ρ_{MR}^{opt}, via the conditional quantum operator,

$$\mathcal{E}_{M|X}^{\mathrm{opt}} = \rho_R^{-1/2} \rho_{MR}^{\mathrm{opt}} \rho_R^{-1/2}. \tag{17}$$

The optimal encoding map is represented as

$$\mathcal{E}_{\mathrm{opt}}(\sigma_X) = \mathrm{tr}_X\left[\left(\mathcal{E}_{M|X}^{\mathrm{opt}}\right)^{T_X}\sigma_X\right]. \tag{18}$$

Since X and R are isomorphic systems, we have introduced a slight abuse of notation by referring to the system X on the left-hand side of Eq. (17) and R on the right-hand side. This simply serves to remind us that $\mathcal{E}_{M|X}$ is used to define an encoding map from X to M.

As shown in Appendix A, an optimal encoding satisfies the following self-consistent equation

$$\mathcal{E}_{M|X}^{\mathrm{opt}} = Z_R^{-1/2} \exp\left(\log_2 \rho_M^{\mathrm{opt}} - \frac{1}{\alpha}H_{MR}\right)Z_R^{-1/2}, \tag{19}$$

where we have introduced

$$H_{MR} = \hat{\mathcal{R}}^\dagger\left(\log_2 \rho_M^{\mathrm{opt}} + \log_2 \rho_Y^{\mathrm{opt}} - \log_2 \rho_{MY}^{\mathrm{opt}}\right), \tag{20}$$

$$Z_R = \mathrm{tr}_M\left[\exp\left(\log_2 \rho_M^{\mathrm{opt}} - \frac{1}{\alpha}H_{MR}\right)\right], \tag{21}$$

and $\hat{\mathcal{R}}^{\dagger Y\to R}$ is the dual map of $\hat{\mathcal{R}}^{R\to Y}$ [39]. We have suppressed explicitly writing out identity operators in tensor products in Eqs. (19)–(21) to keep the notation as simple as possible.

Equation (19) states an implicit relation that any optimal encoding must satisfy. In Sec. III we use this relation to derive several important properties of optimal encodings. Furthermore, Eq. (19) forms the basis for an iterative algorithm to explicitly find $\mathcal{E}_{\mathrm{opt}}^{X\to M}$ for a given input state and relevance channel. In Sec. V we use this algorithm to find optimal encodings for a series of examples and study the properties of these encodings. The iterative algorithm itself is discussed in Appendix C . Equation (19) reduces to the well-known classical result from [26] if the states are purely classically correlated (see Appendix A 3).

III. BASIC PROPERTIES OF OPTIMAL ENCODINGS

A. Structure in the small and large α limits

In the limit of $\alpha \to 0$, the second term in the exponential in Eq. (19) dominates. The operator thus has a form analogous to that of a thermal state, with α playing the role of temperature and H_{MR} the role of Hamiltonian. As shown in Appendix B, the expectation value of H_{MR}, is the negative predictive power:

$$\langle H_{MR}\rangle = \mathrm{tr}[H_{MR}\rho_{MR}] = -I_{\mathrm{pred}}[\rho_{MR}]. \tag{22}$$

We therefore interpret the solution in this limit as a "ground state" that maximizes I_{pred} by minimizing $\langle H_{MR}\rangle$. This is, of course, consistent with the fact that the constraint on I_{mem} in the optimization problem vanishes as $\alpha \to 0$.

In the classical case, the solution found in this limit specifies a deterministic hidden Markov model, which is a minimal representation with sufficient statistics [3,40]. In the quantum case, this solution can be seen as a generalization of the concept of purification: While a purification is a solution that allows for retaining all of the relevant information, there are other, more efficient solutions. A purification would follow from the choice $\mathcal{R}^{R\to Y} = \mathcal{I}_R$. The optimally predictive encoding, in contrast, is maximally correlated with the relevant information, as defined by the second information source, but uncorrelated with the irrelevant information.

In the large α limit, there is an infinite emphasis on compression. Then, the first term in the exponential dominates, and the term proportional to $1/\alpha$ can be neglected. Equation (19) then implies that $\mathcal{E}_{M|X}^{\mathrm{opt}} = \rho_M \otimes I_X$, where ρ_M is arbitrary, and $I_{\mathrm{pred}} = I_{\mathrm{mem}} = 0$ in this limit.

B. No quantum advantage for classical processes

Let $|x\rangle$ and $|y\rangle$ be orthonormal basis sets for the two systems R and Y, respectively (recall that R is isomorphic to X). Recall that the relevance map, $\mathcal{R}^{R\to Y}$, is said to be *classical* if it can be written in terms of a conditional quantum operator of the form

$$\mathcal{R}_{Y|R} = \sum_{x,y} p(y|x)|y\rangle\langle y||x\rangle\langle x|, \tag{23}$$

where $p(y|x)$ is a conditional probability distribution. If, furthermore, the classical relevance channel in Eq. (23) is applied to an initial state ρ_R that is diagonal in the basis $|x\rangle$, $\rho_R = \sum_x p(x)|x\rangle\langle x|$, we call this a *classical process*, since it can be described in terms of a stochastic map from a classical probability distribution to a classical probability distribution.

An optimal encoding channel for classical processes is always of the *classical* form

$$\mathcal{E}_{M|X}^{\mathrm{opt}} = \sum_{m,x} p(m|x)|m\rangle\langle m||x\rangle\langle x|, \tag{24}$$

corresponding to an optimal state of the memory and reference

$$\rho_{MR}^{\mathrm{opt}} = \sum_{m,x} p(x)p(m|x)|m\rangle\langle m||x\rangle\langle x| \tag{25}$$

(this is shown in Appendix B). This means that *if* the relevance map deems only classical information about X to be important, *then* quantum correlations are superfluous and should be filtered out in an optimal data representation.

By encoding information into nonorthogonal states of the memory, it is possible to reduce the entropy, $S(\rho_M)$, of the memory, without loss of predictive power. This was exploited in [41] to construct a memory of smaller entropy than the best classical model, with no loss of information. However, we show in Appendix B that for a classical process, a quantum memory cannot achieve higher predictive power as quantified by I_{pred} than the best classical data representation at the same I_{mem}. Hence, whenever the relevant data are purely classical there is no quantum advantage to predictive inference as we have defined it here.

This raises the question of whether the entropy of the memory is, in itself, a good measure of complexity [41]? Our view is that rather the correlations between the memory and the

source should be considered. Encoding classical information into nonorthogonal quantum states does not require fewer input bits, even if the entropy of the memory alone can be lower. For classical states of the form Eq. (25), it is $I_{mem} = I[M : R]$ that correctly quantifies the encoding rate in classical bits [23,26].

IV. QUANTIFYING QUANTUM ADVANTAGE: THE INFORMATION PLANE

We do expect a quantum memory to generically perform better for nonclassical processes. To quantify the quantum advantage, we compare the optimal solution of the quantum problem to the optimal solution when the memory is restricted to be classical.

More precisely, we say that the memory is classical if the encoding maps leaves MR in a *classical-quantum* state σ_{MR} of the form

$$\sigma_{MR} = \sum_m p(m)|m\rangle\langle m| \otimes \sigma_{R|m}, \qquad (26)$$

for a basis $\{|m\rangle\}$ on M, probability distribution $p(m)$, and arbitrary states $\sigma_{R|m}$ on R, with the constraint $\mathrm{tr}_M \sigma_{MR} = \rho_R$. The optimal classical encoding is thus defined to be the solution to the optimization problem [Eq. (16)], when the optimization is over the restricted set of states given by Eq. (26). This restriction is discussed further in Appendix A 2.

We measure quantum advantage using the information plane [26]. The optimal values of I_{mem} and I_{pred} trace out a convex curve in the plane spanned by I_{mem} and I_{pred} as the Lagrange multiplier α is varied. We can find this curve numerically using the iterative algorithm presented in Appendix C. Whenever we need to distinguish between quantum and classical memories, we denote the values of $I_{mem(pred)}$ by $I^Q_{mem(pred)}$ and $I^C_{mem(pred)}$, evaluated for a quantum and a classical memory, respectively. The optimal curves for classical and quantum memories can, in general, be different, with the quantum curve lying above the classical. This generic situation is illustrated in Fig. 3.

We have the following general bounds on I_{mem} and I_{pred}.

(1) The memory of a purification, $I_{mem}[\psi_{XR}] = 2S[X]$, upper bounds I_{mem} for a any memory, quantum or classical, $I^{C/Q}_{mem} \leqslant I_{mem}[\psi_{XR}]$.

(2) The predictive power of a purification, $I_{pred}[\psi_{XR}]$, upper bounds I_{pred} for any memory, quantum or classical, $I^{C/Q}_{pred} \leqslant I_{pred}[\psi_{XR}]$.

(3) The entropy $S[X]$ of the initial data upper bounds I_{mem} for a classical memory, $I^C_{mem} \leqslant S[X]$.

The last bound follows from writing $I_{mem} = S[R] - S[R|M]$, where $S[R|M] = S(\rho_{MR}) - S(\rho_M)$ is always greater or equal to zero for a classical memory of the form Eq. (26) (recall that $S[R] = S[X]$). A quantum memory can beat this bound by having negative values of the conditional entropy $S[R|M]$ [42,43].

The bounds introduced above, together with the curve traced out by the respective quantum and classical optimal encodings, can be used to define three regions of the information plane: (1) the quantum feasible region containing all achievable points for quantum states satisfying the constraint $\mathrm{tr}_M \sigma_{MR} = \rho_R$; (2) the classically feasible region, a subset thereof, which consists of the points achievable for states
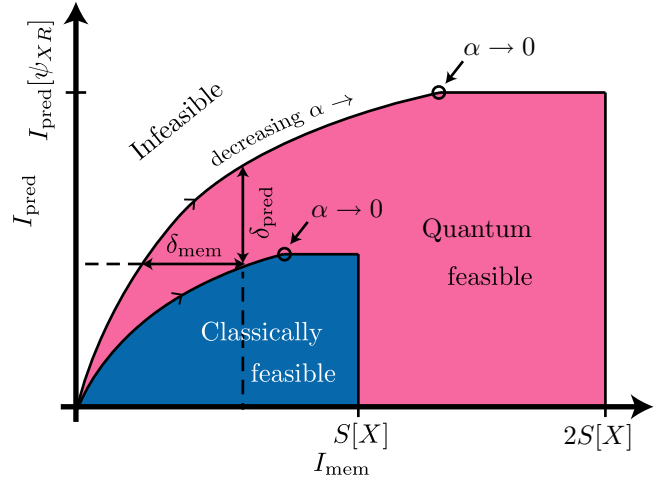

FIG. 3. The information plane is spanned by I_{mem} on the horizontal and I_{pred} on the vertical axis. It can be divided into three regions: the infeasible, the quantum feasible, and the classically feasible region. These regions can be mapped out using the quantum and classical upper bounds on I_{mem} and the optimal curves found using the quantum information bottleneck method. The optimal quantum curve generally lies above the optimal classical curve, and we quantify the gap by the two positive measures δ_{pred} and δ_{mem}, as illustrated in the figure. The circles indicate the $\alpha \to 0$ limits of the optimal quantum and classical solutions.

of the form of Eq. (26) that satisfy the same constraint; (3) the infeasible region containing those points corresponding to values of I_{mem} and I_{pred} that cannot be reached for any state satisfying the constraint. Figure 3 illustrates these regions. The boundaries of the feasible regions in the information plane allow us to read off the maximum predictive power at a given memory cost, as quantified by I_{mem}, for both quantum and classical memories.

We quantify the quantum advantage by comparing the optimal points achievable for a quantum memory to those of a classical memory. We introduce two measures.

(1) $\delta_{pred}[I_{mem}] = I^Q_{pred}[I_{mem}] - I^C_{pred}[I_{mem}] \geqslant 0$, which we define to be the difference between the predictive power of the best quantum model and the best classical model at a given I_{mem}.

(2) $\delta_{mem}[I_{pred}] = I^C_{mem}[I_{pred}] - I^Q_{mem}[I_{pred}] \geqslant 0$, defined to be the difference in memory between the best classical and best quantum model at a given I_{pred}.

δ_{pred} is naturally only defined for $I_{mem} \leqslant S[X]$, and δ_{mem} is defined only for I_{pred} smaller than the maximum achievable value for a classical memory. We emphasize that δ_{pred} and δ_{mem} quantify quantum advantage compared to *any* classical memory of arbitrary dimension d_M. This is in contrast to quantum discord [44,45], which in a sense compares a quantum memory to classical memories of the same dimension: Discord can be thought of as comparing a quantum encoding to an optimal classical encoding achievable through sending the quantum memory through a decoherence channel [12]. Discord thus compares quantum bits to classical bits on a "bit-for-bit" basis, which is not necessarily a fair comparison in the present context.

V. EXAMPLES

We consider three simple examples to illustrate how the quantum information bottleneck method can be used in practice and to showcase some key features of optimal encodings. All optimal encodings are found numerically, using the algorithm in Appendix C.

The first example is a purely classical process. It serves to illustrate the fact that there is no quantum advantage for classical prediction problems, as indicated by our analytical results (Sec. III B). We verify numerically that the optimal solutions for an arbitrary quantum memory are in agreement with those for a classical memory [46] by tracing out the optimal curves in the information plane, showing that they overlap precisely.

The second example illustrates a process with a classical-to-quantum transition for the optimal encoding. A quantum bit is being sent through a phase damping channel. For compression values below the classical maximum, $I_{mem} \leqslant S[X]$, the optimal classical memory performs just as well as the quantum memory. However, the predictive power so achieved does not exceed 65% of the maximum possible predictive power. The optimal quantum memory can break through this classical barrier by taking on negative values of the quantum conditional entropy, $S[R|M]$. Thereby, the quantum memory can achieve *full* predictive power. This illustrates one of the main features of a quantum memory: It can achieve full predictive power on quantum processes where a classical memory cannot.

The third example shows a quantum process with redundant information. In this example, there is a quantum advantage for *any* value of retained memory. We quantify this quantum advantage. The redundant information (containing no utility for prediction) gets filtered out by the iterative algorithm that constructs the optimal encodings. Full predictive power is obtained at substantial compression: The maximum value of $I_{pred}[\psi_{XR}]$ is reached for a memory with almost three times less memory cost, $I_{mem} < I_{mem}[\psi_{XR}]$. We also find that this is achievable for a memory of smaller dimension than the initial data, $d_M < d_X$.

A. Even process: A classical process

Our first example illustrates that there is no quantum advantage for predicting a classical process. The example system is a hidden Markov process called the *even process* [47]. The predictive compressibility characteristics of this process were studied for a classical memory in [46]. We show here that the optimal quantum memory has identical features.

The even process outputs all binary strings consisting of an even number of 1's bounded by 0's and associates a probability to any bit string by choosing either a 0 or a 1 with fair probability after having generated either a 0 or a pair of 1's. Consider the problem of predicting a future string of generated bits, based on having observed previous output strings. We chose the initial data to be the set of all bit strings, x, of length three, and we associate with each of them a probability, $p(x)$, equal to the frequency at which the even process generates the respective string. The relevance map is taken to be the map induced by the even process generating two new output bits from each respective previous sequence
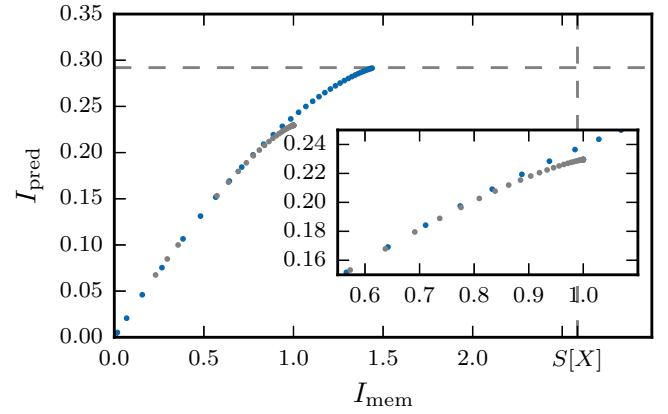

FIG. 4. Optimal curve in the information plane for the classical even process. Classical and quantum solutions coincide exactly. Dots show numerical results for a memory of dimension $d_M = 3$ (blue) and $d_M = 2$ (gray) for comparison. (Inset) Magnification of the region where the solutions for $d_M = 3$ and $d_M = 2$ bifurcate.

of three bits. To quantize this problem, we associate a set of pure orthogonal states $|x\rangle$, one for each history x, and take the state $\rho_X = \sum_x p(x)|x\rangle\langle x|$ as initial data. There are eight possible bit strings of length three, so $d_X = 8$. Similarly, for the output we associate a set of pure orthogonal states $|y\rangle$, one for each possible two-bit output (so that $d_Y = 4$). The relevance map can then be defined by a conditional quantum operator $\mathcal{R}_{Y|R} = \sum_{x,y} p(y|x)|y\rangle\langle y||x\rangle\langle x|$, where $p(y|x)$ is the stochastic map induced by the even process from the set of histories, x, to the set of futures, y.

For both a quantum memory and a classical memory, respectively, we compute the family of optimal encodings with the iterative algorithm and the deterministic annealing scheme described in Appendix C. The optimal values for the two cases exactly coincide, as shown in Fig. 4. The blue dots are the numerically found optimal values of I_{mem} and I_{pred}, for *both* the quantum and the classical memory as α is changed from a large value (lower left corner) to a small value. The maximal value of I_{pred} is reached by a memory of dimension $d_M = 3$. Our results are in agreement with [46].

The gray dots in Fig. 4 show the results for a memory of dimension $d_M = 2$. This memory size is not large enough to reach the maximum value of I_{pred}, and the solution with $d_M = 3$ bifurcates from that with $d_M = 2$ as α is lowered, shown in the inset of Fig. 4. This illustrates that the algorithm can be used not only to find an optimal encoding at a fixed memory dimension, d_M, but also find the smallest possible d_M reaching maximal predictive power.

The maximal value of I_{pred} is reached at a value of $I_{mem} \simeq 1.45$ bits, which is only 56% of the classical entropy of the input data, $S[X] \simeq 2.6$ bits. Hence, significant compression of the input data is possible, without any loss of predictive power [46]. Note that the maximum quantum value is $2S[X] \simeq 5.2$ bits. This large value comes from the low degree of purity of the initial data, allowing a large degree of entanglement with the memory. Such entanglement is wasteful for predicting a classical process, and the algorithm correctly filters out this entanglement.

B. Causally incompressible quantum process

This example serves to illustrate that there is a quantum advantage to predictive inference, even for quantum processes that are not causally compressible.

The even process is an example of a (classical) causally compressible process [48]. If a model can be constructed such that full predictive power is achieved at a memory less than the maximum set by the entropy of the data (in the classical case), then a process is fully causally compressible [48]. This definition is easily extended to quantum memories, with the maximum value for the memory now given by the memory of a purification, $I_{\text{mem}}[\psi_{XR}]$.

We look at a phase-damping qubit channel. Phase damping describes loss of quantum information over time through decay of off-diagonal matrix elements for a quantum state. For a single qubit, the phase-damping channel can be represented by an operator-sum,

$$\mathcal{R}(\sigma_R) = K_1 \sigma_R K_1^\dagger + K_2 \sigma_R K_2^\dagger, \tag{27}$$

$$K_1 = \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{1-\lambda} \end{pmatrix}, \tag{28}$$

$$K_2 = \begin{pmatrix} 0 & 0 \\ 0 & \sqrt{\lambda} \end{pmatrix}, \tag{29}$$

where $\lambda = 1 - \exp(-t/T_2)$ is a probability that grows with time, T_2 being the qubit "coherence time."

As initial data we take the quantum state

$$\rho_X = (1-p)|+\rangle\langle+| + p\frac{I_2}{2}, \tag{30}$$

where $|+\rangle = (|0\rangle + |1\rangle)/\sqrt{2}$ and I_2 is the qubit identity operator. We choose parameters $\lambda = 0.5$ and $p = 0.3$. The qualitative nature of the results does not depend sensitively on the choice of parameters, but of course with a pure ρ_X ($p = 0$) we would necessarily have $I_{\text{mem}}[\psi_{XR}] = 0$, and therefore no predictive compression would be possible.

Numerical results for optimal classical and quantum memories are shown in Fig. 5, with blue and pink dots for the classical and quantum case, respectively. Panel (a) shows the optimal curves in the information plane. The predictive power of a classical memory is limited; it cannot exceed $I_{\text{pred}} \simeq 0.46$ bits. This is only 65% of the maximum possible predictive power of $I_{\text{pred}} \simeq 0.82$ bits. The quantum curve falls exactly on top of the classical curve, as long as $I_{\text{mem}} \leqslant S[X] \simeq 0.61$. Beyond this value, we enter the classically infeasible region. The quantum memory is, however, able to break through this point. The quantum memory reaches the maximum possible predictive power, albeit requiring full complexity, $I_{\text{mem}} \simeq 1.2$ bits—twice the complexity of the best classical solution.

The point where the quantum memory goes beyond what is classically possible is associated with negative values of $S[R|M]$, and a sharp drop in the entropy of the memory, $S[M]$, as shown in panel (b) of Fig. 5. This indicates a classical-quantum transition for the optimal encoding. This transition is further accompanied by a jump in purity at this point, which is shown in panel (c) of Fig. 5. Note that in the limit of maximal predictive power, the optimal quantum solution is a purification of ρ_R, thus keeping all aspects of the original data and hence having full predictive power at full memory cost.
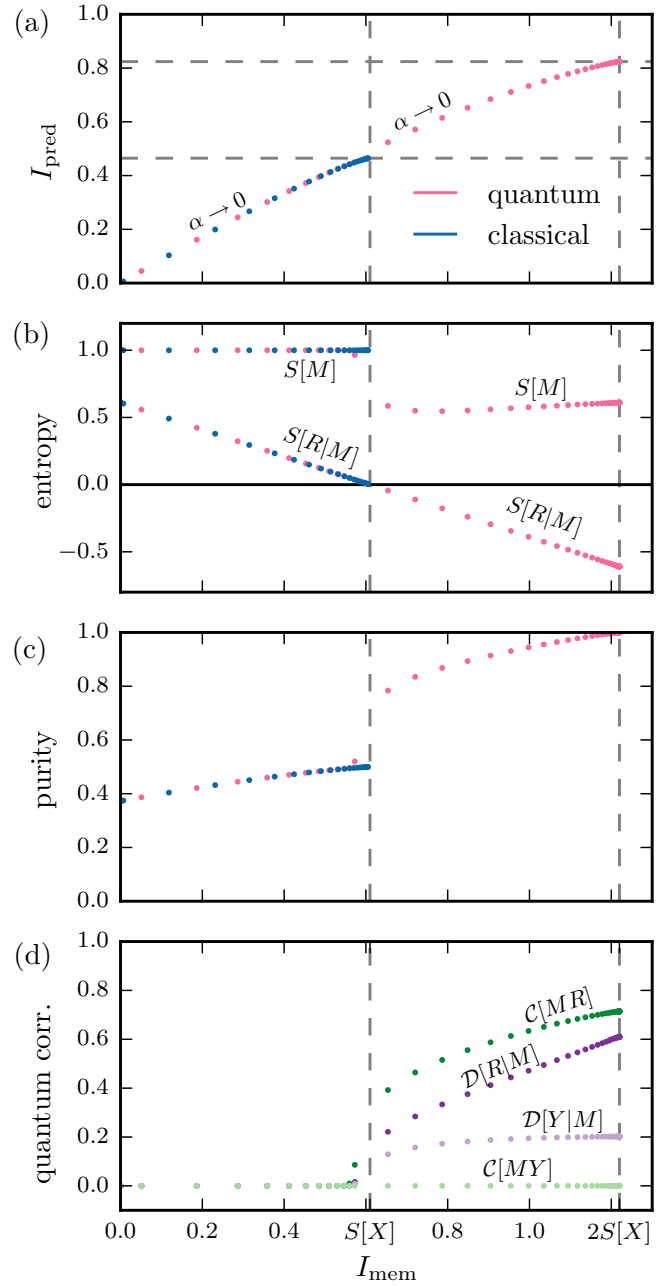

FIG. 5. Characteristics of optimal encodings for a qubit phase damping channel. (a)–(c) Quantum solutions in pink, classical in blue. (a) Optimal solutions depicted in the information plane. (b) Conditional entropy $S[R|M]$ and memory entropy $S[M]$. (c) Purity of the optimal solution ρ_{MR}^{opt}. (d) Quantum correlations in the optimal solutions, quantified by concurrence (green) and quantum discord (purple), for the input (dark) and output (bright) systems, MR and MY, respectively.

The classical-quantum transition also comes along with a change in other quantities. In panel (d) of Fig. 5, we plot two different measures of quantum correlations for the optimal solutions. As a measure of entanglement between the the data qubit and the memory qubit, we plot the concurrence [32],

$$\mathcal{C}[MR] = \mathcal{C}(\rho_{MR}) = \max(0, \lambda_1 - \lambda_2 - \lambda_3 - \lambda_4), \tag{31}$$

where $\lambda_1, \ldots, \lambda_4$ are the eigenvalues, in decreasing order, of the Hermitian matrix $\sqrt{\sqrt{\rho_{MR}}\tilde{\rho}_{MR}\sqrt{\rho_{MR}}}$. Here $\tilde{\rho}_{MR} = (\sigma_y \otimes \sigma_y)\rho_{MR}^*(\sigma_y \otimes \sigma_y)$ and σ_y is the Pauli-y matrix. We also plot the quantum discord [44,45] of the memory,

$$\mathcal{D}[R|M] = \mathcal{D}(\rho_{MR}) = S(\rho_M) - S(\rho_{MR})$$
$$+ \min_{\{P(m)\}} \sum_m p(m) S(\rho_{R|m}), \tag{32}$$

where the minimization is over all sets of rank 1 projectors on M, $\{P(m) = |m\rangle\langle m| \otimes I_R\}$, such that $\sum_m P(m) = I_{MR}$, and $p(m) = \mathrm{tr}[P(m)\rho_{MR}]$, $\rho_{R|m} = \mathrm{tr}_M[P(m)\rho_{MR}P(m)]/p(m)$.

The discord is nonzero only for a nonclassical memory, but is not a measure of nonlocal correlations like entanglement, as it can be created locally [49]. Rather, it stems from encoding information into nonorthogonal states. In panel (d) of Fig. 5 we see that $\mathcal{C}[MR]$ and $\mathcal{D}[R|M]$ are zero for values of I_mem below the classical-quantum transition and nonzero above. In the $\alpha \to 0$ limit, the quantum memory shows a high degree of entanglement and discord. We also plot the concurrence and the discord for the output state ρ_{MY}^opt [replacing R with Y in Eqs. (31) and (32)]. Interestingly, we see that the relevance channel breaks the entanglement such that $\mathcal{C}[MY]$ is zero for all values of I_mem, while the discord, $\mathcal{D}[Y|M]$, is nonzero above the classical-quantum transition. This demonstrates a quantum predictive advantage even for an entanglement breaking process. The nonzero discord of the output state shows that there is a significant degree of "quantumness" of the predictive information encoded in the memory.

C. Quantum process with redundant information

Our last example shows that redundant information is filtered out for fully causally compressible quantum processes. We take the initial data to be represented by two uncorrelated qubits. The relevance map acts on only one of them, while the other is discarded. The relevance channel is chosen in such a way that it can map orthogonal to nonorthogonal states, and is known to be able to create quantum discord [49], even though it is a purely local map. In this example, there is a predictive advantage to having a quantum memory for *any* complexity $I_\mathrm{mem} > 0$: The optimal quantum curve lies strictly above the optimal classical curve in the information plane.

The initial data consists of two qubits, $X = X_1X_2$, where the first is in the state given in Eq. (30) and the second is in a maximally mixed state:

$$\rho_{X_1X_2} = \left[(1-p)|+\rangle\langle+| + p\frac{I_2}{2}\right] \otimes \frac{I_2}{2}. \tag{33}$$

The relevance channel is taken to be an amplitude damping channel on the first qubit, while the second qubit is traced out,

$$\mathcal{R}(\sigma_{R_1R_2}) = \mathrm{tr}_{R_2}\left[K_1\sigma_{R_1R_2}K_1^\dagger + K_2\sigma_{R_1R_2}K_2^\dagger\right], \tag{34}$$

$$K_1 = \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{1-\lambda} \end{pmatrix} \otimes I_2, \tag{35}$$

$$K_2 = \begin{pmatrix} 0 & \sqrt{\lambda} \\ 0 & 0 \end{pmatrix} \otimes I_2. \tag{36}$$

For the parameters we choose $\lambda = 0.7$ and $p = 0.3$. Again, the results do not depend sensitively on this choice of parameters.

Clearly, the second qubit is irrelevant for making an inference about the first qubit. The choice of a maximally mixed state for the second qubit in Eq. (33) is made to have a high degree of mixedness in the initial data, which allows for large degree of correlation with a memory. However, information pertaining to the second qubit is redundant, and we use this example to show that the numerical algorithm filters out this information.

In Fig. 6(a) we show the optimal quantum and classical curves in the information plane, with the quantum curve lying strictly above the classical. We see that a very high degree of compression is possible. Importantly, the optimal quantum memory reaches maximum predictive power for $I_\mathrm{mem} \simeq 1.2$ bits, well below both the quantum and the classical maximal values of roughly 3.2 and 1.6 bits, respectively. The maximum is reached for a memory of dimension $d_M = 2$, showing that the irrelevant quantum bit has been filtered out from the initial data.

The classical memory is limited to a maximum predictive power of $I_\mathrm{pred} \simeq 0.16$ bits, which is roughly 38% of the maximum quantum value of 0.42 bits.

To gain further insight into the origin of the quantum advantage, we plot in panel (b) of Fig. 6 the conditional entropy $S[R|M]$ and the memory entropy $S[M]$. In contrast to the previous example, we now see that the quantum solution has a significantly lower memory entropy than the best classical solution, indicating that information is packed more densely into the quantum memory.

In Fig. 6(c), we plot the purity, $\mathrm{tr}[\rho_{MR}^2]$, of the quantum and the classical solutions. In contrast to the previous example, the optimal quantum solution in the $\alpha \to 0$ limit is no longer a purification of the initial data, due to the presence of the irrelevant data qubit in a highly mixed state. The maximum value of the purity for the quantum solution is found to be 0.5. This degree of mixedness comes solely from the entropy of the irrelevant qubit. Indeed, if we trace out this qubit, we find that the memory and the relevant data qubit are in a pure state: The optimal quantum solution in the $\alpha \to 0$ is found to be $\rho_{MR}^\mathrm{opt}[\alpha \to 0] = |\psi_{MR_1}\rangle\langle\psi_{MR_1}| \otimes I_2/2$, where $|\psi_{MR_1}\rangle$ refers to a purification of the first data qubit. This further illustrates how the irrelevant data qubit has been filtered out, and the numerical algorithm is able to pick out and purify only the relevant information.

The lower entropy of the quantum memory compared to the classical, illustrated in Fig. 6(b), indicates that a predictive advantage is related to information encoded into nonorthogonal states. We therefore expect a nonzero quantum discord for the memory. Since we are not interested in quantum correlations between the memory and the irrelevant second data qubit, we trace this qubit out and consider the discord and the concurrence of the memory and the first, relevant, data qubit, which is denoted R_1. The concurrence and discord are plotted in Fig. 6(d). We show both the quantum correlations of the initial system, MR_1, and the output, MY. A high degree of both entanglement and discord in the initial and the final optimal states is seen to be necessary for an optimal encoding.

Neither discord nor concurrence should be taken as measures of predictive quantum advantage. For this, we consider the measures δ_pred and δ_mem introduced in Sec. IV. These quantities can essentially be read off from panel (a) of Fig. 6
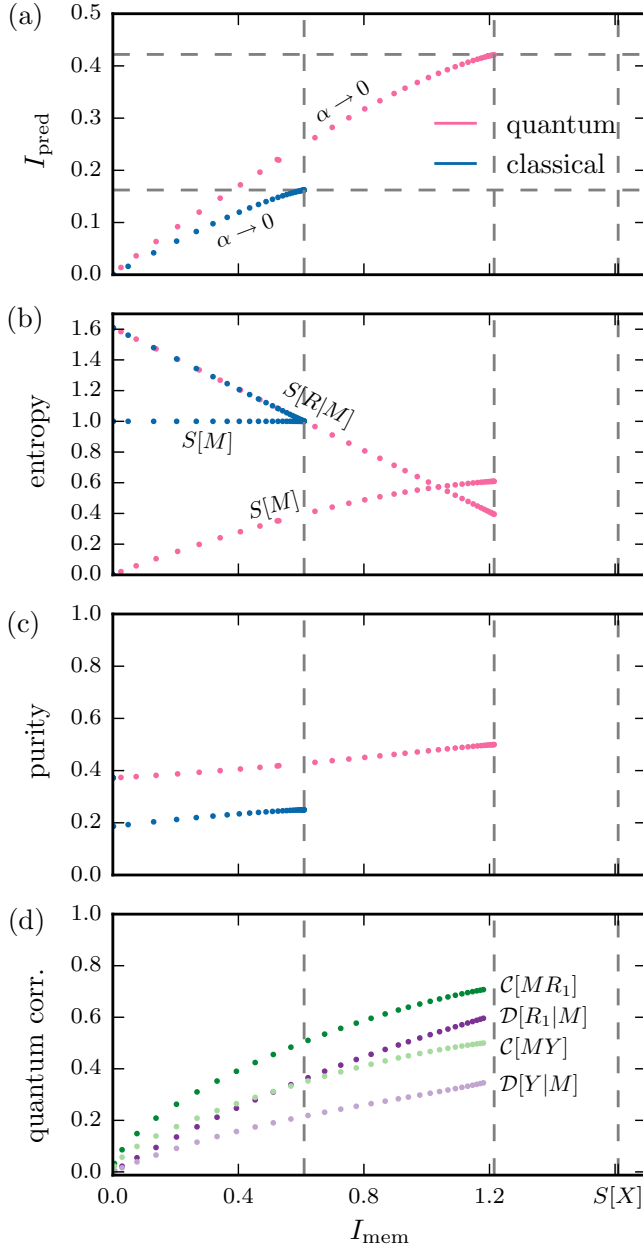
FIG. 6. Optimal encodings for an amplitude damping channel with redundant initial data. (a) The optimal quantum and classical curves in the information plane. (b) The conditional entropy $S[X|M]$ and the memory entropy $S[M]$. (c) The purity of the optimal solution ρ_{MR}^{opt}. (d) Quantum correlations in the optimal solutions, quantified by concurrence and quantum discord, for the input and output systems, MR_1 and MY, respectively, where R_1 refers to the first (relevant) data qubit.

but are displayed more clearly in Fig. 7. The values in this plot are calculated by interpolating the numerical data points displayed in Fig. 6(a).

Quantum predictive advantage increases at the point where the classical limit is reached, Fig. 7 (top). There is up to about 0.1 bits of quantum advantage in the classically feasible regime, and in the classically infeasible regime we find an additional quantum predictive advantage of as much as 0.26 bits. The quantum advantage to compression also becomes
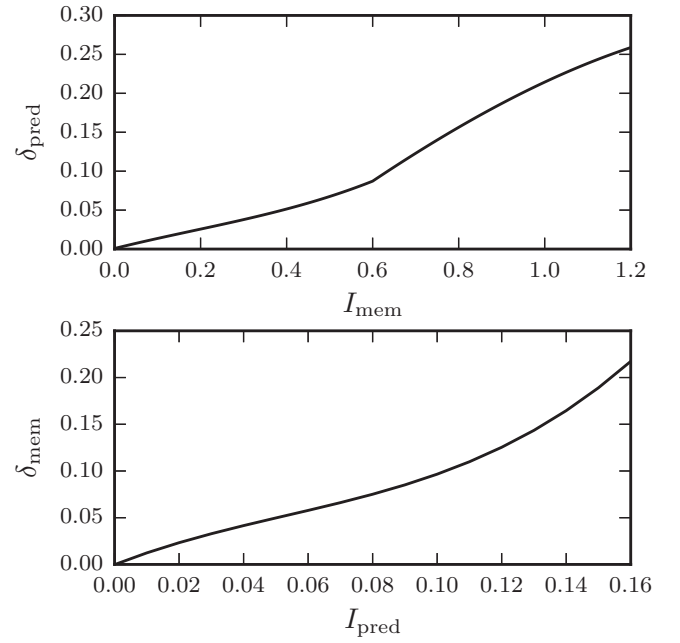

FIG. 7. Predictive quantum advantage for the amplitude damping channel, as quantified by $\delta_{\text{pred}}[I_{\text{mem}}]$ (top) and $\delta_{\text{mem}}[I_{\text{pred}}]$ (bottom).

more pronounced as more predictive information is retained (bottom panel).

VI. CONCLUSIONS

We have introduced and analyzed a quantum generalization of a method introduced by Tishby *et al.* [26] called the information bottleneck method, an approach to lossy compression that extracts relevant information. The coding protocol we introduced can be seen as a generalization of the protocol used in quantum rate-distortion coding.

Both the quality of the encoding and the coding cost are quantified in terms of mutual information. This allowed us to find analytically an implicit relation that any optimal encoding must satisfy.

The approach we have taken allows for choosing how much emphasis should be put on compression versus retention of relevant information. The limit where infinite weight is put on retaining relevant information is particularly interesting. In this limit, an encoding with full predictive information exists, often at significant compression. That is, the compressed data contain as much relevant information as does the initial uncompressed data [50], but it can still be significantly less correlated with the initial data than a purification: The irrelevant information has been filtered out. For a quantum memory, this solution can thus be seen as a generalization of the concept of purification [51]. How much reduction can be achieved depends on the particular data source.

We have introduced a quantum generalization of the information bottleneck algorithm to find optimal encodings numerically. Examples were found exhibiting a substantial quantum predictive advantage. We studied a phase-damping qubit channel and found a classical-to-quantum transition: Both classical and quantum encodings did equally well up to the point where the classical one reached its memory limit.

The quantum encoding was able to capture all of the relevant information in the data (almost a factor two more than the best classical solution). Breaking through the classical memory limit was shown to be associated to a negative quantum conditional entropy. To achieve full predictive power, however, the memory had to be a purification of the initial data.

In contrast, a quantum process with redundant information can be compressed; one need only filter out the redundant bits. Using an example of two qubits, one of which was redundant, we showed that the numerical algorithm successfully picks out and purifies only the relevant qubit. In this example, there was a predictive advantage to having a quantum memory for any value of the memory kept.

We have shown analytically and verified numerically that if the relevant information is purely classical, then there can be no quantum advantage. This result is particularly interesting in the context of biological systems. It might be taken as an indication that an organism living in an environment where it suffices to make predictions about features that are fully approximated by a classical process does not gain from encoding quantum information. This is intuitive, since we expect a possible quantum advantage only when the relevant features of the environment are intrinsically quantum.

ACKNOWLEDGMENTS

APPENDIX A: THE SELF-CONSISTENT EQUATIONS FOR AN OPTIMAL SOLUTION

In this Appendix we show that a solution to the optimization problem, Eq. (15), fulfills Eq. (19) of the main text. We first derive the case for a general memory and then consider the special case where the memory is restricted to be classical.

1. Optimally predictive quantum memories

We wish to solve the optimization problem

$$\max_{\mathcal{E}^{X \to M}} I[M : Y] \quad \text{such that} \quad I[M : R] \leqslant \text{const.,} \quad \text{(A1)}$$

where

$$I[M : R] = S(\sigma_M) + S(\sigma_R) - S(\sigma_{MR}), \quad \text{(A2)}$$

$$I[M : Y] = S(\sigma_M) + S(\sigma_Y) - S(\sigma_{MY}), \quad \text{(A3)}$$

and

$$\sigma_{MR} = \hat{\mathcal{E}}(\psi_{XR}), \quad \text{(A4)}$$

$$\sigma_{MY} = \hat{\mathcal{R}}(\sigma_{MR}). \quad \text{(A5)}$$

Recall that ρ_X, with purification ψ_{XR}, and the relevance map $\mathcal{R}^{R \to Y}$, were assumed to be given. Also note that the reduced state on R, $\sigma_R = \text{tr}_M \sigma_{MR} = \rho_R$, is identical to ρ_X, since $\hat{\mathcal{E}}^{XR \to MR}$ acts trivially on the reference R.

We first remark that optimizing over $\mathcal{E}^{X \to M}$ is equivalent to optimizing over all states, σ_{MR}, on the bipartite system MR,

with the constraint

$$\text{tr}_M \, \sigma_{MR} = \rho_R. \quad \text{(A6)}$$

This is a straightforward consequence of the Choi-Jamiołkowski isomorphism, which gives one-to-one correspondence between channels and positive operators. More specifically, for a *candidate* channel $\mathcal{E}^{X \to M}$, we introduce a representation in terms of a conditional quantum operator [34],

$$\mathcal{E}(\sigma_X) = \text{tr}_X \big[\mathcal{E}_{M|X}^{T_X} \sigma_X \big], \quad \text{(A7)}$$

where $\mathcal{E}_{M|X}$ is a positive operator, satisfying $\text{tr}_M \mathcal{E}_{M|X} = I_X$. T_X denotes the partial transpose on X. We now introduce the following *state* on MR associated with $\mathcal{E}^{X \to M}$:

$$\sigma_{MR} = \rho_R^{1/2} \mathcal{E}_{M|X} \rho_R^{1/2}. \quad \text{(A8)}$$

The slight abuse of notation here is based on the two systems X and R being isomorphic. For ρ_R fixed this uniquely associates a state with the channel. On the other hand, given a state σ_{MR} satisfying $\text{tr}_M \sigma_{MR} = \rho_R$, we introduce the conditional quantum operator

$$\mathcal{E}_{M|X} = \rho_R^{-1/2} \sigma_{MR} \rho_R^{-1/2}, \quad \text{(A9)}$$

which uniquely specifies a quantum channel, $\mathcal{E}^{X \to M}$, through Eq. (A7). For more details on the link between conditional quantum operators, as we define them here, and the Choi-Jamiołkowski isomorphism, we refer the reader to [34].

This means that we can replace the maximization over $\mathcal{E}^{X \to M}$ with one over states, σ_{MX}, in Eq. (A1). We next show explicitly that this is consistent, in the sense that $\hat{\mathcal{E}}(\psi_{XR}) = \rho_R^{1/2} \mathcal{E}_{M|X} \rho_R^{1/2}$, where $\mathcal{E}$ is given by Eq. (A7). First we find for σ_{MR}

$$\sigma_{MR} = \mathcal{E} \otimes \mathcal{I}_R(\psi_{XR}) = \text{tr}_X \big[\mathcal{E}_{M|X}^{T_X} \psi_{XR} \big] \quad \text{(A10)}$$

$$= \sum_{ij} \sqrt{p_i p_j} \text{tr}_X \big[\mathcal{E}_{M|X}^{T_X} |i\rangle\langle j| \big] |i'\rangle\langle j'| \quad \text{(A11)}$$

$$= \sum_{ij} \sqrt{p_i p_j} \langle j | \mathcal{E}_{M|X}^{T_X} | i \rangle |i'\rangle\langle j'| \quad \text{(A12)}$$

$$= \sum_{ij} \sqrt{p_i p_j} \langle i | \mathcal{E}_{M|X} | j \rangle |i'\rangle\langle j'|. \quad \text{(A13)}$$

In the first line we introduced the representation of the map in terms of the conditional state. In the second line, we used Eq. (1) to write ψ_{XR} in terms of basis states $|i\rangle$ and $|i'\rangle$ on the two respective systems, X and R. In the third line we performed the partial trace, and in the last line we performed the partial transpose. On the other hand, we also have that

$$\sigma_{MR} = \rho_R^{1/2} \mathcal{E}_{M|X} \rho_R^{1/2} \quad \text{(A14)}$$

$$= \sum_i \sqrt{p_i} |i\rangle\langle i| \mathcal{E}_{M|X} \sum_j \sqrt{p_j} |j\rangle\langle j| \quad \text{(A15)}$$

$$= \sum_{ij} \sqrt{p_i p_j} \langle i | \mathcal{E}_{M|X} | j \rangle |i\rangle\langle j|, \quad \text{(A16)}$$

where we simply inserted the spectral decomposition $\rho_R = \sum_i p_i |i\rangle\langle i|$. Comparing Eq. (A13) and Eq. (A16), we see that the expressions are identical.

In summary, we now consider the optimization problem

$$\max_{\sigma_{MR}} I[\hat{\mathcal{R}}(\sigma_{MR})] \tag{A17a}$$

$$\text{such that} \quad I(\sigma_{MR}) \leqslant \text{const.}, \tag{A17b}$$

$$\sigma_R = \text{tr}_M[\sigma_{MR}] = \rho_R. \tag{A17c}$$

We solve the constrained optimization problem in Eq. (A17) by introducing Lagrange multipliers. That is, we seek to maximize a Lagrangian

$$\max_{\sigma_{MR}} L(\sigma_{MR}), \tag{A18}$$

defined by

$$L(\sigma_{MR}) = I[\hat{\mathcal{R}}(\sigma_{MR})] - \alpha I(\sigma_{MR}) - \text{tr}_R[\Lambda_R \sigma_R]. \tag{A19}$$

Here $\alpha \geqslant 0$ and Λ_R are the Lagrange multipliers, where Λ_R is a Hermitian operator on R. α is for the constraint on $I(\sigma_{MR})$, and Λ_R for the constraint that $\sigma_R = \text{tr}_M \sigma_{MR} = \rho_R$ is the given initial state. To clarify the latter constraint, it is useful to expand ρ_R in some orthonormal basis of Hermitian operators H_{kl} on $\mathcal{H}_R, \rho_R = \sum_{k,l} x_{kl} H_{kl}, x_{kl} \in \mathbb{R}$. The constraint that $\text{tr}_M \sigma_{MR} = \rho_R$ can then be stated as $\text{tr}_R[H_{kl} \text{tr}_M \sigma_{MR}] = x_{kl}$, for all k,l. Hence, we can write

$$\Lambda_R = \sum_{kl} \lambda_{kl} H_{kl}, \tag{A20}$$

with d_X^2 real Lagrange multipliers λ_{kl}.

We proceed by considering a small variation in σ_{MR}: $\sigma_{MR} \to \sigma_{MR} + \delta_{MR}$ for traceless and Hermitian δ_{MR}. The functional derivative, $\delta L/\delta_{MR}$, is then defined through

$$\delta L = \text{tr}\left[\frac{\delta L}{\delta_{MR}} \delta_{MR}\right], \tag{A21}$$

where δL is the variation in L to first order in δ_{MR}. In particular, we have the following functional derivatives:

$$\frac{\delta S(\sigma_M)}{\delta_{MR}} = -\log_2 \sigma_M \otimes I_R - I_{MR}, \tag{A22}$$

$$\frac{\delta S(\sigma_R)}{\delta_{MR}} = -I_M \otimes \log_2 \sigma_R - I_{MR}, \tag{A23}$$

$$\frac{\delta S(\sigma_{MR})}{\delta_{MR}} = -\log_2 \sigma_{MR} - I_{MR}, \tag{A24}$$

$$\frac{\delta S(\sigma_Y)}{\delta_{MR}} = -\hat{\mathcal{R}}^\dagger(I_M \otimes \log_2 \sigma_Y + I_{MY}), \tag{A25}$$

$$\frac{\delta S(\sigma_{MY})}{\delta_{MR}} = -\hat{\mathcal{R}}^\dagger(\log_2 \sigma_{MY} + I_{MY}). \tag{A26}$$

Note that since $\hat{\mathcal{R}}^{MR \to MY}$ is trace preserving, the dual map, $\hat{\mathcal{R}}^{\dagger MY \to MR}$, is unital, i.e., $\hat{\mathcal{R}}(I_{MY}) = I_{MR}$. We can now write the variational condition for the optimum, $\delta L/\delta_{MR} = 0$,

$$\frac{\delta L(\sigma_{MR})}{\delta_{MR}} = -\log_2 \sigma_M + \hat{\mathcal{R}}^\dagger(\log_2 \sigma_{MY} - \log_2 \sigma_Y)$$

$$+ \alpha(\log_2 \sigma_M + \log_2 \sigma_R - \log_2 \sigma_{MR}) - \tilde{\Lambda}_R = 0, \tag{A27}$$

where we have absorbed an operator proportional to I_{MR} into the Lagrange multiplier $\tilde{\Lambda}_R$ and suppressed the identity operators in tensor products to keep the notation as clean as

possible. Rearranging this, we have

$$\sigma_{MR} = \exp\left(\log_2 \sigma_M + \log_2 \sigma_R - \frac{1}{\alpha} H_{MR} - \tilde{\Lambda}_R\right), \tag{A28}$$

where H_{MR} is defined in Eq. (20).

In Eq. (A28), the Hermitian operator $\tilde{\Lambda}_R$ must be chosen such that $\text{tr}_M \sigma_{MR} = \rho_R$. Since this involves varying d_X^2 real Lagrange multipliers, finding such a $\tilde{\Lambda}_R$ is costly. A better approach is to rewrite the exponential

$$\exp(\log_2 \sigma_M + \log_2 \sigma_R - H_{MR}/\alpha - \tilde{\lambda}_R)$$

$$= C_R \exp(\log_2 \sigma_M - H_{MR}/\alpha)C_R^\dagger,$$

where C_R is an operator acting nontrivially on system R only. To see that this is possible, consider the Zassenhaus formula for the expansion of an exponential [52],

$$e^{X+Y} = e^X e^Y \prod_{n=2}^{\infty} e^{C_n(X,Y)}, \tag{A29}$$

where $C_n(X,Y)$ is a polynomial of degree n in X and Y that can be written in terms of nested commutators of the two operators (with $[X,Y]$ as the innermost commutator). For example, the three first $C_n(X,Y)$ are

$$C_2(X,Y) = -\frac{1}{2}[X,Y],$$

$$C_3(X,Y) = \frac{1}{3}[Y,[X,Y]] + \frac{1}{6}[X,[X,Y]],$$

$$C_4(X,Y) = \frac{1}{24}4[[[X,Y],X],X] + \frac{1}{8}[[[X,Y],X],Y]$$

$$+ \frac{1}{8}[[[X,Y],Y],Y].$$

To apply this to Eq. (A28), we identify $X = \log_2 \sigma_M - H_{MR}/\alpha$ and $Y = \log_2 \sigma_R - \tilde{\Lambda}_R$, so that Y is an operator on system R only. It is then clear that the commutator $[X,Y]$ acts nontrivially on system R only as well. We can therefore write

$$e^{X+Y} = e^{X+Y/2} e^{Y/2} \prod_{n=2}^{\infty} e^{C_n(X+Y/2, Y/2)}$$

$$= \left(\prod_{n=2}^{\infty} e^{C_n(X+Y/2, Y/2)}\right)^\dagger e^{Y/2} e^{X+Y/2}$$

$$= \left(\prod_{n=2}^{\infty} e^{C_n(X, Y/2)}\right)^\dagger e^{Y/2} e^{X+Y/2}$$

$$= \left(\prod_{n=2}^{\infty} e^{C_n(X, Y/2)}\right)^\dagger e^{Y/2} e^X e^{Y/2} \left(\prod_{n=2}^{\infty} e^{C_n(X, Y/2)}\right)$$

$$= C_R e^X C_R^\dagger.$$

In the first line we have applied the Zassenhaus formula. In the second line we have used that X, Y, and $\exp(X + Y)$ are Hermitian. In the third line we have used that $Y/2$ commutes with itself. In the fourth line we applied the Zassenhaus formula again.

The operator C_R must be chosen such that σ_{MR} is positive and the constraint $\text{tr}_M \sigma_{MR} = \rho_R$ holds. This can be guaranteed by choosing

$$C_R = \rho_R^{1/2} Z_R^{-1/2}, \tag{A30}$$

where

$$Z_R^{-1/2} = \left\{ \mathrm{tr}_M \left[\exp \left(\log_2 \rho_M - \frac{1}{\alpha} H_{MR} \right) \right] \right\}^{-1/2}. \quad \text{(A31)}$$

We thus have the following self-consistent equation for an optimal solution σ_{MR}:

$$\sigma_{MR} = \rho_R^{1/2} Z_R^{-1/2} \exp \left(\log_2 \sigma_M - \frac{1}{\alpha} H_{MR} \right) Z_R^{-1/2} \rho_R^{1/2}. \quad \text{(A32)}$$

The associated conditional state, given through Eq. (A9), thus has the form claimed in Eq. (19).

2. Classical memories

We now consider a situation where M is restricted to be classical, while X (R) and Y are not. More precisely, we consider the same optimization problem as before, but now restricted to optimizing over classical-quantum states of the form

$$\sigma_{MR} = \sum_m p(m) |m\rangle\langle m| \otimes \sigma_{R|m}, \quad \text{(A33)}$$

for a fixed basis $\{|m\rangle\}$ on M and arbitrary states $\sigma_{R|m}$ on R. The constraint $\mathrm{tr}_M \sigma_{MR} = \rho_R$ must still be satisfied. Note that the set of all such states is convex when the basis for M is fixed.

The optimization proceeds as before, leading to Eq. (A32), which when using Eq. (A33) gives us

$$p_m \sigma_{R|m} = p_m \rho_R^{1/2} Z_R^{-1/2} \exp \left[\frac{1}{\alpha} \mathcal{R}^\dagger \left(\log_2 \sigma_{Y|m} - \log_2 \sigma_Y \right) \right]$$
$$\times Z_R^{-1/2} \rho_R^{1/2}, \quad \text{(A34)}$$

$$Z_R = \sum_m p_m \exp \left\{ \frac{1}{\alpha} \mathcal{R}^\dagger \left[\log_2 (\sigma_{Y|m}) - \log_2 \sigma_Y \right] \right\}.$$

Together with

$$\begin{aligned} p_m &= \mathrm{tr}[p_m \sigma_{R|m}], \\ \sigma_{Y|m} &= \mathcal{R}(\sigma_{R|m}), \end{aligned} \quad \text{(A35)}$$

this is a self-consistent equation for $p_m \sigma_{R|m}$.

3. Fully classical setup

We now consider a situation where all three systems are entirely classical in the sense that we can write

$$\rho_R = \sum_x p(x) |x\rangle\langle x|, \quad \text{(A36)}$$

$$\rho_{MR} = \sum_{m,x} p(m,x) |m\rangle\langle m| \, |x\rangle\langle x|, \quad \text{(A37)}$$

$$\rho_{MY} = \sum_{m,y} p(m,y) |m\rangle\langle m| \, |y\rangle\langle y|, \quad \text{(A38)}$$

and

$$\mathcal{R}_{Y|R} = \sum_{x,y} p(y|x) |y\rangle\langle y| \langle x|x\rangle\langle x|, \quad \text{(A39)}$$

where $|m\rangle$, $|x\rangle$, and $|y\rangle$ are orthonormal bases for the three respective systems, M, X, and Y, $p(x)$, $p(m,x)$, and $p(m,y)$

are joint probability distributions, and $p(y|x)$ is a conditional probability distribution. By plugging this into Eq. (A32), it is straightforward to show that we can write

$$p(m,x) = \frac{p(m)}{p(x)} \exp \left\{ \frac{1}{\alpha} \sum_y p(y|x) \log_2 \left[\frac{p(m,y)}{p(m)p(y)} \right] \right\} Z_R^{-1}, \quad \text{(A40)}$$

where

$$p(m) = \sum_x p(m,x)$$

$$p(m,y) = \sum_x p(y|x) p(m,x).$$

The normalization factor now reads

$$Z_R = \sum_m p(m) \exp \left\{ \frac{1}{\alpha} \sum_y p(y|x) \log_2 \left[\frac{p(m,y)}{p(m)p(y)} \right] \right\}. \quad \text{(A41)}$$

This is identical to the result for the classical information bottleneck derived in [26].

APPENDIX B: PROPERTIES OF AN OPTIMAL SOLUTION

1. The negative of I_{pred} as "energy" in the low α limit

In Sec. III A we noted that for small α, the conditional quantum operator in Eq. (19) has a form analogous to a thermal state, with α playing the role of temperature and H_{MR} the role of Hamiltonian. We now show that the expectation value of this operator is just the negative of the predictive power:

$$\begin{aligned} \langle H_{MR} \rangle &:= \mathrm{tr}_{MR}[\sigma_{MR} H_{MR}] \\ &= -\mathrm{tr}_{MR}\{\sigma_{MR} \hat{\mathcal{R}}^\dagger [\log_2 \sigma_{MY} - \log_2(\sigma_M \otimes \sigma_Y)]\} \\ &= -\mathrm{tr}_{MY}\{\hat{\mathcal{R}}(\sigma_{MR})[\log_2 \sigma_{MY} - \log_2(\sigma_M \otimes \sigma_Y)]\} \\ &= -\mathrm{tr}_{MY}\{\sigma_e MY[\log_2 \sigma_{MY} - \log_2(\sigma_Y \otimes \sigma_Y)]\} \\ &= -I(\sigma_{MY}) \\ &= -I_{\mathrm{pred}}. \quad \text{(B1)} \end{aligned}$$

In the second line we inserted the definition of H_{MR}, in the third we used the defining property of the dual map, $\hat{\mathcal{R}}^\dagger$, in the fourth that $\sigma_{MY} = \hat{\mathcal{R}}(\sigma_{MR})$, and in the last two lines we simply used the definitions of the mutual information and I_{pred}, respectively.

2. No quantum advantage for classical processes

In this section we prove that there is no quantum advantage for predictive inference on classical processes, as claimed in Sec. III B.

Recall that the relevance channel, $\mathcal{R}^{R\to Y}$, is said to be *classical* if it can be written in terms of a conditional quantum operator of the form

$$\mathcal{R}_{Y|R} = \sum_{x,y} p(y|x) |y\rangle\langle y| \langle x|x\rangle\langle x|, \quad \text{(B2)}$$

where $p(y|x)$ is a conditional probability distribution and $|x\rangle$ and $|y\rangle$ are orthonormal basis sets for the two systems R and Y (recall also that R is isomorphic to X).

We first show that applying a map of this form to an arbitrary state σ_{MR} results in a quantum-classical state on MY, i.e., a state of the form

$$\sigma_{MY} = \hat{\mathcal{R}}(\sigma_{MR}) = \sum_y p(y)\sigma_{M|y}|y\rangle\langle y|, \qquad \text{(B3)}$$

where $p(y)$ is a probability distribution and $\sigma_{M|y}$ are quantum states on M,

$$\hat{\mathcal{R}}(\sigma_{MR}) = \mathrm{tr}_R\left[\sum_{x,y} p(y|x)|y\rangle\langle y|\,|x\rangle\langle x|\sigma_{MR}\right]$$
$$= \sum_{x,y} p(y|x)\langle x|\sigma_{MR}|x\rangle\,|y\rangle\langle y|$$
$$= \sum_y p(y)\sigma_{M|y}|y\rangle\langle y|, \qquad \text{(B4)}$$

where

$$p(y)\sigma_{M|y} = \sum_x p(y|x)\langle x|\sigma_{MR}|x\rangle. \qquad \text{(B5)}$$

We next show that if the initial state, ρ_R, is diagonal in the chosen basis $|x\rangle$,

$$\rho_R = \sum_x p(x)|x\rangle\langle x|, \qquad \text{(B6)}$$

then Eq. (19) implies that the optimal state ρ_{MR}^{opt} corresponding to an optimal encoding is also quantum classical. For ease of notation we drop the superscript "opt" for the rest of this section.

Recall that we showed in Appendix A that the state ρ_{MR} corresponding to an optimal encoding must satisfy

$$\rho_{MR} = \rho_R^{1/2} Z_R^{-1/2} \exp\left(\log_2 \rho_M - \frac{1}{\alpha} H_{MR}\right) Z_R^{-1/2}\rho_R^{1/2}, \qquad \text{(B7)}$$

where

$$H_{MR} = \hat{\mathcal{R}}^\dagger\left(\log_2 \rho_M + \log_2 \rho_Y - \log_2 \rho_{MY}\right), \qquad \text{(B8)}$$

$$Z_R^{-1/2} = \left\{\mathrm{tr}_M\left[\exp\left(\log_2 \rho_M - \frac{1}{\alpha} H_{MR}\right)\right]\right\}^{-1/2}. \qquad \text{(B9)}$$

Using that ρ_{MY} is quantum-classical, we can write

$$\log_2 \rho_{MY} = \log_2\left[\sum_y p(y)\rho_{M|y}|y\rangle\langle y|\right]$$
$$= \sum_y \log_2\left[p(y)\rho_{M|y}\right]|y\rangle\langle y|$$
$$= \sum_y \left[\log_2 p(y) + \log_2 \rho_{M|y}\right]|y\rangle\langle y|$$
$$= \log_2 \rho_Y + \sum_y \log_2\left(\rho_{M|y}\right)|y\rangle\langle y|. \qquad \text{(B10)}$$

Thus,

$$H_{MR} = \hat{\mathcal{R}}^\dagger\left[\log_2 \rho_M - \sum_y \log_2(\rho_{M|y})|y\rangle\langle y|\right]$$
$$= \log_2 \rho_M - \sum_{x,y} p(y|x)\log_2(\rho_{M|y})|x\rangle\langle x|$$
$$= \sum_x\left[\log_2 \rho_M - \sum_y p(y|x)\log_2(\rho_{M|y})\right]|x\rangle\langle x|, \qquad \text{(B11)}$$

where we have used that $\mathcal{R}^\dagger(I_Y) = I_R = \sum_x |x\rangle\langle x|$ and

$$\mathcal{R}^\dagger(|y\rangle\langle y|) = \sum_x p(y|x)|x\rangle\langle x|. \qquad \text{(B12)}$$

Using this, we have that

$$\exp\left(\log_2 \rho_M - \frac{1}{\alpha} H_{MR}\right)$$
$$= \sum_x \exp\left[\left(1 - \frac{1}{\alpha}\right)\log_2 \rho_M \right.$$
$$\left. + \frac{1}{\alpha}\sum_y p(y|x)\log_2(\rho_{M|y})\right]|x\rangle\langle x|$$
$$= \sum_x \tau_{M|x}|x\rangle\langle x|, \qquad \text{(B13)}$$

where we have defined

$$\tau_{M|x} = \exp\left[\left(1 - \frac{1}{\alpha}\right)\log_2 \rho_M + \frac{1}{\alpha}\sum_y p(y|x)\log_2(\rho_{M|y})\right]. \qquad \text{(B14)}$$

This means that we can write

$$\rho_{MR} = \sum_x \tau_{M|x}\rho_R^{1/2} Z_R^{-1/2}|x\rangle\langle x|Z_R^{-1/2}\rho_R^{1/2} \qquad \text{(B15)}$$

and

$$Z_R = \sum_x \mathrm{tr}_M[\tau_{M|x}]|x\rangle\langle x|. \qquad \text{(B16)}$$

Now it just remains to use that $\rho_R = \sum_x p(x)|x\rangle\langle x|$ to arrive at

$$\rho_{MR} = \sum_x p(x)\rho_{M|x}|x\rangle\langle x|, \qquad \text{(B17)}$$

where

$$\rho_{M|x} = \frac{\tau_{M|x}}{\mathrm{tr}[\tau_{M|x}]}. \qquad \text{(B18)}$$

This proves the claim that ρ_{MR} is quantum classical.

The form of Eq. (B17) does not rule out some degree of quantumness of the memory, since the states $\rho_{M|x}$ need not be orthogonal. However, we now show that this does not allow for higher predictive power at fixed memory, thus ruling out any quantum advantage.

First of all, note that the maximum possible value of I_{pred} for a classical process can be reached by a state of the form

$$\rho_{MR} = \sum_{m,x} p(x) p(m|x) |m\rangle\langle m| \, |x\rangle\langle x|, \tag{B19}$$

which we refer to as a *classical state*, since both M and R are classical. This can be achieved by choosing $M \cong R$ and $p(m|x) = \delta_{m,x}$, where $\delta_{m,x}$ is the Kronecker δ. It is easily seen that this gives the same output state ρ_{MY} as would a purification ψ_{XR} sent through the relevance channel. We therefore have that for any achievable predictive power there exists a classical state of the form of Eq. (B19) that reaches this value.

Now consider the memory and predictive power of the state Eq. (B17),

$$I_{\text{mem}}[\rho_{MR}] = S(\rho_M) - \sum_x p(x) S(\rho_{M|x}), \tag{B20}$$

$$I_{\text{pred}}[\rho_{MR}] = S(\rho_M) - \sum_y p(y) S(\rho_{M|y}), \tag{B21}$$

where the output state, after sending through the relevance channel, is

$$\rho_{MY} = \hat{\mathcal{R}}(\rho_{MR}) = \sum_y p(y) \rho_{M|y} |y\rangle\langle y|, \tag{B22}$$

and $p(y) \rho_{M|y} = \sum_x p(y|x) p(x) \rho_{M|x}$. Using Eqs. (B18) and (B14) we find that the Lagrangian, $L(\rho_{MR})$, that we wish to maximize is

$$L(\rho_{MR}) = I_{\text{pred}}[\rho_{MR}] - \alpha I_{\text{mem}}[\rho_{MR}]$$
$$= \sum_x p(x) \log_2(\text{tr} \, \tau_{M|x}). \tag{B23}$$

Since the right-hand side only depends on the trace of $\tau_{M|x}$, the Lagrangian is invariant under a measurement of the memory M in a chosen basis $|m\rangle$. This shows that quantum correlations of the type where $\rho_{M|x}$ are nonorthogonal are superfluous. We conclude that optimal values of the Lagrangian can be reached for classical states in the case of classical processes.

APPENDIX C: AN ITERATIVE ALGORITHM AND DETERMINISTIC ANNEALING

In this Appendix we introduce an algorithm to find optimal encodings for given initial data and relevance channel. This is a direct quantum generalization of the information bottleneck method [26] used to find such optimal encodings in the classical case. The information bottleneck method employs an iterative algorithm where a solution at one iteration is used on the right-hand side of Eq. (19) to compute a solution for the next iteration. This algorithm is similar to the Blahut-Arimoto algorithm, for which each iteration is proven to converge [53,54], but since the functional we are optimizing is not convex over the product space of states on MR and MY uniqueness is not guaranteed [26]. The information bottleneck method has been used in several domains, and it has inspired new algorithms as well as new insights into existing data-analysis methods which can be derived from the elegant conceptual framework (see, e.g., [10,55–65] and references therein).

We introduce the following quantum generalization to find an optimal encoding at a given value of α. Denoting the state at the kth step by $\sigma_{MR}^{(k)}$, the state at the next step is found through the following.

(1) Compute $\sigma_M^{(k)}$, $\sigma_{MY}^{(k)}$, and $\sigma_Y^{(k)}$ according to

$$\sigma_M^{(k)} = \text{tr}_R \sigma_{MR}^{(k)}, \quad \sigma_{MY}^{(k)} = \hat{\mathcal{R}}(\sigma_{MR}^{(k)}), \quad \sigma_Y^{(k)} = \text{tr}_M \sigma_{MY}^{(k)}.$$

(2) Compute the state for the next step according to

$$\sigma_{MR}^{(k+1)} = \rho_R^{1/2} \left(Z_R^{(k)} \right)^{-1/2} \exp\left(\log_2 \sigma_M^{(k)} - \frac{1}{\alpha} H_{MR}^{(k)} \right)$$
$$\times \left(Z_R^{(k)} \right)^{-1/2} \rho_R^{1/2},$$

where

$$H_{MR}^{(k)} = \hat{\mathcal{R}}^\dagger \left(\log_2 \sigma_M^{(k)} + \log_2 \sigma_Y^{(k)} - \log_2 \sigma_{MY}^{(k)} \right)$$

and

$$Z_R^{(k)} = \text{tr}_M \left[\exp\left(\log_2 \sigma_M^{(k)} - \frac{1}{\alpha} H_{MR}^{(k)} \right) \right].$$

(3) Halt if $L^{(k+1)} - L^{(k)} \leqslant \varepsilon$, for some tolerance ε, where $L^{(k)} = I_{\text{pred}}(\sigma_{MR}^{(k)}) - \alpha I_{\text{mem}}(\sigma_{MR}^{(k)})$.

Whenever we restrict to a situation where the memory is classical, as discussed in Appendix A 2, we replace step 3 with the following.

(4) Compute the state for a classical memory for the next step according to

$$p_m^{(k+1)} \sigma_{R|m}^{(k+1)} = p_m^{(k)} \rho_R^{1/2} \left(Z_R^{(k)} \right)^{-1/2}$$
$$\times \exp\left[-\frac{1}{\alpha} \hat{\mathcal{R}}^\dagger \left(\log_2 \sigma_Y^{(k)} - \log_2 \sigma_{Y|m}^{(k)} \right) \right]$$
$$\times \left(Z_R^{(k)} \right)^{-1/2} \rho_R^{1/2},$$

where

$$Z_R^{(k)} = \sum_m p_m^{(k)} \exp\left\{ -\frac{1}{\alpha} \hat{\mathcal{R}}^\dagger \left[\log_2 \sigma_Y^{(k)} - \log_2 \left(\sigma_{Y|m}^{(k)} \right) \right] \right\}.$$

Step 2 is replaced with $p_m^{(k)} = \text{tr}[p_m^{(k)} \sigma_{A|m}^{(k)}]$, and $\sigma_{B|m}^{(k)} = \mathcal{E}(\sigma_{A|m}^{(k)})$. This allows us to compare the optimal solution found for a classical memory to that of a quantum memory.

The iterative algorithm allows us to find a solution at given α, where α parametrizes the tradeoff between I_{pred} and I_{mem}. By varying this parameter, we can trace out a curve in the "information plane" [26,46] with I_{mem} on the x axis and I_{pred} on the y axis. This is analogous to a rate-distortion curve in conventional rate-distortion theory. To trace out this curve, we use the following scheme: We start at a large value of α with the trivial solution $\rho_{MR}^{\text{opt}} = \rho_M \otimes \rho_R$ (ρ_M arbitrary) and gradually "cool" the system by lowering α towards zero in small steps. At each step we use a small perturbation of the solution at the previous step as initial guess. This approach to clustering problems has been dubbed "deterministic annealing," due to the analogy with statistical mechanics [66].

The annealing scheme allows us not only to find the optimal solution at a fixed memory dimension, d_M, but also to find the smallest d_M possible at each value of α. At very large α we have the trivial solution $\rho_M \otimes \rho_R$ for arbitrary ρ_M, and we can

therefore choose M to be a trivial system with Hilbert space dimension $d_M = 1$. As we lower α we compare the solution for a memory of dimension d_M to one with dimension $d_M + 1$, and increase the dimension for the next step, $d_M \leftarrow d_M + 1$, only if the higher-dimensional system outperforms the lower-dimensional one. In this way, the optimal solution will go

through a series of "phase transitions" [66] at critical values of α, where the dimension is increased.

It is particularly interesting that in the $\alpha \to 0$ limit of this annealing scheme one obtains a solution with maximal predictive power equal to that of a purification, but potentially at a lower I_{mem} and memory dimension d_M.

[1] R. Shaw, *The Dripping Faucet as a Model Chaotic System* (Aerial Press, Santa Cruz, CA, 1984).

[2] Peter Grassberger, Toward a quantitative theory of self-generated complexity, Int. J. Theor. Phys. **25**, 907 (1986).

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

[4] I. Nemenman, Ph.D. thesis, Princeton University, 2000.

[5] W. Bialek, I. Nemenman, and N. Tishby, Predictability, complexity and learning, Neural Comput. **13**, 2409 (2001).

[6] William of Ockham, *Philosophical Writings: A Selection, Translated, with an Introduction, by Philotheus Boehner, O.F.M., Late Professor of Philosophy, The Franciscan Institute* (Bobbs-Merrill, Indianapolis, 1964) [first published in various European cities, early 1300s].

[7] H. Jeffreys, *Theory of Probability*, 3rd ed. (Oxford University Press, Oxford, U.K., 1998).

[8] G. J. Chaitin, *Algorithmic Information Theory* (Cambridge University Press, Cambridge, U.K., 2004), Vol. 1.

[9] V. Vapnik, *Statistical Learning Theory* (Wiley & Sons, New York, 1998).

[10] S. Still, Information bottleneck approach to predictive inference, Entropy **16**, 968 (2014).

[11] S. Still, D. A. Sivak, A. J. Bell, and G. E. Crooks, Thermodynamics of Prediction, Phys. Rev. Lett. **109**, 120604 (2012).

[12] A. L. Grimsmo, Quantum correlations in predictive processes, Phys. Rev. A **87**, 060302 (2013).

[13] W. Bialek, *Biophysics: Searching for Principles* (Princeton University Press, Princeton, NJ, 2012).

[14] K. Kinosita, R. Yasuda, H. Noji, and K. Adachi, A rotary molecular motor that can work at near 100% efficiency, Philos. Trans. R. Soc., B **355**, 473 (2000).

[15] Giovanni Cappello, Paolo Pierobon, Clémentine Symonds, Lorenzo Busoni, J. Christof, M. Gebhardt, Matthias Rief, and Jacques Prost, Myosin V stepping mechanism, Proc. Natl. Acad. Sci. USA **104**, 15328 (2007).

[16] A. Hasenstaub, S. Otte, E. Callaway, and T. J. Sejnowski, Metabolic cost as a unifying principle governing neuronal biophysics, Proc. Natl. Acad. Sci. USA **107**, 12329 (2010).

[17] Gregory S. Engel, Tessa R. Calhoun, Elizabeth L. Read, Tae-Kyu Ahn, Tomáš Mančal, Yuan-Chung Cheng, Robert E. Blankenship, and Graham R. Fleming, Evidence for wavelike energy transfer through quantum coherence in photosynthetic systems, Nature (London) **446**, 782 (2007).

[18] Hohjai Lee, Yuan-Chung Cheng, and Graham R. Fleming, Coherence dynamics in photosynthesis: Protein protection of excitonic coherence, Science **316**, 1462 (2007).

[19] Gitt Panitchayangkoon, Dugan Hayes, Kelly A. Fransted, Justin R. Caram, Elad Harel, Jianzhong Wen, Robert E. Blankenship, and Gregory S. Engel, Long-lived quantum coherence in

photosynthetic complexes at physiological temperature, Proc. Natl. Acad. Sci. USA **107**, 12766 (2010).

[20] Masoud Mohseni, Patrick Rebentrost, Seth Lloyd, and Alan Aspuru-Guzik, Environment-assisted quantum walks in photosynthetic energy transfer, J. Chem. Phys. **129**, 174106 (2008).

[21] Susanna F. Huelga and Martin B. Plenio, Vibrations, quanta and biology, Contemp. Phys. **54**, 181 (2013).

[22] A. W. Chin, J. Prior, R. Rosenbach, F. Caycedo-Soler, S. F. Huelga, and M. B. Plenio, The role of non-equilibrium vibrational structures in electronic coherence and recoherence in pigment-protein complexes, Nat. Phys. **9**, 113 (2013).

[23] C. E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. **27**, 379 (1948); **27**, 623 (1948).

[24] A classical text on rate-distortion theory is [25].

[25] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression* (Prentice-Hall, Upper Saddle River, NJ, 1971).

[26] N. Tishby, F. Pereira, and W. Bialek, The information bottleneck method, in *Proceedings of the 37th Annual Allerton Conference*, edited by B. Hajek and R. S. Sreenivas (University of Illinois, Urbana, 1999).

[27] For more details, see [10] and references therein.

[28] Howard Barnum, Quantum rate-distortion coding, Phys. Rev. A **62**, 042309 (2000).

[29] Nilanjana Datta, Min-Hsiu Hsieh, Mark M. Wilde, and Andreas Winter, Quantum-to-classical rate distortion coding, J. Math. Phys. **54**, 042201 (2013).

[30] Mark M. Wilde, Nilanjana Datta, M.-H. Hsieh, and Andreas Winter, Quantum rate-distortion coding with auxiliary resources, IEEE T. Inf. Theory **59**, 6755 (2013).

[31] Nilanjana Datta, Min-Hsiu Hsieh, and Mark M. Wilde, Quantum rate distortion, reverse shannon theorems, and source-channel separation, IEEE T. Inf. Theory **59**, 615 (2013).

[32] William K. Wootters, Entanglement of Formation of an Arbitrary State of Two Qubits, Phys. Rev. Lett. **80**, 2245 (1998).

[33] B. Groisman, S. Popescu, and A. Winter, Quantum, classical, and total amount of correlations in a quantum state, Phys. Rev. A **72**, 032317 (2005).

[34] M. S. Leifer and Robert W. Spekkens, Towards a formulation of quantum theory as a causally neutral theory of Bayesian inference, Phys. Rev. A **88**, 052130 (2013).

[35] The encoding rate and the distortion measure is conventionally defined in a many-copy scenario: One considers n copies of the information source, $X \to X^n$, $\rho_X \to \rho_X^{\otimes n}$, etc. The encoding and decoding channels are, respectively, maps from and back to the n-fold system space. The memory consists of m qubits, and the coding rate is the number of memory qubits per copy of the source, $R = m/n$, as n goes to infinity. The distortion measure introduced in [28] is based on average entanglement fidelity

in the large n limit, with high fidelity meaning low distortion. The optimization problem is to minimize the rate $R(D)$, over encoding and decoding maps, at a constraint on the allowed distortion, D.

[36] The choice of I_{pred} and I_{mem} as, respectively, quality and cost measures represents a direct quantum generalization of the classical information bottleneck method [26]. We note that I_{mem} has an operational meaning as (two times) the qubit rate cost of simulating the channel, $\mathcal{E}^{X \to M}$, in the presence of arbitrary shared entanglement between the two parties X and M [37,38].

[37] I. Devetak, Triangle of Dualities between Quantum Communication Protocols, Phys. Rev. Lett. **97**, 140503 (2006).

[38] Anura Abeyesinghe, Igor Devetak, Patrick Hayden, and Andreas Winter, The mother of all protocols: Restructuring quantum information's family tree, in *Proceedings of the Royal Society of London, Series A, Mathematical and Physical Sciences* (The Royal Society, London, 2009).

[39] The dual map is defined through $\mathrm{tr}_{MR}[\hat{\mathcal{R}}^{\dagger}(A_{MY})B_{MR}] = \mathrm{tr}_{MY}[A_{MY}\hat{\mathcal{R}}(B_{MR})]$, for all operators B_{MR} on MR and A_{MY} on MY.

[40] J. P. Crutchfield and C. R. Shalizi, Thermodynamic depth of causal states: Objective complexity via minimal representations, Phys. Rev. E **59**, 275 (1999).

[41] Mile Gu, Karoline Wiesner, Elisabeth Rieper, and Vlatko Vedral, Quantum mechanics can reduce the complexity of classical models, Nat. Commun. **3**, 762 (2012).

[42] Michał Horodecki, Jonathan Oppenheim, and Andreas Winter, Partial quantum information, Nature (London) **436**, 673 (2005).

[43] Lidia Del Rio, Johan Åberg, Renato Renner, Oscar Dahlsten, and Vlatko Vedral, The thermodynamic meaning of negative entropy, Nature (London) **474**, 61 (2011).

[44] H. Ollivier and W. H. Zurek, Quantum Discord: A Measure of the Quantumness of Correlations, Phys. Rev. Lett. **88**, 017901 (2001).

[45] L. Henderson and V. Vedral, Classical, quantum and total correlations, J. Phys. A **34**, 6899 (2001).

[46] S. Still, J. P. Crutchfield, and C. J. Ellison, Optimal causal inference: Estimating stored information and approximating causal architecture, Chaos **20**, 037111 (2010).

[47] J. P. Crutchfield and D. P. Feldman, Regularities unseen, randomness observed: Levels of entropy convergence, Chaos **13**, 25 (2003).

[48] S. Still and J. P. Crutchfield, Structure or noise?, arXiv:0708.0654.

[49] Francesco Ciccarello and Vittorio Giovannetti, Creating quantum correlations through local nonunitary memoryless channels, Phys. Rev. A **85**, 010102 (2012).

[50] The trivial uncompressed encoding follows from $\mathcal{E}^{X \to M} = \mathcal{I}_X$, leaving the MR in the purification state.

[51] A purification is achieved in this limit with the choice $\mathcal{R} = \mathcal{I}_R$ for the relevance channel.

[52] F. Casas, A. Murua, and M. Nadinic, Efficient computation of the Zassenhaus formula, Comp. Phys. Commun. **183**, 2386 (2012).

[53] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory* (Wiley & Sons, New York, 2012).

[54] Richard E. Blahut, Computation of channel capacity and rate-distortion functions, IEEE T. Inf. Theory **18**, 460 (1972).

[55] N. Friedman, O. Mosenzon, N. Slonim, and N. Tishby, Multivariate information bottleneck, in *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence* (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001), pp. 152–161.

[56] I. S Dhillon, S. Mallela, and D. S. Modha, Information-theoretic co-clustering, in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, New York, 2003), pp. 89–98.

[57] S. Still and W. Bialek, How many clusters? An information theoretic perspective, Neural Comput. **16**, 2483 (2004).

[58] S. Still, W. Bialek, and L. Bottou, Geometric clustering using the information bottleneck method, in *Advances in Neural Information Processing Systems 16*, edited by S. Thrun, L. K. Saul, and B. Schölkopf (MIT Press, Cambridge, MA, 2004).

[59] G. Chechnik, A. Globerson, N. Tishby, and Y. Weiss, Information bottleneck for Gaussian variables, J. Mach. Learn. Res. **6**, 165 (2005).

[60] R. Bekkerman, R. El-Yaniv, and A. McCallum, Multi-way distributional clustering via pairwise interactions, in *Proceedings of the 22nd International Conference on Machine Learning* (Association for Computing Machinery, New York, 2005), pp. 41–48.

[61] S. Still, Information-theoretic approach to interactive learning, Europhys. Lett. **85**, 28005 (2009).

[62] F. Creutzig, A. Globerson, and N. Tishby, The past-future information bottleneck of dynamical systems, Phys. Rev. E **79**, 041925 (2009).

[63] S. Still and D. Precup, An information-theoretic approach to curiosity-driven reinforcement learning, Theory Biosci. **131**, 139 (2012).

[64] N. Jacoby, N. Tishby, and D. Tymoczko, An information theoretic approach to chord categorization and functional harmony, J. New Music Res. **44**, 219 (2015).

[65] N. Tishby and N. Zaslavsky, Deep learning and the information bottleneck principle, IEEE Information Theory Workshop (ITW) 2015, arXiv:1503.02406.

[66] K. Rose, Deterministic annealing for clustering, compression, classification, regression, and related optimization problems, Proc. IEEE **86**, 2210 (1998).