Optimizing Quantum Models of Classical Channels:
The reverse Holevo problem

Samuel P. Loomis,[*] John R. Mahoney,[†] Cina Aghamohammadi,[‡] and James P. Crutchfield[§]
Complexity Sciences Center and Physics Department,
University of California at Davis, One Shields Avenue, Davis, CA 95616

(Dated: October 30, 2019)

Given a classical channel—a stochastic map from inputs to outputs—the input can often be transformed to an intermediate variable that is informationally smaller than the input. The new channel accurately simulates the original but at a smaller transmission rate. Here, we examine this procedure when the intermediate variable is a quantum state. We determine when and how well quantum simulations of classical channels may improve upon the minimal rates of classical simulation. This inverts Holevo's original question of quantifying the capacity of quantum channels with classical resources. We also show that this problem is equivalent to another, involving the local generation of a distribution from common entanglement.

I. INTRODUCTION

One speaks of a quantum advantage when a computational task is performed more efficiently (in memory, time, or both) using quantum mechanical hardware than classical hardware. Quantum advantages appear in the simulation of a variety of classical systems [1]: thermal states [2], fluid flows [3, 4], electromagnetic fields [5], diffusion processes [6, 7], Burger's equation [8], and molecular dynamics [9]. Quantum advantage also has been found in more mathematical contexts. The most well-known problems include the factorization of prime numbers (Shor's integer factoring algorithm [10]), database search (Grover's algorithm [11]), and the efficient solution of linear systems [12].

A recent but rich area of study is the quantum advantage for simulating classical stochastic processes. By stochastic process, we mean a source that probabilistically generates a sequence of symbols $x_0 x_1 \ldots x_t$. When the probability of each new symbol x_t depends on the previous symbols, the process is said to have memory. Computational mechanics has previously studied the memory resources required for simulating and predicting stochastic processes using classical hardware [13–15]. Quantum computational mechanics now proposes to use sequential measurements of a quantum system as a more resource-efficient means of simulating stochastic processes [16–18, 20?].

Recent efforts showed that many important results on optimal simulation and prediction do not carry over from the classical to the quantum domain. For instance, it is a common fact of resource theories that a given resource (say, memory) may have various inequivalent quantifications depending on the specific task at hand (say, asymptotic rates versus single-shot requirements). In classical computational mechanics, in contrast, a stochastic process has a uniquely optimal predictive model—the ϵ-machine. It is minimal according to all quantifications of memory. In this sense, the ϵ-machine *strongly minimizes* memory [21]. However, this no longer holds for quantum models, where inequivalent quantifications have inequivalent minima. For example, a model that is optimal for single-shot implementations is not the optimal model for the asymptotic rate. We call this *weak optimization* [21, 22].

In a memoryful stochastic process, the future is dependent on the past through what is mathematically considered a probabilistic channel. This motivates the exploration of simulating classical channels with quantum resources, which is a fundamental question that sheds light on the more complicated problem of simulating stochastic processes. This question is the main motivator of the following development, although we call-out ancillary results along the way.

Simulating a classical channel with quantum resources is in many ways the inverse of much previous work on channels in quantum information. There, the focus is often on the resources (such as entanglement) required to simulate a fully quantum channel [23–25]. It also often addresses the capacity of a quantum channel to transmit classical information, first considered by Holevo [26–29]. Rather than using classical capacities as a means to study the properties of quantum channels, the following uses quantum resources as a means to study the properties of purely classical channels. Thus, in the spirit of the "reverse Shannon theorem" [24], we consider this a *reverse*

---

[*] sloomis@ucdavis.edu
[†] jrmahoney@ucdavis.edu
[‡] caghamohammadi@ucdavis.edu
[§] chaos@ucdavis.edu

Holevo problem.

Our results also draw from the literature of common information [30, 31], which concerns simulating channels with intermediate classical variables. We adapt those results to the quantum domain by making the intermediate variable quantal.

The following (second) section establishes notation and definitions for our development. The third presents our three main results for the reverse Holevo problem:

1. For every channel, there are quantum models that reduce memory costs across all quantifications of memory (or, in rare cases, do at least as well as the original channel).
2. However, there is generally not a single quantum model that is minimal with respect to all quantifications of memory. This has important implications for single-shot channel simulation.
3. We demonstrate a lower bound for the asymptotic cost of modeling a channel with quantum resources.

We also present a mathematical correspondence between the reverse Holevo problem and another that we call *common entanglement*. The latter asks for the entanglement cost of generating a bivariate classical distribution using only entanglement and local operations (without communication). This correspondence sets the stage for further developments.

II.   NOTATION AND DEFINITIONS

This section adapts results from computational mechanics, as well as from the theory of common information, to describe Shannon communication channels.

Computational mechanics is a subfield of statistical mechanics that addresses the information-theoretic and energetic costs of simulating and predicting stochastic processes [15, 32]. In that setting, a *stochastic process* is defined by an alphabet $\mathcal{A}$ and a probability measure $\mathbb{P}(x_{-\infty:\infty})$ over all bi-infinite words $x_{-\infty:\infty} = \ldots x_{-1}x_0x_1 \ldots$ such that $x_t \in \mathcal{A}$ for all $t \in \mathbb{Z}$.

Random variables are denoted by capital Latin letters X, Y, and so on, and their outcomes by $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and so on. The calligraphic letters $\mathcal{X}, \mathcal{Y}$, and so on represent finite sets.

Classical distributions are denoted by $\mathbb{P}$. For example, the probability of a random variable X taking value $x \in \mathcal{X}$ is given by $\mathbb{P}(X = x)$ if the random variable needs to be specified and is shortened to $\mathbb{P}(x)$ if it can be inferred from context. We say $Y = f(X)$ for a function $f : \mathcal{X} \to \mathcal{Y}$ if $\mathbb{P}(Y = y) = \sum_{x:f(x)=y} \mathbb{P}(X = x)$.

As a measure of its information content, the uncertainty of a random variable is measured by its *Rényi entropies*:

$$H_\alpha(X) := \frac{1}{1-\alpha} \log \left(\sum_{x \in \mathcal{X}} \mathbb{P}(x)^\alpha \right) \;,$$

for $\alpha \in (0,1)$ and $\alpha \geq 1$. In the limits $\alpha \to 0$ and $\alpha \to 1$, we recover the *max-entropy* $H_0(X) := \log|\mathcal{X}|$ (useful for costs in single-shot or zero-error situations) and the *Shannon entropy* $H(X) := -\sum_x \mathbb{P}(x) \log \mathbb{P}(x)$, respectively. An important property of the Rényi entropies is that they are monotonically decreasing under the application of a deterministic function. That is, if $Y = f(X)$, then $H(Y) \leq H(X)$, with equality only when f is bijective.

A.   Classical channels

When discussing channels, we consider an input space $\mathcal{X}$ and output space $\mathcal{Y}$, both (for convenience) assumed to be discrete. On these one may define random variables X and Y, respectively. A *channel* $\mathcal{C}$ is then a conditional probability function $\mathbb{P}_\mathcal{C}(Y = y|X = x)$, which for each outcome $x \in \mathcal{X}$ in the input space defines a probability distribution over the output space $\mathcal{Y}$. We write such a channel as $\mathcal{C} : \mathcal{X} \rightsquigarrow \mathcal{Y}$, distinguishing it from a function via the squiggly arrow.

When it comes to information theory, a channel's associated costs and resources depend on the input distribution $\mathbb{P}(X = x)$. One may vary these inputs to determine maximum capacities or minimum costs, for example. When speaking of a channel $\mathcal{C}$, though, for completeness we assume that an input distribution has been defined and so XY forms a bivariate random variable. XY is defined on the space $\mathcal{X} \times \mathcal{Y}$ with probability distribution $\mathbb{P}(X = x, Y = y) = \mathbb{P}_\mathcal{C}(Y = y|X = x)\mathbb{P}(X = x)$. This assumption does not affect the generality of our results.

For bivariate random variables, the Shannon entropy can be used to define the *conditional entropies* and *mutual information*:

$$H(X|Y) := H(XY) - H(Y) \;,$$
$$H(X|Y) := H(XY) - H(Y) \;, \text{ and}$$
$$I(X : Y) := H(X) + H(Y) - H(XY) \;.$$

These represent the uncertainty remaining in X when Y is known, the uncertainty in Y when X is known, and the information shared between X and Y, respectively.

We wish to *simulate* a particular channel $\mathcal{C}$. That is, we want (i) a random variable Z defined on an intermediate space $\mathcal{Z}$, (ii) an encoding channel $\mathcal{E} : \mathcal{X} \to \mathcal{Z}$ with
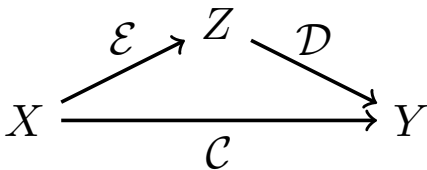
FIG. 1. Information flow when simulating a channel: When $(\mathcal{Z}, \mathcal{E}, \mathcal{D})$ is a model, all paths in the diagram from X to Y represent the same channel—the diagram commutes. When $\mathcal{E} = f$ is a deterministic function, the model is a factor.
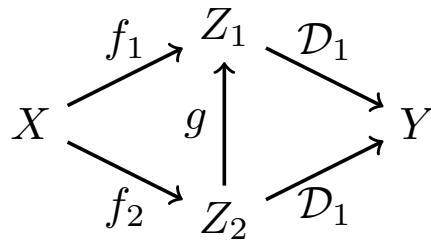

FIG. 2. Channel sufficiency: $F_2 = (\mathcal{Z}_2, f_2, \mathcal{D}_2)$ is sufficient for $F_1 = (\mathcal{Z}_1, f_1, \mathcal{D}_1)$ if, for some deterministic g, this diagram commutes.

conditional distribution $\mathbb{P}_{\mathcal{E}}(Z = z | X = x)$, and (iii) a decoding channel $\mathcal{D} : \mathcal{Z} \rightsquigarrow \mathcal{Y}$ with conditional distribution $\mathbb{P}_{\mathcal{D}}(Y = y | Z = z)$ such that:

$$\mathbb{P}_{\mathcal{C}}(y|x) = \sum_{z \in \mathcal{Z}} \mathbb{P}_{\mathcal{D}}(y|z)\,\mathbb{P}_{\mathcal{E}}(z|x) \ .$$

We write this composition rule as $\mathcal{C} = \mathcal{D} \circ \mathcal{E}$. We call the triplet $M = (\mathcal{Z}, \mathcal{E}, \mathcal{D})$ a *model* of channel $\mathcal{C}$.

Additionally, $X - Z - Y$ form a Markov chain. The *data processing inequality* tells us that $H(Z) \geq I(X : Z) \geq I(X : Y)$. This puts a constraint on models that simulate a given channel $\mathcal{C}$.

It is helpful to consider a particular subset of models called *factorizations* or, simply, *factors*. These are triples $F = (\mathcal{Z}, f, \mathcal{D})$, where the encoder is a deterministic function $f : \mathcal{X} \to \mathcal{Z}$.

Factors and models may be visualized by the commuting diagram of Fig. 1.

What is the use of a model or a factor? Their use arises when considering channels as a resource. If we do not have access to a physical system that directly implements channel $\mathcal{C}$, we may resort to systems that instead implement channels $\mathcal{E}$ and $\mathcal{D}$ in a model $M = (\mathcal{Z}, \mathcal{E}, \mathcal{D})$ of $\mathcal{C}$. The more information required to store the intermediate variable Z, the more "costly" the model.

Models may be compared in quantifiable terms—through measurable features, such as the Rényi entropies of the intermediate variable Z—or they may be compared operationally. On the quantifiable side, one may pick a particular value of α and ask which model has the larger $H_\alpha(Z)$. Given two models $M_1 = (\mathcal{Z}_1, \mathcal{E}_1, \mathcal{D}_1)$ and $M_2 = (\mathcal{Z}_2, \mathcal{E}_2, \mathcal{D}_2)$ one often finds that M_1 is advantageous with respect to some α—$H_\alpha(Z_1) \leq H_\alpha(Z_2)$—while M_2 is advantageous with respect to others—$H_\alpha(Z_2) \leq H_\alpha(Z_1)$.

Sometimes, however, the advantage of M_1 over M_2 is unilateral. Given two models M_1 and M_2 of a channel $\mathcal{C}$, we say that M_1 is *strongly advantageous* with respect to M_2 if, for all $\alpha \geq 0$, $H_\alpha(Z_1) \leq H_\alpha(Z_2)$.

This differs from being *strictly* advantageous, which implies for at least one α we have $H_\alpha(Z_1) < H_\alpha(Z_2)$ without equality.

In contrast, the operational approach asks whether one model of a channel $\mathcal{C}$ can effectively substitute for another. For factorizations, there is a particularly useful operational comparison: given factorizations $F_1 = (\mathcal{Z}_1, f_1, \mathcal{D}_1)$ and $F_2 = (\mathcal{Z}_2, f_2, \mathcal{D}_2)$ of channel $\mathcal{C}$, we say that F_2 is *sufficient* for F_1 if there is a deterministic[1] function $g : \mathcal{Z}_2 \to \mathcal{Z}_1$ such that Fig. 2's diagram commutes.

The idea behind this comparison is that factorization F_2 itself can be factorized to yield the model F_1. In effect, F_2 already uses all the memory required to implement F_1, as well as some additional overhead. The takeaway, then, is that we might as well use F_1, as F_2 costs at least as much.

As a consequence of the concavity of Rényi entropies, the statement "F_2 is sufficient for F_1" necessarily implies that F_1 is strongly advantageous with respect to F_2. If F_2 is sufficient for F_1 and the function $g : \mathcal{Z}_2 \to \mathcal{Z}_1$ is not bijective, then F_1 is strictly advantageous with respect to F_2.

Our focus on factorizations pays off with the following result from computational mechanics: For each channel $\mathcal{C}$, there exists a *causal factorization* $F_\mathcal{C}$, unique up to isomorphism, for which all other factorizations F of $\mathcal{C}$ are sufficient [14]. Consequently, $F_\mathcal{C}$ is strongly minimal with respect to all other factorizations. The rule that $F_\mathcal{C}(x_1) = F_\mathcal{C}(x_2)$ if and only if $\mathbb{P}(Y = y | X = x_1) = \mathbb{P}(Y = y | X = x_2)$ defines $F_\mathcal{C}$ up to an isomorphism [13]. (As a consequence, the image of the factorization $Z_\mathcal{C}$ is the minimal sufficient statistic of X with respect to Y.)

Notice, that the data processing inequality provides the lower bound $H(Z) \geq I(X : Y)$ for all models and factor-

---

[1] The mapping g need not be explicitly deterministic, but if f_1 and f_2 are deterministic, then this will require g to be so as well.

izations of $\mathcal{C}$. This lower bound is generally not achievable with factorizations.

One might conjecture that the nonachievability of $I(X : Y)$ is a consequence of constraining our models to factorizations. However, the problem extends further. The more general problem of optimal *models* (in the broader, stochastic sense defined above) falls under the purview of common information theory [30, 31]. *Common information* refers to the problem of determining when a bivariate distribution XY may be jointly generated from a common variable Z by local stochastic maps. That is, if Z is distributed according to $\mathbb{P}(Z = z)$, then XY is distributed according to:

$$\mathbb{P}(x, y) = \sum_{z \in \mathcal{Z}} \mathbb{P}_A(x|z) \mathbb{P}_B(y|z) \mathbb{P}(z)$$

for channels $\mathbb{P}_A$ and $\mathbb{P}_B$. Any such joint generation scheme can be rewritten as a model of the channel $\mathcal{C}$ from $\mathcal{X}$ to $\mathcal{Y}$. In this way, results in the field of common information carry over to models of channels.

In particular, Wyner introduced a measure of common information that relates the asymptotic problem of modeling to the single-shot problem [30]. In the asymptotic regime, we can discuss a more general type of model. A (n, M, ϵ)-*model* of $\mathcal{C}$ is a pair of channels $\mathcal{E} : X \to Z$ and $\mathcal{D} : Z \to Y$ such that $\mathcal{E}(X)$ is uniform over M outcomes and:

$$\left\| \widehat{\mathbb{P}}(x, y) - \mathbb{P}(x, y) \right\|_1 < \epsilon ,$$

where $\widehat{\mathbb{P}}(x, y) = \mathbb{P}_{\mathcal{D} \circ \mathcal{E}}(y|x) \mathbb{P}(x)$.

That is, (n, M, ϵ) is a model that approximates n independent and identically distributed copies of the channel $\mathcal{C}$ with entropy cost $H(Z) = \log M$ and error ϵ. Error here is measured by the ℓ^1-norm $\|\cdot\|_1$. Wyner's result, strengthened by Winter and Ahlswede [33, 34], is that for any $\epsilon, \delta > 0$ there exists a sufficiently large n such that an (n, M, ϵ) model is Markov with $\frac{1}{n} \log M \le C(X : Y) + \delta$, where:

$$C(X : Y) = \min_{X - Z - Y} I(XY : Z) \qquad (1)$$

and the minimum is taken over all Z such that $X - Z - Y$. Conversely, for any $\delta > 0$, there exists a $\Delta > 0$ such that all (n, M, ϵ)-models with $\frac{1}{n} \log M \le C(X : Y) - \delta$ have error at least $\epsilon > \Delta$. In short, $C(X : Y)$ is the minimal asymptotic cost rate at which a channel can be approximately simulated.

Usually, $C(X : Y) > I(X : Y)$. And so, as in the case of factorizations, stochastic models of channels cannot reduce the memory cost all the way down to the mutual information.

The concepts just reviewed are various and diverse. After introducing equivalent notions for quantum models of channels, we will provide two examples that help elucidate them.

B. Quantum models of channels

The notion of a classical model readily generalizes to one of a *quantum model* for a classical channel. For clarity, however, we start by defining the important concepts from quantum information.

Instead of random variables, we have quantum states—positive operators ρ on a Hilbert space $\mathcal{H}$ such that $\mathrm{Tr}(\rho) = 1$. We assume $\mathcal{H}$ is finite-dimensional with dimension d, so that ρ can be decomposed in the form:

$$\rho = \sum_{i=1}^{d} \lambda_i |i\rangle\langle i| ,$$

for an orthogonal basis $\{|i\rangle\}$ and coefficients $\lambda_i > 0$ satisfying $\sum_i \lambda_i = 1$. The coefficients can be thought of as a probability distribution. Consequently, we can generalize the Rényi entropies to the quantum setting:

$$S_\alpha(\rho) := \frac{1}{1 - \alpha} \log \left(\sum_{i=1}^{d} \lambda_i^\alpha \right)$$
$$= \frac{1}{1 - \alpha} \log \mathrm{Tr}\rho^\alpha .$$

In the limits $\alpha \to 0$ and $\alpha \to 1$, we obtain the *max-entropy* $S_0(\rho) := \log \dim \mathcal{H}$ (useful for costs in single-shot or zero-error situations) and the *von Neumann entropy* $S(\rho) := -\sum_{i=1}^{d} \lambda_i \log \lambda_i = -\mathrm{Tr}(\rho \log \rho)$, respectively.

The interface between the quantum and classical world relies on two classes of operation. The first is an *ensemble*, represented here as a function $\mathcal{E} : \mathcal{X} \to \mathcal{S}(\mathcal{H})$ from a finite set $\mathcal{X}$ to the set $\mathcal{S}(\mathcal{H})$ of states on $\mathcal{H}$. It denotes a preparation of a quantum system using some initial classical information, stored in a random variable X. A *pure-state ensemble* is one for which $\mathcal{E}(x) = |\psi_x\rangle\langle\psi_x|$ for some (not necessarily orthogonal) vectors $\{|\psi_x\rangle\}$.

The second operation is *measurement*, represented as a stochastic map $\mathcal{M} : \mathcal{S}(\mathcal{H}) \to \mathcal{Y}$ from the set of states to a finite set $\mathcal{Y}$, with the constraint that $\mathbb{P}_\mathcal{M}(y|\rho) = \mathrm{Tr}(M_y \rho)$ for some set of positive operators $\{M_y\}$ satisfying $\sum_y M_y = 1$. This is known as a *positive-operator-valued measure* or POVM. We have a *projector-valued measure* or PVM when $M_y = |y\rangle\langle y|$ for some orthogonal basis $\{|y\rangle\}$.

Ensembles and measurements may be composed to create

a classical channel. We say $\mathcal{M} \circ \mathcal{E} = \mathcal{C}$ if:

$$\mathbb{P}_{\mathcal{C}}\left(y|x\right) = \mathrm{Tr}\left(M_y \mathcal{E}(x)\right) \ .$$

We are now ready to define a *quantum model of a channel* $\mathcal{C}$. This is the triplet $\mathcal{Q} = \{\mathcal{H}, \mathcal{E}, \mathcal{M}\}$ containing a Hilbert space $\mathcal{H}$, an ensemble $\mathcal{E} : \mathcal{X} \to \mathcal{B}(\mathcal{H})$, and a POVM $\mathcal{M} : \mathcal{B}(\mathcal{H}) \rightsquigarrow \mathcal{Y}$, such that $\mathcal{M} \circ \mathcal{E} = \mathcal{C}$.

Classically, we found it useful to specify *factors* of channels, which replaced the stochastic encoding map with a deterministic one. Here, the closest analogue is a model $\mathcal{Q} = \{\mathcal{H}, \mathcal{E}, \mathcal{M}\}$ where $\mathcal{E}$ is a pure-state ensemble. We call this a *pure-state model*.

However, each quantum model $\mathcal{Q} = \{\mathcal{H}, \mathcal{E}, \mathcal{M}\}$ also *induces* a classical factorization. The induced factorization of a model $\mathcal{Q}$ is the classical factorization $F_{\mathcal{Q}} = (\mathcal{Z}_{\mathcal{Q}}, f_{\mathcal{Q}}, \mathcal{D}_{\mathcal{Q}})$ of $\mathcal{C}$ such that $f_{\mathcal{Q}}(x) = f_{\mathcal{Q}}(x')$ if and only if $\rho_x = \rho_{x'}$. This factorization is unique up to isomorphisms. Note that each induced factorization must be a *refinement* of the causal factorization $F_{\mathcal{C}}$; that is, $F_Q(x_1) = F_Q(x_2)$ implies that $F_{\mathcal{C}}(x_1) = F_{\mathcal{C}}(x_2)$.

The notion of strongly advantageous carries over to the quantum domain. Let $\mathcal{Q}_1 = \{\mathcal{H}_1, \mathcal{E}_1, \mathcal{M}_1\}$ and $\mathcal{Q}_2 = \{\mathcal{H}_2, \mathcal{E}_2, \mathcal{M}_2\}$ be two quantum models. For a given input $\mathbb{P}(x)$, let $\rho_1 = \sum_x \mathbb{P}(x)\mathcal{E}_1(x)$ and $\rho_2 = \sum_x \mathbb{P}(x)\mathcal{E}_2(x)$ represent the model states. We say that $\mathcal{Q}_1$ is *strongly advantageous* with respect to $\mathcal{Q}_2$ if $S_\alpha(\rho_1) \leq S_\alpha(\rho_2)$ for all $\alpha \geq 0$.

We can also generalize the asymptotic notion of models. An (n, M, ϵ)-quantum model is given by an M-dimensional space $\mathcal{H}$, an ensemble $\mathcal{E} : \mathcal{X}^N \to \mathcal{B}(\mathcal{H})$, and $\mathcal{M} = \{M_{y^n} \in \mathcal{B}(\mathcal{H}) : y^n \in \mathcal{Y}^n\}$, such that $\rho = \sum_{x^n} \mathbb{P}(x^n)\mathcal{E}(x^n)$ is uniform over $\mathcal{H}$ and:

$$\left\| \widehat{\mathbb{P}}\left(x, y\right) - \mathbb{P}(x^n, y^n)\right\|_1 < \epsilon \ ,$$

defining $\widehat{\mathbb{P}}\left(x^n, y^n\right) := \mathbb{P}_{\mathcal{M} \circ \mathcal{E}}\left(y^n | x^n\right) \mathbb{P}(x)$.

C. Examples

The *Redundant Binary Symmetric Channel* [35] (RBSC) is a simple example that compactly illustrates the potential for quantum advantage in simulating classical channels. We define this channel $\mathcal{C}$ as follows:

$$\mathbb{P}(Y = y | X = x) = \begin{array}{c} X \backslash Y \\ \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{cc} 0 & 1 \\ \begin{pmatrix} 1-p & p \\ p & 1-p \\ 1-p & p \\ p & 1-p \end{pmatrix} \end{array}, \qquad (2)$$
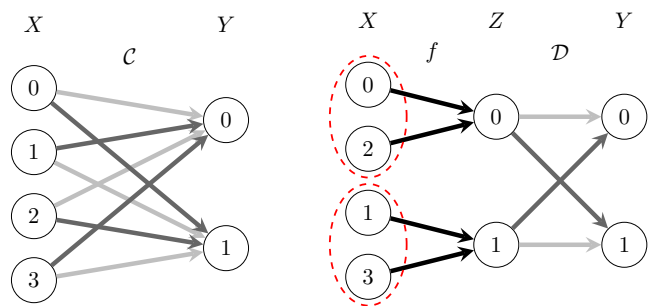

FIG. 3. Factorizing the Redundant Binary Symmetric Channel: (Left) Channel from X to Y. (Right) Intermediate variable Z cleaves channel into two steps—first deterministic, then stochastic. Arrow shade indicates conditional probability: black: 1, dark gray: $1 - p$, and light gray: p. Dashed ellipses show input equivalence classes. (Note re-ordering of $X = 1$ and $X = 2$.)

which is shown in Fig. 3 (left).

To obtain the correct conditional distribution for Y, distinguishing $x = 0$ from $x = 2$ and also $x = 1$ from $x = 3$ are unnecessary. We eliminate this redundancy by mapping to intermediate variable Z with the function $F : \mathcal{X} \to \mathcal{Z}$:

$$F(x) = \begin{cases} 0 \ , & x \in \{0, 2\} \\ 1 \ , & x \in \{1, 3\} \end{cases} \ .$$

The second factor (channel $\mathcal{C}'$) follows directly from this definition of f:

$$\mathbb{P}(Y = y | Z = z) = \begin{array}{c} Z \backslash Y \\ \\ 0 \\ 1 \end{array} \begin{array}{cc} 0 & 1 \\ \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix} \end{array},$$

Figure 3 shows this factorization for the RBSC channel.

It is fairly easy to see that $H_\alpha\left(Z\right) \leq H_\alpha\left(X\right)$ for any input random variable X and all $\alpha \geq 0$. And so, this factorization represents a compression of the input information, without losing information necessary for obtaining the correct output distribution.

Now, consider the quantum map $\mathcal{E} : \mathcal{X} \to \mathcal{B}\left(\mathcal{H}\right)$, given by:

$$\mathcal{E}(x) = \begin{cases} \rho_A = |A\rangle\langle A| \ , & \text{for } x = 0, 2 \\ \rho_B = |B\rangle\langle B| \ , & \text{for } x = 1, 3 \end{cases} \ ,$$

where:

$$|A\rangle \equiv \sqrt{1-p}\,|0\rangle + \sqrt{p}\,|1\rangle \ ,$$
$$|B\rangle \equiv \sqrt{p}\,|0\rangle + \sqrt{1-p}\,|1\rangle \ ,$$

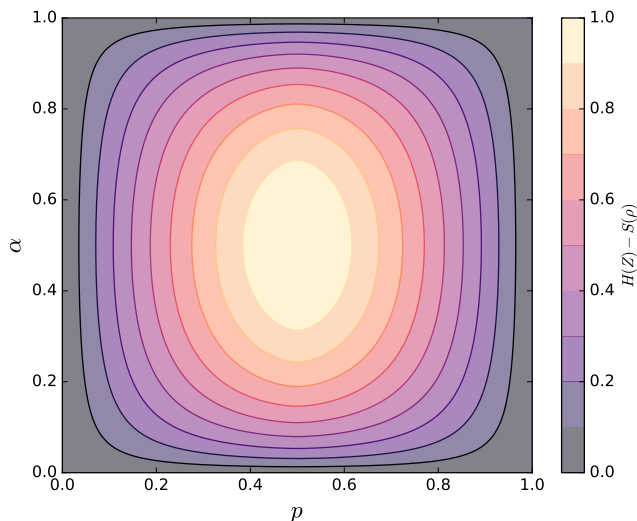with orthonormal basis $\{|0\rangle, |1\rangle\}$. If we use the mea-

FIG. 4. Factorization quantum advantage: $H(Z) - S(\rho)$ for quantum Redundant Binary Symmetric Channel with $\alpha = \mathbb{P}(Z = 0)$; see Fig. 3. Maximum advantage occurs at $p = \alpha = {}^1\!/_2$ and minima when $p = 0$ or $\alpha = 0$.

surement $\mathcal{M} = \{|0\rangle\langle0|, |1\rangle\langle1|\}$, we see that the channel is faithfully represented:

$$\mathbb{P}(Y = 0|X = x \in \{0,2\}) = \langle0|\rho_A|0\rangle = 1 - p ,$$
$$\mathbb{P}(Y = 1|X = x \in \{0,2\}) = \langle1|\rho_A|1\rangle = p ,$$
$$\mathbb{P}(Y = 0|X = x \in \{1,3\}) = \langle0|\rho_B|0\rangle = p ,$$
$$\mathbb{P}(Y = 1|X = x \in \{1,3\}) = \langle1|\rho_B|1\rangle = 1 - p .$$

Note that just as there were two classical intermediate states $\mathcal{Z} = \{0, 1\}$ for the RBSC, there are two quantum signal states ρ_0 and ρ_1. For a given distribution over inputs $\mathbb{P}(X)$, we can also compare the entropy of the classical and quantum intermediate variables. The entropy of the quantum mixed state $\rho = \sum_{x \in \mathcal{X}} \mathbb{P}(x)\mathcal{E}(x)$ is given by the *von Neumann entropy*:

$$S(\rho) = -\operatorname{tr}(\rho \log \rho) .$$

Importantly, this entropy is less than the classical entropy:

$$S(\rho) \leq H(Z) \leq H(X) ,$$

for any input distribution $\mathbb{P}(X)$ [36, 37]. Figure 4 shows the *quantum advantage* $H(Z) - S(\rho)$ as a function of input distribution $\mathbb{P}(X \in \{0, 2\}) = \alpha$ and channel parameter p. *The quantum factorization of the channel is, in this sense, more efficient than the reduced classical channel.*

Consider a second example. Let $\mathcal{X} = \mathcal{Y} = \{A, B, C\}$ and consider the channel $\mathcal{C}$, depicted in Fig. 5, given by the
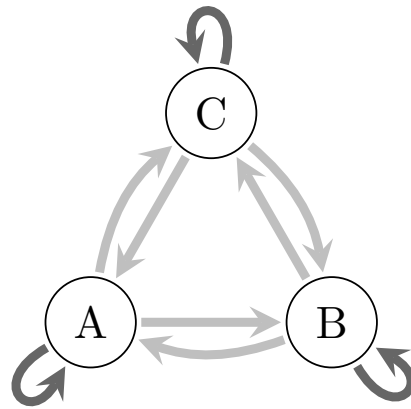

FIG. 5. A channel that maps the set $\{A, B, C\}$ to itself. Dark arrows represent probability $p = 2/3$ of transition and light arrows $p = 1/6$, so that all the outgoing edges of a node sum to 1.

probabilities:

$$\mathbb{P}(Y = y|X = x) = \begin{array}{c|ccc} X\backslash Y & A & B & C \\ \hline A & & & \\ B & & & \\ C & & & \end{array} \begin{pmatrix} 2/3 & 1/6 & 1/6 \\ 1/6 & 2/3 & 1/6 \\ 1/6 & 1/6 & 2/3 \end{pmatrix} .$$

This channel is adapted from a stochastic process in Ref. [21], and we call it the *3-MBW channel* after that process. Note that the channel is, effectively, its own causal factorization, as each $x \in X$ has a unique conditional distribution $\mathbb{P}(y|x)$. Thus, the channel is already its strongly advantageous factorization. With uniform input $\mathbb{P}(x) = 1/3$, we have $H_\alpha(X) = \log_2 3$ for all α.

Two quantum models of this channel may be constructed. The first, which we call $\mathcal{Q}_1$, has a Hilbert space $\mathcal{H}_1$ spanned by the orthogonal basis $\{|A\rangle, |B\rangle, |C\rangle\}$ with ensemble and POVM given by:

$$\mathcal{E}(x) = |\phi_x\rangle\langle\phi_x| ,$$
$$|\phi_x\rangle = \sum_{y \in \mathcal{Y}} \sqrt{\mathbb{P}(y|x)} |y\rangle , \text{ and}$$
$$M_y = |y\rangle\langle y| .$$

Note that this induces the causal factorization. For the uniform input $\mathbb{P}(x) = 1/3$, we get an entropy cost of $S(\rho) \approx 0.69$ qubits and a dimension cost of $S_0(\rho) = \log_2 3$ qubits. Thus, we see the model improves on the classical case in entropy, but not in dimension. This type of model is inspired by the *q-machine* for processes [16] and generalizes easily to other channels.

The second model, which we call $\mathcal{Q}_0$, has a Hilbert space $\mathcal{H}_0$ spanned by the orthogonal basis $\{|0\rangle, |1\rangle\}$ with en-

semble and POVM given by:

$$\rho_x = |\psi_x\rangle\langle\psi_x| \ ,$$

$$|\psi_x\rangle = \begin{cases} |0\rangle & x = \text{A} \\ \frac{1}{2}|0\rangle + \frac{\sqrt{3}}{2}|1\rangle & x = \text{B} \\ \frac{1}{2}|0\rangle - \frac{\sqrt{3}}{2}|1\rangle & x = \text{C} \end{cases} \ , \text{ and}$$

$$M_y = \frac{2}{3}|\psi_y\rangle\langle\psi_y| \ .$$

This, too, induces the causal factorization. For the uniform input, we get an entropy and dimension cost of $S(\rho) = S_0(\rho) = 1$ qubits. In both of these quantifications, this model improves upon the classical case.

The other feature worth noting is that $\mathcal{Q}_0$ is advantageous over $\mathcal{Q}_1$ in the dimension cost S_0, but the opposite is true for the entropy cost S_1. (Indeed, this motivated the subscripts.) Reference [21] demonstrated that, for the Markov process version of this channel, the model $\mathcal{Q}_0$ was actually only two-dimensional model of the process. This result trivially carries over to the channel itself, which means $\mathcal{Q}_0$ is the model that *uniquely* minimizes S_0. However, it clearly does not minimize S_1. This demonstrates that, for at least this channel, no "strongly minimal" pure-state quantum model exists. This contrasts starkly to the causal factorization, which is strongly minimal among classical factorizations.

III. QUANTUM MODEL ADVANTAGE

These examples hint at rather more general results on quantum advantage.

A. Single-shot: Constraints on minimal models

Our first pair of results is motivated by the questions: What models are optimal? And, how do they compare, generally, to the causal factorization for classical models? We begin with the observation, stated earlier, that each quantum model Q of a channel $\mathcal{C}$ induces a classical factorization F_Q. Recall that each induced factorization must be a refinement of the causal factorization $F_{\mathcal{C}}$.

Classically, a refinement to the causal factorization can be *merged*, where we combine causally equivalent inputs until the causal factorization is achieved, minimizing the Rényi entropies. Can we do the same for quantum models? There is, in fact, a process of *state-merging* that allows for the algorithmic reduction of any Rényi entropy individually.

The goal of state-merging is to combine the partitions of causally equivalent states. In this case, state merging

on a quantum model means modifying an ensemble $\mathcal{E}(x)$, where $\mathcal{E}(x_1) \neq \mathcal{E}(x_2)$ for some $x_1, x_2 \in \mathcal{X}$ with $x_1 \sim_C x_2$, into a new $\mathcal{E}'(x)$ so that $\mathcal{E}'(x_1) = \mathcal{E}'(x_2)$. This can be accomplished in at least two ways:

1. $\mathcal{E}'(x_1) = \mathcal{E}(x_2)$ and $\mathcal{E}'(x) = \mathcal{E}(x)$ for $x \neq x_1$ or
2. $\mathcal{E}'(x_2) = \mathcal{E}(x_1)$ and $\mathcal{E}'(x) = \mathcal{E}(x)$ for $x \neq x_2$.

Recalling that the average model state is written as:

$$\rho = \sum_x \mathbb{P}(x)\mathcal{E}(x) \ ,$$

we construct two new average model states:

$$\rho^{(1\to2)} = \sum_{x\neq x_1} \mathbb{P}(x)\mathcal{E}(x) + \mathbb{P}(x_1)\mathcal{E}(x_2) \text{ and}$$

$$\rho^{(2\to1)} = \sum_{x\neq x_2} \mathbb{P}(x)\mathcal{E}(x) + \mathbb{P}(x_2)\mathcal{E}(x_1) \ .$$

Each results from two distinct forms of state merging. It is important to determine if such a technique reduces memory costs. This leads to a lemma that underlies our first main result.

Lemma 1. *Let $f : \mathbb{R} \to \mathbb{R}$ be a concave function and let $S_f(\rho) = \text{Tr}(f(\rho))$. Then:*

$$S_f(\rho) \geq \min\{S_f(\rho^{(1\to2)}), S_f(\rho^{(2\to1)})\} \ . \quad (3)$$

Proof. *Let $f : \mathbb{R} \to \mathbb{R}$ be a concave function and let $S_f(\rho) = \text{Tr}(f(\rho))$. Define:*

$$\alpha = \mathbb{P}(x_1)/(\mathbb{P}(x_1) + \mathbb{P}(x_2))$$
$$1 - \alpha = \mathbb{P}(x_2)/(\mathbb{P}(x_1) + \mathbb{P}(x_2)) \ .$$

One can check that:

$$\rho = \alpha\rho^{(1\to2)} + (1-\alpha)\rho^{(2\to1)} \ .$$

By f's concavity:

$$S_f(\rho) \geq \alpha S_f(\rho^{(1\to2)}) + (1-\alpha)S_f(\rho^{(2\to1)})$$
$$\geq \min\{S_f(\rho^{(1\to2)}), S_f(\rho^{(2\to1)})\} \ . \quad (4)$$

So, either $\rho^{(1\to2)}$ or $\rho^{(2\to1)}$ must be smaller according to f.

From this useful bound follows one of our primary results:

Result 1 (Causal Model Optimality)**.** *A quantum model of a channel $\mathcal{C}$ minimizes $S_f(\rho)$, where ρ is the average of the models' ensemble and f is any concave function, only if it induces the causal factorization of $\mathcal{C}$.*

To see this, suppose otherwise: There exists a quantum model $\mathcal{Q}$ that does not induce the causal factorization, but

that is f-minimal. By Eq. (3), though, we can perform a state-merging that yields a model $\mathcal{Q}'$ with $S_f(\rho') \leq S_f(\rho)$. This is a contradiction.

Of course, this result applies to the Rényi entropies $S_\alpha(\rho)$ as well: $S_\alpha(\rho) = S_f(\rho)/(1-\alpha)$, where $f(x) = x^\alpha$.

There are some caveats to this result. The first is that a minimum does not *necessarily* exist. The space of all possible models contains models of arbitrarily high Hilbert space dimension, and so it is not compact. This is still true even if we restrict to the space of models that induce the causal factorization: There is no restriction on the dimension of the states $\mathcal{E}(x)$.

However, if we restrict to pure-state models that induce the causal factorization, then the states $\mathcal{E}(x)$ can span a space of dimension no larger than $\dim \mathcal{H} = |\mathcal{Z}_\mathcal{C}|$. This makes the model-space compact and ensures the existence of a minimal model for each S_α.

Now, in the classical case, the causal factorization was unique and *strongly minimal*, in that it minimized all H_α. Our result shows that only quantum models that induce the causal factorization can be minimal. However, it does not prove that *strongly* minimal models exist. As we saw in the example of the 3-MBW channel, there may be models that are only *weakly* minimal, in that they only minimize some Rényi entropies and not others.

Our second main result focuses on comparing pure-state quantum models with the classical causal factorization itself. Pure-state quantum models, recall, are a quantum analogue to the factorization. Furthermore, as we just saw, pure-state quantum models that minimize any Rényi entropy must induce the causal factorization. Thus, to determine if and when they improve upon their classical counterpart, it makes sense to compare pure-state models that induce the causal factorization with the causal factorization itself.

Let $\mathcal{Q}$ be a pure-state quantum model that induces the causal factorization. Then its average model state is:

$$\rho = \sum_z \mathbb{P}(z) \, |\psi_z\rangle \, \langle\psi_z| \ ,$$

where the sum is over causal states $z \in \mathcal{Z}_\mathcal{C}$ and the states $|\psi_z\rangle$ correspond to the states of the quantum model: $\mathcal{E}(x) = |\psi_z\rangle\langle\psi_z|$, if $z = f_\mathcal{C}(x)$.

We can also consider the orthogonal decomposition:

$$\rho = \sum_i \lambda_i \, |i\rangle \, \langle i| \ .$$

The vector of coefficients λ_i and the probability distribution $\mathbb{P}(z)$ are related by a unitary matrix U_{iz} such that

[37]:

$$\mathbb{P}(z) = \sum_i |U_{iz}|^2 \, \lambda_i \ .$$

U_{iz} is the identity matrix if and only if the states $\{|\psi_z\rangle\}$ are orthogonal. This establishes a *majorization* relationship between the two vectors, written $\lambda_i \succeq \mathbb{P}(z)$ [38]. The importance of this result here is that, for all concave $\alpha \geq 0$, we have $S_\alpha(\rho) \leq H_\alpha(Z_\mathcal{C})$; with equality for $\alpha > 0$ if and only if the states $\{|\psi_z\rangle\}$ are orthogonal.

This gives our second main result:

Result 2 (Strong/strict advantage of quantum models)**.** *Given any input distribution $\mathbb{P}(x)$, a pure-state quantum model of a channel $\mathcal{C}$ that induces the causal factorization $F_\mathcal{C}$ satisfies $S_\alpha(\rho) \leq H_\alpha(Z_\mathcal{C})$, where $\rho := \sum_x \mathbb{P}(x)\mathcal{E}(x)$ is the average state. For $\alpha \geq 0$, the above bound is strict whenever the ensemble $\mathcal{E}(x)$ is nonorthogonal.*

This result generalizes a similar result from Ref. [21] to channels. In fact, it strengthens it. Rather than $\mathbb{P}(z)$ being determined by a process' unique stationary state distribution, here it can be determined by any input distribution $\mathbb{P}(x)$, and the quantum model state is still strongly advantageous.

B. Asymptotics: A lower bound

The previous section focused on pure states. Now, we return to mixed states to study the general asymptotic case.

Earlier we defined the (n, M, ϵ)-quantum model. We introduce here a single-letter optimization that lower-bounds the rate $\frac{1}{n} \log M$ at which a channel can be quantally modeled, based on the classical Wyner common information. Unlike the classical case, we do not claim this bound is tight.

To set it up, consider a d-dimensional model $(\mathcal{H}, \mathcal{E}, \mathcal{M})$ of the variables XY, with $\rho = \sum_x \mathbb{P}(x)\mathcal{E}(x)$. We define the *Holevo information*:

$$I_\chi(X : \rho) = S(\rho) - \sum_x \mathbb{P}(x) S(\mathcal{E}(x)) \ .$$

For each state in the ensemble, we do an orthogonal decomposition:

$$\mathcal{E}(x) = \sum_{i=1}^d \lambda_{i|x} \, |i, x\rangle\langle i, x| \ ,$$

and define the conditional measurement information:

$$I_\mu \left(Y : \rho | X \right)$$
$$= \sum_{x \in \mathcal{X}} \sum_{i=1}^{d} \mathbb{P}\left(x, i, y \right) \log \left(\frac{\mathbb{P}\left(i, y | x \right)}{\mathbb{P}\left(y | x \right) \lambda_{i|x}} \right)$$

where:

$$\mathbb{P}\left(x, i, y \right) = \mathbb{P}(x) \lambda_{i|x} \left\langle i, x | M_y | i, x \right\rangle \ .$$

(Note that $\mathbb{P}(x, y) = \sum_i \mathbb{P}\left(x, i, y \right)$.) With these two, we define:

$$I \left(XY : \rho \right) = I_\chi \left(X : \rho \right) + I_\mu \left(Y : \rho | X \right) \ .$$

Another way to approach this quantity is to define:

$$\rho_{xy} = \sum_{i=1}^{d} \frac{\lambda_{i|x} \left\langle i, x | M_y | i, x \right\rangle}{\mathbb{P}\left(y | x \right)} | i, x \rangle \langle i, x |$$

as a state representing our knowledge, after the measurement, about what i *was* before measurement. (This is not the final state after measurement, as the measurement itself induces further modification.) We then have:

$$I \left(XY : \rho \right) = S(\rho) - \sum_{x,y} \mathbb{P}\left(x, y \right) S \left(\rho_{xy} \right) \ .$$

This is equivalent to our earlier definition, but expresses the quantity as a kind of Holevo quantity.

We define an analogue to the Wyner common information as:

$$C_{q,l} \left(X \rangle Y \right) = \inf_{X - \rho - Y} I \left(XY : \rho \right) \ .$$

where $X - \rho - Y$ indicates that we are varying over all models of XY. The notation $X \rangle Y$ indicates that this is not a symmetric quantity between X and Y.

Notice that we define this quantity as an infimum and not a minimum, as Wyner did. This is because Wyner's proof [30] demonstrated that the infimum could be attained on a compact subset of the model space. This result relies on being able to decompose $I \left(XY : W \right)$ as $H \left(XY \right) - H \left(XY | W \right)$. The minimization of $I \left(XY : W \right)$ becomes the maximization of $H \left(XY | W \right)$. And this can be written as a linear function of $\mathbb{P}(w)$:

$$H \left(XY | W \right) = - \sum_{w} \mathbb{P}(w) \sum_{x,y} \mathbb{P}\left(x, y | w \right) \log \mathbb{P}\left(x, y | w \right) \ ,$$

if we hold $\mathbb{P}\left(x, y | w \right)$ fixed.

It is the linearity of the optimization function, in the classical case, that allows the assumption of a compact

minimization. However, our quantum generalization $I \left(XY : \rho \right)$ cannot be expressed in a way that allows the optimization to be linearized. This is why we cannot assume the existence of a minimum.

Our third main result can now be stated as follows.

Theorem 1. *Let $\delta > 0$. Then there exists a fixed $\Delta > 0$ such that, for all quantum (n, M, ϵ)-models, $\frac{1}{n} \log M \leq C_{q,l} \left(X : Y \right) - \delta$ implies $\epsilon > \Delta$.*

In other words, if the model's cost rate is bounded below by $C_{q,l}$, then its error is bounded away from zero. To achieve arbitrary accuracy, it must have a rate of at least $C_{q,l}$.

Proof. *The argument follows the same logic as Wyner's [30] and proceeds by contradiction. Suppose with δ fixed that, for any $\Delta > 0$, we can find a (n, M, ϵ)-model with $\frac{1}{n} \log M \leq C_{q,l} \left(X : Y \right) - \delta$ and $\epsilon < \Delta$. For such a model:*

$$\frac{1}{n} \log M \geq \frac{1}{n} I \left(X^n \widehat{Y}^n : \rho \right)$$
$$\geq \frac{1}{n} H \left(X^n \widehat{Y}^n \right) - \frac{1}{n} H \left(X^n \widehat{Y}^n \middle| \rho \right) \ ,$$

where:

$$H \left(X^n \widehat{Y}^n \middle| \rho \right) = H_\chi \left(X^n | \rho \right) + H_\mu \left(Y^n | \rho, X^n \right) \ ,$$
$$H_\chi \left(X^n | \rho \right) = H(X^n) - I_\chi \left(X^n : \rho \right) \ , \ and$$
$$H_\mu \left(\widehat{Y}^n \middle| \rho, X \right) = H \left(\widehat{Y}^n \right) - I_\mu \left(Y^n : \rho | X^n \right) \ .$$

We see that $H_\mu \left(\widehat{Y}^n \middle| \rho, X \right)$ is explicitly a conditional entropy by definition and so is subadditive. The subadditivity of $H_\chi \left(X^n | \rho \right)$ is follows from the strong subadditivity of von Neumann entropy. Subadditivity implies that:

$$H \left(X^n \widehat{Y}^n \middle| \rho \right) \leq \sum_{k} H \left(X_k \widehat{Y}_k \middle| \rho \right) \ .$$

We then have:

$$\frac{1}{n} \log M \geq \frac{1}{n} H \left(X^n \widehat{Y}^n \right) - \frac{1}{n} \sum_{k} H \left(X_k \widehat{Y}_k \middle| \rho \right) \ .$$

By assumption, we can drive Δ low enough that $H \left(X^n \widehat{Y}^n \right) \geq H(X^n Y^n) - O\left(\Delta \right)$ where $O\left(\Delta \right)$ is the standard big-O notation. Since $H \left(X^n Y^n \right)$ is separable, we can write $H \left(X^n \widehat{Y}^n \right) = \sum_k H \left(X_k \widehat{Y}_k \right) - O\left(\Delta \right)$ for some $\eta > 0$. Thus:

$$\frac{1}{n} \log M \geq \frac{1}{n} \sum_{k} I \left(X_k \widehat{Y}_k : \rho \right) - O\left(\Delta \right) \ .$$

This is a sum over a collection of single-letter models of channels that are at least Δ-close to XY.

Each such model is then Δ-close to another model ρ_k that perfectly generates XY. In each case we have $I\left(X_k \widehat{Y}_k : \rho\right) \geq I\left(XY : \rho_k\right) - O\left(\Delta\right)$. And, for each k we have $I\left(XY : \rho_k\right) \geq C_{q,l}\left(X \rangle Y\right)$. Then:

$$\frac{1}{n} \log M \geq \frac{1}{n} \sum_k C_{q,l}\left(X \rangle Y\right) - O\left(\Delta\right)$$

$$= C_{q,l}\left(X \rangle Y\right) - \frac{1}{n} O\left(\Delta/n\right) \ .$$

Now, take Δ small enough so that the error term is less than $n\delta$. We get:

$$\frac{1}{n} \log M \geq C_{q,l}\left(X \rangle Y\right) - \delta \ .$$

However, this contradicts the assumption that $\frac{1}{n} \log M \leq C_{q,l}\left(X \rangle Y\right) - \delta$, completing the proof.

Thus, $C_{q,l}\left(X \rangle Y\right)$ offers a lower-bound on the rate at which one can model the channel $\mathcal{C}$. We now discuss the ways in which this lower-bound is lacking.

First, as we mentioned before, there is no guarantee that this bound can be easily computed via an optimization over a compact set. This arises from the optimization's nonlinearity.

Second, it lacks a direct counterpart—that is, a theorem showing that $C_{q,l}\left(X \rangle Y\right)$ is achievable and not just a lower bound. Again, this stems from a fundamental difference between the setting of Wyner's classical proof and the quantum models we are considering. Wyner's achievability proof turns on showing that, given a single-shot model of a channel $\mathcal{C}$, one can take asymptotically many copies of the single-shot model and use random coding to trim off excess information.

For a quantum model, though, random coding means the random selection of subspaces. If the states ρ_x of the ensemble commute, we may choose subspaces that commute with them all. Otherwise, coherence effects destroy the random coding approximation. In other words, there is a trade-off between minimizing the size of the single-shot model through the use of quantum coherence and, if coherence is present, the inability to further reduce the rate using asymptotic random coding. Finding an achievable lower bound on quantum models requires balancing these two tensions.

Finally, we note simply that $C_{q,l}\left(X \rangle Y\right)$ is not symmetric. While this may not seem necessary—after all, the channel simulation question is not one that is manifestly symmetric—we show in the following section that the quantum simulation of channels is equivalent to another question, involving entanglement, that *is* manifestly symmetric. We propose that this alternative formulation offers a more productive avenue of investigation.

C. Equivalence to common entanglement

Suppose we have a channel $\mathcal{C} : \mathcal{X} \rightarrow \mathcal{Y}$, a quantum model $\mathcal{Q} = (\mathcal{H}, \mathcal{E}, \mathcal{D})$, and an input $X \sim \mathbb{P}(x)$. We will show how to reformulate these elements into a symmetric picture, where the joint random variable XY is generated by taking local measurements on a bipartite entangled state $|\Psi_{AB}\rangle$. Just as the generation of a joint random variable from a classical shared variable with local operations is called common information, the generation of a joint random variable from shared entanglement is here called *common entanglement*.

For a mixed state ρ on a space $\mathcal{H}_B$ with dimension d, we can write its orthogonal decomposition as:

$$\rho = \sum_{i=1}^{d} \lambda_i |i\rangle_B |i\rangle_B \ .$$

We assume without loss of generality that $\mathcal{H}_B$ is entirely spanned by ρ's support; that is, $\lambda_i > 0$ for all $i = 1, \dots, d$. We define ρ's *purification* as a state that exists on the space $\mathcal{H}_A \otimes \mathcal{H}_B$, where $\mathcal{H}_A$ is a copy of $\mathcal{H}_B$ with dimension d as well. ρ's purification is:

$$|\psi\rangle_{AB} = \sum_{i=1}^{d} \sqrt{\lambda_i} |i\rangle_A \otimes |i\rangle_B \ ,$$

where the basis $\{|i\rangle_A\}$ is unitarily equivalent to the original $\{|i\rangle_B\}$. It is a purification in the sense that $\text{Tr}_B\left(|\psi\rangle_{AB}\right) = \rho$. That is, one can imagine the mixed state only exists due to ignorance of a larger system. Up to unitary transformations on $\mathcal{H}_A$, this purification is unique.

Ensembles $\mathcal{E} : \mathcal{X} \rightarrow \mathcal{B}_+\left(\mathcal{H}\right)$ may also be purified. Our scheme here resembles that in Ref. [39] except that we allow the ensemble states $\mathcal{E}\left(x\right)$ to be mixed. Given an input $\mathbb{P}\left(x\right)$, we write:

$$\rho = \sum_{x \in \mathcal{X}} \mathcal{E}(x)$$

$$= \sum_{x \in \mathcal{X}} \mathbb{P}(x) \rho_x$$

as the ensemble decomposition of ρ on $\mathcal{H}_B$. Let $|\psi\rangle_{AB}$ be a purification of ρ and $|\psi_x\rangle_{AB}$ be a purification of the *unnormalized* state $\mathbb{P}(x)\mathcal{E}(x)$ (both onto $\mathcal{H}_A \otimes \mathcal{H}_B$).

Letting $\{|i,x\rangle\}$ be the diagonalizing basis of $\mathbb{P}(x)\mathcal{E}(x)$ and

$\lambda_{i|x}$ be its spectra, we can write:

$$
\begin{aligned}
|\psi_x\rangle_{AB} &= \sum_{i=1}^d \sqrt{\lambda_{i|x}}\,|i,x\rangle_A \otimes |i,x\rangle_B \\
&= \sum_{j=1}^d \sum_{i=1}^d \sqrt{\lambda_{i|x}}\,|i,x\rangle_A \otimes (\langle j|i,x\rangle\,|j\rangle_B) \\
&= \sum_{j=1}^d \sqrt{\lambda_j}\left(\sum_{i=1}^d \sqrt{\frac{\lambda_{i|x}}{\lambda_{i|x}}}\,\langle j|i,x\rangle\,|i,x\rangle_A\right)\otimes |j\rangle_B \\
&= \sum_{\omega\in\Omega} \sqrt{\lambda_j}\,|\Phi_{j,x}\rangle_A \otimes |j\rangle_B \;,
\end{aligned}
$$

where:

$$
|\Phi_{j,x}\rangle_A = \sum_{i=1}^d \sqrt{\frac{\lambda_{i|x}}{\lambda_j}}\,\langle j|i,x\rangle\,|i,x\rangle_A
$$

For each $x \in \mathcal{X}$, there is a unique operator K_x on $\mathcal{H}_A$ such that:

$$
K_x\,|j\rangle_A = |\Phi_{j,x}\rangle_A \;.
$$

Furthermore:

$$
\begin{aligned}
\langle j|\sum_x K_x^\dagger K_x|j'\rangle &= \sum_{x\in\mathcal{X}} \langle\Phi_{j,x}|\Phi_{j',x}\rangle \\
&= \sum_{x\in\mathcal{X}}\sum_{i=1}^d \frac{\lambda_{i|x}}{\sqrt{\lambda_j\lambda_{j'}}}\,\langle j|i,x\rangle\,\langle i,x|j'\rangle \\
&= \frac{1}{\sqrt{\lambda_j\lambda_{j'}}}\,\langle j|\rho|j'\rangle = \delta_{jj'} \;.
\end{aligned}
$$

This shows that the $\{K_x\}$ are the *Kraus operators* of a complete POVM $\mathcal{M} = \{M_x\}$ with $M_x = K_x^\dagger K_x$.

To summarize, an ensemble $\mathcal{E}(x)$ has a unique purification $(|\psi\rangle_{AB}, \mathcal{M}_A)$ satisfying:

$$
\begin{aligned}
\mathrm{Tr}\left(|\psi\rangle\langle\psi|_{AB}\right) &= \rho_B \\
\mathrm{Tr}\left(M_x\,|\psi\rangle\langle\psi|_{AB}\right) &= \mathbb{P}(x)\,\mathcal{E}(x) \;,
\end{aligned}
$$

where $\rho_B = \sum_x \mathbb{P}(x)\mathcal{E}(x)$. This purification is unique up to transformations of the form $K_x \mapsto UK_xV_x^\dagger$ for unitary matrices U and $\{V_x\}$ on $\mathcal{H}_A$. (These arise from the unitary ambiguity of each individual purification of ρ and $\mathcal{E}(x)$.)

Correspondingly, the *common entanglement representation* of a quantum model $\mathcal{Q} = \{\mathcal{H}_B, \mathcal{E}, \mathcal{M}\}$ is the triplet $\Phi = (|\Phi\rangle_{AB}, \mathcal{M}_A, \mathcal{M}_B)$, where $(|\Phi\rangle_{AB}, \mathcal{M}_A)$ is the purification of the ensemble $\mathcal{E}$ and $\mathcal{M}_B = \mathcal{M}$.

In this representation the distribution $\mathbb{P}(x,y)$ is seen as being generated by the local POVM $\mathcal{M}_A \otimes \mathcal{M}_B = \{M_x \otimes N_y\}$ from a shared entangled pair. In the language
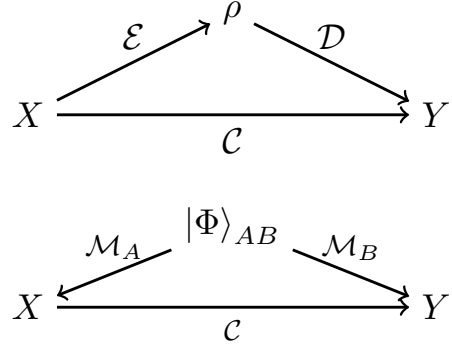

FIG. 6. Two equivalent pictures: (Top) Directional picture modeling a channel $\mathcal{C} : \mathcal{X} \rightsquigarrow \mathcal{Y}$ with an intermediate quantum system ρ. (Bottom) Symmetric picture generating a joint random variable XY from an entangled state $|\Phi\rangle_{AB}$.

of entanglement theory the representation is equivalent to simulating the channel via a quantum ensemble. See Fig. 6.

For every quantum model $\mathcal{Q} = \{\mathcal{H}_B, \mathcal{E}, \mathcal{M}\}$ of the channel $\mathcal{C} : \mathcal{X} \rightsquigarrow \mathcal{Y}$ and input random variable $X \sim \mathbb{P}(x)$, the joint random variable XY has a common entanglement generator $\Phi = (|\Phi\rangle_{AB}, \mathcal{M}_A, \mathcal{M}_B)$. This alternative perspective on the reverse Holevo problem will prove fruitful for future extensions of our results.

IV. CONCLUDING REMARKS

We inverted the question originally posed by Holevo. Rather than using classical resources to quantify the behavior of quantum channels, we analyzed how well quantum resources may simulate the workings of a classical channel.

This question makes sense in the context of wider considerations. We do not claim that it is practically efficient to replace classical channels with quantum ones. Attempts to construct and use quantum channels, such as in quantum teleportation, often require classical communication to work. However, this is a consequence of the difficulties in transmitting quantum information over distances. When we discuss channels as a microcosm of the inner workings of a stochastic process, in contrast, we are invoking information that is retained over time. This information is memory. When we measure the Rényi entropies of the average quantum state of a model, we are subjecting it to various memory metrics.

Thus, our results here are best interpreted as a study not in spatial communication, but in the memory required to simulate a temporal channel. In this setting, we showed that:

1. In the single-shot regime, those quantum models minimizing various memory metrics are those that induce the causal state.

2. Conversely, pure-state models that induce the causal state improve upon the best classical factorization across all memory metrics, and

3. In the asymptotic regime, for arbitrary accuracy to be achieved the quantum memory rate required for simulation is bounded below by the new quantum common information quantity for channels— $C_{q,l}(X\rangle Y)$—introduced here.

Many questions remain. For instance, under what conditions is $C_{q,l}(X\rangle Y)$ achievable? When it is not, can we derive a tighter lower bound? The common entanglement representation of the problem in Section III C should prove useful in furthering study of these questions.

It also remains to rigorously connect our results for channels back to the problem of simulating stochastic processes with quantum resources. The rates achievable for channels can be applied to the *process channel* $\mathbb{P}(x_{0:\infty}|x_{-\infty:0})$ from the past to the future. These would give lower bounds on the rates at which we can *sequentially generate* a process. Likely, they would not be tight. Thus, as results for simulating channels are derived, the additional constraints characterizing sequential generation must also be formulated.

[1] D. A. Meyer. Quantum computing classical physics. *Phil. Trans. Roy. Soc. London A*, 360(1792):395–405 1364–503X, 2002. 1

[2] M. H. Yung, D. Nagaj, J. D. Whitfield, and A. Aspuru-Guzik. Simulation of classical thermal states on a quantum computer: A transfer-matrix approach. *Phys. Rev. A*, 82(6):060302, 2010. 1

[3] J. Yepez. Quantum lattice-gas model for computational fluid dynamics. *Phys. Rev. E*, 63(4):046702, 2001. 1

[4] J. Yepez. Quantum computation of fluid dynamics. In *Quantum Computing and Quantum Communications*, pages 34–60. Springer, 1999. 1

[5] S. Sinha and P. Russer. Quantum computing algorithm for electromagnetic field simulation. *Quant. Info. Processing*, 9(3):385–404 1

[6] J. Yepez. Quantum lattice-gas model for the diffusion equation. *Intl. J. Mod. Physics C*, 12(09):1285–1303 1

[7] G. P. Berman, A. A. Ezhov, D. I. Kamenev, and J. Yepez. Simulation of the diffusion equation on a type-ii quantum computer. *Phys. Rev. A*, 66(1):012310, 2002. 1

[8] J. Yepez. Quantum lattice-gas model for the Burgers equation. *J. Stat. Physics*, 107(1-2):203–224 1

[9] S. A. Harris and V. M. Kendon. Quantum-assisted biomolecular modelling. *Phil. Trans. Roy. Soc. London A*, 368(1924):3581–3592 1364–503X, 2010. 1

[10] P. W. Shor. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM Review*, 41(2):303–332 1

[11] L. K. Grover. A fast quantum mechanical algorithm for database search. In *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing*, pages 212–219 1

[12] A. W. Harrow, A. Hassidim, and S. Lloyd. Quantum algorithm for linear systems of equations. *Phys. Rev. Let.*, 103(15):150502, 2009. 1

[13] J. P. Crutchfield and K. Young. Inferring statistical complexity. *Phys. Rev. Let.*, 63:105–108, 1989. 1, 3

[14] C. R. Shalizi and J. P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *J. Stat. Phys.*, 104:817–879, 2001. 3

[15] J. P. Crutchfield. Between order and chaos. *Nature Physics*, 8:17–24, 2012. 1, 2

[16] M. Gu, K. Wiesner, E. Rieper, and V. Vedral. Quantum mechanics can reduce the complexity of classical models. *Nature Comm.*, 3(762), 2012. 1, 6

[17] R. Tan, D. R. Terno, J. Thompson, V. Vedral, and M. Gu. Towards quantifying complexity with quantum mechanics. *Eur. J. Phys. Plus*, 129:191, 2014.

[18] J. R. Mahoney, C. Aghamohammadi, and J. P. Crutchfield. Occam's quantum strop: Synchronizing and compressing classical cryptic processes via a quantum channel. *Scientific Reports*, 6(20495), 2016. 1

[19] P. M. Riechers, J. R. Mahoney, C. Aghamohammadi, and J. P. Crutchfield. Minimized state-complexity of quantum-encoded cryptic processes. *Phys. Rev. A*, 93(5):052317, 2016.

[20] F. C. Binder, J. Thompson, and M. Gu. A practical, unitary simulator for non-Markovian complex processes. *Phys. Rev. Lett.*, 120:240502, 2017. 1

[21] S. Loomis and J. P. Crutchfield. Strong and weak optimizations in classical and quantum models of stochastic processes. *J. Stat. Phys.*, 176(6):1317–1342, 2019. 1, 6, 7, 8

[22] Q. Liu, T. J. Elliot, F. C. Binder, C. D. Franco, and M. Gu. Optimal stochastic modelling with unitary quantum dynamics. arXiv:1810.09668. 1

[23] C. Bennett, P. W. Shor, J. A. Smolin, and A. V. Thapliyal. Entanglement-assisted classical capacity of noisy quantum channels. *Phys. Rev. Lett.*, 83:3081, 1999. 1

[24] C. Bennett, P. W. Shor, J. A. Smolin, and A. V. Thapliyal. Entanglement-assisted capacity of a quantum channel and the reverse Shannon theorem. *IEEE Trans. Inf. Theory*, 48, 2002. 1

[25] C. Bennett, I. Devetak, A. W. Harrow, P. W. Shor, and A. Winter. The quantum reverse Shannon theorem and resource tradeoffs for simulating quantum channels. *IEEE Trans. Inf. Theory*, 60, 2014. 1

[26] A. S. Holevo. Bounds for the quantity of information transmitted by a quantum communication channel. *Problems of Information Transmission*, 9:177, 1973. 1

[27] B. Schumacher and M. Westmoreland. Sending classical information via noisy quantum channels. *Phys. Rev. A*, 56:131, 97a.

[28] A. S. Holevo. Coding theorems for quantum channels. *Russian Math. Surveys*, 53:1295–1331, 1996.

[29] A. S. Holevo. The capacity of quantum channel with general signal states. *IEEE Trans. Inf. Theory*, 44, 1998. 1

[30] A. D. Wyner. The common information of two dependent random variables. *IEEE Trans. Inf. Theory*, 21, 1975. 2, 4, 9

[31] G. R. Kumar, C. T. Li, and A. El Gamal. Exact common information. In *2014 IEEE ISIT*, page 161. IEEE, 2014. 2, 4

[32] A. B. Boyd, D. Mandal, and J. P. Crutchfield. Thermodynamics of modularity: Structural costs beyond the Landauer bound. *Phys. Rev. X*, 8(3):031036, 2018. 2

[33] R. A. Ahlswede and A. Winter. Strong converse for identification via quantum channels. *IEEE Trans. Inf. Theory*, 48:346, 2002. Addendum: IEEE Trans. Inf. Theory, vol. 49, no. 1, p. 346, 2003. 4

[34] A. Winter. Secret, public and quantum correlation cost of triples of random variables. In *2005 IEEE ISIT*. IEEE, 2005. 4

[35] R. B. Ash. *Information Theory*. Dover, New York, 1990. 5

[36] A. Wehrl. General properties of entropy. *Rev. Mod. Physics*, 50(2):221, 1978. 6

[37] L. P. Hughston, R. Jozsa, and W. K. Wootters. A complete classification of quantum ensembles having a given density matrix. *Phys. Lett. A*, 183(1):14–18, 1993. 6, 8

[38] A. W. Marshall, I. Olkin, and B. C. Arnold. *Inequalities: Theory of Majorization and Its Applications*. Springer, New York, NY, 3 edition, 2011. 8

[39] G. Ghirardi A. Bassi. A general scheme for ensemble purification. *Phys. Lett. A*, 309:24, 2003. 10