JOURNAL OF MATHEMATICAL PHYSICS VOLUME 43, NUMBER 9 SEPTEMBER 2002

Trading quantum for classical resources in quantum data compression

Patrick Hayden[a)]
Institute for Quantum Information, Caltech, Pasadena, California 91125

Richard Jozsa[b)] and Andreas Winter[c)]
Department of Computer Science, University of Bristol, Merchant Venturers Building, Bristol BS8 1UB, United Kingdom

(Received 12 April 2002; accepted for publication 21 May 2002)

We study the visible compression of a source $\mathcal{E}=\{|\varphi_i\rangle,p_i\}$ of pure quantum signal states or, more formally, the minimal resources per signal required to represent arbitrarily long strings of signals with arbitrarily high fidelity, when the compressor is given the identity of the input state sequence as classical information. According to the quantum source coding theorem, the optimal quantum rate is the von Neumann entropy $S(\mathcal{E})$ qubits per signal. We develop a refinement of this theorem in order to analyze the situation in which the states are coded into classical and quantum bits that are quantified separately. This leads to a trade-off curve $Q^*(R)$, where $Q^*(R)$ qubits per signal is the optimal quantum rate for a given classical rate of R bits per signal. Our main result is an explicit characterization of this trade-off function by a simple formula in terms of only single-signal, perfect fidelity encodings of the source. We give a thorough discussion of many further mathematical properties of our formula, including an analysis of its behavior for group covariant sources and a generalization to sources with continuously parametrized states. We also show that our result leads to a number of corollaries characterizing the trade-off between information gain and state disturbance for quantum sources. In addition, we indicate how our techniques also provide a solution to the so-called remote state preparation problem. Finally, we develop a probability-free version of our main result which may be interpreted as an answer to the question: ''How many classical bits does a qubit cost?'' This theorem provides a type of dual to Holevo's theorem, insofar as the latter characterizes the cost of coding classical bits into qubits. © *2002 American Institute of Physics.* [DOI: 10.1063/1.1497184]

TABLE OF CONTENTS

[a)] Electronic mail: patrick@cs.caltech.edu
[b)] Electronic mail: richard@cs.bris.ac.uk
[c)] Electronic mail: winter@cs.bris.ac.uk

0022-2488/2002/43(9)/4404/41/$19.00

I. INTRODUCTION

When the term ‘‘quantum information’’ was first coined, it would have been hard to predict how thorough and fruitful the analogy between quantum mechanics and classical information theory would ultimately prove to be. The general approach, characterized by the treatment of quantum states as resources to be manipulated, has yielded a promising collection of applications, ranging from unconditionally secure cryptographic protocols[1–3] to quantum algorithms.[4–6] Moreover, the analogy, which was initially unavoidably vague, has gradually been filled in by a diverse variety of rigorous theorems describing achievable limits to the manipulation of quantum states, such as the characterization of the classical information capacity of quantum sources,[7,8] the identification of optimal strategies for entanglement concentration and dilution[9] and many more. One of the pivotal results of the emerging theory is the quantum source coding theorem,[10–12] demonstrating that for the task of compressing quantum states, the von Neumann entropy plays a role directly analogous to the Shannon entropy of classical information theory. Indeed, the quantum theorem subsumes the classical one as the special case in which all the quantum states to be compressed are mutually orthogonal.

A quantum source (or ensemble) $\mathcal{E}=\{|\varphi_i\rangle,p_i\}$ is defined by a set of pure quantum signal (or ‘‘letter’’) states $|\varphi_i\rangle$ with given prior probabilities p_i (cf. below for precise definitions of these and other terms used in the Introduction). In this article we will study the so-called *visible* compression of $\mathcal{E}$. More specifically, we wish to characterize the minimal resources per signal that are necessary and sufficient to represent arbitrarily long strings of signals with arbitrarily high fidelity, when the compressor is given the identity of the input state sequence as *classical* information (as the sequence of labels $i_1,\ldots,i_n$ rather than the quantum states $|\varphi_{i_1}\rangle,\ldots,|\varphi_{i_n}\rangle$ themselves, for example). According to the quantum source coding theorem the optimal *quantum* rate in this scenario is the von Neumann entropy $S(\mathcal{E})$ qubits per signal. We will develop a refinement of this theorem in which the states are coded into classical and quantum bits which are quantified *separately*. This leads to a trade-off curve $Q^*(R)$ where $Q^*(R)$ qubits per signal is the optimal quantum rate that suffices for a given classical rate R bits per signal. The quantum source coding theorem implies that $Q^*(0)=S(\mathcal{E})$ and evidently we also have $Q^*(H(p))=0$ where $H(p)$ is the Shannon entropy of the prior distribution of the source. [By standard classical compression, the compressor can represent the full information of the input sequence in $H(p)$ classical bits per signal.] Thus the trade-off curve extends between the limits $0\leqslant R\leqslant H(p)$.

There are various reasons why we might wish to maintain a separation between classical and quantum resources in an encoding.[13] On a purely practical level it seems to be far easier to manufacture classical storage and communication devices than it is to make quantum ones. But perhaps the primary reason is conceptual: classical and quantum information have quite different fundamental characters, with classical information exhibiting special properties not shared by quantum information in general. For example, classical information is robust compared to quantum information—it may be readily stabilized and corrected by repeated measurement that would destroy quantum information. Also, unlike quantum information, it may be cloned or copied. These and other singular properties indicate that for many purposes it may be useful to regard classical information as a separate resource, distinct from quantum information. Classical information is sometimes formally regarded as a special case of quantum information *viz.* the quantum information of a fixed set of orthogonal states. While this characterization is useful for formal analyses, it is unsatisfactory conceptually because it relies on the essentially nonphysical infinite precision of orthogonality. It is, therefore, desirable to view classical information as a separate resource.

Exploring the trade-off possibilities between the two resources will lead to a better understanding of the interrelation of these concepts and the nature of quantum information itself. If bits can always be represented as qubits (and indeed, by Holevo's information bound,[14] one qubit per bit is necessary and sufficient), what are the limitations on representing qubits as bits? Under what conditions is it possible at all? If there is a penalty to be paid, how large is it? In this article we will give answers to these questions.

Our main result is a simple characterization of the trade-off function $Q^*(R)$ which may be paraphrased as follows. Given the ensemble $\mathcal{E}=\{|\varphi_i\rangle,p_i\}$ comprising m states $|\varphi_i\rangle$ we consider decompositions of $\mathcal{E}$ into at most $(m+1)$ ensembles $\mathcal{E}_j$ with associated probabilities q_j, i.e., the ensembles $\mathcal{E}_j=\{|\varphi_i\rangle,q(i|j)\}$ have the same states as $\mathcal{E}$ and their union $\cup_j q_j\mathcal{E}_j$ reproduces $\mathcal{E}$. This is equivalent to the condition

$$p_i=\sum_j q(i|j)q_j \tag{1}$$

on the chosen probabilities q_j and $q(i|j)$ defining the decomposition. Let $\overline{S}=\Sigma_j q_j S(\mathcal{E}_j)$ be the average von Neumann entropy of any such decomposition and let $H(i:j)$ be the classical mutual information of the joint distribution $q(i,j)$. For any R let $\overline{S}_{\min}(R)$ be the least average von Neumann entropy over all decompositions that have $H(i:j)=R$. Then we will prove that the trade-off function is given by $Q^*(R)=\overline{S}_{\min}(R)$.

The prescription of a decomposition $\mathcal{E}=\cup_j q_j\mathcal{E}_j$ may be equivalently given in terms of a visible encoding map E of the states of $\mathcal{E}$:

$$E(i)=|\varphi_i\rangle\langle\varphi_i|\otimes\sum_j p(j|i)|j\rangle\langle j|. \tag{2}$$

Here $p(j|i)$ are chosen freely subject only to the condition that $H(i:j)=R$ and the previous probability distributions are constructed as $q_j=\Sigma_i p(j|i)p_i$ and $q(i|j)=p(j|i)p_i/q_j$. Under this map, i is encoded into a quantum register, simply containing the state $|\varphi_i\rangle$ itself, and a classical register, containing a classical mixture of j values. Note that this is a *single* signal encoding with *perfect* fidelity since the state $|\varphi_i\rangle$ may be regained perfectly from the encoded version by simply discarding the classical register. Hence our result characterizes optimal classical and quantum resources in compression, in terms of very simple single-signal perfect-fidelity encodings, despite the fact that compression is defined asymptotically in terms of arbitrarily long signal strings and fidelities merely *tending* to 1. This is a remarkable and unexpected simplification—even in classical information theory it is by no means the rule that coding problems have solutions that do *not* involve asymptotics (despite a few well-known examples such as Shannon's source and channel coding theorems[15]). The situation is even more tenuous in quantum information theory, which seems to be plagued by further nonadditivity (or unresolved additivity questions) for some of its basic quantities so that, at the present stage, many basic constructions require a limit over optimization problems of exponentially growing size.

Using our formula we will give a thorough discussion of further properties of the trade-off curve including a generalization to group covariant sources and to sources with infinitely many (continuously parametrized) states. We show that our result also leads to a number of corollaries characterizing the trade-off between information gain and state disturbance for quantum sources (yielding the results of Ref. 13 on blind compression as a corollary), and we indicate how our techniques for characterizing $Q^*(R)$ provide a solution to the so-called remote state preparation problem as well. Finally, we develop a probability-free version of our main result which may be interpreted as an answer to the intuitive question: ''How many classical bits does a qubit cost?'' This may also be interpreted as a kind of dual to Holevo's theorem, insofar as the latter characterizes the qubit cost of coding classical information into qubits.

The presentation of these results is organized as follows. At the top level, the article is divided broadly into two parts. The first part, Secs. II–VIII, sets up a precise formulation of the basic definitions and the trade-off problem and gives the proof of the main theorem characterizing $Q^*(R)$, as well as a discussion of some of its important basic properties. The second part, Secs. IX and X, then goes on to provide some further generalizations of the main result. In more detail, the contents of the various sections are as follows.

In Sec. II, we will define the notions of blind and visible compression, the essential difference being that in the blind setting the encoder is given the actual quantum states, while in the visible setting the encoder is given the names of the quantum states as classical data. We then extend these definitions to quantum-classical trade-off coding and introduce the trade-off function $Q^*(R)$.

In Sec. III we will prove a lower bound to the trade-off curve in terms of the simple single-letter formula of the ensemble decomposition construction paraphrased above. In Sec. IV we will, in turn, show that the lower bound is achievable so that the trade-off curve is identical to the single-letter formula. This is our main result, Theorem 4.4.

In Sec. V we use our characterization of the trade-off curve to evaluate $Q^*(R)$ numerically for a selection of particular ensembles, chosen to illustrate various important properties of the trade-off function. In Sec. VI we extend our results to a different asymptotic setting, known as the arbitrarily varying source (AVS), in which there is no (or only limited) knowledge of the prior probability distribution of the states to be compressed. This provides a probability-free generalization of our main result. In Sec. VII we show that our main result can be reinterpreted to provide statements about the trade-off between information gain and state disturbance for blind sources of quantum states (in particular entailing a new proof of the main result of Ref. 13). Finally, in Sec. VIII we indicate how our techniques—developed to study $Q^*(R)$–can also be used to characterize the trade-off curve for the coding problem of remote state preparation posed in Refs. 16 and 17.

Sections IX and X treats two significant further issues. In Sec. IX we show how to apply our results in the setting of group covariant ensembles, which leads to considerable further elegant simplifications. Section X is devoted to the technicalities of generalizing our main result to sources with infinitely many (continuously parametrized) states. Finally, in the Appendix, we collect proofs of various auxiliary propositions that have been quoted in the body of the article.

II. BLIND AND VISIBLE COMPRESSION

We begin by introducing a number of definitions that are required to give a precise statement of the variations of quantum source coding that we will be considering in this article. We will denote an ensemble of quantum states φ_i with prior probabilities p_i as $\mathcal{E}=\{\varphi_i,p_i\}$. In turn, we will write $S(\mathcal{E})=S(\Sigma_i p_i \varphi_i)$ for the von Neumann entropy of the average state of the ensemble: $S(\rho)=-\mathrm{Tr}\rho\log\rho$. (Throughout this article log and exp will denote the logarithm and exponential functions *to base* 2.) Starting from an ensemble $\mathcal{E}$, we can consider the quantum source producing quantum states that are sequentially drawn independently from $\mathcal{E}$. Such a source corresponds to a sequence of ensembles $\mathcal{E}^{\otimes n}=\{\varphi_I,p_I\}$, where

$$I:=i_1\cdots i_n, \tag{3}$$

$$\varphi_I := \varphi_{i_1} \otimes \cdots \otimes \varphi_{i_n}, \tag{4}$$

$$p_I := p_{i_1} \cdots p_{i_n}. \tag{5}$$

This sequence will be referred to as an independent identically distributed (i.i.d.) source and the states of $\mathcal{E}^{\otimes n}$ are called blocks of length n from $\mathcal{E}$. In this article we will focus on sources of pure quantum states $|\varphi_i\rangle$, often making use of the notation $\varphi_i = |\varphi_i\rangle\langle\varphi_i|$. The measure that we will use to determine whether two quantum states are close is the fidelity F. For two mixed states ρ and ω, F is given by the formula

$$F(\rho,\omega) := (\mathrm{Tr}\sqrt{\omega^{1/2}\rho\,\omega^{1/2}})^2. \tag{6}$$

(Note that some authors use the name ''fidelity'' to refer to the square-root of this quantity.) If $\omega = |\omega\rangle\langle\omega|$ is a pure state, then the fidelity has a particularly simple form:

$$F(\rho,\omega) = \langle\omega|\rho|\omega\rangle = \mathrm{Tr}(\rho\omega). \tag{7}$$

Finally, we will use the notation $\mathcal{H}_d$ to denote the Hilbert space of dimension d and $\mathcal{B}_d$ to denote the set of all mixed states on $\mathcal{H}_d$. Likewise, $\mathcal{H}_d^{\otimes n}$ will refer to the n-fold tensor product of $\mathcal{H}_d$ and, in a slight abuse of notation, $\mathcal{B}_d^{\otimes n}$ will refer to the set of density operators on $\mathcal{H}_d^{\otimes n}$. We are now ready to introduce the definition of *blind* quantum compression.

Definition 2.1: A blind coding scheme *for blocks of length n, to R qubits per signal and fidelity* $1-\epsilon$, *comprises the following ingredients:*

(1) a completely positive, trace-preserving (CPTP) encoding map $E_n : \mathcal{B}_d^{\otimes n} \to \mathcal{B}_2^{\otimes nR}$, and

(2) a CPTP decoding map $D_n : \mathcal{B}_2^{\otimes nR} \to \mathcal{B}_d^{\otimes n}$,

such that average fidelity

$$\sum_I p_I \langle\varphi_I| D_n(E_n(\varphi_I)) |\varphi_I\rangle \geq 1-\epsilon. \tag{8}$$

We say that an i.i.d. source $\mathcal{E}$ *can be blindly compressed to R qubits per signal if for all* $\delta,\epsilon > 0$ *and sufficiently large n there exists a blind coding scheme to* $R+\delta$ *qubits per signal with fidelity at least* $1-\epsilon$.

The definition of visible compression is the same except that the (CPTP) restrictions on the encoding map E_n are relaxed; for visible compression E_n can be an arbitrary association of input states to output states. Equivalently, E_n is a mapping from the *names* of the input states to output states. Thus, we write $E_n(I) \in \mathcal{B}_2^{\otimes nR}$. Note that blind and visible compression schemes differ only in the set of encoding maps that are permitted. For blind (respectively visible) compression, the input states are given as quantum (respectively classical) information. In both cases the decoding must be CPTP. In this language, the central result on the compression of quantum information can be expressed as follows.

Theorem 2.2 (Quantum source coding theorem[10–12]): *A source* $\mathcal{E}$ *of pure quantum states can be compressed to* α *qubits per signal if and only if* $\alpha \geq S(\mathcal{E})$. *The result holds for both blind and visible compression.*

It is interesting to study a refinement of quantum source coding in which the states are coded into classical and quantum resources which are quantified separately. Because of restrictions on the manipulation of quantum states such as the no-cloning theorem,[18] blind compression is typically weaker than visible. In Refs. 13 and 19, for example, it was shown that in blind compression it is typically impossible to make use of classical storage. The same is not true in the visible setting, where it is possible to trade classical storage for quantum. In this article we study this trade-off for *visible* compression but, before we begin, we need to recall some basic definitions introduced in Ref. 13.

Consider an encoding operation E_n which maps a signal state $|\varphi_I\rangle$ into a joint state on a quantum register B and a classical register C. If $\{|j\rangle\}$ is the classical orthonormal basis of C, then the most general classical state on C is a probability distribution over j values, implying that the most general form of the encoded state can be written as

$$E_n(I)=\sum_j p(j|I)\,\omega_{I,j}^B\otimes|j\rangle\langle j|^C. \tag{9}$$

The quantum and classical storage requirements (i.e., resources) of the encoding map are simply the sizes of the registers B and C, respectively.

Definition 2.3: The quantum rate *of the encoding map E_n is defined to be*

$$\mathrm{qsupp}(E_n,\mathcal{E}^{\otimes n})=\frac{1}{n}\log\dim\mathcal{H}_B,$$

while the classical rate *of the encoding is defined to be*

$$\mathrm{csupp}(E_n,\mathcal{E}^{\otimes n})=\frac{1}{n}\log\dim\mathcal{H}_C.$$

With these definitions in place, we can make precise the notion of compression with a quantum and a classical part.

Definition 2.4: A source $\mathcal{E}$ can be compressed to R classical bits per signal plus Q qubits per signal if for all $\epsilon,\delta>0$ there exists an $N>0$ such that for all $n>N$ there exists an encoding-decoding scheme (E_n,D_n) with fidelity $1-\epsilon$ satisfying the inequalities

$$\mathrm{csupp}(E_n,\mathcal{E}^{\otimes n})\leqslant R+\delta, \tag{10}$$

$$\mathrm{qsupp}(E_n,\mathcal{E}^{\otimes n})\leqslant Q+\delta. \tag{11}$$

The main result of this article will be a complete characterization of the curve describing the trade-off between R and Q. As mentioned above, for blind encodings there is usually no trade-off to be made: generically, $Q\geqslant S(\mathcal{E})$, regardless of the size of R. The reason is essentially that making effective use of the classical register amounts to extracting classical information from a quantum system in a reversible fashion, which is impossible unless the quantum states of interest obey some orthogonality condition. The more interesting case, therefore, is to study the structure of the trade-off curve for visible encodings. As it turns out, our technique will yield the older results for blind compression as a corollary.

Definition 2.5: For a given source $\mathcal{E}=\{|\varphi_i\rangle,p_i\}$, define the function $Q^*(R)$ to be the infimum over all values of Q for which the source can be visibly compressed to R classical bits per signal and Q quantum bits per signal.

Some properties of the curve $Q^*(R)$ are immediate. For example, the endpoints of the curve are easily found. If $R=0$, then the compression must be fully quantum mechanical and the quantum source coding Theorem 2.2 applies: $Q^*(0)=S(\mathcal{E})$. More generally, the theorem implies that $Q^*(R)+R\geqslant S(\mathcal{E})$ for all R. Similarly, for $R=H(p)$ we have $Q^*(R)=0$, by Shannon's classical source coding theorem. Moreover, for intermediate values of R, the curve is necessarily convex because one method of compressing with classical rate $\lambda_1R_1+\lambda_2R_2$ is simply to timeshare between the optimal protocols for R_1 and R_2 individually, resulting in quantum rate of $\lambda_1Q^*(R_1)+\lambda_2Q^*(R_2)$.

Example (Parametrized BB84 ensemble): Let us consider in more detail the example of a parametrized version of the BB84 ensemble in order to see what sorts of protocols are possible beyond simple timesharing. For $0<\theta\leqslant\pi/4$, let $\mathcal{E}_{BB}(\theta)$ be the ensemble consisting of the states

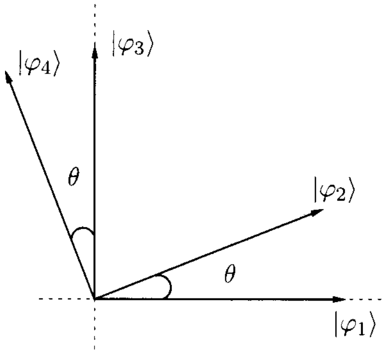

FIG. 1. Parametrized BB84 ensemble $\mathcal{E}_{BB}(\theta)$.

$$|\varphi_1\rangle = |0\rangle, \tag{12}$$

$$|\varphi_2\rangle = \cos\theta|0\rangle + \sin\theta|1\rangle, \tag{13}$$

$$|\varphi_3\rangle = |1\rangle, \tag{14}$$

$$|\varphi_4\rangle = -\sin\theta|0\rangle + \cos\theta|1\rangle, \tag{15}$$

as illustrated in Fig. 1, each occurring with probability $p_i = 1/4$. We then have $S(\mathcal{E}) = 1$ and $H(p) = 2$. From the argument above, we therefore already know two points on the $(R, Q^*(R))$ curve, namely (0,1) and (2,0). To get a better upper bound than the straight line joining these two points, suppose we were to partition the four states into two subsets, $\mathcal{X}_1 = \{|\varphi_1\rangle, |\varphi_2\rangle\}$ and $\mathcal{X}_2 = \{|\varphi_3\rangle, |\varphi_4\rangle\}$. For a given input string $I = i_1 i_2 \cdots i_n$, the classical register could be used to encode, for each k, whether $|\varphi_{i_k}\rangle \in \mathcal{X}_1$ or $|\varphi_{i_k}\rangle \in \mathcal{X}_2$. The classical rate required to do so would be 1 classical bit per signal. Independent of the value of the classical register, the quantum resource required to compress the subensembles is then just the quantum resource required to compress a pair of equiprobable quantum states subtended by the angle θ. Therefore,

$$Q^*(1) \leq S(\tfrac{1}{2}|\varphi_1\rangle\langle\varphi_1| + \tfrac{1}{2}|\varphi_2\rangle\langle\varphi_2|) = H_2(\tfrac{1}{2}(1+\cos\theta)). \tag{16}$$

By timesharing between the point corresponding to this protocol and the two endpoints of the curve that we already calculated, we get a piecewise linear upper bound on Q^*. As we will see later, however, the true curve is strictly below this upper bound. (The impatient reader is allowed to peek at Fig. 5 in Sec. V.)

With this example in mind, let us move on to our analysis of the general case.

III. SINGLE-LETTER LOWER BOUND ON $Q^*(R)$

In this section we will prove a lower bound on the quantum-classical trade-off curve by reducing the asymptotic problem to a single-copy problem. Because compression is only possible asymptotically, however, we need to shift the emphasis away from the quantum and classical resources towards quantum and classical mutual information quantities. In the next section we will then prove that nothing was lost by making this shift—we will show that the resulting lower bound to $Q^*(R)$ is actually achievable.

A. Mutual information and additivity

The information quantities in question will be the mutual information between the name of the state being compressed and the quantum and classical registers containing the output of the encoding map E_n. Thus, we define the state

$$\rho^{ABC}:=\sum_{I,j} p_I|I\rangle\langle I|^A\otimes p(j|I)\,\omega^B_{I,j}\otimes|j\rangle\langle j|^C. \tag{17}$$

The names I are stored in orthogonal states on system A while the quantum and classical encoding registers are labeled B and C, respectively. We can then make the following definitions:

$$S(A{:}C):=S(A)+S(C)-S(AC), \tag{18}$$

$$S(A{:}B|C):=S(AC)+S(BC)-S(ABC)-S(C), \tag{19}$$

where, for any subsystem X, $S(X)$ denotes the von Neumann entropy of the reduced state of X. Note that $S(A{:}C)$ is just the classical mutual information $H(I{:}j)$ between I and j. To interpret $S(A{:}B|C)$, observe that for a given classical output j, we can write down a conditional ensemble

$$\mathcal{E}_j=\{\omega_{I,j},q(I|j)\}, \tag{20}$$

where $q(I|j)$ is calculated using Bayes' rule to be $q(I|j)=p(j|I)p_I/q_j$, with $q_j=\Sigma_I p(j|I)p_I$. The conditional quantum mutual information $S(A{:}B|C)$ is just the average Holevo information χ of the conditional ensembles $\mathcal{E}_j$:

$$S(A{:}B|C)=\sum_j q_j\chi(\mathcal{E}_j), \tag{21}$$

where χ is defined, for an ensemble $\mathcal{E}=\{\rho_k,p_k\}$, as[14]

$$\chi(\mathcal{E}):=S\left(\sum_k p_k\rho_k\right)-\sum_k p_kS(\rho_k). \tag{22}$$

Because $\mathcal{E}_j$ is an ensemble supported on system B, $\chi(\mathcal{E}_j)\leqslant n\text{qsupp}$, which implies that

$$n\text{qsupp}\geqslant S(A{:}B|C). \tag{23}$$

Therefore, roughly speaking, we will derive a lower bound on $Q^*(R)$ by minimizing $S(A{:}B|C)$ subject to the constraint $S(A{:}C)\leqslant nR$ and developing further properties of that minimum. To that end, define $T_\epsilon(\mathcal{E}^{\otimes n},nR)$ to be the set of all encoding maps E for which $S(A{:}C)\leqslant nR$ and there exists a decoding map D satisfying

$$\sum_I p(I)F(\varphi_I,(D\circ E)\varphi_I)\geqslant 1-\epsilon. \tag{24}$$

Next define $M_\epsilon(\mathcal{E}^{\otimes n},nR)$ to be the infimum of $S(A{:}B|C)$ over all $E\in T_\epsilon(\mathcal{E}^{\otimes n},nR)$. We begin by noting the following basic properties of $M_\epsilon(\mathcal{E},R)$.

Lemma 3.1: $M_\epsilon(\mathcal{E},R)$ is a monotonically decreasing function of R. Moreoever, it is jointly convex in ϵ and R, in the sense that, for any set of $\epsilon_k>0$ and $R_k\geqslant 0$ as well as probabilities $\Sigma_k\lambda_k=1$,

$$M_\epsilon(\mathcal{E},R)\leqslant\sum_k \lambda_k M_{\epsilon_k}(\mathcal{E},R_k), \tag{25}$$

where $\epsilon=\Sigma_k\lambda_k\epsilon_k$ and $R=\Sigma_k\lambda_kR_k$.

Proof: Monotonicity follows immediately from the definitions. If $R_1\leqslant R_2$ and $S(A{:}C)\leqslant R_1$, then $S(A{:}C)\leqslant R_2$. Thus the set $T_\epsilon(\mathcal{E},R_1)$ is contained in $T_\epsilon(\mathcal{E},R_2)$ and $M_\epsilon(\mathcal{E},R_1)\geqslant M_\epsilon(\mathcal{E},R_2)$.

To prove joint convexity, let ϵ_k, R_k and λ_k be as in the statement of the lemma and assume that $E_k\in T_{\epsilon_k}(\mathcal{E},R_k)$. Furthermore, suppose that the encoding maps E_k map into orthogonal sectors

C_k of the classical register C. We construct an encoding map with information rate $R \leqslant \Sigma_k \lambda_k R_k$ and fidelity $\epsilon \leqslant \Sigma_k \lambda_k \epsilon_k$ by applying the map E_k with probability λ_k. The first inequality follows from the fact that the sectors C_k are orthogonal:

$$S(A:C) = \sum_k \lambda_k S(A:C_k) \leqslant R. \tag{26}$$

The decoding map for the new encoding consists of first determining which sector C_k was used and then applying the decoding map corresponding to E_k. The output of the encoding-decoding scheme will, therefore, be the average of the outputs of the individual schemes, yielding $1-\epsilon \geqslant \Sigma_k \lambda_k (1-\epsilon_k)$ by the concavity of the fidelity. Finally, if we define $S_k(A:B|C)$ to be the conditional quantum mutual information for the encoding map E_k, then we can calculate the value for the new scheme,

$$S(A:B|C) = \sum_k \lambda_k S_k(A:B|C). \tag{27}$$

Since $M_\epsilon(\mathcal{E},R) \leqslant S(A:B|C)$ by definition and this inequality must hold for all encoding maps E_k, we can conclude that $M_\epsilon(\mathcal{E},R) \leqslant \Sigma_k \lambda_k M_\epsilon(\mathcal{E},R_k)$. □

The particular usefulness of the M_ϵ function derives from an additivity property with respect to the input ensemble given in the next lemma, a property that can be converted into a single-letter lower bound on $Q^*(R)$.

Lemma 3.2: For any ensemble $\mathcal{E}$, numbers $R, \epsilon \geqslant 0$ and non-negative integer n,

$$M_\epsilon(\mathcal{E}^{\otimes n}, nR) \geqslant n M_\epsilon(\mathcal{E},R). \tag{28}$$

Proof: To begin, recall that $I = i_1 i_2 \cdots i_n$ and decompose A into $A_1 A_2 \cdots A_n$, with $|i_k\rangle$ stored on A_k. We will frequently make use of the notation $A_{<k} = A_1 A_2 \cdots A_{k-1}$ and the analogous $I_{<k} = i_1 i_2 \cdots i_{k-1}$, as well the similar $A_{>k}$ and $I_{>k}$. For a fixed $E \in T_\epsilon(\mathcal{E}^{\otimes n}, nR)$, the chain rule for mutual information (cf. Appendix C of Ref. 13) implies that

$$S(A:B|C) = \sum_{k=1}^{n} S(A_k:B|C, A_{<k}). \tag{29}$$

The bulk of the proof will consist of definitions for the purpose of interpreting the individual summands in the chain rule in terms of single-copy encoding maps. Consider one such term, $S(A_k:B|C,A_{<k})$, which we can express as

$$S(A_k:B|C,A_{<k}) = \sum_{I_{<k},j} p(I_{<k},j) \chi(\mathcal{E}_{I_{<k},j}), \tag{30}$$

where $\mathcal{E}_{I_{<k},j}$ is the ensemble of states

$$\mathcal{E}_{I_{<k},j} = \left\{ \sum_{I_{>k}} p(I_{>k}) \omega_{I,j}, q_{I_{<k}}(i_k|j) \right\}, \tag{31}$$

with

$$q_{I_{<k}}(i_k|j) = \frac{\Sigma_{I_{>k}} p(i_k) p(I_{>k}) p(j|I)}{\Sigma_{I_{\geqslant k}} p(I_{\geqslant k}) p(j|I)}. \tag{32}$$

Now define the encoding map $E_{I_{<k}}$ on the ensemble $\mathcal{E}$ to be

$$E_{I_{<k}}(i_k) := \sum_{I_{>k}} p(I_{>k}) E(I) = \sum_{I_{>k}} \sum_{j} p(I_{>k}) p(j|I) \omega_{I,j} \otimes |j\rangle\langle j|. \tag{33}$$

The output of $E_{I_{<k}}$ on the quantum register is described by the set of ensembles $\mathcal{E}_{I_{<k},j}$. Next, define the decoding map $D_k = \mathrm{Tr}_{\neq k} \circ D$ and the fidelity

$$F_{I_{<k}} := 1 - \epsilon_{I_{<k}} := \sum_{i_k} p(i_k) F(\rho_{i_k}, (D_k \circ E_{I_{<k}})(i_k)). \tag{34}$$

We can then calculate that

$$\begin{aligned}
\sum_{I_{<k}} p(I_{<k}) F_{I_{<k}} &= \sum_{I_{<k}} p(I_{<k}) \sum_{i_k} p(i_k) F(\rho_{i_k}, (D_k \circ E_{I_{<k}})(i_k)) \\
&= \sum_{I_{\leqslant k}} p(I_{\leqslant k}) F\left(\rho_{i_k}, \mathrm{Tr}_{\neq k} D\left(\sum_{I_{>k}} p(I_{>k}) E(I) \right) \right) \\
&= \sum_{I_{\leqslant k}} p(I_{\leqslant k}) F\left(\sum_{I_{>k}} p(I_{>k}) \rho_{i_k}, \sum_{I_{>k}} p(I_{>k}) (\mathrm{Tr}_{\neq k} \circ D \circ E)(I) \right) \\
&\geqslant \sum_{I} p(I) F(\mathrm{Tr}_{\neq k} \rho_I, (\mathrm{Tr}_{\neq k} \circ D \circ E)(I)) \\
&\geqslant \sum_{I} p(I) F(\rho_I, (D \circ E)(I)) \geqslant 1 - \epsilon.
\end{aligned} \tag{35}$$

The first three lines are by definition and using linearity to shuffle the terms. The first inequality comes from the joint concavity of the fidelity, the second from its monotonicity under partial trace, and the last from the fidelity condition on $D \circ E$.

Therefore, if we write $j(E_{I_{<k}})$ for the random variable representing the classical output of the encoding map $E_{I_{<k}}$ and $R_{I_{<k}}$ for the corresponding mutual information, then $E_{I_{<k}} \in T_{\epsilon_{I_{<k}}}(\mathcal{E}, R_{I_{<k}})$. Defining $R_k := \Sigma_{I_{<k}} p(I_{<k}) R_{I_{<k}}$ for the average classical information and applying the joint convexity of M then finally yields

$$S(A_k : B | C, A_{<k}) \geqslant M_{\epsilon}(\mathcal{E}, R_k). \tag{36}$$

A simple calculation allows us to bound the R_k from above; however,

$$\sum_k R_k = \sum_k \sum_{I_{<k}} p(I_{<k}) H(i_k : j(E_{I_{<k}})) \tag{37}$$

$$= \sum_k S(A_k : C | A_{<k}) \tag{38}$$

$$= S(A : C) \leqslant nR. \tag{39}$$

Combining Eqs. (36) and (39) with the chain rule, and applying the convexity of M one more time gives the simple inequality

$$S(A : B | C) \geqslant \sum_k M_{\epsilon}(\mathcal{E}, R_k) \geqslant n M_{\epsilon}(\mathcal{E}, R). \tag{40}$$

Since this lower bound must hold for all encoding maps in $T_\epsilon(\mathcal{E}^{\otimes n},R)$, that concludes the proof of the lemma. □

B. Perfect encodings and their properties

Within the set $T_0(\mathcal{E},R)$ of encoding maps with *perfect* fidelity decodings there is a particularly simple subset, in terms of which we will phrase our final bound on $Q^*(R)$. Let $T(\mathcal{E},R)\subset T_0(\mathcal{E},R)$ be the set of all encoding maps E of the form

$$E(i)=|\varphi_i\rangle\langle\varphi_i|^B\otimes\sum_j p(j|i)|j\rangle\langle j|^C. \tag{41}$$

In other words, $T(\mathcal{E},R)$ consists of the encoding maps in which a perfect copy of the state to be compressed is placed in register B. The decoding map is simply to trace over the register C. While such encodings, which simply reproduce the input, are obviously useless for compression, they turn out to be quite sufficient for minimizing $S(A:B|C)$. Indeed, let us define

$$M(\mathcal{E},R)=\inf\{S(A:B|C):E\in T(\mathcal{E},R)\} \tag{42}$$

$$=\inf_{p(\cdot|\cdot)}\{S(A:B|C):S(A:C)\leqslant R\}. \tag{43}$$

By construction, this optimization is no longer over general CPTP maps but only over different possible conditional probability distributions on register C.

Let us collect a few properties of M for later use: First of all, M inherits the convexity of M_ϵ in the variable R. Also, it is clearly nonincreasing, and $M(\mathcal{E},0)=S(\mathcal{E})$ is immediate from the definition. Furthermore, for any choice of $p(\cdot|\cdot)$, we have

$$S(A:C)+S(A:B|C)=S(A:BC)\geqslant S(A:B)=S(\mathcal{E}), \tag{44}$$

from which we conclude that $R+M(\mathcal{E},R)\geqslant S(\mathcal{E})$. This, together with the convexity, implies continuity in R, and the estimates

$$M(\mathcal{E},R)\geqslant M(\mathcal{E},R+\delta)\geqslant M(\mathcal{E},R)-\delta. \tag{45}$$

In what follows, it will also frequently be helpful to use the following fact:

Proposition 3.3:

$$M(\mathcal{E},R)=\inf_{p(\cdot|\cdot)}\{S(A:B|C):S(A:C)=R\}, \tag{46}$$

with an equality condition in the infimum [rather than the inequality of Eq. (43)].

The proof is given in the Appendix, Sec. 1.

In principle one might envisage a limit with larger and larger classical register C. This would constitute a serious obstacle to calculating $M(\mathcal{E},R)$ and carrying through our larger program of evaluating $Q^*(R)$. Fortunately, the next proposition ensures that the range of j's we need to consider in the definition of $M(\mathcal{E},R)$ is bounded universally. Since the mutual informations involved are continuous, the infimum in the definition of $M(\mathcal{E},R)$ can be replaced by a minimum.

Proposition 3.4: In the definition of $M(\mathcal{E},R)$ given in Eq. (43), it suffices to consider encodings of the form Eq. (41) with at most $(m+1)$ j values, where m is the number of states in $\mathcal{E}$.

The proof is given in the Appendix, Sec. 2.

C. Completing the lower bound

Returning to the main argument, we are now prepared to relate $M(\mathcal{E},R)$ to the trade-off curve:

Theorem 3.5: *If a source $\mathcal{E}$ can be visibly compressed to Q qubits per signal and R classical bits per signal, then $Q \geqslant M(\mathcal{E},R)$. Equivalently, $Q^*(R) \geqslant M(\mathcal{E},R)$.*

Proof: By the definition of compression and the previous lemma, we note that, for all $\epsilon,\delta > 0$, the inequality $Q^*(R) \geqslant M_\epsilon(\mathcal{E},R+\delta)$ must hold. We will give a proof that M_ϵ is continuous at $\epsilon=0$, from which the stronger lower bound in terms of $M(\mathcal{E},R)$ will follow.

So, fix ϵ,δ for now and suppose that $E \in T_\epsilon(\mathcal{E},R+\delta)$. Let D be the decoding map associated to E. As usual,

$$E(i) = \sum_j \omega_{i,j}^B \otimes p(j|i)|j\rangle\langle j|^C. \tag{47}$$

For a given j value, the decoding map will produce the ensemble of states $\{\sigma_{i,j}, p(i|j)\}$ where $\sigma_{i,j} = D(\omega_{i,j}^B \otimes |j\rangle\langle j|^B)$. Therefore, applying Markov's inequality (cf. Lemma 6.3 of Ref. 13) and the fidelity condition in the definition of $T_\epsilon(\mathcal{E},R)$, the probability weight of the j's with

$$\sum_i q(i|j) F(\varphi_i, \sigma_{i,j}) \geqslant 1 - \sqrt{\epsilon} \tag{48}$$

is at least $1-\sqrt{\epsilon}$. In other words, for these good j values, the output of the decoding map is close to $\mathcal{E}_j$. Therefore, for these same good j values, by the monotonicity and continuity of χ, we must have

$$\chi(\mathcal{E}_j) \geqslant S\left(\sum_i q(i|j) |\varphi_i\rangle\langle\varphi_i| \right) - f(\epsilon), \tag{49}$$

where we may choose $f(\epsilon) = 4(\sqrt[4]{\epsilon}\log d - \sqrt[4]{\epsilon}\log(2\sqrt[4]{\epsilon}))$ (as shown in Appendix A of Ref. 13). Consequently,

$$S(A:B|C) = \sum_j q_j \chi(\mathcal{E}_j) \geqslant \sum_j q_j S\left(\sum_i q(i|j)|\varphi_i\rangle\langle\varphi_i| \right) - f(\epsilon). \tag{50}$$

Since $f(\epsilon) \to 0$ as $\epsilon \to 0$ we conclude that $\lim_{\epsilon\downarrow 0} M_\epsilon(\mathcal{E},R+\delta) = M_0(\mathcal{E},R+\delta)$ and, moreover, in the limit $\epsilon\to 0$ it suffices to consider encoding maps of the type

$$E(i) = |\varphi_i\rangle\langle\varphi_i|^B \otimes \sum_j p(j|i)|j\rangle\langle j|^C. \tag{51}$$

Thus we obtain $Q^*(R) \geqslant M(\mathcal{E},R+\delta)$, for all $\delta > 0$, which, by Eq. (45) above yields our claim.□

Remark: The estimate $f(\epsilon)$ above may also be derived using Fannes' inequality,[20] which states that for density operators ρ and σ on a d-dimensional space,

$$\|\rho - \sigma\|_1 \leqslant \epsilon \Rightarrow |S(\rho) - S(\sigma)| \leqslant d\,\eta(\epsilon/d). \tag{52}$$

where

$$\eta(x) = \begin{cases} -x\log x & \text{for } x \leqslant \frac{1}{4}, \\ \frac{1}{2} & \text{for } x > \frac{1}{4}. \end{cases} \tag{53}$$

We will use this inequality again later. □

D. On alternative definitions

Inspecting the proofs of Lemma 3.2 and Theorem 3.5 reveals that we do not actually need the block-based fidelity condition

$$\langle F\rangle := \sum_I p_I F(\varphi_I,(D\circ E)(I)) \geqslant 1-\epsilon \tag{54}$$

of Eq. (8), but only the weaker mean letterwise fidelity

$$\langle \overline{F}\rangle := \sum_I p_I \overline{F}_I \geqslant 1-\epsilon, \tag{55}$$

where

$$\overline{F}_I := \frac{1}{n}\left[\sum_{k=1}^{n} F(\varphi_{i_k},(\mathrm{Tr}_{\neq k}\circ D\circ E)(I))\right]. \tag{56}$$

By the monotonicity of the fidelity under partial traces, the latter is directly implied by the former.

The lower bound Eq. (35) is then replaced by $1-\epsilon_k$, with $(1/n)\sum_k \epsilon_k = \epsilon$, and we conclude, instead of Eq. (36), that

$$S(A_k:B|C,A_{<k}) \geqslant M_{\epsilon_k}(\mathcal{E},R_k). \tag{57}$$

The remaining argument is only altered at Eq. (40):

$$S(A:B|C) \geqslant \sum_{k=1}^{n} M_{\epsilon_k}(\mathcal{E},R_k) \geqslant nM_\epsilon(\mathcal{E},R), \tag{58}$$

using joint convexity once more.

Hence, we could define the function $\overline{M}_\epsilon(\mathcal{E},R)$ in a fashion analogous to $M_\epsilon(\mathcal{E},R)$ but using the fidelity function $\overline{F}$ instead of F and Lemma 3.2 would continue to hold for the new function. In fact, $\overline{M}_\epsilon(\mathcal{E},R)$ will be strictly additive, in the sense that

$$\overline{M}_\epsilon(\mathcal{E}^{\otimes n},nR) = n\overline{M}_\epsilon(\mathcal{E},R), \tag{59}$$

because any single-letter encoding with fidelity $1-\epsilon$ repeated n times gives rise to an n-block coding with mean letterwise fidelity $1-\epsilon$.

We also note at this stage that we could have opted for a slightly more sophisticated definition of the quantum resource of the encoding. In particular, if we introduce $\mathrm{qsupp}_j = (1/n)\log \mathrm{Rank}\,\mathcal{E}_j$ as the minimal number of qubits per signal required to support the conditional ensemble $\mathcal{E}_j$, then we could have defined the quantum rate of the encoding map as

$$\overline{\mathrm{qsupp}} = \sum_j q_j \mathrm{qsupp}_j. \tag{60}$$

In this picture, the quantum resource would be the average over classical j values of the minimal number of qubits per signal required to support the quantum portion of the encoded state $E_n(I)$. Such a definition, by treating the classical and quantum storage requirements differently, allows the possibility of variable-length quantum encodings, where the length is a function of the classical message j. Such encodings could potentially be more powerful than the encodings with fixed-sized quantum supports used to define the original qsupp. However, because $\mathrm{qsupp}_j \geqslant \chi(\mathcal{E}_j)$, the analog of Eq. (23) continues to hold. (For a more detailed investigation of the properties of such variable-length quantum memories, see Ref. 21.) More precisely,

$$n\overline{\mathrm{qsupp}} \geqslant S(A:B|C). \tag{61}$$

Therefore, the lower bound of Theorem 3.5 on the trade-off curve $Q^*(R)$ would apply equally well if we had defined $Q^*(R)$ using $\overline{\mathrm{qsupp}}$ instead of qsupp.

Thus, while replacing either F by $\overline{F}$ or qsupp by $\overline{\mathrm{qsupp}}$ in the definition of compression could potentially have reduced the resource requirements, we find that our lower bounds would apply to the modified definitions. Since we will see later in the article that the lower bounds are achievable using the original, restrictive formulation of compression, we can conclude that no advantage can be gained by relaxing the definitions to use $\overline{F}$ and $\overline{\mathrm{qsupp}}$.

IV. ACHIEVING THE LOWER BOUND $M(\mathcal{E},R)$

Recall that the trade-off function $Q^*(R)$ gives the minimal quantum resource Q^* qubits per letter that is sufficient to encode arbitrarily long strings with arbitrarily high fidelity $1-\epsilon$ for any $\epsilon>0$, given a classical resource of R bits per letter. On the other hand, the lower bound $M(\mathcal{E},R)$ is defined as the minimal quantum resource for a particular kind of *single*-letter *perfect* fidelity (i.e., $\epsilon=0$) encoding given in Eq. (51), subject to the constraint that the classical *mutual information* $S(A:C)$ between i and j is R. Hence in the latter case, the classical resource will generally exceed R bits per letter. Thus by implementing the simple encodings of Eq. (51) we can attain $M(\mathcal{E},R)$ as the quantum resource but not generally with a classical resource bounded by R. We now argue that, nevertheless, the classical resource can be reduced to R while retaining the quantum resource at $M(\mathcal{E},R)$ i.e., that the lower bound $M(\mathcal{E},R)$, to $Q^*(R)$ is attainable, so we must then have $Q^*(R)=M(\mathcal{E},R)$.

Our strategy intuitively is the following. We think of the conditional distribution $p(j|i)$ with mutual information $S(A:C)$ in Eq. (51) as a noisy channel from i to j. Then the reverse Shannon theorem[22] states that this noisy channel can be simulated with a noiseless channel of capacity $S(A:C)$ if the receiver and sender have shared randomness, i.e., in the presence of shared randomness, the classical resource can be reduced to $R=S(A:C)$ bits per letter. Finally, we show that only $O(\log n)$ bits of shared randomness suffice to provide a high fidelity encoding-decoding scheme for blocks of length n. Hence this amount of shared randomness can be included in the classical resource of the encoding with asymptotically vanishing cost per letter.

To make the above intuitions mathematically rigorous, we begin by recalling some basic facts from the theory of typical sequences[23,24] and typical subspaces[12,25] in the following two subsections.

A. Typical sequences

For a sequence $I=i_1\cdots i_n\in\mathcal{I}^n$ define the *type* P_I of I as its empirical distribution of letters, i.e.,

$$P_I(i):=\frac{1}{n}N(i|I):=\frac{1}{n}|\{k|i_k=i\}|. \tag{62}$$

The number of types of sequences is polynomial in n: it is $\binom{n+|\mathcal{I}|-1}{|\mathcal{I}|-1}\leqslant(n+1)^{|\mathcal{I}|}$.

The *type class* $\mathcal{T}_P$ of P is the set of all sequences with type P:

$$\mathcal{T}_P:=\{I\in\mathcal{I}^n|P_I=P\}. \tag{63}$$

Consider now any probability distribution P on $\mathcal{I}$, and let $\delta>0$. Then the set of *typical sequences* (with respect to the distribution P and δ) is

$$\mathcal{T}_{P,\delta}:=\{I\in\mathcal{I}:\forall i|P_I(i)-P(i)|\leqslant\delta/\sqrt{n}\}. \tag{64}$$

Note that this set is a union of certain type classes.

The following are standard facts:[23,24]

$$P^{\otimes n}(\mathcal{T}_{P,\delta}) \geq 1 - \frac{1}{\delta^2}, \tag{65}$$

$$(n+1)^{-|\mathcal{I}|} \exp(n(H(P))) \leq |\mathcal{T}_P|, \tag{66}$$

$$\exp(n(H(P))) \geq |\mathcal{T}_P|, \tag{67}$$

$$(n+1)^{-|\mathcal{I}|} \exp(n(H(P) - |\mathcal{I}|\,\eta(\delta/\sqrt{n}))) \leq |\mathcal{T}_{P,\delta}|, \tag{68}$$

$$(n+1)^{|\mathcal{I}|} \exp(n(H(P) + |\mathcal{I}|\,\eta(\delta/\sqrt{n}))) \geq |\mathcal{T}_{P,\delta}|. \tag{69}$$

Note that the latter two follow from the former two by the following well–known explicit estimate on the difference of two entropies[23] [this being a classical case of the Fannes inequality, Eq. (52)]: if P and Q are probability distributions on a set of k elements, then

$$\|P-Q\|_1 \leq \epsilon \Rightarrow |H(P) - H(Q)| \leq k\,\eta\left(\frac{\epsilon}{k}\right), \tag{70}$$

where the function η is given in Eq. (53).

For sequences $I \in \mathcal{I}^n$, $J \in \mathcal{J}^n$, the *conditional type* $W_{J|I}$ of J (conditional on I) is defined as the stochastic matrix given by

$$\forall ij \quad P_I(i) W_{J|I}(j|i) = P_{IJ}(ij), \tag{71}$$

where P_{IJ} is the joint type of $IJ = (i_1 j_1, \ldots, i_n j_n)$. It is undetermined if $P_I(i) = 0$.

The *conditional type class* of W given I is defined as

$$\mathcal{T}_W(I) := \{J : W_{J|I} = W\} = \{J : \forall ij \quad P_{IJ}(ij) = P_I(i) W(j|i)\}. \tag{72}$$

Let W be now an arbitrary stochastic matrix and $\delta > 0$. The *set of conditionally typical sequences* of W given I is defined as

$$\mathcal{T}_{W,\delta}(I) := \{J : \forall ij |W_{J|I}(j|i) - W(j|i)| \leq \delta/\sqrt{N(i|I)}\}. \tag{73}$$

Again, there are a couple of standard facts:

$$W_I(\mathcal{T}_{W,\delta}(I)) \geq 1 - \frac{|\mathcal{I}|}{\delta^2}, \tag{74}$$

for the product distribution $W_I = W_{i_1} \otimes \cdots \otimes W_{i_n}$, and

$$(n+1)^{-|\mathcal{I}||\mathcal{J}|} \exp(nH(W|P_I)) \leq |\mathcal{T}_W(I)|, \tag{75}$$

$$\exp(nH(W|P_I)) \geq |\mathcal{T}_W(I)|, \tag{76}$$

$$(n+1)^{-|\mathcal{I}||\mathcal{J}|} \exp(n(H(W|P_I) - |\mathcal{I}||\mathcal{J}|\,\eta(\delta|\mathcal{I}|/\sqrt{n}))) \leq |\mathcal{T}_{W,\delta}(I)|, \tag{77}$$

$$(n+1)^{|\mathcal{I}||\mathcal{J}|} \exp(n(H(W|P_I) + |\mathcal{I}||\mathcal{J}|\,\eta(\delta|\mathcal{I}|/\sqrt{n}))) \geq |\mathcal{T}_{W,\delta}(I)|, \tag{78}$$

where $H(W|P_I)$ is just the conditional Shannon entropy $\Sigma_i P_I(i) H(W(\cdot|i))$.

B. Typical subspaces

The concepts in the previous subsection translate straightforwardly to their Hilbert space versions via the following recipe:

For a state ρ choose a diagonalization $\rho = \Sigma_{i \in \mathcal{I}} r_i |e_i\rangle\langle e_i|$, with eigenvectors $|e_i\rangle$ and eigenvalues r_i, which define a probability distribution on $\mathcal{I}$. Then we have a diagonalization of $\rho^{\otimes n}$:

$$\rho^{\otimes n} = \sum_{I \in \mathcal{I}} r_I |e_I\rangle\langle e_I|, \tag{79}$$

with

$$|e_I\rangle = |e_{i_1}\rangle \otimes \cdots \otimes |e_{i_n}\rangle, \tag{80}$$

$$r_I = r_{i_1} \cdots r_{i_n}. \tag{81}$$

Now for any subset $\mathcal{A} \subset \mathcal{I}^n$ we can define the subspace spanned by the vectors $\{|e_I\rangle : I \in \mathcal{A}\}$, which is most conveniently described by the subspace projector

$$\Pi_{\mathcal{A}} := \sum_{I \in \mathcal{A}} |e_I\rangle\langle e_I|. \tag{82}$$

In this way we can define, for any distribution P on $\mathcal{I}$,

$$\Pi_P := \sum_{i \in \mathcal{T}_P} |e_I\rangle\langle e_I|, \tag{83}$$

(note that this is not uniquely specified by the distribution P alone, but also requires specification of the basis $|e_i\rangle$), and

$$\Pi_{\rho,\delta} := \sum_{i \in \mathcal{T}_{r,\delta}} |e_I\rangle\langle e_I|. \tag{84}$$

Statements on the cardinality of sets translate into statements on the dimension of the corresponding subspaces (i.e., rank, or equivalently, trace, of the projectors).

Similarly, if we have states W_i with diagonalizations $W_i = \Sigma_j W(j|i) |e_{j|i}\rangle\langle e_{j|i}|$, we can define, for any subset $\mathcal{A} \subset \mathcal{J}^n$ and $I \in \mathcal{I}^n$,

$$\Pi_{\mathcal{A}}(I) := \sum_{J \in \mathcal{A}} |e_{J|I}\rangle\langle e_{J|I}|. \tag{85}$$

This leads to the concept of *conditional typical subspace projector*, for $\delta \geqslant 0$,

$$\Pi_{W,\delta}(I) := \sum_{J \in \mathcal{T}_{W,\delta}} |e_{J|I}\rangle\langle e_{J|I}|, \tag{86}$$

and again probability and cardinality statements about the typical sequences translate into equivalent statements about certain traces.

In particular we shall use the following estimate of the rank of the conditional typical subspace projector:

$$\mathrm{Tr}\Pi_{\rho,\delta}(I) \leqslant (n+1)^{|\mathcal{I}|d} \exp(n(S(\rho|P_I) + |\mathcal{I}| d\, \eta(\delta|\mathcal{I}|/\sqrt{n}))). \tag{87}$$

[Here we make use of the notation $S(\rho|P_I) := \Sigma_i S(W_i)$ in an attempt to match the statements about typical sequences as closely as possible.] We'll also use the important probability estimate

$$\mathrm{Tr}(W_I \mathcal{T}_{W,\delta}(I)) \geqslant 1 - \frac{|\mathcal{I}|}{\delta^2}. \tag{88}$$

C. Trade-off coding

We will use the coding technique that is summarized in the following proposition. The statement is slightly more technical and the estimates more explicit than we would need to prove our main Theorem 4.4. This is because we will reuse it in Secs. VI and X.

Proposition 4.1: For a probability distribution p *on* $\mathcal{I}$ *and a classical noisy channel* $p(\cdot|\cdot):\mathcal{I}\rightarrow\mathcal{J}$ *consider the tripartite state*

$$\rho=\sum_i p_i|i\rangle\langle i|^A\otimes|\varphi_i\rangle\langle\varphi_i|^B\otimes\sum_j p(j|i)|j\rangle\langle j|^C.$$

Then there exists a visible code (E,D) *such that*

$$\forall I\in\mathcal{T}_{p,\delta}\quad F(|\varphi_I\rangle\langle\varphi_I|,(D\circ E)(I))\geqslant 1-\frac{4|\mathcal{I}||\mathcal{J}|}{\delta^2},$$

and having classical and quantum resources

$$nS(A:C)+nK|\mathcal{I}||\mathcal{J}|\eta(\delta/\sqrt{n})+K'|\mathcal{I}||\mathcal{J}|\log(n+1)\quad \textit{classical bits},$$

$$nS(A:B|C)+n\cdot 3d|\mathcal{I}||\mathcal{J}|\eta(2\delta|\mathcal{I}||\mathcal{J}|/\sqrt{n})+d|\mathcal{J}|\log(n+1)\quad \textit{quantum bits},$$

where K *and* K' *are absolute constants.*

Proof: We design an n-block code as follows (typicality conditions throughout are with respect to a previously fixed δ):

(a) Encoding:

(1) Given I generate J according to $p(J|I)$.

(2) Compress (i.e., project) the quantum state $|\varphi_I\rangle\langle\varphi_I|$ to the conditional typical subspace $\Pi_{\tilde{\rho}^{IJ},\delta}(J)$, where $\tilde{\rho}_j^{IJ}=\Sigma_i W_{I|J}(i|j)|\varphi_i\rangle\langle\varphi_i|$.

If I is typical and J is conditionally typical, send J and the joint type of I and J as classical data, and send the projected state on $\Pi_{\tilde{\rho}^{IJ},\delta}(J)$ as quantum data.

(b) Decoding:

Given J, one can isometrically embed the quantum state transmitted back into the ambient Hilbert space.

The fidelity of this scheme is analyzed as follows. (We assume that if, at any point of the above protocol, an ''if'' is not satisfied, then some fixed failure action is taken. Such would be the case when the POVM involving the above subspace projection yields an orthogonal result, for example.) With probability at least $1-|\mathcal{I}|/\delta^2$, J is conditionally typical, and in this case the projection is successful with probability at least $1-|\mathcal{J}|/\delta^2$ [by virtue of Eq. (88)], leaving a state which (cf. Ref. 12) has fidelity $\geqslant 1-2|\mathcal{J}|/\delta^2$ to $|\varphi_I\rangle\langle\varphi_I|$.

Looking at the classical cost of this procedure, we see that it is dominated by sending J, which requires too many, namely $nS(C)$, classical bits. Here the reverse Shannon theorem[22] is invoked. (For a precise statement, see Theorem 4.2 below.) Using this theorem we can simulate the channel p on the typical sequences I sending $nS(A:C)+o(n)$ classical bits, but at the same time needing an amount of shared randomness. The simulation, in fact, has the property that it endows sender and receiver with a common J, the distribution of which is $|\mathcal{I}||\mathcal{J}|/\delta^2$-close to $p(J|I)$. Taking all these points into account, we see that the fidelity of this protocol is at least $1-3|\mathcal{I}||\mathcal{J}|/\delta^2$ for every individual $|\varphi_I\rangle\langle\varphi_I|$ for which I is typical.

The analysis of the quantum resources needed is equally straightforward. By Eq. (87) the number of qubits needed to transmit the projected state is

$$nS(\tilde{\rho}^{IJ}|P_J)+dn|\mathcal{J}|\eta(\delta/\sqrt{n})+d|\mathcal{J}|\log(n+1). \tag{89}$$

Note that the leading term is a conditional von Neumann entropy of the bipartite state

$$\rho = \sum_j \tilde{\rho}_j^{IJ} \otimes P_J(j) |j\rangle\langle j|, \tag{90}$$

which has trace norm distance at most $2\delta|\mathcal{I}||\mathcal{J}|/\sqrt{n}$ from

$$\omega = \sum_{ij} p(i) |\varphi_i\rangle\langle\varphi_i| \otimes p(j|i) |j\rangle\langle j|. \tag{91}$$

(This follows from the typicality of I and conditional typicality of J.) Next, using the Fannes inequality (52), we can upper bound Eq. (89) by

$$nS(\tilde{\rho}|q) + 2dn|\mathcal{J}|\,\eta(2\delta|\mathcal{I}||\mathcal{J}|/\sqrt{n}) + dn|\mathcal{J}|\,\eta(\delta/\sqrt{n}) + d|\mathcal{J}|\log(n+1), \tag{92}$$

with $q_j = \Sigma_i P(i) p(j|i)$ and $\tilde{\rho}_j = q_j^{-1} \Sigma_i P(i) p(j|i) |\varphi_i\rangle\langle\varphi_i|$.

We are left with one remaining feature to address: the protocol uses shared randomness (and to a considerable extent, according to Theorem 4.2). We shall now show that we can reduce this requirement to $O(\log n)$ shared random bits using a technique very much like the derandomization argument in Ref. 26. The proof will then be complete because setting up these bits can be absorbed into the classical communication with asymptotically vanishing cost per letter. (Actually, in order to achieve high average fidelity, no random bits are needed at all, but our goal is to prove that high fidelity can be achieved for every state in the typical subspace, a more stringent requirement that is used later in our study of arbitrarily varying sources.)

Observe that a protocol using shared randomness can be viewed as a probabilistic mixture of ordinary, deterministic protocols. Index these by a variable ν, accompanied by a probability x_ν. For each ν we have a corresponding fidelity $F_I(\nu)$ for each individual I. Our construction shows that for typical I,

$$\sum_\nu x_\nu F_I(\nu) \geqslant 1 - \frac{3|\mathcal{I}||\mathcal{J}|}{\delta^2} =: \mu. \tag{93}$$

Note that the left hand side is exactly the expectation of the random variable F_I. We now choose $\nu_1,\ldots,\nu_L$ independently and identically distributed (i.i.d.), according to the probabilities x_ν. For fixed I the $F_I(\nu_l)$, $l=1,\ldots,L$ are i.i.d. as well, and in the interval [0, 1]. Thus we can apply the Chernoff–Hoeffding bound for their arithmetic mean (Lemma 4.3 below):

$$\Pr\left\{\frac{1}{L}\sum_{l=1}^{L} F_I(\nu_l) < (1-\epsilon)\mu\right\} \leqslant \exp\left(-L\frac{\epsilon^2\mu}{2\ln 2}\right). \tag{94}$$

By the union bound we can estimate the probability that the above event occurs for a single typical I to be less than or equal to

$$\exp\left(-L\frac{\epsilon^2\mu}{2\ln 2}\right)|\mathcal{I}|^n. \tag{95}$$

Choosing $\epsilon = |\mathcal{I}||\mathcal{J}|/\delta^2$, this bound is itself less than 1 if

$$L > \frac{2\delta^4 \ln 2}{|\mathcal{I}|^2|\mathcal{J}|^2\mu} n \log|\mathcal{I}|, \tag{96}$$

in which case we can conclude that there exist values $\nu_1,\ldots,\nu_L$ such that, for all typical I, we have

$$\frac{1}{L}\sum_{l=1}^{L} F_I(\nu_l) \geqslant 1 - \frac{4|\mathcal{I}||\mathcal{J}|}{\delta^2}.$$

Therefore, a shared uniform distribution over the numbers $1,\ldots,L$ is sufficient, where L need only satisfy Eq. (96). This can be accomplished with $O(\log n)$ shared random bits, which is what we wanted. □

Here are the auxiliary results we needed in the proof:

Theorem 4.2 (Reverse Shannon Theorem; see Refs. 22 and 27): *For any channel* $W:\mathcal{I}\to\mathcal{J}$, *distribution* P *on* $\mathcal{I}$, *and* $0<\lambda<1$ *there exist maps*

$$E_\nu:\mathcal{I}^n\to\{1,\ldots,M\},$$

$$D_\nu:\{1,\ldots,M\}\to\mathcal{J}^n,$$

$\nu=1,\ldots,N$, *such that*

$$\forall I\in\mathcal{T}_{P,\delta}\quad \frac{1}{2}\left\| W(I)-\frac{1}{N}\sum_{\nu=1}^{N} D_\nu(E_\nu(I))\right\|_1 \leqslant \frac{|\mathcal{I}||\mathcal{J}|}{\delta^2}.$$

Moreover, with an absolute constant K,

$$\log M\leqslant nH(P:W)+nK|\mathcal{I}||\mathcal{J}|\eta(\delta/\sqrt{n})+K|\mathcal{I}||\mathcal{J}|\log(n+1),$$

$$\log N\leqslant nH(W|P)+nK|\mathcal{I}||\mathcal{J}|\eta(\delta/\sqrt{n})+K|\mathcal{I}||\mathcal{J}|\log(n+1).$$

□

Lemma 4.3 (Chernoff-Hoeffding bound.[28,29]) Let $X_1,\ldots,X_L$ *be independent, identically distri-uted random variables, taking real values in the interval* $[0, 1]$, *and with expectation* $\mathbb{E}X_l\geqslant\mu$. *Then, for* $\epsilon>0$,

$$\Pr\left\{\frac{1}{L}\sum_{l=1}^{L} X_l<(1-\epsilon)\mu\right\}\leqslant\exp\left(-L\frac{\epsilon^2\mu}{2\ln 2}\right).$$

□

With this we are ready to state our main result:

Theorem 4.4: $Q^*(R)=M(\mathcal{E},R)$.

Proof: The inequality ‘‘$\geqslant$’’ is theorem 3.5. For the opposite inequality choose a $p(\cdot|\cdot)$ such that $S(A:C)\leqslant R$ and $S(A:B|C)\leqslant M(\mathcal{E},R)+\epsilon$. Then, according to Proposition 4.1, there exist n-block codes (E,D) with classical and quantum rates bounded by $R+o(1)$ and $M(\mathcal{E},R)+\epsilon+o(1)$, respectively, which have fidelity $1-\epsilon$ for all *typical* I. But since these carry almost all the probability weight (say, larger than $1-\epsilon$) of all sequences, the fidelity of the scheme is at least $1-2\epsilon$, regardless of what is done on nontypical sequences. As ϵ was arbitrary, we get $Q^*(R)=M(\mathcal{E},R)$. □

Remark: The proof of Proposition 4.1, as the eventual ‘‘derandomization’’ shows, does not use the full power of the reverse Shannon theorem, but only a consequence that is actually also used in rate-distortion coding: that one can map the typical sequences I onto $\exp(nH(P:W)+o(n))$ many J’s such that all the pairs $(I,f(I))$ are jointly typical. □

V. EXPLORING THE TRADE-OFF CURVE

In this section we use our formula for the trade-off curve to evaluate $Q^*(R)$ numerically for a selection of particular ensembles chosen to illustrate further important properties of the trade-off function.

To begin, let us consider the simplest possibility, a pair of nonorthogonal states. Figure 2 plots the trade-off curve for the pair $\{|0\rangle,(1/\sqrt{2})(|0\rangle+|1\rangle)\}$, each occurring with probability $\frac{1}{2}$. At first glance, $Q^*(R)$ appears to coincide with the linear upper bound given by interpolating between $(0,S(\mathcal{E}))$ and $(H_2(\frac{1}{2}),0)$. A more detailed examination, however, reveals that the curve is actually

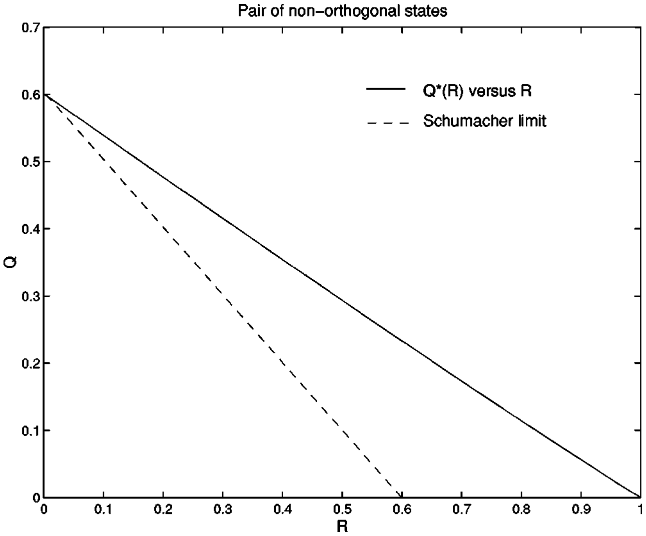

FIG. 2. The trade-off curve for a pair of equiprobable, nonorthogonal states. The dashed line represents the lower bound $Q^*(R)+R \geqslant S(\mathcal{E})$ imposed by the Schumacher limit.

very slightly nonlinear. Therefore, somewhat surprisingly, the simple quantum-classical coding scheme given by timesharing between fully quantum and fully classical coding is nearly optimal but not completely so. As we will see below, this need not always be true.

In general, more complicated ensembles with internal structure will have trade-off curves reflecting that structure. Consider, for example, the three-state ensemble $\mathcal{E}_3$ illustrated in Fig. 3, consisting of the states $|\varphi_1\rangle = |0\rangle$, $|\varphi_2\rangle = (1/\sqrt{2})(|0\rangle + |1\rangle)$ and $|\varphi_3\rangle = |2\rangle$ with equal probabilities. Since the set of states decomposes into two subsets $\mathcal{X}_1 = \{|\varphi_1\rangle, |\varphi_2\rangle\}$ and $\mathcal{X}_2 = \{|\varphi_3\rangle\}$ with mutually orthogonal supports, it is possible to encode whether a given $|\varphi_i\rangle \in \mathcal{X}_1$ or $|\varphi_i\rangle \in \mathcal{X}_2$ efficiently using $H_2(\frac{1}{3})$ classical bits. Indeed, Fig. 4 plots $Q^*(R)$ for this ensemble and we see that the Schumacher limit is achieved for values of $R \leqslant H_2(1/3)$. For values of $R > H_2(\frac{1}{3})$, or once the classical information in the ensemble has been exhausted, the trade-off curve departs from the Schumacher lower bound to meet the point $(H(\frac{1}{3},\frac{1}{3},\frac{1}{3}),0)$.

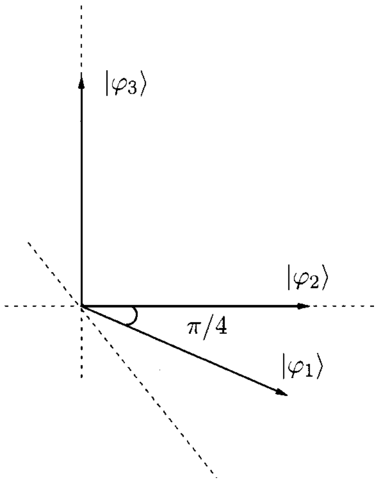

FIG. 3. The three-state ensemble $\mathcal{E}_3$ consists of the states $|\varphi_1\rangle$, $|\varphi_2\rangle$, $|\varphi_3\rangle$ occurring with equal probabilities.

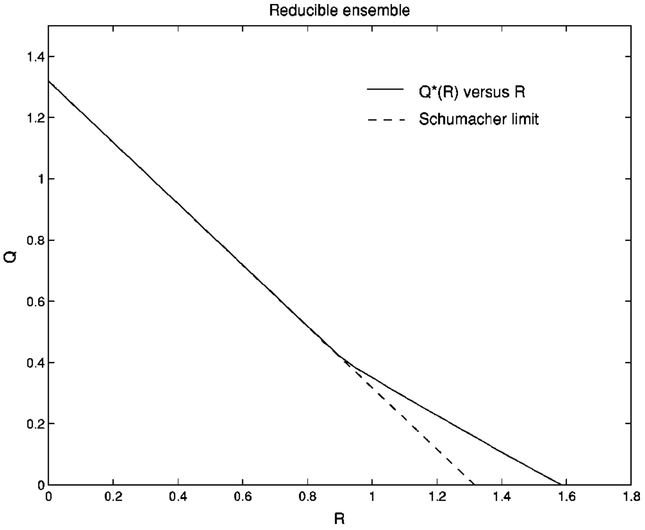

FIG. 4. The trade-off curve for three-state ensemble $\mathcal{E}_3$. The dashed line again represents the Schumacher lower bound, which in this case is achievable for $R \leqslant H(\frac{1}{3})$.

Our third example, the parametrized BB84 ensemble $\mathcal{E}_{BB}(\theta)$ introduced in Sec. II, is an ensemble that, like $\mathcal{E}_3$ above, decomposes naturally into subensembles. On the other hand, unlike for $\mathcal{E}_3$, the subensembles are generally not orthogonal. The trade-off curve for $\theta = \pi/8$ is plotted in Fig. 5. As usual, the dashed lower bound is the Schumacher limit. The dashed-dot line is the piecewise linear upper bound constructed in Sec. II. Squeezed into the intermediate region, we see that $Q^*(R)$ is typically strictly less than the upper bound and, especially in the region $0 < R < 1$, quite strongly curved. The point $(1, H_2(\frac{1}{2}(1+\cos \pi/8)))$ provides another surprise: $Q^*(R)$ and the upper bound coincide there. Therefore, the partitioning scheme is optimal if exactly one bit of classical storage is to be consumed per copy but not otherwise.

We now turn to another interesting property of the trade-off curve. Contrary to what one might expect, the function $M(\mathcal{E}, R)$ is *not concave in the ensemble*, violating the intuition that it should

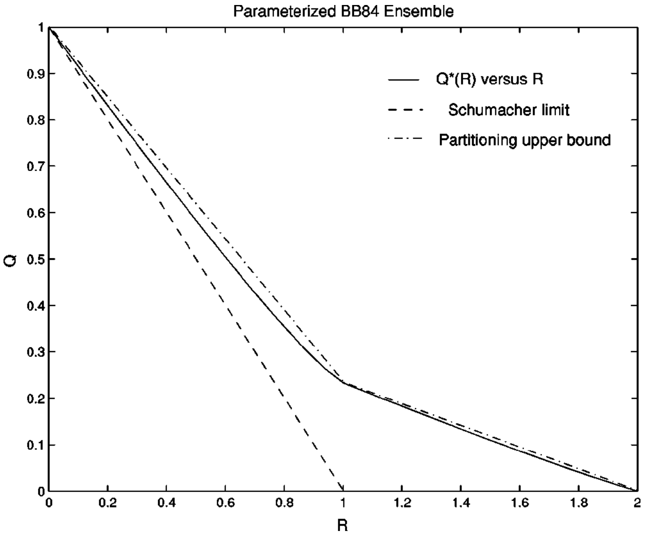

FIG. 5. Trade-off curve for the BB84 ensemble $\mathcal{E}_{BB}(\pi/8)$. The dashed line represents the Schumacher lower bound and the dashed-dot line represents the upper bound from partitioning into the sets $\mathcal{X}_1$ and $\mathcal{X}_2$.

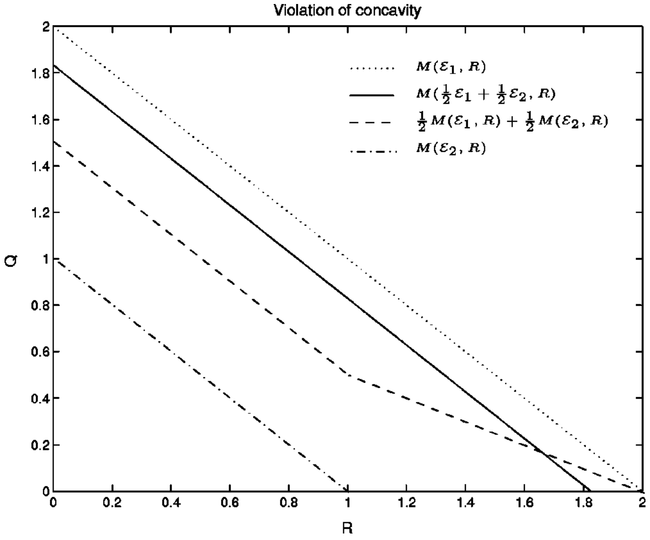

FIG. 6. Violation of concavity in the ensemble. If Q^* were concave in the ensemble, the solid line representing $M(\frac{1}{2}\mathcal{E}_1 + \frac{1}{2}\mathcal{E}_2, R)$ would always exceed the dashed line of $\frac{1}{2}M(\mathcal{E}_1, R)+\frac{1}{2}M(\mathcal{E}_2, R)$. For large values of R we see that is not the case in this example.

be harder to send the mixture of two ensembles than it is to probabilistically send either one. [Note that $M(\mathcal{E},0)$, however, is just the von Neumann entropy $S(\mathcal{E})$ and is, therefore, concave in $\mathcal{E}$.] In fact, counterexamples to concavity can be constructed without even making use of nonorthogonal states. Let $\mathcal{E}_1=\{|i\rangle,\frac{1}{4}\}_{i=0}^{3}$ be an ensemble consisting of four equiprobable orthonormal states and let $\mathcal{E}_2=\{|i\rangle,\frac{1}{2}\}_{i=0}^{1}$. We can also consider the mixture of ensembles

$$\mathcal{E} := \tfrac{1}{2}\mathcal{E}_1 + \tfrac{1}{2}\mathcal{E}_2 = \{(|0\rangle,\tfrac{3}{8}),(|1\rangle,\tfrac{3}{8}),(|2\rangle,\tfrac{1}{8}),(|3\rangle,\tfrac{1}{8})\}. \tag{97}$$

Since each of these ensembles is effectively classical, the Schumacher lower bound is attainable and their trade-off curves are just straight lines with slope -1. From there, we can also evaluate $\frac{1}{2}(M(\mathcal{E}_1,R)+M(\mathcal{E}_1,R))$ and compare it to $M(\mathcal{E},R)$. This is done in Fig. 6, revealing a violation of concavity when R comes close to 2.

In the same spirit, note that an analogous construction shows that, while

$$M(\mathcal{E}_1\otimes\mathcal{E}_2,2R)\leqslant M(\mathcal{E}_1,R)+M(\mathcal{E}_2,R) \tag{98}$$

always holds, equality (i.e., the natural "additivity" property of M under tensor products) may be violated if the ensembles are sufficiently different from each other. More generally we have the following.

Proposition 5.1:

$$M(\mathcal{E}_1\otimes\mathcal{E}_2,R)=\min\{M(\mathcal{E}_1,R_1)+M(\mathcal{E}_2,R_2):R_1+R_2=R\}.$$

Also, while $M(\mathcal{E},R)$ may not be concave in the ensemble $\mathcal{E}$, it does obey a weaker condition analogous to Schur concavity.

Proposition 5.2: Let $\mathcal{E}=\{|\varphi_i\rangle,p_i\}$ be an ensemble. Let $\{a_k\}$ be a set of probabilities with corresponding unitary operators U_k and $\mathcal{F}$ be the ensemble $\mathcal{F}=\{U_k|\varphi_i\rangle,p_ia_k\}$. Then $M(\mathcal{E},R)\leqslant M(\mathcal{F},R)$.

The proofs of these propositions can be found in Appendix Secs. 3 and 4, respectively.

As our last example, we include the trade-off curve for the uniform (unitarily invariant) ensemble on a single qubit as Fig. 7. Devetak and Berger[30] actually calculated an explicit parametrization of the optimal trade-off curve for a restricted class of encodings. Our lower bound of

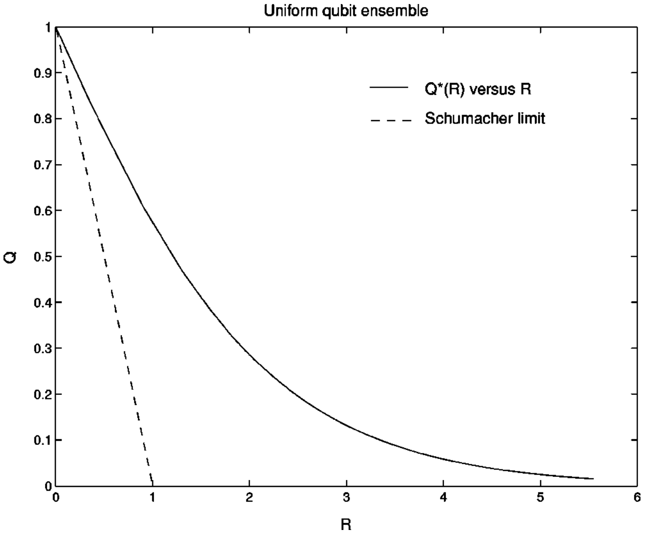

FIG. 7. Trade-off curve for the uniform qubit ensemble. Note that the curve never reaches the $Q=0$ axis, encoding the fact that no finite amount of classical information is sufficient to perfectly transmit an arbitrary qubit state.

Theorem 3.5, or, rather, its infinite source ensemble variant, Theorem 10.1, proves that their construction is optimal within all possible quantum-classical coding strategies. Thus, we can quote their result that, for $\lambda \in (0,\infty)$,

$$R=\frac{\lambda}{\epsilon^{\lambda}-1}-1+\log\left(\frac{\lambda e^{\lambda}}{\epsilon^{\lambda}-1}\right), \tag{99}$$

$$Q^*(R)=H_2\left(\frac{1}{\lambda}-\frac{1}{e^{\lambda}-1}\right), \tag{100}$$

gives a parametrization of $Q^*(R)$. This curve will also play an important role when we construct a probability-free version of our main result in Sec. VI. We will find that, in an extremely strong sense, it describes the cost of a qubit in classical bits.

VI. ARBITRARILY VARYING SOURCES

Our main result does not yet say, however, what a qubit costs in bits because it only supplies the trade-off curve $Q^*(R)$ for a given set of quantum states once a set of prior probabilities have been prescribed. Without the probabilities, the curve is undefined and the rate of exchange between bits and qubits cannot be uniquely identified. However, using the theory of *arbitrarily varying sources (AVS)* (see Ref. 31 for an exposition of this concept in classical information theory), we can develop a probability-independent version of our trade-off curve that will eliminate the ambiguity.

Throughout this section, let $\mathcal{E}$ denote not an ensemble, but just a set of states, and let $\mathbf{P}\subset\mathcal{P}_{\mathcal{E}}$ be a subset of probability distributions on $\mathcal{E}$. For each string $I\in\mathcal{I}^n$ of length n we will consider product distributions

$$p^n(I):=p_1(i_1)\cdots p_n(i_n), \tag{101}$$

where each $p_k\in\mathbf{P}$. An *AVS-code of fidelity* $1-\epsilon$ is defined as a visible code, as before (see Definition II), only that now the fidelity condition is required to hold for all probability distributions in $\mathbf{P}$:

$$\forall p^n \in \mathbf{P}^n \quad \sum_I p^n(I) F(\varphi_I, (D \circ E)(I)) \geqslant 1 - \epsilon. \tag{102}$$

The classical and quantum rates are exactly as in Definition 2.3 and, likewise, Definition 2.4 can be used unchanged to characterize attainable rate pairs (R,Q). This leads to the definition of the trade-off function $Q^*(R,\mathbf{P})$ as the minimum Q such that (R,Q) is attainable.

Intuitively, the encoder-decoder pair plays a game against a clairvoyant adversary whose aim is to minimize their average fidelity and who can control the source mechanism so as to create any of the distributions $p^n \in \mathbf{P}^n$. Their goal is to win by keeping the average fidelity above $1-\epsilon$ against arbitrary strategies of the adversary.

A special case is that of $\mathbf{P} = \mathcal{P}_{\mathcal{E}}$, in which case we have no restriction on the source, so that all possible state strings are to maintain high fidelity.

We shall use the notation $M(\mathcal{E},p,R)$ to designate our earlier function M for the ensemble consisting of the states $\mathcal{E}$ and the probabilities p, and define now

$$M(\mathcal{E},\mathbf{P},R) := \sup_{p \in \mathbf{Q}} M(\mathcal{E},p,R), \tag{103}$$

where $\mathbf{Q} := \mathrm{conv}(\mathbf{P})$ is the convex hull of $\mathbf{P}$.

Theorem 6.1: $Q^*(R,\mathbf{P}) = M(\mathcal{E},\mathbf{P},R)$.

Proof: The inequality "$\geqslant$" follows almost directly from Theorem 3.5: only observe that the adversary can simulate any source ensemble $p \in \mathbf{Q}$, and then Theorem 3.5 applies. [More formally, choose a probability distribution s on $\mathbf{P}$ such that $p = \Sigma_k s_k p_k$, and note that averaging Eq. (102) over the measure $s^{\otimes n}$ gives (102) for $p^{\otimes n}$.]

In the other direction, we only need to exhibit a covering of the union of the "probable sets" of the distributions $p^n \in \mathbf{P}^n$ by appropriate sets of typical sequences, and apply Proposition 4.1. This is done as follows:

For $p^n = p_1 \otimes \cdots \otimes p_n \in \mathbf{P}^n$ observe that the set

$$\mathcal{T}_{p^n} := \left\{ I : \forall i \left| N(i|I) - \sum_{k=1}^{n} p_k(i) \right| \leqslant \delta \sqrt{n} \right\} \tag{104}$$

carries (by Chebyshev's inequality) almost all the weight of the distribution:

$$p^n(\mathcal{T}_{p^n}) \geqslant 1 - \delta^{-2}. \tag{105}$$

Since $\mathcal{T}_{p^n}$ is in fact the same as the set of typical sequences $\mathcal{T}_{\overline{p},\delta}$, for $\overline{p} = (1/n) \Sigma_k p_k \in \mathbf{Q}$, the union $\cup_{p^n} \mathcal{T}_{p^n}$ is actually a union of certain type classes, and hence we may choose $\overline{p_1}, \dots, \overline{p_T}$, $T \leqslant (n+1)^{|\mathcal{I}|}$, such that

$$\mathcal{T} := \bigcup_{p^n \in \mathbf{P}^n} \mathcal{T}_{p^n} = \bigcup_{t=1}^{T} \mathcal{T}_{\overline{p_t},\delta}. \tag{106}$$

The coding is very simple: when $I \in \mathcal{T}$ the encoder chooses t such that $I \in \mathcal{T}_{\overline{p_t},\delta}$. He then communicates t to the decoder, and uses the protocol of Proposition 4.1. (In fact, communication of t is not even necessary, as in the latter protocol the type of I is communicated anyway.) When $I \notin \mathcal{T}$ some fixed default choice is sent.

By construction and by Proposition 4.1, for sufficiently large δ this scheme uses $R + \epsilon$ classical bits and $M(\mathcal{E},\mathbf{P},R) + \epsilon$ qubits per source symbol. For each $p^n \in \mathbf{P}^n$ we obtain high fidelity for all states outside a set of arbitrarily small probability. □

In particular, for the above-mentioned case of no restrictions at all on the probabilities, we get the trade-off function

$$\mathbf{Q}^*(R,\mathcal{P}_{\mathcal{E}}) = \sup_{p\in\mathcal{P}_{\mathcal{E}}} M(\mathcal{E},p,R). \tag{107}$$

which depends only on the states of $\mathcal{E}$. For a finite ensemble it is quite easy to show that $M(\mathcal{E},p,R)$ is continuous in the distribution p. This implies that the suprema in Eqs. (103) and (107) are, in fact, *maxima* (in the former case over the closure of $\mathbf{Q}$).

VII. INFORMATION AND DISTURBANCE

The function $M(\mathcal{E},R)$, in addition to providing the quantum-classical trade-off curve, has a number of other useful interpretations. Recall from Proposition 3.3 that

$$M(\mathcal{E},R) = \inf_{p(\cdot|\cdot)} \{S(A{:}B|C) : S(A{:}C) = R\}, \tag{108}$$

with an equality for $S(A{:}C)$ rather than the inequality we usually use. By the chain rule,

$$S(A{:}C) + S(A{:}B|C) = S(A{:}BC) \tag{109}$$

and $S(A{:}BC)$ is just the Holevo χ quantity of the ensemble

$$\mathcal{F}^{BC} := \left\{ \varphi_i^B \otimes \sum_j p(j|i)|j\rangle\langle j|^C, p_i \right\}. \tag{110}$$

Therefore, if we define the function $X(\mathcal{E},R) := R + M(\mathcal{E},R)$, then we can rewrite Eq. (108) as

$$X(\mathcal{E},R) = \inf_{p(\cdot|\cdot)} \{\chi(\mathcal{F}^{BC}) : S(A{:}C) = R\}. \tag{111}$$

The quantity on the right is now perhaps more familiar than the conditional mutual information $S(A{:}B|C)$: it is a standard measure of the distinguishability present in the ensemble $\mathcal{F}^{BC}$, minimized over all possible ways of including a fixed amount of classical information about the index i in register C. Now suppose that Alice is initially given a state $|\varphi_i\rangle$ from $\mathcal{E}$ (without the name i this time) and, via a CPTP map, manages to extract an amount R of classical information about i without damaging any of the states $|\varphi_i\rangle$. Then her final Holevo χ would necessarily be at least as large as $X(\mathcal{E},R)$, by definition. Typically, however, $X(\mathcal{E},R) > S(\mathcal{E})$ [by the Schumacher lower bound to $Q^*(R) = M(\mathcal{E},R)$], so such an operation will be forbidden by the monotonicity of χ. Therefore, it is impossible for Alice to extract information without disturbing the states.

The simple argument above combined with the additivity of $M_\epsilon(\mathcal{E},R)$ from Sec. III A can be used to prove interesting statements about the trade-off between information gain and state disturbance in an asymptotic and approximate setting. In contrast to the compression problem, however, we can make stronger statements if we use the mean letterwise fidelity measure $\overline{F}$ from Sec. III D instead of the global fidelity measure F. Therefore, we will express our results in terms of the corresponding function $\overline{M}_\epsilon(\mathcal{E}^{\otimes n},nR)$ instead of $M_\epsilon(\mathcal{E}^{\otimes n},nR)$. Recall that these functions are defined identically except that the first uses the mean fidelity function $\overline{F}$ and the second uses the global fidelity F. Likewise, define $\overline{X}_\epsilon(\mathcal{E},R) = R + \overline{M}_\epsilon(\mathcal{E},R)$. Since F and $\overline{F}$ are identical for a single copy, we have $\overline{M}_\epsilon(\mathcal{E},R) = M_\epsilon(\mathcal{E},R)$ and similarly for X and $\overline{X}$. By the discussion in Sec. III D, we know that $\overline{M}_\epsilon(\mathcal{E}^{\otimes n},nR) = n\overline{M}_\epsilon(\mathcal{E},R)$, which in turn implies

$$\overline{X}_\epsilon(\mathcal{E}^{\otimes n},nR) = nX_\epsilon(\mathcal{E},R). \tag{112}$$

Now, generalizing the above single copy argument, suppose that Alice is given a state $|\varphi_I\rangle$ drawn from $\mathcal{E}^{\otimes n}$, which, by a CPTP map Γ, she manages to convert into the state

$$\rho_I = \sum_j \tilde{\varphi}^B_{I,j} \otimes p(j|I)|j\rangle\langle j|^C, \tag{113}$$

with a quantum and classical part such that the mutual information $H(I:j) \geqslant nR$ and the mean letterwise fidelity between Alice's initial states and her final states of system B satisfies

$$\overline{F}(\mathcal{E}^{\otimes n}, \mathrm{Tr}_C \circ \Gamma(\mathcal{E}^{\otimes n})) := \sum_I p_I \frac{1}{n} \sum_{k=1}^{n} F(\varphi_{i_k}, \mathrm{Tr}_{\neq k} \circ \mathrm{Tr}_C(\rho_I)) \geqslant 1-\epsilon. \tag{114}$$

Writing $\mathcal{F}^{BC} = \{\Gamma(\varphi_I), p_I\}$, the monotonicity of χ guarantees that $nS(\mathcal{E}) \geqslant \chi^{BC}$ and it is easy to see that $\chi^{BC} \geqslant \overline{X}_\epsilon(\mathcal{E}^{\otimes n}, nR)$. By applying Eq. (112), we then find

$$S(\mathcal{E}) \geqslant X_\epsilon(\mathcal{E}, R), \tag{115}$$

in which, conspicuously, all dependence on n has vanished. In other words, in order to maximize her information at a given mean letterwise fidelity, Alice should just repeat the optimal single letter strategy for each position; she need not ever apply any collective operations. Summarizing these observations, we have the following.

Theorem 7.1: *Suppose we have a set of states $|\varphi_I\rangle$ drawn from the ensemble $\mathcal{E}^{\otimes n}$ represented on system B and let Γ be a CPTP map from B to the joint system BC, where C is classical, satisfying the following conditions:*

(1) *$H(I:j) \geqslant nR$, where j is the classical output on system C.*
(2) *The mean letterwise fidelity $\overline{F}(\mathcal{E}^{\otimes n}, \mathrm{Tr}_C \circ \Gamma(\mathcal{E}^{\otimes n})) \geqslant 1-\epsilon$.*

Then, for each $\epsilon > 0$, the inequality $S(\mathcal{E}) \geqslant X_\epsilon(\mathcal{E}, R)$ holds. Moreover, the Holevo quantity of the ensemble $\mathcal{F}^{BC} = \{\Gamma(\varphi_I), p_I\}$ satisfies the inequality $\chi(\mathcal{F}^{BC}) \geqslant nX_\epsilon(\mathcal{E}, R)$.

□

One application of the theorem is that it provides an alternative method for analyzing the quantum resources required for blind compression, which was the subject of Ref. 13. The idea is simply to think of the map Γ as the composition $D_n \circ E_n$ of the encoding and decoding maps for blocks of size n. (Because classical information can be copied, we can assume without loss of generality that the decoder keeps his classical information around after the decoding stage has been completed.) Now suppose that the scheme has classical mutual information $H(I:j) \geqslant nR$. If it also has mean letterwise fidelity $1-\epsilon_n$, then, as for the visible case,

$$\mathrm{qsupp} \geqslant \frac{1}{n} \overline{M}_{\epsilon_n}(\mathcal{E}^{\otimes n}, nR) = M_{\epsilon_n}(\mathcal{E}, R). \tag{116}$$

By the previous theorem, however, we must also have the inequality $S(\mathcal{E}) \geqslant X_{\epsilon_n}(\mathcal{E}, R)$. Moreover, if perfect compression is possible asymptotically (using either the block or letterwise fidelity conditions), we get the stronger inequality

$$S(\mathcal{E}) \geqslant \lim_{\epsilon \downarrow 0} X_\epsilon(\mathcal{E}, R) = X_0(\mathcal{E}, R). \tag{117}$$

(The continuity at $\epsilon = 0$ follows from the continuity of M_0, demonstrated earlier.) Because the ensemble $\mathcal{E}$ can always be recovered by tracing over the C register, the monotonicity of χ guarantees that the right hand side is always at least as large as the left, implying $S(\mathcal{E}) = X_0(\mathcal{E}, R)$. We are, therefore, interested in the equality conditions for monotonicity.

Recalling some terminology from Ref. 13, we say an ensemble $\mathcal{E}$ is *reducible* if its states can be partitioned into two nonempty sets with orthogonal supports. An ensemble is said to be irreducible if it is not reducible. Every ensemble, therefore, can be decomposed into orthogonal, irreducible subensembles as

$$\mathcal{E}=\bigcup_{l=1}^{L} a_l \mathcal{E}_l, \tag{118}$$

where a_l is the total probability weight of states in subensemble $\mathcal{E}_l$.

Proposition 7.2: Let $\mathcal{E}=\cup_{l=1}^{L} a_l \mathcal{E}_l$ be a decomposition of the pure-state ensemble $\mathcal{E}$ into irreducible subensembles $\mathcal{E}_l=\{|\varphi_{il}\rangle, p_{i|l}\}$ and let $\mathcal{F}^{BC}=\{\varphi_{il}^{B}\otimes\omega_{il}^{C}, a_l p_{i|l}\}$ be a bipartite extension of the ensemble $\mathcal{E}$. Then $S(\mathcal{E})=\chi(\mathcal{F}^{BC})$ if and only if $\omega_{il}=\omega_{jl}$ for all i, j, and l.

A proof is given in the Appendix, Sec. 5. The meaning of the proposition is essentially that the only information that can be stored on register C without increasing χ is the classical information already present on register B, so that ω_{il} must be a function of l alone. Therefore, in order to satisfy Eq. (117) it is necessary that $R \leqslant H(a_1, \ldots, a_L)$. Conversely, provided the inequality holds, it is possible to extract R bits per signal without disturbance at the encoding stage, at which point the encoding scheme we used for visible compression can be used to achieve the quantum rate $S(\mathcal{E})-R$. Putting these observations together, we obtain an alternative demonstration of the main theorem of Ref. 13:

Theorem 7.3: *Let $\mathcal{E}=\cup_{l=1}^{L} a_l \mathcal{E}_l$ be a decomposition of the ensemble $\mathcal{E}$ into orthogonal, irreducible subensembles. Then blind compression of $\mathcal{E}$ to Q qubits per signal plus auxiliary classical storage is possible if and only if*

$$Q \geqslant \sum_{l} a_l S(\mathcal{E}_l) = S(\mathcal{E}) - H(a_1, \ldots, a_L). \tag{119}$$

□

Thus, the techniques we have introduced to analyze the visible compression problem provide a unified framework for analyzing blind compression as well. In fact, we will see in the next section that the trade-off curve for yet another related problem—remote state preparation—can also be calculated using similar methods.

VIII. APPLICATION TO REMOTE STATE PREPARATION

Remote state preparation, introduced in Ref. 17 following a conjecture of Lo's,[16] is very similar to what we have considered here: it is a visible coding problem for quantum states involving classical resources, in the form of communication, and quantum resources, this time in the form of entanglement. Furthermore, these two types of resources can be traded against each other so it is natural to study the optimal trade-off curve.

Without giving formal definitions, let $E^*(R)$ be the minimum rate of entanglement sufficient for a remote state preparation protocol with classical rate R, such that the average fidelity tends to 1 with growing blocklength.

Given that entanglement can be set up using quantum communication at a cost of one qubit per ebit, and that, on the other hand, quantum communication can be accomplished using teleportation[32] at a cost of two cbits and one ebit per qubit, it is clear that coding methods for the one problem immediately yield (possibly suboptimal) procedures for the other. (In fact, by making use of quantum-classical trade-off coding, this resulted in the ‘‘cap-method’’ of Ref. 17, which was further refined in Ref. 30.)

In Ref. 33 a method of remote state preparation is developed that works for visible coding of product states and is more efficient than teleportation: we really need only to use *one* cbit and one ebit per qubit, asymptotically.

Theorem 8.1 (See Ref. 33): *Given a finite set $\mathcal{X}$ of states (density operators) on $\mathcal{K}$, there is a probabilistic exact (one-shot) remote state preparation protocol working for all states in $\mathcal{X}$ and with failure probability uniformly ϵ, using a maximally entangled state $|\Phi\rangle$ on $\mathcal{K}\otimes\mathcal{K}$ and classical communication of a message out of*

$$M \leqslant 1 + \frac{2\ln 2}{\epsilon^2} \log(2|\mathcal{X}|\dim\mathcal{K})\dim\mathcal{K}.$$

□

This leads immediately to the following.

Theorem 8.2: *For the source $\mathcal{E}=\{|\varphi_i\rangle, p_i\}$ of quantum states, if $R \geqslant 0$ and $Q = Q^*(R)$, then $E^*(R+Q) \leqslant Q$.*

As a consequence, we obtain

$$E^*(R) \leqslant N(\mathcal{E},R) := \min_{p(\cdot|\cdot)} \{S(A{:}B|C) : S(A{:}BC) \leqslant R\},$$

minimization over the same set of tripartite states as in the definition of M.

Proof: We apply Theorem 8.1 to the space $\mathcal{K}$ of *encoded states* of an optimal trade-off coding using R cbits and Q qubits per source symbol, and to the set of all possible encoded states: note that $|\mathcal{X}| \leqslant (|\mathcal{I}||\mathcal{J}|)^n$.

By that result, we need Q ebits to do this, and an additional $Q+o(1)$ cbits to the R cbits from the trade-off coding. □

In fact, in Ref. 33 it is shown, by methods very similar to those in Sec. III, that the above estimate for E^* is in fact an equality, and that our AVS considerations also carry over.

Theorem 8.3: *For the state set $\mathcal{E}$ and AVS* $\mathbf{P}$,

$$E^*(R,\mathbf{P}) = \sup_{p \in \mathbf{Q}} N(\mathcal{E},p,R),$$

with $\mathbf{Q} = \mathrm{conv}(\mathbf{P})$. □

For $\mathbf{P}$ the set of all distributions on the pure states (as indeed for any symmetric family of distributions) we can prove symmetry results like those in the upcoming Sec. IX, and arrive at the conclusion that the *absolute trade-off* between cbits and ebits in remote state preparation is given by the curve $N(\mathcal{P}(\mathcal{H}),u)$, where u is the uniform (i.e., unitarily invariant) measure on the set $\mathcal{P}(\mathcal{H})$ of all pure states on $\mathcal{H}$. Devetak and Berger[30] arrived at a slightly different curve as an upper bound to the true trade-off, starting from $M(\mathcal{P}(\mathcal{H}),u)$ as we did, but employing teleporation instead of the newer technique in Theorem 8.1. For this reason their conjecture that their bound is tight is not correct.

IX. SYMMETRY IN THE ENSEMBLE

Our formulas for the trade-off curve, both in the known and arbitrarily varying source case, can be considerably simplified if there is symmetry in the set of states.

Assume that there is a group G acting on the labels i of the states by a projective unitary representation U_g,

$$\forall g \in G, i \in \mathcal{I} \quad |\varphi_{gi}\rangle\langle\varphi_{gi}| = U_g|\varphi_i\rangle\langle\varphi_i|U_g^\dagger. \tag{120}$$

(We will present the following arguments for a finite group, but they also apply to compact groups: in fact, we only need the existence of an invariant measure, see Ref. 34.) The action of G on $\mathcal{I}$ induces an action on the probability distributions on $\mathcal{I}$ in a natural way: if $p \in \mathcal{P}(\mathcal{I})$ is a distribution, then $p^g(i) = p(g^{-1}i)$ defines the translated distribution. Assume now further that the arbitarily varying source $\mathbf{P}$ is stable under this induced action:

$$\forall p \in \mathbf{P} \quad p^g \in \mathbf{P}. \tag{121}$$

[In the "known source" case, $\mathbf{P} = \{p\}$, this simply means that $p(gi) = p(i)$ for all $i \in \mathcal{I}$ and $g \in G$.]

By the formula for the trade-off curve, Eq. (103), we may assume that $\mathbf{P}$ is convex. Letting

$$\mathbf{P}^G := \{p \in \mathbf{P} : \forall g \in G\, p^g = p\}, \tag{122}$$

we can then prove the following.

Theorem 9.1: *For any G-invariant state set and AVS* **P**,

$$M(\mathcal{E},\mathbf{P},R) = M(\mathcal{E},\mathbf{P}^G,R). \tag{123}$$

Proof: The lhs is by definition greater than or equal than the rhs.

For the opposite inequality we make use of the "restricted concavity" given in proposition 5.2. For the rotations U_g applied with equal probabilities to the ensemble $(\mathcal{E},p)$, we get

$$M\left(\bigcup_g U_g \mathcal{E} U_g^\dagger , \frac{1}{|G|} \sum_g \; p^g,R \right) \geqslant \frac{1}{|G|} M(U_g \mathcal{E} U_g^\dagger ,p^g,R) = M(\mathcal{E},p,R). \tag{124}$$

Note that $(1/|G|)\,\Sigma_g p^g \in \mathbf{P}^G$ and, since the state set is G invariant, we have $\cup_g U_g \mathcal{E} U_g^\dagger = \mathcal{E}$, which proves our claim. □

If G acts *transitively*, this leads to a dramatic simplification of the formula for the AVS trade-off curve (Theorem 6.1): in this case the only G-invariant distribution is the uniform distribution, so from Theorem 6.1 we obtain the following.

Corollary 9.2: For an AVS $(\mathcal{E},\mathbf{P})$ *with transitive group action under which* **P** *is stable, (e.g., for* $\mathbf{P} = \mathcal{P}_{\mathcal{E}}$*), we have*

$$Q^*(R,\mathbf{P}) = M(\mathcal{E},u,R),$$

where u *is the uniform distribution on* $\mathcal{E}$. □

The particular example of $\mathcal{E}$ being the set of all pure states on $\mathcal{H}$ and **P** being the set of all distributions on $\mathcal{E}$ is arguably the setting for *the* trade-off between classical and quantum bits: the trade-off coding becomes a statement solely about states, with no mention of prior probabilities. Of course we have not yet justified the application of our results to infinite state sets. The corresponding but more involved treatment of the coding bounds will be given in Sec. X.

Given this generalization to infinite state sets, we conclude that the *absolute trade-off* for pure states on $\mathcal{H}$ is given by $M(\mathcal{P}(\mathcal{H}),u)$, with the uniform (i.e., unitarily invariant) measure u on the set $\mathcal{P}(\mathcal{H})$ of all pure states. The Devetak-Berger curve introduced earlier corresponds to the case $\mathcal{H} = \mathbb{C}^2$.

Remark: From the proof of Theorem 9.1, we see that we may always restrict the classical encodings $p(\cdot|\cdot)$ to be group covariant as well, in the sense that, for each $j \in \mathcal{J}$, the distribution $q(\cdot|j)$ has the property that for each $g \in G$ there exists a j' satisfying $q_{j'} = q_j$ and $q(gi|j) = q(i|j')$ for all $i \in \mathcal{I}$:

Define a new encoding p' by letting

$$p'(j,g|gi) := \frac{1}{|G|} p(j|i). \tag{125}$$

For a G-invariant distribution p on the ensemble states this does not change the values of $S(A:C)$ and $S(A:B|C)$. However, the resulting probabilities $q'_{j,g} = q_j$ and $q'(gi|j,g) = p_i p(j|i)/q'_{j,g}$ have a useful property: there is a group action of G on the indices (j,g) under which the distribution q' is invariant, and the set of conditional distributions $q'(\cdot|j,g)$ is stable. More precisely, h acts on (j,g) by $h\cdot(j,g) = (j,hg)$. Obviously, q' is invariant under this, and

$$q'(gi|h\cdot(j,g)) = q'(gi|j,hg) = q'(h^{-1}hgi|j,gh), \tag{126}$$

saying that $q'(\cdot|h\cdot(j,g)) = (q'(\cdot|j,hg))^h$.

Hence, when discussing optimal codings given by q_j and $q(\cdot|j)$ such that $\Sigma_j q_j q(\cdot|j) = p$, we may always assume that G also acts on the set of j's, and that

$$\forall j \forall g \quad q_{gj} = q_j \text{ and } q(\cdot|gj) = (q(\cdot|j))^g. \tag{127}$$

□

We close this section by giving a bound on the size of the classical register for a finite ensemble with symmetry, which sometimes improves our earlier result in Proposition 3.4:

Proposition 9.3: Let the group G act on the ensemble $\mathcal{E}=\{\varphi_i, p_i\}_{i\in\mathcal{I}}$ in the way described at the beginning of this section, and assume that p is G-invariant. If the group action partitions $\mathcal{I}$ into t G-orbits, then for every R there exists a classical encoding $p(\cdot|\cdot):\mathcal{I}\rightarrow\mathcal{J}$ which is covariant in the above sense, and satisfies

$$|\mathcal{J}| \leqslant |G|(t+1), \quad S(A:C) \leqslant R, \quad S(A:B|C) = M(\mathcal{E},R).$$

In fact, $\mathcal{J}$ partitions into $t+1$ G-orbits, in the sense described above.

The proof is given in the Appendix, Sec. 6

Example: Let $\mathcal{E}$ consist of any two states: $\mathcal{E}=\{|\varphi_i\rangle\}_{i=1}^2$. By choosing a reflection that swaps $|\varphi_1\rangle$ and $|\varphi_2\rangle$, we get a transitive $\mathbb{Z}_2$ action on the indices i. Therefore, for the AVS $(\mathcal{E},\mathcal{P}_\mathcal{E})$, we have $Q^*(R,\mathbf{P}) = M(\mathcal{E},u,R)$, where u is the uniform distribution $p_i = \frac{1}{2}$. This distribution is clearly G-invariant, so Proposition 9.3 ensures that there is an optimal encoding for which $\mathcal{J}$ partitions into at most $t+1=2$ orbits, each of size either 1 or 2. □

Example: For states in the BB84 ensemble $\mathcal{E}_{BB}(\theta)$, the group $\mathbb{Z}_2\times\mathbb{Z}_2$ acts transitively via reflection along the $\theta/2$ axis and rotation by $\pi/2$. Therefore, once again, the unrestricted AVS can be reduced to the uniform ensemble, for which the optimal encoding can be assumed G-covariant, with $\mathcal{J}$ partitioning into at most two orbits of length 1, 2 or 4. □

X. INFINITE SOURCE ENSEMBLES

It should be noted that, even in the technical parts of our proofs, and, indeed, in the very statements of the *coding theorems*, we assumed that the sets of states under consideration were *finite*.

As there are interesting examples of ensembles with infinite state sets, including perhaps most notably the whole manifold of pure states in a Hilbert space, we show here how a certain approximation technique (used in Ref. 25 to deal with coding for nonstationary quantum channels) can be used to transfer our main results quite directly. The procedure, unfortunately, is not entirely painless; we have to go through the proof of Proposition 4.1 again with a modified and more technical version of the typical subspace. That is why we have chosen to treat the infinite source case separately, confining the details to this section.

A. Formulation of information quantities and the lower bound

To be able to consider infinite ensembles and encodings, we have to reformulate our notions from Secs. II and III in terms of general measure spaces (for the background and terminology see any textbooks on probability, such as Ref. 35, and measure theory[34]):

The source ensemble $\mathcal{E}$ is described by a measure space Ω (with probability measure P), and a measurable map $\varphi:\Omega\rightarrow\mathcal{P}(\mathcal{H})\subset\mathcal{S}(\mathcal{H})$ from Ω into the set of pure states on the Hilbert space $\mathcal{H}$ (which is still of finite dimension d), mapping $\omega\in\Omega$ to $|\varphi_\omega\rangle\langle\varphi_\omega|$. We can then easily define encoding and decoding (E,D) for blocks of length n:

$$E:\Omega^n \rightarrow \mathcal{S}(\mathcal{H}_B)\times\Omega_C, \tag{128}$$

$$D:\mathcal{B}(\mathcal{H}_B)\otimes\mathcal{B}(\ell^2(\Omega_C)) \rightarrow \mathcal{B}_d^{\otimes n}, \tag{129}$$

where E is a Markov kernel, Ω_C is a *finite set*, and D is CPTP. The quantification of classical and quantum resources we adopt unchanged, and the fidelity condition reads as follows: the combined encoding and decoding gives rise to a Markov kernel

$$D\circ E:\Omega^n \rightarrow \mathcal{B}_d^{\otimes n}, \tag{130}$$

and, using the abbreviation

$$(D\circ E)(\omega_1\cdots\omega_n)=\int_{\mathcal{B}(\mathcal{H}_B)}(D\circ E)(\mathrm{d}\sigma|\omega_1\cdots\omega_n)\sigma, \tag{131}$$

we require that

$$F=\int_{\Omega^n}P^{\otimes n}(\mathrm{d}\omega_1\cdots\omega_n)F(\varphi_{\omega_1\cdots\omega_n},(D\circ E)(\omega_1\cdots\omega_n))\geqslant 1-\epsilon. \tag{132}$$

Let us denote by μ the measure induced by P and this Markov kernel on $\Omega\times\mathcal{S}(\mathcal{H}_B)\times\Omega_C$:

$$\mu(F_A\times G_{BC}):=\int_{F_A}P(\mathrm{d}\omega)E(G_{BC}|\omega). \tag{133}$$

We denote its restrictions (marginals) to factors $\Omega_A=\Omega$, $\mathcal{S}(\mathcal{H}_B)$, Ω_C by $P=\mu_A$, μ_B, $q:=\mu_C$, respectively, and analogously μ_{AC}, etc.

With the help of Radon–Nikodym derivatives we can always construct the Bayesian ''inverse'' Markov kernel

$$q:\Omega_C\rightarrow\Omega_A\times\mathcal{S}(\mathcal{H}_B) \tag{134}$$

that gives rise to the same joint distribution:

$$\int_{G_C}\mu_C(\mathrm{d}j)q(F_{AB}|j)=\mu(F_{AB}\times G_C). \tag{135}$$

In fact, μ_C-almost everywhere,

$$q(F_{AB}|j)=\frac{\mathrm{d}\mu(F_{AB}\times\{j\})}{\mathrm{d}\mu_C(j)}. \tag{136}$$

To follow the procedure of Sec. III we have to define the relevant information quantities (for their properties, see Refs. 36 and 37):

First, $S(A:C)$ can be expressed as $D(\mu_{AC}\|\mu_A\otimes\mu_C)$, in terms of the relative entropy (or Kullback–Leibler divergence) of two measures

$$D(\mu\|\lambda):=\int\mu(\mathrm{d}x)\log\left(\frac{\mathrm{d}\mu(x)}{\mathrm{d}\lambda(x)}\right), \tag{137}$$

where $\mathrm{d}\mu(x)/\mathrm{d}\lambda(x)$ denotes the Radon–Nikodym derivative. If this does not exist μ-almost everywhere, we define $D(\mu\|\lambda)=\infty$. It is a fact that in Eq. (137) the Radon–Nikodym derivative always exists, and it can be checked that in the finite case the new definition coincides with the old.

Second, $S(A:B|C)=\int_{\Omega_C}q(\mathrm{d}j)S(A:B|C=j)$, with $S(A:B|C=j)$ denoting the quantum mutual information associated to the conditional probability measure $q(\cdot|j)$ on $\Omega_A\times\mathcal{S}(\mathcal{H}_B)$: for any such distribution λ, with first marginal λ_A and Markov kernel $L:\Omega_A\rightarrow\mathcal{S}(\mathcal{H})$,

$$S_\lambda(A:B)=S\left(\int_{\mathcal{S}(\mathcal{H})}\lambda_B(\mathrm{d}\sigma)\sigma\right)-\int_{\Omega_A}\lambda_A(\mathrm{d}\omega)S\left(\int_{\mathcal{S}(\mathcal{H})}L(\mathrm{d}\sigma|\omega)\sigma\right). \tag{138}$$

Again, it is possible to check that for discrete probability spaces we obtain the same expressions as before.

The proofs of Lemmas 3.1 and 3.2 and of Theorem 3.5 are directly adapted to this language, essentially replacing all sums representing probability averages by integrals. (Note that even the ‘‘continuity in ϵ’’ part in the latter applies as the functions f and g depend only on ϵ and d.) This is possible since the monotonicity and convexity properties we used are still true in the infinite setting.

At the end of the proof we arrive at encodings mapping $\omega\in\Omega$ to $|\varphi_\omega\rangle\langle\varphi_\omega|\otimes\Sigma_j p(j|\omega)|j\rangle \times\langle j|$ (i.e., the corresponding Markov kernel maps i to the point mass at $|\varphi_\omega\rangle\langle\varphi_\omega|$ times a discrete measure on Ω_C). Such encodings we denote ‘‘$p:\Omega_A\to\Omega_C$,’’ and we get

$$Q^*(R)\geqslant \inf_{p:\Omega_A\to\Omega_C,|\Omega_C|<\infty}\{S(A:B|C):S(A:C)\leqslant R\}. \tag{139}$$

Dropping the finiteness of Ω_C can only decrease the lower bound, and we arrive at the following general version of Theorem 3.5:

Theorem 10.1: *For the ensemble* $\mathcal{E}=(\Omega,P,\varphi)$,

$$Q^*(R)\geqslant M(\mathcal{E},R):=\inf_{p:\Omega_A\to\Omega_C}\{S(A:B|C):S(A:C)\leqslant R\},$$

with

$$S(A:C)=D(\mu\|P\otimes q),$$

$$S(A:B|C)=\int_{\Omega_C}q(\mathrm{d}j)S\left(\int_{\Omega_A}q(\mathrm{d}\omega|j)|\varphi_\omega\rangle\langle\varphi_\omega|\right),$$

where μ is the measure on $\Omega_A\times\Omega_C$ induced by P and the Markov kernel $p(\cdot|\cdot)$, q is its marginal on Ω_C and $q(\cdot|\cdot)$ is the Bayesian Markov kernel $\Omega_C\to\Omega_A$. □

B. Adaptation of the coding theorem

The obstacles to an application of our coding scheme, Proposition 4.1, are the potentially infinite range of the source register (Ω) and the classical encoding (Ω_C). Of course, when in the previous subsection we allowed the latter to be infinite, we only made M smaller, and at that point it was not clear that this was a good move.

The purpose of the present subsection is to show that it is possible to approximate the effect of an infinite encoding by a strictly finite one: finitely many possible states on $\mathcal{H}$ and finitely many classical symbols. This will inevitably introduce some error, which we will have to counter by a suitably adapted notion of typical subspace.

Lemma 10.2: For $\epsilon>0$ there exists a partition of $\mathcal{S}(\mathcal{H})$ into $m\leqslant C(d)\epsilon^{-d^2}$ Borel sets each of which has radius at most ϵ: in each part $\mathcal{S}_i$ there exists a state σ_i such that for all $\rho\in\mathcal{S}_i$, $\|\rho-\sigma_i\|_1\leqslant\epsilon$. The constant $C(d)$ depends only on d.

Proof: The set of states on $\mathcal{H}$ is affinely isomorphic to the set of positive complex $d\times d$-matrices with trace 1, which is contained in the set of self-adjoint complex matrices with all d^2 real and imaginary parts of entries in the interval $[-1,1]$: this is a d^2-dimensional hypercube. This can be partitioned into $(2\sqrt{2}d^3)^{d^2}\epsilon^{-d^2}$ many small hypercubes of edge length $\epsilon/(d^3\sqrt{2})$. It is easy to check that for any ρ,σ in the same small cube, $\|\rho-\sigma\|_1\leqslant\epsilon$. □

For a source (Ω,P,φ) such a partition entails a partition $\mathcal{Z}$ of Ω into at most m measurable pieces Z_i, with $\omega_i\in Z_i$ such that $|\varphi_{\omega_i}\rangle\langle\varphi_{\omega_i}|=\sigma_i$. (We need only consider pieces that intersect the image of φ.) A central role will be played by the ‘‘contraction’’ of the infinite ensemble $\mathcal{E}$ to the finite ensemble $\mathcal{E}'=\{\varphi_{\omega_i},\hat{P}(i)=P(Z_i)\}$ which is obtained by identifying all of Z_t to the single state φ_{ω_i}.

We have already defined the set of $\hat{P}$-typical sequences $\mathcal{T}_{\hat{P},\delta}$, and now can define the following typical set for P:

$$\mathcal{T}^{\mathcal{Z}}_{P,\delta} := \bigcup_{I\in\mathcal{T}_{\overline{P},\delta}} Z_{i_1}\times\cdots\times Z_{i_n}. \tag{140}$$

It obviously inherits the large probability property of $\mathcal{T}_{p',\delta}$:

$$P^{\otimes n}(\mathcal{T}^{\mathcal{Z}}_{P,\delta}) \geqslant 1-\frac{1}{\delta^2}. \tag{141}$$

Before we can describe the coding scheme we have to introduce a variant of the conditional typical sequences and subspaces: for a channel $W:\mathcal{I}\rightarrow\mathcal{J}$ and $\delta,\epsilon>0$ define

$$\mathcal{T}^{(\epsilon)}_{W,\delta}(I) := \{J:\forall ij\,|N(ij|IJ)-N(i|I)W(j|i)|\leqslant\delta\sqrt{N(i|I)}+\epsilon N(i|I)\}. \tag{142}$$

(Our previous notion is recovered with $\epsilon=0$, and in the sequel ϵ will be small, compared to δ which we shall choose large.) Observe that this is a union of conditional type classes. Using Eq. (78) it is quite easy to show that

$$\begin{aligned}|\mathcal{T}^{(\epsilon)}_{W,\delta}(I)| &\leqslant (n+1)^{|\mathcal{I}||\mathcal{J}|}\exp\left(nH(W|P_I)+\sum_i N(i|I)|\mathcal{J}|\,\eta(\epsilon+\delta N(i|I)^{-1/2})\right)\\ &\leqslant (n+1)^{|\mathcal{I}||\mathcal{J}|}\exp(nH(W|P_I)+n|\mathcal{J}|\,\eta(\epsilon)+n\,\eta(\delta|\mathcal{I}|/\sqrt{n})),\end{aligned} \tag{143}$$

where we have used the inequality $\eta(x+y)\leqslant\eta(x)+\eta(y)$ and concavity of η.

Similarly, for a collection of states W_i, which we endow with fixed diagonalizations $W_i=\Sigma_{j=1}^d W(j|i)|e_{j|i}\rangle\langle e_{j|i}|$, we can define the projector

$$\Pi^{(\epsilon)}_{W,\delta}(I) := \sum_{J\in\mathcal{T}^{(\epsilon)}_{W,\delta}(I)} |e_{J|I}\rangle\langle e_{J|I}|, \tag{144}$$

and get from Eq. (143) the estimate

$$\mathrm{Tr}\,\Pi^{(\epsilon)}_{W,\delta}(I)\leqslant(n+1)^{d|\mathcal{I}|}\exp(nH(W|P_I)+nd\,\eta(\epsilon)+n\,\eta(\delta|\mathcal{I}|/\sqrt{n})). \tag{145}$$

Its other most important property that we shall use is the following: consider a product state $\sigma=\sigma_1\otimes\cdots\otimes\sigma_n$ such that, with some $I=i_1\cdots i_n$,

$$\forall i\quad \left\|\frac{1}{N(i|I)}\sum_{k:i_k=i}\sigma_k-W_i\right\|_1\leqslant\epsilon. \tag{146}$$

Then we claim that

$$\mathrm{Tr}(\sigma\Pi^{(\epsilon)}_{W,\delta}(I))\geqslant 1-\frac{|\mathcal{I}|}{\delta^2}. \tag{147}$$

The proof goes as follows: the left hand side above does not change if we replace σ_k by $\sigma'_k := \Sigma_j|e_{j|i_k}\rangle\langle e_{j|i_k}|\sigma_k|e_{j|i_k}\rangle\langle e_{j|i_k}|$, because the projector is a sum of one-dimensional projectors $|e_{J|I}\rangle\langle e_{J|I}|$. Thus we may assume that σ_k has diagonal form in the chosen eigenbasis of W_{i_k}: $\sigma_k=\Sigma_j S_k(j)|e_{j|i_k}\rangle\langle e_{j|i_k}|$.

Note that the left hand side of Eq. (147) can be rewritten as $(S_1\otimes\cdots\otimes S_n)(\mathcal{T}^{(\epsilon)}_{W,\delta}(I))$, a classical probability. Now it is immediate from the definition of the latter set [Eq. (142)] and from the condition (146) on σ that

$$\mathcal{T}^{(\epsilon)}_{W,\delta}(I)\supset\mathcal{T}_{\overline{S},\delta}(I), \tag{148}$$

with the channel $\overline{S}(j|i)=[1/N(i|I)]\Sigma_{k:i_k=i}S_k(j)$. Hence

$$(S_1\otimes\cdots\otimes S_n)(\mathcal{T}^{(\epsilon)}_{W,\delta}(I))\geqslant(S_1\otimes\cdots\otimes S_n)(\mathcal{T}_{\overline{S},\delta}(I))\geqslant\left(1-\frac{1}{\delta^2}\right)^{|\mathcal{I}|}\geqslant 1-\frac{|\mathcal{I}|}{\delta^2}, \tag{149}$$

the second line by Chebyshev's inequality.

After these preparations we are ready to prove the infinite source version of Proposition 4.1:

Proposition 10.3: Let $\mathcal{E}=(\Omega_a,P,\varphi)$ *be a source. For a probability distribution* P *on* Ω *and a Markov kernel* $p(\cdot|\cdot):\Omega_A\rightarrow\Omega_C$, $\epsilon>0$, *there exists a partition* $\mathcal{Z}$ *of* Ω_A *into* $m-1<C(d)\epsilon^{-d^2}$ *measurable sets, corresponding to an* ϵ*-fine partition of the state space, and for* $\delta>0$ *a visible code* (E,D) *such that*

$$\forall\,\omega=(\omega_1\cdots\omega_n)\in\mathcal{T}^{\mathcal{Z}}_{P,\delta}\quad F(|\varphi_\omega\rangle\langle\varphi_\omega|,(D\circ E)(\omega))\geqslant 1-\frac{4m^2}{\delta^2}.$$

and sending

$$nS(A:C)+nKm^2\,\eta(\delta/\sqrt{n})+K'm^2\log(n+1)\quad \textit{classical bits,}$$

$$nS(A:B|C)+n(3dm^2\eta(2\,\delta m^2/\sqrt{n})+3d\,\eta(\epsilon))+dm\log(n+1)\quad \textit{quantum bits.}$$

Proof: We can find the partition by Lemma 10.2 and the discussion thereafter.

Consider now the (measurable) coarse-graining map

$$T:\omega\mapsto i\in\{1,\ldots,m-1\}\text{for}\quad\omega\in Z_i. \tag{150}$$

Applying T to Ω_A [and the identity map to $\mathcal{B}(\mathcal{H}_B)$ and Ω_C] leads to a new distribution μ' on $\Omega_{A'}\times\mathcal{B}(\mathcal{H}_B)\times\Omega_C$, with $\Omega_{A'}=\{1,\ldots,m-1\}$. By the data-processing inequality[23,37] we have

$$S(A':C)\leqslant S(A:C)\quad\text{and}\quad S(A':B|C)\leqslant S(A:B|C). \tag{151}$$

Next we change the quantum part of the encoding by collecting all the weight of a piece Z_i into $\varphi_i:=\varphi_{\omega_i}$: we can do this by a similar coarse-graining map

$$\widetilde{T}:\sigma\mapsto|\varphi_i\rangle\langle\varphi_i|\quad\text{for}\quad\sigma\in Z_i. \tag{152}$$

The resulting distribution will be denoted by μ'': it is supported on a finite set $\Omega_{A'}$ and a finite set of states φ_i (in fact, the "contracted" ensemble $\mathcal{E}'$ of the discussion after Lemma 10.2). It is generated by a Markov kernel $\hat{p}:\Omega_{A'}\rightarrow\Omega_C$, which in this case is simply a finite collection of (conditional) distributions $\hat{p}(\cdot|i)$ on Ω_C. Note that this is a valid encoding in the sense of the definition of $M(\mathcal{E}',R)$, in the main section. Let us denote the corresponding conditional quantum mutual information by $S(A':B'|C)$.

By definition of $S(A':B|C)$ and the partition $\mathcal{Z}$, we have

$$S(A':B'|C)\leqslant S(A':B|C)+2d\,\eta(\epsilon/d), \tag{153}$$

using Fannes' inequality (52) twice.

To end this step-by-step discretization, we may change the encoding to a stochastic matrix $p':\Omega_{A'}\rightarrow\{1,\ldots,m\}=:\Omega_{C'}$, by the considerations of Sec. III (see also Proposition 9.3), such that

$$S(A':B'|C')\leqslant S(A':B'|C)\quad\text{and}\quad S(A':C')=S(A':C). \tag{154}$$

So, finally, we are in a position to apply the coding method of Proposition 4.1, with the sole difference that we use for the quantum encoding the projector $\Pi^{(\epsilon)}_{p',\delta}(I)$ instead of our previous conditional typical projector, and I is such that $\omega_1\cdots\omega_n\in Z_I$.

The fidelity estimate is obtained just like there, only using Eq. (147). The classical rate estimate we copy from Proposition 4.1, and for the quantum rate estimate, we follow its derivation in the proof, using Eq. (145) to estimate the range of the projectors $\Pi^{(\epsilon)}_{p',\delta}(I)$: we have to send

$$nS(A':B'|C')+n(3dm^2\eta(2\delta m^2/\sqrt{n})+d\eta(\epsilon))+dm\log(n+1) \tag{155}$$

quantum bits, which, by Eqs. (151)–(154), yields our desired estimate. □

This immediately leads to the result that we wanted:

Theorem 10.4: *For any ensemble* $\mathcal{E}=(\Omega,P,\varphi)$,

$$Q^*(R)=M(\mathcal{E},R).$$

Proof: That $M(\mathcal{E},R)$ is a lower bound to Q^* is proved by Theorem 10.1. For its achievability choose $\epsilon>0$ and a Markov kernel p such that both $S(A:C)\leqslant R$ and $S(A:B|C)\leqslant M(\mathcal{E},R)+\epsilon$.

Choose now a partition $\mathcal{Z}$ according to Proposition 10.3, fixing m. Now choose δ large enough, so that according to that proposition a code exists which has fidelity $1-\epsilon$ on a state set of probability $1-\epsilon$, i.e., it has average fidelity $1-2\epsilon$ on the ensemble. By the proposition it has cbit rate $S(A:C)+o(1)$ and qubit rate

$$S(A:B|C)+2\eta(\epsilon)+o(1)\leqslant M(\mathcal{E},R)+2\eta(\epsilon)+\epsilon+o(1), \tag{156}$$

as $n\to\infty$. As ϵ was arbitrary, our claim is proved. □

C. On the AVS in the infinite setting

With the help of the above Proposition 10.3 the case of an arbitarily varying source of an *infinite* ensemble is dealt with easily, in much the same way as we did in the finite case (see Sec. VI):

Formally, of course, an arbitrarily varying source is a triple $(\Omega,\mathbf{P},\varphi)$, where Ω and φ are a measurable space and a measurable map into states, as before, and $\mathbf{P}$ is a set of probability distributions on Ω.

With the definitions of encoding and decoding from Sec. X A we require

$$\forall P^n\in\mathbf{P}^n\int_{\Omega^n}P^{\otimes n}(\mathrm{d}\omega_1\cdots\omega_n)F(|\varphi_\omega\rangle\langle\varphi_\omega|,(D\circ E)(\omega))\geqslant 1-\epsilon. \tag{157}$$

Denoting the trade-off function as $Q^*(R,\mathbf{P})$, we obtain the expected result:

Theorem 10.5: $Q^*(R,\mathbf{P})=M(\mathbf{P},R)$, *with*

$$M(\mathbf{P},R)=\sup_{P\in\mathbf{Q}}M(P,R),$$

where $\mathbf{Q}=\mathrm{conv}(\mathbf{P})$ *is the convex hull of* $\mathbf{P}$.

Proof: The inequality "$\geqslant$" is obvious, like in the finite case: the adversary can certainly always mock up an i.i.d. source $P\in\mathbf{Q}$, hence Theorem 10.1 applies.

For the opposite inequality, we start by choosing an $\epsilon>0$ and a partition $\mathcal{Z}$ according to Proposition 10.3. Every distribution P in $\mathbf{P}$ gives rise to a distribution $\hat{P}\in\mathcal{P}_{m-1}$, and we denote

$$\hat{\mathbf{P}}:=\{\hat{P}:P\in\mathbf{P}\}. \tag{158}$$

Note that, because the map $P\mapsto\hat{P}$ is affine linear, we get $\hat{\mathbf{Q}}=\mathrm{conv}(\hat{\mathbf{P}})$.

Now for $\delta>0$ we introduce again the set

$$\mathcal{T}:=\bigcup_{\hat{P}\in\hat{\mathbf{Q}}}\mathcal{T}_{\hat{P},\delta}, \tag{159}$$

and it is easy to see [compare Eq. (141)] that

$$\mathcal{T}^{\mathcal{Z}} := \bigcup_{I \in \mathcal{T}} Z_{i_1} \times \cdots \times Z_{i_n} \tag{160}$$

carries $1-\delta^{-2}$ of the probability of every $P^n \in \mathbf{P}^n$. On the other hand, because $\mathcal{T}$ is a union of type classes, we can find "few" $\hat{P}_1,\ldots,\hat{P}_T$, $T \leqslant (n+1)^m$ such that the corresponding $\mathcal{T}_{\hat{P}_t,\delta}$ cover $\mathcal{T}$. The coding is very simple: on seeing a state $\varphi_{\omega_1\ldots\omega_n}$ the encoder finds the index I of the piece Z_I in the partition $\mathcal{Z}^n$ such that $\omega_1\cdots\omega_n \in Z_I$, and the type of I. If $I \in \mathcal{T}$, he looks up t such that $I \in \mathcal{T}_{\hat{P}_t,\delta}$ and uses the coding scheme of Proposition 10.3 for $\hat{P}_t$. (Note that he needs not even send the type of I as that is part of the protocol of Proposition 10.3.) Choosing δ large enough this recipe gives a code with high fidelity for every $P^n \in \mathbf{P}^n$; by construction and Proposition 10.3, it has rates of $R+o(1)$ cbits and $M(\mathbf{P},R)+f(\epsilon)+o(1)$ qubits, with a function $f(\epsilon)$ that tends to 0 as $\epsilon \to 0$. □

To end this discussion, we would like to point out that a similar treatment of remote state preparation can be done: in fact, as we discussed in Sec. VIII, we always use the "1 ebit +1 cbit per qubit" technique (Theorem 8.1) on top of an efficient trade-off coding. To do this for an infinite ensemble one only has to understand that the bound of Theorem 8.1 is strong enough to allow approximation of the set of projected (compressed) product states $\varphi_{\omega_1}\otimes\cdots\otimes\varphi_{\omega_n}$, at negligible additional classical cost.

XI. DISCUSSION AND CONCLUSIONS

Our main result is a simple formula for the trade-off between quantum and classical resources in visible compression. The formula expresses the trade-off curve $Q^*(R)$ in terms of a single-letter optimization over conditional probability distributions of bounded size. This unexpectedly simple resolution places optimal trade-off coding into a small but growing class of problems in quantum information theory whose answers are not only known in principle but can be calculated in practice. (Another notable recent addition is the entanglement-assisted capacity of a quantum channel.[22])

At a conceptual level, for any given ensemble $\mathcal{E}$ of quantum states, $Q^*(R)$ can be thought of as a quantitative description of how "classical" the ensemble is. Any deviation from classicality is captured in the trade-off curve in the form of inefficiency of the classical storage. The amount of information that can be extracted from many copies of $\mathcal{E}$ while causing negligible disturbance, for example, can be read directly off the curve by identifying the point at which classical resources begin to become inefficient as compared to quantum. Much more subtle indicators of classicality are also available in $Q^*(R)$, however. We saw, for instance, that for the parametrized BB84 ensemble, $Q^*(R)$ had a kink at the point corresponding to partitioning the ensemble into nearly orthogonal subensembles.

Going beyond the compression of ensembles, we saw that it is possible to formulate a version of our main result in the setting of arbitrarily varying sources, corresponding to the situation in which the encoder and decoder have only partial or even no knowledge of the distribution of input states. Despite this handicap, compression is frequently still possible and we once again find that the trade-off curve can be calculated via a tractable optimization problem. For ensembles with symmetry, the problem can even often be reduced to calculating $Q^*(R)$ for one particular ensemble. Thus, for any given set of pure states, including the whole manifold of states on a given Hilbert space, these tools allow us to calculate the rate of exchange from qubit storage to classical storage. The answer is given, of course, not in terms of a single number but as the trade-off curve. (Like in any market, the going rate depends on supply.)

Our view that $Q^*(R)$ encodes the balance of quantum and classical information in a given ensemble or set of states is further bolstered by the role it was found to play in optimal remote state preparation. In this context, the minimal amount of classical communication required for any given rate of entanglement consumption can, once again, be read directly off the quantum-classical trade-off curve. That the comparatively exotic process of remote state preparation should reduce, via Theorem 8.1, to visible compression is a tremendous simplification.

Of course, while we have seen that the results of this article resolve some basic questions about trading different types of resources in quantum information, most related questions remain open. To begin, it is possible to trade entanglement, quantum communication and classical communication all together in a generalized type of remote state preparation. Since our results here describe the two extremes when first entanglement and then quantum communication are not permitted, it seems likely that similar techniques could resolve the full trade-off surface. More ambitiously, one could define channel capacities for noisy quantum channels that interpolate between the fully quantum and classical capacities by studying the usefulness of a channel for simultaneously sending quantum and classical information. The problem analogous to the trade-off question studied here would be to determine the achievable *region* of quantum-classical rate pairs. Unfortunately, given that neither the fully classical nor fully quantum extremes are fully understood, it may be a long time before we develop tools capable of analyzing that problem.

Therefore, to end, we offer two related open problems that are perhaps closer to the realm of the tractable. First, it would be useful to have a set of rules for extracting qualitative features of the trade-off curve, such as the location of any kinks and perhaps more detailed differentiability properties, from the structure of the input states (or ensemble). Second, it would be an interesting challenge to apply the observations of Sec. IX on symmetry to the explicit calculation of the trade-off curve for particular examples and, more generally, to find other approaches to simplifying these calculations.

ACKNOWLEDGMENTS

We thank Charles H. Bennett, David P. DiVincenzo, Daniel Gottesman, Debbie W. Leung, Michael A. Nielsen, Dénes Petz, Peter W. Shor and John. A. Smolin for enlightening discussions and helpful suggestions. P.H. was supported by US National Science Foundation Grant No. EIA-0086038 and a Sherman Fairchild Fellowship. R.J. and A.W. are supported by the U.K. Engineering and Physical Sciences Research Council.

APPENDIX: PROOFS OF AUXILIARY PROPOSITIONS

1. Proof of Proposition 3.3

Proof: Suppose the classical register C decomposes into parts C_1 and C_2 with corresponding joint density operator

$$\rho^{ABC_1C_2}=\sum_i p_i|i\rangle\langle i|^A\otimes|\varphi_i\rangle\langle\varphi_i|^B\otimes\sum_{j,k} p(i|j,k)|j\rangle\langle j|^{C_1}\otimes|k\rangle\langle k|^{C_2}. \tag{A1}$$

If we define the conditional ensembles $\mathcal{E}_{jk}$ and $\mathcal{E}_j$, then

$$S(A:B|C_1C_2)=\sum_{jk} q_{jk}S(\mathcal{E}_{jk})\leqslant S(A:B|C_1)=\sum_j q_jS(\mathcal{E}_j) \tag{A2}$$

by the concavity of the von Neumann entropy.

Therefore, for any map with $S(A:C_1)<R\leqslant H(p)$, we can always adjoin a second classical register C_2 such that $S(A:C_1C_2)=R$ without increasing the conditional mutual information. □

2. Proof of Proposition 3.4

Proof: W.l.o.g. let $i\in\{1,\ldots,m\}$. The information quantities in the definition of M can be reexpressed as follows:

$$S(A:B|C)=\sum_j q_jS\left(\sum_i q(i|j)|\varphi_i\rangle\langle\varphi_i|\right), \tag{A3}$$

$$S(A{:}C)=H(p)-\sum_j q_j H(q(\cdot|j)), \tag{A4}$$

with $q_j=\Sigma_i p_i p(j|i)$ and $q_j q(i|j)=p_i p(j|i)$. We read q as a probability distribution on the set $\mathcal{P}_m$ of all probability distributions on $\{1,\ldots,m\}$. Thus the minimization problem in the definition of M can be expressed as finding the infimum of $\Sigma_j q_j S(f(q(\cdot|j)))$ over the set

$$\mathcal{P}(p,R)=\left\{q\,\text{p.d. on } \mathcal{P}_m:\sum_j q_j q(\cdot|j)=p,\sum_j q_j H(q(\cdot|j))\geqslant H(p)-R\right\},$$

where f is an affine linear function on probability distributions, mapping the distribution p to the quantum state $\Sigma_i p_i|\varphi_i\rangle\langle\varphi_i|$.

Now we argue structurally: the set $\mathcal{P}(p,R)$ is convex (as a subset of an infinite dimensional probability simplex with additional linear inequality constraints), and the aim function is linear. Hence the infimum is an infimum over the extreme points of $\mathcal{P}(p,R)$, which are, by Caratheodory's theorem, distributions q with support at most $m+1$, the number of inequalities that define $\mathcal{P}(p,R)\subset\mathcal{P}(\mathcal{P}_m)$ (see, e.g., Ref. 38). In Sec. IX, Proposition 9.3 and Appendix, Sec. 6, we provide a detailed exposition of a more general form of this result. □

3. Proof of Proposition 5.1

Proof: The "$\leqslant$" inequality follows directly by forming the tensor product of two encodings for $\mathcal{E}_1$ and $\mathcal{E}_2$ with classical rates R_1 and R_2 respectively.

The "$\geqslant$" inequality is shown by choosing an encoding for the tensor product with classical rate R and then using the chain rule several times for subdivisions $A=A_1A_2$ and $B=B_1B_2$ as follows. First observe that

$$R\geqslant S(A_1A_2{:}C)=S(A_1{:}C)+S(A_2{:}C|A_1)=:R_1+R_2 \tag{A5}$$

and then

$$\begin{aligned}S(A_1A_2{:}B_1B_2|C)&=S(A_1{:}B_1B_2|C)+S(A_2{:}B_1B_2|C,A_1)\\&\geqslant S(A_1{:}B_1|C)+S(A_2{:}B_2|C,A_1)\\&\geqslant M(\mathcal{E}_1,R_1)+\inf\{S(A_2{:}B_2|C,A_1):S(A_2{:}C|A_1)\leqslant R_2\}\\&\geqslant M(\mathcal{E}_1,R_1)+M(\mathcal{E}_2,R_2)\\&\geqslant \min\{M(\mathcal{E}_1,R_1)+M(\mathcal{E}_2,R_2):R_1+R_2=R\}.\end{aligned} \tag{A6}$$

The second last line is seen as follows: in the line above it, the two mutual informations are conditional on A_1, so they both can be written as averages over the values of A_1. Hence the inequality follows by the convexity of M in R. □

4. Proof of Proposition 5.2

Proof: It is sufficient to verify that any encoding operator

$$\rho^{ABC}=\sum_{ik} p_i a_k |i\rangle\langle i|^A\otimes|k\rangle\langle k|^A\otimes U_k|\varphi_i\rangle\langle\varphi_i|U_k^{\dagger B}\otimes\sum_j p(j|i,k)|j\rangle\langle j|^C \tag{A7}$$

for $\mathcal{F}$ gives rise to a valid encoding operator

$$\sigma^{ABC}=\sum_i p_i|i\rangle\langle i|^A\otimes|\varphi_i\rangle\langle\varphi_i|^B\otimes\sum_{jk} p(j|i,k)a_k|j\rangle\langle j|^C\otimes|k\rangle\langle k|^C \tag{A8}$$

for $\mathcal{E}$ satisfying $S_\sigma(A:B|C) \leqslant S_\rho(A:B|C)$ and $S_\sigma(A:C) \leqslant S_\rho(A:C)$. □

5. Proof of Proposition 7.2

Proof: We will first prove the proposition for irreducible $\mathcal{E}$. Using a trick introduced by Holevo,[14] we can reduce the problem further to the case of a two-state ensemble: for an ensemble $\{\rho_i^B \otimes \sigma_i^C, p_i\}$ of states (we assume that all $p_i > 0$) and two specific indices k and l, define a new index

$$j(i) := \begin{cases} i & i \neq k,l, \\ * & i \in \{k,l\}. \end{cases} \tag{A9}$$

(Of course, in the case we have in mind, the ρ_i are the pure states from the ensemble $\mathcal{E}$, and the σ_i are commuting mixed states representing the classical information.) Then consider the multipartite state

$$\Omega = \sum_i p_i |i\rangle\langle i|^{A_1} \otimes |j(i)\rangle\langle j(i)|^{A_2} \otimes \rho_i^B \otimes \sigma_i^C.$$

The definition of $j(i)$ and the familiar chain rule imply

$$S(A_1:BC) = S(A_1A_2:BC) = S(A_2:BC) + S(A_1:BC|A_2). \tag{A10}$$

Note that the second term is an average over the values of $j(i)$ of Holevo quantities for the corresponding reduced ensembles. Therefore, it has only one nonzero contribution, which is

$$S(A_1:BC|A_2) = (p_k + p_l)\chi(\{\rho_i \otimes \sigma_i, p_i/(p_k+p_l)\}_{i=k,l}). \tag{A11}$$

Then, using Eq. (A10) and monotonicity of χ under partial trace repeatedly,

$$\begin{aligned} \chi(\{p_i, \rho_i \otimes \sigma_i\}) = S(A_1:BC) = S(A_2:BC) + S(A_1:BC|A_2) \\ \geqslant S(A_2:B) + (p_k+p_l)\chi(\{\rho_i \otimes \sigma_i, p_i/(p_k+p_l)\}_{i=k,l}) \\ \geqslant S(A_2:B) + (p_k+p_l)\chi(\{\rho_i, p_i/(p_k+p_l)\}_{i=k,l}) \\ = S(A_2:B) + S(A_1:B|A_2) = S(A_1:B) = \chi(\{\rho_i, p_i\}). \end{aligned}$$

Assuming that the first and the last Holevo quantities have the same value, we must have equality in the third line, implying

$$\chi(\{\rho_i \otimes \sigma_i, q_i\}_{i=k,l}) = \chi(\{\rho_i, q_i\}_{i=k,l}), \tag{A12}$$

with $q_i = p_i/(p_k+p_l)$. Then, applying the general formula

$$\chi(\{\omega_i, p_i\}) = \sum_i p_i D(\omega_i \| \omega) \tag{A13}$$

to Eq. (A12), with $\omega = \Sigma_i p_i \omega_i$ and D the relative entropy function, and using the Lindblad monotonicity once more yields

$$D(\rho_k \otimes \sigma_k \| q_k \rho_k \otimes \sigma_k + q_l \rho_l \otimes \sigma_l) = D(\rho_k \| q_k \rho_k + q_l \rho_l). \tag{A14}$$

(And likewise for l.)

With this we are almost done: invoking a result of Ohya and Petz (see Ref. 37, Theorem 9.12) we conclude that there exists a CPTP map R such that

$$R(\rho_k)=\rho_k\otimes\sigma_k, \tag{A15}$$

$$R(q_k\rho_k+q_l\rho_l)=q_k\rho_k\otimes\sigma_k+q_l\rho_l\otimes\sigma_l, \tag{A16}$$

from which it follows by linearity that

$$R(\rho_l)=\rho_l\otimes\sigma_l. \tag{A17}$$

Since CPTP maps (R and Tr_C) cannot decrease fidelity we thus must have $\rho_k\perp\rho_l$ or $\sigma_k=\sigma_l$.

In the particular case that the initial ensemble is irreducible we conclude that all σ_i must be equal, or else the partial trace over C strictly decreases the Holevo quantity. If the ensemble $\mathcal{E}$ is not irreducible, a simple variation on the previous argument shows that, for each of the irreducible subensembles $\mathcal{E}_l$, $\chi(\mathcal{E}_l)$ must be equal to χ of the corresponding subensemble $\{\varphi_{il}\otimes\sigma_{il},p_{i|l}\}$ of $\mathcal{F}^{BC}$. Applying our conclusions to these subensembles finishes the proof of the proposition. □

6. Proof of Proposition 9.3

Proof: As explained earlier in the proof of Proposition 3.4, any classical encoding map can be viewed as a probability distribution q on the set $\mathcal{P}_{\mathcal{I}}$ of probability distributions on $\mathcal{I}$ with barycenter p: $p=\Sigma_j q_j q(\cdot|j)$.

Covariance of the encoding means invariance of q under the natural action of G on $\mathcal{P}_{\mathcal{I}}$, i.e., $g:p\mapsto p^g$. Hence for each distribution p in the support of q we must have all the p^g in the support as well. On the other hand, we need far fewer conditions to obey, as it will turn out:

Assume that the covariant encoding is given by the distributions

$$(q(\cdot|j))^g \text{ with probability } \frac{1}{|G|}q_j,\ \ g\in G, j=1,\ldots.$$

Now choose representatives $i_1,\ldots,i_t$ of the orbits, and observe that (by G-invariance)

$$\sum_{j,g}\frac{1}{|G|}q_j(q(\cdot|j))^g=p \tag{A18}$$

if and only if

$$\forall\tau=1,\ldots,t\quad \sum_{j,g}\frac{1}{|G|}q_j q(g^{-1}i_\tau|j)=p(i_\tau). \tag{A19}$$

Similarly, $S(A:C)\leqslant R$ if and only if

$$\sum_j q_j H(q(\cdot|j))\geqslant H(p)-R, \tag{A20}$$

and, finally, our aim function reads

$$S(A:B|C)=\sum_{j,g}\frac{1}{|G|}q_j S\left(\sum_i q(i|j)|\varphi_{gi}\rangle\langle\varphi_{gi}|\right). \tag{A21}$$

Now consider the affine linear map from $\mathcal{P}_{\mathcal{I}}$ to $\mathbb{R}^{t+1}$ defined by

$$A:p\mapsto\left(H(p);\frac{1}{|G|}\sum_g p(g^{-1}i_\tau):\tau=1,\ldots,t\right). \tag{A22}$$

Note that the image of this map is in a certain t-dimensional subspace because, if $t-1$ of the conditions (A19) are satisfied, then the tth is also, automatically. Equations (A19) and (A20) are really conditions on the q_j-weighted average of the images $A_j = A(q(\cdot|j))$, $A = \Sigma_j q_j A_j$. By Caratheodory's theorem[38] the same average can be obtained by convex combination of $t+1$ of these, i.e., by a distribution q' on the j's with support containing at most $t+1$ points. In fact, q is easily seen to be expressible as a convex combination of such small support distributions, say $q'^{(a)}$ with weights λ_a.

To conclude, we observe that our aim function in Eq. (A21) is *linear* in the distribution q: hence, it is the λ_a–weighted sum of similar such expressions with $q'^{(a)}$ in place of q. For one value of a at least this is smaller than $S(A:B|C)$, the corresponding $q'^{(a)}$ satisfies $\Sigma_j q'^{(a)} A_j = A$, and hence Eqs. (A19) and (A20). As explained in the remark preceding the statement of Proposition 9.3, to obtain a G–covariant encoding we can split up each $q(\cdot|j)$ (with j in the support of $q'^{(a)}$) into the G translated distributions $(q(\cdot|j))^g$, proving the claim. □

[1] C. H. Bennett and G. Brassard, ''Quantum Cryptography: Public key distribution and coin tossing,'' in *Proceedings of IEEE International Conference on Computers, Systems and Signal Processing* (IEEE, New York, 1984), pp. 175–179.

[2] H. Lo, J. Phys. A **34**, 6957 (2001).

[3] D. Mayers, J. ACM **48**, 351 (2001).

[4] D. Deutsch and R. Jozsa, Proc. R. Soc. London, Ser. A **439**, 553 (1992).

[5] P. Shor, ''Algorithms for quantum computation: discrete logarithms and factoring (Extended abstract),'' in *Proceedings of 35th Annual Symposium on the Foundations of Computer Science* (IEEE Computer Society, Los Alamitos, 1994). Full version in SIAM J. Comput. **26**, 1484 (1997).

[6] D. Simon, ''On the power of quantum computation (Extended abstract),'' in *Proceedings of 35th Annual Symposium on the Foundations of Computer Science* (IEEE Computer Society, Los Alamitos, 1994). Full version in SIAM J. Comput. **26**, 1474 (1997).

[7] A. S. Holevo, IEEE Trans. Inf. Theory **44**, 269 (1998).

[8] B. Schumacher and M. D. Westmoreland, Phys. Rev. A **56**, 131 (1997).

[9] C. H. Bennett, H. J. Bernstein, S. Popescu, and B. Schumacher, Phys. Rev. A **53**, 2046 (1996).

[10] H. Barnum, C. A. Fuchs, R. Jozsa, and B. Schumacher, Phys. Rev. A **54**, 4707 (1996).

[11] R. Jozsa and B. Schumacher, J. Mod. Opt. **41**, 2343 (1994).

[12] B. Schumacher, Phys. Rev. A **51**, 2738 (1995).

[13] H. Barnum, P. Hayden, R. Jozsa, and A. Winter, Proc. R. Soc. London, Ser. A **457**(2012), 2019 (2001).

[14] A. S. Holevo, *Problemy Peredaći Informacii*, 9(3)3-11 (1973). English translation: A. S. Holevo, Probl. Inf. Transm. **9**, 177 (1973).

[15] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948); **27**, 623 (1948).

[16] H.-K. Lo, Phys. Rev. A **62**, 012313 (2000).

[17] C. H. Bennett, D. P. DiVincenzo, P. W. Shor, J. A. Smolin, B. M. Terhal, and W. K. Wootters, Phys. Rev. Lett. **87**, 077902 (2001).

[18] W. K. Wootters and W. H. Zurek, Nature (London) **299**, 802 (1982).

[19] M. Koashi and N. Imoto, Phys. Rev. Lett. **87**, 017902 (2001). Based on: ''What is possible without disturbing partially known quantum states?'' quant-ph/0101144 2001.

[20] M. Fannes, Commun. Math. Phys. **31**, 291 (1973).

[21] G. Kuperberg, ''The capacity of hybrid quantum memory,'' quant-ph/0203105 2002.

[22] C. H. Bennett, P. W. Shor, J. A. Smolin, and A. V. Thapliyal, Phys. Rev. Lett. **83**, 3081 (1999); ''Entanglement-assisted capacity of a quantum channel and the reverse Shannon theoreni,'' quant-ph/0106052 2001.

[23] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems* (Academic, London, 1981).

[24] J. Wolfowitz, *Coding Theorems of Information Theory*, 2nd ed. (Springer-Verlag, Berlin, 1964).

[25] A. Winter, ''Coding Theorems of Quantum Information Theory,'' Ph.D. thesis, Universität Bielefeld, 1999 (unpublished); electronically, http://archiv.ub.uni-bielefeld.de/disshabi/mathe.htm; as e-print, quant-ph/9907077.

[26] R. Ahlswede, Z. Wahrscheinlichkeitstheor. Verwandte Geb. **44**, 159 (1978).

[27] R. Jozsa and A. Winter, ''Compression of sources of probability distributions and density operators,'' in preparation (2002).

[28] H. Chernoff, Ann. Math. Stat. **23**, 493 (1952).

[29] W. Hoeffding, J. Am. Stat. Assoc. **58**, 13 (1963).

[30] I. Devetak and T. Berger, Phys. Rev. Lett. **87**, 197901 (2001).

[31] R. Ahlswede, J. Combin. Inform. System Sci. **4**, 176 (1979).

[32] C. H. Bennett, G. Brassard, C. Crépeau, R. Jozsa, A. Peres, and W. K. Wootters, Phys. Rev. Lett. **70**, 1895 (1993).

[33] C. H. Bennett, P. Hayden, D. W. Leung, P. W. Shor, and A. Winter, ''Remote state preparation,'' in preparation (2002).

[34] P. R. Halmos, *Measure Theory* (Van Nostrand, New York, 1950).

[35] W. Feller, *An Introduction to Probability Theory and Its Applications, Vol. I*, 3rd ed., *Vol. II*, 2nd ed. (Wiley, New York, 1968, 1971).

[36] R. M. Gray, *Entropy and Information Theory* (Springer-Verlag, New York, 1990).

[37] M. Ohya and D. Petz, *Quantum Entropy and Its Use* (Springer-Verlag, Berlin, 1993).

[38] G. M. Ziegler, *Lectures on Polytopes* (Springer-Verlag, New York, 1995).