On quantum coding for ensembles of mixed states

Howard Barnum[1]**, Carlton M Caves**[2]**, Christopher A Fuchs**[3]**,
Richard Jozsa**[1] **and Benjamin Schumacher**[4]

[1] Department of Computer Science, University of Bristol, Merchant Venturers Building, Bristol
BS8 1UB, UK
[2] Department of Physics and Astronomy, University of New Mexico, Albuquerque, NM 87131,
USA
[3] Los Alamos National Laboratory, MS-B285, Los Alamos, NM 87545, USA
[4] Department of Physics, Kenyon College, Gambier, OH 43022, USA

Abstract
We consider the problem of optimal asymptotically faithful compression for
ensembles of mixed quantum states. Although the optimal rate is unknown,
we prove upper and lower bounds and describe a series of illustrative examples
of compression of mixed states. We also discuss a classical analogue of the
problem.

PACS numbers: 03.67.−a, 03.65.Ta

1. Introduction

The emergence of potentially useful theoretical protocols for using quantum states in
cryptography and quantum computation has increased the theoretical (and perhaps ultimately
practical) importance of questions about how quantum states can be compressed, transmitted
across noisy or low-dimensional channels, and recovered, and otherwise manipulated in a
fashion analogous to classical information. Most of the work done on these matters, beginning
with [1], has focused on the manipulation of pure states, with mixed states appearing only in
intermediate stages, as the result of noise. An exception is [2], which considered the copying
or broadcasting of mixed states. When mixed states have appeared as states to be transmitted,
it has usually been required that their potential entanglement with some reference system
be preserved, as in [3]. This focuses attention again on a *pure* state, the entangled state of
system and reference system. As discussed further in [4] there is a close relation between
entanglement transmission and the transmission of pure states of the system itself.

In the present paper, we consider the compression or transmission of mixed states, *without*
any requirement that their entanglement or correlation with other systems be preserved. There
might seem to be good reason to confine oneself to pure-state transmission, since mixed
states, considered apart from any potential entanglement with other systems, might not seem

particularly useful. This may be why the classical analogue of the problem we consider in this paper—the transmission of probability distributions—has not, to our knowledge, been previously studied. Game theory is perhaps the first situation that springs to mind in which one might wish to produce a mixed state intentionally, given that all pure states of which it may be viewed as a mixture are available, since it is well known to game theorists that mixed strategies may be better than any of their component pure strategies in important situations [5, 6]. Thus a 'practical' application of mixed-state compression might be the compression of mixed strategies, where the 'decoding' is done by the player playing the strategy or someone who shares his goals. In cryptographic applications (closely related to game theory, of course) and also in probabilistic classical algorithms, there may be a use for randomness and an interest in compressing it for efficient storage or transmission. Indeed, quantum computation can enable more-efficient-than-classical sampling from probability distributions [7, 8]; there may be relations between these ideas and the work reported here.

The problem of optimal compression for ensembles of pure quantum states has been solved [1, 9, 10], but for sources of mixed states the minimal resources are unknown. This question has also been considered by Horodecki in [11, 12]. In this paper, we consider several variants of the question, depending on the fidelity criteria and encoding/decoding procedures used. Sections 2 and 3 present the problem, in variants depending on whether or not the encoder/compressor knows the identity of each state and can use it to help encode, and depending on whether, in a block-coding setting, a marginal ('local') or total ('global') fidelity criterion is used; section 4 considers relations between these variants of the problem, in general and for the special case of pure states. Section 5 discusses the fact that the entropy of a source ensemble's average density operator provides (as in the pure-state case) an upper bound on the rate at which qubits must be used to represent the source. We also show that under the global fidelity criterion, if decodings are required to be unitary, this is actually the optimal rate. Section 6 formulates a classical version of the problem, which we have not seen treated in classical information theory, and discusses examples. In section 7, we show with several examples that in contrast to the pure state case, it is possible with *general* decodings to compress to below the entropy of the average density operator. This section also introduces a useful preparation-visible technique, that of compression by purifications, which we show does better than our classical methods for some of the classical mixed-state compression problems considered in section 6. Finally, in section 8 we show that the Holevo quantity $S(\sum_i p_i \sigma_i) - \sum_i p_i S(\sigma_i)$ for an ensemble gives a lower bound on the qubit rate required to represent a source. (A different proof is given in [11].) We do not know whether this lower bound is attainable in general.

2. Formulation of the problem

In this paper, $S(\rho)$ will always denote the von Neumann entropy of a density matrix ρ and $H(p_1, \ldots, p_n)$ will denote the Shannon entropy of a probability distribution $p_1, \ldots, p_n$. In both cases logarithms are taken to base 2:

$$S(\rho) := -\operatorname{tr}(\rho \log_2 \rho)$$
$$H(p_1, \ldots, p_n) := -\sum_i p_i \log_2 p_i.$$

Let $\rho_1, \ldots, \rho_n$ be a list of (possibly mixed) d-dimensional quantum states. Each state is assigned a prior probability $p_1, \ldots, p_n$ respectively. We refer to such a list as a source or ensemble of signal states, denoted by $E = \{p_i, \rho_i\}$. Alice is fed an unending sequence of

these signal states, with each successive state chosen randomly and independently from E. At time N she will have the total state $\sigma_N = \rho_{i_1} \otimes \cdots \otimes \rho_{i_N}$ with probability $p_{i_1} p_{i_2} \cdots p_{i_N}$.

Alice wants to perform either of the following tasks (which are equivalent for our considerations):

- Communication: Alice wants to send the signals to Bob using a minimum number of qubits/signal so that Bob can reconstruct long sequences with 'arbitrarily high fidelity'. This involves a "coding procedure" for Alice and a 'decoding procedure' for Bob (cf. later discussion for the precise meaning of all these terms).
- Storage: alternatively, Alice wants to store the signals as efficiently as possible. In this interpretation the coding procedure is used for putting signals into storage, and the decoding procedure for reconstituting them.

We distinguish two fundamental situations for Alice:

- Preparation-blind (blind): Alice is not given the identity of the individual (generally nonorthogonal) signal states (she knows only their prior distribution).
- Preparation-visible (visible): Alice is given the identity of the individual signal states (as well as their prior distribution). Indeed, in this case we may assume that she is simply provided with a sequence of the *names* of the states and she may prepare the states herself if she wishes.

Note that in the blind case, Alice is being fed essentially quantum information, whereas in the visible case she is getting entirely classical information. In both cases, however, Bob on decoding is not required to identify the actual signal states, but only to produce high fidelity representatives of the correct sequence of states. Hence, even in the visible case, the problem is not one of classical coding/information theory. The visible case (for pure states) occurs, for example, in quantum cryptographic protocols (e.g. BB84 [13] and B92 [14]), where the sender (Alice) is also the state preparer.

3. Coding/decoding schemes and their fidelity

Let $\mathcal{H}_d$ denote the space of all d-dimensional states. Given any physical system in state ρ, quantum mechanics allows only the following three types of operations:

- (OP1) A unitary transformation, $\rho \to U\rho U^\dagger$ (U unitary).
- (OP2) Inclusion of an ancilla in a standard state ρ_0 (independent of ρ), $\rho \to \rho \otimes \rho_0$.
- (OP3) Discarding a subsystem (when ρ is a state of a composite system AB), $\rho_{AB} \to \text{tr}_B(\rho_{AB})$.

Note that (OP2) and (OP3) change the value of d.

Consider any length-N string of input states (given either visible or blind):

$$\sigma_N = \rho_{i_1} \otimes \cdots \otimes \rho_{i_N} \in \mathcal{H}_d^{\otimes N} \tag{1}$$

$$\text{Prob}(\sigma_N) = p_{i_1} \cdots p_{i_N}. \tag{2}$$

A coding/decoding scheme, using q qubits/signal, is defined by the following requirements, which are to be specified for all sufficiently large N.

- Blind coding: Alice's coding procedure, if blind, is any specified sequence of the above three operations applied to σ_N, giving a final state ω_N within the required resources of q qubits/signal (i.e. in 2^{qN} dimensions). Mathematically, any such sequence of operations corresponds to a completely positive, trace-preserving map $\mathcal{C}$ on the density operator, i.e. $\omega_N = \mathcal{C}(\rho_{i_1} \otimes \cdots \otimes \rho_{i_N})$, and any completely positive, trace-preserving map corresponds to such a sequence of operations.

- Visible coding: Alice's coding procedure, if visible, corresponds to an *arbitrary* assignment of a state $\omega_N \in \mathcal{H}_{2^{qN}}$ to each σ_N; i.e. Alice can build any state she pleases as the coded version of the input string.
- Finally, Bob's decoding (analogous to blind coding) is any sequence of the above three operations applied to the coded state, yielding a state $\tilde{\sigma}_N$ of N d-dimensional systems. Thus decoding is a completely positive, trace-preserving map from $\mathcal{H}_{2^{qN}}$ to $\mathcal{H}_{d^N}$.

Let us write Bob's decoded state, produced by coding followed by decoding of $\sigma_N = \rho_{i_1} \otimes \cdots \otimes \rho_{i_N}$, as $\tilde{\sigma}_N = \tilde{\sigma}_{i_1 \ldots i_N}$. Let

$$\tilde{\rho}_k = \left\{ \begin{array}{l} \text{trace of } \tilde{\sigma}_{i_1,\ldots,i_N} \text{ over all} \\ \text{signal spaces except the } k\text{th} \end{array} \right\} \qquad k = 1, \ldots, N \tag{3}$$

be the reduced state in the kth signal position after coding and decoding; i.e. $\tilde{\rho}_k$ is the decoded version of the kth transmitted state ρ_{i_k}. Let

$$F(\rho_1, \rho_2) = \left(\text{trace} \left(\rho_1^{1/2} \rho_2 \rho_1^{1/2} \right)^{1/2} \right)^2 \tag{4}$$

denote the Bures–Uhlmann fidelity function [15–17]. The coding/decoding scheme has fidelity $1 - \epsilon$ if it satisfies the following fidelity requirement: there is an N_0 such that for all $N > N_0$,

$$\sum_{\sigma_N} \text{Prob}(\sigma_N) \prod_{k=1}^{N} F(\rho_{i_k}, \tilde{\rho}_k) > 1 - \epsilon \qquad \text{(LOCAL-FID)}. \tag{5}$$

Note that high fidelity according to (LOCAL-FID) allows entanglement to be introduced between output signal states, even though there was no entanglement in the input (1), which was taken to be a product state. This is because we examine $\tilde{\sigma}_N$ only through its partial traces (3), thus reducing the state to each position separately. In view of this we might consider a stronger fidelity criterion (GLOBAL-FID), which replaces (LOCAL-FID) by

$$\sum_{\sigma_N} \text{Prob}(\sigma_N) F(\sigma_N, \tilde{\sigma}_N) > 1 - \epsilon \qquad \text{(GLOBAL-FID)}. \tag{6}$$

For ϵ tending to zero, this eliminates extraneous entanglements in the output sequence. Note that in a continuously varying situation with ϵ tending to zero, (GLOBAL-FID) implies (LOCAL-FID) because (GLOBAL-FID) will require that $\tilde{\sigma}_N$ become arbitrarily close to σ_N and, hence, $F(\rho_{i_k}, \tilde{\rho}_k) \to 1$ for each k, too.

Example 1. Alice wants to send Bob $\sigma_2 = \frac{1}{2}I \otimes \frac{1}{2}I$, so we may take the decoded state to be $\tilde{\sigma}_2 = \frac{1}{2}I \otimes \frac{1}{2}I$ satisfying (LOCAL-FID) and (GLOBAL-FID) with $\epsilon = 0$ or

$$\tilde{\sigma}_2 = \frac{1}{2}(|0\rangle \otimes |0\rangle + |1\rangle \otimes |1\rangle)(\langle 0| \otimes \langle 0| + \langle 1| \otimes \langle 1|)$$

satisfying (LOCAL-FID) with $\epsilon = 0$ but not (GLOBAL-FID).

We will generally adopt the fidelity requirement (GLOBAL-FID) in the following. If it is important that the signal states remain uncorrelated, (GLOBAL-FID) is the appropriate criterion; otherwise it may be too strong.

Remark 2. (LOCAL-FID) has the following awkward feature: If we have a very high fidelity coding/decoding scheme according to (LOCAL-FID) and we repeatedly apply it to a long string, $\sigma_N \to \tilde{\sigma}_N \to \tilde{\tilde{\sigma}}_N$, then we will not necessarily preserve high fidelity in the sequence of reduced states. This is because though σ_N and $\tilde{\sigma}_N$ have essentially the same reduced states at each position, globally they can be very different states (cf. example 1). Since the coding scheme is generally a *block*-coding scheme, it uses the global input state and will work well only if this global state is a *product* state as in (1). Hence $\tilde{\tilde{\sigma}}_N$ will not generally have the correct reduced states.

From the above precise formulations of the notions of coding, decoding and fidelity, we obtain a well defined mathematical problem.

Problem 3. *For a given source E, find the greatest lower bound $q_{\min}$ of all q's with the following property: For all $\epsilon > 0$ there exists a coding/decoding scheme based on q qubits/signal with fidelity $1 - \epsilon$.*

This problem may be considered either in the blind or visible context, with the variation over encodings taken over the appropriate class of maps in each case. Similarly, it may be considered in the case of either of the fidelity criteria, (LOCAL-FID) or (GLOBAL-FID). We will say that the source E can be coded (or compressed) at the rate $q_{\min}$.

Equivalently, the problem may be stated as follows: for a given source E, find $q_{\min}$ with the following property. Given any $\delta > 0$, (a) if $q_{\min} + \delta$ qubits/signal are available, then for every $\epsilon > 0$ there exists a coding scheme with fidelity $1 - \epsilon$, and (b) if $q_{\min} - \delta$ qubits/signal are available, then there exists an $\epsilon > 0$ such that every coding scheme will have fidelity less than $1 - \epsilon$.

4. Comparing the Formulation with Schumacher's coding for pure states

The problem formulated above is intended to be a generalization of the scenario in Schumacher's theorem [1, 18] to the case of mixed input states. Indeed, if the input states happen all to be pure states, then the above formulation reduces precisely to the situation of Schumacher's theorem. It is interesting to note that several of the distinctions made above collapse in the special case of *pure* input states.

Proposition 4. *If the input states are all pure, then there is no distinction between the blind and visible problems.*

Proof. In [18, 10] an optimal coding/decoding scheme for the visible pure-state problem is described. This optimal scheme turns out, remarkably, to be blind; i.e. knowledge of the identities of the individual input signals gives Alice no further benefit in the case of pure states. □

In [10] it is also shown that nonunitary decoding operations are of no advantage (on the criterion (GLOBAL-FID)) in decoding for the pure-state problem. In contrast, for mixed-state signals, nonunitary decodings are generally essential for optimal compression. This follows from theorem 7 and section 7 below.

Finally, the distinction between (LOCAL-FID) and (GLOBAL-FID) also collapses for pure signal states.

Proposition 5. *If the input states are all pure, then the two alternative fidelity criteria, (LOCAL-FID) and (GLOBAL-FID), become equivalent as ϵ is allowed to tend to zero.*

Idea of proof. We already know that the (GLOBAL-FID) criterion implies the (LOCAL-FID) criterion. Suppose that the (LOCAL-FID) criterion holds for a sequence of ϵ values tending to zero. (Here we are thinking of a sequence of coding/decoding schemes which all operate within the resource constraint of q qubits/signal where $q > q_{\min}$.) Then the reduced states $\tilde{\rho}_k$ of $\tilde{\sigma}_N$ become arbitrarily close to the input states ρ_{i_k} which are *pure*. Hence $\tilde{\sigma}_N$ cannot be much entangled since entanglement always shows up as *impurity* in the reduced states $\tilde{\rho}_k$. Thus $\tilde{\sigma}_N$ must approach the product state σ_N and (GLOBAL-FID) holds. □

As a consequence of proposition 5, the awkward feature of (LOCAL-FID) described in remark 2 does not arise in the coding of *pure* states.

5. The $S(\bar{\rho})$ upper bound for $q_{\min}$

Let $\bar{\rho}$ be the average density matrix of the input states:

$$\bar{\rho} = \sum_{i=1}^{n} p_i \rho_i. \tag{7}$$

Proposition 6. $S(\bar{\rho})$ *is an upper bound for* $q_{\min}$ *under the criterion (GLOBAL-FID) (and hence also under the criterion (LOCAL-FID)).*

Proof. For each ρ_i, choose a representative ensemble of pure states corresponding to ρ_i, so that we may view Alice as receiving an overall ensemble of pure states with density matrix $\bar{\rho}$. By Schumacher's theorem this may be transmitted to Bob with arbitrarily high fidelity by compressing to $S(\bar{\rho})$ qubits/signal. $\qquad\square$

Note that this compression preserves too much internal structure: Bob faithfully reconstructs Alice's chosen ensembles of pure states underlying the ρ_i's rather than just the ρ_i's themselves. For our purposes it is sufficient for Bob to decode to any *other* representative ensemble for the ρ_i's. Hence we would expect that further compression is possible, and the examples in section 7 below show that it generally is. Furthermore, the coding in proposition 6 gives high fidelity relative to the stronger criterion (GLOBAL-FID); using the weaker (LOCAL-FID), one might expect even more compression.

In fact we can say more, embodied in the following theorem.

Theorem 7. [9] [19]. *For the stronger fidelity criterion (GLOBAL-FID), if the decoding operation is required to be unitary (i.e. using only OP1 and OP2), then no further compression is possible, i.e.* $q_{\min} = S(\bar{\rho})$.

The proof is given in appendix A.

Note that for the pure-state coding theorem, the decoding may indeed be taken to be unitary and (GLOBAL-FID) is used (being equivalent to (LOCAL-FID) by proposition 5), but we do not necessarily wish to impose these conditions in the mixed-state case.

6. A classical analogue

In the case of Schumacher's pure-state coding theorem, there is a clear classical analogue, which has been well studied and completely solved, namely Shannon's noiseless coding theorem. Though the classical analogue for the case of mixed states appears not to have been studied, it would involve the compression/communication of probability distributions. To formulate the classical problem, let there be a finite number of possible classical states, i.e., distinguishable alternatives (this is the analogue of our assumption of finite-dimensional Hilbert spaces), and identify the input and output classical states with particular orthonormal bases in input and output Hilbert spaces. Write probability weight functions on the sets of orthonormal pure states as column vectors $\boldsymbol{p} = (p_1, \ldots, p_n)^{\mathrm{T}}$ of probabilities. These classical probability distributions then correspond to commuting density operators diagonal in the input and output bases.

We may formulate classical preparation-blind coding or decoding procedures as multiplication of input probability vectors by a stochastic matrix $\boldsymbol{A}$ (one with nonnegative entries whose columns sum to one):

$$\boldsymbol{p}_{\mathrm{out}} = \boldsymbol{A}\boldsymbol{p}_{\mathrm{in}}. \tag{8}$$

The stochasticity ensures that the matrix can be interpreted as a matrix of transition probabilities. As in the quantum case, preparation-visible procedures are described by arbitary maps between the relevant spaces, in this case between the spaces of probability vectors. The stochastic linear maps on the probability distributions correspond to a (convex) subset of the trace-preserving completely positive maps on density operators, and a given classical problem maps onto a corresponding quantum problem of sending commuting density operators. If we allow all possible trace-preserving completely positive maps, instead of just those which correspond to classical dynamics in the diagonalizing bases, we are using quantum means to deal with a classical problem, and we can compare the power of these quantum means to that of the purely classical means defined by restricting the allowable CP-maps to those that act as stochastic matrix multiplication in the given bases.

These notions and comparisons are illustrated in the following examples, which are phrased in terms of the quantum language, i.e., viewing classical distributions as commuting mixed states.

Example 8. We have two input states, $\rho_1, \rho_2 \in \mathcal{H}_2$,

$$\rho_1 \leftrightarrow \mathrm{diag}(\alpha_1, 1 - \alpha_1) \qquad\qquad \rho_2 \leftrightarrow \mathrm{diag}(\alpha_2, 1 - \alpha_2) \tag{9}$$

which are simultaneously diagonal in a basis known to Alice and Bob. Let the prior probabilities for these two states be p_1 and p_2. Classically we may regard the two states as suitably biased coins, C_1 and C_2. A preparer chooses a sequence of coins, C_1 or C_2 with probabilities p_1 and p_2, and tosses each of them once. The sequence of outcomes is passed on to Alice. Since Alice can look at the sequence of outcomes, we can regard the sequence of outcomes as the realization of 'Alice being given an unknown sequence of the two states'. Notice that in the blind case, Alice cannot be given the actual coins that make up the input sequence, for she could then toss each one many times and identify the coins in the sequence, which is impossible to given a single instance of each quantum state in the sequence. In contrast, in the visible case, Alice is given the sequence of coin names (or the actual coins, from which she could generate the sequence of coin names), together with a sequence of outcomes. In both cases, the objective of the protocol is to have Bob generate a sequence of outcomes that are governed by the same probabilities as Alice's input sequence of outcomes. Thus we have the following classical problems.

Blind case. A preparer chooses a sequence of coins, C_1 or C_2 with prior probabilities p_1 and p_2, tosses each of them a single time, and passes the sequence of outcomes on to Alice. Alice 'codes' her sequence of outcomes, and Bob 'decodes' the result, obtaining an output sequence of outcomes. The coding/decoding processes may involve probabilistic processes. As before, Alice would like to compress the input sequence as much as possible for transmission. A perfect coding/decoding scheme would achieve the following: Suppose that in position 1 the preparer has used coin C_2; then, taking into account the probability of outcomes in tossing C_2 and all probabilistic processes involved in coding/decoding, the first entry in Bob's outcome sequence should have a probability distribution which is the *same* as for coin C_2. A similar condition should apply at each position of the sequence.

This condition requires perfect fidelity of transmission of the distributions. In order to allow the usual situation of fidelity that approaches perfection only in an asymptotic limit of longer and longer block coding, we introduce a fidelity function for classical probability distributions. If $\boldsymbol{p} = (p_1, \ldots, p_n)^{\mathrm{T}}$ and $\boldsymbol{q} = (q_1, \ldots, q_n)^{\mathrm{T}}$ are two probability distributions on the same space, then the fidelity is defined by

$$F_{\mathrm{cl}}(\boldsymbol{p}, \boldsymbol{q}) = \left(\sum_{i=1}^{n} \sqrt{p_i} \sqrt{q_i} \right)^2 \tag{10}$$

which is also known as the Bhattacharyya–Wootters distance (or overlap) between the distributions. Notice that $F_{cl}(\boldsymbol{p}, \boldsymbol{q}) = 1$ iff $\boldsymbol{p} = \boldsymbol{q}$. The classical fidelity F_{cl} may be viewed as a special case of the Bures–Uhlmann fidelity (4), i.e., $F_{cl}(\boldsymbol{p}, \boldsymbol{q}) = F(\rho_1, \rho_2)$ for two commuting density operators, ρ_1 and ρ_2, that have $\boldsymbol{p}$ and $\boldsymbol{q}$ on their diagonals.

The problem is then to find the minimum number of bits/signal which suffices to code the input string with asymptotically arbitrarily high fidelity. (A precise formulation is very similar to that given for the quantum problem in section 2.) There is an obvious upper bound on the minimum number of bits/signal: Alice may compress her outcome sequence to the Shannon entropy of the average coin, $H(\bar{\alpha}, 1 - \bar{\alpha}) = S(\bar{\rho})$ bits/signal, where $\bar{\alpha} = p_1\alpha_1 + p_2\alpha_2$ is the average probability for the first outcome; Bob can decode the compressed sequence to produce an output outcome sequence that has asymptotically perfect fidelity. Because we are dealing here with commuting density operators, this upper bound is the same as the $S(\bar{\rho})$ upper bound of section 5.

Visible case. In this case Alice is fed the sequence of actual coin names, C_1 or C_2, in addition to outcomes of tossing each coin once. The blind-case upper bound of $H(\bar{\alpha}, 1 - \bar{\alpha}) = S(\bar{\rho})$ bits/signal applies also to the visible case, but there is an additional clear upper bound in the visible case: Alice may simply send Bob the full information of which coin to use at each stage; she can compress this data by the Shannon entropy of the prior distribution of coin choices, i.e. $H(p_1, p_2)$ bits/signal.

Although we do not know the optimal number of bits/signal for this problem, we now describe a purely classical coding/decoding scheme which beats both bounds for some values of the parameters p_1, p_2, α_1, and α_2.

Example 9. Suppose that $\alpha_2 \geqslant \alpha_1$. Denote the coin toss outcomes by H and T, with H having probability α_i for coin C_i. Alice sends one of three possible messages, M_0, M_1, or M_2, to Bob according to the following (probabilistic) coding scheme:

- Regardless of the input coin (C_1 or C_2), Alice sends M_0 with probability $1 - \alpha_2 + \alpha_1$.
- If the message M_0 is *not* chosen (i.e. with probability $\alpha_2 - \alpha_1$), Alice sends M_1 if the coin is C_1 and M_2 if the coin is C_2.

Bob responds to these signals as follows:

- For M_0 Bob probabilistically generates H or T with prob(H) $= \alpha_1/(1 - \alpha_2 + \alpha_1)$ and prob(T) $= (1 - \alpha_2)/(1 - \alpha_2 + \alpha_1)$.
- For M_1 Bob generates T with probability 1.
- For M_2 Bob generates H with probability 1.

Curiously, in the latter two cases Bob actually learns the identity of the coin yet he responds with a different distribution! It is readily verified that for each position in the sequence, taking into account the probabilistic choices in coding/decoding, Bob's output result correctly represents the result of one toss of the corresponding input coin.

The messages M_0, M_1, and M_2 are sent with probabilities $1 - \alpha_2 + \alpha_1$, $p_1(\alpha_2 - \alpha_1)$, and $p_2(\alpha_2 - \alpha_1)$, so Alice can compress the sequence to

$$\Xi = H(1 - \alpha_2 + \alpha_1, p_1(\alpha_2 - \alpha_1), p_2(\alpha_2 - \alpha_1)) \text{ bits/coin toss.}$$

If $p_1 = p_2 = \frac{1}{2}$ and $\alpha_1 \approx \alpha_2 \approx \frac{1}{4}$, then $H(p_1, p_2) = 1$ and $S(\bar{\rho}) = H(\bar{\alpha}, 1 - \bar{\alpha}) \approx H(\frac{3}{4}, \frac{1}{4})$, whereas $\Xi \approx 0$, thus beating the two bounds. (For some other values of the parameters Ξ exceeds both bounds.)

It has been conjectured that the minimum number of bits/signal for this classical problem (and its natural generalization to many classical distributions) should be the *mutual information*

$H(\bar{\alpha}, 1 - \bar{\alpha}) - p_1 H(\alpha_1, 1 - \alpha_1) - p_2 H(\alpha_2, 1 - \alpha_2)$, even if global fidelity is required, but this has resisted proof/disproof so far. (This would coincide with the lower bound given in section 8.) In example 12 below we will describe a quantum protocol for this problem which is better than all the above protocols.

7. Examples of compression beyond $S(\bar{\rho})$

We now return to our main question of quantum coding for general sources of mixed states. Though the problem of the optimal value of q_{min} remains unsolved, we describe here a series of interesting examples of compression beyond the $S(\bar{\rho})$ upper bound given in section 5. These examples reveal something of the intricacy of this problem. (Notice that example 9 already provides a case of compression beyond the $S(\bar{\rho})$ bound in the classical context.) In the next section we will derive a lower bound for q_{min}.

Example 10. (Trivial cases). The following two situations are blind, but Alice may reliably identify the input states, thus making them visible.
(a) Suppose that there is only one possible input signal $\rho = \frac{1}{2}I$ so $\bar{\rho} = \frac{1}{2}I$ and $S(\bar{\rho}) = 1$ qubit/signal. Yet Alice need not send anything at all; i.e., we may compress to 0 qubits/signal.
(b) Suppose that the input signals ρ_i with prior probabilities p_i are supported on *orthogonal* subspaces. (The support of a mixed state is defined as the subspace spanned by all eigenvectors belonging to *nonzero* eigenvalues.) Thus Alice may reliably measure the identity of the inputs and compress the resulting data to $H(p_1, \ldots, p_n)$ qubits/signal. Now for orthogonally supported states we have generally

$$S(\bar{\rho}) = H(p_1, \ldots, p_n) + \sum p_i S(\rho_i) \geqslant H(p_1, \ldots, p_n). \tag{11}$$

Example 11. (A nontrivial blind example with noncommuting mixed input states). There are two signal states, ρ_1 and ρ_2, in $m + n$ dimensions with prior probabilities p_1 and p_2. The states have a block-diagonal form,

$$\rho_1 = \text{diag}(\epsilon\sigma_1, (1 - \epsilon)\tau_1), \qquad \rho_2 = \text{diag}(\epsilon\sigma_2, (1 - \epsilon)\tau_2)$$

where σ_1 and σ_2 are density matrices of size $m \times m$ and τ_1 and τ_2 are density matrices of size $n \times n$. Writing

$$\bar{\rho} = p_1\rho_1 + p_2\rho_2, \qquad \bar{\sigma} = p_1\sigma_1 + p_2\sigma_2, \qquad \bar{\tau} = p_1\tau_1 + p_2\tau_2$$

ones easily sees that

$$S(\bar{\rho}) = H(\epsilon, 1 - \epsilon) + \epsilon S(\bar{\sigma}) + (1 - \epsilon)S(\bar{\tau}). \tag{12}$$

In the $S(\bar{\rho})$ coding scheme of proposition 6, we may interpret this formula as follows. For a sequence of inputs Alice first measures the σ-space versus the τ-space—projecting the input state into whichever space is the outcome—and she compresses the resulting string of subspace names to $H(\epsilon, 1 - \epsilon)$ bits/name. If the outcome space was 'σ-subspace', a result that occurs a fraction ϵ of the time, she compresses the post-measurement state to $S(\bar{\sigma})$ qubits/signal, and similarly if the outcome was 'τ-subspace', which occurs $(1 - \epsilon)$ of the time, she compresses to $S(\bar{\tau})$ qubits/signal. Thus the total sending resources is the sum of these three terms in (12).

Now suppose that $\sigma_1 \neq \sigma_2$, but that $\tau_1 = \tau_2 \equiv \tau$. Then, in the case that Alice's measurement outcome is 'τ-subspace', a result that also becomes known to Bob through the communication of the subspace names, she need not send the post-measurement state at all, as Bob already knows it (i.e., τ) and can construct it himself. Thus we may drop the last term in (12) and communicate the mixed states (with perfect fidelity) using only $H(\epsilon, 1 - \epsilon) + \epsilon S(\bar{\sigma})$ qubits/signal, which is less than $S(\bar{\rho})$ by an amount $(1 - \epsilon)S(\bar{\tau})$.

Example 12. (Visible coding by purification of the input states). The general idea here (cf. also [11]) is that in the visible situation, Alice may build purifications of the input mixed states and send these purifications (which are *pure* states) to Bob utilizing the compression of Schumacher's pure state coding theorem. On reception Bob regains the mixed states by selecting a suitable subsystem of each decoded purification state.

As a first example, consider a special case of states of the form in example 8. There are two possible input states,

$$\rho_1 = \text{diag}(\epsilon, 1 - \epsilon), \qquad \rho_2 = \text{diag}(1 - \epsilon, \epsilon), \tag{13}$$

with equal prior probabilities $p_1 = p_2 = \frac{1}{2}$. Hence $S(\bar{\rho}) = 1$ and $H(p_1, p_2) = 1$. After constructing purifications $|\psi_1\rangle$ and $|\psi_2\rangle$, Alice's task is to send a 50/50 mixture of $|\psi_1\rangle$ and $|\psi_2\rangle$. Thus to get the greatest benefit from Schumacher compression, the purifications should be chosen so that their ensemble has least von Neumann entropy; i.e., the two purifications should be as parallel as possible. According to Bures and Uhlmann's basic theorem [15–17], the minimum possible angle $\theta_{\min}$ between purifications of ρ_1 and ρ_2 is given by

$$\cos^2 \theta_{\min} = F(\rho_1, \rho_2).$$

Moreover, a 50/50 mixture of states at angle $\theta_{\min}$ has entropy

$$S_{\min} = H\left(\frac{1 + \cos \theta_{\min}}{2}, \frac{1 - \cos \theta_{\min}}{2}\right)$$

which gives the Schumacher limit of qubits/signal in compression by this method.

For the states in (13) we readily compute

$$F(\rho_1, \rho_2) = 4\epsilon(1 - \epsilon)$$

so that Alice may compress the purification ensemble to

$$\Upsilon(\epsilon) = H\left(\tfrac{1}{2} + \sqrt{\epsilon(1 - \epsilon)}, \tfrac{1}{2} - \sqrt{\epsilon(1 - \epsilon)}\right) \quad \text{qubits/signal}$$

which is better than $S(\bar{\rho})$ or $H(p_1, p_2)$, being equal to these only when ϵ is 0 or 1.

Note that the purely classical compression method of example 9 applies to this case, too. The relevant parameters are $p_1 = p_2 = \frac{1}{2}$, $\alpha_1 = \epsilon$, $\alpha_2 = 1 - \epsilon$, and $0 \leqslant \epsilon \leqslant \frac{1}{2}$. The method of example 9 gives compression to

$$\Xi(\epsilon) = H\left(2\epsilon, \tfrac{1}{2} - \epsilon, \tfrac{1}{2} - \epsilon\right) = H(2\epsilon, 1 - 2\epsilon) + 1 - 2\epsilon \quad \text{qubits/signal}$$

and we get

$$\Xi(\epsilon) \geqslant \Upsilon(\epsilon) \qquad \text{for} \quad 0 \leqslant \epsilon \leqslant \tfrac{1}{2}$$

with equality only for $\epsilon = 0$ or $\frac{1}{2}$. Thus, whenever the states ρ_1 and ρ_2 of equation (13) are mixed, the quantum purification compression beats the classical method of example 9.

Remark 13. (A simple construction of optimally parallel purifications for commuting states). Given any mixed state in diagonal form $\rho = \text{diag}(p_1, \ldots, p_n)$, we may immediately write down a canonical purification:

$$|\psi\rangle = \sum_{i=1}^{n} \sqrt{p_i} |e_i\rangle \otimes |e_i\rangle$$

where $\{|e_k\rangle\}$ is the diagonalizing basis for ρ. Given two such states,

$$\rho_1 = \text{diag}(p_1, \ldots, p_n), \qquad \rho_2 = \text{diag}(q_1, \ldots, q_n)$$

the canonical purifications clearly satisfy

$$|\langle\psi_1|\psi_2\rangle|^2 = \left(\sum_{i=1}^n \sqrt{p_i}\sqrt{q_i}\right)^2 = F(\rho_1, \rho_2)$$

$$= \text{(Bures–Uhlmann limit for } \cos^2\theta).$$

Thus for simultaneously diagonal states, the canonical purifications are always optimally parallel.

Notice that the diagonal entries of ρ_1 and ρ_2 are classical distributions $\boldsymbol{p}$ and $\boldsymbol{q}$ and that

$$|\langle\psi_1|\psi_2\rangle|^2 = F(\rho_1, \rho_2) = F_{\text{cl}}(\boldsymbol{p}, \boldsymbol{q}).$$

The construction of optimally parallel purifications converts the Bhattacharyya–Wootters overlap of classical distributions into quantum overlap of pure quantum states. In this way the methods of *quantum* coding may be applied to problems of compression of *classical* probability distributions.

Suppose now that we have two or *more* simultaneously diagonal states,

$$\rho^{(a)} = \text{diag}(p_1^{(a)}, \ldots, p_n^{(a)}) \qquad a = 1, \ldots, K.$$

Then their canonical purifications $|\psi^{(a)}\rangle$ have the remarkable property that they are all simultaneously pairwise maximally parallel. Recall that Uhlmann's theorem gives a limit on how parallel purifications can get for any *pair* of mixed states. It does not follow that this optimal parallelness can be *simultaneously* achieved by purifications of three or more states. Yet for simultaneously diagonal states, this optimal simultaneous parallelism is achieved by the canonical purifications.

It seems unlikely, however, that maximum parallelness gives the best set of purifications for the purpose of mixed-state compression when there are three or more signal states. Jozsa and Schlienz [20] have shown the existence of pairs of pure-state ensembles $\{p_i, |\psi_i\rangle\}$ and $\{p_i, |\chi_i\rangle\}$ for which all homologous pairs in the second ensemble are less parallel (i.e., $\forall i, j|\langle\chi_i|\chi_j\rangle| \geq |\langle\psi_i|\psi_j\rangle|$), but for which the entropy of the second ensemble is nevertheless smaller. This phenomenon is expected to persist under the added constraint that the states involved are purifications of the given mixed states.

Remark 14. If Alice sends Bob the canonical purification of ρ,

$$|\psi\rangle = \sum \sqrt{p_i}|e_i\rangle \otimes |e_i\rangle$$

she is actually supplying him with two copies of ρ—one for each of the two subsystems of the purification. Therefore one suspects that this compression is not optimal, at least when the criterion (LOCAL-FID) is used. To benefit from this observation, we might try to construct purifications each of which codes two signal states, one in each subsystem of the purification. To do this, the two signal states must have the same eigenvalues, but they need not be identical (e.g., as occurs in example 15 below). Thus the signal states would purify each other in pairs at the expense of introducing strong entanglement in the output signal sequence. This construction would have high (LOCAL-FID) fidelity, but low fidelity for the (GLOBAL-FID) criterion. Of course, even with the stronger criterion (GLOBAL-FID), it is not clear that the compression of example 12 is optimal.

Example 15. (The 'photographic negative' example, another application of compression by purification.) Suppose that we have d possible input signals ρ_i, where ρ_i is the $d \times d$ diagonal density matrix with equal entries $1/(d-1)$ along the diagonal except for the ith entry which is zero:

$$\rho_i = \frac{1}{d-1}\text{diag}(1, 1, \ldots, 1, 0, 1, \ldots, 1) \qquad \text{where 0 is in the } i\text{th place.}$$

The signals all have equal prior probabilities $p_i = 1/d$, giving $\bar{\rho} = I/d$.

The canonical purifications in $\mathcal{H}_d \otimes \mathcal{H}_d$ all lie in a d-dimensional subspace spanned by $\{|e_i\rangle \otimes |e_i\rangle\}$, where $\{|e_i\rangle\}$ is the diagonalizing basis of the ρ_i's. A direct calculation shows that the equally weighted mixture of purifications in this d-dimensional subspace is a density matrix

$$\rho = \frac{d-2}{d-1}|\psi\rangle\langle\psi| + \frac{1}{d-1}\frac{I}{d}$$

where I is the identity matrix in the d-dimensional subspace, and

$$|\psi\rangle = \frac{1}{\sqrt{d}}\sum_{i=1}^{d}|e_i\rangle \otimes |e_i\rangle$$

is a maximally entangled state. Thus ρ can be viewed as a mixture of a totally mixed state I/d, with probability $1/(d-1)$, and a maximally entangled pure state $|\psi\rangle\langle\psi|$, with probability $(d-2)/(d-1)$. Changing to a basis in which $|\psi\rangle$ is the first basis vector, we can easily determine the eigenvalues of ρ to be a nondegenerate eigenvalue $(d-1)/d$ and $(d-1)$ degenerate eigenvalues $1/d(d-1)$. A short calculation gives

$$q = S(\rho) = H\left(\frac{d-1}{d}, \frac{1}{d}\right) + \frac{1}{d}\log(d-1)$$

$$= \frac{2}{d}\log(d-1) - \log\left(1 - \frac{1}{d}\right) \quad \text{qubits/signal}$$

for the compression scheme. Note that $q \to 0$ as $d \to \infty$.

Introducing the Holevo quantity for the ensemble $E = \{p_i, \rho_i\}$,

$$\chi(E) = S(\bar{\rho}) - \sum p_i S(\rho_i) = -\log\left(1 - \frac{1}{d}\right)$$

we find

$$q = \chi + \frac{2}{d}\log(d-1)$$

so $q \to \chi$ as $d \to \infty$. Note that although we described this construction in terms of block coding the ensemble of canonical purifications for all the signals, it also provides canonical purifications for the ensemble of N-block mixed states. Nonetheless, for finite d, the above bound remains greater than the Holevo bound. Thus, if the conjecture that the Holevo bound is achievable by visible compression is correct, then, perhaps surprisingly, canonical purification is a suboptimal method of compression.

8. A lower bound on the rate of mixed-state compression

There is a simple argument that the Holevo quantity for an ensemble $E = \{p_i, \rho_i\}$ of mixed states is a lower bound on the rate at which such an ensemble can be coded. Here we use the global fidelity criterion (GLOBAL-FID), and encoding may be blind or visible. This argument uses the result, shown for pure-state ensembles by Hausladen *et al* [21] and for general mixed-state ensembles by Holevo [22] and by Schumacher and Westmoreland [23], that the Holevo quantity for an ensemble E is the capacity for classical information transmission using the states in the ensemble E as an alphabet. The gist of the argument is that if an ensemble of mixed states could be coded at a rate lower than its Holevo quantity, even with preparation-visible encoding, then one could code a Holevo quantity's worth of classical information into those mixed states, compress them to an ensemble on a channel space of size smaller than

the Holevo quantity (per use), recover the original ensemble with high fidelity, and therefore recover the classical information. But since the classical information capacity of an ensemble of states cannot be larger than the log of the dimension of its Hilbert space (since this is greater than or equal to χ for any ensemble), this is impossible.

To formalize this argument, consider an ensemble (or source) $E = \{p_i, \rho_i\}$ of mixed states ρ_i with probabilities p_i on a Hilbert space H_d of dimension d. The Holevo quantity for this ensemble is

$$\chi(E) \equiv S\left(\sum_i p_i \rho_i\right) - \sum_i p_i S(\rho_i). \tag{14}$$

A sequence of N signals from this source gives a state drawn from the ensemble

$$E^{\otimes N} = \{p_{i_1} p_{i_2} \cdots p_{i_N}, \rho_{i_1} \otimes \rho_{i_2} \otimes \cdots \otimes \rho_{i_N}\}. \tag{15}$$

We introduce the notation $f(A) = \{p_i, f(\rho_i)\}$ for the ensemble obtained by applying a map f to the states of the ensemble A, and we write $B(\mathcal{H})$ for the space of bounded operators on a Hilbert space $\mathcal{H}$.

For two ensembles with the same probabilities, $A = \{p_i, \rho_i\}$ and $B = \{p_i, \sigma_i\}$, we define an average fidelity by

$$\bar{F}(A, B) = \sum_i p_i F(\rho_i, \sigma_i). \tag{16}$$

In proving the main theorem of this section, we will need a lemma that bounds the absolute value of the difference in the Holevo quantities for two ensembles in terms of their average fidelity, provided the average fidelity is high enough.

Lemma 16. If $\bar{F}(A, B) > \sqrt{35/36}$, then

$$|\chi(A) - \chi(B)| \leqslant (2 + 2\sqrt{2})\sqrt{1 - \bar{F}(A, B)} \log d + 1 \tag{17}$$

where d is the dimension of the state space of A and B.

The proof is given in appendix B.

Our formulation of the mixed-state compression problem for the fidelity criterion (GLOBAL-FID) can now be stated succinctly. Relative to (GLOBAL-FID), the source E can be coded (or compressed) at a rate q if there exists a channel Hilbert space C with $q = \log(\dim C)$ and encoding/decoding schemes $\{e^{(N)}, \mathcal{D}^{(N)}\}$,

$$\begin{aligned} e^{(N)}: &\quad B(H_d^{\otimes N}) \to B(C^{\otimes N}) \\ \mathcal{D}^{(N)}: &\quad B(C^{\otimes N}) \to B(H_d^{\otimes N}) \end{aligned} \tag{18}$$

such that

$$\lim_{N \to \infty} \bar{F}(E^{\otimes N}, F^{(N)}) = 1 \qquad \text{(GLOBAL-FID)} \tag{19}$$

where

$$F^{(N)} \equiv \mathcal{D}^{(N)} \circ e^{(N)}(E^{\otimes N}) \tag{20}$$

is the ensemble after decoding. We require that the encodings $e^{(N)}$ take density operators to density operators and that the decodings $\mathcal{D}^{(N)}$ be trace-preserving completely positive linear maps. Permitting the encodings to be arbitrary maps on density operators allows for preparation-visible encoding; if $e^{(N)}$ is a trace-preserving completely positive linear map $\mathcal{E}^{(N)}$, then the compression is preparation-blind.

The argument outlined at the beginning of this section can be formalized in the following theorem.

Theorem 17. *For the fidelity criterion (GLOBAL-FID) and for both blind and visible encodings, the Holevo quantity $\chi(E)$ for an ensemble $E = \{p_i, \rho_i\}$ is a lower bound for $q_{\min}$.*

Proof. Suppose that the ensemble $E = \{p_i, \rho_i\}$ can be compressed at a rate $q < \chi(E)$ with asymptotically high fidelity (equation (19)), whether preparation-blind or preparation-visible. Consider the ensemble of channel states (density matrices)

$$W^{(N)} = \{p_{i_1} p_{i_2} \cdots p_{i_N}, w_{i_1 i_2 \ldots i_N}\} \tag{21}$$

where

$$w_{i_1 i_2 \ldots i_n} \equiv e^{(N)} \left(\rho_{i_1} \otimes \rho_{i_2} \otimes \cdots \otimes \rho_{i_N} \right)$$

is the encoded state corresponding to the unencoded source state $\rho_{i_1} \otimes \rho_{i_2} \otimes \cdots \rho_{i_N}$. The Holevo quantity for $W^{(N)}$ satisfies

$$\chi\left(W^{(N)}\right) \leqslant S\left(\begin{array}{c} \text{average density} \\ \text{operator of } W^{(N)} \end{array} \right) \leqslant Nq < N\chi(E) \tag{22}$$

where Nq is the the log of the dimension of the channel Hilbert space for N blocks of channel.

Consider now the following procedure for using the N-block operators $w_{i_1 \ldots i_N}$ as an alphabet to send classical information. Make codewords out of strings of M of these operators. Prune them as one would if one were coding using the operators $\rho_{i_1 \ldots i_N}$ of the ensemble $F^{(N)}$ in the Holevo–Schumacher–Westmoreland procedure for attaining $\chi(F^{(N)})$ as classical capacity. As the first step in the decoding procedure, convert them using the decoding $(\mathcal{D}^{(N)})^{\otimes M}$ into strings of the operators $\rho_{i_1 \ldots i_N}$ of the ensemble $F^{(N)}$. Then apply the decoding measurement appropriate to that ensemble.

This procedure clearly uses the N-block ensemble $W^{(N)}$ to transmit classical information at the rate $\chi(F^{(N)})$ per N blocks. But by assumption (cf. (19)), the ensemble $F^{(N)}$ has, at large enough N, arbitrarily high fidelity to the original ensemble $E^{\otimes N}$. Hence, applying lemma 16 to the ensembles $E^{\otimes N}$ and $F^{(N)}$, whose states lie in d^N-dimensional Hilbert spaces, one finds

$$\left| \chi(E) - \frac{\chi\left(F^{(N)}\right)}{N} \right| \leqslant (2 + \sqrt{2})\sqrt{1 - \bar{F}\left(E^{\otimes N}, F^{(N)}\right)} \log d + \frac{1}{N}. \tag{23}$$

Thus for large enough N, $\chi(F^{(N)})/N$ is arbitrarily close to $\chi(E)$, which is greater than $\chi(W^{(N)})/N$ by at least an amount $\chi(E) - q$, independent of N. So for large enough N, $\chi(F^{(N)})$ exceeds $\chi(W^{(N)})$, contradicting the fact that the classical capacity of the ensemble $W^{(N)}$ is $\chi(W^{(N)})$. We conclude that the compression rate q must satisfy $q \geqslant \chi(E)$. □

Horodecki [11] has independently derived the lower bound of theorem 17, using the nonincrease of the Holevo quantity under completely positive maps. This nonincrease is an easy consequence of the monotonicity of relative entropy under such maps [24, 25], and therefore of Lieb's fundamental concavity theorem [26]. (A good treatment of all of these is to be found in [27].)

A special case of theorem 17 is the lower bound of $S(\bar{\rho})$ qubits per source signal on the rate of compression of ensembles of pure states. This lower bound was established for preparation-blind encodings and unitary decodings in [1]; for arbitrary (preparation-blind or preparation-visible) encodings and unitary decodings in [18]; and, by somewhat technical arguments, for arbitrary encodings and decodings using completely positive trace-preserving maps in [2]. The present result allows for arbitrary encodings and decodings using completely positive trace-preserving maps, so it provides an alternative and perhaps more satisfying derivation of the most general form of the pure-state lower bound.

The lower bound in theorem 17 raises the fundamental open question of whether the bound is achievable (with global fidelity) with either blind or visible encoding. If not, one would like an expression for the achievable rate in both cases. Even for transmitting classical mixed states, the question of the best achievable rate remains open, in both the variant allowing quantum means of compression and that requiring only classical means.

Acknowledgments

Appendix A. Proof of theorem 7

Proposition 6 may be used for part of the proof, but we give a different argument that utilizes properties of the Bures–Uhlmann fidelity function throughout. We first establish two lemmas which are direct Bures–Uhlmann fidelity analogues of lemmas 1 and 2 in [18].

Lemma A1. *Let ρ and ρ' be mixed states on $\mathcal{H}_n$ with ρ' supported on a d-dimensional subspace D. Then $F(\rho, \rho')$ is less than the sum of the d largest eigenvalues of ρ, which we write as $1 - \eta$.*

Proof of lemma A1. We use the fact that

$$F(\rho, \rho') = \inf \operatorname{trace} \rho A \operatorname{trace} \rho' A^{-1} \tag{24}$$

where the infimum is over all strictly positive operators A [28]. Choose

$$A = \begin{cases} I & \text{on } D \\ \epsilon I & \text{on } D^\perp \text{ (for any } \epsilon > 0). \end{cases}$$

Then

$$\operatorname{trace} \rho' A^{-1} = 1$$

and

$$\operatorname{trace} \rho A = \operatorname{trace} \rho D + \epsilon \operatorname{trace} \rho D^\perp \leqslant 1 - \eta + \epsilon \operatorname{trace} \rho D^\perp \leqslant 1 - \eta.$$

Hence

$$F(\rho, \rho') \leqslant \operatorname{trace} \rho A \operatorname{trace} \rho' A^{-1} \leqslant 1 - \eta$$

as required. □

To set the stage for the second lemma, consider a density operator ρ on $\mathcal{H}_n$. Denote the eigenvalues of ρ in decreasing order by λ_i, $i = 1, \ldots, n$. Let D be the d-dimensional subspace spanned by the eigenvectors belonging to the d largest eigenvalues of ρ; denote the sum of these d eigenvalues by $1 - \eta$. Denote the projector onto D by Π, and let $|0\rangle$ be any pure state in D. Now consider the density operator

$$\rho' \equiv \Pi \rho \Pi + \eta |0\rangle\langle 0|.$$

This density operator can be obtained from ρ by first applying the binary measurement that projects onto D (outcome '1') or onto $D^\perp$ (outcome '0') and then, if the outcome is 0, substituting $|0\rangle\langle 0|$ for the post-measurement state. With these preliminaries, the lemma can be stated as follows.

Lemma A2. $F(\rho, \rho') \geqslant 1 - 2\eta$.

Proof of lemma A2. If we write ρ in its orthonormal eigenbasis,

$$\rho = \sum_{i=1}^{n} \lambda_i |e_i\rangle\langle e_i|$$

ρ' becomes

$$\rho' = \sum_{i=1}^{d} \lambda_i |e_i\rangle\langle e_1| + \eta|0\rangle\langle 0|.$$

Introduce the following purifications of ρ and ρ':

$$|\phi\rangle = \sum_{i=1}^{n} \sqrt{\lambda_i} |e_i\rangle \otimes |f_i\rangle$$

$$|\phi'\rangle = \sum_{i=1}^{d} \sqrt{\lambda_i} |e_i\rangle \otimes |f_i\rangle + \sqrt{\eta}|0\rangle \otimes |g\rangle.$$

Here the vectors $|f_i\rangle$, $i = 1, \ldots, d$ are orthonormal, and $|g\rangle$ is orthogonal to each $|f_i\rangle$. Since fidelity is the maximum absolute value of the inner product of purifications, we have

$$F(\rho, \rho') \geqslant |\langle\phi|\phi'\rangle|^2 = \left(\sum_{i=1}^{d} \lambda_i\right)^2 = (1 - \eta)^2 \geqslant 1 - 2\eta$$

as required.   □

Proof of theorem 7. Suppose that we compress to $S(\bar{\rho}) - \delta$ qubits/signal by any coding method whatsoever. Then if the decoding scheme is unitary, the decoded state $\tilde{\sigma}_N$ of an input string σ_N of length N is supported in $N(S(\bar{\rho}) - \delta)$ qubits. Yet the density matrix for strings of length N is $\bar{\rho}^{\otimes N}$, and by a standard typical sequences result (cf. [18]), the sum of the $2^{N(S(\bar{\rho})-\delta)}$ largest eigenvalues of $\bar{\rho}^{\otimes N}$ becomes arbitrarily small with increasing N. Hence, by lemma A1, $F(\sigma_N, \tilde{\sigma}_N)$ is arbitarily small, too, and the fidelity cannot be high by the (GLOBAL-FID) criterion.

On the other hand, if $S(\bar{\rho}) + \delta$ qubits/signal are available, then lemma A2 provides an explicit high-fidelity coding scheme, with D being the $2^{N(S(\bar{\rho})+\delta)}$-dimensional subspace spanned by the $2^{N(S(\bar{\rho})+\delta)}$ weightiest eigenvectors of $\bar{\rho}^{\otimes N}$.   □

Appendix B. Proof of lemma 16

The proof uses the following inequality (proved in [4]):

$$|S(\rho_1) - S(\rho_2)| \leqslant 2\sqrt{1 - F(\rho_1, \rho_2)} \log d + 1 \tag{25}$$

which is valid if

$$2\sqrt{1 - F(\rho_1, \rho_2)} < \tfrac{1}{3}. \tag{26}$$

We use this to obtain a similar relation, but with the average ensemble fidelity in place of the fidelity on the right-hand side. Let $A = \{p_i, \rho_i\}$ and $B = \{p_i, \sigma_i\}$ again denote two mixed-state ensembles having the same probabilities. Letting $\rho \equiv \sum_i p_i \rho_i$ and $\sigma \equiv \sum_i p_i \sigma_i$, we have an inequality involving the error measure $\sqrt{1 - F(\rho, \sigma)}$. We need to convert this into one involving the error measure $1 - \bar{F}(A, B)$. Defining yet another error measure

$\delta \equiv 1 - \sqrt{F(\rho, \sigma)}$, simple algebra gives $\sqrt{1 - F(\rho, \sigma)} = \sqrt{\delta(2 - \delta)}$. The double concavity of $G(\rho_1, \rho_2) \equiv \sqrt{F(\rho_1, \rho_2)}$ (proved in appendix C) gives

$$\sum_i p_i G(\rho_i, \sigma_i) \leqslant G\left(\sum_i p_i \rho_i, \sum_i p_i \sigma_i\right) = G(\rho, \sigma). \tag{27}$$

Hence

$$\delta \equiv 1 - G(\rho, \sigma) \leqslant 1 - \sum_i p_i G(\rho_i, \sigma_i)$$

$$\leqslant 1 - \sum_i p_i F(\rho_i, \sigma_i) = 1 - \bar{F}(A, B) \equiv \epsilon. \tag{28}$$

Therefore, we have the inequality

$$|S(\rho) - S(\sigma)| \leqslant 2\sqrt{\delta(2 - \delta)} \log d + 1$$
$$\leqslant 2\sqrt{\epsilon(2 - \epsilon)} \log d + 1$$
$$\leqslant 2\sqrt{2}\sqrt{1 - \bar{F}(A, B)} \log d + 1. \tag{29}$$

This inequality is valid provided (26) holds, which is certainly true if

$$0 \leqslant \epsilon < 1 - \sqrt{35/36} \qquad \Longleftrightarrow \qquad \bar{F}(A, B) > \sqrt{35/36}. \tag{30}$$

Furthermore, we can also use the inequality (25) to bound the difference in the average entropies for the two ensembles of d-dimensional states,

$$\left| \sum_i p_i S(\rho_i) - \sum_i p_i S(\sigma_i) \right| \leqslant \sum_i p_i |S(\rho_i) - S(\sigma_i)|$$

$$\leqslant \sum_i p_i \left(2\sqrt{1 - F(\rho_i, \sigma_i)} \log d + 1\right)$$

$$\leqslant 2\sqrt{1 - \sum_i p_i F(\rho_i, \sigma_i)} \log d + 1$$

$$\equiv 2\sqrt{1 - \bar{F}(A, B)} \log d + 1. \tag{31}$$

Combining equations (29) and (31) yields the desired result (17). $\qquad \square$

Appendix C. Double concavity of $G(\rho_1, \rho_2)$

In this appendix we show that

$$G(\rho_1, \rho_2) \equiv \sqrt{F(\rho_1, \rho_2)} = \text{trace}\sqrt{\sqrt{\rho_1} \rho_2 \sqrt{\rho_1}} \tag{32}$$

is doubly concave, i.e.,

$$G(\lambda \rho_1 + (1 - \lambda)\sigma_1, \lambda \rho_2 + (1 - \lambda)\sigma_2) \geqslant \lambda G(\rho_1, \rho_2) + (1 - \lambda)G(\sigma_1, \sigma_2). \tag{33}$$

The proof uses a representation of the quantum fidelity in terms of measurement probabilities. Given a measurement described by a positive-operator-valued measure (POVM) with POVM elements E_i, the probability for outcome i is $p_i = \text{trace}\,\rho E_i$. Fuchs and Caves [29] showed that the quantum fidelity of ρ_1 and ρ_2 is the classical fidelity of the measurement probabilities for the measurement that, according to the classical fidelity, best distinguishes the two density operators, i.e.,

$$F(\rho_1, \rho_2) = \min_{\{E_i\}} F_{\text{cl}}(\boldsymbol{p}_1, \boldsymbol{p}_2). \tag{34}$$

Here the minimum is taken over all POVMs $\{E_i\}$, and $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ are the column vectors of measurement probabilities for ρ_1 and ρ_2 generated by the POVM $\{E_i\}$.

The proof begins by noting that for four positive real numbers,

$$0 \leqslant (\sqrt{x_1 y_2} - \sqrt{x_2 y_1})^2 = x_1 y_2 + x_2 y_1 - 2\sqrt{x_1 x_2 y_1 y_2}$$

from which it follows that the function $\sqrt{x_1 x_2}$ is doubly concave, i.e.,

$$\sqrt{[\lambda x_1 + (1-\lambda)y_1][\lambda x_2 + (1-\lambda)y_2]} = \sqrt{\lambda^2 x_1 x_2 + (1-\lambda)^2 y_1 y_2 + \lambda(1-\lambda)(x_1 y_2 + x_2 y_1)}$$

$$\geqslant \sqrt{\lambda^2 x_1 x_2 + (1-\lambda)^2 y_1 y_2 + 2\lambda(1-\lambda)\sqrt{x_1 x_2 y_1 y_2}}$$

$$= \lambda\sqrt{x_1 x_2} + (1-\lambda)\sqrt{y_1 y_2}.$$

The square root of the classical fidelity,

$$G_{\mathrm{cl}}(\boldsymbol{p}, \boldsymbol{q}) = \sqrt{F_{\mathrm{cl}}(\boldsymbol{p}, \boldsymbol{q})} = \sum_{i=1}^{n} \sqrt{p_i q_i} \tag{35}$$

being a sum of such functions, is thus also doubly concave:

$$G_{\mathrm{cl}}(\lambda\boldsymbol{p}_1 + (1-\lambda)\boldsymbol{q}_1, \lambda\boldsymbol{p}_2 + (1-\lambda)\boldsymbol{q}_2) \geqslant \lambda G_{\mathrm{cl}}(\boldsymbol{p}_1, \boldsymbol{p}_2) + (1-\lambda)G_{\mathrm{cl}}(\boldsymbol{q}_1, \boldsymbol{q}_2). \tag{36}$$

Now use the representation (34), written in terms of square roots of fidelities, to show the double concavity of $G(\rho_1, \rho_2)$:

$$G(\lambda\rho_1 + (1-\lambda)\sigma_1, \lambda\rho_2 + (1-\lambda)\sigma_2) = \min_{\{E_j\}} G_{\mathrm{cl}}(\lambda\boldsymbol{p}_1 + (1-\lambda)\boldsymbol{q}_1, \lambda\boldsymbol{p}_2 + (1-\lambda)\boldsymbol{q}_2)$$

$$\geqslant \min_{\{E_j\}} (\lambda G_{\mathrm{cl}}(\boldsymbol{p}_1, \boldsymbol{p}_2) + (1-\lambda)G_{\mathrm{cl}}(\boldsymbol{q}_1, \boldsymbol{q}_2))$$

$$\geqslant \min_{\{E_j\}} \lambda G_{\mathrm{cl}}(\boldsymbol{p}_1, \boldsymbol{p}_2) + \min_{\{F_j\}}(1-\lambda)G_{\mathrm{cl}}(\boldsymbol{q}_1, \boldsymbol{q}_2)$$

$$= \lambda G(\rho_1, \rho_2) + (1-\lambda)G(\sigma_1, \sigma_2). \tag{37}$$

(Another proof, by Nielsen [30], uses the relation of quantum fidelity to purifications.)

References

[1] Schumacher B W 1995 Quantum coding *Phys. Rev.* A **51** 2738–47
[2] Barnum H, Caves C M, Fuchs C A, Jozsa R and Schumacher B W 1996 Noncommuting mixed states cannot be broadcast *Phys. Rev. Lett.* **76** 2738–21
[3] Barnum H, Nielsen M A and Schumacher B W 1998 Information transmission through a noisy quantum channel *Phys. Rev.* A **57** 4153–75
[4] Barnum H, Knill E and Nielsen M A 1998 On quantum fidelities and channel capacities *IEEE Trans. Inform. Theory* **46** 1317–29
[5] von Neumann J and Morgenstern O 1944 *Theory of Games and Economic Behavior* (Princeton, NJ: Princeton University Press)
[6] Luce R D and Raiffa H 1957 *Games and Decisions* (New York: Wiley); reprinted 1989 (New York: Dover)
[7] Grover L 1999 Rapid sampling through quantum computing arXiv *Preprint* quant-ph/9912001
[8] Bernstein E and Vazirani U 1997 Quantum complexity theory *SIAM J Comput.* **26** 1411–73
[9] Jozsa R 1994 Quantum noiseless coding of mixed states *Talk given at Third Santa Fe Workshop on Complexity, Entropy, and the Physics of Information (May 1994)*
[10] Barnum H, Fuchs C A, Jozsa R and Schumacher B 1996 General fidelity limit for quantum channels *Phys. Rev.* A **54** 4707
[11] Horodecki M 1996 Limits for compression of quantum information carried by ensembles of mixed states *Phys. Rev.* A **57** 3364–9
[12] Horodecki M 1999 Towards optimal compression for mixed signal states arXiv *Preprint* quant-ph/9905058
[13] Bennett C H and Brassard G 1985 Quantum cryptography: Public-key distribution and coin-tossing *Proc. IEEE Int. Conf. on Computers, Systems, and Signal Processing (Bangalore)* (New York: IEEE) pp 175–9

[14] Bennett C H 1992 Quantum cryptography using any two nonorthogonal states *Phys. Rev. Lett.* **68** 3121–4

[15] Bures D 1969 An extension of Kakutani's theorem on infinite product measures to the tensor product of semifinite w^*-algebras *Trans. Am. Math. Soc.* **135** 199

[16] Uhlmann A 1976 The 'transition probability' in the state space of a a^*-algebra *Rep. Math. Phys.* **9** 273–9

[17] Jozsa R 1994 Fidelity for mixed quantum states *J. Mod. Opt.* **41** 2314–33

[18] Jozsa R and Schumacher B W 1994 A new proof of the quantum noiseless coding theorem *J. Mod. Opt.* **41** 2343–9

[19] Lo H-K 1995 Quantum coding theorem for mixed states *Opt. Commun.* **19** 552

[20] Jozsa R and Schlienz J 2000 Distinguishability of states and von Neumann entropy *Phys. Rev.* A **62** 012301

[21] Hausladen P, Jozsa R, Schumacher B W, Westmoreland M and Wootters W K 1996 Classical information capacity of a quantum channel *Phys. Rev.* A **54** 1869–76

[22] Holevo A S 1998 The capacity of quantum channel with general signal states *IEEE Trans. Inform. Theory* **44** 269

[23] Schumacher B W and Westmoreland M 1997 Sending classical information via noisy quantum channels *Phys. Rev.* A **56** 131–8

[24] Lindblad G 1973 Entropy, information, and quantum measurements *Commun. Math. Phy.* **33** 305–22

[25] Uhlmann A 1977 Relative entropy and the Wigner–Yanase–Dyson–Lieb concavity in an interpolation theory *Commun. Math. Phy.* **54** 21–32

[26] Lieb E 1973 Convex trace functions and the Wigner–Yanase–Dyson conjecture *Adv. Math.* **11** 267–88

[27] Ahlswede R and Löber P 1999 Quantum data processing arXiv *Preprint* quant-ph/9907081

[28] Alberti P M 1983 A note on the transition probability over C^*-algebras *Lett. Math. Phys.* **7** 25–32

[29] Fuchs C A and Caves C M 1995 Mathematical techniques for quantum communication theory *J. Open Systems Inform. Dynamics* **3** 1

[30] Nielsen M A 1998 Quantum information theory Dissertation (Albuquerque, NM: University of New Mexico)