Compressing Quantum Mixed-State Sources by Sending Classical Information

Emina Soljanin, *Member, IEEE*

***Abstract*—We consider visible compression for discrete memoryless sources of mixed quantum states when only classical information can be sent from Alice to Bob. We assume that Bob knows the source statistics, and that Alice and Bob have access to the same source of random numbers. We put in an information-theoretic framework some recent results on visible compression for sources of states with commuting density operators, and remove the commutativity requirement. We derive a general achievable compression rate, which is for the noncommutative case still higher than the known lower bound. We also present several related problems of classical information theory, and show how they can be used to answer some questions of the mixed-state compression problem.**

***Index Terms*—Data compression, mixed-state sources, quantum information theory.**

I. Introduction

A DISCRETE memoryless source (DMS) of information produces a sequence of independent and identically distributed random variables taking values in a finite set called the *source alphabet*. In quantum systems, source letters are mapped into *quantum states* for quantum transmission or storage. In the simplest case, quantum states correspond to unit length column vectors in a d-dimensional Hilbert space $\mathcal{H}_d$. Such quantum states are called *pure*. When $d = 2$, quantum states are called *qubits*. A column vector is denoted by $|\varphi\rangle$, its complex conjugate transpose by $\langle\varphi|$. A pure state is mathematically described by its *density matrix* equal to the outer product $|\varphi\rangle\langle\varphi|$. In a more complex case, all we know about a quantum state is that it is one of a finite number of possible pure states $|\varphi_i\rangle$ with probability p_i. Such quantum states are called *mixed*. A mixed state is also described by its density matrix which is equal to $\sum_i p_i|\varphi_i\rangle\langle\varphi_i|$. Note that a density matrix is a $d \times d$ Hermitian trace-one positive semidefinite matrix. A classical analog to a mixed state can be a multifaced coin which turns up as any of its faces with the corresponding probability.

Compression algorithms deal with source sequences rather than individual letters. There are two possible scenarios for which algorithms can be designed: *visible* when the encoder Alice knows the source sequence and *blind* when only the quantum state corresponding to the sequence is available to her. The quantum state corresponding to a source sequence of length n has a $d^n \times d^n$ density matrix, equal to the tensor product of density matrices corresponding to the letters in the sequence. In the blind case, lossless quantum compression algorithms map (encode) these product states into states over Hilbert spaces of smaller dimension with arbitrarily high expected reconstruction (decoding) *fidelity* as $n \to \infty$. Operations used for encoding and decoding have to be allowed by quantum mechanics. In the visible case, Alice can as well compress the available classical information, which the decoder Bob can use to prepare a quantum state that (as in the blind case) approximate Alice's with arbitrarily high expected fidelity as $n \to \infty$.

We pose the following question: what is the best compression compatible with the fidelity goal and encoding/decoding constraints for each scenario. The answer to the question was given by Schumacher for discrete memoryless sources of pure quantum states in [1] (see also [2] and [3] for some important extensions of the Schumacher's results). Lossless compression of sources of possibly mixed quantum states is not yet fully understood, and is the subject of current research [4]–[9]. The optimal compression rate for the blind case scenario was discussed in [9]. A lower bound on the compression rate was established by Horodecki in [5] and by Barnum, Caves, Fuchs, Jozsa, and Schumacher in [6]. The optimal compression rates for some special cases were found by Horodecki in [4] and by Barnum, Caves, Fuchs, Jozsa, and Schumacher in [6]. More recently, an algorithm achieving the lower bound on the compression rate for the visible case of states with commuting density operators was presented by Dür, Vidal, and Cirac in [7], and a possibly related classical information theory problem was discussed by Kramer and Savari in [8]. Some of these results will be addressed in more detail after the problem we are dealing with is precisely formulated.

We are concerned with visible compression of discrete memoryless sources when only classical information can be sent from Alice to Bob. We assume that Bob knows the source statistics, and that Alice and Bob have access to the same source of random numbers. This scenario is the one studied by Dür, Vidal, and Cirac for the case of states with commuting density operators [7]. When put in an information-theoretic framework, the commutativity requirement can be easily removed, and an achievable rate can be found in the same manner. However, the derived achievable rate is still higher than the lower bound.

In the second part of the paper, we present several related problems of classical information theory, and show how they can be used to answer some questions of the mixed-state compression problem. This paper is written for both information theorists and physicists, although papers written for two audiences often satisfy neither. Here writing for these two groups of scientists merely means that we tried to keep the paper as self-con-

Manuscript received October 22, 2001; revised March 14, 2002. The material in this paper was presented in part at the American Mathematical Society (AMS) 2001 Spring Eastern Section Meeting, Hoboken, NJ, April 28–29, 2001.

The author is with Bell Labs, Lucent Technologies, Murray Hill, NJ 07974 USA (e-mail: emina@lucent.com).

Communicated by P. W. Shor, Associate Editor for Quantum Information Theory.

Publisher Item Identifier 10.1109/TIT.2002.800500.

0018-9448/02$17.00 © 2002 IEEE

$$\begin{aligned}\rho_1 &= \tfrac{2}{3}|\psi_1\rangle\langle\psi_1| + \tfrac{1}{3}|\psi_2\rangle\langle\psi_2| \\ \rho_2 &= \tfrac{1}{3}|\psi_2\rangle\langle\psi_2| + \tfrac{2}{3}|\psi_3\rangle\langle\psi_3| \\ \rho &= \tfrac{1}{2}\rho_1 + \tfrac{1}{2}\rho_2 \\ &= \tfrac{1}{3}|\psi_1\rangle\langle\psi_1| + \tfrac{1}{3}|\psi_2\rangle\langle\psi_2| + \tfrac{1}{3}|\psi_3\rangle\langle\psi_3| \\ &= \tfrac{1}{2}I\end{aligned}$$

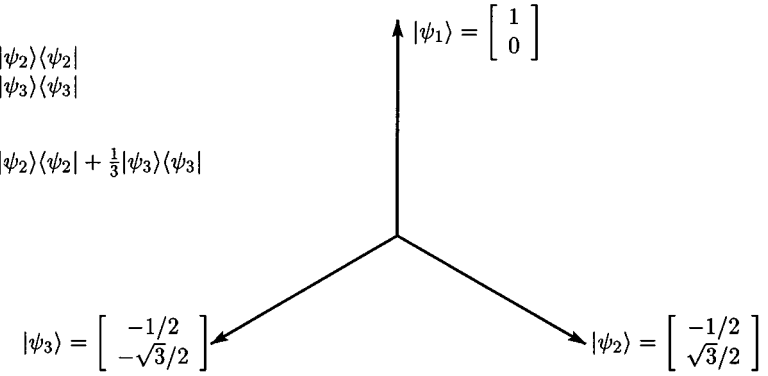


Fig. 1. Mixed-state ensemble.

tained as possible, and presented proofs and other material in an elementary rather than the most efficient way.

A. Problem Formulation

Let $\mathcal{X}$ be a finite set (alphabet), and $\{\rho_a | a \in \mathcal{X}\}$ a set of (possibly mixed) quantum states in a d-dimensional Hilbert space $\mathcal{H}_d$. Let $\mathcal{P}(\mathcal{X})$ be the set of all probability distributions on $\mathcal{X}$, and $P \in \mathcal{P}(\mathcal{X})$ a particular distribution. The set $\mathcal{E} = \{\rho_a, P(a) | a \in \mathcal{X}\}$ is usually referred to as an ensemble of mixed states indexed by the elements of $\mathcal{X}$. The density matrix of the ensemble $\mathcal{E}$, which we shall also refer to as the source density matrix, is given by

$$\rho = \sum_{a \in \mathcal{X}} P(a)\rho_a. \tag{1}$$

We shall assume that states ρ_a are mixtures of known (possibly nonorthogonal) pure states as follows. Let $\mathcal{Y}$ be a finite set, and $\{|\psi_b\rangle\langle\psi_b| | b \in \mathcal{Y}\}$ be a set of pure quantum states in $\mathcal{H}_d$ indexed by the elements of $\mathcal{Y}$. Let W be an $|\mathcal{X}| \times |\mathcal{Y}|$ stochastic matrix with elements $W_{ab} = W(b|a)$, $a \in \mathcal{X}$, $b \in \mathcal{Y}$, where $W(\cdot|a)$ is a probability distribution on $\mathcal{Y}$ for each $a \in \mathcal{X}$. We assume that no two states ρ_a are identical in the sense that no two rows of W are identical. The density matrices in $\mathcal{E}$ are given by

$$\rho_a = \sum_{b \in \mathcal{Y}} W(b|a)\,|\psi_b\rangle\langle\psi_b|, \qquad a \in \mathcal{X}. \tag{2}$$

A source producing mixed states ρ_a, $a \in \mathcal{X}$, independently according to the probability distribution P, effectively produces pure states $|\psi_b\rangle\langle\psi_b|$, $b \in \mathcal{Y}$, independently according to the following probability distribution Q:

$$Q(b) = \sum_{a \in \mathcal{X}} P(a)W(b|a). \tag{3}$$

Thus, the source density matrix (1) can also be expressed in terms of $|\psi_b\rangle\langle\psi_b|$ and $Q(b)$, $b \in \mathcal{Y}$, as

$$\rho = \sum_{b \in \mathcal{Y}} Q(b)|\psi_b\rangle\langle\psi_b|.$$

Example 1: A possible mixed-state ensemble is shown in Fig. 1. Here $d = 2$, $|\mathcal{X}| = 2$, and $|\mathcal{Y}| = 3$.

The memoryless source produces sequences of letters, where each letter is drawn from the set $\mathcal{X}$ independently according to the probability distribution P. Thus, a source sequence $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$ occurs with probability

$$P(\boldsymbol{x}) = P(x_1) \cdot \ldots \cdot P(x_n)$$

and the corresponding state has a density matrix $\boldsymbol{\rho_x} = \rho_{x_1} \otimes \cdots \otimes \rho_{x_n}$. On the transmitting end, the encoder Alice knows $\mathcal{E}$ and $\boldsymbol{x}$. On the receiving end, the decoder Bob knows $\mathcal{E}$. In addition, Alice and Bob have access to the same source of random numbers.

For each source sequence $\boldsymbol{x}$, Alice prepares and sends to Bob Rn bits of classical information, which he uses (together with his prior knowledge of $\mathcal{E}$) to prepare state $\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}$. To measure how faithfully mixed state σ approximates mixed state ω and *vice versa*, we use the so-called *mixed-state fidelity* F defined as

$$F(\sigma, \omega) = \left\{ \mathrm{Tr}\left[\left(\sqrt{\sigma}\omega\sqrt{\sigma} \right)^{1/2} \right] \right\}^2 \tag{4}$$

whose maximum value is 1. We shall say that the mixed-state compression is lossless when the expected value of $F(\boldsymbol{\rho_x}, \hat{\boldsymbol{\rho}}_{\boldsymbol{x}})$ can be made arbitrarily close to 1 by increasing the length n of the source sequence

$$\sum_{\boldsymbol{x} \in \mathcal{X}^n} P(\boldsymbol{x})F(\boldsymbol{\rho_x}, \hat{\boldsymbol{\rho}}_{\boldsymbol{x}}) \to 1, \qquad \text{as } n \to \infty. \tag{5}$$

The fidelity (4), also known as Uhlmann fidelity, is based on so-called Uhlmann's transition probability [10]. Jozsa proposed that it be used to measure the closeness between two mixed states [11], and a good description of its general properties can be found in [15, Ch. 9]. The fidelity (4) is the quantum analog of the classical Bhattacharyya–Wooters overlap between two probability distributions $\{p_i\}$ and $\{q_i\}$ defined as

$$F(\{p_i\}, \{q_i\}) = \left(\sum_i \sqrt{p_i q_i} \right)^2.$$

If two quantum states faithfully approximate each other in the sense of (4), then any measurement performed on those quantum states will give rise to probability distributions which faithfully approximate each other in the classical sense of the Bhattacharyya–Wooters overlap.

B. Information Measures

In compression of mixed-state sources by sending classical information, the well-known classical information measures

will play a role. Entropy $H(Q)$, conditional entropy $H(W/P)$, and mutual information $I(P, W)$ are defined as

$$\begin{aligned} H(Q) &= -\sum_{b\in\mathcal{Y}} Q(b)\log Q(b) \\ H(W/P) &= -\sum_{a\in\mathcal{X}} P(a)\sum_{b\in\mathcal{Y}} W(b|a)\log W(b|a) \\ I(P, W) &= H(Q) - H(W/P). \end{aligned} \tag{6}$$

The following corresponding quantum information measures are the source Von Neumann entropy $S(\rho)$, the expected value of the Von Neumann entropies of the source letters $\overline{S}$, and the Holevo quantity χ:

$$\begin{aligned} S(\rho) &= -\mathrm{Tr}\rho\log\rho \\ \overline{S} &= \sum_{a\in\mathcal{X}} P(a)S(\rho_a) \\ \chi &= S(\rho) - \overline{S}. \end{aligned} \tag{7}$$

When $|\psi_b\rangle\langle\psi_b|$, $b \in \mathcal{Y}$, are orthogonal, the quantum quantities (7) and their classical counterparts (6) are equal

$$\begin{aligned} S(\rho) &= H(Q) \\ \overline{S} &= H(W/P) \\ \chi &= I(P, W). \end{aligned}$$

Notice that the classical information measures (6) are defined as functions of probability distributions as in [14], rather than as functions of random variables as in [21]. The main proofs in this paper use the method of types, and that is the principal reason for this formulation, which in addition makes the parallel between the quantum and classical information measures more apparent, since the quantum information measures are defined in terms of density matrices (quantum analogs of probability distributions). Quantum physicists may find expressions (6) surprising because they commonly use [21] as the standard reference for information theory.

For the classical information theory problems discussed in Section VI, we also need a stochastic matrix U with elements $U_{ba} = U(a|b)$, $a \in \mathcal{X}$, $b \in \mathcal{Y}$, where $U(\cdot|b)$ is a probability distribution on $\mathcal{X}$ for each $b \in \mathcal{Y}$. The elements of U are computed as follows:

$$U(a|b) = P(a)W(b|a)/Q(b).$$

Entropy $H(P)$, conditional entropy $H(U/Q)$, and mutual information $I(Q, U)$ are defined as the corresponding quantities in (6).

C. Known and New Results

With the information measures defined, we can now state more precisely the visible quantum compression results mentioned in the Introduction which are relevant for our problem of visible compression. For sources of pure quantum states, the optimal compression rate is $S(\rho)$; the information sent from Alice to Bob is quantum [1]. (The same compression rate is the optimal for the blind scenario.) Since a source producing mixed states ρ_a, $a \in \mathcal{X}$, independently according to the probability distribution P, effectively produces pure states $|\psi_b\rangle\langle\psi_b|$, $b \in \mathcal{Y}$, independently according to the probability distribution Q of (3), one apparent achievable rate is $S(\rho)$. Since the encoder Alice knows the source sequence $\boldsymbol{x}$, she can pass this information to Bob, who can then prepare $\boldsymbol{\rho_x}$ in his lab. Thus, another apparent achievable rate is $H(P)$. For sources of mixed quantum states and the fidelity criterion (5), a lower bound on the compression rate of any visible compression scheme is the Holevo quantity χ [5], [6]. This lower bound can be achieved by a specific compression algorithm in the case of quantum states with commuting density operators [7]; the information sent from Alice to Bob is classical. The optimal compression rates for some special cases were found in [4] and [6]. Achievable compression rates for sources of quantum states with commuting density operators and a fidelity criterion different than (5) are found in [8] (see Section VI-B).

When the density matrices ρ_a, $a \in \mathcal{X}$, commute, they can be made diagonal in the same basis. Thus, one can assume that they are mixtures of orthogonal pure states $|\psi_b\rangle\langle\psi_b|$, $b \in \mathcal{Y}$. We address the general case, i.e., the one where the $|\psi_b\rangle\langle\psi_b|$, $b \in \mathcal{Y}$, are not necessarily orthogonal. We present a compression algorithm which achieves a compression rate of $I(P, W)$; the information sent from Alice to Bob is classical. Section I-D outlines the main idea of the algorithm, and Sections II–IV are concerned with formal presentation of the algorithm and proofs. Applications of the algorithm to the source described by Fig. 1 and a source with commuting density operators are presented in Section V. In Section VI, we consider three related classical problems. Two deal with the problem of compressing sources of sources, i.e., sources of probability distributions with two different fidelity criteria. One deals with the problem of quantum mixed-state compression by compressing the source sequence $\boldsymbol{x}$. By using ideas for classical channel coding, we show that at compression rates below $H(P)$, the average mixed-state fidelity (5) approaches 0.

D. The Idea for the Compression Algorithm

The main idea is simple to state for the reader already familiar with the notion of typicality as well as the notion of joint and conditional typicality. A rigorous description, given in the proceeding sections, uses the precision provided by the method of types.

For each $\boldsymbol{x}$, Alice's state $\boldsymbol{\rho_x}$ is roughly a uniform mixture of pure states

$$|\Psi_{\boldsymbol{y}}\rangle\langle\Psi_{\boldsymbol{y}}| = |\psi_{y_1}\rangle\langle\psi_{y_1}| \otimes \cdots \otimes |\psi_{y_n}\rangle\langle\psi_{y_n}|$$

where $\boldsymbol{y}$ is conditionally W-typical with respect to $\boldsymbol{x}$ and some unlikely pure states. For each P-typical $\boldsymbol{x}$, there are about $\exp[nH(W/P)]$ such $\boldsymbol{y}$'s, and they are Q-typical. There are about $\exp[nH(Q)]$ Q-typical $\boldsymbol{y}$'s, and a randomly chosen $\boldsymbol{y}$ will be conditionally W-typical with respect to any P-typical $\boldsymbol{x}$ with probability of about

$$\exp[nH(W/P)]/\exp[nH(Q)] = \exp[-nI(P, W)].$$

Therefore, if Bob forms a list of $\exp[nI(P, W)]$ randomly chosen Q-typical $\boldsymbol{y}$'s, then with high probability there will be a conditionally W-typical $\boldsymbol{y}$ with respect to any P-typical $\boldsymbol{x}$ Alice may have. If Alice and Bob use the same random number generator to form a list, Alice (who knows $\boldsymbol{x}$) can identify such $\boldsymbol{y}$ to Bob by sending about $nI(P, W)$ bits of classical information. Bob can then prepare the corresponding $|\Psi_{\boldsymbol{y}}\rangle\langle\Psi_{\boldsymbol{y}}|$, or an error state if no W-typical $\boldsymbol{y}$ was on the list.

Therefore, for every P-typical $\boldsymbol{x}$, Bob's state $\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}$ is with high probability also a uniform mixture of pure states $|\Psi_{\boldsymbol{y}}\rangle\langle\Psi_{\boldsymbol{y}}|$, where $\boldsymbol{y}$ is conditionally W-typical with respect to $\boldsymbol{x}$ and an unlikely error state.

The idea relies on Shannon's famous observation that "it is possible for most purposes to treat long sequences as though there were just 2^{Hn} of them, each with probability 2^{-Hn}" [12]. The limitations of this "typical sequence" approach becomes apparent when one realizes how stringent the fidelity requirement (4) is. For probability distributions (diagonal density matrices), the fidelity is essentially equivalent to the L_1 distance (see, for example, [15, Ch. 9]). In the scheme sketched above, every sequence on Bob's list of randomly chosen Q-typical $\boldsymbol{y}$'s appears with exactly the same probability. Bob's state $\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}$ is with high probability a uniform mixture of pure states $|\Psi_{\boldsymbol{y}}\rangle\langle\Psi_{\boldsymbol{y}}|$, where $\boldsymbol{y}$ is conditionally W-typical with respect to $\boldsymbol{x}$. Alice's state $\boldsymbol{\rho}_{\boldsymbol{x}}$ is also, with high probability, a mixture of the same pure states $|\Psi_{\boldsymbol{y}}\rangle\langle\Psi_{\boldsymbol{y}}|$, but not exactly uniform.

Thus, for formal proofs, we use a simple refinement of the method of typical sequences, known as the method of types [13], [14]. Two sequences over some alphabet $\mathcal{A}$ have the same type if each letter in $\mathcal{A}$ appears in both of them the same number of times. All sequences of the same type form a type class. We partition the set of typical sequences into type classes. Sequences of the same type are equiprobable for a DMS, and Bob can form a list of sequences randomly chosen from the same type class. Now he will be dealing with a single type class at the time rather than the entire set of typical sequences. He has to know which type class to choose, but Alice can send that information to him at no cost to the compression rate asymptotically since the number of type classes is polynomial in n. When two or more sets of sequences are involved (as $\mathcal{X}^n$ and $\mathcal{Y}^n$ above), joint and conditional types have to be considered.

II. Fidelity of Mixed Quantum States

A. Fidelity and Trace Distance

Besides computing the mixed-state fidelity (4), one can measure how close state σ is to state ω by computing the *trace distance*

$$D(\sigma, \omega) = \frac{1}{2}\,\mathrm{Tr}|\sigma - \omega|.$$

Here $|A|$ denotes the positive square root of $A^\dagger A$, i.e., $|A| = \sqrt{A^\dagger A}$. The trace distance and the fidelity are closely related and the following holds:

$$1 - F(\sigma, \omega) \le D(\sigma, \omega) \le \sqrt{1 - F(\sigma, \omega)^2}. \tag{8}$$

The trace distance is a metric on the space of density operators, and therefore the triangle inequality is true as

$$D(\sigma, \omega) \le D(\sigma, \tau) + D(\tau, \omega). \tag{9}$$

It has some other useful properties, as well. When we need one of those properties, we shall switch from the fidelity to the trace distance and back by making use of the inequalities (8).

Since we shall have to estimate the trace distance of a mixture of inputs, the following property, known as *strong convexity*, will be useful. Let $\{p_i\}$ and $\{q_i\}$ be probability distributions over some index set, and ω_i and σ_i density operators also indexed by the same index set. Then

$$D\left(\sum_i p_i\omega_i, \sum_i q_i\sigma_i\right) \le D(\{p_i\}, \{q_i\}) + \sum_i p_i D(\omega_i, \sigma_i). \tag{10}$$

From strong convexity, it directly follows that the trace distance is jointly convex in its arguments

$$D\left(\sum_i p_i\omega_i, \sum_i p_i\sigma_i\right) \le \sum_i p_i D(\omega_i, \sigma_i). \tag{11}$$

All the above properties of the mixed-state fidelity and trace distance and some additional are discussed in the excellent survey [15, Ch. 9].

B. Approximating Density Matrices

The objective of the compression algorithm described in Section IV-A is to leave Bob with states that faithfully approximate Alice's. Only two types of approximations will be used, which we can already demonstrate by just using the above properties of the fidelity and the trace distance.

Let σ and σ_e be two density matrices, $p_{e,n}$ a sequence of numbers such that $p_{e,n} \to 0$ as $n \to \infty$, and ω_n defined as follows:

$$\omega_n = p_{e,n}\sigma_e + (1 - p_{e,n})\sigma.$$

Lemma 1: Let σ and ω_n be as defined above. Then

$$F(\sigma, \omega_n) \to 1, \qquad \text{as } n \to \infty.$$

Proof: By properties (8), and strong convexity of the trace distance (10), we have

$$\begin{aligned} F(\sigma, \omega_n) &\ge 1 - D(\sigma, \omega_n) \\ &\ge 1 - \frac{1}{2}\,|0 - p_{e,n}| - \frac{1}{2}\,|1 - (1 - p_{e,n})| - D(\sigma, \sigma) \\ &\ge 1 - p_{e,n}. \end{aligned}$$

□

Let $\mathcal{Y}$ be a finite set and $\pi_n \in \mathcal{P}(\mathcal{Y}^n)$ a probability distribution on $\mathcal{Y}^n$. Let $\{\sigma_{\boldsymbol{y}}, \pi_n(\boldsymbol{y}) | \boldsymbol{y} \in \mathcal{Y}^n\}$ be an ensemble of (possibly mixed) states over Hilbert space $\mathcal{H}_d^{\otimes n}$. Consider the following density matrix σ_n:

$$\sigma_n = \sum_{\boldsymbol{y} \in \mathcal{Y}^n} \pi_n(\boldsymbol{y})\sigma_{\boldsymbol{y}}.$$

Let $\mathcal{B}_n \subseteq \mathcal{Y}^n$ be a probabilistically large set: $\pi_n(\mathcal{B}_n) = 1 - \epsilon_n$, where $\epsilon_n \to 0$ as $n \to \infty$. It is intuitively clear that if we replace states $\sigma_{\boldsymbol{y}}, \boldsymbol{y} \in \mathcal{Y}^n \setminus \mathcal{B}_n$, in the expression for σ_n by a fixed state σ_e, we obtain a density matrix which faithfully represents σ_n in the sense of (4) when $n \to \infty$. To prove a slightly stronger result (which we shall use in Section IV-C), we proceed as follows.

Consider

$$\begin{aligned} \sigma_n &= \sum_{\boldsymbol{y} \in \mathcal{Y}^n} \pi_n(\boldsymbol{y})\sigma_{\boldsymbol{y}} \\ &= \sum_{\boldsymbol{y} \in \mathcal{B}_n} \pi_n(\boldsymbol{y})\sigma_{\boldsymbol{y}} + \sum_{\boldsymbol{y} \in \mathcal{Y}^n \setminus \mathcal{B}_n} \pi_n(\boldsymbol{y})\sigma_{\boldsymbol{y}}. \end{aligned}$$

Let $p_{e,n}$ be a sequence of numbers such that $p_{e,n} \to 0$ as $n \to \infty$, and $\hat{\sigma}_{\boldsymbol{y}}$, $\boldsymbol{y} \in \mathcal{Y}^n$, a set of density matrices such that $D(\sigma_{\boldsymbol{y}}, \hat{\sigma}_{\boldsymbol{y}}) \le p_{e,n}$ for all $\boldsymbol{y}$. We define a density matrix ω_n as

$$\omega_n = \sum_{\boldsymbol{y} \in \mathcal{B}_n} \pi_n(\boldsymbol{y}) \hat{\sigma}_{\boldsymbol{y}} + \sum_{\boldsymbol{y} \in \mathcal{Y}^n \setminus \mathcal{B}_n} \pi_n(\boldsymbol{y}) \sigma_e.$$

Lemma 2: Let σ_n and ω_n be as defined above. Then

$$F(\sigma_n, \omega_n) \to 1, \qquad \text{as } n \to \infty.$$

Proof: By properties (8), and joint convexity of the trace distance (11), we have

$$\begin{aligned} F(\sigma_n, \omega_n) &\ge 1 - D(\sigma_n, \omega_n) \\ &\ge 1 - \sum_{\boldsymbol{y} \in \mathcal{B}_n} \pi_n(\boldsymbol{y}) D(\sigma_{\boldsymbol{y}}, \hat{\sigma}_{\boldsymbol{y}}) \\ &\quad + \sum_{\boldsymbol{y} \in \mathcal{Y}^n \setminus \mathcal{B}_n} \pi_n(\boldsymbol{y}) D(\sigma_{\boldsymbol{y}}, \sigma_e) \\ &\ge 1 - (1 - \epsilon_n/2) p_{e,n} - \epsilon_n > 1 - p_{e,n} - \epsilon_n. \end{aligned}$$ □

III. The Method of Types

A. Types and Typical Sequences

Let, as before, $\mathcal{X}$ be a finite set and $\mathcal{P}(\mathcal{X})$ the set of all probability distributions on $\mathcal{X}$. Given a sequence

$$\boldsymbol{x} = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$$

and a letter $a \in \mathcal{X}$, let $N(a|\boldsymbol{x})$ denote the number occurrences of a in $\boldsymbol{x}$.

Definition 1: The *type* of a sequence $\boldsymbol{x} \in \mathcal{X}^n$ is the distribution $P_{\boldsymbol{x}} \in \mathcal{P}(\mathcal{X})$ given by

$$P_{\boldsymbol{x}}(a) = \frac{1}{n} N(a|\boldsymbol{x}), \qquad \text{for every } a \in \mathcal{X}.$$

Conversely, the *type class* of a distribution $P \in \mathcal{P}(\mathcal{X})$ is the set T_P^n of all sequences of type P in $\mathcal{X}^n$

$$\mathsf{T}_P^n = \{\boldsymbol{x}\colon \boldsymbol{x} \in \mathcal{X}^n \text{ and } P_{\boldsymbol{x}} = P\}.$$

The subset of $\mathcal{P}(\mathcal{X})$ consisting of the possible types of sequences $\boldsymbol{x} \in \mathcal{X}^n$ is denoted by $\mathcal{P}_n(\mathcal{X})$. It is easy to show by elementary combinatorics that

$$|\mathcal{P}_n(\mathcal{X})| = \binom{n + |\mathcal{X}| - 1}{|\mathcal{X}| - 1} \le (n+1)^{|\mathcal{X}|}.$$

Therefore, there is only a polynomial (in n) number of types.

The size of T_P^n can be bounded as follows.

Lemma 3 [14, p. 30]: For any type $P_{\boldsymbol{x}}$ of sequences in $\mathcal{X}^n$

$$(n+1)^{-|\mathcal{X}|} \exp\{nH(P_{\boldsymbol{x}})\} \le |\mathsf{T}_{P_{\boldsymbol{x}}}| \le \exp\{nH(P_{\boldsymbol{x}})\}.$$

Definition 2: For any distribution P on $\mathcal{X}$, a sequence $\boldsymbol{x} \in \mathcal{X}^n$ is *P-typical* with constant δ if

$$\left| \frac{1}{n} N(a|\boldsymbol{x}) - P(a) \right| \le \delta, \qquad \text{for every } a \in \mathcal{X}$$

and no $a \in \mathcal{X}$ with $P(a) = 0$ occurs in x. The set of such sequences will be denoted by $\mathsf{T}_{P,\delta}^n$, and the set of their types by $\mathcal{P}_n^{P,\delta}(\mathcal{X})$.

Lemma 4 [14, p. 34]: For any distribution P on $\mathcal{X}$, we have

$$P^n\left(\mathsf{T}_{P,\delta}^n\right) \ge 1 - \frac{|\mathcal{X}|}{4n\delta^2}. \tag{12}$$

B. Joint and Conditional Types

If $\mathcal{X}$ and $\mathcal{Y}$ are two finite sets, the *joint type* of a pair of sequences $\boldsymbol{x} \in \mathcal{X}^n$ and $\boldsymbol{y} \in \mathcal{Y}^n$ is defined as a type of the sequence

$$\{(x_1, y_1), \ldots, (x_n, y_n)\} \in \mathcal{X} \times \mathcal{Y}.$$

Namely, it is the distribution $P_{\boldsymbol{x},\boldsymbol{y}} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ given by

$$P_{\boldsymbol{x},\boldsymbol{y}}(a, b) = \frac{1}{n} N(a, b|\boldsymbol{x}, \boldsymbol{y}), \qquad \text{for every } a \in \mathcal{X},\ b \in \mathcal{Y}.$$

Joint types are often given in terms of the type of $\boldsymbol{x}$ and a stochastic matrix $V\colon \mathcal{X} \to \mathcal{Y}$ as

$$P_{\boldsymbol{x},\boldsymbol{y}}(a, b) = P_{\boldsymbol{x}}(a) V(b|a), \qquad \text{for every } a \in \mathcal{X},\ b \in \mathcal{Y}.$$

Definition 3: We say that $\boldsymbol{y} \in \mathcal{Y}^n$ has *conditional type* V given $\boldsymbol{x} \in \mathcal{X}^n$ if

$$N(a, b|\boldsymbol{x}, \boldsymbol{y}) = N(a|\boldsymbol{x}) V(b|a), \quad \text{for every } a \in \mathcal{X},\ b \in \mathcal{Y}.$$

For any given $\boldsymbol{x} \in \mathcal{X}^n$ and a stochastic matrix $V\colon \mathcal{X} \to \mathcal{Y}$, the set of sequences $\boldsymbol{y} \in \mathcal{Y}^n$ having conditional type V given $\boldsymbol{x}$ is called a *V-shell* of $\boldsymbol{x}$, and is denoted by $\mathsf{T}_V^n(\boldsymbol{x})$ or simply by $\mathsf{T}_V(\boldsymbol{x})$. The set of all conditional types of $\boldsymbol{y} \in \mathcal{Y}$ for a given $\boldsymbol{x}$ will be denoted by $\mathcal{V}_n(\mathcal{Y}, \boldsymbol{x})$.

The size of a V-shell can be bounded as follows.

Lemma 5 [14, p. 31]: For any type $P_{\boldsymbol{x}}$ of sequences in $\mathcal{X}^n$ and stochastic matrix V such that $\mathsf{T}_V(\boldsymbol{x})$ is not empty

$$(n+1)^{-|\mathcal{X}||\mathcal{Y}|} \exp\{nH(V|P_{\boldsymbol{x}})\} \le |\mathsf{T}_V(\boldsymbol{x})| \le \exp\{nH(V|P_{\boldsymbol{x}})\}.$$

Clearly, every $\boldsymbol{y}$ in the V-shell of an $\boldsymbol{x}$ in the type class $\mathsf{T}_{P_{\boldsymbol{x}}}^n$ has the same type $P_{\boldsymbol{y}}$

$$P_{\boldsymbol{y}}(b) = \sum_{a \in \mathcal{X}} P_{\boldsymbol{x}}(a) V(b|a).$$

However, by Lemmas 3 and 5, we immediately see that $\mathsf{T}_V(\boldsymbol{x})$ is "exponentially smaller" than T_P, unless all rows of V are equal to $P_{\boldsymbol{y}}$

$$\begin{aligned} (n+1)^{-|\mathcal{X}||\mathcal{Y}|} &\exp\{-nI(P_{\boldsymbol{x}}, V)\} \\ &\le \frac{|\mathsf{T}_V(\boldsymbol{x})|}{|\mathsf{T}_{P_{\boldsymbol{y}}}|} \le (n+1)^{|\mathcal{Y}|} \exp\{-nI(P_{\boldsymbol{x}}, V)\}. \end{aligned} \tag{13}$$

Definition 4: For any given $\boldsymbol{x} \in \mathcal{X}^n$ and a stochastic matrix $W\colon \mathcal{X} \to \mathcal{Y}$, the sequence $\boldsymbol{y} \in \mathcal{Y}^n$ is W-generated by $\boldsymbol{x}$ (or W-typical under the condition $\boldsymbol{x}$) with constant δ' if

$$\left| \frac{1}{n} N(a, b|\boldsymbol{x}, \boldsymbol{y}) - \frac{1}{n} N(a|\boldsymbol{x}) W(b|a) \right| \le \delta', \qquad \text{for every } a \in \mathcal{X},\ b \in \mathcal{Y}$$

and $N(a, b|\boldsymbol{x}, \boldsymbol{y}) = 0$ whenever $W(b|a) = 0$. The set of such sequences will be denoted by $\mathsf{T}_{W,\delta'}^n(\boldsymbol{x})$, and the set of their conditional types by $\mathcal{V}_n^{W,\delta'}(\mathcal{Y}, \boldsymbol{x})$.

Lemma 6 [14, p. 34]: For any stochastic matrix $W\colon \mathcal{X} \to \mathcal{Y}$, we have

$$W^n\left(\mathsf{T}^n_{W,\delta'}(\boldsymbol{x})|\boldsymbol{x}\right) = 1 - \frac{|\mathcal{X}||\mathcal{Y}|}{4n\,\delta'^2}.$$

C. Conditional Typical States

Let ρ_a be the density matrix of mixed state a given by (2). We consider $\boldsymbol{\rho}_{\boldsymbol{x}} = \rho_{x_1} \otimes \cdots \otimes \rho_{x_n}$ for $\boldsymbol{x} \in \mathsf{T}^n_{P_{\boldsymbol{x}}}$

$$\begin{aligned}\boldsymbol{\rho}_{\boldsymbol{x}} &= \left(\sum_{b\in\mathcal{Y}} W(b|x_1)\,|\psi_b\rangle\langle\psi_b|\right)\otimes\cdots \\ &\quad\otimes\left(\sum_{b\in\mathcal{Y}} W(b|x_n)\,|\psi_b\rangle\langle\psi_b|\right) \\ &= \sum_{\boldsymbol{y}\in\mathcal{Y}^n} W(y_1|x_1)\cdot\ldots\cdot W(y_n|x_n) \\ &\quad |\psi_{y_1}\rangle\langle\psi_{y_1}|\otimes\cdots\otimes|\psi_{y_n}\rangle\langle\psi_{y_n}| \\ &= \sum_{\boldsymbol{y}\in\mathcal{Y}^n} W^n(\boldsymbol{y}|\boldsymbol{x})\,|\Psi_{\boldsymbol{y}}\rangle\langle\Psi_{\boldsymbol{y}}|\end{aligned}$$

where $W^n(\boldsymbol{y}|\boldsymbol{x})$ denotes $W(y_1|x_1)\cdot\ldots\cdot W(y_n|x_n)$ and $|\Psi_{\boldsymbol{y}}\rangle\langle\Psi_{\boldsymbol{y}}|$ denotes $|\psi_{y_1}\rangle\langle\psi_{y_1}|\otimes\cdots\otimes|\psi_{y_n}\rangle\langle\psi_{y_n}|$. We define partial density matrices $\boldsymbol{\rho}_{\boldsymbol{x}}(V)$ corresponding to each V-shell in $\mathcal{V}_n(\mathcal{Y}, \boldsymbol{x})$ as follows:

$$\boldsymbol{\rho}_{\boldsymbol{x}}(V) = \sum_{\boldsymbol{y}\in\mathsf{T}_V(\boldsymbol{x})} \frac{1}{|\mathsf{T}_V(\boldsymbol{x})|}\,|\Psi_{\boldsymbol{y}}\rangle\langle\Psi_{\boldsymbol{y}}|.$$

Now we can write $\boldsymbol{\rho}_{\boldsymbol{x}}$ as

$$\begin{aligned}\boldsymbol{\rho}_{\boldsymbol{x}} = &\sum_{V\in\mathcal{V}_n^{W,\delta'}(\mathcal{Y},\boldsymbol{x})} W^n(\mathsf{T}_V(\boldsymbol{x})|\boldsymbol{x})\,\boldsymbol{\rho}_{\boldsymbol{x}}(V) \\ &+ \sum_{V\in\mathcal{V}_n(\mathcal{Y},\boldsymbol{x})\setminus\mathcal{V}_n^{W,\delta'}(\mathcal{Y},\boldsymbol{x})} W^n(\mathsf{T}_V(\boldsymbol{x})|\boldsymbol{x})\,\boldsymbol{\rho}_{\boldsymbol{x}}(V) \quad (14)\end{aligned}$$

where the first term includes only the conditionally typical V-shells (W-generated by $\boldsymbol{x}$), and the second term takes care of the rest.

We are now ready to describe a mixed-state compression algorithm. We shall see that for every typical $\boldsymbol{x}$, the algorithm leaves Bob with a mixed state $\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}$ which differs from Alice's $\boldsymbol{\rho}_{\boldsymbol{x}}$ of (14) only in the following. In the first term of (14), $\boldsymbol{\rho}_{\boldsymbol{x}}(V)$ is approximated by $\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}(V)$ in the sense of Lemma 1, in the second term of (14), $\boldsymbol{\rho}_{\boldsymbol{x}}(V)$ is simply replaced by some fixed error state $\rho_{e,\boldsymbol{x}}$. Consequently, $\boldsymbol{\rho}_{\boldsymbol{x}}$ is approximated by $\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}$ in the sense of Lemma 2.

IV. Mixed-State Compression

A. The Algorithm

Alice and Bob have access to the same source of random numbers.

1) Alice is given a visible source sequence $\boldsymbol{x} \in \mathcal{X}^n$.
2) For every $a \in \mathcal{X}$, Alice determines $N(a|\boldsymbol{x})$, i.e., the type $P_{\boldsymbol{x}}$ of sequence $\boldsymbol{x}$.
3) If $P_{\boldsymbol{x}}$ is not in $\mathcal{P}^{P,\delta}_n(\mathcal{X})$, i.e., $\boldsymbol{x}$ is not P-typical with constant δ, Alice sends an error indicator, and Bob prepares some fixed error-state ρ_e. Otherwise, they proceed as follows.
4) Alice chooses a conditional type for sequence $\boldsymbol{y}$, say V, at random with probability $W^n(\mathsf{T}_V(\boldsymbol{x}))$. If V is not in $\mathcal{V}^{W,\delta'}_n(\mathcal{Y}, \boldsymbol{x})$, i.e., $\boldsymbol{y}$ is not W-generated by $\boldsymbol{x}$ with constant δ', Alice sends an error indicator, and Bob prepares some fixed error-state $\rho_{e,\boldsymbol{x}}$. (Here $\rho_{e,\boldsymbol{x}}$ and $\rho_{e,\boldsymbol{x}}(V)$ below do not depend on $\boldsymbol{x}$ and V since Bob does not have that information. The notation signifies the stage in the algorithm.) Otherwise, they proceed as follows.
5) Alice determines type $P_{\boldsymbol{y}}$ by computing
$$P_{\boldsymbol{y}}(b) = \sum_{a\in\mathcal{X}} P_{\boldsymbol{x}}(a)V(b|a).$$
6) Alice tells the type $P_{\boldsymbol{y}}$ to Bob by sending $\log|\mathcal{P}_n(\mathcal{Y})|$ bits identifying the particular $P_{\boldsymbol{y}}$.
7) Alice and Bob each form a list of N_l sequences $\boldsymbol{y}$ by drawing randomly from the type class $\mathsf{T}_{P_{\boldsymbol{y}}}$. Let
$$R = \frac{\log N_l}{n}.$$
8) If there is one or more $\boldsymbol{y}$'s on the list belonging to the V-shell $\mathsf{T}_V(\boldsymbol{x})$, Alice sends $\log N_l$ bits to Bob identifying the position of first $\boldsymbol{y} \in \mathsf{T}_V(\boldsymbol{x})$ on the list, and Bob prepares $|\Psi_{\boldsymbol{y}}\rangle\langle\Psi_{\boldsymbol{y}}|$. With some probability $p_{e,\boldsymbol{x}}(V)$, no $\boldsymbol{y} \in \mathsf{T}_V(\boldsymbol{x})$ will be on the list that Alice and Bob form. If that is the case, Alice sends an error indicator and Bob prepares some fixed error-state $\rho_{e,\boldsymbol{x}}(V)$.

B. Bob's Density Matrix

For nontypical $\boldsymbol{x}$, Bob's state is ρ_e, while for typical $\boldsymbol{x}$, his state is given by

$$\begin{aligned}\hat{\boldsymbol{\rho}}_{\boldsymbol{x}} = &\sum_{V\in\mathcal{V}_n^{W,\delta'}(\mathcal{Y},\boldsymbol{x})} W^n(\mathsf{T}_V(\boldsymbol{x})|\boldsymbol{x})\,\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}(V) \\ &+ \sum_{V\in\mathcal{V}_n(\mathcal{Y},\boldsymbol{x})\setminus\mathcal{V}_n^{W,\delta'}(\mathcal{Y},\boldsymbol{x})} W^n(\mathsf{T}_V(\boldsymbol{x})|\boldsymbol{x})\,\rho_{e,\boldsymbol{x}}, \qquad \boldsymbol{x}\in\mathsf{T}^n_{P,\delta}.\end{aligned}$$

Here $\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}(V)$ denotes Bob's density matrix when conditional type V is chosen by Alice. Since Bob prepares either the error-state $\rho_{e,\boldsymbol{x}}(V)$ with probability $p_{e,\boldsymbol{x}}(V)$, or one of the states $|\Psi_{\boldsymbol{y}}\rangle\langle\Psi_{\boldsymbol{y}}|$, $\boldsymbol{y}\in\mathsf{T}_V(\boldsymbol{x})$, with probability $1-p_{e,\boldsymbol{x}}(V)$, we have

$$\begin{aligned}\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}(V) = &\,p_{e,\boldsymbol{x}}(V)\rho_{e,\boldsymbol{x}}(V) \\ &+(1-p_{e,\boldsymbol{x}}(V))\sum_{\boldsymbol{y}\in\mathsf{T}_V(\boldsymbol{x})}\frac{1}{|\mathsf{T}_V(\boldsymbol{x})|}\,|\Psi_{\boldsymbol{y}}\rangle\langle\Psi_{\boldsymbol{y}}|.\end{aligned}$$

Note that if $p_{e,\boldsymbol{x}}(V) \to 0$ as $n \to \infty$, then Bob's $\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}(V)$ approximates Alice's $\boldsymbol{\rho}_{\boldsymbol{x}}(V)$ in the sense of Lemma 1, and thus Bob's $\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}$ approximates Alice's $\boldsymbol{\rho}_{\boldsymbol{x}}$ in the sense of Lemma 2.

To see under which conditions $p_{e,\boldsymbol{x}}(V) \to 0$ as $n \to \infty$, we proceed as follows. Clearly, the probability that a sequence $\boldsymbol{y}$ randomly drawn from $\mathsf{T}_{P_{\boldsymbol{y}}}$ is in $\mathsf{T}_V(\boldsymbol{x})$ equals to $|\mathsf{T}_V(\boldsymbol{x})|/|\mathsf{T}_{P_{\boldsymbol{y}}}|$. The probability $p_{e,\boldsymbol{x}}(V)$ that no such sequence

is on the list of length N_l is thus equal to $(1-|\mathsf{T}_V(\boldsymbol{x})|/|\mathsf{T}_{P_{\boldsymbol{y}}}|)^{N_l}$. This quantity can be bounded by applying the inequality

$$(1-x)^k \le e^{-kx}$$

and then the ratio $|\mathsf{T}_V(\boldsymbol{x})|/|\mathsf{T}_{P_{\boldsymbol{y}}}|$ can be bound by applying the inequalities (13)

$$\begin{aligned} p_{e,\boldsymbol{x}}(V) &= \left(1-|\mathsf{T}_V(\boldsymbol{x})|/|\mathsf{T}_{P_{\boldsymbol{y}}}|\right)^{N_l} \\ &\le e^{-N_l|\mathsf{T}_V(\boldsymbol{x})|/|\mathsf{T}_{P_{\boldsymbol{y}}}|} \\ &\le e^{-\exp(n(R-I-\epsilon_n''))} \end{aligned}$$

where I refers to $I(P_{\boldsymbol{x}}, V)$ and $\epsilon_n'' = |\mathcal{X}||\mathcal{Y}|\log(n+1)/n$. Therefore, if

$$R > I(P_{\boldsymbol{x}}, V) + \epsilon_n'' \tag{15}$$

we have $p_{e,\boldsymbol{x}}(V) \to 0$ as $n \to \infty$.

C. Mixed-State Fidelity

We now have all we need to bound the value of

$$\sum_{\boldsymbol{x}\in\mathcal{X}^n} P(\boldsymbol{x})F(\boldsymbol{\rho}_{\boldsymbol{x}}, \hat{\boldsymbol{\rho}}_{\boldsymbol{x}})$$

and thus prove the main result of the compression algorithm.

Theorem 1: Let $R > I(P_{\boldsymbol{x}}, V) + \epsilon_n''$, for all $P_{\boldsymbol{x}} \in \mathcal{P}_n(\mathcal{Y})$ and all $V \in \mathcal{V}_n^{W,\delta'}(\mathcal{Y}, \boldsymbol{x})$. Then

$$\sum_{\boldsymbol{x}\in\mathcal{X}^n} P(\boldsymbol{x})F(\boldsymbol{\rho}_{\boldsymbol{x}}, \hat{\boldsymbol{\rho}}_{\boldsymbol{x}}) \to 1, \qquad \text{as } n \to \infty.$$

Proof: By Lemma 4

$$\begin{aligned} &\sum_{\boldsymbol{x}\in\mathcal{X}^n} P(\boldsymbol{x})F(\boldsymbol{\rho}_{\boldsymbol{x}}, \hat{\boldsymbol{\rho}}_{\boldsymbol{x}}) \\ &\quad\ge 1 - \sum_{\boldsymbol{x}\in\mathcal{X}^n} P(\boldsymbol{x})D(\boldsymbol{\rho}_{\boldsymbol{x}}, \hat{\boldsymbol{\rho}}_{\boldsymbol{x}}) \\ &\quad= 1 - \sum_{\boldsymbol{x}\in\mathsf{T}_{P,\delta}^n} P(\boldsymbol{x})D(\boldsymbol{\rho}_{\boldsymbol{x}}, \hat{\boldsymbol{\rho}}_{\boldsymbol{x}}) - \sum_{\boldsymbol{x}\in\mathcal{X}^n\setminus\mathsf{T}_{P,\delta}^n} P(\boldsymbol{x})D(\boldsymbol{\rho}_{\boldsymbol{x}}, \rho_e) \\ &\quad\ge 1 - \sum_{\boldsymbol{x}\in\mathsf{T}_{P,\delta}^n} P(\boldsymbol{x})D(\boldsymbol{\rho}_{\boldsymbol{x}}, \hat{\boldsymbol{\rho}}_{\boldsymbol{x}}) - \epsilon_n \end{aligned} \tag{16}$$

where $\epsilon_n = |\mathcal{X}|/(2n\delta^2)$. By Lemma 6

$$\begin{aligned} &D(\boldsymbol{\rho}_{\boldsymbol{x}}, \hat{\boldsymbol{\rho}}_{\boldsymbol{x}}) \\ &\quad= \sum_{V\in\mathcal{V}_n^{W,\delta'}(\mathcal{Y},\boldsymbol{x})} W^n\left(\mathsf{T}_V(\boldsymbol{x})|\boldsymbol{x}\right) D\left(\boldsymbol{\rho}_{\boldsymbol{x}}(V), \hat{\boldsymbol{\rho}}_{\boldsymbol{x}}(V)\right) \\ &\qquad+ \sum_{V\in\mathcal{V}_n(\mathcal{Y},\boldsymbol{x})\setminus\mathcal{V}_n^{W,\delta'}(\mathcal{Y},\boldsymbol{x})} W^n\left(\mathsf{T}_V(\boldsymbol{x})|\boldsymbol{x}\right) \\ &\qquad\cdot D\left(\boldsymbol{\rho}_{\boldsymbol{x}}(V), \hat{\rho}_{e,\boldsymbol{x}}\right) \\ &\quad\le \sum_{V\in\mathcal{V}_n^{W,\delta'}(\mathcal{Y},\boldsymbol{x})} W^n(\mathsf{T}_V(\boldsymbol{x})|\boldsymbol{x}) \\ &\qquad\cdot D\left(\boldsymbol{\rho}_{\boldsymbol{x}}(V)\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}(V)\right) + \epsilon_n' \end{aligned} \tag{17}$$

where $\epsilon_n' = |\mathcal{X}||\mathcal{Y}|/(2n\delta'^2)$. By Lemma 1

$$D\left(\boldsymbol{\rho}_{\boldsymbol{x}}(V), \hat{\boldsymbol{\rho}}_{\boldsymbol{x}}(V)\right) \le p_{e,\boldsymbol{x}}(V). \tag{18}$$

Let $p_{e,n}$ denote the maximum of all $p_{e,\boldsymbol{x}}(V)$ over all $\boldsymbol{x} \in \mathsf{T}_{P,\delta}^n$ and $V \in \mathcal{V}_n^{W,\delta'}(\mathcal{Y}, \boldsymbol{x})$. Combining (16)–(18), we obtain

$$\begin{aligned} &\sum_{\boldsymbol{x}\in\mathcal{X}^n} P(\boldsymbol{x})F(\boldsymbol{\rho}_{\boldsymbol{x}}, \hat{\boldsymbol{\rho}}_{\boldsymbol{x}}) \\ &\quad\ge 1 - (1-\epsilon_n)(1-\epsilon_n')p_{e,n} - (1-\epsilon_n)\epsilon_n' - \epsilon_n \\ &\quad> 1 - p_{e,n} - \epsilon_n' - \epsilon_n. \end{aligned}$$

As $n\to\infty$, we know that $\epsilon_n \to 0$ and $\epsilon_n' \to 0$, whereas $p_{e,n} \to 0$ when the compression rate satisfies (15) for all $\boldsymbol{x} \in \mathsf{T}_{P,\delta}^n$ and $V \in \mathcal{V}_n^{W,\delta'}(\mathcal{Y}, \boldsymbol{x})$. Therefore, under the conditions of the Theorem, we have

$$\sum_{\boldsymbol{x}\in\mathcal{X}^n} P(\boldsymbol{x})F(\boldsymbol{\rho}_{\boldsymbol{x}}, \hat{\boldsymbol{\rho}}_{\boldsymbol{x}}) \to 1, \qquad \text{as } n \to \infty. \qquad \square$$

D. Achievable Compression Rate

To show that a compression rate of $I(P, W)$ is achievable, we use the continuity of entropy.

Lemma 7: If $\{p_i\}_{i=1}^N$ and $\{q_i\}_{i=1}^N$ are two probability distributions such that

$$\sum_{i=1}^N |p_i - q_i| \le \theta \le \tfrac{1}{2}$$

then

$$|H(p_1, \ldots, p_N) - H(q_1, \ldots, q_N)| \le -\theta\log\frac{\theta}{N}.$$

We show that for all $\boldsymbol{x} \in \mathsf{T}_{P,\delta}^n$ and $V \in \mathcal{V}_n^{W,\delta'}(\mathcal{Y}, \boldsymbol{x})$

$$|I(P, W) - I(P_{\boldsymbol{x}}, V)| \to 0, \qquad \text{as } \delta, \delta' \to 0.$$

Consider

$$\begin{aligned} &|I(P, W) - I(P_{\boldsymbol{x}}, V)| \\ &\quad\le |H(Q) - H(P_{\boldsymbol{y}})| + |H(W|P) - H(V|P_{\boldsymbol{x}})| \\ &\quad\le |H(Q) - H(P_{\boldsymbol{y}})| + |H(W|P) - H(W|P_{\boldsymbol{x}})| \\ &\qquad+ |H(W|P_{\boldsymbol{x}}) - H(V|P_{\boldsymbol{x}})| \\ &\quad\le -|\mathcal{X}||\mathcal{Y}|(\delta+\delta')\log[(\delta+\delta')|\mathcal{X}|] \\ &\qquad+ \delta\log|\mathcal{Y}| - |\mathcal{X}||\mathcal{Y}|\delta'\log\delta'. \end{aligned} \tag{19}$$

To bound the first and the third term in (19), we used the continuity of entropy (Lemma 7), and to bound the second term, we used the $\log|\mathcal{Y}|$ bound on the entropy of any distribution over $\mathcal{Y}$.

V. Applications

A. The Example of Fig. 1

We consider the system of Example 1 as shown in Fig. 1 Let h denote the binary entropy function

$$h(x) = -x\log(x) - (1-x)\log(1-x).$$

For the classical information measures, we have

$$\begin{aligned} H(Q) &= \log 3 \\ H(P/W) &= h(1/3) = -2/3 + \log 3 \\ I(P, W) &= 2/3. \end{aligned}$$

For the quantum information measures, we have

$$S(\rho)=1$$
$$\overline{S}=h\left(1/2-\sqrt{3}/6\right)$$
$$\chi=1-h\left(1/2-\sqrt{3}/6\right)=0.255\cdots.$$

Notice the gap between the achievable rate under our assumptions $I(P, W)$ and a lower bound on the compression rate of any visible compression scheme χ.

B. Sources of Mixed States With Commuting Density Operators

When the density matrices ρ_a, $a \in \mathcal{X}$, commute, they can be made diagonal in the same basis. Thus, we shall assume that they are mixtures of orthogonal pure states $|\psi_b\rangle\langle\psi_b|$, $b \in \mathcal{Y}$

$$\rho_a=\sum_{b\in\mathcal{Y}} W(b|a)\,|\psi_b\rangle\langle\psi_b|, \qquad a\in\mathcal{X}.$$

Recall that

$$\rho=\sum_{a\in\mathcal{X}} P(a)\rho_a=\sum_{b\in\mathcal{Y}} Q(b)|\psi_b\rangle\langle\psi_b|.$$

Since $|\psi_b\rangle\langle\psi_b|$ are orthogonal, we have $S(\rho)=H(Q)$, and

$$\sum_{a\in\mathcal{X}} P(a)S(\rho_a)=H(W/P).$$

Therefore, the Holevo quantity χ is in this case equal to the mutual information $I(P, W)$

$$\chi=S(\rho)-\sum_{a\in\mathcal{X}} P(a)S(\rho_a)=H(Q)-H(W/P)=I(P, W).$$

A way to ensure that Bob's matrices $\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}$ commute is to assign the uniform mixture of pure states $|\Psi_{\boldsymbol{y}}\rangle\langle\Psi_{\boldsymbol{y}}|$, $\boldsymbol{y}\in\mathcal{Y}^n$, to each error state in the compression algorithm

$$\rho_e=\rho_{e,\boldsymbol{x}}=\rho_{e,n}=\frac{1}{|\mathcal{Y}|^n}\sum_{\boldsymbol{y}\in\mathcal{Y}^n}|\Psi_{\boldsymbol{y}}\rangle\langle\Psi_{\boldsymbol{y}}|. \tag{20}$$

Of course, no particular choice of the error states is required if the only goal is an asymptotically good fidelity. However, commutativity of Bob's matrices keeps the entire system classical, makes it easier to derive an expression for the fidelity, and consequently puts us in a good position to recognize possible related problems of classical information theory.

For each sequence $\boldsymbol{x}$, Alice's density matrix is

$$\boldsymbol{\rho}_{\boldsymbol{x}}=\sum_{\boldsymbol{y}\in\mathcal{Y}^n} W^n(\boldsymbol{y}|\boldsymbol{x})\,|\Psi_{\boldsymbol{y}}\rangle\langle\Psi_{\boldsymbol{y}}|.$$

With assignment (20), the compression algorithm leaves Bob with the density matrix

$$\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}=\sum_{\boldsymbol{y}\in\mathcal{Y}^n} \hat{W}^n(\boldsymbol{y}|\boldsymbol{x})\,|\Psi_{\boldsymbol{y}}\rangle\langle\Psi_{\boldsymbol{y}}|.$$

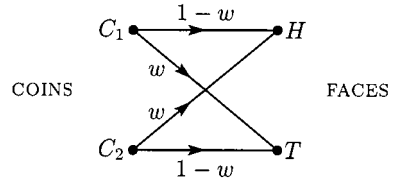


Fig. 2. Source of coins (sources).

Therefore, the mixed-state fidelity between $\boldsymbol{\rho}_{\boldsymbol{x}}$ and $\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}$ is

$$\begin{aligned}F(\boldsymbol{\rho}_{\boldsymbol{x}},\hat{\boldsymbol{\rho}}_{\boldsymbol{x}})&=\left\{\mathrm{Tr}\left[\left(\sqrt{\boldsymbol{\rho}_{\boldsymbol{x}}}\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}\sqrt{\boldsymbol{\rho}_{\boldsymbol{x}}}\right)^{1/2}\right]\right\}^2=\left\{\mathrm{Tr}\left[\left(\boldsymbol{\rho}_{\boldsymbol{x}}\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}\right)^{1/2}\right]\right\}^2\\&=\left\{\mathrm{Tr}\left[\left(\sum_{\boldsymbol{y}\in\mathcal{Y}^n} W^n(\boldsymbol{y}|\boldsymbol{x})\,\hat{W}^n(\boldsymbol{y}|\boldsymbol{x})\,|\Psi_{\boldsymbol{y}}\rangle\langle\Psi_{\boldsymbol{y}}|\right)^{1/2}\right]\right\}^2\\&=\left[\sum_{\boldsymbol{y}\in\mathcal{Y}^n}\sqrt{W^n(\boldsymbol{y}|\boldsymbol{x})\cdot\hat{W}^n(\boldsymbol{y}|\boldsymbol{x})}\right]^2.\end{aligned} \tag{21}$$

The asymptotically perfect fidelity (4) can be achieved by running the compression algorithm described in Section IV. Note that the achievable rate under our assumptions, $I(P, W)$, and a lower bound on the compression rate of any visible compression scheme χ coincide. Thus, the proposed algorithm results in optimal compression.

VI. Connections With Classical Problems

We discuss three problems of classical information theory, each to a certain degree related to the problem of visible mixed-state compression.

A. Sources of Probability Distributions

We consider a discrete memoryless source whose alphabet is a set of $|\mathcal{X}|$ coins with $|\mathcal{Y}|$ faces. When coin C_a is tossed, face b appears with probability $W(b|a)$, $a\in\mathcal{X}$, $b\in\mathcal{Y}$. The source, Alice, produces sequences of coins, i.e., probability distributions, where each coin is drawn independently according to the probability distribution P. A source whose alphabet consists of two probability distribution is described in the following example:

Example 2: A source of two biased coins is shown in Fig. 2. If coin C_1 is tossed, the probability of getting a tail is w, if coin C_2 is tossed, the probability of getting a head is w.

When the n coins in Alice's sequence $C_{\boldsymbol{x}}=\{C_{x_1},\ldots,C_{x_n}\}$ are tossed, the probability of getting sequence $\boldsymbol{y}$ of faces is

$$W^n(\boldsymbol{y}|\boldsymbol{x})=W(y_1|x_1)\cdot\ldots\cdot W(y_n|x_n).$$

Each time Alice is given a sequence of coins $C_{\boldsymbol{x}}$, the reproducing source Bob prepares a sequence of faces $\boldsymbol{y}$ with probability $\hat{W}^n(\boldsymbol{y}|\boldsymbol{x})$ such that

$$\sum_{\boldsymbol{x}\in\mathcal{X}^n} P(\boldsymbol{x})F_{\mathcal{Y}^n}\left(W^n(\cdot|\boldsymbol{x}),\hat{W}^n(\cdot|\boldsymbol{x})\right)\to 1, \qquad \text{as } n\to\infty. \tag{22}$$

Here, $F_{\mathcal{Y}^n}(\cdot,\cdot)$ is the Bhattacharyya–Wooters overlap between two probability distributions over the set $\mathcal{Y}^n$

$$F_{\mathcal{Y}^n}\left(W^n(\cdot|\boldsymbol{x}),\hat{W}^n(\cdot|\boldsymbol{x})\right) = \left[\sum_{\boldsymbol{y}\in\mathcal{Y}^n}\sqrt{W^n(\boldsymbol{y}|\boldsymbol{x})\cdot\hat{W}^n(\boldsymbol{y}|\boldsymbol{x})}\right]^2. \quad (23)$$

Requirement (22) with the fidelity measure (23) ensures that Alice and Bob appear to be identical sources of probability distributions to an observer who can see only the sequences of faces at both ends. More precisely, with probability approaching 1 as n increases, such observer cannot tell the difference between Alice and Bob. Notice that the fidelity measure (23) coincides with the mixed-state fidelity (21) in the commutative quantum case. Consequently, the asymptotically perfect fidelity (22) can be achieved by running the compression algorithm of Section IV. Again the achievable rate $I(P, W)$ and a lower bound on the compression rate of any visible compression scheme χ coincide, and the proposed algorithm results in optimal compression.

B. Type Covering

We again consider the source of the previous section, whose alphabet is a set of $|\mathcal{X}|$ coins with $|\mathcal{Y}|$ faces. But now, for each Alice's sequence $C_{\boldsymbol{x}}$ of coins, Bob prepares a predetermined sequence $\boldsymbol{y}(\boldsymbol{x})$ of faces such that

$$\sum_{\boldsymbol{x}\in\mathcal{X}^n}P(\boldsymbol{x})F_{\mathcal{X}\times\mathcal{Y}}\left(P_{\boldsymbol{x}}W(\cdot|\cdot),P_{\boldsymbol{x},\boldsymbol{y}(\boldsymbol{x})}\right)\to 1 \quad (24)$$

as $n\to\infty$. Here $F_{\mathcal{X}\times\mathcal{Y}}(\cdot,\cdot)$ is the Bhattacharyya–Wooters overlap between two probability distributions over the set $\mathcal{X}\times\mathcal{Y}$

$$\begin{aligned}F_{\mathcal{X}\times\mathcal{Y}}&\left(P_{\boldsymbol{x}}W(\cdot|\cdot),P_{\boldsymbol{x},\boldsymbol{y}(\boldsymbol{x})}\right) \quad (25)\\ &=\left[\sum_{(a,b)\in\mathcal{X}\times\mathcal{Y}}\sqrt{P_{\boldsymbol{x}}(a)W(b|a)\cdot P_{\boldsymbol{x},\boldsymbol{y}(\boldsymbol{x})}(a,b)}\right]^2\\ &=\left[\sum_{(a,b)\in\mathcal{X}\times\mathcal{Y}}\frac{1}{n}\sqrt{N(a|\boldsymbol{x})W(b|a)\cdot N(a,b|\boldsymbol{x},\boldsymbol{y}(\boldsymbol{x}))}\right]^2.\end{aligned}$$

Roughly speaking, Bob prepares a sequence $\boldsymbol{y}(\boldsymbol{x})$ of faces which is conditionally typical given Alice's sequence $C_{\boldsymbol{x}}$ of coins.

The motivation for this problem comes again from compressing sources of quantum states with commuting density operators. Only now the goal is not to achieve the asymptotically perfect fidelity (5), but rather to enable Bob to prepare a state $|\Psi_{\boldsymbol{y}(\boldsymbol{x})}\rangle\langle\Psi_{\boldsymbol{y}(\boldsymbol{x})}|$ which could have been a result of a measurement on $\rho_{\boldsymbol{x}}$. More on the motivation and formulation can be found in [8], where this problem was translated into a rate distortion one, and achievable rates for perfect and imperfect asymptotic fidelity were found. We here use a different method to show that $I(P, W)$ is not only achievable but also the minimum rate for perfect asymptotic fidelity (24). The techniques we use are simple and combinatorial, and only basic information theory presented in Section III is required. Note that the results on the minimum compression rate of [5] and [6] cannot be used here since they were derived for (different) perfect asymptotic fidelity (5).

We first show that the overlap (25) is close to 1 if and only if $\boldsymbol{y}(\boldsymbol{x})$ is W-generated by $\boldsymbol{x}$ with some constant δ close to 0. We prove this claim in the following two lemmas by using the inequalities (8), which bound the fidelity in terms of the trace distance and *vice versa*.

Lemma 8: Let $\boldsymbol{y}(\boldsymbol{x})$ be W-generated by $\boldsymbol{x}$ with constant δ

$$\left|\frac{1}{n}N(a,b|\boldsymbol{x},\boldsymbol{y})-\frac{1}{n}N(a|\boldsymbol{x})W(b|a)\right|\le\delta, \quad \text{for every } a\in\mathcal{X},\ b\in\mathcal{Y}. \quad (26)$$

Then

$$F_{\mathcal{X}\times\mathcal{Y}}\left(P_{\boldsymbol{x}}W(\cdot|\cdot),P_{\boldsymbol{x},\boldsymbol{y}(\boldsymbol{x})}\right)\to 1, \qquad \text{as } \delta\to 0.$$

Proof: By (8), we can bound the Bhattacharyya–Wooters overlap (25) in terms of the corresponding trace distance

$$\begin{aligned}&F_{\mathcal{X}\times\mathcal{Y}}\left(P_{\boldsymbol{x}}W(\cdot|\cdot),P_{\boldsymbol{x},\boldsymbol{y}(\boldsymbol{x})}\right)\\ &=\left[\sum_{(a,b)\in\mathcal{X}\times\mathcal{Y}}\frac{1}{n}\sqrt{N(a|\boldsymbol{x})W(b|a)\cdot N(a,b|\boldsymbol{x},\boldsymbol{y}(\boldsymbol{x}))}\right]^2\\ &\ge 1-\frac{1}{2}\sum_{(a,b)\in\mathcal{X}\times\mathcal{Y}}\left|\frac{1}{n}N(a,b|\boldsymbol{x},\boldsymbol{y})-\frac{1}{n}N(a|\boldsymbol{x})W(b|a)\right|.\end{aligned}$$

Because of (26), we have

$$F_{\mathcal{X}\times\mathcal{Y}}\left(P_{\boldsymbol{x}}W(\cdot|\cdot),P_{\boldsymbol{x},\boldsymbol{y}(\boldsymbol{x})}\right)\ge 1-\delta\cdot|\mathcal{X}||\mathcal{Y}|/2. \qquad \square$$

Lemma 9: Let Bhattacharyya–Wooters overlap (25) be equal to $1-\alpha/2$. Then the sequence $\boldsymbol{y}(\boldsymbol{x})$ is W-generated by $\boldsymbol{x}$ with a constant $2\sqrt{\alpha}$.

Proof: By (8), we can bound the trace distance between the distributions $P_{\boldsymbol{x}}W(\cdot|\cdot)$ and $P_{\boldsymbol{x},\boldsymbol{y}(\boldsymbol{x})}$ in terms of their Bhattacharyya–Wooters overlap (25)

$$\begin{aligned}\frac{1}{2}\sum_{(a,b)\in\mathcal{X}\times\mathcal{Y}}&\left|\frac{1}{n}N(a,b|\boldsymbol{x},\boldsymbol{y})-\frac{1}{n}N(a|\boldsymbol{x})W(b|a)\right|\\ &\le\left[1-F^2_{\mathcal{X}\times\mathcal{Y}}\left(P_{\boldsymbol{x}}W(\cdot|\cdot)P_{\boldsymbol{x},\boldsymbol{y}(\boldsymbol{x})}\right)\right]^{1/2}\le\sqrt{\alpha}.\end{aligned}$$

It follows that

$$\left|\frac{1}{n}N(a,b|\boldsymbol{x},\boldsymbol{y})-\frac{1}{n}N(a|\boldsymbol{x})W(b|a)\right|\le 2\sqrt{\alpha}, \quad \text{for every } a\in\mathcal{X},\ b\in\mathcal{Y}. \qquad \square$$

Therefore, for a given $\boldsymbol{x}$, the fidelity (25) is close to 1 if and only if $\boldsymbol{y}(\boldsymbol{x})$ is W-generated by $\boldsymbol{x}$ with some constant δ close to 0. A compression code $\mathcal{C}\subseteq\mathcal{Y}^n$ will have to contain at least one such $\boldsymbol{y}(\boldsymbol{x})$ for each $\boldsymbol{x}\in\mathsf{T}_{P,\delta_{\boldsymbol{x}}}$, as shown next.

Definition 5: We shall say that code $\mathcal{C}\subseteq\mathcal{Y}^n$ of face sequences *covers* set $\mathcal{B}\subseteq\mathcal{X}^n$ of coin sequences with constant

δ if it contains at least one element of $\mathsf{T}^n_{W,\delta}(\boldsymbol{x})$, for each $\boldsymbol{x} \in \mathcal{B}$, i.e., for each $\boldsymbol{x} \in \mathcal{B}$, we have $\mathcal{C} \cap \mathsf{T}^n_{W,\delta}(\boldsymbol{x}) \neq \emptyset$.

Theorem 2: Let $\mathcal{C}$ be a code which covers the set $\mathsf{T}_{P,\delta_{\boldsymbol{x}}}$. For each $\boldsymbol{x} \in \mathcal{X}^n$, let $\boldsymbol{y}(\boldsymbol{x})$ be an element of $\mathcal{C} \cap \mathsf{T}^n_{W,\delta}(\boldsymbol{x})$ if $\boldsymbol{x} \in \mathsf{T}_{P,\delta_{\boldsymbol{x}}}$, and $\boldsymbol{y}(\boldsymbol{x})$ be an arbitrary $\boldsymbol{y}_e \in \mathcal{Y}^n$, otherwise. Then

$$\sum_{\boldsymbol{x}\in\mathcal{X}^n} P(\boldsymbol{x}) F_{\mathcal{X}\times\mathcal{Y}}\left(P_{\boldsymbol{x}} W(\cdot|\cdot), P_{\boldsymbol{x},\boldsymbol{y}(\boldsymbol{x})}\right) \to 1, \qquad \text{as } n \to \infty.$$

Proof: By Lemmas 4 and 8

$$\sum_{\boldsymbol{x}\in\mathcal{X}^n} P(\boldsymbol{x}) F_{\mathcal{X}\times\mathcal{Y}}\left(P_{\boldsymbol{x}} W(\cdot|\cdot), P_{\boldsymbol{x},\boldsymbol{y}(\boldsymbol{x})}\right) \geq \left(1 - \frac{|\mathcal{X}|}{4n\delta_{\boldsymbol{x}}^2}\right)\left(1 - \frac{\delta|\mathcal{X}||\mathcal{Y}|}{2}\right). \quad \square$$

We assume that Alice and Bob both know the compression code $\mathcal{C}$. To identify $\boldsymbol{y}(\boldsymbol{x})$, Alice has to send to Bob $\log|\mathcal{C}|$ bits of classical information. The compression rate is, therefore, given by

$$R = \frac{\log|\mathcal{C}|}{n}$$

and is determined by the size of the smallest code $\mathcal{C}$ that covers $\mathsf{T}_{P,\delta}$. To bound the size of $\mathcal{C}$, we shall use the following simple general result about coverings, known as Johnson–Stein–Lovász theorem (see, for example, [16, p. 322]).

Theorem 3: Let $\boldsymbol{A}$ be a $0-1$ matrix with N rows and M columns. Assume that each row contains at least v ones and each column at most a ones. Then there exists an $N \times K$ submatrix $\boldsymbol{C}$ of $\boldsymbol{A}$ with

$$K \leq \frac{N}{a} + \frac{M}{v}\log a \leq \frac{M}{v}(1 + \log a)$$

such that $\boldsymbol{C}$ contains no all-zero rows.

We now have all we need to prove the main result of this section.

Theorem 4: The asymptotic compression rate of $I(P, W)$ is both achievable and necessary for perfect asymptotic fidelity (24).

Proof: In order to define a covering code $\mathcal{C}$ and use Theorem 3 to bound its rate R, we construct matrix $\boldsymbol{A}$ as follows. The rows of $\boldsymbol{A}$ are indexed by sequences $\boldsymbol{x}$ that are P-typical with constant $\delta_{\boldsymbol{x}}$, columns by sequences $\boldsymbol{y}$ that are Q-typical with constant $\delta_{\boldsymbol{y}}$. Thus, $\boldsymbol{A}$ has $|\mathsf{T}_{P,\delta_{\boldsymbol{x}}}|$ rows and $|\mathsf{T}_{Q,\delta_{\boldsymbol{y}}}|$ columns. An element of $\boldsymbol{A}$ in row $\boldsymbol{x}$ and column $\boldsymbol{y}$ is set to 1 if $\boldsymbol{x}$ and $\boldsymbol{y}$ are jointly typical with constant $\delta_{\boldsymbol{xy}}$, i.e., if

$$\left|P(a)W(b|a) - \frac{1}{n}N(a, b|\boldsymbol{x}, \boldsymbol{y})\right| = \left|Q(a)U(b|a) - \frac{1}{n}N(a, b|\boldsymbol{x}, \boldsymbol{y})\right| \leq \delta_{\boldsymbol{xy}}$$

otherwise to 0. We first show that all $\boldsymbol{y}$'s corresponding to the 1's in a particular row $\boldsymbol{x}$ are W-generated by $\boldsymbol{x}$ with constant $\delta = \delta_{\boldsymbol{x}} + \delta_{\boldsymbol{xy}}$. For each row $\boldsymbol{x}$ having a 1 in column $\boldsymbol{y}$, we have

$$\begin{aligned}\left|P_{\boldsymbol{x}}(b)W(b|a) - \frac{1}{n}N(a, b|\boldsymbol{x},\boldsymbol{y})\right| &\leq |P_{\boldsymbol{x}}(b)W(b|a) - P(a)W(b|a)| \\ &\quad + \left|P(a)W(b|a) - \frac{1}{n}N(a, b|\boldsymbol{x},\boldsymbol{y})\right| \\ &\leq \delta_{\boldsymbol{x}} + \delta_{\boldsymbol{xy}} = \delta.\end{aligned}$$

Therefore, $\boldsymbol{y} \in \mathsf{T}^n_{W,\delta}(\boldsymbol{x})$. Since $\boldsymbol{C}$ is a submatrix of $\boldsymbol{A}$ with no all-zero rows, the set $\mathcal{C}$ of sequences $\boldsymbol{y}$ indexing the columns of $\boldsymbol{C}$ covers the set $\mathsf{T}^n_{P,\delta_{\boldsymbol{x}}}$. Therefore, by Theorem 2, $\mathcal{C}$ can serve as a compression code asymptotically achieving perfect fidelity.

We find v, a lower bound on the number of 1's in each row as follows. For each $\boldsymbol{x}$, consider all sequences $\boldsymbol{y}$ which are W-generated by $\boldsymbol{x}$ with constant δ'. If δ' is set to be equal to $\delta_{\boldsymbol{xy}} - \delta_{\boldsymbol{x}}$, we have

$$\begin{aligned}\left|P(a)W(b|a) - \frac{1}{n}N(a, b|\boldsymbol{x}, \boldsymbol{y})\right| &\leq |P(a)W(b|a) - P_{\boldsymbol{x}}(a)W(b|a)| \\ &\quad + \left|P_{\boldsymbol{x}}(a)W(b|a) - \frac{1}{n}N(a, b|\boldsymbol{x}, \boldsymbol{y})\right| \\ &\leq \delta_{\boldsymbol{x}} + \delta' = \delta_{\boldsymbol{xy}}.\end{aligned}$$

Thus, if $\boldsymbol{y} \in \mathsf{T}^n_{W,\delta'}(\boldsymbol{x})$, the element of $\boldsymbol{A}$ in row $\boldsymbol{x}$ and column $\boldsymbol{y}$ is set to 1. Therefore, the number of 1's in each row is at least v

$$v = \exp\left[n\left(H(W/P) - \epsilon'\right)\right].$$

We find a, an upper bound on the number of 1's in each column as follows. For each column $\boldsymbol{y}$ having a 1 in row $\boldsymbol{x}$, we have

$$\begin{aligned}\left|P_{\boldsymbol{y}}(b)U(a|b) - \frac{1}{n}N(a, b|\boldsymbol{x}, \boldsymbol{y})\right| &\leq |P_{\boldsymbol{y}}(b)U(a|b) - Q(b)U(a|b)| \\ &\quad + \left|Q(b)U(a|b) - \frac{1}{n}N(a, b|\boldsymbol{x}, \boldsymbol{y})\right| \\ &\leq \delta_{\boldsymbol{y}} + \delta_{\boldsymbol{xy}}.\end{aligned}$$

Therefore, $\boldsymbol{x} \in \mathsf{T}^n_{U,\delta''}(\boldsymbol{y})$, $\delta'' = \delta_{\boldsymbol{y}} + \delta_{\boldsymbol{xy}}$, and thus the number of 1's in each column is at most a

$$a = \exp\left[n\left(H(U/Q) + \epsilon''\right)\right].$$

Theorem 3 gives an upper bound on K, the number of columns in $\mathcal{C}$ and thus the code rate

$$R = \frac{\log K}{n}.$$

Since $M = |\mathsf{T}^n_{Q,\delta_{\boldsymbol{y}}}| \leq \exp[n(H(Q) + \epsilon_{\boldsymbol{y}})]$, we have

$$\begin{aligned}K &\leq \frac{M}{v}(1 + \log a) \\ &\leq \exp\left[n\left(H(Q) - H(W/P) + \epsilon_{\boldsymbol{y}} + \epsilon'\right)\right] \\ &\quad \cdot \left[1 + n\left(H(U/Q) + \epsilon''\right)\right].\end{aligned}$$

Now, for any $N \times K$ submatrix of $\boldsymbol{A}$ with no all-zero rows, we have $K \cdot a \geq N \cdot 1$, and thus,

$$K \geq \frac{N}{a} \geq \exp\left[n\left(H(P) - H\left(U/Q\right) + \epsilon_{\boldsymbol{x}} - \epsilon''\right)\right].$$

Therefore, the compression R is bounded by

$$I(P, W) + \epsilon_{\boldsymbol{x}} - \epsilon'' \leq R \leq I(P, W) + \epsilon_{\boldsymbol{y}} + \epsilon' + \frac{\log\left[1 + n\left(H\left(U/Q\right) + \epsilon''\right)\right]}{n}$$

where $\epsilon_{\boldsymbol{x}}$, $\epsilon_{\boldsymbol{y}}$, ϵ', $\epsilon'' \to 0$, as $n \to \infty$, and the compression rate $I(P, W)$ is asymptotically optimal. □

Let us now compare the compression problem in this section with the earlier one in Section VI-A. In the earlier case, for each Alice's sequence of coins $\boldsymbol{x}$, Bob most likely chooses one of the approximately $\exp\{nH(W/P)\}$ sequences of faces W-generated by $\boldsymbol{x}$, each one with roughly the same probability. In the case we just considered, for each Alice's $\boldsymbol{x}$, Bob's sequence of faces will always be a fixed sequence $\boldsymbol{y}(\boldsymbol{x})$, W-generated by $\boldsymbol{x}$. Note that in both cases, after a $\boldsymbol{y} \in \mathsf{T}^n_{W,\delta}(\boldsymbol{x})$ has been identified for Bob, his uncertainty about $\boldsymbol{x}$ reduces from $H(P)$ to $H(U/Q)$; hence the same compression rate.

To an observer who can see only the sequences of faces at both ends, Alice and Bob now do not appear to be identical sources for any rate of compression R smaller than the entropy $H(Q)$: Bob has about $\exp(nR)$ different, equally likely face sequences of length n whereas Alice has about $\exp(nH(Q))$. For each Alice's $\boldsymbol{x}$, for quantum transmission or storage, Bob can prepare the quantum state $|\Psi_{\boldsymbol{y}}\rangle\langle\Psi_{\boldsymbol{y}}|$ instead of the sequence of faces $\boldsymbol{y}$. In the scenario of Section VI-A, his state is roughly a uniform mixture of pure states $|\Psi_{\boldsymbol{y}}\rangle\langle\Psi_{\boldsymbol{y}}|$ where each $\boldsymbol{y}$ is W-generated by $\boldsymbol{x}$, whereas in the scenario of this section, his state is the pure state $|\Psi_{\boldsymbol{y}(\boldsymbol{x})}\rangle\langle\Psi_{\boldsymbol{y}(\boldsymbol{x})}|$.

C. Lossy Mixed-State Compression and Channel Coding

Since the encoder Alice knows the source sequence $\boldsymbol{x}$, she can with no loss pass this information to Bob by sending roughly $H(P)$ bits of classical information. Bob can then prepare $\boldsymbol{\rho}_{\boldsymbol{x}}$ in his lab, thus achieving perfect asymptotic fidelity (5). From classical data compression results, we know that if Alice tries to compress $\boldsymbol{x}$ at a rate smaller than $H(P)$, Bob will incorrectly decode it with probability approaching 1. To find out what a loss in the average mixed-state fidelity (5) would be at compression rates below $H(P)$, the problem was formulated as a rate-distortion one. We give a proof that at such compression rates, the average mixed-state fidelity (5) approaches 0, and point out the connection between quantum mixed-state compression and classical channel coding.

The Bhattacharyya distance is in classical information theory most commonly known for its role in bounding the error probability of a discrete memoryless channel (DMC). Consider a DMC with input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$, and transition probabilities $W(b|a)$, $a \in \mathcal{X}$, $b \in \mathcal{Y}$. When the sequence $\boldsymbol{x} \in \mathcal{X}^n$ has been transmitted, the probability that the maximum-likelihood detector finds sequence $\boldsymbol{x}' \in \mathcal{X}^n$ more likely is smaller than

$$\sum_{\boldsymbol{y} \in \mathcal{Y}^n} \sqrt{W^n(\boldsymbol{y}|\boldsymbol{x})\, W^n(\boldsymbol{y}|\boldsymbol{x}')}.$$

This bound is known as the *Bhattacharyya bound* and its negative logarithm as the Bhattacharyya distance between sequences $\boldsymbol{x}$ and $\boldsymbol{x}'$. The probability of error for the maximum-likelihood decoder can then be bounded in terms of the rate and the length of the channel code used for transmission. One way to derive such bound is by solving a special rate-distortion problem. We state the problem below and describe its connection with particular lossy mixed-state compression. For its application to deriving error bounds for channel coding, we refer the reader to [17] or textbooks [14, pp. 185, 193] and [18, pp. 408–410].

Consider a lossy mixed-state compression problem where both the original source Alice and the reproduction source Bob have the same alphabet $\mathcal{X}$. Alice's sequences range over $\mathcal{X}^n$, whereas Bob's sequences are limited to some subset $\mathcal{C}$ of $\mathcal{X}^n$, called a *reproduction* code. The fidelity between sequences $\boldsymbol{x}$ and $\boldsymbol{x}'$ is the Bhattacharyya–Wooters overlap between $W^n(\cdot|\boldsymbol{x})$ and $W^n(\cdot|\boldsymbol{x}')$

$$F(\boldsymbol{x}, \boldsymbol{x}') = \left[\sum_{\boldsymbol{y} \in \mathcal{Y}^n} \sqrt{W^n(\boldsymbol{y}|\boldsymbol{x})\, W^n(\boldsymbol{y}|\boldsymbol{x}')}\right]^2 = \prod_{i=1}^{n} \left[\sum_{b \in \mathcal{Y}} \sqrt{W(b|x_i)\, W(b|x'_i)}\right]^2. \quad (27)$$

Let $\mathcal{C} \subseteq \mathcal{X}^n$ be a reproduction code. We encode source sequence $\boldsymbol{x} \in \mathcal{X}^n$ by choosing the codeword $\hat{\boldsymbol{x}}$ which maximizes the fidelity $F(\boldsymbol{x}, \hat{\boldsymbol{x}})$. Let $F(\boldsymbol{x}|\mathcal{C})$ denote this maximum fidelity

$$F(\boldsymbol{x}|\mathcal{C}) = \max_{\hat{\boldsymbol{x}} \in \mathcal{C}} F(\boldsymbol{x}, \hat{\boldsymbol{x}})$$

and $F(\mathcal{C})$, the expected fidelity achieved with code $\mathcal{C}$

$$F(\mathcal{C}) = \sum_{\boldsymbol{x} \in \mathcal{X}^n} P(\boldsymbol{x}) F(\boldsymbol{x}|\mathcal{C}).$$

We are interested in finding out how the fidelity $F(\mathcal{C})$ depends on the rate of code $\mathcal{C}$.

We get the answer to the question through the following rate-distortion problem. Let again the source and reproduction alphabet be $\mathcal{X}$. Note that the fidelity (27) is single-letter multiplicative and should approach 1, and recall that a distortion measure should be single-letter additive and approach 0. Define a single-letter distortion measure between a source letter a and a reproduction letter a' to be the logarithm of the Bhattacharyya distance between the letters as follows:

$$d_W(a, a') = -\log \sum_{b \in \mathcal{Y}} \sqrt{W(b|a)\, W(b|a')}, \qquad a, a' \in \mathcal{X}.$$

To make the distortion finite, we shall assume that any two coins have at least one common face, i.e., for all a, $a' \in \mathcal{X}$, there is a $b \in \mathcal{Y}$ such that $W(b|a) > 0$ and $W(b|a') > 0$. Thus, we have

$$0 \leq d_W(a, a') \leq d_0, \qquad a, a' \in \mathcal{X}.$$

Because of our assumption that W has no identical rows, $d_W(a, a') = 0$ iff $a = a'$.

The distortion between sequences is the average of the per-letter distortion between elements of the sequences

$$\begin{aligned} d_W(\boldsymbol{x}, \boldsymbol{x}') &= \frac{1}{n}\sum_{i=1}^{n} d(x_i, x_i') \\ &= -\frac{1}{n}\log\prod_{i=1}^{n}\left[\sum_{b\in\mathcal{Y}}\sqrt{W(b|x_i)\,W(b|x_i')}\right] \\ &= -\frac{1}{n}\log\sum_{\boldsymbol{y}\in\mathcal{Y}^n}\sqrt{W^n(\boldsymbol{y}|\boldsymbol{x})\,W^n(\boldsymbol{y}|\boldsymbol{x}')}, \\ & \qquad\qquad \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}^n. \end{aligned}$$

The fidelity (27) is, therefore, given by

$$F(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-2n d_W(\boldsymbol{x}, \boldsymbol{x}')\right).$$

Note that if the distortion between two sequences remains strictly positive as n increases, the fidelity between them approaches 0.

Let $\mathcal{C} \subseteq \mathcal{X}^n$ be a reproduction code. We encode source sequence $\boldsymbol{x} \in \mathcal{X}^n$ by choosing the codeword $\hat{\boldsymbol{x}}$ that minimizes the distortion $d(\boldsymbol{x}, \hat{\boldsymbol{x}})$. Let $d(\boldsymbol{x}|\mathcal{C})$ denote this minimum distortion

$$d(\boldsymbol{x}|\mathcal{C}) = \min_{\hat{\boldsymbol{x}}\in\mathcal{C}} d(\boldsymbol{x}, \hat{\boldsymbol{x}})$$

and $d(\mathcal{C})$, the expected distortion achieved with code $\mathcal{C}$

$$d(\mathcal{C}) = \sum_{\boldsymbol{x}\in\mathcal{X}^n} P(\boldsymbol{x}) d(\boldsymbol{x}, \mathcal{C}).$$

Let V be an $|\mathcal{X}|\times|\mathcal{X}|$ stochastic matrix with elements $V_{aa'} = V(a'|a)$, $a, a' \in \mathcal{X}$, and let

$$d(V) = \sum_{a, a'\in\mathcal{X}} P(a) V(a'|a)\, d(a, a')$$

be the *average distortion* associated with V. The *rate-distortion function* of a DMS with generic distribution P is given by

$$R(D) = \max_{V:\, d(V)\le D} I(P, V).$$

Its significance, found by Shannon in [19], is expressed by the source coding theorem and its converse (see [18, pp. 397–400] for the form used here). Before stating the theorem and its application to our problem, we compute the distortion measure $d_W(\cdot, \cdot)$ and the rate distortion function for the source of Example 2.

Example 3: Consider again the source shown in Fig. 2. We have

$$d_W(C_1, C_2) = -\log\sqrt{4p(1-p)}$$

and

$$d_W(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{n} D_H(\boldsymbol{x}, \boldsymbol{x}') \cdot d_W(C_1, C_2)$$

where $D_H(\boldsymbol{x}, \boldsymbol{x}')$ is the Hamming distance between sequences $\boldsymbol{x}$ and $\boldsymbol{x}'$. The rate-distortion function is given by

$$R(D) = H(P) - h\left(D/d_W(C_1, C_2)\right).$$

Note that $R(0) = H(P)$, which is true in general under our assumptions.

Theorem 5: [18, pp. 397–400] Source Coding Theorem and its Converse: For any block length n and rate R, there exists a block code $\mathcal{C} \subseteq \mathcal{X}^n$ with average distortion $d(\mathcal{C})$ satisfying

$$d(\mathcal{C}) \le D + d_0 e^{-nE(R, D)}$$

where $E(R, D) > 0$ for $R > R(D)$. Conversely, no source code for which $d(\mathcal{C}) \le D$ has rate smaller than $R(D)$.

We use this result to show how the fidelity $F(\mathcal{C})$ depends on the rate of code $\mathcal{C}$.

Theorem 6: For any block length n and rate $R > H(P)$, a rate R block code $\mathcal{C} \subseteq \mathcal{X}^n$ exists such that the fidelity $F(\mathcal{C}) \to 1$ as $n \to 0$. Conversely, for any code $\mathcal{C} \subseteq \mathcal{X}^n$ with rate $R < H(P)$, $F(\mathcal{C}) \to 0$ as $n \to 0$.

Proof: By the source coding Theorem 5, we have

$$\begin{aligned} F(\mathcal{C}) &= \sum_{\boldsymbol{x}\in\mathcal{X}^n} P(\boldsymbol{x}) F(\boldsymbol{x}|\mathcal{C}) \\ &= \sum_{\boldsymbol{x}\in\mathcal{X}^n} P(\boldsymbol{x}) \exp\left(2n d_W(\boldsymbol{x}|C)\right) \\ &\ge \sum_{\boldsymbol{x}\in\mathcal{X}^n} P(\boldsymbol{x}) \left(1 - n d_W(\boldsymbol{x}|C)\right) \\ &\ge 1 - n d_0 e^{-nE(R, 0)} \end{aligned}$$

where $E(R, 0) > 0$ for $R > R(0) = H(P)$. Therefore, the fidelity can be made arbitrarily close to 1 by increasing the block length n.

By the Converse to the source coding Theorem 5, no code for which $d(\mathcal{C}) \le 0$ has rate smaller then $H(P)$. Thus, for $R < H(P)$, we have $d(\mathcal{C}) = D > 0$. Therefore, the distortion $d(\boldsymbol{x}|\mathcal{C})$ remains strictly positive as n increases for a probabilistically large set of sequences $\boldsymbol{x}$. Consequently, for the same set the fidelity $F(\boldsymbol{x}|\mathcal{C})$ approaches 0. □

VII. Conclusion

We considered the problem of compressing memoryless sources of quantum mixed states by using only classical communication and shared randomness between the encoder Alice and the decoder Bob. Alice knows source sequences, and it is required that the expected Uhlman fidelity between the Alice's and Bob's sequence states be asymptotically high or, equivalently, the expected trace distance be small.

By using convexity properties of the trace distance, we reduced the compression task to a classical information theory problem involving the discrete memoryless channel. Alice, who has a sequence over the channel input alphabet, should send to Bob enough information (bits) so that he can prepare a sequence over the channel output alphabet with roughly the same probability with which it would appear as a channel output if Alice's sequence was really sent over the channel. In this process, they are allowed to use the same random number generator. The compression rate is defined as the ratio between the number of bits necessary to describe Alice's sequence and the number of bits she actually sends. We showed that the mutual information between the channel input and output is an achievable compression rate. For the sources of mixed states with commuting density operators, this achievable rate coincides with the previously known lower bound on the compression rate of any visible compression scheme.

We considered two more classical information theory problems also involving the discrete memoryless channel. The first problem is similar to the one above, but this time Bob is required to prepare not randomly but deterministically a sequence

over the channel output alphabet which is jointly typical with Alice's sequence over the channel input alphabet. The optimal compression rate is again the mutual information between the channel input and output. The second problem asks what would be the loss in fidelity if Alice encoded her sequence into a reproduction code sequence over the same channel input alphabet, and identified the code sequence to Bob, who can then simulate the channel to obtain an output for this sequence. We showed that if the code rate is smaller than the channel input entropy, the asymptotic fidelity is zero.

REFERENCES

[1] B. W. Schumacher, "Quantum coding," *Phys. Rev. A*, vol. 51, pp. 2738–2747, Apr. 1995.

[2] B. W. Schumacher and R. Jozsa, "A new proof of the quantum noiseless coding theorem," *J. Mod. Opt.*, vol. 41, no. 12, pp. 2343–2349, 1994.

[3] H. Barnum, C. A. Fuchs, R. Jozsa, and B. Schumacher, "General fidelity limit for quantum channels," *Phys. Rev. A*, vol. 54, pp. 4707–4711, 1996.

[4] M. Horodecki, "Limits for compression of quantum information carried by ensembles of mixed states," *Phys. Rev. A*, vol. 57, pp. 3364–3369, May 1998.

[5] ——, "Optimal compression for mixed signal states," *Phys. Rev. A*, vol. 61, pp. 1–4, Apr. 2001. [Online]: Paper no. 052309.

[6] H. Barnum, C. M. Caves, C. A. Fuchs, R. Jozsa, and B. Schumacher, "On quantum coding ensembles of mixed states," *J. Phys. A: Math. Gen.*, vol. 34, pp. 6767–6785, 2001.

[7] W. Dür, G. Vidal, and J. I. Cirac, "Visible compression of commuting mixed states," *Phys. Rev. A*, vol. 64, pp. 1–7, 2001. [Online]: Paper no. 022308.

[8] G. Kramer and S. A. Savari, "Quantum data compression of ensembles of mixed states with commuting density operators," arXiv:quant-ph/0101119.

[9] M. Koashi and N. Imoto, "Compressibility of mixed-state signals," arXiv:quant-ph/0103128.

[10] A. Uhlmann, "The 'transition probability' in the state space of a $*$-algebra," *Rep. Math. Phys.*, vol. 9, no. 2, pp. 273–279, 1976.

[11] R. Jozsa, "Fidelity for mixed quantum states," *J. Mod. Opt.*, vol. 41, no. 12, pp. 2315–2323, 1994.

[12] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 10, pp. 379–423, 623–656, Oct. 1948.

[13] I. Csiszár, "The method of types," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2505–2523, Oct. 1998.

[14] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Budapest, Hungary: Académiai Kiadó, 1986.

[15] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*. New York: Cambridge Univ. Press, 2000.

[16] G. Cohen, I. Honkala, S. Litsyn, and A. Lobstein, *Covering Codes*. Amsterdam, The Netherlands: North Holland, 1997.

[17] J. K. Omura, "Expurgated bounds, Bhattacharyya distance, and rate distortion functions," *Inform. Contr.*, vol. 24, pp. 358–383, 1974.

[18] A. J. Viterbi and J. K. Omura, *Principles of Digital Communication and Coding*. New York: McGraw-Hill, 1979.

[19] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," in *IRE Nat. Conv. Rec.*, 1959, Pt. 4, pp. 142–163.

[20] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ: Prentice-Hall, 1971.

[21] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.