Communicating Probability Distributions

Gerhard Kramer, *Member, IEEE*, and Serap A. Savari, *Member, IEEE*

***Abstract*—A rate distortion problem is solved that is motivated by a quantum data compression problem. The goal is to send information about a source string x so that a receiver can construct a second string y for which the joint empirical probability distribution of x and y is close to some desired distribution. The problem differs from the usual rate distortion problems in that one must consider both remote sources and distortion functions that are not averages of per-letter distortion functions.**

***Index Terms*—Data compression, quantum theory, random sequences, rate distortion theory.**

I. Introduction

CONSIDER a discrete memoryless source (DMS) with probability distribution P_X and finite alphabet $\mathcal{X}$. Suppose the source emits a finite sequence $x^L = x_1, x_2, \ldots, x_L$. The empirical probability of letter i in x^L is

$$Q_{x^L}(i) = n_{x^L}(i)/L \tag{1}$$

where $n_{x^L}(i)$ is the number of times the letter i occurs in x^L. Similarly, let y^L be a sequence of letters from the finite alphabet $\mathcal{Y}$. The empirical probability of the pair (i, j) in x^L and y^L is

$$Q_{x^L y^L}(i,j) = n_{x^L y^L}(i,j)/L \tag{2}$$

where $n_{x^L y^L}(i,j)$ is the number of times (i,j) occurs in the sequence of pairs

$$(x_1, y_1), (x_2, y_2), \ldots, (x_L, y_L).$$

We consider the following type of compression problem. Suppose we are given a DMS with distribution P_X, a set of K indices, and a conditional distribution $P_{Z|X}$. The goal is to map any DMS output x^L to an index u in such a way that 1) K is small and 2) one can recover from u a sequence y^L for which $Q_{x^L y^L}$ is "close" to P_{XZ}. The *fidelity* or *distortion* between $Q_{x^L y^L}$ and P_{XZ} can be measured by using various functions, but such functions are inherently *not* averages of per-letter functions (see, e.g., [1]). This suggests that there is no simple (single-letter) solution to the rate distortion problem. However, we are able to find single-letter solutions for natural distortion functions because we are dealing with empirical probabilities.

The above type of problem is described in [2, Sec. 6] where it is called the compression/communication of probability distributions. The problem is motivated by the problem of compressing quantum states, but it also applies to compressing the outcomes of quantum measurements. We discuss these relations in Section II. We remark that we do not consider the quantum data compression problem itself; this problem is addressed in, e.g., [3]–[14] (see also [15]).

This paper is organized as follows. In Section II, we describe the problems that motivated our results. Section III defines what we mean by communicating probability distributions, and gives a solution to the resulting rate distortion problem. Some properties of this solution are developed. In Section IV, we consider examples illustrating the behavior of the solution. Section V concludes the paper.

II. Preliminaries

A. Motivation: Quantum Sources and Measurements

A *quantum source* is a device that emits a sequence of quantum states. We represent this sequence by

$$\sigma^L = \sigma_1 \otimes \sigma_2 \otimes \cdots \otimes \sigma_L \tag{3}$$

where $\otimes$ is the Kronecker (direct, tensor) product operation, and where each σ_ℓ is a density matrix (or density operator), i.e., σ_ℓ is a complex, positive semi-definite, Hermitian matrix with unit trace. Without loss of generality, we restrict the σ_ℓ to be 2×2 matrices or *qubits*. The source density matrix σ^L thus has size $2^L \times 2^L$. We further assume that the quantum source is *memoryless* in the sense that each σ_ℓ is an independent realization of the random matrix Σ. We consider only finite sources for which Σ takes on values in the set $\mathcal{S} = \{\rho_1, \rho_2, \ldots, \rho_M\}$. We will usually refer to ρ_i simply by its index i. Let X be the random variable defined by $\Sigma = \rho_X$ so that we can associate each σ^L with a sequence of integers $x^L = x_1, \ldots, x_L$. We will deal primarily with x^L rather than σ^L, and we solve classical problems, albeit ones motivated by quantum mechanics.

There are several kinds of quantum operations that can be performed on quantum states. We consider only a quantum *measurement*, also known as a positive operator-valued measurement or POVM (see, e.g., [16, p. 90]). A POVM is a list $\mathcal{Q} = \{Q_1, \ldots, Q_N\}$ of positive semi-definite matrices whose sum is the identity matrix, i.e., we have

$$\sum_{n=1}^{N} Q_n = I. \tag{4}$$

Manuscript received August 9, 2002; revised June 14, 2006. The material in this paper was presented at the IEEE International Symposium on Information Theory, Lausanne, Switzerland, June/July2002.

G. Kramer is with Bell Labs, Alcatel-Lucent, Murray Hill NJ 07974 USA (e-mail: gkr@bell-labs.com).

S. A. Savari was with Bell Labs, Lucent Technologies, Murray Hill NJ 07974 USA. She is now with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: savari@eecs.umich.edu).

Communicated by E. Knill, Associate Editor for Quantum Information Theory.

Digital Object Identifier 10.1109/TIT.2006.889015

0018-9448/$25.00 © 2007 IEEE

Suppose the measurement is applied to a quantum state σ. This means that the measurement has an outcome Z where

$$P_Z(n) = \mathrm{Tr}\,(\sigma\, Q_n) \tag{5}$$

for $n = 1, 2, \ldots, N$. One can check that P_Z is a probability distribution because it evaluates to positive numbers that sum to one (see [17, p. 398, Corollary 7.1.5]).

B. Communicating Joint Distributions

The goal of quantum data compression is for the encoder (Alice) to map σ^L into a state υ^B of B qubits such that B is as small as possible and such that the decoder (Bob) can, on average, recover a good approximation $\widehat{\sigma}^L$ to σ^L. For example (see, e.g., [16, p. 409]), one might require the *Bures–Uhlmann* fidelity

$$F(\sigma^L, \widehat{\sigma}^L) = \left\{ \mathrm{Tr}\left[\left(\sqrt{\sigma^L}\, \widehat{\sigma}^L \sqrt{\sigma^L} \right)^{1/2} \right] \right\}^2 \tag{6}$$

to be close to unity on average, i.e., one desires

$$\overline{F} = \sum_{\sigma^L} P(\sigma^L) F(\sigma^L, \widehat{\sigma}^L) \approx 1. \tag{7}$$

We restrict attention to those $\mathcal{S}$ having commuting density matrices, i.e., the $\rho_1, \ldots, \rho_M$ are simultaneously diagonalizable. Barnum *et al.* discuss this case in detail in [2] where they present *classical analogs* of the so-called *visible* and *blind* problems. We summarize these classical problems.

A classical analog for the *visible* problem has the encoder (Alice) having access to the state indices x^L. For the *blind* case she knows only a sequence z^L where z_ℓ is the outcome of a measurement applied to σ_ℓ. The problem for Alice is to map x^L or z^L to and index u so that the decoder (Bob) can, for sufficiently large L, generate from u a sequence y^L that *could have* been measured from σ^L. More precisely, we require that the empirical probability

$$Q_{x^L y^L}(i,j) = \frac{n_{x^L y^L}(i,j)}{L} \tag{8}$$

be "close" to $P_{XZ}(i,j)$ for all i, j. For example (see [2, eq. (10)]), we might require that the Bhattacharryya–Wootters fidelity or overlap

$$F(P_{XZ}, Q_{x^L y^L}) = \left(\sum_{i,j} \sqrt{P_{XZ}(i,j)\, Q_{x^L y^L}(i,j)} \right)^2 \tag{9}$$

to be approximately unity when averaged over x^L and y^L.

We illustrate the problem by considering some "obvious" encoding strategies. First, in the visible scenario, Alice can transmit x^L to Bob at a rate of $H(X)$ bits per qubit, where $H(X)$ denotes the entropy of X (see, e.g., [18, Ch. 2]). Bob subsequently generates a y^L via $P_{Z|X}$, and the resulting $Q_{x^L y^L}$ will be close to P_{XZ} if L is large. Second, in either the visible or blind scenario, Alice can transmit z^L to Bob at a rate of $H(Z)$ bits per qubit, and Bob sets $y^L = z^L$. However, neither of these approaches is optimal in general. For example, suppose that Z is independent of X, and that $H(X) > 0$ and $H(Z) > 0$. The best code has Alice transmit nothing—Bob simply generates y^L using P_Z.

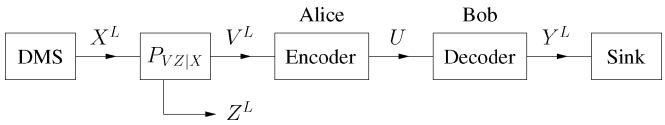


Fig. 1. Model for the problem of communicating probability distributions.

C. Communicating Conditional Distributions

The classical analog described in [2, Sec. 6] can alternatively be interpreted as considering not joint probabilities, as we have done above, but the *conditional* empirical probabilities

$$Q_{y^L|x^L}(j|i) = \frac{n_{x^L y^L}(i,j)}{n_{x^L}(i)}. \tag{10}$$

One might now require the Bhattacharryya–Wootters fidelity

$$F(P_{Z|X}, Q_{y^L|x^L}) = \left(\sum_{i,j:\, P_X(i)>0} \sqrt{P_{Z|X}(j|i)\, Q_{y^L|x^L}(j|i)} \right)^2 \tag{11}$$

to be approximately $|\mathcal{X}|^2$ when averaged over x^L and y^L (recall that $\mathcal{X}$ is the alphabet of X). Note that we deliberately permit sequences x^L with $n_{x^L}(i) = 0$ in (11) to leave open the possibility of choosing $Q_{y^L|x^L}(j|i)$ when $P_X(i) > 0$ but $n_{x^L}(i) = 0$. For instance, one might choose such $Q_{y^L|x^L}(j|i)$ to give $\mathrm{E}[F(P_{Z|X}, Q_{Y^L|X^L})]$ special properties. We defer a treatment of fidelities based on conditional distributions to Appendix C because of this subtlety. However, for large L, the theory for joint and conditional distributions turns out to be similar.

D. Compressing Outcomes of Quantum Measurements

As a different problem, suppose we are given a measurement $\mathcal{Q}$ whose elements are 2×2 matrices. Suppose further that Alice applies $\mathcal{Q}$ to σ^L on a qubit-by-qubit basis. By this we mean that the measurement of σ^L is the list $\mathcal{Q}^{\otimes L}$ consisting of all N^L Kronecker products of L matrices from $\mathcal{Q}$. There are therefore N^L possible outcomes, and outcome $z^L = z_1, z_2, \ldots, z_L$ occurs with probability $\prod_{\ell=1}^{L} P_Z(z_\ell)$ (see [26, Sec. 4.2]).

Suppose that Alice applies $\mathcal{Q}^{\otimes L}$, obtains the outcome sequence z^L, and would like to send her knowledge about σ^L to Bob. It is now natural to consider communicating probability distributions in the sense of Section I. The reason is that it is immaterial whether Alice sends z^L or any other $\widetilde{z}^L$ that could have typically been measured. For example, if Z^L is independent of X^L then Alice need not send Bob anything: Bob can generate typical z^L via P_Z by himself.

III. Problem and Solution

The model we consider is depicted in Fig. 1. A DMS with distribution P_X and finite alphabet $\mathcal{X}$ puts out the sequence x^L. This sequence is applied to a discrete memoryless channel $P_{VZ|X}$, i.e., v_ℓ and z_ℓ are noisy versions of x_ℓ for

$\ell = 1, 2, \ldots, L$. Let the alphabets of v_ℓ and z_ℓ be $\mathcal{V}$ and $\mathcal{Z}$, respectively. The sequence v^L plays different roles depending on the problem: in the visible case, we have $v^L = x^L$ and in the blind case, we have $v^L = z^L$. More generally, v^L might be neither x^L nor z^L. The encoder (Alice) maps v^L to an index u that takes on values in the set $\{1, 2, \ldots, K\}$. The decoder (Bob) maps u to a sequence y^L where the alphabet of y_ℓ is $\mathcal{Z}$ for all ℓ. Note that Y^L is not memoryless in general.

We continue by describing the problem in general terms, and later make our description precise. Alice and Bob's goal is as follows: choose an encoder with small K and for which there is a decoder so that, on average, the joint empirical distribution $Q_{x^L y^L}$ is "close" to P_{XZ}. For example, we might require that $\mathrm{E}[d(Q_{X^L Y^L})]$ is small where $d(\cdot)$ is the Bhattacharyya–Wootters distortion

$$\begin{aligned} d(Q_{x^L y^L}) &= 1 - F(P_{XZ}, Q_{x^L y^L}) \\ &= 1 - \left(\sum_{i,j} \sqrt{P_{XZ}(i,j)\, Q_{x^L y^L}(i,j)} \right)^2 \end{aligned} \tag{12}$$

or the *variational* distortion (see [18, p. 299] and [19, p. 58])

$$d(Q_{x^L y^L}) = \sum_{i,j} \left| P_{XZ}(i,j) - Q_{x^L y^L}(i,j) \right|. \tag{13}$$

Note that we write $d(Q_{x^L y^L})$ rather than $d(P_{XZ}, Q_{x^L y^L})$ for economy of notation. Also note that both (12) and (13) are nonnegative, and they are 0 if and only if $Q_{x^L y^L}(i,j) = P_{XZ}(i,j)$ for all i and j.

We consider distortion functions with the following properties: 1) $d(P_{AB})$ is nonnegative, 2) $d(P_{AB})$ is bounded from above by a number $d_{\max}$, 3) $d(P_{AB})$ is continuous in P_{AB} at P_{XZ} in the sense that for any $\gamma > 0$ there is a $\mu > 0$ such that

$$\sum_{i,j} |P_{XZ}(i,j) - P_{AB}(i,j)| < \mu \tag{14}$$

implies that

$$|d(P_{XZ}) - d(P_{AB})| < \gamma \tag{15}$$

and 4) $d(P_{AB})$ is convex in P_{AB}, i.e., we have

$$\lambda\, d(P_{AB}) + (1-\lambda)\, d(P_{CD}) \geq d(\lambda\, P_{AB} + (1-\lambda)\, P_{CD}). \tag{16}$$

The property (16) is perhaps unusual, but many natural distortion functions including (12) and (13) possess it. The import of convexity is that one can use Jensen's inequality [18, p. 25] to bound the expected distortion as

$$\mathrm{E}\left[d(Q_{X^L Y^L})\right] \geq d\left(\mathrm{E}\left[Q_{X^L Y^L}\right]\right) = d\left(P_{\overline{X}\,\overline{Y}}\right) \tag{17}$$

where $P_{\overline{X}\,\overline{Y}} = \sum_{\ell=1}^{L} P_{X_\ell Y_\ell}/L$ and $P_{X_\ell Y_\ell}$ is the joint probability distribution of (X_ℓ, Y_ℓ). The equality in (17) is a vector version of the following result:

$$\begin{aligned} \mathrm{E}\left[Q_{A^L}(i)\right] &= \sum_{a^L} P_{A^L}(a^L) \frac{n_{a^L}(i)}{L} \\ &= \sum_{a^L} P_{A^L}(a^L) \frac{1}{L} \sum_{\ell=1}^{L} 1(a_\ell = i) \\ &= \frac{1}{L} \sum_{\ell=1}^{L} P_A(i) \end{aligned} \tag{18}$$

where $1(a = b)$ is the indicator function that is 1 if its argument is true and is 0 otherwise. Note that (18) is valid even if A^L has memory. We remark that the choice of $d(\cdot)$ is not especially crucial if one requires that $d(Q_{x^L y^L}) = 0$ if and only if $Q_{x^L y^L} = P_{XZ}$ and if the average distortion must approach zero.

We can now define the problem of communicating probability distributions: for each nonnegative Δ find the infimum (over code books, code book sizes K, and block lengths L) of rates $R = \log_2(K)/L$ bits per symbol such that $\mathrm{E}[d(Q_{X^L Y^L})] \leq \Delta$.

Theorem 1: The rate distortion function for the problem of communicating probability distributions is

$$R(\Delta) = \min_{\substack{P_{Y|V}: \\ d(P_{XY}) \leq \Delta}} I(V;Y) \tag{19}$$

assuming the set of $P_{Y|V}$ satisfying $d(P_{XY}) \leq \Delta$ is nonempty.

Proof: We first show that the right-hand side of (19) is a lower bound to $R(\Delta)$. The other direction of the result, that it is also an upper bound to $R(\Delta)$, is shown in Appendix B by using lemmas listed in Appendix A.

Observe that $H(Y^L) \leq \log_2(K) = LR$. We can further bound $H(Y^L) \geq I(V^L; Y^L)$ by the nonnegativity of $H(Y^L|V^L)$. Expanding $I(V^L; Y^L)$, we have

$$\begin{aligned} LR &\geq \sum_{\ell=1}^{L} H(V_\ell|V^{\ell-1}) - H(V_\ell|Y^L V^{\ell-1}) \\ &= \sum_{\ell=1}^{L} H(V_\ell) - H(V_\ell|Y^L V^{\ell-1}) \\ &\geq \sum_{\ell=1}^{L} I(V_\ell; Y_\ell) \\ &\geq L \cdot I\left(\overline{V}; \overline{Y}\right) \end{aligned} \tag{20}$$

where $P_{\overline{V}\,\overline{Y}} = \sum_{\ell=1}^{L} P_{V_\ell Y_\ell}/L$. The equality in (20) follows because the V_ℓ are independent, the second inequality because conditioning cannot increase entropy, and the third inequality because the mutual information $I(V;Y)$ is convex in $P_{Y|V}$ when P_V is fixed [18, p. 31]. The combination of (20) and (17) shows that $R(\Delta)$ is lower-bounded by the expression on the right-hand side of (19). □

Remark 1: We obtain a single-letter solution even though our sequence distortion is *not* the average of per-letter distortions (see, e.g., [1]).

Remark 2: $R(\Delta)$ is nonincreasing with Δ, and is convex in Δ. To prove convexity, consider the points $(\Delta_1, R(\Delta_1))$ and $(\Delta_2, R(\Delta_2))$ obtained with (19) and the respective channels $P_{Y_1|V}$ and $P_{Y_2|V}$. Consider the mixture

$$P_{Y_3|V} = \lambda P_{Y_1|V} + (1-\lambda) P_{Y_2|V}$$

and let $\Delta_3 = d(P_{XY_3})$. We have

$$\begin{aligned} \Delta_3 &= d\left(\lambda P_{XY_1} + (1-\lambda) P_{XY_2}\right) \\ &\leq \lambda\, d(P_{XY_1}) + (1-\lambda)\, d(P_{XY_2}) \\ &= \lambda\, \Delta_1 + (1-\lambda)\, \Delta_2 \end{aligned}$$

where the inequality follows by (16). Observe that if we increase the distortion from Δ_3 to $\lambda\, \Delta_1 + (1-\lambda)\, \Delta_2$ then we

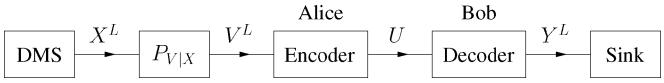


Fig. 2. Model for the Dobrushin/Tsybakov rate distortion problem with a remote source.

can achieve a rate less than or equal to $R(\Delta_3)$. Furthermore, the convexity of mutual information (see, e.g., [18, p. 31]) implies that $I(V;Y_3) \leq \lambda I(V;Y_1)+(1-\lambda)I(V;Y_2)$. Combining these two results, we have

$$\begin{aligned} R(\lambda\,\Delta_1 + (1-\lambda)\,\Delta_2) &\leq R(\Delta_3) \\ &\leq \lambda\,R(\Delta_1) + (1-\lambda)\,R(\Delta_2). \end{aligned}$$

Remark 3: Theorem 1 does not require $V^L = X^L$ or $V^L = Z^L$. Thus, (19) gives the rate distortion function for the general problem where Alice has any kind of partial information V^L about X^L via the memoryless channel $P_{VZ|X}$.

Remark 4: The blind case ($V = Z$) is closely related to a problem solved by Dobrushin and Tsybakov in [20] (see also [21]–[23]). This problem has Alice observing V^L as in Fig. 1. However, now there is no Z^L (see Fig. 2) and the sequence distortion is an average of the per-letter distortions $d(x_\ell, y_\ell)$, $\ell = 1, 2, \ldots, L$. The DMS in this problem is sometimes called a *remote source* [19, p. 136], [24, p. 78], and the rate distortion function is

$$R(\Delta) = \min_{\substack{P_{Y|V}: \\ \mathrm{E}[d(X,Y)] \leq \Delta}} I(V;Y) \tag{21}$$

assuming the set of $P_{Y|V}$ satisfying $\mathrm{E}[d(X,Y)] \leq \Delta$ is nonempty. Observe the similarity of (19) to (21).

IV. Examples

We give two examples to illustrate the behavior of $R(\Delta)$ with the Bhattacharyya–Wootters distortion (12). Consider Example 8 of [2] for which the density operators $\rho_1 = \mathrm{diag}(\alpha_1, 1-\alpha_1)$ and $\rho_2 = \mathrm{diag}(\alpha_2, 1-\alpha_2)$ have prior probabilities p and $1-p$, respectively, where $\mathrm{diag}(a,b)$ is a diagonal matrix with entries a and b. Barnum *et al.* explain in [2] that one may regard the two states as biased coins c_1 and c_2 whose tossing outcomes are H (for heads) with respective probabilities α_1 and α_2, and T (for tails) with respective probabilities $1-\alpha_1$ and $1-\alpha_2$. Adapting this to Fig. 1, we let X^L be the sequence of coins and Z^L a sequence of outcomes of coin tosses, i.e., $P_X(c_1) = p$, $P_{Z|X}(H|c_1) = \alpha_1$, and $P_{Z|X}(H|c_2) = \alpha_2$.

Consider the visible case where $V = X$ so that

$$R(\Delta) = \min_{\substack{P_{Y|X}: \\ d(P_{XY}) \leq \Delta}} I(X;Y). \tag{22}$$

If $\Delta = 0$ then we choose $P_{XY} = P_{XZ}$, and we have

$$\begin{aligned} I(X;Y) &= I(X;Z) \\ &= h\,(p\,\alpha_1 + (1-p)\,\alpha_2) - [p\,h(\alpha_1) + (1-p)\,h(\alpha_2)] \end{aligned}$$

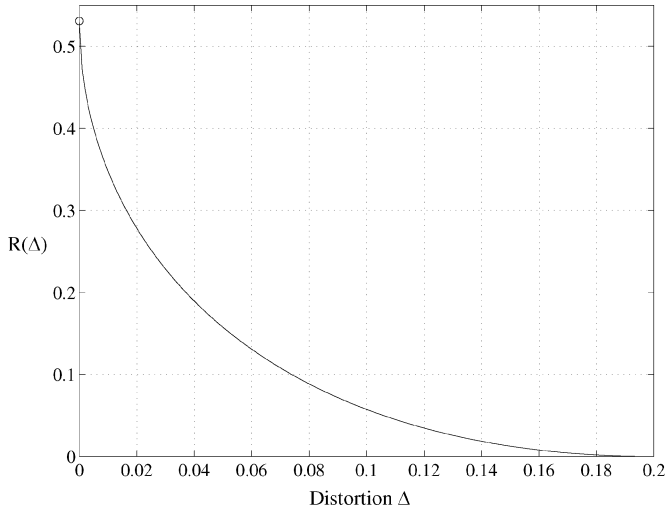


Fig. 3. Rate distortion function for a visible example.

where $h(\alpha) = -\alpha\log_2(\alpha) - (1-\alpha)\log_2(1-\alpha)$ is the binary entropy function. For a concrete example, set $p = 1/2$, $\alpha_1 = 1/10$, and $\alpha_2 = 9/10$. Fig. 3 shows the resulting rate distortion curve $R(\Delta)$. One can compute that $R(0) \approx 0.5310$.

Consider next the blind case where $V = Z$ and

$$R(\Delta) = \min_{\substack{P_{Y|Z}: \\ d(P_{XY}) \leq \Delta}} I(Z;Y). \tag{23}$$

Again, if $\Delta = 0$ then we choose $P_{XY} = P_{XZ}$ and find that $R(0)$ can be a *discontinuous* function of α_2: for $\alpha_2 \neq \alpha_1$ we have $R(0) = H(Z) = h(p\,\alpha_1 + (1-p)\,\alpha_2)$ and for $\alpha_2 = \alpha_1$ we have $R(0) = 0$. For example, suppose that $p = 1/2$ and $\alpha_1 = 1/3$. We plot $R(\Delta)$ as a function of α_2 for various Δ in Fig. 4. Observe that as $\Delta \to 0$, we will have a discontinuity at $\alpha_2 = 1/3$. Furthermore, even for relatively small Δ, e.g., $\Delta = 10^{-3}$, one can achieve substantially better compression rates than $R(0)$ for many α_2.

V. Concluding Remarks

We developed information theory for a problem known as communicating probability distributions. By using a general model, we solved the visible and blind variants of the problem simultaneously, and also included lossy compression.

Appendix A
Properties of Typical Sequences

This appendix reviews standard results that we use to prove Theorem 1. We define the set of *δ-typical sequences* with respect to P_X as (see [25])

$$T_\delta^L(P_X) = \{\, x^L : |Q_{x^L}(i) - P_X(i)| \leq \delta \cdot P_X(i), \text{ for all } i\}. \tag{24}$$

Define the *support set* of X to be $\mathcal{S}_X = \{a \in \mathcal{X} : P_X(a) > 0\}$. Let μ_X be the smallest positive probability of $P_X(\cdot)$.

The set of *jointly* δ-typical sequences with respect to P_{XY} is

$$\begin{aligned} T_\delta^L(P_{XY}) = \{\, (x^L, y^L) : &|Q_{x^L y^L}(i,j) - P_{XY}(i,j)| \\ &\leq \delta \cdot P_{XY}(i,j) \text{ for all } i,j\,\}. \end{aligned} \tag{25}$$

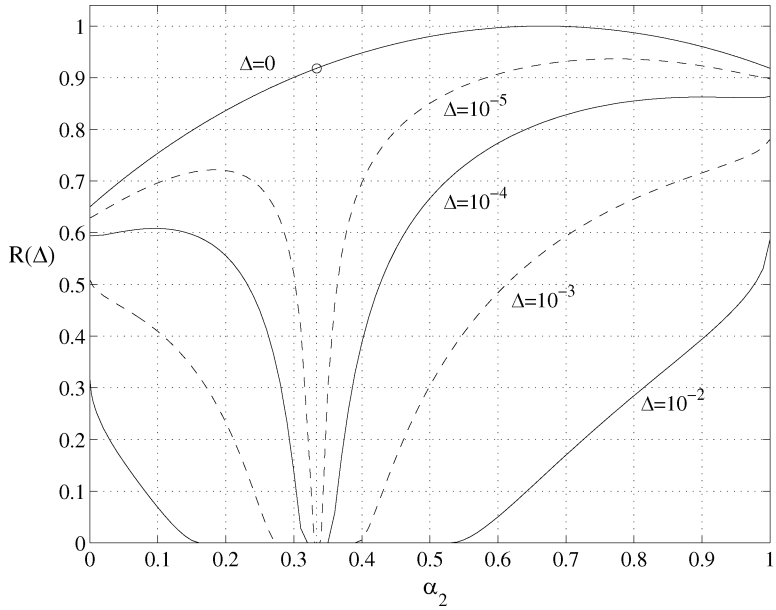


Fig. 4. Rate distortion functions for a blind example.

Define $\mathcal{S}_{X,Y} = \{(a,b) \in \mathcal{X} \times \mathcal{Y} : P_{XY}(a,b) > 0\}$ and let μ_{XY} be the smallest positive probability of $P_{XY}(\cdot)$.

The following properties of typical sequences and sets are proved in the appendix of [25].

Lemma 1: Let $0 < \delta \leq 1$ and

$$\epsilon_\delta(L) = 2|\mathcal{S}_X| e^{-\delta^2 \cdot \mu_X \cdot L/3}. \tag{26}$$

Suppose X^L is put out by the DMS P_X. We have

$$\Pr[X^L \notin T_\delta^L(P_X)] \leq \epsilon_\delta(L). \tag{27}$$

Lemma 2: $T_\delta^L(P_{XY}) \subseteq T_\delta^L(P_X) \times T_\delta^L(P_Y)$.

Lemma 3: Let $0 < \delta_1 < \delta_2 \leq 1$ and

$$\epsilon_{\delta_1,\delta_2}(L) = 2|\mathcal{S}_{X,Y}| e^{-\frac{(\delta_2-\delta_1)^2}{1+\delta_1} \cdot \mu_{X,Y} \cdot L/3}. \tag{28}$$

Suppose Y^L is put out by the DMS P_Y. We have

$$\begin{aligned}&(1-\epsilon_{\delta_1,\delta_2}(L)) \cdot 2^{-L(I(X;Y)+2\delta_2 H(Y))} \\ &\leq \Pr[(x^L, Y^L) \in T_{\delta_2}^L(P_{XY})] \leq 2^{-L(I(X;Y)-2\delta_2 H(Y))}\end{aligned} \tag{29}$$

where the lower bound holds for every $x^L \in T_{\delta_1}^L(P_X)$ and the upper bound holds for every $x^L \in \mathcal{X}^L$.

Lemma 4: Let $W - X - Y$ be a Markov chain. For every $(w^L, x^L) \in T_{\delta_1}^L(P_{WX})$ we have

$$\Pr[(w^L, x^L, Y^L) \in T_{\delta_2}^L(P_{WXY}) | X^L = x^L] \geq 1 - \epsilon_{\delta_1,\delta_2}(L). \tag{30}$$

Appendix B
Proof of Theorem 1

The proof of achievability of the rate (19) follows mostly by standard arguments (cf. [18, Ch. 13]). We begin by choosing a channel $P_{Y|V}$ and compute P_Y. We next generate a code book with K codewords y_k^L, $k = 1, 2, \ldots, K$, by choosing every symbol of each y_k^L independently using P_Y. Given v^L, the encoder tries to find an index k^* such that

$$(v^L, y_{k^*}^L) \in T_\delta^L(P_{VY}). \tag{31}$$

If successful, the encoder transmits k^* to the decoder, otherwise it transmits $k^* = 1$. The decoder puts out the sequence $y_{k^*}^L$.

To analyze the performance, consider the respective typical and encoder-error events

$$\mathcal{T} = \{(X^L, V^L, Z^L) \in T_\delta^L(P_{XVZ})\} \tag{32}$$

$$\mathcal{E} = \bigcap_{k=1}^{K} \{(V^L, Y_k^L) \notin T_{\delta_2}^L(P_{VY})\} \tag{33}$$

where $\delta_2 > \delta$. We bound the average distortion over the ensemble of all code books as

$$\begin{aligned}\mathrm{E}[d(Q_{X^L Y^L})] \leq\ & \left(\Pr[\mathcal{T}^C] + \Pr[\mathcal{T}]\ \Pr[\mathcal{E}|\mathcal{T}]\right) d_{\max} \\ &+ \Pr[\mathcal{T}]\ \Pr[\mathcal{E}^C|\mathcal{T}]\, \mathrm{E}[d(Q_{X^L Y^L})|\mathcal{T} \cap \mathcal{E}^C]\end{aligned} \tag{34}$$

where $\mathcal{E}^C$ is the complement of $\mathcal{E}$.

We proceed to bound the terms in (34). From Lemma 1, we know that $\Pr[\mathcal{T}^C] \leq \epsilon_\delta(L)$ which is small for large L. From Lemma 2, the event $\mathcal{T}$ implies $V^L \in T_\delta^L(P_V)$. Further, from Lemma 3 we find that for any $v^L \in T_\delta^L(P_V)$ and k we have

$$\begin{aligned}&\Pr[(v^L, Y_k^L) \in T_{\delta_2}^L(P_{VY})] \\ &\qquad \geq (1 - \epsilon_{\delta,\delta_2}(L)) \cdot 2^{-L(I(V;Y)+2\delta_2 H(Y))}.\end{aligned} \tag{35}$$

Since the events in the intersection in (33) are statistically independent given $\mathcal{T}$, we have

$$\begin{aligned}\Pr[\mathcal{E}|\mathcal{T}] &\leq \left[1 - (1 - \epsilon_{\delta,\delta_2}(L)) 2^{-L(I(V;Y)+2\delta_2 H(Y))}\right]^K \\ &\leq \exp\left(-(1-\epsilon_{\delta,\delta_2}(L)) 2^{L(R-I(V;Y)-2\delta_2 H(Y))}\right)\end{aligned} \tag{36}$$

where we have used $(1-x)^m \le e^{-mx}$ for $x \le 1$ (cf. [18, p. 353]) and $K = 2^{RL}$. The right-hand side of (36) is small for large L as long as

$$R > I(V;Y) + 2\delta_2 H(Y). \tag{37}$$

Finally, note that $Y-V-X$ forms a Markov chain (see Fig. 1) and the event $\mathcal{E}^C$ implies that there is some k^* with $(v^L, y^L_{k^*}) \in T^L_{\delta_2}(P_{VY})$. From Lemmas 2 and 4, the event $\mathcal{T} \cap \mathcal{E}^C$ implies that, for any δ_3 with $\delta_3 > \delta_2$, we have $(x^L, y^L_{k^*}) \in T^L_{\delta_3}(P_{XY})$ with probability close to one for sufficiently large L. But any such typical (x^L, y^L) has

$$\sum_{i,j} |Q_{x^L y^L}(i,j) - P_{XY}(i,j)| \le \delta_3. \tag{38}$$

Further, because $d(P_{AB})$ is continuous in P_{AB} at P_{XZ}, we also have that for any $\gamma > 0$ there exists a $\delta_3 > 0$ such that (38) implies

$$|d(Q_{x^L y^L}) - d(P_{XY})| \le \gamma. \tag{39}$$

Let $\mathcal{F} = \{(V^L, Y^L_{k^*}, X^L) \in T^L_{\delta_3}(P_{VYX})\}$. We have

$$\begin{aligned} &\mathrm{E}[d(Q_{X^L Y^L})|\mathcal{T} \cap \mathcal{E}^C] \\ &\quad \le \Pr[\mathcal{F}|\mathcal{T} \cap \mathcal{E}^C]\,(d(P_{XY}) + \gamma) + \Pr[\mathcal{F}^C|\mathcal{T} \cap \mathcal{E}^C]\, d_{\max} \\ &\quad \le d(P_{XY}) + \gamma + \epsilon_{\delta_2,\delta_3}(L)\, d_{\max}. \end{aligned} \tag{40}$$

Inserting (36) and (40) into (34), we find that as long as R satisfies (37) we can make the average distortion as close to $d(P_{XY})$ as desired by choosing sufficiently small γ, δ, δ_2, and δ_3, and large enough L. This proves that an ensemble-average of code books can approach the rate and distortion of (19) if there is a $P_{Y|V}$ satisfying $d(P_{XY}) \le \Delta$. Thus, there exists a code book that achieves this rate and distortion. We remark that this achievability result does not require $d(P_{AB})$ to be convex in P_{AB}.

Appendix C
Conditional Empirical Probabilities

As in Section II-C, for conditional distributions we define $d(P_{B|A})$ by using only those $i \in \mathcal{X}$ for which $P_X(i) > 0$. For example, for the Bhattacharryya–Wootters distortion (cf. (12)) we use

$$d\left(P_{B|A}\right) = 1 - \frac{1}{|\mathcal{X}|^2}\left(\sum_{i,j:\, P_X(i)>0} \sqrt{P_{Z|X}(j|i)\, P_{B|A}(j|i)}\right)^2. \tag{41}$$

Similarly, for the variational distortion (cf. (13)) we use

$$d(P_{B|A}) = \sum_{i,j:\, P_X(i)>0} \left|P_{Z|X}(j|i) - P_{B|A}(j|i)\right|. \tag{42}$$

To illustrate the effect of choosing $P_{B|A}(j|i)$ when $P_X(i) > 0$ but $P_A(i) = 0$, observe that to minimize (42) we would choose such $P_{B|A}(j|i)$ to be $P_{Z|X}(j|i)$. (For (41) one might be tempted to choose the ambiguous $P_{B|A}(j|i)$ to be unity. However, the distortion can then become negative which may be undesirable.)

One can prove achievability results for conditional distributions by using the same code books as in Appendix B. The necessary changes begin with (38). In particular, from (24) we find that $0 < \delta < 1$, $x^L \in T^L_\delta(P_X)$, and $P_X(i) > 0$ imply $n_{x^L}(i) > 0$. Similarly, from (25), we find that $0 < \delta < 1$, $(x^L, y^L) \in T^L_\delta(P_{XY})$, and $P_X(i) > 0$ imply

$$\left|\frac{Q_{x^L}(i)}{P_X(i)} Q_{y^L|x^L}(j|i) - P_{Y|X}(j|i)\right| \le \delta P_{Y|X}(j|i) \tag{43}$$

$$\left|\frac{Q_{x^L}(i)}{P_X(i)} - 1\right| \le \delta. \tag{44}$$

The bounds (43) and (44) together imply

$$|Q_{y^L|x^L}(j|i) - P_{Y|X}(j|i)| \le \delta(P_{Y|X}(j|i) + Q_{y^L|x^L}(j|i)). \tag{45}$$

Any typical (x^L, y^L) for $0 < \delta < 1$ therefore has

$$\sum_{i,j:\, P_X(i)>0} |Q_{y^L|x^L}(j|i) - P_{Y|X}(j|i)| \le 2\delta|\mathcal{S}_X|. \tag{46}$$

Now, suppose that $d(P_{B|A})$ is continuous in $P_{B|A}$ at $P_{Z|X}$ in the sense that for any $\gamma > 0$ there exists a $\mu > 0$ such that (cf. (14) and (15))

$$\sum_{i,j:\, P_X(i)>0} |P_{Z|X}(j|i) - P_{B|A}(j|i)| < \mu \tag{47}$$

implies that

$$|d(P_{Z|X}) - d(P_{B|A})| < \gamma \tag{48}$$

for some choice of $P_{B|A}(j|i)$ when $P_X(i) > 0$ but $P_A(i) = 1$. We can then make the distortion close to $d(P_{Y|X})$ by following the same steps as in Appendix B.

The converse is more subtle. We can certainly use (20), and the analog of (17) is

$$\mathrm{E}\left[d(Q_{Y^L|X^L})\right] \ge d\left(\mathrm{E}\left[Q_{Y^L|X^L}\right]\right). \tag{49}$$

We would now like to write

$$\mathrm{E}\left[Q_{Y^L|X^L}\right] = \sum_{\ell=1}^{L} P_{Y_\ell|X_\ell}/L \tag{50}$$

in which case the rate distortion function would be

$$R(\Delta) = \min_{\substack{P_{Y|V}:\\ d(P_{Y|X}) \le \Delta}} I(V;Y). \tag{51}$$

However, there is some freedom in evaluating $\mathrm{E}[Q_{Y^L|X^L}]$ since we can choose $Q_{y^L|x^L}(j|i)$ when $P_X(i) > 0$ but $n_{x^L}(i) = 0$. A simple approach is to not include such x^L in the expectation $\mathrm{E}[Q_{Y^L|X^L}]$. However, it then turns out that (50) is not necessarily true. A potential fix to this problem, but an artificial and unsatisfying one, is to try to choose those $Q_{y^L|x^L}(j|i)$ with $n_{x^L}(i) = 0$ so as to ensure that (50) is true. Instead, we show that for large L, the desired relation (50) is approximately true for any reasonable choice for the ambiguous $Q_{y^L|x^L}(j|i)$.

Lemma 5: Suppose X^L is output from a DMS P_X and Y^L is output from the channel $P_{Y^L|X^L}$. Let

$$P_{\overline{Y}|\overline{X}} = \sum_{\ell=1}^{L} P_{Y_\ell|X_\ell}/L$$

and $0 < \delta < 1$. We then have

$$\frac{P_{\overline{Y}|\overline{X}}(j|i)}{1+\delta} - \epsilon_\delta(L)/\mu_X \le \mathrm{E}\left[Q_{Y^L|X^L}(j|i)\right] \le \frac{P_{\overline{Y}|\overline{X}}(j|i)}{1-\delta} + \epsilon_\delta(L) \quad (52)$$

for all j and i with $P_X(i) > 0$, where the $Q_{y^L|x^L}(j|i)$ with $n_{x^L}(i) = 0$ are chosen as any probabilities, i.e., numbers between 0 and 1 inclusive, in the expectation in (52).

Proof: Consider i with $P_X(i) > 0$. We have

$$\begin{aligned}
&\mathrm{E}[Q_{Y^L|X^L}(j|i)] \\
&\le \left(\sum_{x^L \in T_\delta^L(P_X)} P(x^L) \sum_{y^L} P(y^L|x^L) \frac{n_{x^L y^L}(i,j)}{n_{x^L}(i)}\right) + \epsilon_\delta(L) \\
&\le \left(\sum_{x^L, y^L} P(x^L, y^L) \frac{n_{x^L y^L}(i,j)}{LP_X(i)(1-\delta)}\right) + \epsilon_\delta(L) \\
&= \frac{P_{\overline{Y}|\overline{X}}(j|i)}{1-\delta} + \epsilon_\delta(L) \qquad (53)
\end{aligned}$$

where the first step follows by (27) and the last step follows by (see (18))

$$\begin{aligned}
\mathrm{E}[Q_{X^L Y^L}(i,j)] &= \sum_{x^L, y^L} P(x^L, y^L)\, n_{x^L y^L}(i,j)/L \\
&= P_{\overline{X}\,\overline{Y}}(i,j) \\
&= P_X(i) \cdot P_{\overline{Y}|\overline{X}}(j|i). \qquad (54)
\end{aligned}$$

Note that, for convenience, we removed the subscripts of some of the probability distributions in (53) and (54). For the lower bound in (52), we have

$$\begin{aligned}
&\mathrm{E}[Q_{Y^L|X^L}(j|i)] \\
&\ge \left(\sum_{x^L \in T_\delta^L(P_X)} P(x^L) \sum_{y^L} P(y^L|x^L) \frac{n_{x^L y^L}(i,j)}{n_{x^L}(i)}\right) \\
&\ge \left(\sum_{x^L, y^L} P(x^L, y^L) \frac{n_{x^L y^L}(i,j)}{LP_X(i)(1+\delta)}\right) \\
&\quad - \left(\sum_{x^L \notin T_\delta^L(P_X)} P(x^L) \sum_{y^L} P(y^L|x^L) \frac{n_{x^L y^L}(i,j)}{LP_X(i)(1+\delta)}\right) \\
&\ge \frac{P_{\overline{Y}|\overline{X}}(j|i)}{1+\delta} - \left(\sum_{x^L \notin T_\delta^L(P_X)} P(x^L) \frac{L}{LP_X(i)(1+\delta)}\right) \\
&\ge \frac{P_{\overline{Y}|\overline{X}}(j|i)}{1+\delta} - \epsilon_\delta(L)/\mu_X. \qquad (55)
\end{aligned}$$

where the third step follows by (54) and $n_{x^L y^L}(i,j) \le L$. □

Lemma 5 shows that (50) can be made as accurate as desired for large L. Thus, if $d(P_{B|A})$ is continuous in $P_{B|A}$ as described in (47) and (48), then the rate distortion function for large L is (51).

As a special case, consider the Bhattacharryya–Wootters distortion (41) and suppose we desire $\mathrm{E}[d(Q_{Y^L|X^L})]$ to be close to zero. Suppose also that we take a "worst case" approach and set the $Q_{y^L|x^L}(j|i)$ with $n_{x^L}(i) = 0$ to be zero. But for any finite L such x^L have nonzero probability and will cause $\mathrm{E}[d(Q_{Y^L|X^L})]$ to be bounded from below by a positive number. We thus need to use large L, which in turn means that the minimum rate is given by (51) with $\Delta = 0$.

ACKNOWLEDGMENT

The authors are grateful to C. Fuchs for bringing the problem to their attention. They would also like to thank I. Cirac for providing them with a draft of [11] and E. Knill as well as the anonymous referees for their useful remarks.

REFERENCES

[1] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE Nat. Conv. Rec.*, pt. 4, pp. 142–163, 1959.
[2] H. Barnum, C. M. Caves, C. A. Fuchs, R. Jozsa, and B. Schumacher, "On quantum coding for ensembles of mixed states," *J. Phys. A.: Math. Gen.*, vol. 34, pp. 6767–6785, 2001.
[3] B. W. Schumacher, "Quantum coding," *Phys. Rev. A*, vol. 51, pp. 2738–2747, 1995.
[4] R. Jozsa and B. W. Schumacher, "A new proof of the quantum noiseless coding theorem," *J. Mod. Opt.*, vol. 41, pp. 2343–2349, 1994.
[5] H. Barnum, C. A. Fuchs, R. Jozsa, and B. Schumacher, "General fidelity limits for quantum channels," *Phys. Rev. A*, vol. 54, p. 4707, 1996.
[6] R. Jozsa, "Quantum noiseless coding of mixed states," presented at the Third Santa Fe Workshop on Complexity, Entropy, and the Physics of Information, Santa Fe, NM, May 1994.
[7] H.-K. Lo, Quantum Coding Theorem for Mixed States, 1995 [Online]. Available: LANL ArXiV.org e-print quant-ph/9504004
[8] M. Horodecki, "Limits for compression of quantum information carried by ensembles of mixed states," *Phys. Rev. A*, vol. 57, pp. 3364–3369, 1998.
[9] ——, "Toward optimal compression for mixed signal states," *Phys. Rev. A*, vol. 61, p. 052309, 2000.
[10] M. Koashi and N. Imoto, "Compressibility of quantum mixed-state signals," *Phys. Rev. Lett.*, vol. 87, no. 1, 2001.
[11] W. Dür, G. Vidal, and J. I. Cirac, "Visible compression of commuting mixed states," *Phys. Rev. A*, vol. 64, no. ms. 022308, Jul. 2001.
[12] E. Soljanin, "Compressing quantum mixed-state source by sending classical information," *IEEE Trans. Inf. Theory*, vol. 48, no. 8, pp. 2263–2275, Aug. 2002.
[13] H. Barnum, "Quantum rate-distortion coding," *Phys. Rev. A*, vol. 62, p. 042309, 2000.
[14] I. Devetak and T. Berger, "Quantum rate-distortion theory for i.i.d. sources," in *Proc. IEEE Int. Symp. Information Theory*, Washington, DC, Jun. 2001, p. 276.
[15] G. Kramer and S. A. Savari, Quantum Data Compression of Ensembles of Mixed States with Commuting Density Operators, 2001 [Online]. Available: LANL ArXiV.org e-print quant-ph/quant-ph/0101119
[16] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*. Cambridge, U.K.: Cambridge Univ. Press, 2000.
[17] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985.
[18] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[19] I. Csiszár and J. Körner, *Information Theory, Coding Theorems for Discrete Memoryless Systems*. Budapest, Hungary: Akadémiai Kiadó, 1981.
[20] R. L. Dobrushin and B. S. Tsybakov, "Information transmission with additional noise," *IEEE Trans. Inf. Theory*, vol. IT-8, no. 5, pp. 293–304, Sep. 1962.

[21] J. K. Wolf and J. Ziv, "Transmission of noisy information to a noisy receiver with minimum distortion," *IEEE Trans. Inf. Theory*, vol. IT-16, no. 4, pp. 406–411, Jul. 1970.
[22] H. Witsenhausen, "Indirect rate distortion problems," *IEEE Trans. Inf. Theory*, vol. IT-26, no. 5, pp. 518–521, Sep. 1980.
[23] T. Linder, G. Lugosi, and K. Zeger, "Empirical quantizer design in the presence of source noise or channel noise," *IEEE Trans. Inf. Theory*, vol. 43, no. 2, pp. 612–623, Mar. 1997.
[24] T. Berger, *Rate Distortion Theory: A Mathematical basis for Data Compression*. Englewood Cliffs, NJ: Prentice-Hall, 1971.
[25] A. Orlitsky and J. R. Roche, "Coding for computing," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 903–917, Mar. 2001.
[26] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1991.