The Rate-Distortion Function for Source Coding with Side Information at the Decoder

AARON D. WYNER, FELLOW, IEEE, AND JACOB ZIV, FELLOW, IEEE

Abstract—Let $\{(X_k,Y_k)\}_{k=1}^{\infty}$ be a sequence of independent drawings of a pair of dependent random variables X,Y. Let us say that X takes values in the finite set $\mathscr{X}$. It is desired to encode the sequence $\{X_k\}$ in blocks of length n into a binary stream of rate R, which can in turn be decoded as a sequence $\{\hat{X}_k\}$, where $\hat{X}_k \in \hat{\mathscr{X}}$, the reproduction alphabet. The average distortion level is $(1/n)\sum_{k=1}^{n} E[D(X_k,\hat{X}_k)]$, where $D(x,\hat{x}) \geq 0$, $x \in \mathscr{X}$, $\hat{x} \in \hat{\mathscr{X}}$, is a pre-assigned distortion measure. The special assumption made here is that the decoder has access to the side information $\{Y_k\}$. In this paper we determine the quantity $R^*(d)$, defined as the infimum of rates R such that (with $\varepsilon > 0$ arbitrarily small and with suitably large n) communication is possible in the above setting at an average distortion level (as defined above) not exceeding $d + \varepsilon$. The main result is that $R^*(d) = \inf[I(X;Z) - I(Y;Z)]$, where the infimum is with respect to all auxiliary random variables Z (which take values in a finite set $\mathscr{Z}$) that satisfy: i) Y,Z conditionally independent given X; ii) there exists a function $f: \mathscr{Y} \times \mathscr{Z} \rightarrow \hat{\mathscr{X}}$, such that $E[D(X,f(Y,Z))] \leq d$.

Let $R_{X|Y}(d)$ be the rate-distortion function which results when the encoder as well as the decoder has access to the side information $\{Y_k\}$. In nearly all cases it is shown that when $d > 0$ then $R^*(d) > R_{X|Y}(d)$, so that knowledge of the side information at the encoder permits transmission of the $\{X_k\}$ at a given distortion level using a smaller transmission rate. This is in contrast to the situation treated by Slepian and Wolf [5] where, for arbitrarily accurate reproduction of $\{X_k\}$, i.e., $d = \varepsilon$ for any $\varepsilon > 0$, knowledge of the side information at the encoder does not allow a reduction of the transmission rate.

I. Introduction, Problem Statement, and Results

A. Introduction

In this paper we consider the problem of source encoding with a fidelity criterion in a situation where the decoder has access to side information about the source. To put the problem in perspective, consider the system shown in Fig. 1.

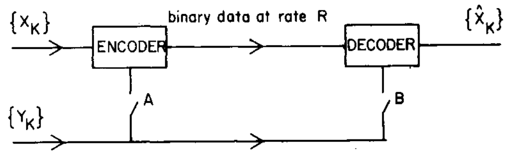

Fig. 1.

The sequence $\{(X_k,Y_k)\}_{k=1}^{\infty}$ represents independent copies of a pair of dependent random variables (X,Y) which take values in the finite sets $\mathscr{X},\mathscr{Y}$, respectively. The encoder output is a binary sequence which appears at a rate R bits per input symbol. The decoder output is a sequence $\{\hat{X}_k\}_1^{\infty}$ which take values in the finite reproduction alphabet $\hat{\mathscr{X}}$. The encoding and decoding is done in blocks of length n, and the fidelity criterion is the expectation of

$$\frac{1}{n}\sum_{k=1}^{n} D(X_k,\hat{X}_k),$$

where $D(x,\hat{x}) \geq 0$, $x \in \mathscr{X}$, $\hat{x} \in \hat{\mathscr{X}}$, is a given distortion function. If switch A and/or B is closed then the encoder and/or decoder, respectively, are assumed to have knowledge of the side information sequence $\{Y_k\}$. If switch A and/or B is open, then the side information is not available to the encoder and/or decoder, respectively.

Now consider the following cases:

i) switches A and B are open, i.e., there is no available side information;

ii) switches A and B are closed, i.e., both the encoder and the decoder have access to the side information $\{Y_k\}$;

iii) switch A is open and switch B is closed, i.e., only the decoder has access to the side information.

We define $R_X(d)$, $R_{X|Y}(d)$, and $R^*(d)$ as the minimum rates for which the system of Fig. 1 can operate in cases i), ii), and iii), respectively, when n is large and the average distortion $E[1/n \sum_{k=1}^{n} D(X_k,\hat{X}_k)]$ is arbitrarily close to d. The first two of these quantities can be characterized as follows.

For $d \geq 0$, define $\mathscr{M}_0(d)$ as the set of probability distributions $p(x,y,\hat{x})$, $x \in \mathscr{X}$, $y \in \mathscr{Y}$, $\hat{x} \in \hat{\mathscr{X}}$, such that the marginal distribution $\sum_{\hat{x} \in \hat{\mathscr{X}}} p(x,y,\hat{x})$ is the given distribution for (X,Y), and

$$\sum_{x,\hat{x},y} D(x,\hat{x})p(x,y,\hat{x}) \leq d. \qquad (1)$$

Then the classical Shannon theory yields for case i), cf. [1], [3], [4], that

$$R_X(d) = \min_{p \in \mathscr{M}_0(d)} I(X;\hat{X}), \qquad (2)$$

and for case ii), cf. [1, sec. 6.1.1], that

$$R_{X|Y}(d) = \min_{p \in \mathscr{M}_0(d)} I(X;\hat{X} \mid Y). \qquad (3)$$

The random variables $X,Y,\hat{X}$ corresponding to $p \in \mathscr{M}_0(d)$ are defined in the obvious way, and $I(\cdot)$ denotes the ordinary Shannon mutual information [3].

We now turn to case iii) and the determination of $R^*(d)$. For the very large and important class of situations when $\mathscr{X} = \hat{\mathscr{X}}$ and

$$D(x,x) = 0,$$
$$D(x,\hat{x}) > 0, \qquad x \neq \hat{x}, \qquad (4)$$

Manuscript received January 3, 1975; revised July 3, 1975.

A. D. Wyner is with Bell Laboratories, Murray Hill, N.J. 07974.
J. Ziv is with the Faculty of Electrical Engineering, Technion—Israel Institute of Technology, Haifa, Israel.

it is easy to show that

$$R_X(0) = H(X) \qquad R_{X|Y}(0) = H(X \mid Y) \tag{5}$$

where H denotes entropy [3]. In this case Slepian and Wolf [5] have established that

$$R^*(0) = R_{X|Y}(0) = H(X \mid Y). \tag{6}$$

The main result of this paper is the determination of $R^*(d)$, for $d \geq 0$, in the general case. In particular, it follows from our result that usually $R^*(d) > R_{X|Y}(d)$, $d > 0$.

At this point we pause to give some of the history of our problem. The characterization of $R^*(d)$ was first attempted by T. Goblick (Ph.D. dissertation, M.I.T., 1962) and later by Berger [1, sec. 6.1]. It should be pointed out that Theorem 6.1.1 in [1], which purports to give a characterization of $R^*(d)$, is not correct. After the discovery of his error, Berger (private communication) did succeed in giving an upper bound on $R^*(d)$ for the special case studied in Section II of the present paper. In fact our results show that Berger's bound is tight.

An outline of the remainder of this paper is as follows. In Section I-B we give a formal and precise statement of the problem. In Section I-C we state our results including the characterization of $R^*(d)$. Section II contains an evaluation of $R^*(d)$ for a special binary source. The proofs follow in Sections III and IV.

B. Formal Statement of Problem

In this section we will give a precise statement of the problem which we stated informally in Section I-A.

First, a word about notation: Let $\mathscr{U}$ be an arbitrary finite set, and consider $\mathscr{U}^n$, the set of n-vectors with elements in $\mathscr{U}$. The members of $\mathscr{U}^n$ will be written as $\boldsymbol{u}^n = (u_1, u_2, \cdots, u_n)$, where the subscripted letters denote the coordinates and boldface superscripted letters denote vectors. A similar convention will apply to random variables and vectors, which will be denoted by upper case letters. When the dimension n of a vector $\boldsymbol{u}^n$ is clear from the context, we will omit the superscript. Next for $k = 1, 2, \cdots$, define the set

$$I_k = \{0, 1, 2, \cdots, k-1\}. \tag{7}$$

Finally for random variables X, Y, etc., the notation $H(X)$, $H(X \mid Y)$, $I(X;Y)$, etc., will denote the standard information-theoretic quantities as defined in Gallager [3]. All logarithms in this paper are taken to the base 2.

Let $\mathscr{X}, \mathscr{Y}, \hat{\mathscr{X}}$ be finite sets and let $\{(X_k, Y_k)\}_1^\infty$ be a sequence of independent drawings of a pair of dependent random variables X, Y which take values in $\mathscr{X}, \mathscr{Y}$, respectively. The probability distribution for X, Y is

$$Q(x,y) = \Pr\{X = x,\ Y = y\}, \qquad x \in \mathscr{X}, y \in \mathscr{Y}. \tag{8}$$

Let $D: \mathscr{X} \times \hat{\mathscr{X}} \to [0, \infty)$ be a distortion function. A *code* (n, M, Δ) is defined by two mappings F_E, F_D, an "encoder" and a "decoder," respectively, where

$$F_E: \mathscr{X}^n \to I_M, \tag{9a}$$

$$F_D: \mathscr{Y}^n \times I_M \to \hat{\mathscr{X}}^n, \tag{9b}$$

and

$$E \frac{1}{n} \sum_{k=1}^{n} D(X_k, \hat{X}_k) = \Delta, \tag{9c}$$

where $\hat{\boldsymbol{X}}^n = F_D(\boldsymbol{Y}^n, F_E(\boldsymbol{X}^n))$. The correspondence between a code as defined here and the system of Fig. 1 with switch A open and switch B closed should be clear.

A pair (R,d) is said to be *achievable* if, for arbitrary $\varepsilon > 0$, there exists (for n sufficiently large) a code (n, M, Δ) with

$$M \leq 2^{n(R+\varepsilon)}, \qquad \Delta \leq d + \varepsilon. \tag{01}$$

We define $\mathscr{R}$ as the set of achievable (R,d) pairs, and define

$$R^*(d) = \min_{(R,d) \in \mathscr{R}} R. \tag{11}$$

Since from the definition, $\mathscr{R}$ is closed, the indicated minimum exists. Our main problem is the determination of $R^*(d)$.

We pause at this point to observe the following. Since $R^*(d)$ is nonincreasing in d, we have $R^*(0) \geq \lim_{d \to 0} R^*(d)$. Furthermore, from (11), for all $d \geq 0$, the pair $(R^*(d), d) \in \mathscr{R}$. Since $\mathscr{R}$ is closed, $(\lim_{d \to 0} R^*(d), 0) \in \mathscr{R}$, so that $R^*(0) \leq \lim_{d \to 0} R^*(d)$. We conclude that $R^*(d)$ is continuous at $d = 0$.

C. Summary of Results

Let X, Y, etc., be as above. Let $p(x,y,z)$, $x \in \mathscr{X}$, $y \in \mathscr{Y}$, $z \in \mathscr{Z}$, where $\mathscr{Z}$ is an arbitrary finite set, be a probability distribution which defines random variables X, Y, Z, such that the marginal distribution for X, Y

$$\sum_{z \in \mathscr{Z}} p(x,y,z) = Q(x,y), \tag{12a}$$

and such that

$$Y, Z \text{ are conditionally independent given } X. \tag{12b}$$

An alternative way of expressing (12) is

$$p(x,y,z) = Q(x,y) p_t(z \mid x), \tag{13}$$

where $p_t(z \mid x)$ can be thought of as the transition probability of a "test channel" whose input is X and whose output is Z. Now, for $d > 0$, define $\mathscr{M}(d)$ as the set of $p(x,y,z)$ which satisfy (12) (or equivalently (13)) and which have the property that there exists a function $f: \mathscr{Y} \times \mathscr{Z} \to \hat{\mathscr{X}}$ such that

$$E[D(X, \hat{X})] \leq d, \qquad \text{where } \hat{X} = f(Y,Z). \tag{14}$$

As a mnemonic for remembering the above, we can think of $X, Y, Z, \hat{X}$ as being generated by the configuration in Fig. 2.

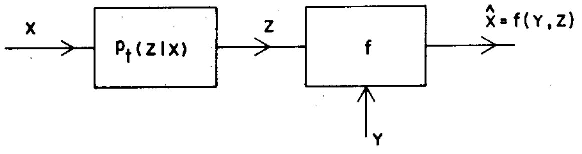

Fig. 2.

Next define, for $d > 0$, the quantity

$$\bar{R}(d) \triangleq \inf_{p \in \mathscr{M}(d)} [I(X;Z) - I(Y;Z)]. \tag{15a}$$

Since $\mathscr{M}(d)$ is nondecreasing in d, $\bar{R}(d)$ is nonincreasing, for $d \in (0,\infty)$. Thus we can meaningfully define

$$\bar{R}(0) = \lim_{d \to 0} \bar{R}(d). \tag{15b}$$

Our main result is the following.

Theorem 1: For $d \geq 0$, $R^*(d) = \bar{R}(d)$.

Remarks: 1) We remarked (following (11)) that $R^*(d)$ is continuous at $d = 0$. Since $\bar{R}(d)$ is, by construction, also continuous at $d = 0$, it will suffice to prove Theorem 1 for $d > 0$.

2) Let X,Y,Z satisfy (12). Then

$$\begin{aligned} I(X;Z) - I(Y;Z) &= H(Z \mid Y) - H(Z \mid X) \\ &\overset{(*)}{=} H(Z \mid Y) - H(Z \mid X,Y) \\ &= I(X;Z \mid Y), \end{aligned} \tag{16}$$

where step $(*)$ follows from (12b). Thus (15a) can be written, for $d > 0$,

$$\bar{R}(d) = \inf_{p \in \mathscr{M}(d)} I(X;Z \mid Y).$$

3) Let D satisfy (4). Let $\delta \triangleq \min_{x \neq \hat{x}} D(x,\hat{x}) > 0$. Thus if $X,Y,Z,\hat{X}$ correspond to $p \in \mathscr{M}(d)$,

$$\lambda \triangleq \Pr\{X \neq \hat{X}\} \leq ED(X,\hat{X})/\delta \leq d/\delta.$$

Now since $\hat{X}$ is a function of Z,Y, Fano's inequality [3] implies that

$$\begin{aligned} H(X \mid ZY) &\leq -\lambda \log \lambda - (1-\lambda)\log(1-\lambda) \\ &\quad + \lambda \log(\operatorname{card} \mathscr{X}) \\ &\triangleq \varepsilon(\lambda), \end{aligned}$$

so that

$$\begin{aligned} I(X;Z \mid Y) &= H(X \mid Y) - H(X \mid ZY) \\ &\geq H(X \mid Y) - \varepsilon\left(\frac{d}{\delta}\right) \\ &\to H(X \mid Y), \qquad \text{as } d \to 0. \end{aligned}$$

Thus $\bar{R}(0) \geq H(X \mid Y)$. Furthermore, since setting $Z \equiv X$ and $f(Y,Z) = Z = X$, results in a distribution in $\mathscr{M}(d)$, for all $d > 0$, we have

$$R(0) \leq I(X;X \mid Y) = H(X \mid Y).$$

Thus $\bar{R}(0) = H(X \mid Y)$, and Theorem 1 is consistent with the Slepian–Wolf result given in (6).

4) The following is shown in the Appendix (Theorem A2):

a) $R^*(d) = \bar{R}(d)$ is a continuous convex function of d, $0 \leq d < \infty$;

b) in evaluating $\bar{R}(d)$ from (15) it suffices to consider only sets $\mathscr{Z}$ with card $\mathscr{Z} \leq$ (card $\mathscr{X}$) + 1;

c) the infimum in (15) is in fact a minimum.

5) Let $p \in \mathscr{M}(d)$ define $X,Y,Z,\hat{X} = f(Y,Z)$. Then from (16)

$$I(X;Z) - I(Y;Z) = I(X;Z \mid Y).$$

Furthermore, given that $Y = y$, the random variables X and $\hat{X} = f(y,Z)$ are conditionally independent given Z. Thus the data-processing theorem [3] yields

$$I(X;Z \mid Y = y) \geq I(X;\hat{X} \mid Y = y),$$

so that

$$I(X;Z \mid Y) \geq I(X;\hat{X} \mid Y). \tag{17}$$

Furthermore, equality follows in (17) if and only if

$$I(X;Z \mid \hat{X}Y) = 0. \tag{18}$$

Finally, the distribution defining $X,Y,\hat{X}$ belongs to $\mathscr{M}_0(d)$, so that remark 2), (3), (17), and Theorem 1 imply that

$$R^*(d) \geq R_{X|Y}(d), \qquad d \geq 0. \tag{19}$$

The equality holds in (19), for $d > 0$, if and only if the distribution for $X,Y,\hat{X}$ which achieves the minimization in (3) can be represented as in Fig. 2, with $X,Y,Z,\hat{X}$ satisfying (12) and (18). This is, in fact, an extremely severe condition and seems hardly ever to be satisfied. In particular, it is not satisfied in the binary example discussed in Section II. See remark 6) below.

6) Although the discussion in this paper has been restricted to the case where $\mathscr{X}$ and $\mathscr{Y}$ are finite sets, it will be shown elsewhere that Theorem 1 is valid in a more general setting which includes the case where X is Gaussian and $Y = X + U$, where U is also Gaussian and is independent of X. The distortion is $D(x,\hat{x}) = (x - \hat{x})^2$. In this case, it turns out that for all $d > 0$

$$R^*(d) = R_{X|Y}(d)$$
$$= \begin{cases} \frac{1}{2} \log \dfrac{\sigma_U^2 \sigma_X^2}{(\sigma_X^2 + \sigma_U^2)d}, & 0 < d \leq \dfrac{\sigma_U^2 \sigma_X^2}{\sigma_X^2 + \sigma_U^2}, \\ 0, & d \geq \dfrac{\sigma_U^2 \sigma_X^2}{\sigma_X^2 + \sigma_U^2}, \end{cases}$$

where $\sigma_X^2 = \operatorname{var} X$, $\sigma_U^2 = \operatorname{var} U$. Thus in this case the condition for equality of $R_{X|Y}$ and R^* given in remark 5) holds.

II. Example: Doubly Symmetric Binary Source

A. Evaluation of $R^*(d)$

In this section we evaluate $R^*(d)$ from Theorem 1 for the special case where $\mathscr{X} = \mathscr{Y} = \hat{\mathscr{X}} = \{0,1\}$ and for $x,y = 0,1$,

$$Q(x,y) = \frac{(1-p_0)}{2}\delta_{x,y} + \frac{p_0}{2}(1 - \delta_{x,y}) \tag{20}$$

where $0 \leq p_0 \leq \frac{1}{2}$. We can think of X as being the unbiased input to a binary symmetric channel (BSC) with crossover probability p_0, and Y as the corresponding output, or vice versa. The distortion measure is, for $x,\hat{x} = 0,1$,

$$D(x,\hat{x}) = \begin{cases} 0, & x = \hat{x}, \\ 1, & x \neq \hat{x}. \end{cases} \tag{21}$$

It is known [1, pp. 46–47] that

$$R_{X|Y}(d) = \begin{cases} h(p) - h(d), & 0 \leq d \leq p, \\ 0, & d \geq p, \end{cases} \tag{22}$$

where $h(\lambda) = -\lambda \log \lambda - (1 - \lambda) \log (1 - \lambda), 0 \leq \lambda \leq 1$, is the usual entropy function. We now give the formula for $R^*(d)$.

For $0 \leq d \leq p_0$, let $g(d)$ be defined by

$$g(d) = \begin{cases} h(p_0 * d) - h(d), & 0 \leq d < p_0, \\ 0, & d = p_0, \end{cases} \tag{23a}$$

where for $0 \leq u, v \leq 1$,

$$u * v \triangleq u(1 - v) + v(1 - u). \tag{23b}$$

Also define

$$g^*(d) = \inf_{\theta,\beta_1,\beta_2} [\theta g(\beta_1) + (1 - \theta) g(\beta_2)] \tag{24}$$

where the infimum is with respect to all $\theta \in [0,1]$ and $\beta_1,\beta_2 \in [0,p_0]$ such that $d = \theta\beta_1 + (1 - \theta)\beta_2$. The function $g^*(d)$ is seen to be the lower convex envelope of $g(d)$.

We will show below that, for $0 \leq d < p_0$, $g(d)$ is convex. Thus, in performing the minimization in (24), we can take $\beta_2 = p_0$, and (24) becomes

$$g^*(d) = \inf_{\theta,\beta} [\theta(h(p_0 * \beta) - h(\beta))], \qquad 0 \leq d \leq p_0 \tag{25a}$$

where the infimum is with respect to all θ,β, such that

$$0 \leq \theta \leq 1, \qquad 0 \leq \beta < p_0 \tag{25b}$$

and

$$d = \theta\beta + (1 - \theta)p_0. \tag{25c}$$

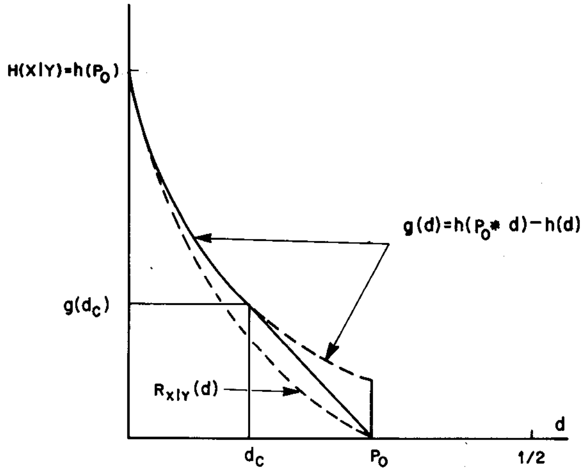

Fig. 3. Graph of $g^*(d)$ versus d (solid line). Lower dashed curve is $R_{X|Y}(d) = h(p_0) - h(d)$.

Referring to Fig. 3, we see that $g^*(d) = g(d)$, for $d \leq d_C$. For $d_C \leq d \leq p_0$, the graph of $g^*(d)$ is the straight line which is tangent to the graph of $g(d)$ and which passes through the point $(p_0,0)$. The point of tangency is $(d_C,g(d_C))$. Thus d_C is the solution to

$$\frac{g(d_C)}{d_C - p_0} = g'(d_C). \tag{26}$$

Of course, $g^*(d)$ is nonincreasing in d.

In the remainder of this section we establish that

$$R^*(d) = g^*(d), \tag{27}$$

as given in (25). It should be observed that, for $d > 0$, $R^*(d) = g^*(d) > R_{X|Y}(d)$, as given in (22). Thus we see that knowledge of the side information at the encoder does effect the required code rate. When $d = 0$, however, $R_{X|Y}(0) = R^*(0) = h(p_0)$.

Before proceeding with the proof of (27), we pause to establish the convexity of $g(d)$, $0 \leq d < p_0$. We shall fulfill this task by establishing the following.

Lemma A: For $0 \leq u \leq 1$, $0 \leq p_0 \leq \frac{1}{2}$, the function

$$G(u) \triangleq h(p_0 * u) - h(u) \tag{28}$$

is convex in u.

Proof: Let $a = (1 - 2p_0)$, so that $p_0 * u = au + p_0$, and

$$\frac{d^2G}{du^2}(u) = a^2h''(au + p_0) - h''(u)$$

where $h''(u) = (\log e)(-1/u(1 - u))$. Continuing we have

$$\begin{aligned} \frac{d^2G}{du^2}(u) &= \frac{\log e}{(au + p_0)(1 - au - p_0)u(1 - u)} \\ &\quad \cdot [(au + p_0)(1 - au - p_0) - a^2u(1 - u)] \\ &= \frac{\log e}{(au + p_0)(1 - au - p_0)u(1 - u)} \\ &\quad \cdot [p_0(1 - p_0)] \geq 0, \qquad 0 \leq u \leq 1, \end{aligned}$$

so that $G(u)$ is convex, completing the proof.

In Section II-B we show that $R^*(d) \leq g^*(d)$, and in Section II-C that $R^*(d) \geq g^*(d)$.

B. Proof that $R^*(d) \leq g^*(d)$

We begin by obtaining an upper bound on $R^*(d)$. Consider the following situations.

a) Let Z be the output obtained from a BSC with crossover probability β $(0 \leq \beta \leq \frac{1}{2})$, when the input is X. Setting $\hat{X} = f(Y,Z) = Z$, we have $E[D(X,\hat{X})] = \beta$, so that the distribution for X,Y,Z belongs to $\mathcal{M}(\beta)$. Now Y,Z can be thought of as being connected by the channel which is the cascade of a BSC with crossover probability p_0 with a BSC with crossover probability β. This cascade is a BSC with crossover probability $p_0 * \beta = p_0(1 - \beta) + (1 - p_0)\beta$. Thus

$$\begin{aligned} I(X;Z) - I(Y;Z) &= [1 - h(\beta)] - [1 - h(p_0 * \beta)] \\ &= h(p_0 * \beta) - h(\beta). \end{aligned}$$

Thus, from Theorem 1,

$$R^*(\beta) \le h(p_0 * \beta) - h(\beta), \qquad 0 \le \beta \le \tfrac{1}{2}. \tag{29}$$

b) Let Z be degenerate, let us say $Z \equiv 0$, and let $\hat{X} = f(Y,Z) = Y$. Then we have $E[D(X,\hat{X})] = p_0$, so that this distribution belongs to $\mathscr{M}(p_0)$. Since $I(X;Z) - I(Y;Z) = 0$, we have from Theorem 1,

$$R^*(p_0) = 0. \tag{30}$$

Now let d, $0 \le d \le p_0$, be given and say that θ,β are such that

$$d = \theta\beta + (1-\theta)p_0, \qquad 0 \le \theta \le 1, \quad 0 \le \beta < p_0. \tag{31}$$

[Equation (31) is always satisfied for some θ,β.] Since $R^*(d)$ is convex,

$$\begin{aligned} R^*(d) &= R^*(\theta\beta + (1-\theta)p_0) \le \theta R^*(\beta) + (1-\theta)R^*(p_0) \\ &\le \theta[h(p_0 * \beta) - h(\beta)] \end{aligned} \tag{32}$$

where the last inequality follows from (29) and (30). To get the tightest bound we minimize (32) with respect to all θ,β satisfying (31) or (25b), (25c), yielding $R^*(d) \le g^*(d)$.

C. Proof that $R^*(d) \ge g^*(d)$

Let $X,Y,Z,\hat{X} = f(Y,Z)$ define a distribution in $\mathscr{M}(d)$ $(d < p_0)$. We will show that $R^*(d) \ge g^*(d)$ by showing that

$$I(X;Z) - I(Y;Z) \ge g^*(d) \tag{33}$$

and invoking Theorem 1. Define the set

$$A = \{z : f(0,z) = f(1,z)\} \tag{34a}$$

so that its complement

$$A^C = \mathscr{Z} - A = \{z : f(0,z) \ne f(1,z)\}, \tag{34b}$$

by hypothesis,

$$\begin{aligned} E[D(X,\hat{X})] &= \Pr\{Z \in A\}E[D(X,\hat{X}) \mid Z \in A] \\ &\quad + \Pr\{Z \in A^C\}E[D(X,\hat{X}) \mid Z \in A^C] \\ &\le d. \end{aligned} \tag{35}$$

We first show that

$$E[D(X,\hat{X}) \mid Z \in A^C] \ge p_0. \tag{36}$$

To do this, we write

$$E[D \mid Z \in A^C] = \sum_{z \in A^C} \frac{\Pr\{Z = z\}}{\Pr\{Z \in A^C\}} E[D \mid Z = z]. \tag{37}$$

Next, note that if $z \in A^C$ and if $f(0,z) = 0$, then $f(1,z) = 1$. Therefore, for such z,

$$\begin{aligned} &E[D \mid Z = z] \\ &\quad = \Pr\{X = 1, Y = 0 \mid Z = z\} \\ &\qquad + \Pr\{X = 0, Y = 1 \mid Z = z\} \\ &\quad \overset{(*)}{=} \Pr\{X = 1 \mid Z = z\}\Pr\{Y = 0 \mid X = 1\} \\ &\qquad + \Pr\{X = 0 \mid Z = z\}\Pr\{Y = 1 \mid X = 0\} = p_0. \end{aligned} \tag{38a}$$

Equality (*) follows from (12b). If, however, $z \in A^C$, but $f(0,z) = 1$, then we have

$$E[D \mid Z = z] = 1 - p_0 \ge p_0, \tag{38b}$$

since $p_0 \le \frac{1}{2}$. Inequality (36) follows from (37) and (38). Therefore, using

$$E[D \mid Z \in A] = \sum_{z \in A} \frac{\Pr\{Z = z\}}{\Pr\{Z \in A\}} E[D \mid Z = z]$$

(35) and (36) yield

$$d' \triangleq \theta \sum_{z \in A} \lambda_z d_z + (1-\theta)p_0 \le d \tag{39a}$$

where $\theta = \Pr\{Z \in A\}$, $\lambda_z = \Pr\{Z = z\}/\Pr\{Z \in A\}$, and

$$d_z = E[D(X,\hat{X}) \mid Z = z]. \tag{39b}$$

Next, consider

$$\begin{aligned} &I(X;Z) - I(Y;Z) \\ &\quad = H(Y \mid Z) - H(X \mid Z) \\ &\quad \ge \sum_{z \in A} [H(Y \mid Z = z) - H(X \mid Z = z)] \Pr\{Z = z\} \\ &\quad = \theta \sum_{z \in A} \lambda_z [H(Y \mid Z = z) - H(X \mid Z = z)]. \end{aligned} \tag{40}$$

Now, for $z \in A$, define $\gamma(z) = f(0,z) = f(1,z)$. Then

$$d_z = E[D(X,\hat{X}) \mid Z = z] = \Pr\{X \ne \gamma(z) \mid Z = z\},$$

so that

$$H(X \mid Z = z) = h(d_z) \tag{41a}$$

and from (12b),

$$H(Y \mid Z = z) = h(p_0 * d_z). \tag{41b}$$

Thus (40) and (41) yield

$$\begin{aligned} I(X;Z) - I(Y;Z) &\ge \theta \sum_{z \in A} \lambda_z [h(p_0 * d_z) - h(d_z)] \\ &= \theta \sum_{z \in A} \lambda_z G(d_z), \end{aligned} \tag{42}$$

where G is defined by (28). Since, by Lemma A, G is convex, and $\sum_{z \in A} \lambda_z = 1$,

$$I(X;Z) - I(Y;Z) \ge \theta G\left(\sum_z \lambda_z d_z\right) = \theta[h(p_0 * \beta) - h(\beta)] \tag{43a}$$

where

$$\beta = \sum_{z \in A} \lambda_z d_z. \tag{43b}$$

Thus we have shown that, for any distribution in $\mathscr{M}(d)$, there exists $0 \le \theta \le 1$ and $0 \le \beta < p_0$ such that (43a) holds and (from (39a)),

$$\theta\beta + (1-\theta)p_0 = d'. \tag{44}$$

Comparison of (43a) and (44) with (25) yields

$$I(X;Z) - I(Y;Z) \ge g^*(d'). \tag{45}$$

Now, from (39a), $d' \le d$, and since $g^*(d)$ is nonincreasing in d, we have that (45) yields (33), completing the proof.

III. Converse Theorem

In this section we establish the converse theorem.

Theorem 2: $R^*(d) \ge \bar{R}(d)$, $d \ge 0$.

Proof: Let (F_E,F_D) define a code with parameters (n,M,Δ). We will show that

$$\frac{1}{n}\log M \ge \bar{R}(\Delta). \tag{46}$$

Thus if $(R,d) \in \mathscr{R}$, then, for arbitrary $\varepsilon > 0$, with n sufficiently large, there exists a code (n,M,Δ) with $M \le 2^{n(R+\varepsilon)}$, $\Delta \le d + \varepsilon$. Inequality (46) and the monotonicity of $\bar{R}(\Delta)$ imply

$$R + \varepsilon \ge \bar{R}(\Delta) \ge \bar{R}(d+\varepsilon). \tag{47}$$

Letting $\varepsilon \to 0$ and invoking the continuity of $\bar{R}(d)$ (see Theorem A2), we have $R \ge \bar{R}(d)$, whenever $(R,d) \in \mathscr{R}$, which implies Theorem 2. It remains to establish (46).

Let $W = F_E(X^n)$, so that $\hat{X}^n = (\hat{X}_1,\cdots,\hat{X}_n) = F_D(Y^n,W)$. Let

$$\Delta_k = ED(X_k,\hat{X}_k) \tag{48}$$

so that

$$\Delta = \frac{1}{n}\sum_{k=1}^{n} \Delta_k. \tag{49}$$

Now

$$\log M \overset{(1)}{\ge} H(W) \overset{(2)}{\ge} I(X^n;W) \overset{(3)}{=} H(X^n) - H(X^n \mid W). \tag{50}$$

Step (1) follows from $W \in I_M$ and steps (2) and (3) from standard identities. Now

$$\begin{aligned} H(X^n \mid W) &\overset{(1)}{=} I(X^n;Y^n \mid W) + H(X^n \mid Y^n,W) \\ &\overset{(2)}{=} H(Y^n \mid W) - H(Y^n \mid X^nW) + H(X^n \mid Y^nW) \\ &\overset{(3)}{\le} H(Y^n) - H(Y^n \mid X^n) + H(X^n \mid Y^nW) \\ &\overset{(4)}{=} I(X^n;Y^n) + H(X^n \mid Y^nW). \end{aligned} \tag{51}$$

Steps (1), (2), and (4) are standard identities, and step (3) follows from $H(Y \mid W) \le H(Y)$, and $H(Y^n \mid X^nW) = H(Y^n \mid X^n)$ since $W = F_E(X^n)$. Substituting (51) into (50) we have

$$\begin{aligned} \log M &\ge H(X^n) - H(X^n \mid Y^nW) - I(X^n;Y^n) \\ &= \sum_{k=1} [H(X_k) - H(X_k \mid X^{k-1}Y^nW) - I(X_k;Y_k)] \end{aligned} \tag{52}$$

where $X^{k-1} = (X_1,\cdots,X_{k-1})$. Here we have used the independence of the $\{(X_k,Y_k)\}$ and standard identities. Now define, for $k = 1,2,\cdots,n$,

$$Z_k = (X^{k-1},Y_1,Y_2,\cdots,Y_{k-1},Y_{k+1},\cdots,Y_n,W) \tag{53}$$

so that (52) becomes

$$\log M \ge \sum_{k=1}^{n} [H(X_k) - H(X_k \mid Y_kZ_k) - I(X_k;Y_k)]. \tag{54}$$

We pause to point out two facts about Z_k:

a) $\hat{X}_k$ is the kth coordinate of $F_D(Y^n,W)$ so that we can write $\hat{X}_k$ as a deterministic function of Y_k and Z_k, let us say $\hat{X}_k = f(Y_k,Z_k)$; of course, (48) still holds, so that $E[D(X_k,\hat{X}_k)] = \Delta_k$;

b) Y_k,Z_k are conditionally independent given X_k.

Facts a), b) imply that the distribution which defines X_k,Y_k,Z_k belongs to $\mathscr{M}(\Delta_k)$, so that from the definition (15),

$$I(X_k;Z_k) - I(Y_k;Z_k) \ge \bar{R}(\Delta_k). \tag{55}$$

Now, returning to (54) we can write the second term in the summand as

$$\begin{aligned} H(X_k \mid Y_k, Z_k) &= H(X_k, Y_k \mid Z_k) - H(Y_k \mid Z_k) \\ &= H(X_k \mid Z_k) + H(Y_k \mid X_kZ_k) - H(Y_k \mid Z_k) \\ &= H(X_k \mid Z_k) + H(Y_k \mid X_k) - H(Y_k \mid Z_k). \end{aligned} \tag{56}$$

The above follows from fact b), which implies that $H(Y_k \mid X_kZ_k) = H(Y_k \mid X_k)$ and standard identities. Substituting (56) into (54) we have

$$\begin{aligned} \log M &\ge \sum_{k=1}^{n} [H(X_k) - H(X_k \mid Z_k) - H(Y_k \mid X_k) \\ &\qquad + H(Y_k \mid Z_k) - H(Y_k) + H(Y_k \mid X_k)] \\ &= \sum_{k=1}^{n} [I(X_k;Z_k) - I(Y_k;Z_k)] \\ &\overset{(1)}{\ge} \sum_{k=1}^{n} \bar{R}(\Delta_k) \overset{(2)}{\ge} n\bar{R}\left(\frac{1}{n}\sum_{1}^{n} \Delta_k\right) \overset{(3)}{=} n\bar{R}(\Delta). \end{aligned} \tag{57}$$

Step (1) follows from (55), step (2) from the convexity of $\bar{R}(\Delta)$ established in the Appendix, and step (3) from (49). This establishes (46) and completes the proof of Theorem 2.

IV. Direct Theorem

We begin by stating precisely a version of the result of Slepian and Wolf [1] which was mentioned in Section I. Let X,Y be random variables as above which take values in $\mathscr{X},\mathscr{Y}$, respectively. Let the reproduction alphabet $\hat{\mathscr{X}} = \mathscr{X}$, and the distortion measure be the Hamming metric, for x, $\hat{x} \in \mathscr{X}$,

$$D(x,\hat{x}) = D_H(x,\hat{x}) = \begin{cases} 1, & x \ne \hat{x}, \\ 0, & x = \hat{x}. \end{cases} \tag{58}$$

Their result is that $R^*(0)$ (as defined in Section I–B) is given by

$$R^*(0) = H(X \mid Y). \tag{59}$$

This means that, for arbitrary $\varepsilon_1,\varepsilon_2 > 0$, the sequence $\{X_k\}$ can be encoded into blocks of length n (sufficiently large) and that the decoder (which has access to $\{Y_k\}$) can produce a sequence $\{\hat{X}_k\}$ such that

$$\frac{1}{n}\sum_{k=1}^{n} \Pr\{\hat{X}_k \ne X_k\} \le \varepsilon_1.$$

Furthermore, the code with block length n has at most $\exp_2\{n(H(X \mid Y)+\varepsilon_2)\}$ codewords.

The following is a corollary of the Slepian–Wolf theorem.

Lemma 3: Let X,Y, etc., be as in Section I–B, and let $(F_E^{(0)},F_D^{(0)})$ be a code (n_0,M_0,Δ_0), as defined in that section. Define R_0, by

$$R_0=\frac{1}{n_0}H(W \mid Y^{n_0}),$$

where $W=F_E^{(0)}(X^{n_0})$. Then, for arbitrary $\delta>0$, there exists an n_1 (sufficiently large) so that there is a code (n,M,Δ) with

$$n=n_0n_1 \tag{60a}$$

$$M \leq 2^{n_1(n_0R_0+\delta)} \leq 2^{n(R_0+\delta)} \tag{60b}$$

$$\Delta \leq \Delta_0+\delta. \tag{60c}$$

Remark: The essence of Lemma 3 is the following. Ordinarily, the rate of any code is about $(1/n)H(W)$, where $W=F_E(X^n)$. However, should a code be such that $(1/n)H(W \mid Y^n)$ is significantly less than $(1/n)H(W)$, the rate can be reduced to about $(1/n)H(W \mid Y^n)$. This can be done by further encoding the W corresponding to successive blocks (of length n_0) using the Slepian–Wolf scheme. It follows that for an optimal code $(1/n)H(W) \approx (1/n) \cdot H(W \mid Y^n)$ or $(1/n)I(W;Y^n) \approx 0$. Thus the encoded information W and the side information Y are approximately independent.

Proof of Lemma 3: Let $(F_E^{(0)},F_D^{(0)})$ satisfy the hypotheses of the lemma. We can consider the independent repetitions of (W,Y^{n_0}) as a new "supersource." Denote the sequence of successive repetitions of W by $\{W_j\}$.

Now let $\delta>0$ be given. For n_1 sufficiently large, there exists (by Slepian–Wolf) an encoding of $(W_1,\cdots,W_{n_1})$ into a code with no more than

$$\exp_2\{n_1(H(W \mid Y^{n_0})+\delta)\}=2^{n_1(n_0R_0+\delta)} \tag{61}$$

codewords such that the decoder (which knows $\{Y_k\}$) can recover a sequence, let us say $(\hat{W}_1,\hat{W}_2,\cdots,\hat{W}_{n_1})$, where

$$\frac{1}{n_1}\sum_{j=1}^{n_1}\Pr\{W_j \neq \hat{W}_j\} \leq \delta/\max_{x,\hat{x}} D(x,\hat{x}). \tag{62}$$

The decoder can then apply $F_D^{(0)}$ to $\hat{W}_j$ and $(Y_{(j-1)n_0+1},\cdots,Y_{jn_0})$ to obtain say $(X^*_{(j-1)n_0+1},\cdots,X^*_{jn_0})$ as a decoded message. The combination of the given code $(F_E^{(0)},F_D^{(0)})$ and the Slepian–Wolf code results in a new code (n,M,Δ), where $n=n_0n_1$ (satisfying (60a)), and M satisfies (60b). Furthermore,

$$\begin{aligned}\Delta&=\frac{1}{n_0n_1}\sum_{k=1}^{n_0n_1}E[D(X_k,X_k^*)]\\&=\frac{1}{n_1}\sum_{j=1}^{n_1}E\left\{\frac{1}{n_0}\sum_{k=(j-1)n_0+1}^{jn_0}D(X_k,X_k^*)\right\}\\&\overset{(1)}{\leq}\frac{1}{n_1}\sum_{j=1}^{n_1}[\Delta_0+\max_{x,\hat{x}}D(x,\hat{x})\Pr\{W_j \neq \hat{W}_j\}]\\&\overset{(2)}{\leq}\Delta_0+\delta.\end{aligned} \tag{63}$$

To verify inequality (1), set the term in brackets in the left member of inequality (1) equal to ϕ_j and define the event $\mathscr{E}=\{W_j \neq \hat{W}_j\}$. Now write

$$\begin{aligned}E\phi_j&=\Pr\{\mathscr{E}^C\}E[\phi_j \mid \mathscr{E}^C]+\Pr\{\mathscr{E}\}E[\phi_j \mid \mathscr{E}]\\&\leq \Pr\{\mathscr{E}^C\}E[\phi_j \mid \mathscr{E}^C]+\Pr\{\mathscr{E}\}\max_{x,\hat{x}}D(x,\hat{x}),\end{aligned}$$

and observe that if $\mathscr{E}^C$ occurs, then

$$\phi_j=\frac{1}{n_0}\sum_k D(X_k,\hat{X}_k),$$

so that (since $D \geq 0$)

$$\Pr\{\mathscr{E}^C\}E[\phi_j \mid \mathscr{E}^C] \leq E\left[\frac{1}{n_0}\sum_k D(X_k,\hat{X}_k)\right]=\Delta_0.$$

Thus

$$E[\phi_j] \leq \Delta_0+\Pr\{W_j \neq \hat{W}_j\}\max_{x,\hat{x}}D(x,\hat{x}),$$

which is inequality (1). Inequality (2) follows from (62). Since (63) is (60c), the proof of Lemma 3 is complete.

We now state the direct theorem.

Theorem 4: For $d \geq 0$, $R^*(d) \leq \bar{R}(d)$.

Proof: As indicated in remark 1) following Theorem 1, it will suffice to establish Theorem 4 for $d>0$. We will do this by showing that if X,Y,Z, and $\hat{X}=f(Y,Z)$ correspond to a distribution $p(x,y,z)$ in $\mathscr{M}(d)$, for some $d>0$, then $(R_0,d) \in \mathscr{R}$, where

$$R_0=I(X;Z)-I(Y;Z). \tag{64}$$

We will do this by means of the following.

Lemma 5: With X,Y,Z,f as above, and $\varepsilon_0>0$ arbitrary, there exists, for n_0 sufficiently large, a code $(F_E^{(0)},F_D^{(0)})$ with parameters (n_0,M_0,Δ_0), such that

$$\Delta_0 \leq d+\varepsilon_0 \tag{65}$$

and

$$\frac{1}{n_0}H(W \mid Y^{n_0}) \leq R_0+\varepsilon_0 \tag{66}$$

with $R_0=I(X;Z)-I(Y;Z)$, as in (64).

Theorem 4 now follows from Lemmas 3 and 5, which together assert, for arbitrary ε_0, $\delta>0$, the existence of a code (n,M,Δ) with $M \leq 2^{n(R_0+\varepsilon_0+\delta)}$ and $\Delta \leq d+\varepsilon_0+\delta$. Thus $(R,d) \in \mathscr{R}$. We now give the proof of Lemma 5.

Proof of Lemma 5: With $\varepsilon_0>0$ and $p(x,y,z) \in \mathscr{M}(d)$ given, let $p^{(n)}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})$ define the probability distribution on $\mathscr{X}^n \times \mathscr{Y}^n \times \mathscr{Z}^n$ corresponding to n independent repetitions of $p(x,y,z)$, $n=1,2,\cdots$. *All probabilities in this proof will be computed with respect to* $p^{(n)}$. For $\alpha>0$ and $n=1,2,\cdots$, let $T(n,\alpha)$, the "typical" sequences, be defined as the set of all $\boldsymbol{y},\boldsymbol{z} \in \mathscr{Y}^n \times \mathscr{Z}^n$ for which

$$\left|-\frac{1}{n}\log \Pr\{Y^n=\boldsymbol{y},\ Z^n=\boldsymbol{z}\}-H(Y,Z)\right| \leq \alpha \tag{67}$$

and

$$\left| -\frac{1}{n} \log \Pr \{Z^n = z\} - H(Z) \right| \le \alpha. \tag{68}$$

It follows from the weak law of large numbers that, with $\alpha > 0$ held fixed,

$$\Pr \{(Y^n,Z^n) \notin T(n,\alpha)\} \to 0, \qquad \text{as } n \to \infty. \tag{69}$$

The set $T(n,\alpha)$ also has the property that for all $z \in \mathscr{Z}^n$,

$$\operatorname{card} \{y: (y,z) \in T(n,\alpha)\} \le 2^{n(H(Y|Z)+2\alpha)}. \tag{70}$$

To verify (70), note that if $(y,z) \in T(n,\alpha)$, then

$$\begin{aligned}\Pr \{Y^n = y \mid Z^n = z\} &= \frac{\Pr \{Y^n = y,\, Z^n = z\}}{\Pr \{Z^n = z\}} \\ &\ge \exp_2 \{-n(H(Y\mid Z) + 2\alpha)\}.\end{aligned} \tag{71}$$

Inequality (70) follows from (71) on writing

$$\begin{aligned}1 &\ge \sum_{y:\,(y,z) \in T(n,\alpha)} \Pr \{Y^n = y \mid Z^n = z\} \\ &\ge 2^{n(H(Y|Z)+2\alpha)} \operatorname{card} \{y: (y,z) \in T(n,\alpha)\}.\end{aligned}$$

Next, for $n = 1,2,\cdots$, define the function $D_n: \mathscr{X}^n \times \hat{\mathscr{X}}^n \to [0,\infty)$ by

$$D_n(x^n,\hat{x}^n) = \frac{1}{n} \sum_{k=1}^{n} D(x_k,\hat{x}_k). \tag{72}$$

Also for $y,z \in \mathscr{Y}^n \times \mathscr{Z}^n$, set

$$f_n(y,z) = (f(y_1,z_1),f(y_2,z_2),\cdots,f(y_n,z_n)) \in \mathscr{X}^n \tag{73}$$

so that we have

$$\begin{aligned}D_n(X^n,f_n(Y^n,Z^n)) &= \frac{1}{n} \sum_{k=1}^{n} D(X_k,f(Y_k,Z_k)) \\ &= \frac{1}{n} \sum_{k=1}^{n} D(X_k,\hat{X}_k).\end{aligned} \tag{74}$$

Again, the law of large numbers yields, with $\alpha > 0$ held fixed,

$$\Pr \{D_n(X^n,f_n(Y^n,Z^n)) \ge (d + \alpha)\} \to 0, \qquad \text{as } n \to \infty. \tag{75}$$

Now define the function $\psi_n: \mathscr{X}^n \times \mathscr{Y}^n \times \mathscr{Z}^n \to [0,1]$, by

$$\psi_n(x,y,z) = \begin{cases} 1, & \text{if } D_n(x,f_n(y,z)) \ge d + \alpha, \\ & \quad \text{or } (y,z) \notin T(n,\alpha) \\ 0, & \text{otherwise.} \end{cases} \tag{76}$$

We have from (69) and (75),

$$E[\psi_n(X^n,Y^n,Z^n) \to 0], \qquad \text{as } n \to \infty. \tag{77}$$

We now apply a lemma in [7]. This lemma asserts that, for arbitrary $\varepsilon > 0$, there exists (for n_0 sufficiently large) a mapping $F: \mathscr{X}^{n_0} \to \{z_i\}_{i=1}^{M_0} \subseteq \mathscr{Z}^{n_0}$, such that

$$M_0 \le 2^{n_0(I(X:Z)+\varepsilon)} \tag{78a}$$

$$E[\psi_n(X^{n_0},Y^{n_0},F(X^{n_0}))] \le \varepsilon. \tag{78b}$$

We show how to obtain (78) from the result in [7] in Appendix A2. The mapping F will define the code $(F_E{}^{(0)},F_D{}^{(0)})$ which will establish Lemma 5 as follows.

Let $\alpha,\varepsilon > 0$ be arbitrary and let F be a mapping which satisfies (78). Define $F_E{}^{(0)}$ by $F_E{}^{(0)}(x) = i$, when $F(x) = z_i$. Define $F_D{}^{(0)}$ by $F_D{}^{(0)}(y,i) = f_{n_0}(y,z_i)$, where f_{n_0} is defined by (73). Let the parameters of $(F_E{}^{(0)},F_D{}^{(0)})$ be (n_0,M_0,Δ_0). Set $W = F_E{}^{(0)}(X^{n_0})$. From (76) and (78) we have

$$\begin{aligned}\Delta_0 &= E[D_{n_0}(X^{n_0},F_D(Y^{n_0},W))] \\ &= E[D_{n_0}(X^{n_0},f_{n_0}(Y^{n_0},F(X^{n_0})))] \\ &\le (d + \alpha) + [\max_{x,\hat{x}} D(x,\hat{x})] \\ &\qquad \cdot \Pr \{D_{n_0}(X^{n_0},f_{n_0}(Y^{n_0},F(X^{n_0}))) \ge d + \alpha\} \\ &\le (d + \alpha) + [\max D(x,\hat{x})]E[\psi_{n_0}(X^{n_0},Y^{n_0},F(X^{n_0}))] \\ &\le d + \alpha + \varepsilon[\max D(x,\hat{x})]\end{aligned} \tag{79}$$

with α,ε sufficiently small, $\Delta_0 \le d + \varepsilon_0$, which is (65). It remains to verify (66). Write

$$\begin{aligned}\frac{1}{n_0} H(W \mid Y^{n_0}) &= \frac{1}{n_0} [H(W,Y^{n_0}) - H(Y^{n_0})] \\ &= \frac{1}{n_0} [H(W) + H(Y^{n_0} \mid W)] - H(Y) \\ &\le I(X;Z) + \varepsilon - H(Y) + \frac{1}{n_0} H(Y^{n_0} \mid W).\end{aligned} \tag{80}$$

The inequality follows from $H(W) \le \log M_0$, and (78a). Now

$$\begin{aligned}H(Y^{n_0} \mid W) &\le H(Y^{n_0},\psi_{n_0} \mid W) \\ &= H(\psi_{n_0} \mid W) + H(Y^{n_0} \mid W\psi_{n_0}) \\ &= H(\psi_{n_0} \mid W) \\ &\qquad + \Pr \{\psi_{n_0} = 0\}H(Y^{n_0} \mid W, \psi_{n_0} = 0) \\ &\qquad + \Pr \{\psi_{n_0} = 1\}H(Y^{n_0} \mid W, \psi_{n_0} = 1) \\ &\le H(\psi_{n_0}) + H(Y^{n_0} \mid W, \psi_{n_0} = 0) \\ &\qquad + (E\psi_{n_0})H(Y^{n_0} \mid W, \psi_{n_0} = 1) \\ &\overset{(1)}{\le} h(\varepsilon) + n_0(H(Y \mid Z) + 2\alpha) \\ &\qquad + \varepsilon n_0 \log \operatorname{card} \mathscr{Y} \\ &\le n_0[H(Y \mid Z) + 2\alpha + \varepsilon \log \operatorname{card} \mathscr{Y} + h(\varepsilon)].\end{aligned} \tag{81}$$

Inequality (1) follows from the definition of ψ_{n_0} (76) and from (70) (assuming $\varepsilon \le \frac{1}{2}$). Substituting (81) into (80) yields

$$\begin{aligned}\frac{1}{n_0} H(W \mid Y^{n_0}) &\le I(X;Z) - I(Y;Z) \\ &\qquad + 2\alpha + \varepsilon + \varepsilon \log \operatorname{card} \mathscr{Y} + h(\varepsilon) \\ &\le R_0 + \varepsilon_0\end{aligned}$$

for ε,α sufficiently small. This is (66). This completes the proof of Lemma 5.

Appendix

A1. Some Facts About $\bar{R}(d)$

The techniques in this section are similar to those in [6]. We begin by giving an alternative, though equivalent, formulation of the definition of $\bar{R}(d)$. Let the set $\mathscr{X} = \{1,2,\cdots,K\}$ and the set $\mathscr{Y} = \{1,2,\cdots,J\}$. Then, for $k = 1,2,\cdots,K$, set

$$Q_k = \Pr\{X = k\} = \sum_{j=1}^{J} Q(k,j), \tag{A1}$$

where the given distribution Q defines X,Y. Also let

$$t_{jk} = \Pr\{Y = j \mid X = k\} = \frac{Q(k,j)}{Q_k}, \qquad 1 \le k \le K, \quad 1 \le j \le J. \tag{A2}$$

Let T be the $J \times K$ matrix, with (j,k)th entry t_{jk}. Also, for $m = 1,2,\cdots$, let Δ_m be the simplex of probability m-vectors. Then $Q = (Q_1,Q_2,\cdots,Q_K)^t \in \Delta_K$, and $TQ \in \Delta_J$. Of course, T defines a channel.

Now let $\{q(z)\}_{z\in\mathscr{Z}}$ be a finite set of vectors in Δ_K, indexed by the finite set $\mathscr{Z}$. Also let $\{\lambda_z\}_{z\in\mathscr{Z}}$ satisfy

$$\lambda_z \ge 0 \qquad \sum_{z\in\mathscr{Z}} \lambda_z = 1. \tag{A3}$$

Let Z be the random variable which takes the value $z \in \mathscr{Z}$ with probability λ_z. Furthermore, suppose that Z is the input to a channel with output (let us say X') taking values in $\mathscr{X}$ with transition probability

$$\Pr\{X' = k \mid Z = z\} = q_k(z), \qquad 1 \le k \le K \tag{A4}$$

where $q_k(z)$ is the kth component of $q(z)$. Let Y' be the output of the channel defined by T when X' is the input. The random variables X',Y',Z satisfy (12) if and only if

$$\sum_z \lambda_z q(z) = Q. \tag{A5}$$

(In other words, X' and Y' have marginal distribution Q.) Assuming that (A5) is satisfied, we have

$$I(X';Z) - I(Y';Z) = H(X) - H(Y) + \sum_{z\in\mathscr{Z}} \lambda_z[\mathscr{H}(Tq(z)) - \mathscr{H}(q(z))], \tag{A6}$$

where $\mathscr{H}(p_1,\cdots,p_m) = -\sum_{i=1}^m p_i \log p_i$. Thus we can write

$$I(X';Z) - I(Y';Z) = \sum_{z\in\mathscr{Z}} \lambda_z \Gamma(q(z)) \tag{A7a}$$

where

$$\Gamma(q) \triangleq H(X) - H(Y) + \mathscr{H}(Tq) - \mathscr{H}(q). \tag{A7b}$$

Finally, assume that $\{\lambda_z,q(z)\}_z$ satisfy (A5), and let X',Y',Z be the corresponding random variables. Let $f: \mathscr{Y} \times \mathscr{Z} \to \hat{\mathscr{X}}$. Then, with $\hat{X} = f(Y,Z)$,

$$\begin{aligned} E[D(X',\hat{X})] &= \sum_{z\in\mathscr{Z}} \lambda_z \sum_{j,k} \Pr\{Y' = j \mid X' = k\} \\ &\quad \cdot \Pr\{X' = k \mid Z = z\} D(k,f(j,z)) \\ &= \sum_z \lambda_z \sum_{j,k} t_{jk} q_k(z) D(k,f(j,z)). \end{aligned} \tag{A8}$$

Now $E[D]$ is minimized with respect to f, if for all j,z we take $f(j,z)$ as that value of $\hat{x} \in \hat{\mathscr{X}}$ which minimizes

$$\sum_k t_{jk} q_k(z) D(k,\hat{x}).$$

Thus, since t_{jk} is given and fixed, the function which minimizes $E[D]$ depends only on $q(z)$. If we assume that f is always this minimizing function, we can write

$$E[D(X',\hat{X})] = \sum_z \lambda_z \Delta(q(z)) \tag{A9a}$$

where

$$\Delta(q(z)) = \sum_{j,k} t_{jk} q_k(z) D(k,f(j,z)). \tag{A9b}$$

Now let $\hat{\mathscr{M}}(d)$ be the family of $\{\lambda_z,q(z)\}_{z\in\mathscr{Z}}$ (where $\mathscr{Z}$ is a finite set and $q(z) \in \Delta_K$) such that (A5) is satisfied and $E[D(X,\hat{X})]$ as given by (A9) does not exceed d. Each member of $\hat{\mathscr{M}}(d)$ generates a distribution for X',Y',Z in $\mathscr{M}(d)$. Furthermore, each distribution for X',Y',Z in $\mathscr{M}(d)$ generates a $\{\lambda_z,q(z)\}$ in $\hat{\mathscr{M}}(d)$ by

$$\lambda_z = \Pr\{Z = z\}$$
$$q_k(z) = \Pr\{X' = k \mid Z = z\}.$$

Thus we have shown the following.

Lemma A1: For $d > 0$,

$$\bar{R}(d) = \inf \sum_{z\in\mathscr{Z}} \lambda_z \Gamma(q(z))$$

where the infimum is with respect to all $\{\lambda_z,q(z)\}$ in $\hat{\mathscr{M}}(d)$.

Next consider the polytope $\Delta_K \times [0,\infty] \times [0,\infty]$, which is $(K+1)$-dimensional. The mapping $q \to (q,\Gamma(q),\Delta(q))$ assigns a point in this polytope to each point in Δ_K. Let S be the image of Δ_K under this mapping and let C be the convex hull of S. Let $C_Q = \{(\rho,\delta): (Q,\rho,\delta) \in C\}$. C_Q is also convex. A pair $(\rho,\delta) \in C_Q$, if and only if for some $\{\lambda_z,q(z)\}_z$ satisfies (A5), and the corresponding X',Y',Z satisfy

$$I(X';Z) - I(Y';Z) = \sum_{z\in\mathscr{Z}} \lambda_z \Gamma(q(z)) = \rho \tag{A10a}$$

and

$$E[D(X,\hat{X})] = \sum_z \lambda_z \Delta(q(z)) = \delta. \tag{A10b}$$

Thus from Lemma A1, for $d > 0$,

$$\bar{R}(d) = \inf_{\substack{(\rho,\delta)\in C_Q \\ \delta \le d}} \rho. \tag{A11}$$

We can now establish the following.

Theorem A2: $\bar{R}(d)$ is convex and, therefore, continuous in d, $0 < d < \infty$. Furthermore, in the calculation of $\bar{R}(d)$ in (15), we can assume that card $\mathscr{Z} \le$ card $\mathscr{X} + 1 = K + 1$, and we can assert that the "infimum" is a minimum.

Proof: The convexity of $\bar{R}(d)$ follows from (A11) and the convexity of C_Q. Next observe that, since S is the continuous image of the compact set Δ_K, S is connected and compact. It then follows from the Fenchel–Eggleston strengthening of Caratheodory's theorem[1] that any point in $(\rho,\delta) \in C_Q$ can be expressed as in (A10) with card $\mathscr{Z} \le$ dimension $S \le K + 1$. Finally, since S is compact, C and C_Q are also compact, so that the infimum in (A11) and, therefore, in (15) is in fact a minimum. This establishes Theorem A2.

A2. Application of [7]

In Section IV of the present paper we used a lemma from [7] to deduce the existence of the function "F." The result from [7]

[1] If S is a connected subset of an n-dimensional linear space, any point in the convex hull of S can be expressed as a convex combination of at most n points of S [2, p. 35].

which we used is the following. Let the random vectors U^n, V^n, W^n be independent copies of U, V, W which take values in the finite sets $\mathscr{U}, \mathscr{V}, \mathscr{W}$, respectively. Assume further that U and W are conditionally independent given V. Let $\psi_n: \mathscr{U}^n \times \mathscr{W}^n \rightarrow [0,1]$ be an arbitrary function and let $\{\psi_n\}_{n=1}^{\infty}$ be sequence of such functions. Assume that

$$\lim_{n\rightarrow\infty} E[\psi_n(U^n, W^n)] = 0. \tag{A12}$$

A *PB-code* (n_0, M_0) is a mapping

$$F: \mathscr{V}^n \rightarrow \{w_j\}_{j=1}^{M_0} \subseteq \mathscr{W}^n.$$

Lemma 4.3 in [7] states that, for arbitrary $\varepsilon > 0$ and for $n_0 = n_0(\varepsilon)$ sufficiently large, there exists a PB-code (n_0, M_0) such that

$$M_0 \leq \exp_2 \{n_0[I(V;W) + \varepsilon]\} \tag{A13a}$$

$$E[\psi_{n_0}(U^n, F(V^n))] \leq \varepsilon. \tag{A13b}$$

To apply this lemma to our problem, let $\mathscr{U} = \mathscr{X} \times \mathscr{Y}, \mathscr{V} = \mathscr{X}$, and $\mathscr{W} = \mathscr{Z}$, and set $U = (X,Y)$, $V = X$, and $W = Z$. Since X, Y, Z satisfy (12b) and U and W are conditionally independent given V, as required. Now the functions $\{\psi_n\}$, defined in (76) satisfy (A12) (by (77)), so that Lemma 4.3 in [7] can be applied to deduce the existence of a function F which satisfies (A13). Since (A13a) and (A13b) are identical to (78a) and (78b), respectively, our task is completed.

References

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

[2] H. G. Eggleston, *Convexity.* Cambridge, England: Cambridge Univ. Press, 1958, p. 35.

[3] R. G. Gallager, *Information Theory and Reliable Communication.* New York: Wiley, 1968.

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

[5] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 471–480, July 1973.

[6] H. S. Witsenhausen and A. D. Wyner, "A conditional entropy bound for a pair of discrete random variables," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 493–501, Sept. 1975.

[7] A. D. Wyner, "On source coding with side-information at the decoder," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 294–300, May 1975.