Rate Distortion when Side Information May Be Absent

CHRIS HEEGARD, MEMBER, IEEE, AND TOBY BERGER, FELLOW, IEEE

Abstract—**The problem is considered of encoding a discrete memoryless source when correlated side information may or may not be available to the decoder. It is assumed that the side information is not available to the encoder. The rate-distortion function $R(D_1, D_2)$ is determined where D_1 is the distortion achieved with side information and D_2 is the distortion achieved without it. A generalization is made to the case of m decoders, each of which is privy to its own side information. An appropriately defined D-admissible rate for this general case is shown to equal $R(\boldsymbol{D})$ when the side information sources satisfy a specified degradedness condition. Explicit results are obtained in the quadratic Gaussian case and in the binary Hamming case.**

I. INTRODUCTION

IN A CELEBRATED PAPER Wyner and Ziv [1] extended rate-distortion theory to the case in which side information is present at the decoder. They knew from Slepian and Wolf's earlier treatment of distortionless coding of correlated information sources [2] that each cell of the encoder's partition of the set of typical source sequences should consist of widely dispersed elements.

Suppose the side information failed to reach the decoder in the Wyner–Ziv problem. Then the decoder would know only the partition cell index. Because the elements of the cell are so widely dispersed, this knowledge would be virtually useless for reproducing the source output. This contingency has led us to consider the rate-distortion problem for cases in which it is not known whether side information will be present. An equivalent formulation of this problem has two decoders; one of these receives the side information, and the other does not (cf. Fig. 1).

We determine the rate-distortion function $R(D_1, D_2)$, where D_1 is the distortion achieved with the side information, and D_2 is the distortion achieved without it. This is accomplished by proving both a source coding theorem and its converse. We also generalize to cases in which there are m decoders, each of which is privy to its own side information. An appropriately defined $\boldsymbol{D}$-admissible rate for this general case is shown to equal $R(\boldsymbol{D})$ when the side information sources satisfy a specified degradedness condition. Explicit results are obtained in the quadratic Gaussian case and the binary Hamming case.

Manuscript received February 2, 1984; revised March 10, 1985. This work was supported in part by the National Science Foundation under Grants ENG79-21976 and ECS-8204886 and the Department of Energy under Contract DEAC-0180RA50419.

The authors are with the School of Electrical Engineering, Cornell University, Ithaca, NY 14853, USA.

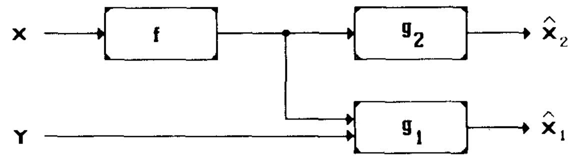

Fig. 1. Coding system.

II. PROBLEM STATEMENT AND THEOREM STATEMENT

Let $(\mathscr{X}, \mathscr{Y}, p(x, y))$ be a discrete memoryless 2-source with generic random variables X and Y. For $i \in \{1, 2\}$ let $\hat{\mathscr{X}}_i$ be a reconstruction alphabet and let

$$d_i: \mathscr{X} \times \hat{\mathscr{X}}_i \to [0, \infty)$$

be a distortion measure. An (n, M, D_1, D_2) code consists of an encoder

$$f: \mathscr{X}^n \to \{0, 1, \cdots, M-1\}$$

and two decoders

$$g_1: \{0, 1, \cdots, M-1\} \times \mathscr{Y}^n \to \hat{\mathscr{X}}_1^n$$
$$g_2: \{0, 1, \cdots, M-1\} \to \hat{\mathscr{X}}_2^n.$$

The expected distortion $\boldsymbol{D} = (D_1, D_2)$ for the code is given by

$$D_i = Ed_i(\boldsymbol{X}, \hat{\boldsymbol{X}}_i) \equiv E\frac{1}{n}\sum_{k=1}^{n} d_i(X_k, \hat{X}_{ik}), \qquad i = 1, 2,$$

where

$$\hat{X}_1 = g_1(f(\boldsymbol{X}), \boldsymbol{Y}) \qquad \hat{X}_2 = g_2(f(\boldsymbol{X})).$$

The rate R is said to be $\boldsymbol{D}$-admissible if for every $\epsilon > 0$ there exists for some n an $(n, M, D_1 + \epsilon, D_2 + \epsilon)$ code with $n^{-1}\log M \le R + \epsilon$. We shall determine the rate-distortion function $R(D_1, D_2) = R(\boldsymbol{D})$ defined by

$$R(\boldsymbol{D}) = \inf\{R: R \text{ is } \boldsymbol{D}\text{-admissible}\}.$$

Define

$$R_0(\boldsymbol{D}) \equiv \min_{P(\boldsymbol{D})}[I(X; W) + I(X; U|Y, W)],$$

where $P(\boldsymbol{D})$ is the set of all random variables $(W, U) \in \mathscr{W} \times \mathscr{U}$ jointly distributed with the generic random variables (X, Y), such that the following conditions are satisfied.

1) $Y \ominus X \ominus (W, U)$ is a Markov string.

2) There exist functions $\hat{X}_1(W, U, Y)$ and $\hat{X}_2(W)$ such that $Ed_i(X, \hat{X}_i) \le D_i$, $i = 1, 2$.

3) $|\mathscr{W}| \le |\mathscr{X}| + 2$ and $|\mathscr{U}| \le (|\mathscr{X}| + 1)^2$.

0018-9448/85/1100-0727$01.00

Theorem 1: $R(\boldsymbol{D}) = R_0(\boldsymbol{D})$.

Proof: The proof of Theorem 1 is divided into two parts. The inequality $R_0(\boldsymbol{D}) \geq R(\boldsymbol{D})$, which says that $R_0(\boldsymbol{D})$ is $\boldsymbol{D}$-admissible, is proved in Section III. Then the converse, $R_0(\boldsymbol{D}) \leq R(\boldsymbol{D})$, is proved in Section IV.

Remark: The informed decoder g_1 is faced with a Wyner–Ziv problem, while the uninformed decoder g_2 is faced with an ordinary rate-distortion problem. Therefore, our prescription for calculating $R_0(\boldsymbol{D})$ must reduce accordingly in the appropriate special cases. In this regard note that if only the informed decoder were present, then no loss of generality would result from setting $\boldsymbol{W} =$ constant, thereby reducing the task of calculating $R_0(\boldsymbol{D})$ to minimization of $I(X; U|Y)$ over all U such that $Y \ominus X \ominus U$ and there exists $\hat{X}(U, Y)$ satisfying $Ed_1(X, \hat{X}) \leq D_1$. This is the prescription [1] for the Wyner–Ziv rate-distortion function, $R_{W-Z}(D_1)$. Similarly, if only the uninformed decoder were present, no loss of generality would result from setting $U =$ constant and letting $\hat{X}(W)$ simply equal W. These steps reduce $R_0(\boldsymbol{D})$ to the ordinary rate-distortion function $R(D_2)$, namely, the minimum of $I(X; \hat{X})$ over all $\hat{X}$ such that $ED_2(X, \hat{X}) \leq D_2$. It is clear that $R_0(D_1, D_2) \leq R_{W-Z}(D_1) + R(D_2)$, in general. The question of whether or not there are situations in which equality holds when both $R_{W-Z}(D_1)$ and $R(D_2)$ are positive remains to be investigated.

III. Admissibility

We begin the proof that $R_0(\boldsymbol{D}) \geq R(\boldsymbol{D})$ by fixing $\epsilon > 0$ and choosing $(W, U) \in P(\boldsymbol{D})$. Next we specify the domains and ranges of some functions a, a', b, and b', which will be used to describe the encoder, namely,

$$a: \mathscr{X}^n \to \{0, 1, \cdots, M_0' - 1\}$$
$$a': \{0, 1, \cdots, M_0' - 1\} \to \{0, 1, \cdots, M_0 - 1\}$$
$$b: \mathscr{X}^n \times \{0, 1, \cdots, M_0' - 1\} \to \{0, 1, \cdots, M_1' - 1\}$$
$$b': \{0, 1, \cdots, M_1' - 1\} \to \{0, 1, \cdots, M_1 - 1\}.$$

The encoder $f(\cdot)$ is defined by formula

$$f(\boldsymbol{X}) = I + JM_0,$$

where

$$I = a'(a(\boldsymbol{X})) \qquad J = b'(b(\boldsymbol{X}, a(\boldsymbol{X})).$$

(Note: $M = M_0 M_1$, and I and J can be calculated uniquely from $f(\boldsymbol{X})$.) Similarly, we specify the domains and ranges of the functions c', c, d', and d, which will be used to describe the two decoders, namely,

$$c': \{0, 1, \cdots, M_0 - 1\} \to \{0, 1, \cdots, M_0' - 1\}$$
$$c: \{0, 1, \cdots, M_0' - 1\} \to \hat{\mathscr{X}}_2^n$$
$$d': \{0, 1, \cdots, M_0 - 1\} \times \{0, 1, \cdots, M_1 - 1\} \times \mathscr{Y}^n$$
$$\to \{0, 1, \cdots, M_0' - 1\} \times \{0, 1, \cdots, M_1' - 1\}$$
$$d: \{0, 1, \cdots, M_0' - 1\} \times \{0, 1, \cdots, M_1' - 1\} \times \mathscr{Y}^n \to \hat{\mathscr{X}}_1^n.$$

The decoders are defined by

$$g_1(f(\boldsymbol{X}), \boldsymbol{Y}) = d(d'(I, J, \boldsymbol{Y}), \boldsymbol{Y})$$

and

$$g_2(f(\boldsymbol{X})) = c(c'(I)).$$

We shall now describe a procedure for randomly constructing the maps a, a', b, b', c, c', d, and d' such that

$$n^{-1} \log M \leq I(X; W) + I(X; U|Y, W) + \epsilon$$

and

$$Ed_i(X, g_i(X)) \leq D_i + \epsilon$$

for n sufficiently large. This will imply the existence of an $(n, M, D_1 + \epsilon, D_2 + \epsilon)$ code with $\boldsymbol{M}$ satisfying the above inequality, thereby allowing us to draw the desired conclusion that $R_0(\boldsymbol{D}) \geq R(\boldsymbol{D})$.

Let $\delta > 0$ be a small positive number to be specified later. In what follows we shall adopt the notation and conventions of Csiszár and Körner [3]. Let $T^n_{[W]_\delta}$ be the set of δ-typical W-vectors with length n. Choose vectors $\boldsymbol{W}_i$, $1 \leq i \leq M_0' - 1$, independently, according to a uniform distribution over $T^n_{[W]_\delta}$. For each such $\boldsymbol{W}_i$ choose vectors $\boldsymbol{U}_{ij}$, $1 \leq j \leq M_1' - 1$, independently according to a uniform distribution over $T^n_{[U|W]_\delta}(\boldsymbol{W}_i)$.

For each $\boldsymbol{x} \in T^n_{[X]_\delta}$, if a value of $1 \leq i \leq M_0' - 1$ can be found such that $\boldsymbol{W}_i \in T^n_{[W|X]_\delta}(x)$, set $a(\boldsymbol{x}) = i$ and let $\boldsymbol{w}_i$ denote the value assumed by $\boldsymbol{W}_i$. Otherwise (i.e., if $\boldsymbol{x} \notin T^n_{[X]_\delta}$ or $\boldsymbol{W}_i \notin T^n_{[W|X]_\delta}(x)$ for every $1 \leq i \leq M_0' - 1$), set $a(\boldsymbol{x}) = 0$. If $a(\boldsymbol{x}) = i > 0$ and $(\boldsymbol{x}, \boldsymbol{w}_i) \in T^n_{[X, W]_\delta}$, look for a value of $1 \leq j \leq M_1' - 1$ such that $\boldsymbol{U}_{ij} \in T^n_{[U|X, W]_\delta}(\boldsymbol{x}, \boldsymbol{w}_i)$ and set $b(\boldsymbol{x}, i) = j$; otherwise, set $b(\boldsymbol{x}, i) = 0$. Next, choose the maps a' and b' randomly over the set of all maps from $\{0, 1, \cdots, M_0' - 1\}$ to $\{0, 1, \cdots, M_0 - 1\}$ and from $\{0, 1, \cdots, M_1' - 1\}$ to $\{0, 1, \cdots, M_1 - 1\}$, respectively, that partition the domain into "equal"-size subsets (often referred to as "bins"). For example, in the case of a' we require

$$\left\lfloor \frac{M_0'}{M_0} \right\rfloor \leq |\{i: a'(i) = k\}| \leq \left\lceil \frac{M_0'}{M_0} \right\rceil$$

for every $0 \leq k \leq M_0 - 1$.

The decoding function $g_1 = d(d'(\cdot), \cdot)$ is constructed by specifying d' and d as follows. (The construction of the simpler decoder function g_2 will be discussed subsequently.) For each $(I, J, \boldsymbol{y}) \in \{0, 1, \cdots, M_0 - 1\} \times \{0, 1, \cdots, M_1 - 1\} \times T^n_{[Y]_\delta}$, search for a unique pair $(i, j) \in \{1, \cdots, M_0' - 1\} \times \{1, \cdots, M_1' - 1\}$ such that $a'(i) = I$, $b'(j) = J$, $\boldsymbol{W}_i \in T^n_{[W|Y]_\delta}(\boldsymbol{y})$ and $\boldsymbol{U}_{ij} \in T^n_{[U|Y, W]_\delta}(\boldsymbol{y}, \boldsymbol{W}_i)$. If such a pair exists, set $d'(I, J, \boldsymbol{y}) = (i, j)$; otherwise, set $d'(I, J, \boldsymbol{y}) = (0, 0)$. Finally, for each $(i, j, \boldsymbol{y}) \in \{1, \cdots, M_0 - 1\} \times \{1, \cdots, M_1 - 1\} \times T^n_{[Y]_\delta}$ such that $(\boldsymbol{W}_i, \boldsymbol{U}_{ij}, \boldsymbol{y}) \in T^n_{[W, U, Y]_\delta}$, define $d(i, j, \boldsymbol{y})$ to be the vector $\hat{\boldsymbol{X}}_1 \in \hat{\mathscr{X}}_1^n$ whose kth component is $\hat{X}_1(W_{ik}, U_{ijk}, y_k)$, where $\hat{X}_1(\cdot)$ is the function referred to in 2) of the definition of $P(\boldsymbol{D})$. Otherwise, set $d(i, j, \boldsymbol{y}) =$ constant $\in \hat{\mathscr{X}}_1^n$.

The construction of $g_2 = c(c'(\cdot))$ parallels that of g_1 in the special case in which Y is a degenerate random variable, say $Y =$ constant. In this simple case the random variables that play the role analogous to that of U_{ij} in the construction of g_1 also are degenerate, so they are excised from the argument. In the last step $\hat{X}_2$ replaces $\hat{X}_1$. We omit the details.

The distortion achieved by decoder g_1 can be bounded from above as follows:

$$\begin{aligned} Ed_1(\boldsymbol{X}, \hat{\boldsymbol{X}}_1) &\le \Pr(F)(D_1 + \delta D_{\max}) + \Pr(F^c) D_{\max} \\ &\le D_1 + (\delta + \Pr(F^c)) D_{\max}, \end{aligned}$$

where $D_{\max} = \max_{(x,\hat{x})} d_1(x, \hat{x})$ and F is the event that the encoding and g_1—decoding operations never involve any of the "otherwise" contingencies cited above and result in an (i, j) pair for which $(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{W}_i, \boldsymbol{U}_{ij}) \in T^n_{[Y,X,W,U]_\delta}$. That is,

$$F = G \cap A \cap B \cap C,$$

where

$$\begin{aligned} G &= \left\{(\boldsymbol{X}, \boldsymbol{Y}) \in T^n_{[X,Y]_\delta}\right\} \\ A &= \{a(\boldsymbol{X}) \ne 0\}, \; B = \{b(\boldsymbol{X}, a(\boldsymbol{X})) \ne 0\} \\ C &= \Big\{ d'(a'(a(\boldsymbol{X})), b'(b(\boldsymbol{X}, a(\boldsymbol{X}))), \boldsymbol{Y}) \\ &\qquad = (a(\boldsymbol{X}), b(\boldsymbol{X}, a(\boldsymbol{X}))) = (i_0, j_0) \\ &\qquad \text{such that } \left(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{W}_{i_0}, \boldsymbol{U}_{i_0 j_0}\right) \in T^n_{[Y,X,W,U]_\delta} \Big\}. \end{aligned}$$

It will suffice to show that $P(F^c) \to 0$ as $n \to \infty$ for

$$R \equiv n^{-1} \log M = n^{-1} \log M_0 + n^{-1} \log M_1 \equiv R_0 + R_1$$

exceeding $I(X; W) + I(X; U|Y, W)$ by less than ϵ. We have

$$\begin{aligned} \Pr(F^c) &= \Pr(G^c \cup A^c \cup B^c \cup C^c) \\ &\le \Pr(G^c) + \Pr(A^c|G) + \Pr(B^c|A \cap G) \\ &\quad + \Pr(C^c|A \cap B \cap G). \end{aligned}$$

By the law of large numbers $\Pr(G^c) < \epsilon'/5$, for $n \ge n_0(\delta, \epsilon')$,

$$\begin{aligned} \Pr(A^c|G) &= \Pr\left(\bigcap_{i=1}^{M_0'-1} \left\{ \boldsymbol{W}_i \in T^n_{[W|X]_\delta}(\boldsymbol{X}) \right\} | G \right) \\ &= \left(1 - \Pr\left(\boldsymbol{W}_1 \in T^n_{[W|X]_\delta}(\boldsymbol{X}) | G\right)\right)^{M_0'-1} \end{aligned}$$

and

$$\begin{aligned} &\Pr\left(\boldsymbol{W}_1 \in T^n_{[W|X]_\delta}(\boldsymbol{X}) | G\right) \\ &\qquad \ge \frac{\min_{\boldsymbol{x} \in T^n_{[X]_\delta}} |T^n_{[W|X]_\delta}(\boldsymbol{x})|}{|T^n_{[W]_\delta}|} \\ &\qquad \ge 2^{-n[I(X;W)+\gamma]}, \text{ where } \gamma \to 0 \text{ as } \delta \to 0. \end{aligned}$$

Thus, using $(1-x)^y < e^{-xy}$ for $x, y > 0$ and requiring $R_0' \equiv n^{-1} \log_2 M_0' \ge I(X; W) + 2\gamma$ uniformly in n, we deduce that

$$\Pr(A^c|G) \le e^{-(M_0'-1) 2^{-n[I(X;W)+\gamma]}} < \epsilon'/5$$

for $n \ge n_1(\gamma, \epsilon')$. A similar argument shows

$$\Pr(B^c|A \cap G) < \epsilon'/5$$

provided $R_1' \equiv n^{-1} \log_2 M_1' \ge I(X; U|W) + 2\gamma$ and $n \ge n_1(\gamma, \epsilon')$.

Next,

$$\begin{aligned} &\Pr(C^c|A \cap B \cap G) \\ &\quad = \Pr\Big(C^c_{i_0} \cup D^c_{i_0 j_0} \cup T^c_0 \cup \bigcup_{i \ne i_0} C_i \\ &\qquad \cup \bigcup_{j \ne j_0} D_{i_0 j} | A \cap B \cap G, i_0, j_0 \Big) \\ &\quad i_0 \equiv a(\boldsymbol{X}) \qquad j_0 \equiv b(\boldsymbol{X}, a(\boldsymbol{X})), \end{aligned}$$

where

$$\begin{aligned} C_i &= \left\{ \boldsymbol{W}_i \in T^n_{[W|Y]_\delta}(\boldsymbol{Y}), \; a'(i) = a'(i_0) \right\}, \\ D_{ij} &= \left\{ \boldsymbol{U}_{ij} \in T^n_{[U|Y,W]_\delta}(\boldsymbol{Y}, \boldsymbol{W}_i), \; b'(j) = b'(j_0) \right\}, \end{aligned}$$

and

$$T_0 = \left\{ \left(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{W}_{i_0}, \boldsymbol{U}_{i_0 j_0}\right) \in T^n_{[Y,X,W,U]_\delta} \right\}.$$

Therefore,

$$\begin{aligned} &\Pr(C^c|A \cap B \cap G) \\ &\quad \le \Pr\left(C^c_{i_0} \cup D^c_{i_0 j_0} \cup T^c_0 | A \cap B \cap G, i_0, j_0 \right) \\ &\qquad + \sum_{\substack{i \ne i_0 \\ a'(i) = a'(i_0)}} \Pr\left(C_i | A \cap B \cap G \cap C_{i_0}\right) \\ &\qquad + \sum_{\substack{j \ne j_0 \\ b'(j) = b'(j_0)}} \Pr\left(D_{i_0 j} | A \cap B \cap G \cap C_{i_0} \cap D_{i_0 j_0} \right). \end{aligned}$$

We now invoke the familiar result (cf. Wyner [4] or Berger [5]) that if

$$Y \ominus X \ominus (W, U) \tag{1}$$

$$\Pr\left((\boldsymbol{X}, \boldsymbol{Y}) \in T^n_{[X,Y]_\delta}\right) \to 1 \qquad \text{as } n \to \infty, \tag{2}$$

and

$$\Pr\left(\left(\boldsymbol{X}, \boldsymbol{W}_{i_0}, \boldsymbol{U}_{i_0 j_0}\right) \in T^n_{[X,W,U]_\delta}\right) \to 1 \qquad \text{as } n \to \infty, \tag{3}$$

then

$$\Pr\left(\left(Y, X, W_{i_0}, U_{i_0 j_0}\right) \in T^n_{[Y,X,W,U]_\delta}\right) \to 1 \qquad \text{as } n \to \infty.$$

In this case this result implies that for $n \ge n_2(\delta, \epsilon')$

$$\Pr\left(C^c_{i_0} \cup D^c_{i_0} \cup T^c_0 | A \cap B \cap G, i_0, j_0 \right) < \epsilon'/10.$$

Also, for $i \ne i_0$

$$\begin{aligned} \Pr\left(C_i | A \cap B \cap G \cap C_{i_0}\right) &\le \frac{\max_{\boldsymbol{y} \in T^n_{[Y]_\delta}} |T^n_{[W|Y]_\delta}(\boldsymbol{y})|}{|T^n_{[W]_\delta}|} \\ &\le 2^{-n[I(Y;W)-\gamma]}. \end{aligned}$$

Similarly, for $j \ne j_0$

$$\begin{aligned} &\Pr\left(D_{i_0 j} | A \cap B \cap G \cap C_{i_0} \cap D_{i_0 j_0} \right) \\ &\qquad \le \frac{\max_{(\boldsymbol{y}, \boldsymbol{w}) \in T^n_{[Y,W]_\delta}} |T^n_{[U|Y,W]_\delta}(\boldsymbol{y}, \boldsymbol{w})|}{\min_{\boldsymbol{w} \in T^n_{[W]_\delta}} |T^n_{[U|W]_\delta}(\boldsymbol{w})|} \\ &\qquad \le 2^{-n[I(Y;U|W)-\gamma]}. \end{aligned}$$

Thus

$$\Pr(C^c|A\cap B\cap G)\le \epsilon'/10+\frac{M_0'}{M_0}2^{-n[I(Y;W)-\gamma]}+\frac{M_1'}{M_1}2^{-n[I(Y;U|W)-\gamma]}\le \epsilon'/5$$

provided $n\ge n_1(\gamma,\epsilon)$ and

$$R_0\equiv n^{-1}\log M_0\ge n^{-1}\log M_0'-I(Y;W)+2\gamma\equiv R_0'-I(Y;W)+2\gamma$$

$$R_1\equiv n^{-1}\log M_1\ge n^{-1}\log M_1'-I(Y;U|W)+2\gamma\equiv R_1'-I(Y;U|W)+2\gamma.$$

Temporarily suppose that decoder 2 also was provided with side information in the form of another source $\{Z_k\}$ jointly distributed with $\{(X_k,Y_k)\}$; previously, we had been considering only the degenerate case $Z=\varnothing$. Then an argument paralleling that given above for decoder 1 would yield the analogous bounds

$$R_0\ge R_0'-I(Z;W)+2\gamma$$
$$R_2\ge R_2'-I(Z;V|W)+2\gamma,$$

where R_2 and R_2' are defined in the obvious way, V is another auxiliary random variable satisfying $(Y,Z)\ominus X\ominus(W,U,V)$ and $Ed_2(X,\hat{X}(W,V,Z))\le D_2$. The overall rate now is $R=n^{-1}\log M=n^{-1}\log M_0M_1M_2=R_0+R_1+R_2$. To complete the argument, set

$$R_0'=I(X;W)+2\gamma,$$
$$R_1'=I(X;U|W)+2\gamma,$$
$$R_2'=I(X;V|W)+2\gamma,$$

Then, the requirements on R_0, R_1, and R_2 are

$$R_0\ge I(X;W)-\min(I(Y;W),I(Z;W))+4\gamma=\max(I(X;W|Y),I(X;W|Z))+4\gamma,$$

$$R_1\ge I(X;U|W)-I(Y;U|W)+4\gamma+4\gamma=I(X;U|Y,W)+4\gamma,$$

and

$$R_2\ge I(X;V|W)-I(Z;V|W)+4\gamma=I(X;V|Z,W)+4\gamma,$$

yielding

$$R\ge\max(I(X;W|Y),I(X;W|Z))+I(X;U|Y,W)+I(X;V|Z,W)+12\gamma.$$

Taking $13\gamma<\epsilon$, $\delta<\epsilon'/5$ and $\epsilon'<\epsilon/D_{\max}$, we conclude that for $n\ge\max(n_0(\delta,\epsilon'),n_1(\gamma,\epsilon'),n_2(\gamma,\epsilon'))$ we will have

$$Ed_i(\hat{\boldsymbol{X}},\hat{\boldsymbol{X}}_i)<D_i+\epsilon,\qquad i=1,2$$

for

$$R<\max(I(X;W|Y),I(X;W|Z))+I(X;U|Y,W)+I(X;V|Z,W)+\epsilon.$$

In the special case $Z=\varnothing$ originally under consideration, we need not introduce V, so we conclude that for any $(W,U)\in P(\boldsymbol{D})$ there exists a $\boldsymbol{D}$-admissible rate satisfying

$$R\le I(X;W)+I(X;U|Y,W)+\epsilon.$$

The inequality $R(D)\le R_0(\boldsymbol{D})$ is established. (We generalized to $Z\ne\varnothing$ during the proof both in the interest of symmetry and in an attempt to generate appreciation for the form of Theorem 2 of Section VII, which treats a still more general situation.)

It remains to establish that the bounds on $|\mathscr{W}|$ and $|\mathscr{U}|$ specified by condition 3) in the definition of $P(\boldsymbol{D})$ do not affect the minimization. Toward that end we invoke the support lemma [3, p. 310] in order to deduce that $\mathscr{W}$ must have $|\mathscr{X}|-1$ letters in order to ensure preservation of $p(x|w)$ plus three more to preserve the constraints on D_1, D_2 and $I(X;W)$, so $|\mathscr{W}|=|\mathscr{X}|+2$ suffices. Similarly, $\mathscr{U}$ must have $|\mathscr{X}||\mathscr{W}|-1$ letters in order to ensure preservation of $p(x,w|u)$ plus two more to preserve D_1 and $I(X;U|Y,W)$. Thus, it suffices to have

$$|U|\le|X||W|-1+2=|X||W|+1\le|X|(|X|+2)+1=(|X|+1)^2.$$

IV. The Converse

Toward proving that $R(\boldsymbol{D})\ge R_0(\boldsymbol{D})$, we first show that $R_0(\boldsymbol{D})$ is convex. Let

$$\overline{R}_0(\boldsymbol{D})=\min_{\overline{P}(\boldsymbol{D})}[I(X;W|T)+I(X;U|Y,W,T)],$$

where $\overline{P}(\boldsymbol{D})$ is the set of all random variables (W,U,T) jointly distributed with the generic source variables (X,Y) such that the following hold.

1) $Y\ominus X\ominus(W,U,T)$ is a Markov string.

2) There exists $\hat{X}_1(W,U,T,Y)$ and $\hat{X}_2(W,T)$ such that $Ed_i(X,\hat{X}_i)\le D_i$, $i\in\{1,2\}$.

3) T is independent of (X,Y).

Note that $\overline{R}_0(\boldsymbol{D})$ is the lower convex envelope of $R_0(\boldsymbol{D})$. Since $I(X;T)=0$ we have $I(X;W|T)=I(X;W,T)$. Upon defining a new random variable $W'=(W,T)$, we see that $\overline{R}_0(\boldsymbol{D})\ge R_0(\boldsymbol{D})$. Thus $R_0(\boldsymbol{D})$ lies on or below its lower convex envelope, so $R_0(\boldsymbol{D})$ must be convex.

We now shall show that if an (n,M,D_0,D_1) code exists, then

$$n^{-1}\log M\ge I(X;W)+I(X;U|Y,W)$$

for random variables $(W,U)\in P(\boldsymbol{D})$. If we define $J=f(\boldsymbol{X})$, then

$$nR\equiv\log M\ge H(J)\ge I(\boldsymbol{X};J)=I(\boldsymbol{X};J,\boldsymbol{Y})-I(\boldsymbol{X};\boldsymbol{Y}|J)=\sum_{k=1}^{n}[I(X_k;J,\boldsymbol{Y}|\boldsymbol{X}_k^-)-I(\boldsymbol{X};Y_k|J,\boldsymbol{Y}_k^-)],$$

where $\boldsymbol{X}_1^-\equiv\varnothing$ and $\boldsymbol{X}_k^-\equiv(X_1,X_2,\cdots,X_{k-1})$ for $k>1$; similarly, $\boldsymbol{Y}_n^+\equiv\varnothing$ and $\boldsymbol{Y}_k^+=(Y_{k+1},Y_{k+2},\cdots,Y_n)$ for $i<n$, etc. Since the source is memoryless, X_k is independent of $\boldsymbol{X}_k^-$, so $I(X_k;J,\boldsymbol{Y}|\boldsymbol{X}_k^-)=I(X_k;J,\boldsymbol{Y},\boldsymbol{X}_k^-)\ge I(X_k;J,\boldsymbol{Y})$. Also $Y_k\ominus X_k\ominus(\boldsymbol{X}_k^-,X_k,J,\boldsymbol{Y}_k^-)$, so

$I(X; Y_k|J, Y_k^-) = I(X_k; Y_k|J, Y_k^-)$. Thus

$$\begin{aligned} R &\geq n^{-1} \sum_{k=1}^{n} \left[I(X_k; J, Y) - I(X_k; Y_k|J, Y_k^-) \right] \\ &= n^{-1} \sum_{k=1}^{n} \left[I(X_k; J, Y_k^-) + I(X_k; Y_k^+|J, Y_k^-, Y_k) \right] \\ &= n^{-1} \sum_{k=1}^{n} \left[I(X_k; W_k) + I(X_k; U_k|W_k, Y_k) \right], \end{aligned}$$

where $W_k \equiv (J, Y_k^-)$ and $U_k \equiv Y_k^+$. Note that $Y_k \ominus X_k \ominus (W_k, U_k)$. It follows via standard arguments invoking the convexity of $R_0(D)$ that $R \geq R_0(\boldsymbol{D})$ and therefore that $R(\boldsymbol{D}) \geq R_0(\boldsymbol{D})$. This completes the proof of the converse, thereby establishing Theorem 1.

V. The Quadratic Gaussian Case

Suppose $Y = X + Z$, where X and Z are independent, (zero-mean) Gaussian random variables with respective variances σ_X^2 and σ_Z^2. Since the source alphabets $\mathscr{X}$ and $\mathscr{Y}$ are no longer finite, a formal justification of the extension of the results to the case at hand is in order. We omit this justification, since it directly parallel's Wyner's [7] extension of his work with Ziv [1]. The minimization defining $R(D_1, D_2)$ can be carried out explicitly in this case. Not surprisingly, it occurs when $U = X + Z_1$ and $W = X + Z_2$ with Z, Z_1, and Z_2 all independent, zero mean, and Gaussian; i.e., U, W, and Y correspond to observations of X via independent channels with additive Gaussian noise. Decoder 1 forms the optimum estimate of X given (U, W, Y), which is known to be a linear combination of the form

$$\hat{X}_1 = \alpha U + \beta W + \gamma Y.$$

Decoder 2 sees only W and therefore forms

$$\hat{X}_2 = \delta W.$$

It is easy to see that the minimum of $I(X; W) + I(X; U|Y, W)$ subject to $E(X - \hat{X}_i)^2 = D_i$, $i = 1, 2$, occurs when

$$I(X; W) = I(X; \hat{X}_2) = \frac{1}{2} \log\left(\sigma_X^2 / D_2\right)$$

and

$$\begin{aligned} I(X; U|Y, W) &= I(X; \hat{X}_1|Y, W) \\ &= \frac{1}{2} \log\left(\frac{\sigma^2(X|Y, W)}{D_1} \right), \end{aligned}$$

where $\sigma^2(X|Y, W)$ is the conditional variance of X given Y and W. When W already is known, receiving knowledge of Y reduces the conditional variance of X from D_2 to $D_2\sigma_Z^2/(D_2 + \sigma_Z^2)$. It follows that

$$R(D_1, D_2) = \begin{cases} \dfrac{1}{2} \ln\left(\dfrac{\sigma_X^2 \sigma_Z^2}{D_1(D_2 + \sigma_Z^2)} \right), & \\ \qquad \text{if } D_1 \leq \dfrac{D_2\sigma_Z^2}{D_2 + \sigma_Z^2},\ D_2 \leq \sigma_X^2 & \\ \dfrac{1}{2} \ln\left(\dfrac{\sigma_X^2}{D_2} \right), \qquad \text{if } D_1 \geq \dfrac{D_2\sigma_Z^2}{D_2 + \sigma_Z^2},\ D_2 \leq \sigma_X^2 & \\ \dfrac{1}{2} \ln\left(\dfrac{\sigma_X^2 \sigma_Z^2}{D_1(\sigma_X^2 + \sigma_Z^2)} \right), & \\ \qquad \text{if } D_1 \leq \dfrac{D_2\sigma_Z^2}{D_2 + \sigma_Z^2},\ D_2 > \sigma_X^2 & \\ 0, \qquad \text{if } D_1 \geq \dfrac{D_2\sigma_Z^2}{D_2 + \sigma_Z^2},\ D_2 > \sigma_X^2. & \end{cases}$$

A sketch of $R(D_1, D_2)$ when $\sigma_X^2 = \sigma_Z^2 = 1$ is shown in Fig. 2.

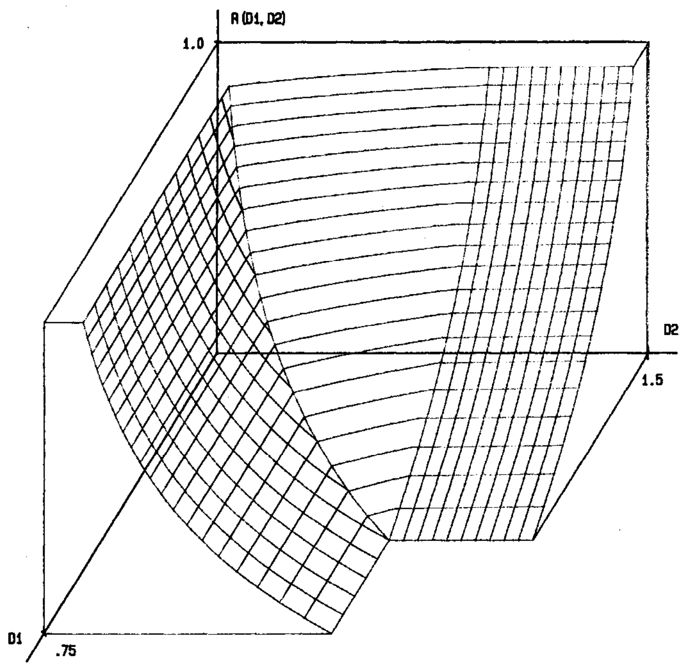

Fig. 2. $R(D_1, D_2)$ in quadratic Gaussian case when $\sigma_x^2 = \sigma_z^2 = 1$.

VI. The Binary Hamming Case

Now suppose that X is the input to and Y the output of a hypothetical binary symmetric channel (BSC) of crossover probability $p < 1/2$. Assume that $P(X = 0) = P(X = 1) = 1/2$ and that distortion is measured by Hamming distance, $d(x, \hat{x}) = (x + \hat{x})$ modulo 2. Let the random variable U' be the result of passing X through another BSC with crossover probability β that operates independently of that which connects X and Y, where $\beta \leq D_0 \equiv \min(p, D_2)$. Let U be the result of passing U' through a binary erasure channel (BEC) with erasure probability e that operates independently of both the aforementioned BSC's. Finally, let W be the result of passing U' through another independent BSC whose crossover probability is such that $Ed(X, W) = D_2$ (see Fig. 3).

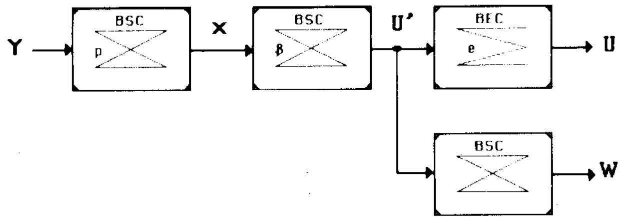

Fig. 3. Joint distribution of (X, Y, U, W) in binary Hamming case.

It is clear that the requirement $Y \ominus X \ominus (U, W)$ is satisfied. Moreover, the $\hat{X}_1(U, W, Y)$ that minimizes $Ed(X, \hat{X}_1)$ is

$$\hat{X}_1 = \begin{cases} U, & \text{if no erasure occurs} \\ W, & \text{if an erasure occurs and } D_2 < p \\ Y, & \text{if an erasure occurs and } D_2 \geq p, \end{cases}$$

and the minimized average distortion that results is

$$D_1 = (1 - e)\beta + e \min(p, D_2).$$

For given (D_1, D_2) and β this equation specifies $e = e(\beta)$. When we evaluate $I(X; W) + I(X; U|WY)$, it equals

$$1 - h(D_2) + [1 - e(\beta)][g(\beta) - g(D_2)],$$

where $h(x)$ is the binary entropy function, $g(x) = h(p \cdot x) - h(x)$, and $p \cdot x = p(1 - x) + (1 - p)x$. It follows that

$$R(D_1, D_2) \leq 1 - h(D_2) + \min_{\beta}[1 - e(\beta)][g(\beta) - g(D_2)].$$

The minimum is over $\beta \in [0, \min(p, D_2)]$, and $e(\beta)$ is specified by the above equation for D_1.

We believe that $R(D_1, D_2)$ may equal the right-hand side of the previous equation, but we have not yet attempted to extend the argument of Wyner and Ziv [1] in order to prove that our choice of (W, U) is the optimum one in $P(\boldsymbol{D})$. In the special cases that correspond to ordinary rate distortion and to the Wyner–Ziv problem, our candidate expression reduces to the known correct answer.

VII. Generalization to Several Informed Decoders

Let $(\mathscr{X}, \mathscr{Y}_1, \mathscr{Y}_2, \cdots, \mathscr{Y}_m, p(x, y_1, y_2, \cdots, y_m))$ be a discrete memoryless multisource with generic random variables $X, Y_1, Y_2, \cdots, Y_m$. For $1 \leq i \leq m$ let $\hat{\mathscr{X}}_i$ be a reconstruction alphabet and, let

$$d_i: \mathscr{X} \times \hat{\mathscr{X}}_i \to [0, \infty)$$

be a distortion measure. An $(n, M, \boldsymbol{D})$ code consists of an encoder

$$f: \mathscr{X}^n \to \{0, 1, \cdots, M - 1\}$$

and decoders $g = (g_1, g_2, \cdots, g_m)$ satisfying

$$g_i: \{0, 1, \cdots, M - 1\} \times \mathscr{Y}_i^n \to \hat{\mathscr{X}}_i^n, \qquad 1 \leq i \leq m.$$

The expected distortion $\boldsymbol{D} = (D_1, \cdots, D_m)$ for the code (f, g) is given by

$$D_i = Ed_i(X, \hat{X}_i) \equiv \frac{1}{n} \sum_{k=1}^{n} d_i(X_k, \hat{X}_{ik}),$$

where

$$\hat{X}_i = g_i(f(X), Y_i).$$

The rate R is $\boldsymbol{D}$-admissible if for every $\epsilon > 0$ there exists for some n an $(n, M, \boldsymbol{D} + \epsilon)$ code with $n^{-1} \log M \leq R + \epsilon$. The rate-distortion function is defined by

$$R(\boldsymbol{D}) = \inf\{R: R \text{ is } \boldsymbol{D}\text{-admissible}\}.$$

To upper bound $R(\boldsymbol{D})$ for this generalized problem, we must introduce an auxiliary random variable U_T for each of the $2^m - 1$ nonempty subsets $T \subset \{1, \cdots, m\}$. In the associated generalization of the random coding argument of Section III, n-vectors U_T will be selected at random from $T^n_{[U_T]\delta}$. Decoder i is able to recover correctly the $\boldsymbol{U}_T$ designated by the encoder with probability approaching 1 as $n \to \infty$ if $i \in T$. Let

$$V_T \equiv (U_S: T \subset S, S \neq T).$$

Before proceeding, we provide verbal interpretations for U_T and V_T. For this purpose let "group T" refer to those decoders whose indices belong to T, group $T = \{\text{decoder } i: i \in T\}$.

1) U_T is block recoverable only by the decoders in group T. That is, U_T is the *private information* that the members of group T share only with one another.

2) V_T is the set of all auxiliary random variables that are block recoverable by all the decoders in group T and also by one or more decoders not in group T. That is, $V_T \equiv \{U_S: U_S \in V_T\}$ is the *common information* that the members of group T share not only with one another but also with one or more outsiders.

Clearly, (V_T, U_T) represents the *total information* shared by all the members of group T. In the special case $T = \{i\}$, $U_{\{i\}}$ is the information known only to decoder i, $V_{\{i\}}$ is the information decoder i shares with others, and $(V_{\{i\}}, U_{\{i\}})$ is the totality of blocks of auxiliary random variables that decoder i is capable of recovering.

Let $P(\boldsymbol{D})$ denote the set of all discrete random variables $(U_T, \varnothing \neq T \subset \{1, \cdots, m\})$ jointly distributed with the generic source variables (X, Y) such that the following two conditions are satisfied.

1) $Y \ominus X \ominus (U_T, \varnothing \neq T \subset \{1, \cdots, m\})$.

2) There exists $\hat{X}_i(V_{\{i\}}, U_{\{i\}}, Y_i)$ such that $Ed_i(X, \hat{X}_i) \leq D_i$, $1 \leq i \leq m$.

Theorem 2: $R(\boldsymbol{D}) \le R_0(\boldsymbol{D})$, where

$$R_0(\boldsymbol{D}) \equiv \min_{P(D)} \sum_{\varnothing \neq T \subset \{1,\cdots,m\}} \max_{i \in T} I(X; U_T | Y_i, V_T).$$

Proof: One may employ a multiterminal random coding argument that, albeit involved, nonetheless can be considered relatively straightforward by today's standards [3]. In the interest of conservation and simplicity, we omit the details.

We have not been able to prove the converse of Theorem 2 for the general case. The special case in which $\boldsymbol{D} = \boldsymbol{0}$ is subsumed by a theorem of Sgarro [6]. In that case it suffices to set $U_{\{1,\cdots,m\}} = U$ and $U_T =$ constant for all other values of T. Then, Theorem 2 yields $R(\boldsymbol{D}) \le \max_i H(X|Y_i)$, which Sgarro has shown to be tight.

In the case in which the generic variables of the multisource satisfy the degradedness condition

$$X \ominus Y_1 \ominus Y_2 \ominus \cdots \ominus Y_m,$$

we can show that $R(\boldsymbol{D}) = R_0(\boldsymbol{D})$. First, let us note the simplifications that result from said degradedness. If $i < j$ then decoder i can block-recover any U_T that decoder j can recover. This implies that there are only m nontrivial auxiliary random variables, namely, $U_{\{1\}}, U_{\{1,2\}}, \cdots$ and $U_{\{1,\cdots,m\}}$. For simplicity of notation we denote these subsequently by $W_1, W_2, \cdots,$ and W_m, respectively. It is easy to see that for $1 \le i \le m$

$$V_{\{1,\cdots,i\}} = (W_{i+1}, \cdots, W_m)$$
$$(V_{\{i\}}, U_{\{i\}}) = (W_i, W_{i+1}, \cdots, W_m).$$

It follows that in the degraded case

$$R_0(\boldsymbol{D}) = \min_{P(D)} \sum_{i=1}^{m} I(X; W_i | Y_i, W_{i+1}, \cdots, W_m),$$

where $P(\boldsymbol{D})$ is the set of all W jointly distributed with (X, Y) such that

1) $Y \ominus X \ominus W$.

2) There exists $\hat{X}_i(W_i, W_{i+1}, \cdots, W_m, Y_i)$ such that $Ed_i(X, \hat{X}_i) \le D_i$, $1 \le i \le m$.

In deducing this simplified form of $R_0(\boldsymbol{D})$, we have used the fact that the degradedness condition and condition 1) together imply that the maximum of $I(X; W_i | Y_j, W_{i+1}, \cdots, W_m)$ over $j \le i$ occurs for $j = i$.

Remark: To imbed the problem treated in Sections II and III into the more general problem currently under consideration, make the following associations: $m = 2$, $Y_1 = Y$, $Y_2 =$ constant, $W_1 = U$, and $W_2 = W$.

Theorem 3: For the degraded multisource with $X \ominus Y_1 \ominus Y_2 \ominus \cdots \ominus Y_m$, we have $R(\boldsymbol{D}) = R_0(\boldsymbol{D})$.

Proof: From Theorem 2 we know that $R(\boldsymbol{D}) \le R_0(\boldsymbol{D})$. Given an $(n, M, \boldsymbol{D})$ code, let $J = f(\boldsymbol{X})$ denote the output of the encoder. Let $\boldsymbol{Y}_i = (Y_{i,j}, 1 \le j \le n)$, $\boldsymbol{Y}_{i,k}^- = (Y_{i,j}, 1 \le j < k)$ and $\boldsymbol{Y}_{i,k}^+ = (Y_{i,j}, k < j \le n)$; whenever no ambiguity results, contract notation from $\boldsymbol{Y}_{i,k}^{\pm}$ to $\boldsymbol{Y}_i^{\pm}$. We shall show that $n^{-1} \log M \ge R_0(\boldsymbol{D})$.

$$\begin{aligned}
nR &\equiv \log M \ge H(J) \ge I(\boldsymbol{X}; J | \boldsymbol{Y}_m) \\
&= I(\boldsymbol{X}; J, \boldsymbol{Y}_1, \boldsymbol{Y}_2, \cdots, \boldsymbol{Y}_{m-1} | \boldsymbol{Y}_m) - I(\boldsymbol{X}; \boldsymbol{Y}_{m-1} | J, \boldsymbol{Y}_m) \\
&\quad - I(\boldsymbol{X}; \boldsymbol{Y}_{m-2} | J, \boldsymbol{Y}_{m-1}, \boldsymbol{Y}_m) - \cdots \\
&\quad - I(\boldsymbol{X}; \boldsymbol{Y}_1 | J, \boldsymbol{Y}_2, \cdots, \boldsymbol{Y}_m) \\
&= \sum_{k=1}^{n} \left[I(X_k; J, \boldsymbol{Y}_1, \boldsymbol{Y}_2, \cdots, \boldsymbol{Y}_{m-1} | \boldsymbol{Y}_m, \boldsymbol{X}^-) \right. \\
&\quad - I(\boldsymbol{X}; Y_{m-1,k} | J, \boldsymbol{Y}_{m-1}^-, \boldsymbol{Y}_m) \\
&\quad - I(\boldsymbol{X}; Y_{m-2,k} | J, \boldsymbol{Y}_{m-2}^-, \boldsymbol{Y}_{m-1}, \boldsymbol{Y}_m) - \cdots \\
&\quad \left. - I(\boldsymbol{X}; Y_{1,k} | J, \boldsymbol{Y}_1^-, \boldsymbol{Y}_2, \cdots, \boldsymbol{Y}_m) \right].
\end{aligned}$$

Since $(X_k, Y_{m,k})$ is independent of $(\boldsymbol{X}_k^-, \boldsymbol{Y}_{m,k}^-, \boldsymbol{Y}_{m,k}^+)$, we have

$$\begin{aligned}
I(X_k; &J, \boldsymbol{Y}_1, \boldsymbol{Y}_2, \cdots, \boldsymbol{Y}_{m-1} | \boldsymbol{Y}_m, \boldsymbol{X}^-) \\
&= I(X_k; J, \boldsymbol{Y}_1, \boldsymbol{Y}_2, \cdots, \boldsymbol{Y}_{m-1}, \boldsymbol{Y}_m^-, \boldsymbol{Y}_m^+, \boldsymbol{X}^{-1} Y_{m,k}) \\
&\ge I(X_k; J, \boldsymbol{Y}_1, \boldsymbol{Y}_2, \cdots, \boldsymbol{Y}_{m-1}, \boldsymbol{Y}_m^-, \boldsymbol{Y}_m^+ | Y_{m,k}).
\end{aligned}$$

The Markov string

$$Y_{m-1,k} \ominus (X_k, Y_{m,k}) \ominus (\boldsymbol{X}_k^-, \boldsymbol{X}_k^+, J, \boldsymbol{Y}_{m-1,k}^-, \boldsymbol{Y}_{m,k}^-, \boldsymbol{Y}_{m,k}^+)$$

implies

$$I(\boldsymbol{X}; Y_{m-1,k} | J, \boldsymbol{Y}_{m-1}^-, \boldsymbol{Y}_m) = I(X_k; Y_{m-1,k} | J, \boldsymbol{Y}_{m-1}^-, \boldsymbol{Y}_m).$$

More generally, for $1 \le i \le m-1$ the Markov string

$$Y_{i,k} \ominus (X_k, Y_{i+1,k}, Y_{i+2,k}, \cdots, Y_{m,k}) \ominus (\boldsymbol{X}^-, \boldsymbol{X}^+, J, \boldsymbol{Y}_i^-, \boldsymbol{Y}_{i+1}^-, \boldsymbol{Y}_{i+1}^+, \cdots, \boldsymbol{Y}_m^-, \boldsymbol{Y}_m^+)$$

implies

$$\begin{aligned}
I(\boldsymbol{X}; &Y_{i,k} | J, \boldsymbol{Y}_i^-, \boldsymbol{Y}_{i+1}, \boldsymbol{Y}_{i+2}, \cdots, \boldsymbol{Y}_m) \\
&= I(X_k; Y_{i,k} | J, \boldsymbol{Y}_i^-, \boldsymbol{Y}_{i+1}, \boldsymbol{Y}_{i+2}, \cdots, \boldsymbol{Y}_m).
\end{aligned}$$

Thus

$$\begin{aligned}
R &\ge \frac{1}{n} \sum_{k=1}^{n} \left[I(X_k; J, \boldsymbol{Y}_1, \boldsymbol{Y}_2, \cdots, \boldsymbol{Y}_{m-1}, \boldsymbol{Y}_m^-, \boldsymbol{Y}_m^+ | Y_{m,k}) \right. \\
&\quad - I(X_k; Y_{m-1,k} | J, \boldsymbol{Y}_{m-1}^-, \boldsymbol{Y}_m) \\
&\quad - I(X_k; Y_{m-2,k} | J, \boldsymbol{Y}_{m-2}^-, \boldsymbol{Y}_{m-1}, \boldsymbol{Y}_m) \\
&\quad \vdots \\
&\quad \left. - I(X_k; Y_{1,k} | J, \boldsymbol{Y}_1^-, \boldsymbol{Y}_2, \cdots, \boldsymbol{Y}_m) \right] \\
&= \frac{1}{n} \sum_{k=1}^{n} \left[I(X_k; J, \boldsymbol{Y}_{m-1}^-, \boldsymbol{Y}_m^-, \boldsymbol{Y}_m^+ | Y_{m,k}) \right. \\
&\quad + I(X_k; \boldsymbol{Y}_1, \boldsymbol{Y}_2, \cdots, \boldsymbol{Y}_{m-2}, \boldsymbol{Y}_{m-1}^+ | Y_{m-1,k}, J, \boldsymbol{Y}_{m-1}^-, \boldsymbol{Y}_m) \\
&\quad - I(X_k; Y_{m-2,k} | J, \boldsymbol{Y}_{m-2}^-, \boldsymbol{Y}_{m-1}, \boldsymbol{Y}_m) \\
&\quad \vdots \\
&\quad \left. - I(X_k; Y_{1,k} | J, \boldsymbol{Y}_1^-, \boldsymbol{Y}_2, \cdots, \boldsymbol{Y}_m) \right].
\end{aligned}$$

Again, the first negative term in the summation may be canceled by a portion of the positive term immediately preceding it. Proceeding in this manner yields

$$\begin{aligned} R \geq \frac{1}{n} \sum_k & I(X_k; J, Y_{m-1}^-, Y_m^-, Y_m^+ | Y_{m,k}) \\ & + I(X_k; Y_{m-2}^-, Y_{m-1}^+ | Y_{m-1,k}, J, Y_{m-1}^-, Y_m) \\ & + I(X_k; Y_{m-3}^-, Y_{m-2}^+ | Y_{m-2,k}, J, Y_{m-2}^-, Y_{m-1}, Y_m) \\ & \vdots \\ & + I(X_k; Y_1^+ | Y_{1,k}, J, Y_1^-, Y_2, \cdots, Y_m). \end{aligned}$$

Degradedness gives the Markov string

$$Y_{m,k} \ominus Y_{m-1,k} \ominus (X_k, J, Y_{m-1}^-, Y_m^-, Y_m^+),$$

which implies that

$$\begin{aligned} I(X_k; & Y_{m-2}^-, Y_{m-1}^+ | Y_{m-1,k}, J, Y_{m-1}^-, Y_m) \\ & = I(X_k; Y_{m-2}^-, Y_{m-1}^+, Y_{m,k} | Y_{m-1,k} J, Y_{m-1}^-, Y_m^-, Y_m^+) \\ & \geq I(X_k; Y_{m-2}^-, Y_{m-1}^+ | Y_{m-1,k}, J, Y_{m-1}^-, Y_m^-, Y_m^+). \end{aligned}$$

Similarly, degradedness yields

$$\begin{aligned} (Y_{m-1,k}, Y_{m,k}) & \ominus Y_{m-2,k} \\ & \ominus (X_k, J, Y_{m-2}^-, Y_{m-1}^-, Y_{m-1}^+, Y_m^-, Y_m^+), \end{aligned}$$

which implies that

$$\begin{aligned} I(X_k; & Y_{m-3}^-, Y_{m-2}^+ | Y_{m-2,k}, J, Y_{m-2}^-, Y_{m-1}, Y_m) \\ & \geq I(X_k; Y_{m-3}^-, Y_{m-2}^+ | Y_{m-2,k}, J, \\ & \qquad Y_{m-2}^-, Y_{m-1}^-, Y_{m-1}^+, Y_m^-, Y_m^+). \end{aligned}$$

We continue iteratively in this vein until finally observing that the Markov string

$$\begin{aligned} (Y_{2,k}, Y_{3,k}, \cdots, Y_{m,k}) \ominus Y_{1,k} \ominus X_k, J, \\ Y_1^-, Y_2^-, Y_2^+, \cdots, Y_m^-, Y_m^+ \end{aligned}$$

implies

$$\begin{aligned} I(X_k; & Y_1^+ | Y_{1,k}, J, Y_1^-, Y_2, \cdots, Y_m) \\ & \geq I(X_k; Y_1^+ | Y_{1,k}, J, Y_1^-, Y_2^-, Y_2^+, \cdots, Y_m^-, Y_m^+). \end{aligned}$$

Thus, if we set

$$\begin{aligned} W_{m,k} &= (J, Y_m^-, Y_{m-1}^-, Y_m^+) \\ W_{m-1,k} &= (Y_{m-2}^-, Y_{m-1}^+) \\ W_{m-2,k} &= (Y_{m-3}^-, Y_{m-2}^+) \\ & \vdots \\ W_{1,k} &= Y_1^+, \end{aligned}$$

we may write

$$\begin{aligned} R \geq \frac{1}{n} \sum_{k=1}^{n} & [I(X_k; W_{m,k} | Y_{m,k}) \\ & + I(X_k; W_{m-1,k} | Y_{m-1,k}, W_{m,k}) \\ & + I(X_k; W_{m-2,k} | Y_{m-2,k}, W_{m-1,k}, W_{m,k}) \\ & + \cdots + I(X_k; W_{1,k} | Y_{1,k}, W_{2,k}, \cdots, W_{m,k})] \\ = \frac{1}{n} \sum_{k=1}^{n} \sum_{i=1}^{m} & I(X_k; W_{i,k} | Y_{i,k}, W_{i+1,k}, \cdots, W_{m,k}). \end{aligned}$$

Observe that

$$\begin{aligned} (Y_{i,k}, 1 \leq i \leq m) \ominus X_k \ominus (W_{i,k}, 1 \leq i \leq m) \\ = (J, Y_1^-, Y_1^+, Y_2^-, Y_2^+, \cdots, Y_m^-, Y_m^+), \end{aligned}$$

so the $W_{i,k}$ satisfy 1) in the definition of $P(\boldsymbol{D})$ for each $1 \leq k \leq n$. Also,

$$\begin{aligned} (W_{i,k}, & W_{i+1,k}, \cdots, W_{m,k}, Y_{i,k}) \\ & = (Y_{i-1}^-, Y_i^+, Y_i^-, Y_{i+1}^+, \cdots, Y_{m-2}^-, Y_{m-1}^+, \\ & \qquad J, Y_{m-1}^-, Y_m^-, Y_m^+, Y_{i,k}) \end{aligned}$$

contains (J, Y_i), which is the totality of information available to decoder i. Hence $\hat{X}_{i,k}(W_{i,k}, \cdots, W_{m,k}, Y_{i,k})$ exists such that $Ed_i(X_k, \hat{X}_{i,k}) \leq D_{i,k}$, where $D_{i,k}$ is the average distortion with which decoder i reproduces X_k. It follows that

$$R \geq \frac{1}{n} \sum_{k=1}^{n} R_0(\boldsymbol{D}_k),$$

where $\boldsymbol{D}_k = (D_{1,k}, \cdots, D_{m,k})$. An argument similar to that employed in Section III shows that $R_0(\cdot)$ is convex, so

$$R \geq R_0\left(\frac{1}{n} \sum_{k=1}^{n} \boldsymbol{D}_k\right) = R_0(\boldsymbol{D}),$$

and Theorem 3 is proved.

The degradedness condition $X \ominus Y_1 \ominus Y_2 \ominus \cdots \ominus Y_m$ need not be physical. Also, $R(\boldsymbol{D})$ clearly depends on $p(x, y)$ only through its second-order marginals $\{p(x, y_i), 1 \leq i \leq m\}$.

Acknowledgment

We are grateful to Dr. János Körner for pointing out that Theorem 2 subsumes Sgarro's result, to Dr. Amiram Kaspi for suggestions regarding the alphabet cardinality bounds, to Ken Kerpez for computing the quadratic Gaussian example, and to the referees for their meticulousness.

References

[1] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 1–10, Jan. 1976.

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

[3] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.

[4] A. D. Wyner, "On source coding with side information at the decoder," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 163–179, 1975.

[5] T. Berger, "Multiterminal source coding," in *Information Theory Approach to Communications*, G. Longo, Ed. CISM Courses and Lecture Notes 229. New York: Springer, pp. 172–231, 1977.

[6] A. Sgarro, "Source coding with side information at several decoders," *IEEE Trans. Inform. Theory*, vol. IT-23, pp. 179–182, 1977.

[7] A. D. Wyner, "The rate-distortion function for source coding with side information at the decoder-II: General sources," *Inform. Contr.*, vol. 38, pp. 60–80, 1978.