INFORMATION AND CONTROL 38, 60–80 (1978)

The Rate-Distortion Function for Source Coding with Side Information at the Decoder–II: General Sources

A. D. WYNER

Bell Telephone Laboratories, Inc., Murray Hill, New Jersey 07974

In this paper we generalize (to nondiscrete sources) the results of a previous paper (Wyner and Ziv, 1976) on source coding with a fidelity criterion in a situation where the decoder (but not the encoder) has access to side information about the source. We define $R^*(d)$ as the minimum rate (in the usual Shannon sense) required for encoding the source at a distortion level about d. The main result is the characterization of $R^*(d)$ by an information theoretic minimization. In a special case in which the source and the side information are jointly Gaussian, it is shown that $R^*(d)$ is equal to the rate which would be required if the encoder (as well as the decoder) is informed of the side information.

1. INTRODUCTION

In this paper we generalize (to nondiscrete sources) the results of Wyner and Ziv (1976) on source coding with a fidelity criterion in a situation where the decoder has access to side information about the source. Our problem concerns the system shown in Fig. 1. The sequence $\{(X_k, Y_k)\}_{k=1}^{\infty}$ represents independent

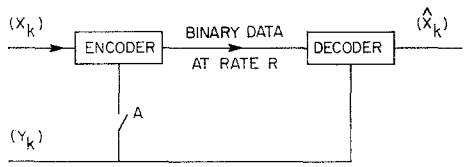

FIGURE 1

copies of a pair of dependent random variables (X, Y) which take values in the arbitrary (i.e., discrete or nondiscrete) spaces $\mathcal{X}$, $\mathcal{Y}$, respectively. The *encoder* output is a binary sequence which appears at rate R bits per input symbol. The *decoder* output is a sequence $\{\hat{X}_k\}_1^{\infty}$ which takes values in an arbitrary "reproduction" space $\hat{\mathcal{X}}$. The encoding and decoding is done in blocks of length n, and the fidelity criterion is $E(1/n)\sum_{k=1}^{\infty} D(X_k, \hat{X}_k)$, where $D: \mathcal{X} \times \hat{\mathcal{X}} \to [0, \infty)$ is a given distortion function. If switch A is closed, then the encoder,

0019-9958/78/0381-0060$02.00/0

as well as the decoder is assumed to have knowledge of the side information sequence $\{Y_k\}$. If switch A is open, then only the decoder has access to the side information. For $d \geqslant 0$, we are interested in the minimum rate R for which the system of Fig. 1 can operate with n large and average distortion $E(1/n)\sum_1^n D(X_k, \hat{X}_k)$ arbitrarily close to d. We define this minimum rate to be $R_{X|Y}(d)$ when switch A is closed, and $R^*(d)$ when switch A is open.

For the special case where the spaces $\mathscr{X}$, $\mathscr{Y}$, $\hat{\mathscr{X}}$ are finite, $R_{X|Y}(d)$ was found by Berger [1971, Section 6.1, Case 4], and $R^*(d)$ by Slepian and Wolf (1973) and Wyner and Ziv (1976). The relatively simple generalization of the coding theorems concerning $R_{X|Y}(d)$ to arbitrary $\mathscr{X}$, $\mathscr{Y}$, $\hat{\mathscr{X}}$ is discussed in Appendix A. The generalization of the theorems concerning $R^*(d)$ (which are given for the finite alphabet case in Wyner and Ziv) is, we as shall see, somewhat more delicate. In fact, this generalization is the main contribution of this paper.

We remark at this point that the proof of the (direct) coding theorem given in Wyner and Ziv for the discrete case depends rather heavily on the finiteness of the alphabets $\mathscr{X}$ and $\mathscr{Y}$. In many other situations in the Shannon theory, such proofs can be easily adapted to the nondiscrete case by finding appropriate discrete approximations to nondiscrete random objects. In the present problem, however, this approach is particularly difficult. Among the reasons is the following: Let X, Y, Z be a "chain" of nondiscrete random variables, i.e., X, Z are conditionally independent given Y. Let $\tilde{X}$, $\tilde{Y}$, $\tilde{Z}$ be finite approximations (i.e., the results of quantizations) to X, Y, Z, respectively. Then $\tilde{X}$, $\tilde{Y}$, $\tilde{Z}$ is not necessarily a chain. Since chains play a vital role in the proofs here and in other multiple-user Shannon theory problems, it is clear that we must proceed with care.

In Section 2 we give a formal statement of our problem and state our main results. In Section 3 we calculate $R_{X|Y}(d)$ and $R^*(d)$ in the special case where X, Y are jointly Gaussian. In this case it turns out that $R_{X|Y}(d) = R^*(d)$, so that knowledge of the side information at the decoder (switch A closed) does not allow a reduction in the transmission rate R necessary to achieve a given distortion level d. In Sections 4 and 5 we give the proofs of our theorems.

2. Formal Statement of the Problem and Results

We begin with some words about notation. Let $\mathscr{U}$ be an arbitrary set. The elements of $\mathscr{U}^n$, the set of n-vectors with elements in $\mathscr{U}$, will be written as $\mathbf{u}^n = (u_1, \ldots, u_n)$, where the subscripted letters denote 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 is clear from the context, we will omit the superscript. For $k = 1, 2, \ldots$, define the set

$$\mathscr{I}_k = \{0, 1, \ldots, k-1\}.$$

Next, let $(\Omega, \mathcal{A}, P)$ be a probability space and let $\mathcal{B} \subseteq \mathcal{A}$ be a sub-σ-field. Then for $A \in \mathcal{A}$, let $P(A \mid \mathcal{B})$ denote the conditional probability of A given $\mathcal{B}$. Of course, $P(A \mid \mathcal{B})$ is a $\mathcal{B}$-measurable function on Ω. Next, let $U_j: \Omega \to \mathcal{U}_j$ $(1 \leqslant j \leqslant 3)$, where $\mathcal{U}_j$ is a arbitrary measurable space and U_j is assumed to be a measurable mapping. Let $\mathcal{A}_j \subseteq \mathcal{A}$ be the σ-algebra induced by U_j, $1 \leqslant j \leqslant 3$. We say that U_1, U_2, U_3 is a *chain* if U_1 and U_3 are conditionally independent given U_2. In other words, for all $S \in \mathcal{A}_3$,

$$P(S \mid \mathcal{A}_1, \mathcal{A}_2) = P(S \mid \mathcal{A}_2), \qquad \text{a.s.} \tag{2.1a}$$

An equivalent condition to (2.1a) is the condition that for all $S_1 \in \mathcal{A}_1$, $S_3 \in \mathcal{A}_3$,

$$P(S_1 \cap S_3 \mid \mathcal{A}_2) = P(S_1 \mid \mathcal{A}_2)\, P(S_3 \mid \mathcal{A}_2), \qquad \text{a.s.} \tag{2.1b}$$

For a complete discussion of conditional independence and chains, the reader referred to Loeve (1955) or Ash (1972).

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 (1968) or Pinsker (1964). A discussion of the definition of the conditional mutual information for arbitrary (nondiscrete) random variables is contained in the companion paper (Wyner, 1978). All logarithms in this paper are taken to the base 2.

We are now ready to define the problem. Let $(\Omega_0, \mathcal{A}_0 \ P_0)$ be an (underlying) probability space, and let X, Y be functions $X: \Omega_0 \to \mathcal{X}$, $Y: \Omega_0 \to \mathcal{Y}$, where $\mathcal{X}$, $\mathcal{Y}$ are arbitrary measurable spaces, and X, Y are measurable mappings. In addition we make the assumption throughout this paper that

$$I(X; Y) < \infty. \tag{2.2}$$

Let (X_k, Y_k), $k = 1, 2, \ldots$, be independent copies of (X, Y). Let $\hat{\mathcal{X}}$ be another measurable space, and let $D: \mathcal{X} \times \hat{\mathcal{X}} \to [0, \infty)$ be a measurable "distortion" function.

A *code* (n, M, Δ) is defined by two measurable mappings F_E, F_D, an "encoder" and a "decoder," respectively, where

$$F_E: \mathcal{X}^n \to \mathcal{J}_M, \tag{2.3a}$$

$$F_D: \mathcal{Y}^n \times \mathcal{J}_M \to \hat{\mathcal{X}}^n, \tag{2.3b}$$

and

$$E \frac{1}{n} \sum_{k=1}^{n} D(X_k, \hat{X}_k) = \Delta, \tag{2.3c}$$

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

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

$$M \leqslant 2^{n(R+\epsilon)}, \qquad \Delta \leqslant d + \epsilon. \tag{2.4}$$

We define $\mathscr{R}$ as the set of achievable (R, d) pairs, and define for $d \geqslant 0$,

$$R^*(d) = \min_{(R,d)\in\mathscr{R}} R. \tag{2.5}$$

If for some d, there is no $R < \infty$ such that $(R, d) \in \mathscr{R}$, we take $R^*(d) = \infty$. It follows from the definition $\mathscr{R}$ that if $(R_k, d) \in \mathscr{R}$, $k = 1, 2, \ldots$, then $(\lim_k R_k, d) \in \mathscr{R}$. Thus we conclude that the indicated minimum in (2.5) exists. Our main problem is the determination of $R^*(d)$.

We pause at this point to show that

$$R^*(0) = \lim_{d\to 0} R^*(d). \tag{2.6}$$

Since $R^*(d)$ is nonincreasing in d, we have $R^*(0) \geqslant \lim_{d\to 0} R^*(d)$. If this limit is infinite, then (2.6) follows. Assume that $\lim_{d\to 0} R^*(d) = R_0 < \infty$. If we show that $(R_0, 0) \in \mathscr{R}$, it will follow that $\lim_{d\to 0} R^*(d) = R_0 \geqslant R^*(0)$, completing the verification of (2.6). To show that $(R_0, 0) \in \mathscr{R}$, it will suffice to show that for any $\epsilon > 0$, there exists a code (n, M, Δ) with $M \leqslant 2^{n(R_0+\epsilon)}$, and $\Delta \leqslant \epsilon$. To do this let d be sufficiently small so that $| R^*(d) - R_0 | \leqslant \epsilon/2$ and $d \leqslant \epsilon/2$. Then, since $(R^*(d), d)$ is achievable, we can find a code (n, M, Δ) satisfying (2.4) with $R = R(d)$ and ϵ replaced by $\epsilon/2$, i.e.,

$$M \leqslant 2^{n(R(d)+\epsilon/2)} \leqslant 2^{n(R_0+\epsilon)}$$

and

$$\Delta \leqslant d + \epsilon/2 \leqslant \epsilon.$$

Thus we conclude that $(R_0, 0) \in \mathscr{R}$.

Summary of Results

Let $(\Omega_0, \mathscr{A}_0, P_0)$, X, Y, etc., be as above. Let $(\Omega_1, \mathscr{A}_1, P_1)$ be another probability space and let $(\Omega, \mathscr{A}, P) = (\Omega_0 \times \Omega_1, \mathscr{A}_0 \times \mathscr{A}_1, P_0 \times P_1)$ be the product space with product measure. Of course we can assume that X, Y are defined on Ω, e.g., let $X(\omega_0, \omega_1)$ be given by $X(\omega_0)$, for $(\omega_0, \omega_1) \in \Omega_0 \times \Omega_1 = \Omega$. Let $\mathscr{A}_X$, $\mathscr{A}_Y$ be the sub-σ-algebras (of $\mathscr{A}$) induced by X, Y, respectively. Let $\mathscr{Z}$ be another arbitrary measurable space, and let $Z: \Omega \to \mathscr{Z}$ be an arbitrary measurable function. Let $\mathscr{A}_Z$ be the σ-field induced by Z. Assume that the triple Y, X, Z is a chain, i.e., for all $B \in \mathscr{A}_Z$,

$$P(B \mid \mathscr{A}_Y \mathscr{A}_X) = P(B \mid \mathscr{A}_X), \qquad \text{a.s.} \tag{2.7}$$

Intuitively we can think of Z being realized as the output of a "test channel," P_t, the input of which is X. Ω_1 corresponds to the "noise" in the test channel. Now for $d > 0$, define $\mathcal{M}(d)$ as the set of functions $Z: \Omega \to \mathcal{Z}$, which satisfy (2.7), and have the property that there exists a measurable function $f: \mathcal{Y} \times \mathcal{Z} \to \hat{\mathcal{X}}$, such that

$$E\, D(X, \hat{X}) \leqslant d, \qquad \text{where} \qquad \hat{X} = f(Y, Z). \tag{2.8}$$

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.

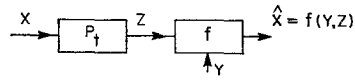

FIGURE 2

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

$$\bar{R}(d) = \inf_{Z \in \mathcal{M}(d)} [I(X; Z) - I(Y; Z)]. \tag{2.9a}$$

From the data processing theorem (inequality (3.13) of Wyner (1978)), $I(Y; Z) \leqslant I(X; Y) < \infty$. Thus (2.9a) is meaningful.[1] Further, since $\mathcal{M}(d)$ is nondecreasing in d, $\bar{R}(d)$ is nonincreasing for $d \in (0, \infty)$. Thus, we can meaningfully define

$$\bar{R}(0) = \lim_{d \to 0} \bar{R}(d). \tag{2.9b}$$

Our main results are Theorems 2.1 and 2.2, the proofs of which are given in Sections 4 and 5, respectively. We first state

THEOREM 2.1 (Converse). *For $d \geqslant 0$, $R^*(d) \geqslant \bar{R}(d)$.*

The "direct" theorem, i.e., the reverse of the inequality in Theorem 2.1, will also be shown to hold. We will however, have to make the following two technical assumptions about X, the space $\hat{\mathcal{X}}$, and the distortion function D.

(i) The first assumption, one that is sometimes made in source coding theory, is that for all $\hat{x} \in \hat{\mathcal{X}}$,

$$E\, D(X, \hat{x}) < \infty. \tag{2.10}$$

(ii) The second assumption concerns the "smoothness" of the function D. It is as follows. For all measurable functions $\hat{X}: \Omega \to \hat{\mathcal{X}}$, such that

[1] That is, the right member of (2.9a) is not "$\infty - \infty$."

$0 < E\,D(X, \hat{X}) < \infty$, and all $\epsilon > 0$, there exists a finite subset $\{\hat{I}_j\}_{j=1}^{N} \subseteq \hat{\mathscr{X}}$, and a "quantization" mapping $f_Q: \hat{\mathscr{X}} \to \{\hat{I}_j\}$, such that

$$E\,D(X, f_Q(\hat{X})) \leqslant (1 + \epsilon) E\,D(X, \hat{X}). \tag{2.11}$$

We show in Remark (E) below that when $\mathscr{X} = \hat{\mathscr{X}} =$ the reals, and $D(x, \hat{x}) = |x - \hat{x}|^r$, $r > 0$, and if $E\,|X|^r < \infty$, then conditions (i) and (ii) hold. We now state

THEOREM 2.2 (Direct Theorem). *If conditions* (i), (ii) *above hold, then* $R^*(d) \leqslant \bar{R}(d)$, $0 \leqslant d < \infty$.

Remarks. (A) From (2.6) and (2.9b), it will suffice to prove Theorems 2.1, 2.2 for $d > 0$.

(B) In Appendix B we show that $\bar{R}(d)$ is a convex function of d.

(C) Let X, Y, Z be a chain. From Lemma 3.1 of Wyner, $I(Y; Z \mid X) = 0$. Thus, using Lemma 3.2 of Wyner, we have

$$\begin{aligned} I(X; Z) - I(Y; Z) &= I(XY; Z) - I(Y; Z \mid X) - I(Y; Z) \\ &= I(XY; Z) - I(Y; Z) = I(X; Z \mid Y). \end{aligned} \tag{2.12}$$

Thus it follows that

$$\bar{R}(d) = \inf_{Z \in \mathscr{M}(d)} I(X; Z \mid Y). \tag{2.13}$$

(D) Let $Z \in \mathscr{M}(d)$, and let $\hat{X} = f(Y, Z)$. Then from the data-processing theorem (use Lemma 3.4 of Wyner, with $U_1 = X$, $U_2 = Z$, $U_3 = \hat{X}$, $U_4 = Y$),

$$I(X; Z \mid Y) \geqslant I(X; \hat{X} \mid Y) \geqslant R_{X|Y}(d), \tag{2.14}$$

where the last inequality follows from the discussion in the Appendix where $R_{X|Y}(d)$ is defined (since $\hat{X} \in \mathscr{M}_0(d)$). Minimizing (2.14) with respect to $Z \in \mathscr{M}(d)$, we have

$$R^*(d) \geqslant R_{X|Y}(d),$$

which is, of course, obvious from the "physical" situation. Inequality (2.14), however, also tells us that $R^*(d) = R_{X|Y}(d)$ if the $\hat{X} \in \mathscr{M}_0(d)$ which achieves $I(X; \hat{X} \mid Y) = R_{X|Y}(d)$ can be generated as in Fig. 2 with $I(X; Z \mid Y) = I(X; \hat{X} \mid Y)$. This occurs (see (3.11) of Wyner) if and only if $I(X; Z \mid \hat{X}Y) = 0$. In Section 3 we give an example of a source for which this rather severe condition holds and $R^*(d) = R_{X|Y}(d)$.

(E) Condition (i) is a rather common assumption for source coding theorems. Furthermore, neither condition (i) nor (ii) is especially restrictive.

We will now show that when $\mathscr{X} = \hat{\mathscr{X}} =$ the reals, and $D(x, \hat{x}) = |x - \hat{x}|^r$, $r > 0$, and if $E|X|^r < \infty$, then conditions (i) and (ii) hold. To do this we exploit the Minkowski inequality, which states that for arbitrary random variables U, V,

$$E^{1/r}|U + V|^r \leqslant E^{1/r}|U|^r + E^{1/r}|V|^r. \tag{2.15}$$

Thus, if $E|X|^r < \infty$, then

$$E|X - \hat{x}|^r \leqslant (E^{1/r}|X|^r + |\hat{x}|)^r < \infty,$$

which is condition (i). We now turn to condition (ii). Suppose that $0 < E|X - \hat{X}|^r < \infty$, and $E|X|^r < \infty$. It follows that

$$E|\hat{X}|^r \leqslant (E^{1/r}|X|^r + E^{1/r}|X - \hat{X}|^r)^r < \infty.$$

Now, for $N = 1, 2, \ldots$, define

$$\begin{aligned} f_Q(\hat{x}) &= (N+1)\Delta, && \hat{x} \geqslant (N+1)\Delta, \\ &= -N\Delta, && \hat{x} < -N\Delta, \\ &= n\Delta, && n\Delta \leqslant \hat{x} < (n+1)\Delta, \quad 0 \leqslant n \leqslant N, \\ &= (n+1)\Delta, && n\Delta \leqslant \hat{x} < (n+1)\Delta, \quad -N \leqslant n \leqslant -1, \end{aligned} \tag{2.16}$$

where $\Delta = N^{-1/2}$. Let $\hat{X}_N = f_Q(\hat{X})$. Note that, as $N \to \infty$, $\hat{X}_N \to \hat{X}$. Also, since $|\hat{X}_N - \hat{X}| \leqslant 2|\hat{X}|$, we have from the dominated convergence theorem,

$$E|\hat{X} - \hat{X}_N|^r \to 0, \qquad \text{as} \quad N \to \infty.$$

Now let $\epsilon > 0$ be given. Choose N sufficiently large so that

$$\begin{aligned} E^{1/r}|\hat{X} - \hat{X}_N|^r &= E^{1/r}|\hat{X} - f_Q(\hat{X})|^r \\ &\leqslant \delta(\epsilon)\, E^{1/r}|X - \hat{X}|^r, \end{aligned}$$

where $\delta(\epsilon) = (1 + \epsilon)^{1/r} - 1$. Then

$$\begin{aligned} E^{1/r}|X - f_Q(\hat{X})|^r &\leqslant E^{1/r}|X - \hat{X}|^r + E^{1/r}|\hat{X} - f_Q(\hat{X})|^r \\ &\leqslant (1 + \delta(\epsilon))\, E^{1/r}|X - \hat{X}|^r \\ &= [(1 + \epsilon)\, E|X - \hat{X}|^r]^{1/r}. \end{aligned}$$

Raising both sides to the rth power yields

$$E|X - f_Q(\hat{X})|^r \leqslant (1 + \epsilon)E|X - \hat{X}|^r,$$

which is (2.11). Thus condition (ii) holds.

3. An Example: X, Y Jointly Gaussian

In this section we consider the special case where $\mathscr{X} = \mathscr{Y} = \hat{\mathscr{X}} =$ the reals, $D(x, \hat{x}) = (x - \hat{x})^2$, and X, Y are jointly Gaussian with zero mean. With no loss of generality we can write

$$Y = \alpha(X + U), \tag{3.1}$$

where $\alpha > 0$, and X, U are independent Gaussian variates with $EX = EU = 0$, and $EX^2 = \sigma_X^2$, $EU^2 = \sigma_U^2$. Thus the side information Y, is a noisy version of X. A straightforward application of Bayes' rule yields that given $Y = y$, X is normally distributed with conditional mean

$$E(X \mid Y = y) = (c/\alpha)\, y, \tag{3.2a}$$

and conditional variance

$$\mathrm{Var}(X \mid Y = y) = c\, \sigma_U^2, \tag{3.2b}$$

where

$$c = \sigma_X^2/(\sigma_X^2 + \sigma_U^2). \tag{3.2c}$$

By analogy with the conventional situation where there is no side information, we are inclined to guess that

$$\begin{aligned} R_{X|Y}(d) &= \tfrac{1}{2} \log c\, \sigma_U^2/d, \qquad 0 < d < c\, \sigma_U^2 \\ &= 0, \qquad\qquad\quad d \geqslant c\, \sigma_U^2. \end{aligned} \tag{3.3}$$

The random variable $\hat{X} \in \mathscr{M}(d)$ which would achieve $I(X; \hat{X} \mid Y) = R_{X|Y}(d)$ can be realized by the system given in Fig. 3, with ψ a zero-mean Gaussian variate with variance $c\, d\sigma_U^2/(c\, \sigma_U^2 - d)$. Fig. 3 is the same as Fig. 9.7.3 in Gallager (1968) with the conditional mean, cY/α, subtracted out at the input and then added in at the output. In fact we shall give a proof that $R_{X|Y}(d)$ is given by (3.3) at the conclusion of this section.

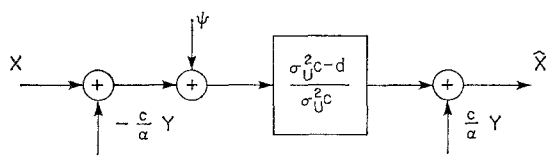

Figure 3

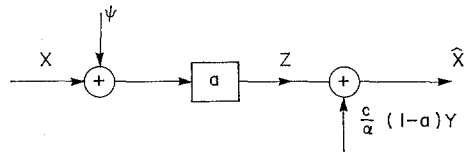

FIGURE 4

We now observe that the system of Fig. 3 can be redrawn as in Fig. 4 (where $a = (\sigma_U{}^2 c - d)/\sigma_U{}^2 c$). Note that Fig. 4 is of the same form as Fig. 2, with $\hat{X} = f(Y, Z) = Z + (c/\alpha)(1 - a)Y$. Observe that if we are given the values of $\hat{X}$ and Y, then we can calculate Z. Thus given the values of $(\hat{X}, Y)$, the random variable Z is a constant, which implies that

$$I(X; Z \mid \hat{X}Y) = 0. \tag{3.5}$$

This is the condition given in Section 2, Remark (D), for the equality of $R^*(d)$ and $R_{X|Y}(d)$. Thus we conclude that for X, Y as in (3.1),

$$\begin{aligned} R^*(d) = R_{X|Y}(d) &= \frac{1}{2}\log\frac{\sigma_X{}^2\sigma_U{}^2}{(\sigma_X{}^2 + \sigma_U{}^2)d}, & 0 < d < \frac{\sigma_X{}^2\sigma_U{}^2}{\sigma_X{}^2 + \sigma_Y{}^2} \\ &= 0, & d \geqslant \frac{\sigma_X{}^2\sigma_U{}^2}{\sigma_X{}^2 + \sigma_U{}^2}. \end{aligned} \tag{3.6}$$

It remains to verify (3.3). A straightforward calculation yields that with $\hat{X}$ as in Fig. 3, $E(X - \hat{X})^2 = d$ and $I(X; \hat{X} \mid Y) =$ the right member of (3.3). Thus we must show that for arbitrary $\hat{X} \in \mathscr{M}_0(d)$,

$$I(X; \hat{X} \mid Y) \geqslant \frac{1}{2}\log\frac{c\,\sigma_U{}^2}{d}, \qquad 0 < d < c\,\sigma_U{}^2. \tag{3.7}$$

Now inequality (3.7) follows from a standard bounding technique in which the information is written as the difference of the differential entropy and the conditional differential entropy. There is, however, a question as to the existence of the probability density functions which define these differential entropies. We therefore digress to establish a simple lemma about conditional densities.

Let U, V be a pair of real-valued random variables which define a joint probability measure on the plane. Let P_U be the marginal probability measure corresponding to the random variable U. Suppose that P_U is absolutely continuous with respect to Lebesgue measure, denoted μ_L. Also assume that V takes its values only on the set $\{v_j\}_{j=1}^N$ $(1 \leqslant N < \infty)$, and that $\Pr\{V = v_j\} > 0$, $1 \leqslant j \leqslant N$. We now state

LEMMA 3.1. *Let U, V be random variables as above. Then for each $j = 1$,*

2,..., N, there exists a Borel measurable (density) function $q_j(\cdot)$ such that for all Borel sets B,

$$P_j(B) \triangleq \Pr\{U \in B \mid V = v_j\} = \frac{\Pr\{U \in B,\ V = v_j\}}{\Pr\{V = v_j\}} = \int_B q_j(x)\, dx.$$

Proof. Observe first that $P_j(\cdot)$ is a probability measure on the Borel sets. Our result will follow from the Radon–Nikodym theorem if we show that P_j is absolutely continuous with respect to μ_L. Let B be such that $P_j(B) > 0$. Then

$$P_U(B) = \sum_{j'=1}^{N} \Pr\{V = j'\}\, P_{j'}(B) \geqslant \Pr\{V = v_j\}\, P_j(B) > 0,$$

so that the hypothesis implies $\mu_L(B) > 0$. ∎

We now return to (3.7). Let $\hat{X} \in \mathscr{M}_0(d)$, $d > 0$, and let $\epsilon > 0$ be arbitrary. Since $D(x, \hat{x}) = (x - \hat{x})^2$ satisfies condition (ii) in Section 2, there exists a function

$$f_Q\colon \hat{\mathscr{X}} \to \{\hat{x}_j\}_1^N \subseteq \hat{\mathscr{X}}, \tag{3.8a}$$

such that

$$E(X - \hat{X}_Q)^2 \leqslant (1 + \epsilon)\, E(X - \hat{X})^2 \leqslant d(1 + \epsilon), \tag{3.8b}$$

where $\hat{X}_Q = f_Q(\hat{X})$. It follows from the data-processing theorem (use Lemma 3.4 in Wyner, with $U_1 = X$, $U_2 = \hat{X}$, $U_3 = \hat{X}_Q$, $U_4 = Y$) that

$$I(X; \hat{X} \mid Y) \geqslant I(X; \hat{X}_Q \mid Y). \tag{3.9}$$

Now we can apply exactly the same reasoning which we used to obtain (A.6) to assert the existence of a random variable $\tilde{Y}$ which takes but a finite number of values for which

$$|\, I(X; \hat{X}_Q \mid Y) - I(X; \hat{X}_Q \mid \tilde{Y})| \leqslant \epsilon, \tag{3.10a}$$

and

$$|\, I(X; Y) - I(X; \tilde{Y})| \leqslant \epsilon. \tag{3.10b}$$

Next, we define the differential entropy. Let U, V be random variables with joint probability density function $p_{12}(u, v)$. Let

$$p_1(u) = \int_{-\infty}^{\infty} p_{12}(u, v)\, dv,$$

and

$$p_{2|1}(v \mid u) = \frac{p_{12}(u, v)}{p_1(u)}, \qquad p_1(u) > 0.$$

Then the *differential entropy* of U and the *differential entropy* of V given U are defined by

$$H_d(U) = -E \log p_1(U),$$
$$H_d(V \mid U) = -E \log p_{1|2}(V \mid U),$$

respectively. Note that $H_d(X)$, $H_d(X \mid Y)$, and $H_d(X \mid \tilde{Y})$ are meaningful (by Lemma 3.1 the conditional density for X given $\tilde{Y}$ exists). Further we can write

$$I(X; Y) = H_d(X) - H_d(X \mid Y) \tag{3.11a}$$

and

$$I(X; \tilde{Y}) = H_d(X) - H_d(X \mid \tilde{Y}), \tag{3.11b}$$

so that (3.10b) and (3.11) yield

$$H_d(X \mid \tilde{Y}) \geqslant H_d(X \mid Y) - \epsilon = \tfrac{1}{2}[\log 2\pi ec\sigma_U{}^2] - \epsilon, \tag{3.12}$$

where c is given by (3.2c).

We now establish (3.7) by writing

$$\begin{aligned} I(X; \hat{X} \mid Y) &\overset{(a)}{\geqslant} I(X; \hat{X}_Q \mid \tilde{Y}) - \epsilon \\ &\overset{(b)}{=} H_d(X \mid \tilde{Y}) - H_d(X \mid \hat{X}_Q\tilde{Y}) - \epsilon \\ &\overset{(c)}{\geqslant} \tfrac{1}{2} \log 2\pi ec\, \sigma_U{}^2 - H_d(X \mid \hat{X}_Q\tilde{Y}) - 2\epsilon. \end{aligned} \tag{3.13}$$

Step (a) follows from (3.9) and (3.10a). Step (b) is meaningful in the light of Lemma 3.1, and step (c) follows from (3.12). Now, denote the range of $\tilde{Y}$ by $\{i\}$, and let

$$d_{ij} = E[(X - \hat{X}_Q)^2 \mid \tilde{Y} = i, \hat{X}_Q = \hat{y}_j]. \tag{3.14a}$$

Also let

$$a_{ij} = \Pr\{\tilde{Y} = i, \hat{X}_Q = \hat{x}_j\}. \tag{3.14b}$$

Note that

$$d_{ij} \geqslant \operatorname{Var}(X \mid \tilde{Y} = i, \hat{X}_Q = \hat{x}_j),$$

so that a standard inequality (Ash, 1965, Theorem 8.3.8) yields

$$H(X \mid \tilde{Y} = i, \hat{X}_Q = \hat{x}_j) \leqslant \tfrac{1}{2} \log 2\pi e\, d_{ij}\,. \tag{3.15a}$$

Also note that, from (3.8b),

$$\sum a_{ij}d_{ij} = E(X - \hat{X}_Q)^2 \leqslant d(1 + \epsilon). \tag{3.15b}$$

Inequalities (3.15) and the concavity of the logarithm imply that

$$\begin{aligned} H(X \mid \hat{Y}, \hat{X}_Q) &= \sum_{i,j} a_{ij} H(X \mid \hat{Y} = i, \hat{X}_Q = \hat{x}_j) \\ &\leqslant \sum_{i,j} a_{ij}(\tfrac{1}{2} \log 2\pi e\, d_{ij}) \\ &\leqslant \tfrac{1}{2} \log \Big(2\pi e \sum_{i,j} a_{ij} d_{ij}\Big) \leqslant \tfrac{1}{2} \log(2\pi e\, d(1 + \epsilon)). \end{aligned} \tag{3.16}$$

Substituting (3.16) into (3.13) we obtain

$$I(X; \hat{X} \mid Y) \geqslant \tfrac{1}{2} \log 2\pi e c \sigma_U^2 - \tfrac{1}{2} \log 2\pi e d(1 + \epsilon)) - 2\epsilon,$$

and letting $\epsilon \to 0$, we have (3.7).

4. The Converse Theorem

In this section we will give a proof of Theorem 2.1 which asserts that $R^*(d) \geqslant \bar{R}(d)$, $d \geqslant 0$. As pointed out in Remark (A) following Theorem 2.2, we need only establish this result for $d > 0$. The ideas in the proof are essentially identical to those used in the proof for the discrete case in Wyner and Ziv. The merchanics of the proof are, however, slightly *simpler* here than in that reference.

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

$$(1/n) \log M \geqslant \bar{R}(\Delta). \tag{4.1}$$

If $d > 0$ and $(R, d) \in \mathscr{R}$, then for arbitrary $\epsilon > 0$, with n sufficiently large, there exists a code (n, M, Δ) with $M \leqslant 2^{n(R+\epsilon)}$, and $\Delta \leqslant d + \epsilon$. Inequality (4.1) and the monotonicity of $\bar{R}(\cdot)$ imply that

$$R + \epsilon \geqslant \bar{R}(\Delta) \geqslant \bar{R}(d + \epsilon). \tag{4.2}$$

Letting $\epsilon \to 0$, and invoking the continuity of $\bar{R}(d)$ (which follows from the convexity of $\bar{R}(d)$, which is established in Appendix C), we have $R \geqslant \bar{R}(d)$ for $(R, d) \in \mathscr{R}$, $d > 0$. This implies Theorem 2.1. It remains to establish (4.1).

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

$$\Delta_k = E\, D(X_k, \hat{X}_k), \tag{4.3}$$

so that

$$\Delta = \frac{1}{n} \sum_{k=1}^{n} \Delta_k . \tag{4.4}$$

Now

$$\begin{aligned} \log M &\overset{(a)}{\geq} I(\mathbf{X}^n; W \mid \mathbf{Y}^n) \\ &\overset{(b)}{=} I(\mathbf{X}^n; \mathbf{Y}^n W) - I(\mathbf{X}^n; \mathbf{Y}^n) \\ &\overset{(c)}{=} \sum_{k=1}^{n} [I(X_k ; \mathbf{Y}W \mid \mathbf{X}^{k-1}) - I(X_k ; Y_k)]. \end{aligned} \tag{4.5}$$

Step (a) follows from $W \in \mathscr{J}_M$. Step (b) follows from Lemma 3.2 of Wyner, and is meaningful since (by (2.2)) $I(\mathbf{X}^n; \mathbf{Y}^n) < \infty$. Step (c) follows from the independence of $\{(X_k, Y_k)\}_k$ and from Lemma 3.3 of Wyner (repeated n times).

Further, since X_k and $\mathbf{X}^{k-1}$ are independent

$$\begin{aligned} I(X_k ; \mathbf{Y}W \mid \mathbf{X}^{k-1}) &= I(X_k ; \mathbf{Y}W\mathbf{X}^{k-1}) - I(X_k ; \mathbf{X}^{k-1}) \\ &= I(X_k ; \mathbf{Y}W\mathbf{X}^{k-1}) = I(X_k ; Y_k Z_k), \end{aligned} \tag{4.6a}$$

where

$$Z_k = (\mathbf{X}^{k-1}, Y_1, \ldots, Y_{k-1}, Y_{k+1}, \ldots, Y_n, W). \tag{4.6b}$$

Substituting (4.6) into (4.5) we obtain, using Lemma A.2,

$$\begin{aligned} \log M &\geq \sum_{k=1}^{n} [I(X_k ; Y_k Z_k) - I(X_k ; Y_k)] \\ &= \sum_{k=1}^{n} [I(X_k ; Z_k \mid Y_k)]. \end{aligned} \tag{4.7}$$

We now point out two facts about Z_k: (a) $\hat{X}_k$ is the kth coordinate of $F_D(\mathbf{Y}^n, W)$, so that we can write $\hat{X}_k$ as a deterministic function of Y_k and Z_k, say $\hat{X}_k = f(Y_k, Z_k)$. Of course (4.3) 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 $Z_k: \Omega \to \mathscr{X}^{k-1} \times \mathscr{Y}^{n-1} \times \mathscr{J}_M$ belongs to $\mathscr{M}(\Delta_k)$, so that from the definition,

$$I(X_k ; Z_k \mid Y_k) = I(X_k ; Z_k) - I(Y_k ; Z_k) \geq \bar{R}(\Delta_k). \tag{4.8}$$

Substituting (4.8) into (4.7) and invoking the convexity of $\bar{R}(\cdot)$ (Appendix B), we have

$$\log M \geq \sum_{k=1}^{n} \bar{R}(\Delta_k) \geq n\bar{R}\left(\frac{1}{n}\sum_{k=1}^{n} \Delta_k\right) = n\bar{R}(\Delta),$$

where the last step follows from (4.4). This establishes (4.1), and completes the proof of Theorem 2.1.

5. The Direct Half

In this section we will outline the proof of Theorem 2.2 which asserts that, subject to conditions (i) and (ii) of Section 2, $R^*(d) \leqslant \bar{R}(d)$. The proof leans very heavily on the proof given in Wyner and Ziv for $\mathscr{X}$, $\mathscr{Y}$, $\hat{\mathscr{X}}$ finite. In fact, the discussion in this section will be devoted to showing how the proof for the finite case be modified to hold in the general case.

We begin by establishing a simple lemma.

Lemma 5.1. *Let $(\Omega_0, \mathscr{A}_0, P_0)$ be a probability space and $U: \Omega_0 \to [0, \infty)$ a random variable such that $EU < \infty$. Then for all $\delta > 0$, there exists a $\nu = \nu(\delta, U)$ such that*

$$S \in \mathscr{A}_0, \quad P_0(S) < \nu \Rightarrow E I_S U = \int_S U \, dP_0 < \delta, \tag{5.1}$$

where I_S is the indicator function of S.

Proof. Let $A = \{U \leqslant a\}$, and write

$$\begin{aligned} E I_S U &= E I_S I_A U + E I_S I_{A^c} U \\ &\leqslant a E I_S + E I_{A^c} U = a P_0(S) + E I_{A^c} U. \end{aligned}$$

Since $E U < \infty$, $\lim_{a \to \infty} E I_{A^c} U = 0$. Thus, with $\delta > 0$ specified, let a be sufficiently large so that $E I_{A^c} U \leqslant \delta/2$, and then set $\nu = \delta/2a$. This choice of ν satisfies (5.1). ∎

We now give

Lemma 5.2. *If $\mathscr{Y}$, $\hat{\mathscr{X}}$ are finite sets, and conditions* (i) *and* (ii) *of Section 2 are satisfied ($\mathscr{X}$ is* not *assumed finite), and if $\mathscr{M}(d)$ is defined as above except that $\mathscr{X}$ is required to be finite, then $Z \in \mathscr{M}(d)$ implies that $(R, d) = (I(X; Z) - I(Y; Z), d) \in \mathscr{R}$.*

Proof. A careful examination of the proof of the direct half given in Section 4 of Wyner and Ziv will indicate that the only places which the finiteness of $\mathscr{X}$ is exploited are in (63) and (79) of that reference. In a number of other places in that proof, it is of course necessary to give the appropriate measure-theoretic interpretations to various expressions.

It is easy to see that we can rewrite (63) in Wyner and Ziv as

$$\begin{aligned} \Delta &= \frac{1}{n_1} \sum_{j=1}^{n_1} E \left\{ \frac{1}{n_0} \sum_{k=(j-1)n_0+1}^{jn_0} D(X_k, X^*_k) \right\} \\ &= \frac{1}{n_1} \sum_{j=1}^{n_1} E \varphi_j \leqslant \Delta_0 + \frac{1}{n_1} \sum_{j=1}^{n_1} E I_{\mathscr{E}_1} \varphi_j, \end{aligned} \tag{5.2}$$

where ϕ_j is (as in Wyner and Ziv) the term in brackets in (5.2), and the event $\mathscr{E}_j = \{W_j \neq \hat{W}_j\}$ and $I_{\mathscr{E}_j}$ is its indicator. Note that, for $1 \leqslant j \leqslant n_1$,

$$\begin{aligned} \varphi_j &= \frac{1}{n_0} \sum_{k=(j-1)n_0+1}^{jn_0} D(X_k, X^*_k) \\ &\leqslant \max_{\substack{\mathbf{y}\in\mathscr{Y}^{n_0} \\ 1\leqslant i\leqslant M_0}} \frac{1}{n_0} \sum_k D(X_k, x^*_k(\mathbf{y}, i)) \triangleq U(\mathbf{X}_j), \end{aligned} \tag{5.3}$$

where $x^*_k(\mathbf{y}, i)$ is the kth coordinate of $F_D^{(0)}(\mathbf{y}, i)$, and $\mathbf{X}_j = (X_{(j-1)n_0+1}, \ldots, X_{jn_0})$. Note that $E\,U(\mathbf{X}_j) < \infty$. Substituting (5.3) into (5.2) yields

$$\Delta \leqslant \Delta_0 + \frac{1}{n_1} \sum_{j=1}^{n_1} E\, I_{\mathscr{E}_j} U(\mathbf{X}_j). \tag{5.4}$$

Now we would like to show that the conclusion of (63) of Wyner and Ziv, i.e. $\Delta \leqslant \Delta_0 + \delta$ can be made to hold. Applying Lemma 5.1 to $U(X_j)$, we have that $E\, I_{\mathscr{E}_j} U(\mathbf{X}_j) \leqslant \delta$ if $\Pr\{\mathscr{E}_j\} \leqslant \nu(\delta, U)$. This will be satisfied if condition (62) of Wyner and Ziv is replaced by

$$\Pr\{\mathscr{E}_j\} = \Pr\{W_j \neq \hat{W}_j\} \leqslant \nu(\delta, U). \tag{5.5}$$

A check of the Slepian and Wolf (1973) result, indicates that with n_1 sufficiently large, (5.5) can be made to hold.

It remains to find a replacement for (79) in Wyner and Ziv. It is easy to see that it will suffice to show that, as $n_0 \to \infty$,

$$Q_{n_0} \triangleq E\, \psi_{n_0} D_{n_0}(\mathbf{X}^{n_0}, f_{n_0}(\mathbf{Y}^{n_0}, F(\mathbf{X}^{n_0})) < \delta. \tag{5.6}$$

To do this, put $\hat{\mathbf{Z}}^{n_0} = F(\mathbf{X}^{n_0})$, and write

$$\begin{aligned} D_{n_0}(\mathbf{X}^{n_0}, f_{n_0}(\mathbf{Y}^{n_0}, \hat{\mathbf{Z}}^{n_0})) &= \frac{1}{n_0} \sum_{k=1}^{n_0} D(X_k, f(Y_k, \hat{Z}_k)) \\ &\leqslant \frac{1}{n_0} \sum_{k=1}^{n_0} \max_{\substack{y\in\mathscr{Y} \\ z\in\mathscr{Z}}} D(X_k, f(y, z)) \triangleq \frac{1}{n_0} \sum_{k=1}^{n_0} U(X_k), \end{aligned}$$

where $E\,U(X_k) < \infty$. Thus

$$Q_{n_0} \leqslant \frac{1}{n_0} \sum_{k=1}^{n_0} E\, \psi_{n_0} U(X_k). \tag{5.7}$$

Since ψ_{n_0} is the indicator of a sequence of events whose probabilities vanish, (5.6) follows from Lemma 5.1. This completes the proof of Lemma 5.2. ∎

Theorem 2.2 will follow from Lemma 5.2 and the following lemma, which asserts the existence of discrete approximations to arbitrary Y, Z.

LEMMA 5.3. *Assume that conditions* (i) *and* (ii) *of Section* 2 *are satisfied. Let* $Z \in \mathscr{M}(d)$. *Then for all* $\epsilon > 0$, *there exist finite partitions* $\{A_i\}_{i=1}^{N_1}$ *and* $\{B_j\}_{j=1}^{N_2}$ *of* $\mathscr{U}$ *and* $\mathscr{Z}$, *respectively, and a function*

$$f_1: \mathscr{U} \times \mathscr{Z} \to \hat{\mathscr{X}}$$

which satisfies

(a) $$E\, D(X, f_1(Y, Z)) \leqslant d + \epsilon \tag{5.8a}$$

(b) f_1 *is constant on the rectangles* $A_i \times B_j$,

$$1 \leqslant i \leqslant N_1, \qquad 1 \leqslant j \leqslant N_2. \tag{5.8b}$$

(c) $$I(X; \tilde{Z}) - I(\tilde{Y}; \tilde{Z}) \leqslant I(X; Z) - I(Y; Z) + \epsilon, \tag{5.8c}$$

where

$$\tilde{Y} = i \text{ for } Y \in A_i, \qquad \tilde{Z} = j \text{ for } Z \in B_j, \tag{5.8d}$$

$1 \leqslant i \leqslant N_1$, $1 \leqslant j \leqslant N_2$.

We now show to obtain Theorem 2.2. From (5.8b) and (5.8c) we can think of f_1 as a function of $\tilde{Y}$, $\tilde{Z}$, and from (5.8a)

$$E\, D(X, f_1(\tilde{Y}, \tilde{Z})) \leqslant d + \epsilon.$$

Further, $\tilde{Y}$, X, $\tilde{Z}$ is a chain, so that $Z \in \mathscr{M}(d)$ for the source $(X, \tilde{Y})$. Finally since $(R, d) \in \mathscr{R}$ for the source $(X, \tilde{Y})$ implies that $(R, d) \in \mathscr{R}$ for source (X, Y), we conclude from Lemma 5.2 $(R, d) = (I(X; Z) - I(Y; Z) + \epsilon, d + \epsilon)$ is achievable for source (X, Y). Since the region $\mathscr{R}$ is closed, letting $\epsilon \to 0$ yields Theorem 2.2. It remains to give the

Proof of Lemma 5.3. Since $Z \in \mathscr{M}(d)$, there exists a function $f: \mathscr{U} \times \mathscr{Z} \to \hat{\mathscr{X}}$ such that with $\hat{X} = f(Y, Z)$,

$$E\, D(X, \hat{X}) \leqslant d.$$

Condition (ii) of Section 2 asserts that for all $\epsilon > 0$, there exists a finite set $\{\hat{x}_j\}_1^N \subseteq \hat{\mathscr{X}}$ and a mapping $f_0 \triangleq f_Q \circ f: \mathscr{Y} \times \mathscr{Z} \to \{\hat{I}_j\}$ such that

$$E\, D(X, f_0(Y, Z)) \leqslant d + \epsilon/2. \tag{5.9}$$

Let $C_j = \{(y, z): f_0(y, z) = \hat{x}_j\}$, $1 \leqslant j \leqslant N$. At the conclusion of this proof

we will show that for arbitrary $\epsilon_1 > 0$, there exists a collection of disjoint sets $\{S_j\}_1^N$, where each $S_j \subseteq \mathscr{Y} \times \mathscr{Z}$ is a finite union of rectangles, for which

$$\Pr\{C_j \triangle S_j\} \leqslant \epsilon_1, \qquad 1 \leqslant j \leqslant N. \tag{5.10}$$

Defining $f_1: \mathscr{Y} \times \mathscr{Z} \to \hat{\mathscr{X}}$ by

$$\begin{aligned} f_1(y, z) &= \hat{x}_j, \qquad (y, z) \in S_j, \quad 1 \leqslant j \leqslant N, \\ &= \hat{x}_1, \qquad (y, z) \notin \bigcup_1^N S_j, \end{aligned} \tag{5.11}$$

we write

$$E\, D(X, f_1(Y, Z)) = E\, I_{\mathscr{E}} D + E\, I_{\mathscr{E}^c} D, \tag{5.12}$$

where $\mathscr{E} = \bigcup_{j=1}^N (C_j \cap S_j)$. When $\mathscr{E}$ occurs, $f_1(Y, Z) = f_0(U, Z)$, so that

$$E\, I_{\mathscr{E}} \leqslant E\, D(X, f_0(Y, Z)) \leqslant d + \epsilon/2. \tag{5.13}$$

Also

$$E\, I_{\mathscr{E}^c} D \leqslant E\, I_{\mathscr{E}^c} \max_{1 \leqslant j \leqslant N} D(X, \hat{x}_j) \triangleq E\, I_{\mathscr{E}^c}.$$

Since $E\, U < \infty$, Lemma 5.1 and (5.10) imply that $E\, I$–U can be made $\leqslant \epsilon/2$, if ϵ_1 is sufficiently small. With such a choice of ϵ_1, we have using (5.12) and (5.13),

$$E\, D(X, f_1(Y, Z)) \leqslant d + \epsilon/2 + \epsilon/2 = d + \epsilon, \tag{5.14}$$

which is (5.8a). Furthermore since f_1 is constant on each S_j (a finite union of rectangles), we can find finite partitions say $\{A_i^{(1)}\}$ and $\{B_k^{(1)}\}$ which satisfy (5.8b). Finally, $(I(X; Z), I(Y; Z))$ can be approximated arbitrarily closely by $(I(X; \tilde{Z}), I(X; \tilde{Y}))$ for $\tilde{Y}$, $\tilde{Z}$ defined as in (5.8d) with a suitable pair of partitions say $\{A_i^{(2)}\}\{B_j^{(2)}\}$. Lemma 5.3 then follows on letting partition $\{A_i\}$ be the common refinement of $\{A_i^{(1)}\}$ and $\{A_i^{(2)}\}$, and letting $\{B_j\}$ be the common refinement of $\{B_j^{(1)}\}$ and $\{B_j^{(2)}\}$. To complete the proof of Lemma 5.3 we must verify (5.10).

Since the field of finite unions of rectangles generates the product σ-field which corresponds to $\mathscr{Y} \times \mathscr{Z}$, (5.10) will follow from

LEMMA 5.4. *Let $(\Omega_0, \mathscr{A}_0, P_0)$ be a probability space. Let $\{A_j\}_1^N$, $A_j \in \mathscr{A}_0$, be a partition of Ω_0. Let $\mathscr{B}$ be a field of subsets of Ω_0 which generates $\mathscr{A}_0$. Then for all $\epsilon_1 > 0$, there exists a collection $\{B_j\}_1^N$ of disjoint sets, where $B_j \in \mathscr{B}$, such that*

$$P_0(A_j \triangle B_j) \leqslant \epsilon_1, \qquad 1 \leqslant j \leqslant N. \tag{5.15}$$

Proof. A standard result from measure theory (Ash, 1972, p. 20) asserts the existence of a collection $\{B'_j\}_1^N$, $B'_j \in \mathscr{B}$, such that

$$P_0(B'_j \triangle A_j) \leqslant \epsilon_1/N, \qquad 1 \leqslant j \leqslant N. \tag{5.16}$$

The $\{B'_j\}$ are not necessarily disjoint, however. Let

$$\begin{aligned} B_1 &= B'_1 \\ B_2 &= B'_2 B'^c_1 \\ B_3 &= B'_3 B'^c_1 B'^c_2 \\ &\vdots \\ B_N &= B'_N B'^c_1 B'^c_2 \cdots B'^c_{N-1}. \end{aligned} \tag{5.17}$$

The $\{B_j\}$ are disjoint, and

$$\begin{aligned} P_0(A_j \Delta B_j) &= P_0(A_j \Delta B'_j B'^c_1 B'^c_2 \cdots B'^c_{j-1}) \\ &\leqslant P_0(A_j^c B'_j) + P_0(A_j B'^c_j) + \sum_{i=1}^{j-1} P_0(A_j B'_i) \\ &= P_0(A_j \Delta B'_j) + \sum_{i=1}^{j-1} P_0(A_j B'_i). \end{aligned} \tag{5.18}$$

Now since the $\{A_j\}$ are disjoint, $A_j \subseteq A_i^c$ $(i < j)$, so that $P_0(A_j B'_i) \leqslant P_0(A_i^c B'_i) \leqslant P_0(A_i \triangle B'_i) \leqslant \epsilon_1/N$. Substitution of this and (5.16) into (5.18) yields

$$P_0(A_j \Delta B_j) \leqslant \frac{\epsilon_1}{N} + (N-1)\frac{\epsilon_1}{N} = \epsilon_1. \quad \blacksquare$$

APPENDIX A: Discussion of $R_{X|Y}(d)$

In Section 1 we defined $R_{X|Y}(d)$ as the minimum rate R required in the system of Fig. 1 with switch A closed for reproduction at an average distortion $E\,D(X, \hat{X})$ of about d. In this appendix we will show how to generalize Berger's characterization of $R_{X|Y}(d)$ for discrete X Y to the case where X Y are arbitrary random objects.

We begin with some definitions. For the case where switch A is closed we define the *source* (X, Y) exactly as in Section 2. A *code* (n, M, Δ) is also defined as in Section 2, except that in this case the encoder operates on $\mathbf{X}^n$ and $\mathbf{Y}^n$. Thus (2.2a) is replaced by

$$F_E: \mathscr{X}^n \times \mathscr{Y}^n \to \mathscr{I}_M. \tag{A.1}$$

The set of *achievable* rate pairs (R, d) is also as defined in Section 2. Letting $\mathscr{R}_1$ be the set of achievable rate pairs. we define

$$R_{X|Y}(d) = \min_{(R,d)\in\mathscr{R}_1} R. \tag{A.2}$$

The characterization of $R_{X|Y}(d)$ is given as follows: For $d \geqslant 0$. let $\mathscr{M}_0(d)$ be the set of measurable functions $\hat{X}: \Omega \to \hat{\mathscr{X}}$. such that $E\,D(X, \hat{X}) \leqslant d$. Then for $d > 0$.

$$R_{X|Y}(d) = \inf_{\hat{X} \in \mathscr{M}_0(d)} I(X; \hat{X} \mid Y) \triangleq \bar{R}_1(d). \tag{A.3}$$

Since $R_{X|Y}(d)$ is continuous at $d = 0$, (A.3) characterizes $R_{X|Y}(0)$ also.

Berger (1971) established (A.3) for the case where $\mathscr{X}$, $\mathscr{Y}$, $\hat{\mathscr{X}}$ are discrete. (See Gray (1972, 1973) for a more complete discussion of $R_{X|Y}$.) The extention of (A.3) to the general case is a fairly straightforward task. The converse, i.e., $R_{X|Y}(d) \geqslant \bar{R}_1(d)$, is proved by writing for any code $(n\ M, \Delta)$

$$\begin{aligned}
\frac{1}{n} \log M &\overset{(a)}{\geqslant} \frac{1}{n} I(\mathbf{X}^n; F_e(\mathbf{X}^n, \mathbf{Y}^n) \mid \mathbf{Y}) \overset{(b)}{\geqslant} \frac{1}{n} I(\mathbf{X}^n; \tilde{\mathbf{X}}^n \mid \mathbf{Y}) \\
&= I(\mathbf{X}; \tilde{\mathbf{X}}\mathbf{Y}) - I(\mathbf{X}; \mathbf{Y}) \overset{(c)}{\geqslant} \frac{1}{n} \sum_{k=1}^{n} I(X_k\,; \hat{X}_k Y_k) - I(X_k\,; Y_k) \qquad \text{(A.4)} \\
&= \sum_{k=1}^{n} I(X_k\,; \hat{X}_k \mid Y_k) \overset{(d)}{\geqslant} \frac{1}{n} \sum_{k=1}^{n} \bar{R}_1(\Delta_k) \overset{(e)}{\geqslant} \bar{R}_1(\Delta),
\end{aligned}$$

where $\Delta_k = E\,D(X_k\,, \hat{X}_k)$, so that $\Delta = (1/n)\sum_k \Delta_k$. These steps are justified as follows: (a) from the fact that $F_E(\mathbf{X}^n, \mathbf{Y}^n) \in \mathscr{I}_M$; (b) from the data-processing theorem (Lemma 3.4 of Wyner); (c) from a standard inequality which follows from the independence of the pairs $\{(X_k\,, Y_k)\}$ (see Kadota, 1971); (d) from the definition of $\bar{R}_1$; (e) from the concavity of $\bar{R}_1(d)$ which can easily be verified. The converse follows on applying the definition of $R_{X|Y}(d)$.

To prove the direct half, i.e., $R_{X|Y}(d) \leqslant \bar{R}_1(d)$, we first observe that Berger's result, goes over exactly when $\mathscr{Y}$ is finite. For arbitrary $\mathscr{Y}$, let $\hat{X} \in \mathscr{M}_0(d)$. Then write

$$I(X; \hat{X} \mid Y) = I(X; Y\hat{X}) - I(X; Y), \tag{A.4}$$

which is meaningful since $I(X; Y) < \infty$ (2.2). Next, let $\epsilon > 0$ be arbitrary and let $\mathscr{P}_Y$ be a finite partition such that the corresponding $\tilde{Y}$ satisfies

$$|\, I(X; \tilde{Y}\hat{X}) - I(X; Y\hat{X})| \leqslant \epsilon/2, \tag{A.5a}$$

$$|\, I(X; \tilde{Y}) - I(X; Y)| < \epsilon/2. \tag{A.5b}$$

It follows that

$$I(X; \hat{X} \mid \tilde{Y}) \leqslant I(X; \hat{X} \mid Y) + \epsilon. \tag{A.6}$$

We then apply Berger's proof with Y replaced by $\tilde{Y}$.

APPENDIX B: Convexity of $\bar{R}(d)$

We must show that for $d_1, d_2 \geqslant 0$ and $0 \leqslant \theta \leqslant 1$,

$$\bar{R}(\theta d_1 + (1-\theta) d_2) \leqslant \theta \bar{R}(d_1) + (1-\theta) \bar{R}(d_2). \tag{B.1}$$

To begin with, let $\epsilon > 0$ be given, and let $Z_1 \in \mathscr{M}(d_1)$, $Z_2 \in \mathscr{M}(d_2)$ be such that (use (2.13))

$$I(X; Z_i \mid Y) \leqslant R(d_i) + \epsilon, \qquad i = 1, 2. \tag{B.2}$$

and

$$E\, D(X, f_i(Y, Z_i)) \leqslant d_i, \qquad i = 1, 2. \tag{B.3}$$

Here $f_i(Y, Z_i)$, $i = 1, 2$, are the functions whose existance is guaranteed by the definition of $\mathscr{M}(d)$ (see (2.8)). Now let V be a random variable which is independent of X, Y, Z_1, Z_2, with

$$\Pr\{V = 1\} = \theta, \qquad \Pr\{V = 2\} = 1 - \theta.$$

Then set

$$\begin{aligned} Z &= (Z_1, V), \qquad V = 1, \\ &= (Z_2, V), \qquad V = 2. \end{aligned}$$

Letting

$$\begin{aligned} f(Y, Z) &= f_1(Y, Z_1), \qquad \text{when} \quad V = 1, \\ &= f_2(Y, Z_2), \qquad \text{when} \quad V = 2, \end{aligned}$$

we have, using (B.3),

$$\begin{aligned} E\, D(X, f(Y, Z)) &= \theta\, E\, D(X, f_1(Y\ Z_1)) + (1-\theta) E\, D(X, f_2(Y, Z_2)) \\ &\leqslant \theta d_1 + (1-\theta) d_2 . \end{aligned}$$

Thus $Z \in \mathscr{M}(\theta d_1 + (1-\theta) d_2)$, and $I(X; Z \mid Y) \geqslant R(\theta d_1 + (1-\theta) d_2)$. Further,

$$\begin{aligned} &R(\theta d_1 + (1-\theta) d_2) \\ &\quad \leqslant I(X; Z \mid Y) \overset{(a)}{=} I(X; ZV \mid Y) - I(X; V \mid YZ) \\ &\qquad \overset{(b)}{=} I(X; V \mid Y) + I(X; Z \mid YV) \\ &\qquad \overset{(c)}{=} I(X; Z \mid YV) = \Pr(V = 1)\, I(X; Z \mid Y, V = 1) \\ &\qquad\quad + \Pr(V = 2)\, I(X; Z \mid Y, V = 2) \\ &\qquad = \theta I(X; Z_1 \mid Y) + (1-\theta)\, I(X; Z_2 \mid Y) \\ &\qquad \overset{(d)}{\leqslant} \theta(R(d_1) + \epsilon) + (1-\theta)(R(D_2) + \epsilon). \end{aligned}$$

Step (a) follows from Lemma 3.3 of Wyner (1977) and $I(X; V \mid YZ) < \infty$; step (b) from the same lemma and the fact that V is determined by Z so that $I(X; V \mid YZ) = 0$; step (c) from the independence of (X, Y) and V so that $I(X; V \mid Y) = 0$, and step (d) from (B.2). Letting $\epsilon \to 0$, we have (B.1). ▮

Acknowledgments

We acknowledge with thanks the help of T. T. Kadota, S. P. Lloyd, and J. Ziv. In particular, Dr. Ziv contributed significantly to the streamlining of the proof of converse theorem (Theorem 2.1).

Received: December 5, 1975; revised: September 23, 1977

References

Ash, R. B. (1965), "Information Theory," Interscience, New York.

Ash, R. B. (1972), "Real Analysis and Probability," Academic Press, New York.

Berger, Toby (1971), "Rate-Distortion Theory," Prentice–Hall, Englewood Cliffs, N.J.

Gallager, R. G. (1968), "Information Theory and Reliable Communication," Wiley, New York.

Gray, R. M. (1973), A new class of lower bounds to information rates of stationary sources via conditional rate-distortion functions, *IEEE Trans. Information Theory* **IT-19**, 480–489.

Gray, R. M. (1972), "Conditional Rate Distortion Theory," Stanford Electronics Labs., Stanford, Calif., Tech. Rep. 6502-2.

Kadota, T. T. (1971), On the mutual information of memoryless channels, *IEEE Trans. Information Theory* **IT-17**, 140–143.

Loeve, M. (1955), "Probability Theory," Van Nostrand, Princeton, N.J.

Pinsker, M. (1964), "Information and Information Stability of Random Variables and Processes," Holden–Day, San Francisco.

Slepian, D., and Wolf, J. K. (1973), Noiseless coding of correlated information sources, *IEEE Trans. Information Theory* **IT-19**, 471–480.

Wyner, A. D. (1978), A definition of conditional mutual information for arbitrary ensembles, *Inform. Contr.* **38**, 51–59.

Wyner, A. D., and Ziv, J. (1973), Rate-distortion function for source coding with side information at the decoder, *IEEE Trans. Information Theory* **IT-22**, 1–11.