Foundations and Trends® in
Communications and Information Theory
Vol. 4, Nos. 4–5 (2007) 265–444

DOI: 10.1561/0100000028

now
the essence of knowledge

Topics in Multi-User Information Theory

Gerhard Kramer

Bell Laboratories, Alcatel-Lucent, 600 Mountain Avenue, Murray Hill, New Jersey, 07974, USA, gkr@bell-labs.com

Abstract

This survey reviews fundamental concepts of multi-user information theory. Starting with typical sequences, the survey builds up knowledge on random coding, binning, superposition coding, and capacity converses by introducing progressively more sophisticated tools for a selection of source and channel models. The problems addressed include: Source Coding; Rate-Distortion and Multiple Descriptions; Capacity-Cost; The Slepian–Wolf Problem; The Wyner-Ziv Problem; The Gelfand-Pinsker Problem; The Broadcast Channel; The Multiaccess Channel; The Relay Channel; The Multiple Relay Channel; and The Multiaccess Channel with Generalized Feedback. The survey also includes a review of basic probability and information theory.

Notations and Acronyms

We use standard notation for probabilities, random variables, entropy, mutual information, and so forth. Table 1 lists notation developed in the appendices of this survey, and we use this without further explanation in the main body of the survey. We introduce the remaining notation as we go along. The reader is referred to the appendices for a review of the relevant probability and information theory concepts.

Table 1 Probability and information theory notation.

Sequences, Vectors, Matrices	
x^n	the finite sequence $x_1, x_2, \ldots, x_n$
$x^n y^m$	sequence concatenation: $x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_m$
$\underline{x}$	the vector $[x_1, x_2, \ldots, x_n]$
$\mathbf{H}$	a matrix
$\lvert\mathbf{Q}\rvert$	determinant of the matrix $\mathbf{Q}$

(*Continued*)

Table 1 (*Continued*)

Probability	
$\Pr[\mathcal{A}]$	probability of the event $\mathcal{A}$
$\Pr[\mathcal{A}\|\mathcal{B}]$	probability of event $\mathcal{A}$ conditioned on event $\mathcal{B}$
$P_X(\cdot)$	probability distribution of the random variable X
$P_{X\|Y}(\cdot)$	probability distribution of X conditioned on Y
$\mathrm{supp}(P_X)$	support of P_X
$p_X(\cdot)$	probability density of the random variable X
$p_{X\|Y}(\cdot)$	probability density of X conditioned on Y
$E[X]$	expectation of the real-valued X
$E[X\|\mathcal{A}]$	expectation of X conditioned on event $\mathcal{A}$
$\mathsf{Var}[X]$	variance of X
$\mathbf{Q}_{\underline{X}}$	covariance matrix of $\underline{X}$
Information Theory	
$H(X)$	entropy of the discrete random variable X
$H(X\|Y)$	entropy of X conditioned on Y
$I(X;Y)$	mutual information between X and Y
$I(X;Y\|Z)$	mutual information between X and Y conditioned on Z
$D(P_X\|\|P_Y)$	informational divergence between P_X and P_Y
$h(X)$	differential entropy of X
$h(X\|Y)$	differential entropy of X conditioned on Y
$H_2(\cdot)$	binary entropy function

1

Typical Sequences and Source Coding

1.1 Typical Sequences

Shannon introduced the notion of a "typical sequence" in his 1948 paper "A Mathematical Theory of Communication" [55]. To illustrate the idea, consider a discrete memoryless source (DMS), which is a device that emits symbols from a discrete and finite alphabet $\mathcal{X}$ in an independent and identically distributed (i.i.d.) manner (see Figure 1.1). Suppose the source probability distribution is $P_X(\cdot)$ where

$$P_X(0) = 2/3 \quad \text{and} \quad P_X(1) = 1/3. \tag{1.1}$$

Consider the following experiment: we generated a sequence of length 18 by using a random number generator with the distribution (1.1). We write this sequence below along with three other sequences that we generated artificially.

(a) 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
(b) 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0
(c) 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0
(d) 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1.

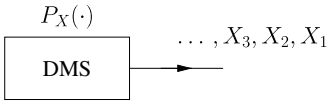

Fig. 1.1 A discrete memoryless source with distribution $P_X(\cdot)$.

If we compute the probabilities that these sequences were emitted by the source (1.1), we have

(a) $(2/3)^{18} \cdot (1/3)^0 \approx 6.77 \cdot 10^{-4}$
(b) $(2/3)^9 \cdot (1/3)^9 \approx 1.32 \cdot 10^{-6}$
(c) $(2/3)^{11} \cdot (1/3)^7 \approx 5.29 \cdot 10^{-6}$
(d) $(2/3)^0 \cdot (1/3)^{18} \approx 2.58 \cdot 10^{-9}$.

Thus, the first sequence is the most probable one by a large margin. However, the reader will likely *not* be surprised to find out that it is sequence (c) that was actually put out by the random number generator. Why is this intuition correct? To explain this, we must define more precisely what one might mean by a "typical" sequence.

1.2 Entropy-Typical Sequences

Let x^n be a finite sequence whose ith entry x_i takes on values in $\mathcal{X}$. We write $\mathcal{X}^n$ for the Cartesian product of the set $\mathcal{X}$ with itself n times, i.e., we have $x^n \in \mathcal{X}^n$. Let $N(a|x^n)$ be the number of positions of x^n having the letter a, where $a \in \mathcal{X}$.

There are several natural definitions for typical sequences. Shannon in [55, Sec. 7] chose a definition based on the entropy of a random variable X. Suppose that X^n is a sequence put out by the DMS $P_X(\cdot)$, which means that $P_{X^n}(x^n) = \prod_{i=1}^n P_X(x_i)$ is the probability that x^n was put out by the DMS $P_X(\cdot)$. More generally, we will use the notation

$$P_X^n(x^n) = \prod_{i=1}^n P_X(x_i). \tag{1.2}$$

We further have

$$P_X^n(x^n) = \begin{cases} \prod_{a \in \mathrm{supp}(P_X)} P_X(a)^{N(a|x^n)} & \text{if } N(a|x^n) = 0 \text{ whenever } P_X(a) = 0 \\ 0 & \text{else} \end{cases} \tag{1.3}$$

and intuitively one might expect that the letter a occurs about $N(a|x^n) \approx nP_X(a)$ times, so that $P_X^n(x^n) \approx \Pi_{a\in\mathrm{supp}(P_X)} P_X(a)^{nP_X(a)}$ or

$$-\frac{1}{n}\log_2 P_X^n(x^n) \approx \sum_{a\in\mathrm{supp}(P_X)} -P_X(a)\log_2 P_X(a).$$

Shannon therefore defined a sequence x^n to be typical with respect to ϵ and $P_X(\cdot)$ if

$$\left|\frac{-\log_2 P_X^n(x^n)}{n} - H(X)\right| < \epsilon \tag{1.4}$$

for some small positive ϵ. The sequences satisfying (1.4) are sometimes called *weakly* typical sequences or *entropy*-typical sequences [19, p. 40]. We can equivalently write (1.4) as

$$2^{-n[H(X)+\epsilon]} < P_X^n(x^n) < 2^{-n[H(X)-\epsilon]}. \tag{1.5}$$

Example 1.1. If $P_X(\cdot)$ is uniform then for any x^n we have

$$P_X^n(x^n) = |\mathcal{X}|^{-n} = 2^{-n\log_2|\mathcal{X}|} = 2^{-nH(X)} \tag{1.6}$$

and *all* sequences in $\mathcal{X}^n$ are entropy-typical.

Example 1.2. The source (1.1) has $H(X) \approx 0.9183$ and the above four sequences are entropy-typical with respect to $P_X(\cdot)$ if

(a) $\epsilon > 1/3$
(b) $\epsilon > 1/6$
(c) $\epsilon > 1/18$
(d) $\epsilon > 2/3$.

Note that sequence (c) requires the smallest ϵ.

We remark that *entropy* typicality applies to *continuous* random variables with a density if we replace the probability $P_X^n(x^n)$ in (1.4) with the density value $p_X^n(x^n)$. In contrast, the next definition can be used only for discrete random variables.

1.3 Letter-Typical Sequences

A perhaps more natural definition for *discrete* random variables than (1.4) is the following. For $\epsilon \geq 0$, we say a sequence x^n is ϵ-*letter* typical with respect to $P_X(\cdot)$ if

$$\left|\frac{1}{n}N(a|x^n) - P_X(a)\right| \leq \epsilon \cdot P_X(a) \quad \text{for all } a \in \mathcal{X} \tag{1.7}$$

The set of x^n satisfying (1.7) is called the ϵ-*letter-typical set* $T_\epsilon^n(P_X)$ with respect to $P_X(\cdot)$. The letter typical x^n are thus sequences whose *empirical* probability distribution is close to $P_X(\cdot)$.

Example 1.3. If $P_X(\cdot)$ is uniform then ϵ-letter typical x^n satisfy

$$\frac{(1-\epsilon)n}{|\mathcal{X}|} \leq N(a|x^n) \leq \frac{(1+\epsilon)n}{|\mathcal{X}|} \tag{1.8}$$

and if $\epsilon < |\mathcal{X}| - 1$, as is usually the case, then *not* all x^n are letter-typical. The definition (1.7) is then more restrictive than (1.4) (see Example 1.1).

We will generally rely on letter typicality, since for discrete random variables it is just as easy to use as entropy typicality, but can give stronger results.

We remark that one often finds small variations of the conditions (1.7). For example, for *strongly* typical sequences one replaces the $\epsilon P_X(a)$ on the right-hand side of (1.7) with ϵ or $\epsilon/|\mathcal{X}|$ (see [19, p. 33], and [18, pp. 288, 358]). One further often adds the condition that $N(a|x^n) = 0$ if $P_X(a) = 0$ so that typical sequences cannot have zero-probability letters. Observe, however, that this condition is included in (1.7). We also remark that the letter-typical sequences are simply called "typical sequences" in [44] and "robustly typical sequences" in [46]. In general, by the label "letter-typical" we mean any choice of typicality where one performs a per-alphabet-letter test on the empirical probabilities. We will focus on the definition (1.7).

We next develop the following theorem that describes some of the most important properties of letter-typical sequences and sets.

Let $\mu_X = \min_{x \in \mathrm{supp}(P_X)} P_X(x)$ and define

$$\delta_\epsilon(n) = 2|\mathcal{X}| \cdot e^{-n\epsilon^2 \mu_X}. \tag{1.9}$$

Observe that $\delta_\epsilon(n) \to 0$ for fixed ϵ, $\epsilon > 0$, and $n \to \infty$.

Theorem 1.1. Suppose $0 \le \epsilon \le \mu_X$, $x^n \in T_\epsilon^n(P_X)$, and X^n is emitted by a DMS $P_X(\cdot)$. We have

$$2^{-n(1+\epsilon)H(X)} \le P_X^n(x^n) \le 2^{-n(1-\epsilon)H(X)} \tag{1.10}$$

$$(1 - \delta_\epsilon(n))\, 2^{n(1-\epsilon)H(X)} \le |T_\epsilon^n(P_X)| \le 2^{n(1+\epsilon)H(X)} \tag{1.11}$$

$$1 - \delta_\epsilon(n) \le \Pr[X^n \in T_\epsilon^n(P_X)] \le 1. \tag{1.12}$$

Proof. Consider (1.10). For $x^n \in T_\epsilon^n(P_X)$, we have

$$\begin{aligned} P_X^n(x^n) &= \prod_{a \in \mathrm{supp}(P_X)} P_X(a)^{N(a|x^n)} \\ &\le \prod_{a \in \mathrm{supp}(P_X)} P_X(a)^{nP_X(a)(1-\epsilon)} \\ &= 2^{\sum_{a \in \mathrm{supp}(P_X)} nP_X(a)(1-\epsilon)\log_2 P_X(a)} \\ &= 2^{-n(1-\epsilon)H(X)}, \end{aligned} \tag{1.13}$$

where the inequality follows because, by the definition (1.7), typical x^n satisfy $N(a|x^n)/n \ge P_X(a)(1-\epsilon)$. One can similarly prove the left-hand side of (1.10).

Next, consider (1.12). In the appendix of this section, we prove the following result using the Chernoff bound:

$$\Pr\left[\left|\frac{N(a|X^n)}{n} - P_X(a)\right| > \epsilon\, P_X(a)\right] \le 2 \cdot e^{-n\epsilon^2 \mu_X}, \tag{1.14}$$

where $0 \le \epsilon \le \mu_X$. We thus have

$$\begin{aligned} \Pr[X^n \notin T_\epsilon^n(P_X)] &= \Pr\left[\bigcup_{a \in \mathcal{X}} \left\{\left|\frac{N(a|X^n)}{n} - P_X(a)\right| > \epsilon\, P_X(a)\right\}\right] \\ &\le \sum_{a \in \mathcal{X}} \Pr\left[\left|\frac{N(a|X^n)}{n} - P_X(a)\right| > \epsilon\, P_X(a)\right] \\ &\le 2|\mathcal{X}| \cdot e^{-n\epsilon^2 \mu_X}, \end{aligned} \tag{1.15}$$

where we have used the union bound (see (A.5)) for the second step. This proves the left-hand side of (1.12).

Finally, for (1.11) observe that

$$\begin{aligned}\Pr[X^n \in T_\epsilon^n(P_X)] &= \sum_{x^n \in T_\epsilon^n(P_X)} P_X^n(x^n) \\ &\le |T_\epsilon^n(P_X)|\, 2^{-n(1-\epsilon)H(X)}, \end{aligned} \tag{1.16}$$

where the inequality follows by (1.13). Using (1.15) and (1.16), we thus have

$$|T_\epsilon^n(P_X)| \ge (1-\delta_\epsilon(n))\, 2^{n(1-\epsilon)H(X)}. \tag{1.17}$$

We similarly derive the right-hand side of (1.11). □

1.4 Source Coding

The source coding problem is depicted in Figure 1.2. A DMS $P_X(\cdot)$ emits a sequence x^n of symbols that are passed to an encoder. The source encoder "compresses" x^n into an index w and sends w to the decoder. The decoder reconstructs x^n from w as $\hat{x}^n(w)$, and is said to be successful if $\hat{x}^n(w) = x^n$.

The source encoding can be done in several ways:

- Fixed-length to fixed-length coding (or block-to-block coding).
- Fixed-length to variable-length coding (block-to-variable-length coding).
- Variable-length to fixed-length coding (variable-length-to-block coding).
- Variable-length to variable-length coding.

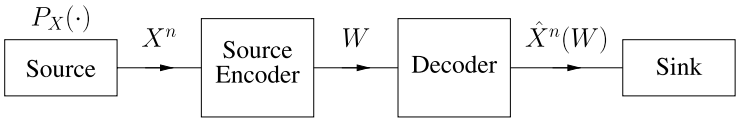

Fig. 1.2 The source coding problem.

We will here consider only the first two approaches. For a block-to-variable-length scheme, the number of bits transmitted by the encoder depends on x^n. We will consider the case where every source sequence is assigned a *unique* index w. Hence, one can reconstruct x^n perfectly. Let $L(x^n)$ be the number of bits transmitted for x^n. The goal is to minimize the *average* rate $R = E\left[L(X^N)\right]/n$.

For a block-to-block encoding scheme, the index w takes on one of 2^{nR} indexes w, $w = 1, 2, \ldots, 2^{nR}$, and we assume that 2^{nR} is a positive integer. The encoder sends exactly nR bits for every source sequence x^n, and the goal is to make R as small as possible. Observe that block-to-block encoding might require the encoder to send the *same* w for two *different* source sequences.

Suppose first that we permit no error in the reconstruction. We use the block-to-variable-length encoder, choose an n and an ϵ, and assign each sequence in $T_\epsilon^n(P_X)$ a unique positive integer w. According to (1.11), these indexes w can be represented by at most $n(1+\epsilon)H(X) + 1$ bits. Next, the encoder collects a sequence x^n. If $x^n \in T_\epsilon^n(P_X)$, then the encoder sends a "0" followed by the $n(1+\epsilon)H(X)+1$ bits that represent this sequence. If $x^n \notin T_\epsilon^n(P_X)$, then the encoder sends a "1" followed by $n\log_2|\mathcal{X}| + 1$ bits that represent x^n. The average number of bits per source symbol is the compression rate R, and it is upper bounded by

$$\begin{aligned} R &\le \Pr\left[X^n \in T_\epsilon^n(P_X)\right]\left[(1+\epsilon)H(X) + 2/n\right] \\ &\quad + \Pr\left[X^n \notin T_\epsilon^n(P_X)\right](\log_2|\mathcal{X}| + 2/n) \\ &\le (1+\epsilon)H(X) + 2/n + \delta_\epsilon(n)(\log_2|\mathcal{X}| + 2/n). \end{aligned} \tag{1.18}$$

But since $\delta_\epsilon(n) \to 0$ as $n \to \infty$, we can transmit at any rate above $H(X)$ bits per source symbol. For example, if the DMS is binary with $P_X(0) = 1 - P_X(1) = 2/3$, then we can transmit the source outputs in a lossless fashion at any rate above $H(X) \approx 0.9183$ bits per source symbol.

Suppose next that we must use a block-to-block encoder, but that we permit a small error probability in the reconstruction. Based on the above discussion, we can transmit at any rate above $(1+\epsilon)H(X)$ bits

per source symbol with an error probability $\delta_\epsilon(n)$. By making n large, we can make $\delta_\epsilon(n)$ as close to zero as desired.

But what about a converse result? Can one compress with a small error probability, or even zero error probability, at rates below $H(X)$? We will prove a converse for block-to-block encoders only, since the block-to-variable-length case requires somewhat more work.

Consider Fano's inequality (see Section A.10) which ensures us that

$$H_2(P_e) + P_e \log_2(|\mathcal{X}|^n - 1) \geq H(X^n|\hat{X}^n), \tag{1.19}$$

where $P_e = \Pr[\hat{X}^n \neq X^n]$. Recall that there are at most 2^{nR} different sequences $\hat{x}^n$, and that $\hat{x}^n$ is a function of x^n. We thus have

$$\begin{aligned} nR &\geq H(\hat{X}^n) \\ &= H(\hat{X}^n) - H(\hat{X}^n|X^n) \\ &= I(X^n;\hat{X}^n) \\ &= H(X^n) - H(X^n|\hat{X}^n) \\ &= nH(X) - H(X^n|\hat{X}^n) \\ &\geq n\left[H(X) - \frac{H_2(P_e)}{n} - P_e \log_2|\mathcal{X}|\right], \end{aligned} \tag{1.20}$$

where the last step follows by (1.19). Since we require that P_e be zero, or approach zero with n, we find that $R \geq H(X)$ for block-to-block encoders with arbitrarily small positive P_e. This is the desired converse.

1.5 Jointly and Conditionally Typical Sequences

Let $N(a,b|x^n,y^n)$ be the number of times the pair (a,b) occurs in the sequence of pairs $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$. The *jointly* typical set with respect to $P_{XY}(\cdot)$ is simply

$$\begin{aligned} T_\epsilon^n(P_{XY}) = \Bigg\{ (x^n,y^n) : &\left|\frac{1}{n}N(a,b|x^n,y^n) - P_{XY}(a,b)\right| \\ &\leq \epsilon \cdot P_{XY}(a,b) \quad \text{for all } (a,b) \in \mathcal{X}\times\mathcal{Y}\Bigg\}. \end{aligned} \tag{1.21}$$

The reader can easily check that $(x^n,y^n) \in T_\epsilon^n(P_{XY})$ implies both $x^n \in T_\epsilon^n(P_X)$ and $y^n \in T_\epsilon^n(P_Y)$.

Consider the conditional distribution $P_{Y|X}(\cdot)$ and define

$$P^n_{Y|X}(y^n|x^n) = \prod_{i=1}^{n} P_{Y|X}(y_i|x_i) \tag{1.22}$$

$$T^n_\epsilon(P_{XY}|x^n) = \{y^n : (x^n, y^n) \in T^n_\epsilon(P_{XY})\}. \tag{1.23}$$

Observe that $T^n_\epsilon(P_{XY}|x^n) = \emptyset$ if x^n is not in $T^n_\epsilon(P_X)$. We shall further need the following counterpart of $\delta_\epsilon(n)$ in (1.9):

$$\delta_{\epsilon_1,\epsilon_2}(n) = 2|\mathcal{X}||\mathcal{Y}| \exp\left(-n \cdot \frac{(\epsilon_2 - \epsilon_1)^2}{1+\epsilon_1} \cdot \mu_{XY}\right), \tag{1.24}$$

where $\mu_{XY} = \min_{(a,b)\in \mathrm{supp}(P_{XY})} P_{XY}(a,b)$ and $0 \le \epsilon_1 < \epsilon_2 \le 1$. Note that $\delta_{\epsilon_1,\epsilon_2}(n) \to 0$ as $n \to \infty$. In the Appendix, we prove the following theorem that generalizes Theorem 1.1 to include conditioning.

Theorem 1.2. Suppose $0 \le \epsilon_1 < \epsilon_2 \le \mu_{XY}$, $(x^n, y^n) \in T^n_{\epsilon_1}(P_{XY})$, and (X^n, Y^n) was emitted by the DMS $P_{XY}(\cdot)$. We have

$$2^{-nH(Y|X)(1+\epsilon_1)} \le P^n_{Y|X}(y^n|x^n) \le 2^{-nH(Y|X)(1-\epsilon_1)} \tag{1.25}$$

$$(1 - \delta_{\epsilon_1,\epsilon_2}(n))\, 2^{nH(Y|X)(1-\epsilon_2)} \le |T^n_{\epsilon_2}(P_{XY}|x^n)| \le 2^{nH(Y|X)(1+\epsilon_2)} \tag{1.26}$$

$$1 - \delta_{\epsilon_1,\epsilon_2}(n) \le \Pr\left[Y^n \in T^n_{\epsilon_2}(P_{XY}|x^n) \,|\, X^n = x^n\right] \le 1. \tag{1.27}$$

The following result follows easily from Theorem 1.2 and will be extremely useful to us.

Theorem 1.3. Consider a joint distribution $P_{XY}(\cdot)$ and suppose $0 \le \epsilon_1 < \epsilon_2 \le \mu_{XY}$, Y^n is emitted by a DMS $P_Y(\cdot)$, and $x^n \in T^n_{\epsilon_1}(P_X)$. We have

$$\begin{aligned}(1 - \delta_{\epsilon_1,\epsilon_2}(n))\; & 2^{-n[I(X;Y)+2\epsilon_2 H(Y)]} \\ & \le \Pr\left[Y^n \in T^n_{\epsilon_2}(P_{XY}|x^n)\right] \le 2^{-n[I(X;Y)-2\epsilon_2 H(Y)]}.\end{aligned} \tag{1.28}$$

Proof. The upper bound follows by (1.25) and (1.26):

$$\begin{aligned}\Pr\left[Y^n \in T_{\epsilon_2}^n(P_{XY}|x^n)\right] &= \sum_{y^n \in T_{\epsilon_2}(P_{XY}|x^n)} P_Y^n(y^n) \\ &\le 2^{nH(Y|X)(1+\epsilon_2)}\, 2^{-nH(Y)(1-\epsilon_2)} \\ &\le 2^{-n[I(X;Y)-2\epsilon_2 H(Y)]}. \end{aligned} \tag{1.29}$$

The lower bound also follows from (1.25) and (1.26). □

For small ϵ_1 and ϵ_2, large n, typical (x^n, y^n), and (X^n, Y^n) emitted by a DMS $P_{XY}(\cdot)$, we thus have

$$P_{Y|X}^n(y^n|x^n) \approx 2^{-nH(Y|X)} \tag{1.30}$$

$$|T_{\epsilon_2}^n(P_{XY}|x^n)| \approx 2^{nH(Y|X)} \tag{1.31}$$

$$\Pr\left[Y^n \in T_{\epsilon_2}^n(P_{XY}|x^n)\,|\,X^n = x^n\right] \approx 1 \tag{1.32}$$

$$\Pr\left[Y^n \in T_{\epsilon_2}^n(P_{XY}|x^n)\right] \approx 2^{-nI(X;Y)}. \tag{1.33}$$

We remark that the probabilities in (1.27) and (1.28) (or (1.32) and (1.33)) differ only in whether or not one conditions on $X^n = x^n$.

Example 1.4. Suppose X and Y are independent, in which case the approximations (1.32) and (1.33) both give

$$\Pr\left[Y^n \in T_{\epsilon_2}^n(P_{XY}|x^n)\right] \approx 1. \tag{1.34}$$

Note, however, that the precise version (1.28) of (1.33) is trivial for large n. This example shows that one must exercise caution when working with the approximations (1.30)–(1.33).

Example 1.5. Suppose that $X = Y$ so that (1.33) gives

$$\Pr\left[Y^n \in T_{\epsilon_2}^n(P_{XY}|x^n)\right] \approx 2^{-nH(X)}. \tag{1.35}$$

This result should not be surprising because $|T_{\epsilon_2}^n(P_X)| \approx 2^{nH(X)}$ and we are computing the probability of the event $X^n = x^n$ for some $x^n \in T_{\epsilon_1}^n(P_{XY})$ (the fact that ϵ_2 is larger than ϵ_1 does not play a role for large n).

1.6 Appendix: Proofs

Proof of Inequality (1.14)

We prove the bound (1.14). Consider first $P_X(a) = 0$ for which we have

$$\Pr\left[\frac{N(a|X^n)}{n} > P_X(a)(1+\epsilon)\right] = 0. \tag{1.36}$$

Next, suppose that $P_X(a) > 0$. Using the Chernoff bound, we have

$$\begin{aligned}
\Pr\left[\frac{N(a|X^n)}{n} > P_X(a)(1+\epsilon)\right] &\le \Pr\left[\frac{N(a|X^n)}{n} \ge P_X(a)(1+\epsilon)\right] \\
&\le E\left[e^{\nu N(a|X^n)/n}\right] e^{-\nu P_X(a)(1+\epsilon)} \\
&= \left[\sum_{m=0}^{n} \Pr\left[N(a|X^n) = m\right] e^{\nu m/n}\right] e^{-\nu P_X(a)(1+\epsilon)} \\
&= \left[\sum_{m=0}^{n} \binom{n}{m} P_X(a)^m (1-P_X(a))^{n-m} e^{\nu m/n}\right] e^{-\nu P_X(a)(1+\epsilon)} \\
&= \left[(1-P_X(a)) + P_X(a)e^{\nu/n}\right]^n e^{-\nu P_X(a)(1+\epsilon)}. \qquad (1.37) \\
& \qquad (1.38)
\end{aligned}$$

Optimizing (1.38) with respect to ν, we find that

$$\begin{array}{ll} \nu = \infty & \text{if } P_X(a)(1+\epsilon) \ge 1 \\ e^{\nu/n} = \frac{(1-P_X(a))(1+\epsilon)}{1-P_X(a)(1+\epsilon)} & \text{if } P_X(a)(1+\epsilon) < 1. \end{array} \tag{1.39}$$

In fact, the Chernoff bound correctly identifies the probabilities to be 0 and $P_X(a)^n$ for the cases $P_X(a)(1+\epsilon) > 1$ and $P_X(a)(1+\epsilon) = 1$, respectively. More interestingly, for $P_X(a)(1+\epsilon) < 1$ we insert (1.39) into (1.38) and obtain

$$\Pr\left[\frac{N(a|X^n)}{n} \ge P_X(a)(1+\epsilon)\right] \le 2^{-nD(P_B\|P_A)}, \tag{1.40}$$

where A and B are binary random variables with

$$\begin{aligned} P_A(0) &= 1 - P_A(1) = P_X(a) \\ P_B(0) &= 1 - P_B(1) = P_X(a)(1+\epsilon). \end{aligned} \tag{1.41}$$

We can write $P_B(0) = P_A(0)(1+\epsilon)$ and hence

$$D(P_B\|P_A) = P_A(0)(1+\epsilon)\log_2(1+\epsilon) + [1 - P_A(0)(1+\epsilon)]\log_2\left(\frac{1-P_A(0)(1+\epsilon)}{1-P_A(0)}\right). \quad (1.42)$$

We wish to further simplify (1.42). The first two derivatives of (1.42) with respect to ϵ are

$$\frac{dD(P_B\|P_A)}{d\epsilon} = P_A(0)\log_2\left(\frac{(1-P_A(0))(1+\epsilon)}{(1-P_A(0))(1+\epsilon)}\right) \quad (1.43)$$

$$\frac{d^2D(P_B\|P_A)}{d\epsilon^2} = \frac{P_A(0)\log_2(e)}{(1+\epsilon)[1-P_A(0)(1+\epsilon)]}. \quad (1.44)$$

We find that (1.43) is zero for $\epsilon = 0$ and we can lower bound (1.44) by $P_X(a)\log_2(e)$ for $0 \le \epsilon \le \mu_X$. The second derivative of $D(P_B\|P_A)$ with respect to ϵ is thus larger than $P_X(a)\log_2(e)$ and so we have

$$D(P_B\|P_A) \ge \epsilon^2 \cdot P_A(0)\log_2(e) \quad (1.45)$$

for $0 \le \epsilon \le \mu_X$. Combining (1.40) and (1.45) we arrive at

$$\Pr\left[\frac{N(a|X^n)}{n} \ge P_X(a)(1+\epsilon)\right] \le e^{-n\epsilon^2 P_X(a)}. \quad (1.46)$$

One can similarly bound

$$\Pr\left[\frac{N(a|X^n)}{n} \le P_X(a)(1-\epsilon)\right] \le e^{-n\epsilon^2 P_X(a)}. \quad (1.47)$$

Note that (1.46) and (1.47) are valid for all $a \in \mathcal{X}$ including a with $P_X(a) = 0$. However, the event in (1.14) has a strict inequality so we can improve the above bounds for the case $P_X(a) = 0$ (see (1.36)). This observation lets us replace $P_X(a)$ in (1.46) and (1.47) with μ_X and the result is (1.14).

Proof of Theorem 1.2

Suppose that $(x^n, y^n) \in T^n_{\epsilon_1}(P_{XY})$. We prove (1.25) by bounding

$$\begin{aligned} P^n_{Y|X}(y^n|x^n) &= \prod_{(a,b)\in \mathrm{supp}(P_{XY})} P_{Y|X}(b|a)^{N(a,b|x^n,y^n)} \\ &\le \prod_{(a,b)\in \mathrm{supp}(P_{XY})} P_{Y|X}(b|a)^{nP_{XY}(a,b)(1-\epsilon_1)} \\ &= 2^{n(1-\epsilon_1)\sum_{(a,b)\in \mathrm{supp}(P_{XY})} P_{XY}(a,b)\log_2 P_{Y|X}(b|a)} \\ &= 2^{-n(1-\epsilon_1)H(Y|X)}. \end{aligned} \tag{1.48}$$

This gives the lower bound in (1.25) and the upper bound is proved similarly.

Next, suppose that $(x^n, y^n) \in T^n_{\epsilon}(P_{XY})$ and (X^n, Y^n) was emitted by the DMS $P_{XY}(\cdot)$. We prove (1.27) as follows.

Consider first $P_{XY}(a,b) = 0$ for which we have

$$\Pr\left[\frac{N(a,b|X^n,Y^n)}{n} > P_{XY}(a,b)(1+\epsilon)\right] = 0. \tag{1.49}$$

Now consider $P_{XY}(a,b) > 0$. If $N(a|x^n) = 0$, then $N(a,b|x^n,y^n) = 0$ and

$$\Pr\left[\frac{N(a,b|X^n,Y^n)}{n} > P_{XY}(a,b)(1+\epsilon)\middle| X^n = x^n\right] = 0. \tag{1.50}$$

More interestingly, if $N(a|x^n) > 0$ then the Chernoff bound gives

$$\begin{aligned} &\Pr\left[\frac{N(a,b|X^n,Y^n)}{n} > P_{XY}(a,b)(1+\epsilon)\middle| X^n = x^n\right] \\ &\le \Pr\left[\frac{N(a,b|X^n,Y^n)}{n} \ge P_{XY}(a,b)(1+\epsilon)\middle| X^n = x^n\right] \\ &= \Pr\left[\frac{N(a,b|X^n,Y^n)}{N(a|x^n)} \ge \frac{P_{XY}(a,b)}{N(a|x^n)/n}(1+\epsilon)\middle| X^n = x^n\right] \end{aligned}$$

$$
\begin{aligned}
&\le E\left[e^{\nu N(a,b|X^n,Y^n)/N(a|x^n)}\middle| X^n=x^n\right] e^{-\nu\frac{P_{XY}(a,b)(1+\epsilon)}{N(a|x^n)/n}} \\
&= \left[\sum_{m=0}^{N(a|x^n)} \binom{N(a|x^n)}{m} P_{Y|X}(b|a)^m (1-P_{Y|X}(b|a))^{N(a|x^n)-m}\right. \\
&\qquad \left. e^{\nu m/N(a|x^n)}\right] e^{-\nu\frac{P_{XY}(a,b)(1+\epsilon)}{N(a|x^n)/n}} \\
&= \left[(1-P_{Y|X}(b|a)) + P_{Y|X}(b|a)e^{\nu/N(a|x^n)}\right]^{N(a|x^n)} e^{-\nu\frac{P_{XY}(a,b)(1+\epsilon)}{N(a|x^n)/n}}.
\end{aligned}
\tag{1.51}
$$

Minimizing (1.51) with respect to ν, we find that

$$
\begin{aligned}
&\nu=\infty && \text{if } P_{XY}(a,b)(1+\epsilon) \ge N(a|x^n)/n \\
&e^{\nu/N(a|x^n)} = \tfrac{P_X(a)(1-P_{Y|X}(b|a))(1+\epsilon)}{N(a|x^n)/n - P_{XY}(a,b)(1+\epsilon)} && \text{if } P_{XY}(a,b)(1+\epsilon) < N(a|x^n)/n.
\end{aligned}
\tag{1.52}
$$

Again, the Chernoff bound correctly identifies the probabilities to be 0 and $P_{Y|X}(b|a)^n$ for the cases $P_{XY}(a,b)(1+\epsilon) > N(a|x^n)/n$ and $P_{XY}(a,b)(1+\epsilon) = N(a|x^n)/n$, respectively. More interestingly, for $P_{XY}(a,b)(1+\epsilon) < N(a|x^n)/n$ we insert (1.52) into (1.51) and obtain

$$
\Pr\left[\frac{N(a,b|X^n)}{n} \ge P_{XY}(a,b)(1+\epsilon)\middle| X^n=x^n\right] \le 2^{-N(a|x^n)\,D(P_B\|P_A)},
\tag{1.53}
$$

where A and B are binary random variables with

$$
\begin{aligned}
P_A(0) &= 1-P_A(1) = P_{Y|X}(b|a) \\
P_B(0) &= 1-P_B(1) = \frac{P_{XY}(a,b)}{N(a|x^n)/n}(1+\epsilon).
\end{aligned}
\tag{1.54}
$$

We would like to have the form $P_B(0) = P_A(0)(1+\tilde{\epsilon})$ and compute

$$
\tilde{\epsilon} = \frac{P_X(a)}{N(a|x^n)/n}(1+\epsilon) - 1.
\tag{1.55}
$$

We can now use (1.41)–(1.46) to arrive at

$$
\begin{aligned}
&\Pr\left[\frac{N(a,b|X^n,Y^n)}{n} \ge P_{XY}(a,b)(1+\epsilon)\middle| X^n=x^n\right] \\
&\qquad \le e^{-N(a|x^n)\tilde{\epsilon}^2 P_{Y|X}(b|a)}
\end{aligned}
\tag{1.56}
$$

as long as $\epsilon \leq \min_{b:(a,b)\in\text{supp}(P_{XY})} P_{Y|X}(b|a)$. Now to guarantee that $\tilde{\epsilon}^2$ is positive, we must require that x^n is "more than" ϵ-letter typical, i.e., we must choose $x^n \in T_{\epsilon_1}(P_X)$, where $0 \leq \epsilon_1 < \epsilon$. Inserting $N(a|x^n)/n \geq (1+\epsilon_1)P_X(a)$ into (1.56), we have

$$\Pr\left[\frac{N(a,b|X^n,Y^n)}{n} \geq P_{XY}(a,b)(1+\epsilon)\middle| X^n = x^n\right] \leq e^{-n\frac{(\epsilon-\epsilon_1)^2}{1+\epsilon_1}P_{XY}(a,b)} \tag{1.57}$$

for $0 \leq \epsilon_1 < \epsilon \leq \mu_{XY}$ (we could allow ϵ to be up to $\min_{b:(a,b)\in\text{supp}(P_{XY})} P_{Y|X}(b|a)$ but we ignore this subtlety). One can similarly bound

$$\Pr\left[\frac{N(a,b|X^n,Y^n)}{n} \leq P_{XY}(a,b)(1-\epsilon)\middle| X^n = x^n\right] \leq e^{-n\frac{(\epsilon-\epsilon_1)^2}{1+\epsilon_1}P_{XY}(a,b)}. \tag{1.58}$$

As for the unconditioned case, note that (1.57) and (1.58) are valid for all (a,b) including (a,b) with $P_{XY}(a,b) = 0$. However, the event we are interested in has a strict inequality so that we can improve the above bounds for the case $P_{XY}(a,b) = 0$ (see (1.49)). We can thus replace $P_{XY}(a,b)$ in (1.57) and (1.58) with μ_{XY} and the result is

$$\Pr\left[\left|\frac{N(a,b|X^n,Y^n)}{n} - P_{XY}(a,b)\right| > \epsilon\, P_{XY}(a,b)\middle| X^n = x^n\right] \leq 2\cdot e^{-n\frac{(\epsilon-\epsilon_1)^2}{1+\epsilon_1}\mu_{XY}}. \tag{1.59}$$

for $0 \leq \epsilon_1 < \epsilon \leq \mu_{XY}$ (we could allow ϵ to be up to $\mu_{Y|X} = \min_{(a,b)\in\text{supp}(P_{XY})} P_{Y|X}(b|a)$ but, again, we ignore this subtlety). We thus have

$$\begin{aligned}&\Pr[Y^n \notin T_\epsilon^n(P_{XY}|x^n)|\, X^n = x^n] \\ &\quad = \Pr\left[\bigcup_{a,b}\left\{\left|\frac{N(a,b|X^n)}{n} - P_{XY}(a,b)\right| > \epsilon\, P_{XY}(a,b)\right\}\middle| X^n = x^n\right]\end{aligned}$$

$$\leq \sum_{a,b} \Pr\left[\left|\frac{N(a,b|X^n,Y^n)}{n} - P_{XY}(a,b)\right| > \epsilon\, P_{XY}(a,b) \middle| X^n = x^n\right]$$

$$\leq 2|\mathcal{X}||\mathcal{Y}| \cdot e^{-n\frac{(\epsilon-\epsilon_1)^2}{1+\epsilon_1}\mu_{XY}}, \tag{1.60}$$

where we have used the union bound for the last inequality. The result is the left-hand side of (1.27).

Finally, for $x^n \in T^n_{\epsilon_1}(P_X)$ and $0 \leq \epsilon_1 < \epsilon \leq \mu_{XY}$ we have

$$\begin{aligned}\Pr[Y^n \in T^n_\epsilon(P_{XY}|x^n)\,|\,X^n = x^n] &= \sum_{y^n \in T^n_\epsilon(P_{XY}|x^n)} P^n_{Y|X}(y^n|x^n) \\ &\leq |T^n_\epsilon(P_{XY}|x^n)|\, 2^{-n(1-\epsilon)H(Y|X)},\end{aligned} \tag{1.61}$$

where the inequality follows by (1.48). We thus have

$$|T^n_\epsilon(P_{XY}|x^n)| \geq (1 - \delta_{\epsilon_1,\epsilon}(n))\, 2^{n(1-\epsilon)H(Y|X)}. \tag{1.62}$$

We similarly have

$$|T^n_\epsilon(P_{XY}|x^n)| \leq 2^{n(1+\epsilon)H(Y|X)}. \tag{1.63}$$

2

Rate-Distortion and Multiple Descriptions

2.1 Problem Description

Rate distortion theory is concerned with quantization or lossy compression. Consider the problem shown in Figure 2.1. A DMS $P_X(\cdot)$ with alphabet $\mathcal{X}$ emits a sequence x^n that is passed to a source encoder. The encoder "quantizes" x^n into one of 2^{nR} sequences $\hat{x}^n(w)$, $w = 1, 2, \ldots 2^{nR}$, and sends the index w to the decoder (we assume that 2^{nR} is a positive integer in the remainder of this survey). Finally, the decoder puts out $\hat{x}^n(w)$ that is called a *reconstruction* of x^n. The letters $\hat{x}_i$ take on values in the alphabet $\hat{\mathcal{X}}$, which is often the same as $\mathcal{X}$ but could be different. The goal is to ensure that a non-negative and real-valued distortion $d^n(x^n, \hat{x}^n)$ is within some specified value D. A less restrictive version of the problem requires only that the *average* distortion $E\big[d^n(X^n, \hat{X}^n)\big]$ is at most D.

The choice of distortion function $d^n(\cdot)$ depends on the application. For example, for a DMS a natural distortion function is the normalized Hamming distance, i.e., we set

$$d^n(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i), \tag{2.1}$$

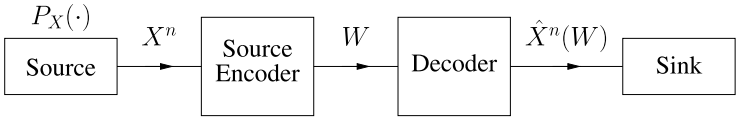

Fig. 2.1 The rate distortion problem.

where $d(x,\hat{x}) = 0$ if $\hat{x} = x$, and $d(x,\hat{x}) = 1$ if $\hat{x} \neq x$. For real sources, a natural choice might be the mean squared error

$$d^n(x^n,\hat{x}^n) = \frac{1}{n}\sum_{i=1}^{n}(x - x_i)^2. \tag{2.2}$$

Note that for binary $(0,1)$ sources both (2.1) and (2.2) are the same. Note further that both (2.1) and (2.2) are averages of *per-letter* distortion functions. Such a choice is not appropriate for many applications, but we will consider only such distortion functions. We do this for simplicity, tractability, and to gain insight into what can be accomplished in general. We further assume that $d(\cdot)$ is upper-bounded by some number $d_{\max}$.

The *rate distortion* (RD) problem is the following: find the set of pairs (R,D) that one can approach with source encoders for sufficiently large n (see [55, Part V], [57]). Note that we ignore the practical difficulties associated with large block lengths. However, the theory developed below provides useful bounds on the distortion achieved by *finite* length codes as well. The smallest rate R as a function of the distortion D is called the *rate distortion function*. The smallest D as a function of R is called the *distortion rate function*.

2.2 An Achievable RD Region

We present a random code construction in this section, and analyze the set of (R,D) that it can achieve. Suppose we choose a "channel" $P_{\hat{X}|X}(\cdot)$ and compute $P_{\hat{X}}(\cdot)$ as the marginal distribution of $P_{X\hat{X}}(\cdot)$.

Code Construction: Generate 2^{nR} codewords $\hat{x}^n(w)$, $w = 1,2,\ldots,2^{nR}$, by choosing each of the $n \cdot 2^{nR}$ symbols $\hat{x}_i(w)$ in the code book independently at random using $P_{\hat{X}}(\cdot)$ (see Figure 2.2).

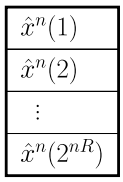

Fig. 2.2 A code book for the RD problem.

Encoder: Given x^n, try to find a codeword $\hat{x}^n(w)$ such that $(x^n, \hat{x}^n(w)) \in T_\epsilon^n(P_{X\hat{X}})$. If one is successful, send the corresponding index w. If one is unsuccessful, send $w = 1$.
(Note: the design of the code book has so far ignored the distortion function $d(\cdot)$. The design will include $d(\cdot)$ once we optimize over the choice of $P_{\hat{X}|X}(\cdot|\cdot)$.)

Decoder: Put out the reconstruction $\hat{x}^n(w)$.

Analysis: We bound $E\big[d^n(X^n, \hat{X}^n)\big]$ as follows: we partition the sample space into three disjoint events

$$\mathcal{E}_1 = \big\{X^n \notin T_{\epsilon_1}^n(P_X)\big\} \tag{2.3}$$

$$\mathcal{E}_2 = \mathcal{E}_1^c \bigcap \left\{ \bigcap_{w=1}^{2^{nR}} \Big\{(X^n, \hat{X}^n(w)) \notin T_\epsilon(P_{X\hat{X}})\Big\} \right\} \tag{2.4}$$

$$\mathcal{E}_3 = (\mathcal{E}_1 \cup \mathcal{E}_2)^c, \tag{2.5}$$

where $\mathcal{E}_1^c$ is the complement of $\mathcal{E}_1$. Next, we apply the Theorem on Total Expectation (see Section A.3)

$$E\big[d^n(X^n, \hat{X}^n)\big] = \sum_{i=1}^{3} \Pr[\mathcal{E}_i]\, E\big[d^n(X^n, \hat{X}^n)|\mathcal{E}_i\big]. \tag{2.6}$$

Let $0 < \epsilon_1 < \epsilon \leq \mu_{X\hat{X}}$, where we recall from Section 1.5 that $\mu_{X\hat{X}} = \min_{(a,b)\in \mathrm{supp}(P_{X\hat{X}})} P_{X\hat{X}}(a,b)$.

(1) Suppose that $X^n \notin T_{\epsilon_1}^n(P_X)$, in which case we upper bound the average distortion by $d_{\max}$. But recall that $\Pr\big[X^n \notin T_{\epsilon_1}^n(P_X)\big] \leq \delta_{\epsilon_1}(n)$, and $\delta_{\epsilon_1}(n)$ approaches zero exponentially in n if $\epsilon_1 > 0$.

(2) Suppose that $X^n = x^n$ and $x^n \in T^n_{\epsilon_1}(P_X)$ but none of the $\hat{X}^n(w)$ satisfies

$$(x^n, \hat{X}^n(w)) \in T^n_\epsilon(P_{X\hat{X}}). \tag{2.7}$$

We again upper bound the average distortion by $d_{\max}$. The events (2.7), $w = 1, 2, \ldots, 2^{nR}$, are independent since each $\hat{x}^n(w)$ was generated without considering x^n or the other codewords. The probability $P_e(x^n)$ that none of the codewords are satisfactory is thus

$$\begin{aligned} P_e(x^n) &= \Pr\left[\bigcap_{w=1}^{2^{nR}} \left\{(x^n, \hat{X}^n(w)) \notin T_\epsilon(P_{X\hat{X}})\right\}\right] \\ &= \left[1 - \Pr\left[(x^n, \hat{X}^n) \in T^n_\epsilon(P_{X\hat{X}})\right]\right]^{2^{nR}} \\ &\le \left[1 - (1 - \delta_{\epsilon_1,\epsilon}(n))\, 2^{-n[I(X;\hat{X}) + 2\epsilon H(\hat{X})]}\right]^{2^{nR}} \\ &\le \exp\left(-(1 - \delta_{\epsilon_1,\epsilon}(n))\, 2^{n[R - I(X;\hat{X}) - 2\epsilon H(\hat{X})]}\right), \end{aligned} \tag{2.8}$$

where the first inequality follows by Theorem 1.3, and the second inequality by $(1-x)^m \le e^{-mx}$. Inequality (2.8) implies that we can choose large n and

$$R > I(X;\hat{X}) + 2\epsilon H(\hat{X}) \tag{2.9}$$

to drive the error probability to zero. In addition, observe that the bound is valid for *any* x^n in $T^n_{\epsilon_1}(P_X)$, and the error probability decreases *doubly* exponentially in n. Denote the resulting error probability as $\delta_{\epsilon_1,\epsilon}(n, R)$.

(3) Suppose $X^n = x^n$, $x^n \in T^n_{\epsilon_1}(P_X)$, and we find a $\hat{x}^n(w)$ with $(x^n, \hat{x}^n(w)) \in T^n_\epsilon(P_{X\hat{X}})$. The distortion is

$$\begin{aligned} d^n(x^n, \hat{x}^n(w)) &= \frac{1}{n}\sum_{i=1}^n d(x_i, \hat{x}_i(w)) \\ &= \frac{1}{n}\sum_{a,b} N(a,b|x^n, \hat{x}^n(w))\, d(a,b) \end{aligned}$$

$$\begin{aligned} &\leq \sum_{a,b} P_{X\hat{X}}(a,b)(1+\epsilon)\, d(a,b) \\ &\leq E\big[d(X,\hat{X})\big] + \epsilon\, d_{\max}, \end{aligned} \tag{2.10}$$

where the first inequality follows by the definition (1.21).

Combining the above results using (2.6), we have

$$E\big[d^n(X^n,\hat{X}^n)\big] \leq E\big[d(X,\hat{X})\big] + (\delta_{\epsilon_1}(n) + \delta_{\epsilon_1,\epsilon}(n,R) + \epsilon)\, d_{\max}. \tag{2.11}$$

As a final step, we choose small ϵ, large n, R satisfying (2.9), and $P_{X\hat{X}}$ for which $E\big[d(X,\hat{X})\big] < D$. A random code thus *achieves* the rates R satisfying

$$R > \min_{P_{\hat{X}|X}:\ E\left[d(X,\hat{X})\right]<D} I(X;\hat{X}). \tag{2.12}$$

Alternatively, we say that a random code *approaches* the rate

$$R(D) = \min_{P_{\hat{X}|X}:\ E\left[d(X,\hat{X})\right]\leq D} I(X;\hat{X}). \tag{2.13}$$

The words *achieves* and *approaches* are often used interchangeably both here and in the literature.

We remark that there is a subtlety in the above argument: the expectation $E\left[d^n(X^n,\hat{X}^n)\right]$ is performed over both the source sequence and the code book. The reader might therefore wonder whether there is *one particular* code book for which the average distortion is D if the average distortion over all code books is D. A simple argument shows that this is the case: partition the sample space into the set of possible code books, and the Theorem on Total Expectation tells us that at least one of the codebooks must have a distortion at most the average.

2.3 Discrete Alphabet Examples

As an example, consider the binary symmetric source (BSS) with the Hamming distortion function and desired average distortion D, where

$D \leq 1/2$. We then require $\Pr\big[X \neq \hat{X}\big] \leq D$, and can bound

$$\begin{aligned} I(X;\hat{X}) &= H(X) - H(X|\hat{X}) \\ &= 1 - H(X \oplus \hat{X}|\hat{X}) \\ &\geq 1 - H(X \oplus \hat{X}) \\ &\geq 1 - H_2(D), \end{aligned} \tag{2.14}$$

where the last step follows because $E = X \oplus \hat{X}$ is binary with $P_E(1) \leq D$, and we recall that $H_2(x) = -x\log_2(x) - (1-x)\log_2(1-x)$ is the binary entropy function. Furthermore, we can "achieve" $R(D) = 1 - H_2(D)$ by choosing $P_{\hat{X}|X}(\cdot)$ to be the binary symmetric channel (BSC) with crossover probability D.

As a second example, consider again the BSS but with $\hat{\mathcal{X}} = \{0, 1, \Delta\}$, where Δ represents an erasure, and where we use the *erasure* distortion function

$$d(x,\hat{x}) = \begin{cases} 0, & \text{if } x = \hat{x} \\ 1, & \text{if } \hat{x} = \Delta \\ \infty, & \text{if } \hat{x} = x \oplus 1. \end{cases} \tag{2.15}$$

(Note that we are permitting an unbounded distortion; this causes no difficulties here.) To achieve finite distortion D, we must choose $P_{\hat{X}|X}(1|0) = P_{\hat{X}|X}(0|1) = 0$ and $\Pr\big[\hat{X} = \Delta\big] \leq D$. We thus have

$$\begin{aligned} I(X;\hat{X}) &= 1 - H(X|\hat{X}) \\ &= 1 - \sum_{b \in \hat{\mathcal{X}}} P_{\hat{X}}(b) H(X|\hat{X} = b) \\ &\geq 1 - D. \end{aligned} \tag{2.16}$$

We can achieve $R(D) = 1 - D$ by simply sending $w = x^{(1-D)n}$. The decoder puts out as its reconstruction $\hat{x}^n = [x^{(1-D)n}\ \Delta^{Dn}]$, where$\Delta^{\ m}$ is a string of m successiveΔs.

2.4 Gaussian Source and Mean Squared Error Distortion

Suppose that we can approach the rate (2.13) for the memoryless Gaussian source with mean squared error distortion (we will not prove this

here, see [18, Sec. 9]). We require $E\big[(X-\hat{X})^2\big] \le D$, and bound

$$\begin{aligned} I(X;\hat{X}) &= h(X) - h(X|\hat{X}) \\ &= \frac{1}{2}\log(2\pi e\sigma^2) - h(X-\hat{X}|\hat{X}) \\ &\ge \frac{1}{2}\log(2\pi e\sigma^2) - h(X-\hat{X}) \\ &\ge \frac{1}{2}\log(2\pi e\sigma^2) - \frac{1}{2}\log\big(2\pi e\, E[(X-\hat{X})^2]\big) \\ &\ge \frac{1}{2}\log(2\pi e\sigma^2) - \frac{1}{2}\log(2\pi e D) \\ &= \frac{1}{2}\log(\sigma^2/D), \end{aligned} \tag{2.17}$$

where σ^2 is the source variance, and where the second inequality follows by the maximum entropy theorem (see Section B.5.3 and [18, p. 234]). We can achieve $R(D) = \frac{1}{2}\log(\sigma^2/D)$ by choosing $P_{X|\hat{X}}(\cdot)$ (note that this is not $P_{\hat{X}|X}(\cdot)$) to be the additive white Gaussian noise (AWGN) channel with noise variance D. Alternatively, we can achieve the distortion $D(R) = \sigma^2\exp(-2R)$, i.e., we can gain 6 dB per quantization bit.

2.5 Two Properties of $R(D)$

We develop two properties of the function $R(D)$ in (2.13). First, it is clear that $R(D)$ is a non-increasing function with D because the set of $P_{\hat{X}|X}(\cdot)$ does not shrink by increasing D. Second, we prove that $R(D)$ is *convex* in D [57], [18, Lemma 13.4.1 on p. 349].

Consider two distinct points (R_1, D_1) and (R_2, D_2) on the boundary of $R(D)$, and suppose the channels $P_{\hat{X}_1|X}(\cdot)$ and $P_{\hat{X}_2|X}(\cdot)$ achieve these respective points. Consider also the distribution defined by

$$P_{\hat{X}_3|X}(a|b) = \lambda P_{\hat{X}_1|X}(a|b) + (1-\lambda) P_{\hat{X}_2|X}(a|b) \tag{2.18}$$

for all a, b, where $0 \le \lambda \le 1$. The distortion with $P_{\hat{X}_3|X}$ is simply $D_3 = \lambda D_1 + (1-\lambda) D_1$. The new mutual information, however, is less than the convex combination of mutual informations, i.e., we have (see

Section A.11)

$$I(X;\hat{X}_3) \le \lambda I(X;\hat{X}_1) + (1-\lambda) I(X;\hat{X}_2) \tag{2.19}$$

as follows by the convexity of $I(X;Y)$ in $P_{Y|X}(\cdot)$ when $P_X(\cdot)$ is held fixed [18, p. 31]. We thus have

$$\begin{aligned} R(\lambda D_1 + (1-\lambda) D_2) &= R(D_3) \\ &\le I(X;\hat{X}_3) \\ &\le \lambda I(X;\hat{X}_1) + (1-\lambda) I(X;\hat{X}_2) \\ &= \lambda R(D_1) + (1-\lambda) R(D_2). \end{aligned} \tag{2.20}$$

Thus, $R(D)$ is a convex function of D.

2.6 A Lower Bound on the Rate given the Distortion

We show that $R(D)$ in (2.13) is the rate distortion function. Thus, the random coding scheme described in Section 2.2 is rate-optimal given D.

Suppose we are using some encoder and decoder for which $E\left[d^n(X^n,\hat{X}^n)\right] \le D$. Recall that the code book has 2^{nR} sequences $\hat{x}^n$, and that $\hat{x}^n$ is a function of x^n. We thus have

$$\begin{aligned} nR &\ge H(\hat{X}^n) \\ &= H(\hat{X}^n) - H(\hat{X}^n|X^n) \\ &= I(X^n;\hat{X}^n) \\ &= H(X^n) - H(X^n|\hat{X}^n) \\ &= \sum_{i=1}^{n} H(X_i) - H(X_i|\hat{X}^n X^{i-1}) \\ &\ge \sum_{i=1}^{n} H(X_i) - H(X_i|\hat{X}_i) \\ &= \sum_{i=1}^{n} I(X_i;\hat{X}_i). \end{aligned} \tag{2.21}$$

We use (2.13) and the convexity (2.20) to continue the chain of inequalities (2.21):

$$\begin{aligned} nR &\geq \sum_{i=1}^{n} R\Big(E\big[d(X_i,\hat{X}_i)\big]\Big) \\ &\geq nR\left(\frac{1}{n}\sum_{i=1}^{n} E\big[d(X_i,\hat{X}_i)\big]\right) \\ &= nR\Big(E\big[d^n(X^n,\hat{X}^n)\big]\Big) \\ &\geq nR(D). \end{aligned} \tag{2.22}$$

Thus, the rate must be larger than $R(D)$, and this is called a *converse* result. But we can also achieve $R(D)$ by (2.13), so the rate distortion function is $R(D)$.

2.7 The Multiple Description Problem

A generalization of the RD problem is depicted in Figure 2.3, and is known as the *multiple-description* (MD) problem. A DMS again puts out a sequence of symbols x^n, but now the source encoder has *two or more* channels through which to send indexes $W_1, W_2, \ldots, W_L$ (also called "descriptions" of x^n). We will concentrate on two channels only, since the following discussion can be extended in a straightforward way to more than two channels. For two channels, the encoder might quantize x^n to one of 2^{nR_1} sequences $\hat{x}_1^n(w_1)$, $w_1 = 1, 2, \ldots, 2^{nR_1}$, and to one of 2^{nR_2} sequences $\hat{x}_2^n(w_2)$, $w_2 = 1, 2, \ldots, 2^{nR_2}$. The indexes w_1 and w_2 are sent over the respective channels 1 and 2. As another possibility, the encoder might quantize x^n to one of $2^{n(R_1+R_2)}$ sequences $\hat{x}_{12}^n(w_1, w_2)$

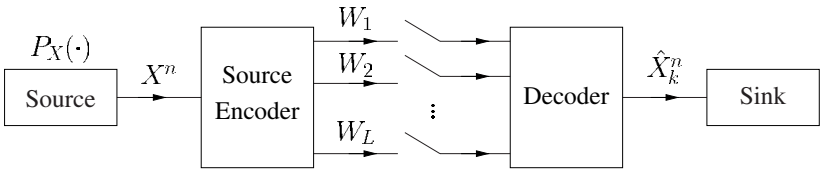

Fig. 2.3 The multiple description problem.

and send w_1 and w_2 over the respective channels 1 and 2. There are, in fact, many other strategies that one can employ.

Suppose both w_1 and w_2 are always received by the decoder. We then simply have the RD problem. The MD problem becomes different than the RD problem by modeling the individual channels through which W_1 and W_2 pass as being either noise-free or completely noisy, i.e., the decoder either receives W_1 (or W_2) over channel 1 (or channel 2), or it does not. This scenario models the case where a source sequence, e.g., an audio or video file, is sent across a network in several packets. These packets may or may not be received, or may be received with errors, in which case the decoder discards the packet.

The decoder encounters one of three interesting situations: either W_1 or W_2 is received, or both are received. There are, therefore, three interesting average distortions:

$$D_1 = \frac{1}{n}\sum_{i=1}^{n} E\big[d_1(X_i, \hat{X}_{1i}(W_1))\big] \tag{2.23}$$

$$D_2 = \frac{1}{n}\sum_{i=1}^{n} E\big[d_2(X_i, \hat{X}_{2i}(W_2))\big] \tag{2.24}$$

$$D_{12} = \frac{1}{n}\sum_{i=1}^{n} E\big[d_{12}(X_i, \hat{X}_{(12)i}(W_1, W_2)\big], \tag{2.25}$$

where $\hat{X}_k^n$, $k = 1, 2, 12$, is the reconstruction of X^n when only W_1 is received ($k = 1$), only W_2 is received ($k = 2$), and both W_1 and W_2 are received ($k = 12$). Observe that the distortion functions might depend on k.

The source encoder usually does not know ahead of time which W_ℓ will be received. The MD problem is, therefore, determining the set of 5-tuples $(R_1, R_2, D_1, D_2, D_{12})$ that can be approached with source encoders for any length n (see [22]).

2.8 A Random Code for the MD Problem

We present a random code construction that generalizes the scheme of Section 2.2. We choose a $P_{\hat{X}_1\hat{X}_2\hat{X}_{12}|X}(\cdot)$ and compute $P_{\hat{X}_1}(\cdot)$, $P_{\hat{X}_2}(\cdot)$, and $P_{\hat{X}_{12}|\hat{X}_1\hat{X}_2}(\cdot)$ as marginal distributions of $P_{X\hat{X}_1\hat{X}_2\hat{X}_{12}}(\cdot)$.

Code Construction: Generate 2^{nR_1} codewords $\hat{x}_1^n(w_1)$, $w_1 = 1,2,\ldots,2^{nR_1}$, by choosing each of the $n \cdot 2^{nR_1}$ symbols $\hat{x}_{1i}(w_1)$ at random according to $P_{\hat{X}_1}(\cdot)$. Similarly, generate 2^{nR_2} codewords $\hat{x}_2^n(w_2)$, $w_2 = 1,2,\ldots,2^{nR_2}$, by using $P_{\hat{X}_2}(\cdot)$. Finally, for each pair (w_1,w_2), generate *one* codeword $\hat{x}_{12}^n(w_1,w_2)$ by choosing its ith symbol at random according to $P_{\hat{X}_{12}|\hat{X}_1\hat{X}_2}(\cdot|\hat{x}_{1i},\hat{x}_{2i})$.

Encoder: Given x^n, try to find a triple $(\hat{x}_1^n(w_1),\hat{x}_2^n(w_2),\hat{x}_{12}^n(w_1,w_2))$ such that

$$(x^n,\hat{x}_1^n(w_1),\hat{x}_2^n(w_2),\hat{x}_{12}^n(w_1,w_2)) \in T_\epsilon^n(P_{X\hat{X}_1\hat{X}_2\hat{X}_{12}}).$$

If one finds such a codeword, send w_1 across the first channel and w_2 across the second channel. If one is unsuccessful, send $w_1 = w_2 = 1$.

Decoder: Put out $\hat{x}_1^n(w_1)$ if only w_1 is received. Put out $\hat{x}_2^n(w_2)$ if only w_2 is received. Put out $\hat{x}_{12}^n(w_1,w_2)$ if both w_1 and w_2 are received.

Analysis: One can again partition the sample space as in Section 2.2. There is one new difficulty: one cannot claim that the triples $(\hat{x}_1^n(w_1'),\hat{x}_2^n(w_2'),\hat{x}_{12}^n(w_1',w_2'))$ and $(\hat{x}_1^n(w_1),\hat{x}_2^n(w_2),\hat{x}_{12}^n(w_1,w_2))$ are independent if $(w_1',w_2') \neq (w_1,w_2)$. The reason is that one might encounter $w_1' = w_1$, $w_2' \neq w_2$ or $w_2' = w_2$, $w_1' \neq w_1$. We refer to Section 7.10 and [64] for one approach for dealing with this problem. The resulting MD region is the set of (R_1,R_2,D_1,D_2,D_{12}) satisfying

$$\begin{aligned} R_1 &\geq I(X;\hat{X}_1) \\ R_2 &\geq I(X;\hat{X}_2) \\ R_1 + R_2 &\geq I(X;\hat{X}_1\hat{X}_2\hat{X}_{12}) + I(\hat{X}_1;\hat{X}_2) \\ D_k &\geq E\big[d_k(X;\hat{X}_k)\big] \quad \text{for } k = 1,2,12, \end{aligned} \tag{2.26}$$

where $P_{\hat{X}_1\hat{X}_2\hat{X}_{12}|X}(\cdot)$ is arbitrary. This region was shown to be achievable by El Gamal and Cover in [22]. The current best region for two channels is due to Zhang and Berger [75].

As an example, consider again the BSS and the erasure distortion function (2.15). An outer bound on the MD region is

$$\begin{aligned} R_1 &\geq I(X;\hat{X}_1) \geq 1 - D_1 \\ R_2 &\geq I(X;\hat{X}_2) \geq 1 - D_2 \\ R_1 + R_2 &\geq I(X;\hat{X}_1\hat{X}_2\hat{X}_{12}) \geq 1 - D_{12}, \end{aligned} \tag{2.27}$$

which can be derived from the RD function, and the same steps as in (2.16). But for any D_1, D_2, and D_{12}, we can achieve all rates and distortions in (2.27) as follows. If $1 - D_{12} \leq (1 - D_1) + (1 - D_2)$, send $w_1 = x^{(1-D_1)n}$ and $w_2 = x_i^j = [x_i, x_{i+1}, \ldots, x_j]$, where $i = (1 - D_1)\, n + 1$ and $j = (1 - D_1)n + (1 - D_2)n$. If $1 - D_{12} > (1 - D_1) + (1 - D_2)$, choose one of two strategies to achieve the two corner points of (2.27). The first strategy is to send $w_1 = x^{(1-D_1)n}$ and $w_2 = x_i^j$, where $i = (1 - D_1)n + 1$ and $j = (1 - D_{12})n$. For the second strategy, swap the indexes 1 and 2 of the first strategy. One can achieve any point inside (2.27) by time-sharing these two strategies.

Finally, we remark that the MD problem is still open, even for only two channels! Fortunately, the entire MD region is known for the Gaussian source and squared error distortion [47]. But even for this important source and distortion function the problem is still open for more than two channels [50, 51, 64, 65].

3

Capacity–Cost

3.1 Problem Description

The discrete memoryless channel (DMC) is the basic model for channel coding, and it is depicted in Figure 3.1. A source sends a message w, $w \in \{1, 2, \ldots, 2^{nR}\}$, to a receiver by mapping it into a sequence x^n in $\mathcal{X}^n$. We assume that the messages are equiprobable for now. The channel $P_{Y|X}(\cdot)$ puts out y^n, $y^n \in \mathcal{Y}^n$, and the decoder maps y^n to its estimate $\hat{w}$ of w. The goal is to find the maximum rate R for which one can make $P_e = \Pr[\hat{W} \neq W]$ arbitrarily close to zero (but not necessarily exactly zero). This maximum rate is called the capacity C.

We refine the problem by adding a *cost constraint*. Suppose that transmitting the sequence x^n and receiving the sequence y^n incurs a cost of $s^n(x^n, y^n)$ units. In a way reminiscent of the rate-distortion problem, we require the *average* cost $E[s^n(X^n, Y^n)]$ to be at most some specified value S. We further consider only real-valued cost functions $s^n(\cdot)$ that are averages of a per-letter cost function $s(\cdot)$:

$$s^n(x^n, y^n) = \frac{1}{n} \sum_{i=1}^{n} s(x_i, y_i). \tag{3.1}$$

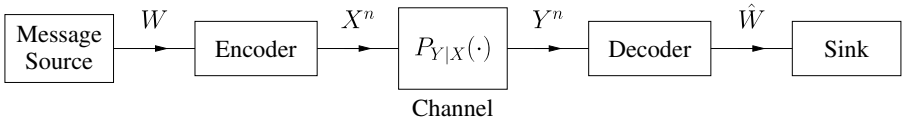

Fig. 3.1 The capacity–cost problem.

We further assume that $s(\cdot)$ is upper-bounded by some number $s_{\max}$. The largest rate C as a function of the cost S is called the *capacity cost* function, and is denoted $C(S)$.

As an example, suppose we are transmitting data over an optical channel with binary $(0,1)$ inputs and outputs, and where the transmission of a 1 costs $s(1,y) = E$ units of energy for any y, while transmitting a 0 costs $s(0,y) = 0$ units of energy for any y. A cost constraint with $0 \le S < E$ will bias the best transmission scheme toward sending the symbol 1 less often.

3.2 Data Processing Inequalities

Suppose $X - Y - Z$ forms a Markov chain, i.e., we have $I(X;Z|Y) = 0$. Then the following *data processing inequalities* are valid and are proved in the appendix of this section:

$$I(X;Z) \le I(X;Y) \quad \text{and} \quad I(X;Z) \le I(Y;Z). \tag{3.2}$$

Second, suppose Y_1 and Y_2 are the respective outputs of a channel $P_{Y|X}(\cdot)$ with inputs X_1 and X_2. In the appendix of this section, we show that

$$D(P_{Y_1} \| P_{Y_2}) \le D(P_{X_1} \| P_{X_2}). \tag{3.3}$$

3.3 Applications of Fano's Inequality

Suppose that we have a message W with $H(W) = nR$ so that we can represent W by a string of nR bits $V_1, V_2, \ldots, V_{nR}$ (as usual, for simplicity we assume that nR is an integer). Consider any channel coding problem where W (or V^{nR}) is to be transmitted to a sink, and is estimated as $\hat{W}$ (or $\hat{V}^{nR}$). We wish to determine properties of, and relations

between, the *block* error probability

$$P_e = \Pr[\hat{W} \neq W] \tag{3.4}$$

and the *average bit* error probability

$$P_{\rm b} = \frac{1}{nR}\sum_{i=1}^{nR} \Pr[\hat{V}_i \neq V_i]. \tag{3.5}$$

We begin with P_e. Using Fano's inequality (see Section A.10) we have

$$H_2(P_e) + P_e \log_2(|\mathcal{W}| - 1) \geq H(W|\hat{W}), \tag{3.6}$$

where the alphabet size $|\mathcal{W}|$ can be assumed to be at most 2^{nR} because V^{nR} represents W. We thus have

$$H_2(P_e) + P_e nR \geq H(W) - I(W;\hat{W}) \tag{3.7}$$

and, using $H(W) = nR$, we have

$$nR \leq \frac{I(W;\hat{W}) + H_2(P_e)}{1 - P_e}. \tag{3.8}$$

This simple bound shows that we require $nR \leq I(W;\hat{W})$ if P_e is to be made small. Of course, (3.8) is valid for any choice of P_e.

Consider next $P_{\rm b}$ for which we bound

$$\begin{aligned} H_2(P_{\rm b}) &= H_2\left(\frac{1}{nR}\sum_{i=1}^{nR} \Pr[\hat{V}_i \neq V_i]\right) \\ &\geq \frac{1}{nR}\sum_{i=1}^{nR} H_2(\Pr[\hat{V}_i \neq V_i]) \\ &\geq \frac{1}{nR}\sum_{i=1}^{nR} H(V_i|\hat{V}_i), \end{aligned} \tag{3.9}$$

where the second step follows by the concavity of $H_2(\cdot)$, and the third step by Fano's inequality. We continue the chain of inequalities as

$$\begin{aligned} H_2(P_{\mathrm{b}}) &\geq \frac{1}{nR}\sum_{i=1}^{nR} H(V_i|V^{i-1}\hat{V}^{nR}) \\ &= \frac{1}{nR}H(V^{nR}|\hat{V}^{nR}) \\ &= \frac{1}{nR}\big(H(V^{nR}) - I(V^{nR};\hat{V}^{nR})\big) \\ &= 1 - \frac{I(W;\hat{W})}{nR}. \end{aligned} \tag{3.10}$$

Alternatively, we have the following counterpart to (3.8):

$$nR \leq \frac{I(W;\hat{W})}{1 - H_2(P_{\mathrm{b}})}. \tag{3.11}$$

We thus require $nR \leq I(W;\hat{W})$ if P_{b} is to be made small. We further have the following relation between P_{b} and the average *block* error probability P_{e}:

$$P_{\mathrm{b}} \leq P_e \leq n\,P_{\mathrm{b}}. \tag{3.12}$$

Thus, if P_{b} is lower bounded, so is P_e. Similarly, if P_e is small, so is P_{b}. This is why *achievable* coding theorems should upper bound P_e, while *converse* theorems should lower bound P_{b}. For example, a code that has large P_e might have very small P_{b}.

3.4 An Achievable Rate

We construct a random code book for the DMC with cost constraint S. We begin by choosing a distribution $P_X(\cdot)$.

Code Construction: Generate 2^{nR} codewords $x^n(w)$, $w = 1, 2, \ldots, 2^{nR}$, by choosing the $n \cdot 2^{nR}$ symbols $x_i(w)$, $i = 1, 2, \ldots, n$, independently using $P_X(\cdot)$.

Encoder: Given w, transmit $x^n(w)$.

Decoder: Given y^n, try to find a $\tilde{w}$ such that $(x^n(\tilde{w}), y^n) \in T_\epsilon^n(P_{XY})$. If there is one or more such $\tilde{w}$, then choose one as $\hat{w}$. If there is no such $\tilde{w}$, then put out $\hat{w} = 1$.

Analysis: We split the analysis into several parts and use the Theorem on Total Expectation as in Section 2.2. Let $0 < \epsilon_1 < \epsilon_2 < \epsilon \leq \mu_{XY}$.

(1) Suppose that $X^n(w) \notin T^n_{\epsilon_1}(P_X)$, in which case we upper bound the average cost by $s_{\max}$. Recall from Theorem 1.1 that $\Pr\left[X^n(w) \notin T^n_{\epsilon_1}(P_X)\right] \leq \delta_{\epsilon_1}(n)$, and $\delta_{\epsilon_1}(n)$ approaches zero exponentially in n if $\epsilon_1 > 0$.

(2) Suppose that $X^n(w) = x^n(w)$ and $x^n(w) \in T^n_{\epsilon_1}(P_X)$ but $(x^n(w), Y^n) \notin T^n_{\epsilon_2}(P_{XY})$. We again upper bound the average cost by $s_{\max}$. Using Theorem 1.2, the probability of this event is upper bounded by $\delta_{\epsilon_1,\epsilon_2}(n)$, and $\delta_{\epsilon_1,\epsilon_2}(n)$ approaches zero exponentially in n if $\epsilon_1 \geq 0$ and $\epsilon_2 > 0$.

(3) Suppose $(x^n(w), y^n) \in T^n_{\epsilon_2}(P_{XY})$, but that we also find a $\tilde{w} \neq w$ such that $(x^n(\tilde{w}), y^n) \in T^n_{\epsilon}(P_{XY})$. Using Theorem 1.3, the probability of this event is

$$\begin{aligned} P_e(w) &= \Pr\left[\bigcup_{\tilde{w}\neq w} \{(X^n(\tilde{w}), y^n) \in T_\epsilon(P_{XY})\}\right] \\ &\leq \sum_{\tilde{w}\neq w} \Pr\left[(X^n(\tilde{w}), y^n) \in T_\epsilon(P_{XY})\right] \\ &\leq (2^{nR} - 1)\, 2^{-n[I(X;Y)-2\epsilon H(X)]}, \end{aligned} \tag{3.13}$$

where the first inequality follows by the union bound (see (A.5)) and the second inequality follows by Theorem 1.3. Inequality (3.13) implies that we can choose large n and

$$R < I(X;Y) - 2\epsilon H(X) \tag{3.14}$$

to drive $P_e(w)$ to zero.

(4) Finally, we compute the average cost of transmission if $(x^n(w), y^n) \in T_\epsilon(P_{XY})$:

$$\begin{aligned} s^n(x^n(w), y^n) &= \frac{1}{n}\sum_{i=1}^{n} s(x_i(w), y_i) \\ &= \frac{1}{n}\sum_{a,b} N(a,b|x^n(w), y^n)\; s(a,b) \\ &\le \sum_{a,b} P_{XY}(a,b)(1+\epsilon)\; s(a,b) \\ &\le E[s(X,Y)] + \epsilon\, s_{\max}, \end{aligned} \tag{3.15}$$

where the first inequality follows by the definition (1.21).

Combining the above results, there is a code in the random ensemble of codes that approaches the rate

$$C(S) = \max_{P_X(\cdot):\; E[s(X,Y)] \le S} I(X;Y). \tag{3.16}$$

We will later show that (3.16) is the capacity–cost function. If there is no cost constraint, we achieve

$$C = \max_{P_X(\cdot)} I(X;Y). \tag{3.17}$$

3.5 Discrete Alphabet Examples

As an example, consider the binary symmetric channel (BSC) with $\mathcal{X} = \mathcal{Y} = \{0,1\}$ and $\Pr[Y \neq X] = p$. Suppose the costs $s(X)$ depend on X only and are $s(0) = 0$ and $s(1) = E$. We compute

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H_2(P_X(1) * p) - H_2(p) \qquad (3.18) \\ E[s(X)] &= P_X(1) \cdot E, \qquad (3.19) \end{aligned}$$

where $q * p = q(1-p) + (1-q)p$. The capacity cost function is thus

$$C(S) = H_2(\min(S/E, 1/2) * p) - H_2(p) \tag{3.20}$$

and for $S \ge E/2$ we have $C = 1 - H_2(p)$.

As a second example, consider the binary erasure channel (BEC) with $\mathcal{X} = \{0,1\}$ and $\mathcal{Y} = \{0,1,\Delta\}$, and where $\Pr[Y = X] = 1 - p$ and $\Pr[Y = \Delta] = p$. For no cost constraint, we compute

$$\begin{aligned} C &= \max_{P_X(\cdot)} H(X) - H(X|Y) \\ &= \max_{P_X(\cdot)} H(X)\,(1-p) \\ &= 1 - p. \end{aligned} \tag{3.21}$$

3.6 Gaussian Examples

Consider the additive white Gaussian noise (AWGN) channel with

$$Y = X + Z, \tag{3.22}$$

where Z is a zero-mean, variance N, Gaussian random variable that is statistically independent of X. We further choose the cost function $s(x) = x^2$ and $S = P$ for some P.

One can generalize the information theory for discrete alphabets to continuous channels in several ways. First, we could quantize the input and output alphabets into fine discrete alphabets and compute the resulting capacity. We could repeat this procedure using progressively finer and finer quantizations, and the capacity will increase and converge if it is bounded. Alternatively, we could use the theory of entropy-typical sequences (see Sections 1.2 and B.6) to develop a capacity theorem directly from the channel model.

Either way, the resulting capacity turns out to be precisely (3.16). We thus compute

$$\begin{aligned} C(P) &= \max_{P_X(\cdot):\, E[X^2] \le P} h(Y) - h(Y|X) \\ &= \max_{P_X(\cdot):\, E[X^2] \le P} h(Y) - \frac{1}{2}\log(2\pi e N) \\ &\le \frac{1}{2}\log(2\pi e(P+N)) - \frac{1}{2}\log(2\pi e N) \\ &= \frac{1}{2}\log(1 + P/N), \end{aligned} \tag{3.23}$$

where the inequality follows by the maximum entropy theorem. We achieve the $C(P)$ in (3.23) by choosing X to be a zero-mean, variance P, Gaussian random variable.

Next, consider the following channel with a vector output:

$$Y = [H\,X + Z,\ H], \tag{3.24}$$

where Z is Gaussian as before, and H is a random variable with density $p_H(\cdot)$ that is independent of X and Z. This problem models a *fading* channel where the receiver, but not the transmitter, knows the fading coefficient H. We choose the cost function $s(x) = x^2$ with $S = P$, and compute

$$\begin{aligned}
C(P) &= \max_{P_X(\cdot):\, E[X^2]\le P} I(X;[H\,X + Z,\ H]) \\
&= \max_{P_X(\cdot):\, E[X^2]\le P} I(X;H) + I(X;H\,X + Z|H) \\
&= \max_{P_X(\cdot):\, E[X^2]\le P} I(X;H\,X + Z|H) \\
&= \max_{P_X(\cdot):\, E[X^2]\le P} \int_a p_H(a)\, h(a\,X + Z)\, da - \frac{1}{2}\log(2\pi e N) \\
&\le \int_a p_H(a) \cdot \frac{1}{2}\log(1 + a^2 P/N)\, da,
\end{aligned} \tag{3.25}$$

where the last step follows by the maximum entropy theorem (see Appendix B.5.3). One can similarly compute $C(P)$ if H is discrete. For example, suppose H takes on one of the three values: $P_H(1/2) = 1/4$, $P_H(1) = 1/2$, and $P_H(2) = 1/4$. The capacity is then

$$C(P) = \frac{1}{8}\log\left(1 + \frac{P}{4N}\right) + \frac{1}{4}\log\left(1 + \frac{P}{N}\right) + \frac{1}{8}\log\left(1 + \frac{4P}{N}\right).$$

Finally, consider the channel with $n_t \times 1$ input $\underline{X}$, $n_r \times n_t$ fading matrix $\mathbf{H}$, $n_r \times 1$ output $\underline{Y}$, and

$$\underline{Y} = \mathbf{H}\,\underline{X} + \underline{Z}, \tag{3.26}$$

where $\underline{Z}$ is an $n_r \times 1$ Gaussian vector with i.i.d. entries of unit variance, and H is a fixed matrix. This problem is known as a vector (or multi-antenna, or multi-input, multi-output, or MIMO) AWGN

channel. We choose the cost function $s(x) = \|\underline{x}\|^2$ with $S = P$, and compute

$$\begin{aligned} C(P) &= \max_{P_{\underline{X}}(\cdot):\ E[\|\underline{X}\|^2] \le P} I(\underline{X}; \mathbf{H}\underline{X} + \underline{Z}) \\ &= \max_{P_{\underline{X}}(\cdot):\ E[\|\underline{X}\|^2] \le P} h(\mathbf{H}\underline{X} + \underline{Z}) - \frac{n_r}{2}\log(2\pi e) \\ &= \max_{\mathrm{tr}[\mathbf{Q}_{\underline{X}}] \le P} \frac{1}{2}\log\left|I + \mathbf{H}\mathbf{Q}_{\underline{X}}\mathbf{H}^T\right|, \end{aligned} \tag{3.27}$$

where the last step follows by the maximum entropy theorem (see Appendix B.5.3). But note that one can write $\mathbf{H} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, where $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices (with $\mathbf{U}\mathbf{U}^T = \mathbf{I}$ and $\mathbf{V}\mathbf{V}^T = \mathbf{I}$) and where $\mathbf{D}$ is a diagonal $n_r \times n_t$ matrix with the *singular values* of $\mathbf{H}$ on the diagonal. We can rewrite (3.27) as

$$\begin{aligned} C(P) &= \max_{\mathrm{tr}[\mathbf{Q}_{\underline{X}}] \le P} \frac{1}{2}\log\left|\mathbf{I} + \mathbf{D}\mathbf{Q}_{\underline{X}}\mathbf{D}^T\right| \\ &= \max_{\sum_{i=1}^{\min(n_t,n_r)} \lambda_i \le P} \sum_{i=1}^{\min(n_t,n_r)} \frac{1}{2}\log\left(1 + d_i^2\lambda_i\right), \end{aligned} \tag{3.28}$$

where the d_i, $i = 1, 2, \ldots, \min(n_t, n_r)$, are the singular values of $\mathbf{H}$, and where we have used Hadamard's inequality (see [18, p. 233]) for the second step. The remaining optimization problem is the same as that of parallel Gaussian channels with different gains. One can solve for the λ_i by using *waterfilling* [18, Sec. 10.4], and the result is (see [62, Sec. 3.1])

$$\lambda_i = \left(\mu - \frac{1}{d_i^2}\right)^+, \tag{3.29}$$

where $(x)^+ = \max(0, x)$ and μ is chosen so that

$$\sum_{i=1}^{\min(n_t,n_r)} \left(\mu - \frac{1}{d_i^2}\right)^+ = P. \tag{3.30}$$

3.7 Two Properties of $C(S)$

We develop two properties of $C(S)$ in (3.16). First, $C(S)$ is a non-decreasing function with S because the set of permissible $P_X(\cdot)$ does not

shrink by increasing S. Second, we show that $C(S)$ is a concave function of S. Consider two distinct points (C_1, S_1) and (C_2, S_2) on the boundary of $C(S)$, and suppose the distributions $P_{X_1}(\cdot)$ and $P_{X_2}(\cdot)$ achieve these respective points. Consider also the distribution defined by

$$P_{X_3}(a) = \lambda P_{X_1}(a) + (1 - \lambda) P_{X_2}(a), \tag{3.31}$$

for all a, where $0 \leq \lambda \leq 1$. The cost with $P_{X_3}(\cdot)$ is simply $S_3 = \lambda S_1 + (1 - \lambda) S_2$. The new mutual information, however, is larger than the convex combination of mutual informations, i.e., we have

$$I(X_3;Y) \geq \lambda I(X_1;Y) + (1 - \lambda) I(X_2;Y) \tag{3.32}$$

as follows by the concavity of $I(X;Y)$ in $P_X(\cdot)$ when $P_{Y|X}(\cdot)$ is fixed (see Section A.11 and [18, p. 31]). We thus have

$$\begin{aligned} C(\lambda S_1 + (1 - \lambda) S_2) &= C(S_3) \\ &\geq I(X_3;Y) \\ &\geq \lambda I(X_1;Y) + (1 - \lambda) I(X_2;Y) \\ &= \lambda C(S_1) + (1 - \lambda) C(S_2). \end{aligned} \tag{3.33}$$

3.8 Converse

We show that $C(S)$ in (3.16) is the capacity–cost function. We bound

$$\begin{aligned} I(W;\hat{W}) &\leq I(X^n;Y^n) \\ &= \sum_{i=1}^{n} H(Y_i|Y^{i-1}) - H(Y_i|X_i) \\ &\leq \sum_{i=1}^{n} H(Y_i) - H(Y_i|X_i) \\ &= \sum_{i=1}^{n} I(X_i;Y_i). \end{aligned} \tag{3.34}$$

We use (3.16) and the concavity (3.33) to continue the chain of inequalities (3.34):

$$\begin{aligned} I(W;\hat{W}) &\le \sum_{i=1}^{n} C\left(E[s(X_i,Y_i)]\right) \\ &\le nC\left(\frac{1}{n}\sum_{i=1}^{n} E[s(X_i,Y_i)]\right) \\ &= nC(E[s^n(X^n,Y^n)]) \\ &\le nC(S), \end{aligned} \tag{3.35}$$

where the last step follows because we require $E[s^n(X^n,Y^n)] \le S$ and because $C(S)$ is non-decreasing. Inserting (3.35) into (3.8) and (3.11) we have

$$R \le \frac{C(S) + H_2(P_e)/n}{1 - P_e}, \tag{3.36}$$

and

$$R \le \frac{C(S)}{1 - H_2(P_{\mathrm{b}})}. \tag{3.37}$$

Thus, we find that R can be at most $C(S)$ for reliable communication and $E[s^n(X^n,Y^n)] \le S$.

3.9 Feedback

Suppose we have a DMC *with feedback* in the sense that X_i can be a function of the message W and some noisy function of the *past* channel outputs Y^{i-1}. One might expect that feedback can increase the capacity of the channel. To check this, we study the best type of feedback: suppose Y^{i-1} is passed through a noise-free channel as shown in Figure 3.2. We slightly modify (3.34) and bound

$$\begin{aligned} I(W;\hat{W}) &\le I(W;Y^n) \\ &= \sum_{i=1}^{n} H(Y_i|Y^{i-1}) - H(Y_i|WY^{i-1}) \end{aligned}$$

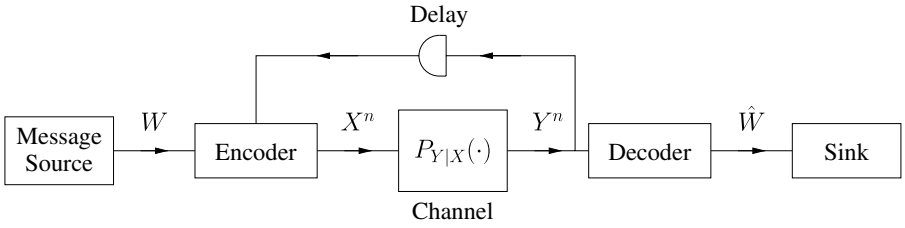

Fig. 3.2 The capacity–cost problem with feedback.

$$
\begin{aligned}
&= \sum_{i=1}^{n} H(Y_i|Y^{i-1}) - H(Y_i|WY^{i-1}X^i) \\
&= \sum_{i=1}^{n} H(Y_i|Y^{i-1}) - H(Y_i|Y^{i-1}X^i) \\
&= \sum_{i=1}^{n} I(X^i;Y_i|Y^{i-1}), \qquad (3.38)
\end{aligned}
$$

where the third step follows because X^i is a function of W and Y^{i-1}, and the fourth step because the channel is memoryless. The last quantity in (3.38) is known as the *directed information* flowing from X^n to Y^n and is written as $I(X^n \to Y^n)$ (see [45]). The directed information is the "right" quantity to study for many types of channels including multi-user channels (see [37]).

Continuing with (3.38), we have

$$
\begin{aligned}
I(X^n \to Y^n) &= \sum_{i=1}^{n} H(Y_i|Y^{i-1}) - H(Y_i|Y^{i-1}X^i) \\
&= \sum_{i=1}^{n} H(Y_i|Y^{i-1}) - H(Y_i|X_i) \\
&\le \sum_{i=1}^{n} H(Y_i) - H(Y_i|X_i) \\
&= \sum_{i=1}^{n} I(X_i;Y_i), \qquad (3.39)
\end{aligned}
$$

where the second step follows because the channel is memoryless. We have thus arrived at (3.34) and find the surprising result that feedback does *not* improve the capacity–cost function of a discrete memoryless channel [20, 56].

3.10 Appendix: Data Processing Inequalities

Proof. We prove the data processing inequalities. We have

$$\begin{aligned} I(X;Z) &= H(X) - H(X|Z) \\ &\le H(X) - H(X|ZY) \\ &= H(X) - H(X|Y) \\ &= I(X;Y). \end{aligned} \tag{3.40}$$

One can prove $I(X;Z) \le I(Y;Z)$ in the same way. Next, by the log-sum inequality (A.78) we have

$$\begin{aligned} D(P_{Y_1}\|P_{Y_2}) &= \sum_y P_{Y_1}(y) \log \frac{P_{Y_1}(y)}{P_{Y_2}(y)} \\ &= \sum_y \left(\sum_x P_{X_1}(x) P_{Y|X}(y|x) \right) \log \frac{\sum_x P_{X_1}(x) P_{Y|X}(y|x)}{\sum_x P_{X_2}(x) P_{Y|X}(y|x)} \\ &\le \sum_y \sum_x P_{X_1}(x) P_{Y|X}(y|x) \log \frac{P_{X_1}(x) P_{Y|X}(y|x)}{P_{X_2}(x) P_{Y|X}(y|x)} \\ &= \sum_x P_{X_1}(x) \left(\sum_y P_{Y|X}(y|x) \right) \log \frac{P_{X_1}(x)}{P_{X_2}(x)} \\ &= D(P_{X_1}\|P_{X_2}). \end{aligned} \tag{3.41}$$

□

4

The Slepian–Wolf Problem, or Distributed Source Coding

4.1 Problem Description

The distributed source coding problem is the first *multi*-terminal problem we consider, in the sense that there is more than one encoder or decoder. Suppose a DMS $P_{XY}(\cdot)$ with alphabet $\mathcal{X} \times \mathcal{Y}$ emits *two* sequences x^n and y^n, where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$ for all i (see Figure 4.1). There are two encoders: one encoder maps x^n into one of 2^{nR_1} indexes w_1, and the other encoder maps y^n into one of 2^{nR_2} indexes w_2. A decoder receives both w_1 and w_2 and produces the sequences $\hat{x}^n(w_1, w_2)$ and $\hat{y}^n(w_1, w_2)$, where $\hat{x}_i \in \mathcal{X}$ and $\hat{y}_i \in \mathcal{Y}$ for all i. The problem is to find the set of rate pairs (R_1, R_2) for which one can, for sufficiently large n, design encoders and a decoder so that the error probability

$$P_e = \Pr\big[(\hat{X}^n, \hat{Y}^n) \neq (X^n, Y^n)\big] \tag{4.1}$$

can be made an arbitrarily small positive number.

This type of problem might be a simple model for a scenario involving two *sensors* that observe dependent measurement streams X^n and Y^n, and that must send these to a "fusion center." The sensors usually have limited energy to transmit their data, so they are interested in communicating both *efficiently* and *reliably*. For example, an

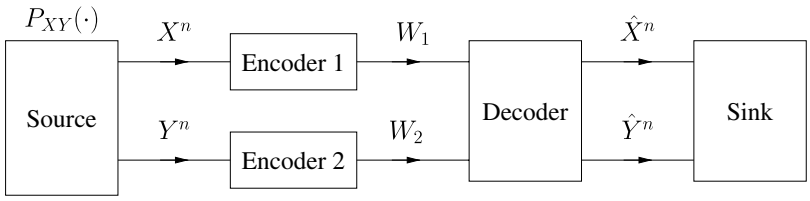

Fig. 4.1 A distributed source coding problem.

obvious strategy is for both encoders to compress their streams to entropy so that one achieves $(R_1, R_2) \approx (H(X), H(Y))$. On the other hand, an obvious *outer* bound on the set of achievable rate-pairs is $R_1 + R_2 \geq H(XY)$, since this is the smallest possible sum-rate if both encoders cooperate.

The problem of Figure 4.1 was solved by Slepian and Wolf in an important paper in 1973 [60]. They found the rather surprising result that the sum-rate $R_1 + R_2 = H(XY)$ is, in fact, approachable! Moreover, their encoding technique involves a simple and effective trick similar to hashing, and this trick has since been applied to many other communication problems. The Slepian–Wolf encoding scheme can be generalized to ergodic sources [14], and is now widely known as partitioning or *binning*.

4.2 Preliminaries

Recall (see Theorem 1.3) that for $0 \leq \epsilon_1 < \epsilon \leq \mu_{XY}$, $x^n \in T^n_{\epsilon_1}(P_X)$, and Y^n emitted from a DMS $P_Y(\cdot)$, we have

$$\Pr[(x^n, Y^n) \in T^n_\epsilon(P_{XY})] \leq 2^{-n[I(X;Y)-2\epsilon H(Y)]}. \tag{4.2}$$

It is somewhat easier to prove a random version of (4.2) rather than a conditional one. That is, if X^n and Y^n are output by the respective $P_X(\cdot)$ and $P_Y(\cdot)$, then we can use Theorem 1.1 to bound

$$\begin{aligned}\Pr[(X^n, Y^n) \in T^n_\epsilon(P_{XY})] &= \sum_{(x^n,y^n)\in T^n_\epsilon(P_{XY})} P^n_X(x^n) P^n_Y(y^n) \\ &\leq 2^{nH(XY)(1+\epsilon)}\, 2^{-nH(X)(1-\epsilon)}\, 2^{-nH(Y)(1-\epsilon)} \\ &\leq 2^{-n[I(X;Y)-3\epsilon H(XY)]}. \end{aligned} \tag{4.3}$$

We similarly use Theorem 1.1 to compute

$$\Pr[(X^n, Y^n) \in T_\epsilon^n(P_{XY})] \geq (1 - \delta_\epsilon(n)2^{-n[I(X;Y)+3\epsilon H(XY)]}, \quad (4.4)$$

where

$$\delta_\epsilon(n) = 2|\mathcal{X}||\mathcal{Y}| \cdot e^{-n\epsilon^2 \mu_{XY}}. \quad (4.5)$$

4.3 An Achievable Region

We present a random code construction that makes use of binning. We will consider only block-to-block encoders, although one could also use variable-length encoders. The code book construction is depicted in Figures 4.2 and 4.3 (see also [18, p. 412]).

Code Construction: Generate $2^{n(R_1+R_1')}$ codewords $x^n(w_1, v_1)$, $w_1 = 1, 2, \ldots, 2^{nR_1}$, $v_1 = 1, 2, \ldots, 2^{R_1'}$, by choosing the $n \cdot 2^{n(R_1+R_1')}$ symbols $x_i(w_1, v_1)$ independently at random using $P_X(\cdot)$. Similarly, generate $2^{n(R_2+R_2')}$ codewords $y^n(w_2, v_2)$, $w_2 = 1, 2, \ldots, 2^{nR_2}$, $v_2 = 1, 2, \ldots, 2^{R_2'}$, by choosing the $n \cdot 2^{n(R_2+R_2')}$ symbols $y_i(w_2, v_2)$ independently at random using $P_Y(\cdot)$.

Encoders: Encoder 1 tries to find a pair (w_1, v_1) such that $x^n = x^n(w_1, v_1)$. If successful, Encoder 1 transmits the bin index w_1. If

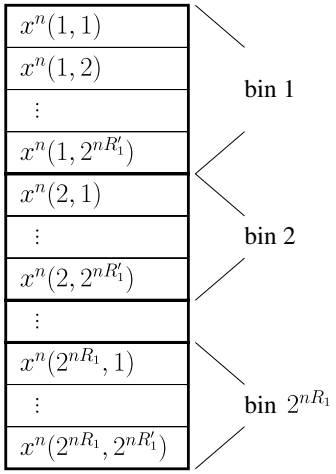

Fig. 4.2 Binning for the x^n sequences.

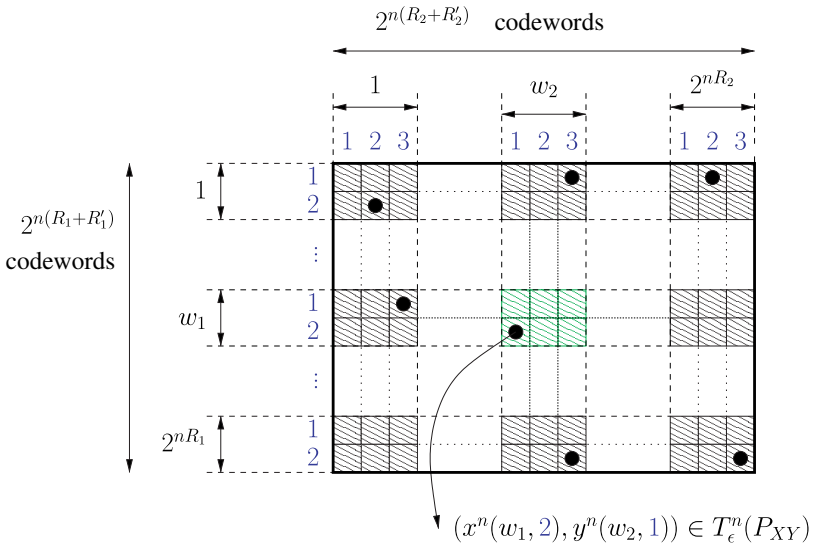

Fig. 4.3 Binning for the x^n and y^n sequences. A dot indicates a pair (x^n, y^n) in $T_\epsilon^n(P_{XY})$. There should be at most one dot for every bin pair (w_1, w_2).

unsuccessful, encoder 1 transmits $w_1 = 1$. Encoder 2 proceeds in the same way with y^n and transmits w_2.

Decoder: Given (w_1, w_2), try to find a pair $(\hat{v}_1, \hat{v}_2)$ such that $(x^n(w_1, \hat{v}_1), y^n(w_2, \hat{v}_2)) \in T_\epsilon^n(P_{XY})$. If successful, put out the corresponding sequences. If unsuccessful, put out $(x^n(w_1, 1), y^n(w_2, 1))$.

Analysis: We consider five events. Let $0 < \epsilon_1 < \epsilon \leq \mu_{XY}$.

(1) Suppose that $(x^n, y^n) \notin T_{\epsilon_1}^n(P_{XY})$. The probability of this event is at most $\delta_{\epsilon_1}(n)$ where

$$\delta_{\epsilon_1}(n) = 2|\mathcal{X}||\mathcal{Y}| \cdot e^{-n\epsilon_1^2 \mu_{XY}} \tag{4.6}$$

since we are considering X and Y together. As usual, $\delta_{\epsilon_1}(n) \to 0$ for $n \to \infty$ and $\epsilon_1 > 0$.

(2) Suppose for the remaining steps that $(x^n, y^n) \in T_{\epsilon_1}^n(P_{XY})$. Encoder 1 makes an error if x^n is not a codeword. Using $(1 - x) \leq e^{-x}$, the probability that this happens is upper

bounded by

$$\begin{aligned}[1 - P_X^n(x^n)]^{2^{n(R_1+R_1')}} &\le \exp\left(-2^{n(R_1+R_1')} \cdot P_X^n(x^n)\right) \\ &< \exp\left(-2^{n[R_1+R_1'-H(X)(1+\epsilon_1)]}\right).\end{aligned} \tag{4.7}$$

A similar bound can be derived for the probability of the event that y^n is not a codeword. We thus choose

$$\begin{aligned}R_1' &= H(X) - R_1 + 2\epsilon_1 H(XY) \\ R_2' &= H(Y) - R_2 + 2\epsilon_1 H(XY).\end{aligned} \tag{4.8}$$

(3) Suppose that $x^n = x^n(w_1, v_1)$ and $y^n = y^n(w_2, v_2)$. Consider the event that there is a $(\tilde{v}_1, \tilde{v}_2) \neq (v_1, v_2)$ with $(x^n(w_1, \tilde{v}_1), y^n(w_2, \tilde{v}_2)) \in T_\epsilon^n(P_{XY})$. Since the $x^n(w_1, v_1)$ were chosen independently via $P_X(\cdot)$, the probability of this event is

$$\begin{aligned}&\Pr\left[\bigcup_{(\tilde{v}_1,\tilde{v}_2)\neq(v_1,v_2)} \{(X^n(w_1,\tilde{v}_1), Y^n(w_2,\tilde{v}_2)) \in T_\epsilon^n(P_{XY})\}\right] \\ &\quad\le \sum_{\tilde{v}_1 \neq v_1} \Pr\left[(X^n, y^n(w_2, v_2)) \in T_\epsilon^n(P_{XY})\right] \\ &\qquad + \sum_{\tilde{v}_2 \neq v_2} \Pr\left[(x^n(w_1, v_1), Y^n) \in T_\epsilon^n(P_{XY})\right] \\ &\qquad + \sum_{\tilde{v}_1 \neq v_1} \sum_{\tilde{v}_2 \neq v_2} \Pr\left[(X^n, Y^n) \in T_\epsilon^n(P_{XY})\right] \\ &\quad< 2^{nR_1'} 2^{-n[I(X;Y)-2\epsilon H(X)]} + 2^{nR_2'} 2^{-n[I(X;Y)-2\epsilon H(Y)]} \\ &\qquad + 2^{n(R_1'+R_2')} 2^{-n[I(X;Y)-3\epsilon H(XY)]} \\ &\le 2^{n[H(X|Y)-R_1+4\epsilon H(XY)]} + 2^{n[H(Y|X)-R_2+4\epsilon H(XY)]} \\ &\qquad + 2^{n[H(XY)-R_1-R_2+7\epsilon H(XY)]},\end{aligned} \tag{4.9}$$

where we have used the union bound for the first step, (4.2) and (4.3) for the second step, and (4.8) for the third step.

The bound (4.9) implies that we can choose large n and

$$R_1 > H(X|Y) + 4\epsilon H(XY) \tag{4.10}$$
$$R_2 > H(Y|X) + 4\epsilon H(XY) \tag{4.11}$$
$$R_1 + R_2 > H(XY) + 7\epsilon H(XY) \tag{4.12}$$

to drive the probability of this event to zero.

Combining the above results, for large n we can approach the rate pairs (R_1, R_2) satisfying

$$\begin{aligned} R_1 &\geq H(X|Y) \\ R_2 &\geq H(Y|X) \\ R_1 + R_2 &\geq H(XY). \end{aligned} \tag{4.13}$$

The form of this region is depicted in Figure 4.4. We remark again that separate encoding of the sources achieves the point $(R_1, R_2) = (H(X), H(Y))$, and the resulting achievable region is shown as the shaded region in Figure 4.4. Note also, the remarkable fact that one can approach $R_1 + R_2 = H(XY)$, which is the minimum sum-rate even if both encoders could cooperate!

4.4 Example

As an example, suppose $P_{XY}(\cdot)$ is defined via

$$Y = X \oplus Z, \tag{4.14}$$

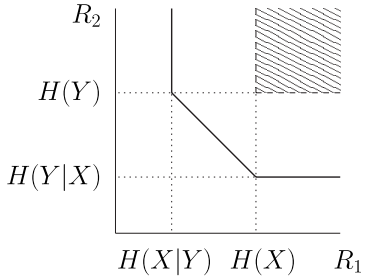

Fig. 4.4 The Slepian–Wolf source coding region.

where $P_X(0) = P_X(1) = 1/2$, and Z is independent of X with $P_Z(0) = 1 - p$ and $P_Z(1) = p$. The region of achievable (R_1, R_2) is therefore

$$\begin{aligned} R_1 &\geq H_2(p) \\ R_2 &\geq H_2(p) \\ R_1 + R_2 &\geq 1 + H_2(p). \end{aligned} \tag{4.15}$$

For example, if $p \approx 0.11$ we have $H_2(p) = 0.5$. The equal rate boundary point is $R_1 = R_2 = 0.75$, which is substantially better than the $R_1 = R_2 = 1$ achieved with separate encoding.

Continuing with this example, suppose we wish to approach the corner point $(R_1, R_2) = (1, 0.5)$. We can use the following encoding procedure: transmit x^n without compression to the decoder, and compress y^n by multiplying y^n on the right by a $n \times (n/2)$ sparse binary matrix H (we use matrix operations over the Galois field GF(2)). The matrix H can be considered to be a parity-check matrix for a *low-density parity-check* (LDPC) code. The encoding can be depicted in graphical form as shown in Figure 4.5. Furthermore, the decoder can consider the x^n to be outputs from a binary symmetric channel (BSC) with inputs y^n and crossover probability $p \approx 0.11$. One must, therefore, design the LDPC code to approach capacity on such a channel, and techniques for doing this are known [52]. This example shows how *channel coding* techniques can be used to solve a *source coding* problem.

4.5 Converse

We show that the rates of (4.13) are, in fact, the best rates we can hope to achieve for block-to-block encoding. Recall that there are 2^{nR_1} indexes w_1, and that w_1 is a function of x^n. We thus have

$$\begin{aligned} nR_1 &\geq H(W_1) \\ &\geq H(W_1|Y^n) \\ &= H(W_1|Y^n) - H(W_1|X^nY^n) \\ &= I(X^n; W_1|Y^n) \\ &= H(X^n|Y^n) - H(X^n|Y^nW_1). \end{aligned} \tag{4.16}$$

Next, note that $H(X^n|Y^n) = nH(X|Y)$, that w_2 is a function of y^n, and that $\hat{x}^n$ and $\hat{y^n}$ are functions of w_1 and w_2. We continue the above

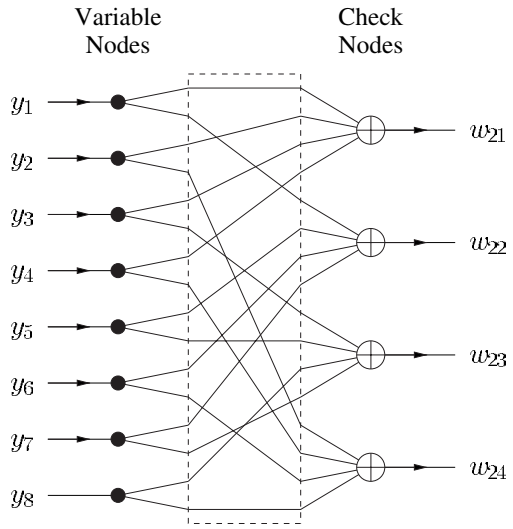

Fig. 4.5 A linear source encoder for binary y^n.

chain of inequalities as

$$\begin{aligned} nR_1 &\geq nH(X|Y) - H(X^n|Y^nW_1) \\ &= nH(X|Y) - H(X^n|Y^nW_1W_2\hat{X}^n\hat{Y}^n) \\ &\geq nH(X|Y) - H(X^nY^n|\hat{X}^n\hat{Y}^n) \\ &\geq nH(X|Y) - n\left[P_e \log_2(|\mathcal{X}| \cdot |\mathcal{Y}|) + H_2(P_e)/n\right], \end{aligned} \tag{4.17}$$

where the final step follows by using $P_e = \Pr[(X^n, Y^n) \neq (\hat{X}^n, \hat{Y}^n)]$ and applying Fano's inequality. We thus find that $R_1 \geq H(X|Y)$ for (block-to-block) encoders with arbitrarily small positive P_e. Similar steps show that

$$\begin{aligned} R_2 &\geq H(Y|X) - [P_e \log_2(|\mathcal{X}| \cdot |\mathcal{Y}|) + H_2(P_e)/n] \\ R_1 + R_2 &\geq H(XY) - [P_e \log_2(|\mathcal{X}| \cdot |\mathcal{Y}|) + H_2(P_e)/n]. \end{aligned} \tag{4.18}$$

This completes the converse.

5

The Wyner–Ziv Problem, or Rate Distortion with Side Information

5.1 Problem Description

Consider again the model of Figure 4.1 that is depicted in Figure 5.1. However, we now permit $\hat{X}^n \in \hat{\mathcal{X}}^n$ and $\hat{Y}^n \in \hat{\mathcal{Y}}^n$ to be *distorted* versions of the respective X^n and Y^n. The goal is to design the encoders and decoder so the average distortions $E\big[d_1^n(X^n, \hat{X}^n)\big]$ and $E\big[d_2^n(Y^n, \hat{Y}^n)\big]$ are smaller than the respective D_1 and D_2.

It might seem remarkable, but this *distributed* source coding problem is still open even if both distortion functions are averages of per-letter distortion functions. That is, the best (known) achievable region of four-tuples (R_1, R_2, D_1, D_2) is *not* the same as the best (known) outer bound on the set of such four-tuples. The problem has, however, been solved for several important special cases. One of these is the RD problem, where Y could be modeled as being independent of X. A second case is the Slepian–Wolf problem that has $D_1 = D_2 = 0$. A third case is where $R_2 \geq H(Y)$ (or $R_1 \geq H(X)$), in which case the decoder can be made to recover Y^n with probability 1 as n becomes large. This problem is known as the *Wyner–Ziv* problem that we will treat here (see [71]).

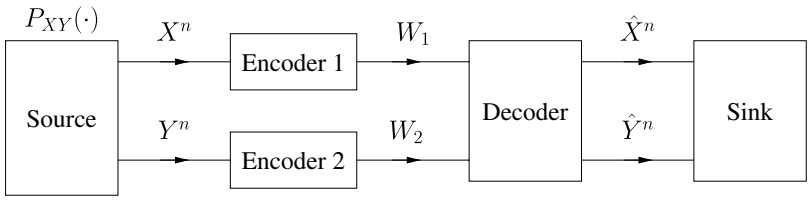

Fig. 5.1 A distributed source coding problem.

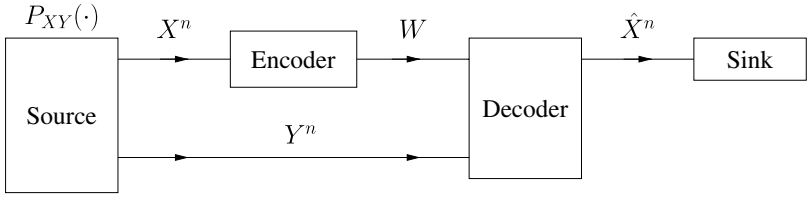

Fig. 5.2 The Wyner–Ziv problem.

Consider, then, the Wyner–Ziv problem, depicted in a simpler form in Figure 5.2. This problem is also referred to as *rate distortion with side information*, where Y^n is the "side information." The index w takes on one of 2^{nR} values, and the average distortion

$$\frac{1}{n}\sum_{i=1}^{n} E\big[d\big(X_i, \hat{X}_i(W, Y^n)\big)\big]$$

should be at most D. The problem is to find the set of pairs (R, D) that can be approached with source encoders and decoders.

This problem has practical import in some, perhaps, unexpected problems. Consider, e.g., a wireless *relay* channel with a transmitter, relay, and destination. The relay might decide to pass on to the destination its noisy observations Y^n of the transmitter signal X^n. The destination would then naturally view Y^n as side information. There are many other problems where side information plays an important role.

5.2 Markov Lemma

We need a result concerning Markov chains $X - Y - Z$ that is known as the *Markov Lemma* [6]. Let μ_{XYZ} be the smallest positive value of $P_{XYZ}(\cdot)$ and $0 \le \epsilon_1 < \epsilon_2 \le \mu_{XYZ}$. Suppose that $(x^n, y^n) \in T_{\epsilon_1}(P_{XY})$

and (X^n, Y^n, Z^n) was emitted by the DMS $P_{XYZ}(\cdot)$. Theorem 1.2 immediately gives

$$\begin{aligned}
&\Pr\left[Z^n \in T^n_{\epsilon_2}(P_{XYZ}|x^n, y^n)\,|Y^n = y^n\right] \\
&\quad = \Pr\left[Z^n \in T^n_{\epsilon_2}(P_{XYZ}|x^n, y^n)\,|X^n = x^n, Y^n = y^n\right] \\
&\quad \geq 1 - \delta_{\epsilon_1,\epsilon_2}(n),
\end{aligned} \tag{5.1}$$

where the first step follows by Markovity, and the second step by (1.27) where

$$\delta_{\epsilon_1,\epsilon_2}(n) = 2|\mathcal{X}||\mathcal{Y}||\mathcal{Z}|\exp\left(-n \cdot \frac{(\epsilon_2 - \epsilon_1)^2}{1+\epsilon_1} \cdot \mu_{XYZ}\right). \tag{5.2}$$

Observe that the right-hand side of (5.1) approaches 1 as $n \to \infty$.

5.3 An Achievable Region

The coding and analysis will be somewhat trickier than for the RD or Slepian–Wolf problems. We introduce a new random variable U, often called an *auxiliary* random variable, to the problem. Let $P_{U|X}(\cdot)$ be a "channel" from X to U, where the alphabet of U is $\mathcal{U}$. U represents a codeword sent from the encoder to the decoder. We further define a function $f(\cdot)$ that maps symbols in $\mathcal{U} \times \mathcal{Y}$ to $\hat{\mathcal{X}}$, i.e., the reconstruction $\hat{x}^n$ has $\hat{x}_i = f(u_i, y_i)$ for all i (recall that y^n is one of the two output sequences of the source). We write the corresponding sequence mapping as $\hat{x}^n = f^n(u^n, y^n)$.

Code Construction: Generate $2^{n(R+R')}$ codewords $u^n(w,v)$, $w = 1, 2, \ldots, 2^{nR}$, $v = 1, 2, \ldots, 2^{nR'}$, by choosing the $n \cdot 2^{n(R+R')}$ symbols $u_i(w,v)$ in the code book independently at random according to $P_U(\cdot)$ (computed from $P_{XU}(\cdot)$). Observe that we are using the same type of binning as for the Slepian–Wolf problem.

Encoder: Given x^n, try to find a pair (w,v) such that $(x^n, u^n(w,v)) \in T^n_\epsilon(P_{XU})$. If one is successful, send the index w. If one is unsuccessful, send $w = 1$.

Decoder: Given w and y^n, try to find a $\tilde{v}$ such that $(y^n, u^n(w,\tilde{v})) \in T^n_\epsilon(P_{YU})$. If there is one or more such $\tilde{v}$, choose one as $\hat{v}$ and put out

the reconstruction $\hat{x}^n(w, y^n) = f^n(u^n(w, \hat{v}), y^n)$. If there is no such $\tilde{v}$, then put out $\hat{x}^n(w, y^n) = f^n(u^n(w, 1), y^n)$.

Analysis: We divide the analysis into several parts, and upper bound the average distortion for all but the last part by $d_{\max}$ (see [18, pp. 442–443]). Let $0 < \epsilon_1 < \epsilon_2 < \epsilon \leq \mu_{UXY}$.

(1) Suppose that $(x^n, y^n) \notin T^n_{\epsilon_1}(P_{XY})$. The probability of this event approaches zero with n.

(2) Suppose that $(x^n, y^n) \in T^n_{\epsilon_1}(P_{XY})$ but the encoder cannot find a pair (w, v) such that $(x^n, u^n(w, v)) \in T^n_{\epsilon_2}(P_{XU})$. This event is basically the same as that studied for the RD encoder in (2.8). That is, the probability of this event is small if ϵ_2 is small, n is large and

$$R + R' > I(X;U). \tag{5.3}$$

(3) Suppose $(x^n, y^n) \in T^n_{\epsilon_1}(P_{XY})$ and the encoder finds a (w, v) with $(x^n, u^n(w, v)) \in T^n_{\epsilon_2}(P_{XU})$. However, suppose the decoder finds a $\tilde{v} \neq v$ such that $(y^n, u^n(w, \tilde{v})) \in T^n_{\epsilon_2}(P_{YU})$. The probability of this event is upper bounded by

$$\begin{aligned} &\Pr\left[\bigcup_{\tilde{v}\neq v} \left\{(y^n, U^n(w, \tilde{v})) \in T^n_{\epsilon_2}(P_{YU})\right\}\right] \\ &\quad \leq \sum_{\tilde{v}\neq v} \Pr\left[(y^n, U^n) \in T^n_{\epsilon_2}(P_{YU})\right] \\ &\quad < 2^{n[R' - I(U;Y) + 2\epsilon_2 H(U)]}. \end{aligned} \tag{5.4}$$

Thus, we require that ϵ_2 is small, n is large, and

$$R' < I(Y;U). \tag{5.5}$$

(4) Suppose $(x^n, y^n) \in T^n_{\epsilon_1}(P_{XY})$, the encoder finds a (w, v) with $(x^n, u^n(w, v)) \in T^n_{\epsilon_2}(P_{XU})$, but the decoder cannot find an appropriate $\tilde{v}$. That is, y_i was chosen via $P_{Y|X}(\cdot|x_i) = P_{Y|XU}(\cdot|x_i, u_i)$ for all i and any u_i, and $U - X - Y$ forms a Markov chain, but we have $(y^n, x^n, u^n(w, v)) \notin T^n_{\epsilon}(P_{YXU})$. The bound (5.1) states that the probability of this event is small for large n.

(5) Finally, consider the case $(x^n, u^n(w,v), y^n) \in T_\epsilon^n(P_{XUY})$ and $\hat{v} = v$. The distortion is bounded by

$$\begin{aligned} D(x^n, y^n) &= \frac{1}{n}\sum_{i=1}^{n} d(x_i, \hat{x}_i(w, y^n)) \\ &= \frac{1}{n}\sum_{i=1}^{n} d(x_i, f(u_i, y_i)) \\ &= \frac{1}{n}\sum_{a,b,c} N(a,b,c|x^n,u^n,y^n)\; d(a, f(b,c)) \\ &\le \sum_{a,b,c} P_{XUY}(a,b,c)(1+\epsilon)\; d(a, f(b,c)) \\ &= E\,[d(X, f(U,Y))] + \epsilon\, d_{\max}, \end{aligned} \tag{5.6}$$

where we have assumed that $d(\cdot)$ is upper bounded by $d_{\max}$.

Combining the above results, we can achieve the rate

$$R_{WZ}(D) = \min_{P_{U|X}(\cdot), f(\cdot):\ E[d(X,f(U,Y))] \le D} I(X;U) - I(Y;U). \tag{5.7}$$

One can use the Fenchel–Eggleston strengthening of Carthéodory's Theorem to show that one can restrict attention to U whose alphabet $\mathcal{U}$ satisfies $|\mathcal{U}| \le |\mathcal{X}| + 1$ [71, Proof of Thm. A2 on p. 9]. We remark that one could replace $f(\cdot)$ by a probability distribution $P_{\hat{X}|UY}(\cdot)$, but it suffices to use a deterministic mapping $\hat{X} = f(U,Y)$.

Observe that one can alternatively write the mutual information expression in (5.7) as

$$\begin{aligned} I(X;U) - I(Y;U) &= H(U|Y) - H(U|X) \\ &= H(U|Y) - H(U|XY) \\ &= I(X;U|Y). \end{aligned} \tag{5.8}$$

The formulation (5.8) is intuitively appealing from the *decoder's* perspective if we regard U as representing the index W in Figure 5.2. However, the interpretation is not fitting from the *encoder's* perspective because the encoder does not know Y. Moreover, note that

$$I(X;U|Y) = I(X;U\hat{X}|Y) \ge I(X;\hat{X}|Y) \tag{5.9}$$

with equality if and only if $I(X;U|Y\hat{X}) = 0$. It is the expression on the right in (5.9) that corresponds to the case where the encoder also sees Y. That is, the RD function for the problem where *both* the encoder *and* decoder have access to the side information Y is

$$R_{X|Y}(D) = \min_{P_{\hat{X}|XY}(\cdot):\, E\left[d(X,\hat{X})\right] \le D} I(X;\hat{X}|Y), \tag{5.10}$$

which is less than $R_{WZ}(D)$ for most common sources and distortion functions.

5.4 Discrete Alphabet Example

As an example, consider the BSS $P_X(\cdot)$ with Hamming distortion. Suppose Y is the output of a BSC that has input X and crossover probability p. We use two encoding strategies and time-share between them. For the first strategy, we choose U as the output of a BSC with input X and crossover probability β (note that $|\mathcal{U}| \le |\mathcal{X}| + 1$). We further choose $\hat{X} = f(Y,U) = U$ and compute

$$\begin{aligned} I(X;U) - I(Y;U) &= [1 - H_2(\beta)] - [1 - H_2(p * \beta)] \\ &= H_2(p * \beta) - H_2(\beta), \end{aligned} \tag{5.11}$$

Where $p * \beta = p(1-\beta) + (1-p)\beta$ and $E[d(X,\hat{X})] = \beta$. For the second strategy, we choose $U = 0$ and $\hat{X} = f(Y,U) = Y$. This implies $I(X;U) - I(Y;U) = 0$ and $E[d(X,\hat{X})] = p$. Finally, we use the first and second strategies a fraction λ and $1-\lambda$ of the time, respectively. We achieve the rates

$$R'_{WZ}(D) = \min_{\lambda,\beta:\, \lambda\beta + (1-\lambda)p \le D} \lambda\,[H_2(p * \beta) - H_2(\beta)]\,. \tag{5.12}$$

This achievable region is, in fact, the rate distortion function for this problem (see [71, Sec. II]).

Recall that, without side information, the RD function for the BSS and Hamming distortion is $1 - H_2(p)$. One can check that this rate is larger than (5.12) unless $D = 1/2$ or $p = 1/2$, i.e., unless $R(D) = 0$ or X and Y are independent. Consider also the case where the encoder

has access to Y^n. For the BSS and Hamming distortion, we compute

$$R_{X|Y}(D) = \begin{cases} h(p) - h(D) & 0 \le D < p \\ 0 & p \le D. \end{cases} \tag{5.13}$$

We find that $R_{X|Y}(D)$ is less than (5.12) unless $D = 0$, $p \le D$, or $p = 1/2$.

5.5 Gaussian Source and Mean Squared Error Distortion

As a second example, suppose X and Y are Gaussian random variables with variances σ_X^2 and σ_Y^2, respectively, and with correlation coefficient $\rho = E[XY]/(\sigma_X \sigma_Y)$. For the Gaussian distortion function, we require $E[(X - \hat{X})^2] \le D$. Clearly, if $D \ge \sigma_X^2(1 - \rho^2)$, then $R(D) = 0$. So suppose that $D < \sigma_X^2(1 - \rho^2)$. We choose $U = X + Z$, where Z is a Gaussian random variable with variance σ_Z^2 and $\hat{X} = f(Y,U) = E[X|Y,U]$, i.e., $\hat{X}$ is the minimum mean-square error (MMSE) estimate of X given Y and U. We use (5.8) to compute

$$\begin{aligned} I(X;U|Y) &= h(X|Y) - h(X|YU) \\ &= h(X|Y) - h(X - \hat{X}|YU) \\ &= h(X|Y) - h(X - \hat{X}) \\ &= \frac{1}{2} \log \left(\frac{\sigma_X^2(1 - \rho^2)}{D} \right), \end{aligned} \tag{5.14}$$

where the third step follows by the orthogonality principle of MMSE estimation, and where the fourth step follows by choosing Z so that $E\left[(X - \hat{X})^2\right] = D$. The rate (5.14) turns out to be optimal, and it is generally smaller than the RD function $R(D) = \log(\sigma_X^2/D)/2$ that we computed in Section 2.4. However, one can check that $R_{X|Y}(D) = R_{WZ}(D)$. Thus, for the Gaussian source and squared error distortion the encoder can compress at the same rate whether or not it sees Y!

5.6 Two Properties of $R_{WZ}(D)$

The function $R_{WZ}(D)$ in (5.7) is clearly non-increasing with D. We prove that $R_{WZ}(D)$ is *convex* in D [18, Lemma 14.9.1 on p. 439].

Consider two distinct points (R_1, D_1) and (R_2, D_2) on the boundary of $R_{WZ}(D)$, and suppose the channels and functions $P_{U_1|X}(\cdot)$, $\hat{X}_1 = f_1(U_1, Y)$ and $P_{U_2|X}(\cdot)$, $\hat{X}_2 = f_2(U_2, Y)$ achieve these respective points. Let Q be a random variable with $P_Q(1) = 1 - P_Q(2) = \lambda$ that is independent of X and Y. Define $U_3 = [Q, \tilde{U}_3]$ and consider the distribution

$$P_{[Q,\tilde{U}_3]|X}([q,a]|b) = P_Q(q) P_{U_q|X}(a|b) \quad \text{for all } q,\, a,\, b, \tag{5.15}$$

i.e., we have $\tilde{U}_3 = U_1$ if $Q = 1$ and $\tilde{U}_3 = U_2$ if $Q = 2$. We consider U_3 as our auxiliary random variable. Consider also $f_3(\cdot)$ with $f_3([Q, \tilde{U}_3], Y) = (2 - Q)\, f_1(\tilde{U}_3, Y) + (Q - 1)\, f_2(\tilde{U}_3, Y)$. The distortion with $P_{[Q,\tilde{U}_3]|X}(\cdot)$ is simply $D_3 = \lambda D_1 + (1 - \lambda) D_2$. We thus have

$$\begin{aligned} R_{WZ}(\lambda D_1 + (1-\lambda) D_2) &= R_{WZ}(D_3) \\ &\le I(X; Q\tilde{U}_3|Y) \\ &= I(X; \tilde{U}_3|YQ) \\ &= \lambda\, I(X; U_1|Y) + (1-\lambda)\, I(X; U_2|Y) \\ &= \lambda\, R_{WZ}(D_1) + (1-\lambda)\, R_{WZ}(D_2). \end{aligned} \tag{5.16}$$

5.7 Converse

We show that $R_{WZ}(D)$ in (5.7) is the RD function for the Wyner–Ziv problem. Let $\hat{X}^n = g(W, Y^n)$ and $D = \frac{1}{n}\sum_{i=1}^n E[d(X_i, \hat{X}_i)]$. Recall that there are 2^{nR} indexes w. We thus have

$$\begin{aligned} nR &\ge H(W) \\ &\ge I(X^n; W|Y^n) \\ &= H(X^n|Y^n) - H(X^n|WY^n) \\ &= \sum_{i=1}^n H(X_i|Y_i) - H(X_i|Y_i\ (WY^{i-1}Y_{i+1}^n)\ X^{i-1}) \\ &\ge \sum_{i=1}^n H(X_i|Y_i) - H(X_i|Y_i\ (WY^{i-1}Y_{i+1}^n)) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} H(X_i|Y_i) - H(X_i|Y_i U_i) \\
&= \sum_{i=1}^{n} I(X_i; U_i|Y_i), \qquad (5.17)
\end{aligned}
$$

where the second last step follows by setting $U_i = [W Y^{i-1} Y_{i+1}^n]$. Note that $U_i - X_i - Y_i$ forms a Markov chain for all i, and that $\hat{X}_i = g_i(W, Y^n) = f_i(U_i, Y_i)$ for some $g_i(\cdot)$ and $f_i(\cdot)$. We use the definition (5.7), the alternative formulation (5.8), and the convexity (5.16) to continue the chain of inequalities (5.17):

$$
\begin{aligned}
nR &\geq \sum_{i=1}^{n} R_{WZ}\left(E[d(X_i, f_i(U_i, Y_i))]\right) \\
&\geq nR_{WZ}\left(\frac{1}{n}\sum_{i=1}^{n} E[d(X_i, f_i(U_i, Y_i))]\right) \\
&= nR_{WZ}\left(\frac{1}{n}\sum_{i=1}^{n} E\left[d(X_i, \hat{X}_i)\right]\right) \\
&\geq nR_{WZ}(D). \qquad (5.18)
\end{aligned}
$$

Thus, the random coding scheme described in Section 5.3 is rate-optimal.

6

The Gelfand–Pinsker Problem, or Coding for Channels with State

6.1 Problem Description

The Gelfand–Pinsker problem is depicted in Figure 6.1. A source sends a message w, $w \in \{1, 2, \ldots, 2^{nR}\}$, to a receiver by mapping it into a sequence x^n. However, as an important change to a DMC, the channel $P_{Y|XS}(\cdot)$ has *interference* in the form of a sequence s^n that is output from a DMS $P_S(\cdot)$. Moreover, the encoder has access to the interference s^n in a *noncausal* fashion, i.e., the encoder knows s^n *ahead* of time. The receiver does not know s^n. The goal is to design the encoder and decoder to maximize R while ensuring that $P_{\mathrm{e}} = \Pr[\hat{W} \neq W]$ can be made an arbitrarily small positive number. The capacity C is the supremum of the achievable rates R.

The problem might seem strange at first glance. Why should interference be known noncausally? However, such a situation can arise in practice. Consider, for example, the encoder of a broadcast channel with two receivers. The two messages for the receivers might be mapped to sequences s_1^n and s_2^n, respectively, and s_1^n can be thought of as being interference for s_2^n. Furthermore, the encoder *does* have noncausal knowledge of s_1^n. We will develop such a coding scheme later on.

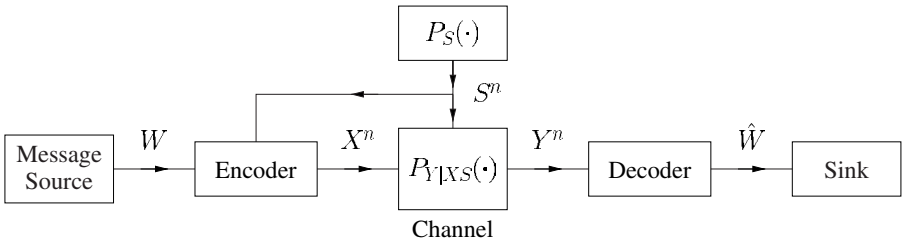

Fig. 6.1 The Gelfand–Pinsker problem.

As a second example, suppose we are given a memory device that has been imprinted, such as a compact disc. We wish to encode new data on this "old" disc in order to reuse it. We can read the data s^n already in the memory, and we can view s^n as interference that we know noncausally. We might further wish to model the effect of errors that an imprinting device can make during imprinting by using a probabilistic channel $P_{Y|XS}(\cdot)$.

6.2 An Achievable Region

The Gelfand–Pinsker problem was solved in [28] by using binning. We begin by introducing an auxiliary random variable U with alphabet $\mathcal{U}$, and we consider U to be the output of a "channel" $P_{U|S}(\cdot)$. We also define a function $f(\cdot)$ that maps symbols in $\mathcal{U} \times \mathcal{S}$ to $\mathcal{X}$, i.e., the sequence x^n will have $x_i = f(u_i, s_i)$ for all i. We write the corresponding sequence mapping as $x^n = f^n(u^n, s^n)$.

Code Construction: Generate $2^{n(R+R')}$ codewords $u^n(w,v)$, $w = 1,2,\ldots,2^{nR}$, $v = 1,2,\ldots,2^{nR'}$, by choosing the $n \cdot 2^{n(R+R')}$ symbols $u_i(w,v)$ in the code book independently at random according to $P_U(\cdot)$.

Encoder: Given w and s^n, try to find a v such that $(u^n(w,v), s^n) \in T_\epsilon^n(P_{US})$. That is, w chooses the bin with codewords $u^n(w,1), u^n(w,2), \ldots, u^n(w,2^{nR'})$, and the interference "selects" $u(w,v)$ from this bin. If one finds an appropriate codeword $u(w,v)$, transmit $x^n = f^n(u^n(w,v), s^n)$. If not, transmit $x^n = f^n(u^n(w,1), s^n)$.

Decoder: Given y^n, try to find a pair $(\tilde{w}, \tilde{v})$ such that $(u^n(\tilde{w}, \tilde{v}), y^n) \in T_\epsilon^n(P_{UY})$. If there is one or more such pair, then choose one and put

out the corresponding $\tilde{w}$ as $\hat{w}$. If there is no such pair, then put out $\hat{w} = 1$.

Analysis: We proceed in several steps. Let $0 < \epsilon_1 < \epsilon_2 < \epsilon_3 < \epsilon \leq \mu_{USXY}$, where μ_{USXY} is the smallest positive value of $P_{USXY}(\cdot)$.

(1) Suppose that $s^n \notin T^n_{\epsilon_1}(P_S)$. The probability of this event approaches zero with n.
(2) Suppose $s^n \in T^n_{\epsilon_1}(P_S)$ but the encoder cannot find a v such that $(u^n(w,v), s^n) \in T^n_{\epsilon_2}(P_{US})$. This event is basically the same as that studied for the rate-distortion problem. That is, the probability of this event is small if ϵ_2 is small, n is large and

$$R' > I(U;S). \tag{6.1}$$

(3) Suppose $(u^n(w,v), s^n) \in T^n_{\epsilon_2}(P_{US})$ which implies $(u^n(w,v), s^n, x^n) \in T^n_{\epsilon_2}(P_{USX})$ (to see this, write $N(a,b,c|u^n,s^n,x^n)$ as a function of $N(a,b|u^n,s^n)$). Suppose further that $(u^n(w,v), y^n) \notin T^n_{\epsilon_3}(P_{UY})$, i.e., y_i was chosen using $P_{Y|SX}(\cdot|s_i, x_i(u_i,s_i))$ for all i, and $Y - [S,X] - U$, but we have $(y^n, [s^n,x^n], u^n) \notin T^n_{\epsilon_3}(P_{Y[S,X]U})$. The Markov Lemma in Section 5.2 ensures that the probability of this event is small for large n.
(4) Suppose $y^n \in T^n_{\epsilon_3}(P_Y)$ and the decoder finds a $(\tilde{w}, \tilde{v})$ with $\tilde{w} \neq w$ and $(u^n(\tilde{w}, \tilde{v}), y^n) \in T^n_{\epsilon}(P_{UY})$. By Theorem 1.3, the probability of this event for any of the $(2^{nR} - 1) \cdot 2^{nR'}$ codewords outside of w's bin is upper bounded by $2^{-n[I(U;Y) - 2\epsilon H(U)]}$. Thus, we require that ϵ is small, n is large, and

$$R + R' < I(U;Y). \tag{6.2}$$

Combining (6.1) and (6.2), we can approach the rate

$$R_{\mathrm{GP}} = \max_{P_{U|S}(\cdot), f(\cdot)} I(U;Y) - I(U;S), \tag{6.3}$$

where $U - [S,X] - Y$ forms a Markov chain. As shown below, R_{GP} is the capacity of the Gelfand–Pinsker problem.

We list a few properties of $R_{\rm GP}$. First, Carthéodory's theorem guarantees that one can restrict attention to U whose alphabet $\mathcal{U}$ satisfies $|\mathcal{U}| \leq |\mathcal{X}| \cdot |\mathcal{S}| + 1$ [28, Prop. 1]. Second, one achieves R_{GP} *without* obtaining S^n at the receiver. Observe also that

$$\begin{aligned} I(U;Y) - I(U;S) &= H(U|S) - H(U|Y) \\ &\leq H(U|S) - H(U|YS) \\ &= I(U;Y|S) \\ &= I(X;Y|S). \end{aligned} \tag{6.4}$$

Thus, $R_{\rm GP}$ is less than the capacity if both the encoder and decoder have access to S^n, namely

$$R_S = \max_{P_{X|S}(\cdot)} I(X;Y|S). \tag{6.5}$$

Next, observe that if Y is independent of S given X then we can choose $[U,X,Y]$ to be independent of S and arrive at

$$\begin{aligned} R_{\rm GP} &= \max_{P_U(\cdot), f(\cdot)} I(U;Y) \\ &= \max_{P_X(\cdot)} I(X;Y) \\ &= \max_{P_X(\cdot)} I(X;Y|S). \end{aligned} \tag{6.6}$$

Finally, the rate expression in (6.3) has convexity properties developed in Section 6.5.

6.3 Discrete Alphabet Example

As an example, suppose $P_{Y|XS}(\cdot)$ has binary X and Y, and ternary S. Suppose that if $S=0$ we have $P_{Y|XS}(1|x,0) = q$ for $x = 0,1$, if $S=1$ we have $P_{Y|XS}(1|x,0) = 1-q$ for $x=0,1$, and if $S=2$ we have $P_{Y|XS}(x|x,0) = 1-p$ for $x=0,1$. Suppose further that $P_S(0) = P_S(1) = \lambda$ and $P_S(2) = 1-2\lambda$. We wish to design $P_{U|S}(\cdot)$ and $f(\cdot)$. We should consider $|\mathcal{U}| \leq 7$, but we here concentrate on binary U. Consider $S=0$ and $S=1$ for which $P_{Y|XS}(\cdot)$ does not depend on X, so we may as well choose $X=S$. We further choose $P_{U|S}(0|0) = P_{U|S}(1|1) = \alpha$. For $S=2$, we choose $X=U$ and $P_X(0) = P_X(1) = 1/2$. We compute

the achievable rate to be

$$R(\alpha) = I(U;Y) - I(U;S) = [1 - H_2(\Pr[Y = U])] - 2\lambda[1 - H_2(\alpha)], \tag{6.7}$$

where

$$\Pr[Y = U] = 2\lambda[\alpha(1-q) + (1-\alpha)q] + (1-2\lambda)(1-p). \tag{6.8}$$

The final step is to optimize over the parameter α. The resulting rate turns out to be the capacity of this channel (see [28, Sec. 5]).

6.4 Gaussian Channel

Suppose S is a (possibly non-Gaussian) random variable with finite variance [12, Sec. II-D, 13],. Suppose further that

$$Y = X + S + Z,$$

where Z is additive white Gaussian noise (AWGN) with variance N, and that we have the power constraint $E\left[X_i^2\right] \leq P$ for $i = 1, 2, \ldots, n$. This problem has become known as "writing on dirty paper" [13]. We define U and X via $U = X + \alpha S$, where X is Gaussian, has variance P, and is statistically independent of S. (N.B. This does *not* necessarily mean that X^n is statistically independent of S^n.) We further choose $\alpha = P/(P+N)$ to make $X + Z$ and $(1-\alpha)X - \alpha Z$ uncorrelated, and hence statistically independent since they are Gaussian. We follow the approach of [12, Sec. II-D] and compute

$$\begin{aligned} h(U|Y) &= h(X + \alpha S \mid X + S + Z) \\ &= h(X + \alpha S - \alpha(X + S + Z) \mid X + S + Z) \\ &= h((1-\alpha)X - \alpha Z \mid X + Z + S) \\ &= h((1-\alpha)X - \alpha Z \mid X + Z) \\ &= h(X \mid X + Z), \end{aligned} \tag{6.9}$$

where the fourth step follows because $X + Z$, $(1-\alpha)X - \alpha Z$, and S are jointly statistically independent. We similarly compute

$$h(U|S) = h(X + \alpha S \mid S) = h(X|S) = h(X). \tag{6.10}$$

We can thus achieve the rate

$$\begin{aligned} I(U;Y) - I(U;S) &= h(U|S) - h(U|Y) \\ &= h(X) - h(X \mid X+Z) \\ &= I(X\,;\,X+Z) \\ &= \frac{1}{2}\log\left(1+\frac{P}{N}\right). \end{aligned} \tag{6.11}$$

But (6.11) is the same as the capacity if the interference is known at both the transmitter and receiver (or the capacity *without* interference). Thus, for the AWGN channel with additive interference, the encoder can transmit at the same rate as if the interference was not present! We generalize this result to *vector* channels in the appendix of this section.

6.5 Convexity Properties

We prove the following proposition (see [28, Prop. 1]). This result is useful, e.g., for optimizing the distributions $P_{U|S}(\cdot)$ and $P_{X|SU}(\cdot)$.

Proposition 6.1. Consider the expression

$$R(S,U,X,Y) = I(U;Y) - I(U;S), \tag{6.12}$$

where the joint distribution of the random variables factors as

$$P_{SUXY}(a,b,c,d) = P_{SUX}(a,b,c) \cdot P_{Y|SX}(d|a,c) \tag{6.13}$$

for all a,b,c,d. $R(S,U,X,Y)$ is a concave (or convex-$\cap$) function of $P_{U|S}(\cdot)$ if $P_S(\cdot)$, $P_{X|SU}(\cdot)$, and $P_{Y|SX}(\cdot)$ are fixed. Similarly, $R(S,U,X,Y)$ is a convex (or convex-$\cup$) function of $P_{X|SU}(\cdot)$ if $P_S(\cdot)$, $P_{U|S}(\cdot)$, and $P_{Y|SX}(\cdot)$ are fixed.

Proof. We begin with the second case where $I(U;S)$ is fixed. We know that $I(U;Y)$ is a convex function of $P_{Y|U}(\cdot)$ if $P_U(\cdot)$ is fixed. But we have

$$P_{Y|U}(d|b) = \sum_{a,c} P_{S|U}(a|b)\, P_{X|SU}(c|a,b)\, P_{Y|SX}(d|a,c), \tag{6.14}$$

i.e., $P_{Y|U}(\cdot)$ is a linear function of $P_{X|SU}(\cdot)$. Thus, $R(S,U,X,Y)$ is a convex function of $P_{X|SU}(\cdot)$.

For the first case, $-I(U;S)$ is clearly concave in $P_{U|S}(\cdot)$. Furthermore, we have $I(U;Y) = H(Y) - H(Y|U)$, where $H(Y)$ is concave in $P_Y(\cdot)$ and $H(Y|U)$ is linear in $P_{U|S}(\cdot)$. But $P_Y(\cdot)$ is also linear in $P_{U|S}(\cdot)$, and the sum of two concave functions is concave, so we have the desired result. □

6.6 Converse

We show that R_{GP} in (6.3) is the capacity of the Gelfand–Pinsker problem. We use Fano's inequality to bound the rate for reliable communication as (see (3.8) and (3.11))

$$\begin{aligned} nR &\le I(W;\hat{W}) \\ &\le I(W;Y^n) \\ &= \sum_{i=1}^{n} I(WS_{i+1}^n;Y^i) - I(WS_i^n;Y^{i-1}), \end{aligned} \tag{6.15}$$

where the second step follows by the data processing theorem, and the third step by expanding the sum, canceling terms pair-wise, and setting $S_i^j = [S_i, S_{i+1}, \ldots, S_j]$ and $Y_0 = 0$. We continue the chain of (in)equalities (6.15):

$$\begin{aligned} nR &\le \sum_{i=1}^{n} \big[I(WS_{i+1}^n;Y^{i-1}) + I(WS_{i+1}^n;Y_i|Y^{i-1})\big] \\ &\quad - \big[I(WS_{i+1}^n;Y^{i-1}) + I(S_i;Y^{i-1}|WS_{i+1}^n)\big] \\ &= \sum_{i=1}^{n} I(WS_{i+1}^n;Y_i|Y^{i-1}) - I(S_i;Y^{i-1}|WS_{i+1}^n) \\ &= \sum_{i=1}^{n} \big[H(Y_i|Y^{i-1}) - H(Y_i|U_i)\big] - \big[H(S_i|WS_{i+1}^n) - H(S_i|U_i)\big] \\ &\le \sum_{i=1}^{n} [H(Y_i) - H(Y_i|U_i)] - [H(S_i) - H(S_i|U_i)] \\ &= \sum_{i=1}^{n} I(U_i;Y_i) - I(U_i;S_i), \end{aligned} \tag{6.16}$$

where for the second step we have defined $U_i = [W, S_{i+1}^n, Y^{i-1}]$, and the third step follows because S_i is independent of W and S_{i+1}^n. We further have that $U_i - [X_i, S_i] - Y_i$ forms a Markov chain. We can bound the sum (6.16) by n times its maximum term to obtain

$$\begin{aligned} R &\le \max_i [I(U_i;Y_i) - I(U_i;S_i)] \\ &\le \max_{P_{U|S}(\cdot), P_{X|SU}(\cdot)} I(U;Y) - I(U;S). \end{aligned} \tag{6.17}$$

The final step is to use Proposition 6.1. Because $I(U;Y) - I(U;S)$ is convex in $P_{X|SU}(\cdot)$, one should choose X to be a deterministic function of U and S, i.e., $X = f(U,S)$ for some $f(\cdot)$.

6.7 Appendix: Writing on Dirty Paper with Vector Symbols

Suppose the channel output is an $N \times 1$ vector

$$\underline{Y} = \mathbf{H}_X \underline{X} + \mathbf{H}_S \underline{S} + \underline{Z}, \tag{6.18}$$

where $\underline{X}$ is a random $M \times 1$ vector, $\underline{S}$ is a random $L \times 1$ vector, $\mathbf{H}_X$ and $\mathbf{H}_S$ are $N \times M$ and $N \times L$ matrices, respectively, and $\underline{Z}$ is a $N \times 1$ Gaussian vector that is statistically independent of $\underline{S}$ and $\underline{H}$, and has zero mean and nonsingular covariance matrix $\mathbf{Q}_{\underline{Z}}$. Suppose $\underline{S}$ is a (possibly non-Gaussian) random vector (see [74]) and that $E\left[\|\underline{X}_i\|^2\right] \le P$ for $i = 1, 2, \ldots, n$. We define $\underline{U} = \underline{X} + \mathbf{A}\mathbf{H}_S\underline{S}$, where

$$\mathbf{A} = \mathbf{Q}_{\underline{X}} \mathbf{H}_X^T \left(\mathbf{Q}_{\underline{Z}} + \mathbf{H}_X \mathbf{Q}_{\underline{X}} \mathbf{H}_X^T\right)^{-1} \tag{6.19}$$

is an $M \times N$ matrix, and where $\underline{X}$ is Gaussian, statistically independent of $\underline{S}$, and has covariance matrix $\mathbf{Q}_{\underline{X}}$ with trace $\mathrm{tr}\left[\mathbf{Q}_{\underline{X}}\right] = P$. One can check that $\mathbf{H}_X\underline{X} + \underline{Z}$ and $(I - \mathbf{A}\mathbf{H}_X)\underline{X} - \mathbf{A}\underline{Z}$ are uncorrelated, and hence statistically independent since they are Gaussian. We follow the same steps as for the scalar Gaussian example to compute

$$I(U;Y) - I(U;S) = \frac{1}{2} \log \left| I + \mathbf{H}_X \mathbf{Q}_{\underline{X}} \mathbf{H}_X^T \mathbf{Q}_{\underline{Z}}^{-1} \right|. \tag{6.20}$$

The expression (6.20) is the information rate across the vector channel if there is no interference, i.e., $\underline{S} = \underline{0}$ or $\mathbf{H}_S = 0$. The final step is to maximize (6.20) over all choices of $\mathbf{Q}_{\underline{X}}$. We can do this as for the vector

AWGN channel in Section 3.6. Observe that we can factor the positive definite matrix $\mathbf{Q}_{\underline{Z}}^{-1}$ as $\mathbf{Q}_{\underline{Z}}^{-1/2} \cdot \mathbf{Q}_{\underline{Z}}^{-1/2}$, where $\mathbf{Q}_{\underline{Z}}^{-1/2}$ is a positive definite matrix [31, p. 406]. Equation (6.20) thus gives

$$I(U;Y) - I(U;S) = \frac{1}{2} \log \left| I + \mathbf{Q}_{\underline{Z}}^{-1/2} \mathbf{H}_X \mathbf{Q}_{\underline{X}} \mathbf{H}_X^T \mathbf{Q}_{\underline{Z}}^{-1/2} \right|. \tag{6.21}$$

The resulting optimization problem has the same form as (3.27) with $\mathbf{H} = \mathbf{Q}_{\underline{Z}}^{-1/2} \mathbf{H}_X$.

7

The Broadcast Channel

7.1 Problem Description

The broadcast channel is depicted in Figure 7.1. There are three sources, one encoder, and two decoders and sinks. The sources put out the statistically independent messages W_0, W_1, W_2 with nR_0, nR_1, nR_2 bits, respectively. The message W_0 is destined for *both* sinks, and is sometimes called the *common* or *public* message. The messages W_1 and W_2 are destined for sinks 1 and 2, respectively, and are sometimes called *private* messages. The encoder maps (w_0, w_1, w_2) to a sequence $x^n \in \mathcal{X}^n$, and the channel $P_{Y_1Y_2|X}(\cdot)$ puts out two sequences $y_1^n \in \mathcal{Y}_1^n$ and $y_2^n \in \mathcal{Y}_2^n$. Decoder 1 uses y_1^n to compute its estimate $(\hat{w}_0(1), \hat{w}_1)$ of (w_0, w_1), and decoder 2 similarly uses y_2^n to compute its estimate $(\hat{w}_0(2), \hat{w}_2)$ of (w_0, w_2). The problem is to find the set of rate-tuples (R_0, R_1, R_2) for which one can make

$$P_e = \Pr\big[(\hat{W}_0(1), \hat{W}_0(2), \hat{W}_1, \hat{W}_2) \neq (W_0, W_0, W_1, W_2)\big] \tag{7.1}$$

an arbitrarily small positive number. The closure of the region of achievable (R_0, R_1, R_2) is the broadcast channel capacity region $\mathcal{C}_{\mathrm{BC}}$.

The broadcast channel has important applications. For example, consider the design of a *base station* for a cellular radio system. If the

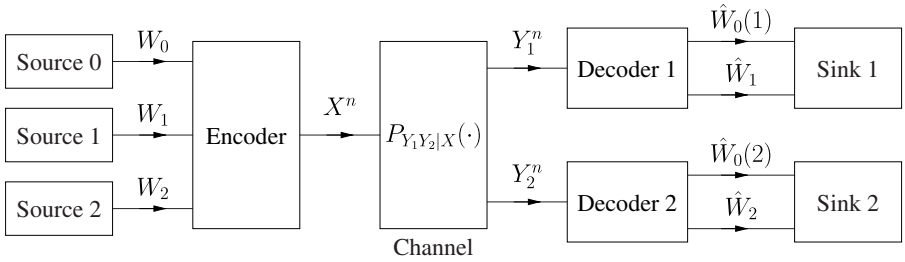

Fig. 7.1 The two-receiver broadcast channel.

base station transmits to two *mobile stations*, the model of Figure 7.1 describes the essence of the coding problem. One can easily extend the model to include three or more mobile stations, but we will study only the two-receiver problem. Despite intense research activity on broadcast channels spanning over three decades, the problem is still open! We will study the problem by focusing on several special cases. The theory for each of these cases gives insight into how one should code in general.

7.2 Preliminaries

7.2.1 Basic Properties

The broadcast channel was studied by Cover in [15], who described several interesting properties and methods for this channel. One simple property is that one can convert some fraction of the R_0 bits to R_1 and/or R_2 bits. Thus, if (R_0, R_1, R_2) is achievable, then so is $(\alpha_0 R_0, R_1 + \alpha_1 R_0, R_2 + \alpha_2 R_0)$, where $\alpha_i \geq 0$, $i = 0, 1, 2$, and $\alpha_0 + \alpha_1 + \alpha_2 = 1$.

A second important property is that the capacity region $\mathcal{C}_{BC}$ depends only on the *marginals* $P_{Y_1|X}(\cdot)$ and $P_{Y_2|X}(\cdot)$. That is, $\mathcal{C}_{BC}$ is the same for the channels $P_{\tilde{Y}_1\tilde{Y}_2|X}(\cdot)$ and $P_{Y_1Y_2|X}(\cdot)$ if

$$\begin{aligned} P_{\tilde{Y}_1|X}(b|a) &= P_{Y_1|X}(b|a) \quad \text{for all } (a,b) \in \mathcal{X} \times \mathcal{Y}_1 \\ P_{\tilde{Y}_2|X}(c|a) &= P_{Y_2|X}(c|a) \quad \text{for all } (a,c) \in \mathcal{X} \times \mathcal{Y}_2. \end{aligned} \tag{7.2}$$

To prove this claim, suppose the channel is $P_{Y_1Y_2|X}(\cdot)$ and let

$$\begin{aligned}\mathcal{E}_1 &= \left\{(\hat{W}_0(1), \hat{W}_1) \neq (W_0, W_1)\right\} \\ \mathcal{E}_2 &= \left\{(\hat{W}_0(2), \hat{W}_2) \neq (W_0, W_2)\right\}\end{aligned} \tag{7.3}$$

so that $P_{e1} = \Pr[\mathcal{E}_1]$ and $P_{e2} = \Pr[\mathcal{E}_2]$ are the respective error probabilities of decoders 1 and 2. We have $P_e = \Pr[\mathcal{E}_1 \cup \mathcal{E}_2]$ and, by elementary set inclusion, we also have (see [18, p. 454])

$$\max(P_{e1}, P_{e2}) \leq P_e \leq P_{e1} + P_{e2}. \tag{7.4}$$

Thus, P_e is small if and only if both P_{e1} and P_{e2} are small. But P_{e1} and P_{e2} depend only on the respective $P_{Y_1|X}(\cdot)$ and $P_{Y_2|X}(\cdot)$, so the same code for any $P_{Y_1Y_2|X}(\cdot)$ with marginals $P_{Y_1|X}(\cdot)$ and $P_{Y_2|X}(\cdot)$ gives the same P_{e1} and P_{e2}. This proves the claim.

The above property lets one restrict attention to broadcast channels where, for example, $Y_1 - X - Y_2$ forms a Markov chain. However, to prove capacity theorems it is sometimes useful to carefully choose the dependencies between Y_1 and Y_2 given X.

7.2.2 A Bound on Binning Rates

We consider a channel coding "dual" of Slepian–Wolf binning. Suppose we generate codewords $x^n(w_1, v_1)$ and $y^n(w_2, v_2)$ exactly as in Section 4.3. Recall that for source coding we required the bins to be *small* enough so that there is at *most* one typical (x^n, y^n) in each bin. We now ask a different question: how *large* must the bins be to ensure that there is at *least* one typical (x^n, y^n) in each bin? The probability that there is no typical (x^n, y^n) in bin (w_1, w_2) is

$$P_{\text{bin},e}(w_1, w_2) = \Pr\left[\bigcap_{v_1, v_2} \{(X^n(w_1, v_1), Y^n(w_2, v_2)) \notin T_\epsilon^n(P_{XY})\}\right]. \tag{7.5}$$

The difficulty in upper bounding (7.5) is that it involves an *intersection* of dependent events, rather than a union. One approach for treating such problems is the *second moment method* [3] . We use this method in the appendix of this section to show that $P_{\text{bin},e}(w_1, w_2)$ is small if n

is large, ϵ is small, and the binning rates satisfy (see [23])

$$R_1' + R_2' > I(X;Y). \tag{7.6}$$

7.2.3 A Conditional Typicality Bound

We will need a result related to the Markov Lemma in Section 5.2.

Theorem 7.1. Suppose $0 \le \epsilon_1 < \epsilon_2 \le \mu_{UXY}$, X_i is emitted by a DMS $P_{X|U}(\cdot|u_i)$ for $i = 1, 2, \ldots, n$, and $(u^n, y^n) \in T_{\epsilon_1}^n(P_{UY})$. We have

$$\begin{aligned}&(1 - \delta_{\epsilon_1,\epsilon_2}(n))\ 2^{-n[I(X;Y|U)+2\epsilon_2 H(X|U)]} \\ &\le \Pr\left[X^n \in T_{\epsilon_2}^n(P_{UXY}|u^n, y^n) \middle| U^n = u^n\right] \le 2^{-n[I(X;Y|U)-2\epsilon_2 H(X|U)]},\end{aligned} \tag{7.7}$$

where

$$\delta_{\epsilon_1,\epsilon_2}(n) = 2|\mathcal{U}||\mathcal{X}||\mathcal{Y}| \exp\left(-n \cdot \frac{(\epsilon_2 - \epsilon_1)^2}{1 + \epsilon_1} \cdot \mu_{UXY}\right). \tag{7.8}$$

Proof. The upper bound follows by (1.25) and (1.26):

$$\begin{aligned}&\Pr\left[X^n \in T_{\epsilon_2}^n(P_{UXY}|u^n, y^n) \middle| U^n = u^n\right] \\ &= \sum_{x^n \in T_{\epsilon_2}^n(P_{UXY}|u^n,y^n)} P_{X|U}^n(x^n|u^n) \\ &\le 2^{nH(X|UY)(1+\epsilon_2)}\, 2^{-nH(X|U)(1-\epsilon_2)} \\ &\le 2^{-n[I(X;Y|U)-2\epsilon_2 H(X|U)]}.\end{aligned} \tag{7.9}$$

The lower bound also follows from (1.25) and (1.26). □

7.3 The Capacity for $R_1 = R_2 = 0$

Suppose one is interested in broadcasting in the usual sense that there is only one message W_0. This problem is essentially the same as coding for a DMC in Section 3.4.

Code Construction: Generate 2^{nR_0} codewords $x^n(w_0)$, $w_0 = 1, 2, \ldots, 2^{nR_0}$, by choosing the $n \cdot 2^{nR_0}$ symbols $x_i(w)$ independently using a distribution $P_X(\cdot)$.

Encoder: Given w_0, transmit $x^n(w_0)$.

Decoder 1: Given y_1^n, try to find a $\tilde{w}_0$ such that $(x^n(\tilde{w}_0), y_1^n) \in T_\epsilon^n(P_{XY_1})$. If there is one or more such index, then choose one and put out the corresponding $\tilde{w}_0$ as $\hat{w}_0(1)$. If there is no such index, then put out $\hat{w}_0(1) = 1$.

Decoder 2: Proceed as decoder 1, but with y_2^n, $T_\epsilon^n(P_{XY_2})$, and $\hat{w}_0(2)$.

Analysis: Virtually the same analysis as in Section 3.4 establishes that one can achieve rates up to

$$R_0 = \max_{P_X(\cdot)} \min\left(I(X;Y_1), I(X;Y_2)\right). \tag{7.10}$$

For the converse, from Section 3.8 we know that reliable communication requires

$$\begin{aligned} nR_0 &\le I(W_0; \hat{W}_0(1)) \\ &\le I(X^n; Y_1^n) \\ &= \sum_{i=1}^n H(Y_{1i}|Y_1^{i-1}) - H(Y_{1i}|X_i) \\ &\le \sum_{i=1}^n H(Y_{1i}) - H(Y_{1i}|X_i) \\ &= n \sum_{i=1}^n \frac{1}{n} I(X_i; Y_{1i}) \\ &\le nI(\bar{X}; \bar{Y}_1), \end{aligned} \tag{7.11}$$

where the last step follows by the convexity of mutual information and by setting

$$P_{\bar{X}\bar{Y}_1\bar{Y}_2}(a,b,c) = \left[\frac{1}{n}\sum_{i=1}^n P_{X_i}(a)\right] P_{Y_1Y_2|X}(b,c|a) \tag{7.12}$$

for all appropriate a, b, c. We similarly have $nR_0 \le nI(\bar{X}; \bar{Y}_2)$ so that

$$R_0 \le \max_{P_X(\cdot)} \min\left(I(X;Y_1), I(X;Y_2)\right). \tag{7.13}$$

The rate (7.10) is thus the capacity. Note that the capacity is in general smaller than $\min(C_1, C_2)$, where C_1 and C_2 are the capacities of the respective channels $P_{Y_1|X}$ and $P_{Y_2|X}$.

7.4 An Achievable Region for $R_0 = 0$ via Binning

We construct a codebook for the case where there is no common message. Consider a distribution $P_{U_1U_2}(\cdot)$, and let $\mathcal{U}_1$ and $\mathcal{U}_2$ be the respective alphabets of U_1 and U_2. Consider also a function $f(\cdot)$ that maps symbols in $\mathcal{U}_1 \times \mathcal{U}_2$ to symbols in $\mathcal{X}$. We write $X = f(U_1, U_2)$ and $X^n = f^n(U_1^n, U_2^n)$ to mean that $X_i = f(U_{1i}, U_{2i})$ for $i = 1, 2, \ldots, n$. We generate codewords as for the Slepian–Wolf problem.

Code Construction: Generate $2^{n(R_1+R_1')}$ codewords $u_1^n(w_1, v_1)$, $w_1 = 1, 2, \ldots, 2^{nR_1}$, $v_1 = 1, 2, \ldots, 2^{nR_1'}$, by choosing the symbols $u_{1i}(w_1, v_1)$ independently using $P_{U_1}(\cdot)$. Similarly generate $2^{n(R_2+R_2')}$ codewords $u_2^n(w_2, v_2)$, $w_2 = 1, 2, \ldots, 2^{nR_2}$, $v_2 = 1, 2, \ldots, 2^{nR_2'}$, using $P_{U_2}(\cdot)$.

Encoder: Given w_1 and w_2, try to find a pair (v_1, v_2) such that $(u_1^n(w_1, v_1), u_2^n(w_2, v_2)) \in T_\epsilon^n(P_{U_1U_2})$. If there is one or more such (v_1, v_2), choose one and transmit $x^n = f^n(u_1^n(w_1, v_1), u_2^n(w_2, v_2))$. In practice, the decoder might know ahead of time which (v_1, v_2) is chosen. However, this is not necessary since the receivers will discard these indexes, as shown in the next step. One can, in fact, choose the (v_1, v_2) ahead of time for all bins, i.e., the pair (v_1, v_2) is a function of (w_1, w_2).

Decoder 1: Given y_1^n, try to find a pair $(\tilde{w}_1, \tilde{v}_1)$ such that $(u_1^n(\tilde{w}_1, \tilde{v}_1), y_1^n) \in T_\epsilon^n(P_{U_1Y_1})$. If there is one or more such pair, then choose one and put out the corresponding $\tilde{w}_1$ as $\hat{w}_1$. If there is no such pair, then put out $\hat{w}_1 = 1$.

Decoder 2: Proceed as decoder 1, except replace the index "1" by "2" everywhere.

Analysis: Let $0 < \epsilon_1 < \epsilon < \mu_{U_1U_2XY_1Y_2}$, where $\mu_{U_1U_2XY_1Y_2}$ is defined as usual to be the minimum positive probability of $P_{U_1U_2XY_1Y_2}(\cdot)$. Using (7.6), we find that the encoder finds an appropriate pair (v_1, v_2) with probability close to 1 as long as n is large, ϵ_1 is small, and

$$R_1' + R_2' > I(U_1; U_2). \tag{7.14}$$

So suppose the encoder was successful, and the likely event that

$$(u_1^n(w_1, v_1), u_2^n(w_2, v_2), x^n(u_1^n, u_2^n), y_1^n, y_2^n) \in T_{\epsilon_1}^n(P_{U_1U_2XY_1Y_2}) \tag{7.15}$$

occurred. (We remark that the event (7.15) is likely to occur only if $P_{U_1U_2XY_1Y_2}(\cdot)$ factors as $P_{U_1U_2X}(\cdot)P_{Y_1Y_2|X}(\cdot)$.) Decoder 1 is likely to make an error if there is a pair $\tilde{w}_1, \tilde{v}_1$ with $\tilde{w}_1 \neq w_1$ such that $(u_1^n(\tilde{w}_1, \tilde{v}_1), y_1^n) \in T_\epsilon^n(P_{U_1Y_1})$. But the probability of this event can be made small if

$$R_1 + R_1' < I(U_1;Y_1). \tag{7.16}$$

The corresponding event for decoder 2 can be made to have small probability if

$$R_2 + R_2' < I(U_2;Y_2). \tag{7.17}$$

To see what rates (R_1, R_2) are achievable with (7.14)–(7.17), suppose we choose $R_1' = \alpha I(U_1;U_2)$ for $0 \leq \alpha \leq 1$. We then achieve

$$(R_1, R_2) = (\ I(U_1;Y_1) - \alpha I(U_1;U_2), I(U_2;Y_2) - (1-\alpha)I(U_1;U_2)\ ).$$

Alternatively, the achievable rate region is defined by the pentagon

$$\begin{aligned} &0 \leq R_1 \leq I(U_1;Y_1) \\ &0 \leq R_2 \leq I(U_2;Y_2) \\ &R_1 + R_2 \leq I(U_1;Y_1) + I(U_2;Y_2) - I(U_1;U_2), \end{aligned} \tag{7.18}$$

where $[U_1, U_2] - X - [Y_1, Y_2]$ forms a Markov chain. This result is due to Marton [43] and the region is depicted in Figure 7.2.

Consider, e.g., the corner point with $\alpha = 1$. Note that the rate $R_1 = I(U_1;Y_1) - I(U_1;U_2)$ is identical to the Gelfand-Pinsker rate R_{GP} if we

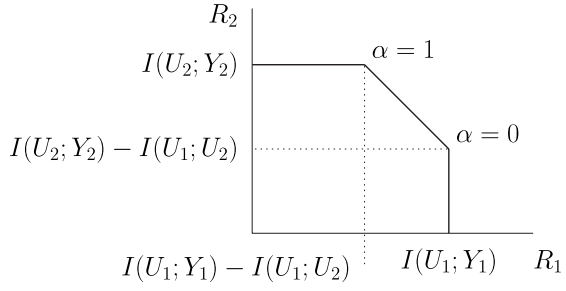

Fig. 7.2 An achievable region for $R_0 = 0$.

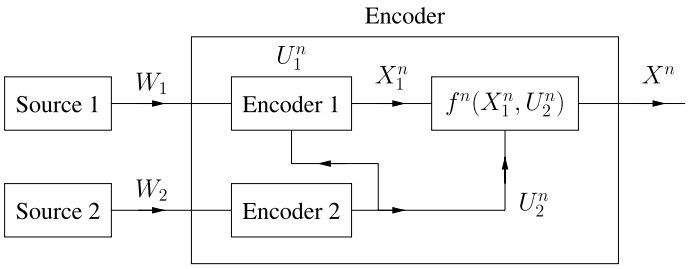

Fig. 7.3 An encoder structure inspired by the Gelfand-Pinsker problem and solution.

consider u_2^n to be interference known noncausally at the transmitter. This suggests designing an encoder as shown in Figure 7.3. The overall encoder consists of two encoders: one for u_2^n and one for u_1^n based on Gelfand-Pinsker coding. The output x_1^n is a function of u_1^n and u_2^n as in the Gelfand-Pinsker problem. We remark that, in general, decoders 1 and 2 will be able to decode *only* the messages w_1 and w_2, respectively. The next coding scheme we consider, superposition coding, has one decoder decoding both messages.

Example 7.1. Suppose the broadcast channel is *deterministic* in the sense that

$$Y_1 = f_1(X) \quad \text{and} \quad Y_2 = f_2(X) \tag{7.19}$$

for some functions $f_1(\cdot)$ and $f_2(\cdot)$. We can choose $U_1 = Y_1$ and $U_2 = Y_2$ since $[U_1, U_2] - X - [Y_1, Y_2]$ forms a Markov chain. The bounds (7.18) are thus

$$\begin{aligned} 0 &\le R_1 \le H(Y_1) \\ 0 &\le R_2 \le H(Y_2) \\ R_1 + R_2 &\le H(Y_1Y_2). \end{aligned} \tag{7.20}$$

The resulting region turns out to be the capacity region $\mathcal{C}_{BC}$ for this problem. Thus, binning is optimal for deterministic broadcast channels.

7.5 Superposition Coding

We next introduce *superposition coding* that is a method for "stacking" codebooks. This method turns out to be optimal for an important

class of channels known as *degraded* broadcast channels. Furthermore, a judicious combination of superposition coding and binning gives the currently best achievable rate region for broadcast channels. We develop this region in Section 7.8.

For simplicity, suppose for now that $R_2 = 0$. Consider a distribution $P_{UXY_1Y_2}(\cdot)$ that factors as $P_{UX}(\cdot)P_{Y_1Y_2|X}(\cdot)$, and where the alphabet of U is $\mathcal{U}$.

Code Construction: Consider $P_{UX}(\cdot)$. Generate 2^{nR_0} codewords $u^n(w_0)$, $w_0 = 1, 2, \ldots, 2^{nR_0}$, by using $P_U(\cdot)$. Next, for every $u^n(w_0)$, generate 2^{nR_1} codewords $x^n(w_0, w_1)$ by choosing the symbols $x_i(w_0, w_1)$ independently at random according to $P_{X|U}(\cdot|u_i(w_0))$ for $i = 1, 2, \ldots, n$ and $w_1 = 1, 2, \ldots, 2^{nR_1}$. This second step is called *superposition coding*, and it is depicted in Figure 7.4. In the "space" of all codewords, one can view the $u^n(w_0)$ as cloud centers, and the $x^n(w_0, w_1)$ as satellites (see Figure 7.5).

Encoder: Given w_0 and w_1, transmit $x^n(w_0, w_1)$.

Decoder 1: Given y_1^n, try to find a pair $(\tilde{w}_0, \tilde{w}_1)$ such that $(u^n(\tilde{w}_0), x^n(\tilde{w}_0, \tilde{w}_1), y_1^n) \in T_\epsilon^n(P_{UXY_1})$. If there is one or more such pair, then choose one and call it $(\hat{w}_0(1), \hat{w}_1)$. If there is no such pair, then put out $(\hat{w}_0(1), \hat{w}_1) = (1, 1)$.

Decoder 2: Given y_2^n, try to find a $\tilde{w}_0$ such that $(u^n(\tilde{w}_0), y_2^n) \in T_\epsilon^n(P_{UY_2})$. If there is one or more such index, then choose one and call it $\hat{w}_0(2)$. If there is no such index, then put out $\hat{w}_0(2) = 1$.

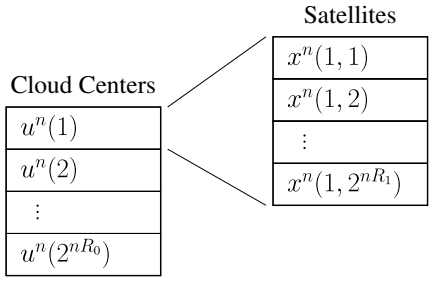

Fig. 7.4 Codebooks for superposition coding.

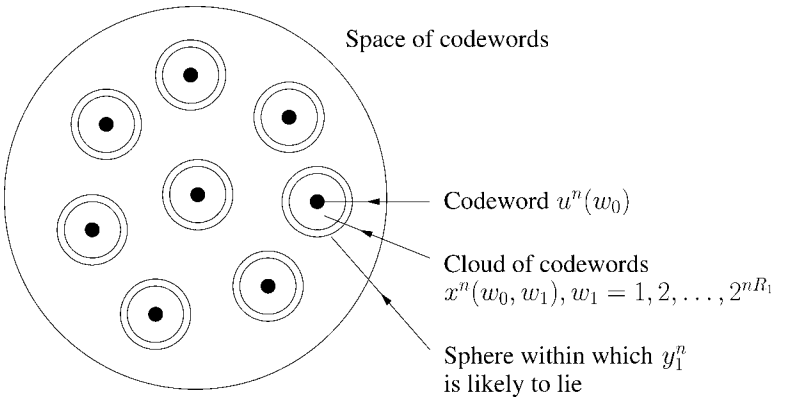

Fig. 7.5 Space of codewords for superposition coding (reproduced with modifications from [15, Figure 4]).

Analysis: Let $0 < \epsilon_1 < \epsilon < \mu_{UXY_1Y_2}$. We have $(u^n(w_0), x^n(w_0, w_1), y_1^n, y_2^n) \in T_{\epsilon_1}^n(P_{UXY_1Y_2})$ with probability close to one for large n. Consider first decoder 2. The probability that this decoder finds an incorrect $\tilde{w}_0$ can be made small if

$$R_0 < I(U; Y_2). \tag{7.21}$$

Next, consider decoder 1 which must decode both w_0 and w_1. We split the potential error events into three disjoint parts: first, that $\hat{w}_0(1) \neq w_0, \hat{w}_1 = w_1$; second, that $\hat{w}_0(1) = w_0, \hat{w}_1 \neq w_1$; and finally that $\hat{w}_0(1) \neq w_0, \hat{w}_1 \neq w_1$. The probability of the first event is

$$\begin{aligned} &\Pr\left[\bigcup_{\tilde{w}_0 \neq w_0} \{(U^n(\tilde{w}_0), X^n(\tilde{w}_0, w_1), y_1^n) \in T_\epsilon^n(P_{UXY_1})\}\right] \\ &\quad \leq 2^{n[R_0 - I(UX;Y_1) + 2\epsilon H(UX)]}. \end{aligned} \tag{7.22}$$

The probability of the second event is

$$\begin{aligned} &\Pr\left[\bigcup_{\tilde{w}_1 \neq w_1} \{(u^n(w_0), X^n(w_0, \tilde{w}_1), y_1^n) \in T_\epsilon^n(P_{UXY_1})\}\right] \\ &\quad \leq 2^{n[R_1 - I(X;Y_1|U) + 2\epsilon H(X|U)]}, \end{aligned} \tag{7.23}$$

where we have used (7.7). The probability of the third event is

$$\Pr\left[\bigcup_{\tilde{w}_0 \neq w_0} \bigcup_{\tilde{w}_1 \neq w_1} \{(U^n(\tilde{w}_0), X^n(\tilde{w}_0, \tilde{w}_1), y_1^n) \in T_\epsilon^n(P_{UXY_1})\}\right] \leq 2^{n[R_0+R_1-I(UX;Y_1)+2\epsilon H(UX)]}. \tag{7.24}$$

Note that (7.24) makes (7.22) unnecessary. Combining (7.21)–(7.24), we have that $(R_0, R_1, 0) \in \mathcal{C}_{BC}$ if

$$\begin{aligned} 0 &\leq R_0 \leq I(U;Y_2) \\ 0 &\leq R_1 \leq I(X;Y_1|U) \\ R_0 + R_1 &\leq I(X;Y_1), \end{aligned} \tag{7.25}$$

where $U - X - [Y_1, Y_2]$ forms a Markov chain. The above coding scheme is due to Bergmans [7] and is based on work by Cover [15]. One can restrict attention to $\mathcal{U}$ satisfying $|\mathcal{U}| \leq |\mathcal{Y}| + 2$. Moreover, the region (7.25) turns out to be the capacity region when $R_2 = 0$ (see [35]). To get a general achievable region, we use the first property described in Section 7.2.1: we convert some of the R_0 bits into R_2 bits. The resulting region of (R_0, R_1, R_2) is simply (7.25) with R_0 replaced by $R_0 + R_2$.

7.6 Degraded Broadcast Channels

Consider next a class of channels called *degraded channels* that have the special property that decoder 1 can decode anything that decoder 2 can decode. A broadcast channel is said to be *physically degraded* if

$$X - Y_1 - Y_2,$$

forms a Markov chain. A broadcast channel $P_{Y_1Y_2|X}(\cdot)$ is said to be *degraded* or *stochastically degraded* if it has the same marginals $P_{Y_1|X}(\cdot)$ and $P_{Y_2|X}(\cdot)$ as some physically degraded channel. Another way of stating this is that

$$P_{Y_2|X}(c|a) = \sum_{b \in \mathcal{Y}_1} P_{Y_1|X}(b|a)\, P^*_{Y_2|Y_1}(c|b), \tag{7.26}$$

for some $P^*_{Y_2|Y_1}(\cdot)$. The capacity region of a degraded broadcast channel is thus the same as its physically degraded counterpart, and we will study this physically degraded channel.

Consider, then, a physically degraded broadcast channel $P_{Y_1Y_2|X}(\cdot)$. Suppose we encode using superposition coding as described above. We have

$$I(U;Y_2) \leq I(U;Y_1) \tag{7.27}$$

because $U - X - Y_1 - Y_2$ forms a Markov chain. We thus also have

$$I(U;Y_2) + I(X;Y_1|U) \leq I(U;Y_1) + I(X;Y_1|U) = I(X;Y_1), \tag{7.28}$$

which means that the third bound in (7.25) is unnecessary. The resulting achievable region is the set of non-negative (R_0, R_1, R_2) satisfying

$$\begin{aligned} R_1 &\leq I(X;Y_1|U) \\ R_0 + R_2 &\leq I(U;Y_2), \end{aligned} \tag{7.29}$$

where $U - X - Y_1 - Y_2$ forms a Markov chain. The union of these achievable regions over all $P_{UX}(\cdot)$ turns out to be the capacity region of the degraded broadcast channel [27]. We prove the converse theorem in the appendix of this section.

Example 7.2. Consider the (physically degraded) binary symmetric broadcast channel (BSBC). This channel has $\mathcal{X} = \mathcal{Y}_1 = \mathcal{Y}_2 = \{0,1\}$,

$$Y_1 = X \oplus Z_1 \quad \text{and} \quad Y_2 = X \oplus Z_2, \tag{7.30}$$

where $P_{Z_1}(1) = 1 - P_{Z_1}(0) = p_1$, $P_{Z_2}(1) = 1 - P_{Z_2}(0) = p_2$, and X, Z_1, and Z_2 are statistically independent. Suppose that $p_1 \leq p_2 \leq 1/2$. We can convert (7.30) to a *physically degraded* channel by writing

$$Y_1 = X \oplus Z_1 \quad \text{and} \quad Y_2 = X \oplus Z_1 \oplus Z_2', \tag{7.31}$$

where $P_{Z_2'}(1) = 1 - P_{Z_2'}(0) = p_2 - p_1$ and Z_2' is independent of X and Z_1. We choose $P_U(0) = P_U(1) = 1/2$ and set $\Pr[X \neq U] = q$. Evaluating (7.29), we have

$$\begin{aligned} R_1 &\leq H_2(q * p_1) - H_2(p_1) \\ R_0 + R_2 &\leq 1 - H_2(q * p_2), \end{aligned} \tag{7.32}$$

where $p * q = p(1-q) + (1-p)q$. This region is depicted in Figure 7.6, and it defines the capacity region of this channel [69, 70].

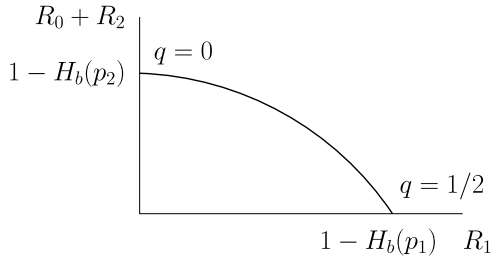

Fig. 7.6 The capacity region of a BSBC.

7.7 Coding for Gaussian Channels

This section describes two coding methods for scalar additive white Gaussian noise (AWGN) broadcast channels and one method for vector AWGN broadcast channels. The methods are based on superposition coding and binning, and the motivation for describing the different methods is to illustrate two main points. First, we show that one can achieve capacity in more than one way for scalar AWGN channels. Second, we show how to apply binning to vector AWGN channels.

7.7.1 Superposition Coding for Scalar AWGN Channels

Suppose we have a (scalar) AWGN broadcast channel

$$\begin{aligned} Y_1 &= X + Z_1 \\ Y_2 &= X + Z_2, \end{aligned} \tag{7.33}$$

where $\mathcal{X}$ is the set of real numbers, we have the per-symbol power (or energy) constraint $E\left[X^2\right] \leq P$, and Z_1 and Z_2 are (possibly correlated) Gaussian random variables with respective variances σ_1^2 and σ_2^2. Suppose that $\sigma_1^2 \leq \sigma_2^2$. We can convert (7.33) to a *physically degraded* channel by writing

$$\begin{aligned} Y_1 &= X + Z_1 \\ Y_2 &= X + Z_1 + Z_2', \end{aligned} \tag{7.34}$$

where Z_2' is Gaussian, independent of X and Z_1, and has variance $\sigma_2^2 - \sigma_1^2$. For superposition coding, we choose

$$X = U + V, \tag{7.35}$$

where U and V are *independent* Gaussian random variables with respective variances αP and $(1-\alpha)P$ for some α satisfying $0 \le \alpha \le 1$. We consider (7.29) and compute

$$\begin{aligned} I(U;Y_2) &= h(Y_2) - h(Y_2|U) \\ &= \frac{1}{2}\log\left(2\pi e\,[P+\sigma_2^2]\right) - \frac{1}{2}\log\left(2\pi e\,[(1-\alpha)P+\sigma_2^2]\right) \\ &= \frac{1}{2}\log\left(1+\frac{\alpha P}{(1-\alpha)P+\sigma_2^2}\right) \qquad (7.36) \\ I(X;Y_1|U) &= h(Y_1|U) - h(Y_1|X) \\ &= \frac{1}{2}\log\left(2\pi e\,[(1-\alpha)P+\sigma_1^2]\right) - \frac{1}{2}\log\left(2\pi e\sigma_1^2\right) \\ &= \frac{1}{2}\log\left(1+\frac{(1-\alpha)P}{\sigma_1^2}\right). \qquad (7.37) \end{aligned}$$

The achievable (R_1, R_2) are determined by varying α, and they are depicted in Figure 7.7. Observe that the region dominates the *time-sharing* region, whose boundary is given by the dashed line. One can show that (7.36) and (7.37) define the capacity region by using Shannon's *entropy power inequality* (see the appendix of this section).

Finally, we point out the following interesting fact about (7.35). We can encode by generating two code books of sizes 2^{nR_1} and 2^{nR_2} with codewords $v^n(w_1)$, $w_1 = 1,2,\ldots,2^{nR_1}$, and $u^n(w_2)$, $w_2 = 1,2,\ldots,2^{nR_2}$, and by using $x^n = f^n(v^n(w_1), u^n(w_2))$ for some per-letter function $f(\cdot)$. This superposition coding scheme is closely related to the scheme described above, but it is simpler. Superposition coding is often done either as in Section 7.5 or as suggested by (7.35).

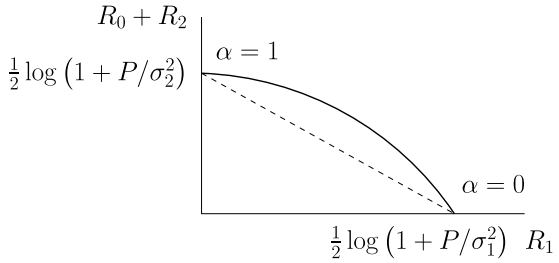

Fig. 7.7 The capacity region of an AWGN broadcast channel.

7.7.2 Binning for Scalar AWGN Channels

We use *binning* for the AWGN broadcast channel, as described in Section 7.4. Consider the encoder structure depicted in Figure 7.3. We choose four random variables:

$$\begin{aligned}
&U_2 \text{ Gaussian with variance } \alpha P \\
&X_1 \text{ Gaussian, independent of } U_2, \text{ with variance } (1-\alpha)P \\
&U_1 = X_1 + \beta U_2 \\
&\beta = (1-\alpha)P \big/ [(1-\alpha)P + \sigma_1^2] \\
&X = X_1 + U_2.
\end{aligned}$$

Using (7.18) and Section 6.4, we compute

$$\begin{aligned}
I(U_2;Y_2) &= \frac{1}{2}\log\left(1 + \frac{\alpha P}{(1-\alpha)P + \sigma_2^2}\right) \\
I(U_1;Y_1) - I(U_1;U_2) &= \frac{1}{2}\log\left(1 + \frac{(1-\alpha)P}{\sigma_1^2}\right).
\end{aligned} \tag{7.38}$$

That is, we can achieve all points inside the capacity region of the AWGN broadcast channel by using a Gelfand-Pinsker encoder!

7.7.3 Binning for Vector AWGN Channels

Motivated by the above result, we use the same approach for AWGN *vector* broadcast channels. We use column vectors to represent the channel input and outputs, and write $\mathbf{Q}_{\underline{X}}$ for $E\left[\underline{X}\,\underline{X}^T\right]$. The model is

$$\begin{aligned}
\underline{Y}_1 &= \mathbf{H}_1\,\underline{X} + \underline{Z}_1 \\
\underline{Y}_2 &= \mathbf{H}_2\,\underline{X} + \underline{Z}_2,
\end{aligned} \tag{7.39}$$

where $\underline{X}$ has length M, $\underline{Y}_1$ has length N_1, $\underline{Y}_2$ has length N_2, $\mathbf{H}_1$ has dimension $N_1 \times M$, $\mathbf{H}_2$ has dimension $N_2 \times M$, and $\underline{Z}_1$ and $\underline{Z}_2$ are (possible correlated) Gaussian random vectors with respective lengths N_1 and N_2, and with respective positive-definite covariance matrices $\mathbf{Q}_{\underline{Z}_1}$ and $\mathbf{Q}_{\underline{Z}_2}$. We consider the case $M \geq N_1$ and $M \geq N_2$. The power constraint is $E\left[\|\underline{X}\|^2\right] \leq P$ or, equivalently, $\mathrm{tr}\left[\mathbf{Q}_{\underline{X}}\right] \leq P$.

Note that the channel is *not* necessarily degraded, so one cannot necessarily expect superposition coding to be optimal. However, recall

from the appendix of Section 6 that one *can* operate a Gelfand-Pinsker encoder at the same rate as if the interference was not present. We choose four random variables:

$$\begin{aligned}
&\underline{U}_2 \text{ Gaussian} \\
&\underline{X}_1 \text{ Gaussian, independent of } \underline{U}_2 \\
&\underline{U}_1 = \underline{X}_1 + \mathbf{B}\underline{U}_2 \\
&\mathbf{B} = \mathbf{Q}_{\underline{X}_1}\mathbf{H}_1^T\left[\mathbf{Q}_{\underline{Z}_1} + \mathbf{H}_1\mathbf{Q}_{\underline{X}_1}\mathbf{H}_1^T\right]^{-1} \\
&\underline{X} = \underline{X}_1 + \underline{U}_2.
\end{aligned}$$

Using (7.18) and the appendix of Section 6, we compute

$$I(\underline{U}_2;\underline{Y}_2) = \frac{1}{2}\log\frac{\det\left(\mathbf{Q}_{\underline{Z}_2} + \mathbf{H}_2(\mathbf{Q}_{\underline{X}_1} + \mathbf{Q}_{\underline{U}_2})\mathbf{H}_2^T\right)}{\det\left(\mathbf{Q}_{\underline{Z}_2} + \mathbf{H}_2\mathbf{Q}_{\underline{X}_1}\mathbf{H}_2^T\right)}$$

$$I(\underline{U}_1;\underline{Y}_1) - I(\underline{U}_1;\underline{U}_2) = \frac{1}{2}\log\frac{\det\left(\mathbf{Q}_{\underline{Z}_1} + \mathbf{H}_1\mathbf{Q}_{\underline{X}_1}\mathbf{H}_1^T\right)}{\det\mathbf{Q}_{\underline{Z}_1}}. \tag{7.40}$$

It remains to optimize over $\mathbf{Q}_{\underline{X}_1}$ and $\mathbf{Q}_{\underline{U}_2}$. Recent research [66] has shown that (7.40) defines the capacity region for this channel when $R_0 = 0$.

7.8 Marton's Achievable Region

The best known achievable-rate region for broadcast channels is due to Marton [19, 41, 43] and it requires using superposition coding and binning. We briefly develop this region here. A discussion of some of the subtleties of Marton's region is given in [41, Sec. III.A].

Code Construction: Consider a distribution $P_{TU_1U_2}(\cdot)$ and a function $f(\cdot)$ mapping symbols in $\mathcal{T} \times \mathcal{U}_1 \times \mathcal{U}_2$ to symbols in $\mathcal{X}$. Generate 2^{nR_0} codewords $t^n(w_0)$, $w_0 = 1, 2, \ldots, 2^{nR_0}$, by using $P_T(\cdot)$. Next, for every $t^n(w_0)$, use the code construction of Section 7.4. Generate $2^{n(R_1+R_1')}$ codewords $u_1^n(w_0, w_1, v_1)$, $w_1 = 1, 2, \ldots, 2^{nR_1}$, $v_1 = 1, 2, \ldots, 2^{nR_1'}$, by choosing the symbols $u_{1i}(w_0, w_1, v_1)$ independently using $P_{U_1|T}(\cdot|t_i(w_0))$. Similarly generate $2^{n(R_2+R_2')}$ codewords $u_2^n(w_0, w_2, v_2)$, $w_2 = 1, 2, \ldots, 2^{nR_2}$, $v_2 = 1, 2, \ldots, 2^{nR_2'}$, by using $P_{U_2|T}(\cdot|t_i(w_0))$.

Encoder: Given w_0, w_1, and w_2, try to find a pair (v_1, v_2) such that $(t^n(w_0), u_1^n(w_0, w_1, v_1), u_2^n(w_0, w_2, v_2)) \in T_\epsilon^n(P_{TU_1U_2})$. If there is one or more such (v_1, v_2), choose one and transmit $x^n = f^n(t^n(w_0), u_1^n(w_0, w_1, v_1), u_2^n(w_0, w_2, v_2))$.

Decoder 1: Given y_1^n, try to find a triple $(\tilde{w}_0, \tilde{w}_1, \tilde{v}_1)$ such that $(t^n(\tilde{w}_0), u_1^n(\tilde{w}_0, \tilde{w}_1, \tilde{v}_1), y_1^n) \in T_\epsilon^n(P_{TU_1Y_1})$. If there is one or more such triple, then choose one and put out the corresponding $(\tilde{w}_0, \tilde{w}_1)$ as $(\hat{w}_0(1), \hat{w}_1)$. If there is no such pair, then put out $(\hat{w}_0(1), \hat{w}_1) = (1, 1)$.

Decoder 2: Proceed as decoder 1, except replace the index "1" by "2" everywhere.

Using the analysis procedure that we are by now accustomed to, the rate bounds are (see (7.14), (7.16), (7.17), and (7.25))

$$R_1' + R_2' > I(U_1; U_2|T) \tag{7.41}$$
$$R_1 + R_1' < I(U_1; Y_1|T) \tag{7.42}$$
$$R_0 + R_1 + R_1' < I(TU_1; Y_1) \tag{7.43}$$
$$R_2 + R_2' < I(U_2; Y_2|T) \tag{7.44}$$
$$R_0 + R_2 + R_2' < I(TU_2; Y_2). \tag{7.45}$$

We can remove R_1' and R_2' to get the bounds

$$R_1 < I(U_1; Y_1|T) \tag{7.46}$$
$$R_2 < I(U_2; Y_2|T) \tag{7.47}$$
$$R_1 + R_2 < I(U_1; Y_1|T) + I(U_2; Y_2|T) - I(U_1; U_2|T) \tag{7.48}$$
$$R_0 + R_1 < I(TU_1; Y_1) \tag{7.49}$$
$$R_0 + R_2 < I(TU_2; Y_2) \tag{7.50}$$
$$R_0 + R_1 + R_2 < I(TU_1; Y_1) + I(U_2; Y_2|T) - I(U_1; U_2|T) \tag{7.51}$$
$$R_0 + R_1 + R_2 < I(U_1; Y_1|T) + I(TU_2; Y_2) - I(U_1; U_2|T). \tag{7.52}$$
$$2R_0 + R_1 + R_2 < I(TU_1; Y_1) + I(TU_2; Y_2) - I(U_1; U_2|T). \tag{7.53}$$

However, it turns out that we can do better. Recall from Section 7.2.1 that if (R_0, R_1, R_2) is achievable, then so is $(\alpha_0 R_0, R_1 + \alpha_1 R_0, R_2 + \alpha_2 R_0)$, where $\alpha_i \geq 0$, $i = 0, 1, 2$, and $\alpha_0 + \alpha_1 + \alpha_2 = 1$. Applying this

idea to (7.46)–(7.52), we get the rate bounds (see [41, Thm. 5])

$$R_0 + R_1 < I(TU_1; Y_1) \tag{7.54}$$
$$R_0 + R_2 < I(TU_2; Y_2) \tag{7.55}$$
$$R_0 + R_1 + R_2 < I(TU_1; Y_1) + I(U_2; Y_2|T) - I(U_1; U_2|T) \tag{7.56}$$
$$R_0 + R_1 + R_2 < I(U_1; Y_1|T) + I(TU_2; Y_2) - I(U_1; U_2|T) \tag{7.57}$$
$$2R_0 + R_1 + R_2 < I(TU_1; Y_1) + I(TU_2; Y_2) - I(U_1; U_2|T). \tag{7.58}$$

Finally, we can take the closure of the union over all $P_{TU_1U_2}(\cdot)$ and all $f(\cdot)$ of the rates satisfying (7.54)–(7.58). The resulting region is Marton's achievable-rate region.

7.9 Capacity Region Outer Bounds

A simple outer bound on $\mathcal{C}_{BC}$ was given by Cover [15]. Clearly, based on our results for a DMC, one must have $R_0 + R_1 \leq \max_{P_X(\cdot)} I(X; Y_1)$ and $R_0 + R_2 \leq \max_{P_X(\cdot)} I(X; Y_2)$. However, rather than optimizing $P_X(\cdot)$ for each mutual information separately, the same steps as in (7.11) and (7.12) can be used to show that one can consider the same $P_X(\cdot)$ for both bounds simultaneously. One can further add a bound based on letting the receivers cooperate. Summarizing the result, let $\mathcal{R}(P_X)$ be the set of (R_0, R_1, R_2) permitted by

$$\begin{aligned} R_0 + R_1 &\leq I(X; Y_1) \\ R_0 + R_2 &\leq I(X; Y_2) \\ R_0 + R_1 + R_2 &\leq I(X; Y_1 Y_2), \end{aligned} \tag{7.59}$$

when the distribution $P_X(\cdot)$ is fixed. The result is

$$\mathcal{C}_{BC} \subseteq \bigcup_{P_X(\cdot)} \mathcal{R}(P_X). \tag{7.60}$$

7.9.1 Sato's Outer Bound

Another simple but useful bound on $\mathcal{C}_{BC}$ was determined by Sato [54]. Let $\mathcal{P}(P_{Y_1|X}, P_{Y_2|X})$ be the set of broadcast channels that have the marginals $P_{Y_1|X}(\cdot)$ and $P_{Y_2|X}(\cdot)$. Suppose we let the receivers cooperate

for any channel in $\mathcal{P}(P_{Y_1|X}, P_{Y_2|X})$. Sato's sum-rate bound is

$$R_0 + R_1 + R_2 \le \min \max_{P_X(\cdot)} I(X; Y_1 Y_2), \tag{7.61}$$

where the minimization is over all $P_{Y_1 Y_2|X}(\cdot) \in \mathcal{P}(P_{Y_1|X}, P_{Y_2|X})$.

7.9.2 Körner and Marton's Outer Bound

Yet a third outer bound is due to Körner and Marton [43, Thm. 5]. Following similar steps as in Section 6.6, reliable communication requires

$$\begin{aligned} n(R_1 + R_2) &\le I(W_1; \hat{W}_1) + I(W_2; \hat{W}_2) \\ &\le I(W_1; Y_1^n W_2) + I(W_2; Y_2^n) \\ &= \sum_{i=1}^{n} I(W_1; Y_{1i} | W_2 Y_{1(i+1)}^n) \\ &\quad + \sum_{i=1}^{n} I(W_2 Y_{1(i+1)}^n; Y_2^i) - I(W_2 Y_{1i}^n; Y_2^{i-1}), \end{aligned} \tag{7.62}$$

where the third step follows by setting $Y_{20} = 0$. We continue the chain of (in)equalities (7.62):

$$\begin{aligned} n(R_1 + R_2) &\le \sum_{i=1}^{n} I(W_1; Y_{1i} | W_2 Y_{1(i+1)}^n) \\ &\quad + \big[I(W_2 Y_{1(i+1)}^n; Y_2^{i-1}) + I(W_2 Y_{1(i+1)}^n; Y_{2i} | Y_2^{i-1}) \big] \\ &\quad - \big[I(W_2 Y_{1(i+1)}^n; Y_2^{i-1}) + I(Y_{1i}; Y_2^{i-1} | W_2 Y_{1(i+1)}^n) \big] \\ &= \sum_{i=1}^{n} I(W_1; Y_{1i} | W_2 Y_{1(i+1)}^n) \\ &\quad + \big[I(W_2 Y_{1(i+1)}^n; Y_{2i} | Y_2^{i-1}) - I(Y_{1i}; Y_2^{i-1} | W_2 Y_{1(i+1)}^n) \big] \\ &= \sum_{i=1}^{n} \big[H(Y_{2i} | Y_2^{i-1}) - H(Y_{2i} | U_i) \big] + [H(Y_{1i} | U_i) - H(Y_{1i} | X_i)] \\ &\le \sum_{i=1}^{n} [H(Y_{2i}) - H(Y_{2i} | U_i)] + [H(Y_{1i} | U_i) - H(Y_{1i} | X_i U_i)] \\ &= \sum_{i=1}^{n} I(U_i; Y_{2i}) + I(X_i; Y_{1i} | U_i), \end{aligned} \tag{7.63}$$

where for the second step we have defined $U_i = [W_2, Y_{1(i+1)}^n, Y_2^{i-1}]$. We further have that $U_i - X_i - [Y_{1i}, Y_{2i}]$ forms a Markov chain. Combining (7.63) with a few more steps, one can show that, in the plane defined by $R_0 = 0$, $\mathcal{C}_{BC}$ is inside the set of non-negative (R_1, R_2) satisfying

$$\begin{aligned} R_1 &\leq I(X;Y_1) \\ R_2 &\leq I(U;Y_2) \\ R_1 + R_2 &\leq I(X;Y_1|U) + I(U;Y_2), \end{aligned} \tag{7.64}$$

for some $P_{UXY_1Y_2}(\cdot)$ that factors as $P_{UX}(\cdot)P_{Y_1Y_2|X}(\cdot)$, and where $|\mathcal{U}| \leq |\mathcal{X}| + 2$.

7.10 Appendix: Binning Bound and Capacity Converses

7.10.1 Bound on Binning Rates

Consider $P_{\text{bin},e}(w_1, w_2)$ given in (7.5). Let $I(v_1, v_2)$ be the indicator random variable that the event

$$\{(X^n(w_1, v_1), Y^n(w_2, v_2)) \in T_\epsilon^n(P_{XY})\} \tag{7.65}$$

occurred. Let $S = \sum_{v_1, v_2} I(v_1, v_2)$, $\bar{S} = E[S]$, and $\mathsf{Var}[S] = E\left[(S - \bar{S})^2\right] = E\left[S^2\right] - \bar{S}^2$. We bound

$$\begin{aligned} P_{\text{bin},e}(w_1, w_2) &= \Pr[S = 0] \\ &\leq \Pr\left[(S - \bar{S})^2 \geq \bar{S}^2\right] \\ &\leq \mathsf{Var}[S] \big/ \bar{S}^2, \end{aligned} \tag{7.66}$$

where the last step follows by the Markov inequality for non-negative random variables: $\Pr[W \geq \alpha] \leq E[W]/\alpha$. We bound

$$\begin{aligned} \bar{S} &= \sum_{v_1, v_2} E[I(v_1, v_2)] \\ &= \sum_{v_1, v_2} \Pr[(X^n, Y^n) \in T_\epsilon^n(P_{XY})] \\ &\geq \sum_{v_1, v_2} (1 - \delta_\epsilon(n)) \cdot 2^{-n[I(X;Y) + 3\epsilon H(XY)]} \\ &= (1 - \delta_\epsilon(n)) \cdot 2^{n[R_1' + R_2' - I(X;Y) - 3\epsilon H(XY)]}. \end{aligned} \tag{7.67}$$

We also have

$$\mathsf{Var}[S] = \sum_{v_1,v_2} \sum_{\tilde{v}_1,\tilde{v}_2} \{E[I(v_1,v_2)I(\tilde{v}_1,\tilde{v}_2)] - E[I(v_1,v_2)]\,E[I(\tilde{v}_1,\tilde{v}_2)]\}. \tag{7.68}$$

Observe that if $\tilde{v}_1 \neq v_1$ and $\tilde{v}_2 \neq v_2$, then $I(v_1,v_2)$ and $I(\tilde{v}_1,\tilde{v}_2)$ are independent, and the summand in (7.68) is zero. Next, if $\tilde{v}_1 \neq v_1$ but $\tilde{v}_2 = v_2$, then we can bound

$$\begin{aligned} E[I(v_1,v_2)I(\tilde{v}_1,\tilde{v}_2)] &= \Pr[\{I(v_1,v_2)=1\} \cap \{I(\tilde{v}_1,v_2)=1\}] \\ &= \Pr[I(v_1,v_2)=1] \cdot \Pr\big[I(\tilde{v}_1,v_2)=1 \big| I(v_1,v_2)=1\big] \\ &\leq 2^{-n[I(X;Y)-3\epsilon H(XY)]} \cdot \Pr\big[(X^n,y^n) \in T_\epsilon^n(P_{XY}) \big| y^n \in T_\epsilon^n(P_Y)\big] \\ &\leq 2^{-n[2I(X;Y)-3\epsilon H(XY)-2\epsilon H(X)]}. \end{aligned} \tag{7.69}$$

By symmetry, we can derive a similar bound for $\tilde{v}_1 = v_1$ and $\tilde{v}_2 \neq v_2$. Finally, if $\tilde{v}_1 = v_1$ and $\tilde{v}_2 = v_2$, then we have

$$\begin{aligned} E[I(v_1,v_2)I(\tilde{v}_1,\tilde{v}_2)] &= E[I(v_1,v_2)] \\ &= \Pr[(X^n,Y^n) \in T_\epsilon^n(P_{XY})] \\ &\leq 2^{-n[I(X;Y)-3\epsilon H(XY)]}. \end{aligned} \tag{7.70}$$

Combining the results, we have

$$\begin{aligned} \mathsf{Var}[S] \leq 2^{n(R_1'+R_2')}\big(&2^{-n[I(X;Y)-3\epsilon H(XY)]} \\ &+ (2^{nR_1'} + 2^{nR_2'}) \cdot 2^{-n[2I(X;Y)-5\epsilon H(XY)]}\big). \end{aligned} \tag{7.71}$$

Using (7.66), we also have

$$\begin{aligned} &P_{\mathrm{bin},e}(w_1,w_2) \\ &\leq \frac{2^{-n(R_1'+R_2')}}{(1-\delta_\epsilon(n))^2} \cdot \big(2^{n[I(X;Y)+9\epsilon H(XY)]} + (2^{nR_1'} + 2^{nR_2'}) \cdot 2^{n11\epsilon H(XY)}\big) \\ &\leq \frac{2^{-n[R_1'+R_2'-I(X;Y)-9\epsilon H(XY)]}}{(1-\delta_\epsilon(n))^2} + \frac{2^{nR_1'} + 2^{nR_2'}}{(1-\delta_\epsilon(n))^2\, 2^{n(R_1'+R_2')}} \cdot 2^{n11\epsilon H(XY)}. \end{aligned} \tag{7.72}$$

The second term in (7.72) is small if $R_1' > 0$, $R_2' > 0$, $\min(R_1', R_2') > 11\epsilon H(XY)$, $\epsilon > 0$, and n is large. We thus find that $P_{\text{bin},e}(w_1, w_2)$ can be made small for $R_1' > 0$ and $R_2' > 0$ if

$$R_1' + R_2' > I(X;Y). \tag{7.73}$$

It remains to consider the cases $R_1' = 0$ and $R_2' = 0$. For $R_1' = 0$, we have

$$P_{\text{bin},e}(w_1, w_2) = \Pr\left[\bigcap_{v_2} \{(X^n(w_1, 1), Y^n(w_2, v_2)) \notin T_\epsilon^n(P_{XY})\}\right]. \tag{7.74}$$

But (7.74) is identical to the probability that a rate distortion encoder does not find an appropriate codeword $Y^n(w_2, v_2)$ that is typical with the "source" sequence $X^n(w_1, 1)$. We thus require

$$R_2' > I(X;Y), \tag{7.75}$$

which is the same as (7.73) with $R_1' = 0$. By symmetry, we also get (7.73) for $R_2' = 0$. This completes the proof.

7.10.2 Converse for Degraded Channels

Consider a physically degraded broadcast channel [27]. For reliable communication, we have

$$\begin{aligned} nR_1 &\le I(W_1; Y_1^n) \\ &\le I(W_1; Y_1^n W_0 W_2) \\ &= I(W_1; Y_1^n | W_0 W_2) \\ &= \sum_{i=1}^n H(Y_{1i} | W_0 W_2 Y_1^{i-1}) - H(Y_{1i} | X_i W_0 W_2 Y_1^{i-1} W_1) \\ &= \sum_{i=1}^n I(X_i; Y_{1i} | U_i'), \end{aligned} \tag{7.76}$$

where we have set $U_i' = [W_0, W_2, Y_1^{i-1}]$. Note that $U_i' - X_i - Y_{1i} - Y_{2i}$ forms a Markov chain. We similarly bound

$$\begin{aligned} n(R_0 + R_2) &\le I(W_0 W_2; Y_2^n) \\ &= \sum_{i=1}^{n} H(Y_{2i}|Y_2^{i-1}) - H(Y_{2i}|W_0 W_2 Y_2^{i-1}) \\ &\le \sum_{i=1}^{n} H(Y_{2i}) - H(Y_{2i}|W_0 W_2 Y_2^{i-1} Y_1^{i-1}) \\ &= \sum_{i=1}^{n} H(Y_{2i}) - H(Y_{2i}|W_0 W_2 Y_1^{i-1}) \\ &= \sum_{i=1}^{n} I(U_i'; Y_{2i}), \end{aligned} \tag{7.77}$$

where the fourth step follows because $Y_{2i} - [W_0, W_2, Y_1^{i-1}] - Y_2^{i-1}$ forms a Markov chain for every i. Finally, let I be a random variable that is independent of all other random variables and that takes on the value i, $i = 1, 2, \ldots, n$, with probability $1/n$. Furthermore, let $U = [U_I', I]$, so we can write (7.76) and (7.77) as

$$\begin{aligned} R_1 &\le I(X_I; Y_{1I}|U) \\ R_0 + R_2 &\le I(U_I'; Y_{2I}|I) \\ &\le I(U; Y_{2I}), \end{aligned} \tag{7.78}$$

where the first inequality follows because U includes I. Combining these results, we find that (R_0, R_1, R_2) must satisfy

$$\begin{aligned} R_1 &\le I(X_I; Y_{1I}|U) \\ R_0 + R_2 &\le I(U; Y_{2I}), \end{aligned} \tag{7.79}$$

where $U - X_I - Y_{1I} - Y_{2I}$ forms a Markov chain. This proves the converse.

7.10.3 Converse for the Scalar AWGN Channel

The entropy power inequality (see Appendix B.7) can be used to show that the region of (7.36)–(7.38) gives the capacity region of the scalar

AWGN broadcast channel. The original proof of this result is due to Bergmans [8]. The following proof is due to El Gamal (unpublished).

Fano's inequality assures us that for reliable communication, we must have

$$\begin{aligned} nR_1 &\le I(W_1; Y_1^n | W_0 W_2) \\ n(R_0 + R_2) &\le I(W_0 W_2; Y_2^n). \end{aligned} \tag{7.80}$$

We further have

$$\begin{aligned} I(W_0 W_2; Y_2^n) &= h(Y_2^n) - h(Y_2^n | W_0 W_2) \\ &\le \left[\sum_{i=1}^{n} h(Y_{2i}) \right] - h(Y_2^n | W_0 W_2) \\ &\le \frac{n}{2} \log(2\pi e\,(P + \sigma_2^2)) - h(Y_2^n | W_0 W_2), \end{aligned} \tag{7.81}$$

where the last step follows by the maximum entropy theorem. But we also have

$$\begin{aligned} \frac{n}{2} \log(2\pi e \sigma_2^2) = h(Z_2^n) = h(Y_2^n | X^n) &\le h(Y_2^n | W_0 W_2) \\ &\le h(Y_2^n) \le \frac{n}{2} \log(2\pi e (P + \sigma_2^2)) \end{aligned} \tag{7.82}$$

so there must exist an α, $0 \le \alpha \le 1$, such that

$$h(Y_2^n | W_0 W_2) = \frac{n}{2} \log\left(2\pi e\,[(1-\alpha)P + \sigma_2^2]\right). \tag{7.83}$$

Consider now $Y_2^n = Y_1^n + (Z_2'')^n$, where Z_{2i}'' has variance $\sigma_2^2 - \sigma_1^2$. Using a conditional version of the entropy power inequality, we bound

$$\begin{aligned} I(W_1; Y_1^n | W_0 W_2) &= h(Y_1^n | W_0 W_2) - \frac{n}{2} \log(2\pi e \sigma_1^2) \\ &\le \frac{n}{2} \log_2\left(2^{\frac{2}{n} h(Y_2^n | W_0 W_2)} - 2\pi e\,(\sigma_2^2 - \sigma_1^2)\right) - \frac{n}{2} \log(2\pi e \sigma_1^2) \\ &= \frac{n}{2} \log_2\left(2\pi e\,[(1-\alpha)P + \sigma_2^2] - 2\pi e\,(\sigma_2^2 - \sigma_1^2)\right) - \frac{n}{2} \log(2\pi e \sigma_1^2) \\ &= \frac{n}{2} \log_2\left(2\pi e\,[(1-\alpha)P + \sigma_1^2]\right) - \frac{n}{2} \log(2\pi e \sigma_1^2). \end{aligned} \tag{7.84}$$

Combining (7.80), (7.81), (7.83), and (7.84), we have the desired region:

$$R_1 \le \frac{1}{2}\log_2\left(1+\frac{(1-\alpha)P}{\sigma_1^2}\right)$$
$$R_0 + R_2 \le \frac{1}{2}\log_2\left(1+\frac{\alpha P}{(1-\alpha)P+\sigma_2^2}\right). \tag{7.85}$$

8

The Multiaccess Channel

8.1 Problem Description

The multiaccess channel (MAC) with two transmitters and three sources is depicted in Figure 8.1. The sources put out statistically independent messages W_0, W_1, W_2 with nR_0, nR_1, nR_2 bits, respectively. The message W_0 is seen by both encoders, and is called the *common* message. The messages W_1 and W_2 appear only at the respective encoders 1 and 2. Encoder 1 maps (w_0, w_1) to a sequence $x_1^n \in \mathcal{X}_1^n$, encoder 2 maps (w_0, w_2) to a sequence $x_2^n \in \mathcal{X}_2^n$, and the channel $P_{Y|X_1X_2}(\cdot)$ puts out the sequence $y^n \in \mathcal{Y}^n$. The decoder uses y^n to compute its estimate $(\hat{w}_0, \hat{w}_1, \hat{w}_2)$ of (w_0, w_1, w_2), and the problem is to find the set of rate-tuples (R_0, R_1, R_2) for which one can make

$$P_e = \Pr\big[(\hat{W}_0, \hat{W}_1, \hat{W}_2) \neq (W_0, W_1, W_2)\big] \tag{8.1}$$

an arbitrarily small positive number. The closure of the region of achievable (R_0, R_1, R_2) is the MAC capacity region $\mathcal{C}_{\text{MAC}}$.

The MAC can be viewed as being the *reverse link* of a cellular radio system, if one views the broadcast channel as being the *forward link* (other popular names are *uplink* for the MAC and *downlink* for the broadcast channel). If there are two *mobile stations* transmitting

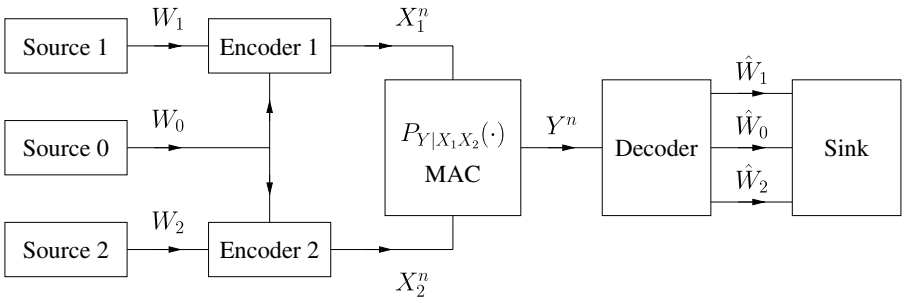

Fig. 8.1 The two-transmitter MAC with a common message.

to the base station, the model of Figure 8.1 describes the essence of the coding problem. One can easily extend the model to include three or more mobile stations, but we will study only the two-transmitter problem. The common message might represent a common time reference that lets the mobile stations *synchronize* their transmissions, in which case we have $R_0 = 0$. Alternatively, this message might represent information the mobile stations are "relaying" from one base station to the next.

8.2 An Achievable Rate Region

The MAC with $R_0 = 0$ was first considered by Shannon in [58, Sec. 17]. The capacity region of the MAC with $R_0 = 0$ was developed by Ahlswede [1] and Liao [42]. (We remark that Shannon wrote in [58, Sec. 17] that he had found a "complete and simple solution of the capacity region."). The capacity region with $R_0 > 0$ was found by Slepian and Wolf [59], who used superposition coding. We consider the general problem, where the main trick is to introduce an auxiliary random variable U that represents the code book for W_0 (see Figure 8.2). Consider a distribution $P_{UX_1X_2Y}$ that factors as $P_U P_{X_1|U} P_{X_2|U} P_{Y|X_1X_2}$.

Code Construction: Consider $P_U(\cdot)$, where the alphabet of U is $\mathcal{U}$. Generate 2^{nR_0} codewords $u^n(w_0)$, $w_0 = 1, 2, \ldots, 2^{nR_0}$, by choosing the $u_i(w_0)$ independently using $P_U(\cdot)$ for $i = 1, 2, \ldots, n$. For each $u^n(w_0)$, generate 2^{nR_1} codewords $x_1^n(w_0, w_1)$, $w_1 = 1, 2, \ldots, 2^{nR_1}$, by choosing the $x_{1i}(w_0, w_1)$ independently using $P_{X_1|U}(\cdot|u_i(w_0))$ for $i = 1, 2, \ldots, n$.

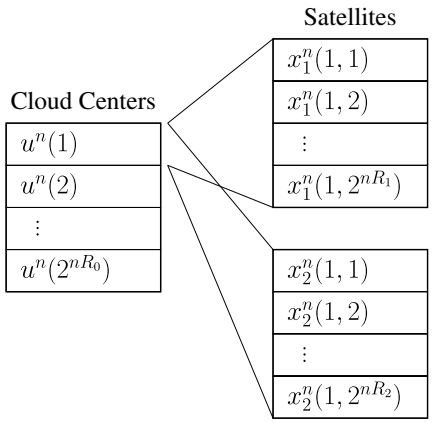

Fig. 8.2 A code book for the MAC with a common message.

Similarly, generate 2^{nR_2} codewords $x_2^n(w_0, w_2)$ by using $P_{X_2|U}(\cdot|u_i(w_0))$ for $i = 1, 2, \ldots, n$.

Encoders: Given (w_0, w_1), encoder 1 transmits $x_1^n(w_0, w_1)$. Given (w_0, w_2), encoder 2 transmits $x_2^n(w_0, w_2)$.

Decoder: Given y^n, try to find a triple $(\tilde{w}_0, \tilde{w}_1, \tilde{w}_2)$ such that

$$(u^n(\tilde{w}_0), x_1^n(\tilde{w}_0, \tilde{w}_1), x_2^n(\tilde{w}_0, \tilde{w}_2), y^n) \in T_\epsilon^n(P_{UX_1X_2Y}). \tag{8.2}$$

If one or more such triple is found, choose one and call it $(\hat{w}_0, \hat{w}_1, \hat{w}_2)$. If no such triple is found, set $(\hat{w}_0, \hat{w}_1, \hat{w}_2) = (1, 1, 1)$.

Analysis: Let $0 \le \epsilon_1 < \epsilon < \mu_{UX_1X_2Y}$. We know that, with probability close to one, we will have

$$(u^n(w_0), x_1^n(w_0, w_1), x_2^n(w_0, w_2), y^n) \in T_{\epsilon_1}^n(P_{UX_1X_2Y}) \tag{8.3}$$

for the transmitted triple (w_0, w_1, w_2) as long as $P_{UX_1X_2Y}(\cdot)$ factors as specified above. The remaining analysis is similar to that of the degraded broadcast channel, i.e., one splits the error probability into seven disjoint events that correspond to the seven different ways in which one or more of the $\hat{w}_i$, $i = 0, 1, 2$, is not equal to w_i.

For example, consider the event that there was a $\tilde{w}_0 \neq w_0$ such that

$$(u^n(\tilde{w}_0), x_1^n(\tilde{w}_0, w_1), x_2^n(\tilde{w}_0, w_2), y^n) \in T_\epsilon^n(P_{UX_1X_2Y}). \tag{8.4}$$

Note that *all three* codewords in (8.4) were chosen independent of the actually transmitted codewords. We can upper bound the probability of the event (8.4) by

$$\sum_{\tilde{w}_0 \neq w_0} 2^{-n[I(UX_1X_2;Y)-2\epsilon H(UX_1X_2)]} < 2^{n[R_0-I(UX_1X_2;Y)+2\epsilon H(UX_1X_2)]}. \tag{8.5}$$

We leave the details of the remaining (and by now familiar) analysis to the reader, and simply state the seven rate bounds for reliable communication:

$$R_0 \leq I(X_1X_2;Y) \tag{8.6}$$

$$R_0 + R_1 \leq I(X_1X_2;Y) \tag{8.7}$$

$$R_0 + R_2 \leq I(X_1X_2;Y) \tag{8.8}$$

and

$$R_1 \leq I(X_1;Y|X_2U) \tag{8.9}$$

$$R_2 \leq I(X_2;Y|X_1U) \tag{8.10}$$

$$R_1 + R_2 \leq I(X_1X_2;Y|U) \tag{8.11}$$

$$R_0 + R_1 + R_2 \leq I(X_1X_2;Y), \tag{8.12}$$

where $X_1 - U - X_2$ and $U - [X_1, X_2] - Y$ form Markov chains. Note that we are stating the bounds with *non-strict* inequalities, so we are already considering approachable rates. Note also that the bounds (8.6)–(8.8) are redundant because of (8.12), so that we need consider only (8.9)–(8.12). One can further restrict attention to $|\mathcal{U}| \leq \min(|\mathcal{Y}| + 3, |\mathcal{X}_1| \cdot |\mathcal{X}_2| + 2)$ (see [19, p. 293 and pp. 310–312], [68, Appendix B], [67, p. 18]).

The bounds (8.9)–(8.12) describe a region $\mathcal{R}(P_U, P_{X_1|U}, P_{X_2|U})$ with seven faces, four of which arise from (8.9)–(8.12), and three of which are non-negativity constraints on the rates (see Figure 8.3). We can further achieve the union of such regions, i.e., we can achieve

$$\mathcal{C}_{\text{MAC}} = \bigcup_{P_U, P_{X_1|U}, P_{X_2|U}} \mathcal{R}(P_U, P_{X_1|U}, P_{X_2|U}), \tag{8.13}$$

where $|\mathcal{U}| \leq \min(|\mathcal{Y}| + 3, |\mathcal{X}_1| \cdot |\mathcal{X}_2| + 2)$. We show that (8.13) is the capacity region in Section 8.4.

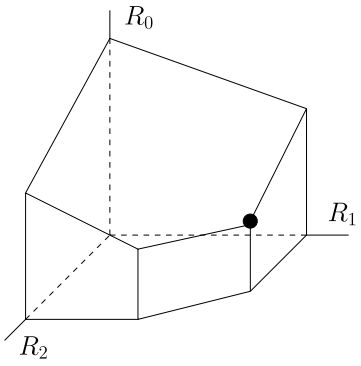

Fig. 8.3 The form of $\mathcal{R}(P_U, P_{X_1|U}, P_{X_2|U})$.

8.3 Gaussian Channel

As an example, consider the additive white Gaussian noise (AWGN) MAC with

$$Y = X_1 + X_2 + Z, \tag{8.14}$$

where Z is Gaussian, zero mean, unit variance, and independent of the real random variables X_1 and X_2. We impose the power (or energy) constraints $\sum_{i=1}^{n} E\left[X_{1i}^2\right]/n \le P_1$ and $\sum_{i=1}^{n} E\left[X_{2i}^2\right]/n \le P_2$. One can show that the best choice for the random variables in (8.9)–(8.12) is jointly Gaussian [10]. Let U, V_1, and V_2 be independent, unit variance, Gaussian random variables, and define

$$X_1 = (\sqrt{P_1}\rho_1)\,U + \sqrt{P_1(1-\rho_1^2)}\,V_1 \tag{8.15}$$

$$X_2 = (\sqrt{P_2}\rho_2)\,U + \sqrt{P_2(1-\rho_2^2)}\,V_2. \tag{8.16}$$

We have $E[UX_1]/\sqrt{P_1} = \rho_1$ and $E[UX_2]/\sqrt{P_2} = \rho_2$, and compute

$$I(X_1;Y|X_2U) = \frac{1}{2}\log\left(1 + P_1(1-\rho_1^2)\right) \tag{8.17}$$

$$I(X_2;Y|X_1U) = \frac{1}{2}\log\left(1 + P_2(1-\rho_2^2)\right) \tag{8.18}$$

$$I(X_1X_2;Y|U) = \frac{1}{2}\log\left(1 + P_1(1-\rho_1^2) + P_2(1-\rho_2^2)\right) \tag{8.19}$$

$$I(X_1X_2;Y) = \frac{1}{2}\log\left(1 + P_1 + P_2 + 2\sqrt{P_1P_2}\,\rho_1\rho_2\right). \tag{8.20}$$

The resulting capacity region is found by considering all ρ_1 and ρ_2 with $0 \le \rho_1 \le 1$ and $0 \le \rho_2 \le 1$.

8.4 Converse

For reliable communication, the rate R_1 must satisfy

$$\begin{aligned}
nR_1 &\le I(W_1;\hat{W}_1) \\
&\le I(W_1;Y^n) \\
&\le I(W_1;Y^n W_0 W_2) \\
&= I(W_1;Y^n|W_0W_2) \\
&= \sum_{i=1}^{n} H(Y_i|Y^{i-1}W_0W_2) - H(Y_i|Y^{i-1}W_0W_1W_2) \\
&= \sum_{i=1}^{n} H(Y_i|Y^{i-1}W_0W_2X_2^n) - H(Y_i|X_{1i}X_{2i}W_0) \\
&\le \sum_{i=1}^{n} H(Y_i|X_{2i}W_0) - H(Y_i|X_{1i}X_{2i}W_0) \\
&= \sum_{i=1}^{n} I(X_{1i};Y_i|X_{2i}W_0). \qquad (8.21)
\end{aligned}$$

We introduce the random variable $U = [W_0, I]$, where I is independent of all other random variables (except U) and has distribution $P_I(a) = 1/n$ for $a = 1, 2, \ldots, n$. We further define $X_1 = X_{1I}$, $X_2 = X_{2I}$, and $Y = Y_I$ so that $P_{UX_1X_2Y}(\cdot)$ factors as

$$P_U([a,i])P_{X_1|U}(b\,|\,[a,i])P_{X_2|U}(c\,|\,[a,i])P_{Y|X_1X_2}(d\,|\,b,c) \qquad (8.22)$$

for all a, b, c, d. We can now write the bound (8.21) as

$$R_1 \le I(X_1;Y|X_2U). \qquad (8.23)$$

We similarly have

$$R_2 \le I(X_2;Y|X_1U) \qquad (8.24)$$

$$R_1 + R_2 \le I(X_1X_2;Y|U) \qquad (8.25)$$

$$R_0 + R_1 + R_2 \le I(X_1X_2;Y). \qquad (8.26)$$

The expressions (8.22)–(8.26) specify that every achievable (R_0, R_1, R_2) must lie in $\mathcal{C}_{\text{MAC}}$. Thus, $\mathcal{C}_{\text{MAC}}$ is the capacity region.

We remark that $\mathcal{C}_{\text{MAC}}$ must be convex since time-sharing is permitted in the converse, i.e, one can use one code book for some fraction of the time and another code book for another fraction of the time. One can check that the union of regions (8.13) is indeed convex (see [67, Appendix A]).

8.5 The Capacity Region with $R_0 = 0$

The MAC is usually treated with $R_0 = 0$, in which case the capacity region reduces to

$$\mathcal{C}_{\text{MAC}} = \bigcup \left\{ (R_1, R_2): \begin{array}{c} 0 \le R_1 \le I(X_1; Y|X_2U) \\ 0 \le R_2 \le I(X_2; Y|X_1U) \\ R_1 + R_2 \le I(X_1X_2; Y|U) \end{array} \right\}, \tag{8.27}$$

where the union is over joint distributions that factor as

$$P_{UX_1X_2Y} = P_U \, P_{X_1|U} \, P_{X_2|U} \, P_{Y|X_1X_2} \tag{8.28}$$

and where $|\mathcal{U}| \le \min(|\mathcal{Y}| + 3, |\mathcal{X}_1| \cdot |\mathcal{X}_2| + 2)$ (one can, in fact, restrict attention to $|\mathcal{U}| \le 2$ [19, p. 278]). However, one often encounters the following equivalent formulation of $\mathcal{C}_{\text{MAC}}$:

$$\mathcal{R}_{\text{MAC}} = \text{co}\left(\bigcup \left\{ (R_1, R_2): \begin{array}{c} 0 \le R_1 \le I(X_1; Y|X_2) \\ 0 \le R_2 \le I(X_2; Y|X_1) \\ R_1 + R_2 \le I(X_1X_2; Y) \end{array} \right\} \right), \tag{8.29}$$

where the union is over joint distributions that factor as

$$P_{X_1X_2Y} = P_{X_1} \, P_{X_2} \, P_{Y|X_1X_2} \tag{8.30}$$

and where $\text{co}(\mathcal{S})$ is the convex hull of a set $\mathcal{S}$. Proving that $\mathcal{R}_{\text{MAC}} = \mathcal{C}_{\text{MAC}}$ requires some additional work, and we refer to [67, sec. 3.5] for a discussion on this topic. Some authors prefer (8.29) for historical reasons, and because (8.29) has no U. Other authors prefer (8.27) because it requires no convex hull operation. We do point out, however, that for some channels (other than MACs) a time-sharing random variable U gives larger regions than the convex hull operator (see [19, pp. 288–290]).

Consider two examples. First, consider the AWGN MAC with block or per-symbol power constraints P_1 and P_2 for the respective transmitters 1 and 2. The maximum entropy theorem ensures that

$$\mathcal{C}_{\text{MAC}} = \left\{ (R_1, R_2) : \begin{array}{c} 0 \le R_1 \le \frac{1}{2}\log(1+P_1) \\ 0 \le R_2 \le \frac{1}{2}\log(1+P_2) \\ R_1 + R_2 \le \frac{1}{2}\log(1+P_1+P_2) \end{array} \right\}. \tag{8.31}$$

The resulting region is plotted in Figure 8.4. We remark that an alternative coding method for block power constraints is to use time-division multiplexing (TDM) or frequency-division multiplexing (FDM). For example, suppose that transmitters 1 and 2 use the fractions α and $1-\alpha$ of the available bandwidth, respectively. The resulting rates are

$$R_1 = \frac{\alpha}{2}\log\left(1 + \frac{P_1}{\alpha}\right) \tag{8.32}$$

$$R_2 = \frac{1-\alpha}{2}\log\left(1 + \frac{P_2}{1-\alpha}\right), \tag{8.33}$$

where the transmitters boost their powers in their frequency bands. The resulting rate pairs are plotted in Figure 8.4. In particular, by choosing $\alpha = P_1/(P_1+P_2)$ one achieves a boundary point with

$$R_1 + R_2 = \log(1+P_1+P_2). \tag{8.34}$$

This shows that TDM and FDM can be effective techniques for the MAC.

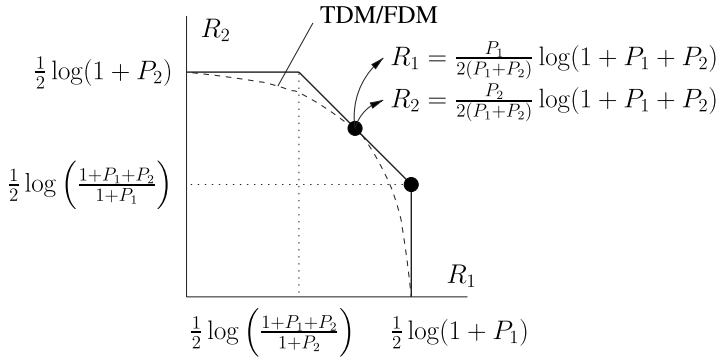

Fig. 8.4 $\mathcal{C}_{\text{MAC}}$ for the AWGN MAC.

Second, consider the *binary adder channel* or BAC with $\mathcal{X}_1 = \mathcal{X}_2 = \{0,1\}$, $\mathcal{Y} = \{0,1,2\}$, and

$$Y = X_1 + X_2, \tag{8.35}$$

where "+" refers to integer addition. The best X_1 and X_2 are uniformly distributed and we compute

$$\mathcal{C}_{\text{MAC}} = \left\{ \begin{array}{c} 0 \le R_1 \le 1 \\ (R_1, R_2) : 0 \le R_2 \le 1 \\ R_1 + R_2 \le 1.5 \end{array} \right\}. \tag{8.36}$$

The resulting region has a similar form as that shown in Figure 8.4.

8.6 Decoding Methods

8.6.1 Single-User Decoding and Rate-Splitting

The capacity expression (8.29) is suggestive for code design. Consider, e.g., the AWGN MAC and the marked corner point in Figure 8.4. The decoder can proceed in two stages: first, decode w_2 by considering $x_1^n(w_1)$ as AWGN with variance P_1; second, subtract $x_2^n(w_2)$ from y^n and decode w_1. The capacities of the second and first channels are the respective

$$\begin{aligned} R_1 &= \frac{1}{2}\log(1 + P_1) \\ R_2 &= \frac{1}{2}\log\left(1 + \frac{P_2}{1 + P_1}\right). \end{aligned} \tag{8.37}$$

This type of decoding is known as *single-user* decoding, *stripping*, *onion peeling*, or *step-by-step* decoding.

One difficulty with (this form of) single-user decoding is that one can achieve only the corner points of the pentagon in Figure 8.4. The other points of the face with maximal $R_1 + R_2$ must be achieved by time-sharing between these two corner points. However, there is a simple trick known as *rate-splitting* by which one can achieve the other rate points by single-user decoding [29, 53]. The idea is to split encoder 2 into *two* encoders operating at the respective rates R_{21} and R_{22} with $R_2 = R_{21} + R_{22}$. Suppose these encoders transmit with respective powers P_{21} and P_{22}, where $P_2 = P_{21} + P_{22}$, and that the output of the

second transmitter is the sum of the two encoded signals. The decoder performs single-user decoding in three stages: first, decode the R_{21} code; second, decode the R_1 code; third, decode the R_{22} code. The rates are

$$\begin{aligned} R_1 &= \frac{1}{2}\log\left(1+\frac{P_1}{1+P_{22}}\right) \\ R_2 &= R_{21}+R_{22} = \frac{1}{2}\log\left(1+\frac{P_{21}}{1+P_1+P_{22}}\right)+\frac{1}{2}\log\left(1+P_{22}\right). \end{aligned} \tag{8.38}$$

Note that by choosing $P_{22}=0$ we recover (8.37), while if we choose $P_{22}=P_2$ we obtain the other corner point of the pentagon in Figure 8.4. By varying P_{22} from 0 to P_2, we thus achieve any rate point on the boundary of that face of the pentagon with maximum sum-rate.

8.6.2 Joint Decoding

Joint decoding refers to decoding both messages simultaneously. For the MAC, an "optimal" joint decoder is much more complex than an "optimal" single-user decoder because one must consider all codeword *pairs*. However, by using iterative decoding, joint decoders can be implemented almost as easily as single-user decoders [4]. Suppose, for

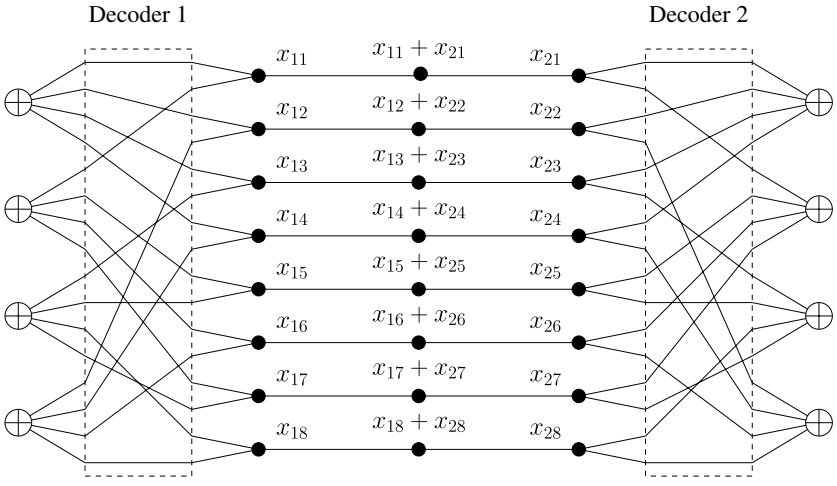

Fig. 8.5 Graph for an iterative joint decoder for the AWGN MAC.

example, that both messages are encoded with a low-density parity-check (LDPC) code. An example of a decoding graph (or *factor graph*) for the decoders *and* the MAC is depicted in Figure 8.5. The iterative decoder is initialized by giving the nodes labeled $x_{1i} + x_{2i}$ a log-likelihood ratio (LLR) based on the y_i, $i = 1, 2, \ldots, n$. The remaining operation of the decoder is similar to that for a DMC or a point-to-point AWGN channel.

9

The Relay Channel

9.1 Problem Description

The relay channel is a multi-terminal problem where a source terminal transmits a message to a sink terminal with the help of one or more relays. We begin by considering the model of Figure 9.1 that has one relay. The message W with entropy nR bits is transmitted from the source terminal (terminal 1), with the help of a relay terminal (terminal 2), to the sink terminal (terminal 3) via a channel $P_{Y_2Y_3|X_1X_2}(\cdot)$. We model the transmissions as taking place *synchronously*, i.e., there is a central *clock* that governs the operation of the terminals. The clock ticks n times, and terminals 1 and 2 apply the respective inputs X_{1i} and X_{2i} to the channel *after* clock tick $i-1$ and *before* clock tick i. The receiving terminals 2 and 3 see their respective channel outputs Y_{2i} and Y_{3i} *at* clock tick i. Thus, there is a small delay before reception that ensures the system operates in a *causal* fashion. The alphabets of X_{1i}, X_{2i}, Y_{2i}, and Y_{3i} are $\mathcal{X}_1$, $\mathcal{X}_2$, $\mathcal{Y}_2$, and $\mathcal{Y}_3$, respectively.

The synchronism we require is somewhat restrictive, and a more realistic model might be to view time as being continuous, and to permit each terminal to transmit a waveform of duration T seconds.

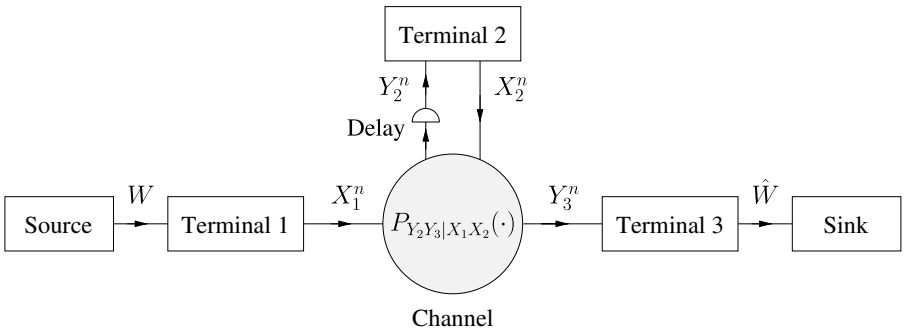

Fig. 9.1 The relay channel.

However, for such scenarios many more issues must be considered carefully, such as bandwidth (are the waveforms band-limited?), the channel (is it linear? time varying?), the receiver processing (what kind of filters and samplers are used?), and so on. We do not wish to consider these issues here. We study the simpler model because it will help us understand how to design codes for more complex problems.

We return to our discrete-time and synchronous model, and add a few more constraints. We require the input sequence X_1^n to be a function of W, the input symbol X_{2i} to be a function of Y_2^{i-1}, $i = 2, 3, \ldots, n$, and $\hat{W}$ to be a function of Y_3^n. The joint probability distribution of the random variables thus factors as

$$\begin{aligned} &P(w, x_1^n, x_2^n, y_2^n, y_3^n, \hat{w}) \\ &= P(w)P(x_1^n|w)\left[\prod_{i=1}^{n} P(x_{2i}|y_2^{i-1})P_{Y_2Y_3|X_1X_2}(y_{2i}, y_{3i}|x_{1i}, x_{2i})\right] P(\hat{w}|y_3^n), \end{aligned} \tag{9.1}$$

where $P(x_1^n|w)$, $P(x_{2i}|y_2^{i-1})$, and $P(\hat{w}|y_3^n)$ take on the values 0 and 1 only. Note that in (9.1) we have adopted the convention of dropping subscripts on probability distributions if the arguments are lower-case versions of the random variables. This is commonly done in the literature, but it is often wise to keep the subscripts to avoid confusing oneself and the readers. The capacity C of the relay channel is the supremum of rates R for which one can design encoders $P(x_1^n|w)$ and

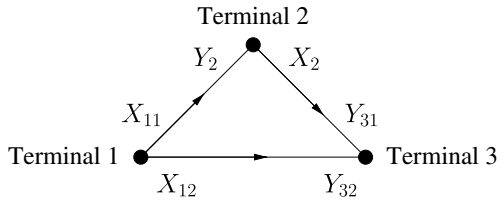

Fig. 9.2 A network of DMCs.

$P(x_{2i}|y_2^{i-1})$, and a decoder $P(\hat{w}|y_3^n)$, so that $\Pr\left[W \neq \hat{W}\right] < \epsilon$ for any positive ϵ.

The above model includes a wide variety of practical problems. For example, consider the *wired* network of discrete memoryless channels (DMCs) shown in Figure 9.2. The channel input of terminal 1 is a *vector* $X_1 = [X_{11}, X_{12}]$, where the meaning is that X_{11} is the input of the DMC from terminal 1 to terminal 2, and X_{12} is the input of the DMC from terminal 1 to terminal 3. The input of the DMC from terminal 2 to terminal 3 is X_2. Similarly, the two relay channel outputs are Y_2 and $Y_3 = [Y_{31}, Y_{32}]$. The channel probability distribution thus factors as

$$P(y_2, y_{31}, y_{32}|x_{11}, x_{12}, x_2) = P(y_2|x_{11})P(y_{31}|x_2)P(y_{32}|x_{12}). \tag{9.2}$$

Suppose that X_{11}, X_{12}, and X_2 are binary, and that $Y_2 = X_{11}$, $Y_{31} = X_2$, and $Y_{32} = X_{12}$. The capacity is known to be 2 bits per clock tick, as follows from Ford and Fulkerson's Max-flow, Min-cut Theorem [24] (the book [2] gives a good introduction to network flow problems). The achievability of 2 bits per clock tick is obvious, and the converse follows by observing that terminal 1 can send (and terminal 3 receive) at most 2 bits per clock tick.

As another example, consider the additive white Gaussian noise (AWGN) relay channel depicted in Figure 9.3. The channel $P_{Y_2Y_3|X_1X_2}(\cdot)$ is defined by

$$\begin{aligned} Y_2 &= X_1 + Z_2 \\ Y_3 &= X_1 + X_2 + Z_3, \end{aligned} \tag{9.3}$$

where Z_2 and Z_3 are Gaussian random variables of variance σ_2^2 and σ_3^2, respectively, and are independent of each other and all other random

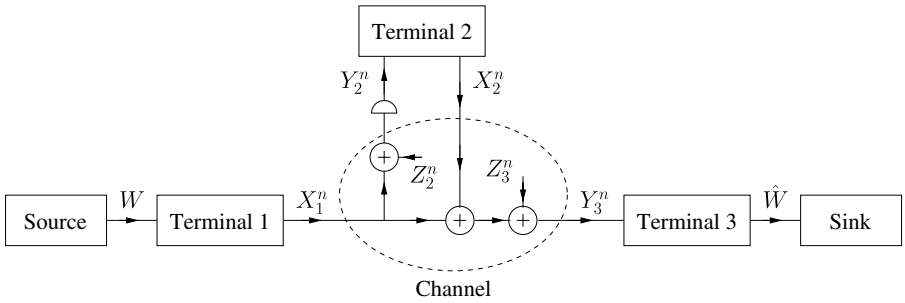

Fig. 9.3 The AWGN relay channel.

variables. There are power constraints on the two input sequences X_1^n and X_2^n, namely

$$\frac{1}{n}\sum_{i=1}^{n} E\left[X_{ti}^2\right] \leq P_t, \quad t = 1, 2. \tag{9.4}$$

As a third example, for *wireless* networks the relay can often not transmit and receive at the same time. In this case, one should add the following constraints to the model:

$$Y_2 = 0 \text{ if } X_2 \neq 0 \tag{9.5}$$

$$\beta \leq \Pr[X_2 = 0] \leq \gamma \tag{9.6}$$

for some β and γ with $0 \leq \beta \leq 1$ and $0 \leq \gamma \leq 1$. The constraint (9.6) puts limits on how often the relay can transmit.

9.2 Decode-and-Forward

The relay channel was studied early on in [63]. The capacity of the relay channel is still not known in general! We will develop three coding strategies for this channel, and show that these can sometimes achieve capacity. The first strategy uses a technique called *block-Markov superposition encoding* and is now often called Decode-and-Forward (DF). The second strategy adds partial decoding, and the third strategy combines block-Markov coding with binning. The second and third strategies are described in the appendix of this section. All three strategies are due to Cover and El Gamal [16].

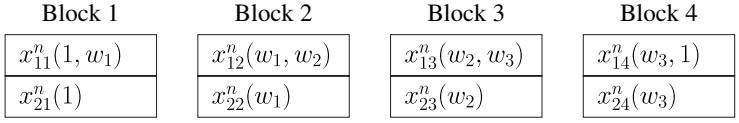

Fig. 9.4 Block-Markov superposition encoding for the relay channel assuming the relay decodes correctly.

Code Construction: Consider a distribution $P_{X_1X_2}(\cdot)$. Encoding is performed in $B+1$ blocks, and for ease of analysis we will generate a *separate* code book for each block (see Figure 9.4 where $B+1=4$). That is, for block b, $b=1,2,\ldots,B+1$, generate 2^{nR} codewords $x_{2b}^n(v)$, $v=1,2,\ldots,2^{nR}$, by choosing the symbols $x_{2bi}(v)$ independently using $P_{X_2}(\cdot)$. Next, for every $x_{2b}^n(v)$, use superposition coding and generate 2^{nR} codewords $x_{1b}^n(v,w)$, $w=1,2,\ldots,2^{nR}$, by choosing the $x_{1bi}(v,w)$ independently using $P_{X_1|X_2}(\cdot|x_{2bi}(v))$.

Source Terminal: The message w of nRB bits is split into B equally-sized blocks $w_1,w_2,\ldots,w_B$ of nR bits each. In block b, $b=1,2,\ldots,B+1$, the source transmits $x_{1b}^n(w_{b-1},w_b)$, where $w_0=w_{B+1}=1$. This type of transmission is called *block Markov superposition encoding*.

Relay Terminal: After the transmission of block b is completed, the relay has seen y_{2b}^n. The relay tries to find a $\tilde{w}_b$ such that

$$(x_{1b}^n(\hat{w}_{b-1}(2),\tilde{w}_b),x_{2b}^n(\hat{w}_{b-1}(2)),y_{2b}^n)\in T_\epsilon^n(P_{X_1X_2Y_2}), \tag{9.7}$$

where $\hat{w}_{b-1}(2)$ is the relay terminal's estimate of w_{b-1}. If one or more such $\tilde{w}_b$ are found, then the relay chooses one of them, calls this choice $\hat{w}_b(2)$, and transmits $x_{2(b+1)}^n(\hat{w}_b(2))$ in block $b+1$. If no such $\tilde{w}_b(2)$ is found, the relay sets $\hat{w}_b(2)=1$ and transmits $x_{2(b+1)}^n(1)$.

Sink Terminal: The sink decodes by using a *sliding window* decoding method [11, 72]. After block b, the receiver has seen $y_{3(b-1)}^n$ and y_{3b}^n, and tries to find a $\tilde{w}_{b-1}$ such that

$$(x_{1(b-1)}^n(\hat{w}_{b-2}(3),\tilde{w}_{b-1}),x_{2(b-1)}^n(\hat{w}_{b-2}(3)),y_{3(b-1)}^n)\in T_\epsilon^n(P_{X_1X_2Y_3})$$

and

$$(x_{2b}^n(\tilde{w}_{b-1}),y_{3b}^n)\in T_\epsilon^n(P_{X_2Y_3}), \tag{9.8}$$

where $\hat{w}_{b-2}(3)$ is the sink terminal's estimate of w_{b-2}. For example, after block 2 terminal 3 decodes w_1 by using y_{31}^n and y_{32}^n (see Figure 9.4).

If one or more such $\tilde{w}_{b-1}$ are found, then the sink chooses one of them, and puts out this choice as $\hat{w}_{b-1}(3)$. If no such $\tilde{w}_{b-1}$ is found, the sink puts out $\hat{w}_{b-1}(3) = 1$.

Analysis: Let $\mathcal{E}_{2b}^{0}$ and $\mathcal{E}_{2b}^{2+}$ be the respective events that the *relay* finds no appropriate $\tilde{w}_b$ and that it finds a $\tilde{w}_b \neq w_b$ that satisfies (9.7). Similarly, let $\mathcal{E}_{3b}^{0}$ and $\mathcal{E}_{3b}^{2+}$ be the respective events that the *sink* finds no appropriate $\tilde{w}_{b-1}$ and that it finds a $\tilde{w}_{b-1} \neq w_{b-1}$ that satisfies (9.8). We further define $\mathcal{F}_{b-1}$ to be the event that *no* errors have been made up to block b. We can write the overall probability of error as

$$\begin{aligned} P_B &= \Pr\left[\bigcup_{b=1}^{B}[\mathcal{E}_{2b}^{0} \cup \mathcal{E}_{2b}^{2+}] \ \cup \ \bigcup_{b=2}^{B+1}[\mathcal{E}_{3b}^{0} \cup \mathcal{E}_{3b}^{2+}]\right] \\ &= \Pr\left[\mathcal{E}_{21}^{0} \cup \mathcal{E}_{21}^{2+}\right] + \sum_{b=2}^{B}\Pr\left[[\mathcal{E}_{2b}^{0} \cup \mathcal{E}_{2b}^{2+}] \ \cup \ [\mathcal{E}_{3b}^{0} \cup \mathcal{E}_{3b}^{2+}] \,|\mathcal{F}_{b-1}\right] \\ &\quad + \Pr[\mathcal{E}_{3(B+1)}^{0} \cup \mathcal{E}_{3(B+1)}^{2+} \,|\mathcal{F}_B]. \end{aligned} \tag{9.9}$$

The expression (9.9) specifies that we can consider each block separately by assuming that no errors were made in the previous blocks. The overall block error probability P_B will then be upper-bounded by $B + 1$ times the maximum error probability of any block.

So suppose that no errors were made up to block b. We divide the error analysis into several parts. Let $0 < \epsilon_1 < \epsilon < \mu_{X_1X_2Y_2Y_3}$.

(1) Suppose that $(x_{1b}^n(w_{b-1}, w_b), x_{2b}^n(w_{b-1}), y_{2b}^n, y_{3b}^n) \notin T_{\epsilon_1}^n(P_{X_1X_2Y_2Y_3})$ for any b, where $\hat{w}_{b-1}(2) = w_{b-1}$ and $\hat{w}_{b-2}(3) = w_{b-2}$ since $\mathcal{F}_{b-1}$ has occurred. The probability of this event approaches zero with n. Thus, with probability close to one, both the relay and sink will find at least one $\tilde{w}_b$ and $\tilde{w}_{b-1}$ that satisfy (9.7) and (9.8), respectively.

(2) Suppose the relay finds a $\tilde{w}_b \neq w_b$ satisfying (9.7), where in (9.7) we set $\hat{w}_{b-1}(2) = w_{b-1}$. The erroneous $x_{1b}^n(w_{b-1}, \tilde{w}_b)$ was chosen using $P_{X_1|X_2}(\cdot|x_{2bi}(w_{b-1}))$. We can thus use

Theorem 7.1 to write

$$\Pr\left[\mathcal{E}_{2b}^{2+}\,\middle|\,\mathcal{F}_{b-1}\cap\bar{\mathcal{E}}_{2b}^{0}\right]\le\sum_{\tilde{w}_b\neq w_b}2^{-n[I(X_1;Y_2|X_2)-2\epsilon H(X_1|X_2)]}$$
$$<2^{n[R-I(X_1;Y_2|X_2)+2\epsilon H(X_1|X_2)]},\quad(9.10)$$

where $\bar{\mathcal{E}}_{2b}^{0}$ is the complement of $\mathcal{E}_{2b}^{0}$.

(3) Suppose the sink finds a $\tilde{w}_{b-1}\neq w_{b-1}$ satisfying (9.8), where in (9.8) we set $\hat{w}_{b-2}(3)=w_{b-2}$. The erroneous $x_{1(b-1)}^{n}(w_{b-2},\tilde{w}_{b-1})$ was chosen using $P_{X_1|X_2}(\cdot|x_{2(b-1)i}(w_{b-2}))$. Furthermore, the erroneous $x_{2b}^{n}(\tilde{w}_{b-1})$ was chosen *independent* of the erroneous $x_{1(b-1)}^{n}(w_{b-2},\tilde{w}_{b-1})$ and independent of all other past events. The result is

$$\Pr\left[\mathcal{E}_{3b}^{2+}\,\middle|\,\mathcal{F}_{b-1}\cap\bar{\mathcal{E}}_{3b}^{0}\right]$$
$$\le\sum_{\tilde{w}_{b-1}\neq w_{b-1}}2^{-n[I(X_1;Y_3|X_2)-2\epsilon H(X_1|X_2)]}\cdot 2^{-n[I(X_2;Y_3)-2\epsilon H(X_2)]}$$
$$<2^{n[R-I(X_1X_2;Y_3)+2\epsilon H(X_1X_2)]},\quad(9.11)$$

where $\bar{\mathcal{E}}_{3b}^{0}$ is the complement of $\mathcal{E}_{3b}^{0}$.

Combining (9.10) and (9.11), and letting B become large, we can approach the rate

$$R=\max_{P_{X_1X_2}(\cdot)}\min\left[I(X_1;Y_2|X_2),I(X_1X_2;Y_3)\right].\quad(9.12)$$

The mutual information $I(X_1;Y_2|X_2)$ in (9.12) represents the information transfer on the source-to-relay link, while the mutual information $I(X_1X_2;Y_3)$ represents the combined information transfer from the source and relay to the destination.

We will later show that the following is an *upper bound* on the relay channel capacity:

$$C\le\max_{P_{X_1X_2}(\cdot)}\min\left[I(X_1;Y_2Y_3|X_2),I(X_1X_2;Y_3)\right].\quad(9.13)$$

Note that an additional Y_3 appears in $I(X_1;Y_2Y_3|X_2)$ in (9.13) as compared to (9.12).

We remark that (9.12) can be achieved in several ways [38]. For instance, the book [18, Sec. 14.7 on pp. 428–432] follows the approach

of [16] by combining block-Markov superposition encoding with partitioning or binning (see also [5, 30], where this method is extended to multiple relays). Yet a third approach is to replace sliding window decoding with a *backward* decoding technique described in [67, Sec. 7].

9.2.1 Examples

As a first example, consider the relay channel of Figure 9.2. The rate (9.12) is only 1 bit per clock tick because we require the relay to decode the message w. Clearly, the above strategy is not very good for such a network. Both of the strategies in the appendix of this section remedy this problem.

Consider next the AWGN relay channel of Figure 9.3. We specialize the model: consider the geometry of Figure 9.5 for which the channel is

$$\begin{aligned} Y_2 &= \frac{X_1}{d} + Z_2 \\ Y_3 &= X_1 + \frac{X_2}{1-d} + Z_3, \end{aligned} \tag{9.14}$$

where Z_2 and Z_3 are *unit*-variance Gaussian random variables. We choose $p_{X_1X_2}(\cdot)$ to be zero-mean Gaussian with $E\left[X_1^2\right] = P_1$, $E\left[X_2^2\right] = P_2$, and $E[X_1X_2] = \rho\sqrt{P_1P_2}$. We compute (9.12) to be

$$\begin{aligned} R = \max_{0\le\rho\le 1} \min \Bigg[&\frac{1}{2}\log\left(1 + \frac{(1-\rho^2)P_1}{d^2}\right), \\ &\frac{1}{2}\log\left(1 + P_1 + \frac{P_2}{(1-d)^2} + 2\rho\frac{\sqrt{P_1P_2}}{|1-d|}\right)\Bigg]. \end{aligned} \tag{9.15}$$

The resulting optimized ρ and rates are plotted in Figure 9.6 as the curves labeled "strategy 1." For instance, suppose that $d = 1$, in which case the optimum ρ is 0 and the best achievable rate is $\log_2(1+10)/2 \approx 1.73$ bits per clock tick. This is the same rate that

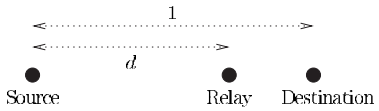

Fig. 9.5 A single relay on a line.

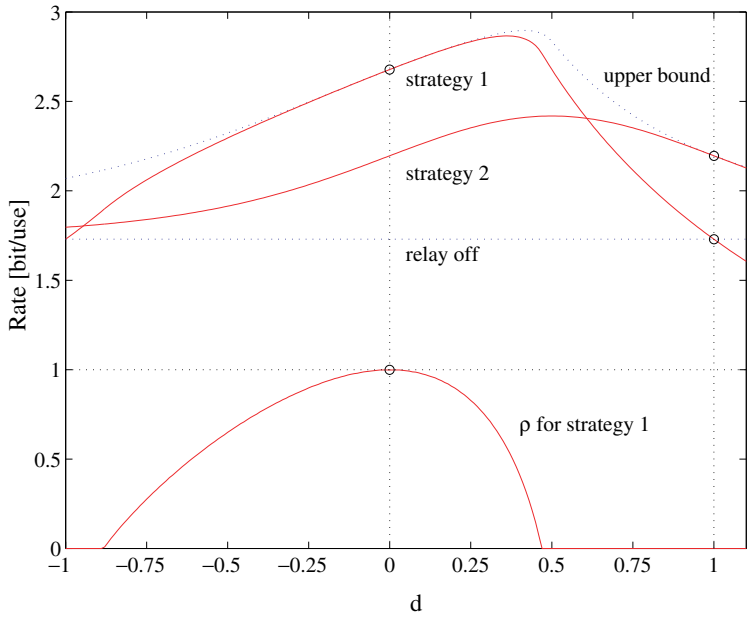

Fig. 9.6 Rates for an AWGN relay channel with $P_1 = P_2 = 10$.

we can achieve without a relay. However, for $d \to 0$ we have $\rho \to 1$ and $R \to \log_2(1 + 40) \approx 2.68$ bits per clock tick. Now the relay boosts the rate substantially. The curve labeled "strategy 2" gives the rates of the compress-and-forward strategy described in the appendix of this section. (The partial-decode-and-forward strategy of the appendix gives the same rates as "strategy 1" for this problem.) The curve labeled "upper bound" in Figure 9.6 gives an upper bound on C. We show how to compute this curve later.

9.3 Physically Degraded Relay Channels

Recall that the capacity region of the broadcast channel is still not known, but for physically or stochastically degraded broadcast channels we know that superposition encoding achieves capacity. One might therefore suspect that the same is true for relay channels. Unfortunately, this is not quite the case.

Consider the AWGN relay channel of (9.3). A natural definition for a stochastically degraded relay channel is that $\sigma_2^2 \leq \sigma_3^2$, or perhaps some other relation between P_1, P_2, σ_2^2, and σ_3^2. However, as we have seen in Figure 9.6, the block-Markov superposition encoding scheme developed above does not achieve capacity except in trivial cases. This seems discouraging.

Consider, then, the more restrictive *physically* degraded model

$$\begin{aligned} Y_2 &= X_1 + Z_2 \\ Y_3 &= X_1 + X_2 + Z_2 + \tilde{Z}_3 \\ &= X_2 + Y_2 + \tilde{Z}_3, \end{aligned} \tag{9.16}$$

where $\tilde{Z}_3$ is a Gaussian random variable of variance $\tilde{\sigma}_3^2$ that is independent of all other random variables. We now have that $X_1 - [Y_2, X_2] - Y_3$ forms a Markov chain, and therefore

$$\begin{aligned} I(X_1; Y_2 Y_3 | X_2) &= I(X_1; Y_2 | X_2) + I(X_1; Y_3 | X_2 Y_2) \\ &= I(X_1; Y_2 | X_2) \end{aligned} \tag{9.17}$$

for any input distribution $P_{X_1 X_2}(\cdot)$. That is, the capacity lower bound (9.12) and upper bound (9.13) are identical and block Markov superposition encoding is optimal. One can obviously extend this result to any relay channels for which $X_1 - [Y_2, X_2] - Y_3$ forms a Markov chain [16, Sec. 14.7]. This example shows that, unlike for broadcast channels, *physical* degradation is not "equivalent" to *stochastic* degradation, in the sense that the capacities can be different.

Consider next a "reversely" physically degraded relay channel, i.e., we have that $X_1 - [Y_3, X_2] - Y_2$ forms a Markov chain. We now compute

$$I(X_1; Y_2 Y_3 | X_2) = I(X_1; Y_3 | X_2) \leq I(X_1 X_2; Y_3) \tag{9.18}$$

for any input distribution $P_{X_1 X_2}(\cdot)$. This implies that the upper bound (9.13) is

$$C \leq \max_{a \in \mathcal{X}_2} \max_{P_{X_1}(\cdot)} I(X_1; Y_3 | X_2 = a), \tag{9.19}$$

where $\mathcal{X}_2$ is the alphabet of X_2 [16, Thm. 2]. The rate (9.19) is certainly achievable, so we have equality in (9.19).

9.4 Appendix: Other Strategies

9.4.1 A Partial Decoding Strategy

One of the limitations of the strategy developed in Section 9.2 is that the relay decodes all the message bits. To circumvent this problem, we split W into two parts W' and W'' with respective rates R' and R'', and demand that the relay decode only W'. Such a *partial decoding* strategy can be designed by introducing an auxiliary random variable U and creating a separate codebook for W'. The following strategy is often called *Partial-Decode-and-Forward* or *Multipath Decode-and-Forward* (see [39, Sec. 4.2.7]).

Code Construction: Consider a distribution $P_{UX_1X_2}(\cdot)$. Encoding is again performed in $B+1$ blocks (see Figure 9.7 where $B+1=4$). For block b, generate $2^{nR'}$ codewords $x_{2b}^n(v)$, $v=1,2,\ldots,2^{nR'}$, by choosing the $x_{2bi}(v)$ independently using $P_{X_2}(\cdot)$. Next, for every $x_{2b}^n(v)$, use superposition coding and generate $2^{nR'}$ codewords $u_b^n(v,w)$, $w=1,2,\ldots,2^{nR'}$, by choosing the $u_{bi}(v,w)$ independently using $P_{U|X_2}(\cdot|x_{2bi}(v))$. Finally, for every $(x_{2b}^n(v),u_b^n(v,w))$ choose $2^{nR''}$ codewords $x_{1b}^n(v,w,t)$, $t=1,2,\ldots,2^{nR''}$, by choosing the $x_{1bi}(v,w)$ independently using $P_{X_1|X_2U}(\cdot|x_{2bi}(v),u_{bi}(v,w))$.

Source Terminal: The message w' of $nR'B$ bits is split into B equally sized blocks $w_1,w_2,\ldots,w_B$ of nR' bits each. Similarly, w'' of $nR''B$ bits is split into B equally sized blocks $t_1,t_2,\ldots,t_B$ of nR'' bits each. In block b, $b=1,2,\ldots,B+1$, the source transmits $x_{1b}^n(w_{b-1},w_b,t_b)$, where $w_0=w_{B+1}=t_{B+1}=1$.

Block 1	Block 2	Block 3	Block 4
$x_{11}^n(1,w_1,t_1)$	$x_{12}^n(w_1,w_2,t_2)$	$x_{13}^n(w_2,w_3,t_3)$	$x_{14}^n(w_3,1,1)$
$u_{21}^n(1,w_1)$	$u_{22}^n(w_1,w_2)$	$u_{23}^n(w_2,w_3)$	$u_{24}^n(w_3,1)$
$x_{21}^n(1)$	$x_{22}^n(w_1)$	$x_{23}^n(w_2)$	$x_{24}^n(w_3)$

Fig. 9.7 A partial decoding strategy for the relay channel assuming the relay decodes correctly.

Relay Terminal: After the transmission of block b is completed, the relay has seen y_{2b}^n. The relay tries to find a $\tilde{w}_b$ such that

$$(u_{1b}^n(\hat{w}_{b-1}(2), \tilde{w}_b), x_{2b}^n(\hat{w}_{b-1}(2)), y_{2b}^n) \in T_\epsilon^n(P_{UX_2Y_2}). \tag{9.20}$$

where $\hat{w}_{b-1}(2)$ is the relay's estimate of w_{b-1}. If one or more such $\tilde{w}_b$ are found, then the relay chooses one of them, calls this choice $\hat{w}_b(2)$, and transmits $x_{2(b+1)}^n(\hat{w}_b(2))$ in block $b+1$. If no such $\tilde{w}_b$ is found, the relay sets $\hat{w}_b(2) = 1$ and transmits $x_{2(b+1)}^n(1)$.

Sink Terminal: After block b, the receiver has seen $y_{3(b-1)}^n$ and y_{3b}^n, and tries to find a pair $(\tilde{w}_{b-1}, \tilde{t}_{b-1})$ such that

$$\begin{aligned} &(u_{1(b-1)}^n(\hat{w}_{b-2}(3), \tilde{w}_{b-1}), x_{1(b-1)}^n(\hat{w}_{b-2}(3), \tilde{w}_{b-1}, \tilde{t}_{b-1}), \\ &x_{2(b-1)}^n(\hat{w}_{b-2}(3)), y_{3(b-1)}^n) \in T_\epsilon^n(P_{UX_1X_2Y_3}) \\ &\text{and} \quad (x_{2b}^n(\tilde{w}_{b-1}), y_{3b}^n) \in T_\epsilon^n(P_{X_2Y_3}), \end{aligned} \tag{9.21}$$

where $\hat{w}_{b-2}(3)$ is the sink terminal's estimate of w_{b-2}. If one or more such pair is found, then the sink chooses one of them, and puts out this choice as $(\hat{w}_{b-1}(3), \hat{t}_{b-1}(3))$. If no such pair is found, then the sink puts out $(\hat{w}_{b-1}(3), \hat{t}_{b-1}(3)) = (1, 1)$.

Analysis: We use the same approach as in Section 9.2, and suppose that no errors were made up to block b. We again divide the error analysis into several parts, and summarize the results. Let $0 < \epsilon_1 < \epsilon < \mu_{UX_1X_2Y_2Y_3}$.

(1) With probability close to 1, for every b we have

$$(u_b^n(w_b), x_{1b}^n(w_{b-1}, w_b, t_b), x_{2b}^n(w_{b-1}), y_{2b}^n, y_{3b}^n) \in T_{\epsilon_1}^n(P_{X_1X_2Y_2Y_3}).$$

(2) The relay decoding step requires

$$R' < I(U; Y_2|X_2). \tag{9.22}$$

(3) The sink decoding step requires

$$\begin{aligned} R' &< I(UX_1; Y_3|X_2) + I(X_2; Y_3) \\ R'' &< I(X_1; Y_3|X_2U) \\ R' + R'' &< I(UX_1; Y_3|X_2) + I(X_2; Y_3). \end{aligned} \tag{9.23}$$

We have $R = R' + R''$. Combining (9.22) and (9.23), for large B we can approach the rate

$$R = \max_{P_{UX_1X_2}(\cdot)} \min\left[I(U;Y_2|X_2) + I(X_1;Y_3|X_2U), I(X_1X_2;Y_3)\right]. \quad (9.24)$$

The rate (9.24) is the same as (9.12) if $U = X_1$.

Example 9.1. Consider the relay channel of Figure 9.2, and recall that the rate (9.12) is only 1 bit per clock tick because we require the relay to decode w. Suppose we instead use the partial-decode-and-forward with $U = X_{11}$, and where X_{11}, X_2, and X_{12} are statistically independent coin-flipping random variables. We compute

$$\begin{aligned} &I(U;Y_2|X_2) = H(Y_2|X_2) = H(X_{11}|X_2) = H(X_{11}) = 1 \\ &I(X_1;Y_3|X_2U) = H(Y_3|X_2U) = H(X_2X_{12}|X_2U) = H(X_{12}) = 1 \\ &I(X_1X_2;Y_3) = H(Y_3) = H(X_{11}X_{12}) = 2. \end{aligned} \quad (9.25)$$

Thus, we find that $R = 2$ bits per clock tick are achievable, which is clearly optimal.

Example 9.2. Suppose the relay channel is *semi-deterministic* in the sense that $Y_2 = f(X_1, X_2)$ for some function $f(\cdot)$. We can then choose $U = Y_2$ without violating the Markov chain $U - [X_1, X_2] - [Y_2, Y_3]$ and find that (9.24) reduces to

$$R = \max_{P_{X_1X_2}(\cdot)} \min\left[H(Y_2|X_2) + I(X_1;Y_3|X_2Y_2), I(X_1X_2;Y_3)\right]. \quad (9.26)$$

But the capacity upper bound (9.13) is

$$\begin{aligned} C &\leq \max_{P_{X_1X_2}(\cdot)} \min\left[I(X_1;Y_2|X_2) + I(X_1;Y_3|X_2Y_2), I(X_1X_2;Y_3)\right] \\ &= \max_{P_{X_1X_2}(\cdot)} \min\left[H(Y_2|X_2) + I(X_1;Y_3|X_2Y_2), I(X_1X_2;Y_3)\right]. \end{aligned} \quad (9.27)$$

Partial-decode-and-forward therefore achieves the capacity of semi-deterministic relay channels [21] and this capacity is given by (9.26).

9.4.2 Compress-and-Forward

We next develop a strategy that uses block Markov encoding, superposition, *and* binning (see [16, Thm. 6]). This strategy is now often called Compress-and-Forward (CF).

Code Construction: Encoding is performed in $B+1$ blocks, and we again generate a *separate* code book for each block (see Figure 9.8 where $B+1=4$). For block b, $b=1,2,\ldots,B+1$, generate 2^{nR} codewords $x_{1b}^n(w)$, $w=1,2,\ldots,2^{nR}$, by choosing the $x_{1bi}(w)$ independently using $P_{X_1}(\cdot)$. Similarly, generate 2^{nR_2} codewords $x_{2b}^n(v)$, $v=1,2,\ldots,2^{nR_2}$, by choosing the $x_{2bi}(v)$ independently using $P_{X_2}(\cdot)$. Finally, introduce an auxiliary random variable $\hat{Y}_2$ that represents a quantized and compressed version of Y_2, and consider a distribution $P_{\hat{Y}_2|X_2}(\cdot)$. For each $x_{2b}^n(v)$, generate a "quantization" code book by generating $2^{n(R_2+R_2')}$ codewords $\hat{y}_{2b}^n(v,t,u)$, $t=1,2,\ldots,2^{nR_2'}$, $u=1,2,\ldots,2^{nR_2}$, by choosing the $\hat{y}_{2bi}(v,t,u)$ independently using $P_{\hat{Y}_2|X_2}(\cdot|x_{2bi}(v))$.

Source Terminal: The message w of 2^{nRB} bits is split into B equally sized blocks $w_1,w_2,\ldots,w_B$ of 2^{nR} bits each. In block b, $b=1,2,\ldots,B+1$, the source transmits $x_{1b}(w_b)$, where $w_{B+1}=1$.

Relay Terminal: In block $b=1$, the relay transmits $x_2^n(1)$. After block b, the relay has seen y_{2b}^n. The relay tries to find a $(\tilde{t}_b,\tilde{u}_b)$ such that

$$(\hat{y}_{2b}^n(v_b,\tilde{t}_b,\tilde{u}_b),x_{2b}^n(v_b),y_{2b}^n)\in T_\epsilon^n(P_{\hat{Y}_2X_2Y_2}). \tag{9.28}$$

If one or more such $(\tilde{t}_b,\tilde{u}_b)$ are found, then the relay chooses one of them, sets $v_{b+1}=\tilde{u}_b$, and transmits $x_{2(b+1)}(v_{b+1})$. If no such pair is found, the relay sets $v_{b+1}=1$ and transmits $x_{2(b+1)}(1)$.

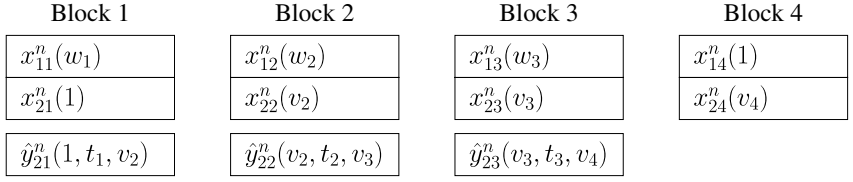

Fig. 9.8 A compress-and-forward strategy for the relay channel.

Sink Terminal: After block b, $b = 2, 3, \ldots, B+1$, the receiver has seen the sequence of outputs $y^n_{3(b-1)}$ and y^n_{3b}, and tries to find a $\tilde{v}_b$ such that

$$(x^n_{2b}(\tilde{v}_b), y^n_{3b}) \in T^n_\epsilon(P_{X_2Y_3}). \tag{9.29}$$

If one or more such $\tilde{v}_b$ are found, then the sink chooses one of them, and puts out this choice as $\hat{v}_b(3)$. If no such $\tilde{v}_b$ is found, the sink puts out $\hat{v}_b(3) = 1$. Next, the sink considers $y^n_{3(b-1)}$ and tries to find a $\tilde{t}_{b-1}$ such that

$$(\hat{y}^n_{2(b-1)}(\hat{v}_{b-1}(3), \tilde{t}_{b-1}, \hat{v}_b(3)), x^n_{2(b-1)}(\hat{v}_{b-1}(3)), y^n_{3(b-1)}) \in T^n_\epsilon(P_{\hat{Y}_2X_2Y_3}), \tag{9.30}$$

where $\hat{v}_{b-1}(3)$ is the sink terminal's estimate of v_{b-1}. If one or more such $\tilde{t}_{b-1}$ are found, then the sink chooses one of them, and calls this choice $\hat{t}_{b-1}(3)$. If no such $\tilde{t}_{b-1}$ is found, the sink sets $\hat{t}_{b-1}(3) = 1$. Finally, the sink tries to find a $\tilde{w}_{b-1}$ such that

$$\begin{aligned}&(x^n_{1(b-1)}(\tilde{w}_{b-1}), \hat{y}^n_{2b}(\hat{v}_{b-1}(3), \hat{t}_{b-1}(3), \hat{v}_b(3)), x^n_{2(b-1)}(\hat{v}_{b-1}(3)), y^n_{3(b-1)})\\ &\in T^n_\epsilon(P_{X_1\hat{Y}_2X_2Y_3}).\end{aligned} \tag{9.31}$$

If one or more such $\tilde{w}_{b-1}$ are found, then the sink chooses one of them, and calls this choice $\hat{w}_{b-1}$. If no such $\tilde{w}_{b-1}$ is found, the sink sets $\hat{w}_{b-1} = 1$.

Analysis: The analysis follows familiar steps, and we summarize the results.

(1) The relay quantization step requires

$$R_2 + R'_2 > I(\hat{Y}_2; Y_2|X_2). \tag{9.32}$$

(2) The sink's three decoding steps require

$$R_2 < I(X_2; Y_3) \tag{9.33}$$

$$R'_2 < I(\hat{Y}_2; Y_3|X_2) \tag{9.34}$$

$$\begin{aligned}R &< I(X_1; \hat{Y}_2X_2Y_3)\\ &= I(X_1; \hat{Y}_2Y_3|X_2).\end{aligned} \tag{9.35}$$

For the bounds (9.32) and (9.35), we choose $R_2' = I(\hat{Y}_2; Y_3|X_2) - \delta$ for appropriate δ, and require that $Y_3 - [X_2, Y_2] - \hat{Y}_2$ forms a Markov chain. We thus have, using (9.32),

$$\begin{aligned} R_2 &> I(\hat{Y}_2; Y_2|X_2) - I(\hat{Y}_2; Y_3|X_2) + \delta \\ &= I(\hat{Y}_2; Y_2|X_2 Y_3) + \delta. \end{aligned} \tag{9.36}$$

Combining (9.35) and (9.36), we have the achievable rate

$$R = I(X_1; \hat{Y}_2 Y_3|X_2), \tag{9.37}$$

where the joint distribution of the random variables factors as

$$P_{X_1}(a) P_{X_2}(b) P_{Y_2 Y_3|X_1 X_2}(c, d|a, b) P_{\hat{Y}_2|X_2 Y_2}(f|b, c) \tag{9.38}$$

for all a, b, c, d, f, and the joint distribution satisfies

$$I(\hat{Y}_2; Y_2|X_2 Y_3) \leq I(X_2; Y_3). \tag{9.39}$$

The rate (9.37) reminds one of a MIMO system with one transmit antenna and two receive antennas. After all, the destination receives both Y_3 and an approximation $\hat{Y}_2$ of Y_2.

Example 9.3. Consider again the relay channel of Figure 9.2 but now with the compress-and-forward strategy. We choose $\hat{Y}_2 = Y_2 = X_{11}$, and choose X_{11}, X_{12}, and X_2 to be independent coin-flipping random variables. We compute

$$\begin{aligned} I(X_1; \hat{Y}_2 Y_3|X_2) &= H(\hat{Y}_2 Y_3|X_2) = H(X_{11} X_{12} X_2|X_2) = H(X_{11} X_{12}) = 2 \\ I(\hat{Y}_2; Y_2|X_2 Y_3) &= H(\hat{Y}_2|X_2 Y_3) = H(X_{11}|X_2 X_{12}) = H(X_{11}) = 1 \\ I(X_2; Y_3) &= H(Y_3) - H(Y_3|X_2) \\ &= H(X_2 X_{12}) - H(X_2 X_{12}|X_2) = 1 \end{aligned} \tag{9.40}$$

and again find that $R = 2$ bits per clock tick are achievable. Thus, both the partial-decode-and-forward and compress-and-forward strategies achieve capacity.

Example 9.4. Consider the AWGN relay channel of Figure 9.3. We use compress-and-forward with X_1 and X_2 Gaussian, and $\hat{Y}_2 = Y_2 + \hat{Z}_2$, where $\hat{Z}_2$ is a Gaussian random variable with zero-mean, variance $\hat{N}_2$, and that is independent of all other random variables. The rate (9.37) is then

$$R = \frac{1}{2}\log\left(1 + \frac{P_1}{d^2(1+\hat{N}_2)} + P_1\right), \tag{9.41}$$

where the choice

$$\hat{N}_2 = \frac{P_1(1/d^2+1)+1}{P_2/(1-d)^2} \tag{9.42}$$

satisfies (9.39) with equality. The rate (9.41) is plotted in Figure 9.6 as the curve labeled "strategy 2." Observe from (9.41) and (9.42) that compress-and-forward achieves capacity as $P_2 \to \infty$ or $d \to 1$.

10

The Multiple Relay Channel

10.1 Problem Description and An Achievable Rate

We extend the relay channel of Figure 9.1 to include two or more relays, and we generalize the multi-hopping strategy of Section 9.2. Consider the two-relay model of Figure 10.1, and recall that the basic idea of the strategy of Section 9.2 is to "hop" the message blocks $w_1, w_2, \ldots, w_B$ successively to the relay, and then to the source. One can generalize this approach in a natural way as shown in Figure 10.2. This technique appeared in [38, 72, 73], and it generalizes the strategy of [11].

Code Construction: Consider a joint distribution $P_{X_1X_2X_3}(\cdot)$ and generate codewords $x_{3b}^n(w_1)$, $x_{2b}^n(w_1, w_2)$, and $x_{1b}^n(w_1, w_2, w_3)$ using $P_{X_3}(\cdot)$, $P_{X_2|X_3}(\cdot|x_{3bi}(w_1))$, and $P_{X_1|X_2X_3}(\cdot|x_{2bi}(w_1, w_2), x_{3bi}(w_1))$, respectively, for $b = 1, 2, \ldots, B+2$, $w_t = 1, 2, \ldots, 2^{nR}$ for $t = 1, 2, 3$, and $i = 1, 2, \ldots, n$.

Note that transmission is performed in $B+2$ blocks. The $x_{3b}^n(w_1)$ can be viewed as cloud centers, the $x_{2b}^n(w_1, w_2)$ as satellites, and the $x_{1b}^n(w_1, w_2, w_3)$ as satellites of the satellites.

Terminal 1: The message w of nRB bits is divided into B equally sized blocks $w_1, w_2, \ldots, w_B$ of nR bits each. In block b, $b = 1, 2, \ldots, B+2$,

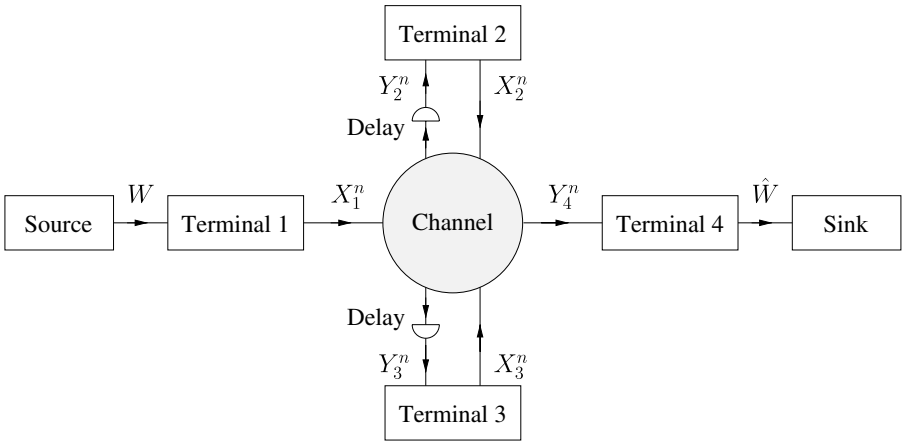

Fig. 10.1 The relay channel with two relays.

Block 1	Block 2	Block 3	Block 4
$x_{11}^n(1,1,w_1)$	$x_{12}^n(1,w_1,w_2)$	$x_{13}^n(w_1,w_2,w_3)$	$x_{14}^n(w_2,w_3,w_4)$
$x_{21}^n(1,1)$	$x_{22}^n(1,w_1)$	$x_{23}^n(w_1,w_2)$	$x_{24}^n(w_2,w_3)$
$x_{31}^n(1)$	$x_{32}^n(1)$	$x_{33}^n(w_1)$	$x_{34}^n(w_2)$

Block 5	Block 6	Block 7	Block 8
$x_{15}^n(w_3,w_4,w_5)$	$x_{16}^n(w_4,w_5,w_6)$	$x_{17}^n(w_5,w_6,1)$	$x_{18}^n(w_6,1,1)$
$x_{25}^n(w_3,w_4)$	$x_{26}^n(w_4,w_5)$	$x_{27}^n(w_5,w_6)$	$x_{28}^n(w_6,1)$
$x_{35}^n(w_3)$	$x_{36}^n(w_4)$	$x_{37}^n(w_5)$	$x_{38}^n(w_6)$

Fig. 10.2 Block-Markov superposition encoding for the multiple relay channel assuming the relays decode correctly.

the source terminal transmits $x_{1b}(w_{b-2}, w_{b-1}, w_b)$, where $w_{-1} = w_0 = w_{B+1} = w_{B+2} = 1$.

Terminal 2: After the transmission of block b is completed, relay terminal 2 uses y_{2b}^n and its past estimates $\hat{w}_{b-2}(2)$ and $\hat{w}_{b-1}(2)$, and tries to find a $\tilde{w}_b$ such that

$$(x_{1b}^n(\hat{w}_{b-2}(2), \hat{w}_{b-1}(2), \tilde{w}_b), \hat{x}_{2b}^n, \hat{x}_{3b}^n, y_{2b}^n) \in T_\epsilon^n(P_{X_1X_2X_3Y_2}), \qquad (10.1)$$

where $\hat{x}_{2b}^n$ and $\hat{x}_{3b}^n$ are the codewords corresponding to $\hat{w}_{b-2}(2)$ and $\hat{w}_{b-1}(2)$. If one or more such $\tilde{w}_b$ are found, then the relay chooses one

of them, calls this choice $\hat{w}_b(2)$, and transmits $x_{2(b+1)}(\hat{w}_{b-1}(2), \hat{w}_b(2))$ in block $b+1$. If no such $\tilde{w}_{2b}$ is found, the relay sets $\hat{w}_b(2) = 1$ and transmits $x^n_{2(b+1)}(\hat{w}_{b-1}(2), 1)$.

Terminal 3: After block b, relay terminal 3 uses $y^n_{3(b-1)}$, y^n_{3b}, and its past estimates $\hat{w}_{b-3}(3)$, $\hat{w}_{b-2}(3)$, and tries to find a $\tilde{w}_{b-1}$ such that

$$\begin{aligned}&(x^n_{1(b-1)}(\hat{w}_{b-3}(3), \hat{w}_{b-2}(3), \tilde{w}_{b-1}), \hat{x}^n_{2(b-1)}, \hat{x}^n_{3(b-1)}, y^n_{3(b-1)})\\ &\in T^n_\epsilon(P_{X_1X_2X_3Y_3})\end{aligned}$$

and

$$(x^n_{2b}(\hat{w}_{b-3}(3), \tilde{w}_{b-1}), \hat{x}^n_{3b}, y^n_{3b}) \in T^n_\epsilon(P_{X_2X_3Y_3}), \tag{10.2}$$

where $\hat{x}^n_{2(b-1)}$, $\hat{x}^n_{3(b-1)}$, and $\hat{x}^n_{3b}$ are the codewords corresponding to $\hat{w}_{b-3}(3)$ and $\hat{w}_{b-2}(3)$. If one or more such $\tilde{w}_{b-1}$ are found, then the relay chooses one of them, calls this choice $\hat{w}_{b-1}(3)$, and transmits $x^n_{3(b+1)}(\hat{w}_{b-1}(3))$ in block $b+1$. If no such $\tilde{w}_{b-1}$ is found, the relay sets $\hat{w}_{b-1}(3) = 1$ and transmits $x^n_{3(b+1)}(1)$.

Terminal 4: After block b, terminal 4 uses $y^n_{3(b-2)}$, $y^n_{3(b-1)}$, y^n_{3b}, and $\hat{w}_{b-4}(4)$, $\hat{w}_{b-3}(4)$, and tries to find a $\tilde{w}_{b-2}$ such that

$$\begin{aligned}&(x^n_{1(b-2)}(\hat{w}_{b-4}(4), \hat{w}_{b-3}(4), \tilde{w}_{b-2}), \hat{x}^n_{2(b-2)}, \hat{x}^n_{3(b-2)}, y^n_{4(b-2)})\\ &\qquad \in T^n_\epsilon(P_{X_1X_2X_3Y_4})\end{aligned}$$

and

$$(x^n_{2(b-1)}(\hat{w}_{b-3}(4), \tilde{w}_{b-2}), \hat{x}^n_{3(b-1)}, y^n_{4(b-1)}) \in T^n_\epsilon(P_{X_2X_3Y_4})$$

and

$$(x^n_{3b}(\tilde{w}_{b-2}), y^n_{4b}) \in T^n_\epsilon(P_{X_3Y_4}). \tag{10.3}$$

If one or more such $\tilde{w}_{b-2}$ are found, then the sink chooses one of them, and puts out this choice as $\hat{w}_{b-2}(4)$. If no such $\tilde{w}_{b-2}$ is found, the sink puts out $\hat{w}_{b-2}(4) = 1$.

Analysis: The analysis is similar to that in Section 9.2. Summarizing the result, we find that terminals 2, 3, and 4 can decode reliably if the following respective conditions hold:

$$R < I(X_1; Y_2|X_2X_3) \tag{10.4}$$

$$R < I(X_1; Y_3|X_2X_3) + I(X_2; Y_3|X_3) \tag{10.5}$$

$$R < I(X_1; Y_4|X_2X_3) + I(X_2; Y_4|X_3) + I(X_3; Y_4). \tag{10.6}$$

Combining (10.4)–(10.6), and letting B become large, we can approach the rate

$$R = \max_{P_{X_1X_2X_3}(\cdot)} \min\left[I(X_1;Y_2|X_2X_3), I(X_1X_2;Y_3|X_3), I(X_1X_2X_3;Y_4)\right]. \tag{10.7}$$

We remark that one can exchange the roles of terminals 2 and 3 and achieve a rate that might be larger than (10.7). One can further generalize the above approach to more than two relays in a natural way. That is, we will have *one bound per hop* or one bound per decoder. Moreover, there is a delay of *one block per hop* before the message w_b is decoded at the destination.

10.2 Cut-set Bounds

We wish to develop a capacity upper bound for relay channels. However, this bound is just as easy to develop for networks with multiple sources and sinks, so we take a more general approach (see also [18, Sec. 14.10]).

Consider the *Discrete Memoryless Network (DMN)* depicted in Figure 10.3. There are three messages, each destined for one or more sinks, and four terminals. We see that this network has multiple accessing (terminals 1 and 2 to terminal 3), broadcasting (terminal 2 to terminals 1 and 3), and relaying (terminal 1 to terminal 4 with the help

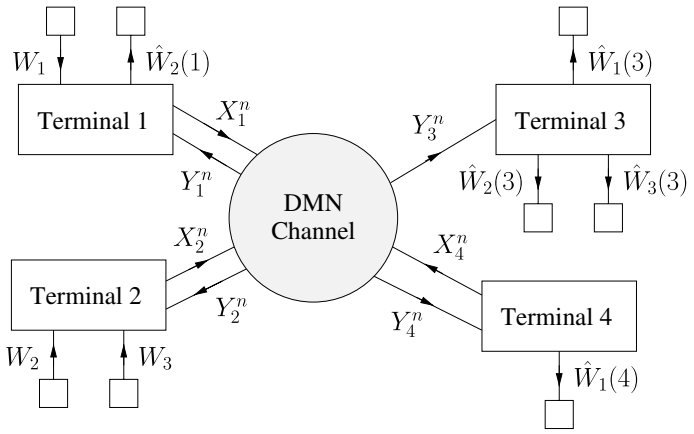

Fig. 10.3 A DMN with four terminals.

of terminal 2). More generally, a DMN has T terminals and a channel defined by a conditional probability distribution

$$P_{Y^T|X^T}\left(b^T|a^T\right), \tag{10.8}$$

where $X^T = X_1, X_2, \ldots, X_T$, $Y^T = Y_1, Y_2, \ldots, Y_T$, and X_t and Y_t are the respective inputs and outputs of terminal t. The other elements and rules of a DMN are similar to those already described in Section 9.1 for the relay channel, and we list them below.

- The network is *synchronous* in the sense that a universal *clock* governs the transmissions of the X_{ti} and Y_{ti}. The clock ticks n times and terminal t can transmit X_{ti} *after* clock tick $i-1$ and *before* clock tick i for $i = 1, 2, \ldots, n$. Terminal t receives Y_{ti} *at* clock tick i.
- There are M statistically independent messages W_m, $m = 1, 2, \ldots, M$. Message W_m has entropy nR_m bits so the rate of W_m is R_m bits per clock tick. Each message originates at exactly one vertex, but this message can be destined for any of the other $T-1$ vertices. Thus, each vertex has up to $2^{T-1}-1$ messages, one for each of the $2^{T-1}-1$ non-empty subsets of the other $T-1$ vertices.
- Let $\mathcal{M}(t)$ be the set of indexes of the messages originating at terminal t and define $W_{\mathcal{S}} = \{W_m : m \in \mathcal{S}\}$. The input X_{ti} is a function of $W_{\mathcal{M}(t)}$ and the channel outputs Y_t^{i-1}.
- The channel outputs Y_{ti} are noisy functions of the channel inputs X_{ti}, i.e., we have

$$Y_{ti} = f_t(X_{1i}, X_{2i}, \ldots, X_{Ti}, Z_i) \tag{10.9}$$

  for some functions $f_t(\cdot)$, $t = 1, 2, \ldots, T$, and for some noise random variable Z_i that is statistically independent of all other noise and message random variables.
- Let $\mathcal{D}_m$ be the set of terminals that decode W_m, and let $\hat{W}_m(t)$ be the estimate of W_m at node t, $t \in \mathcal{D}_m$. The *capacity region* $\mathcal{C}$ is the closure of the set of rate-tuples $(R_1, R_2, \ldots, R_M)$ for which, for sufficiently large n, there are

encoders and decoders so that the error probability

$$\Pr\left[\bigcup_{m=1}^{M}\bigcup_{t\in\mathcal{D}_m}\left\{\hat{W}_m(t)\neq W_m\right\}\right] \tag{10.10}$$

can be made as close to 0 as desired (but not necessarily exactly 0).

We return to our bound and partition the set of terminals $\mathcal{T}=\{1,2,\ldots,T\}$ into two sets $\mathcal{S}$ and $\bar{\mathcal{S}}$. We call the pair $(\mathcal{S},\bar{\mathcal{S}})$ a *cut*. We remark that the terminology "cut" usually refers to a set of *edges* of a network graph [24] and one can unify this approach with what follows (see [39, Sec. 3.7.1]).

We say that the cut $(\mathcal{S},\bar{\mathcal{S}})$ *separates* a message W_m and its estimate $\hat{W}_m(t)$ if W_m originates at a terminal in $\mathcal{S}$ and $t\in\bar{\mathcal{S}}$. Let $\mathcal{M}(\mathcal{S})$ be the set of messages separated from one of their estimates by the cut $(\mathcal{S},\bar{\mathcal{S}})$, and let $R_{\mathcal{M}(\mathcal{S})}$ be the sum of the rates of these messages. We further define $X_{\mathcal{S}}^n=\{X_t^n : t\in\mathcal{S}\}$, and similarly for $Y_{\mathcal{S}}^n$, $X_{\mathcal{S}}$, and $Y_{\mathcal{S}}$. The rates for reliable communication are bounded by

$$\begin{aligned}
nR_{\mathcal{M}(\mathcal{S})} &\overset{(a)}{\leq} I(W_{\mathcal{M}(\mathcal{S})};Y_{\bar{\mathcal{S}}}^n W_{\mathcal{M}(\bar{\mathcal{S}})})\\
&= I(W_{\mathcal{M}(\mathcal{S})};Y_{\bar{\mathcal{S}}}^n \,|\, W_{\mathcal{M}(\bar{\mathcal{S}})})\\
&= \sum_{i=1}^{n} H(Y_{\bar{\mathcal{S}}i}|Y_{\bar{\mathcal{S}}}^{i-1}W_{\mathcal{M}(\bar{\mathcal{S}})}) - H(Y_{\bar{\mathcal{S}}i}|Y_{\bar{\mathcal{S}}}^{i-1}W_{\mathcal{M}(\mathcal{T})})\\
&= \sum_{i=1}^{n} H(Y_{\bar{\mathcal{S}}i}|Y_{\bar{\mathcal{S}}}^{i-1}W_{\mathcal{M}(\bar{\mathcal{S}})}X_{\bar{\mathcal{S}}}^{i}) - H(Y_{\bar{\mathcal{S}}i}|Y_{\bar{\mathcal{S}}}^{i-1}W_{\mathcal{M}(\mathcal{T})}X_{\bar{\mathcal{S}}}^{i})\\
&\leq \sum_{i=1}^{n} H(Y_{\bar{\mathcal{S}}i}|X_{\bar{\mathcal{S}}i}) - H(Y_{\bar{\mathcal{S}}i}|X_{\bar{\mathcal{S}}i}X_{\mathcal{S}i})\\
&= \sum_{i=1}^{n} I(X_{\mathcal{S}i};Y_{\bar{\mathcal{S}}i}|X_{\bar{\mathcal{S}}i})\\
&\overset{(b)}{=} n\cdot I(X_{\mathcal{S}I};Y_{\bar{\mathcal{S}}I}|X_{\bar{\mathcal{S}}I}I) \qquad (10.11)\\
&\overset{(c)}{\leq} n\cdot I(X_{\mathcal{S}I};Y_{\bar{\mathcal{S}}I}|X_{\bar{\mathcal{S}}I}), \qquad (10.12)
\end{aligned}$$

where (a) follows by Fano's inequality, (b) by choosing I to be uniformly distributed over $\{1,2,\ldots,n\}$, and (c) because conditioning cannot increase entropy, and because

$$P_{IX_I^T Y_I^T}(i,a^T,b^T) = P_I(i)\; P_{X_I^T|I}(a^T|i)\; P_{Y^T|X^T}(b^T|a^T) \tag{10.13}$$

for all i, a^T and b^T. Note that in (10.13) we have used the channel distribution (10.8).

Let $\mathcal{R}(P_{X_I^T},\mathcal{S})$ be the set of non-negative rate-tuples $(R_1,R_2\ldots,R_M)$ that are permitted by (10.12). We note the following important fact: the distribution (10.13) is the *same* for all $\mathcal{S}$. We thus find that, for a given P_{X^T}, the reliably achievable rate-tuples must lie in the set

$$\mathcal{R}(P_{X^T}) = \bigcap_{\mathcal{S}\subseteq\mathcal{T}} \mathcal{R}(P_{X^T},\mathcal{S}). \tag{10.14}$$

Thus, the capacity region $\mathcal{C}$ must satisfy

$$\mathcal{C} \subseteq \bigcup_{P_{X^T}} \bigcap_{\mathcal{S}\subseteq\mathcal{T}} \mathcal{R}(P_{X^T},\mathcal{S}). \tag{10.15}$$

We emphasize that (10.15) involves first an intersection of regions and then a union, and not the other way around. We further remark that the intersection in (10.15) involves many regions for every $P_{X^T}(\cdot)$. However, we do not need to evaluate all of them: we can choose any subset of the regions, and we will still have a capacity outer bound given $P_{X^T}(\cdot)$. However, we *must* optimize (10.15) over *all* $P_{X^T}(\cdot)$. Fortunately, this is a convex optimization problem, since the mutual informations (10.12) are concave functions of $P_{X^T}(\cdot)$, and the set of $P_{X^T}(\cdot)$ is convex.

10.3 Examples

For example, consider the relay channel of Figure 9.1. The bound (10.15) on the capacity C is

$$C \leq \max_{P_{X_1X_2}(\cdot)} \min\left[I(X_1;Y_2Y_3|X_2), I(X_1X_2;Y_3)\right]. \tag{10.16}$$

For the Gaussian relay channel, the maximization over $P_{X_1X_2}(\cdot)$ becomes a maximization over densities $p_{X_1X_2}(\cdot)$ satisfying $E\left[X_1^2\right] \leq P_1$

and $E\left[X_2^2\right] \le P_2$. A conditional maximum entropy theorem ensures that $p_{X_1X_2}(\cdot)$ should be Gaussian. The resulting capacity upper bound is

$$R = \max_{0\le\rho\le 1} \min\left[\frac{1}{2}\log\left(1+(1-\rho^2)P_1\left(\frac{1}{d^2}+1\right)\right), \right. \\ \left. \frac{1}{2}\log\left(1+P_1+\frac{P_2}{(1-d)^2}+2\rho\frac{\sqrt{P_1P_2}}{1-d}\right)\right] \qquad (10.17)$$

and is plotted in Figure 9.6.

As a second example, consider the two-relay channel of Figure 10.1. There are four cuts to consider, namely $\mathcal{S}=\{1\}$, $\mathcal{S}=\{1,2\}$, $\mathcal{S}=\{1,3\}$, and $\mathcal{S}=\{1,2,3\}$. The bound (10.15) on the capacity C is

$$C \le \max_{P(x_1,x_2,x_3)} \min[I(X_1;Y_2Y_3Y_4|X_2X_3), I(X_1X_2;Y_3Y_4|X_3), \\ I(X_1X_3;Y_2Y_4|X_2), I(X_1X_2X_3;Y_4)]. \qquad (10.18)$$

As a third example, consider a broadcast channel $P_{Y_1Y_2|X}(\cdot)$. There are three cuts $\mathcal{S}=\{1\}$, $\mathcal{S}=\{1,2\}$, and $\mathcal{S}=\{1,3\}$, and the cut-set bound is the union over $P_X(\cdot)$ of the regions $\mathcal{R}(P_X)$ defined by

$$\begin{aligned} R_1 &\le I(X;Y_1) \\ R_2 &\le I(X;Y_2) \\ R_1+R_2 &\le I(X;Y_1Y_2). \end{aligned} \qquad (10.19)$$

For *deterministic* broadcast channels, the cut-set bound thus defines the capacity region. (As shown in Section 7.4, we can achieve any (R_1,R_2) satisfying $R_1 \le H(Y_1)$, $R_2 \le H(Y_2)$, and $R_1+R_2 \le H(Y_1Y_2)$ for any $P_X(\cdot)$ for such channels.)

Finally, consider a MAC $P_{Y|X_1X_2}(\cdot)$. The cut-set bound is

$$\begin{aligned} R_1 &\le I(X_1;Y|X_2) \\ R_2 &\le I(X_2;Y|X_1) \\ R_1+R_2 &\le I(X_1X_2;Y), \end{aligned} \qquad (10.20)$$

where *all* joint distributions $P_{X_1X_2}(\cdot)$ are permitted. The resulting outer bound is *not* the capacity region of the MAC in general, although it does give the right mutual information expressions.

11

The Multiaccess Channel with Generalized Feedback

11.1 Problem Description

The multiaccess channel with generalized feedback (MAC-GF) and with two transmitters (or users) and three sources is depicted in Figure 11.1. The sources put out statistically independent messages W_0, W_1, W_2 with nR_0, nR_1, nR_2 bits, respectively. The common message W_0 is seen by both encoders. The messages W_1 and W_2 appear only at the respective encoders 1 and 2. At time i, $i = 1, 2, \ldots, n$, encoder 1 maps (w_0, w_1) and its past received symbols $y_1^{i-1} = y_{11}, y_{12}, \ldots, y_{1(i-1)}$ to the channel input x_{1i}. Encoder 2 similarly maps (w_0, w_2) and y_2^{i-1} to its channel input x_{2i}. The channel $P_{Y_1Y_2Y|X_1X_2}(\cdot)$ has two inputs and three outputs. The decoder uses its output sequence y^n to compute its estimate $(\hat{w}_0, \hat{w}_1, \hat{w}_2)$ of (w_0, w_1, w_2), and the problem is to find the set of rate-tuples (R_0, R_1, R_2) for which one can make

$$P_e = \Pr\left[(\hat{W}_0, \hat{W}_1, \hat{W}_2) \neq (W_0, W_1, W_2)\right] \tag{11.1}$$

an arbitrarily small positive number. The closure of the region of achievable (R_0, R_1, R_2) is the MAC-GF capacity region $\mathcal{C}_{\text{MAC-GF}}$.

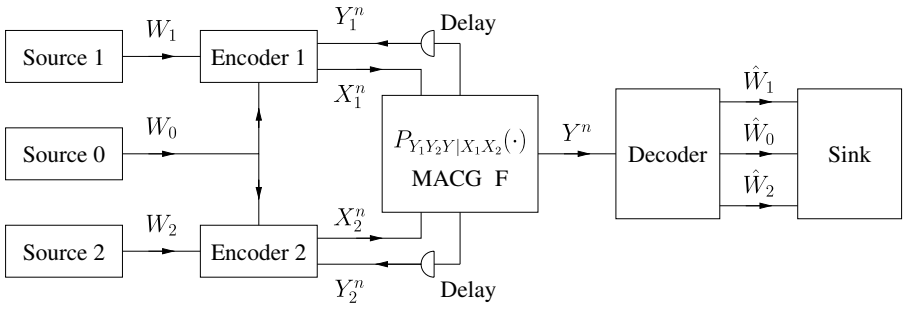

Fig. 11.1 The two-transmitter MAC with generalized feedback.

The terminology "generalized feedback" refers to the wide range of possible situations the model of Figure 11.1 encompasses. We list a few cases that have been studied in the past.

(1) The MAC without feedback has Y_1 and Y_2 being constants.
(2) The MAC with output feedback has $Y_1 = Y_2 = Y$. This model might be appropriate if the receiver has a high capacity link to the transmitters.
(3) The MAC with degraded output feedback has

$$Y_1 = f_1(Y, Z_{12}) \tag{11.2}$$
$$Y_2 = f_2(Y, Z_{12}), \tag{11.3}$$

where Z_{12} is a noise random variable. This model limits the capacity of the feedback links.
(4) The MAC-GF with independent noise has

$$Y_1 = f_1(X_1, X_2, Z_1) \tag{11.4}$$
$$Y_2 = f_2(X_1, X_2, Z_2) \tag{11.5}$$
$$Y = f(X_1, X_2, Z), \tag{11.6}$$

where Z_1, Z_2, and Z are statistically independent noise random variables. This model might fit a scenario where two mobile terminals cooperate to transmit their data to an access point or base station.
(5) The MAC with conferencing encoders has two noise-free links between the transmitters, as depicted in Figure 11.2. The link

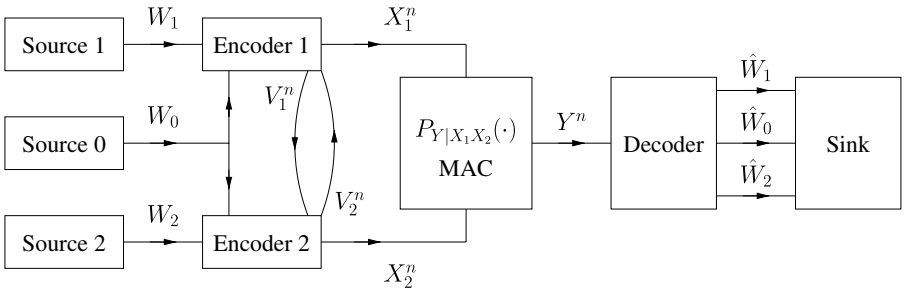

Fig. 11.2 The two-transmitter MAC with conferencing encoders.

from transmitter 1 to transmitter 2 has capacity C_{12}, and from transmitter 2 to transmitter 1 has capacity C_{21}. We can include this scenario in our MAC-GF model as follows. We abuse notation and write $\underline{X}_1 = [V_1, X_1]$ and $\underline{X}_2 = [V_2, X_2]$, where V_1 and V_2 have alphabet sizes $\log_2(C_{12})$ and $\log_2(C_{21})$, respectively. We further set $Y_1 = V_2$ and $Y_2 = V_1$ by defining the MAC-GF channel distribution to be

$$P_{Y_1Y_2Y|\underline{X}_1\underline{X}_2}(y_1, y_2, y\big|[v_1, x_1], [v_2, x_2]) \\ = 1(y_1 = v_2) \cdot 1(y_2 = v_1) \cdot P_{Y|X_1X_2}(y|x_1, x_2). \qquad (11.7)$$

(6) The relay channel is a special type of MAC-GF with $R_0 = R_2 = 0$ and Y_1 a constant.

We will derive an achievable rate region for the MAC-GF by using block-Markov superposition coding. We then specialize this region to the above cases.

11.2 An Achievable Rate Region

Variations of the MAC-GF were studied in [5, 11, 17, 25, 34, 48, 67]. We use block-Markov superposition coding where one new trick is to introduce *three* auxiliary random variables U, V_1, V_2. This seems rather complicated, but these random variables have natural interpretations. The random variable U represents information that is common to both transmitters, e.g., the message W_0. The random variable V_1 represents information that transmitter 1 sends to transmitter 2 to enable

cooperation. Similarly, V_2 represents information that transmitter 2 sends to transmitter 1. One might alternatively interpret the random variables as representing different *paths* through the network: a direct path U to the destination for W_0, two paths V_1, X_1 to the destination for W_1, where V_1 represents the path through encoder 2 and X_1 the direct path, and two paths V_2, X_2 to the destination for W_2, where V_2 is the path through encoder 1 and X_2 is the direct path. Another important trick is to use a *backward* decoding technique that was invented by Willems [67].

Code Construction: As for the relay channel, encoding is performed in $B+1$ blocks but we now use the *same* code books for each block (see Figure 11.3 where $B+1=3$). Consider a distribution $P_{UV_1V_2X_1X_2}$ that factors as $P_U P_{V_1X_1|U} P_{V_2X_2|U}$. We generate codebooks as depicted in Figure 11.4.

- Split the rates as $R_1 = R_1' + R_1''$ and $R_2 = R_2' + R_2''$, where all rate values are non-negative.
- Generate $2^{n(R_0+R_1'+R_2')}$ codewords $u^n(w_0, \tilde{w}_1, \tilde{w}_2)$, $w_0 = 1, 2, \ldots, 2^{nR_0}$, $\tilde{w}_1' = 1, 2, \ldots, 2^{nR_1'}$, $\tilde{w}_2' = 1, 2, \ldots, 2^{nR_2'}$, by choosing the $u_i(w_0, \tilde{w}_1', \tilde{w}_2')$ independently using $P_U(\cdot)$ for $i = 1, 2, \ldots, n$.
- Let $w = (w_0, \tilde{w}_1', \tilde{w}_2')$ and generate $2^{nR_1'}$ codewords $v_1^n(w, w_1')$, $w_1' = 1, 2, \ldots, 2^{nR_1'}$, by choosing the $v_{1i}(w, w_1')$ independently using $P_{V_1|U}(\cdot|u_i(w))$ for $i = 1, 2, \ldots, n$.
- For each tuple (w, w_1'), generate $2^{nR_1''}$ codewords $x_1^n(w, w_1', w_1'')$, $w_1'' = 1, 2, \ldots, 2^{nR_1''}$, by choosing the

Block 1	Block 2	Block 3
$u^n(w_{01}, 1, 1)$	$u^n(w_{02}, w_{11}', w_{21}')$	$u^n(w_{03}, w_{12}', w_{22}')$
$v_1^n([w_{01}, 1, 1], w_{11}')$	$v_1^n([w_{02}, w_{11}', w_{21}'], w_{12}')$	$v_1^n([w_{03}, w_{12}', w_{22}'], 1)$
$x_1^n([w_{01}, 1, 1], w_{11}', w_{11}'')$	$x_1^n([w_{02}, w_{11}', w_{21}'], w_{12}', w_{12}'')$	$x_1^n([w_{03}, w_{12}', w_{22}'], 1, w_{13}'')$
$v_2^n([w_{01}, 1, 1], w_{21}')$	$v_2^n([w_{02}, w_{11}', w_{21}'], w_{22}')$	$v_2^n([w_{03}, w_{12}', w_{22}'], 1)$
$x_2^n([w_{01}, 1, 1], w_{21}', w_{21}'')$	$x_2^n([w_{02}, w_{11}', w_{21}'], w_{22}', w_{22}'')$	$x_2^n([w_{03}, w_{12}', w_{22}'], 1, w_{23}'')$

Fig. 11.3 Block-Markov superposition encoding for a MAC-GF.

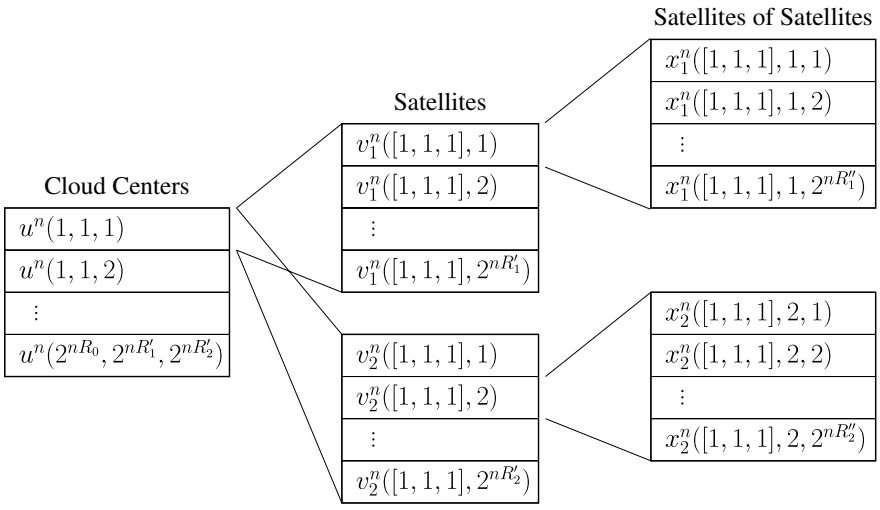

Fig. 11.4 A codebook for the MAC-GF with a common message.

$x_{1i}(w, w'_1, w''_1)$ independently using $P_{X_1|UV_1}(\cdot|u_i(w), v_{1i}(w, w'_1))$ for $i = 1, 2, \ldots, n$.

The codebooks for transmitter 2 are generated in the same way, except that there are now $2^{nR'_2}$ and $2^{nR''_2}$ codewords in each of the respective $v_2^n(\cdot)$ and $x_2^n(\cdot)$ codebooks.

Encoders: We use the block-Markov encoding strategy depicted in Figure 11.3. The message w_0 has $nR_0(B+1)$ bits. The message w_1 has $n(R_1B + R''_1)$ bits and is split into two parts w'_1 with nR'_1B bits and w''_1 with $nR''_1(B+1)$ bits, respectively (w_1 and w''_1 have an extra nR''_1 bits to make the decoding symmetric across blocks). The message w_2 is similarly divided into w'_2 and w''_2. Each of the five messages w_0, w'_1, w''_1, w'_2, and w''_2 is further divided into B sub-blocks of equal lengths for each message. We use the notation w_{0b} to refer to sub-block b of message w_0, and similarly for the other messages.

Let $w_b = (w_{0b}, w'_{1(b-1)}, w'_{2(b-1)})$ and suppose that transmitter 1 has somehow obtained $w'_{2(b-1)}$ before block b. In block b, $b = 1, 2, \ldots, B+1$,

encoder 1 transmits

$$x_1^n(w_b, w'_{1b}, w''_{1b}),$$

where $w'_{10} = w'_{1(B+1)} = 1$. Encoder 2 operates in the same fashion.

Decoders of Users 1 and 2: After the transmission of block b is completed, user 1 has seen y_{1b}^n. User 1 tries to find a $\tilde{w}'_{2b}$ such that

$$\begin{aligned}(u^n(\hat{w}_b), v_1^n(\hat{w}_b, w'_{1b}), x_1^n(\hat{w}_b, w'_{1b}, w''_{1b}), v_2^n(\hat{w}_b, \tilde{\mathbf{w}}'_{\mathbf{2b}}), y_{1b}^n) \\ \in T_\epsilon^n(P_{UV_1X_1V_2Y_1}),\end{aligned} \tag{11.8}$$

where $\hat{w}_b$ is user 1's estimate of w_b that contains an estimate of $w'_{2(b-1)}$. If one or more such $\tilde{w}'_{2b}$ are found, then user 1 chooses one of them, calls this choice $\hat{w}'_{2b}$. If no such $\tilde{w}'_{2b}$ is found, then user 1 sets $\hat{w}'_{2b} = 1$. User 2 operates in the same way.

Decoder: The decoder waits until the last block of transmission is completed. Given y_{B+1}^n, it tries to find a tuple $(\tilde{w}_{B+1}, \tilde{w}''_{1(B+1)}, \tilde{w}''_{2(B+1)})$ such that

$$\begin{aligned}(u^n(\tilde{w}_{B+1}), v_1^n(\tilde{w}_{B+1}, 1), x_1^n(\tilde{w}_{B+1}, 1, \tilde{w}''_{1(B+1)}), \\ v_2^n(\tilde{w}_{B+1}, 1), x_2^n(\tilde{w}_{B+1}, 1, \tilde{w}''_{2(B+1)}), y_{B+1}^n) \in T_\epsilon^n(P_{UV_1X_1V_2X_2Y}).\end{aligned} \tag{11.9}$$

If one or more such tuple is found, choose one and call it $(\hat{w}_{B+1}, \hat{w}''_{1(B+1)}, \hat{w}''_{2(B+1)})$ (note that $\hat{w}_{B+1} = [\hat{w}_{0(B+1)}, \hat{w}'_{1B}, \hat{w}'_{2B}]$). If no such triple is found, set $(\hat{w}_{B+1}, \hat{w}''_{1(B+1)}, \hat{w}''_{2(B+1)}) = (1, 1, 1)$.

Suppose the decoding for transmission block $B + 1$ is correct. The decoder next considers y_B^n and performs the same decoding step as above except that the first two "1"s in the arguments of (11.9) are replaced by $\hat{w}'_{1B}$, and the second two "1"s by $\hat{w}'_{2B}$. The decoder continues in this fashion until it reaches the first block. It should now be clear why this is called *backward* decoding.

Analysis: Consider block 1 and let $0 < \epsilon_1 < \epsilon < \mu_{UV_1X_2V_2X_2Y_1Y_2Y}$. We know that, with probability close to one, we will have

$$\begin{aligned}(u^n(w_1), v_1^n(w_1, w'_{11}), x_1^n(w_1, w'_{11}, w''_{11}), \\ v_2^n(w_1, w'_{21}), x_2^n(w_2, w'_{21}, w''_{22}), y_{11}^n, y_{21}^n, y_1^n) \in T_{\epsilon_1}^n(P_{UV_1X_1V_2X_2Y_1Y_2Y}).\end{aligned} \tag{11.10}$$

Consider user 1 and suppose that there was a $\tilde{w}'_{21} \neq w'_{21}$ such that

$$(u^n(w_1), v_1^n(w_1, w'_{11}), x_1^n(w_1, w'_{11}, w''_{11}), v_2^n(w_1, \tilde{\mathbf{w}}'_{\mathbf{21}}), y_{11}^n) \in T_\epsilon^n(P_{UV_1X_1V_2Y_1}). \tag{11.11}$$

We upper bound the probability of the event (11.11) by

$$\sum_{\tilde{w}'_{21} \neq w'_{21}} 2^{-n[I(V_2;Y_1|UV_1X_1)-2\epsilon H(V_2|UV_1X_1)]} < 2^{n[R'_1 - I(V_2;Y_1|UV_1X_1)+2\epsilon H(V_2|UV_1X_1)]}. \tag{11.12}$$

A similar bound can be derived for user 2.

Consider next the decoder and block $B+1$. We split the "overall" error event into 31 disjoint events that correspond to the 31 different ways in which one or more of the five messages is decoded incorrectly. For example, consider the event that there was a $\tilde{w}''_{0(B+1)} \neq w''_{0(B+1)}$ such that

$$(u^n(\tilde{w}_{B+1}), v_1^n(\tilde{w}_{B+1}, 1), x_1^n(\tilde{w}_{B+1}, 1, w''_{1(B+1)}), v_2^n(\tilde{w}_{B+1}, 1), x_2^n(\tilde{w}_{B+1}, 1, w''_{2(B+1)}), y_{B+1}^n) \in T_\epsilon^n(P_{UV_1X_1V_2X_2Y}). \tag{11.13}$$

Note that in this case *all five* codewords in (11.13) were chosen independent of the actually transmitted codewords. We can thus upper bound the probability of the event (11.13) by

$$\sum_{\tilde{w}_0 \neq w_0} 2^{-n[I(X_1X_2;Y)-2\epsilon H(UV_1X_1V_2X_2)]} < 2^{n[R_0 - I(X_1X_2;Y)+2\epsilon H(UV_1X_1V_2X_2)]}, \tag{11.14}$$

where we have taken advantage of the fact that

$$[U, V_1, V_2] - [X_1, X_2] - Y$$

forms a Markov chain. Fortunately, this rate bound is redundant, and so are many of the other bounds on the 31 possible error events. We leave the details of the analysis to the reader, and simply state the

decoder's four resulting rate bounds for reliable communication:

$$R_1'' \le I(X_1;Y|UV_1V_2X_2) \tag{11.15}$$

$$R_2'' \le I(X_2;Y|UV_1V_2X_1) \tag{11.16}$$

$$R_1'' + R_2'' \le I(X_1X_2;Y|UV_1V_2) \tag{11.17}$$

$$R_0 + R_1 + R_2 \le I(X_1X_2;Y). \tag{11.18}$$

In fact, 28 of the 31 rate bounds are dominated by (11.18). Finally, we combine the bounds (11.15)–(11.18) with the bound (11.12), and with the counterpart of (11.12) for user 2. The result is that the non-negative triples (R_0, R_1, R_2) satisfying the following four bounds are achievable:

$$R_1 \le I(X_1;Y|UV_1X_2) + I(V_1;Y_2|UX_2) \tag{11.19}$$

$$R_2 \le I(X_2;Y|UV_2X_1) + I(V_2;Y_1|UX_1) \tag{11.20}$$

$$\begin{aligned} R_1 + R_2 \le\ & I(X_1X_2;Y|UV_1V_2) \\ & + I(V_1;Y_2|UX_2) + I(V_2;Y_1|UX_1) \end{aligned} \tag{11.21}$$

$$R_0 + R_1 + R_2 \le I(X_1X_2;Y), \tag{11.22}$$

where $[V_1, X_1] - U - [V_2, X_2]$ forms a Markov chain. It is rather remarkable that our region requires only four rate bounds despite having used a complicated encoding and decoding procedure. Note that, by Markovity, we have been able to remove either V_1 or V_2 from most of the mutual information expressions in (11.19)–(11.21). The above bounds describe a region $\mathcal{R}(P_U, P_{V_1X_1|U}, P_{V_2X_2|U})$ with seven faces, four of which arise from (11.19)–(11.22), and three of which are non-negativity constraints on the rates (see Figure 11.5). We can further achieve rates in the union of such regions, i.e., we can achieve rates in

$$\mathcal{R} = \bigcup_{P_U, P_{V_1X_1|U}, P_{V_2X_2|U}} \mathcal{R}(P_U, P_{V_1X_1|U}, P_{V_2X_2|U}). \tag{11.23}$$

The methods of [67, Appendix A] can be used to show that this region is convex.

11.3 Special Cases

11.3.1 MAC Without Feedback

The MAC without feedback has Y_1 and Y_2 being constants and the reader can check that we may as well set $V_1 = V_2 = 0$ in (11.19)–(11.22).

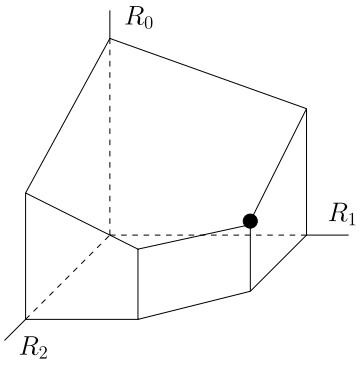

Fig. 11.5 The form of $\mathcal{R}(P_U, P_{V_1X_1|U}, P_{V_2X_2|U})$.

The resulting region turns out to be the capacity region derived in Section 8.3.

11.3.2 MAC with Output Feedback

Consider a MAC with output feedback, i.e., we have $Y_1 = Y_2 = Y$. Unlike the point-to-point transmission problem (see Section 3.9), now feedback can sometimes enlarge the capacity region. The bounds (11.19), (11.20), and (11.22) have no V_1 and V_2, and one can further check that the bound (11.21) is made redundant by choosing $V_1 = X_1$ and $V_2 = X_2$. This choice is therefore best. The region (11.19)–(11.22) is thus

$$R_1 \le I(X_1; Y|UX_2) \tag{11.24}$$

$$R_2 \le I(X_2; Y|UX_1) \tag{11.25}$$

$$R_0 + R_1 + R_2 \le I(X_1X_2; Y), \tag{11.26}$$

where $X_1 - U - X_2$ forms a Markov chain. The capacity region of the MAC with output feedback is still not known in general. Furthermore, for the AWGN channel

$$Y = X_1 + X_2 + Z \tag{11.27}$$

one can show that the region defined by (11.24)–(11.26) is strictly inside the capacity region. In fact, the capacity region for the AWGN channel

with $R_0 = 0$ turns out to be given by (11.24)–(11.26) but *without* the requirement that $X_1 - U - X_2$ forms a Markov chain. That is, the capacity region is the set of rate pairs (R_1, R_2) satisfying

$$R_1 \le I(X_1; Y|X_2) = \frac{1}{2}\log\left(1 + P_1(1-\rho^2)\right) \tag{11.28}$$

$$R_2 \le I(X_2; Y|X_1) = \frac{1}{2}\log\left(1 + P_2(1-\rho^2)\right) \tag{11.29}$$

$$R_1 + R_2 \le I(X_1X_2; Y) = \frac{1}{2}\log\left(1 + P_1 + P_2 + 2\sqrt{P_1P_2}\,\rho\right), \tag{11.30}$$

where $\rho = E[X_1X_2]/\sqrt{P_1P_2}$ takes on any value in $0 \le \rho \le 1$. Observe that for $\rho = 0$ the sum of (11.28) and (11.29) is larger than (11.30). Moreover, as we increase ρ from 0 to 1, there is a unique ρ^* for which the sum of (11.28) and (11.29) is the same as (11.30). We next describe how to achieve this boundary point of the capacity region.

11.3.3 Ozarow's Strategy

We develop a simple encoding strategy for the AWGN MAC with output feedback. Suppose we map W_1 with B_1 uniformly distributed bits to a point θ_1 in the interval $(-1/2, 1/2)$ by placing a (binary) decimal point in front of the bit string W_1 and interpreting the result as a (binary) fraction minus $(1/2 - 1/2^{B_1+1})$. This means that θ_1 has zero mean. For instance, if $W_1 = 0,1,0,0,1$ then we map W_1 to the point $(1/4 + 1/32) - (1/2 - 1/64)$. We similarly map W_2 to a point θ_2 in $(-1/2, 1/2)$.

Consider the first channel use. Users 1 and 2 transmit the respective

$$X_{11} = \sqrt{\frac{P_1}{\sigma_{10}^2}}\,\theta_1 \tag{11.31}$$

$$X_{21} = \sqrt{\frac{P_2}{\sigma_{20}^2}}\,\theta_2, \tag{11.32}$$

where $\sigma_{10}^2 = E[\theta_1^2]$ and $\sigma_{20}^2 = E[\theta_2^2]$ are both $1/12$. We have $E[X_{11}^2] = P_1$ and $E[X_{21}^2] = P_2$ by construction.

Consider now the receiver that computes linear minimum-mean square error (LMMSE) estimates of θ_1 and θ_2 given Y_1:

$$\hat{\theta}_{11} = \frac{E[\theta_1 Y_1]}{E[Y_1^2]}\, Y_1 \tag{11.33}$$

$$\hat{\theta}_{21} = \frac{E[\theta_2 Y_2]}{E[Y_1^2]}\, Y_1. \tag{11.34}$$

Note that $\hat{\theta}_{11}$ and $\hat{\theta}_{21}$ are identical. The transmitters can also generate this estimate because they have output feedback. Let the errors in the estimates after symbol i be

$$\epsilon_{1i} = \theta_1 - \hat{\theta}_{1i} \tag{11.35}$$

$$\epsilon_{2i} = \theta_2 - \hat{\theta}_{2i}. \tag{11.36}$$

In subsequent steps, the users correct the receiver's estimates by sending

$$X_{1i} = \sqrt{\frac{P_1}{\sigma^2_{1(i-1)}}}\, \epsilon_{1(i-1)} \tag{11.37}$$

$$X_{2i} = \sqrt{\frac{P_2}{\sigma^2_{2(i-1)}}}\, \epsilon_{2(i-1)}, \cdot m_{2i} \tag{11.38}$$

where $\sigma^2_{1i} = E[\epsilon^2_{1i}]$, $\sigma^2_{2i} = E[\epsilon^2_{2i}]$, and m_{2i} is a *modulation coefficient* taken to be either $+1$ or -1. Again, we have $E[X^2_{1i}] = P_1$ and $E[X^2_{2i}] = P_2$ by construction. The receiver computes the LMMSE estimate $\hat{\epsilon}_{k(i-1)}$ of $\epsilon_{k(i-1)}$ given Y_i and forms

$$\hat{\theta}_{1i} = \hat{\theta}_{1(i-1)} + \hat{\epsilon}_{1(i-1)} = \hat{\theta}_{1(i-1)} + \frac{E[\epsilon_{1(i-1)} Y_i]}{E[Y_i^2]}\, Y_i \tag{11.39}$$

$$\hat{\theta}_{2i} = \hat{\theta}_{2(i-1)} + \hat{\epsilon}_{2(i-1)} = \hat{\theta}_{2(i-1)} + \frac{E[\epsilon_{2(i-1)} Y_i]}{E[Y_i^2]}\, Y_i. \tag{11.40}$$

We outline an analysis of the convergence of the error variances σ^2_{ki} when one chooses the modulation coefficients. Consider first

$$\begin{aligned} \sigma^2_{1i} &= E[\epsilon^2_{1i}] \\ &= E[(\epsilon_{1(i-1)} - \hat{\epsilon}_{1(i-1)})^2] \end{aligned}$$

$$
\begin{aligned}
&= E[\epsilon_{1(i-1)}^2] - E[\hat{\epsilon}_{1(i-1)}^2] \\
&= E[\epsilon_{1(i-1)}^2] - \frac{E[\epsilon_{1(i-1)} Y_i]^2}{E[Y_i^2]} \\
&= E[\epsilon_{1(i-1)}^2] \cdot \left[1 - \frac{E[\epsilon_{1(i-1)} Y_i]^2}{E[Y_i^2] E[\epsilon_{1(i-1)}^2]} \right] \\
&= \sigma_{1(i-1)}^2 \cdot \left[1 - \frac{E[X_{1i} Y_i]^2}{E[Y_i^2] P_1} \right], \qquad (11.41)
\end{aligned}
$$

where the third step follows by the orthogonality principle. We remark that

$$
R_{1i} = \log \left(\sigma_{1(i-1)}^2 / \sigma_{1i}^2 \right) \qquad (11.42)
$$

is directly related to the rate of user 1.

Consider next the correlation

$$
\begin{aligned}
E[\epsilon_{1i}\, \epsilon_{2i}] &= E[(\epsilon_{1(i-1)} - \hat{\epsilon}_{1(i-1)})\, (\epsilon_{2(i-1)} - \hat{\epsilon}_{2(i-1)})] \\
&= E[\epsilon_{1(i-1)}\, \epsilon_{2(i-1)}] - \frac{E[\epsilon_{1(i-1)} Y_i] E[\epsilon_{2(i-1)} Y_i]}{E[Y_i^2]} \,. \qquad (11.43)
\end{aligned}
$$

We can rewrite this as

$$
\begin{aligned}
E[X_{1(i+1)} X_{2(i+1)}] = &\sqrt{\frac{\sigma_{1(i-1)}^2}{\sigma_{1i}^2} \cdot \frac{\sigma_{2(i-1)}^2}{\sigma_{2i}^2}} \cdot \frac{m_{2(i+1)}}{m_{2i}} \\
&\times \left[E[X_{1i} X_{2i}] - \frac{E[X_{1i} Y_i] E[X_{2i} Y_i]}{E[Y_i^2]} \right]. \qquad (11.44)
\end{aligned}
$$

We convert the above to a matrix recursion as follows. Let K_i be the covariance matrix of $[X_{1i}, X_{2i}]^T$. We then have

$$
E[X_{ki} Y_i] = (K_i \underline{1})_k \qquad (11.45)
$$

$$
E[Y_i^2] = \underline{1}^T K_i \underline{1} + 1 \qquad (11.46)
$$

$$
R_{ki} = \log \left(\frac{P_k (\underline{1}^T K_i \underline{1} + 1)}{P_k (\underline{1}^T K_i \underline{1} + 1) - (K_i \underline{1})_k} \right), \qquad (11.47)
$$

where $\underline{1} = [1, 1]^T$ and $(V)_k$ is the kth entry of the vector V. Using $m_{2i} = (-1)^{i-1}$, we further have

$$
K_{i+1} = \begin{bmatrix} e^{R_{1i}/2} & 0 \\ 0 & -e^{R_{2i}/2} \end{bmatrix} \left[K_i - \frac{(K_i \underline{1})(K_i \underline{1})^T}{\underline{1}^T K_i \underline{1} + 1} \right] \begin{bmatrix} e^{R_{1i}/2} & 0 \\ 0 & -e^{R_{2i}/2} \end{bmatrix} \qquad (11.48)
$$

that is a matrix recursion related to a discrete-time algebraic Riccati equation (DARE). One can show that (11.48) has a unique fixed point K. We take the determinant of both sides of (11.48) and find that this fixed point satisfies

$$\det K = \frac{e^{R_1} e^{R_2}}{\underline{1}^T K \underline{1} + 1} \det K, \tag{11.49}$$

where we have dropped the index i for fixed point values. Taking logarithms of both sides, we find that we have

$$R_1 + R_2 = \log\left(\underline{1}^T K \underline{1} + 1\right), \tag{11.50}$$

which implies that the fixed point $\rho = E[X_1 X_2]/\sqrt{P_1 P_2}$ described after (11.30) is the same as ρ^*.

11.3.4 MAC-GF with Independent Noise

Consider a MAC-GF with AWGN and the channel outputs

$$Y = X_1/d_1 + X_2/d_2 + Z \tag{11.51}$$

$$Y_1 = X_2/d_{21} + Z_1 \tag{11.52}$$

$$Y_2 = X_1/d_{12} + Z_2, \tag{11.53}$$

where the Z, Z_1, Z_2 are Gaussian, zero mean, unit variance, and independent of each other and the X_1 and X_2. The d_i and d_{ij} represent *distances* between the terminals, and they add a geometric component to the model. We again impose the constraints $E\left[X_1^2\right] \leq P_1$ and $E\left[X_2^2\right] \leq P_2$. Let V_1, V_2, X_1, X_2 be jointly Gaussian with

$$V_1 = (\sqrt{P_1}\rho_1)U + \sqrt{P_1'(1-\rho_1^2)}U_1' \tag{11.54}$$

$$V_2 = (\sqrt{P_2}\rho_2)U + \sqrt{P_1'(1-\rho_2^2)}U_2' \tag{11.55}$$

$$X_1 = V_1 + \sqrt{P_1''(1-\rho_1^2)}U_1'' \tag{11.56}$$

$$X_2 = V_2 + \sqrt{P_1''(1-\rho_2^2)}U_2'' \tag{11.57}$$

where U, U_1', U_2', U_1'', and U_2'' are independent, unit variance, Gaussian, and where $P_1 = P_1' + P_1''$, $P_2 = P_2' + P_2''$. We compute

$$I(V_1;Y_2|UX_2) = \frac{1}{2}\log\left(1 + \frac{P_1'(1-\rho_1^2)/d_{12}^2}{1 + P_1''(1-\rho_1^2)/d_{12}^2}\right) \tag{11.58}$$

$$I(V_2;Y_1|UX_1) = \frac{1}{2}\log\left(1 + \frac{P_2'(1-\rho_2^2)/d_{21}^2}{1 + P_2''(1-\rho_2^2)/d_{21}^2}\right) \tag{11.59}$$

$$I(X_1;Y|UV_1X_2) = \frac{1}{2}\log\left(1 + P_1''(1-\rho_1^2)/d_1^2\right) \tag{11.60}$$

$$I(X_2;Y|UV_2X_1) = \frac{1}{2}\log\left(1 + P_2''(1-\rho_2^2)/d_2^2\right) \tag{11.61}$$

$$I(X_1X_2;Y|UV_1V_2) = \frac{1}{2}\log\left(1 + P_1''(1-\rho_1^2)/d_1^2 + P_2''(1-\rho_2^2)/d_2^2\right) \tag{11.62}$$

$$I(X_1X_2;Y) = \frac{1}{2}\log\left(1 + P_1/d_1^2 + P_2/d_2^2 + 2\sqrt{(P_1/d_1^2)(P_2/d_2^2)}\,\rho_1\rho_2\right). \tag{11.63}$$

The achievable-rate bounds (11.19)–(11.22) are therefore

$$R_1 \le \frac{1}{2}\log\left(\left(1 + P_1''(1-\rho_1^2)/d_1^2\right)\left(1 + \frac{P_1'(1-\rho_1^2)/d_{12}^2}{1 + P_1''(1-\rho_1^2)/d_{12}^2}\right)\right) \tag{11.64}$$

$$R_2 \le \frac{1}{2}\log\left(\left(1 + P_2''(1-\rho_2^2)/d_2^2\right)\left(1 + \frac{P_2'(1-\rho_2^2)/d_{21}^2}{1 + P_2''(1-\rho_2^2)/d_{21}^2}\right)\right) \tag{11.65}$$

$$R_1 + R_2 \le \frac{1}{2}\log\Big(\left(1 + P_1''(1-\rho_1^2)/d_1^2 + P_2''(1-\rho_2^2)/d_2^2\right) \times \left(1 + \frac{P_1'(1-\rho_1^2)/d_{12}^2}{1 + P_1''(1-\rho_1^2)/d_{12}^2}\right)\left(1 + \frac{P_2'(1-\rho_2^2)/d_{21}^2}{1 + P_2''(1-\rho_2^2)/d_{21}^2}\right)\Big) \tag{11.66}$$

$$R_0 + R_1 + R_2 \le \frac{1}{2}\log\left(1 + P_1/d_1^2 + P_2/d_2^2 + 2\sqrt{(P_1/d_1^2)(P_2/d_2^2)}\,\rho_1\rho_2\right). \tag{11.67}$$

For example, suppose that $d_{12} = d_{21}$ and $d_{12} \ge d_1$ and $d_{12} \ge d_2$. One can check that the achievable mutual informations in (11.64)–(11.66) are then not larger than their corresponding mutual informations

(8.17)–(8.19) with P_1 and P_2 replaced by the respective P_1/d_1^2 and P_2/d_2^2. This means that we may as well set $V_1 = V_2 = 0$ and cooperate only through the U. This makes intuitive sense: if the users are both farther from each other than from the receiver, they need not decode each other's data. On the other hand, if $d_{12} \leq d_1$ or $d_{12} \leq d_2$ then cooperation by "multi-hopping" part of the message can help. This geometric insight is especially useful when performing resource allocation (power allocation) when the channels between the users and the receiver are time-varying.

11.3.5 MAC with Conferencing Encoders

Recall that the MAC with conferencing encoders has two noise-free links between the transmitters (see Figure 11.1). The link from transmitter 1 to transmitter 2 has capacity C_{12}, and from transmitter 2 to transmitter 1 has capacity C_{21}. We simply equate the V_1 and V_2 in Figure 11.2 with those in (11.19)–(11.22), make V_1 and V_2 independent of U, X_1, and X_2, and arrive at the achievable region

$$R_1 \leq I(X_1;Y|UX_2) + C_{12} \tag{11.68}$$

$$R_2 \leq I(X_2;Y|UX_1) + C_{21} \tag{11.69}$$

$$R_1 + R_2 \leq I(X_1X_2;Y|U) + C_{12} + C_{21} \tag{11.70}$$

$$R_0 + R_1 + R_2 \leq I(X_1X_2;Y), \tag{11.71}$$

where $X_1 - U - X_2$ forms a Markov chain. Willems [67, Sec. 8] showed that these expressions give the capacity region after taking the union described in (11.23).

A

Discrete Probability and Information Theory

A.1 Discrete Probability

We begin with basic definitions. A discrete *sample space* $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$ is the set of possible outcomes of a random experiment. An *event* is a subset of Ω including the empty set $\emptyset$ and the certain eventΩ . The *probability measure* $\Pr[\cdot]$ assigns each event a number in the interval $[0,1] = \{x : 0 \leq x \leq 1\}$ such that

$$\Pr[\Omega] = 1 \tag{A.1}$$

$$\Pr[\mathcal{A} \cup \mathcal{B}] = \Pr[\mathcal{A}] + \Pr[\mathcal{B}] \quad \text{if } \mathcal{A} \cap \mathcal{B} = \emptyset. \tag{A.2}$$

The *atomic events* are the events $\{\omega_i\}$, $i = 1, 2, \ldots, N$, so we have

$$\Pr[\mathcal{A}] = \sum_{\omega_i \in \mathcal{A}} \Pr[\omega_i], \tag{A.3}$$

where we have written $\Pr[\omega_i]$ as a shorthand for $\Pr[\{\omega_i\}]$. The *complement* $\mathcal{A}^c$ (or $\bar{\mathcal{A}}$) of event $\mathcal{A}$ is the set of all ω_i that are not in $\mathcal{A}$.

Example A.1. Consider a six-sided die and define $\Omega = \{1,2,3,4,5,6\}$ (see Figure A.1). A fair die has $\Pr[\omega_i] = 1/6$ for all i. The probability of

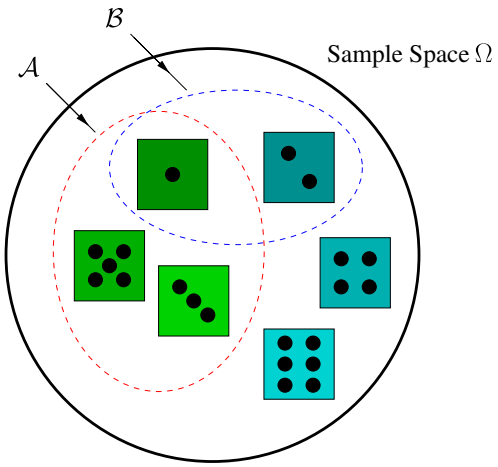

Fig. A.1 A sample space with six atomic events.

the event $\mathcal{A}$ is therefore $|\mathcal{A}|/|\Omega|$, where $|\mathcal{A}|$ be the number of elements in $\mathcal{A}$.

We say that "event $\mathcal{A}$ implies event $\mathcal{B}$," or $\mathcal{A} \Rightarrow \mathcal{B}$, if and only if $\mathcal{A} \subseteq \mathcal{B}$. By using (A.3), we thus find that $\mathcal{A} \Rightarrow \mathcal{B}$ gives $\Pr[\mathcal{A}] \leq \Pr[\mathcal{B}]$. Equation (A.3) also implies that

$$\Pr[\mathcal{A} \cup \mathcal{B}] = \Pr[\mathcal{A}] + \Pr[\mathcal{B}] - \Pr[\mathcal{A} \cap \mathcal{B}]. \tag{A.4}$$

We thus have

$$\Pr[\mathcal{A} \cup \mathcal{B}] \leq \Pr[\mathcal{A}] + \Pr[\mathcal{B}], \tag{A.5}$$

which is known as the *union bound*.

The *conditional* probability of the event $\mathcal{B}$ given the occurrence of the event $\mathcal{A}$ with $\Pr[\mathcal{A}] > 0$ is

$$\Pr[\mathcal{B}|\mathcal{A}] = \frac{\Pr[\mathcal{A} \cap \mathcal{B}]}{\Pr[\mathcal{A}]}. \tag{A.6}$$

The events $\mathcal{A}$ and $\mathcal{B}$ are said to be *independent* if

$$\Pr[\mathcal{A} \cap \mathcal{B}] = \Pr[\mathcal{A}] \cdot \Pr[\mathcal{B}]. \tag{A.7}$$

Thus, if $\Pr[\mathcal{A}] > 0$ then using (A.6) the events $\mathcal{A}$ and $\mathcal{B}$ are independent if $\Pr[\mathcal{B}|\mathcal{A}] = \Pr[\mathcal{B}]$. On the other hand, from (A.3) we have

$$\Pr[\mathcal{A} \cap \mathcal{B}] \leq \Pr[\mathcal{A}] \tag{A.8}$$

so that if $\Pr[\mathcal{A}] = 0$ then $\Pr[\mathcal{A} \cap \mathcal{B}] = 0$ and (A.7) is satisfied. Thus, if $\Pr[\mathcal{A}] = 0$ then $\mathcal{A}$ and $\mathcal{B}$ are always independent.

Example A.2. Consider our fair die and the events $\mathcal{A} = \{1,3,5\}$ and $\mathcal{B} = \{1,2\}$ in Figure A.1. We find that (A.7) is satisfied so $\mathcal{A}$ and $\mathcal{B}$ are independent.

A.2 Discrete Random Variables

A *discrete random variable* X is a mapping from Ω into a discrete and finite set $\mathcal{X}$ and its range is denoted by $X(\Omega)$. (More generally, $\Omega$and $\mathcal{X}$ might both be countably infinite.) The *probability distribution* $P_X(\cdot)$ is a mapping from $X(\Omega)$ into the interval $[0,1]$ such that

$$P_X(a) = \Pr[\omega : X(\omega) = a] \tag{A.9}$$

or simply $P_X(a) = \Pr[X = a]$. We thus have

$$P_X(a) \geq 0 \quad \text{for all } a \in \mathcal{X} \tag{A.10}$$

$$\sum_{a \in X(\Omega)} P_X(a) = 1. \tag{A.11}$$

Consider next n random variables $X^n = X_1, X_2, \ldots, X_n$ with domain Ω and range $X^n(\Omega) = X_1(\Omega) \times X_2(\Omega) \times \cdots \times X_n(\Omega)$. The *joint* probability distribution $P_{X^n}(\cdot)$ of these random variables is the mapping from $X^n(\Omega)$ into the interval $[0,1]$ such that

$$P_{X^n}(a^n) = \Pr\left[\bigcap_{i=1}^{n} \{X_i = a_i\}\right]. \tag{A.12}$$

We thus have

$$P_{X^n}(a^n) \geq 0 \quad \text{for all } a^n \in X^n(\Omega) \tag{A.13}$$

$$\sum_{a^n \in X^n(\Omega)} P_{X^n}(a^n) = 1. \tag{A.14}$$

We further have

$$P_{X^{n-1}}(a^{n-1}) = P_{X_1,X_2,\ldots,X_{n-1}}(a_1,a_2,\ldots,a_{n-1}) = \sum_{a_n \in X_n(\Omega)} P_{X_1,X_2,\ldots,X_{n-1}X_n}(a_1,a_2,\ldots,a_{n-1},a_n). \quad \text{(A.15)}$$

The random variables $X_1, X_2, \ldots, X_n$ are *statistically independent* if

$$P_{X^n}(a^n) = \prod_{i=1}^{n} P_{X_i}(a_i) \quad \text{for all } a^n \in X^n(\Omega). \quad \text{(A.16)}$$

Similarly, $X_1, X_2, \ldots, X_n$ are statistically independent conditioned on the event $\mathcal{A}$ with $\Pr[A] > 0$ if, for all $a^n \in X^n(\Omega)$, we have

$$\Pr\left[\bigcap_{i=1}^{n}\{X_i = a_i\} \middle| \mathcal{A}\right] = \prod_{i=1}^{n} \Pr[X_i = a_i|\mathcal{A}]. \quad \text{(A.17)}$$

The *support* of a random variable X is the set

$$\text{supp}(P_X) = \{a : a \in \mathcal{X}, P_X(a) > 0\}. \quad \text{(A.18)}$$

The *conditional* probability distribution $P_{Y|X}(\cdot)$ is a mapping from $\text{supp}(P_X) \times Y(\Omega)$ into the interval $[0,1]$ such that

$$P_{Y|X}(b|a) = \frac{P_{XY}(a,b)}{P_X(a)}. \quad \text{(A.19)}$$

Thus, using (A.16) we find that X and Y are statistically independent if and only if

$$P_{Y|X}(b|a) = P_Y(b) \quad \text{for all } (a,b) \in \text{supp}(P_X) \times Y(\Omega). \quad \text{(A.20)}$$

Similarly, we say that X and Y are statistically independent conditioned on Z if

$$P_{XY|Z}(a,b|c) = P_{X|Z}(a|c)P_{Y|Z}(b|c), \quad \text{(A.21)}$$

for all $(a,b,c) \in X(\Omega) \times Y(\Omega) \times \text{supp}(P_Z)$. Thus, we find that X and Y are statistically independent conditioned on Z if and only if

$$P_{Y|XZ}(b|a,c) = P_{Y|Z}(b|c), \quad \text{(A.22)}$$

for all $(a,b,c) \in \text{supp}(P_X) \times Y(\Omega) \times \text{supp}(P_Z)$. Alternatively, X and Y are statistically independent conditioned on Z if and only if

$$P_{X|YZ}(a|b,c) = P_{X|Z}(a|c), \quad \text{(A.23)}$$

for all $(a,b,c) \in X(\Omega) \times \text{supp}(P_{YZ})$.

A.3 Expectation

Consider a real-valued function $f(\cdot)$ with domain $X(\Omega)$. The *expectation* of the random variable $Y = f(X)$ is

$$E[Y] = E[f(X)] = \sum_{a \in \mathrm{supp}(P_X)} P_X(a)\, f(a). \tag{A.24}$$

One sometimes encounters the notation $\mathrm{E}_X[Y]$ if it is unclear which of the letters in the argument of $E[\cdot]$ are random variables. The *conditional* expectation of $f(X)$ given that the event $\mathcal{A}$ with $\Pr[\mathcal{A}] > 0$ occurred is

$$E[f(X)|\mathcal{A}] = \sum_{a:\Pr[\{X=a\}\cap\mathcal{A}]>0} \Pr[X = a|\mathcal{A}]\, f(a), \tag{A.25}$$

where the conditional probability $\Pr[X = a|\mathcal{A}]$ is defined as in (A.6). In particular, if $\mathcal{A} = \{Z = c\}$ and $P_Z(c) > 0$ we have

$$E[f(X)|Z = c] = \sum_{a \in \mathrm{supp}(P_{X|Z}(\cdot|c))} P_{X|Z}(a|c)\, f(a). \tag{A.26}$$

We can re-write the above definitions in a slightly different way. Let $\{\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_M\}$ be a collection of events that *partition* the sample space, i.e., we have

$$\bigcup_{m=1}^{M} \mathcal{B}_m = \Omega \text{and} \qquad \mathcal{B}_i \cap \mathcal{B}_j = \emptyset,\ i \neq j. \tag{A.27}$$

We can then write (A.24) as

$$\begin{aligned} E[f(X)] &= \sum_{i,a:\Pr[\mathcal{B}_i\cap\{X=a\}]>0} \Pr[\mathcal{B}_i \cap \{X = a\}]\, f(a) \\ &= \sum_{i:\Pr[\mathcal{B}_i]>0} \Pr[\mathcal{B}_i] \sum_{a:\Pr[\mathcal{B}_i\cap\{X=a\}]>0} \frac{\Pr[\mathcal{B}_i \cap \{X = a\}]}{\Pr[\mathcal{B}_i]}\, f(a) \\ &= \sum_{i:\Pr[\mathcal{B}_i]>0} \Pr[\mathcal{B}_i] \sum_{a:\Pr[\mathcal{B}_i\cap\{X=a\}]>0} \Pr[X = a|\mathcal{B}_i]\, f(a) \\ &= \sum_{i:\Pr[\mathcal{B}_i]>0} \Pr[\mathcal{B}_i]\, E[f(X)|\mathcal{B}_i] \end{aligned} \tag{A.28}$$

and (A.25) as

$$E[f(X)|\mathcal{A}] = \sum_{i:\Pr[\mathcal{B}_i\cap\mathcal{A}]>0} \Pr[\mathcal{B}_i|\mathcal{A}]\, E[f(X)|\mathcal{B}_i \cap \mathcal{A}]\,. \tag{A.29}$$

Example A.3. For a discrete random variable Y we can choose $\mathcal{B}_b = \{Y = b\}$ and write

$$E[f(X)] = \sum_{b \in \mathrm{supp}(P_Y)} P_Y(b)\, E[f(X)|Y = b] \tag{A.30}$$

$$E[f(X)|\mathcal{A}] = \sum_{b:\Pr[\{Y=b\}\cap\mathcal{A}]>0} \Pr[Y = b|\mathcal{A}]\, E[f(X,Y)|\{Y = b\} \cap \mathcal{A}]\,. \tag{A.31}$$

The identities (A.28)–(A.31) are known as the *Theorem on Total Expectation*.

A.4 Entropy

The *entropy* or *uncertainty* of the discrete random variable X is (see [26, 44, 19, 18] for more details)

$$H(X) = \sum_{a \in \mathrm{supp}(P_X)} -P_X(a) \log_2 P_X(a). \tag{A.32}$$

Alternatively, we can write

$$H(X) = E[-\log_2 P_X(X)]\,. \tag{A.33}$$

One sometimes encounters the notation $H(P_X)$ rather than $H(X)$ in order to simplify notation and/or to avoid confusion.

Note that we have chosen to evaluate the logarithm using the base 2, and we continue to follow this convention for discrete random variables below. Our entropy units are, therefore, *bits*. One can extend the definition (A.32) to *continuous* alphabets and certain continuous random variables by taking appropriate limits. We will often simply assume that the results carry over in a natural way to "well-behaved" continuous random variables (see Appendix B).

Example A.4. Suppose that $\mathcal{X} = \{0, 1\}$ and $P_X(0) = p$. The entropy of X is

$$H_2(p) = -p \log_2 p - (1 - p) \log_2 (1 - p) \tag{A.34}$$

and $H_2(\cdot)$ is called the *binary entropy function*. Note that $H_2(0) = H_2(1) = 0$, $H_2(0.11) \approx 1/2$, $H_2(1/2) = 1$, and $H_2(p)$ is maximized by $p = 1/2$. More generally, we have the following important result where we recall that $|\mathcal{X}|$ is the number of values in $\mathcal{X}$.

Theorem A.1.

$$0 \leq H(X) \leq \log_2 |\mathcal{X}| \tag{A.35}$$

with equality on the left if and only if there is one letter a in $\mathcal{X}$ with $P_X(a) = 1$, and with equality on the right if and only if $P_X(a) = 1/|\mathcal{X}|$ for all $a \in \mathcal{X}$, i.e., X is *uniform* over $\mathcal{X}$.

Proof. Consider first the left-hand side of (A.35) and note that for $0 < p \leq 1$ we have $-p \log_2 p \geq 0$ with equality if and only if $p = 1$. Thus, we have $H(X) \geq 0$ with equality if and only if there is one letter a in $\mathcal{X}$ with $P_X(a) = 1$. Consider next the right-hand side of (A.35) and observe that we have

$$0 \leq H(X) = E\left[\log_2 \frac{1}{|\mathcal{X}| P_X(X)}\right] + \log_2 |\mathcal{X}|. \tag{A.36}$$

But we have the inequality

$$\log_2(x) \leq \frac{x-1}{\ln(2)}, \tag{A.37}$$

where $\ln(x)$ is the natural logarithm of x, and where equality holds for $x > 0$ if and only if $x = 1$. Applying (A.37) to (A.36), we find that equality holds on the right in (A.35) if and only if $P_X(a) = 1/|\mathcal{X}|$ for all $a \in \mathcal{X}$. □

Example A.5. Consider $\mathcal{X} = \{0, 1, 2\}$ and $P_X(0) = P_X(1) = p/2$ and $P_X(2) = 1 - p$. We have

$$\begin{aligned} H(X) &= -\frac{p}{2}\log_2 \frac{p}{2} - \frac{p}{2}\log_2 \frac{p}{2} - (1-p)\log_2(1-p) \\ &= p + H_2(p) \end{aligned} \tag{A.38}$$

and $H(X) = \log_2(3)$ if $p = 2/3$.

Another interesting property is that $H_2(p)$ is concave in p since

$$\frac{d}{dp}H_2(p) = \log_2 \frac{1-p}{p} \tag{A.39}$$

$$\frac{d^2}{dp^2}H_2(p) = \frac{-1}{\ln(2)\, p(1-p)}. \tag{A.40}$$

We extend this property to random variables with larger alphabets in Section A.11.

A.5 Conditional Entropy

Consider a joint distribution $P_{XY}(\cdot)$, where the random variable Y takes on values in a discrete and finite alphabet $\mathcal{Y}$. The *conditional* entropy of X given the event $Y = b$ with probability $\Pr[Y = b] > 0$ is

$$\begin{aligned} H(X|Y=b) &= \sum_{a \in \operatorname{supp}(P_{X|Y}(\cdot|b))} -P_{X|Y}(a|b) \log_2 P_{X|Y}(a|b) \\ &= E\left[-\log_2 P_{X|Y}(X|Y) \middle| Y = b\right]. \end{aligned} \tag{A.41}$$

Using the same steps as in the previous section, one can show that

$$0 \le H(X|Y=b) \le \log_2 |\mathcal{X}| \tag{A.42}$$

with equality on the left if and only if $P_{X|Y}(a|b) = 1$ for some a, and with equality on the right if and only if $P_{X|Y}(a|b) = 1/|\mathcal{X}|$ for all a.

The conditional entropy of X given Y is the average of the values (A.41), i.e., we define

$$\begin{aligned} H(X|Y) &= \sum_{b \in \operatorname{supp}(P_Y)} P_Y(b) H(X|Y=b) \\ &= \sum_{(a,b) \in \operatorname{supp}(P_{XY})} -P_{XY}(a,b) \log_2 P_{X|Y}(a|b) \\ &= E\left[-\log_2 P_{X|Y}(X|Y)\right]. \end{aligned} \tag{A.43}$$

Again, one can show that

$$0 \le H(X|Y) \le \log_2 |\mathcal{X}| \tag{A.44}$$

with equality on the left if and only if for every b in $\operatorname{supp}(P_Y)$ there is an a such that $P_{X|Y}(a|b) = 1$, and with equality on the right if and

only if for every b in $\mathrm{supp}(P_Y)$ we have $P_{X|Y}(a|b) = 1/|\mathcal{X}|$ for all a. We say that Y *essentially determines* X if $H(X|Y) = 0$.

The above definitions and bounds extend naturally to more than two random variables. For example, consider the distribution $P_{XYZ}(\cdot)$. We define the conditional entropy of X given Y and the event $Z = c$ with $\Pr[Z = c] > 0$ as

$$\begin{aligned} H(X|Y, Z = c) &= \sum_{(a,b) \in \mathrm{supp}(P_{XY|Z}(\cdot|c))} -P_{XY|Z}(a,b|c) \log_2 P_{X|YZ}(a|b,c) \\ &= E\left[-\log_2 P_{X|YZ}(X|Y,Z) \middle| Z = c \right]. \end{aligned} \tag{A.45}$$

A.6 Joint Entropy

The *joint* entropy of X and Y is defined by considering the concatenation XY of X and Y as a new discrete random variable, i.e., we have

$$\begin{aligned} H(XY) &= \sum_{(a,b) \in \mathrm{supp}(P_{XY})} -P_{XY}(a,b) \log_2 P_{XY}(a,b) \\ &= E[-\log_2 P_{XY}(X,Y)]. \end{aligned} \tag{A.46}$$

Alternatively, one can represent XY by the vector $[X,Y]$ and write $H(X,Y)$ in place of $H(XY)$ or $H([X,Y])$. Theorem A.1 gives

$$0 \leq H(XY) \leq \log_2(|\mathcal{X}| \cdot |\mathcal{Y}|) \tag{A.47}$$

with equality on the left if and only if $P_{XY}(a,b) = 1$ for some (a,b), and with equality on the right if and only if $P_{XY}(a,b) = 1/(|\mathcal{X}|\,|\mathcal{Y}|)$ for all (a,b). Note that we have written the two variables in $H(XY)$ without punctuation and the reader should not confuse XY with "X multiplied by Y." Some authors prefer to write $H(X,Y)$ instead of $H(XY)$ and this is a matter of taste. We will follow the convention of not using punctuation if no confusion arises.

Using Bayes' rule for expanding joint probability distributions, one can expand the joint entropy using conditional entropies as

$$\begin{aligned} H(XY) &= H(X) + H(Y|X) \\ &= H(Y) + H(X|Y). \end{aligned} \tag{A.48}$$

More generally, we have

$$\begin{aligned} H(X_1X_2\ldots X_n) &= H(X_1) + H(X_2|X_1) + \cdots + H(X_n|X_1X_2\cdots X_{n-1}) \\ &= \sum_{i=1}^{n} H(X_i|X^{i-1}), \end{aligned} \tag{A.49}$$

where, as before, we have used the notation $X^j = X_1, X_2, \ldots, X_j$. Expansions such as (A.48) and (A.49) are called the *chain rule* for entropy.

Finally, we often use the following simple rule for manipulating conditional and joint entropies. Let $f(\cdot)$ and $g(\cdot)$ be functions whose domains are the ranges of $[X,Y]$ and Y, respectively. We have

$$H(X|Y) = H(Xf(X,Y)|Yg(Y)). \tag{A.50}$$

To prove (A.50), observe that the chain rule for entropy gives

$$\begin{aligned} &H(Xf(X,Y)|Yg(Y)) \\ &\quad = H(Xf(X,Y)g(Y)|Y) - H(g(Y)|Y) \\ &\quad = H(X|Y) + H(f(X,Y)g(Y)|XY) - H(g(Y)|Y). \end{aligned} \tag{A.51}$$

But the last two entropies in (A.51) are zero because $[X,Y]$ determines $f(X,Y)$ and $g(Y)$, and Y determines $g(Y)$.

A.7 Informational Divergence

The *informational divergence* (or *relative entropy* or *Kullback–Leibler distance*) between two distributions $P_X(\cdot)$ and $P_Y(\cdot)$ whose domains are the same alphabet $\mathcal{X}$ is defined as

$$\begin{aligned} D(P_X\|P_Y) &= \sum_{a\in\mathrm{supp}(P_X)} P_X(a)\log_2\frac{P_X(a)}{P_Y(a)} \\ &= E\left[\log_2\frac{P_X(X)}{P_Y(X)}\right] \end{aligned} \tag{A.52}$$

and we define $D(P_X\|P_Y) = \infty$ if $P_Y(a) = 0$ for some $P_X(a) > 0$. Note that, in general, we have $D(P_X\|P_Y) \neq D(P_Y\|P_X)$. Next, we prove the following fundamental result.

Theorem A.2.

$$D(P_X \| P_Y) \geq 0 \tag{A.53}$$

with equality if and only if $P_X(a) = P_Y(a)$ for all $a \in \text{supp}(P_X)$.

Proof. Write $D(P_X \| P_Y) = E[-\log_2(P_Y(X)/P_X(X))]$ and apply the inequality (A.37). □

Example A.6. Consider $\mathcal{X} = \{0,1\}$ and $P_X(0) = P_Y(0)(1+\epsilon)$, where $0 \leq \epsilon \leq 1/P_Y(0) - 1$. We compute

$$\begin{aligned} D(P_X \| P_Y) &= P_Y(0)(1+\epsilon)\log_2(1+\epsilon) \\ &\quad + [1 - P_Y(0)(1+\epsilon)]\log_2\left(\frac{1 - P_Y(0)(1+\epsilon)}{1 - P_Y(0)}\right) \end{aligned} \tag{A.54}$$

and we have $D(P_X \| P_Y) \geq 0$ with equality if and only if $\epsilon = 0$. We remark that $D(P_X \| P_Y)$ in (A.54) is convex in ϵ.

As in (A.52), given a third discrete random variable Z, we define the *conditional* informational divergence between $P_{X|Z}(\cdot)$ and $P_{Y|Z}(\cdot)$ as

$$\begin{aligned} D(P_{X|Z} \| P_{Y|Z} | P_Z) &= \sum_{b \in \text{supp}(P_Z)} P_Z(b) D\left(P_{X|Z}(\cdot|b) \| P_{Y|Z}(\cdot|b)\right) \\ &= \sum_{(a,b) \in \text{supp}(P_{XZ})} P_Z(b) P_{X|Z}(a|b) \log_2 \frac{P_{X|Z}(a|b)}{P_{Y|Z}(a|b)} \\ &= E\left[\log_2 \frac{P_{X|Z}(X|Z)}{P_{Y|Z}(X|Z)}\right]. \end{aligned} \tag{A.55}$$

Similar to (A.54), we have $D(P_{X|Z} \| P_{Y|Z} | P_Z) \geq 0$ with equality if and only if $P_{X|Z}(a|b) = P_{Y|Z}(a|b)$ for all $(a,b) \in \text{supp}(P_{XZ})$.

A.8 Mutual Information

The *mutual information* $I(X;Y)$ between two random variables X and Y with respective discrete and finite alphabets $\mathcal{X}$ and $\mathcal{Y}$ is defined as

$$I(X;Y) = H(X) - H(X|Y). \tag{A.56}$$

The name "mutual" describes the symmetry in the arguments of $I(X;Y)$, i.e., we have

$$I(X;Y) = H(Y) - H(Y|X). \tag{A.57}$$

Furthermore, using the chain rule (A.48) and the definition of informational divergence (A.52) we have

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(XY) \\ &= H(XY) - H(X|Y) - H(Y|X) \\ &= D(P_{XY} \| P_X P_Y) \\ &= \sum_{(a,b)\in \mathrm{supp}(P_{XY})} P_{XY}(a,b) \log_2 \frac{P_{XY}(a,b)}{P_X(a) P_Y(b)}. \end{aligned} \tag{A.58}$$

The last identity in (A.58) and Theorem A.2 imply the following inequalities.

Theorem A.3.

$$I(X;Y) \geq 0 \tag{A.59}$$

$$H(X|Y) \leq H(X) \tag{A.60}$$

$$H(XY) \leq H(X) + H(Y), \tag{A.61}$$

with equality in (A.59)–(A.61) if and only if X and Y are statistically independent.

The inequality (A.60) means that *conditioning cannot increase entropy*, or colloquially that *conditioning reduces entropy*. Note, however, that $H(X|Y=b)$ can be larger than $H(X)$.

Example A.7. Suppose X and Y are binary and $P_{XY}(0,0) = P_{XY}(0,1) = 0.11/2$, $P_{XY}(1,0) = 0.78$, and $P_{XY}(1,1) = 0$. We then have $H(X) = H_2(0.11) \approx 1/2$ but $H(X|Y=0) = 1$ and $H(X|Y=1) = 0$.

We can expand mutual information in a similar way as joint entropies, namely

$$\begin{aligned} I(X_1X_2\cdots X_n;Y) &= I(X_1;Y) + I(X_2;Y|X_1) \\ &\quad + \cdots + I(X_n;Y|X_1X_2\cdots X_{n-1}) \\ &= \sum_{i=1}^{n} I(X_i;Y|X^{i-1}). \end{aligned} \tag{A.62}$$

The expansions (A.62) is called the *chain rule* for mutual information.

The *conditional* mutual information between X and Y given a random variable Z is defined as

$$I(X;Y|Z) = H(X|Z) - H(X|YZ). \tag{A.63}$$

From the definition of conditional informational divergence in (A.55), we can also write

$$\begin{aligned} I(X;Y|Z) &= D(P_{XY|Z}\|P_{X|Z}P_{Y|Z}|P_Z) \\ &= \sum_{c\in\mathrm{supp}(P_Z)} P_Z(c) I(X;Y|Z=c), \end{aligned} \tag{A.64}$$

where

$$I(X;Y|Z=z) = H(X|Z=z) - H(X|Y,Z=z). \tag{A.65}$$

We further have

$$0 \le I(X;Y|Z) \le \min\left(H(X|Z), H(Y|Z)\right) \tag{A.66}$$

with equality on the left if and only if X and Y are independent given Z. If equality holds on the left, we say that

$$X - Z - Y \tag{A.67}$$

forms a Markov chain. Equality holds on the right in (A.66) if and only if $[Y,Z]$ essentially determines X, or $[X,Z]$ essentially determines Y, or both.

We can expand

$$I(X^n;Y|Z) = \sum_{i=1}^{n} I(X_i;Y|ZX^{i-1}). \tag{A.68}$$

Finally, let $f(\cdot)$, $g(\cdot)$, and $h(\cdot)$ be functions whose domains are the ranges of $[X,Z]$, $[Y,Z]$, and Z, respectively. We have

$$I(X;Y|Z) = I(Xf(X,Z);Yg(Y,Z)|Zh(Z)). \tag{A.69}$$

The proof of (A.69) follows easily from (A.63) and (A.50):

$$\begin{aligned} I(Xf(X,Z);Yg(Y,Z)|Zh(Z)) &= H(Xf(X,Z)|Zh(Z)) \\ &\quad - H(Xf(X,Z)|YZg(Y,Z)h(Z)) \\ &= H(X|Z) - H(X|YZ) \\ &= I(X;Y|Z). \end{aligned} \tag{A.70}$$

A.9 Establishing Conditional Statistical Independence

The random variables of multi-user problems are often related to each other in a complicated manner. It turns out that graphs are useful to ease the understanding of these relationships, and even to prove conditional statistical independence results.

A useful graphical tool in this respect is known as a *functional dependence graph* or FDG. An FDG is a graph where the vertices represent random variables and the edges represent the functional dependencies between the random variables [36, 37, 40]. For instance, suppose we have N_{RV} random variables that are defined by S_{RV} independent (or source) random variables by N_{RV} functions. An FDG $\mathcal{G}$ is a directed graph having $N_{RV} + S_{RV}$ vertices representing the random variables and in which edges are drawn from one vertex to another if the random variable of the former vertex is an argument of the function defining the random variable of the latter vertex.

Example A.8. Figure A.2 depicts the FDG for the first three uses of a channel with feedback. In this graph the channel input symbol X_i, $i = 1, 2, 3$, is a function of the message W and the past channel outputs Y^{i-1}. We have drawn the feedback links using dashed lines to emphasize the role that feedback plays. The output Y_i is a function of X_i and a noise random variable Z_i. The graph has $N_{RV} = 6$ random variables defined by $S_{RV} = 4$ independent random variables. The S_{RV} vertices

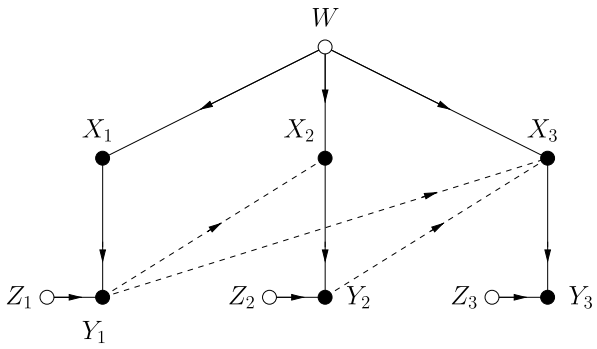

Fig. A.2 The FDG for the first three uses of a memoryless channel with feedback.

representing the independent W, Z_1, Z_2, and Z_3 are distinguished by drawing them with a hollow circle.

It turns out that the precise structure of FDGs lets one establish the conditional statistical independence of sets of random variables by using graphical procedures called d-separation and fd-separation ("d" for *dependence* and "fd" for *functional dependence*). By d-separation we mean the following reformulation of a definition in [49, p. 117] that is described in [36, 37].

Definition A.1. Let $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ be disjoint subsets of the vertices of an FDG $\mathcal{G}$. $\mathcal{Z}$ is said to *d-separate* $\mathcal{X}$ from $\mathcal{Y}$ if there is no path between a vertex in $\mathcal{X}$ and a vertex in $\mathcal{Y}$ after the following manipulations of the graph have been performed.

(1) Consider the subgraph $\mathcal{G}_{\mathcal{XYZ}}$ of $\mathcal{G}$ consisting of the vertices in $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$, as well as the edges and vertices encountered when moving *backward* one or more edges starting from any of the vertices in $\mathcal{X}$ or $\mathcal{Y}$ or $\mathcal{Z}$.
(2) In $\mathcal{G}_{\mathcal{XYZ}}$ delete all edges coming *out* of the vertices in $\mathcal{Z}$. Call the resulting graph $\mathcal{G}_{\mathcal{XY}|\mathcal{Z}}$.

(3) Remove the arrows on the remaining edges of $\mathcal{G}_{\mathcal{XY}|\mathcal{Z}}$ to obtain an undirected graph.

A fundamental result of [49, Sec. 3.3] is that d-separation establishes conditional independence in FDGs *having no directed cycles*. That is, if $\mathcal{G}$ is acyclic, $\mathcal{Z}$ d-separates $\mathcal{X}$ from $\mathcal{Y}$ in $\mathcal{G}$, and we collect the random variables of the vertices in $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ in the respective vectors $\underline{X}$, $\underline{Y}$ and $\underline{Z}$, then $I(\underline{X};\underline{Y}|\underline{Z}) = 0$ and $\underline{X} - \underline{Z} - \underline{Y}$ forms a Markov chain.

Example A.9. Consder Figure A.2 and choose $\mathcal{X} = \{W\}$, $\mathcal{Y} = \{Y_2\}$, and $\mathcal{Z} = \{X_1, X_2\}$. We find that $\mathcal{Z}$ d-separates $\mathcal{X}$ from $\mathcal{Y}$ so that $I(W;Y_2|X_1,X_2) = 0$.

A simple extension of d-separation is known as fd-separation which uses the fact that the FDG represents *functional* relations, and not only Markov relations as in Bayesian networks (see [36, Ch. 2],[40]). For fd-separation, after the second step above one removes all edges coming out of vertices that are disconnected from the S_{RV} source vertices in an undirected sense. We remark that fd-separation applies to an FDG $\mathcal{G}$ with cycles, as long as all subgraphs of $\mathcal{G}$ are also FDGs (see [36, Sec. 2]).

A.10 Inequalities

We state and prove several useful inequalities.

Markov Inequality: Let X be a *non-negative* real-valued random variable with mean $E[X]$. For $a > 0$, we have

$$\Pr[X \geq a] \leq \frac{E[X]}{a}. \tag{A.71}$$

Proof. We have $\Pr[X \geq a] = E[1(X \geq a)]$, where $1(\cdot)$ is the indicator function that takes on the value 1 if its argument is true and is 0 otherwise. We further note that $a\,1(X \geq a) \leq X$. We thus have $a \Pr[X \geq a] = E[a\,1(X \geq a)] \leq E[X]$. □

Example A.10. Suppose we set $X = |Y - E[Y]|$. Markov's inequality then gives *Tchebycheff's* inequality

$$\begin{aligned} \Pr[|Y - E[Y]| \geq a] &= \Pr\left[|Y - E[Y]|^2 \geq a^2\right] \\ &\leq \frac{\mathsf{Var}[Y]}{a^2}, \end{aligned} \tag{A.72}$$

where $\mathsf{Var}[Y]$ is the variance of Y and $a > 0$.

Example A.11. Suppose we set $X = e^{\nu Y}$ and $a = e^{\nu b}$. Markov's inequality then gives the *Chernoff bounds*

$$\begin{aligned} &\Pr[Y \geq b] \leq E\left[e^{\nu Y}\right] e^{-\nu b} \quad \text{for } \nu \geq 0 \\ &\Pr[Y \leq b] \leq E\left[e^{\nu Y}\right] e^{-\nu b} \quad \text{for } \nu \leq 0. \end{aligned} \tag{A.73}$$

Jensen's Inequality: We say that a real-valued function $f(\cdot)$ with domain interval $\mathcal{I}$ of non zero length on the real line is *convex* (or convex-$\cup$) on $\mathcal{I}$ if, for every interior point x_0 of $\mathcal{I}$, there exists a real number m (that may depend on x_0) such that

$$f(x) \geq f(x_0) + m(x - x_0) \quad \text{for all } x \in \mathcal{I}. \tag{A.74}$$

The convexity is *strict* if the inequality (A.74) is strict whenever $x \neq x_0$. One can show that an alternative and equivalent definition is that $f(\cdot)$ is convex on $\mathcal{I}$ if for every x_1 and x_2 in $\mathcal{I}$ we have

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda) f(x_2) \quad \text{for } 0 < \lambda < 1. \tag{A.75}$$

We say that $f(\cdot)$ is *concave* (or convex-$\cap$) on $\mathcal{I}$ if $-f(\cdot)$ is convex on $\mathcal{I}$. Observe that we are here considering functions of one variable, but the above definitions and the following results extend readily to many variables.

Let X be a real-valued random variable taking values in $\mathcal{I}$ and let $f(\cdot)$ be convex on $\mathcal{I}$. Jensen's inequality states that

$$f(E[X]) \leq E[f(X)]. \tag{A.76}$$

To prove (A.76), choose $x_0 = E[X]$ in (A.74), choose an m that satisfies (A.74) for this x_0, replace x with the random variable X, and take expectations of both sides of (A.74). Alternatively, if $f(\cdot)$ is concave on $\mathcal{I}$, then we have

$$f(E[X]) \geq E[f(X)]. \tag{A.77}$$

Furthermore, if $f(\cdot)$ is strictly convex (or concave), equality holds in (A.76) (or (A.77)) if and only if X is a constant.

Log-sum Inequality: For any non-negative a_i and positive b_i, $i = 1, 2, \ldots, n$, we have [19, p. 48], [18, p. 29]

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i \right) \log \frac{(\sum_{i=1}^{n} a_i)}{(\sum_{i=1}^{n} b_i)} \tag{A.78}$$

with equality if and only if a_i/b_i is the same for all i.

Proof. We choose $f(x) = x \log(x)$, and one can check that $f(\cdot)$ is strictly convex for positive x. We further choose X so that $X = a_i/b_i$ with probability $b_i/(\sum_j b_j)$. We thus have

$$E[f(X)] = \sum_{i=1}^{n} \frac{b_i}{\sum_j b_j} \cdot \frac{a_i}{b_i} \log \frac{a_i}{b_i}$$

$$f(E[X]) = \left(\sum_{i=1}^{n} \frac{b_i}{\sum_j b_j} \cdot \frac{a_i}{b_i} \right) \log \left(\sum_{i=1}^{n} \frac{b_i}{\sum_j b_j} \cdot \frac{a_i}{b_i} \right)$$

and Jensen's inequality (A.76) gives the desired result. □

Fano's Inequality: Fano's inequality gives a useful lower bound on error probability based on conditional entropy (see [18, p. 280]). Suppose both X and $\hat{X}$ take on values in the alphabet $\mathcal{X}$, and let $P_e = \Pr\left[\hat{X} \neq X\right]$. We have

$$H_2(P_e) + P_e \log_2(|\mathcal{X}| - 1) \geq H(X|\hat{X}). \tag{A.79}$$

We can interpret (A.79) as follows: P_e is bounded from below by some positive number if $H(X|\hat{X})$ is bounded from below by some positive number.

Proof. Let $E = 1(\hat{X} \neq X)$, where $1(\cdot)$ is the indicator function. We use the chain rule to expand $H(EX|\hat{X})$ in two ways as

$$\begin{aligned}
H(EX|\hat{X}) &= H(X|\hat{X}) + H(E|\hat{X}X) = H(X|\hat{X}) \\
H(EX|\hat{X}) &= H(E|\hat{X}) + H(X|\hat{X}E) \\
&= H(E|\hat{X}) + \Pr[E=0]\, H(X|\hat{X}, E=0) \\
&\quad + \Pr[E=1]\, H(X|\hat{X}, E=1) \\
&= H(E|\hat{X}) + \Pr[E=1]\, H(X|\hat{X}, E=1) \\
&\leq H(E|\hat{X}) + P_e \log_2(|\mathcal{X}| - 1) \\
&\leq H(E) + P_e \log_2(|\mathcal{X}| - 1) \\
&= H_2(P_e) + P_e \log_2(|\mathcal{X}| - 1),
\end{aligned}$$

where the first inequality follows because, given $\hat{X}$ and $E = 1$, X takes on at most $|\mathcal{X}| - 1$ values. □

Example A.12. Consider $\mathcal{X} = \{0, 1\}$ for which Fano's inequality is

$$H_2(P_e) \geq H(X|\hat{X}). \tag{A.80}$$

One can check that equality holds if $X = \hat{X} + Z$, where Z is independent of $\hat{X}$ and "+" denotes addition modulo-2.

Example A.13. Consider $\mathcal{X} = \{0, 1, 2\}$ and $X = \hat{X} + Z$, where Z is independent of $\hat{X}$, "+" denotes addition modulo-3, and $P_Z(i) = p_i$, $i = 0, 1, 2$. Fano's inequality is

$$H_2(1 - p_0) + (1 - p_0) \geq H(X|\hat{X}), \tag{A.81}$$

and one can check that equality holds if and only if $p_1 = p_2$ (see (A.38)).

A.11 Convexity Properties

Entropy, informational divergence, and mutual information have convexity properties that are useful for proving capacity theorems. We list and prove some of these below.

Convexity of Informational Divergence: $D(P_X \| P_Y)$ is convex (or convex-$\cup$) in the pair $(P_X(\cdot), P_Y(\cdot))$.

Proof. We use the log-sum inequality to write

$$\begin{aligned}
&\lambda P_X(a) \log_2 \frac{\lambda P_X(a)}{\lambda P_Y(a)} + (1-\lambda) Q_X(a) \log_2 \frac{(1-\lambda) Q_X(a)}{(1-\lambda) Q_Y(a)} \\
&\qquad \geq [\lambda P_X(a) + (1-\lambda) Q_X(a)] \log_2 \frac{\lambda P_X(a) + (1-\lambda) Q_X(a)}{\lambda P_Y(a) + (1-\lambda) Q_Y(a)},
\end{aligned}$$

where $0 \leq \lambda \leq 1$. Summing both sides over all appropriate $a \in \mathcal{X}$, we obtain the desired

$$\begin{aligned}
&\lambda D(P_X \| P_Y) + (1-\lambda) D(Q_X \| Q_Y) \\
&\qquad \geq D\left(\lambda P_X + (1-\lambda) Q_X \| \lambda P_Y + (1-\lambda) Q_Y\right).
\end{aligned}$$

□

Concavity of Entropy: $H(X)$ is concave (or convex-$\cap$) in $P_X(\cdot)$.

Proof. We again use the log-sum inequality to write

$$\begin{aligned}
&\lambda P_X(a) \log_2 \frac{\lambda P_X(a)}{\lambda} + (1-\lambda) Q_X(a) \log_2 \frac{(1-\lambda) Q_X(a)}{1-\lambda} \\
&\qquad \geq [\lambda P_X(a) + (1-\lambda) Q_X(a)] \log_2 (\lambda P_X(a) + (1-\lambda) Q_X(a)),
\end{aligned}$$

where $0 \leq \lambda \leq 1$. Summing both sides over all appropriate $a \in \mathcal{X}$, and multiplying by -1, we obtain the desired

$$\lambda H(P_X) + (1-\lambda) H(Q_X) \leq H\left(\lambda P_X + (1-\lambda) Q_X\right),$$

where we have written $H(X)$ as $H(P_X)$ to simplify the expression. □

Convexity of Mutual Information: $I(X;Y)$ is concave in $P_X(\cdot)$ if $P_{Y|X}(\cdot)$ is fixed, and $I(X;Y)$ is convex in $P_{Y|X}(\cdot)$ if $P_X(\cdot)$ is fixed.

Proof. Suppose $P_{Y|X}(\cdot)$ is fixed, and consider $I(X;Y) = H(Y) - H(Y|X)$. Note that $H(Y)$ is concave in $P_Y(\cdot)$. But $P_Y(\cdot)$ and $H(Y|X)$ are linear in $P_X(\cdot)$. Thus, $I(X;Y)$ is concave in $P_X(\cdot)$.

Suppose next that $P_X(\cdot)$ is fixed, and consider $I(X;Y) = D(P_X P_{Y|X} \| P_X P_Y)$. Note that $P_Y(\cdot)$ is linear in $P_{Y|X}(\cdot)$, so that $D(P_X P_{Y|X} \| P_X P_Y)$ is convex in $P_{Y|X}(\cdot)$. □

B

Differential Entropy

B.1 Definitions

The *differential entropy* of a real-valued and continuous random variable with density $p_X(\cdot)$ is defined in a similar manner as the entropy of a discrete random variable:

$$h(X) = \int_{\text{supp}(p_X)} -p_X(a) \log p_X(a)\, da. \tag{B.1}$$

Formally, one ofen adds "if this integral exists" but we shall permit differential entropies to take on the values $-\infty$ or $+\infty$. We can alternatively write

$$h(X) = E\left[-\log p_X(X)\right]. \tag{B.2}$$

Similarly, the *joint* differential entropy of real-valued and continuous random variables $X_1, X_2, \ldots, X_n$ with joint density $p_{X^n}(\cdot)$ is defined as

$$h(X^n) = \int_{\text{supp}(p_{X^n})} -p_{X^n}(\underline{a}) \log p_{X^n}(\underline{a})\, d\underline{a}. \tag{B.3}$$

We can alternatively write (B.3) as $h(\underline{X})$, where $\underline{X} = [X_1, X_2, \ldots, X_n]$.

Simple exercises show that for a nonzero real number c we have

$$\begin{array}{ll} \text{Translation rule:} & h(X+c) = h(X) \\ \text{Scaling rule:} & h(cX) = h(X) + \log|c|. \end{array} \tag{B.4}$$

Similarly, for a real-valued column vector $\underline{c}$ of dimension n and an invertible $n \times n$ matrix $\mathbf{C}$ we have

$$\begin{array}{ll} \text{Translation rule:} & h(\underline{X}+\underline{c}) = h(\underline{X}) \\ \text{Scaling rule:} & h(\mathbf{C}\underline{X}) = h(X) + \log|\det \mathbf{C}|, \end{array} \tag{B.5}$$

where $\det \mathbf{C}$ is the determinant of $\mathbf{C}$. We will, however, use the notation $|\mathbf{C}|$ for the determinant of $\mathbf{C}$ in the rest of the document.

Next, consider a joint density $p_{XY}(\cdot)$, and consider its conditional density $p_{Y|X}(\cdot) = p_{XY}(\cdot)/p_X(\cdot)$. We define

$$h(Y|X) = \int_{\text{supp}(p_{XY})} -p_{XY}(a,b) \log p_{Y|X}(b|a)\, da\, db. \tag{B.6}$$

We thus have $h(Y|X) = h(XY) - h(X)$. Note that we can define $h(Y|X)$ similar to (B.6) if the density $p_{Y|X}(\cdot|a)$ exists for every a but X does not have a density. Note further that, by conditioning on $X = a$ and using the translation rule in (B.4), for any real constant c we obtain

$$h(Y + cX|X) = h(Y|X). \tag{B.7}$$

B.2 Uniform Random Variables

An interesting observation is that, in contrast to $H(X)$, the differential entropy $h(X)$ can be *negative*. For example, consider the *uniform* density with $p_X(a) = 1/A$ for $a \in [0, A)$, where $[0, A) = \{x : 0 \le x < A\}$. We compute

$$h(X) = \log(A) \tag{B.8}$$

so that $h(X) \to -\infty$ as $A \to 0$. We can interpret such limiting densities as consisting of "Dirac-δ" (generalized) functions, and as representing discrete random variables. For instance, suppose that $p_X(a) = p_i/A$ for some integers i, $a \in [i, i + A)$, and $0 \le A \le 1$. As $A \to 0$, this density represents a discrete random variable $\tilde{X}$ with $P_{\tilde{X}}(i) = p_i$. We compute

$$h(X) = \sum_i -p_i \log(p_i/A) = \log(A) + H(\tilde{X}) \tag{B.9}$$

so $h(X)$ has increased as compared to (B.8). However, $h(X)$ still approaches $-\infty$ for small A.

In general, one must exercise caution when dealing with $h(X)$, where X might be discrete or have discrete components. For example, we have $h(Xf(X)) = h(X) + h(f(X)|X)$ but $h(f(X)|X) = -\infty$.

B.3 Gaussian Random Variables

Consider the Gaussian density

$$p_X(a) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2\sigma^2}(a-m)^2}, \tag{B.10}$$

where $m = E[X]$ and $\sigma^2 = \mathsf{Var}[X]$ is the variance of X. Inserting (B.10) into (B.1), we compute

$$h(X) = \frac{1}{2}\log\left(2\pi e\sigma^2\right). \tag{B.11}$$

We find that $h(X) < 0$ if $\sigma^2 < 1/(2\pi e)$. In fact, we have $h(X) \to -\infty$ as $\sigma^2 \to 0$.

More generally, consider a random column vector $\underline{X}$ of dimension n, mean $\underline{m}_{\underline{X}}$, and covariance matrix

$$\mathbf{Q}_{\underline{X}} = E\left[(\underline{X} - \underline{m}_{\underline{X}})(\underline{X} - \underline{m}_{\underline{X}})^T\right], \tag{B.12}$$

where the superscript "T" denotes transposition. Suppose $\underline{X}$ is Gaussian distributed, i.e., the density of $\underline{X}$ is

$$p_{\underline{X}}(\underline{a}) = \frac{1}{(2\pi)^{n/2}\left|\mathbf{Q}_{\underline{X}}\right|^{1/2}} \exp\left(-\frac{1}{2}(\underline{a} - \underline{m})^T \mathbf{Q}_{\underline{X}}^{-1}(\underline{a} - \underline{m})\right), \tag{B.13}$$

where $\left|\mathbf{Q}_{\underline{X}}\right|$ is the determinant of $\mathbf{Q}_{\underline{X}}$. Inserting (B.13) into (B.1), we compute

$$h(\underline{X}) = \frac{1}{2}\log\left((2\pi e)^n \left|\mathbf{Q}_{\underline{X}}\right|\right). \tag{B.14}$$

Note that $h(\underline{X})$ is negative for small $\left|\mathbf{Q}_{\underline{X}}\right|$.

Finally, suppose $p_{\underline{X}\,\underline{Y}}(\cdot)$ is Gaussian, where $\underline{X}$ has dimension n and $\underline{Y}$ has dimension m. We compute

$$h(\underline{Y}|\underline{X}) = h(\underline{X}\,\underline{Y}) - h(\underline{X}) = \frac{1}{2}\log\left((2\pi e)^m \left|\mathbf{Q}_{\underline{X}\,\underline{Y}}\right| / \left|\mathbf{Q}_{\underline{X}}\right|\right). \tag{B.15}$$

B.4 Informational Divergence

The informational divergence for continuous random variables X and Y is

$$D(p_X \| p_Y) = \int_{\text{supp}(p_X)} p_X(a) \log \frac{p_X(a)}{p_Y(a)} \, da \tag{B.16}$$

This definition extends to continuous random vectors $\underline{X}$ and $\underline{Y}$ that have the same dimension in the obvious way. The mutual information between X and Y is

$$\begin{aligned} I(X;Y) &= h(X) - h(X|Y) \\ &= D(p_{XY} \| p_X p_Y). \end{aligned} \tag{B.17}$$

We can derive similar relations for the continuous random variable versions of the other quantities in Appendix A. The bound $\ln(x) \leq x - 1$ again implies that

$$D(p_{\underline{X}} \| p_{\underline{Y}}) \geq 0 \tag{B.18}$$

with equality if and only if $p_X(a) = p_Y(a)$ for all $a \in \text{supp}(p_X)$. This further means that

$$I(X;Y) \geq 0 \tag{B.19}$$

$$h(X|Y) \leq h(X) \tag{B.20}$$

$$h(XY) \leq h(X) + h(Y) \tag{B.21}$$

with equality if and only if X and Y are independent.

B.5 Maximum Entropy

B.5.1 Alphabet Constraint

Recall that the uniform distribution maximizes the entropy of discrete random variables with alphabet $\mathcal{X}$. Similarly, the uniform density maximizes the differential entropy of continuous random variables with a support of finite volume. To prove this, suppose that $\underline{X}$ is confined to a set $\mathcal{S}$ in $\mathbb{R}^n$. Let $|\mathcal{S}|$ be the volume of $\mathcal{S}$ and let $\underline{U}$ be uniform over $\mathcal{S}$.

We use (B.18) and compute

$$\begin{aligned} 0 \le D(p_{\underline{X}} \| p_{\underline{U}}) &= \int_{\mathrm{supp}(p_{\underline{X}})} p_{\underline{X}}(\underline{a}) \log \left(p_{\underline{X}}(\underline{a}) |\mathcal{S}| \right) \, d\underline{a} \\ &\le -h(\underline{X}) + \log |\mathcal{S}|. \end{aligned} \tag{B.22}$$

We thus find that if $\underline{X}$ is limited to $\mathcal{S}$ then $h(\underline{X})$ is maximum and equal to $\log |\mathcal{S}|$ if and only if $p_{\underline{X}}(\underline{a}) = 1/|\mathcal{S}|$ for $\underline{a} \in \mathcal{S}$.

B.5.2 First Moment Constraint

For continuous random variables, one is often interested in *moment* constraints rather than volume constraints. For example, suppose that the alphabet of $\underline{X}$ is all of $\mathbb{R}^n$ and we wish to maximize $h(\underline{X})$ under the first-moment constraint (B.23)

$$E[\underline{X}] \le \underline{m}, \tag{B.23}$$

where the inequality $\underline{a} \le \underline{b}$ means that $a_i \le b_i$ for all entries a_i and b_i of the respective $\underline{a}$ and $\underline{b}$.

Observe that, without further constraints, we can choose $\underline{X}$ to be uniform over the interval $[-A, 0)$ for large positive A and make $h(\underline{X})$ arbitrarily large. We hence further restrict attention to *non-negative* $\underline{X}$, i.e., every entry X_i of $\underline{X}$ must be non-negative.

Let $\underline{E}$ have independent entries E_i that are exponentially distributed with mean m_i, i.e., we choose

$$p_{E_i}(a) = \begin{cases} \dfrac{1}{m_i} e^{-a/m_i} & a \ge 0 \\ 0 & a < 0. \end{cases} \tag{B.24}$$

We use the same approach as in (B.22) to compute

$$\begin{aligned} 0 \le D(p_{\underline{X}} \| p_{\underline{E}}) &= \int_{\mathrm{supp}(p_{\underline{X}})} p_{\underline{X}}(\underline{a}) \log \frac{p_{\underline{X}}(\underline{a})}{p_{\underline{E}}(\underline{a})} \, d\underline{a} \\ &= -h(\underline{X}) - \int_{\mathrm{supp}(p_{\underline{X}})} p_{\underline{X}}(\underline{a}) \log p_{\underline{E}}(\underline{a}) \, d\underline{a} \\ &= -h(\underline{X}) + \sum_i \log(e\, m_i) \end{aligned} \tag{B.25}$$

with equality in the first step if $\underline{X} = \underline{E}$. This proves the desired result, namely that (independent) exponential random variables maximize (differential) entropy under first moment and non-negativity constraints.

B.5.3 Second Moment Constraint

Suppose we wish to maximize $h(\underline{X})$ under the second-moment constraint

$$|\mathbf{Q}_{\underline{X}}| \leq D, \tag{B.26}$$

where D is some constant. For example, the constraint (B.26) occurs if we are restricting attention to $\underline{X}$ that satisfy

$$\mathbf{Q}_{\underline{X}} \preceq \mathbf{Q} \tag{B.27}$$

for some positive semidefinite $\mathbf{Q}$, where $\mathbf{A} \preceq \mathbf{B}$ means that $\mathbf{B} - \mathbf{A}$ is positive semi-definite (and hence $|\mathbf{A}| \leq |\mathbf{B}|$; see [31, p. 471]).

Let $\underline{G}$ be Gaussian with the same covariance matrix $\mathbf{Q}_{\underline{X}}$ as $\underline{X}$. We repeat the approach of (B.22) and (B.25) and compute

$$\begin{aligned} 0 \leq D(p_{\underline{X}} \| p_{\underline{G}}) &= \int_{\mathrm{supp}(p_{\underline{X}})} p_{\underline{X}}(\underline{a}) \log \frac{p_{\underline{X}}(\underline{a})}{p_{\underline{G}}(\underline{a})} \, d\underline{a} \\ &= -h(\underline{X}) - \int_{\mathrm{supp}(p_{\underline{X}})} p_{\underline{X}}(\underline{a}) \log p_{\underline{G}}(\underline{a}) \, d\underline{a} \\ &= -h(\underline{X}) + \frac{1}{2} \log \left((2\pi e)^n |\mathbf{Q}_{\underline{X}}| \right) \end{aligned} \tag{B.28}$$

with equality in the first step if $\underline{X} = \underline{G}$. This proves the desired result, namely that Gaussian random variables maximize (differential) entropy under the second moment constraints (B.26) or (B.27).

Finally, we prove a conditional version of the maximum entropy theorem. Suppose we have densities $p_{\underline{X}\,\underline{Y}}(\cdot)$ and $p_{\tilde{\underline{X}}\,\tilde{\underline{Y}}}(\cdot)$ with respective conditional densities $p_{\underline{Y}|\underline{X}}(\cdot)$ and $p_{\tilde{\underline{Y}}|\tilde{\underline{X}}}(\cdot)$. We define

$$D\left(p_{\underline{Y}|\underline{X}} \middle\| p_{\tilde{\underline{Y}}|\tilde{\underline{X}}} \middle| p_{\underline{X}}\right) = \int_{\mathrm{supp}(p_{\underline{X}\,\underline{Y}})} p_{\underline{X}\,\underline{Y}}(\underline{a},\underline{b}) \log \frac{p_{\underline{Y}|\underline{X}}(\underline{b}|\underline{a})}{p_{\tilde{\underline{Y}}|\tilde{\underline{X}}}(\underline{b}|\underline{a})} \, d\underline{a} \, d\underline{b}, \tag{B.29}$$

which one can show is non-negative. Suppose that $(\tilde{\underline{X}}, \tilde{\underline{Y}})$ is Gaussian with the same covariance matrix $\mathbf{Q}_{\underline{X}\,\underline{Y}}$ as $(\underline{X}, \underline{Y})$. We compute

$$\begin{aligned} & D\left(p_{\underline{Y}|\underline{X}} \| p_{\tilde{\underline{Y}}|\tilde{\underline{X}}} | p_{\underline{X}}\right) \\ &= -h(\underline{Y}|\underline{X}) - \int_{\mathrm{supp}(p_{\underline{X}\,\underline{Y}})} p_{\underline{X}\,\underline{Y}}(\underline{a}, \underline{b}) \log p_{\tilde{\underline{Y}}|\tilde{\underline{X}}}(\underline{b}|\underline{a}) \, d\underline{a} \, d\underline{b}. \\ &= -h(\underline{Y}|\underline{X}) + \frac{1}{2} \log\left((2\pi e)^m \left|\mathbf{Q}_{\underline{X}\,\underline{Y}}\right| / \left|\mathbf{Q}_{\underline{X}}\right|\right). \end{aligned} \tag{B.30}$$

This proves that, for fixed $\mathbf{Q}_{\underline{X}\,\underline{Y}}$, $h(\underline{Y}|\underline{X})$ is maximized by jointly Gaussian $\underline{X}$ and $\underline{Y}$.

B.6 Entropy Typicality

It turns out that we cannot use letter-typicality for continuous random variables. For example, consider the Gaussian random variable (B.10) with $m = 0$. The trouble with applying a letter-typicality test is that the probability mass function $P_X(x)$ is zero for of any letter x. However, we can use entropy-typicality if we replace the distribution $P_X(\cdot)$ in (1.4) with the density $p_X(\cdot)$. For example, we find that x^n is entropy-typical with respect to the density in (B.10) if

$$\left| \left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 \right) - \sigma^2 \right| < 2\sigma^2 \epsilon. \tag{B.31}$$

We can interpret (B.31) as follows: the average *energy* of an entropy-typical x^n is close to σ^2.

B.7 Entropy-Power Inequality

The (vector) entropy power inequality states that for independent random vectors $\underline{Y}$ and $\underline{Z}$ of dimension n, we have

$$2^{\frac{2}{n}h(\underline{Y}+\underline{Z})} \geq 2^{\frac{2}{n}h(\underline{Y})} + 2^{\frac{2}{n}h(\underline{Z})} \tag{B.32}$$

with equality if $\underline{Y}$ and $\underline{Z}$ are jointly Gaussian with proportional covariance matrices, i.e., $\mathbf{Q}_{\underline{Y}} = c\mathbf{Q}_{\underline{Z}}$ for some scalar c. The original result is due to Shannon [55, sec. 23] with further results by Stam and Blachman [9, 61]. Recent references on this inequality are [32, 33].

Acknowledgments

This survey began with lectures I gave at the ETH in Zurich in May 2004. Since that time, I was fortunate to be able to teach this course, or parts thereof, in several other locations, including Murray Hill, Bangalore, Adelaide, New York, Rutgers, Vienna, Seoul, and Lund. One pleasant surprise for me was how interested students were in what must have come across as a very abstract topic (it certainly was for me when I started!). But the positive feedback that I received encouraged me to "keep going" and refine the notes over time.

The choice of topics is, naturally, biased by my own interests. For example, I have focused more on channel coding rather than source coding. The theoretical development in the text reflects how my own understanding progressed during writing, and the organization is meant to "build up" knowledge from chapter to chapter rather than collect all closely related facts in one place. The reader is warned that the bibliography does not provide an exhaustive list of references on the subject matter, nor does it attempt to. Rather, the survey should motivate the student to study the subject further for himself or herself.

Unfortunatetly, space and time constraints prevented me from adding all the material that I had originally planned to. For instance, I

had hoped to add sections on source–channel coding, interference channels, routing, network coding, and a few other topics. Perhaps this will happen in the future.

I would like to thank Sergio Verdú for his interest in these notes and for encouraging me to publish in the Foundations and Trends series. The two anonymous reviewers provided very thorough feedback and helped me correct many errors. I also received detailed comments on the course by the students attending my lectures, and by Roy Yates and Bo Bernhardsson. Thank you all for your help.

I am especially grateful to Debasis Mitra who led the Math Center at Bell Labs from 2000–2007. This survey would not have appeared without his consistent support for my teaching assignments. My trips abroad were also made possible by the support of the Board of Trustees of the University of Illinois Subaward no. 04-217 under National Science Foundation Grant CCR-0325673. I hope that this survey will serve the NSF's charter of promoting the progress of science, at least the science of information theory.

References

[1] R. Ahlswede, "Multi-way communication channels," in *Proceedings of 2nd International Symposium Information Theory (1971)*, pp. 23–52, Tsahkadsor, Armenian S.S.R.: Publishing House of the Hungarian Academy of Sciences, 1973.

[2] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, *Network Flows: Theory, Algorithms, and Applications*. Upper Saddle River, New Jersey: Prentice Hall, 1993.

[3] N. Alon and J. H. Spencer, *The Probabilistic Method*. New York: Wiley, Second ed., 2000.

[4] A. Amraoui, S. Dusad, and R. Urbanke, "Achieving general points in the 2-user Gaussian MAC without time-sharing or rate-splitting by means of iterative coding," in *Proceedings of IEEE International Symposium on Information Theory*, p. 334, Lausanne, Switzerland, June 30–July 5 2002.

[5] M. R. Aref, *Information Flow in Relay Networks*. PhD thesis, Stanford, CA: Stanford University, October 1980.

[6] T. Berger, "Multiterminal source coding," in *The Information Theory Approach to Communications*, (G. Longo, ed.), pp. 171–231, Berlin, Germany: Springer Verlag, 1978.

[7] P. P. Bergmans, "Random coding theorem for broadcast channels with degraded components," *IEEE Transactions on Information Theory*, vol. 19, no. 2, pp. 197–207, March 1973.

[8] P. P. Bergmans, "A simple converse for broadcast channels with additive white Gaussian noise," *IEEE Transactions on Information Theory*, vol. 20, no. 2, pp. 279–280, March 1974.

[9] N. Blachman, "The convolution inequality for entropy powers," *IEEE Transactions on Information Theory*, vol. 11, no. 2, pp. 267–271, April 1965.

[10] S. I. Bross, A. Lapidoth, and M. A. Wigger, "The Gaussian MAC with conferencing encoders," in *Proceedings of IEEE International Symposium on Information Theory*, Toronto, Canada, July 6–11 2008.
[11] A. B. Carleial, "Multiple-access channels with different generalized feedback signals," *IEEE Transactions on Information Theory*, vol. 28, no. 6, pp. 841–850, November 1982.
[12] A. S. Cohen and A. Lapidoth, "The Gaussian watermarking game," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1639–1667, June 2002.
[13] M. H. M. Costa, "Writing on dirty paper," *IEEE Transactions on Information Theory*, vol. 29, no. 3, pp. 439–441, May 1983.
[14] T. Cover, "A proof of the data compression theorem of Slepian and Wolf for ergodic sources," *IEEE Transactions on Information Theory*, vol. 21, no. 2, pp. 226–228, March 1975.
[15] T. M. Cover, "Broadcast channels," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 2–14, January 1972.
[16] T. M. Cover and A. El Gamal, "Capacity theorems for the relay channel," *IEEE Transactions on Information Theory*, vol. 25, no. 5, pp. 572–584, September 1979.
[17] T. M. Cover and C. Leung, "An achievable rate region for the multiple-access channel with feedback," *IEEE Transactions on Information Theory*, vol. 27, no. 3, pp. 292–298, May 1981.
[18] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: John Wiley & Sons, 1991.
[19] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Channels*. Budapest: Akadémiai Kiadó, 1981.
[20] R. L. Dobrushin, "Information transmission in a channel with feedback," *Theory of Probabilistic Applications*, vol. 34, pp. 367–383, December 1958.
[21] A. El Gamal and M. Aref, "The capacity of the semideterministic relay channel," *IEEE Transactions on Information Theory*, vol. 28, no. 3, p. 536, May 1982.
[22] A. El Gamal and T. M. Cover, "Achievable rates for multiple descriptions," *IEEE Transactions on Information Theory*, vol. 28, no. 6, pp. 851–857, November 1982.
[23] A. El Gamal and E. C. van der Meulen, "A proof of Marton's coding theorem for the discrete memoryless broadcast channel," *IEEE Transactions on Information Theory*, vol. 27, no. 1, pp. 120–122, January 1981.
[24] L. R. Ford and D. R. Fulkerson, "Maximal flow through a network," *Canadian Journal of Mathematics*, vol. 8, pp. 399–404, 1956.
[25] N. Gaarder and J. Wolf, "The capacity region of a multiple-access discrete memoryless channel can increase with feedback," *IEEE Transactions on Information Theory*, vol. 21, no. 1, pp. 100–102, January 1975.
[26] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.
[27] R. G. Gallager, "Capacity and coding for degraded broadcast channels," *Problemy Peredachi Informatsii*, vol. 10, no. 3, pp. 3–14, July–September 1974.

[28] S. I. Gel'fand and M. S. Pinsker, "Coding for channels with random parameters," *Problems of Control and Information Theory*, vol. 9, no. 1, pp. 19–31, 1980.

[29] A. J. Grant, B. Rimoldi, R. L. Urbanke, and P. A. Whiting, "Rate-splitting multiple-access for discrete memoryless channels," *IEEE Transactions on Information Theory*, vol. 47, no. 3, pp. 873–890, March 2001.

[30] P. Gupta and P. R. Kumar, "Towards an information theory of large networks: An achievable rate region," *IEEE Transactions on Information Theory*, vol. 49, no. 8, pp. 1877–1894, August 2003.

[31] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge: Cambridge University Press, 1985.

[32] O. Johnson, "A conditional entropy power inequality for dependent random variables," *IEEE Transactions on Information Theory*, vol. 50, no. 8, pp. 1581–1583, August 2004.

[33] O. Johnson, *Information Theory and the Central Limit Theorem*. London, UK: Imperial College Press, 2004.

[34] R. C. King, *Multiple Access Channels with Generalized Feedback*. PhD thesis, Stanford, CA: Stanford University, March 1978.

[35] J. Körner and K. Marton, "General broadcast channels with degraded message sets," *IEEE Transactions on Information Theory*, vol. 23, no. 1, pp. 60–64, January 1977.

[36] G. Kramer, *Directed Information for Channels with Feedback,* volume ETH Series in Information Processing. Vol. 11, Konstanz, Germany: Hartung-Gorre Verlag, 1998.

[37] G. Kramer, "Capacity results for the discrete memoryless network," *IEEE Transactions on Information Theory*, vol. 49, no. 1, pp. 4–21, January 2003.

[38] G. Kramer, M. Gastpar, and P. Gupta, "Cooperative strategies and capacity theorems for relay networks," *IEEE Transactions on Information Theory*, vol. 51, no. 9, pp. 3037–3063, September 2005.

[39] G. Kramer, I. Marić, and R. D. Yates, "Cooperative communications," *Foundations and Trends in Networking*, vol. 1, no. 3–4, pp. 271–425, 2006.

[40] G. Kramer and S. A. Savari, "Edge-cut bounds on network coding rates," *Journal of Network and Systems Management*, vol. 14, no. 1, pp. 49–67, March 2006.

[41] Y. Liang and G. Kramer, "Rate regions for relay broadcast channels," *IEEE Transactions on Information Theory*, vol. 53, no. 10, pp. 3517–3535, October 2007.

[42] H. Liao, "A coding theorem for multiple access communications," in *Proceedings of IEEE International Symposium on Information Theory*, Asilomar, CA, 1972.

[43] K. Marton, "A coding theorem for the discrete memoryless broadcast channel," *IEEE Transactions on Information Theory*, vol. 25, no. 3, pp. 306–311, May 1979.

[44] J. L. Massey, *Applied Digital Information Theory*. Zurich, Switerland: ETH Zurich, 1980–1998.

[45] J. L. Massey, "Causality, feedback and directed information," in *Proceedings of IEEE International Symposium on Information Theory Applications*, pp. 27–30, Hawaii, USA, November 1990.

[46] A. Orlitsky and J. R. Roche, "Coding for computing," *IEEE Transactions on Information Theory*, vol. 47, no. 3, pp. 903–917, March 2001.

[47] L. Ozarow, "On a source-coding problem with two channels and three receivers," *Bell System Technical Journal*, vol. 59, no. 10, pp. 1909–1921, December 1980.

[48] L. Ozarow, "The capacity of the white Gaussian multiple access channel with feedback," *IEEE Transactions on Information Theory*, vol. 30, no. 4, pp. 623–629, July 1984.

[49] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann, 1988.

[50] S. S. Pradhan, R. Puri, and K. Ramchandran, "n-channel symmetric multiple-descriptions — Part I: (n,k) source-channel erasure codes," *IEEE Transactions on Information Theory*, vol. 50, no. 1, pp. 47–61, January 2004.

[51] R. Puri, S. S. Pradhan, and K. Ramchandran, "n-channel symmetric multiple-descriptions — Part II: an achievable rate-distortion region," *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1377–1392, April 2005.

[52] T. J. Richardson, A. Shokrollahi, and R. L. Urbanke, "Design of capacity-approaching low-density parity-check codes," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 619–637, February 2001.

[53] B. Rimoldi and R. Urbanke, "A rate-splitting approach to the Gaussian multiple-access channel," *IEEE Transactions on Information Theory*, vol. 42, no. 2, pp. 364–375, March 1996.

[54] H. Sato, "An outer bound to the capacity region of broadcast channels," *IEEE Transactions on Information Theory*, vol. 24, no. 3, pp. 374–377, May 1978.

[55] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, July and October 1948, (Reprinted in *Claude Elwood Shannon: Collected Papers*, pp. 5–83, (N. J. A. Sloane and A. D. Wyner, eds.) Piscataway: IEEE Press, 1993).

[56] C. E. Shannon, "The zero error capacity of a noisy channel," *IRE Transaction Information Theory*, vol. 2, pp. 221–238, September 1956, (Reprinted in *Claude Elwood Shannon: Collected Papers*, (N. J. A. Sloane and A. D. Wyner, eds.) pp. 221–238, Piscataway: IEEE Press, 1993).

[57] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," in *IRE International Convention Record*, pp. 142–163, March 1959. (Reprinted in *Claude Elwood Shannon: Collected Papers*, (N. J. A. Sloane and A. D. Wyner, eds.) pp. 325–350, Piscataway: IEEE Press, 1993.

[58] C. E. Shannon, "Two-way communication channels," in *Proceedings of 4th Berkeley Symposium on Mathematical Statistics and Probability*, (J. Neyman, ed.), pp. 611–644, Berkeley, CA: University California Press, 1961. (Reprinted in *Claude Elwood Shannon: Collected Papers*, (N. J. A. Sloane and A. D. Wyner, eds.), pp. 351–384, Piscataway: IEEE Press, 1993).

[59] D. Slepian and J. K. Wolf, "A coding theorem for multiple access channels with correlated sources," *Bell System Technical Journal*, vol. 52, pp. 1037–1076, September 1973.

[60] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Transactions on Information Theory*, vol. 19, no. 9, pp. 471–480, July 1973.

[61] A. Stam, "Some inequalities satisfied by the quantities of information of Fisher and Shannon," *Information Control*, vol. 2, pp. 101–112, July 1959.

[62] E. Telatar, "Capacity of multi-antenna Gaussian channels," *European Transactions on Telecommunication*, vol. 10, no. 6, pp. 585–595, November–December 1999.

[63] E. C. van der Meulen, *Transmission of Information in a T-Terminal Discrete Memoryless Channel*. PhD thesis, Berkeley, CA: University of California, January 1968.

[64] R. Venkataramani, G.Kramer, and V. K. Goyal, "Multiple description coding with many channels," *IEEE Transactions on Information Theory*, vol. 49, no. 9, pp. 2106–2114, September 2003.

[65] H. Wang and P. Viswanath, "Vector Gaussian multiple-description for individual and central receivers," *IEEE Transactions on Information Theory*, vol. 53, no. 6, pp. 2133–2153, June 2007.

[66] H. Weingarten, Y. Steinberg, and S. Shamai (Shitz), "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Transactions on Information Theory*, vol. 52, no. 9, pp. 3936–3964, September 2006.

[67] F. M. J. Willems, *Information Theoretical Results for the Discrete Memoryless Multiple Access Channel*. PhD thesis, Leuven, Belgium: Katholieke Universiteit, October 1982.

[68] F. M. J. Willems and E. C. van der Meulen, "The discrete memoryless multiple-access channel with cribbing encoders," *IEEE Transactions on Information Theory*, vol. 31, no. 3, pp. 313–327, May 1985.

[69] A. D. Wyner, "A theorem on the entropy of certain binary sequences and applications: Part II," *IEEE Transactions on Information Theory*, vol. 19, no. 6, pp. 772–777, November 1973.

[70] A. D. Wyner and J. Ziv, "A theorem on the entropy of certain binary sequences and applications: Part I," *IEEE Transactions on Information Theory*, vol. 19, no. 6, pp. 769–772, November 1973.

[71] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Transactions on Information Theory*, vol. 22, no. 1, pp. 1–10, January 1976.

[72] L.-L. Xie and P. R. Kumar, "A network information theory for wireless communication: scaling laws and optimal operation," *IEEE Transactions on Information Theory*, vol. 50, no. 5, pp. 748–767, May 2004.

[73] L.-L. Xie and P. R. Kumar, "An achievable rate for the multiple-level relay channel," *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1348–1358, April 2005.

[74] W. Yu, A. Sutivong, D. Julian, T. M. Cover, and M. Chiang, "Writing on colored paper," in *Proceedings of 2001 IEEE International Symposium Information Theory*, p. 302, Washington, DC, June 24–29 2001.

[75] Z. Zhang and T. Berger, "New results in binary multiple descriptions," *IEEE Transactions on Information Theory*, vol. 33, no. 4, pp. 502–521, July 1987.