Technische Universität München
Lehrstuhl für Nachrichtentechnik
Prof. Dr. sc. techn. Gerhard Kramer

Lecture Notes for

Advanced Information Theory

Course at Technische Universität Graz

March-April 2015

Prof. Dr. sc. techn. Gerhard Kramer

Institute for Communications Engineering
Technische Universität München
Building: N4, 2nd – 4th Floor

Edition: SS 2015
Authors: G. Kramer

Contents

Notation

We use standard notation for probabilities, random variables, entropy, mutual information, and so forth. Table 0.1 lists notation, some of it developed in the appendices, and we use this without further explanation in the main body of the text. We introduce the remaining notation as we go along. The reader is referred to the appendices for a review of relevant probability theory concepts.

Table 0.1.: Probability and Information Theory Notation.

Sets	
Ω	sample space
$\mathcal{A}^c$	complement of the set $\mathcal{A}$ in Ω
$\mathbb{N}_0$ and $\mathbb{N}_1$	natural numbers $\{0, 1, 2, \ldots\}$ and $\{1, 2, \ldots\}$
$\mathbb{Z}$, $\mathbb{R}$, $\mathbb{C}$	integers, real numbers, complex numbers
Strings, Vectors, Matrices	
x^n	the string $x_1 x_2 \ldots x_n$ or $x_1, x_2, \ldots, x_n$
$x^n y^m$	string concatenation: $x_1 x_2 \ldots x_n y_1 y_2 \ldots y_m$
$\underline{x}$	the vector $[x_1, x_2, \ldots, x_n]$
$\underline{x}^T$, $\underline{x}^\dagger$	transpose of $\underline{x}$, complex-conjugate transpose of $\underline{x}$
$\mathbf{H}$, $\mathbf{H}^T$, $\mathbf{H}^\dagger$	a matrix, its transpose, and its complex-conjugate transpose
$\lvert\mathbf{Q}\rvert$	determinant of the matrix $\mathbf{Q}$
Probability	
$\Pr[\mathcal{A}]$	probability of the event $\mathcal{A}$
$\Pr[\mathcal{A}\lvert\mathcal{B}]$	probability of event $\mathcal{A}$ conditioned on event $\mathcal{B}$
$P_X(\cdot)$	probability distribution of the random variable X
$P_{X\lvert Y}(\cdot)$	probability distribution of X conditioned on Y
$\mathrm{supp}(P_X)$	support of P_X, i.e., the set of a such that $P_X(a) > 0$
$p_X(\cdot)$	probability density of the random variable X
$p_{X\lvert Y}(\cdot)$	probability density of X conditioned on Y
Expectation and Variance	
$\mathsf{E}[X]$	expectation of the real-valued X
$\mathsf{E}[X\lvert\mathcal{A}]$	expectation of X conditioned on event $\mathcal{A}$
$\mathsf{E}[X\lvert Y]$	random variable that takes on the value $\mathsf{E}[X\lvert Y = y]$ if $Y = y$
$\mathsf{Var}[X]$	variance of the real-valued X
$\mathsf{Var}[X\lvert\mathcal{A}]$	variance of X conditioned on event $\mathcal{A}$
$\mathsf{Var}[X\lvert Y]$	random variable that takes on the value $\mathsf{Var}[X\lvert Y = y]$ if $Y = y$
$\mathsf{Cov}[X, Y]$	covariance of X and Y
$\mathbf{Q}_{\underline{X}}$	covariance matrix of $\underline{X}$
$\mathbf{Cov}[\underline{X}, \underline{Y}]$	covariance matrix of $\underline{X}$ and $\underline{Y}$
Information Theory	
$H(X)$	entropy of the discrete random variable X
$H(X\lvert Y)$	entropy of X conditioned on Y
$I(X;Y)$	mutual information between X and Y
$I(X;Y\lvert Z)$	mutual information between X and Y conditioned on Z
$D(P_X\Vert P_Y)$	informational divergence between P_X and P_Y
$D(P_X\Vert P_Y\lvert P_Z)$	informational divergence between P_X and P_Y conditioned on Z
$h(X)$	differential entropy of X
$h(X\lvert Y)$	differential entropy of X conditioned on Y
$H_2(\cdot)$	binary entropy function

Chapter 1.

Information Theory for Discrete Variables

1.1. Message Sets

Information theory was born as *A Mathematical Theory of Communication* as developed by Shannon in [1]. Shannon was particularly interested in *messages* and he wrote that:

> The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point.

But what is a message? Shannon suggested that a message has to do with *choice* and *sets*. In his words:

> The significant aspect is that the actual message is one *selected from a set* of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.

For example, suppose we wish to communicate the local weather to a friend. Suppose that our message set is

$$\{\text{sunny, cloudy, rain, thunderstorm}\}.$$

We could, e.g., communicate the weather "sunny" by transmitting this word. However, from an engineering perspective this approach is inefficient. We could instead represent the four elements in the message set by using only binary digits or *bits*:

$$\{00, 01, 10, 11\}$$

where 00 represents "sunny", 01 represents "cloudy", and so forth. The main point is, again in Shannon's words, that although

> the messages have meaning ... these semantic aspects of communication are irrelevant to the engineering problem.

1.2. Measuring Choice

Shannon was interested in defining

> a quantity which will measure, in some sense, how much information is "produced" by

choosing messages [1, Sec. 6]. He suggested that the *logarithm* of the number of elements in the message set

> can be regarded as a measure of the information produced when one message is chosen from the set, all choices being equally likely.

This logarithmic measure of information was also suggested by Hartley [2]. For example, consider a fair coin that takes on values in {Heads, Tails} and suppose we flip the coin n times. The string of coin flips takes on one of 2^n values, all equally likely, and the information produced is

$$\log_2 2^n = n \text{ bits.}$$

Note that the base of the logarithm simply defines the units of measurement.

For non-equally likely choices Shannon developed a measure that has the same form as the *entropy* in statistical mechanics. For example, consider a *biased* coin that takes on the value Heads with probability p and Tails with probability $1-p$. Shannon proposed that the information produced when flipping the coin n times is

$$n\left(p\log_2\frac{1}{p}+(1-p)\log_2\frac{1}{1-p}\right) \text{ bits.}$$

One motivation for this choice is as follows: approximately np of the n coin flips should take on the value Heads. Furthermore, each string of coin flips with approximately np Heads will be equally likely. If we take the logarithm of the number of such strings we obtain

$$\log_2\binom{n}{np} \approx n\left(p\log_2\frac{1}{p}+(1-p)\log_2\frac{1}{1-p}\right) \text{ bits.} \tag{1.1}$$

where we have used Stirling's approximation[1] to write

$$\binom{n}{np} = \frac{n!}{(np)!(n(1-p))!} \approx \frac{1}{\sqrt{2\pi p(1-p)n}}\cdot\frac{1}{p^{pn}(1-p)^{(1-p)n}}.$$

For instance, if $p=1/2$ then the right-hand side of (1.1) gives n bits, as expected. But if $p=0.11$ then we compute $n/2$ bits.

More generally, consider an experiment that has m possible outcomes with probabilities p_i, $i=1,2,\ldots,m$. Suppose we repeat the experiment n times.

[1] Stirling's approximation is $n! \approx \sqrt{2\pi n}(n/e)^n$.

We expect outcome i to occur approximately np_i times for all i, and the logarithm of the number of expected strings is

$$\log_2 \binom{n}{np_1, np_2, \ldots, np_m} \approx n \sum_{i=1}^{m} -p_i \log_2 p_i. \tag{1.2}$$

The amount of information obtained by observing the outcomes of the experiments thus seems to grow at the rate $\sum_{i=1}^{m} -p_i \log_2 p_i$.

1.3. Entropy

The word "entropy" was invented by Clausius [3, p. 390] from the Greek ἡ τροπὴ for "turn" or "change". Clausius was attempting to find a Greek word

- that was related to the German "Verwandlungsinhalt" which translates literally to "transformation content";
- for which the German version, in this case "Entropie", sounds similar to the German word for *energy*, in this case "Energie".

Calusius had a *physical* notion of "transformation content." Shannon instead related entropy to messages and sets. In other words, Shannon entropy is not necessarily physical: it has to do with *choice* and *uncertainty* in a broader sense than envisioned by Clausius.

We now become more formal in our treatment. Let $\mathrm{supp}(f)$ be the *support set* of the function f, i.e., the set of a such that $f(a) > 0$. We define the *entropy* or *uncertainty* of the discrete random variable X as

$$\boxed{H(X) = \sum_{a \in \mathrm{supp}(P_X)} -P_X(a) \log_2 P_X(a)}. \tag{1.3}$$

We remark that by taking the sum over the support of $P_X(\cdot)$ we avoid dealing with the expression $0 \cdot \log_2 0$. Many authors simply define $0 \cdot \log_2 0$ to be zero because $\lim_{x \to 0^+} x \log_2 x = 0$, where the notation "$0^+$" means that limit is taken from above. However, it is often instructive to be concerned about events with probability zero, see Sec. 1.7 below.

Example 1.1. The entropy $H(X)$ depends only on the probability distribution $P_X(\cdot)$, and not on what we call the letters a in the alphabet $\mathcal{X}$ of X. This idea is consistent with our discussion in Sec. 1.1. We thus have

$$H(X) = H(g(X)) \tag{1.4}$$

for any invertible function $g(\cdot)$, since $g(\cdot)$ simply relabels the letters in $\mathcal{X}$.

We can express (1.3) in an insightful way by considering $Y = -\log_2 P_X(X)$ to be a random variable that is a function of X. The entropy $H(X)$ is then

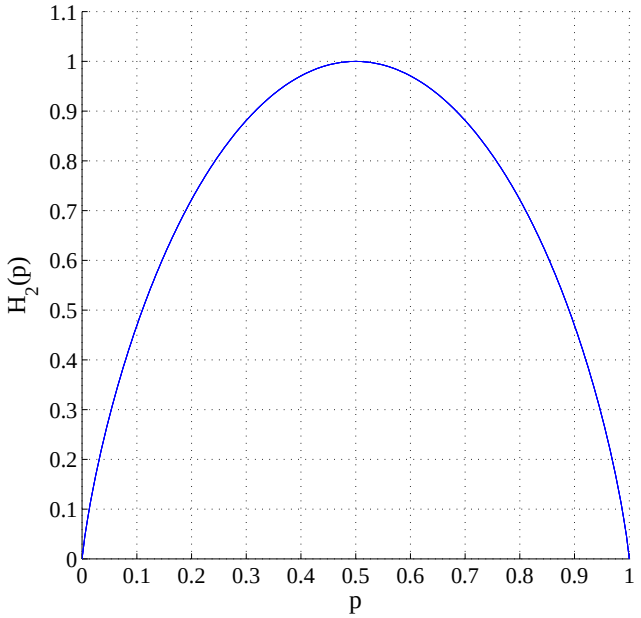

Figure 1.1.: The binary entropy function $H_2(\cdot)$.

the expectation of Y:

$$H(X) = \mathsf{E}[-\log_2 P_X(X)]\,. \tag{1.5}$$

In cases where the distribution P_X is a variable, we may use the notation $H(P_X)$ rather than $H(X)$ (see Sec. 1.10). We have chosen to evaluate the logarithm using the base 2, and we continue to follow this convention for *discrete* random variables. Our entropy units are, therefore, *bits*.

Example 1.2. Consider the Bernoulli distribution that has $\mathcal{X} = \{0, 1\}$ and $P_X(0) = 1 - p$. The entropy of X is

$$H(X) = H_2(p) = -p \log_2 p - (1-p) \log_2(1-p) \tag{1.6}$$

and $H_2(\cdot)$ is called the *binary entropy function*.

The binary entropy function is plotted in Fig. 1.1. Observe that $H_2(0) = H_2(1) = 0$, $H_2(0.11) = H_2(0.89) \approx 1/2$, $H_2(1/2) = 1$, 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 $\mathcal{X}$.

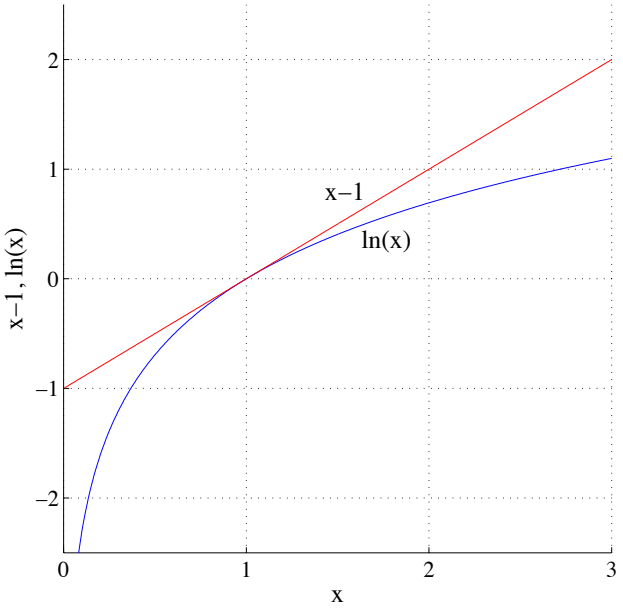

Figure 1.2.: Illustration that $\ln(x) \leq x - 1$ for $x \geq 0$.

Theorem 1.1.

$$0 \leq H(X) \leq \log_2 |\mathcal{X}| \tag{1.7}$$

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 (1.3) 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 (1.5) and observe that

$$H(X) = \mathsf{E}\left[\log_2 \frac{1}{|\mathcal{X}| P_X(X)}\right] + \log_2 |\mathcal{X}|. \tag{1.8}$$

But we have the inequality (see Fig. 1.2)

$$\log_2(x) = \ln(x) \log_2(e) \leq (x - 1) \log_2(e) \tag{1.9}$$

where $\ln(x)$ is the natural logarithm of x, and where equality holds if and only if $x = 1$. Applying (1.9) to (1.8) we have

$$\begin{aligned} H(X) &\overset{(a)}{\leq} \mathsf{E}\left[\frac{1}{|\mathcal{X}|P_X(X)} - 1\right] \log_2(e) + \log_2|\mathcal{X}| \\ &= \sum_{a \in \mathrm{supp}(P_X)} P_X(a) \left(\frac{1}{|\mathcal{X}|P_X(a)} - 1\right) \log_2(e) + \log_2|\mathcal{X}| \\ &= \left(\frac{|\mathrm{supp}(P_X)|}{|\mathcal{X}|} - 1\right) \log_2(e) + \log_2|\mathcal{X}| \\ &\overset{(b)}{\leq} \log_2|\mathcal{X}| \end{aligned} \tag{1.10}$$

with equality in (a) if and only if X is uniform over $\mathcal{X}$, and with equality in (b) if and only if $\mathrm{supp}(P_X) = \mathcal{X}$. □

The two bounds in (1.7) are intuitively pleasing: uncertainty should not be negative (what would negative uncertainty mean?) and the maximum uncertainty is when all possibilities are equally likely.

Example 1.3. Consider a fair die with $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$ and $P_X(a) = 1/6$ for all $a \in \mathcal{X}$. Theorem 1.1 tells us that

$$H(X) = \log_2 6 \approx 2.585 \text{ bits}. \tag{1.11}$$

Furthermore, an unfair six-sided die has smaller entropy or uncertainty.

Example 1.4. Consider $\mathcal{X} = \{0, 1, 2\}$ and $P_X(0) = P_X(1) = p/2$ and $P_X(2) = 1 - p$. We have

$$\begin{aligned} H(X) &= -\frac{p}{2}\log_2\frac{p}{2} - \frac{p}{2}\log_2\frac{p}{2} - (1-p)\log_2(1-p) \\ &= p + H_2(p) \end{aligned} \tag{1.12}$$

and entropy is maximized at $H(X) = \log_2(3)$ if $p = 2/3$.

1.4. Example Distributions

1.4.1. Binomial Distribution

Consider the random variable

$$X = \sum_{i=1}^{n} X_i \tag{1.13}$$

where the X_i, $i = 1, 2, \ldots, n$, are independent binary random variables with the distribution of Example 1.2. The alphabet of X is $\mathcal{X} = \{0, 1, 2, \ldots, n\}$ and $P_X(\cdot)$ is called the *binomial distribution* with

$$P_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k \in \mathcal{X}. \tag{1.14}$$

A plot of the distribution with $p = 0.5$ and $n = 20$ is shown in Fig. 1.3. We compute $\mathsf{E}[X] = n\,p$, $\mathsf{Var}[X] = n\,p\,(1-p)$ and

$$H(X) = \sum_{k=0}^{n} -\binom{n}{k} p^k (1-p)^{n-k} \log_2 \left(\binom{n}{k} p^k (1-p)^{n-k} \right). \tag{1.15}$$

The expression (1.15) seems not to have a simple closed form for general n. We instead use a Gaussian approximation for large n:

$$P_X(k) \approx \frac{1}{\sqrt{2\pi \mathsf{Var}[X]}} e^{-\frac{(k-\mathsf{E}[X])^2}{2\mathsf{Var}[X]}} \tag{1.16}$$

which gives

$$\begin{aligned} H(X) &\approx \mathsf{E}\left[-\log_2 \left(\frac{1}{\sqrt{2\pi \mathsf{Var}[X]}} e^{-\frac{-(X-\mathsf{E}[X])^2}{2\mathsf{Var}[X]}} \right) \right] \\ &= \frac{1}{2} \log_2 \left(2\pi e \cdot np(1-p) \right). \end{aligned} \tag{1.17}$$

1.4.2. Poisson Distribution

The *Poisson distribution* is[2]

$$P_X(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k \in \mathbb{N}_0 = \{0, 1, 2, \ldots\}. \tag{1.18}$$

Note that the range of X is discrete but *infinite*. A plot of the distribution with $\lambda = 5$ is shown in Fig. 1.4. We compute $\mathsf{E}[X] = \lambda$ and $\mathsf{Var}[X] = \lambda$. We define $H(X)$ as in (1.3), and $H(X)$ again seems not to have a simple closed

[2]The notation $\mathbb{N}$ refers to the so-called *natural* numbers $\{1, 2, 3, \ldots\}$ or $\{0, 1, 2, 3, \ldots\}$. To avoid ambiguity, we refer to the former set as $\mathbb{N}_1$ and to the latter as $\mathbb{N}_0$.

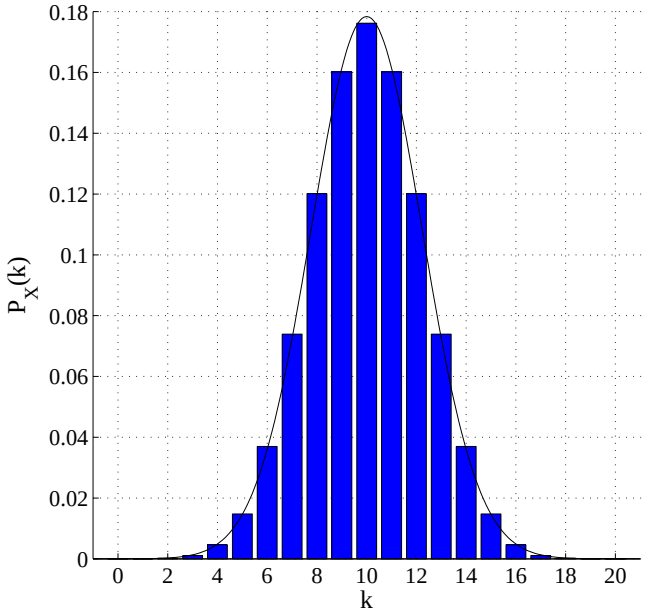

Figure 1.3.: The binomial distribution for $p = 0.5$ and $n = 20$. The Gaussian approximation with $\mathsf{E}[X] = 10$ and $\mathsf{Var}[X] = 5$ is the solid curve.

form for general n. However, if λ is large then $P_X(\cdot)$ is approximately the binomial distribution with variance λ and mean shifted to λ so we have

$$H(X) \approx \frac{1}{2} \log_2 (2\pi e \lambda) . \tag{1.19}$$

1.4.3. Geometric Distribution

The *geometric distribution* arises when counting the number of Bernoulli trials needed for observing a 1. The distribution is

$$P_X(k) = p\,(1-p)^{k-1}, \quad k \in \mathbb{N}_1 = \{1, 2, 3, \ldots\}. \tag{1.20}$$

A plot of the distribution with $p = 0.25$ is shown in Fig. 1.5. We compute (see Problem 1.5) $\mathsf{E}[X] = 1/p$, $\mathsf{Var}[X] = (1-p)/p^2$, and

$$H(X) = H_2(p)\,\mathsf{E}[X] . \tag{1.21}$$

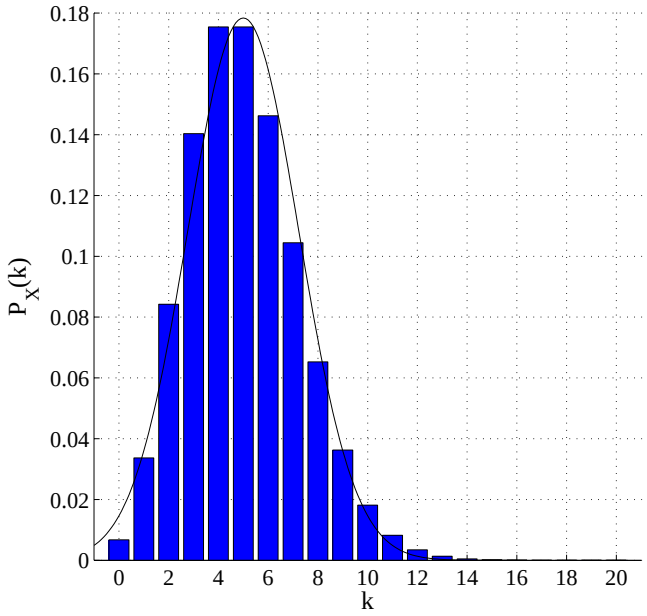

Figure 1.4.: The Poisson distribution for $\lambda = 5$. The Gaussian approximation with $\mathsf{E}[X] = 5$ and $\mathsf{Var}[X] = 5$ is the solid curve.

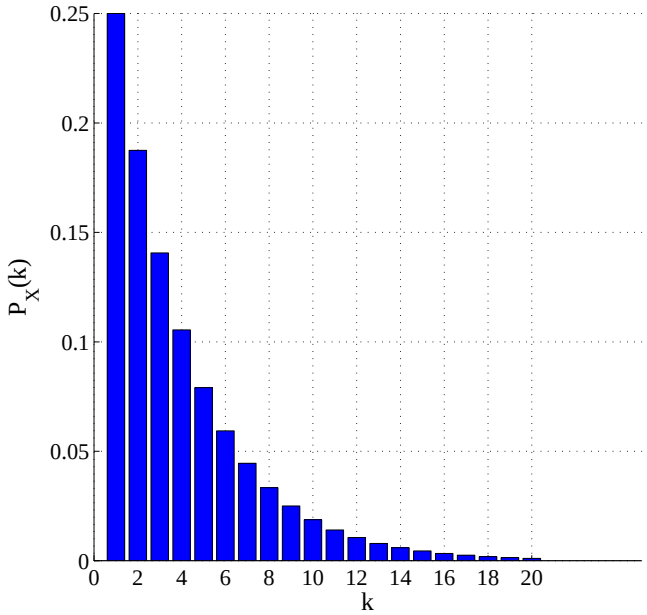

Figure 1.5.: The geometric distribution for $p = 0.25$.

1.5. Conditional Entropy

Consider a joint distribution $P_{XY}(\cdot)$ where the random variables X and Y take on values in the discrete and finite alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively. The *conditional* entropy of X given the event $Y = b$ with probability $\Pr[Y = b] > 0$ is

$$\begin{aligned} H(X|Y=b) &= \sum_{a \in \mathrm{supp}(P_{X|Y}(\cdot|b))} -P_{X|Y}(a|b) \log_2 P_{X|Y}(a|b) \\ &= \mathsf{E}\Big[-\log_2 P_{X|Y}(X|Y)\Big|\, Y=b\Big]. \end{aligned} \tag{1.22}$$

Using the same steps as in the proof of Theorem 1.1, one can show that

$$0 \le H(X|Y=b) \le \log_2 |\mathcal{X}| \tag{1.23}$$

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 (1.22), i.e., we define

$$\boxed{H(X|Y) = \sum_{b \in \mathrm{supp}(P_Y)} P_Y(b) H(X|Y=b)}. \tag{1.24}$$

Alternatively, we have

$$\begin{aligned} H(X|Y) &= \sum_{(a,b) \in \mathrm{supp}(P_{XY})} -P_{XY}(a,b) \log_2 P_{X|Y}(a|b) \\ &= \mathsf{E}\Big[-\log_2 P_{X|Y}(X|Y)\Big]. \end{aligned} \tag{1.25}$$

Example 1.5. Consider the joint distribution

$P_{XY}(a,b)$	$a=0$	$a=1$
$b=0$	$1/3$	$1/3$
$b=1$	$1/3$	0

(1.26)

We compute $H(X|Y=0) = 1$, $H(X|Y=1) = 0$, and $H(X|Y) = 2/3$.

One can show that (try Exercise 1.7)

$$0 \le H(X|Y) \le \log_2 |\mathcal{X}| \tag{1.27}$$

with equality on the left if and only if for every b in $\mathrm{supp}(P_Y)$ 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

$$\begin{aligned} H(X|Y, Z = c) &= \sum_{(a,b)\in \mathrm{supp}(P_{XY|Z}(\cdot|c))} -P_{XY|Z}(a, b|c) \log_2 P_{X|YZ}(a|b, c) \\ &= \mathsf{E}\Big[-\log_2 P_{X|YZ}(X|Y, Z)\Big| Z = c\Big]. \end{aligned} \tag{1.28}$$

1.6. 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

$$\begin{aligned} H(XY) &= \sum_{(a,b)\in \mathrm{supp}(P_{XY})} -P_{XY}(a,b)\log_2 P_{XY}(a,b) \\ &= \mathsf{E}[-\log_2 P_{XY}(X,Y)]\,. \end{aligned} \tag{1.29}$$

Note that we have written $H(XY)$ rather than $H(X,Y)$ and the reader should not confuse XY with "X multiplied by Y". Some authors prefer $H(X,Y)$ and this is a matter of taste. For example, if one considers XY as a vector $[X,Y]$ then the notation $H(X,Y)$ makes sense. We will follow the convention of not using punctuation if no confusion arises.[3]

Using Bayes' rule in (1.29) we have

$$\begin{aligned} H(XY) &= \mathsf{E}\Big[-\log_2\Big(P_X(X)P_{Y|X}(Y|X)\Big)\Big] \\ &= \mathsf{E}[-\log_2 P_X(X)] + \mathsf{E}\Big[-\log_2 P_{Y|X}(Y|X)\Big] \\ &= H(X) + H(Y|X). \end{aligned} \tag{1.30}$$

We similarly have

$$H(XY) = H(Y) + H(X|Y). \tag{1.31}$$

More generally, we have the *chain rule* for entropy

$$\begin{aligned} H(X_1X_2\ldots X_n) &= H(X_1) + H(X_2|X_1) + \ldots + H(X_n|X_1X_2\cdots X_{n-1}) \\ &= \sum_{i=1}^{n} H(X_i|X^{i-1}). \end{aligned} \tag{1.32}$$

where $X^j = X_1X_2\ldots X_j$ (see Sec. A.2) and X^0 is a constant.

Example 1.6. Consider the joint distribution of Example 1.5. We compute

$$\begin{aligned} &H(XY) = \log_2(3) \approx 1.585 \text{ bits} \\ &H(X) = H(Y) = H_2(1/3) \approx 0.918 \text{ bits} \\ &H(X|Y) = H(Y|X) = 2/3 \text{ bits} \end{aligned}$$

and one may check that (1.30) and (1.31) are satisfied.

Theorem 1.1 and (1.27) give

$$\max\left(H(X), H(Y)\right) \le H(XY) \le \log_2(|\mathcal{X}|\cdot|\mathcal{Y}|) \tag{1.33}$$

[3] We are not entirely consistent with this approach, e.g., we write $P_{XY}(a,b)$ without punctuation in the subscript and with punctuation in the argument.

with equality on the left if and only if X essential determines Y, or Y essentially determines X, or both, and with equality on the right if and only if $P_{XY}(a,b) = 1/(|\mathcal{X}|\,|\mathcal{Y}|)$ for all (a,b).

Example 1.7. Using (1.33), (1.30), and that X essentially determines $f(X)$ we have

$$\boxed{H(f(X)) \leq H(Xf(X)) = H(X)}. \tag{1.34}$$

1.7. Informational Divergence

Suppose we wish to measure how close two distributions are to each other. A natural approach is to use the ℓ_1 distance

$$\boxed{d(P_X, P_Y) = \sum_{a \in \mathcal{X}} |P_X(a) - P_Y(a)|} \tag{1.35}$$

where $P_X(\cdot)$ and $P_Y(\cdot)$ have the same domain $\mathcal{X}$. This distance is sometimes called the *variational* distance and it measures *additive* differences.

Another useful approach is the following that considers *logarithmic* differences. The *informational divergence* (or *relative entropy* or *Kullback-Leibler distance*) between $P_X(\cdot)$ and $P_Y(\cdot)$ is defined as

$$\boxed{D(P_X \| P_Y) = \sum_{a \in \mathrm{supp}(P_X)} P_X(a) \log_2 \frac{P_X(a)}{P_Y(a)}} \tag{1.36}$$

which is the same as

$$D(P_X \| P_Y) = \mathsf{E}\left[\log_2 \frac{P_X(X)}{P_Y(X)}\right]. \tag{1.37}$$

We define $D(P_X \| P_Y) = \infty$ if $P_Y(a) = 0$ for some $a \in \mathrm{supp}(P_X)$. Observe that the definition is not symmetric in P_X and P_Y, i.e., we have $D(P_X \| P_Y) \neq D(P_Y \| P_X)$ in general.

Example 1.8. Consider $\mathcal{X} = \{0, 1\}$ and the Bernoulli distributions with $P_X(0) = p$, $P_Y(0) = q$. If $0 < p < 1$ then we have

$$D(P_X \| P_Y) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}. \tag{1.38}$$

If $q = 1/2$ then we compute $D(P_X \| P_Y) = 1 - H_2(p)$. If $q = 0$ or $q = 1$ then we have $D(P_X \| P_Y) = \infty$.

To avoid the case $D(P_X \| P_Y) = \infty$ we introduce the notation $P_X \ll P_Y$ (or $P_Y \gg P_X$) to mean that $P_Y(a) = 0 \Rightarrow P_X(a) = 0$ for all $a \in \mathcal{X}$. In other words, $P_X \ll P_Y$ is the same as saying that $\mathrm{supp}(P_X) \subseteq \mathrm{supp}(P_Y)$, or that $D(P_X \| P_Y) < \infty$ for finite sets. The measure-theoretic terminology is that P_X is *absolutely continuous* with respect to P_Y.

Example 1.9. Consider a joint distribution P_{XY} and its marginals P_X and P_Y. We use the notation $P_X P_Y(\cdot)$ to refer to the distribution for which

$$P_X P_Y(a, b) = P_X(a) P_Y(b) \text{ for all } a \text{ and } b. \tag{1.39}$$

Observe that $P_X(a) = 0$ implies $P_{XY}(a, b) = 0$. However, we may have $P_{XY}(a, b) = 0$ even though $P_X(a) > 0$ and $P_Y(b) > 0$ (see Example 1.5). In

other words, we always have

$$P_{XY} \ll P_X P_Y \tag{1.40}$$

but the statement $P_{XY} \gg P_X P_Y$ is not necessarily true.

The following result is fundamentally important.

Theorem 1.2.

$$D(P_X \| P_Y) \geq 0 \tag{1.41}$$

with equality if and only if $P_X(a) = P_Y(a)$ for all $a \in \operatorname{supp}(P_X)$.[4]

Proof. Apply the inequality (1.9) to the negative of (1.37):

$$\begin{aligned} -D(P_X \| P_Y) &= \mathsf{E}\left[\log_2 \frac{P_Y(X)}{P_X(X)}\right] \\ &\overset{(a)}{\leq} \mathsf{E}\left[\frac{P_Y(X)}{P_X(X)} - 1\right] \log_2(e) \\ &= \sum_{a \in \operatorname{supp}(P_X)} P_X(a) \left[\frac{P_Y(a)}{P_X(a)} - 1\right] \log_2(e) \\ &= \left[\left(\sum_{a \in \operatorname{supp}(P_X)} P_Y(a)\right) - 1\right] \log_2(e) \\ &\overset{(b)}{\leq} 0 \end{aligned} \tag{1.42}$$

with equality in (a) if and only if $P_Y(a) = P_X(a)$ for all $a \in \operatorname{supp}(P_X)$, and with equality in (b) if and only if $P_Y \ll P_X$. □

Example 1.10. Suppose $P_Y(\cdot)$ is uniform on $\mathcal{X}$. We compute

$$D(P_X \| P_Y) = \log_2 |\mathcal{X}| - H(X) \tag{1.43}$$

and so Theorem 1.2 gives $H(X) \leq \log_2 |\mathcal{X}|$ with equality if and only if P_X is uniform. This reproves the interesting part of Theorem 1.1.

Example 1.11. If $P_X(\cdot)$ is uniform on $\mathcal{X}$ then we have

$$D(P_X \| P_Y) = -\log_2 |\mathcal{X}| - \frac{1}{|\mathcal{X}|} \sum_{a \in \mathcal{X}} \log_2 P_Y(a). \tag{1.44}$$

[4] This is the same as requiring $P_X(a) = P_Y(a)$ for all $a \in \mathcal{X}$.

Example 1.12. Suppose $P_X(a) = 1$ for some $a \in \mathcal{X}$. We then have

$$D(P_X \| P_Y) = -\log_2 P_Y(a). \tag{1.45}$$

As in (1.37), given a third discrete random variable Z, we define the *conditional* informational divergence between $P_{X|Z}(\cdot)$ and $P_{Y|Z}(\cdot)$ as

$$\begin{aligned} D(P_{X|Z} \| P_{Y|Z} | P_Z) &= \sum_{b \in \mathrm{supp}(P_Z)} P_Z(b) D\left(P_{X|Z}(\cdot|b) \| P_{Y|Z}(\cdot|b)\right) \\ &= \sum_{(a,b) \in \mathrm{supp}(P_{XZ})} P_{XZ}(a,b) \log_2 \frac{P_{X|Z}(a|b)}{P_{Y|Z}(a|b)} \\ &= \mathsf{E}\left[\log_2 \frac{P_{X|Z}(X|Z)}{P_{Y|Z}(X|Z)}\right]. \end{aligned} \tag{1.46}$$

Similar to (1.41), we have

$$D(P_{X|Z} \| P_{Y|Z} | P_Z) \geq 0 \tag{1.47}$$

with equality if and only if $P_{X|Z}(a|b) = P_{Y|Z}(a|b)$ for all $(a,b) \in \mathrm{supp}(P_{XZ})$.

Problem 1.16 develops a chain rule for informational divergence similar to the chain rule (1.32) for entropy, namely

$$D(P_{X^n} \| P_{Y^n}) = \sum_{i=1}^{n} D(P_{X_i|X^{i-1}} \| P_{Y_i|Y^{i-1}} | P_{X^{i-1}}). \tag{1.48}$$

1.8. 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

$$\boxed{I(X;Y) = D(P_{XY} \| P_X P_Y)}. \tag{1.49}$$

More explicitly, we have

$$I(X;Y) = \sum_{(a,b)\in \mathrm{supp}(P_{XY})} P_{XY}(a,b) \log_2 \frac{P_{XY}(a,b)}{P_X(a) P_Y(b)}. \tag{1.50}$$

Note that $P_{XY} = P_X P_Y$ means that X and Y are statistically independent. Thus, $I(X;Y)$ measures the dependence of X and Y. Recall from Example 1.9 that $P_{XY} \ll P_X P_Y$ so that $I(X;Y)$ is finite for finite-alphabet random variables.

The term "mutual" describes the symmetry in the arguments of $I(X;Y)$. By using Bayes' rule and expanding the logarithm in (1.50) in various ways we may write

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(XY) \\ &= H(XY) - H(X|Y) - H(Y|X) \end{aligned} \tag{1.51}$$

The definition (1.49) and Theorem 1.2 imply the following inequalities.

Theorem 1.3. We have the bounds

$$I(X;Y) \geq 0 \tag{1.52}$$
$$H(X|Y) \leq H(X) \tag{1.53}$$
$$H(XY) \leq H(X) + H(Y) \tag{1.54}$$

with equality in (1.52)-(1.54) if and only if X and Y are statistically independent.

The inequality (1.53) 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 1.13. Consider Example 1.5 for which we compute

$$H(X) = H(Y) = H_2(1/3) \approx 0.918.$$

We thus have $H(X|Y) \leq H(X)$ in accordance with (1.53), but observe that $H(X|Y=0) > H(X)$.

The *conditional* mutual information between X and Y given a random variable Z is defined as

$$I(X;Y|Z) = \sum_{c \in \mathrm{supp}(P_Z)} P_Z(c)\, I(X;Y|Z=c) \tag{1.55}$$

where we define

$$I(X;Y|Z=c) = D\left(P_{XY|Z}(\cdot|c) \,\middle\|\, P_{X|Z}(\cdot|c) P_{Y|Z}(\cdot|c) \right). \tag{1.56}$$

From (1.46) we may write

$$I(X;Y|Z) = D(P_{XY|Z} \| P_{X|Z} P_{Y|Z} | P_Z). \tag{1.57}$$

Using (1.56) and Theorem 1.2, we have

$$I(X;Y|Z=c) \geq 0 \tag{1.58}$$

$$I(X;Y|Z) \geq 0 \tag{1.59}$$

with equality in (1.58) if and only if X and Y are statistically independent conditioned on the event $Z=c$, and with equality in (1.59) if and only if X and Y are statistically independent when conditioned on *any* event $Z=c$ with positive probability. Equality in (1.59) is thus the same as saying that $X - Z - Y$ forms a Markov chain (see Sec. A.3).

We compute

$$\begin{aligned} I(X;Y|Z) &= H(X|Z) - H(X|YZ) \\ &= H(Y|Z) - H(Y|XZ). \end{aligned} \tag{1.60}$$

We thus have

$$\boxed{0 \leq I(X;Y|Z) \leq \min\left(H(X|Z), H(Y|Z)\right)} \tag{1.61}$$

with equality on the left if and only if $X-Z-Y$ forms a Markov chain, and with equality on the right if and only if YZ essentially determines X, or XZ essentially determines Y, or both. The left-hand side of (1.61) implies

$$\boxed{H(X|YZ) \leq H(X|Z)} \tag{1.62}$$

with equality if and only if $X - Z - Y$ forms a Markov chain.

We can expand mutual information as follows:

$$\begin{aligned} I(X_1 X_2 \ldots X_n; Y) &= I(X_1;Y) + I(X_2;Y|X_1) + \ldots \\ &\quad + I(X_n;Y|X_1 X_2 \cdots X_{n-1}) \\ &= \sum_{i=1}^{n} I(X_i;Y|X^{i-1}). \end{aligned} \tag{1.63}$$

The expansion (1.63) is called the *chain rule* for mutual information.

Example 1.14. Using the chain rule for mutual information and the left-hand side of (1.61), we have

$$\begin{aligned} I(X;Y) &\leq I(X;YZ) \\ I(X;Y) &\leq I(XZ;Y). \end{aligned} \tag{1.64}$$

1.9. Inequalities

1.9.1. Log-Sum Identity and Inequality

Theorem 1.4. (Log-sum Identity and Inequality) Consider positive a_i and non-negative b_i, $i = 1, 2, \ldots, n$, and suppose that at least one of the b_i is positive. Let $S_a = \sum_{i=1}^{n} a_i$, $S_b = \sum_{i=1}^{n} b_i$, and define $P_X(i) = a_i/S_a$ and $P_Y(i) = b_i/S_b$ for $i = 1, 2, \ldots, n$. We have

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} = S_a D(P_X \| P_Y) + S_a \log \frac{S_a}{S_b} \tag{1.65}$$

$$\geq S_a \log \frac{S_a}{S_b} \tag{1.66}$$

with equality if and only if $a_i/b_i = S_a/S_b$ for all i.

Proof. If $b_i = 0$ for some i then the left-hand side of (1.65) is infinity and the identity and the inequality are valid. Now suppose that $b_i > 0$ for all i. The identity (1.65) follows by substitution. The inequality (1.66) and the condition for equality follow by Theorem 1.2. □

Example 1.15. Consider $n = 2$ for which the log-sum inequality is

$$a_1 \log \frac{a_1}{b_1} + a_2 \log \frac{a_2}{b_2} \geq (a_1 + a_2) \log \frac{a_1 + a_2}{b_1 + b_2} \tag{1.67}$$

with equality if and only if $a_1/b_1 = a_2/b_2 = (a_1 + a_2)/(b_1 + b_2)$.

1.9.2. Data Processing Inequalities

Theorem 1.5. (Data Processing Inequalities) If $X - Y - Z$ forms a Markov chain, then we have

$$I(X;Z) \leq I(X;Y) \quad \text{and} \quad I(X;Z) \leq I(Y;Z). \tag{1.68}$$

If Y_1 and Y_2 are the outputs of a channel $P_{Y|X}(\cdot)$ with inputs X_1 and X_2, respectively, then we have

$$D(P_{Y_1} \| P_{Y_2}) \leq D(P_{X_1} \| P_{X_2}). \tag{1.69}$$

Proof. We have

$$\begin{aligned} I(X;Z) &\overset{(a)}{\leq} I(X;YZ) \\ &\overset{(b)}{=} I(X;Y) + I(X;Z|Y) \\ &\overset{(c)}{=} I(X;Y). \end{aligned} \tag{1.70}$$

where (a) follows by (1.64), (b) follows by the chain rule (1.63), and (c) follows because $X-Y-Z$ forms a Markov chain. One can prove $I(X;Z) \leq I(Y;Z)$ in the same way. Next, by the chain rule (1.48) we have the two expansions

$$D(P_{X_1Y_1}\|P_{X_2Y_2}) = D(P_{X_1}\|P_{X_2}) + \underbrace{D(P_{Y_1|X_1}\|P_{Y_2|X_2}|P_{X_1})}_{=0} \tag{1.71}$$

$$D(P_{X_1Y_1}\|P_{X_2Y_2}) = D(P_{Y_1}\|P_{Y_2}) + D(P_{X_1|Y_1}\|P_{X_2|Y_2}|P_{Y_1}) \tag{1.72}$$

which gives

$$D(P_{Y_1}\|P_{Y_2}) = D(P_{X_1}\|P_{X_2}) - D(P_{X_1|Y_1}\|P_{X_2|Y_2}|P_{Y_1}). \tag{1.73}$$

and which implies (1.69). □

Example 1.16. Suppose $P_{Y|X}$ has the special property that a P_X uniform on $\mathcal{X}$ produces a P_Y uniform on $\mathcal{Y}$. So choosing P_{X_2} uniform in (1.69) makes P_{Y_2} uniform. Using (1.43) we find that

$$H(X_1) \leq H(Y_1). \tag{1.74}$$

In other words, such a channel $P_{Y|X}$ does not decrease entropy.

1.9.3. Fano's Inequality

A useful lower bound on error probability is the following.

Theorem 1.6. (Fano's Inequality) 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

$$H_2(P_e) + P_e \log_2(|\mathcal{X}| - 1) \geq H(X|\hat{X}) \tag{1.75}$$

with equality if and only if, for all a and b in $\mathcal{X}$, we have

$$P_{X|\hat{X}}(a|b) = \begin{cases} 1 - P_e & \text{if } b = a \\ \frac{P_e}{|\mathcal{X}|-1} & \text{if } b \neq a \end{cases} . \tag{1.76}$$

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

$$\begin{aligned} H(EX|\hat{X}) &= H(X|\hat{X}) + H(E|\hat{X}X) \\ &\overset{(a)}{=} H(X|\hat{X}) \\ H(EX|\hat{X}) &= H(E|\hat{X}) + H(X|\hat{X}E) \\ &= H(E|\hat{X}) + \Pr[E=0]\, H(X|\hat{X}, E=0) \\ &\quad + \Pr[E=1]\, H(X|\hat{X}, E=1) \\ &= H(E|\hat{X}) + P_e\, H(X|\hat{X}, E=1) \\ &\overset{(b)}{\leq} H(E|\hat{X}) + P_e \log_2(|\mathcal{X}| - 1) \\ &\overset{(c)}{\leq} H(E) + P_e \log_2(|\mathcal{X}| - 1) \\ &= H_2(P_e) + P_e \log_2(|\mathcal{X}| - 1) \end{aligned}$$

where (a) follows because $\hat{X}$ and X essentially determine E. Step (b) follows because, given $\hat{X}$ and $E = 1$, X takes on at most $|\mathcal{X}| - 1$ values. Furthermore, by Theorem 1.1 equality holds in (b) if and only if, conditioned on $\hat{X}$, X is uniform over the set of $\mathcal{X}$ not including $\hat{X}$, i.e., if and only if (1.76) is satisfied for all a and b in $\mathcal{X}$. Inequality (c) holds with equality if (1.76) is satisfied for all a and b in $\mathcal{X}$. □

Example 1.17. Consider $\mathcal{X} = \{0, 1\}$ for which Fano's inequality is

$$H_2(P_e) \geq H(X|\hat{X}). \tag{1.77}$$

Equality holds if and only if $X = \hat{X} \oplus Z$ where Z is independent of $\hat{X}$, $P_Z(0) = 1 - P_e$, $P_Z(1) = P_e$, and "$\oplus$" denotes addition modulo-2 (or XOR).

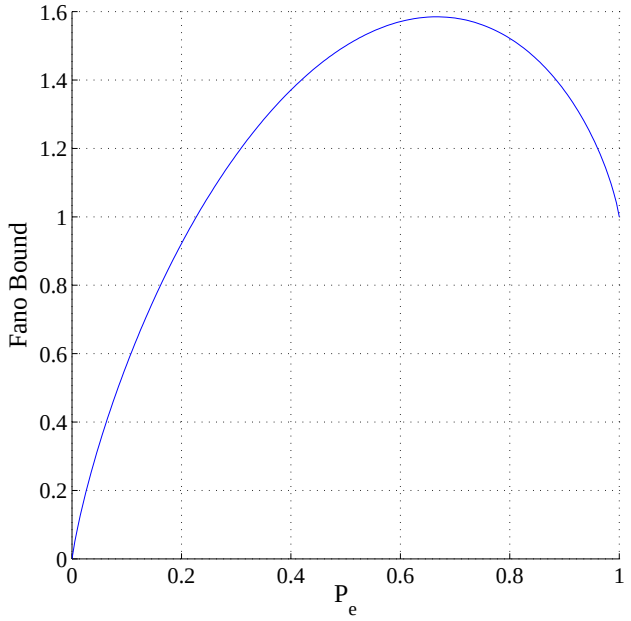

Figure 1.6.: Fano upper bound on $H(X|\hat{X})$ for $|\mathcal{X}| = 3$.

Example 1.18. 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

$$H_2(1 - p_0) + (1 - p_0) \geq H(X|\hat{X}) \tag{1.78}$$

and one can check that equality holds if and only if $p_1 = p_2$ (see (1.12)).

A plot of the left-hand side of (1.75) as a function of P_e is shown in Fig. 1.6 for $|\mathcal{X}| = 3$. We can interpret (1.75) as follows: P_e cannot be driven to zero if $H(X|\hat{X})$ is bounded from below by some positive number.

1.9.4. Pinsker's Inequality

Suppose $P(\cdot)$ and $Q(\cdot)$ are probability distributions with discrete and finite domain $\mathcal{X}$. Recall that the variational distance between $P(\cdot)$ and $Q(\cdot)$ is

$$d(P,Q) = \sum_{a\in\mathcal{X}} |P(a) - Q(a)| \,. \tag{1.79}$$

Oberve that

$$0 \le d(P,Q) \le 2 \tag{1.80}$$

where the right inequality follows by $|P(a) - Q(a)| \le P(a) + Q(a)$. We have equality on the left if and only if $P(a) = Q(a)$ for all a, and we have equality on the right if and only if $\mathrm{supp}(P) \cap \mathrm{supp}(Q) = \emptyset$.

Theorem 1.7. (Pinsker's Inequality)

$$D(P\|Q) \ge \frac{1}{2\ln(2)}\, d^2(P,Q) \tag{1.81}$$

Proof. Partition $\mathcal{X}$ into two sets:

$$\mathcal{X}_1 = \{a : P(a) \ge Q(a)\} \tag{1.82}$$
$$\mathcal{X}_2 = \{a : P(a) < Q(a)\}. \tag{1.83}$$

Define

$$p = \sum_{a\in\mathcal{X}_1} P(a) \quad \text{and} \quad q = \sum_{a\in\mathcal{X}_1} Q(a) \tag{1.84}$$

so that $p \ge q$. We have $d(P,Q) = 2(p-q)$ and the log-sum inequality gives

$$D(P\|Q) \ge p\log_2\frac{p}{q} + (1-p)\log_2\frac{1-p}{1-q}. \tag{1.85}$$

We thus have

$$D(P\|Q) - \frac{1}{2\ln(2)}\, d(P,Q)^2 \ge \left[p\log_2\frac{p}{q} + (1-p)\log_2\frac{1-p}{1-q}\right] - \frac{2}{\ln(2)}\,(p-q)^2. \tag{1.86}$$

If $q = p$ then the right-hand side of (1.86) is zero, so suppose $q < p$. The derivative of (1.86) with respect to q is

$$\frac{p-q}{\ln(2)}\left(\frac{-1}{q(1-q)} + 4\right) \le 0. \tag{1.87}$$

Thus, as q decreases the right-hand side of (1.86) increases from its (minimum) value of 0. □

1.10. 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.

Recall (see Appendix A.12) that a real-valued function $f(\cdot)$ whose domain is a non-empty convex set $\mathcal{S}$ in $\mathbb{R}^n$ is convex on $\mathcal{S}$ if for every point every $\underline{x}_1$ and $\underline{x}_2$ in $\mathcal{S}$ we have

$$\lambda f(\underline{x}_1) + (1-\lambda) f(\underline{x}_2) \geq f\left(\lambda \underline{x}_1 + (1-\lambda)\underline{x}_2\right) \quad \text{for } 0 < \lambda < 1. \tag{1.88}$$

We say that $f(\cdot)$ is *concave* (or convex-$\cap$) on $\mathcal{S}$ if $-f(\cdot)$ is convex on $\mathcal{S}$. A useful tool for convex and concave functions is *Jensen's inequality* (see Theorem A.3).

For the following, it is useful to think of P_X as being a real-valued n-dimensional vector that is in the convex set of vectors with non-negative entries and for which $\sum_a P_X(a) = 1$. Similarly, we view $P_{Y|X}$ as being a $n \times m$ real matrix, or rather a long vector of length $n \cdot m$, that is in the convex set of matrices, or long vectors, with non-negative entries and for which $\sum_b P_{Y|X}(b|a) = 1$ for all a.

Theorem 1.8. (Convexity of Informational Divergence) $D(P_X \| P_Y)$ is convex (or convex-$\cup$) in the pair (P_X, P_Y). That is, for distributions P_X, P_Y, Q_X, Q_Y with the same domain $\mathcal{X}$ we have

$$\begin{aligned} &\lambda D(P_X \| P_Y) + (1-\lambda) D(Q_X \| Q_Y) \\ &\quad \geq D\left(\lambda P_X + (1-\lambda) Q_X \| \lambda P_Y + (1-\lambda) Q_Y\right) \end{aligned} \tag{1.89}$$

for any λ satisfying $0 < \lambda < 1$.

Proof. Consider $P_X(a) > 0$ and $Q_X(a) > 0$. The log-sum inequality gives

$$\begin{aligned} &\lambda P_X(a) \log_2 \frac{\lambda P_X(a)}{\lambda P_Y(a)} + (1-\lambda) Q_X(a) \log_2 \frac{(1-\lambda) Q_X(a)}{(1-\lambda) Q_Y(a)} \\ &\quad \geq \left[\lambda P_X(a) + (1-\lambda) Q_X(a)\right] \log_2 \frac{\lambda P_X(a) + (1-\lambda) Q_X(a)}{\lambda P_Y(a) + (1-\lambda) Q_Y(a)} \end{aligned}$$

where $0 < \lambda < 1$. If $P_X(a) > 0$ and $Q_X(a) = 0$ then we have

$$\lambda P_X(a) \log_2 \frac{\lambda P_X(a)}{\lambda P_Y(a)} \geq \lambda P_X(a) \log_2 \frac{\lambda P_X(a)}{\lambda P_Y(a) + (1-\lambda) Q_Y(a)}.$$

A similar bound results when $P_X(a) = 0$ and $Q_X(a) > 0$. Now sum the appropriate bounds over $\mathrm{supp}(P_X)$ and $\mathrm{supp}(Q_X)$ to obtain (1.89). $\square$

Theorem 1.9. (Concavity of Entropy) We write $H(X)$ as $H(P_X)$. The entropy $H(P_X)$ is concave (or convex-$\cap$) in P_X. That is, for two distributions P_X and Q_X with the same domain $\mathcal{X}$ we have

$$\lambda H(P_X) + (1-\lambda)H(Q_X) \leq H\left(\lambda P_X + (1-\lambda)Q_X\right) \tag{1.90}$$

for any λ satisfying $0 < \lambda < 1$.

Proof. From (1.43) we can write $H(X) = \log_2|\mathcal{X}| - D(P_X\|P_Y)$ where P_Y is uniform. Theorem 1.8 thus gives the desired result by fixing P_Y as the uniform distribution, i.e., use (1.89) with $Q_Y = P_Y$. □

Example 1.19. For $H_2(p)$ we compute

$$\frac{d}{dp}H_2(p) = \log_2\frac{1-p}{p} \tag{1.91}$$

$$\frac{d^2}{dp^2}H_2(p) = \frac{-1}{\ln(2)\,p(1-p)} < 0. \tag{1.92}$$

$H_2(p)$ is therefore concave in p.

Theorem 1.10. (Convexity of Mutual Information) We write $I(X;Y)$ as $I(P_X, P_{Y|X})$. The function $I(P_X, P_{Y|X})$ is concave in P_X if $P_{Y|X}$ is fixed, and $I(P_X, P_{Y|X})$ is convex in $P_{Y|X}$ if P_X is fixed. That is, we have

$$\lambda I(P_X, P_{Y|X}) + (1-\lambda)I(Q_X, P_{Y|X}) \leq I\left(\lambda P_X + (1-\lambda)Q_X, P_{Y|X}\right) \tag{1.93}$$

$$\lambda I(P_X, P_{Y|X}) + (1-\lambda)I(P_X, Q_{Y|X}) \geq I\left(P_X, \lambda P_{Y|X} + (1-\lambda)Q_{Y|X}\right) \tag{1.94}$$

or any λ satisfying $0 < \lambda < 1$.

Proof. Suppose $P_{Y|X}$ is fixed and consider $I(X;Y) = H(Y) - H(Y|X)$. Note that $H(Y)$ is concave in P_Y. But P_Y and $H(Y|X)$ are linear in P_X. Thus, $I(X;Y)$ is concave in P_X.

Suppose next that P_X is fixed and consider $I(X;Y) = D(P_X P_{Y|X}\|P_X P_Y)$. P_Y is linear in $P_{Y|X}$ so that $D(P_X P_{Y|X}\|P_X P_Y)$ is convex in $P_{Y|X}$. □

1.11. Problems

1.1. Expectation

Consider random variables X and Y with

$$P_X(0) = P_X(1) = \frac{1}{2}, \quad P_X(2) = 0$$
$$P_Y(0) = P_Y(1) = P_Y(2) = \frac{1}{3}.$$

a) Determine $\mathsf{E}\Big[\frac{1}{P_X(X)}\Big]$, $\mathsf{E}\Big[\frac{1}{P_X(Y)}\Big]$, $\mathsf{E}\Big[\frac{1}{P_Y(X)}\Big]$, and $\mathsf{E}\Big[\frac{1}{P_Y(Y)}\Big]$.

b) Compute $\mathsf{E}[-\log_2 P_X(X)]$ and $\mathsf{E}[-\log_2 P_Y(X)]$.

c) Compute $\mathsf{E}\Big[\log_2 \frac{P_X(X)}{P_Y(X)}\Big]$ and $\mathsf{E}\Big[\log_2 \frac{P_Y(Y)}{P_X(Y)}\Big]$.

1.2. Maximum Entropy

Prove the inequality on the right-hand side of (1.7) by using Jensen's inequality (A.71) and the concavity of $\log_2(x)$ in x for $x > 0$. What are the conditions for equality?

1.3. Binomial Distribution 1

Show that

$$(a+b)^n = \sum_{i=0}^{n} \binom{n}{i} a^i \, b^{n-i} \tag{1.95}$$

and use this identity to show that the $P_X(k)$, $k = 0, 1, 2, \ldots, n$ specified in (1.14) give a probability distribution.

1.4. Binomial Distribution 2

Consider a binomial probability distribution $P_X(k)$ for $n = 5$ and $p = 1/3$.

a) Plot P_X.

b) Compute $\mathsf{E}[X]$ and $\mathsf{Var}[X]$ and overlay the Gaussian approximation (1.16) on the plot in part (a).

c) Compute $H(X)$ and compare with the approximation (1.17).

1.5. Geometric Distribution

Consider a geometric distribution P_X and verify the equations for $\mathsf{E}[X]$, $\mathsf{Var}[X]$, and $H(X)$ given in Sec. 1.4.3.

1.6. Conditional Entropy

Consider a jar in which there is a fair coin c_1 and an unfair coin c_2. The fair coin takes on the values H and T with probability 1/2 each and the unfair coin always takes on the value H. Suppose we reach into the jar, choose one of the coins, and flip it. Let X be the random variable that represents the outcome H or T.

a) Let C be a random variable that represents the choice of coin, i.e., C takes on the values c_1 and c_2 each with probability 1/2. Determine $H(X|C = c_1)$ and $H(X|C = c_2)$.

b) Compute $H(X|C)$.

c) Determine $P_X(\cdot)$ and compute $H(X)$.

d) Compare the four entropies you computed by ordering them. Can conditioning increase entropy?

1.7. Bounds on Conditional Entropy

Verify (1.27) including the conditions for equality.

1.8. Entropy of a Single Parity Check Code

A single parity check code of length 3 has codewords $X_1X_2X_3$ that take on the values 000, 011, 101, 110 each with probability 1/4.

a) Compute $H(X_i)$ for $i = 1, 2, 3$.

b) Compute $H(X_1X_2)$, $H(X_1X_3)$, and $H(X_2X_3)$.

c) Compute $H(X_1X_2X_3)$.

d) Compute $H(X_1|X_2X_3 = ab)$ for $ab = 00, 01, 01, 11$. Now compute $H(X_1|X_2X_3)$.

1.9. Functions of Variables

Let $f(\cdot)$ and $g(\cdot)$ be functions whose domains are the ranges of $[X, Y]$ and Y, respectively. Show that

$$H(Xf(X,Y)|Yg(Y)) = H(X|Y). \tag{1.96}$$

Interpret the result.

1.10. Calculation of Joint Entropy

Suppose X is binary with alphabet $\mathcal{X} = \{0, 1\}$. Consider the *Binary Symmetric Channel* or BSC for which $\mathcal{Y} = \{0, 1\}$ and

$$P_{Y|X}(b|a) = \begin{cases} 1-p, & \text{if } b = a \\ p, & \text{if } b \neq a. \end{cases} \tag{1.97}$$

a) Show that

$$H(X) = H_2(P_X(0))$$
$$H(Y|X) = H_2(p).$$

b) Defining $q * p = q(1-p) + (1-q)p$, verify that

$$H(Y) = H_2\left(P_X(0) * p\right)$$
$$H(X|Y) = H_2(P_X(0)) + H_2(p) - H_2\left(P_X(0) * p\right).$$

c) Show that $H(Y)$ is maximized by $P_X(0) = 1/2$.

1.11. Binary Channels

Verify the following claims where the inverse function $H_2^{-1}(\cdot)$ is defined with domain the interval $[0, 1]$ and range the interval $[0, 1/2]$.

a) Show that $H_2\left(H_2^{-1}(x)\right) = x$ for $0 \leq x \leq 1$ but $H_2^{-1}\left(H_2(x)\right) \neq x$ can happen (for which x?).

b) Show that $[\lambda a + (1-\lambda)b] * p = \lambda(a * p) + (1-\lambda)(b * p)$, where $q * p = q(1-p) + (1-q)p$.

c) Show that $H_2(a) * p \leq H_2(a * p)$ for $0.11 \leq a \leq 0.5$ and $0 \leq p \leq 1$.

d) Using the above two results and $H_2\left(H_2^{-1}(a)\right) = a$, show that

$$[\lambda a + (1-\lambda)b] * p \leq H_2\left(\lambda H_2^{-1}(a) * p + (1-\lambda) H_2^{-1}(b) * p\right)$$

for the range of a and p of problem (c).

1.12. Mrs. Gerber's Lemma

The following problem makes use of the fact that $H_2\left(p * H_2^{-1}(h)\right)$ is convex in h, $0 \leq h \leq 1$, for p satisfying $0 \leq p \leq 1$ [4, Sec. 2].

a) Consider the BSC of Problem 1.10. Show that for any discrete random variable U for which $U - X - Y$ forms a Markov chain we have

$$H(Y|U) \geq H_2(p * H_2^{-1}(H(X|U)). \tag{1.98}$$

b) Suppose the BSC is used n times with input X_i and output Y_i for $i = 1, 2, \ldots, n$. The $\{X_i\}_{i=1}^n$ could be dependent. Show that

$$H(Y^n) \geq \sum_{i=1}^{n} H(Y_i|X^{i-1}) \tag{1.99}$$

with equality if and only if the $\{X_i\}_{i=1}^n$ are independent or $p = 1/2$.

c) Use (1.98) and (1.99) to show that

$$H(Y^n)/n \geq H_2\left(p * H_2^{-1}(H(X^n)/n)\right) \tag{1.100}$$

with equality if and only if the $\{X_i\}_{i=1}^n$ are independent or $p = 1/2$. This result is known as Mrs. Gerber's Lemma [4].

d) Now show that for any discrete random variable V we have

$$H(Y^n|V)/n \geq H_2\left(p * H_2^{-1}(H(X^n|V)/n)\right). \tag{1.101}$$

1.13. Informational Divergence 1

Prove the inequality (1.41) by using Jensen's inequality (A.71) and the concavity of $\log_2(x)$ in x for $x > 0$. What are the conditions for equality?

1.14. Informational Divergence 2

a) Verify that $D(P_1P_2\|Q_1Q_2) = D(P_1\|Q_1) + D(P_2\|Q_2)$.

b) Verify the "parallelogram identity" for probability distributions P, Q, R:

$$\begin{aligned} D(P\|R) + D(Q\|R) = {} & D\left(P\|\frac{P+Q}{2}\right) + D\left(Q\|\frac{P+Q}{2}\right) \\ & + 2D\left(\frac{P+Q}{2}\|R\right). \end{aligned} \tag{1.102}$$

1.15. Informational Divergence 3

Consider $\mathcal{X} = \{0, 1\}$ and $P_X(0) = P_Y(0)(1+\epsilon)$ where $0 \le \epsilon \le 1/P_Y(0) - 1$.

a) Verify that

$$D(P_X\|P_Y) = P_Y(0)(1+\epsilon)\log_2(1+\epsilon) + [1 - P_Y(0)(1+\epsilon)]\log_2\left(\frac{1-P_Y(0)(1+\epsilon)}{1-P_Y(0)}\right) \quad (1.103)$$

and we have $D(P_X\|P_Y) \ge 0$ with equality if and only if $\epsilon = 0$.

b) Show that $D(P_X\|P_Y)$ in (1.103) is convex in ϵ.

1.16. Chain Rule for Informational Divergence

Verify the chain rule (1.48) for informational divergence.

1.17. Mutual Information

a) Verify the following identities:

$$\begin{aligned} I(X;Y) &= D(P_{X|Y}\|P_X|P_Y) && (1.104)\\ &= D(P_{Y|X}\|P_Y|P_X) && (1.105)\\ &= D(P_{Y|X}\|Q_Y|P_X) - D(P_Y\|Q_Y) && (1.106)\\ &= \mathsf{E}\left[\log\frac{Q_{Y|X}(Y|X)}{Q_Y(Y)}\right] + D(P_{XY}\|P_Y R_{X|Y}) && (1.107) \end{aligned}$$

where Q_Y is *any* distribution on $\mathcal{Y}$, $Q_{Y|X}$ is an *auxiliary* channel with $Q_{Y|X} = Q_Y/P_X$, $R_{X|Y} = P_X Q_{Y|X}/Q_Y$ is a *reverse* channel, and the expectation in (1.107) is over P_{XY}.

b) What happens if $Q_{Y|X} = P_{Y|X}$? What happens if $Q_{Y|X} \ne P_{Y|X}$?

c) Argue that if $P_{Y|X}$ can be simulated, but not computed, then one can compute a lower bound on $I(X;Y)$ as follows. Generate a long string of independent and identically distributed (i.i.d.) inputs x^n by using P_X, simulate the outputs y^n by passing x^n through the channel $P_{Y|X}$, choose a $Q_{Y|X}$ to compute $Q_{Y|X}(y_i|x_i)$ for $i = 1, 2, \ldots, n$, and then estimate the expectation in (1.107) by averaging.

d) Similarly, argue that one can compute an upper bound on $I(X;Y)$ by using (1.106).

1.18. Bounds on Conditional Mutual Information

Verify (1.61) including the conditions for equality.

1.19. Chain Rule for Mutual Information 1

Show that

$$I(X^n;Y|Z) = \sum_{i=1}^{n} I(X_i;Y|ZX^{i-1}). \quad (1.108)$$

1.20. Chain Rule for Mutual Information 2

Define $X_{i+1}^n = X_{i+1}X_{i+2}\ldots X_n$ and consider X^0 and X_{n+1}^n to be constants.

a) Establish the *telescoping identity*

$$\sum_{i=1}^{n} I(X^i; Y_{i+1}^n) = \sum_{i=1}^{n} I(X^{i-1}; Y_i^n). \tag{1.109}$$

b) Establish the *Csiszár sum identity*

$$\sum_{i=1}^{n} I(X_i; Y_{i+1}^n | X^{i-1}) = \sum_{i=1}^{n} I(X^{i-1}; Y_i | Y_{i+1}^n). \tag{1.110}$$

Hint: Use (1.109) and use the chain rule (1.63) to show that

$$I(X^i; Y_{i+1}^n) = I(X^{i-1}; Y_{i+1}^n) + I(X_i; Y_{i+1}^n | X^{i-1})$$
$$I(X^{i-1}; Y_i^n) = I(X^{i-1}; Y_{i+1}^n) + I(X^{i-1}; Y_i | Y_{i+1}^n).$$

1.21. Functional Dependence and Mutual Information

Let $f(\cdot)$, $g(\cdot)$, and $h(\cdot)$ be functions whose domains are the ranges of $[X, Z]$, $[Y, Z]$, and Z, respectively. Show that

$$I(X; Y|Z) = I(Xf(X, Z); Yg(Y, Z)|Zh(Z)). \tag{1.111}$$

1.22. Conditional Convexity

This exercise develops conditional versions of Theorems 1.8-1.10.

a) Verify that the conditional informational divergence $D(P_{X|Z}\|P_{Y|Z}|P_Z)$ is convex in the pair $(P_{X|Z}, P_{Y|Z})$.
Is $D(P_{X|Z}\|P_{Y|Z}|P_Z)$ convex in (P_{XZ}, P_{YZ})?
Is $D(P_{X|Z}\|P_{Y|Z}|P_Z)$ convex in P_{XYZ}?

b) Verify that $H(X|Z)$ is concave in $P_{X|Z}$. Is $H(X|Z)$ concave in P_{XZ}?

c) Verify that $I(X; Y|Z)$ is concave in $P_{X|Z}$ for fixed $P_{Y|XZ}$, and is convex in $P_{Y|XZ}$ for fixed $P_{X|Z}$.

1.23. Data Processing Inequality

Prove (1.69) by using the log-sum inequality (see Theorem 1.4).

References

[1] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 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.

[2] R. V. L. Hartley. Transmission of information. *Bell Syst. Tech. J.*, page 535, July 1928.

[3] R. Clausius. Ueber verschiedene für die Anwendung bequeme Formen der Hauptgleichungen der mechanischen Wärmetheorie. *Annalen der Physik und Chemie*, CXXV(7):353–400, 1865.

[4] A. D. Wyner and J. Ziv. A theorem on the entropy of certain binary sequences and applications: Part i. *IEEE Trans. Inf. Theory*, 19(6):769–772, November 1973.

Chapter 2.

Information Theory for Continuous Variables

2.1. Differential Entropy

Consider a real-valued random variable X with density $p_X(\cdot)$. The *differential entropy* of X is defined in a similar manner as the entropy of a discrete random variable:

$$\boxed{h(X) = \int_{\mathrm{supp}(p_X)} -p_X(a) \log p_X(a)\, da} \tag{2.1}$$

assuming this integral exists. We can alternatively write

$$h(X) = \mathsf{E}[-\log p_X(X)]\,. \tag{2.2}$$

Example 2.1. Consider the *uniform* distribution with $p_X(a) = 1/A$ for $a \in [0, A)$ where $[0, A) = \{x : 0 \leq x < A\}$ (see Fig. 2.1). We compute

$$\boxed{h(X) = \log(A).} \tag{2.3}$$

We find that $h(X)$ is *negative* for $A < 1$. A natural next question is how we should interpret $h(X)$: what does "negative uncertainty" mean? We shall give one interpretation in Sec 2.5.1.

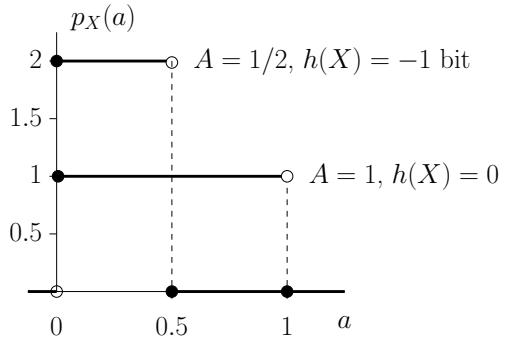

Figure 2.1.: Densities of uniform distributions.

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

$$h(X^n) = \int_{\mathrm{supp}(p_{X^n})} -p_{X^n}(\underline{a}) \log p_{X^n}(\underline{a}) \, d\underline{a}. \tag{2.4}$$

We can alternatively write (2.4) as $h(\underline{X})$ where $\underline{X} = [X_1, X_2, \ldots, X_n]$.

Simple exercises show that for a non-zero real number c we have

$$\begin{array}{ll} \text{Translation rule:} & h(X + c) = h(X) \\ \text{Scaling rule:} & h(cX) = h(X) + \log |c|. \end{array} \tag{2.5}$$

Similarly, for a real-valued column vector $\underline{c}$ of dimension n and an invertible $n \times n$ matrix $\mathbf{C}$ we have

$$\begin{array}{ll} \text{Translation rule:} & h(\underline{X} + \underline{c}) = h(\underline{X}) \\ \text{Scaling rule:} & h(\mathbf{C}\underline{X}) = h(X) + \log |\det \mathbf{C}| \end{array} \tag{2.6}$$

where $\det \mathbf{C}$ is the determinant of $\mathbf{C}$. We will, however, use the notation $|\mathbf{C}|$ for the determinant of $\mathbf{C}$ below.

Consider a joint density $p_{XY}(\cdot)$ and its conditional density $p_{Y|X}(\cdot) = p_{XY}(\cdot)/p_X(\cdot)$. We define

$$h(Y|X = a) = \int_{\mathrm{supp}(p_{Y|X}(\cdot|a))} -p_{Y|X}(b|a) \log p_{Y|X}(b|a) \, db \tag{2.7}$$

and if X has a density then

$$\boxed{h(Y|X) = \int_{\mathrm{supp}(p_X)} p_X(a) \, h(Y|X = a) \, da}. \tag{2.8}$$

We thus have $h(Y|X) = h(XY) - h(X)$. If X is discrete, then we define

$$h(Y|X) = \sum_{a \in \mathrm{supp}(P_X)} P_X(a) \, h(Y|X = a). \tag{2.9}$$

Note that by conditioning on $X = a$ and using the translation rule in (2.5), for any real constant c we obtain

$$h(Y + cX|X) = h(Y|X). \tag{2.10}$$

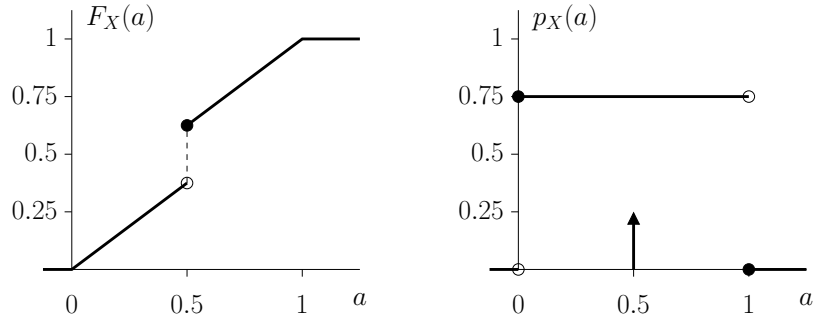

Figure 2.2.: A cumulative distribution and its density.

2.2. Mixed Distributions

One often encounters problems with random variables having both discrete and continuous components. For example, consider a real-valued random variable X with cumulative distribution (see Fig. 2.2)

$$F_X(a) = \begin{cases} 0, & a < 0 \\ 3a/4, & 0 \le a < 1/2 \\ (3a+1)/4, & 1/2 \le a < 1 \\ 1, & a \ge 1. \end{cases} \tag{2.11}$$

Observe that $F_X(\cdot)$ has a discontinuity that represents a probability mass. A common way of writing the "density" for such a mixed distribution is as a sum of a discrete part and a continuous part, in this case (see Fig. 2.2)

$$p_X(a) = \frac{3}{4} p_U(a) + \frac{1}{4}\delta(a - 1/2) \tag{2.12}$$

where $p_U(\cdot)$ is the uniform density in the interval $[0, 1)$ and where $\delta(\cdot)$ is the "Dirac-δ" (generalized) function defined indirectly by taking integrals over proper[1] intervals I:

$$\int_I f(x)\delta(x)dx = \begin{cases} f(0), & 0 \in I \\ 0, & \text{otherwise.} \end{cases} \tag{2.13}$$

More generally, if there are probability mass points at a_i, $i = 1, 2, \ldots, J$, with probabilities $P_X(a_i)$, then we may write the "density" as

$$p_X(x) = p_{\tilde{X}}(x) + \sum_{i=1}^{J} P_X(a_i)\delta(x - a_i) \tag{2.14}$$

where $p_{\tilde{X}}(\cdot)$ is a normalized density of a continuous random variable $\tilde{X}$.

We will see below that $h(X) = -\infty$ for a mixed random variable with a non-zero probability mass. Hence one should exercise caution when dealing with such random variables and using differential entropy.

[1] An interval is *proper* if it is non-empty and does not have just one point.

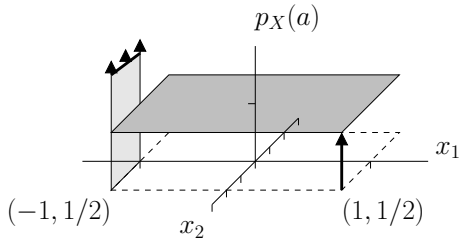

Figure 2.3.: A probability density function for a two-dimensional random vector.

One can extend the above ideas to random vectors. For instance, consider $\underline{X} = [X_1, X_2]$ that takes on values in $\mathbb{R}^2$, and suppose $\underline{X}$ has a

- 0-dimensional part: probability mass $1/4$ at the point $(1, 1/2)$;
- 1-dimensional part: probability mass $1/2$ on the interval defined by $x_1 = -1$ and $x_2 \in [0, 1/2)$;
- 2-dimensional part: uniform probability density on the rectangle $(x_1, x_2) \in [-1, 1) \times [-1/2, 1/2)$ (so the 2-dimensional probability mass is $1/4$).

A plot of the "density" is shown in Fig. 2.3. We have $h(\underline{X}) = -\infty$ due to the 0-dimensional and 1-dimensional parts. However, if we condition on the event $\mathcal{E} = \{X_1 = -1, X_2 \in [0, 1/2)\}$ then we compute $h(\underline{X}|\mathcal{E}) = -1$ bit because $\underline{X}$ is uniform given $\mathcal{E}$.

2.3. Example Distributions

2.3.1. Discrete Random Variables

Consider again the *uniform* density with $p_X(a) = 1/A$ for $a \in [0, A)$ and let $A \to 0$. We thus have $h(X) \to -\infty$. We can interpret such limiting densities as Dirac-δ functions representing discrete random variables. For instance, suppose that $p_X(a) = p_i/A$ for some integers i, $a \in [i, i + A)$, and $0 \leq A \leq 1$. As $A \to 0$, this density represents a discrete random variable $\tilde{X}$ with $P_{\tilde{X}}(i) = p_i$. We compute

$$h(X) = \sum_i -p_i \log\left(p_i/A\right) = \log(A) + H\left(\tilde{X}\right) \tag{2.15}$$

so $h(X)$ has increased as compared to (2.3). However, $h(X)$ still approaches $-\infty$ for small A.

In general, one must exercise caution when dealing with $h(X)$ where X has discrete components. For example, we have $h(Xf(X)) = h(X)+h(f(X)|X)$ and $h(f(X)|X) = -\infty$.

2.3.2. Exponential Distribution

The exponential density is

$$p_X(a) = \begin{cases} \frac{1}{m} e^{-a/m} & a \geq 0 \\ 0 & a < 0 \end{cases} \tag{2.16}$$

where $m = \mathsf{E}[X]$. We compute $\mathsf{Var}[X] = m^2$ and

$$\boxed{h(X) = \log(e\, m)}. \tag{2.17}$$

We find that $h(X) < 0$ if $m < 1/e$ and $h(X) \to -\infty$ as $m \to 0$.

2.3.3. Gaussian Distribution

The scalar Gaussian density is

$$p_X(a) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2\sigma^2}(a-m)^2} \tag{2.18}$$

where $m = \mathsf{E}[X]$ and $\sigma^2 = \mathsf{Var}[X]$ is the variance of X. Inserting (2.18) into (2.1), we compute

$$\boxed{h(X) = \frac{1}{2} \log\left(2\pi e \sigma^2\right).} \tag{2.19}$$

We find that $h(X) < 0$ if $\sigma^2 < 1/(2\pi e)$. We further have $h(X) \to -\infty$ as $\sigma^2 \to 0$.

More generally, consider a random column vector $\underline{X}$ of dimension n, mean $\mathsf{E}[\underline{X}] = \underline{m}$ and covariance matrix

$$\mathbf{Q}_{\underline{X}} = \mathsf{E}\Big[(\underline{X} - \underline{m})(\underline{X} - \underline{m})^T\Big] \tag{2.20}$$

where the superscript "T" denotes transposition. Suppose $\underline{X}$ is Gaussian distributed, i.e., the density of $\underline{X}$ is

$$p_{\underline{X}}(\underline{a}) = \frac{1}{(2\pi)^{n/2}\left|\mathbf{Q}_{\underline{X}}\right|^{1/2}} \exp\left(-\frac{1}{2}(\underline{a} - \underline{m})^T \mathbf{Q}_{\underline{X}}^{-1}(\underline{a} - \underline{m})\right) \tag{2.21}$$

where we recall that $\left|\mathbf{Q}_{\underline{X}}\right|$ is the determinant of $\mathbf{Q}_{\underline{X}}$. Inserting (2.21) into (2.1), we compute (see Problem 2.2)

$$\boxed{h(\underline{X}) = \frac{1}{2}\log\left((2\pi e)^n \left|\mathbf{Q}_{\underline{X}}\right|\right)}. \tag{2.22}$$

Note that $h(\underline{X})$ is negative for small $\left|\mathbf{Q}_{\underline{X}}\right|$.

Example 2.2. Consider (2.21) with $\underline{X} = [X\, Y]^T$ and

$$\mathbf{Q}_{XY} = \begin{bmatrix} \sigma_X^2 & \rho\,\sigma_X\sigma_Y \\ \rho\,\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix} \tag{2.23}$$

so that $\rho = \mathsf{E}[XY]/(\sigma_X\sigma_Y)$ is the correlation coefficient of X and Y. The differential entropy (2.22) is

$$h(XY) = \frac{1}{2}\log\left((2\pi e)^2\,\sigma_X^2\sigma_Y^2(1-\rho^2)\right). \tag{2.24}$$

X and Y are uncorrelated if $\rho = 0$, in which case X and Y are independent and $h(XY) = h(X) + h(Y)$. On the other hand, if $X = a \cdot Y + b$ for some constants a and b then we have $\rho = 1$ and $h(XY) = -\infty$. We must therefore be cautious with Gaussian vectors: if one entry is a function of the other entries then we have $|\mathbf{Q}_{\underline{X}}| = 0$ and $h(\underline{X}) = -\infty$.

Finally, suppose $p_{\underline{X}\,\underline{Y}}(\cdot)$ is Gaussian, where $\underline{X}$ has dimension n and $\underline{Y}$ has dimension m. Let $\mathbf{Q}_{\underline{X}\,\underline{Y}}$ be the covariance matrix of the stacked vector $[\underline{X}^T\,\underline{Y}^T]^T$. We compute

$$h(\underline{Y}|\underline{X}) = h(\underline{X}\,\underline{Y}) - h(\underline{X}) = \frac{1}{2}\log\left((2\pi e)^m \left|\mathbf{Q}_{\underline{X}\,\underline{Y}}\right| \Big/ \left|\mathbf{Q}_{\underline{X}}\right|\right). \tag{2.25}$$

Example 2.3. Consider (2.23) for which we compute

$$h(Y|X) = \frac{1}{2}\log\left(2\pi e\,\sigma_Y^2(1-\rho^2)\right). \tag{2.26}$$

2.4. Informational Divergence

The informational divergence between the densities p_X and p_Y is

$$\boxed{D(p_X\|p_Y) = \int_{\mathrm{supp}(p_X)} p_X(a) \log \frac{p_X(a)}{p_Y(a)}\, da} \tag{2.27}$$

assuming this integral exists. If p_X and p_Y are mixed distributions of the form (2.14), then we *partition* the events that can occur into events with probability mass on points (0-dimensional parts), curves (1-dimensional parts), and so forth. The informational divergence is defined by summing informational divergences of the parts. For example, if p_X has 0-dimensional and 1-dimensional parts then we write

$$\begin{aligned} D(p_X\|p_Y) = &\left[\sum_{a\in\mathrm{supp}(P_X)} P_X(a) \log \frac{P_X(a)}{P_Y(a)}\right] \\ &+ \int_{\mathrm{supp}(p_{\tilde{X}})} p_{\tilde{X}}(a) \log \frac{p_{\tilde{X}}(a)}{p_{\tilde{Y}}(a)}\, da. \end{aligned} \tag{2.28}$$

A similar generalization is possible for vectors $\underline{X}$ and $\underline{Y}$.

The mutual information between continuous random variables X and Y is

$$\boxed{I(X;Y) = D(p_{XY}\|p_X p_Y)}. \tag{2.29}$$

We can derive natural bounds on informational divergence for continuous random variables. For instance, the bound $\ln(x) \le x - 1$ implies

$$\boxed{D(p_{\underline{X}}\|p_{\underline{Y}}) \ge 0} \tag{2.30}$$

with equality if and only if $p_X(a) = p_Y(a)$ for (almost) all $a \in \mathrm{supp}(p_X)$. This means that

$$I(X;Y) \ge 0 \tag{2.31}$$
$$h(X|Y) \le h(X) \tag{2.32}$$
$$h(XY) \le h(X) + h(Y) \tag{2.33}$$

with equality if and only if X and Y are independent.

The conditional informational divergence between $p_{X|Z}(\cdot)$ and $p_{Y|Z}(\cdot)$ is defined as (see (1.46))

$$\begin{aligned} D(p_{X|Z}\|p_{Y|Z}|p_Z) &= \int_{\mathrm{supp}(p_Z)} p_Z(b) D(p_{X|Z}(\cdot|b)\|p_{Y|Z}(\cdot|b))\, db \\ &= \int_{\mathrm{supp}(p_{XZ})} p_{XZ}(a,b) \log \frac{p_{X|Z}(a|b)}{p_{Y|Z}(a|b)}\, db \\ &= \mathsf{E}\left[\log \frac{p_{X|Z}(X|Z)}{p_{Y|Z}(X|Z)}\right] \end{aligned} \tag{2.34}$$

assuming the integrals exist. We have $D(p_{X|Z}\|p_{Y|Z}|p_Z) \ge 0$ with equality

if and only if $p_{X|Z}(a|b) = p_{Y|Z}(a|b)$ for (almost) all $(a, b) \in \text{supp}(p_{XZ})$.

2.5. Maximum Entropy

2.5.1. Alphabet or Volume 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 $\mathcal{S}$, i.e., $|\mathcal{S}| = \int_{\mathrm{supp}(p_{\underline{X}})} 1 \, dx$, and let $\underline{U}$ be uniform over $\mathcal{S}$. We use (2.30) and compute

$$\begin{aligned} 0 \leq D(p_{\underline{X}} \| p_{\underline{U}}) &= \int_{\mathrm{supp}(p_{\underline{X}})} p_{\underline{X}}(\underline{a}) \log \left(p_{\underline{X}}(\underline{a}) |\mathcal{S}| \right) \, d\underline{a} \\ &= -h(\underline{X}) + \log |\mathcal{S}|. \end{aligned} \tag{2.35}$$

We thus find that if $\underline{X}$ is limited to $\mathcal{S}$ then

$$\boxed{h(\underline{X}) \leq \log |\mathcal{S}|} \tag{2.36}$$

with equality if and only if $p_{\underline{X}}(\underline{a}) = 1/|\mathcal{S}|$ for (almost) all $\underline{a} \in \mathcal{S}$.

Alternatively, we have $2^{h(\underline{X})} \leq |\mathcal{S}|$. This bound justifies having negative differential entropy, namely that $2^{h(\underline{X})}$ for a uniform $\underline{X}$ measures the volume of the support set $\mathcal{S}$.

2.5.2. First Moment Constraint

For continuous random variables, one is often interested in *moment* constraints rather than alphabet 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 (2.37)

$$\mathsf{E}[\underline{X}] \leq \underline{m} \tag{2.37}$$

where the inequality $\underline{a} \leq \underline{b}$ means that $a_i \leq b_i$ for all entries a_i and b_i of the respective $\underline{a}$ and $\underline{b}$.

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

$$p_{E_i}(a) = \begin{cases} \frac{1}{m_i} e^{-a/m_i} & a \geq 0 \\ 0 & a < 0. \end{cases} \tag{2.38}$$

We use the same approach as in (2.35) to compute

$$\begin{aligned} 0 \le D(p_{\underline{X}} \| p_{\underline{E}}) &= \int_{\mathrm{supp}(p_{\underline{X}})} p_{\underline{X}}(\underline{a}) \log \frac{p_{\underline{X}}(\underline{a})}{p_{\underline{E}}(\underline{a})} \, d\underline{a} \\ &= -h(\underline{X}) - \int_{\mathrm{supp}(p_{\underline{X}})} p_{\underline{X}}(\underline{a}) \log p_{\underline{E}}(\underline{a}) \, d\underline{a} \\ &= -h(\underline{X}) - \int_{\mathrm{supp}(p_{\underline{X}})} p_{\underline{X}}(\underline{a}) \sum_i \left[-\frac{a_i}{m_i} \log e - \log m_i \right] \, d\underline{a} \\ &= -h(\underline{X}) + \sum_i \log(e\, m_i) \end{aligned} \tag{2.39}$$

We thus have

$$\boxed{h(\underline{X}) \le \sum_i \log(e\, m_i)} \tag{2.40}$$

with equality if and only if $p_{\underline{X}}(\underline{x}) = p_{\underline{E}}(\underline{x})$ for almost all $\underline{x}$. Independent exponential random variables therefore maximize (differential) entropy under first moment and non-negativity constraints.

2.5.3. Second Moment Constraint

Suppose we wish to maximize $h(\underline{X})$ under the second-moment constraint

$$|\mathbf{Q}_{\underline{X}}| \le D \tag{2.41}$$

where D is some constant. For example, the constraint (2.41) occurs if we are restricting attention to $\underline{X}$ that satisfy

$$\mathbf{Q}_{\underline{X}} \preceq \mathbf{Q} \tag{2.42}$$

for some positive semidefinite $\mathbf{Q}$ with $|\mathbf{Q}| = D$, where $\mathbf{A} \preceq \mathbf{B}$ means that $\mathbf{B} - \mathbf{A}$ is positive semi-definite (and hence $|\mathbf{A}| \le |\mathbf{B}|$; see [1, p. 471]).

Let $\underline{G}$ be Gaussian with the same mean $\underline{m}$ and covariance matrix $\mathbf{Q}_{\underline{X}}$ as $\underline{X}$. We repeat the approach of (2.35) and (2.39) and compute

$$\begin{aligned} 0 \le D(p_{\underline{X}} \| p_{\underline{G}}) &= \int_{\mathrm{supp}(p_{\underline{X}})} p_{\underline{X}}(\underline{a}) \log \frac{p_{\underline{X}}(\underline{a})}{p_{\underline{G}}(\underline{a})} \, d\underline{a} \\ &= -h(\underline{X}) - \int_{\mathrm{supp}(p_{\underline{X}})} p_{\underline{X}}(\underline{a}) \log p_{\underline{G}}(\underline{a}) \, d\underline{a} \\ &= -h(\underline{X}) - \int_{\mathrm{supp}(p_{\underline{X}})} p_{\underline{X}}(\underline{a}) \left[-\frac{1}{2} \log \left((2\pi)^n |\mathbf{Q}_{\underline{X}}| \right) \right. \\ &\qquad \left. -\frac{1}{2} (\underline{a} - \underline{m})^T \mathbf{Q}_{\underline{X}}^{-1} (\underline{a} - \underline{m}) \log e \right] \, d\underline{a} \\ &= -h(\underline{X}) + \frac{1}{2} \log \left((2\pi e)^n |\mathbf{Q}_{\underline{X}}| \right). \end{aligned} \tag{2.43}$$

We thus have

$$\boxed{h(\underline{X}) \leq \frac{1}{2} \log \left((2\pi e)^n |\mathbf{Q}_{\underline{X}}| \right)} \tag{2.44}$$

with equality if and only if $p_{\underline{X}}(\underline{x}) = p_{\underline{G}}(\underline{x})$ for almost all $\underline{x}$. Gaussian random variables thus maximize (differential) entropy under the second moment constraint (2.41).

Example 2.4. The constraint (2.41) for scalars ($n = 1$) is $\mathsf{Var}[X] \leq D$ and the bound (2.44) implies

$$h(X) \leq \frac{1}{2} \log (2\pi e\, D) \,. \tag{2.45}$$

A Gaussian random variable X with variance D (and any mean) thus maximizes differential entropy under the variance constraint.

Finally, we prove a conditional version of the maximum entropy theorem. Consider a density $p_{\underline{X}\,\underline{Y}}(\cdot)$ that has the conditional density $p_{\underline{Y}|\underline{X}}(\cdot)$ and covariance matrix $\mathbf{Q}_{\underline{X}\,\underline{Y}}$. Suppose that $(\underline{G}, \underline{H})$ is Gaussian with the same covariance matrix. We compute

$$\begin{aligned} 0 &\leq D\left(p_{\underline{Y}|\underline{X}} \middle\| p_{\underline{H}|\underline{G}} \middle| p_{\underline{X}}\right) \\ &= -h(\underline{Y}|\underline{X}) - \int_{\mathrm{supp}(p_{\underline{X}\,\underline{Y}})} p_{\underline{X}\,\underline{Y}}(\underline{a}, \underline{b}) \log p_{\underline{H}|\underline{G}}(\underline{b}|\underline{a}) \; d\underline{a}\, d\underline{b}. \\ &= -h(\underline{Y}|\underline{X}) + \frac{1}{2} \log \left((2\pi e)^m \left|\mathbf{Q}_{\underline{X}\,\underline{Y}}\right| \Big/ \left|\mathbf{Q}_{\underline{X}}\right| \right). \end{aligned} \tag{2.46}$$

We thus have

$$\boxed{h(\underline{Y}|\underline{X}) \leq \frac{1}{2} \log \left((2\pi e)^m \frac{\left|\mathbf{Q}_{\underline{X}\,\underline{Y}}\right|}{|\mathbf{Q}_{\underline{X}}|} \right)} \tag{2.47}$$

with equality if and only if $p_{\underline{Y}|\underline{X}}(\underline{x}) = p_{\underline{H}|\underline{G}}(\underline{x})$ for almost all $\underline{x}$.

2.6. Problems

2.1. Translation and Scaling

Verify equations (2.5) and (2.6).

2.2. Entropy for Gaussian Random Vectors

Verify equation (2.22).
Hint: use $\mathrm{tr}\,(AB) = \mathrm{tr}\,(BA)$ where $\mathrm{tr}\,(A)$ is the trace of matrix A.

2.3. Informational Divergence

a) Verify (2.30) and show that this bound holds with equality if and only if $p_X(a) = p_Y(a)$ for (almost) all $a \in \mathrm{supp}(p_X)$.

b) Verify that $D(p_{X|Z}\|p_{Y|Z}|p_Z) \geq 0$ with equality if and only if $p_{X|Z}(a|b) = p_{Y|Z}(a|b)$ for (almost) all $(a,b) \in \mathrm{supp}(p_{XZ})$.

2.4. Conditional Entropy-Power Inequality

The *entropy power* of a real random variable X with differential entropy $h(X)$ is defined as $e^{2h(X)}$. The entropy power inequality states that the sum $X+Y$ of two independent random variables X and Y having differential entropies satisfies

$$2^{2h(X+Y)} \geq 2^{2h(X)} + 2^{2h(Y)} \tag{2.48}$$

Prove that $f(x) = \log(e^x + c)$, where c is a non-negative constant, is convex-$\cup$ in x. Now use the entropy power inequality to prove that

$$2^{2h(X+Y|U)} \geq 2^{2h(X|U)} + 2^{2h(Y|U)} \tag{2.49}$$

if U and Y are independent (assume that U has a density).

2.5. Polar Coordinates

Consider a complex random variable $X = X_R + jX_I$ where $j = \sqrt{-1}$ and X_R and X_I are real random variables. We can view X as a vector $[X_R, X_I]$ (or as a string $X_R X_I$) and we define

$$h(X) = h(X_R X_I). \tag{2.50}$$

Suppose that we wish to represent X in polar coordinates via the vector $[A_X, \Phi_X]$ where $A_X = |X|$ and Φ_X is the phase of X. Show that

$$\begin{aligned} h(X) &= h(A_X\Phi_X) + \mathsf{E}[\log A_X] && (2.51)\\ &= h(A_X^2) + h(\Phi_X|A_X) - \log 2. && (2.52) \end{aligned}$$

2.7. Appendix: Table of Differential Entropies

The constant $\gamma \approx 0.57721566$ is Euler's constant.

Distribution	Density $p(x)$	Support	Entropy (in nats)
Uniform	$\frac{1}{b-a}$	$a \leq x < b$	$\ln(b-a)$
Exponential	$\lambda e^{-\lambda x}$, $\lambda > 0$	$x \geq 0$	$1 - \ln \lambda$
Gaussian	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}}$	$x \in \mathbb{R}$	$\frac{1}{2}\ln(2\pi e\sigma^2)$
Rayleigh	$\frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}}$	$x \geq 0$	$1 + \frac{\gamma}{2} + \ln \frac{\sigma}{\sqrt{2}}$

References

[1] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, UK, 1985.

Chapter 3.

Channel Coding

3.1. Rate, Reliability, and Cost

Channel coding is concerned with transmitting data reliably from a source to a sink. One can anticipate that by reducing rate, one can increase reliability. For example, if your conversation partner does not understand what you say, simply repeat the message until you receive an acknowledgment. We thus expect a rate-reliability tradeoff and would like to determine the frontier of this tradeoff.

As a great surprise, Shannon showed that the tradeoff exhibits a sharp behavior: below a certain rate he called *capacity* one can achieve as reliable communication as desired! And above capacity one has a positive lower bound on reliability that increases very rapidly with increasing rate. Moreover, reliability is in general possible only by *coding* over long sequences of symbols.

But Shannon discovered more. First, one can control reliability by applying a lossy compression code followed by a channel code. Second, this *separation* of compression and reliability coding incurs no loss in rate. Third, if one introduces a *cost* S associated with transmission then there is a capacity-cost tradeoff. For example, a symbol with large energy generally costs more to transmit than one with small energy. Capacity naturally increases with increasing cost, and this suggests capacity-cost shapes such as those depicted in Fig. 3.1. We shall find that the uppermost shape is the correct one for the cost functions that we are interested in.

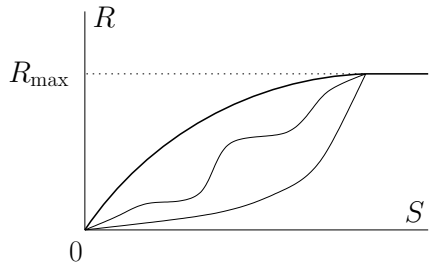

Figure 3.1.: Three possible shapes for increasing curves.

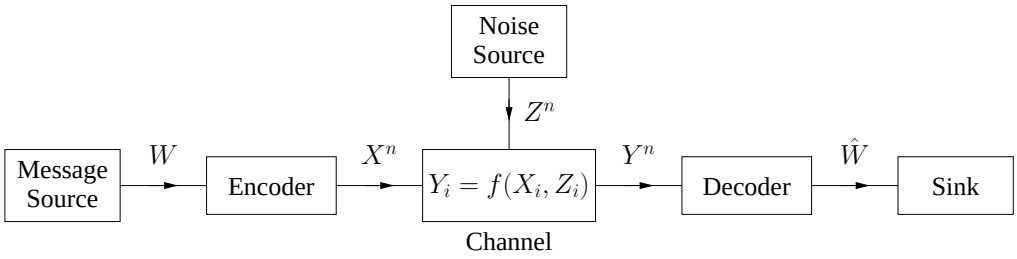

Figure 3.2.: The channel coding problem: the channel is defined by a function $f(\cdot)$ and a noise source $P_Z(\cdot)$.

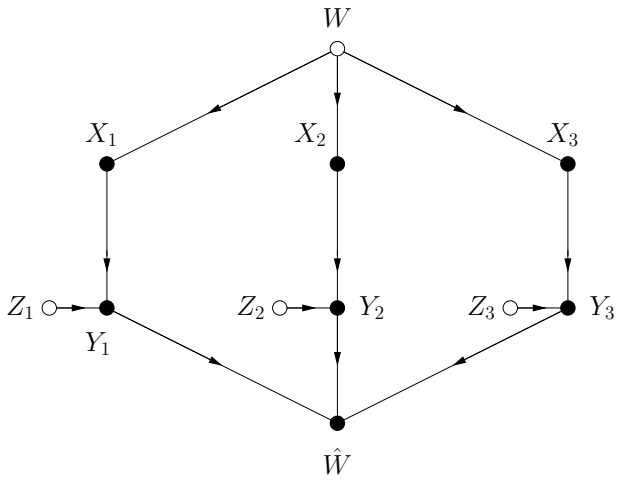

Figure 3.3.: FDG for a memoryless channel with $n = 3$. The hollow vertices represent mutually statistically independent random variables.

3.2. Memoryless Channels

A *memoryless channel* is the basic model for channel coding, and it is depicted in Fig. 3.2. The functional dependence graph (FDG) shown in Fig. 3.3 specifies the relationships between the random variables.[1] There are five types of variables: a source message W, channel inputs X_i, channel outputs Y_i, noise Z_i, $i = 1, 2, \ldots, n$, and a message estimate $\hat{W}$.

A *source* puts out the message w, $w \in \{1, 2, \ldots, M\}$. An *encoder* maps w to a string x^n in $\mathcal{X}^n$. We assume that $H(W) = nR$ and nR is an integer for simplicity. We may thus view W as being a string V^{nR} of independent bits V_i, $i = 1, 2, \ldots, nR$, where $P_{V_i}(0) = P_{V_i}(1) = 1/2$. The rate R measures how many information bits are sent per channel input symbol.

The *channel* puts out

$$y_i = f(x_i, z_i), \quad i = 1, 2, \ldots, n \tag{3.1}$$

for some function $f(\cdot)$ with range $\mathcal{Y}$, and where each z_i is a different realization of a noise random variable Z with alphabet $\mathcal{Z}$. In other words,

[1] A FDG is a graph where the vertices represent random variables and the edges represent functional dependencies, see Sec. A.5.

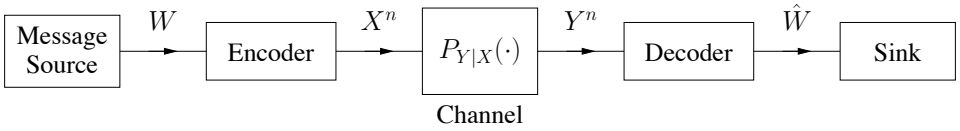

Figure 3.4.: The channel coding problem: the channel is a conditional probability distribution $P_{Y|X}(\cdot)$.

the $Z_1, Z_2, \ldots, Z_n$ all have the same distribution P_Z and are statistically independent. Furthermore, the noise string Z^n is statistically independent of W and X^n. The channel is *memoryless* because y_i is a function of x_i and z_i only. The channel is *time invariant* because $f(\cdot)$ and $P_Z(\cdot)$ do not depend on time. If the alphabets $\mathcal{X}$ and $\mathcal{Y}$ are discrete and finite, then the channel is called a *discrete memoryless channel (DMC)*.

The *decoder* maps y^n to an estimate $\hat{w}$ of w, where $\hat{\mathcal{W}} = \mathcal{W}$. The goal is to find the maximum rate R for which, by choosing n sufficiently large, one can make $P_e = \Pr\left[\hat{W} \neq W\right]$ arbitrarily close to zero (but not necessarily exactly zero). This maximum rate is called the *capacity* C.

Observe that capacity does not account for the *delay* due to encoding and decoding, or for the *complexity* of encoding and decoding. Delay and complexity are, of course, of great engineering relevance. We are therefore trying to find only the limits that we should try to approach if we we are willing to delay information transfer, and if we are willing to build complex devices.

The above functional definitions can alternatively be written by using probability distributions. In particular, for a DMC we may write the channel (3.1) as a conditional probability distribution $P_{Y|X}(\cdot)$ as in Fig. 3.4. The joint distribution of the random variables (other than the noise) satisfies

$$P(w, x^n, y^n, \hat{w}) = P(w)P(x^n|w)\left[\prod_{i=1}^{n} P_{Y|X}(y_i|x_i)\right] P(\hat{w}|y^n) \tag{3.2}$$

where both $P(x^n|w)$ and $P(\hat{w}|y^n)$ are either 0 or 1.

As done here, we will often remove the subscripts on the probability distributions if the arguments are lower-case versions of the random variables. For instance, we write $P(w)$ for $P_W(w)$. Similarly, we could write $P(y_i|x_i)$ for $P_{Y_i|X_i}(y_i|x_i)$ but we prefer to write this as $P_{Y|X}(y_i|x_i)$ to emphasize that the channel $P_{Y|X}$ is time-invariant.

Example 3.1. A *binary symmetric channel* (BSC) has $\mathcal{X} = \mathcal{Y} = \mathcal{Z} = \{0, 1\}$ and

$$f(x, z) = x + z \tag{3.3}$$

where addition is modulo-2. A diagram representing the BSC is shown in Fig. 3.5 where $P_Z(1) = p$. The parameter p is called the channel *crossover probability* and we usually have $p \leq 1/2$. The channel conditional probabil-

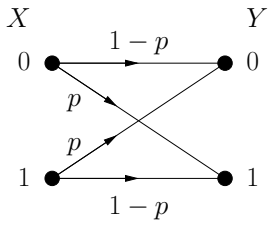

Figure 3.5.: BSC with crossover probability p.

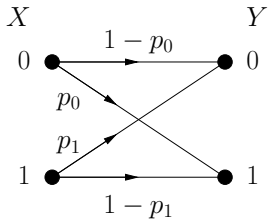

Figure 3.6.: Binary channel with asymmetric crossover probabilities.

ity distribution is

$$P(y|x) = \left\{ \begin{array}{ll} 1-p & \text{if } y = x \\ p & \text{if } y \neq x \end{array} \right\}. \tag{3.4}$$

Example 3.2. An asymmetric version of the BSC has $\mathcal{Z} = \{00, 01, 10, 11\}$ and

$$f(x, z_0 z_1) = \left\{ \begin{array}{ll} z_0 & \text{if } x = 0 \\ 1 + z_1 & \text{if } x = 1 \end{array} \right\} = x + z_x \tag{3.5}$$

where the noise is now a pair $Z = Z_0 Z_1$ of binary random variables. The variables X and Z are independent, but X and Z_X are dependent in general. A diagram representing the BSC is shown in Fig. 3.6 where $p_x = P_{Z_x}(1)$. The channel conditional probability distribution is

$$P(y|x) = \left\{ \begin{array}{ll} 1-p_x & \text{if } y = x \\ p_x & \text{if } y \neq x \end{array} \right\}. \tag{3.6}$$

Example 3.3. A *binary erasure channel* (BEC) has $\mathcal{X} = \{0, 1\}$, $\mathcal{Y} = \{0, 1, \Delta\}$, $\mathcal{Z} = \{0, \Delta\}$, and

$$f(x, z) = \left\{ \begin{array}{ll} x & \text{if } z = 0 \\ \Delta & \text{if } z = \Delta \end{array} \right\}. \tag{3.7}$$

A diagram representing the BEC is shown in Fig. 3.7 where $P_Z(\Delta) = p$. Observe that receiving $y = 0$ or $y = 1$ gives full knowledge of x. The channel

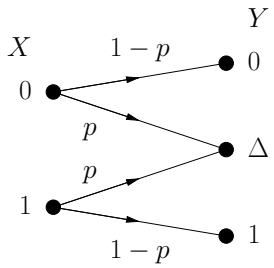

Figure 3.7.: BEC with erasure probability p.

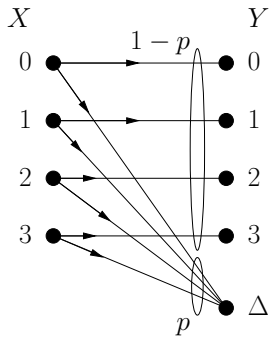

Figure 3.8.: PEC with $b = 2$ and erasure probability p.

conditional probability distribution is

$$P(y|x) = \left\{ \begin{array}{ll} 1-p & \text{if } y = x \\ p & \text{if } y = \Delta \\ 0 & \text{if } x = 0 \text{ and } y = 1 \text{ or if } x = 1 \text{ and } y = 0 \end{array} \right\}. \tag{3.8}$$

Example 3.4. A *packet erasure channel* (PEC) with packets of length b bits has $\mathcal{X} = \{0, 1, \ldots, 2^b - 1\}$, $\mathcal{Y} = \{0, 1, \ldots, 2^b - 1, \Delta\}$, $\mathcal{Z} = \{0, \Delta\}$, and $f(x, z)$ having the same form as (3.7). The PEC with $b = 1$ is thus a BEC. A diagram representing the PEC with $b = 2$ is shown in Fig. 3.8 where $P_Z(\Delta) = p$. Receiving y, $y \neq \Delta$, again gives full knowledge of x. The channel conditional probability distribution is

$$P(y|x) = \left\{ \begin{array}{ll} 1-p & \text{if } y = x \\ p & \text{if } y = \Delta \\ 0 & \text{if } y \neq \Delta \text{ and } y \neq x \end{array} \right\}. \tag{3.9}$$

Example 3.5. An *additive channel* has

$$f(x, z) = x + z \tag{3.10}$$

where the "+" denotes addition in a field with alphabets $\mathcal{X} = \mathcal{Y} = \mathcal{Z}$. For example, a BSC is additive over the Galois field GF(2). If the field is the set

of real numbers and Z is a Gaussian random variable with zero mean then we have an *additive white Gaussian noise (AWGN)* channel. Of course, an AWGN channel is not a DMC. We must therefore replace $P_{Y|X}(y|x)$ in (3.2) with a channel conditional probability *density* function $p(y|x)$. If Z has variance N then we have

$$p(y|x) = p_Z(y-x) = \frac{1}{\sqrt{2\pi N}} e^{\frac{-(y-x)^2}{2N}}. \tag{3.11}$$

3.3. Cost Functions

The transmission and reception of symbols often incurs *costs*, e.g., power or energy costs. We therefore refine the capacity problem by adding a *cost constraint*. Suppose that transmitting x^n and receiving y^n incurs a cost of $s^n(x^n, y^n)$ units. We require that the *average* cost satisfy

$$\mathsf{E}[s^n(X^n, Y^n)] \leq S. \tag{3.12}$$

We consider $s^n(\cdot)$ that are averages of a per-letter cost function $s(\cdot)$:

$$s^n(x^n, y^n) = \frac{1}{n}\sum_{i=1}^{n} s(x_i, y_i). \tag{3.13}$$

The largest rate C as a function of the cost S is called the *capacity cost* function, and is denoted $C(S)$.

Example 3.6. Suppose $\mathcal{X} = \{0, 1\}$ and

$$s(x, y) = \begin{cases} 0 & \text{if } x = 0 \\ E & \text{if } x = 1 \end{cases} \tag{3.14}$$

so that sending $x = 1$ costs E units of *energy*. This situation might occur for an optical channel where transmitting a 1 represents light while transmitting a 0 represents dark. A cost constraint with $0 \leq S < E/2$ will bias the best transmission scheme towards sending the symbol 1 less often.

Example 3.7. Suppose $\mathcal{X} = \mathbb{R}$ and $s(x, y) = x^2$ so that

$$\mathsf{E}[s^n(X^n, Y^n)] = \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}\left[X_i^2\right] \leq P \tag{3.15}$$

where we have chosen $S = P$ to emphasize that we interpret P as the *average transmit power*. The *energy* is thus at most nP. The constraint (3.15) is called an *average block power* constraint. Another interesting choice is $s(x, y) = y^2$ which limits the *average receive* power.

Example 3.8. A more stringent constraint than (3.12) is that the cost be bounded with probability 1:

$$\Pr\left[s^n(X^n, Y^n) \leq S\right] = 1. \tag{3.16}$$

For example, with $s(x, y) = x^2$ and $S = P$ we obtain the power constraints

$$\frac{1}{n}\sum_{i=1}^{n} x_i(w)^2 \leq P, \quad \text{for all } w = 1, 2, \ldots, M. \tag{3.17}$$

3.4. Block and Bit Error Probability

Consider $W = V^b$ where the V_i are independent bits with $P_{V_i}(0) = P_{V_i}(1) = 1/2$, $i = 1, 2, \ldots, b$. We thus have $M = 2^b$ and $R = b/n$ bits per channel symbol. We have defined the channel coding by requiring that the *block* error probability

$$P_e = \Pr\left[\hat{W} \neq W\right] \tag{3.18}$$

be small. However, one is sometimes interested in minimizing the *average bit* error probability

$$P_{\rm b} = \frac{1}{b}\sum_{i=1}^{b} \Pr\left[\hat{V}_i \neq V_i\right]. \tag{3.19}$$

We have the following relations between $P_{\rm b}$ and $P_{\rm e}$:

$$P_{\rm b} \overset{(a)}{\leq} P_{\rm e} \overset{(b)}{\leq} b\, P_{\rm b}. \tag{3.20}$$

The bound (a) is because a bit error implies a block error, and the bound (b) is because a block error implies at least 1 bit error for the b bits. One has equality on the left if *all* bits in an erroneous block are incorrect, and one has equality on the right if *exactly one* bit is incorrect in each erroneous block. The bounds (3.20) imply that if $P_{\rm b}$ is positive so is $P_{\rm e}$. Similarly, if $P_{\rm e}$ is small so is $P_{\rm b}$. This is why *coding* theorems should upper bound $P_{\rm e}$ and *converse* theorems should lower bound $P_{\rm b}$. For example, a code with large $P_{\rm e}$ can have small $P_{\rm b}$.

We next develop lower bounds on $P_{\rm e}$ and $P_{\rm b}$. Consider first $P_{\rm e}$. Using Fano's inequality and $|\mathcal{W}| = M = 2^{nR}$ we have

$$\begin{aligned} H(W|\hat{W}) &\leq H_2(P_e) + P_e \log_2(|\mathcal{W}| - 1) \\ &< H_2(P_e) + P_e\, nR \end{aligned} \tag{3.21}$$

We also have $H(W|\hat{W}) = H(W) - I(W;\hat{W})$ and $H(W) = nR$ so that[2]

$$\boxed{nR < \frac{I(W;\hat{W}) + H_2(P_e)}{1 - P_e}.} \tag{3.22}$$

The bound (3.22) gives a rate-reliability tradeoff, i.e., nR is at most $I(W;\hat{W})$ if reliability is good (P_e is small) and nR can become larger if reliability is poor (P_e is large). The tradeoff is parameterized by $I(W;\hat{W})$. We emphasize that (3.22) is valid for any choice of P_e.

[2]The bounds (3.22) and (3.25) are valid for any channel: discrete or continuous, memoryless or with memory. An improved lower bound on the block error probability P_e for memoryless channels is called a *strong* converse and is developed in Problem 3.6.

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

$$\begin{aligned} H_2(P_{\rm b}) &= H_2\left(\frac{1}{b}\sum_{i=1}^{b}\Pr\left[\hat{V}_i \neq V_i\right]\right) \\ &\overset{(a)}{\geq} \frac{1}{b}\sum_{i=1}^{b} H_2\left(\Pr\left[\hat{V}_i \neq V_i\right]\right) \\ &\overset{(b)}{\geq} \frac{1}{b}\sum_{i=1}^{b} H(V_i|\hat{V}_i) \end{aligned} \tag{3.23}$$

where (a) follows by the concavity of $H_2(\cdot)$, and (b) by Fano's inequality. We continue the chain of inequalities by using $H(V_i|\hat{V}_i) \geq H(V_i|V^{i-1}\hat{V}^b)$ so that

$$\begin{aligned} H_2(P_{\rm b}) &\geq \frac{1}{b}\sum_{i=1}^{b} H(V_i|V^{i-1}\hat{V}^b) \\ &= \frac{1}{b}H(V^b|\hat{V}^b) \\ &= \frac{1}{b}\left(H(V^b) - I(V^b;\hat{V}^b)\right) \\ &= 1 - \frac{I(W;\hat{W})}{nR}. \end{aligned} \tag{3.24}$$

We thus have the following counterpart to (3.22):

$$\boxed{nR \leq \frac{I(W;\hat{W})}{1-H_2(P_{\rm b})}.} \tag{3.25}$$

We again have a rate-reliability tradeoff with $0 \leq P_b \leq 1/2$, and we essentially require $nR \leq I(W;\hat{W})$ if $P_{\rm b}$ is small.

Observe that we can write

$$P_e = \sum_{w=1}^{M} P(w)\Pr\left[\hat{W} \neq w \,\middle|\, W = w\right] \tag{3.26}$$

and the probabilities $\Pr\left[\hat{W} \neq w \,\middle|\, W = w\right]$ can be different. An interesting parameter is therefore the *maximum* block error probability

$$P_m = \max_{1\leq w\leq M} \Pr\left[\hat{W} \neq w \,\middle|\, W = w\right]. \tag{3.27}$$

We clearly have $P_b \leq P_e \leq P_m$ so that a lower bound on P_b is a lower bound on P_m and an upper bound on P_m is also an upper bound on P_b and P_e.

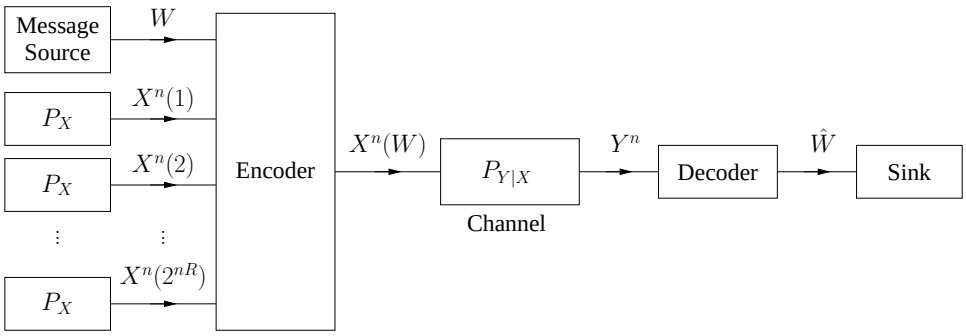

Figure 3.9.: Random coding experiment for channel coding.

3.5. Random Coding

We construct a *random* code book for the DMC with a cost constraint. We begin by choosing a distribution P_X.

Code Construction: Generate $M = 2^{nR}$ codewords $x^n(w)$, $w = 1, 2, \ldots, 2^{nR}$, by choosing the $n \cdot 2^{nR}$ symbols $x_i(w)$ in the code book

$$\underline{x}^n = [x^n(1), x^n(2), \ldots, x^n(M)] \tag{3.28}$$

independently using P_X.

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

Decoder: Given y^n, choose $\hat{w}$ as (one of) the message(s) w that maximizes

$$P(w|y^n) = \frac{P(w)P(y^n|w)}{P(y^n)} = \frac{P(w)P(y^n|x^n(w))}{P(y^n)}. \tag{3.29}$$

This decoder is called a *maximimum a-posteriori probability* (MAP) decoder. Since all messages are equally likely, the MAP decoder is the same as the *maximum likelihood* (ML) decoder that chooses $\hat{w}$ as (one of) the message(s) w that maximizes the $P(y^n|x^n(w))$.

Analysis: The random coding experiment is shown in Fig. 3.9 where the random variables W, $X^n(1), X^n(2), \ldots, X^n(2^{nR})$ are mutually statistically independent. The joint distribution of the random variables is

$$\begin{aligned} &P(w, x^n(1), \ldots, x^n(2^{nR}), y^n, \hat{w}) \\ &= P(w) \left[\prod_{i=1}^{2^{nR}} P_X^n\left(x^n(i)\right) \right] P_{Y|X}^n\left(y^n|x^n(w)\right) 1(\hat{w} = f(y^n)) \end{aligned} \tag{3.30}$$

where $1(\cdot)$ is the indicator function that takes on the value 1 if its argument is true and is 0 otherwise. We compute the error probability for this experiment. We have two error events

$$\mathcal{E} = \{\hat{W} \neq W\} \tag{3.31}$$

$$\mathcal{F} = \{\mathsf{E}[s^n(X^n, Y^n)] > S\} \tag{3.32}$$

and the error probability can be written as the code book average

$$\Pr[\mathcal{E} \cup \mathcal{F}] = \sum_{\underline{x}^n} P(\underline{x}^n) \Pr[\mathcal{E} \cup \mathcal{F} \,|\, \underline{X}^n = \underline{x}^n]. \tag{3.33}$$

We wish to find conditions so that for any positive δ there is a sufficiently large n such that $\Pr[\mathcal{E} \cup \mathcal{F}] \leq \delta$. If we are successful, then (3.33) guarantees there must exist a code book $\underline{x}^n$ for which $\Pr[\mathcal{E} \cup \mathcal{F} \,|\, \underline{X}^n = \underline{x}^n] \leq \delta$.

3.5.1. Block Error Probability

The union bound gives $\Pr[\mathcal{E} \cup \mathcal{F}] \leq \Pr[\mathcal{E}] + \Pr[\mathcal{F}]$ so we may upper bound the probability of each error event separately. We begin with

$$\Pr[\mathcal{E}] = \sum_{w} P(w) \Pr\left[\hat{W} \neq w \,|\, W = w\right]. \tag{3.34}$$

By symmetry, we have $\Pr\left[\hat{W} \neq w \,|\, W = w\right] = \Pr\left[\hat{W} \neq 1 \,|\, W = 1\right]$ so we proceed to bound $\Pr\left[\hat{W} \neq 1 \,|\, W = 1\right]$.

Consider the events

$$\mathcal{E}(\tilde{w}) = \{P_{Y^n|X^n}(Y^n|X^n(1)) \leq P_{Y^n|X^n}(Y^n|X^n(\tilde{w}))\} \tag{3.35}$$

for $\tilde{w} \neq 1$. The event $\{\hat{W} \neq 1\}$ occurs only if $\mathcal{E}(\tilde{w})$ occurs for at least one $\tilde{w}$ with $\tilde{w} \neq 1$. Furthermore, for such a $\tilde{w}$ the likelihood ratio

$$L(\tilde{w}) = \frac{P_{Y^n|X^n}(Y^n|X^n(\tilde{w}))}{P_{Y^n|X^n}(Y^n|X^n(1))} \tag{3.36}$$

must be at least 1. We thus have $L(\tilde{w})^s \geq 1$ for any $s \geq 0$ and also $\{\sum_{\tilde{w}\neq 1} L(\tilde{w})^s\}^\rho \geq 1$ for any $s \geq 0$ and any $\rho \geq 0$. These results imply

$$\begin{aligned}\Pr\left[\hat{W} \neq 1 \,|\, W = 1\right] &\overset{(a)}{\leq} \Pr\left[\left\{\sum_{\tilde{w}\neq 1} L(\tilde{w})^s\right\}^\rho \geq 1 \,\middle|\, W = 1\right] \\ &\overset{(b)}{\leq} \mathsf{E}\left[\left\{\sum_{\tilde{w}\neq 1} L(\tilde{w})^s\right\}^\rho \,\middle|\, W = 1\right]\end{aligned} \tag{3.37}$$

where (a) is because $\mathcal{A} \Rightarrow \mathcal{B}$ implies $\Pr[A] \leq \Pr[B]$, and (b) follows by the Markov inequality (A.65).

The expectation in (3.37) is with respect to the $X^n(w)$, $w = 1, 2, \ldots, M$, and Y^n. Since we have $W = 1$, the pair $(X^n(1), Y^n)$ is jointly distributed according to the channel distribution $P_{Y|X}$. The $X^n(\tilde{w})$, $\tilde{w} \neq 1$, are statistically independent of $(X^n(1), Y^n)$. If we take the expectation over the pair

$(X^n(1), Y^n)$, then the right-hand side of (3.37) is

$$\sum_{x^n,y^n} P(x^n,y^n)\cdot \mathsf{E}\left[\left\{\sum_{\tilde{w}\neq 1}\left(\frac{P_{Y^n|X^n}(y^n|X^n(\tilde{w}))}{P(y^n|x^n)}\right)^s\right\}^\rho \middle| X^n(1)=x^n, Y^n=y^n\right]$$

$$\stackrel{(a)}{=} \sum_{x^n,y^n} P(x^n)P(y^n|x^n)^{1-s\rho}\cdot \mathsf{E}\left[\left\{\sum_{\tilde{w}\neq 1} P_{Y^n|X^n}(y^n|X^n(\tilde{w}))^s\right\}^\rho\right] \tag{3.38}$$

where we removed the conditioning on $W=1$ because $(X^n(1), Y^n)$ and the $X^n(\tilde{w})$, $\tilde{w}\neq 1$, are independent of $\{W=1\}$. Step (a) follows because the $X^n(\tilde{w})$, $\tilde{w}\neq 1$, are independent of $(X^n(1), Y^n)$. For $0\leq\rho\leq 1$, Jensen's inequality gives $\mathsf{E}[X^\rho]\leq \mathsf{E}[X]^\rho$ so the expectation in (3.38) is upper bounded by

$$\left\{\sum_{\tilde{w}\neq 1}\mathsf{E}\left[P_{Y^n|X^n}(y^n|X^n(\tilde{w}))^s\right]\right\}^\rho = \left\{(M-1)\sum_{x^n}P(x^n)P(y^n|x^n)^s\right\}^\rho. \tag{3.39}$$

Now substitute $s=1/(1+\rho)$ and insert (3.39) into (3.38) to obtain

$$\Pr\left[\hat{W}\neq 1\,|\,W=1\right] \leq (M-1)^\rho \sum_{y^n}\left\{\sum_{x^n}P(x^n)P(y^n|x^n)^{\frac{1}{1+\rho}}\right\}^{1+\rho}. \tag{3.40}$$

We remark that (3.40) is valid for channels *with* memory and any $P(x^n)$. It is also valid for continuous channels by replacing $P(y^n|x^n)$ with the conditional density $p(y^n|x^n)$.

Using $(M-1)^\rho \leq M^\rho = 2^{nR\rho}$, random coding with $P(x^n)=\prod_{i=1}^n P_X(x_i)$, and the memoryless property $P(y^n|x^n)=\prod_{i=1}^n P_{Y|X}(y_i|x_i)$, the bound (3.40) becomes

$$\Pr\left[\hat{W}\neq 1\,|\,W=1\right] \leq 2^{-n[E_0(\rho,P_X)-\rho R]} \tag{3.41}$$

for all $0\leq\rho\leq 1$ where

$$E_0(\rho,P_X) = -\log_2 \sum_y \left\{\sum_x P(x)P(y|x)^{\frac{1}{1+\rho}}\right\}^{1+\rho}. \tag{3.42}$$

Optimizing over ρ, we have

$$\Pr\left[\hat{W}\neq 1\,|\,W=1\right] \leq 2^{-nE_G(R,P_X)} \tag{3.43}$$

where $E_G(R,P_X)$ is the *Gallager exponent*[3]

$$\boxed{E_G(R,P_X) = \max_{0\leq\rho\leq 1}\left[E_0(\rho,P_X)-\rho R\right].} \tag{3.44}$$

One may prove the following properties of $E_G(R,P_X)$ (see Problem 3.5):

a) $E_G(0,P_X) = E_0(1,P_X)$.

[3]The Gallager exponent optimized over P_X is written as $E_G(R)=\max_{P_X} E_G(R,P_X)$.

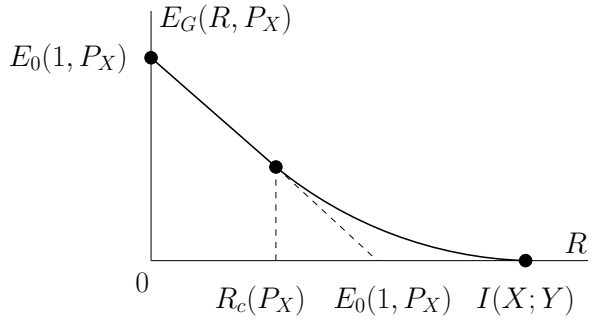

Figure 3.10.: Gallager exponent $E_G(R, P_X)$ for a fixed P_X.

b) $E_G(R, P_X)$ is linear with slope -1 for $0 \leq R \leq R_c(P_X)$ where

$$R_c(P_X) = \left.\frac{\partial E_0(\rho, P_X)}{\partial \rho}\right|_{\rho=1}. \tag{3.45}$$

c) $E_G(R, P_X)$ is convex and positive in the interval $0 \leq R < I(X;Y)$ if $I(X;Y)$ is positive.

The shape of $E_G(R, P_X)$ is shown in Fig. 3.10. We see that as long as we choose R and P_X so that

$$\boxed{R < I(X;Y)} \tag{3.46}$$

then we can make $\Pr\left[\hat{W} \neq W\right] \leq \delta$ for any $\delta > 0$ by choosing large n.

3.5.2. Capacity-Cost Function

For the cost constraint we have

$$\Pr[\mathcal{F}] = 1\left(\mathsf{E}[s^n(X^n(W), Y^n)] > S\right) \tag{3.47}$$

where 1(.) is the indicator function that takes on the value 1 if its argument is true and is zero otherwise. We must therefore choose P_X so that $\mathsf{E}[s^n(X^n, Y^n)] \leq S$ over the ensemble of code books. We easily compute

$$\mathsf{E}[s^n(X^n(W), Y^n)] = \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}[s(X_i(W), Y_i)] = \mathsf{E}[s(X, Y)]\,. \tag{3.48}$$

Thus, we must choose P_X so that

$$\boxed{\mathsf{E}[s(X, Y)] \leq S}. \tag{3.49}$$

Combining the above results, for large n there is a code in the random ensemble of codes that has small P_e, expected cost at most S, and with a

rate that approaches the *capacity-cost function*

$$\boxed{C(S) = \max_{P_X:\, \mathsf{E}[s(X,Y)]\leq S} I(X;Y).} \tag{3.50}$$

If there is no cost constraint, then we can approach the rate

$$\boxed{C = \max_{P_X} I(X;Y).} \tag{3.51}$$

3.5.3. Maximum Error Probability

Consider the maximum error probability (3.27). Suppose we have a *fixed* code for which

$$\Pr\left[\mathcal{E} \cup \mathcal{F}\right] \leq \delta. \tag{3.52}$$

For example, we might have found this code by using the random coding method described above. We may write

$$\Pr\left[\mathcal{E} \cup \mathcal{F}\right] = \sum_{w=1}^{M} P(w) \Pr\left[\mathcal{E} \cup \mathcal{F} \,|\, W = w\right]. \tag{3.53}$$

Now re-index the codewords by ordering them so that the error probability increases as w increases, i.e., re-index so that

$$\Pr\left[\mathcal{E} \cup \mathcal{F} \,|\, W = w_1\right] \leq \Pr\left[\mathcal{E} \cup \mathcal{F} \,|\, W = w_2\right] \tag{3.54}$$

if $w_1 \leq w_2$. Now consider the code book having only the first $M/2$ messages (suppose M is even) of the re-indexed code book. The rate is reduced by $1/n$ bits as compared to the original code book, but the maximum error probability satisfies

$$\begin{aligned} P_m &= \max_{1\leq w\leq M/2} \Pr\left[\mathcal{E} |\; W = w\right] \\ &\leq \max_{1\leq w\leq M/2} \Pr\left[\mathcal{E} \cup \mathcal{F} |\; W = w\right] \\ &= \Pr\left[\mathcal{E} \cup \mathcal{F} |\; W = M/2\right] \\ &\overset{(a)}{\leq} 2\delta \end{aligned} \tag{3.55}$$

where (a) follows by (3.53) since no more than half of the terms in the sum can be larger than 2δ. In other words, by *expurgating* our codes we obtain codes with a maximum error probability as close to zero as desired and rates as close to $C(S)$ as desired.

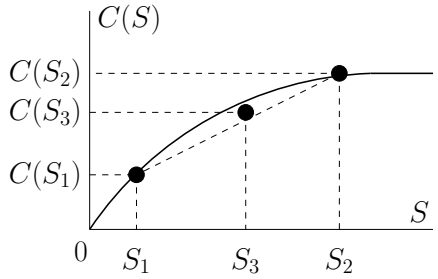

Figure 3.11.: Concavity of the capacity-cost function.

3.6. Concavity and Converse

The function $C(S)$ in (3.50) is non-decreasing in S because increasing S permits using a larger class of P_X. We show that $C(S)$ is concave in S.

Consider two distinct points $(S_1, C(S_1))$ and $(S_2, C(S_2))$ and suppose the distributions P_{X_1} and P_{X_2} achieve these respective points (see Fig. 3.11). That is, we have

$$\begin{aligned} S_1 &= \mathsf{E}[s(X_1,Y_1)]\,, \quad C(S_1) = I(X_1;Y_1) \\ S_2 &= \mathsf{E}[s(X_2,Y_2)]\,, \quad C(S_2) = I(X_2;Y_2) \end{aligned} \tag{3.56}$$

where Y_1 and Y_2 are the respective outputs of the channel $P_{Y|X}$ when the input is X_1 and X_2. Consider the mixture distribution

$$P_{X_3}(x) = \lambda\, P_{X_1}(x) + (1-\lambda)\, P_{X_2}(x) \tag{3.57}$$

for all x, where $0 \le \lambda \le 1$. We have

$$\begin{aligned} S_3 &= \sum_{(x,y)\in \mathrm{supp} P_{X_3Y}} P_{X_3}(x)\; P(y|x)\; s(x,y) \\ &= \sum_{(x,y)\in \mathrm{supp} P_{X_3Y}} (\lambda P_{X_1}(x) + (1-\lambda) P_{X_2}(x))\;\; P(y|x)\; s(x,y) \\ &= \lambda S_1 + (1-\lambda) S_2. \end{aligned} \tag{3.58}$$

We thus have

$$\begin{aligned} C(\lambda\, S_1 + (1-\lambda)\, S_2) &= C(S_3) \\ &\overset{(a)}{\ge} I(X_3;Y_3) \\ &\overset{(b)}{\ge} \lambda\, I(X_1;Y_1) + (1-\lambda)\, I(X_2;Y_2) \\ &= \lambda\, C(S_1) + (1-\lambda)\, C(S_2). \end{aligned} \tag{3.59}$$

where (a) follows because P_{X_3} might not give the maximum mutual information for the cost S_3, and (b) follows by the concavity of $I(X;Y)$ in P_X when $P_{Y|X}$ is held fixed (see Thm. 1.10). The bound (3.59) means that $C(S)$ is concave in S.

We now show that $C(S)$ in (3.50) is the capacity cost function. We have

$$\begin{aligned} I(W;\hat{W}) &\overset{(a)}{\leq} I(X^n;Y^n) \\ &= \sum_{i=1}^{n} H(Y_i|Y^{i-1}) - H(Y_i|X_i) \\ &\leq \sum_{i=1}^{n} H(Y_i) - H(Y_i|X_i) \\ &= \sum_{i=1}^{n} I(X_i;Y_i). \end{aligned} \tag{3.60}$$

where (a) follows by the data processing inequality. If there is no cost constraint then we can use the simple bound

$$I(X_i;Y_i) \leq \max_{P_X} I(X;Y) = C. \tag{3.61}$$

Inserting (3.61) into (3.60) and then into (3.22) and (3.25) we obtain $R \leq C$ for small P_e and P_b.

More generally, with a cost constraint we must use the concavity of $C(S)$ in S. We have $I(X_i;Y_i) \leq C(\mathsf{E}[s(X_i,Y_i)])$ because $C(\mathsf{E}[s(X_i,Y_i)])$ maximizes mutual information for the cost $\mathsf{E}[s(X_i,Y_i)]$. We thus have

$$\begin{aligned} \sum_{i=1}^{n} I(X_i;Y_i) &\leq n \sum_{i=1}^{n} \frac{1}{n} C\left(\mathsf{E}[s(X_i,Y_i)]\right) \\ &\overset{(a)}{\leq} n\, C\left(\frac{1}{n}\sum_{i=1}^{n} \mathsf{E}[s(X_i,Y_i)]\right) \\ &= n\, C(\mathsf{E}[s^n(X^n,Y^n)]) \\ &\overset{(b)}{\leq} n\, C(S) \end{aligned} \tag{3.62}$$

where (a) follows by the concavity of $C(S)$ and (b) follows because we require $\mathsf{E}[s^n(X^n,Y^n)] \leq S$ and because $C(S)$ is non-decreasing in S. Inserting (3.62) into (3.60) and then into (3.22) and (3.25) we have

$$\boxed{R < \frac{C(S) + H_2(P_e)/n}{1 - P_e}.} \tag{3.63}$$

and

$$\boxed{R \leq \frac{C(S)}{1 - H_2(P_\mathrm{b})}.} \tag{3.64}$$

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

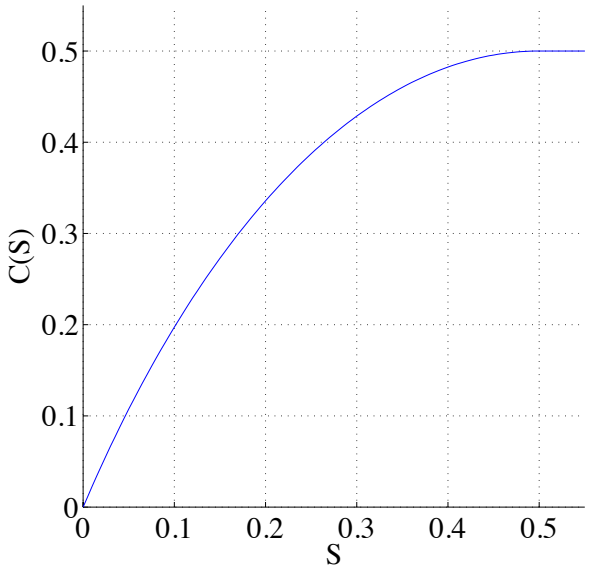

Figure 3.12.: Capacity-cost function for a BSC with $p = 0.11$ and $E = 1$.

3.7. Discrete Alphabet Examples

3.7.1. Binary Symmetric Channel

The binary symmetric channel (BSC) has $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ and $\Pr[Y \neq X] = p$. In the absence of a cost constraint, we have

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H_2(P_X(1) * p) - H_2(p) \end{aligned} \tag{3.65}$$

where $q * p = q(1-p) + (1-q)p$. The best choice for P_X is $P_X(0) = P_X(1) = 1/2$ so that

$$\boxed{C = 1 - H_2(p)}. \tag{3.66}$$

Suppose next that we have the cost function (3.14). We compute

$$\mathsf{E}[s(X)] = P_X(1) \cdot E. \tag{3.67}$$

The capacity cost function is thus

$$C(S) = H_2(\min(S/E, 1/2) * p) - H_2(p) \tag{3.68}$$

and for $S \geq E/2$ we have $C = 1 - H_2(p)$. A plot of $C(S)$ is shown in Fig. 3.12 for the case $p = 0.11$ and $E = 1$.

3.7.2. Binary Erasure Channel

The binary erasure channel (BEC) has $\mathcal{X} = \{0, 1\}$, $\mathcal{Y} = \{0, 1, \Delta\}$, and $\Pr[Y = X] = 1-p$ and $\Pr[Y = \Delta] = p$. For no cost constraint, we compute

$$\begin{aligned} C &= \max_{P_X} \; H(X) - H(X|Y) \\ &= \max_{P_X} \; H(X)\,(1-p) \end{aligned} \tag{3.69}$$

and choosing P_X to be coin-flipping we have

$$\boxed{C = 1 - p}. \tag{3.70}$$

3.7.3. Strongly Symmetric Channels

Many practical channels exhibit symmetries. Two symmetries that we consider in detail are:

- *Uniformly dispersive*: for every input letter x the list of probabilities $\{P(y|x) : y \in \mathcal{Y}\}$, is the same.
- *Uniformly focusing*: for every output letter y the list of probabilities $\{P(y|x) : x \in \mathcal{X}\}$, is the same.

For example, a BSC is both uniformly dispersive and uniformly focusing. A BEC is uniformly dispersive but not uniformly focusing.

Uniformly dispersive channels have the special property that $H(Y|X)$ does not depend on P_X. To see this, observe that

$$H(Y|X = x) = \sum_{y \in \mathcal{Y}} -P(y|x) \log_2 P(y|x) \tag{3.71}$$

is the same for all x. Hence $H(Y|X)$ is given by (3.71) and determining the capacity-cost function reduces to a constrained maximum-entropy problem:

$$C(S) = \left[\max_{P_X \colon \mathsf{E}[s(X,Y)] \le S} \; H(Y) \right] - H(Y|X). \tag{3.72}$$

On the other hand, uniformly focusing channels have the special property that a uniform P_X results in a uniform P_Y. To see this, observe that if $P(x) = 1/|\mathcal{X}|$ for all x then we have

$$P(y) = \sum_{x \in \mathcal{X}} \frac{1}{|\mathcal{X}|} P(y|x) \tag{3.73}$$

which is the same for all b. Thus, P_Y is uniform which maximizes $H(Y)$.

A channel is said to be *strongly symmetric* if it is both uniformly dispersive and uniformly focusing. In the absence of a cost constraint, this means that

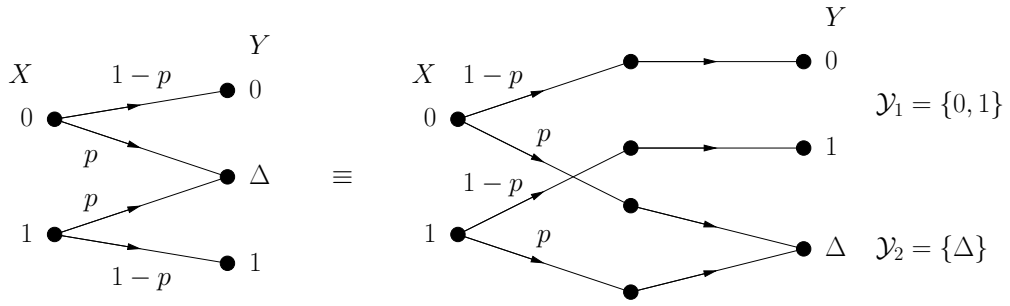

Figure 3.13.: A BEC decomposed into two strongly symmetric channels.

a strongly symmetric channel has capacity

$$\boxed{C = \log_2 |\mathcal{Y}| - \sum_{y \in \mathcal{Y}} -P(y|x) \log_2 P(y|x)} \tag{3.74}$$

for any choice of x, and capacity is achieved by a uniform P_X. For example, a BSC is strongly symmetric and has capacity $C = 1 - H_2(p)$ which matches (3.74).

3.7.4. Symmetric Channels

Strongly symmetric channels are somewhat restrictive, e.g., they do not include the BEC. We now consider the more interesting class of *symmetric* channels that have the special property that they can be decomposed into strongly symmetric channels. By this we mean that one can partition the output symbols in $\mathcal{Y}$ into L sets $\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_L$ such that the channel from $\mathcal{X}$ to $\mathcal{Y}_i$ is strongly symmetric (uniformly dispersive and uniformly focusing) for all $i = 1, 2, \ldots, L$.

Example 3.9. The BEC can be decomposed into $L = 2$ strongly symmetric channels with $\mathcal{Y}_1 = \{0, 1\}$ and $\mathcal{Y}_2 = \{\Delta\}$, i.e., we have

$$\text{Channel 1}: \begin{cases} \{P(y|x) : y \in \mathcal{Y}_1\} = \{0, 1-p\} \text{ is the same for all } x \in \mathcal{X} \\ \{P(y|x) : x \in \mathcal{X}\} = \{0, 1-p\} \text{ is the same for all } y \in \mathcal{Y}_1 \end{cases}$$

$$\text{Channel 2}: \begin{cases} \{P(y|x) : y \in \mathcal{Y}_2\} = \{p\} \text{ is the same for all } x \in \mathcal{X} \\ \{P(y|x) : x \in \mathcal{X}\} = \{p, p\} \text{ is the same for all } y \in \mathcal{Y}_2. \end{cases}$$

The decomposition is shown in Fig. 3.13. Observe that each input symbol X goes through either a BSC with crossover probability 0 (Channel 1) or through a channel that maps both inputs to Δ with probability 1 (Channel 2). The former channel is chosen with probability $1 - p$ and the latter with probability p. Observe that Y specifies which sub-channel was chosen.

Many channels in practice are symmetric. Furthermore, it turns out that (in the absence of a cost constraint) the capacity of a symmetric DMC is

easily computed from the capacities of the sub-channels:

$$\boxed{C = \sum_{i=1}^{L} q_i \, C_i} \tag{3.75}$$

where $q_i = \sum_{y \in \mathcal{Y}_i} P(y|x)$ for any $x \in \mathcal{X}$ is the probability that sub-channel i is chosen, and C_i is the capacity of this sub-channel. Moreover, a uniform P_X achieves capacity. For example, the BEC has $q_1 = 1-p$, $q_2 = p$, $C_1 = 1$, and $C_2 = 0$.

To prove (3.75), let A be a random variable that represents the sub-channel, i.e., we have $P_A(i) = q_i$ for $i = 1, 2, \ldots, L$. We have

$$\begin{aligned} I(X;Y) &\overset{(a)}{=} I(X;YA) \\ &\overset{(b)}{=} I(X;A) + I(X;Y|A) \\ &\overset{(c)}{=} I(X;Y|A) \end{aligned} \tag{3.76}$$

where (a) is because A is a function of Y (see (1.111)), (b) follows by the chain rule for mutual information (see (1.63)), and (c) is because X and A are independent. We further have

$$\begin{aligned} I(X;Y|A) &= \sum_{i=1}^{L} q_i \, I(X_i;Y_i|A=i) \\ &\overset{(a)}{\leq} \sum_{i=1}^{L} q_i \, C_i \end{aligned} \tag{3.77}$$

where (a) follows by the definition of C_i. Finally, a uniform P_X simultaneously maximizes $I(X_i;Y_i|A_i)$ for all $i = 1, 2, \ldots, L$ so that we can achieve equality in (a).

Example 3.10. Consider an AWGN channel where Z is Gaussian with zero mean and variance N. Suppose we use *binary phase shift keying (BPSK)* with $\mathcal{X} = \{-\sqrt{P}, \sqrt{P}\}$ and a *uniform quantizer* that maps Y to the nearest value in the set $\tilde{\mathcal{Y}} = \{-L+1, -L+3, \ldots, L-3, L-1\}$. Call the resulting discrete output $\tilde{Y}$. The X-to-$\tilde{Y}$ channel is a symmetric DMC that can be decomposed into $\lceil L/2 \rceil$ strongly symmetric channels. If L is even then all sub-channels are BSCs. If L is odd then all sub-channels are BSCs except one sub-channel that maps both inputs to $\tilde{Y} = 0$. Either way, the capacity is given by (3.75).

For instance, suppose we have $\sqrt{P} \leq 2$. We choose $L = 4$ so that $\tilde{Y}$ takes on values in the set $\tilde{\mathcal{Y}} = \{-3, -1, 1, 3\}$ (see Fig. 3.14 where the quantizer boundaries are shown with dashed lines). There are two sub-channels whose

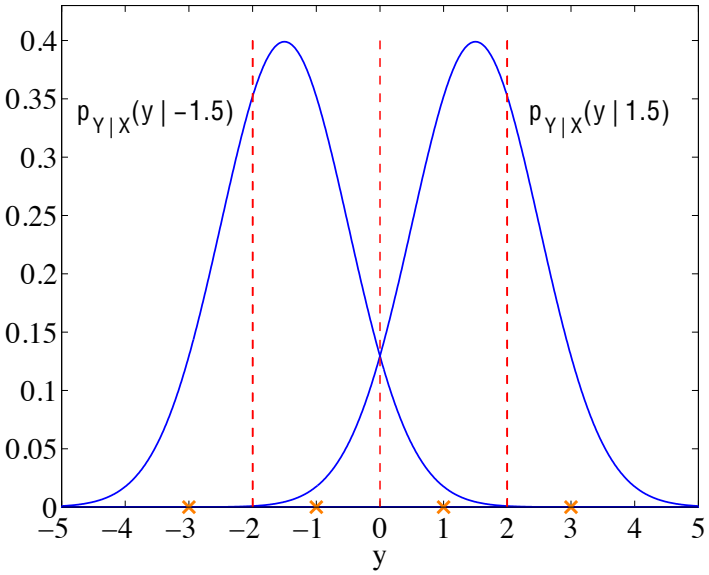

Figure 3.14.: Conditional probability densities for an AWGN channel with $N = 1$ and BPSK with $\sqrt{P} = 1.5$. The channel output is quantized to values in the set $\{-3, -1, 1, 3\}$.

probabilities and capacities are:

$$\begin{aligned} q_1 &= Q\left((2-\sqrt{P})/\sqrt{N}\right) + Q\left((2+\sqrt{P})/\sqrt{N}\right) \\ q_2 &= 1 - q_1 \\ C_1 &= 1 - H_2\left(\frac{Q\left((2+\sqrt{P})/\sqrt{N}\right)}{q_1}\right) \\ C_2 &= 1 - H_2\left(\frac{Q\left(\sqrt{P/N}\right) - Q\left((2+\sqrt{P})/\sqrt{N}\right)}{q_2}\right) \end{aligned} \tag{3.78}$$

where

$$Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}}\, e^{-y^2/2}\, dy. \tag{3.79}$$

3.8. Continuous Alphabet Examples

3.8.1. AWGN Channel

Consider the additive white Gaussian noise (AWGN) channel with

$$Y = X + Z \tag{3.80}$$

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

At this point, we have not shown that the capacity-cost function (3.50) gives a rate that we can approach for continuous-alphabet channels. However, one can check that the arguments in Sec. 3.5 extend to AWGN and other continuous-alphabet channels. We compute

$$\begin{aligned} C(P) &= \max_{P_X:\, \mathsf{E}[X^2]\le P} [h(Y) - h(Y|X)] \\ &= \left[\max_{P_X:\, \mathsf{E}[X^2]\le P} h(Y)\right] - \frac{1}{2}\log(2\pi eN) \\ &\overset{(a)}{=} \frac{1}{2}\log(2\pi e(P+N)) - \frac{1}{2}\log(2\pi eN) \end{aligned} \tag{3.81}$$

where (a) follows by

$$\mathsf{Var}[Y] = \mathsf{Var}[X] + N \le \mathsf{E}\left[X^2\right] + N \le P + N \tag{3.82}$$

and the maximum entropy result (2.44). We thus have

$$\boxed{C(P) = \frac{1}{2}\log\left(1 + \frac{P}{N}\right).} \tag{3.83}$$

Furthermore, we achieve $C(P)$ by choosing X to be Gaussian with zero-mean and variance P. Observe that $C(P)$ depends only on the *signal-to-noise ratio* (SNR) P/N. The function (3.83) is plotted in Fig. 3.15 for $N = 1$ (and therefore $P = P/N$) as the curve labeled "Gauss". We have taken the logarithm to the base 2 so that the rate units are bits per channel symbol.

3.8.2. AWGN Channel with BPSK

Consider the additive white Gaussian noise (AWGN) channel (3.80) but suppose we use BPSK with $\mathcal{X} = \{-\sqrt{P}, \sqrt{P}\}$. From Example 3.10 it is clear that we should choose P_X to be uniform even if Y is not quantized. The capacity is therefore

$$C(P) = I(X;Y) = h(Y) - \frac{1}{2}\log(2\pi eN) \tag{3.84}$$

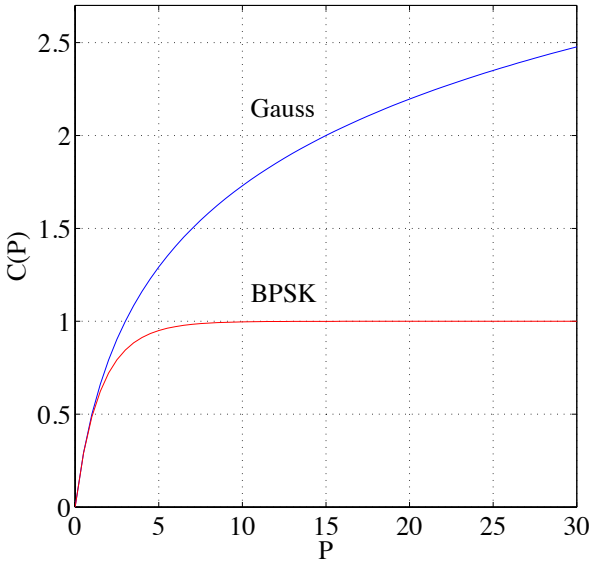

Figure 3.15.: Capacity-cost functions for an AWGN channel with $N = 1$. The rate units are bits per channel symbol.

where the density of y is

$$\begin{aligned} p(y) &= \frac{1}{2} p_{Y|X}\left(y \middle| -\sqrt{P}\right) + \frac{1}{2} p_{Y|X}\left(y \middle| \sqrt{P}\right) \\ &= \frac{1}{2}\frac{1}{\sqrt{2\pi N}} e^{-(y+\sqrt{P})^2/(2N)} + \frac{1}{2}\frac{1}{\sqrt{2\pi N}} e^{-(y-\sqrt{P})^2/(2N)}. \end{aligned} \tag{3.85}$$

The differential entropy $h(Y)$ can be computed numerically from (3.85). The function (3.84) is plotted in Fig. 3.15 as the curve labeled "BPSK". As should be expected, the capacity saturates at 1 bit per symbol for large P. A more interesting result is that BPSK almost achieves capacity for small P (say $P \leq 2$).

We may alternatively compute capacity by expanding $I(X;Y)$ differently:

$$\begin{aligned} C(P) &= H(X) - H(X|Y) \\ &= 1 - \int_{-\infty}^{\infty} p(y) H_2\left(P_{X|Y}\left(-\sqrt{P}\middle| y\right)\right) dy \end{aligned} \tag{3.86}$$

where

$$\begin{aligned} P_{X|Y}\left(-\sqrt{P}\middle| y\right) &= \frac{P_X\left(-\sqrt{P}\right)\, p_{Y|X}\left(y \middle| -\sqrt{P}\right)}{p(y)} \\ &= \frac{1}{2p(y)}\frac{1}{\sqrt{2\pi N}} e^{-(y+\sqrt{P})^2/(2N)}. \end{aligned} \tag{3.87}$$

The integral (3.86) must again be computed numerically.

3.8.3. Complex AWGN Channel

The *complex* AWGN channel has

$$Y = X + Z \tag{3.88}$$

where $X = X_R + jX_I$, $Y = Y_R + jY_I$, and $Z = Z_R + jZ_I$ are complex random variables with $j = \sqrt{-1}$. The noise variables Z_R and Z_I are Gaussian with zero mean. One usually chooses Z_R and Z_I to be independent with the same variance in which case the noise Z is said to be *proper complex* or *circularly symmetric*. We choose $\mathsf{E}[Z_R^2] = \mathsf{E}[Z_I^2] = N/2$ so that $\mathsf{E}[|Z|^2] = N$.

In communications engineering, complex AWGN channels usually model *bandpass* channels. The symbol X_R represents the sign and amplitude of the transmit component that is *in-phase* with a carrier $\sqrt{2}\cos(2\pi f_c t)$ at frequency f_c. The symbol X_I represents the sign and amplitude of the transmit component in *quadrature* with this carrier, i.e., the component that is in-phase with $\sqrt{2}\sin(2\pi f_c t)$.

The power constraint is usually taken to be $\mathsf{E}[|X|^2] = \mathsf{E}[X_R^2] + \mathsf{E}[X_I^2] \leq P$. We may view this channel as a *parallel* Gaussian channel with two sub-channels and a sum power constraint. We define $P_R = \mathsf{E}[X_R^2]$ and $P_I = \mathsf{E}[X_I^2]$ and compute

$$\begin{aligned} C(P) &= \max_{P_{X_R X_I}:\, P_R + P_I \leq P} [h(Y_R Y_I) - h(Y_R Y_I | X_R X_I)] \\ &= \left[\max_{P_{X_R X_I}:\, P_R + P_I \leq P} h(Y_R Y_I) \right] - \log(\pi e N) \\ &\overset{(a)}{\leq} \left[\max_{P_{X_R} P_{X_I}:\, P_R + P_I \leq P} h(Y_R) + h(Y_I) \right] - \log(\pi e N) \\ &\overset{(b)}{\leq} \max_{P_R + P_I \leq P} \left(\frac{1}{2}\log(1 + 2P_R/N) + \frac{1}{2}\log(1 + 2P_I/N) \right) \end{aligned} \tag{3.89}$$

with equality in (a) if X_R and X_I are independent, and with equality in (b) if X_R and X_I are Gaussian with zero mean. It is clearly best to choose $P_R + P_I = P$ and direct differentiation shows that the best choice of power allocation is $P_R = P_I = P/2$. We thus have

$$\boxed{C(P) = \log\left(1 + \frac{P}{N}\right).} \tag{3.90}$$

$C(P)$ again depends only on the SNR and the capacity is twice that of the real case (3.83) where $\mathsf{E}[X^2] \leq P$ and $\mathsf{E}[Z^2] = N$. We achieve $C(P)$ by choosing X_R and X_I to be independent Gaussian random variables with zero-mean and variance $P/2$.

The capacity-cost function $C(P)$ is clearly concave in P. However, there are two different and more common ways of plotting $C(P)$. First, one considers a *bandlimited* channel with bandwidth W Hz and transmits $2W$ real symbols per second that are represented as W complex symbols per

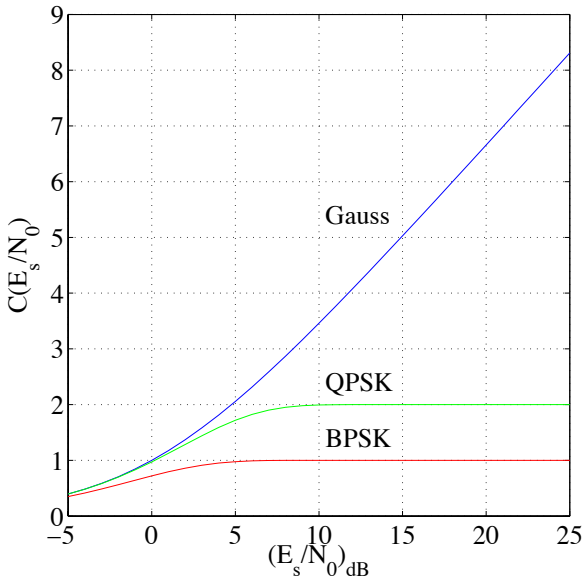

Figure 3.16.: $C(E_s/N_0)$ for an AWGN channel in terms of $(E_s/N_0)_{\rm dB}$. The rate units are bits per channel symbol.

second. The received signal is sampled at the rate W complex symbols per second. If the transmit signal has a power constraint P and the noise is modeled as AWGN with power $N = N_0 W$, where N_0 is a constant with units of Watts/Hz, then the capacity is

$$\boxed{\tilde{C}(P) = W \log_2 \left(1 + \frac{P}{N_0 W}\right) \text{ bits/second.}} \tag{3.91}$$

The tilde on $\tilde{C}(P)$ emphasizes that the units are bits per *second*. The transmit *energy* with transmit intervals of length $1/W$ seconds is $E_s = P/W$ Joules. We can also express the capacity in units of *bits per second per Hertz* (which are units of spectral efficiency) or in units of *bits per symbol*:

$$\boxed{C(E_s/N_0) = \log_2 \left(1 + \frac{E_s}{N_0}\right) \text{ bits/symbol.}} \tag{3.92}$$

The expression (3.92) is basically the same as (3.90) but now the SNR is measured in terms of an energy ratio rather than a power ratio. One usually plots $C(E_s/N_0)$ by measuring the SNR in decibels:

$$(E_s/N_0)_{\rm dB} = 10 \log_{10} E_s/N_0. \tag{3.93}$$

The resulting curve is labeled "Gauss" in Fig. 3.16 The two other curves show the capacities of BPSK and quaternary phase shift keying (QPSK), respectively, where QPSK has the *modulation* alphabet

$$\mathcal{X} = \{\sqrt{P}, \sqrt{P}\, j, -\sqrt{P}, -\sqrt{P}\, j\}. \tag{3.94}$$

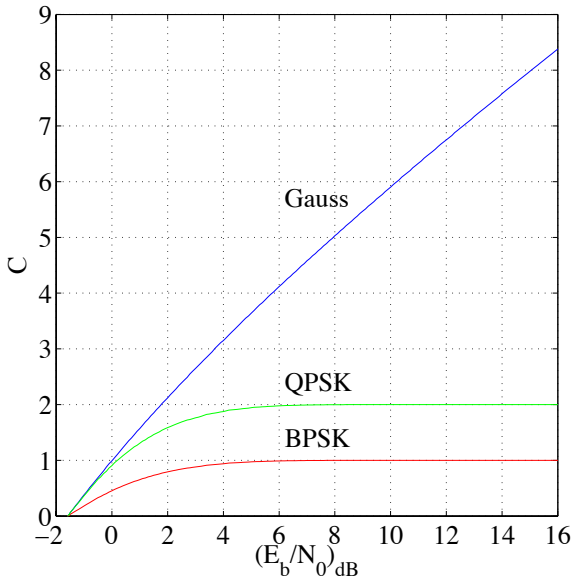

Figure 3.17.: Capacity for an AWGN channel in terms of $(E_b/N_0)_{\text{dB}}$. The rate units are bits per channel symbol.

The QPSK capacity is computed by doubling the BPSK capacity and then shifting the curve to the right by 3 dB (since the energy per dimension is $E_s/2$). Both modulations almost achieve capacity at low SNR.

Yet another way of plotting the capacity is to measure the SNR in terms of the energy per information bit: $E_b = E_s/R$ where R is the rate in bits per symbol. We require

$$R \leq \log_2\left(1 + \frac{E_s}{N_0}\right) = \log_2\left(1 + \frac{E_b}{N_0} \cdot R\right) \tag{3.95}$$

so we say that the capacity is the largest R that satisfies (3.95).[4] One usually plots this capacity against

$$(E_b/N_0)_{\text{dB}} = 10 \log_{10} E_b/N_0. \tag{3.96}$$

The resulting curve is shown in Fig. 3.17. Observe that the smallest E_b/N_0 as $R \to 0$ is $E_b/N_0 = \ln(2)$ which is

$$(E_b/N_0)_{\text{dB}} = 10 \log_{10}(\ln 2) \approx -1.6 \text{ dB}. \tag{3.97}$$

Thus, there is an ultimate minimum energy (per information bit) required to transmit reliably over an AWGN channel. Moreover, we now find that the QPSK capacity is exactly double the BPSK capacity. BPSK is poor at low SNR because it uses only 1 of the 2 complex dimensions available for each symbol. One trick to improve BPSK is to use *single sideband modulation* (SSB) to reduce the number of dimensions per symbol back to 1. Thus, BPSK with SSB achieves the same rates as QPSK.

[4] Alternatively, the smallest SNR that satisfies (3.95) is $E_b/N_0 = (2^R - 1)/R$.

3.8.4. Parallel AWGN Channels

Suppose we have L AWGN sub-channels:

$$Y_i = h_i X_i + Z_i, \quad i = 1, 2, \ldots, L, \tag{3.98}$$

where the h_i are constants and the Z_i are Gaussian random variables with zero-mean and variance N_i. The Z_i, $i = 1, 2, \ldots, L$, are mutually statistically independent of each other and of the input variables X_i, $i = 1, 2, \ldots, L$. For example, the complex Gaussian channel treated in Sec. 3.8.3 effectively has $L = 2$, $h_1 = h_2 = 1$, and $N_1 = N_2 = N/2$. We may write the overall channel in vector form as

$$\underline{Y} = \mathbf{H}\,\underline{X} + \underline{Z} \tag{3.99}$$

where $\mathbf{H}$ is a $L \times L$ diagonal matrix, and where the $\underline{X}$, $\underline{Y}$, and $\underline{Z}$ have as their ith entries the random variables X_i, Y_i, and Z_i, respectively.

If each sub-channel had its own power constraint $s_i(x_i, y_i) = x_i^2$ then the capacity is simply the sum of the capacities of the L sub-channels. The situation is more interesting if we have a sum power constraint

$$s(\underline{x}, \underline{y}) = \|\underline{x}\|^2 = \sum_{i=1}^{L} x_i^2. \tag{3.100}$$

We define $P_i = \mathsf{E}[X_i^2]$ and compute

$$\begin{aligned} C(P) &= \max_{P_{\underline{X}}:\, \mathsf{E}[\|\underline{X}\|^2] \le P} I(\underline{X}; \mathbf{H}\,\underline{X} + \underline{Z}) \\ &\overset{(a)}{=} \max_{\sum_{i=1}^{L} P_i \le P} \sum_{i=1}^{L} \frac{1}{2} \log\left(1 + \frac{h_i^2 P_i}{N_i}\right) \end{aligned} \tag{3.101}$$

where (a) follows by the maximum entropy result (2.44) (see Problem 3.4). We have thus arrived at a *power allocation* problem.

We may solve the problem by using standard optimization methods. The problem is concave in $\underline{P} = [P_1, P_2, \ldots, P_L]$, and the Lagrangian is

$$\Lambda = \left[\sum_{i=1}^{L} \frac{1}{2} \log\left(1 + \frac{h_i^2 P_i}{N_i}\right)\right] + \left[\sum_{i=1}^{L} \gamma_i P_i\right] + \lambda \left[P - \sum_{i=1}^{L} P_i\right] \tag{3.102}$$

where the γ_i, $i = 1, 2, \ldots, L$, and λ are Lagrange multipliers. The Karush-Kuhn-Tucker (KKT) conditions for the optimal solution are

$$\begin{aligned} &\gamma_i \ge 0, \quad P_i \ge 0, \quad \gamma_i \cdot P_i = 0, \quad i = 1, 2, \ldots, L \\ &\lambda \ge 0, \quad \sum_{i=1}^{L} P_i \le P, \quad \lambda \cdot \left[P - \sum_{i=1}^{L} P_i\right] = 0 \\ &\frac{\partial \Lambda}{\partial P_i} = \frac{1}{2} \frac{1}{N_i/h_i^2 + P_i} + \gamma_i - \lambda = 0, \quad i = 1, 2, \ldots, L \end{aligned} \tag{3.103}$$

Table 3.1.: Sub-channel parameters

i	1	2	3	4
h_i	2	$\sqrt{2}$	1	$\sqrt{2}$
N_i	1	1	2	2
N_i/h_i^2	1/4	1/2	2	1

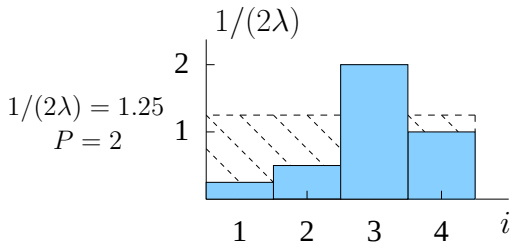

Figure 3.18.: Waterfilling for $L = 4$ sub-channels.

where we have used the natural logarithm for the derivative. The γ_i are slack variables that we may choose to achieve equality in the last L conditions above. Furthermore, we cannot choose $\lambda = 0$ so we must have $\sum_{i=1}^{L} P_i = P$. The KKT conditions thus reduce to

$$P_i \geq 0, \quad \left[\lambda - \frac{1}{2}\frac{1}{N_i/h_i^2 + P_i}\right] \cdot P_i = 0, \quad i = 1, 2, \ldots, L \tag{3.104}$$

$$\lambda > 0, \quad \sum_{i=1}^{L} P_i = P. \tag{3.105}$$

Now if $\lambda > h_i^2/(2N_i)$ or $1/(2\lambda) < N_i/h_i^2$ then (3.104) requires $P_i = 0$. We must thus choose

$$\boxed{P_i = \max\left(\frac{1}{2\lambda} - \frac{N_i}{h_i^2}, 0\right), \quad i = 1, 2, \ldots, L.} \tag{3.106}$$

This solution is often called *waterfilling* because one visualizes pouring water up to a level $1/(2\lambda)$ when the ground level is at N_i/h_i^2. One stops pouring once the water "volume" is P. The capacity is given by (3.101), namely

$$\boxed{C(P) = \sum_{i=1}^{L} \frac{1}{2} \log\left(1 + \frac{h_i^2}{N_i} \cdot P_i\right).} \tag{3.107}$$

Example 3.11. Consider $L = 4$ sub-channels with the parameters given in Table 3.1. Suppose $P = 2$ in which case we need to set $1/(2\lambda) = 5/4$ so that $P_1 = 1$, $P_2 = 3/4$, $P_3 = 0$, $P_4 = 1/4$ (see Fig. 3.18). The capacity is

$$\begin{aligned} C(2) &= \frac{1}{2}\log\left(1+4P_1\right) + \frac{1}{2}\log\left(1+2P_2\right) + \frac{1}{2}\log\left(1+P_4\right) \\ &= \frac{1}{2}\log(5) + \frac{1}{2}\log(2.5) + \frac{1}{2}\log(1.25) \\ &\approx 1.98 \text{ bits.} \end{aligned} \tag{3.108}$$

Example 3.12. Suppose the L sub-channels are *proper complex* AWGN channels, i.e., Z_i is circularly symmetric for all i. Repeating the above analysis requires only slight modifications to (3.106) and (3.107): choose λ so that $\sum_{i=1}^{L} P_i = P$ where

$$P_i = \max\left(\frac{1}{\lambda} - \frac{N_i}{|h_i|^2}, 0\right), \quad i = 1, 2, \ldots, L. \tag{3.109}$$

The resulting capacity is

$$C(P) = \sum_{i=1}^{L} \log\left(1 + \frac{|h_i|^2}{N_i} \cdot P_i\right). \tag{3.110}$$

3.8.5. Vector AWGN Channels

Consider the complex *vector* AWGN channel with $n_t \times 1$ input $\underline{X}$, $n_r \times n_t$ matrix $\mathbf{H}$, and $n_r \times 1$ output

$$\underline{Y} = \mathbf{H}\,\underline{X} + \underline{Z} \tag{3.111}$$

where $\underline{Z}$ is a $n_r \times 1$ Gaussian vector with independent and identically distributed (i.i.d.) proper complex entries of unit variance. This problem is also known as a multi-antenna, or multi-input, multi-output (MIMO) AWGN channel. We choose the cost function $s(\underline{x}, \underline{y}) = \|\underline{x}\|^2$ and compute

$$\begin{aligned} C(P) &= \max_{P_{\underline{X}}\colon \mathsf{E}[\|\underline{X}\|^2] \le P} I(\underline{X}; \mathbf{H}\,\underline{X} + \underline{Z}) \\ &= \left[\max_{P_{\underline{X}}\colon \mathsf{E}[\|\underline{X}\|^2] \le P} h(\mathbf{H}\,\underline{X} + \underline{Z}) \right] - n_r \log(\pi e) \\ &\overset{(a)}{=} \max_{\mathrm{tr}\left(\mathbf{Q}_{\underline{X}}\right) \le P} \log \left| \mathbf{I} + \mathbf{H}\mathbf{Q}_{\underline{X}}\mathbf{H}^\dagger \right| \end{aligned} \tag{3.112}$$

where (a) follows by the maximum entropy result (2.44). Suppose $\mathbf{H}$ has the singular-value decomposition $\mathbf{H} = \mathbf{U}\mathbf{D}\mathbf{V}^\dagger$ where $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices (with $\mathbf{U}\mathbf{U}^\dagger = \mathbf{I}$ and $\mathbf{V}\mathbf{V}^\dagger = \mathbf{I}$) and where $\mathbf{D}$ is a real, diagonal $n_r \times n_t$ matrix with the *singular values* of $\mathbf{H}$ on the diagonal. We write (3.112) as

$$\begin{aligned} C(P) &= \max_{\mathrm{tr}\left(\mathbf{Q}_{\underline{X}}\right) \le P} \log \left| \mathbf{I} + \mathbf{D}\mathbf{Q}_{\underline{X}}\mathbf{D}^T \right| \\ &\overset{(a)}{=} \max_{\sum_{i=1}^{\min(n_t,n_r)} P_i \le P} \sum_{i=1}^{\min(n_t,n_r)} \log\left(1 + d_i^2 P_i\right) \end{aligned} \tag{3.113}$$

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 for matrices for (a). The remaining optimization problem is the same as for parallel Gaussian channels with $N_i = 1$ for all i. The waterfilling solution is to choose the water level $\tilde{\lambda}$ so that $\sum_{i=1}^{\min(n_t,n_r)} P_i = P$ where

$$P_i = \max\left(\tilde{\lambda} - \frac{1}{d_i^2}, 0\right). \tag{3.114}$$

The capacity is

$$C(P) = \sum_{i=1}^{\min(n_t,n_r)} \log\left(1 + d_i^2 \cdot P_i\right). \tag{3.115}$$

3.8.6. AWGN Channels with Receiver Channel Information

Consider the following complex channel with a vector output:

$$Y = [H\,X + Z,\ H] \tag{3.116}$$

where Z is proper complex Gaussian, and H is a random variable with density p_H 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 $s(x, y) = |x|^2$ and compute

$$\begin{aligned} C(P) &= \max_{P_X:\,\mathsf{E}[|X|^2]\le P} I(X;[H\,X+Z,\ H]) \\ &= \max_{P_X:\,\mathsf{E}[|X|^2]\le P} I(X;H) + I(X;H\,X+Z|H) \\ &= \max_{P_X:\,\mathsf{E}[|X^2]\le P} I(X;H\,X+Z|H) \\ &= \max_{P_X:\,\mathsf{E}[|X|^2]\le P} \int_a p_H(a)\,h(a\,X+Z)\,da - \log(\pi e N) \\ &= \int_a p_H(a)\cdot\log(1+|a|^2P/N)\,da \end{aligned} \tag{3.117}$$

where the last step follows by the maximum entropy result (2.44), i.e., a Gaussian X with zero mean and variance P simultaneously maximizes $h(aX + Z)$ for all values of a.

If the fading coefficient H does not have a density then the analysis hardly changes. For example, suppose that H is a discrete random variable with: $P_H(1/2) = 1/4$, $P_H(1) = 1/2$, and $P_H(2) = 1/4$. The capacity is

$$C(P) = \frac{1}{4}\log\left(1+\frac{P}{4N}\right) + \frac{1}{2}\log\left(1+\frac{P}{N}\right) + \frac{1}{4}\log\left(1+\frac{4P}{N}\right). \tag{3.118}$$

We remark that QPSK may perform poorly for fading channels because the information rate $I(X; aX + Z)$ saturates at 2 bits per symbol even if the fading coefficient $H = a$ has a large amplitude. Thus, for fading channels it can be important to use a large modulation set $\mathcal{X}$ to approach the capacity (3.117) that is achieved with a Gaussian distribution.

3.9. Source and Channel Coding

3.9.1. Separate Coding

Suppose we wish to communicate the output of a DMS P_U with entropy $H(U)$ across a DMC with capacity-cost function $C(S)$. A natural approach is to *separate* source and channel coding: compress the DMS output to a rate close to $H(U)$ bits per source symbol and then communicate these bits across the DMC reliably at a rate close to $C(S)$ bits per channel symbol. The overall rate is $C(S)/H(U)$ source symbols per channel symbol.

We wish to prove that other techniques cannot do better. Suppose the source puts out m symbols U^m and we communicate over n channel uses. A simple modification of Fano's inequality (3.21) gives

$$\begin{aligned} H(U^m|\hat{U}^m) &\leq H_2(P_e) + P_e \log_2(|\mathcal{U}|^m - 1) \\ &< H_2(P_e) + P_e\, m \log_2 |\mathcal{U}|. \end{aligned} \tag{3.119}$$

We also have $H(U^m|\hat{U}^m) = H(U^m) - I(U^m;\hat{U}^m)$ and $H(U^m) = mH(U)$ so that

$$\begin{aligned} m &< \frac{I(U^m;\hat{U}^m) + H_2(P_e)}{H(U) - P_e \log_2 |\mathcal{U}|} \\ &\overset{(a)}{\leq} \frac{I(X^n;Y^n) + H_2(P_e)}{H(U) - P_e \log_2 |\mathcal{U}|} \\ &\overset{(b)}{\leq} \frac{nC(S) + H_2(P_e)}{H(U) - P_e \log_2 |\mathcal{U}|} \end{aligned} \tag{3.120}$$

where (a) follows by the data processing inequality and (b) by (3.60)-(3.62). Now if $P_e \to 0$ then we have the desired bound

$$\frac{m}{n} \leq \frac{C(S)}{H(U)} \tag{3.121}$$

source symbols per channel symbol. Thus, separate source and channel coding achieves the best possible performance (if we can permit $n \to \infty$).

3.9.2. Rates Beyond Capacity

Suppose we permit an average bit error probability

$$P_b = \frac{1}{m} \sum_{i=1}^{m} \Pr\left[\hat{U}_i \neq U_i\right]. \tag{3.122}$$

It turns out that there are source codes that can compress coin-tossing bits U^m to $m(1 - H_2(P_b))$ bits from which we can recover a string $\hat{U}^m$ that has an average bit error probability P_b. But we can send the $m(1 - H_2(P_b))$ bits reliably (with average bit error probability near zero) over the channel

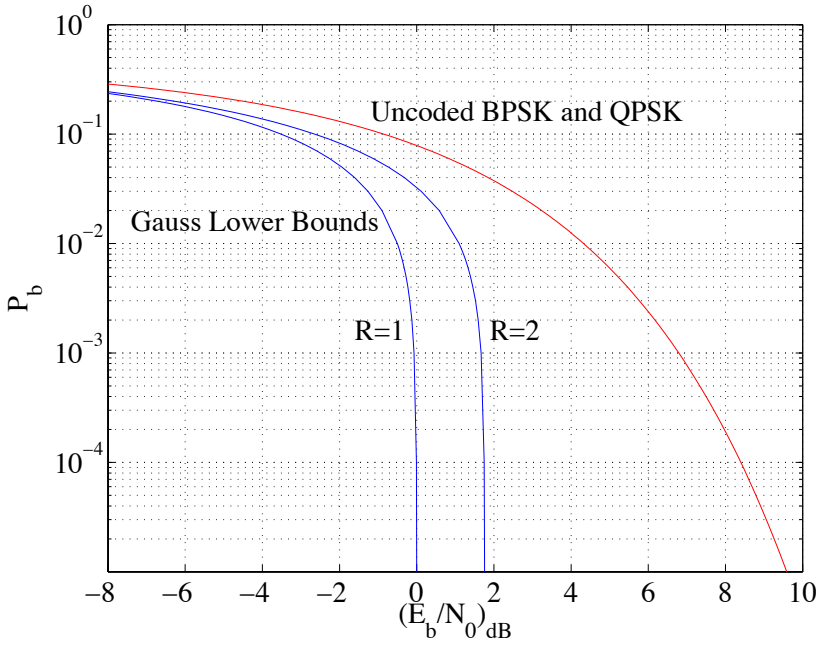

Figure 3.19.: Average bit error probability P_b as a function of $(E_b/N_0)_{\text{dB}}$ for a complex AWGN channel. The blue curves are lower bounds on the best possible P_b given R.

in approximately $n = m(1 - H_2(P_b))/C(S)$ channel uses. The overall rate $R = m/n$ is

$$R = \frac{C(S)}{1 - H_2(P_{\text{b}})} \tag{3.123}$$

so that we can approach equality in (3.64). For instance, if we permit $P_b = 0.11$ then we can approach $R = 2C(S)$ as closely as desired. Moreover, we can accomplish this task by *separate* source coding and channel coding.

We further explore the impact of this result. Consider a complex AWGN channel and suppose we wish to communicate R bits per symbol. We use (3.64), (3.92), and $E_s = E_b R$ to compute

$$\boxed{\frac{E_b}{N_0} \geq \frac{1}{R}\left(2^{R(1-H_2(P_b))} - 1\right).} \tag{3.124}$$

As argued above, we can approach equality in (3.124) by using separate source and channel coding. For example, for $P_b = 0$ and $R = 1$ we compute $E_b/N_0 = 1$ or $(E_b/N_0)_{\text{dB}} = 0$ dB. The minimum $(E_b/N_0)_{\text{dB}}$ for positive P_b and $R = 1$ and $R = 2$ are plotted in Fig. 3.19 as the curves labeled "Gauss Lower Bounds". For instance, if we permit $P_b \approx 0.1$ then for $R = 1$ we can save over 3 dB, or over half, the energy as compared to $P_b \to 0$. Alternatively, for a fixed E_b/N_0, if we increase P_b we may increase R.

Now suppose we transmit using BPSK without coding. We have $R = 1$ and

$$P_b = Q\left(\sqrt{\frac{P}{N/2}}\right) = Q\left(\sqrt{2\frac{E_b}{N_0}}\right) \tag{3.125}$$

since $E_b = E_s = P/W$ and $N = N_0 W$. Similarly, QPSK has $R = 2$ and

$$P_b = Q\left(\sqrt{\frac{P/2}{N/2}}\right) = Q\left(\sqrt{2\frac{E_b}{N_0}}\right) \tag{3.126}$$

since $E_b = E_s/2 = P/(2W)$. Thus, QPSK achieves the same P_b as BPSK but at double the rate. The resulting P_b are shown in Fig. 3.19 as the curve labeled "Uncoded BPSK and QPSK". For instance, uncoded BPSK and QPSK require $(E_b/N_0)_{\text{dB}} \approx 9.6$ dB for $P_b = 10^{-5}$, while coding and Gaussian modulation require only $(E_b/N_0)_{\text{dB}} = 0$ dB and $(E_b/N_0)_{\text{dB}} \approx 1.8$ dB for $R = 1$ and $R = 2$, respectively. The potential energy savings of coding and modulation are therefore up to 9.6 dB, or over a factor of 9, as compared to uncoded BPSK, and up to 7.8 dB, or over a factor of 6, as compared to uncoded QPSK.

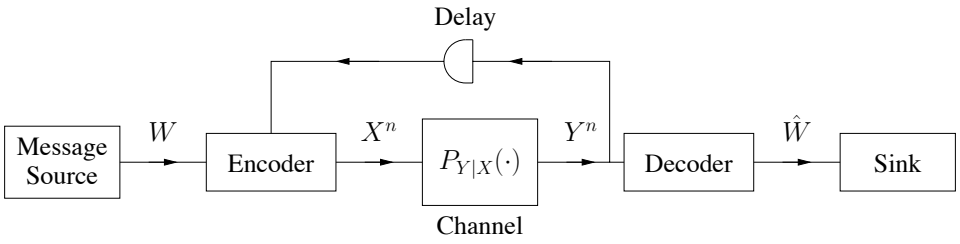

Figure 3.20.: The capacity-cost problem with feedback.

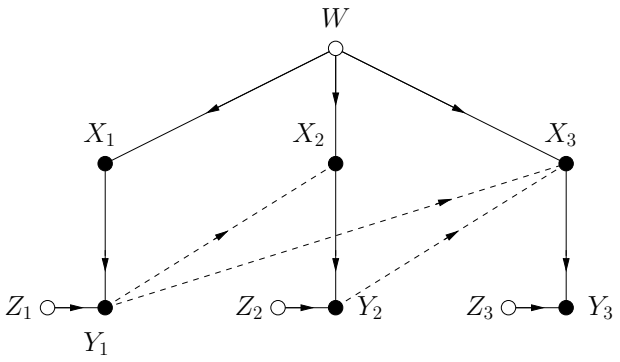

Figure 3.21.: FDG for a memoryless channel with feedback. The message estimate $\hat{W}$ is not shown.

3.10. Feedback

Feedback is usually used in communications to improve performance. One might expect that feedback can increase capacity. To check this, consider a memoryless channel *with feedback* in the sense that X_i can be a function of the message W and a function of the *past* channel outputs Y^{i-1}. The most informative feedback would thus be that the transmitter is aware of Y^{i-1}, as shown in Fig. 3.20 and Fig. 3.21 (the latter figure is also in Sec. A.5). We slightly modify (3.60) and bound

$$\begin{aligned} I(W;\hat{W}) &\le I(W;Y^n) \\ &= \sum_{i=1}^{n} H(Y_i|Y^{i-1}) - H(Y_i|WY^{i-1}) \\ &\overset{(a)}{=} \sum_{i=1}^{n} H(Y_i|Y^{i-1}) - H(Y_i|WY^{i-1}X_i) \\ &\overset{(b)}{=} \sum_{i=1}^{n} H(Y_i|Y^{i-1}) - H(Y_i|X_i) \\ &\le \sum_{i=1}^{n} I(X_i;Y_i) \end{aligned} \tag{3.127}$$

where (a) follows because X_i is a function of W and Y^{i-1}, and (b) is because the channel is memoryless. We have thus arrived at (3.60) and find the surprising result that feedback does *not* improve the capacity-cost function of a discrete memoryless channel [1, 2]. However, we emphasize that feed-

back can help to reduce complexity and improve reliability, as the following example shows. Feedback can also increase the capacity of channels with memory and the capacity of multi-user channels.

Example 3.13. Consider the BEC with feedback of the channel outputs. The encoder can use a *variable-length* encoding scheme where each message bit is repeated until the decoder receives this bit without erasure. The number N of times that each bit must be transmitted has the geometric distribution

$$P_N(k) = (1-p)p^{k-1}, \quad k = 1, 2, 3, \ldots \tag{3.128}$$

whose mean is $\mathsf{E}[N] = 1/(1-p)$. The average rate $\overline{R}$ is the number of information bits transmitted divided by the average number of trials, giving $\overline{R} = 1 - p = C$. The benefit of feedback is therefore not in increasing capacity but in giving a simple (variable-length) coding method that achieves capacity. In fact, the error probability is zero if one permits infinite delay.

3.11. Problems

3.1. Block and Bit Error Probability

Prove the relations (3.20).
Hint: Observe that $\{\hat{W} \neq W\} = \cup_{i=1}^{b} \{\hat{V}_i \neq V_i\}$. Now use the union bound to upper bound P_e. Finally, try to upper bound $1 - P_e$.

3.2. BSC with Output Cost

a) Compute the capacity-cost function $C(S)$ for the binary symmetric channel (BSC) with alphabets $\mathcal{X} = \mathcal{Y} = \{0, 1\}$, crossover probability $\Pr[Y \neq X] = p$, and the cost function $s(x, y) = 2y$.

b) Plot $C(S)$ as a function of S for $p = 1/3$. Is $C(S)$ concave? Explain what happens when $S \in [0, 2/3]$.

3.3. Z-Channel with Sum Cost

a) Compute the capacity-cost function $C(S)$ for the binary "Z-channel" with

$$\begin{array}{ll} P_{Y|X}(0|0) = 1, & P_{Y|X}(1|0) = 0 \\ P_{Y|X}(0|1) = 1/2, & P_{Y|X}(1|1) = 1/2 \end{array} \tag{3.129}$$

and the cost constraint $\mathsf{E}[X + Y] \leq S$.

b) Plot $C(S)$ as a function of S.

3.4. Parallel Channel Power Allocation

Prove step (a) in (3.101).

3.5. Gallager Exponent

Prove the three properties of $E_G(R, P_X)$ given after (3.44).
Hint: Use the result:

$$\frac{d}{dx} f(x)^{g(x)} = \frac{d}{dx} e^{g(x) \ln f(x)} = \left[\frac{g(x)}{f(x)} \frac{df(x)}{dx} + \frac{dg(x)}{dx} \ln f(x) \right] f(x)^{g(x)}.$$

3.6. Strong Converse

Consider a code book $\underline{x}^n$ as in (3.28). The ML decoder chooses as its estimate $\hat{w} = i$ one of the messages that maximizes $P_{Y^n|X^n}(y^n|x^n(i))$. The ML decoder thus partitions the set of channel output sequences y^n into M disjoint *decoding regions*.

a) Show that the error probability of an ML decoder for any code with M equally-likely codewords is given by

$$\Pr[\mathcal{E}] = 1 - \sum_{y^n} \frac{1}{M} \max_i P_{Y^n|X^n}(y^n|x^n(i)). \tag{3.130}$$

b) Now use $a = \{a^s\}^{1/s}$ to show that for any $s > 0$ we have

$$\Pr[\mathcal{E}] \geq 1 - \sum_{y^n} \frac{1}{M} \left\{ \sum_{i=1}^{M} P_{Y^n|X^n}(y^n|x^n(i))^s \right\}^{1/s}. \tag{3.131}$$

c) Argue that $\Pr[\mathcal{E}]$ does not change by permuting the indexes $i = 1, 2, \ldots, M$. Take the expectation of $\Pr[\mathcal{E}]$ over all permutations, each with probability $1/M!$, to show that for $s \geq 1$ we have

$$\begin{aligned} \Pr[\mathcal{E}] &\geq 1 - \sum_{y^n} \frac{1}{M} \left\{ \sum_{i=1}^{M} \mathsf{E}\left[P_{Y^n|X^n}(Y^n|X^n(i))^s \right] \right\}^{1/s} \\ &\geq 1 - M^{(1-s)/s} \max_{P_{X^n}} \sum_{y^n} \left\{ \sum_{x^n} P_{X^n}(x^n) P_{Y^n|X^n}(y^n|x^n)^s \right\}^{1/s}. \end{aligned} \tag{3.132}$$

d) Since the channel is memoryless, show that (3.132) reduces to

$$\begin{aligned} \Pr[\mathcal{E}] &\geq 1 - M^{(1-s)/s} \max_{P_X} \left[\sum_{y} \left\{ \sum_{x} P_X(x) P(y|x)^s \right\}^{1/s} \right]^n \\ &\overset{(a)}{=} 1 - 2^{-n\left[-\rho R + \min_{P_X} E_0(\rho, P_X)\right]} \end{aligned} \tag{3.133}$$

where (a) follows by choosing $s = 1/(1+\rho)$ for $-1 \leq \rho < 0$ and defining $\min_{P_X} E_0(-1, P_X) = \lim_{\rho \to -1} \min_{P_X} E_0(\rho, P_X)$.

e) Define

$$E(R) = \max_{-1 \leq \rho < 0} \left[-\rho R + \min_{P_X} E_0(\rho, P_X) \right] \tag{3.134}$$

Show that if $R > C$ then we have $E(R) > 0$. In other words, for large n the block error probability approaches 1. This result is known as a *strong* converse because it shows that the block error probability must approach 1 as n grows.

3.7. Frame Error Rate

A *frame* is a block $\underline{V} = [V_1, V_2, \ldots, V_b]$ of b bits. We map each frame into a block X^n of n channel symbols. Suppose we send L blocks $\underline{V}_i$, $i = 1, 2, \ldots, L$, so the total number of bits and channel symbols is $B = L\,b$ and $N = L\,n$, respectively. The rate is $R = B/N = b/n$ and the *frame*

error rate (FER) is

$$P_F = \frac{1}{L}\sum_{i=1}^{L} \Pr\left[\underline{\hat{V}}_i \neq \underline{V}_i\right]. \tag{3.135}$$

a) Suppose we permit coding across frames, i.e., X^N is a function of $W = \underline{V}^L$. Use Fano's inequality as in Sec. 3.4 to show that

$$NR \leq \frac{I(W;\hat{W})}{1 - H_2(P_F)/b - P_F \log_2(2^b - 1)/b}. \tag{3.136}$$

We recover (3.25) for $b = 1$, as should be expected. Show that for large b we essentially have

$$R \lesssim \frac{I(W;\hat{W})/N}{1 - P_F} \leq \frac{C(S)}{1 - P_F}. \tag{3.137}$$

b) Show that on can approach equality in (3.137) by discarding a fraction P_F of the frames and transmitting the remaining fraction $1 - P_F$ of the frames reliably. Is coding across frames required?

c) The above shows that one may transmit at rates above capacity without having P_F approach one. How does this result relate to the strong converse of Problem 3.6? Which result is relevant in practice?

References

[1] C. E. Shannon. The zero error capacity of a noisy channel. *IRE Trans. Inf. Theory*, 2:221–238, September 1956. Reprinted in *Claude Elwood Shannon: Collected Papers*, pp. 221-238, (N.J.A. Sloane and A.D. Wyner, eds.) Piscataway: IEEE Press, 1993.

[2] R. L. Dobrushin. Information transmission in a channel with feedback. *Theory Prob. Appl.*, 34:367–383, December 1958.

Chapter 4.

Typical Sequences and Sets

4.1. Typical Sequences

Shannon considered "typical sequences" in his 1948 paper [1]. 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 Fig. 4.1). Suppose the source probability distribution is $P_X(\cdot)$ where

$$P_X(0) = 2/3 \quad \text{and} \quad P_X(1) = 1/3. \tag{4.1}$$

Consider the following experiment: we generated a sequence of length 18 by using a random number generator with the distribution (4.1). We write this sequence below along with three other sequences that we generated artificially.

$$\begin{array}{ll} (a) & 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0 \\ (b) & 1,0,1,1,0,1,0,1,1,1,0,0,0,0,1,0,1,0 \\ (c) & 0,0,0,1,1,0,0,1,0,0,1,1,0,0,0,1,1,0 \\ (d) & 1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1. \end{array} \tag{4.2}$$

If we compute the probabilities that these sequences were emitted by the source (4.1), we have

$$\begin{array}{ll} (a) & (2/3)^{18} \cdot (1/3)^0 \approx 6.77 \cdot 10^{-4} \\ (b) & (2/3)^{9} \cdot (1/3)^9 \approx 1.32 \cdot 10^{-6} \\ (c) & (2/3)^{11} \cdot (1/3)^7 \approx 5.29 \cdot 10^{-6} \\ (d) & (2/3)^{0} \cdot (1/3)^{18} \approx 2.58 \cdot 10^{-9}. \end{array} \tag{4.3}$$

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

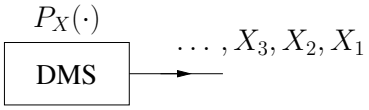

Figure 4.1.: A discrete memoryless source with distribution $P_X(\cdot)$.

(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.

4.2. 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 $a \in \mathcal{X}$. For instance, the sequence (c) in (4.2) has $N(0|x^n) = 11$ and $N(1|x^n) = 7$.

There are several natural definitions for typical sequences. Shannon in [1, §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

$$P_X^n(x^n) = \prod_{i=1}^n P_X(x_i). \tag{4.4}$$

We have $P_X^n(x^n) = 0$ if $N(a|x^n) > 0$ for some $a \notin \operatorname{supp}(P_X)$ and otherwise

$$P_X^n(x^n) = \prod_{a \in \operatorname{supp}(P_X)} P_X(a)^{N(a|x^n)} \tag{4.5}$$

Intuitively (see also Example A.15), we know that $N(a|x^n) \approx nP_X(a)$ so that $P_X^n(x^n) \approx \Pi_{a \in \operatorname{supp}(P_X)} P_X(a)^{nP_X(a)}$ or

$$-\frac{1}{n} \log_2 P_X^n(x^n) \approx H(X).$$

We make this intuition precise by observing that

$$-\frac{1}{n} \log_2 P_X^n(X^n) = \frac{1}{n} \sum_{i=1}^n Y_i \tag{4.6}$$

where $Y_i = -\log_2 P_X(X_i)$. We compute $\mathsf{E}[Y_i] = H(X)$ and note that $\mathsf{Var}[Y_i]$ is a finite number, let's call it σ^2. The quantitative version of the weak law of large numbers (A.78) gives

$$\Pr\left[\left|-\frac{1}{n} \log_2 P_X^n(X^n) - H(X)\right| < \epsilon\right] \geq 1 - \frac{\sigma^2}{n\epsilon^2} \tag{4.7}$$

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

$$\left|-\frac{1}{n} \log_2 P_X^n(x^n) - H(X)\right| < \epsilon \tag{4.8}$$

for some small positive ϵ. Alternatively, we can write

$$H(X) - \epsilon < -\frac{1}{n} \log_2 P_X^n(x^n) < H(X) + \epsilon \tag{4.9}$$

The sequences satisfying (4.8) or (4.9) are sometimes called *weakly* typical sequences or *entropy*-typical sequences [2, p. 40].

Example 4.1. Consider a binary source such as (4.1) but with $P_X(0) = p$ and $P_X(1) = 1 - p$. The test (4.8) can be written as

$$\begin{aligned} \epsilon &> \left| \frac{-N(0|x^n) \log_2 p - N(1|x^n) \log_2(1-p)}{n} - H_2(p) \right| \\ &= \left| \left(\frac{N(0|x^n)}{n} - p \right) \log_2 \frac{1-p}{p} \right|. \end{aligned} \tag{4.10}$$

where we have used $N(1|x^n) = n - N(0|x^n)$. For instance, for $p = 2/3$ as in (4.1) the bound (4.10) is

$$\left| \frac{N(0|x^n)}{n} - \frac{2}{3} \right| < \epsilon. \tag{4.11}$$

For example, if $n = 18$ and $\epsilon = 1/9$ then only x^{18} with $N(0|x^{18}) = 11, 12, 13$ are entropy-typical. If instead $\epsilon = 1/18$ then only x^{18} with $N(0|x^{18}) = 12$ are entropy-typical.

Example 4.2. The source (4.1) has $H(X) \approx 0.9183$ and the four sequences in (4.2) are entropy-typical with respect to ϵ and $P_X(\cdot)$ if (4.11) is satisfied with $n = 18$. The required values of ϵ are as follows:

$$\begin{array}{lllll} (a) & N(0|x^{18}) = 18 & \Rightarrow & \epsilon > 1/3 \\ (b) & N(0|x^{18}) = 9 & \Rightarrow & \epsilon > 1/6 \\ (c) & N(0|x^{18}) = 11 & \Rightarrow & \epsilon > 1/18 \\ (d) & N(0|x^{18}) = 0 & \Rightarrow & \epsilon > 2/3. \end{array} \tag{4.12}$$

Note that sequence (c) permits the smallest ϵ. The ϵ values required for other values of $N(0|x^{18})$ can be inferred from the last column of Table 4.1. Observe that if $N(0|x^{18}) = 12$ then we may choose ϵ as close to zero as desired (but not exactly zero).

Example 4.3. If $P_X(\cdot)$ is uniform then for any x^n we have

$$P_X^n(x^n) = |\mathcal{X}|^{-n} = 2^{-n \log_2 |\mathcal{X}|} = 2^{-nH(X)} \tag{4.13}$$

and *all* sequences in $\mathcal{X}^n$ are entropy-typical. For example, one can check that for $p = 1/2$ the condition (4.10) is always satisfied for positive ϵ.

Let $\tilde{T}_\epsilon^n(P_X)$ be the set of entropy-typical sequences. The following theorem describes some of the most important properties of these sequences.

Theorem 4.1. Suppose $\epsilon > 0$, $x^n \in \tilde{T}_\epsilon^n(P_X)$, and X^n is emitted by the DMS $P_X(\cdot)$. Let $\sigma^2 = \mathsf{Var}[-\log_2 P_X(X)]$. We have

$$2^{-n(H(X)+\epsilon)} < P_X^n(x^n) < 2^{-n(H(X)-\epsilon)} \tag{4.14}$$

$$\left(1 - \frac{\sigma^2}{n\epsilon^2}\right) 2^{n(H(X)-\epsilon)} < |\tilde{T}_\epsilon^n(P_X)| < 2^{n(H(X)+\epsilon)} \tag{4.15}$$

$$1 - \frac{\sigma^2}{n\epsilon^2} \leq \Pr\left[X^n \in \tilde{T}_\epsilon^n(P_X)\right] \leq 1. \tag{4.16}$$

Proof. The expression (4.14) is the same as (4.8), and the left-hand side of (4.16) is simply (4.7). For (4.15) observe that

$$\begin{aligned}\Pr\left[X^n \in \tilde{T}_\epsilon^n(P_X)\right] &= \sum_{x^n \in \tilde{T}_\epsilon^n(P_X)} P_X^n(x^n) \\ &< |\tilde{T}_\epsilon^n(P_X)|\, 2^{-n(H(X)-\epsilon)}\end{aligned} \tag{4.17}$$

where the inequality follows by (4.14). Using (4.16) we thus have

$$|\tilde{T}_\epsilon^n(P_X)| > \left(1 - \frac{\sigma^2}{n\epsilon^2}\right) 2^{n(H(X)-\epsilon)}. \tag{4.18}$$

We may similarly derive the right-hand side of (4.15). □

Theorem 4.1 gives *quantitative* bounds on typical sequences and sets. For example, the source (4.1) has $H(X) \approx 0.9183$ and $\sigma^2 = 2/11$, and for $n = 18$ and $\epsilon = 1/6$ the bounds of Theorem 4.1 are

$$0.00000132 < P_X^n(x^n) < 0.0000846 \tag{4.19}$$

$$6,568 < |\tilde{T}_\epsilon^n(P_X)| < 756,681 \tag{4.20}$$

$$5/11 \leq \Pr\left[X^n \in \tilde{T}_\epsilon^n(P_X)\right] \leq 1. \tag{4.21}$$

However, it is usually easier to remember the *qualitative* statements:

$$P_X^n(x^n) \approx 2^{-nH(X)} \tag{4.22}$$

$$|\tilde{T}_\epsilon^n(P_X)| \approx 2^{nH(X)} \tag{4.23}$$

$$\Pr\left[X^n \in \tilde{T}_\epsilon^n(P_X)\right] \approx 1. \tag{4.24}$$

Of course, these qualitative statements should be used to guide intuition only; they are no substitutes for the precise quantitative bounds.

Finally, we remark that *entropy* typicality applies to *continuous* random variables with a density with finite variance if we replace the probability $P_X^n(x^n)$ in (4.8) with the density value $p_X^n(x^n)$, and if we replace the entropy $H(X)$ with the differential entropy $h(X)$ treated in Chapter 2. In contrast, the next definition can be used only for discrete random variables.

Table 4.1.: Example with $k = N(0|x^n)$, $n = 18$, $p = 2/3$.

Seq.	k	$\binom{n}{k}$	$\approx p^k(1-p)^{n-k}$	$\approx \binom{n}{k}p^k(1-p)^{n-k}$	$-\frac{1}{n}\log_2 P_X^n(x^n) - H(X)$
(d)	0	1	0.00000000258	0.00000000258	2/3
	1	18	0.00000000516	0.0000000929	11/18
	2	153	0.0000000103	0.00000158	5/11
	3	816	0.0000000206	0.0000169	1/2
	4	3,060	0.0000000413	0.000126	4/11
	5	8,568	0.0000000826	0.000708	7/18
	6	18,564	0.000000165	0.00307	1/3
	7	31,824	0.000000330	0.0105	5/18
	8	43,758	0.000000661	0.0289	2/11
(b)	9	48,620	0.00000132	0.0643	1/6
	10	43,758	0.00000264	0.116	1/11
(c)	11	31,824	0.00000529	0.168	1/18
$*$	12	18,564	0.0000106	**0.196**	**0**
	13	8,568	0.0000211	0.181	-1/18
	14	3,060	0.0000432	0.129	-1/11
	15	816	0.0000846	0.0690	-1/6
	16	153	0.000169	0.0259	-2/11
	17	18	0.000338	0.00609	-5/18
(a)	18	1	**0.000677**	0.000677	-1/3

4.3. Letter-Typical Sequences

A perhaps more natural definition for *discrete* random variables than (4.8) is the following. For $\epsilon \geq 0$, we say that a sequence x^n is *ϵ-letter* typical with respect to ϵ and $P_X(\cdot)$ if the *empirical* probability distribution $N(\cdot|x^n)/n$ is close to $P_X(\cdot)$. More precisely, we require

$$\left|\frac{N(a|x^n)}{n} - P_X(a)\right| \leq \epsilon \cdot P_X(a) \quad \text{for all } a \in \mathcal{X} \tag{4.25}$$

or, alternatively, we require that for all $a \in \mathcal{X}$ we have

$$(1-\epsilon) \cdot P_X(a) \leq \frac{N(a|x^n)}{n} \leq (1+\epsilon) \cdot P_X(a). \tag{4.26}$$

The set of x^n satisfying (4.25) is called the *ϵ-letter-typical set* with respect to ϵ and $P_X(\cdot)$ and is denoted $T_\epsilon^n(P_X)$.

Example 4.4. Consider again a binary source with $P_X(0) = 1-p$ and $P_X(1) = p$. The two tests (4.25) are

$$\left|\frac{N(1|x^n)}{n} - p\right| \leq \epsilon \cdot \min\{1-p, p\} \tag{4.27}$$

where we have used $N(0|x^n) = n - N(1|x^n)$. The rule (4.27) looks similar to (4.10). For instance, for $p = 1/3$ as in (4.1) the sequences in (4.2) are letter-typical with respect to ϵ and $P_X(\cdot)$ if

$$\begin{array}{ll} (a) & \epsilon \geq 1 \\ (b) & \epsilon \geq 1/2 \\ (c) & \epsilon \geq 1/6 \\ (d) & \epsilon \geq 2. \end{array} \tag{4.28}$$

The numbers are thus increased by a factor of 3 as compared to (4.12).

Example 4.5. If $P_X(\cdot)$ is uniform then ϵ-letter typical x^n satisfy

$$\frac{(1-\epsilon)n}{|\mathcal{X}|} \leq N(a|x^n) \leq \frac{(1+\epsilon)n}{|\mathcal{X}|}, \quad \text{for all } a \in \mathcal{X}. \tag{4.29}$$

But if $\epsilon < |\mathcal{X}| - 1$, as is usually the case, then *not* all x^n are letter-typical. The definition (4.25) is then more restrictive than (4.8) (see Example 4.3).

Example 4.6. If $P_X(a) = 0$ then (4.25) requires $N(a|x^n) = 0$.

Example 4.7. If $P_X(a) > 0$ but $N(a|x^n) = 0$ then (4.25) requires

$$P_X(a) \leq \epsilon \cdot P_X(a). \tag{4.30}$$

Typical sequences with $\epsilon < 1$ must therefore have *all* non-zero probability letters appearing *at least once* in x^n.

We will rely on letter typicality for discrete random variables and entropy typicality for continuous random variables. We remark that one often finds small variations of the conditions (4.25). For example, for *strongly* typical sequences one replaces the $\epsilon P_X(a)$ on the right-hand side of (4.25) with ϵ or $\epsilon/|\mathcal{X}|$ (see [2, p. 33] and [3, p. 288 and p. 358]). One further adds the condition that $N(a|x^n) = 0$ if $P_X(a) = 0$ so that typical sequences cannot have zero-probability letters. Observe that this condition is already included in (4.25) (see Example 4.6). Letter-typical sequences are simply called "typical sequences" in [4] and "robustly typical sequences" in [5]. 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 next develop a counterpart to Theorem 4.1 but with a bound on the probability that sequences are typical that is exponential in n. Let $\mu_X = \min_{a \in \mathrm{supp}(P_X)} P_X(a)$ and define

$$\delta_\epsilon(P_X, n) = 2|\mathcal{X}| \cdot e^{-2n\epsilon^2 \mu_X^2}. \tag{4.31}$$

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

Theorem 4.2. Suppose $\epsilon \geq 0$, $x^n \in T_\epsilon^n(P_X)$, and X^n is emitted by the DMS $P_X(\cdot)$. We have

$$2^{-n(1+\epsilon)H(X)} \leq P_X^n(x^n) \leq 2^{-n(1-\epsilon)H(X)} \tag{4.32}$$

$$(1 - \delta_\epsilon(P_X, n))\, 2^{n(1-\epsilon)H(X)} \leq |T_\epsilon^n(P_X)| \leq 2^{n(1+\epsilon)H(X)} \tag{4.33}$$

$$1 - \delta_\epsilon(P_X, n) \leq \Pr\left[X^n \in T_\epsilon^n(P_X)\right] \leq 1. \tag{4.34}$$

Proof. Consider (4.32). For $x^n \in T_\epsilon^n(P_X)$, we have

$$\begin{aligned} P_X^n(x^n) &\overset{(a)}{=} \prod_{a \in \mathrm{supp}(P_X)} P_X(a)^{N(a|x^n)} \\ &\overset{(b)}{\leq} \prod_{a \in \mathrm{supp}(P_X)} P_X(a)^{nP_X(a)(1-\epsilon)} \\ &= 2^{\sum_{a \in \mathrm{supp}(P_X)} n(1-\epsilon)P_X(a)\log_2 P_X(a)} \\ &= 2^{-n(1-\epsilon)H(X)} \end{aligned} \tag{4.35}$$

where (a) follows because a with $P_X(a) = 0$ cannot appear in typical x^n, and where (b) follows because x^n satisfies $N(a|x^n)/n \geq P_X(a)(1-\epsilon)$. One

can similarly prove the left-hand side of (4.32). We prove (4.34) in Appendix 4.10. For (4.33) we may use the same steps as in (4.17). □

As does Theorem 4.1, Theorem 4.2 gives *quantitative* bounds on typical sequences and sets. The corresponding *qualitative* statements are basically the same as in (4.22)-(4.24):

$$P_X^n(x^n) \approx 2^{-nH(X)} \tag{4.36}$$

$$|T_\epsilon^n(P_X)| \approx 2^{nH(X)} \tag{4.37}$$

$$\Pr\left[X^n \in T_\epsilon^n(P_X)\right] \approx 1. \tag{4.38}$$

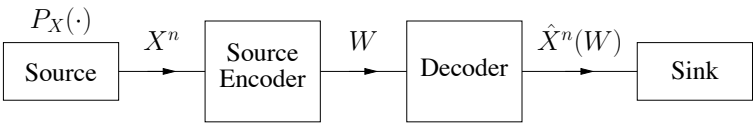

Figure 4.2.: The source coding problem.

4.4. Source Coding with Typical Sequences

The source coding problem is depicted in Fig. 4.2. A DMS $P_X(\cdot)$ emits a string x^n of symbols that are passed to an encoder. The source encoder puts out 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, as we have already seen in earlier chapters. For example, we may use

- block-to-block coding
- block-to-variable-length coding
- variable-length-to-block coding
- variable-length to variable-length coding.

We here consider the first two approaches. Let $L(x^n)$ be the number of bits transmitted for x^n. The goal is to minimize the *average* rate $R = \mathsf{E}[L(X^n)]/n$.

Consider first a block-to-variable-length encoder. We treat the sequences x^n differently depending on whether they are typical or not.

- Assign each sequence in $T_\epsilon^n(P_X)$ a unique positive integer w. According to (4.33), w can be represented by at most $\lceil n(1+\epsilon)H(X) \rceil$ bits. If a typical x^n is put out by the source, then the encoder sends a "0" followed by the $\lceil n(1+\epsilon)H(X) \rceil$ bits that represent x^n.
- If a non-typical x^n is put out by the source, then the encoder sends a "1" followed by $\lceil n \log_2 |\mathcal{X}| \rceil$ bits that represent x^n.

The idea is that x^n is typical with high probability, and there are about $2^{nH(X)}$ such sequences that we can represent with $H(X)$ bits per source symbol. In fact, the average compression rate is upper bounded by

$$\begin{aligned} R = \mathsf{E}[L]/n &= \Pr\left[X^n \in T_\epsilon^n(P_X)\right] \mathsf{E}[L|X^n \in T_\epsilon^n(P_X)]/n \\ &\quad + \Pr\left[X^n \notin T_\epsilon^n(P_X)\right] \mathsf{E}[L|X^n \notin T_\epsilon^n(P_X)]/n \\ &\le \Pr\left[X^n \in T_\epsilon^n(P_X)\right] [(1+\epsilon)H(X) + 2/n] \\ &\quad + \Pr\left[X^n \notin T_\epsilon^n(P_X)\right] (\log_2 |\mathcal{X}| + 2/n) \\ &\le (1+\epsilon)H(X) + 2/n + \delta_\epsilon(P_X, n)(\log_2 |\mathcal{X}| + 2/n). \end{aligned} \tag{4.39}$$

But since $\delta_\epsilon(P_X, n) \to 0$ as $n \to \infty$, we can transmit at any rate above $H(X)$ bits per source symbol, as expected.

Suppose next that we use a block-to-block encoder. We use the same encoding as above if x^n is typical, but we now declare an error if x^n is not

typical, and for this we reserve the all-zeros sequence. The rate is therefore bounded by

$$\begin{aligned} R &\leq \frac{1}{n}\left(\lceil n(1+\epsilon)H(X)\rceil + 1\right) \\ &\leq (1+\epsilon)H(X) + 2/n. \end{aligned} \tag{4.40}$$

The error probability is upper bounded by $\delta_\epsilon(P_X, n)$. By making ϵ small and n large, we can transmit at any rate above $H(X)$ bits per source symbol with a vanishing error probability .

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 which ensures us that

$$H_2(P_e) + P_e \log_2(|\mathcal{X}|^n - 1) \geq H(X^n|\hat{X}^n) \tag{4.41}$$

where $P_e = \Pr\left[\hat{X}^n \neq X^n\right]$. 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

$$\begin{aligned} nR &\geq H(\hat{X}^n) \\ &= H(\hat{X}^n) - H(\hat{X}^n|X^n) \\ &= I(X^n; \hat{X}^n) \\ &= H(X^n) - H(X^n|\hat{X}^n) \\ &= nH(X) - H(X^n|\hat{X}^n) \\ &\overset{(a)}{\geq} n\left[H(X) - \frac{H_2(P_e)}{n} - P_e \log_2|\mathcal{X}|\right] \end{aligned} \tag{4.42}$$

where (a) follows by (4.41). Since we require that P_e be zero, or to be very small, we find that $R \geq H(X)$ for block-to-block encoders. This is the desired converse.

We remark that the above method constructs a code that achieves the best-possible compression rate. However, implementing such a code requires a look-up table of size $2^{\lceil n(1+\epsilon)H(X)\rceil}$ that grows exponentially with n, and an exponential growth rate is completely impractical even for relatively small n. For example, if $H(X) = 1/2$ and $n = 1000$ symbols, then the table has over 2^{500} entries. As a comparison, an estimate of the number of hydrogen atoms in the visible universe is "only" about $10^{80} \approx 2^{266}$. This example demonstrates the importance of having *simple* encoders (functions mapping x^n to w) and simple decoders (functions mapping w to x^n) that approach the rate of $H(X)$ bits per source symbol.

Yet another important task is to develop *universal* encoders and decoders that can compress a sequence X^n with *memory* and *unknown* statistics to a small number of bits per source symbol. A natural approach to such

problems is to have the encoders and decoders effectively learn the source statistics before or during compression and decompression, respectively.

4.5. Jointly 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* ϵ-letter typical set with respect to $P_{XY}(\cdot)$ is simply

$$T_\epsilon^n(P_{XY}) = \Big\{(x^n,y^n) : \Big|\frac{1}{n}N(a,b|x^n,y^n) - P_{XY}(a,b)\Big| \leq \epsilon \cdot P_{XY}(a,b) \quad \text{for all } (a,b)\in\mathcal{X}\times\mathcal{Y}\Big\}. \tag{4.43}$$

The reader can check that $(x^n,y^n)\in T_\epsilon^n(P_{XY})$ implies both $x^n\in T_\epsilon^n(P_X)$ and $y^n\in T_\epsilon^n(P_Y)$ (see Problem 4.3).

Example 4.8. Consider the joint distribution

$$\begin{aligned} P_{XY}(0,0) &= 1/3, \quad P_{XY}(0,1) = 1/3 \\ P_{XY}(1,0) &= 1/3, \quad P_{XY}(1,1) = 0 \end{aligned} \tag{4.44}$$

and choose $\epsilon = 0$. The joint typicality tests in (4.43) are simply

$$\frac{N(a,b|x^n,y^n)}{n} = P_{XY}(a,b) \tag{4.45}$$

and the empirical distribution must match the desired distribution *exactly*. Thus n must be a multiple of 3 and $T_0^n(P_{XY})$ is the set of all $n!/[(n/3)!]^3$ strings of pairs with exactly $n/3$ pairs being $(0,0)$, $(0,1)$, and $(1,0)$. For example, for $n=18$ this gives $|T_0^{18}(P_{XY})|$=17,153,136 strings of pairs.

Joint typicality is a special case of the usual typicality since we can view the word XY as a random variable Z. However, an interesting experiment is the following. Suppose $\tilde{X}^n$ and $\tilde{Y}^n$ are output by the *statistically independent* sources $P_X(\cdot)$ and $P_Y(\cdot)$, as shown in Fig. 4.3. We are interested in the probability that $(\tilde{X}^n,\tilde{Y}^n)\in T_\epsilon^n(P_{XY})$ for some joint distribution $P_{XY}(\cdot)$ that has marginals $P_X(\cdot)$ and $P_Y(\cdot)$. We use Theorem 4.2 to bound

$$\begin{aligned} \Pr\Big[(\tilde{X}^n,\tilde{Y}^n)\in T_\epsilon^n(P_{XY})\Big] &= \sum_{(\tilde{x}^n,\tilde{y}^n)\in T_\epsilon^n(P_{XY})} P_X^n(\tilde{x}^n)P_Y^n(\tilde{y}^n) \\ &\leq 2^{n(1+\epsilon)H(XY)}\,2^{-n(1-\epsilon)H(X)}\,2^{-n(1-\epsilon)H(Y)} \\ &= 2^{-n[I(X;Y)-\epsilon(H(XY)+H(X)+H(Y))]}. \end{aligned} \tag{4.46}$$

We similarly have

$$\begin{aligned} &\Pr\Big[(\tilde{X}^n,\tilde{Y}^n)\in T_\epsilon^n(P_{XY})\Big] \\ &\quad\geq (1-\delta_\epsilon(P_{XY},n))\,2^{-n[I(X;Y)+\epsilon(H(XY)+H(X)+H(Y))]} \end{aligned} \tag{4.47}$$

so that

$$\Pr\Big[(\tilde{X}^n,\tilde{Y}^n)\in T_\epsilon^n(P_{XY})\Big] \approx 2^{-nI(X;Y)}. \tag{4.48}$$

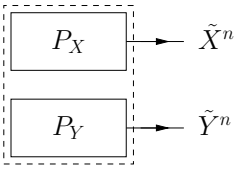

Figure 4.3.: A random experiment with two independent sources.

The result (4.48) has important consequences for source and channel coding.

4.6. Conditionally Typical Sequences

We next study *conditional* typicality which is more subtle than joint typicality. Consider the conditional distribution $P_{Y|X}(\cdot)$ and define

$$P_{Y|X}^n(y^n|x^n) = \prod_{i=1}^n P_{Y|X}(y_i|x_i) \tag{4.49}$$

$$T_\epsilon^n(P_{XY}|x^n) = \{y^n : (x^n, y^n) \in T_\epsilon^n(P_{XY})\}. \tag{4.50}$$

Observe that $T_\epsilon^n(P_{XY}|x^n) = \emptyset$ if x^n is not in $T_\epsilon^n(P_X)$.

Example 4.9. Consider the joint distribution of Example 4.8. The distribution $P_X(\cdot)$ is (4.1) and for $\epsilon = 0$ and $n = 18$ the set $T_0^{18}(P_X)$ has the $\binom{18}{6}$ sequences with 12 zeros and 6 ones. Consider the typical sequence

$$x^{18} = 000000000000111111. \tag{4.51}$$

We find that $T_0^{18}(P_{XY}|x^{18})$ is the set of $\binom{12}{6}$ strings y^{18} with 6 zeros and 6 ones in the first twelve positions, followed by 6 zeros. For example, the following string is in $T_0^{18}(P_{XY}|x^{18})$:

$$y^{18} = 000000111111000000. \tag{4.52}$$

The joint empirical distribution $N(\cdot|x^{18}, y^{18})/18$ is thus exactly $P_{XY}(\cdot)$.

We shall need the following counterpart of $\delta_\epsilon(P_X, n)$ in (4.31):

$$\delta_{\epsilon_1,\epsilon_2}(P_{XY}, n) = 2|\mathcal{X}||\mathcal{Y}| \exp\left(-2n \cdot (1-\epsilon_1) \cdot \left(\frac{\epsilon_2 - \epsilon_1}{1+\epsilon_1}\right)^2 \cdot \mu_{XY}^2\right) \tag{4.53}$$

where $\mu_{XY} = \min_{(a,b)\in \mathrm{supp}(P_{XY})} P_{XY}(a,b)$ and $0 \le \epsilon_1 < \epsilon_2$. Note that $\delta_{\epsilon_1,\epsilon_2}(P_{XY}, n) \to 0$ as $n \to \infty$. Also note from (4.31) that

$$\delta_{\epsilon_2}(P_{XY}, n) \le \delta_{\epsilon_1,\epsilon_2}(P_{XY}, n). \tag{4.54}$$

In the Appendix, we prove the following theorem that generalizes Theorem 4.2 to include conditioning.

Theorem 4.3. Suppose $0 \le \epsilon_1 < \epsilon$, $(x^n, y^n) \in T_{\epsilon_1}^n(P_{XY})$, and (X^n, Y^n) was emitted by the DMS $P_{XY}(\cdot)$. We have

$$2^{-nH(Y|X)(1+\epsilon_1)} \le P_{Y|X}^n(y^n|x^n) \le 2^{-nH(Y|X)(1-\epsilon_1)} \tag{4.55}$$

$$(1 - \delta_{\epsilon_1,\epsilon}(P_{XY}, n))\, 2^{nH(Y|X)(1-\epsilon)} \le |T_\epsilon^n(P_{XY}|x^n)| \le 2^{nH(Y|X)(1+\epsilon)} \tag{4.56}$$

$$1 - \delta_{\epsilon_1,\epsilon}(P_{XY}, n) \le \Pr\left[Y^n \in T_\epsilon^n(P_{XY}|x^n) \,|\, X^n = x^n\right] \le 1. \tag{4.57}$$

Consider now the probability in (4.57) except *without* conditioning on the event $X^n = x^n$. This means that Y^n is generated independent of x^n.

Theorem 4.4. Consider $P_{XY}(\cdot)$ and suppose $0 \leq \epsilon_1 < \epsilon$, Y^n is emitted by a DMS $P_Y(\cdot)$, and $x^n \in T_{\epsilon_1}^n(P_X)$. We have

$$\begin{aligned}&(1 - \delta_{\epsilon_1,\epsilon}(P_{XY}, n))\ 2^{-n[I(X;Y)+2\epsilon H(Y)]} \\ &\leq \Pr\left[Y^n \in T_\epsilon^n(P_{XY}|x^n)\right] \leq 2^{-n[I(X;Y)-2\epsilon H(Y)]}.\end{aligned} \tag{4.58}$$

Proof. The upper bound follows by (4.55) and (4.56):

$$\begin{aligned}\Pr\left[Y^n \in T_\epsilon^n(P_{XY}|x^n)\right] &= \sum_{y^n \in T_\epsilon(P_{XY}|x^n)} P_Y^n(y^n) \\ &\leq 2^{nH(Y|X)(1+\epsilon)}\, 2^{-nH(Y)(1-\epsilon)} \\ &\leq 2^{-n[I(X;Y)-2\epsilon H(Y)]}\end{aligned} \tag{4.59}$$

The lower bound also follows from (4.55) and (4.56). □

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

$$P_{Y|X}^n(y^n|x^n) \approx 2^{-nH(Y|X)} \tag{4.60}$$

$$|T_\epsilon^n(P_{XY}|x^n)| \approx 2^{nH(Y|X)} \tag{4.61}$$

$$\Pr\left[Y^n \in T_\epsilon^n(P_{XY}|x^n)\,|\,X^n = x^n\right] \approx 1 \tag{4.62}$$

$$\Pr\left[Y^n \in T_\epsilon^n(P_{XY}|x^n)\right] \approx 2^{-nI(X;Y)}. \tag{4.63}$$

We remark that the probabilities in (4.57) and (4.58) (or (4.62) and (4.63)) differ only in whether or not one conditions on $X^n = x^n$.

Example 4.10. Suppose X and Y are independent, in which case the approximations (4.62) and (4.63) both give

$$\Pr\left[Y^n \in T_\epsilon^n(P_{XY}|x^n)\right] \approx 1. \tag{4.64}$$

However, the precise version (4.58) of (4.63) is trivial for $I(X;Y) = 0$ and large n. One must therefore exercise caution when working with approximations such as (4.36)-(4.38) or (4.60)-(4.63).

Example 4.11. Suppose that $X = Y$ so that (4.63) gives

$$\Pr\left[Y^n \in T_\epsilon^n(P_{XY}|x^n)\right] \approx 2^{-nH(X)}. \tag{4.65}$$

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

4.7. Mismatched Typicality

Consider the event that a DMS $P_X(\cdot)$ happens to put out a sequence $\tilde{x}^n$ in $T_\epsilon^n(P_{\tilde{X}})$ where $P_{\tilde{X}}(\cdot)$ could be different than $P_X(\cdot)$. We refer to this situation as *mismatched typicality.*

Suppose first there is a letter a with $P_X(a) = 0$ but $P_{\tilde{X}}(a) > 0$. Example 4.7 shows that if $\epsilon < 1$ then $\tilde{x}^n \in T_\epsilon^n(P_{\tilde{X}})$ implies that $N(a|x^n) > 0$. But such $\tilde{x}^n$ have zero-probability of being put out by the DMS $P_X(\cdot)$. Hence we have $\Pr[X^n \in T_\epsilon^n(P_{\tilde{X}})] = 0$ for $\epsilon < 1$. More interestingly, we next consider $P_{\tilde{X}} \ll P_X$ (see Sec. 1.7).

Theorem 4.5. Suppose $\epsilon \geq 0$, X^n is emitted by the DMS $P_X(\cdot)$, and $P_{\tilde{X}} \ll P_X$. We then have

$$\begin{aligned}&(1 - \delta_\epsilon(P_{\tilde{X}}, n))\ 2^{-n[D(P_{\tilde{X}}\|P_X) - \epsilon \log_2(\mu_{\tilde{X}}\mu_X)]} \\ &\leq \Pr[X^n \in T_\epsilon^n(P_{\tilde{X}})] \leq 2^{-n[D(P_{\tilde{X}}\|P_X) + \epsilon \log_2(\mu_{\tilde{X}}\mu_X)]}.\end{aligned} \tag{4.66}$$

Proof. Consider $\tilde{x}^n \in T_\epsilon^n(P_{\tilde{X}})$ and compute

$$\begin{aligned}P_X^n(\tilde{x}^n) &= \prod_{a \in \mathrm{supp}(P_{\tilde{X}})} P_X(a)^{N(a|\tilde{x}^n)} \\ &\leq \prod_{a \in \mathrm{supp}(P_{\tilde{X}})} P_X(a)^{nP_{\tilde{X}}(a)(1-\epsilon)} \\ &= 2^{\sum_{a \in \mathrm{supp}(P_{\tilde{X}})} n(1-\epsilon)P_{\tilde{X}}(a)\log_2 P_X(a)}.\end{aligned} \tag{4.67}$$

Note that the sum is over $\mathrm{supp}(P_{\tilde{X}})$ and not $\mathrm{supp}(P_X)$. A similar lower bound on $P_X^n(\tilde{x}^n)$ follows as above. We use (4.33) and (4.67) to bound

$$\begin{aligned}\Pr[X^n \in T_\epsilon^n(P_{\tilde{X}})] &= \sum_{\tilde{x}^n \in T_\epsilon^n(P_{\tilde{X}})} P_X^n(\tilde{x}^n) \\ &\leq 2^{n(1+\epsilon)H(\tilde{X})}\, 2^{\sum_{a \in \mathrm{supp}(P_{\tilde{X}})} n(1-\epsilon)P_{\tilde{X}}(a)\log_2 P_X(a)} \\ &\leq 2^{-n\sum_{a \in \mathrm{supp}(P_{\tilde{X}})}(1+\epsilon)P_{\tilde{X}}(a)\log_2 P_{\tilde{X}}(a) - (1-\epsilon)P_{\tilde{X}}(a)\log_2 P_X(a)} \\ &\leq 2^{-n\left[D(P_{\tilde{X}}\|P_X) + \epsilon \log_2(\mu_{\tilde{X}}\mu_X)\right]}.\end{aligned} \tag{4.68}$$

The lower bound in (4.66) follows similarly. □

For small ϵ, large n, and X^n emitted by a DMS $P_X(\cdot)$, we thus have the *qualitative* result

$$\boxed{\Pr[X^n \in T_\epsilon^n(P_{\tilde{X}})] \approx 2^{-nD(P_{\tilde{X}}\|P_X)}}. \tag{4.69}$$

Note that (4.69) is true for small ϵ whether $P_{\tilde{X}} \ll P_X$ or $P_{\tilde{X}} \gg P_X$.

Example 4.12. Suppose $P_{\tilde{X}} = P_X$ so that (4.69) gives

$$\Pr\left[X^n \in T_\epsilon^n(P_{\tilde{X}})\right] \approx 1 \tag{4.70}$$

as expected. Note, however, that the precise version (4.66) is trivial for large n. This example shows that one must exercise caution when working with the approximation (4.69).

Example 4.13. Suppose a vector source $P_X P_Y$ puts out the pair $X^n Y^n$. We wish to determine the probability that $X^n Y^n$ is typical with respect to the joint distribution P_{XY}. Theorem 4.5 gives

$$\begin{aligned}&(1-\delta_\epsilon(P_{XY},n))\ 2^{-n[I(X;Y)-\epsilon\log_2(\mu_{XY}\mu_X\mu_Y)]}\\ &\le \Pr\left[(X^n,Y^n)\in T_\epsilon^n(P_{XY})\right] \le 2^{-n[I(X;Y)+\epsilon\log_2(\mu_{XY}\mu_X\mu_Y)]}.\end{aligned} \tag{4.71}$$

Thus, Theorem 4.5 effectively generalizes (4.46)-(4.47) and Theorem 4.4.

Example 4.14. Suppose P_X is uniform so that $P_{\tilde{X}} \ll P_X$ for any $P_{\tilde{X}}$. We have

$$D(P_{\tilde{X}}\|P_X) = \log_2|\mathcal{X}| - H(\tilde{X}) \tag{4.72}$$

and (4.69) gives

$$\Pr\left[X^n \in T_\epsilon^n(P_{\tilde{X}})\right] \approx \frac{2^{nH(\tilde{X})}}{|\mathcal{X}|^n}. \tag{4.73}$$

The entropy $H(\tilde{X})$ is thus a simple measure for the probability that a uniform X is empirically like $\tilde{X}$.

4.8. Entropy-Typicality for Gaussian Variables

As mentioned at the end of Sec. 4.2, we cannot use letter-typicality for continuous random variables. For example, consider a Gaussian random variable X with density (see (2.18))

$$p_X(a) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2\sigma^2}(a-m)^2} \tag{4.74}$$

where $m = \mathsf{E}[X]$ and $\sigma^2 = \mathsf{Var}[X]$ is the variance of X. The trouble with applying a letter-typicality test is that the probability mass function $P_X(x)$ is zero for any letter x. However, we can use entropy-typicality if we replace distributions $P_X(\cdot)$ with densities $p_X(\cdot)$, and if we replace the entropy $H(X)$ with the differential entropy $h(X)$.

For example, we find that x^n is entropy-typical with respect to the Gaussian density (4.74) if

$$\begin{aligned} &\left| -\frac{1}{n}\log\left(\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x_i-m)^2}{2\sigma^2}}\right) - \frac{1}{2}\log(2\pi e\sigma^2)\right| < \epsilon \\ &\Leftrightarrow \left|\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-m)^2\right) - \sigma^2\right| < \frac{2\sigma^2}{\log(e)}\epsilon. \end{aligned} \tag{4.75}$$

We can interpret (4.75) as follows for small ϵ: an "*empirical variance*" of x^n is close to σ^2.[1] Alternatively, if $m = 0$ the *power* of x^n is close to σ^2.

Consider next joint entropy-typicality. We will say that (x^n, y^n) is jointly entropy-typical with respect to the density $p_{XY}(\cdot)$ if the following *three* conditions are satisfied:

$$\left|\frac{-\log p_X^n(x^n)}{n} - h(X)\right| < \epsilon \tag{4.76}$$

$$\left|\frac{-\log p_Y^n(y^n)}{n} - h(Y)\right| < \epsilon \tag{4.77}$$

$$\left|\frac{-\log p_{XY}^n(x^n, y^n)}{n} - h(XY)\right| < \epsilon \tag{4.78}$$

The reason we require three conditions rather than one is because (4.78) does not necessarily imply (4.76) or (4.77).

Consider a Gaussian density (see (2.21))

$$p_{XY}(\underline{a}) = \frac{1}{2\pi\left|\mathbf{Q}_{XY}\right|^{1/2}} \exp\left(-\frac{1}{2}(\underline{a}-\underline{m})^T\mathbf{Q}_{XY}^{-1}(\underline{a}-\underline{m})\right) \tag{4.79}$$

[1] The name "empirical variance" is not really a good choice, because this would naturally be interpreted to be $\frac{1}{n}\sum_{i=1}^{n}(x_i - m(x^n))^2$ where $m(x^n) = \frac{1}{n}\sum_{i=1}^{n} x_i$. Instead, in (4.75) we are using the random variable mean $m = \mathsf{E}[X]$ rather than the empirical mean $m(x^n)$.

with covariance matrix (see (2.23))

$$\mathbf{Q}_{XY} = \begin{bmatrix} \sigma_X^2 & \rho\,\sigma_X\sigma_Y \\ \rho\,\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}. \tag{4.80}$$

Suppose that ϵ is small. We know that (4.76) means $\sum_{i=1}^n x_i^2 \approx n\sigma_X^2$, and similarly (4.77) means $\sum_{i=1}^n y_i^2 \approx n\sigma_Y^2$. For (4.78) we compute

$$\left| \left(\frac{1}{n} \sum_{i=1}^{n} [x_i\ y_i]\, \mathbf{Q}_{XY}^{-1} \begin{bmatrix} x_i \\ y_i \end{bmatrix} \right) - 2 \right| < \frac{2}{\log(e)}\epsilon. \tag{4.81}$$

For example, suppose that $Y = X + Z$ where Z is independent of X and Z is a Gaussian random variable with zero mean and variance σ_Z^2. We have

$$\mathbf{Q}_{XY} = \begin{bmatrix} \sigma_X^2 & \sigma_X^2 \\ \sigma_X^2 & \sigma_X^2 + \sigma_Z^2 \end{bmatrix} \tag{4.82}$$

and (4.81) becomes

$$\left| \frac{1}{n} \sum_{i=1}^{n} \left(\frac{(y_i - x_i)^2}{\sigma_Z^2} + \frac{x_i^2}{\sigma_X^2} \right) - 2 \right| < \frac{2}{\log(e)}\epsilon. \tag{4.83}$$

We thus find that, in combination with (4.76), the sequences x^n and y^n must have a Euclidean distance $\sum_{i=1}^n (y_i - x_i)^2 \approx n\sigma_z^2$ (or $\sum_{i=1}^n z_i^2 \approx n\sigma_z^2$).

Example 4.15. Consider $n = 1$, $m_X = m_Z = 0$, $\sigma_X^2 = \sigma_Z^2 = 1$, and $\epsilon = 0.4$. We use the logarithm to the base 2. The conditions (4.76)-(4.78) are satisfied if

$$|x^2 - 1| < 0.55 \quad \text{... see (4.75)} \tag{4.84}$$
$$|y^2 - 2| < 1.1 \tag{4.85}$$
$$|(y - x)^2 + x^2 - 2| < 0.55 \quad \text{... see (4.83).} \tag{4.86}$$

The contours of the first two regions are shown in Fig. 4.4 as vertical and horizontal lines. The contours of the third region are ellipses. The jointly typical points (x, y) where all three conditions are satisfied are shaded. These are the points that we expect to observe.

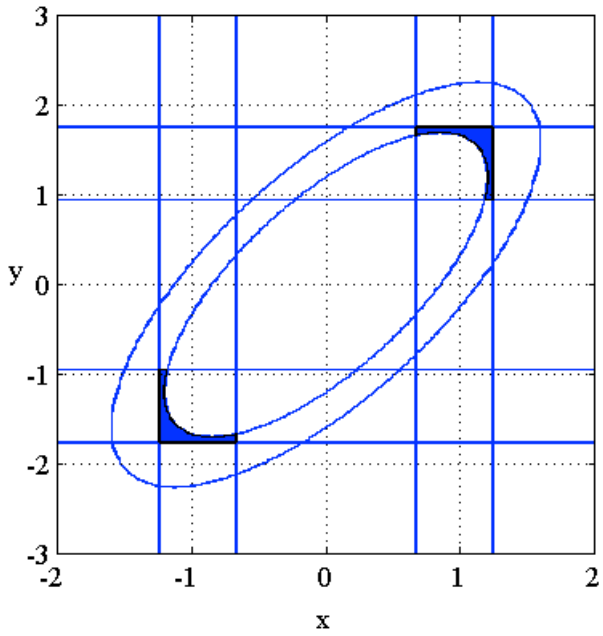

Figure 4.4.: Contour plot for the entropy-typicality conditions.

4.9. Problems

4.1. Entropy Typicality 1

Let $\tilde{T}_\epsilon^n(P_X)$ and $\tilde{T}_\epsilon^n(P_{XY})$ be the sets of **entropy**-typical sequences:

$$\tilde{T}_\epsilon^n(P_X) = \left\{\tilde{x}^n : \left|\frac{-\log_2 P_X^n(\tilde{x}^n)}{n} - H(X)\right| < \epsilon\right\} \tag{4.87}$$

$$\tilde{T}_\epsilon^n(P_{XY}) = \left\{(\tilde{x}^n, \tilde{y}^n) : \left|\frac{-\log_2 P_{XY}^n(\tilde{x}^n, \tilde{y}^n)}{n} - H(XY)\right| < \epsilon\right\}. \tag{4.88}$$

Consider $y_i = f(x_i)$ for some function $f(\cdot)$ and all $i = 1, 2, \ldots, n$. Let $Y = f(X)$.

a) Show that, for entropy-typical sequences, joint typicality (4.88) does not always imply marginal typicality (4.87), i.e., show that there are $(x^n, y^n) \in \tilde{T}_\epsilon^n(P_{XY})$ for which $x^n \notin \tilde{T}_\epsilon^n(P_X)$.
Hint: Try binary X and Y and $n = 1$.

b) Prove that $x^n \in \tilde{T}_\epsilon^n(P_X)$ if and only if $(x^n, y^n) \in \tilde{T}_\epsilon^n(P_{XY})$.
Also prove that $\tilde{y}^n \neq y^n$ implies $(x^n, \tilde{y}^n) \notin \tilde{T}_\epsilon^n(P_{XY})$.

4.2. Entropy Typicality 2

Verify (4.81)-(4.83).

4.3. Letter Typicality

a) Let $T_\epsilon^n(P_X)$ and $T_\epsilon^n(P_{XY})$ be ϵ-letter typical sets. Consider $\mathcal{X} = \{0, 1\}$ and $P_X(0) = P_X(1) = 1/2$. What is $T_{1/3}^1(P_X)$, $T_{1/3}^2(P_X)$, and $T_{1/3}^3(P_X)$?

b) Prove that joint typicality implies marginal typicality.

c) Let y^n satisfy $y_i = f(x_i)$ for some function $f(\cdot)$ and all $i = 1, 2, \ldots, n$. Prove that $x^n \in T_\epsilon^n(P_X)$ if and only if $(x^n, y^n) \in T_\epsilon^n(P_{XY})$ where $Y = f(X)$.

4.4. Letter Typicality with ϵ instead of $\epsilon\, P_X(a)$

Let $\tilde{T}_\epsilon^n(P_X)$ and $\tilde{T}_\epsilon^n(P_{XY})$ now be the sets

$$\tilde{T}_\epsilon^n(P_X) = \left\{x^n : \left|\frac{N(a|x^n)}{n} - P_X(a)\right| < \epsilon\right\} \tag{4.89}$$

$$\tilde{T}_\epsilon^n(P_{XY}) = \left\{(x^n, y^n) : \left|\frac{N(a, b|x^n, y^n)}{n} - P_{XY}(a, b)\right| < \epsilon\right\}. \tag{4.90}$$

Show that joint typicality does not necessarily imply marginal typicality, i.e., show that there are $(x^n, y^n) \in \tilde{T}_\epsilon^n(P_{XY})$ for which $x^n \notin \tilde{T}_\epsilon^n(P_X)$.
Hint: Try binary X and Y, uniform $P_{XY}(\cdot)$, $n = 2$, and $\epsilon = 1/4$.

4.5. Sequence Typicality

A per-letter typicality test can be converted into a single test by adding the tests over all letters. For instance, consider the sum of (4.25) over all letters a which is

$$\sum_{a\in\mathcal{X}} \left|\frac{1}{n}N(a|x^n) - P_X(a)\right| \leq \epsilon. \tag{4.91}$$

Does property (4.25) (or (4.32)) remain valid for x^n satisfying (4.91)?

4.6. Typicality is Typical

Verify the steps of the proof of (4.34) given in Appendix 4.10.
Similarly verify the steps of the proof of Theorem 4.3 given in Appendix 4.10.

4.7. A Simpler Bound

Use Tchebycheff's inequality to prove another version of (4.34), namely

$$\Pr\left[X^n \in T_\epsilon^n(P_X)\right] \geq 1 - \frac{|\mathcal{X}|}{n\epsilon^2\mu_X}. \tag{4.92}$$

Next, use (4.92) and (4.32) to derive a bound on $|T_\epsilon^n(P_X)|$ and compare this with (4.33).
Hint: To prove (4.92), you can use $N(a|x^n) = \sum_{i=1}^n 1(X_i = a)$ where $1(\cdot)$ is the indicator function that takes on the value 1 if its argument is true, and otherwise is 0. Now show that $\mathsf{Var}[1(X_i = a)] = P_X(a)(1 - P_X(a))$.

4.8. Source Coding with Side Information

Consider the source coding problem of Fig. 4.2 but now suppose the DMS is P_{XY}, the source encoder is given X^n and Y^n, and the decoder is given Y^n. Use Theorem 4.3 to show that there are variable-length encoders for which R is close to $H(X|Y)$ and the decoder can recover X^n. Next, show that $H(X|Y)$ is the best possible rate for low-error block-to-block encoding. (Note: the Y^n for this problem is often called "side information".)

4.9. Mismatched Typicality

Suppose X^n, Y^n, Z^n are output by the *statistically independent* sources $P_X(\cdot)$, $P_Y(\cdot)$, $P_Z(\cdot)$.

a) Bound the probability that $(X^n, Y^n, Z^n) \in T_\epsilon^n(P_{XYZ})$ if $P_{XYZ}(\cdot)$ has marginals $P_X(\cdot)$, $P_Y(\cdot)$, $P_Z(\cdot)$.
Hint: Use Theorem 4.5.

b) Simplify the expression if $P_{XYZ}(\cdot) = P_X P_{Y|X} P_{Z|X}$.

4.10. Typicality-Testing Decoders

The proof of the channel coding theorem is often done by using decoders that test for typicality rather than ML decoders. Suppose that, given y^n, the decoder chooses $\hat{w}$ as (one of) the message(s) w for which

$$(x^n(w), y^n) \in T_\epsilon^n(P_{XY}). \tag{4.93}$$

If no such w exists, then the decoder puts out the default value $\hat{w} = 1$.

a) Consider the events

$$\mathcal{E}(\tilde{w}) = \{(X^n(\tilde{w}), Y^n) \in T_\epsilon^n(P_{XY})\}. \tag{4.94}$$

Show that

$$\Pr[\mathcal{E}(1)|W = 1] \geq 1 - \delta_\epsilon(P_{XY}, n) \tag{4.95}$$

$$\Pr[\mathcal{E}(\tilde{w})|W = 1] \leq 2^{-n[I(X;Y)-\epsilon(H(XY)+H(X)+H(Y))]} \tag{4.96}$$

for $\tilde{w} \neq 1$, where (see (4.31))

$$\delta_\epsilon(P_{XY}, n) = 2|\mathcal{X}||\mathcal{Y}| \cdot e^{-2n\epsilon^2\mu_{XY}^2}. \tag{4.97}$$

b) Now show that

$$\Pr[\mathcal{E}] \leq \delta_\epsilon(P_{XY}, n) + (2^{nR} - 1) \cdot 2^{-n[I(X;Y)-\epsilon(H(XY)+H(X)+H(Y))]}. \tag{4.98}$$

Thus, we may replace the ML decoder with a decoder that tests (4.93). Both decoders show that if (3.46) is satisfied then one can make the block error probability small for large n.

c) Using (4.98), find the largest $E_T(R, P_X)$ for which we can write

$$\Pr[\mathcal{E}] \leq a \cdot 2^{-nE_T(R,P_X)} \tag{4.99}$$

for some constant a that is independent of R and P_X. Compare this $E_T(R, P_X)$ to $E_G(R, P_X)$. Which is better?

4.10. Appendix: Proofs

Proof of (4.34)

The right-hand side of (4.34) is trivial. For the left-hand side, consider first $P_X(a) = 0$ for which we have

$$\Pr\left[\frac{N(a|X^n)}{n} > P_X(a)(1+\epsilon)\right] = 0. \tag{4.100}$$

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

$$\begin{aligned}
&\Pr\left[\frac{N(a|X^n)}{n} > P_X(a)(1+\epsilon)\right] \le \Pr\left[\frac{N(a|X^n)}{n} \ge P_X(a)(1+\epsilon)\right] \\
&\le \mathsf{E}\left[e^{\nu N(a|X^n)/n}\right] e^{-\nu P_X(a)(1+\epsilon)} \quad \nu > 0 \\
&= \left[\sum_{m=0}^{n} \Pr\left[N(a|X^n) = m\right] e^{\nu m/n}\right] e^{-\nu P_X(a)(1+\epsilon)} \\
&= \left[\sum_{m=0}^{n} \binom{n}{m} P_X(a)^m (1-P_X(a))^{n-m} e^{\nu m/n}\right] e^{-\nu P_X(a)(1+\epsilon)} \\
&= \left[(1-P_X(a)) + P_X(a)e^{\nu/n}\right]^n e^{-\nu P_X(a)(1+\epsilon)}.
\end{aligned} \tag{4.101}$$

Optimizing (4.101) with respect to ν, we find that

$$\begin{array}{ll} \nu = \infty & \text{if } P_X(a)(1+\epsilon) \ge 1 \\ e^{\nu/n} = \frac{(1-P_X(a))(1+\epsilon)}{1-P_X(a)(1+\epsilon)} & \text{if } P_X(a)(1+\epsilon) < 1. \end{array} \tag{4.102}$$

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 (4.102) into (4.101) and obtain

$$\Pr\left[\frac{N(a|X^n)}{n} \ge P_X(a)(1+\epsilon)\right] \le 2^{-nD(P_B\|P_A)} \tag{4.103}$$

where A and B are binary random variables with

$$\begin{aligned} P_A(0) &= 1 - P_A(1) = P_X(a) \\ P_B(0) &= 1 - P_B(1) = P_X(a)(1+\epsilon). \end{aligned} \tag{4.104}$$

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

$$\begin{aligned} D\left(P_B\|P_A\right) = {} & P_A(0)(1+\epsilon)\log_2(1+\epsilon) \\ & + \left[1 - P_A(0)(1+\epsilon)\right]\log_2\left(\frac{1-P_A(0)(1+\epsilon)}{1-P_A(0)}\right). \end{aligned} \tag{4.105}$$

We wish to lower bound (4.105). A convenient bound is Pinsker's inequality

(see Theorem 1.7) which states that for any distributions P_A and P_B with the same alphabet $\mathcal{X}$ we have

$$D(P_B\|P_A) \geq \frac{\log_2(e)}{2}\left[\sum_{a\in\mathcal{X}}|P_B(a)-P_A(a)|\right]^2 \tag{4.106}$$

Using (4.106) for our problem, we have

$$D\left(P_B\|P_A\right) \geq 2\epsilon^2 P_X(a)^2\log_2(e) \tag{4.107}$$

and combining (4.103) and (4.107) we arrive at

$$\Pr\left[\frac{N(a|X^n)}{n} \geq P_X(a)(1+\epsilon)\right] \leq e^{-2n\epsilon^2 P_X(a)^2}. \tag{4.108}$$

One can similarly bound

$$\Pr\left[\frac{N(a|X^n)}{n} \leq P_X(a)(1-\epsilon)\right] \leq e^{-2n\epsilon^2 P_X(a)^2}. \tag{4.109}$$

Note that (4.108) and (4.109) are valid for all $a \in \mathcal{X}$ including a with $P_X(a)=0$, but we can improve the above bounds for the case $P_X(a)=0$ (see (4.100)). This observation lets us replace $P_X(a)$ in (4.108) and (4.109) with μ_X. Therefore, we get

$$\boxed{\Pr\left[\left|\frac{N(a|X^n)}{n}-P_X(a)\right| > \epsilon\, P_X(a)\right] \leq 2\cdot e^{-2n\epsilon^2\mu_X^2}} \tag{4.110}$$

where $\mu_X = \min_{a\in\mathrm{supp}(P_X)} P_X(a)$. The union bound further gives

$$\begin{aligned}\Pr\left[X^n \notin T_\epsilon^n(P_X)\right] &= \Pr\left[\bigcup_{a\in\mathcal{X}}\left\{\left|\frac{N(a|X^n)}{n}-P_X(a)\right| > \epsilon\, P_X(a)\right\}\right]\\ &\leq \sum_{a\in\mathcal{X}}\Pr\left[\left|\frac{N(a|X^n)}{n}-P_X(a)\right| > \epsilon\, P_X(a)\right]\\ &\leq 2|\mathcal{X}|\cdot e^{-2n\epsilon^2\mu_X^2}.\end{aligned} \tag{4.111}$$

Proof of Theorem 4.3

Suppose that $(x^n,y^n)\in T_{\epsilon_1}^n(P_{XY})$. We prove (4.55) by bounding

$$\begin{aligned}P_{Y|X}^n(y^n|x^n) &= \prod_{(a,b)\in\mathrm{supp}(P_{XY})} P_{Y|X}(b|a)^{N(a,b|x^n,y^n)}\\ &\leq \prod_{(a,b)\in\mathrm{supp}(P_{XY})} P_{Y|X}(b|a)^{nP_{XY}(a,b)(1-\epsilon_1)}\\ &= 2^{n(1-\epsilon_1)\sum_{(a,b)\in\mathrm{supp}(P_{XY})}P_{XY}(a,b)\log_2 P_{Y|X}(b|a)}\\ &= 2^{-n(1-\epsilon_1)H(Y|X)}.\end{aligned} \tag{4.112}$$

This gives the lower bound in (4.55) and the upper bound is proved similarly.

Next, suppose that $x^n \in T^n_{\epsilon_1}(P_X)$ and (X^n, Y^n) was emitted by the DMS $P_{XY}(\cdot)$. We prove (4.57) with $0 \leq \epsilon_1 < \epsilon$ as follows (here ϵ plays the role of ϵ_2 in Theorem 4.3). Consider first $P_{XY}(a,b) = 0$ for which we have

$$\Pr\left[\frac{N(a,b|X^n,Y^n)}{n} > P_{XY}(a,b)(1+\epsilon)\middle| X^n = x^n\right] = 0. \tag{4.113}$$

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

$$\Pr\left[\frac{N(a,b|X^n,Y^n)}{n} > P_{XY}(a,b)(1+\epsilon)\middle| X^n = x^n\right] = 0. \tag{4.114}$$

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

$$\begin{aligned}
&\Pr\left[\frac{N(a,b|X^n,Y^n)}{n} > P_{XY}(a,b)(1+\epsilon)\middle| X^n = x^n\right] \\
&\leq \Pr\left[\frac{N(a,b|X^n,Y^n)}{n} \geq P_{XY}(a,b)(1+\epsilon)\middle| X^n = x^n\right] \\
&= \Pr\left[\frac{N(a,b|X^n,Y^n)}{N(a|x^n)} \geq \frac{P_{XY}(a,b)}{N(a|x^n)/n}(1+\epsilon)\middle| X^n = x^n\right] \\
&\leq \mathsf{E}\left[e^{\nu N(a,b|X^n,Y^n)/N(a|x^n)}\middle| X^n = x^n\right] e^{-\nu\frac{P_{XY}(a,b)(1+\epsilon)}{N(a|x^n)/n}} \\
&= \left[\sum_{m=0}^{N(a|x^n)} \binom{N(a|x^n)}{m} P_{Y|X}(b|a)^m (1-P_{Y|X}(b|a))^{N(a|x^n)-m}\right. \\
&\qquad \left. e^{\nu m/N(a|x^n)}\right] e^{-\nu\frac{P_{XY}(a,b)(1+\epsilon)}{N(a|x^n)/n}} \\
&= \left[(1-P_{Y|X}(b|a)) + P_{Y|X}(b|a)e^{\nu/N(a|x^n)}\right]^{N(a|x^n)} e^{-\nu\frac{P_{XY}(a,b)(1+\epsilon)}{N(a|x^n)/n}}.
\end{aligned} \tag{4.115}$$

Minimizing (4.115) with respect to ν, we find that

$$\begin{aligned}
&\nu = \infty && \text{if } P_{XY}(a,b)(1+\epsilon) \geq N(a|x^n)/n \\
&e^{\nu/N(a|x^n)} = \tfrac{P_X(a)(1-P_{Y|X}(b|a))(1+\epsilon)}{N(a|x^n)/n - P_{XY}(a,b)(1+\epsilon)} && \text{if } P_{XY}(a,b)(1+\epsilon) < N(a|x^n)/n.
\end{aligned} \tag{4.116}$$

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 (4.116) into (4.115) and obtain

$$\Pr\left[\frac{N(a,b|X^n,Y^n)}{n} \geq P_{XY}(a,b)(1+\epsilon)\middle| X^n = x^n\right] \leq 2^{-N(a|x^n)\,D(P_B\|P_A)} \tag{4.117}$$

where A and B are binary random variables with

$$\begin{aligned} P_A(0) &= 1 - P_A(1) = P_{Y|X}(b|a) \\ P_B(0) &= 1 - P_B(1) = \frac{P_{XY}(a,b)}{N(a|x^n)/n}(1+\epsilon). \end{aligned} \tag{4.118}$$

We again use Pinsker's inequality (4.106) to bound

$$D\left(P_B \| P_A\right) \geq 2\left(\frac{\epsilon - \epsilon_1}{1+\epsilon_1}\right)^2 P_{Y|X}(b|a)^2 \log_2(e) \tag{4.119}$$

where we have used $(1-\epsilon_1)P_X(a) \leq \frac{N(a|x^n)}{n} \leq (1+\epsilon_1)P_X(a)$ since $x^n \in T^n_{\epsilon_1}(P_X)$.

Combining (4.117) and (4.119) we arrive at

$$\Pr\left[\frac{N(a,b|X^n,Y^n)}{n} \geq P_{XY}(a,b)(1+\epsilon)\middle| X^n = x^n\right] \leq e^{-2N(a|x^n)\left(\frac{\epsilon-\epsilon_1}{1+\epsilon_1}\right)^2 P_{Y|X}(b|a)^2}. \tag{4.120}$$

One can similarly bound

$$\Pr\left[\frac{N(a,b|X^n,Y^n)}{n} \leq P_{XY}(a,b)(1-\epsilon)\middle| X^n = x^n\right] \leq e^{-2N(a|x^n)\left(\frac{\epsilon-\epsilon_1}{1+\epsilon_1}\right)^2 P_{Y|X}(b|a)^2}. \tag{4.121}$$

For simplicity, we use $P_{Y|X}(b|a) \geq P_{XY}(a,b)$ to replace $P_{Y|X}(b|a)$ with $P_{XY}(a,b)$ so that our bound becomes

$$\begin{aligned} &\Pr\left[\left|\frac{N(a,b|X^n,Y^n)}{n} - P_{XY}(a,b)\right| > \epsilon\, P_{XY}(a,b)\middle| X^n = x^n\right] \\ &\leq 2e^{-2n(1-\epsilon_1)P_X(a)\left(\frac{\epsilon-\epsilon_1}{1+\epsilon_1}\right)^2 P_{Y|X}(b|a)^2} \\ &= 2e^{-2n(1-\epsilon_1)\left(\frac{\epsilon-\epsilon_1}{1+\epsilon_1}\right)^2 P_{XY}(a,b)P_{Y|X}(b|a)} \\ &\leq 2e^{-2n(1-\epsilon_1)\left(\frac{\epsilon-\epsilon_1}{1+\epsilon_1}\right)^2 P_{XY}(a,b)^2}. \end{aligned} \tag{4.122}$$

We thus have

$$\begin{aligned} &\Pr\left[Y^n \notin T^n_\epsilon(P_{XY}|x^n)\middle| X^n = x^n\right] \\ &= \Pr\left[\bigcup_{a,b}\left\{\left|\frac{N(a,b|X^n,Y^n)}{n} - P_{XY}(a,b)\right| > \epsilon\, P_{XY}(a,b)\right\}\middle| X^n = x^n\right] \\ &\leq \sum_{(a,b)\in \mathrm{supp}(P_{XY})} \Pr\left[\left|\frac{N(a,b|X^n,Y^n)}{n} - P_{XY}(a,b)\right| > \epsilon\, P_{XY}(a,b)\middle| X^n = x^n\right] \\ &\leq 2|\mathcal{X}||\mathcal{Y}| \cdot e^{-2n(1-\epsilon_1)\left(\frac{\epsilon-\epsilon_1}{1+\epsilon_1}\right)^2 \mu^2_{XY}}. \end{aligned} \tag{4.123}$$

where we have used the union bound and $P_{XY}(a,b) \geq \mu_{XY}$ for all $(a,b) \in \text{supp}(P_{XY})$. The result is the left-hand side of (4.57).

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

$$\begin{aligned}\Pr\left[Y^n \in T_\epsilon^n(P_{XY}|x^n)\,|\,X^n = x^n\right] &= \sum_{y^n \in T_\epsilon^n(P_{XY}|x^n)} P_{Y|X}^n(y^n|x^n) \\ &\leq |T_\epsilon^n(P_{XY}|x^n)|\, 2^{-n(1-\epsilon)H(Y|X)} \end{aligned} \tag{4.124}$$

where the inequality follows by (4.112). We thus have

$$|T_\epsilon^n(P_{XY}|x^n)| \geq (1 - \delta_{\epsilon_1,\epsilon}(P_{XY}, n))\, 2^{n(1-\epsilon)H(Y|X)}. \tag{4.125}$$

We similarly have

$$|T_\epsilon^n(P_{XY}|x^n)| \leq 2^{n(1+\epsilon)H(Y|X)}. \tag{4.126}$$

References

[1] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 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.

[2] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Channels*. Akadémiai Kiadó, Budapest, 1981.

[3] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, 1991.

[4] J. L. Massey. *Applied Digital Information Theory*. ETH Zurich, Zurich, Switerland, 1980-1998.

[5] A. Orlitsky and J. R. Roche. Coding for computing. *IEEE Trans. Inf. Theory*, 47(3):903–917, March 2001.

Chapter 5.

Lossy Source Coding

5.1. Quantization

Digital communications relies on converting source signals into bits. If the source signal is a waveform, then this waveform is *digitized*, i.e., it is *sampled* in time and *quantized* in value. Sampling theorems guarantee that if the waveform energy is located in a limited amount of (contiguous or discontiguous) bandwidth, then we can sample the signal at twice this bandwidth without losing information about the signal. However, one inevitably loses information by quantizing analog sampled values to bits. *Rate distortion theory* characterizes the limits of quantization.

Quantization is the process of representing fine (analog) signals by a finite number of bits. From this perspective, quantization is the same as *lossy source coding* or lossy compression. For example, suppose we have a speech signal with a bandwidth of 4 kHz. We sample this signal at 8 kHz and represent each sample with 8 bits, giving a bit rate of 64 kbit/s. In fact, the signal can often be represented with adequate distortion at a much slower rate. As a second example, suppose we transmit a string of bits but we permit some fraction, say 11%, of the bits to be received in error. It turns out that this *distortion* of the original bit stream lets us cut the transmission rate in half. This remarkable rate reduction is achieved by *coding* over long sequences of bits.

The smallest required rate decreases with increasing distortion, and this suggests shapes such as those shown in Fig. 5.1. We have a rate-distortion tradeoff and we are interested in determining the frontier of this tradeoff.

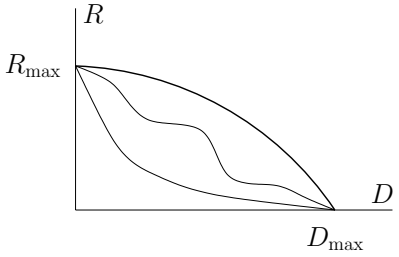

Figure 5.1.: Possible shapes of rate-distortion curves.

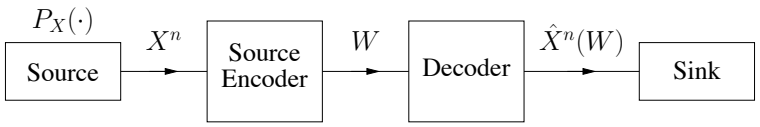

Figure 5.2.: The rate distortion problem.

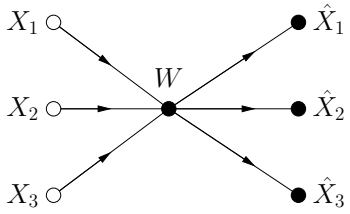

Figure 5.3.: FDG for the rate distortion problem.

We shall find that the lowest shape is the correct one for the distortion functions that we are interested in.

5.2. Problem Description

Consider the problem shown in Fig. 5.2. The FDG is depicted in Fig. 5.3 for $n = 3$ source symbols. 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). Finally, the decoder puts out the *reconstruction* $\hat{x}^n(w)$ of x^n. The letters $\hat{x}_i$ take on values in the alphabet $\hat{\mathcal{X}}$, which is often the same as $\mathcal{X}$. 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 $\mathsf{E}\left[d^n(X^n, \hat{X}^n)\right]$ is at most D.

Example 5.1. Suppose that we have $\mathcal{X} = \hat{\mathcal{X}}$. A commonly-used distortion function is the *Hamming distance*

$$d^n(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} 1(x_i \neq \hat{x}_i) \tag{5.1}$$

where $1(\cdot)$ is the indicator function that takes on the value 1 if the argument is true and is 0 otherwise. The distortion (5.1) measures the *average symbol error probability* of x^n due to the reconstruction $\hat{x}^n$.

Example 5.2. Suppose that $\mathcal{X} = \hat{\mathcal{X}} = \mathbb{R}$. A commonly-used distortion function is the *Euclidean distance* or *squared error distortion*

$$d^n(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{x}_i)^2. \tag{5.2}$$

As in Examples 5.1 and 5.2, we consider sequence distortion functions of the form

$$d^n(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i) \tag{5.3}$$

so that $d^n(\cdot)$ is the *average* of a *per-letter* distortion function $d(\cdot)$. For instance, for Hamming distance we have

$$d(x, \hat{x}) = \begin{cases} 0 & \text{if } \hat{x} = x \\ 1 & \text{else.} \end{cases} \tag{5.4}$$

For real sources and squared error distortion we have

$$d(x, \hat{x}) = (x - \hat{x})^2. \tag{5.5}$$

The functions (5.4) and (5.5) happen to be the same for binary $(0, 1)$ sources. Choosing distortion functions of the form (5.3) is not appropriate for all applications, but we consider only such distortion functions.

As a technicality, we assume that there is a letter $a \in \hat{\mathcal{X}}$ such that $\mathsf{E}[d(X, a)] = d_{\max} \leq \infty$. For example, for a finite-variance source and squared error distortion we may choose $a = \mathsf{E}[X]$ to get $\mathsf{E}[d(X, \mathsf{E}[X])] = \mathsf{Var}[X] < \infty$.

The *rate distortion* (RD) problem is the following: find the set of pairs (D, R) that one can approach with source encoders for sufficiently large n (see [1, Part V], [2]). We ignore the practical difficulties associated with large n but the theory will provide useful bounds on the distortion achieved by *finite* length codes too. 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*.

$\hat{x}^n(1) = \hat{x}_1(1)\,\hat{x}_2(1)\,\ldots\,\hat{x}_n(1)$
$\hat{x}^n(2) = \hat{x}_1(2)\,\hat{x}_2(2)\,\ldots\,\hat{x}_n(2)$
$\vdots \qquad\qquad \vdots$
$\hat{x}^n(2^{nR}) = \hat{x}_1(2^{nR})\,\ldots\,\hat{x}_n(2^{nR})$

Figure 5.4.: A code book for the RD problem.

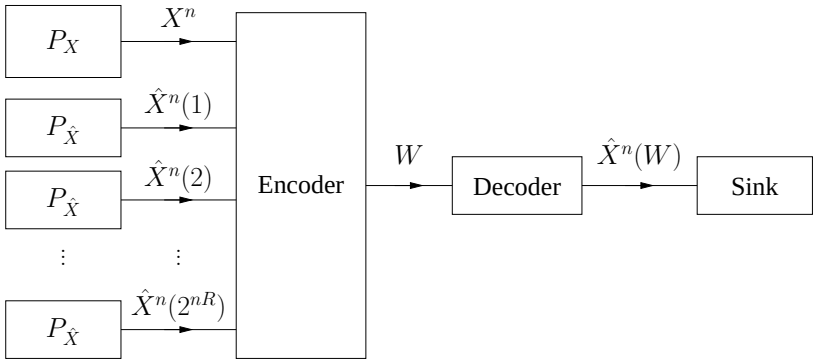

Figure 5.5.: Random coding experiment for lossy source coding.

5.3. Achievable Region for Discrete Sources

We construct a random code book as for the channel coding problem. We begin by choosing 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 $\hat{x}_i(w)$ independently using $P_{\hat{X}}(\cdot)$ (see Fig. 5.4). Choose a default codeword $\hat{x}_i(2^{nR}+1) = a$ for all i, where a is the letter such that $\mathsf{E}[d(X,a)] = d_{\max}$. The overall rate is therefore reduced by the factor $2^{nR}/(2^{nR}+1)$ which rapidly approaches 1 for large n.

Encoder: Given x^n, choose w as (one of) the message(s) that minimizes $d^n(x^n, \hat{x}^n(w))$. Send this w to the decoder.

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

Analysis: The random coding experiment is shown in Fig. 5.5 where the random variables X^n, $\hat{X}^n(1), \hat{X}^n(2), \ldots, \hat{X}^n(2^{nR})$ are mutually statistically independent. We will work with typical sequences and choose the code book size (in terms of R) sufficiently large so that there is a $\hat{x}^n(w)$ that is jointly typical with x^n. We then show that this $\hat{x}^n(w)$ gives a bounded distortion. The encoder that minimizes distortion gives the same or smaller distortion.

So consider the event that there is no $\hat{X}^n(w)$ that is jointly typical with X^n:

$$\mathcal{E} = \bigcap_{w=1}^{2^{nR}+1} \left\{ (X^n, \hat{X}^n(w)) \notin T_\epsilon(P_{X\hat{X}}) \right\}. \tag{5.6}$$

The Theorem on Total Expectation gives

$$\mathsf{E}\left[d^n(X^n,\hat{X}^n)\right] = \Pr[\mathcal{E}]\,\mathsf{E}\left[d^n(X^n,\hat{X}^n)|\mathcal{E}\right] + \Pr[\mathcal{E}^c]\,\mathsf{E}\left[d^n(X^n,\hat{X}^n)|\mathcal{E}^c\right] \tag{5.7}$$

where $\mathcal{E}^c$ is the complement of $\mathcal{E}$. We consider each term in (5.7) separately.

- Let $0 < \epsilon_1 < \epsilon$. We have

$$\begin{aligned}
&\Pr[\mathcal{E}] = \sum_{x^n} P_X^n(x^n)\Pr[\mathcal{E}\,|X^n = x^n] \\
&= \Pr\left[X^n \notin T_{\epsilon_1}^n(P_X)\right]\cdot\Pr\left[\mathcal{E}\,\middle|X^n \notin T_{\epsilon_1}^n(P_X)\right] \\
&+ \sum_{x^n\in T_{\epsilon_1}^n(P_X)} P_X^n(x^n)\cdot\Pr\left[\bigcap_{w=1}^{2^{nR}+1}\left\{(x^n,\hat{X}^n(w)) \notin T_\epsilon^n(P_{X\hat{X}})\right\}\middle| X^n = x^n\right].
\end{aligned} \tag{5.8}$$

  We may upper bound the first product in (5.8) by $\delta_{\epsilon_1}(P_X,n)$, and for simplicity we write this as $\delta_{\epsilon_1}(n)$. Observe that the conditioning on $X^n = x^n$ can be removed in (5.8) because $\hat{X}^n(w)$ is statistically independent of X^n for all w. Moreover, the 2^{nR} events in the intersection in (5.8) are independent because each $\hat{X}^n(w)$ is generated independently. The probability of the intersection of events in (5.8) is thus upper bounded as

$$\begin{aligned}
&\left[1-\Pr\left[(x^n,\hat{X}^n)\in T_\epsilon^n(P_{X\hat{X}})\right]\right]^{2^{nR}} \\
&\overset{(a)}{\leq}\left[1-(1-\delta_{\epsilon_1,\epsilon}(n))\,2^{-n[I(X;\hat{X})+2\epsilon H(\hat{X})]}\right]^{2^{nR}} \\
&\overset{(b)}{\leq}\exp\left(-(1-\delta_{\epsilon_1,\epsilon}(n))\,2^{n[R-I(X;\hat{X})-2\epsilon H(\hat{X})]}\right)
\end{aligned} \tag{5.9}$$

  where (a) follows by Theorem 4.4, and (b) follows by $(1-x)^m \leq e^{-mx}$. Inequality (5.9) implies that we can choose large n and

$$R > I(X;\hat{X}) + 2\epsilon H(\hat{X}) \tag{5.10}$$

  to drive the right-hand side of (5.9) to zero. In addition, observe that the bound is valid for *any* x^n in $T_{\epsilon_1}^n(P_X)$, and the error probability decreases *doubly* exponentially in n. Denote the quantity on the right-hand side of (5.9) as $\delta_{\epsilon_1,\epsilon}(n,R)$.

- If $\mathcal{E}$ occurs then the encoder could send $w = 2^{nR}+1$ and achieve

$$\mathsf{E}\left[d^n(X^n,\hat{X}^n)|\mathcal{E}\right] = d_{\max} \tag{5.11}$$

  But our encoder chooses a codeword that minimizes $d^n(x^n,\hat{x}^n(w))$ and thus achieves an average distortion at most $d_{\max}$.

- For $\mathsf{E}\Big[d^n(X^n,\hat{X}^n)|\mathcal{E}^c\Big]$, observe that $(x^n,\hat{x}^n(w)) \in T_\epsilon^n(P_{X\hat{X}})$ implies

$$\begin{aligned} d^n(x^n,\hat{x}^n(w)) &= \frac{1}{n}\sum_{i=1}^{n} d(x_i,\hat{x}_i(w)) \\ &= \frac{1}{n}\sum_{a,b} N(a,b|x^n,\hat{x}^n(w))\, d(a,b) \\ &\overset{(a)}{\leq} \sum_{a,b} P_{X\hat{X}}(a,b)(1+\epsilon)\, d(a,b) \\ &= \mathsf{E}\Big[d(X,\hat{X})\Big]\,(1+\epsilon) \end{aligned} \tag{5.12}$$

  where (a) follows by the definition of typical pairs. Thus, if $\mathcal{E}^c$ occurs then the message w that minimizes $d^n(x^n,\hat{x}^n(w))$ has distortion at most $d^n(x^n,\hat{x}^n(w)) = \mathsf{E}\Big[d(X,\hat{X})\Big]\,(1+\epsilon)$.

Combining the above results using (5.7), we have

$$\mathsf{E}\Big[d^n(X^n,\hat{X}^n)\Big] \leq (\delta_{\epsilon_1}(n) + \delta_{\epsilon_1,\epsilon}(n,R))\; d_{\max} + \mathsf{E}\Big[d(X,\hat{X})\Big]\,(1+\epsilon). \tag{5.13}$$

As a final step, we choose small ϵ, large n, R satisfying (5.10), and $P_{\hat{X}|X}$ so that $\mathsf{E}\Big[d(X,\hat{X})\Big] < D$, assuming this is possible. A random code thus *achieves* the rates R satisfying

$$R > \min_{P_{\hat{X}|X}\colon\, \mathsf{E}\left[d(X,\hat{X})\right]<D} I(X;\hat{X}) \tag{5.14}$$

as long as the constraint $\mathsf{E}\Big[d(X,\hat{X})\Big] < D$ can be satisfied by some $P_{\hat{X}|X}$. Alternatively, we say that a random code *approaches* the rate

$$\boxed{R(D) = \min_{P_{\hat{X}|X}\colon\, \mathsf{E}\left[d(X,\hat{X})\right]\leq D} I(X;\hat{X})} \tag{5.15}$$

as long as the constraint can be satisfied, and $R(D)$ is undefined otherwise. 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 in (5.7) is over source string X^n and the code book $\hat{X}^n(1),\hat{X}^n(2),\ldots,\hat{X}^n(2^{nR})$. 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 based on the 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.

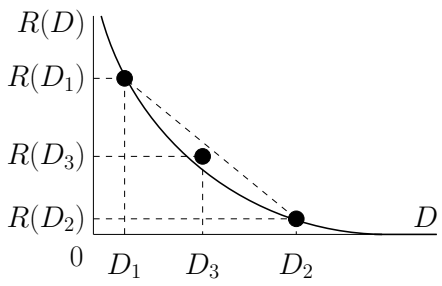

Figure 5.6.: Convexity of the rate-distortion function.

5.4. Convexity and Converse

The function $R(D)$ in (5.15) is non-increasing in D because increasing D lets us use a larger class of channels $P_{\hat{X}|X}$. We show that $R(D)$ is convex in D [2]. Consider two distinct points $(D_1, R(D_1))$ and $(D_2, R(D_2))$ and suppose the channels $P_{\hat{X}_1|X}(\cdot)$ and $P_{\hat{X}_2|X}(\cdot)$ achieve these respective points (see Fig. 5.6). That is, we have

$$\begin{aligned} D_1 &= \mathsf{E}\Big[d(X,\hat{X}_1)\Big], \quad R(D_1) = I(X;\hat{X}_1) \\ D_2 &= \mathsf{E}\Big[d(X,\hat{X}_2)\Big], \quad R(D_2) = I(X;\hat{X}_2). \end{aligned} \tag{5.16}$$

Consider the mixture distribution

$$P_{\hat{X}_3|X}(\hat{x}|x) = \lambda\, P_{\hat{X}_1|X}(\hat{x}|x) + (1-\lambda)\, P_{\hat{X}_2|X}(\hat{x}|x) \tag{5.17}$$

for all $x, \hat{x}$, where $0 \le \lambda \le 1$. We have

$$\begin{aligned} D_3 &= \sum_{(x,\hat{x})\in \mathrm{supp} P_{X\hat{X}_3}} P_X(x)\; P_{\hat{X}_3|X}(\hat{x}|x)\; d(x,\hat{x}) \\ &= \sum_{(x,\hat{x})\in \mathrm{supp} P_{X\hat{X}_3}} P_X(x)\left(\lambda P_{\hat{X}_1|X}(\hat{x}|x) + (1-\lambda) P_{\hat{X}_2|X}(\hat{x}|x)\right) d(x,\hat{x}) \\ &= \lambda D_1 + (1-\lambda) D_2. \end{aligned} \tag{5.18}$$

We thus have

$$\begin{aligned} R(\lambda\, D_1 + (1-\lambda)\, D_2) &= R(D_3) \\ &\overset{(a)}{\le} I(X;\hat{X}_3) \\ &\overset{(b)}{\le} \lambda\, I(X;\hat{X}_1) + (1-\lambda)\, I(X;\hat{X}_2) \\ &= \lambda\, R(D_1) + (1-\lambda)\, R(D_2). \end{aligned} \tag{5.19}$$

where (a) follows because $P_{\hat{X}_3|X}$ might not minimize the mutual information for the distortion D_3, and (b) follows by the convexity of $I(X;Y)$ in $P_{Y|X}$ when P_X is held fixed (see Thm. 1.10). Thus, $R(D)$ is convex in D.

We now show that $R(D)$ in (5.15) is the rate distortion function. The code

book has 2^{nR} sequences $\hat{x}^n$, and $\hat{x}^n$ is a function of x^n. We thus have

$$\begin{aligned} nR &\geq H(\hat{X}^n) \\ &= H(\hat{X}^n) - H(\hat{X}^n|X^n) \\ &= I(X^n;\hat{X}^n) \\ &= H(X^n) - H(X^n|\hat{X}^n) \\ &= \sum_{i=1}^{n} \Big[H(X_i) - H(X_i|\hat{X}^n X^{i-1})\Big] \\ &\geq \sum_{i=1}^{n} \Big[H(X_i) - H(X_i|\hat{X}_i)\Big] \\ &= \sum_{i=1}^{n} I(X_i;\hat{X}_i). \end{aligned} \tag{5.20}$$

But we have $I(X_i;\hat{X}_i) \geq R(\mathsf{E}\Big[d(X_i,\hat{X}_i)\Big])$ because $R(\mathsf{E}\Big[d(X_i,\hat{X}_i)\Big])$ minimizes mutual information for the distortion $\mathsf{E}\Big[d(X_i,\hat{X}_i)\Big]$. We thus have

$$\begin{aligned} R &\geq \sum_{i=1}^{n} \frac{1}{n} R\left(\mathsf{E}\Big[d(X_i,\hat{X}_i)\Big]\right) \\ &\overset{(a)}{\geq} R\left(\frac{1}{n}\sum_{i=1}^{n} \mathsf{E}\Big[d(X_i,\hat{X}_i)\Big]\right) \\ &= R(\mathsf{E}\Big[d^n(X^n,\hat{X}^n)\Big]) \\ &\overset{(b)}{\geq} R(D) \end{aligned} \tag{5.21}$$

where (a) follows by (5.19) and (b) follows because $R(D)$ is non-increasing in D. Thus, R must be larger than $R(D)$, and this is called a *converse*. But we can achieve $R(D)$ so the rate distortion function is $R(D)$.

5.5. Discrete Alphabet Examples

5.5.1. Binary Symmetric Source and Hamming Distortion

Consider the binary symmetric source (BSS) with the Hamming distortion function and desired average distortion D, where $D \leq 1/2$. We then require $\Pr\left[X \neq \hat{X}\right] \leq D$, and can bound

$$\begin{aligned} I(X;\hat{X}) &= H(X) - H(X|\hat{X}) \\ &= 1 - H(X \oplus \hat{X}|\hat{X}) \\ &\geq 1 - H(X \oplus \hat{X}) \\ &\geq 1 - H_2(D) \end{aligned} \tag{5.22}$$

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. But we "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.

5.5.2. Scalar Quantization

Scalar quantization has $\hat{X} = f(X)$ for some function $f(\cdot)$. The "rate-distortion" function with scalar quantization is therefore

$$R_1(D) = \min_{f:\ \mathsf{E}[d(X,f(X))] \leq D} H(f(X)). \tag{5.23}$$

Scalar quantizers designed to approach the rate (5.23), or alternatively to minimize distortion under a rate constraint, are called *entropy-coded quantizers*. For example, consider a BSS and Hamming distance. The interesting functions are $f(0) = 0, f(1) = 1$ and $f(0) = f(1) = 0$ (or $f(0) = f(1) = 1$) so that we have

$$R_1(D) = \begin{cases} 1, & 0 \leq D < 1/2 \\ 0, & D \geq 1/2. \end{cases} \tag{5.24}$$

We find that $R(D) \leq R_1(D)$, as should be expected. Furthermore, $R_1(D)$ is *discontinuous* in D.

5.5.3. Binary Symmetric Source and Erasure Distortion

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

$$d(x, \hat{x}) = \begin{cases} 0, & \text{if } x = \hat{x} \\ 1, & \text{if } \hat{x} = \Delta \\ \infty, & \text{if } \hat{x} = x \oplus 1. \end{cases} \tag{5.25}$$

Note that the letter $\Delta \in \hat{\mathcal{X}}$ gives $\mathsf{E}[d(X, \Delta)] = 1 < \infty$ so that $d_{\max} = 1$. To achieve finite distortion D, we must choose $P_{\hat{X}|X}(1|0) = P_{\hat{X}|X}(0|1) = 0$ and $\Pr\left[\hat{X} = \Delta\right] \leq D$. We thus have

$$\begin{aligned} I(X; \hat{X}) &= 1 - H(X|\hat{X}) \\ &= 1 - \sum_{b \in \hat{\mathcal{X}}} P_{\hat{X}}(b) H(X|\hat{X} = b) \\ &\geq 1 - D. \end{aligned} \tag{5.26}$$

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 Δ^m is a string of m successive Δs.

5.6. Gaussian Source and Squared Error Distortion

We can approach the rate (5.15) for the memoryless Gaussian source with squared error distortion. We will not prove this here, see [3, Ch. 9]. We require $\mathsf{E}\left[(X - \hat{X})^2\right] \leq D$, and bound

$$\begin{aligned} I(X; \hat{X}) &= h(X) - h(X|\hat{X}) \\ &= \frac{1}{2} \log_2(2\pi e \sigma^2) - h(X - \hat{X}|\hat{X}) \\ &\geq \frac{1}{2} \log_2(2\pi e \sigma^2) - h(X - \hat{X}) \\ &\overset{(a)}{\geq} \frac{1}{2} \log_2(2\pi e \sigma^2) - \frac{1}{2} \log_2\left(2\pi e \, \mathsf{E}\left[(X - \hat{X})^2\right]\right) \\ &\geq \frac{1}{2} \log_2(2\pi e \sigma^2) - \frac{1}{2} \log_2(2\pi e D) \\ &= \frac{1}{2} \log_2(\sigma^2 / D) \end{aligned} \tag{5.27}$$

where σ^2 is the source variance, and where (a) follows by the maximum entropy theorem (see [3, p. 234]). We can achieve $R(D) = \frac{1}{2} \log_2(\sigma^2/D)$ bits by choosing $P_{X|\hat{X}}$ (note that this is not $P_{\hat{X}|X}$) to be the additive white

Gaussian noise (AWGN) channel with noise variance D. Alternatively, we can achieve the distortion

$$D(R) = \sigma^2 2^{-2R}, \tag{5.28}$$

i.e., we can gain 6 dB per quantization bit.

5.7. Problems

5.1. Gaussian Source, Mean-Squared-Error Distortion

a) For the Gaussian example of Sec. 2.4 with $D \leq \sigma^2$, show that choosing $X = \hat{X} + Z$ where Z is Gaussian, zero-mean, variance D, and independent of $\hat{X}$, gives $I(X;\hat{X}) = \frac{1}{2}\log(\sigma^2/D)$ and $\mathsf{E}\Big[(X-\hat{X})^2\Big] = D$.

b) For the Gaussian example of Sec. 2.4 with $D \leq \sigma^2$, suppose we choose $\hat{X} = aX + Z$ where a is a constant and Z is independent of X. Find a and Z that achieve $I(X;\hat{X}) = \frac{1}{2}\log(\sigma^2/D)$ and $\mathsf{E}\Big[(X-\hat{X})^2\Big] = D$. Is your choice for a and Z unique?

5.2. Vector Gaussian Source

Consider a vector Gaussian source that puts out $\underline{X} = [X_1, X_2, \ldots, X_M]$ where the X_m, $m = 1, 2, \ldots, M$, are independent Gaussian random variables, and where X_m has variance N_m. Suppose the per-source-letter distortion function is $d(\underline{x}, \underline{\hat{x}}) = \sum_{m=1}^{M}(x_m - \hat{x}_m)^2$. Determine the rate-distortion function $R(D)$.

References

[1] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 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.

[2] C. E. Shannon. Coding theorems for a discrete source with a fidelity criterion. In *IRE Int. Conv. Rec.*, pages 142–163, March 1959. Reprinted in *Claude Elwood Shannon: Collected Papers*, pp. 325-350, (N.J.A. Sloane and A.D. Wyner, eds.) Piscataway: IEEE Press, 1993.

[3] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley & Sons, New York, 1991.

Chapter 6.

Distributed Source Coding

6.1. 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 Fig. 6.1 and the FDG in Fig. 6.2).

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

$$P_e = \Pr\left[(\hat{X}^n, \hat{Y}^n) \neq (X^n, Y^n)\right] \tag{6.1}$$

can be made an arbitrarily small positive number.

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

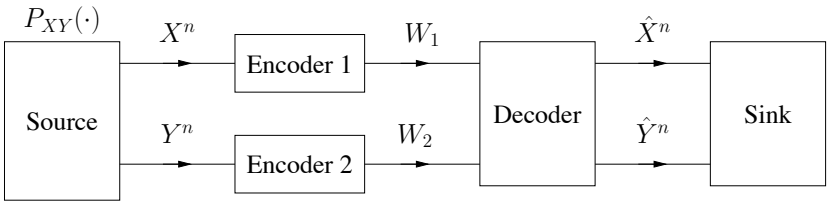

Figure 6.1.: A distributed source coding problem.

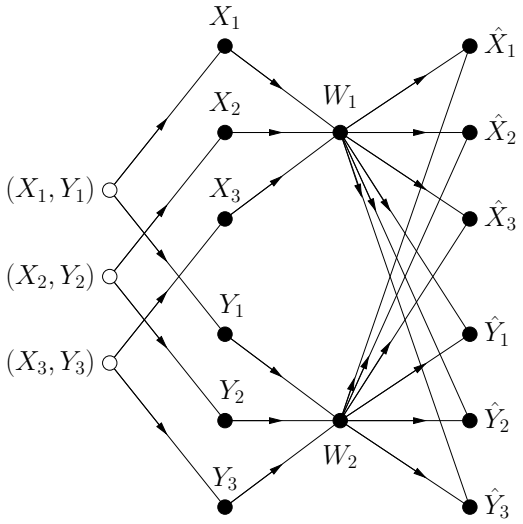

Figure 6.2.: FDG for the distributed source coding problem.

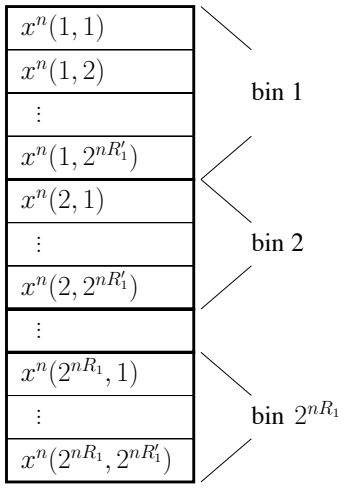

Figure 6.3.: Binning for the x^n sequences.

The problem of Fig. 6.1 was solved by D. Slepian and J. K. Wolf in a fundamental paper in 1973 [1]. 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 [2], and is now widely known as partitioning, binning, or hashing.

6.2. An Achievable Region

We will consider only block-to-block encoders, although one could also use variable-length encoders. The code construction is depicted in Fig. 6.3 and Fig. 6.4 (see also [3, p. 412]). We use random coding that makes use of a method sometimes called *binning*.

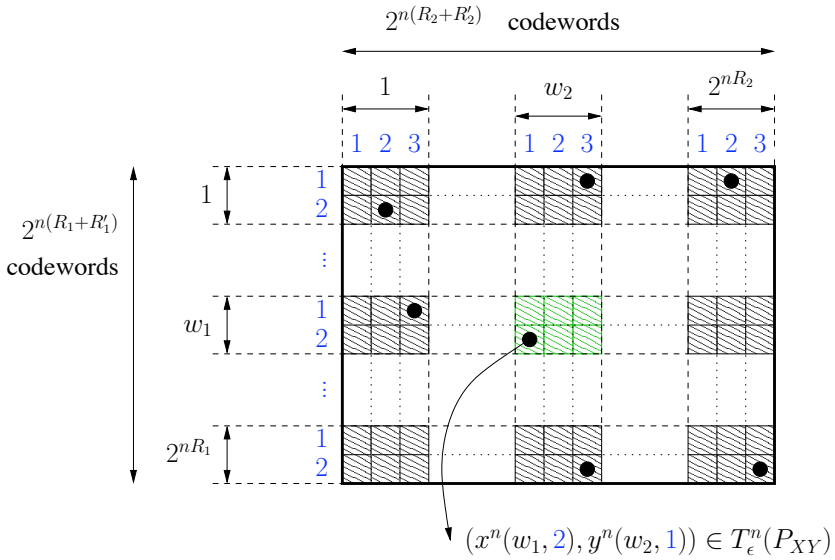

Figure 6.4.: Binning for x^n and y^n. A dot indicates (x^n, y^n) in $T_\epsilon^n(P_{XY})$. There should be at most one dot for every bin pair (w_1, w_2).

Code Construction: Let R_1, R_1', R_2, R_2', be fixed rates, the choice of which will be specified later. Generate $2^{n(R_1+R_1')}$ codewords $x^n(w_1, v_1)$, $w_1 = 1, 2, \ldots, 2^{nR_1}$, $v_1 = 1, 2, \ldots, 2^{nR_1'}$, by choosing the $n \cdot 2^{n(R_1+R_1')}$ symbols $x_i(w_1, v_1)$ independently using $P_X(\cdot)$. Similarly, generate $2^{n(R_2+R_2')}$ codewords $y^n(w_2, v_2)$, $w_2 = 1, 2, \ldots, 2^{nR_2}$, $v_2 = 1, 2, \ldots, 2^{nR_2'}$, by choosing the $n \cdot 2^{n(R_2+R_2')}$ symbols $y_i(w_2, v_2)$ independently using $P_Y(\cdot)$.

Encoders: Encoder 1 tries to find a pair $(\tilde{w}_1, \tilde{v}_1)$ such that $x^n = x^n(\tilde{w}_1, \tilde{v}_1)$. If there is one or more such pair, then Encoder 1 chooses one by using a pre-defined function with $(w_1, v_1) = f_1(x^n, x^n(\cdot)))$. If unsuccessful, Encoder 1 chooses $(w_1, v_1) = (1, 1)$. Encoder 2 proceeds in the same way with y^n and a pre-defined function $f_2(\cdot)$ and transmits w_2.

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

Analysis: The random coding experiment has the joint probability distribution

$$\begin{aligned} & P_{XY}^n(x^n, y^n) \left[\prod_{\tilde{w}_1, \tilde{v}_1} P_X^n(x^n(\tilde{w}_1, \tilde{v}_1)) \right] \left[\prod_{\tilde{w}_2, \tilde{v}_2} P_Y^n(y^n(\tilde{w}_2, \tilde{v}_2)) \right] \\ & \cdot 1((W_1, V_1) = f_1(x^n, x^n(\cdot))) \cdot 1((W_2, V_2) = f_2(y^n, y^n(\cdot))) \\ & \cdot 1((\hat{V}_1, \hat{V}_2) = g(W_1, W_2, x^n(\cdot), y^n(\cdot))) \\ & \cdot 1\left(\left(\hat{X}^n(W_1, \hat{V}_1), \hat{Y}^n(W_2, \hat{V}_2) \right) = (x^n(W_1, \hat{V}_1), y^n(W_2, \hat{V}_2))) \right) \end{aligned} \tag{6.2}$$

where $f_1(\cdot)$ and $f_2(\cdot)$ are the encoding functions at Encoder 1 and Encoder 2, respectively, and $g(\cdot)$ is a decoder function. We wish to bound $P_e = \Pr\left[(\hat{X}^n, \hat{Y}^n) \neq (X^n, Y^n)\right]$, which is the average error probability over all code books. We again use the Theorem on Total Expectation in two ways

based on

a) events $\mathcal{E}_i$: $P_e = \sum_i \Pr[\mathcal{E}_i] \Pr\left[(\hat{X}^n, \hat{Y}^n) \neq (X^n, Y^n) | \mathcal{E}_i\right]$
b) code books $\mathcal{C}_j$: $P_e = \sum_j \Pr[\mathcal{C}_j] \Pr\left[(\hat{X}^n, \hat{Y}^n) \neq (X^n, Y^n) | \mathcal{C}_j\right]$.

For the first approach, we consider four error events.

a) The source sequences are not jointly typical:

$$\mathcal{E}_1 = \{(X^n, Y^n) \notin T_\epsilon^n(P_{XY})\}. \tag{6.3}$$

b) Encoder 1 cannot find a $X^n(\tilde{w}_1, \tilde{v}_1)$ that is X^n:

$$\mathcal{E}_2 = \bigcap_{\tilde{w}_1, \tilde{v}_1} \{X^n(\tilde{w}_1, \tilde{v}_1) \neq X^n\}. \tag{6.4}$$

c) Encoder 2 cannot find a $Y^n(\tilde{w}_2, \tilde{v}_2)$ that is Y^n:

$$\mathcal{E}_3 = \bigcap_{\tilde{w}_2, \tilde{v}_2} \{Y^n(\tilde{w}_2, \tilde{v}_2) \neq Y^n\}. \tag{6.5}$$

d) The decoder may choose the wrong pair $(\tilde{v}_1, \tilde{v}_2)$. For this case we consider the event

$$\mathcal{E}_4(w_1, w_2, v_1, v_2) = \bigcup_{(\tilde{v}_1, \tilde{v}_2) \neq (v_1, v_2)} \{(X^n(w_1, \tilde{v}_1), Y^n(w_2, \tilde{v}_2)) \in T_\epsilon^n(P_{XY})\}. \tag{6.6}$$

The overall error event is $\bigcup_{i=1}^4 \mathcal{E}_i$ and using the union bound we have

$$P_e \leq \Pr[\mathcal{E}_1] + \Pr[\mathcal{E}_2 \cap \mathcal{E}_1^c] + \Pr[\mathcal{E}_3 \cap \mathcal{E}_1^c] + \Pr[\mathcal{E}_4]. \tag{6.7}$$

We next consider the four probabilities on the right-hand side of (6.7).

a) We already know that $\Pr[\mathcal{E}_1] \leq \delta_\epsilon(n, P_{XY})$.
b) The event $\mathcal{E}_1^c$ implies the event $\{X^n \in T_\epsilon^n(P_X)\}$. We use the Theorem on Total Expectation to write

$$\begin{aligned}
&\Pr[\mathcal{E}_2 \cap \mathcal{E}_1^c] \leq \Pr[\mathcal{E}_2 \cap \{X^n \in T_\epsilon^n(P_X)\}] \\
&= \sum_{x^n \in T_\epsilon(P_X)} P_X^n(x^n) \Pr[\mathcal{E}_2 | X^n = x^n] \\
&= \sum_{x^n \in T_\epsilon(P_X)} P_X^n(x^n) \Pr\left[\bigcap_{\tilde{w}_1, \tilde{v}_1} \{X^n(\tilde{w}_1, \tilde{v}_1) \neq x^n\} \middle| X^n = x^n\right].
\end{aligned} \tag{6.8}$$

We can remove the conditioning on $X^n = x^n$ because X^n and $X^n(\tilde{w}_1, \tilde{v}_1)$ are statistically independent for all $(\tilde{w}_1, \tilde{v}_1)$. Furthermore, the events in the intersection are mutually statistically independent because the $X^n(\tilde{w}_1, \tilde{v}_1)$ are mutually statistically independent for all $(\tilde{w}_1, \tilde{v}_1)$. The

probability of the intersection above is thus upper bounded by

$$\begin{aligned}&\Pr\left[X^n(\tilde{w}_1,\tilde{v}_1)\neq x^n\right]^{2^{n\left(R_1+R_1'\right)}}\\&=\left(1-P_X(x^n)\right)^{2^{n\left(R_1+R_1'\right)}}\\&\leq\exp\left(-2^{n[R_1+R_1'-H(X)(1+\epsilon)]}\right)\end{aligned}\tag{6.9}$$

where we have used $(1-\alpha)\leq e^{-\alpha}$. We may thus drive $\Pr\left[\mathcal{E}_2\cap\mathcal{E}_1^c\right]$ to zero by choosing large n and

$$R_1+R_1'>H(X)(1+\epsilon)\tag{6.10}$$

c) Similarly, $\Pr\left[\mathcal{E}_3\cap\mathcal{E}_1^c\right]$ becomes small by choosing large n and

$$R_2+R_2'>H(Y)(1+\epsilon).\tag{6.11}$$

d) We have

$$\begin{aligned}&\Pr\left[\mathcal{E}_4(W_1,W_2,V_1,V_2)\right]\\&=\sum_{w_1,w_2,v_1,v_2}\Pr\left[\mathcal{E}_4(w_1,w_2,v_1,v_2)\cap\{(W_1,W_2,V_1,V_2)=(w_1,w_2,v_1,v_2)\}\right]\\&=\sum_{w_1,w_2,v_1,v_2}\Pr\left[\mathcal{E}_4(w_1,w_2,v_1,v_2)\right]\\&\quad\cdot\Pr\left[(W_1,W_2,V_1,V_2)=(w_1,w_2,v_1,v_2)\,|\mathcal{E}_4(w_1,w_2,v_1,v_2)\right].\end{aligned}\tag{6.12}$$

We further have

$$\begin{aligned}&\Pr\left[\mathcal{E}_4(w_1,w_2,v_1,v_2)\right]\leq\Pr\left[\bigcup_{\tilde{v}_1,\tilde{v}_2}\{(X^n(w_1,\tilde{v}_1),Y^n(w_2,\tilde{v}_2))\in T_\epsilon^n(P_{XY})\}\right]\\&\overset{(a)}{\leq}\sum_{\tilde{v}_1,\tilde{v}_2}\Pr\left[(X^n(w_1,\tilde{v}_1),Y^n(w_2,\tilde{v}_2))\in T_\epsilon^n(P_{XY})\right]\\&\overset{(b)}{\leq}2^{n(R_1'+R_2')}2^{-n[I(X;Y)-3\epsilon H(XY)]}\end{aligned}\tag{6.13}$$

where (a) follows by the union bound and (b) follows by (4.46). Inserting (6.13) into (6.12), we find that $\Pr\left[\mathcal{E}_4(W_1,W_2,V_1,V_2)\right]$ approaches zero by choosing large n and

$$R_1'+R_2'<I(X;Y)-3\epsilon H(XY)\tag{6.14}$$

The bounds (6.10), (6.11), and (6.14) imply that we can choose large n and

$$R_1>H(X|Y)+4\epsilon H(XY)\tag{6.15}$$
$$R_2>H(Y|X)+4\epsilon H(XY)\tag{6.16}$$
$$R_1+R_2>H(XY)+5\epsilon H(XY)\tag{6.17}$$

and thereby ensure that all error events have small probability. We can thus

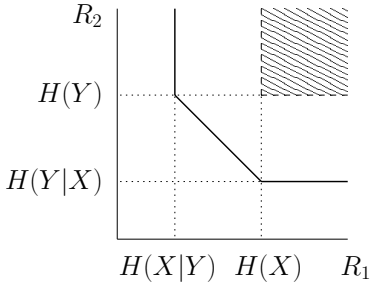

Figure 6.5.: The Slepian-Wolf source coding region.

approach the rate pairs (R_1, R_2) in the region

$$\boxed{\mathcal{R} = \left\{ (R_1, R_2) : \begin{array}{c} R_1 \geq H(X|Y) \\ R_2 \geq H(Y|X) \\ R_1 + R_2 \geq H(XY) \end{array} \right\}}. \tag{6.18}$$

The form of this region is depicted in Fig. 6.5. 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 Fig. 6.5. Note 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!

Finally, we remark again that we can find a specific code book, encoders, and a decoder, that achieves the above rate region. One can see this by expanding P_e as an average over the code books, and observing that at least one code book must give an error probability at most the average.

6.3. Example

Suppose $P_{XY}(\cdot)$ is defined via

$$Y = X \oplus Z \tag{6.19}$$

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

$$\begin{aligned} R_1 &\geq H_2(p) \\ R_2 &\geq H_2(p) \\ R_1 + R_2 &\geq 1 + H_2(p). \end{aligned} \tag{6.20}$$

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 and decoding.

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

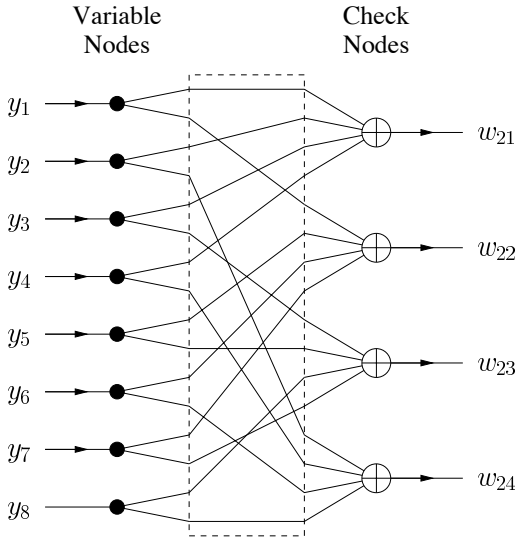

Figure 6.6.: A linear source encoder for binary y^n.

on the right by H^T where H is a $(n/2) \times n$ parity-check matrix of a binary linear code (we use matrix operations over the Galois field GF(2)). The encoding can be depicted in graphical form as shown in Fig. 6.6. 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 linear code to approach capacity on such a channel, and techniques for doing this are known [4, 5]. This example shows how *channel coding* techniques can be used to solve a *source coding* problem.

6.4. Converse

We show that the rates of (6.18) 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

$$\begin{aligned} nR_1 &\geq H(W_1) \\ &\geq H(W_1|Y^n) \\ &= H(W_1|Y^n) - H(W_1|X^nY^n) \\ &= I(X^n; W_1|Y^n) \\ &= H(X^n|Y^n) - H(X^n|Y^nW_1). \end{aligned} \tag{6.21}$$

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 chain of

inequalities as

$$\begin{aligned} nR_1 &\geq nH(X|Y) - H(X^n|Y^nW_1) \\ &= nH(X|Y) - H(X^n|Y^nW_1W_2\hat{X}^n\hat{Y}^n) \\ &\geq nH(X|Y) - H(X^nY^n|\hat{X}^n\hat{Y}^n) \\ &\geq nH(X|Y) - n\left[P_e \log_2(|\mathcal{X}| \cdot |\mathcal{Y}|) + H_2(P_e)/n\right] \end{aligned} \tag{6.22}$$

where the final step follows by 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

$$\begin{aligned} R_2 &\geq H(Y|X) - \left[P_e \log_2(|\mathcal{X}| \cdot |\mathcal{Y}|) + H_2(P_e)/n\right] \\ R_1 + R_2 &\geq H(XY) - \left[P_e \log_2(|\mathcal{X}| \cdot |\mathcal{Y}|) + H_2(P_e)/n\right]. \end{aligned} \tag{6.23}$$

This completes the converse.

6.5. Problems

6.1. Multiplicative Noise

Compute and plot the Slepian-Wolf region for the source $P_{XY}(\cdot)$ where $Y = X \cdot Z$, P_X and P_Z are independent binary symmetric sources, and "$\cdot$" denotes multiplication modulo-2.

6.2. Three Encoders

Suppose the source $P_{XYZ}(\cdot)$ puts out three sequences x^n, y^n, z^n, each of which is available to a different encoder. Encoders 1, 2, and 3 put out the respective indexes W_1, W_2, W_3 as functions of x^n, y^n, z^n. The three rates are R_1, R_2, R_3. The decoder receives all three indexes.

Show that the region of triples (R_1, R_2, R_3) at which the decoder can recover the three sequences with high probability is given by

$$\mathcal{R} = \left\{ (R_1, R_2, R_3) : \begin{array}{l} R_1 \geq H(X|YZ) \\ R_2 \geq H(Y|XZ) \\ R_3 \geq H(Z|XY) \\ R_1 + R_2 \geq H(XY|Z) \\ R_1 + R_3 \geq H(XZ|Y) \\ R_2 + R_3 \geq H(YZ|X) \\ R_1 + R_2 + R_3 \geq H(XYZ) \end{array} \right\}. \tag{6.24}$$

What do you expect the region for K sequences to be for $K > 3$?

References

[1] D. Slepian and J. K. Wolf. Noiseless coding of correlated information sources. *IEEE Trans. Inf. Theory*, 19(9):471–480, July 1973.

[2] T. Cover. A proof of the data compression theorem of Slepian and Wolf for ergodic sources. *IEEE Trans. Inf. Theory*, 21(2):226–228, March 1975.

[3] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, 1991.

[4] E. Arikan. Channel polarization: a method for constructing capacity-achieving codes for symmetric binary-input memoryless channels. *IEEE Trans. Inf. Theory*, 55(7):3051–3073, July 2009.

[5] T. J. Richardson, A. Shokrollahi, and R. L. Urbanke. Design of capacity-approaching low-density parity-check codes. *IEEE Trans. Inf. Theory*, 47(2):619–637, February 2001.

Chapter 7.

Multiaccess Channels

7.1. Problem Description

The multiaccess channel (MAC) with two transmitters and three sources is depicted in Fig. 7.1 and the FDG for $n = 3$ channel uses in Fig. 7.2. 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 $x_1^n \in \mathcal{X}_1^n$, encoder 2 maps (w_0, w_2) to $x_2^n \in \mathcal{X}_2^n$, and the channel $P_{Y|X_1X_2}(\cdot)$ puts out the sequence $y^n \in \mathcal{Y}^n$.

We now must be concerned with the *synchronization* of transmission since we require that y_i is a noisy function of x_{1i} and x_{2i} only. In other words, we are modeling the transmissions as taking place *synchronously*. We take the point of view that there is a central *clock* that governs the operation of the nodes. The clock ticks n times, and nodes 1 and 2 apply the respective inputs X_{1i} and X_{2i} to the channel at clock tick i. The receiving node sees its channel output Y_i at clock tick i, or perhaps shortly thereafter.

We remark that the common message W_0 might seem strange. One may view this message as being the clock information (in which case one may set $R_0 = 0$) or simply another message that is available to both nodes.

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

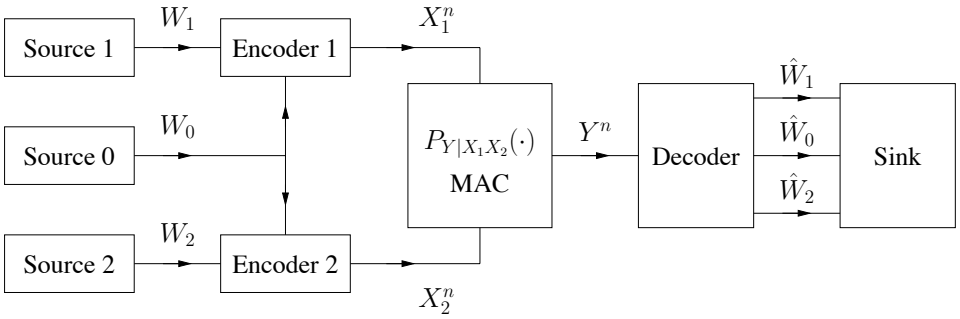

Figure 7.1.: The two-transmitter MAC with a common message.

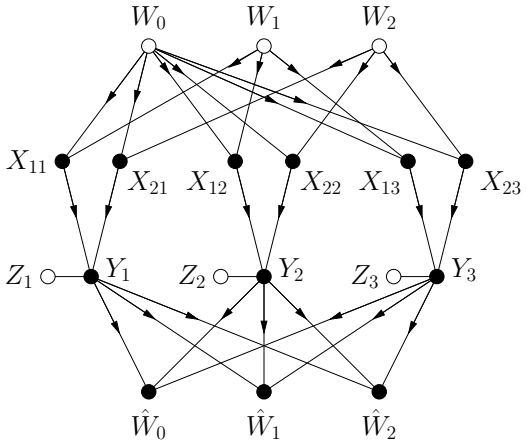

Figure 7.2.: FDG for the two-transmitter MAC with a common message.

can make

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

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*, the model of Fig. 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-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.

7.2. The MAC Capacity Region

The MAC was first considered by Shannon in [1, §17]. The capacity region of the MAC with $R_0 = 0$ was found by Ahlswede [2] and Liao [3]. The capacity region with $R_0 > 0$ was found by Slepian and Wolf [4], 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 Fig. 7.3). 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 $x_{1i}(w_0, w_1)$

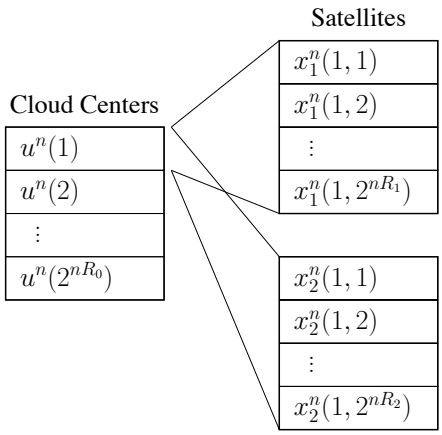

Figure 7.3.: A codebook for the MAC with a common message.

independently using $P_{X_1|U}(\cdot|u_i(w_0))$ for $i = 1, 2, \ldots, n$. 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

$$(u^n(\tilde{w}_0), x_1^n(\tilde{w}_0, \tilde{w}_1), x_2^n(\tilde{w}_0, \tilde{w}_2), y^n) \in T_\epsilon^n(P_{UX_1X_2Y}). \tag{7.2}$$

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: We know that, with probability close to one, we will have

$$(u^n(w_0), x_1^n(w_0, w_1), x_2^n(w_0, w_2), y^n) \in T_\epsilon^n(P_{UX_1X_2Y}) \tag{7.3}$$

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 for 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

$$(u^n(\tilde{w}_0), x_1^n(\tilde{w}_0, w_1), x_2^n(\tilde{w}_0, w_2), y^n) \in T_\epsilon^n(P_{UX_1X_2Y}). \tag{7.4}$$

Note that *all three* codewords in (7.4) were chosen independent of the actually transmitted codewords. We can upper bound the probability of the event (7.4) by

$$\sum_{\tilde{w}_0 \neq w_0} 2^{-n[I(UX_1X_2;Y)-\delta]} < 2^{n[R_0 - I(UX_1X_2;Y)+\delta]} \tag{7.5}$$

where $\delta \to 0$ as $\epsilon \to 0$. We leave the details of the remaining (and by now familiar) analysis to the reader, and simply state the seven rate bounds for

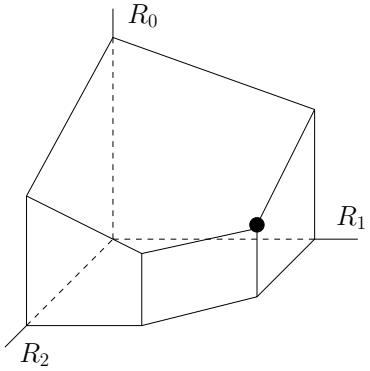

Figure 7.4.: The form of $\mathcal{R}(P_U, P_{X_1|U}, P_{X_2|U})$.

reliable communication:

$$R_0 \leq I(X_1X_2;Y) \tag{7.6}$$
$$R_0 + R_1 \leq I(X_1X_2;Y) \tag{7.7}$$
$$R_0 + R_2 \leq I(X_1X_2;Y) \tag{7.8}$$

and

$$R_1 \leq I(X_1;Y|X_2U) \tag{7.9}$$
$$R_2 \leq I(X_2;Y|X_1U) \tag{7.10}$$
$$R_1 + R_2 \leq I(X_1X_2;Y|U) \tag{7.11}$$
$$R_0 + R_1 + R_2 \leq I(X_1X_2;Y) \tag{7.12}$$

where $X_1 - U - X_2$ and $U - X_1X_2 - Y$ 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 (7.6)–(7.8) are redundant because of (7.12), so that we need consider only (7.9)–(7.12). One can further restrict attention to $|\mathcal{U}| \leq \min(|\mathcal{Y}| + 3, |\mathcal{X}_1| \cdot |\mathcal{X}_2| + 2)$ (see [5, p. 293 and p. 310–312], [6, Appendix B], [7, p. 18]).

The bounds (7.9)–(7.12) describe a region $\mathcal{R}(P_U, P_{X_1|U}, P_{X_2|U})$ with seven faces, four of which arise from (7.9)–(7.12), and three of which are non-negativity constraints on the rates (see Fig. 7.4). We can further achieve the union of such regions, i.e., we can achieve

$$\boxed{\mathcal{C}_{MAC} = \bigcup_{P_U, P_{X_1|U}, P_{X_2|U}} \mathcal{R}(P_U, P_{X_1|U}, P_{X_2|U})} \tag{7.13}$$

where $U - X_1X_2 - Y$ forms a Markov chain and $|\mathcal{U}| \leq \min(|\mathcal{Y}| + 3, |\mathcal{X}_1| \cdot |\mathcal{X}_2| + 2)$. We show below that (7.13) is the capacity region.

7.3. Converse

For reliable communication, the rate R_1 must satisfy

$$\begin{aligned} nR_1 &\leq I(W_1; Y^n) \\ &\leq I(W_1; Y^n W_0 W_2) \\ &= I(W_1; Y^n | W_0 W_2) \\ &= \sum_{i=1}^{n} H(Y_i|Y^{i-1}W_0W_2) - H(Y_i|Y^{i-1}W_0W_1W_2) \\ &= \sum_{i=1}^{n} H(Y_i|Y^{i-1}W_0W_2X_2^n) - H(Y_i|X_{1i}X_{2i}W_0) \\ &\leq \sum_{i=1}^{n} H(Y_i|X_{2i}W_0) - H(Y_i|X_{1i}X_{2i}W_0) \\ &= \sum_{i=1}^{n} I(X_{1i}; Y_i|X_{2i}W_0). \end{aligned} \tag{7.14}$$

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

$$P_U([a,i])P_{X_1|U}(b\,|\,[a,i])P_{X_2|U}(c\,|\,[a,i])P_{Y|X_1X_2}(d\,|\,b,c) \tag{7.15}$$

for all a, b, c, d. In other words, both $X_1 - U - X_2$ and $U - X_1X_2 - Y$ are Markov chains. We can now write the bound (7.14) as

$$R_1 \leq I(X_1; Y|X_2U). \tag{7.16}$$

We similarly have

$$R_2 \leq I(X_2; Y|X_1U) \tag{7.17}$$

$$R_1 + R_2 \leq I(X_1X_2; Y|U) \tag{7.18}$$

$$R_0 + R_1 + R_2 \leq I(X_1X_2; Y). \tag{7.19}$$

The expressions (7.15)–(7.19) specify that every achievable (R_0, R_1, R_2) must lie in $\mathcal{C}_{MAC}$. Thus, $\mathcal{C}_{MAC}$ is the capacity region.

We remark that $\mathcal{C}_{MAC}$ must be convex since time-sharing is permitted, i.e, one can use one codebook for some fraction of the time and another codebook for another fraction of the time. One can check that the union of regions (7.13) is indeed convex (see Problem 7.1).

7.4. Gaussian MAC

Consider the additive white Gaussian noise (AWGN) MAC with

$$Y = X_1 + X_2 + Z \tag{7.20}$$

where Z is Gaussian, zero mean, has variance N, and is independent of the real random variables X_1 and X_2. We impose the power (or energy) constraints $\sum_{i=1}^{n} \mathsf{E}[X_1^2] \leq nP_1$ and $\sum_{i=1}^{n} \mathsf{E}[X_2^2] \leq nP_2$.

Our first task is to show that the best choice for UX_1X_2 is jointly Gaussian. A natural approach is to replace the (perhaps non-Gaussian) UX_1X_2 with a Gaussian triple $U_GX_{1G}X_{2G}$ having the same covariance matrix $Q_{UX_1X_2}$. However, it turns out that $X_1 - U - X_2$ being a Markov chain does not imply that $X_{1G} - U_G - X_{2G}$ is a Markov chain. Hence it is not clear that the usual Gaussian substitution results in an achievable region.

Instead, consider $V = \mathsf{E}[X_1|U]$ that is a function of U and note that

$$h(Y|X_1U) \overset{(a)}{=} h(Y|X_1UV) \leq h(Y|X_1V) \tag{7.21}$$

$$h(Y|X_2U) \overset{(a)}{=} h(Y|X_2UV) \leq h(Y|X_2V) \tag{7.22}$$

where (a) follows because V is a function of U. We thus find that

a) replacing U with V does not shrink the rate region;

b) replacing VX_1X_2 with $V_GX_{1G}X_{2G}$ having the same covariance matrix $Q_{VX_1X_2}$ does not shrink the rate region either;

c) the chain $X_1 - V - X_2$ is not necessarily Markov but $X_{1G} - V_G - X_{2G}$ is Markov.

Summarizing, we find that Gaussian UX_1X_2 are optimal. So let U, V_1 and V_2 be independent, unit variance, Gaussian random variables, and define

$$X_1 = (\sqrt{P_1}\rho_1)\, U + \sqrt{P_1(1-\rho_1^2)}\, V_1 \tag{7.23}$$

$$X_2 = (\sqrt{P_2}\rho_2)\, U + \sqrt{P_2(1-\rho_2^2)}\, V_2. \tag{7.24}$$

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

$$I(X_1;Y|X_2U) = \frac{1}{2}\log\left(1 + \frac{P_1(1-\rho_1^2)}{N}\right) \tag{7.25}$$

$$I(X_2;Y|X_1U) = \frac{1}{2}\log\left(1 + \frac{P_2(1-\rho_2^2)}{N}\right) \tag{7.26}$$

$$I(X_1X_2;Y|U) = \frac{1}{2}\log\left(1 + \frac{P_1(1-\rho_1^2) + P_2(1-\rho_2^2)}{N}\right) \tag{7.27}$$

$$I(X_1X_2;Y) = \frac{1}{2}\log\left(1 + \frac{P_1 + P_2 + 2\sqrt{P_1P_2}\,\rho_1\rho_2}{N}\right). \tag{7.28}$$

$\mathcal{C}_{MAC}$ is found by considering all ρ_1 and ρ_2 with $0 \leq \rho_1 \leq 1$ and $0 \leq \rho_2 \leq 1$.

7.5. The MAC with $R_0 = 0$

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

$$\boxed{\mathcal{C}_{MAC} = \bigcup \left\{ (R_1, R_2): \begin{array}{r} 0 \le R_1 \le I(X_1; Y|X_2U) \\ 0 \le R_2 \le I(X_2; Y|X_1U) \\ R_1 + R_2 \le I(X_1X_2; Y|U) \end{array} \right\}} \tag{7.29}$$

where the union is over joint distributions that factor as

$$P_{UX_1X_2Y} = P_U \, P_{X_1|U} \, P_{X_2|U} \, P_{Y|X_1X_2}. \tag{7.30}$$

The third inequality in (7.29) follows from (7.11) and (7.12) with $R_0 = 0$; since $U - X_1X_2 - Y$ is Markov, $I(X_1X_2; Y|U) \le I(UX_1X_2; Y) = I(X_1X_2; Y)$ so inequality (7.12) is redundant when $R_0 = 0$.
Recall that we have $|\mathcal{U}| \le \min(|\mathcal{Y}| + 3, |\mathcal{X}_1| \cdot |\mathcal{X}_2| + 2)$. However, for $R_0 = 0$ it turns out that one can restrict attention to $|\mathcal{U}| \le 2$ [5, p. 278].

One often encounters the following equivalent formulation of $\mathcal{C}_{MAC}$:

$$\mathcal{R}_{MAC} = \mathrm{co}\left(\bigcup \left\{ (R_1, R_2): \begin{array}{r} 0 \le R_1 \le I(X_1; Y|X_2) \\ 0 \le R_2 \le I(X_2; Y|X_1) \\ R_1 + R_2 \le I(X_1X_2; Y) \end{array} \right\} \right) \tag{7.31}$$

where the union is over joint distributions that factor as

$$P_{X_1X_2Y} = P_{X_1} \, P_{X_2} \, P_{Y|X_1X_2} \tag{7.32}$$

and where $\mathrm{co}(\mathcal{S})$ is the convex hull of a set $\mathcal{S}$. Proving that $\mathcal{R}_{MAC} = \mathcal{C}_{MAC}$ requires some additional work, and we refer to [7, §3.5] for a discussion on this topic. Some authors prefer (7.31) for historical reasons, and because (7.31) has no U. Other authors prefer (7.29) 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 [5, pp. 288-290]).

Example 7.1. The *binary adder channel* or BAC has $\mathcal{X}_1 = \mathcal{X}_2 = \{0, 1\}$, $\mathcal{Y} = \{0, 1, 2\}$, and $Y = X_1 + X_2$. The channel is deterministic so that the mutual information expressions in (7.30) become conditional entropies. One can easily check that the best X_1 and X_2 are uniformly distributed and

$$\mathcal{C}_{MAC} = \left\{ (R_1, R_2): \begin{array}{r} 0 \le R_1 \le 1 \\ 0 \le R_2 \le 1 \\ R_1 + R_2 \le 1.5 \end{array} \right\}. \tag{7.33}$$

Example 7.2. 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

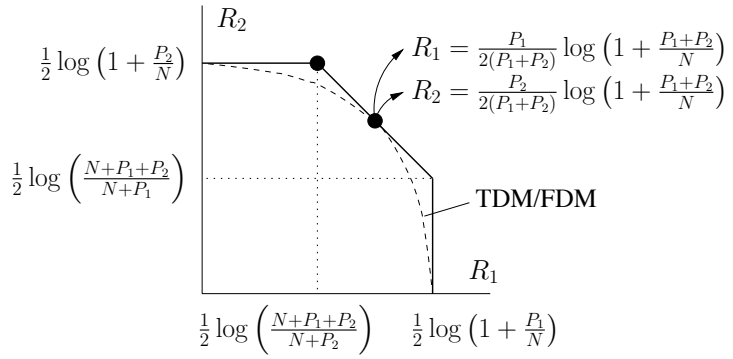

Figure 7.5.: $\mathcal{C}_{MAC}$ for the AWGN MAC with $P_1 \approx P_2$.

entropy theorem ensures that

$$\boxed{\mathcal{C}_{MAC} = \left\{ \begin{array}{r} 0 \le R_1 \le \frac{1}{2}\log\left(1+\frac{P_1}{N}\right) \\ (R_1, R_2): \ 0 \le R_2 \le \frac{1}{2}\log\left(1+\frac{P_2}{N}\right) \\ R_1 + R_2 \le \frac{1}{2}\log\left(1+\frac{P_+P_2}{N}\right) \end{array} \right\}}. \tag{7.34}$$

The resulting region is plotted in Fig. 7.5.

Example 7.3. An important 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

$$\begin{aligned} R_1 &= \frac{\alpha}{2}\log\left(1+\frac{P_1}{\alpha N}\right) \\ R_2 &= \frac{1-\alpha}{2}\log\left(1+\frac{P_2}{(1-\alpha)N}\right) \end{aligned} \tag{7.35}$$

where the transmitters boost their powers in their frequency bands. The resulting rate pairs are plotted in Fig. 7.5. In particular, by choosing $\alpha = P_1/(P_1+P_2)$ one achieves a boundary point with

$$R_1 + R_2 = \frac{1}{2}\log\left(1+\frac{P_1+P_2}{N}\right). \tag{7.36}$$

Example 7.4. The above example shows that TDM and FDM can be effective techniques for the MAC. However, Fig. 7.5 can be misleading because it is plotted with $P_1 \approx P_2$. Suppose instead that there is a 20 dB difference in the received powers, which we can model with a 20 dB difference in P_1 and P_2. For example, suppose that $P_1/N = 100$, $P_2/N = 1$, a situation that could very well occur in wireless problems. The resulting rate region (7.34) and TDM/FDM rates (7.35) are shown in Fig. 7.6.

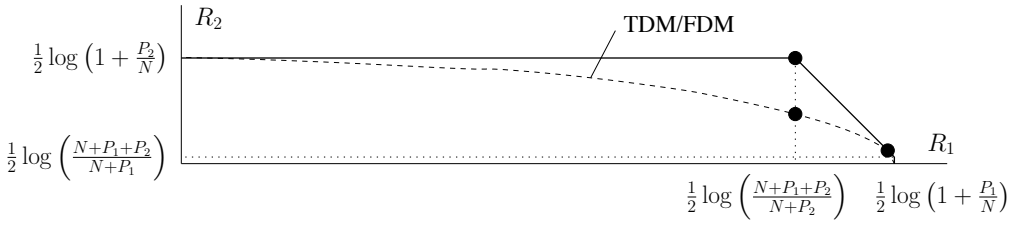

Figure 7.6.: $\mathcal{C}_{MAC}$ for the AWGN MAC with $P_1/N = 100$ and $P_2/N = 1$.

TDM/FDM achieves the sum-rate capacity at $(R_1, R_2) \approx (3.303, 0.033)$ with $\alpha = 100/101$ which makes R_2 is rather small. But if transmitter 1 is willing to give up 1/2 bit of its throughput, we can choose $\alpha = 0.8$ and the TDM/FDM rate is $(R_1, R_2) \approx (2.79, 0.26)$. Transmitter 2's rate has increased by about a factor of 8. The "user experience" of transmitter 2 at this point is much preferable over the maximum sum-rate point for many applications.

We can improve the situation further by using better codes and decoders, e.g., the methods described in Sec. 7.6. The corner point marked with a solid circle in Fig. 7.6 has $(R_1, R_2) \approx (2.84, 0.50)$ which almost doubles the rate of transmitter 2 over TDM/FDM.

7.6. Decoding Methods

7.6.1. Single-User Decoding and Rate-Splitting

The capacity expression (7.31) is suggestive for code design. Consider, e.g., the AWGN MAC and the marked corner points in Fig. 7.5 and Fig. 7.6. The decoder can proceed in two stages: first, decode w_1 by considering $x_2^n(w_2)$ as AWGN with variance P_2; second, subtract $x_1^n(w_1)$ from y^n and decode w_2. The capacities of the second and first channels are the respective

$$\begin{aligned} R_1 &= \frac{1}{2}\log\left(1+\frac{P_1}{N+P_2}\right) \\ R_2 &= \frac{1}{2}\log\left(1+\frac{P_2}{N}\right). \end{aligned} \tag{7.37}$$

This type of decoding is known as *single-user* decoding via *interference cancellation* (or also as *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 Fig. 7.5 and Fig. 7.6. 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 [9, 10]. The idea is to split encoder 1 into *two* encoders operating at the respective rates R_{11} and R_{12} with $R_1 = R_{11} + R_{12}$. Suppose these

encoders transmit with respective powers P_{11} and P_{12}, where $P_1 = P_{11}+P_{12}$, and that the output of the first transmitter is the sum of the two encoded signals. The decoder performs single-user decoding in three stages: first, decode the R_{11} code; second, decode the R_2 code; third, decode the R_{12} code. The rates are

$$R_1 = R_{11} + R_{12} = \frac{1}{2}\log\left(1 + \frac{P_{11}}{N + P_{12} + P_2}\right) + \frac{1}{2}\log\left(1 + \frac{P_{12}}{N}\right)$$
$$R_2 = \frac{1}{2}\log\left(1 + \frac{P_2}{N + P_{12}}\right)$$

Note that by choosing $P_{12} = 0$ we recover (7.37), while if we choose $P_{12} = P_1$ we obtain the other corner points of the pentagons in Fig. 7.5 and Fig. 7.6. By varying P_{12} from 0 to P_1, we thus achieve any rate point on the boundary of that face of the pentagon with maximum sum-rate.

7.6.2. 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 [11].

For example, suppose 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 Fig. 7.7. 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. This type of approach is also called *soft* interference cancellation.

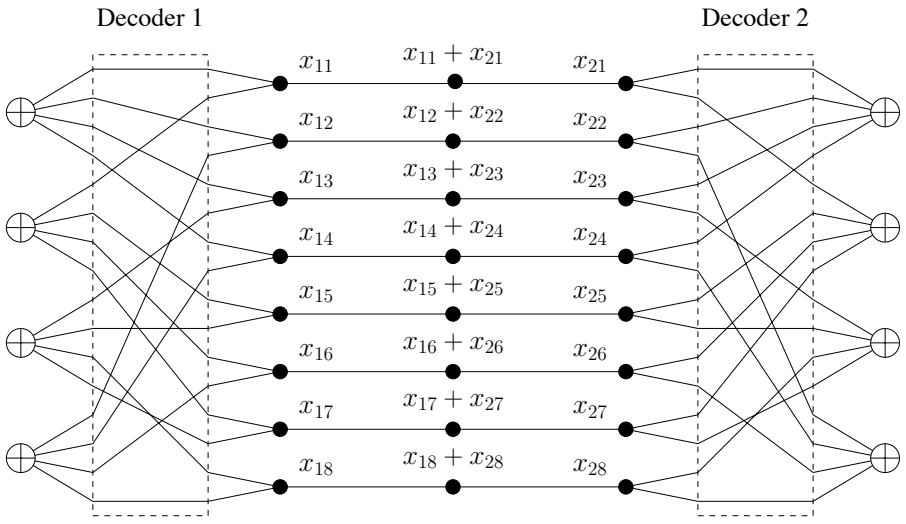

Figure 7.7.: Graph for an iterative joint decoder for the AWGN MAC.

7.7. Problems

7.1. $\mathcal{C}_{MAC}$ is Convex

Show that the region $\mathcal{C}_{MAC}$ in (7.13) is a convex set.

7.2. AWGN MAC and BC Capacity Regions

a) Consider the additive-white Gaussian noise (AWGN) multiaccess channel

$$Y = \sqrt{\frac{N}{N_1}}X_1 + \sqrt{\frac{N}{N_2}}X_2 + Z \tag{7.38}$$

where Z is zero-mean Gaussian noise with variance N, N_1 and N_2 are some positive numbers with $N_1 \leq N_2$, and Z is independent of X_1 and X_2. However, suppose now that there is a **sum** power constraint $\mathsf{E}[X_1^2 + X_2^2] \leq P$ (and **not** the constraints $\mathsf{E}[X_1^2] \leq P_1$ and $\mathsf{E}[X_2^2] \leq P_2$). Compute the two-dimensional capacity region when $R_0 = 0$.

b) Plot this region and explain how it is related to the capacity region of the AWGN broadcast channel with the noise variances $\mathsf{E}[Z_1^2] = N_1$ and $\mathsf{E}[Z_2^2] = N_2$, $N_1 \leq N_2$, and with the power constraint $\mathsf{E}[X^2] \leq P$.

References

[1] C. E. Shannon. Two-way communication channels. In J. Neyman, editor, *Proc. of 4th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 611–644. Berkeley, CA: University of California Press, 1961. Reprinted in *Claude Elwood Shannon: Collected Papers*, pp. 351–384, (N. J. A. Sloane and A. D. Wyner, eds.) Piscataway: IEEE Press, 1993.

[2] R. Ahlswede. Multi-way communication channels. In *Proc. 2nd Int. Symp. Inform. Theory (1971)*, pages 23–52. Tsahkadsor, Armenian S.S.R., Publishing House of the Hungarian Academy of Sciences, 1973.

[3] H. Liao. A coding theorem for multiple access communications. In *Proc. of IEEE International Symposium on Information Theory*. Asilomar, CA, 1972.

[4] D. Slepian and J. K. Wolf. A coding theorem for multiple access channels with correlated sources. *Bell Syst. Tech. J.*, 52:1037–1076, September 1973.

[5] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Channels*. Akadémiai Kiadó, Budapest, 1981.

[6] F. M. J. Willems and E. C. van der Meulen. The discrete memoryless multiple-access channel with cribbing encoders. *IEEE Trans. Inf. Theory*, 31(3):313–327, May 1985.

[7] F. M. J. Willems. *Informationtheoretical Results for the Discrete Memoryless Multiple Access Channel*. Doctor in de wetenschappen proefschrift, Katholieke Universiteit Leuven, Leuven, Belgium, October 1982.

[8] S. I. Bross, A. Lapidoth, and M. A. Wigger. The Gaussian MAC with conferencing encoders. In *Proc. IEEE Int. Symp. Inform. Theory*, Toronto, Canada, July 6-11 2008.

[9] B. Rimoldi and R. Urbanke. A rate-splitting approach to the Gaussian multiple-access channel. *IEEE Trans. Inf. Theory*, 42(2):364–375, March 1996.

[10] A. J. Grant, B. Rimoldi, R. L. Urbanke, and P. A. Whiting. Rate-splitting multiple-access for discrete memoryless channels,. *IEEE Trans. Inf. Theory*, 47(3):873–890, March 2001.

[11] 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 *Proc. IEEE Int. Symp. Inform. Theory*, page 334, Lausanne, Switzerland, June 30 - July 5 2002.

Appendix A.

Discrete Probability

A.1. Events, Sample Space, and Probability Measure

We begin with basic definitions. A discrete *sample space* $\Omega = \{\omega_1, \ldots, \omega_N\}$ 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 event Ω. 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 Fig. A.1). A fair die has $\Pr[\omega_i] = 1/6$ for all i. The probability of the event $\mathcal{A}$ is therefore $|\mathcal{A}|/|\Omega|$, where $|\mathcal{A}|$ is 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*. Equality holds in (A.5) if and only if $\Pr[\mathcal{A} \cap \mathcal{B}] = 0$ (this does not necessarily mean that $\mathcal{A} \cap \mathcal{B} = \emptyset$).

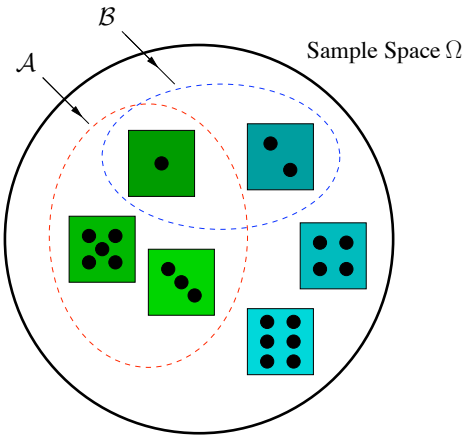

Figure A.1.: A sample space with six atomic events.

The events $\mathcal{A}$ and $\mathcal{B}$ are said to be *independent* if

$$\Pr[\mathcal{A} \cap \mathcal{B}] = \Pr[\mathcal{A}] \cdot \Pr[\mathcal{B}]. \tag{A.6}$$

More generally, the sets $\mathcal{A}_i$, $i = 1, 2, \ldots, n$, are independent if

$$\Pr\left[\bigcap_{i=1}^{n} \mathcal{A}_i\right] = \prod_{i=1}^{n} \Pr[\mathcal{A}_i]. \tag{A.7}$$

The *conditional* probability of the event $\mathcal{B}$ given the occurrence of the event $\mathcal{A}$ with $\Pr[\mathcal{A}] > 0$ is

$$\Pr[\mathcal{B}|\mathcal{A}] = \frac{\Pr[\mathcal{A} \cap \mathcal{B}]}{\Pr[\mathcal{A}]}. \tag{A.8}$$

Thus, if $\Pr[\mathcal{A}] > 0$ then using (A.8) the events $\mathcal{A}$ and $\mathcal{B}$ are independent if $\Pr[\mathcal{B}|\mathcal{A}] = \Pr[\mathcal{B}]$.[1]

Example A.2. Consider our fair die and the events $\mathcal{A} = \{1, 3, 5\}$ and $\mathcal{B} = \{1, 2\}$ in Fig. A.1. We find that (A.6) is satisfied so $\mathcal{A}$ and $\mathcal{B}$ are independent. We further have $\Pr[\mathcal{A}] > 0$ and compute

$$\Pr[\mathcal{B}|\mathcal{A}] = \frac{\Pr[\{1\}]}{\Pr[\mathcal{A}]} = \frac{1/6}{1/2} = \frac{1}{3}. \tag{A.9}$$

---

[1] The reader may now wonder what happens if $\Pr[\mathcal{A}] = 0$. We then have $\Pr[\mathcal{A} \cap \mathcal{B}] = 0$ so that (A.6) is satisfied. Thus, if $\Pr[\mathcal{A}] = 0$ then $\mathcal{A}$ and $\mathcal{B}$ are always 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 $X(\Omega)$ (we usually consider random variables with $X(\Omega) = \mathcal{X}$). The *preimage* (or inverse image) of a point a, $a \in \mathcal{X}$, is written as

$$X^{-1}(a) = \{\omega : X(\omega) = a\}. \tag{A.10}$$

More generally, for a subset $\mathcal{A}$ of $\mathcal{X}$ we write $X^{-1}(\mathcal{A}) = \{\omega : X(\omega) \in \mathcal{A}\}$. The *probability distribution* $P_X(\cdot)$ is a mapping from $X(\Omega)$ into the interval $[0, 1]$ such that

$$P_X(a) = \Pr\left[X^{-1}(a)\right] \tag{A.11}$$

or simply $P_X(a) = \Pr[X = a]$. We thus have

$$P_X(a) \geq 0 \quad \text{for all } a \in X(\Omega) \tag{A.12}$$

$$\sum_{a \in X(\Omega)} P_X(a) = 1. \tag{A.13}$$

Example A.3. Consider the sample space of Example A.1 and choose $\mathcal{X} = \{\text{odd}, \text{even}\}$. We define the mapping $X(\cdot)$ as follows:

$$X(1) = X(3) = X(5) = \text{odd}$$
$$X(2) = X(4) = X(6) = \text{even}.$$

We compute $P_X(\text{odd}) = P_X(\text{even}) = 1/2$.

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

$$P_{X^n}(a^n) = \Pr\left[\bigcap_{i=1}^{n} \{X_i = a_i\}\right]. \tag{A.14}$$

We thus have

$$P_{X^n}(a^n) \geq 0 \quad \text{for all } a^n \in X^n(\Omega) \tag{A.15}$$

$$\sum_{a^n \in X^n(\Omega)} P_{X^n}(a^n) = 1. \tag{A.16}$$

We further have

$$\begin{aligned} P_{X^{n-1}}(a^{n-1}) &= P_{X_1 X_2 \ldots X_{n-1}}(a_1, a_2, \ldots, a_{n-1}) \\ &= \sum_{a_n \in X_n(\Omega)} P_{X_1 X_2 \ldots X_{n-1} X_n}(a_1, a_2, \ldots, a_{n-1}, a_n). \end{aligned} \tag{A.17}$$

The *marginal* distributions of P_{X^n} are the distributions $P_{X_1}, P_{X_2}, \ldots, P_{X_n}$. The process of "removing" a random variable as in (A.17) by summing a joint distribution over the range of the random variable is called *marginalization*.

The *support* of a random variable X is the set

$$\operatorname{supp}(P_X) = \{a : a \in \mathcal{X}, P_X(a) > 0\}. \tag{A.18}$$

The *conditional* probability distribution $P_{Y|X}(\cdot)$ is a mapping from $X(\Omega) \times Y(\Omega)$ into the interval $[0, 1]$ such that

$$P_{Y|X}(b|a) = \begin{cases} \frac{P_{XY}(a,b)}{P_X(a)}, & P_X(a) > 0 \\ \text{undefined}, & \text{else.} \end{cases} \tag{A.19}$$

The value "undefined" is sometimes chosen as a number in the interval $[0, 1]$. An alternative (but indirect) way of defining $P_{Y|X}(\cdot)$ is as *any* function from $X(\Omega) \times Y(\Omega)$ into the interval $[0, 1]$ such that

$$P_{Y|X}(b|a)P_X(a) = P_{XY}(a, b) \tag{A.20}$$

for all $(a, b) \in X(\Omega) \times Y(\Omega)$. This alternative approach recovers the usual definition if $P_X(a) > 0$, and it motivates choosing $P_{Y|X}(b|a)$ to be "undefined" otherwise.

A.3. Independent Random Variables

The random variables $X_1, X_2, \ldots, X_n$ are *statistically independent* if

$$P_{X^n}(a^n) = \prod_{i=1}^{n} P_{X_i}(a_i) \quad \text{for all } a^n \in X^n(\Omega). \tag{A.21}$$

Similarly, $X_1, X_2, \ldots, X_n$ are statistically independent conditioned on the event $\mathcal{A}$ with $\Pr[\mathcal{A}] > 0$ if, for all $a^n \in X^n(\Omega)$, we have

$$\Pr\left[\bigcap_{i=1}^{n}\{X_i = a_i\}\middle|\,\mathcal{A}\right] = \prod_{i=1}^{n}\Pr\left[X_i = a_i|\mathcal{A}\right]. \tag{A.22}$$

Thus, using (A.21) we find that X and Y are statistically independent if and only if

$$P_{Y|X}(b|a) = P_Y(b) \quad \text{for all } (a,b) \in \operatorname{supp}(P_X) \times Y(\Omega). \tag{A.23}$$

Similarly, we say that X and Y are statistically independent conditioned on Z if

$$P_{XY|Z}(a,b|c) = P_{X|Z}(a|c)P_{Y|Z}(b|c) \tag{A.24}$$

for all $(a,b,c) \in X(\Omega) \times Y(\Omega) \times \operatorname{supp}(P_Z)$. Thus, we find that X and Y are statistically independent conditioned on Z if and only if

$$P_{Y|XZ}(b|a,c) = P_{Y|Z}(b|c) \tag{A.25}$$

for all $(a,b,c) \in \operatorname{supp}(P_X) \times Y(\Omega) \times \operatorname{supp}(P_Z)$. Alternatively, X and Y are statistically independent conditioned on Z if and only if

$$P_{X|YZ}(a|b,c) = P_{X|Z}(a|c) \tag{A.26}$$

for all $(a,b,c) \in X(\Omega) \times \operatorname{supp}(P_{YZ})$.

A common way of expressing that X and Y are statistically independent given Z is to say that

$$X - Z - Y \tag{A.27}$$

forms a Markov chain.

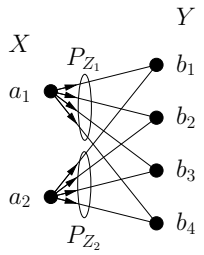

Figure A.2.: Graph representation of $P_{XY}(\cdot)$. The statistics of Z_i depend on i but $Z = Z_1 Z_2$ is statistically independent of X.

A.4. Probabilistic Dependence via Functional Dependence

Consider a joint distribution $P_{XY}(\cdot)$. We claim that there is a random variable $\tilde{Y}$ having $P_{X\tilde{Y}}(\cdot) = P_{XY}(\cdot)$ and where

$$\tilde{Y} = f(X, Z) \tag{A.28}$$

for some function $f(\cdot)$ and some random variable Z that is statistically independent of X and that has alphabet $\mathcal{Z}$ of size at most $|\mathcal{Z}| = |\mathcal{X}| \cdot |\mathcal{Y}|$.

Suppose $\mathcal{X} = \{a_1, a_2, \ldots, a_{|\mathcal{X}|}\}$ and $\mathcal{Y} = \{b_1, b_2, \ldots, b_{|\mathcal{Y}|}\}$. Consider the graph representation of $P_{XY}(\cdot)$, as depicted in Fig. A.2 for $|\mathcal{X}| = 2$ and $|\mathcal{Y}| = 4$. Let Z be a word with $|\mathcal{X}|$ letters whose ith letter Z_i, $i = 1, 2, \ldots, |\mathcal{X}|$, takes on the value b_j with probability $P_{Z_i}(b_j) = P_{Y|X}(b_j|a_i)$ as long as $P_X(a_i) > 0$. If $P_X(a_i) = 0$ then we leave $P_{Z_i}(\cdot)$ unspecified. We define the function $\text{index}(a_i) = i$, $i = 1, 2, \ldots, |\mathcal{X}|$, and choose Z independent of X (the Z_i, $i = 1, 2, \ldots, |\mathcal{X}|$, could be dependent). We claim that the function

$$\tilde{Y} = Z_{\text{index}(X)} \tag{A.29}$$

makes $X\tilde{Y}$ have the joint distribution $P_{XY}(\cdot)$. Indeed, by construction have

$$P_{\tilde{Y}|X}(b_j|a_i) = P_{Z_i}(b_j) = P_{Y|X}(b_j|a_i). \tag{A.30}$$

The purpose of the above exercise is to show that we may as well represent any channel $P_{Y|X}(\cdot)$ by a functional relation $Y = f(X, Z)$ where the "noise" Z is independent of the channel input X. This result forms the basis of the ideas in the next section.

Example A.4. Consider binary X and Y for which

$$\begin{aligned} P_{XY}(0,0) &= p_{00}, \quad P_{XY}(0,1) = p_{01}, \\ P_{XY}(1,0) &= p_{10}, \quad P_{XY}(1,1) = p_{11}. \end{aligned} \tag{A.31}$$

We define $a_1 = 0, a_2 = 1$, $b_1 = 0, b_2 = 1$, and $Z = Z_1 Z_2$ where

$$\begin{aligned} P_{Z_1}(0) &= P_{Y|X}(0|0) = \tfrac{p_{00}}{p_{00}+p_{01}} \\ P_{Z_2}(0) &= P_{Y|X}(0|1) = \tfrac{p_{10}}{p_{10}+p_{11}}. \end{aligned} \tag{A.32}$$

We choose $\tilde{Y} = Z_1$ if $X = 0$ and $\tilde{Y} = Z_2$ if $X = 1$.

Example A.5. Consider Example A.4 with $p_{00} = p_{11}$ and $p_{01} = p_{10}$. The symmetry lets us simplify $f(\cdot)$ and Z. We may define $\tilde{Y} = X \oplus Z$ where $P_X(0) = P_X(1) = 1/2$, Z is independent of X, and $P_Z(0) = p_{00} + p_{11}$, $P_Z(1) = 1 - P_Z(0)$. We see that $|\mathcal{Z}| = 2$ suffices in this case.

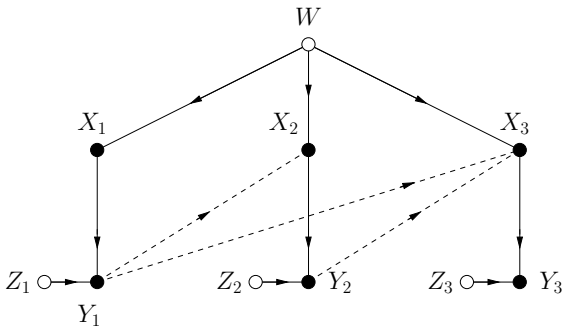

Figure A.3.: FDG for a memoryless channel with feedback.

A.5. Establishing Conditional Statistical Independence

Graphs can help us to understand relationships between random variables and to prove conditional statistical independence results. We will use *functional dependence graphs* or FDGs.

An FDG has vertices that represent random variables and edges that represent the functional dependencies between the random variables. For instance, suppose we have N_{RV} random variables that are defined by S_{RV} independent (source) random variables by N_{RV} functions. The corresponding FDG $\mathcal{G}$ is a directed graph having $N_{RV} + S_{RV}$ vertices and where 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.6. Fig. A.3 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 representing the independent W, Z_1, Z_2 and Z_3 are distinguished by drawing them with a hollow circle.

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.

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.*

a) 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 $\mathcal{Z}$.*

b) In $\mathcal{G}_{\mathcal{XYZ}}$ delete all edges coming out *of the vertices in $\mathcal{Z}$. Call the resulting graph $\mathcal{G}_{\mathcal{XY}|\mathcal{Z}}$.*

c) Remove the arrows on the remaining edges of $\mathcal{G}_{\mathcal{XY}|\mathcal{Z}}$ to obtain an undirected graph.

A fundamental result of [1, 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 $P_{\underline{X}\,\underline{Y}|\underline{Z}} = P_{\underline{X}|\underline{Z}} P_{\underline{Y}|\underline{Z}}$. This is the same as saying that $\underline{X} - \underline{Z} - \underline{Y}$ forms a Markov chain.

Example A.7. Consider Fig. A.3 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 $W - X_1 X_2 - Y_2$ forms a Markov chain.

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 [2, Ch. 2],[3]. For fd-separation, after the second step above one removes all edges coming out of vertices to which there is no path (in a directed sense) from the S_{RV} source vertices. 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 [2, Ch. 2]).

A.6. Expectation

Expectation is an integral and usually involves continuous real-valued random variables. However, (real-valued) discrete random variables have all the important properties we shall need for the continuous cases.

Consider a real-valued function $f(\cdot)$ with domain $X(\Omega)$. The *expectation* of the random variable $Y = f(X)$ is

$$\mathsf{E}[Y] = \mathsf{E}[f(X)] = \sum_{a \in \mathrm{supp}(P_X)} P_X(a)\, f(a). \tag{A.33}$$

One sometimes encounters the notation $\mathsf{E}_X[\cdot]$ if it is unclear which of the letters in the argument of $\mathsf{E}[\cdot]$ are random variables.

Example A.8. The random variable $Y = f(X) = 1(X = a)$ has

$$\mathsf{E}[Y] = P_X(a). \tag{A.34}$$

Similarly, $Y = f(X) = 1(X \in \mathcal{A})$ for $\mathcal{A} \subseteq X(\Omega)$ has

$$\mathsf{E}[Y] = \sum_{a \in \mathcal{A}} P_X(a). \tag{A.35}$$

The *conditional* expectation of $f(X)$ given that the event $\mathcal{A}$ with $\Pr[\mathcal{A}] > 0$ occurred is

$$\mathsf{E}[f(X)|\mathcal{A}] = \sum_{a:\Pr[\{X=a\}\cap\mathcal{A}]>0} \Pr[X = a|\mathcal{A}]\ f(a) \tag{A.36}$$

where the conditional probability $\Pr[X = a|\mathcal{A}]$ is defined as in (A.8). In particular, if $\mathcal{A} = \{Z = c\}$ and $P_Z(c) > 0$ we have

$$\mathsf{E}[f(X)|Z = c] = \sum_{a \in \mathrm{supp}(P_{X|Z}(\cdot|c))} P_{X|Z}(a|c)\, f(a). \tag{A.37}$$

We can re-write the above definitions slightly differently. 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

$$\bigcup_{m=1}^{M} \mathcal{B}_m = \Omega \quad \text{and} \quad \mathcal{B}_i \cap \mathcal{B}_j = \emptyset,\ i \neq j. \tag{A.38}$$

We can then write (A.33) as

$$\begin{aligned}\mathsf{E}[f(X)] &= \sum_{i,a:\Pr[\mathcal{B}_i\cap\{X=a\}]>0} \Pr\left[\mathcal{B}_i \cap \{X=a\}\right]\, f(a) \\ &= \sum_{i:\Pr[\mathcal{B}_i]>0} \Pr\left[\mathcal{B}_i\right] \sum_{a:\Pr[\mathcal{B}_i\cap\{X=a\}]>0} \frac{\Pr\left[\mathcal{B}_i \cap \{X=a\}\right]}{\Pr\left[\mathcal{B}_i\right]}\, f(a) \\ &= \sum_{i:\Pr[\mathcal{B}_i]>0} \Pr\left[\mathcal{B}_i\right] \sum_{a:\Pr[\mathcal{B}_i\cap\{X=a\}]>0} \Pr\left[X=a|\mathcal{B}_i\right]\, f(a) \\ &= \sum_{i:\Pr[\mathcal{B}_i]>0} \Pr\left[\mathcal{B}_i\right] \mathsf{E}[f(X)|\mathcal{B}_i] \end{aligned} \tag{A.39}$$

and (A.36) as

$$\mathsf{E}[f(X)|\mathcal{A}] = \sum_{i:\Pr[\mathcal{B}_i\cap\mathcal{A}]>0} \Pr\left[\mathcal{B}_i|\mathcal{A}\right]\, \mathsf{E}[f(X)|\mathcal{B}_i\cap\mathcal{A}]\,. \tag{A.40}$$

Example A.9. For a discrete random variable Y we can choose $\mathcal{B}_b = \{Y = b\}$ and write

$$\mathsf{E}[f(X)] = \sum_{b\in\mathrm{supp}(P_Y)} P_Y(b)\, \mathsf{E}[f(X)|Y=b] \tag{A.41}$$

$$\mathsf{E}[f(X)|\mathcal{A}] = \sum_{b:\Pr[\{Y=b\}\cap\mathcal{A}]>0} \Pr\left[Y=b|\mathcal{A}\right]\, \mathsf{E}[f(X)|\{Y=b\}\cap\mathcal{A}]\,. \tag{A.42}$$

The identities (A.39)-(A.42) are called the *Theorem on Total Expectation.*

A.7. Second-Order Statistics for Scalars

The m'th moment, $m = 1, 2, 3, \ldots$, of a real-valued random variable Y is $\mathsf{E}[Y^m]$. The *variance* of Y is

$$\mathsf{Var}[Y] = \mathsf{E}\Big[(Y - \mathsf{E}[Y])^2\Big] = \mathsf{E}\Big[Y^2\Big] - \mathsf{E}[Y]^2 \,. \tag{A.43}$$

The *covariance* of real-valued X and Y is

$$\begin{aligned} \mathsf{Cov}[X, Y] &= \mathsf{E}[(X - \mathsf{E}[X])(Y - \mathsf{E}[Y])] \\ &= \mathsf{E}[XY] - \mathsf{E}[X]\,\mathsf{E}[Y] \,. \end{aligned} \tag{A.44}$$

We thus have $\mathsf{Var}[X] = \mathsf{Cov}[X, X]$. We say that X and Y are *uncorrelated* if $\mathsf{Cov}[X, Y] = 0$ or, alternatively, if $\mathsf{E}[XY] = \mathsf{E}[X]\,\mathsf{E}[Y]$.

Example A.10. Statistically independent X and Y are uncorrelated.

Example A.11. Consider random variables $X_1, X_2, \ldots, X_n$. We compute

$$\mathsf{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathsf{Cov}[X_i, X_j] \,. \tag{A.45}$$

If the $\{X_i\}_{i=1}^n$ are pairwise uncorrelated then (A.45) is $\sum_{i=1}^n \mathsf{Var}[X_i]$.

The *correlation coefficient* of X and Y is the normalized covariance

$$\rho = \frac{\mathsf{E}[(X - \mathsf{E}[X])(Y - \mathsf{E}[Y])]}{\sqrt{\mathsf{Var}[X]\,\mathsf{Var}[Y]}} \tag{A.46}$$

as long as the variances are positive; otherwise we say $\rho = 0$. One can show that $-1 \le \rho \le 1$.

A.8. Second-Order Statistics for Vectors

The definitions of the previous section extend naturally to random vectors. The *covariance matrix* of the real-valued column vectors $\underline{X} = [X_1 \ \ldots \ X_\ell]^T$ and $\underline{Y} = [Y_1 \ \ldots \ Y_m]^T$ is

$$\begin{aligned}\mathbf{Cov}[\underline{X}, \underline{Y}] &= \mathsf{E}\Big[(\underline{X} - \mathsf{E}[\underline{X}])(\underline{Y} - \mathsf{E}[\underline{Y}])^T\Big] \\ &= \mathsf{E}\Big[\underline{X}\,\underline{Y}^T\Big] - \mathsf{E}[\underline{X}]\,\mathsf{E}[\underline{Y}]^T\,. \end{aligned} \tag{A.47}$$

The covariance matrix of $\underline{X}$ is $\mathbf{Q}_{\underline{X}} = \mathbf{Cov}[\underline{X}, \underline{X}]$. We say that $\underline{X}$ and $\underline{Y}$ are *uncorrelated* if $\mathbf{Cov}[\underline{X}, \underline{Y}] = 0$.

Example A.12. For the random vectors $\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_n$ we compute

$$\mathbf{Q}_{\sum_{i=1}^n \underline{X}_i} = \sum_{i=1}^n \sum_{j=1}^n \mathbf{Cov}\Big[\underline{X}_i, \underline{X}_j\Big]\,. \tag{A.48}$$

If the $\{\underline{X}_i\}_{i=1}^n$ are pairwise uncorrelated then (A.48) is $\sum_{i=1}^n \mathbf{Q}_{\underline{X}_i}$.

The *correlation matrix* of $\underline{X}$ and $\underline{Y}$ is the normalized covariance matrix

$$\mathbf{R} = \mathsf{E}\Big[\tilde{\underline{X}}\,\tilde{\underline{Y}}^T\Big] \tag{A.49}$$

where the ith entry of $\tilde{\underline{X}}$ is

$$\tilde{X}_i = \frac{X_i - \mathsf{E}[X_i]}{\sqrt{\mathsf{Var}[X_i]}} \tag{A.50}$$

as long as the variance is positive; otherwise we set $\tilde{X}_i = 0$. The entries of $\tilde{\underline{Y}}$ are similarly defined. In other words, the entry of row i and column j of $\mathbf{R}$ is simply the correlation coefficient of X_i and Y_j.

A.9. Conditional Expectation Random Variables

The conditional expectation (A.37) motivates defining the random variable $\mathsf{E}[Y|X]$ that takes on the value $\mathsf{E}[Y|X=a]$ when $X=a$. $\mathsf{E}[Y|X]$ is therefore a deterministic function of X and we have

$$P_{\mathsf{E}[Y|X]}(b) = \sum_{a:\,\mathsf{E}[Y|X=a]=b} P_X(a). \tag{A.51}$$

We similarly define $\mathsf{Var}[Y|X]$ as the random variable that takes on the value $\mathsf{Var}[Y|X=a]$ when $X=a$.

We next list several simple properties of conditional expectation.

- $\mathsf{E}[f(X)|X] = f(X)$
- $\mathsf{E}[\mathsf{E}[Y|X]] = \mathsf{E}[Y]$ and more generally $\mathsf{E}[\mathsf{E}[Y|X]\,|Z] = \mathsf{E}[Y|Z]$
- $\mathsf{E}[YZ] = \mathsf{E}[Y\mathsf{E}[Z|Y]]$ and more generally $\mathsf{E}[YZ|X] = \mathsf{E}[Y\mathsf{E}[Z|XY]\,|X]$
- If $Y - X - Z$ forms a Markov chain then $\mathsf{E}[Z|XY] = \mathsf{E}[Z|X]$
- If $Y - X - Z$ forms a Markov chain then $\mathsf{E}[YZ|X] = \mathsf{E}[Y|X]\,\mathsf{E}[Z|X]$
- $\mathsf{Var}[Y|X] = \mathsf{E}[Y^2|X] - \mathsf{E}[Y|X]^2$ and $\mathsf{E}[\mathsf{Var}[Y|X]] = \mathsf{E}[Y^2] - \mathsf{E}\left[\mathsf{E}[Y|X]^2\right]$

An important property of conditional expectation concerns *minimum-mean square-error* (MMSE) estimation. Consider the error $S = X - \hat{X}$ and suppose that, given $Y=y$, we wish to find the estimate $\hat{X}(y)$ that minimizes

$$\mathsf{Var}[S|Y=y] = \mathsf{E}\left[\left(X - \hat{X}(y)\right)^2 \middle| Y=y\right]. \tag{A.52}$$

A simple optimization gives $\hat{X}(y) = \mathsf{E}[X|Y=y]$ and the MMSE is

$$\mathsf{Var}[S|Y=y] = \mathsf{E}\left[X^2|Y=y\right] - \mathsf{E}[X|Y=y]^2. \tag{A.53}$$

Hence, the MMSE estimate is the random variable

$$\boxed{\hat{X}(Y) = \mathsf{E}[X|Y]} \tag{A.54}$$

and the MMSE is the random variable $\mathsf{Var}[S|Y] = \mathsf{E}[X^2|Y] - \mathsf{E}[X|Y]^2$. We caution that $\mathsf{E}[\mathsf{Var}[S|Y]] \neq \mathsf{Var}[S]$ in general. In fact, using Jensen's inequality and the strict convexity of the function $f(x) = x^2$, we have

$$\begin{aligned} \mathsf{E}[\mathsf{Var}[S|Y]] &= \mathsf{E}\left[X^2\right] - \mathsf{E}\left[\mathsf{E}[X|Y]^2\right] \\ &\leq \mathsf{E}\left[X^2\right] - \mathsf{E}[X]^2 \\ &= \mathsf{Var}[S] \end{aligned} \tag{A.55}$$

with equality if and only if $\mathsf{E}[X|Y]$ is the constant $\mathsf{E}[X]$.

A.10. Linear Estimation

This section is concerned with *linear* minimum-mean square-error (LMMSE) estimators. Consider the zero-mean random variables X and Y. (For non-zero mean X and Y, simply subtract the means to give zero-mean random variables. One may add $\mathsf{E}[X]$ to the final estimate without affecting the estimator performance.) Given Y, we estimate $\hat{X} = c\,Y$ where c is chosen to minimize

$$\mathsf{E}\left[\left(X - \hat{X}\right)^2\right]. \tag{A.56}$$

Simple calculations give

$$c = \mathsf{E}[XY]\,/\mathsf{E}\left[Y^2\right] \tag{A.57}$$

$$\mathsf{E}\left[\left(X - \hat{X}\right)^2\right] = \mathsf{E}\left[X^2\right] - \mathsf{E}\left[\hat{X}^2\right] \tag{A.58}$$

$$\mathsf{E}\left[\left(X - \hat{X}\right) \cdot Y\right] \stackrel{(a)}{=} 0 \tag{A.59}$$

and (a) is called the *orthogonality principle*.

More generally, for (zero-mean column) vectors the LMMSE estimator is $\underline{\hat{X}} = \mathbf{C}\,\underline{Y}$ where $\mathbf{C}$ minimizes

$$\mathsf{E}\left[\left\|\underline{X} - \underline{\hat{X}}\right\|^2\right]. \tag{A.60}$$

Suppose that $\underline{X}$ and $\underline{Y}$ have zero mean and $\mathbf{Q}_{\underline{Y}} = \mathsf{E}\left[\underline{Y}\,\underline{Y}^T\right]$ is invertible. We compute

$$\mathbf{C} = \mathsf{E}\left[\underline{X}\,\underline{Y}^T\right]\mathbf{Q}_{\underline{Y}}^{-1} \tag{A.61}$$

$$\mathsf{E}\left[\left\|\underline{X} - \underline{\hat{X}}\right\|^2\right] = \mathsf{E}\left[\|\underline{X}\|^2\right] - \mathsf{E}\left[\left\|\underline{\hat{X}}\right\|^2\right] \tag{A.62}$$

$$\mathsf{E}\left[\left(\underline{X} - \underline{\hat{X}}\right) \cdot \underline{Y}^T\right] \stackrel{(a)}{=} \mathbf{0} \tag{A.63}$$

and (a) is again called the *orthogonality principle*.

Summarizing, the vector LMMSE estimator can be written as

$$\boxed{\underline{\hat{X}}(\underline{Y}) = \mathsf{E}\left[\underline{X}\,\underline{Y}^T\right]\mathbf{Q}_{\underline{Y}}^{-1}\,\underline{Y}}. \tag{A.64}$$

A.11. Markov Inequalities

We state and prove several useful inequalities.

Theorem A.2. (Markov Inequality) Let X be a *non-negative* real-valued random variable with mean $\mathsf{E}[X]$. For $a > 0$, we have

$$\Pr[X \geq a] \leq \frac{\mathsf{E}[X]}{a}. \tag{A.65}$$

Proof. We have $\Pr[X \geq a] = \mathsf{E}[1(X \geq a)]$, 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 $a\Pr[X \geq a] = \mathsf{E}[a\,1(X \geq a)] \leq \mathsf{E}[X]$. □

Example A.13. Suppose we set $X = |Y - \mathsf{E}[Y]|$. Markov's inequality then gives *Tchebycheff's* inequality

$$\begin{aligned}\Pr[|Y - \mathsf{E}[Y]| \geq a] &= \Pr\left[|Y - \mathsf{E}[Y]|^2 \geq a^2\right] \\ &\leq \frac{\mathsf{Var}[Y]}{a^2}\end{aligned} \tag{A.66}$$

where $\mathsf{Var}[Y]$ is the variance of Y and $a > 0$.

Example A.14. Suppose we set $X = e^{\nu Y}$ and $a = e^{\nu b}$. Markov's inequality then gives the *Chernoff bounds*

$$\begin{aligned}\Pr[Y \geq b] &\leq \mathsf{E}\left[e^{\nu Y}\right] e^{-\nu b} \quad \text{for } \nu \geq 0 \\ \Pr[Y \leq b] &\leq \mathsf{E}\left[e^{\nu Y}\right] e^{-\nu b} \quad \text{for } \nu \leq 0.\end{aligned} \tag{A.67}$$

A.12. Jensen Inequalities

A real-valued function $f(\cdot)$ with domain an 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

$$f(x) \geq f(x_0) + m(x - x_0) \quad \text{for all } x \in \mathcal{I}. \tag{A.68}$$

The convexity is *strict* if the inequality (A.68) is strict whenever $x \neq x_0$. 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

$$f\left(\lambda x_1 + (1-\lambda)x_2\right) \leq \lambda f(x_1) + (1-\lambda) f(x_2) \quad \text{for } 0 < \lambda < 1. \tag{A.69}$$

We say that $f(\cdot)$ is *concave* (or convex-$\cap$) on $\mathcal{I}$ if $-f(\cdot)$ is convex on $\mathcal{I}$.

Theorem A.3. (Jensen's Inequality) Let X be a real-valued random variable taking values in $\mathcal{I}$ and suppose $f(\cdot)$ is convex on $\mathcal{I}$. We have

$$f(\mathsf{E}[X]) \leq \mathsf{E}[f(X)]\,. \tag{A.70}$$

Similarly, if $f(\cdot)$ is concave on $\mathcal{I}$ then we have

$$f(\mathsf{E}[X]) \geq \mathsf{E}[f(X)]\,. \tag{A.71}$$

If $f(\cdot)$ is strictly convex (or concave) then equality holds in (A.70) (or (A.71)) if and only if X is a constant.

Instead of proving Theorem A.3, we prove a more general result for vectors. Consider a real-valued function $f(\cdot)$ whose domain is a non-empty convex set $\mathcal{S}$ in the n-dimensional vector space $\mathbb{R}^n$. We say that $f(\cdot)$ is *convex* (or convex-$\cup$) on $\mathcal{S}$ if, for every interior point $\underline{x}_0$ of $\mathcal{S}$, there exists a real vector $\underline{m}$ (that may depend on $\underline{x}_0$) such that

$$f(\underline{x}) \geq f(\underline{x}_0) + \underline{m}^T(\underline{x} - \underline{x}_0) \quad \text{for all } \underline{x} \in \mathcal{S}. \tag{A.72}$$

The convexity is *strict* if the inequality (A.72) is strict whenever $\underline{x} \neq \underline{x}_0$. An alternative and equivalent definition is that $f(\cdot)$ is convex on $\mathcal{S}$ if for every $\underline{x}_1$ and $\underline{x}_2$ in $\mathcal{S}$ we have

$$f\left(\lambda \underline{x}_1 + (1-\lambda)\underline{x}_2\right) \leq \lambda f(\underline{x}_1) + (1-\lambda) f(\underline{x}_2) \quad \text{for } 0 < \lambda < 1. \tag{A.73}$$

We say that $f(\cdot)$ is *concave* (or convex-$\cap$) on $\mathcal{S}$ if $-f(\cdot)$ is convex on $\mathcal{S}$.

Theorem A.4. (Jensen's Inequality for Vectors) Let $\underline{X}$ be a vector-valued random variable taking values in $\mathcal{S}$ and suppose $f(\cdot)$ is convex on $\mathcal{S}$. We have

$$f(\mathsf{E}[\underline{X}]) \leq \mathsf{E}[f(\underline{X})]\,. \tag{A.74}$$

Similarly, if $f(\cdot)$ is concave on $\mathcal{S}$ then we have

$$f(\mathsf{E}[\underline{X}]) \geq \mathsf{E}[f(\underline{X})]\,. \tag{A.75}$$

If $f(\cdot)$ is strictly convex (or concave) then equality holds in (A.74) (or (A.75)) if and only if $\underline{X}$ is a constant.

Proof. Choose $\underline{x}_0 = \mathsf{E}[\underline{X}]$ in (A.72), choose an $\underline{m}$ that satisfies (A.72) for this $\underline{x}_0$, replace $\underline{x}$ with the random variable $\underline{X}$, and take expectations of both sides of (A.72). The result is (A.74). If $f(\cdot)$ is concave on $\mathcal{S}$, then we similarly obtain (A.75). Furthermore, if $f(\cdot)$ is strictly convex (or concave), equality holds in (A.74) (or (A.75)) if and only if $\underline{X}$ is a constant. □

A.13. Weak Law of Large Numbers

Laws of large numbers are concerned with n repeated trials of the same random experiment. Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (i.i.d.) real-valued random variables with $P_{X_i}(\cdot) = P_X(\cdot)$ for all i. The *sample mean* S_n is

$$S_n = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{A.76}$$

We clearly have $\mathsf{E}[S_n] = \mathsf{E}[X]$ and $\mathsf{Var}[S_n] = \mathsf{Var}[X]/n$. Applying Tchebycheff's inequality (A.66), we have

$$\Pr\left[|S_n - \mathsf{E}[X]| \geq \epsilon\right] \leq \frac{\mathsf{Var}[X]}{n\epsilon^2}. \tag{A.77}$$

Alternatively, we have

$$\boxed{\Pr\left[|S_n - \mathsf{E}[X]| < \epsilon\right] \geq 1 - \frac{\mathsf{Var}[X]}{n\epsilon^2}} \tag{A.78}$$

which is a quantitative version of the *Weak Law of Large Numbers*. The Weak Law thus states that the sample mean of n independent samples of X is almost certain to be near $\mathsf{E}[X]$ when n is large. Qualitatively, we can write (A.78) as a "limit in probability":

$$\boxed{\operatorname*{plim}_{n\to\infty} S_n = \mathsf{E}[X]}. \tag{A.79}$$

Example A.15. Consider a (perhaps non-real) random variable X with alphabet $\mathcal{X} = \{a_1, a_2, \ldots, a_{|\mathcal{X}|}\}$. Let $1(\cdot)$ be the *indicator* function that is 1 if its argument is true and is 0 otherwise. Let $Y = 1(X_i = a_j)$ for some $1 \leq j \leq |\mathcal{X}|$ for which we compute

$$\mathsf{E}[Y] = P_X(a_j) \tag{A.80}$$

$$\mathsf{Var}[Y] = P_X(a_j)\left(1 - P_X(a_j)\right). \tag{A.81}$$

Now consider n independent trials $X_1, X_2, \ldots, X_n$ of X which generate n independent trials of real-valued $Y_1, Y_2, \ldots, Y_n$ of Y. We form the sum (A.76) with Y_i replacing X_i. Note that this sum is $1/n$ multiplying the number $N(a_j|X^n)$ of times that the letter a_j occurs in X^n. Using the weak law (A.78) we find that

$$\Pr\left[\left|\frac{N(a_j|X^n)}{n} - P_X(a_j)\right| < \epsilon\right] \geq 1 - \frac{P_X(a_j)\left(1 - P_X(a_j)\right)}{n\epsilon^2}. \tag{A.82}$$

In other words, for large n the letter a_j occurs about $nP_X(a_j)$ times in X^n.

A.14. Strong Law of Large Numbers

The Weak Law of Large Numbers required only the Tchebycheff bound that is based on $\mathsf{Var}[X]$. We next wish to derive a stronger result that concerns the semi-infinite sequence of sums $S_1, S_2, S_3, \ldots$. The Chernoff bound (A.67) with $Y = S_n - \mathsf{E}[X]$ gives

$$\begin{aligned}\Pr\left[(S_n - \mathsf{E}[X]) \geq \epsilon\right] &\leq \mathsf{E}\left[e^{\nu(S_n - \mathsf{E}[X])}\right] e^{-\nu\epsilon} \\ &\overset{(a)}{=} \mathsf{E}\left[e^{(\nu/n)(X - \mathsf{E}[X])}\right]^n e^{-\nu\epsilon} \end{aligned} \tag{A.83}$$

where $\nu \geq 0$ and (a) follows by the independence of the trials. We should now optimize over ν but instead choose $\nu = n\,\delta$ for some small constant δ. Suppose the magnitudes of all moments of X are bounded from above as follows:

$$\left|\mathsf{E}\left[(X - \mathsf{E}[X])^i\right]\right| \leq m^i \tag{A.84}$$

for $i = 2, 3, 4, \ldots$ and for some positive m. The right-hand side of (A.83) is

$$\begin{aligned}\mathsf{E}\left[e^{(\nu/n)(X - \mathsf{E}[X])}\right]^n e^{-\nu\epsilon} &= \mathsf{E}\left[1 + \sum_{i=2}^{\infty} \frac{\delta^i}{i!} (X - \mathsf{E}[X])^i\right]^n e^{-n\delta\epsilon} \\ &\leq \left(1 + \sum_{i=2}^{\infty} \frac{\delta^i}{i!} m^i\right)^n e^{-n\delta\epsilon} \\ &= \left(e^{m\delta} - m\delta\right)^n e^{-n\delta\epsilon} \\ &= e^{n\left[\log(e^{m\delta} - m\delta) - \delta\epsilon\right]}. \end{aligned} \tag{A.85}$$

The term in square brackets in (A.85) evaluates to zero at $\delta = 0$ and its derivative with respect to δ is $-\epsilon$ at $\delta = 0$. Hence we can find a small δ for which the term in square brackets is negative. We thus find that

$$\Pr\left[(S_n - \mathsf{E}[X]) \geq \epsilon\right] \leq \beta^n \tag{A.86}$$

where β depends on ϵ and m and satisfies $0 < \beta < 1$. Combining (A.86) with a similar bound on $\Pr\left[S_n - \mathsf{E}[X] \leq -\epsilon\right]$ we have

$$\Pr\left[|S_n - \mathsf{E}[X]\,| \geq \epsilon\right] \leq 2\beta^n. \tag{A.87}$$

Now consider the following result on the behavior of $\{S_i\}_{i=1}^{\infty}$ for $i \geq n$:

$$\begin{aligned}\Pr\left[\sup_{i \geq n} |S_i - \mathsf{E}[X]\,| \geq \epsilon\right] &= \Pr\left[\bigcup_{i \geq n} \{|S_i - \mathsf{E}[X]\,| \geq \epsilon\}\right] \\ &\overset{(a)}{\leq} \sum_{i=n}^{\infty} \Pr\left[|S_n - \mathsf{E}[X]\,| \geq \epsilon\right] \\ &\overset{(b)}{\leq} \frac{2\beta^n}{1 - \beta} \end{aligned} \tag{A.88}$$

where (a) follows by the union bound, and (b) follows by the Chernoff bound (A.87). Note that step (b) does not work with the Tchebycheff bound (A.77) because the probabilities decrease only as $1/n$ rather than exponentially with n. The bound (A.88) implies the *Strong Law of Large Numbers*, namely

$$\boxed{\lim_{n\to\infty} \Pr\left[\sup_{i\geq n} |S_i - \mathsf{E}[X]\,| < \epsilon\right] = 1}. \tag{A.89}$$

A.15. Problems

A.1. Borel-Kolmogorov "Paradox"

Consider a 3-dimensional sphere centered at the origin and label the points on its surface by using the spherical coordinates longitude ϕ, $-\pi \leq \phi < \pi$, and latitude θ, $-\pi/2 \leq \theta \leq \pi/2$. Suppose X is a point that is uniformly distributed on the sphere, i.e., we write $X(\phi, \theta)$ and consider the joint density

$$p_{\Phi\Theta}(\phi, \theta) = \frac{1}{4\pi} \cos\theta. \tag{A.90}$$

a) Determine the marginal densities $p_\Phi(\cdot)$ and $p_\Theta(\cdot)$.

b) Determine the density of a point on the great circle defined by $\theta = 0$, i.e., determine $p_{\Phi|\Theta}(\cdot|0)$.

c) Determine the density of a point on the great circle defined by $\phi = 0$, i.e., determine $p_{\Theta|\Phi}(\cdot|0)$. Compare your answer to the previous result and interpret.

A.2. Strong Law with Fourth Central Moment Constraint

This exercise proves the strong law of large numbers but replaces (A.84) with the weaker constraint that X has a finite fourth central moment

$$\mathsf{E}\Big[(X - \mathsf{E}[X])^4\Big] \leq m \tag{A.91}$$

for some non-negative m.

a) Use Jensen's inequality to show that

$$\mathsf{Var}[X]^2 \leq \mathsf{E}\Big[(X - \mathsf{E}[X])^4\Big] \tag{A.92}$$

so that (A.91) implies $\mathsf{Var}[X] \leq \sqrt{m}$.

b) Use the Markov inequality to show that

$$\Pr\left[(S_n - \mathsf{E}[X]) \geq \epsilon\right] \leq \frac{\mathsf{E}[(S_n - \mathsf{E}[X])^4]}{\epsilon^4}. \tag{A.93}$$

c) Show by direct computation or induction that for *zero-mean* and i.i.d. $Y_1, Y_2, Y_3, \ldots$ with common distribution P_Y, we have

$$\mathsf{E}\left[\left(\sum_{i=1}^{n} Y_i\right)^4\right] = n\mathsf{E}\left[Y^4\right] + 3n(n-1)\mathsf{E}\left[Y^2\right]^2. \tag{A.94}$$

d) Now use $Y = X - \mathsf{E}[X]$ and show that

$$\mathsf{E}\left[(S_n - \mathsf{E}[X])^4\right] \leq \frac{(3n^2 - 2n)m}{n^4} \leq \frac{3m}{n^2}. \tag{A.95}$$

e) Insert (A.95) into (A.93) and modify the steps (A.86)-(A.88) to show that

$$\Pr\left[\sup_{i \geq n} |S_i - \mathsf{E}[X]| \geq \epsilon\right] \leq \sum_{i=n}^{\infty} \frac{3m}{i^2\epsilon^4} \leq \frac{3m}{\epsilon^4} \int_n^{\infty} \frac{1}{x^2} dx = \frac{3m}{n\epsilon^4} \tag{A.96}$$

Explain why the result (A.96) proves the strong law of large numbers.

References

[1] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988.

[2] G. Kramer. *Directed Information for Channels with Feedback*. Hartung-Gorre Verlag, Konstanz, Germany, 1998. ETH Series in Information Processing, Vol. 11.

[3] G. Kramer and S. A. Savari. Edge-cut bounds on network coding rates. *Journal of Network and Systems Management*, 14(1):49–67, March 2006.