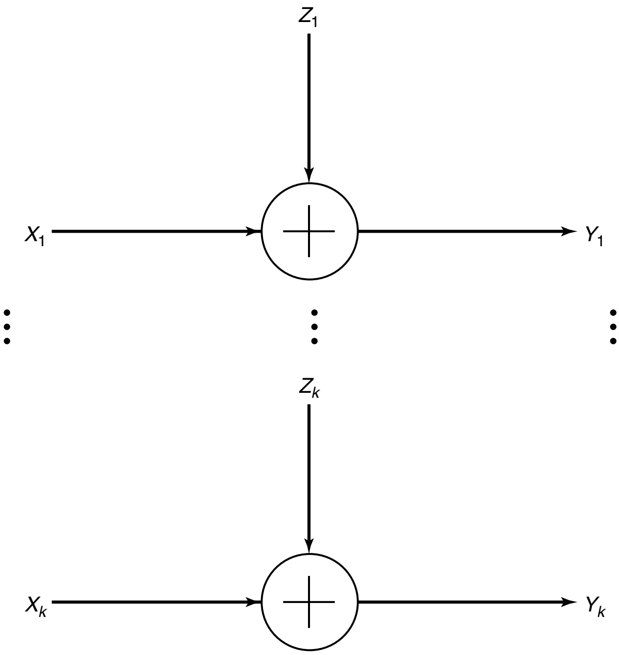


FIGURE 9.3. Parallel Gaussian channels.

We calculate the distribution that achieves the information capacity for this channel. The fact that the information capacity is the supremum of achievable rates can be proved by methods identical to those in the proof of the capacity theorem for single Gaussian channels and will be omitted.

Since $Z_1, Z_2, \ldots, Z_k$ are independent,

$$\begin{aligned}
&I(X_1, X_2, \ldots, X_k;\, Y_1, Y_2, \ldots, Y_k) \\
&\quad = h(Y_1, Y_2, \ldots, Y_k) - h(Y_1, Y_2, \ldots, Y_k|X_1, X_2, \ldots, X_k) \\
&\quad = h(Y_1, Y_2, \ldots, Y_k) - h(Z_1, Z_2, \ldots, Z_k|X_1, X_2, \ldots, X_k) \\
&\quad = h(Y_1, Y_2, \ldots, Y_k) - h(Z_1, Z_2, \ldots, Z_k) && (9.68) \\
&\quad = h(Y_1, Y_2, \ldots, Y_k) - \sum_i h(Z_i) && (9.69) \\
&\quad \le \sum_i h(Y_i) - h(Z_i) && (9.70) \\
&\quad \le \sum_i \frac{1}{2} \log\left(1 + \frac{P_i}{N_i}\right), && (9.71)
\end{aligned}$$

where $P_i = EX_i^2$, and $\sum P_i = P$. Equality is achieved by

$$(X_1, X_2, \ldots, X_k) \sim \mathcal{N}\left(0, \begin{bmatrix} P_1 & 0 & \cdots & 0 \\ 0 & P_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & P_k \end{bmatrix}\right). \tag{9.72}$$

So the problem is reduced to finding the power allotment that maximizes the capacity subject to the constraint that $\sum P_i = P$. This is a standard optimization problem and can be solved using Lagrange multipliers. Writing the functional as

$$J(P_1, \ldots, P_k) = \sum \frac{1}{2} \log\left(1 + \frac{P_i}{N_i}\right) + \lambda \left(\sum P_i\right) \tag{9.73}$$

and differentiating with respect to P_i, we have

$$\frac{1}{2} \frac{1}{P_i + N_i} + \lambda = 0 \tag{9.74}$$

or

$$P_i = \nu - N_i. \tag{9.75}$$

However, since the P_i's must be nonnegative, it may not always be possible to find a solution of this form. In this case, we use the Kuhn–Tucker conditions to verify that the solution

$$P_i = (\nu - N_i)^+ \tag{9.76}$$

is the assignment that maximizes capacity, where ν is chosen so that

$$\sum (\nu - N_i)^+ = P. \tag{9.77}$$

Here $(x)^+$ denotes the positive part of x:

$$(x)^+ = \begin{cases} x & \text{if } x \geq 0, \\ 0 & \text{if } x < 0. \end{cases} \tag{9.78}$$

This solution is illustrated graphically in Figure 9.4. The vertical levels indicate the noise levels in the various channels. As the signal power is increased from zero, we allot the power to the channels with the lowest

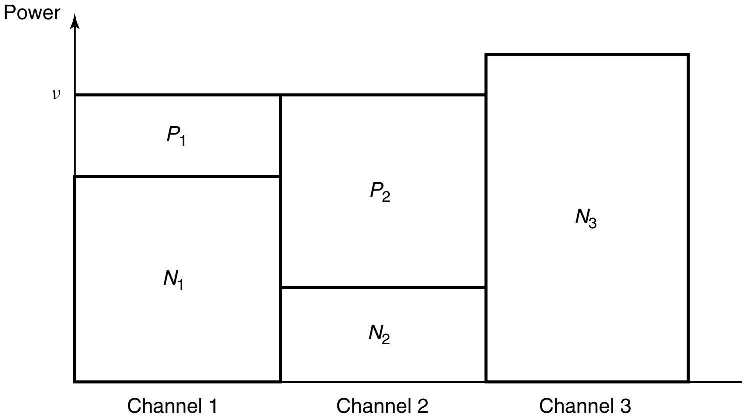


FIGURE 9.4. Water-filling for parallel channels.

noise. When the available power is increased still further, some of the power is put into noisier channels. The process by which the power is distributed among the various bins is identical to the way in which water distributes itself in a vessel, hence this process is sometimes referred to as *water-filling*.

9.5 CHANNELS WITH COLORED GAUSSIAN NOISE

In Section 9.4, we considered the case of a set of parallel independent Gaussian channels in which the noise samples from different channels were independent. Now we will consider the case when the noise is dependent. This represents not only the case of parallel channels, but also the case when the channel has Gaussian noise with memory. For channels with memory, we can consider a block of n consecutive uses of the channel as n channels in parallel with dependent noise. As in Section 9.4, we will calculate only the information capacity for this channel.

Let K_Z be the covariance matrix of the noise, and let K_X be the input covariance matrix. The power constraint on the input can then be written as

$$\frac{1}{n}\sum_i EX_i^2 \le P, \tag{9.79}$$

or equivalently,

$$\frac{1}{n}\mathrm{tr}(K_X) \le P. \tag{9.80}$$

Unlike Section 9.4, the power constraint here depends on n; the capacity will have to be calculated for each n.

Just as in the case of independent channels, we can write

$$I(X_1, X_2, \ldots, X_n; Y_1, Y_2, \ldots, Y_n) = h(Y_1, Y_2, \ldots, Y_n) - h(Z_1, Z_2, \ldots, Z_n). \tag{9.81}$$

Here $h(Z_1, Z_2, \ldots, Z_n)$ is determined only by the distribution of the noise and is not dependent on the choice of input distribution. So finding the capacity amounts to maximizing $h(Y_1, Y_2, \ldots, Y_n)$. The entropy of the output is maximized when Y is normal, which is achieved when the input is normal. Since the input and the noise are independent, the covariance of the output Y is $K_Y = K_X + K_Z$ and the entropy is

$$h(Y_1, Y_2, \ldots, Y_n) = \frac{1}{2} \log \left((2\pi e)^n |K_X + K_Z|\right). \tag{9.82}$$

Now the problem is reduced to choosing K_X so as to maximize $|K_X + K_Z|$, subject to a trace constraint on K_X. To do this, we decompose K_Z into its diagonal form,

$$K_Z = Q\Lambda Q^t, \quad \text{where } QQ^t = I. \tag{9.83}$$

Then

$$|K_X + K_Z| = |K_X + Q\Lambda Q^t| \tag{9.84}$$

$$= |Q||Q^t K_X Q + \Lambda||Q^t| \tag{9.85}$$

$$= |Q^t K_X Q + \Lambda| \tag{9.86}$$

$$= |A + \Lambda|, \tag{9.87}$$

where $A = Q^t K_X Q$. Since for any matrices B and C,

$$\operatorname{tr}(BC) = \operatorname{tr}(CB), \tag{9.88}$$

we have

$$\operatorname{tr}(A) = \operatorname{tr}(Q^t K_X Q) \tag{9.89}$$

$$= \operatorname{tr}(QQ^t K_X) \tag{9.90}$$

$$= \operatorname{tr}(K_X). \tag{9.91}$$

Now the problem is reduced to maximizing $|A + \Lambda|$ subject to a trace constraint $\text{tr}(A) \le nP$.

Now we apply Hadamard's inequality, mentioned in Chapter 8. Hadamard's inequality states that the determinant of any positive definite matrix K is less than the product of its diagonal elements, that is,

$$|K| \le \prod_i K_{ii} \tag{9.92}$$

with equality iff the matrix is diagonal. Thus,

$$|A + \Lambda| \le \prod_i (A_{ii} + \lambda_i) \tag{9.93}$$

with equality iff A is diagonal. Since A is subject to a trace constraint,

$$\frac{1}{n} \sum_i A_{ii} \le P, \tag{9.94}$$

and $A_{ii} \ge 0$, the maximum value of $\prod_i (A_{ii} + \lambda_i)$ is attained when

$$A_{ii} + \lambda_i = \nu. \tag{9.95}$$

However, given the constraints, it may not always be possible to satisfy this equation with positive A_{ii}. In such cases, we can show by the standard Kuhn–Tucker conditions that the optimum solution corresponds to setting

$$A_{ii} = (\nu - \lambda_i)^+, \tag{9.96}$$

where the water level ν is chosen so that $\sum A_{ii} = nP$. This value of A maximizes the entropy of Y and hence the mutual information. We can use Figure 9.4 to see the connection between the methods described above and water-filling.

Consider a channel in which the additive Gaussian noise is a stochastic process with finite-dimensional covariance matrix $K_Z^{(n)}$. If the process is stationary, the covariance matrix is Toeplitz and the density of eigenvalues on the real line tends to the power spectrum of the stochastic process [262]. In this case, the above water-filling argument translates to water-filling in the spectral domain.

Hence, for channels in which the noise forms a stationary stochastic process, the input signal should be chosen to be a Gaussian process with a spectrum that is large at frequencies where the noise spectrum is small.

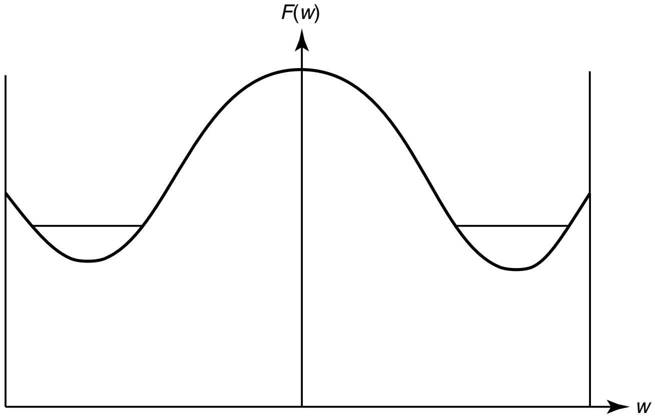


FIGURE 9.5. Water-filling in the spectral domain.

This is illustrated in Figure 9.5. The capacity of an additive Gaussian noise channel with noise power spectrum $N(f)$ can be shown to be [233]

$$C = \int_{-\pi}^{\pi} \frac{1}{2} \log \left(1 + \frac{(\nu - N(f))^+}{N(f)}\right) df, \tag{9.97}$$

where ν is chosen so that $\int (\nu - N(f))^+ \, df = P$.

9.6 GAUSSIAN CHANNELS WITH FEEDBACK

In Chapter 7 we proved that feedback does not increase the capacity for discrete memoryless channels, although it can help greatly in reducing the complexity of encoding or decoding. The same is true of an additive noise channel with white noise. As in the discrete case, feedback does not increase capacity for memoryless Gaussian channels.

However, for channels with memory, where the noise is correlated from time instant to time instant, feedback does increase capacity. The capacity without feedback can be calculated using water-filling, but we do not have a simple explicit characterization of the capacity with feedback. In this section we describe an expression for the capacity in terms of the covariance matrix of the noise Z. We prove a converse for this expression for capacity. We then derive a simple bound on the increase in capacity due to feedback.

The Gaussian channel with feedback is illustrated in Figure 9.6. The output of the channel Y_i is

$$Y_i = X_i + Z_i, \quad Z_i \sim \mathcal{N}(0, K_Z^{(n)}). \tag{9.98}$$

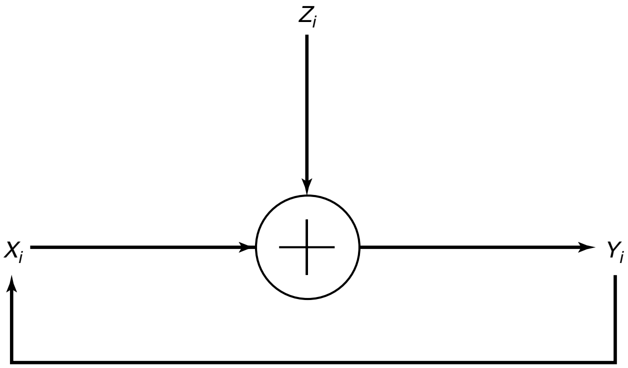


FIGURE 9.6. Gaussian channel with feedback.

The feedback allows the input of the channel to depend on the past values of the output.

A $(2^{nR}, n)$ code for the Gaussian channel with feedback consists of a sequence of mappings $x_i(W, Y^{i-1})$, where $W \in \{1, 2, \ldots, 2^{nR}\}$ is the input message and Y^{i-1} is the sequence of past values of the output. Thus, $x(W, \cdot)$ is a code function rather than a codeword. In addition, we require that the code satisfy a power constraint,

$$E\left[\frac{1}{n}\sum_{i=1}^{n} x_i^2(w, Y^{i-1})\right] \le P, \quad w \in \{1, 2, \ldots, 2^{nR}\}, \tag{9.99}$$

where the expectation is over all possible noise sequences.

We characterize the capacity of the Gaussian channel is terms of the covariance matrices of the input X and the noise Z. Because of the feedback, X^n and Z^n are not independent; X_i depends causally on the past values of Z. In the next section we prove a converse for the Gaussian channel with feedback and show that we achieve capacity if we take X to be Gaussian.

We now state an informal characterization of the capacity of the channel with and without feedback.

1. *With feedback*. The capacity $C_{n,\text{FB}}$ in bits per transmission of the time-varying Gaussian channel with feedback is

$$C_{n,\text{FB}} = \max_{\frac{1}{n}\text{tr}(K_X^{(n)})\le P} \frac{1}{2n} \log \frac{|K_{X+Z}^{(n)}|}{|K_Z^{(n)}|}, \tag{9.100}$$

where the maximization is taken over all X^n of the form

$$X_i = \sum_{j=1}^{i-1} b_{ij} Z_j + V_i, \qquad i = 1, 2, \ldots, n, \tag{9.101}$$

and V^n is independent of Z^n. To verify that the maximization over (9.101) involves no loss of generality, note that the distribution on $X^n + Z^n$ achieving the maximum entropy is Gaussian. Since Z^n is also Gaussian, it can be verified that a jointly Gaussian distribution on $(X^n, Z^n, X^n + Z^n)$ achieves the maximization in (9.100). But since $Z^n = Y^n - X^n$, the most general jointly normal causal dependence of X^n on Y^n is of the form (9.101), where V^n plays the role of the innovations process. Recasting (9.100) and (9.101) using $X = BZ + V$ and $Y = X + Z$, we can write

$$C_{n,\text{FB}} = \max \frac{1}{2n} \log \frac{|(B+I)K_Z^{(n)}(B+I)^t + K_V|}{|K_Z^{(n)}|}, \tag{9.102}$$

where the maximum is taken over all nonnegative definite K_V and strictly lower triangular B such that

$$\operatorname{tr}(BK_Z^{(n)}B^t + K_V) \le nP. \tag{9.103}$$

Note that B is 0 if feedback is not allowed.

2. *Without feedback*. The capacity C_n of the time-varying Gaussian channel without feedback is given by

$$C_n = \max_{\frac{1}{n}\operatorname{tr}(K_X^{(n)}) \le P} \frac{1}{2n} \log \frac{|K_X^{(n)} + K_Z^{(n)}|}{|K_Z^{(n)}|}. \tag{9.104}$$

This reduces to water-filling on the eigenvalues $\{\lambda_i^{(n)}\}$ of $K_Z^{(n)}$. Thus,

$$C_n = \frac{1}{2n} \sum_{i=1}^{n} \log\left(1 + \frac{(\lambda - \lambda_i^{(n)})^+}{\lambda_i^{(n)}}\right), \tag{9.105}$$

where $(y)^+ = \max\{y, 0\}$ and where λ is chosen so that

$$\sum_{i=1}^{n} (\lambda - \lambda_i^{(n)})^+ = nP. \tag{9.106}$$

We now prove an upper bound for the capacity of the Gaussian channel with feedback. This bound is actually achievable [136], and is therefore the capacity, but we do not prove this here.

Theorem 9.6.1 *For a Gaussian channel with feedback, the rate R_n for any sequence of $(2^{nR_n}, n)$ codes with $P_e^{(n)} \to 0$ satisfies*

$$R_n \le C_{n,FB} + \epsilon_n, \tag{9.107}$$

with $\epsilon_n \to 0$ as $n \to \infty$, where $C_{n,FB}$ is defined in (9.100).

Proof: Let W be uniform over 2^{nR}, and therefore the probability of error $P_e^{(n)}$ is bounded by Fano's inequality,

$$H(W|\hat{W}) \le 1 + nR_n P_e^{(n)} = n\epsilon_n, \tag{9.108}$$

where $\epsilon_n \to 0$ as $P_e^{(n)} \to 0$. We can then bound the rate as follows:

$$\begin{aligned}
nR_n &= H(W) && (9.109)\\
&= I(W;\hat{W}) + H(W|\hat{W}) && (9.110)\\
&\le I(W;\hat{W}) + n\epsilon_n && (9.111)\\
&\le I(W;Y^n) + n\epsilon_n && (9.112)\\
&= \sum I(W;Y_i|Y^{i-1}) + n\epsilon_n && (9.113)\\
&\overset{(a)}{=} \sum \big(h(Y_i|Y^{i-1}) - h(Y_i|W, Y^{i-1}, X_i, X^{i-1}, Z^{i-1})\big) + n\epsilon_n && (9.114)\\
&\overset{(b)}{=} \sum \big(h(Y_i|Y^{i-1}) - h(Z_i|W, Y^{i-1}, X_i, X^{i-1}, Z^{i-1})\big) + n\epsilon_n && (9.115)\\
&\overset{(c)}{=} \sum \big(h(Y_i|Y^{i-1}) - h(Z_i|Z^{i-1})\big) + n\epsilon_n && (9.116)\\
&= h(Y^n) - h(Z^n) + n\epsilon_n, && (9.117)
\end{aligned}$$

where (a) follows from the fact that X_i is a function of W and the past Y_i's, and Z^{i-1} is $Y^{i-1} - X^{i-1}$, (b) follows from $Y_i = X_i + Z_i$ and the fact that $h(X + Z|X) = h(Z|X)$, and (c) follows from the fact Z_i and (W, Y^{i-1}, X^i) are conditionally independent given Z^{i-1}. Continuing the chain of inequalities after dividing by n, we have

$$R_n \leq \frac{1}{n}\big(h(Y^n) - h(Z^n)\big) + \epsilon_n \tag{9.118}$$

$$\leq \frac{1}{2n} \log \frac{|K_Y^{(n)}|}{|K_Z^{(n)}|} + \epsilon_n \tag{9.119}$$

$$\leq C_{n,FB} + \epsilon_n, \tag{9.120}$$

by the entropy maximizing property of the normal. □

We have proved an upper bound on the capacity of the Gaussian channel with feedback in terms of the covariance matrix $K_{X+Z}^{(n)}$. We now derive bounds on the capacity with feedback in terms of $K_X^{(n)}$ and $K_Z^{(n)}$, which will then be used to derive bounds in terms of the capacity without feedback. For simplicity of notation, we will drop the superscript n in the symbols for covariance matrices.

We first prove a series of lemmas about matrices and determinants.

Lemma 9.6.1 *Let X and Z be n-dimensional random vectors. Then*

$$K_{X+Z} + K_{X-Z} = 2K_X + 2K_Z. \tag{9.121}$$

Proof

$$K_{X+Z} = E(X+Z)(X+Z)^t \tag{9.122}$$

$$= EXX^t + EXZ^t + EZX^t + EZZ^t \tag{9.123}$$

$$= K_X + K_{XZ} + K_{ZX} + K_Z. \tag{9.124}$$

Similarly,

$$K_{X-Z} = K_X - K_{XZ} - K_{ZX} + K_Z. \tag{9.125}$$

Adding these two equations completes the proof. □

Lemma 9.6.2 *For two $n \times n$ nonnegative definite matrices A and B, if $A - B$ is nonnegative definite, then $|A| \geq |B|$.*

Proof: Let $C = A - B$. Since B and C are nonnegative definite, we can consider them as covariance matrices. Consider two independent normal random vectors $\mathbf{X}_1 \sim \mathcal{N}(0, B)$ and $\mathbf{X}_2 \sim \mathcal{N}(0, C)$. Let $\mathbf{Y} = \mathbf{X}_1 + \mathbf{X}_2$.

Then

$$h(\mathbf{Y}) \geq h(\mathbf{Y}|\mathbf{X}_2) \tag{9.126}$$

$$= h(\mathbf{X}_1|\mathbf{X}_2) \tag{9.127}$$

$$= h(\mathbf{X}_1), \tag{9.128}$$

where the inequality follows from the fact that conditioning reduces differential entropy, and the final equality from the fact that $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent. Substituting the expressions for the differential entropies of a normal random variable, we obtain

$$\frac{1}{2}\log(2\pi e)^n|A| > \frac{1}{2}\log(2\pi e)^n|B|, \tag{9.129}$$

which is equivalent to the desired lemma. □

Lemma 9.6.3 *For two n-dimensional random vectors X and Z,*

$$|K_{X+Z}| \leq 2^n|K_X + K_Z|. \tag{9.130}$$

Proof: From Lemma 9.6.1,

$$2(K_X + K_Z) - K_{X+Z} = K_{X-Z} \succeq 0, \tag{9.131}$$

where $A \succeq 0$ means that A is nonnegative definite. Hence, applying Lemma 9.6.2, we have

$$|K_{X+Z}| \leq |2(K_X + K_Z)| = 2^n|K_X + K_Z|, \tag{9.132}$$

which is the desired result. □

Lemma 9.6.4 *For A, B nonnegative definite matrices and* $0 \leq \lambda \leq 1$*,*

$$|\lambda A + (1-\lambda)B| \geq |A|^\lambda |B|^{1-\lambda}. \tag{9.133}$$

Proof: Let $\mathbf{X} \sim \mathcal{N}_n(0, A)$ and $\mathbf{Y} \sim \mathcal{N}_n(0, B)$. Let $\mathbf{Z}$ be the mixture random vector

$$\mathbf{Z} = \begin{cases} \mathbf{X} & \text{if } \theta = 1 \\ \mathbf{Y} & \text{if } \theta = 2, \end{cases} \tag{9.134}$$

where

$$\theta = \begin{cases} 1 & \text{with probability } \lambda \\ 2 & \text{with probability } 1-\lambda. \end{cases} \tag{9.135}$$

Let **X**, **Y**, and θ be independent. Then

$$K_Z = \lambda A + (1-\lambda)B. \tag{9.136}$$

We observe that

$$\frac{1}{2}\ln(2\pi e)^n|\lambda A + (1-\lambda)B| \geq h(\mathbf{Z}) \tag{9.137}$$

$$\geq h(\mathbf{Z}|\theta) \tag{9.138}$$

$$= \lambda h(\mathbf{X}) + (1-\lambda)h(\mathbf{Y}) \tag{9.139}$$

$$= \frac{1}{2}\ln(2\pi e)^n|A|^\lambda|B|^{1-\lambda}, \tag{9.140}$$

which proves the result. The first inequality follows from the entropy maximizing property of the Gaussian under the covariance constraint. □

Definition We say that a random vector X^n is causally related to Z^n if

$$f(x^n, z^n) = f(z^n)\prod_{i=1}^{n} f(x_i|x^{i-1}, z^{i-1}). \tag{9.141}$$

Note that the feedback codes necessarily yield causally related (X^n, Z^n).

Lemma 9.6.5 *If X^n and Z^n are causally related, then*

$$h(X^n - Z^n) \geq h(Z^n) \tag{9.142}$$

and

$$|K_{X-Z}| \geq |K_Z|, \tag{9.143}$$

where K_{X-Z} and K_Z are the covariance matrices of $X^n - Z^n$ and Z^n, respectively.

Proof: We have

$$h(X^n - Z^n) \overset{(a)}{=} \sum_{i=1}^{n} h(X_i - Z_i|X^{i-1} - Z^{i-1}) \tag{9.144}$$

$$\overset{\text{(b)}}{\geq} \sum_{i=1}^{n} h(X_i - Z_i|X^{i-1}, Z^{i-1}, X_i) \tag{9.145}$$

$$\overset{\text{(c)}}{=} \sum_{i=1}^{n} h(Z_i|X^{i-1}, Z^{i-1}, X_i) \tag{9.146}$$

$$\overset{\text{(d)}}{=} \sum_{i=1}^{n} h(Z_i|Z^{i-1}) \tag{9.147}$$

$$\overset{\text{(e)}}{=} h(Z^n). \tag{9.148}$$

Here (a) follows from the chain rule, (b) follows from conditioning $h(A|B) \geq h(A|B, C)$, (c) follows from the conditional determinism of X_i and the invariance of differential entropy under translation, (d) follows from the causal relationship of X^n and Z^n, and (e) follows from the chain rule.

Finally, suppose that X^n and Z^n are causally related and the associated covariance matrices for Z^n and $X^n - Z^n$ are K_Z and K_{X-Z}. There obviously exists a multivariate normal (causally related) pair of random vectors $\tilde{X}^n$, $\tilde{Z}^n$ with the same covariance structure. Thus, from (9.148), we have

$$\frac{1}{2}\ln(2\pi e)^n|K_{X-Z}| = h(\tilde{X}^n - \tilde{Z}^n) \tag{9.149}$$

$$\geq h(\tilde{Z}^n) \tag{9.150}$$

$$= \frac{1}{2}\ln(2\pi e)^n|K_Z|, \tag{9.151}$$

thus proving (9.143). □

We are now in a position to prove that feedback increases the capacity of a nonwhite Gaussian additive noise channel by at most half a bit.

Theorem 9.6.2

$$C_{n,\text{FB}} \leq C_n + \frac{1}{2} \qquad \textit{bits per transmission.} \tag{9.152}$$

Proof: Combining all the lemmas, we obtain

$$C_{n,\text{FB}} \leq \max_{\text{tr}(K_X)\leq nP} \frac{1}{2n}\log\frac{|K_Y|}{|K_Z|} \tag{9.153}$$

$$\leq \max_{\operatorname{tr}(K_X)\leq nP} \frac{1}{2n} \log \frac{2^n|K_X + K_Z|}{|K_Z|} \tag{9.154}$$

$$= \max_{\operatorname{tr}(K_X)\leq nP} \frac{1}{2n} \log \frac{|K_X + K_Z|}{|K_Z|} + \frac{1}{2} \tag{9.155}$$

$$\leq C_n + \frac{1}{2} \qquad \text{bits per transmission,} \tag{9.156}$$

where the inequalities follow from Theorem 9.6.1, Lemma 9.6.3, and the definition of capacity without feedback, respectively. □

We now prove Pinsker's statement that feedback can at most double the capacity of colored noise channels.

Theorem 9.6.3 $C_{n,\text{FB}} \leq 2C_n$.

Proof: It is enough to show that

$$\frac{1}{2}\frac{1}{2n} \log \frac{|K_{X+Z}|}{|K_Z|} \leq \frac{1}{2n} \log \frac{|K_X + K_Z|}{|K_Z|}, \tag{9.157}$$

for it will then follow that by maximizing the right side and then the left side that

$$\frac{1}{2} C_{n,\text{FB}} \leq C_n. \tag{9.158}$$

We have

$$\frac{1}{2n} \log \frac{|K_X + K_Z|}{|K_Z|} \overset{(a)}{=} \frac{1}{2n} \log \frac{|\frac{1}{2}K_{X+Z} + \frac{1}{2}K_{X-Z}|}{|K_Z|} \tag{9.159}$$

$$\overset{(b)}{\geq} \frac{1}{2n} \log \frac{|K_{X+Z}|^{\frac{1}{2}}|K_{X-Z}|^{\frac{1}{2}}}{|K_Z|} \tag{9.160}$$

$$\overset{(c)}{\geq} \frac{1}{2n} \log \frac{|K_{X+Z}|^{\frac{1}{2}}|K_Z|^{\frac{1}{2}}}{|K_Z|} \tag{9.161}$$

$$\overset{(d)}{=} \frac{1}{2}\frac{1}{2n} \log \frac{|K_{X+Z}|}{|K_Z|} \tag{9.162}$$

and the result is proved. Here (a) follows from Lemma 9.6.1, (b) is the inequality in Lemma 9.6.4, and (c) is Lemma 9.6.5 in which causality is used. □

Thus, we have shown that Gaussian channel capacity is not increased by more than half a bit or by more than a factor of 2 when we have feedback; feedback helps, but not by much.

SUMMARY

Maximum entropy. $\max_{EX^2=\alpha} h(X) = \frac{1}{2}\log 2\pi e\alpha$.

Gaussian channel. $Y_i = X_i + Z_i$; $Z_i \sim \mathcal{N}(0, N)$; power constraint $\frac{1}{n}\sum_{i=1}^n x_i^2 \le P$; and

$$C = \frac{1}{2}\log\left(1 + \frac{P}{N}\right) \qquad \text{bits per transmission.} \tag{9.163}$$

Bandlimited additive white Gaussian noise channel. Bandwidth W; two-sided power spectral density $N_0/2$; signal power P; and

$$C = W\log\left(1 + \frac{P}{N_0 W}\right) \qquad \text{bits per second.} \tag{9.164}$$

Water-filling (k parallel Gaussian channels). $Y_j = X_j + Z_j$, $j = 1, 2, \ldots, k$; $Z_j \sim \mathcal{N}(0, N_j)$; $\sum_{j=1}^k X_j^2 \le P$; and

$$C = \sum_{i=1}^k \frac{1}{2}\log\left(1 + \frac{(\nu - N_i)^+}{N_i}\right), \tag{9.165}$$

where ν is chosen so that $\sum(\nu - N_i)^+ = nP$.

Additive nonwhite Gaussian noise channel. $Y_i = X_i + Z_i$; $Z^n \sim \mathcal{N}(0, K_Z)$; and

$$C = \frac{1}{n}\sum_{i=1}^n \frac{1}{2}\log\left(1 + \frac{(\nu - \lambda_i)^+}{\lambda_i}\right), \tag{9.166}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of K_Z and ν is chosen so that $\sum_i(\nu - \lambda_i)^+ = P$.

Capacity without feedback

$$C_n = \max_{\mathrm{tr}(K_X)\le nP} \frac{1}{2n}\log\frac{|K_X + K_Z|}{|K_Z|}. \tag{9.167}$$

Capacity with feedback

$$C_{n,\mathrm{FB}} = \max_{\mathrm{tr}(K_X)\le nP} \frac{1}{2n} \log \frac{|K_{X+Z}|}{|K_Z|}. \tag{9.168}$$

Feedback bounds

$$C_{n,\mathrm{FB}} \le C_n + \frac{1}{2}. \tag{9.169}$$

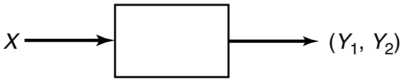
$$C_{n,\mathrm{FB}} \le 2C_n. \tag{9.170}$$

PROBLEMS

9.1 *Channel with two independent looks at Y.* Let Y_1 and Y_2 be conditionally independent and conditionally identically distributed given X.

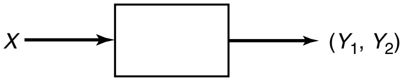
(a) Show that $I(X;Y_1,Y_2) = 2I(X;Y_1) - I(Y_1;Y_2)$.

(b) Conclude that the capacity of the channel



is less than twice the capacity of the channel

9.2 *Two-look Gaussian channel*



Consider the ordinary Gaussian channel with two correlated looks at X, that is, $Y = (Y_1, Y_2)$, where

$$Y_1 = X + Z_1 \tag{9.171}$$

$$Y_2 = X + Z_2 \tag{9.172}$$

with a power constraint P on X, and $(Z_1, Z_2) \sim \mathcal{N}_2(0, K)$, where

$$K = \begin{bmatrix} N & N\rho \\ N\rho & N \end{bmatrix}. \tag{9.173}$$

Find the capacity C for

(a) $\rho = 1$

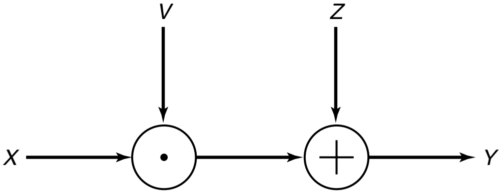
(b) $\rho = 0$

(c) $\rho = -1$

9.3 *Output power constraint*. Consider an additive white Gaussian noise channel with an expected *output* power constraint P. Thus, $Y = X + Z$, $Z \sim N(0, \sigma^2)$, Z is independent of X, and $EY^2 \le P$. Find the channel capacity.

9.4 *Exponential noise channels*. $Y_i = X_i + Z_i$, where Z_i is i.i.d. exponentially distributed noise with mean μ. Assume that we have a mean constraint on the signal (i.e., $EX_i \le \lambda$). Show that the capacity of such a channel is $C = \log(1 + \frac{\lambda}{\mu})$.

9.5 *Fading channel*. Consider an additive noise fading channel



$$Y = XV + Z,$$

where Z is additive noise, V is a random variable representing fading, and Z and V are independent of each other and of X. Argue that knowledge of the fading factor V improves capacity by showing that

$$I(X; Y|V) \ge I(X; Y).$$

9.6 *Parallel channels and water-filling*. Consider a pair of parallel Gaussian channels:

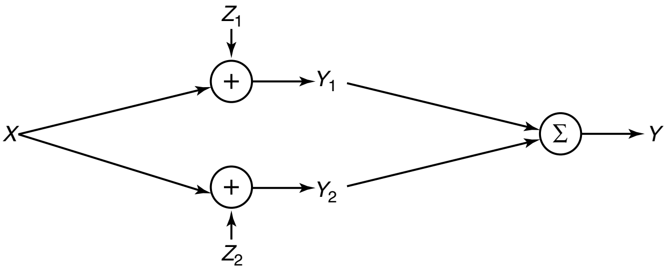
$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} + \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}, \tag{9.174}$$

where

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}\right), \tag{9.175}$$

and there is a power constraint $E(X_1^2 + X_2^2) \le 2P$. Assume that $\sigma_1^2 > \sigma_2^2$. At what power does the channel stop behaving like a single channel with noise variance σ_2^2, and begin behaving like a pair of channels?

9.7 *Multipath Gaussian channel*. Consider a Gaussian noise channel with power constraint P, where the signal takes two different paths and the received noisy signals are added together at the antenna.

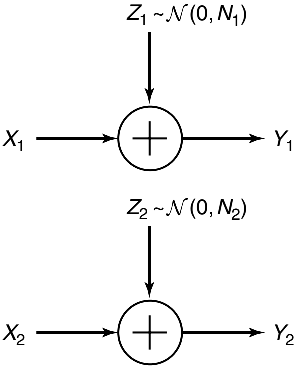


(a) Find the capacity of this channel if Z_1 and Z_2 are jointly normal with covariance matrix

$$K_Z = \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}.$$

(b) What is the capacity for $\rho = 0$, $\rho = 1$, $\rho = -1$?

9.8 *Parallel Gaussian channels*. Consider the following parallel Gaussian channel:



where $Z_1 \sim \mathcal{N}(0,N_1)$ and $Z_2 \sim \mathcal{N}(0,N_2)$ are independent Gaussian random variables and $Y_i = X_i + Z_i$. We wish to allocate power to the two parallel channels. Let β_1 and β_2 be fixed. Consider a total cost constraint $\beta_1 P_1 + \beta_2 P_2 \leq \beta$, where P_i is the power allocated to the ith channel and β_i is the cost per unit power in that channel. Thus, $P_1 \geq 0$ and $P_2 \geq 0$ can be chosen subject to the cost constraint β.

(a) For what value of β does the channel stop acting like a single channel and start acting like a pair of channels?

(b) Evaluate the capacity and find P_1 and P_2 that achieve capacity for $\beta_1 = 1$, $\beta_2 = 2$, $N_1 = 3$, $N_2 = 2$, and $\beta = 10$.

9.9 *Vector Gaussian channel*. Consider the vector Gaussian noise channel

$$Y = X + Z,$$

where $X = (X_1, X_2, X_3)$, $Z = (Z_1, Z_2, Z_3)$, $Y = (Y_1, Y_2, Y_3)$, $E\|X\|^2 \leq P$, and

$$Z \sim \mathcal{N}\left(0, \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 2 \end{bmatrix}\right).$$

Find the capacity. The answer may be surprising.

9.10 *Capacity of photographic film*. Here is a problem with a nice answer that takes a little time. We're interested in the capacity of photographic film. The film consists of silver iodide crystals, Poisson distributed, with a density of λ particles per square inch. The film is illuminated without knowledge of the position of the silver iodide particles. It is then developed and the receiver sees only the silver iodide particles that have been illuminated. It is assumed that light incident on a cell exposes the grain if it is there and otherwise results in a blank response. Silver iodide particles that are not illuminated and vacant portions of the film remain blank. The question is: What is the capacity of this film?

We make the following assumptions. We grid the film very finely into cells of area dA. It is assumed that there is at most one silver iodide particle per cell and that no silver iodide particle is intersected by the cell boundaries. Thus, the film can be considered to be a large number of parallel binary asymmetric channels with crossover probability $1 - \lambda dA$. By calculating the capacity of this binary asymmetric channel to first order in dA (making the necessary approximations), one can calculate the capacity of the film in bits per square inch. It is, of course, proportional to λ. The question is: What is the multiplicative constant?

The answer would be λ bits per unit area if both illuminator and receiver knew the positions of the crystals.

9.11 *Gaussian mutual information.* Suppose that (X, Y, Z) are jointly Gaussian and that $X \to Y \to Z$ forms a Markov chain. Let X and Y have correlation coefficient ρ_1 and let Y and Z have correlation coefficient ρ_2. Find $I(X; Z)$.

9.12 *Time-varying channel.* A train pulls out of the station at constant velocity. The received signal energy thus falls off with time as $1/i^2$. The total received signal at time i is

$$Y_i = \frac{1}{i} X_i + Z_i,$$

where $Z_1, Z_2, \ldots$ are i.i.d. $\sim N(0, N)$. The transmitter constraint for block length n is

$$\frac{1}{n} \sum_{i=1}^{n} x_i^2(w) \le P, \quad w \in \{1, 2, \ldots, 2^{nR}\}.$$

Using Fano's inequality, show that the capacity C is equal to zero for this channel.

9.13 *Feedback capacity.* Let $(Z_1, Z_2) \sim N(0, K)$, $K = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$. Find the maximum of $\frac{1}{2} \log \frac{|K_{X+Z}|}{|K_Z|}$ with and without feedback given a trace (power) constraint $\mathrm{tr}(K_X) \le 2P$.

9.14 *Additive noise channel.* Consider the channel $Y = X + Z$, where X is the transmitted signal with power constraint P, Z is independent additive noise, and Y is the received signal. Let

$$Z = \begin{cases} 0 & \text{with probability } \frac{1}{10} \\ Z^* & \text{with probability } \frac{9}{10}, \end{cases}$$

where $Z^* \sim N(0, N)$. Thus, Z has a mixture distribution that is the mixture of a Gaussian distribution and a degenerate distribution with mass 1 at 0.

(a) What is the capacity of this channel? This should be a pleasant surprise.

(b) How would you signal to achieve capacity?

9.15 *Discrete input, continuous output channel.* Let $\Pr\{X = 1\} = p$, $\Pr\{X = 0\} = 1 - p$, and let $Y = X + Z$, where Z is uniform over the interval $[0, a]$, $a > 1$, and Z is independent of X.

(a) Calculate

$$I(X; Y) = H(X) - H(X|Y).$$

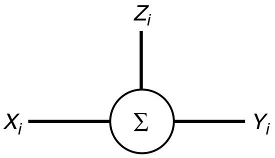
(b) Now calculate $I(X; Y)$ the other way by

$$I(X; Y) = h(Y) - h(Y|X).$$

(c) Calculate the capacity of this channel by maximizing over p.

9.16 *Gaussian mutual information.* Suppose that (X, Y, Z) are jointly Gaussian and that $X \to Y \to Z$ forms a Markov chain. Let X and Y have correlation coefficient ρ_1 and let Y and Z have correlation coefficient ρ_2. Find $I(X; Z)$.

9.17 *Impulse power.* Consider the additive white Gaussian channel



where $Z_i \sim N(0, N)$, and the input signal has average power constraint P.

(a) Suppose that we use all our power at time 1 (i.e., $EX_1^2 = nP$ and $EX_i^2 = 0$ for $i = 2, 3, \ldots, n$). Find

$$\max_{f(x^n)} \frac{I(X^n; Y^n)}{n},$$

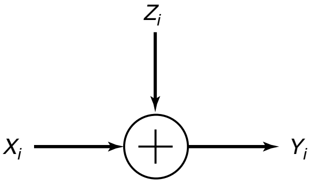
where the maximization is over all distributions $f(x^n)$ subject to the constraint $EX_1^2 = nP$ and $EX_i^2 = 0$ for $i = 2, 3, \ldots, n$.

(b) Find

$$\max_{f(x^n):\ E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right)\le P} \frac{1}{n} I(X^n; Y^n)$$

and compare to part (a).

9.18 *Gaussian channel with time-varying mean.* Find the capacity of the following Gaussian channel:



Let $Z_1, Z_2, \ldots$ be independent and let there be a power constraint P on $x^n(W)$. Find the capacity when:

(a) $\mu_i = 0$, for all i.

(b) $\mu_i = e^i, \quad i = 1, 2, \ldots$. Assume that μ_i is known to the transmitter and receiver.

(c) μ_i unknown, but μ_i i.i.d. $\sim N(0, N_1)$ for all i.

9.19 *Parametric form for channel capacity.* Consider m parallel Gaussian channels, $Y_i = X_i + Z_i$, where $Z_i \sim N(0, \lambda_i)$ and the noises X_i are independent random variables. Thus, $C = \sum_{i=1}^{m} \frac{1}{2} \log(1 + \frac{(\lambda-\lambda_i)^+}{\lambda_i})$, where λ is chosen to satisfy $\sum_{i=1}^{m} (\lambda - \lambda_i)^+ = P$. Show that this can be rewritten in the form

$$\begin{aligned} P(\lambda) &= \textstyle\sum_{i:\lambda_i \le \lambda} (\lambda - \lambda_i) \\ C(\lambda) &= \textstyle\sum_{i:\lambda_i \le \lambda} \frac{1}{2} \log \frac{\lambda}{\lambda_i}. \end{aligned}$$

Here $P(\lambda)$ is piecewise linear and $C(\lambda)$ is piecewise logarithmic in λ.

9.20 *Robust decoding.* Consider an additive noise channel whose output Y is given by

$$Y = X + Z,$$

where the channel input X is average power limited,

$$EX^2 \le P,$$

and the noise process $\{Z_k\}_{k=-\infty}^{\infty}$ is i.i.d. with marginal distribution $p_Z(z)$ (not necessarily Gaussian) of power N,

$$EZ^2 = N.$$

(a) Show that the channel capacity, $C = \max_{EX^2 \le P} I(X;Y)$, is lower bounded by C_G, where

$$C_G = \frac{1}{2} \log \left(1 + \frac{P}{N}\right)$$

(i.e., the capacity C_G corresponding to white Gaussian noise).

(b) Decoding the received vector to the codeword that is closest to it in Euclidean distance is in general suboptimal if the noise is non-Gaussian. Show, however, that the rate C_G is achievable even if one insists on performing nearest-neighbor decoding (minimum Euclidean distance decoding) rather than the optimal maximum-likelihood or joint typicality decoding (with respect to the true noise distribution).

(c) Extend the result to the case where the noise is not i.i.d. but is stationary and ergodic with power N.

(*Hint for b and c:* Consider a size 2^{nR} random codebook whose codewords are drawn independently of each other according to a uniform distribution over the n-dimensional sphere of radius $\sqrt{nP}$.)

(a) Using a symmetry argument, show that conditioned on the noise vector, the ensemble average probability of error depends on the noise vector only via its Euclidean norm $\|\mathbf{z}\|$.

(b) Use a geometric argument to show that this dependence is monotonic.

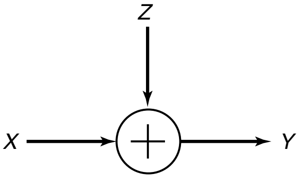
(c) Given a rate $R < C_G$, choose some $N' > N$ such that

$$R < \frac{1}{2} \log \left(1 + \frac{P}{N'}\right).$$

Compare the case where the noise is i.i.d. $\mathcal{N}(0, N')$ to the case at hand.

(d) Conclude the proof using the fact that the above ensemble of codebooks can achieve the capacity of the Gaussian channel (no need to prove that).

9.21 *Mutual information game*. Consider the following channel:



Throughout this problem we shall constrain the signal power

$$EX = 0, \qquad EX^2 = P, \tag{9.176}$$

and the noise power

$$EZ = 0, \qquad EZ^2 = N, \tag{9.177}$$

and assume that X and Z are independent. The channel capacity is given by $I(X; X + Z)$.
Now for the game. The noise player chooses a distribution on Z to minimize $I(X; X + Z)$, while the signal player chooses a distribution on X to maximize $I(X; X + Z)$. Letting $X^* \sim \mathcal{N}(0, P)$, $Z^* \sim \mathcal{N}(0, N)$, show that Gaussian X^* and Z^* satisfy the saddlepoint conditions

$$I(X; X + Z^*) \le I(X^*; X^* + Z^*) \le I(X^*; X^* + Z). \tag{9.178}$$

Thus,

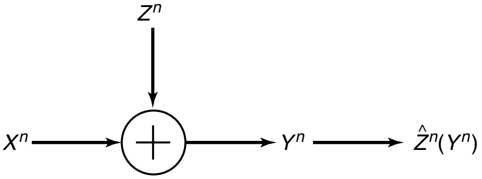
$$\min_Z \max_X I(X; X + Z) = \max_X \min_Z I(X; X + Z) \tag{9.179}$$

$$= \frac{1}{2} \log \left(1 + \frac{P}{N}\right), \tag{9.180}$$

and the game has a value. In particular, a deviation from normal for either player worsens the mutual information from that player's standpoint. Can you discuss the implications of this?
Note: Part of the proof hinges on the entropy power inequality from Section 17.8, which states that if $\mathbf{X}$ and $\mathbf{Y}$ are independent random n-vectors with densities, then

$$2^{\frac{2}{n}h(\mathbf{X}+\mathbf{Y})} \ge 2^{\frac{2}{n}h(\mathbf{X})} + 2^{\frac{2}{n}h(\mathbf{Y})}. \tag{9.181}$$

9.22 *Recovering the noise*. Consider a standard Gaussian channel $Y^n = X^n + Z^n$, where Z_i is i.i.d. $\sim \mathcal{N}(0, N)$, $i = 1, 2, \ldots, n$, and $\frac{1}{n}\sum_{i=1}^n X_i^2 \le P$. Here we are interested in recovering the noise Z^n and we don't care about the signal X^n. By sending $X^n = (0, 0, \ldots, 0)$, the receiver gets $Y^n = Z^n$ and can fully determine the value of Z^n. We wonder how much variability there can be in X^n and still recover the Gaussian noise Z^n. Use of the channel looks like



Argue that for some $R > 0$, the transmitter can arbitrarily send one of 2^{nR} different sequences of x^n without affecting the recovery of the noise in the sense that

$$\Pr\{\hat{Z}^n \neq Z^n\} \to 0 \qquad \text{as } n \to \infty.$$

For what R is this possible?

HISTORICAL NOTES

The Gaussian channel was first analyzed by Shannon in his original paper [472]. The water-filling solution to the capacity of the colored noise Gaussian channel was developed by Shannon [480] and treated in detail by Pinsker [425]. The time-continuous Gaussian channel is treated in Wyner [576], Gallager [233], and Landau, Pollak, and Slepian [340, 341, 500].

Pinsker [421] and Ebert [178] argued that feedback at most doubles the capacity of a nonwhite Gaussian channel; the proof in the text is from Cover and Pombra [136], who also show that feedback increases the capacity of the nonwhite Gaussian channel by at most half a bit. The most recent feedback capacity results for nonwhite Gaussian noise channels are due to Kim [314].

CHAPTER 10

RATE DISTORTION THEORY

The description of an arbitrary real number requires an infinite number of bits, so a finite representation of a continuous random variable can never be perfect. How well can we do? To frame the question appropriately, it is necessary to define the "goodness" of a representation of a source. This is accomplished by defining a distortion measure which is a measure of distance between the random variable and its representation. The basic problem in rate distortion theory can then be stated as follows: Given a source distribution and a distortion measure, what is the minimum expected distortion achievable at a particular rate? Or, equivalently, what is the minimum rate description required to achieve a particular distortion?

One of the most intriguing aspects of this theory is that joint descriptions are more efficient than individual descriptions. It is simpler to describe an elephant and a chicken with one description than to describe each alone. This is true even for independent random variables. It is simpler to describe X_1 and X_2 together (at a given distortion for each) than to describe each by itself. Why don't independent problems have independent solutions? The answer is found in the geometry. Apparently, rectangular grid points (arising from independent descriptions) do not fill up the space efficiently.

Rate distortion theory can be applied to both discrete and continuous random variables. The zero-error data compression theory of Chapter 5 is an important special case of rate distortion theory applied to a discrete source with zero distortion. We begin by considering the simple problem of representing a single continuous random variable by a finite number of bits.

10.1 QUANTIZATION

In this section we motivate the elegant theory of rate distortion by showing how complicated it is to solve the quantization problem exactly for a single

Elements of Information Theory, Second Edition, By Thomas M. Cover and Joy A. Thomas

random variable. Since a continuous random source requires infinite precision to represent exactly, we cannot reproduce it exactly using a finite-rate code. The question is then to find the best possible representation for any given data rate.

We first consider the problem of representing a single sample from the source. Let the random variable be represented be X and let the representation of X be denoted as $\hat{X}(X)$. If we are given R bits to represent X, the function $\hat{X}$ can take on 2^R values. The problem is to find the optimum set of values for $\hat{X}$ (called the *reproduction points* or *code points*) and the regions that are associated with each value $\hat{X}$.

For example, let $X \sim \mathcal{N}(0, \sigma^2)$, and assume a squared-error distortion measure. In this case we wish to find the function $\hat{X}(X)$ such that $\hat{X}$ takes on at most 2^R values and minimizes $E(X - \hat{X}(X))^2$. If we are given one bit to represent X, it is clear that the bit should distinguish whether or not $X > 0$. To minimize squared error, each reproduced symbol should be the conditional mean of its region. This is illustrated in Figure 10.1. Thus,

$$\hat{X}(x) = \begin{cases} \sqrt{\frac{2}{\pi}}\sigma & \text{if } x \geq 0, \\ -\sqrt{\frac{2}{\pi}}\sigma & \text{if } x < 0. \end{cases} \tag{10.1}$$

If we are given 2 bits to represent the sample, the situation is not as simple. Clearly, we want to divide the real line into four regions and use

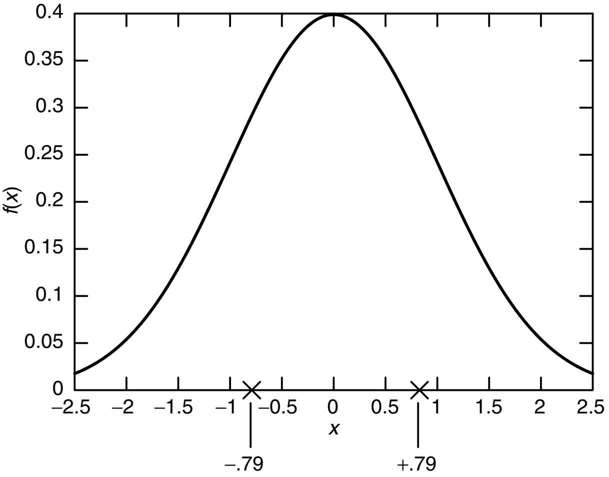


FIGURE 10.1. One-bit quantization of Gaussian random variable.

a point within each region to represent the sample. But it is no longer immediately obvious what the representation regions and the reconstruction points should be. We can, however, state two simple properties of optimal regions and reconstruction points for the quantization of a single random variable:

- Given a set $\{\hat{X}(w)\}$ of reconstruction points, the distortion is minimized by mapping a source random variable X to the representation $\hat{X}(w)$ that is closest to it. The set of regions of $\mathcal{X}$ defined by this mapping is called a *Voronoi* or *Dirichlet partition* defined by the reconstruction points.
- The reconstruction points should minimize the conditional expected distortion over their respective assignment regions.

These two properties enable us to construct a simple algorithm to find a "good" quantizer: We start with a set of reconstruction points, find the optimal set of reconstruction regions (which are the nearest-neighbor regions with respect to the distortion measure), then find the optimal reconstruction points for these regions (the centroids of these regions if the distortion is squared error), and then repeat the iteration for this new set of reconstruction points. The expected distortion is decreased at each stage in the algorithm, so the algorithm will converge to a local minimum of the distortion. This algorithm is called the *Lloyd algorithm* [363] (for real-valued random variables) or the *generalized Lloyd algorithm* [358] (for vector-valued random variables) and is frequently used to design quantization systems.

Instead of quantizing a single random variable, let us assume that we are given a set of n i.i.d. random variables drawn according to a Gaussian distribution. These random variables are to be represented using nR bits. Since the source is i.i.d., the symbols are independent, and it may appear that the representation of each element is an independent problem to be treated separately. But this is not true, as the results on rate distortion theory will show. We will represent the entire sequence by a single index taking 2^{nR} values. This treatment of entire sequences at once achieves a lower distortion for the same rate than independent quantization of the individual samples.

10.2 DEFINITIONS

Assume that we have a source that produces a sequence $X_1, X_2, \ldots, X_n$ i.i.d. $\sim p(x), x \in \mathcal{X}$. For the proofs in this chapter, we assume that the

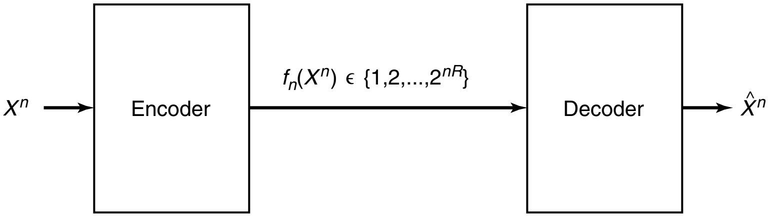


FIGURE 10.2. Rate distortion encoder and decoder.

alphabet is finite, but most of the proofs can be extended to continuous random variables. The encoder describes the source sequence X^n by an index $f_n(X^n) \in \{1, 2, \ldots, 2^{nR}\}$. The decoder represents X^n by an estimate $\hat{X}^n \in \hat{\mathcal{X}}$, as illustrated in Figure 10.2.

Definition A *distortion function* or *distortion measure* is a mapping

$$d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathcal{R}^+ \tag{10.2}$$

from the set of source alphabet-reproduction alphabet pairs into the set of nonnegative real numbers. The distortion $d(x, \hat{x})$ is a measure of the cost of representing the symbol x by the symbol $\hat{x}$.

Definition A distortion measure is said to be *bounded* if the maximum value of the distortion is finite:

$$d_{\max} \overset{\text{def}}{=} \max_{x \in \mathcal{X}, \hat{x} \in \hat{X}} d(x, \hat{x}) < \infty. \tag{10.3}$$

In most cases, the reproduction alphabet $\hat{\mathcal{X}}$ is the same as the source alphabet $\mathcal{X}$.

Examples of common distortion functions are

- *Hamming (probability of error) distortion*. The Hamming distortion is given by

$$d(x, \hat{x}) = \begin{cases} 0 & \text{if } x = \hat{x} \\ 1 & \text{if } x \neq \hat{x}, \end{cases} \tag{10.4}$$

  which results in a probability of error distortion, since $\mathrm{E}d(X, \hat{X}) = \Pr(X \neq \hat{X})$.

- *Squared-error distortion*. The squared-error distortion,

  $$d(x, \hat{x}) = (x - \hat{x})^2, \tag{10.5}$$

  is the most popular distortion measure used for continuous alphabets. Its advantages are its simplicity and its relationship to least-squares prediction. But in applications such as image and speech coding, various authors have pointed out that the mean-squared error is not an appropriate measure of distortion for human observers. For example, there is a large squared-error distortion between a speech waveform and another version of the same waveform slightly shifted in time, even though both would sound the same to a human observer.

Many alternatives have been proposed; a popular measure of distortion in speech coding is the *Itakura–Saito distance*, which is the relative entropy between multivariate normal processes. In image coding, however, there is at present no real alternative to using the mean-squared error as the distortion measure.

The distortion measure is defined on a symbol-by-symbol basis. We extend the definition to sequences by using the following definition:

Definition The *distortion between sequences x^n* and $\hat{x}^n$ is defined by

$$d(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i). \tag{10.6}$$

So the distortion for a sequence is the average of the per symbol distortion of the elements of the sequence. This is not the only reasonable definition. For example, one may want to measure the distortion between two sequences by the maximum of the per symbol distortions. The theory derived below does not apply directly to this more general distortion measure.

Definition A $(2^{nR}, n)$*-rate distortion code* consists of an encoding function,

$$f_n : \mathcal{X}^n \to \{1, 2, \ldots, 2^{nR}\}, \tag{10.7}$$

and a decoding (reproduction) function,

$$g_n : \{1, 2, \ldots, 2^{nR}\} \to \hat{\mathcal{X}}^n. \tag{10.8}$$

The distortion associated with the $(2^{nR}, n)$ code is defined as

$$D = \mathrm{E}d(X^n, g_n(f_n(X^n))), \tag{10.9}$$

where the expectation is with respect to the probability distribution on X:

$$D = \sum_{x^n} p(x^n) d(x^n, g_n(f_n(x^n))). \tag{10.10}$$

The set of n-tuples $g_n(1), g_n(2), \ldots, g_n(2^{nR})$, denoted by $\hat{X}^n(1), \ldots, \hat{X}^n(2^{nR})$, constitutes the *codebook*, and $f_n^{-1}(1), \ldots, f_n^{-1}(2^{nR})$ are the associated *assignment regions.*

Many terms are used to describe the replacement of X^n by its quantized version $\hat{X}^n(w)$. It is common to refer to $\hat{X}^n$ as the *vector quantization*, *reproduction*, *reconstruction*, *representation*, *source code*, or *estimate* of X^n.

Definition A rate distortion pair (R, D) is said to be *achievable* if there exists a sequence of $(2^{nR}, n)$-rate distortion codes (f_n, g_n) with $\lim_{n\to\infty} \mathrm{E}d(X^n, g_n(f_n(X^n))) \le D$.

Definition The *rate distortion region* for a source is the closure of the set of achievable rate distortion pairs (R, D).

Definition The *rate distortion function* $R(D)$ is the infimum of rates R such that (R, D) is in the rate distortion region of the source for a given distortion D.

Definition The *distortion rate function* $D(R)$ is the infimum of all distortions D such that (R, D) is in the rate distortion region of the source for a given rate R.

The distortion rate function defines another way of looking at the boundary of the rate distortion region. We will in general use the rate distortion function rather than the distortion rate function to describe this boundary, although the two approaches are equivalent.

We now define a mathematical function of the source, which we call the *information rate distortion function.* The main result of this chapter is the proof that the information rate distortion function is equal to the rate distortion function defined above (i.e., it is the infimum of rates that achieve a particular distortion).

Definition The *information rate distortion function* $R^{(I)}(D)$ for a source X with distortion measure $d(x, \hat{x})$ is defined as

$$R^{(I)}(D) = \min_{p(\hat{x}|x): \sum_{(x,\hat{x})} p(x)p(\hat{x}|x)d(x,\hat{x}) \leq D} I(X; \hat{X}), \tag{10.11}$$

where the minimization is over all conditional distributions $p(\hat{x}|x)$ for which the joint distribution $p(x, \hat{x}) = p(x)p(\hat{x}|x)$ satisfies the expected distortion constraint.

Paralleling the discussion of channel capacity in Chapter 7, we initially consider the properties of the information rate distortion function and calculate it for some simple sources and distortion measures. Later we prove that we can actually achieve this function (i.e., there exist codes with rate $R^{(I)}(D)$ with distortion D). We also prove a converse establishing that $R \geq R^{(I)}(D)$ for any code that achieves distortion D.

The main theorem of rate distortion theory can now be stated as follows:

Theorem 10.2.1 *The rate distortion function for an i.i.d. source X with distribution $p(x)$ and bounded distortion function $d(x, \hat{x})$ is equal to the associated information rate distortion function. Thus,*

$$R(D) = R^{(I)}(D) = \min_{p(\hat{x}|x): \sum_{(x,\hat{x})} p(x)p(\hat{x}|x)d(x,\hat{x}) \leq D} I(X; \hat{X}) \tag{10.12}$$

is the minimum achievable rate at distortion D.

This theorem shows that the operational definition of the rate distortion function is equal to the information definition. Hence we will use $R(D)$ from now on to denote both definitions of the rate distortion function. Before coming to the proof of the theorem, we calculate the information rate distortion function for some simple sources and distortions.

10.3 CALCULATION OF THE RATE DISTORTION FUNCTION

10.3.1 Binary Source

We now find the description rate $R(D)$ required to describe a Bernoulli(p) source with an expected proportion of errors less than or equal to D.

Theorem 10.3.1 *The rate distortion function for a Bernoulli(p) source with Hamming distortion is given by*

$$R(D) = \begin{cases} H(p) - H(D), & 0 \leq D \leq \min\{p, 1-p\}, \\ 0, & D > \min\{p, 1-p\}. \end{cases} \tag{10.13}$$

Proof: Consider a binary source $X \sim \text{Bernoulli}(p)$ with a Hamming distortion measure. Without loss of generality, we may assume that $p < \frac{1}{2}$. We wish to calculate the rate distortion function,

$$R(D) = \min_{p(\hat{x}|x):\sum_{(x,\hat{x})} p(x)p(\hat{x}|x)d(x,\hat{x})\le D} I(X;\hat{X}). \tag{10.14}$$

Let $\oplus$ denote modulo 2 addition. Thus, $X \oplus \hat{X} = 1$ is equivalent to $X \ne \hat{X}$. We do not minimize $I(X;\hat{X})$ directly; instead, we find a lower bound and then show that this lower bound is achievable. For any joint distribution satisfying the distortion constraint, we have

$$I(X;\hat{X}) = H(X) - H(X|\hat{X}) \tag{10.15}$$

$$= H(p) - H(X \oplus \hat{X}|\hat{X}) \tag{10.16}$$

$$\ge H(p) - H(X \oplus \hat{X}) \tag{10.17}$$

$$\ge H(p) - H(D), \tag{10.18}$$

since $\Pr(X \ne \hat{X}) \le D$ and $H(D)$ increases with D for $D \le \frac{1}{2}$. Thus,

$$R(D) \ge H(p) - H(D). \tag{10.19}$$

We now show that the lower bound is actually the rate distortion function by finding a joint distribution that meets the distortion constraint and has $I(X;\hat{X}) = R(D)$. For $0 \le D \le p$, we can achieve the value of the rate distortion function in (10.19) by choosing $(X, \hat{X})$ to have the joint distribution given by the binary symmetric channel shown in Figure 10.3.

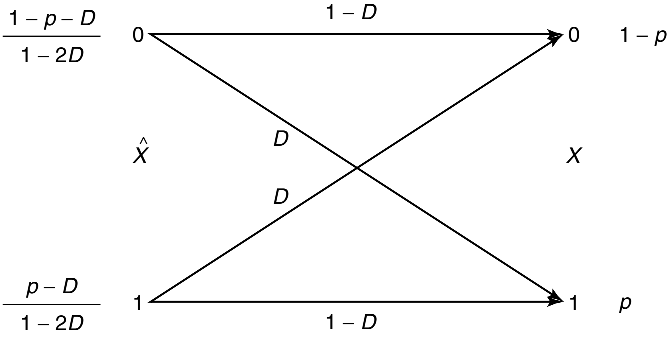


FIGURE 10.3. Joint distribution for binary source.

We choose the distribution of $\hat{X}$ at the input of the channel so that the output distribution of X is the specified distribution. Let $r = \Pr(\hat{X} = 1)$. Then choose r so that

$$r(1 - D) + (1 - r)D = p, \tag{10.20}$$

or

$$r = \frac{p - D}{1 - 2D}. \tag{10.21}$$

If $D \le p \le \frac{1}{2}$, then $\Pr(\hat{X} = 1) \ge 0$ and $\Pr(\hat{X} = 0) \ge 0$. We then have

$$I(X; \hat{X}) = H(X) - H(X|\hat{X}) = H(p) - H(D), \tag{10.22}$$

and the expected distortion is $\Pr(X \ne \hat{X}) = D$.

If $D \ge p$, we can achieve $R(D) = 0$ by letting $\hat{X} = 0$ with probability 1. In this case, $I(X; \hat{X}) = 0$ and $D = p$. Similarly, if $D \ge 1 - p$, we can achieve $R(D) = 0$ by setting $\hat{X} = 1$ with probability 1. Hence, the rate distortion function for a binary source is

$$R(D) = \begin{cases} H(p) - H(D), & 0 \le D \le \min\{p, 1 - p\}, \\ 0, & D > \min\{p, 1 - p\}. \end{cases} \tag{10.23}$$

This function is illustrated in Figure 10.4. □

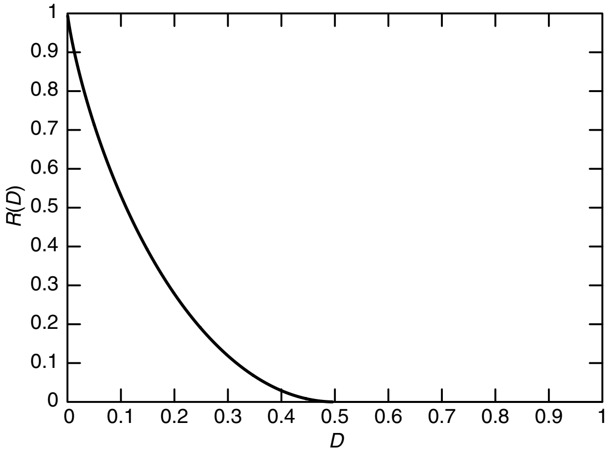


FIGURE 10.4. Rate distortion function for a Bernoulli $(\frac{1}{2})$ source.

The above calculations may seem entirely unmotivated. Why should minimizing mutual information have anything to do with quantization? The answer to this question must wait until we prove Theorem 10.2.1.

10.3.2 Gaussian Source

Although Theorem 10.2.1 is proved only for discrete sources with a bounded distortion measure, it can also be proved for well-behaved continuous sources and unbounded distortion measures. Assuming this general theorem, we calculate the rate distortion function for a Gaussian source with squared-error distortion.

Theorem 10.3.2 *The rate distortion function for a $\mathcal{N}(0, \sigma^2)$ source with squared-error distortion is*

$$R(D) = \begin{cases} \frac{1}{2}\log\frac{\sigma^2}{D}, & 0 \le D \le \sigma^2, \\ 0, & D > \sigma^2. \end{cases} \tag{10.24}$$

Proof: Let X be $\sim \mathcal{N}(0, \sigma^2)$. By the rate distortion theorem extended to continuous alphabets, we have

$$R(D) = \min_{f(\hat{x}|x): E(\hat{X}-X)^2 \le D} I(X;\hat{X}). \tag{10.25}$$

As in the preceding example, we first find a lower bound for the rate distortion function and then prove that this is achievable. Since $E(X - \hat{X})^2 \le D$, we observe that

$$I(X;\hat{X}) = h(X) - h(X|\hat{X}) \tag{10.26}$$

$$= \frac{1}{2}\log(2\pi e)\sigma^2 - h(X - \hat{X}|\hat{X}) \tag{10.27}$$

$$\ge \frac{1}{2}\log(2\pi e)\sigma^2 - h(X - \hat{X}) \tag{10.28}$$

$$\ge \frac{1}{2}\log(2\pi e)\sigma^2 - h(\mathcal{N}(0, E(X - \hat{X})^2)) \tag{10.29}$$

$$= \frac{1}{2}\log(2\pi e)\sigma^2 - \frac{1}{2}\log(2\pi e)E(X - \hat{X})^2 \tag{10.30}$$

$$\ge \frac{1}{2}\log(2\pi e)\sigma^2 - \frac{1}{2}\log(2\pi e)D \tag{10.31}$$

$$= \frac{1}{2}\log\frac{\sigma^2}{D}, \tag{10.32}$$

where (10.28) follows from the fact that conditioning reduces entropy and (10.29) follows from the fact that the normal distribution maximizes the entropy for a given second moment (Theorem 8.6.5). Hence,

$$R(D) \ge \frac{1}{2}\log\frac{\sigma^2}{D}. \tag{10.33}$$

To find the conditional density $f(\hat{x}|x)$ that achieves this lower bound, it is usually more convenient to look at the conditional density $f(x|\hat{x})$, which is sometimes called the *test channel* (thus emphasizing the duality of rate distortion with channel capacity). As in the binary case, we construct $f(x|\hat{x})$ to achieve equality in the bound. We choose the joint distribution as shown in Figure 10.5. If $D \le \sigma^2$, we choose

$$X = \hat{X} + Z, \quad \hat{X} \sim \mathcal{N}(0, \sigma^2 - D), \quad Z \sim \mathcal{N}(0, D), \tag{10.34}$$

where $\hat{X}$ and Z are independent. For this joint distribution, we calculate

$$I(X;\hat{X}) = \frac{1}{2}\log\frac{\sigma^2}{D}, \tag{10.35}$$

and $E(X-\hat{X})^2 = D$, thus achieving the bound in (10.33). If $D > \sigma^2$, we choose $\hat{X} = 0$ with probability 1, achieving $R(D) = 0$. Hence, the rate distortion function for the Gaussian source with squared-error distortion is

$$R(D) = \begin{cases} \frac{1}{2}\log\frac{\sigma^2}{D}, & 0 \le D \le \sigma^2, \\ 0, & D > \sigma^2, \end{cases} \tag{10.36}$$

as illustrated in Figure 10.6. □

We can rewrite (10.36) to express the distortion in terms of the rate,

$$D(R) = \sigma^2 2^{-2R}. \tag{10.37}$$

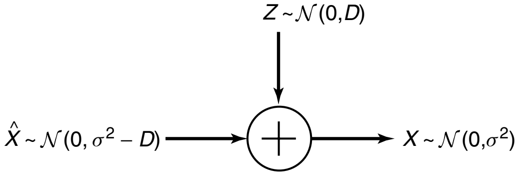

$Z \sim \mathcal{N}(0,D)$
$\hat{X} \sim \mathcal{N}(0,\sigma^2 - D)$
$X \sim \mathcal{N}(0,\sigma^2)$

FIGURE 10.5. Joint distribution for Gaussian source.

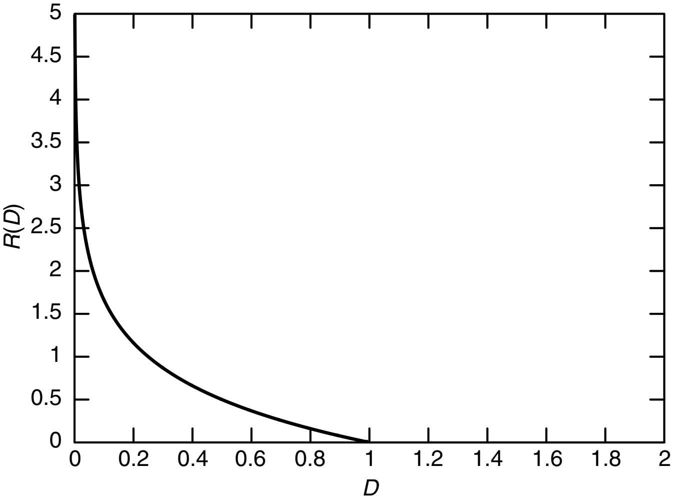


FIGURE 10.6. Rate distortion function for a Gaussian source.

Each bit of description reduces the expected distortion by a factor of 4. With a 1-bit description, the best expected square error is $\sigma^2/4$. We can compare this with the result of simple 1-bit quantization of a $\mathcal{N}(0, \sigma^2)$ random variable as described in Section 10.1. In this case, using the two regions corresponding to the positive and negative real lines and reproduction points as the centroids of the respective regions, the expected distortion is $\frac{(\pi-2)}{\pi}\sigma^2 = 0.3633\sigma^2$ (see Problem 10.1). As we prove later, the rate distortion limit $R(D)$ is achieved by considering long block lengths. This example shows that we can achieve a lower distortion by considering several distortion problems in succession (long block lengths) than can be achieved by considering each problem separately. This is somewhat surprising because we are quantizing independent random variables.

10.3.3 Simultaneous Description of Independent Gaussian Random Variables

Consider the case of representing m independent (but not identically distributed) normal random sources $X_1, \ldots, X_m$, where X_i are $\sim \mathcal{N}(0, \sigma_i^2)$, with squared-error distortion. Assume that we are given R bits with which to represent this random vector. The question naturally arises as to how we should allot these bits to the various components to minimize the total distortion. Extending the definition of the information rate distortion function to the vector case, we have

$$R(D) = \min_{f(\hat{x}^m|x^m):\mathrm{E}d(X^m,\hat{X}^m)\le D} I(X^m;\hat{X}^m), \tag{10.38}$$

where $d(x^m, \hat{x}^m) = \sum_{i=1}^m (x_i - \hat{x}_i)^2$. Now using the arguments in the preceding example, we have

$$\begin{aligned}
I(X^m;\hat{X}^m) &= h(X^m) - h(X^m|\hat{X}^m) && (10.39)\\
&= \sum_{i=1}^m h(X_i) - \sum_{i=1}^m h(X_i|X^{i-1},\hat{X}^m) && (10.40)\\
&\ge \sum_{i=1}^m h(X_i) - \sum_{i=1}^m h(X_i|\hat{X}_i) && (10.41)\\
&= \sum_{i=1}^m I(X_i;\hat{X}_i) && (10.42)\\
&\ge \sum_{i=1}^m R(D_i) && (10.43)\\
&= \sum_{i=1}^m \left(\frac{1}{2}\log\frac{\sigma_i^2}{D_i}\right)^+, && (10.44)
\end{aligned}$$

where $D_i = E(X_i - \hat{X}_i)^2$ and (10.41) follows from the fact that conditioning reduces entropy. We can achieve equality in (10.41) by choosing $f(x^m|\hat{x}^m) = \prod_{i=1}^m f(x_i|\hat{x}_i)$ and in (10.43) by choosing the distribution of each $\hat{X}_i \sim \mathcal{N}(0, \sigma_i^2 - D_i)$, as in the preceding example. Hence, the problem of finding the rate distortion function can be reduced to the following optimization (using nats for convenience):

$$R(D) = \min_{\sum D_i = D} \sum_{i=1}^m \max\left\{\frac{1}{2}\ln\frac{\sigma_i^2}{D_i}, 0\right\}. \tag{10.45}$$

Using Lagrange multipliers, we construct the functional

$$J(D) = \sum_{i=1}^m \frac{1}{2}\ln\frac{\sigma_i^2}{D_i} + \lambda\sum_{i=1}^m D_i, \tag{10.46}$$

and differentiating with respect to D_i and setting equal to 0, we have

$$\frac{\partial J}{\partial D_i} = -\frac{1}{2}\frac{1}{D_i} + \lambda = 0 \tag{10.47}$$

or

$$D_i = \lambda'. \tag{10.48}$$

Hence, the optimum allotment of the bits to the various descriptions results in an equal distortion for each random variable. This is possible if the constant λ' in (10.48) is less than σ_i^2 for all i. As the total allowable distortion D is increased, the constant λ' increases until it exceeds σ_i^2 for some i. At this point the solution (10.48) is on the boundary of the allowable region of distortions. If we increase the total distortion, we must use the Kuhn–Tucker conditions to find the minimum in (10.46). In this case the Kuhn–Tucker conditions yield

$$\frac{\partial J}{\partial D_i} = -\frac{1}{2}\frac{1}{D_i} + \lambda, \tag{10.49}$$

where λ is chosen so that

$$\frac{\partial J}{\partial D_i}\begin{cases} = 0 & \text{if } D_i < \sigma_i^2 \\ \le 0 & \text{if } D_i \ge \sigma_i^2. \end{cases} \tag{10.50}$$

It is easy to check that the solution to the Kuhn–Tucker equations is given by the following theorem:

Theorem 10.3.3 *(Rate distortion for a parallel Gaussian source) Let $X_i \sim \mathcal{N}(0, \sigma_i^2)$, $i = 1, 2, \ldots, m$, be independent Gaussian random variables, and let the distortion measure be $d(x^m, \hat{x}^m) = \sum_{i=1}^m (x_i - \hat{x}_i)^2$. Then the rate distortion function is given by*

$$R(D) = \sum_{i=1}^m \frac{1}{2}\log\frac{\sigma_i^2}{D_i}, \tag{10.51}$$

where

$$D_i = \begin{cases} \lambda & \textit{if } \lambda < \sigma_i^2, \\ \sigma_i^2 & \textit{if } \lambda \ge \sigma_i^2, \end{cases} \tag{10.52}$$

where λ is chosen so that $\sum_{i=1}^m D_i = D$.

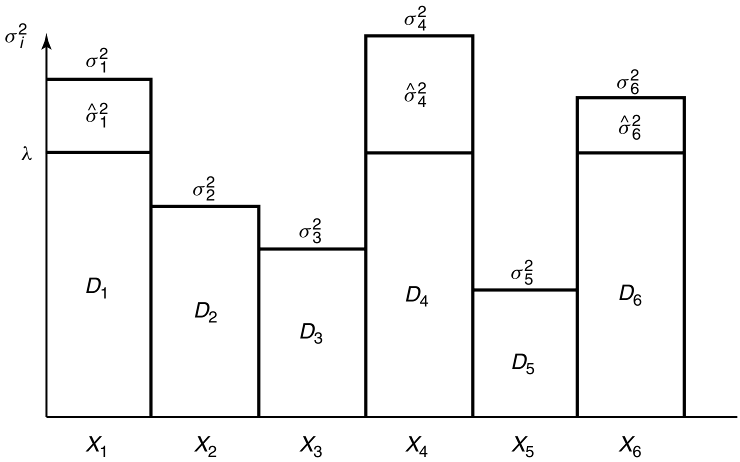


FIGURE 10.7. Reverse water-filling for independent Gaussian random variables.

This gives rise to a kind of reverse water-filling, as illustrated in Figure 10.7. We choose a constant λ and only describe those random variables with variances greater than λ. No bits are used to describe random variables with variance less than λ. Summarizing, if

$$X \sim \mathcal{N}(0, \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_m^2 \end{bmatrix}), \text{ then } \hat{X} \sim \mathcal{N}(0, \begin{bmatrix} \hat{\sigma}_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \hat{\sigma}_m^2 \end{bmatrix}),$$

and $E(X_i - \hat{X}_i)^2 = D_i$, where $D_i = \min\{\lambda, \sigma_i^2\}$. More generally, the rate distortion function for a multivariate normal vector can be obtained by reverse water-filling on the eigenvalues. We can also apply the same arguments to a Gaussian stochastic process. By the spectral representation theorem, a Gaussian stochastic process can be represented as an integral of independent Gaussian processes in the various frequency bands. Reverse water-filling on the spectrum yields the rate distortion function.

10.4 CONVERSE TO THE RATE DISTORTION THEOREM

In this section we prove the converse to Theorem 10.2.1 by showing that we cannot achieve a distortion of less than D if we describe X at a rate less than $R(D)$, where

$$R(D) = \min_{p(\hat{x}|x): \sum_{(x,\hat{x})} p(x)p(\hat{x}|x)d(x,\hat{x}) \le D} I(X; \hat{X}). \tag{10.53}$$

The minimization is over all conditional distributions $p(\hat{x}|x)$ for which the joint distribution $p(x, \hat{x}) = p(x)p(\hat{x}|x)$ satisfies the expected distortion constraint. Before proving the converse, we establish some simple properties of the information rate distortion function.

Lemma 10.4.1 *(Convexity of $R(D)$) The rate distortion function $R(D)$ given in (10.53) is a nonincreasing convex function of D.*

Proof: $R(D)$ is the minimum of the mutual information over increasingly larger sets as D increases. Thus, $R(D)$ is nonincreasing in D. To prove that $R(D)$ is convex, consider two rate distortion pairs, (R_1, D_1) and (R_2, D_2), which lie on the rate distortion curve. Let the joint distributions that achieve these pairs be $p_1(x, \hat{x}) = p(x)p_1(\hat{x}|x)$ and $p_2(x, \hat{x}) = p(x)p_2(\hat{x}|x)$. Consider the distribution $p_\lambda = \lambda p_1 + (1 - \lambda)p_2$. Since the distortion is a linear function of the distribution, we have $D(p_\lambda) = \lambda D_1 + (1 - \lambda)D_2$. Mutual information, on the other hand, is a convex function of the conditional distribution (Theorem 2.7.4), and hence

$$I_{p_\lambda}(X; \hat{X}) \le \lambda I_{p_1}(X; \hat{X}) + (1 - \lambda)I_{p_2}(X; \hat{X}). \tag{10.54}$$

Hence, by the definition of the rate distortion function,

$$\begin{aligned} R(D_\lambda) &\le I_{p_\lambda}(X; \hat{X}) && (10.55)\\ &\le \lambda I_{p_1}(X; \hat{X}) + (1 - \lambda)I_{p_2}(X; \hat{X}) && (10.56)\\ &= \lambda R(D_1) + (1 - \lambda)R(D_2), && (10.57) \end{aligned}$$

which proves that $R(D)$ is a convex function of D. □

The converse can now be proved.

Proof: (*Converse in Theorem 10.2.1*). We must show for any source X drawn i.i.d. $\sim p(x)$ with distortion measure $d(x, \hat{x})$ and any $(2^{nR}, n)$ rate distortion code with distortion $\le D$, that the rate R of the code satisfies $R \ge R(D)$. In fact, we prove that $R \ge R(D)$ even for randomized mappings f_n and g_n, as long as f_n takes on at most 2^{nR} values.

Consider any $(2^{nR}, n)$ rate distortion code defined by functions f_n and g_n as given in (10.7) and (10.8). Let $\hat{X}^n = \hat{X}^n(X^n) = g_n(f_n(X^n))$ be the reproduced sequence corresponding to X^n. Assume that $Ed(X^n, \hat{X}^n) \ge D$

for this code. Then we have the following chain of inequalities:

$$
\begin{aligned}
nR &\overset{(a)}{\geq} H(f_n(X^n)) && (10.58)\\
&\overset{(b)}{\geq} H(f_n(X^n)) - H(f_n(X^n)|X^n) && (10.59)\\
&= I(X^n; f_n(X^n)) && (10.60)\\
&\overset{(c)}{\geq} I(X^n; \hat{X}^n) && (10.61)\\
&= H(X^n) - H(X^n|\hat{X}^n) && (10.62)\\
&\overset{(d)}{=} \sum_{i=1}^{n} H(X_i) - H(X^n|\hat{X}^n) && (10.63)\\
&\overset{(e)}{=} \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}^n, X_{i-1}, \ldots, X_1) && (10.64)\\
&\overset{(f)}{\geq} \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}_i) && (10.65)\\
&= \sum_{i=1}^{n} I(X_i; \hat{X}_i) && (10.66)\\
&\overset{(g)}{\geq} \sum_{i=1}^{n} R(Ed(X_i, \hat{X}_i)) && (10.67)\\
&= n\left(\frac{1}{n}\sum_{i=1}^{n} R(Ed(X_i, \hat{X}_i))\right) && (10.68)\\
&\overset{(h)}{\geq} nR\left(\frac{1}{n}\sum_{i=1}^{n} Ed(X_i, \hat{X}_i)\right) && (10.69)\\
&\overset{(i)}{=} nR(Ed(X^n, \hat{X}^n)) && (10.70)\\
&\overset{(j)}{=} nR(D), && (10.71)
\end{aligned}
$$

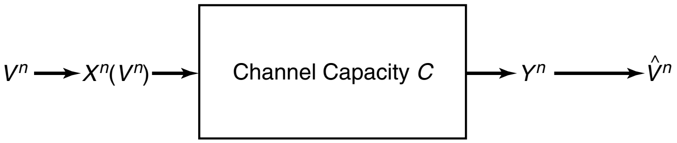
where
(a) follows from the fact that the range of f_n is at most 2^{nR}
(b) follows from the fact that $H(f_n(X^n)|X^n) \geq 0$

(c) follows from the data-processing inequality
(d) follows from the fact that the X_i are independent
(e) follows from the chain rule for entropy
(f) follows from the fact that conditioning reduces entropy
(g) follows from the definition of the rate distortion function
(h) follows from the convexity of the rate distortion function (Lemma 10.4.1) and Jensen's inequality
(i) follows from the definition of distortion for blocks of length n
(j) follows from the fact that $R(D)$ is a nonincreasing function of D and $Ed(X^n, \hat{X}^n) \leq D$

This shows that the rate R of any rate distortion code exceeds the rate distortion function $R(D)$ evaluated at the distortion level $D = Ed(X^n, \hat{X}^n)$ achieved by that code. □

A similar argument can be applied when the encoded source is passed through a noisy channel and hence we have the equivalent of the source channel separation theorem with distortion:

Theorem 10.4.1 (*Source–channel separation theorem with distortion*) *Let $V_1, V_2, \ldots, V_n$ be a finite alphabet i.i.d. source which is encoded as a sequence of n input symbols X^n of a discrete memoryless channel with capacity C. The output of the channel Y^n is mapped onto the reconstruction alphabet $\hat{V}^n = g(Y^n)$. Let $D = Ed(V^n, \hat{V}^n) = \frac{1}{n}\sum_{i=1}^{n} Ed(V_i, \hat{V}_i)$ be the average distortion achieved by this combined source and channel coding scheme. Then distortion D is achievable if and only if $C > R(D)$.*



Proof: See Problem 10.17. □

10.5 ACHIEVABILITY OF THE RATE DISTORTION FUNCTION

We now prove the achievability of the rate distortion function. We begin with a modified version of the joint AEP in which we add the condition that the pair of sequences be typical with respect to the distortion measure.

Definition Let $p(x, \hat{x})$ be a joint probability distribution on $\mathcal{X} \times \hat{\mathcal{X}}$ and let $d(x, \hat{x})$ be a distortion measure on $\mathcal{X} \times \hat{\mathcal{X}}$. For any $\epsilon > 0$, a pair of sequences $(x^n, \hat{x}^n)$ is said to be *distortion ϵ-typical* or simply *distortion typical* if

$$\left| -\frac{1}{n} \log p(x^n) - H(X) \right| < \epsilon \tag{10.72}$$

$$\left| -\frac{1}{n} \log p(\hat{x}^n) - H(\hat{X}) \right| < \epsilon \tag{10.73}$$

$$\left| -\frac{1}{n} \log p(x^n, \hat{x}^n) - H(X, \hat{X}) \right| < \epsilon \tag{10.74}$$

$$|d(x^n, \hat{x}^n) - Ed(X, \hat{X})| < \epsilon. \tag{10.75}$$

The set of distortion typical sequences is called the *distortion typical set* and is denoted $A_{d,\epsilon}^{(n)}$.

Note that this is the definition of the jointly typical set (Section 7.6) with the additional constraint that the distortion be close to the expected value. Hence, the distortion typical set is a subset of the jointly typical set (i.e., $A_{d,\epsilon}^{(n)} \subset A_{\epsilon}^{(n)}$). If $(X_i, \hat{X}_i)$ are drawn i.i.d $\sim p(x, \hat{x})$, the distortion between two random sequences

$$d(X^n, \hat{X}^n) = \frac{1}{n} \sum_{i=1}^{n} d(X_i, \hat{X}_i) \tag{10.76}$$

is an average of i.i.d. random variables, and the law of large numbers implies that it is close to its expected value with high probability. Hence we have the following lemma.

Lemma 10.5.1 *Let $(X_i, \hat{X}_i)$ be drawn i.i.d. $\sim p(x, \hat{x})$. Then* $\Pr(A_{d,\epsilon}^{(n)}) \to$ 1 *as* $n \to \infty$.

Proof: The sums in the four conditions in the definition of $A_{d,\epsilon}^{(n)}$ are all normalized sums of i.i.d random variables and hence, by the law of large numbers, tend to their respective expected values with probability 1. Hence the set of sequences satisfying all four conditions has probability tending to 1 as $n \to \infty$. □

The following lemma is a direct consequence of the definition of the distortion typical set.

Lemma 10.5.2 *For all* $(x^n, \hat{x}^n) \in A_{d,\epsilon}^{(n)}$,

$$p(\hat{x}^n) \geq p(\hat{x}^n|x^n) 2^{-n(I(X;\hat{X})+3\epsilon)}. \tag{10.77}$$

Proof: Using the definition of $A_{d,\epsilon}^{(n)}$, we can bound the probabilities $p(x^n)$, $p(\hat{x}^n)$ and $p(x^n, \hat{x}^n)$ for all $(x^n, \hat{x}^n) \in A_{d,\epsilon}^{(n)}$, and hence

$$\begin{aligned} p(\hat{x}^n|x^n) &= \frac{p(x^n, \hat{x}^n)}{p(x^n)} && (10.78)\\ &= p(\hat{x}^n)\frac{p(x^n, \hat{x}^n)}{p(x^n)p(\hat{x}^n)} && (10.79)\\ &\leq p(\hat{x}^n)\frac{2^{-n(H(X,\hat{X})-\epsilon)}}{2^{-n(H(X)+\epsilon)}2^{-n(H(\hat{X})+\epsilon)}} && (10.80)\\ &= p(\hat{x}^n)2^{n(I(X;\hat{X})+3\epsilon)}, && (10.81) \end{aligned}$$

and the lemma follows immediately. □

We also need the following interesting inequality.

Lemma 10.5.3 *For* $0 \leq x, y \leq 1$, $n > 0$,

$$(1-xy)^n \leq 1 - x + e^{-yn}. \tag{10.82}$$

Proof: Let $f(y) = e^{-y} - 1 + y$. Then $f(0) = 0$ and $f'(y) = -e^{-y} + 1 > 0$ for $y > 0$, and hence $f(y) > 0$ for $y > 0$. Hence for $0 \leq y \leq 1$, we have $1 - y \leq e^{-y}$, and raising this to the nth power, we obtain

$$(1-y)^n \leq e^{-yn}. \tag{10.83}$$

Thus, the lemma is satisfied for $x = 1$. By examination, it is clear that the inequality is also satisfied for $x = 0$. By differentiation, it is easy to see that $g_y(x) = (1-xy)^n$ is a convex function of x, and hence for $0 \leq x \leq 1$, we have

$$\begin{aligned} (1-xy)^n &= g_y(x) && (10.84)\\ &\leq (1-x)g_y(0) + xg_y(1) && (10.85)\\ &= (1-x)1 + x(1-y)^n && (10.86)\\ &\leq 1 - x + xe^{-yn} && (10.87)\\ &\leq 1 - x + e^{-yn}. \quad \square && (10.88) \end{aligned}$$

We use the preceding proof to prove the achievability of Theorem 10.2.1.

Proof: (*Achievability in Theorem 10.2.1*). Let $X_1, X_2, \ldots, X_n$ be drawn i.i.d. $\sim p(x)$ and let $d(x, \hat{x})$ be a bounded distortion measure for this source. Let the rate distortion function for this source be $R(D)$. Then for any D, and any $R > R(D)$, we will show that the rate distortion pair (R, D) is achievable by proving the existence of a sequence of rate distortion codes with rate R and asymptotic distortion D. Fix $p(\hat{x}|x)$, where $p(\hat{x}|x)$ achieves equality in (10.53). Thus, $I(X; \hat{X}) = R(D)$. Calculate $p(\hat{x}) = \sum_x p(x)p(\hat{x}|x)$. Choose $\delta > 0$. We will prove the existence of a rate distortion code with rate R and distortion less than or equal to $D + \delta$.

Generation of codebook: Randomly generate a rate distortion codebook $\mathcal{C}$ consisting of 2^{nR} sequences $\hat{X}^n$ drawn i.i.d. $\sim \prod_{i=1}^n p(\hat{x}_i)$. Index these codewords by $w \in \{1, 2, \ldots, 2^{nR}\}$. Reveal this codebook to the encoder and decoder.

Encoding: Encode X^n by w if there exists a w such that $(X^n, \hat{X}^n(w)) \in A_{d,\epsilon}^{(n)}$, the distortion typical set. If there is more than one such w, send the least. If there is no such w, let $w = 1$. Thus, nR bits suffice to describe the index w of the jointly typical codeword.

Decoding: The reproduced sequence is $\hat{X}^n(w)$.

Calculation of distortion: As in the case of the channel coding theorem, we calculate the expected distortion over the random choice of codebooks $\mathcal{C}$ as

$$\overline{D} = E_{X^n,\mathcal{C}} d(X^n, \hat{X}^n), \tag{10.89}$$

where the expectation is over the random choice of codebooks and over X^n.

For a fixed codebook $\mathcal{C}$ and choice of $\epsilon > 0$, we divide the sequences $x^n \in \mathcal{X}^n$ into two categories:

- Sequences x^n such that there exists a codeword $\hat{X}^n(w)$ that is distortion typical with x^n [i.e., $d(x^n, \hat{x}^n(w)) < D + \epsilon$]. Since the total probability of these sequences is at most 1, these sequences contribute at most $D + \epsilon$ to the expected distortion.
- Sequences x^n such that there does not exist a codeword $\hat{X}^n(w)$ that is distortion typical with x^n. Let P_e be the total probability of these sequences. Since the distortion for any individual sequence is bounded by $d_{\max}$, these sequences contribute at most $P_e d_{\max}$ to the expected distortion.

Hence, we can bound the total distortion by

$$Ed(X^n, \hat{X}^n(X^n)) \leq D + \epsilon + P_e d_{\max}, \tag{10.90}$$

which can be made less than $D + \delta$ for an appropriate choice of ϵ if P_e is small enough. Hence, if we show that P_e is small, the expected distortion is close to D and the theorem is proved.

Calculation of P_e: We must bound the probability that for a random choice of codebook $\mathcal{C}$ and a randomly chosen source sequence, there is no codeword that is distortion typical with the source sequence. Let $J(\mathcal{C})$ denote the set of source sequences x^n such that at least one codeword in $\mathcal{C}$ is distortion typical with x^n. Then

$$P_e = \sum_{\mathcal{C}} P(\mathcal{C}) \sum_{x^n : x^n \notin J(\mathcal{C})} p(x^n). \tag{10.91}$$

This is the probability of all sequences not well represented by a code, averaged over the randomly chosen code. By changing the order of summation, we can also interpret this as the probability of choosing a codebook that does not well represent sequence x^n, averaged with respect to $p(x^n)$. Thus,

$$P_e = \sum_{x^n} p(x^n) \sum_{\mathcal{C} : x^n \notin J(\mathcal{C})} p(\mathcal{C}). \tag{10.92}$$

Let us define

$$K(x^n, \hat{x}^n) = \begin{cases} 1 & \text{if } (x^n, \hat{x}^n) \in A_{d,\epsilon}^{(n)}, \\ 0 & \text{if } (x^n, \hat{x}^n) \notin A_{d,\epsilon}^{(n)}. \end{cases} \tag{10.93}$$

The probability that a single randomly chosen codeword $\hat{X}^n$ does not well represent a fixed x^n is

$$\Pr((x^n, \hat{X}^n) \notin A_{d,\epsilon}^{(n)}) = \Pr(K(x^n, \hat{X}^n) = 0) = 1 - \sum_{\hat{x}^n} p(\hat{x}^n) K(x^n, \hat{x}^n), \tag{10.94}$$

and therefore the probability that 2^{nR} independently chosen codewords do not represent x^n, averaged over $p(x^n)$, is

$$P_e = \sum_{x^n} p(x^n) \sum_{\mathcal{C} : x^n \notin J(\mathcal{C})} p(\mathcal{C}) \tag{10.95}$$

$$= \sum_{x^n} p(x^n) \left[1 - \sum_{\hat{x}^n} p(\hat{x}^n) K(x^n, \hat{x}^n) \right]^{2^{nR}}. \tag{10.96}$$

We now use Lemma 10.5.2 to bound the sum within the brackets. From Lemma 10.5.2, it follows that

$$\sum_{\hat{x}^n} p(\hat{x}^n)K(x^n, \hat{x}^n) \ge \sum_{\hat{x}^n} p(\hat{x}^n|x^n)2^{-n(I(X;\hat{X})+3\epsilon)}K(x^n, \hat{x}^n), \quad (10.97)$$

and hence

$$P_e \le \sum_{x^n} p(x^n)\left(1 - 2^{-n(I(X;\hat{X})+3\epsilon)}\sum_{\hat{x}^n} p(\hat{x}^n|x^n)K(x^n, \hat{x}^n)\right)^{2^{nR}}. \quad (10.98)$$

We now use Lemma 10.5.3 to bound the term on the right-hand side of (10.98) and obtain

$$\left(1 - 2^{-n(I(X;\hat{X})+3\epsilon)}\sum_{\hat{x}^n} p(\hat{x}^n|x^n)K(x^n, \hat{x}^n)\right)^{2^{nR}}$$
$$\le 1 - \textstyle\sum_{\hat{x}^n} p(\hat{x}^n|x^n)K(x^n, \hat{x}^n) + e^{-(2^{-n(I(X;\hat{X})+3\epsilon)}2^{nR})}. \quad (10.99)$$

Substituting this inequality in (10.98), we obtain

$$P_e \le 1 - \sum_{x^n}\sum_{\hat{x}^n} p(x^n)p(\hat{x}^n|x^n)K(x^n, \hat{x}^n) + e^{-2^{-n(I(X;\hat{X})+3\epsilon)}2^{nR}}. \quad (10.100)$$

The last term in the bound is equal to

$$e^{-2^{n(R-I(X;\hat{X})-3\epsilon)}}, \quad (10.101)$$

which goes to zero exponentially fast with n if $R > I(X;\hat{X}) + 3\epsilon$. Hence if we choose $p(\hat{x}|x)$ to be the conditional distribution that achieves the minimum in the rate distortion function, then $R > R(D)$ implies that $R > I(X;\hat{X})$ and we can choose ϵ small enough so that the last term in (10.100) goes to 0.

The first two terms in (10.100) give the probability under the joint distribution $p(x^n, \hat{x}^n)$ that the pair of sequences is not distortion typical. Hence, using Lemma 10.5.1, we obtain

$$1 - \sum_{x^n}\sum_{\hat{x}^n} p(x^n, \hat{x}^n)K(x^n, \hat{x}^n) = \Pr((X^n, \hat{X}^n) \notin A_{d,\epsilon}^{(n)}) < \epsilon \quad (10.102)$$

for n sufficiently large. Therefore, by an appropriate choice of ϵ and n, we can make P_e as small as we like.

So, for any choice of $\delta > 0$, there exists an ϵ and n such that over all randomly chosen rate R codes of block length n, the expected distortion is less than $D + \delta$. Hence, there must exist at least one code $\mathcal{C}^*$ with this rate and block length with average distortion less than $D + \delta$. Since δ was arbitrary, we have shown that (R, D) is achievable if $R > R(D)$. □

We have proved the existence of a rate distortion code with an expected distortion close to D and a rate close to $R(D)$. The similarities between the random coding proof of the rate distortion theorem and the random coding proof of the channel coding theorem are now evident. We will explore the parallels further by considering the Gaussian example, which provides some geometric insight into the problem. It turns out that channel coding is sphere packing and rate distortion coding is sphere covering.

Channel coding for the Gaussian channel. Consider a Gaussian channel, $Y_i = X_i + Z_i$, where the Z_i are i.i.d. $\sim \mathcal{N}(0, N)$ and there is a power constraint P on the power per symbol of the transmitted codeword. Consider a sequence of n transmissions. The power constraint implies that the transmitted sequence lies within a sphere of radius $\sqrt{nP}$ in $\mathcal{R}^n$. The coding problem is equivalent to finding a set of 2^{nR} sequences within this sphere such that the probability of any of them being mistaken for any other is small—the spheres of radius $\sqrt{nN}$ around each of them are almost disjoint. This corresponds to filling a sphere of radius $\sqrt{n(P+N)}$ with spheres of radius $\sqrt{nN}$. One would expect that the largest number of spheres that could be fit would be the ratio of their volumes, or, equivalently, the nth power of the ratio of their radii. Thus, if M is the number of codewords that can be transmitted efficiently, we have

$$M \le \frac{(\sqrt{n(P+N)})^n}{(\sqrt{nN})^n} = \left(\frac{P+N}{N}\right)^{\frac{n}{2}}. \tag{10.103}$$

The results of the channel coding theorem show that it is possible to do this efficiently for large n; it is possible to find approximately

$$2^{nC} = \left(\frac{P+N}{N}\right)^{\frac{n}{2}} \tag{10.104}$$

codewords such that the noise spheres around them are almost disjoint (the total volume of their intersection is arbitrarily small).

Rate distortion for the Gaussian source. Consider a Gaussian source of variance σ^2. A $(2^{nR}, n)$ rate distortion code for this source with distortion D is a set of 2^{nR} sequences in $\mathcal{R}^n$ such that most source sequences of length n (all those that lie within a sphere of radius $\sqrt{n\sigma^2}$) are within a distance $\sqrt{nD}$ of some codeword. Again, by the sphere-packing argument, it is clear that the minimum number of codewords required is

$$2^{nR(D)} = \left(\frac{\sigma^2}{D}\right)^{\frac{n}{2}}. \tag{10.105}$$

The rate distortion theorem shows that this minimum rate is asymptotically achievable (i.e., that there exists a collection of spheres of radius $\sqrt{nD}$ that cover the space except for a set of arbitrarily small probability).

The above geometric arguments also enable us to transform a good code for channel transmission into a good code for rate distortion. In both cases, the essential idea is to fill the space of source sequences: In channel transmission, we want to find the largest set of codewords that have a large minimum distance between codewords, whereas in rate distortion, we wish to find the smallest set of codewords that covers the entire space. If we have any set that meets the sphere packing bound for one, it will meet the sphere packing bound for the other. In the Gaussian case, choosing the codewords to be Gaussian with the appropriate variance is asymptotically optimal for both rate distortion and channel coding.

10.6 STRONGLY TYPICAL SEQUENCES AND RATE DISTORTION

In Section 10.5 we proved the existence of a rate distortion code of rate $R(D)$ with average distortion close to D. In fact, not only is the average distortion close to D, but the total probability that the distortion is greater than $D + \delta$ is close to 0. The proof of this is similar to the proof in Section 10.5; the main difference is that we will use strongly typical sequences rather than weakly typical sequences. This will enable us to give an upper bound to the probability that a typical source sequence is not well represented by a randomly chosen codeword in (10.94). We now outline an alternative proof based on strong typicality that will provide a stronger and more intuitive approach to the rate distortion theorem.

We begin by defining strong typicality and quoting a basic theorem bounding the probability that two sequences are jointly typical. The properties of strong typicality were introduced by Berger [53] and were explored in detail in the book by Csiszár and Körner [149]. We will define strong typicality (as in Chapter 11) and state a fundamental lemma (Lemma 10.6.2).

Definition A sequence $x^n \in \mathcal{X}^n$ is said to be *ϵ-strongly typical* with respect to a distribution $p(x)$ on $\mathcal{X}$ if:

1. For all $a \in \mathcal{X}$ with $p(a) > 0$, we have
$$\left|\frac{1}{n}N(a|x^n) - p(a)\right| < \frac{\epsilon}{|\mathcal{X}|}. \tag{10.106}$$
2. For all $a \in \mathcal{X}$ with $p(a) = 0$, $N(a|x^n) = 0$.

$N(a|x^n)$ is the number of occurrences of the symbol a in the sequence x^n.

The set of sequences $x^n \in \mathcal{X}^n$ such that x^n is strongly typical is called the *strongly typical* set and is denoted $A_\epsilon^{*(n)}(X)$ or $A_\epsilon^{*(n)}$ when the random variable is understood from the context.

Definition A pair of sequences $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ is said to be *ϵ-strongly typical* with respect to a distribution $p(x, y)$ on $\mathcal{X} \times \mathcal{Y}$ if:

1. For all $(a, b) \in \mathcal{X} \times \mathcal{Y}$ with $p(a, b) > 0$, we have
$$\left|\frac{1}{n}N(a, b|x^n, y^n) - p(a, b)\right| < \frac{\epsilon}{|\mathcal{X}||\mathcal{Y}|}. \tag{10.107}$$
2. For all $(a, b) \in \mathcal{X} \times \mathcal{Y}$ with $p(a, b) = 0$, $N(a, b|x^n, y^n) = 0$.

$N(a, b|x^n, y^n)$ is the number of occurrences of the pair (a, b) in the pair of sequences (x^n, y^n).

The set of sequences $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ such that (x^n, y^n) is strongly typical is called the *strongly typical* set and is denoted $A_\epsilon^{*(n)}(X, Y)$ or $A_\epsilon^{*(n)}$. From the definition, it follows that if $(x^n, y^n) \in A_\epsilon^{*(n)}(X, Y)$, then $x^n \in A_\epsilon^{*(n)}(X)$. From the strong law of large numbers, the following lemma is immediate.

Lemma 10.6.1 *Let (X_i, Y_i) be drawn i.i.d. $\sim p(x, y)$. Then* $\Pr(A_\epsilon^{*(n)}) \to 1$ *as $n \to \infty$.*

We will use one basic result, which bounds the probability that an independently drawn sequence will be seen as jointly strongly typical with a given sequence. Theorem 7.6.1 shows that if we choose X^n and Y^n independently, the probability that they will be weakly jointly typical is $\approx 2^{-nI(X;Y)}$. The following lemma extends the result to strongly typical sequences. This is stronger than the earlier result in that it gives a lower bound on the probability that a randomly chosen sequence is jointly typical with a fixed typical x^n.

Lemma 10.6.2 *Let $Y_1, Y_2, \ldots, Y_n$ be drawn i.i.d. $\sim p(y)$. For $x^n \in A_\epsilon^{*(n)}(X)$, the probability that $(x^n, Y^n) \in A_\epsilon^{*(n)}$ is bounded by*

$$2^{-n(I(X;Y)+\epsilon_1)} \le \Pr((x^n, Y^n) \in A_\epsilon^{*(n)}) \le 2^{-n(I(X;Y)-\epsilon_1)}, \qquad (10.108)$$

where ϵ_1 goes to 0 as $\epsilon \to 0$ and $n \to \infty$.

Proof: We will not prove this lemma, but instead, outline the proof in Problem 10.16 at the end of the chapter. In essence, the proof involves finding a lower bound on the size of the conditionally typical set. □

We will proceed directly to the achievability of the rate distortion function. We will only give an outline to illustrate the main ideas. The construction of the codebook and the encoding and decoding are similar to the proof in Section 10.5.

Proof: Fix $p(\hat{x}|x)$. Calculate $p(\hat{x}) = \sum_x p(x)p(\hat{x}|x)$. Fix $\epsilon > 0$. Later we will choose ϵ appropriately to achieve an expected distortion less than $D + \delta$.

Generation of codebook: Generate a rate distortion codebook $\mathcal{C}$ consisting of 2^{nR} sequences $\hat{X}^n$ drawn i.i.d. $\sim \prod_i p(\hat{x}_i)$. Denote the sequences $\hat{X}^n(1), \ldots, \hat{X}^n(2^{nR})$.

Encoding: Given a sequence X^n, index it by w if there exists a w such that $(X^n, \hat{X}^n(w)) \in A_\epsilon^{*(n)}$, the strongly jointly typical set. If there is more than one such w, send the first in lexicographic order. If there is no such w, let $w = 1$.

Decoding: Let the reproduced sequence be $\hat{X}^n(w)$.

Calculation of distortion: As in the case of the proof in Section 10.5, we calculate the expected distortion over the random choice of codebook as

$$D = E_{X^n,\mathcal{C}} d(X^n, \hat{X}^n) \qquad (10.109)$$

$$= E_{\mathcal{C}} \sum_{x^n} p(x^n) d(x^n, \hat{X}^n(x^n)) \qquad (10.110)$$

$$= \sum_{x^n} p(x^n) E_{\mathcal{C}} d(x^n, \hat{X}^n), \qquad (10.111)$$

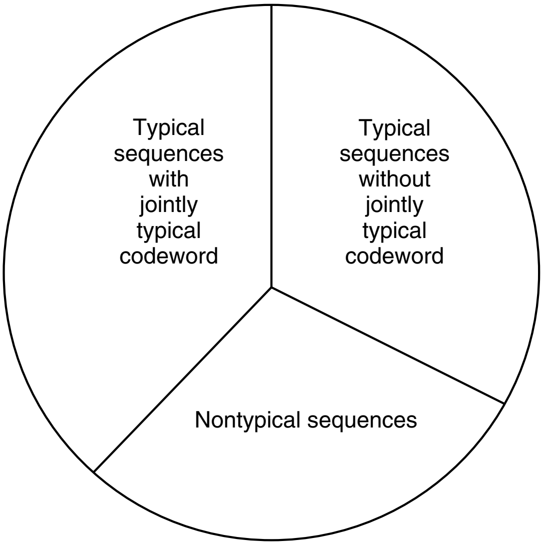

Typical sequences with jointly typical codeword
Typical sequences without jointly typical codeword
Nontypical sequences

FIGURE 10.8. Classes of source sequences in rate distortion theorem.

where the expectation is over the random choice of codebook. For a fixed codebook $\mathcal{C}$, we divide the sequences $x^n \in \mathcal{X}^n$ into three categories, as shown in Figure 10.8.

- *Nontypical sequences* $x^n \notin A_\epsilon^{*(n)}$. The total probability of these sequences can be made less than ϵ by choosing n large enough. Since the individual distortion between any two sequences is bounded by $d_{\max}$, the nontypical sequences can contribute at most $\epsilon d_{\max}$ to the expected distortion.
- *Typical sequences* $x^n \in A_\epsilon^{*(n)}$ *such that there exists a codeword* $\hat{X}^n(w)$ *that is jointly typical with* x^n. In this case, since the source sequence and the codeword are strongly jointly typical, the continuity of the distortion as a function of the joint distribution ensures that they are also distortion typical. Hence, the distortion between these x^n and their codewords is bounded by $D + \epsilon d_{\max}$, and since the total probability of these sequences is at most 1, these sequences contribute at most $D + \epsilon d_{\max}$ to the expected distortion.
- *Typical sequences* $x^n \in A_\epsilon^{*(n)}$ *such that there does not exist a codeword* $\hat{X}^n$ *that is jointly typical with* x^n. Let P_e be the total probability of these sequences. Since the distortion for any individual sequence is bounded by $d_{\max}$, these sequences contribute at most $P_e d_{\max}$ to the expected distortion.

The sequences in the first and third categories are the sequences that may not be well represented by this rate distortion code. The probability of the first category of sequences is less than ϵ for sufficiently large n. The probability of the last category is P_e, which we will show can be made small. This will prove the theorem that the total probability of sequences that are not well represented is small. In turn, we use this to show that the average distortion is close to D.

Calculation of P_e: We must bound the probability that there is no codeword that is jointly typical with the given sequence X^n. From the joint AEP, we know that the probability that X^n and any $\hat{X}^n$ are jointly typical is $\doteq 2^{-nI(X;\hat{X})}$. Hence the expected number of jointly typical $\hat{X}^n(w)$ is $2^{nR}2^{-nI(X;\hat{X})}$, which is exponentially large if $R > I(X;\hat{X})$.

But this is not sufficient to show that $P_e \to 0$. We must show that the probability that there is no codeword that is jointly typical with X^n goes to zero. The fact that the expected number of jointly typical codewords is exponentially large does not ensure that there will at least one with high probability. Just as in (10.94), we can expand the probability of error as

$$P_e = \sum_{x^n \in A_\epsilon^{*(n)}} p(x^n)\left[1 - \Pr((x^n, \hat{X}^n) \in A_\epsilon^{*(n)})\right]^{2^{nR}}. \tag{10.112}$$

From Lemma 10.6.2 we have

$$\Pr((x^n, \hat{X}^n) \in A_\epsilon^{*(n)}) \geq 2^{-n(I(X;\hat{X})+\epsilon_1)}. \tag{10.113}$$

Substituting this in (10.112) and using the inequality $(1-x)^n \leq e^{-nx}$, we have

$$P_e \leq e^{-(2^{nR}2^{-n(I(X;\hat{X})+\epsilon_1)})}, \tag{10.114}$$

which goes to 0 as $n \to \infty$ if $R > I(X;\hat{X}) + \epsilon_1$. Hence for an appropriate choice of ϵ and n, we can get the total probability of all badly represented sequences to be as small as we want. Not only is the expected distortion close to D, but with probability going to 1, we will find a codeword whose distortion with respect to the given sequence is less than $D + \delta$. □

10.7 CHARACTERIZATION OF THE RATE DISTORTION FUNCTION

We have defined the information rate distortion function as

$$R(D) = \min_{q(\hat{x}|x):\sum_{(x,\hat{x})} p(x)q(\hat{x}|x)d(x,\hat{x}) \leq D} I(X;\hat{X}), \tag{10.115}$$

where the minimization is over all conditional distributions $q(\hat{x}|x)$ for which the joint distribution $p(x)q(\hat{x}|x)$ satisfies the expected distortion constraint. This is a standard minimization problem of a convex function over the convex set of all $q(\hat{x}|x) \geq 0$ satisfying $\sum_{\hat{x}} q(\hat{x}|x) = 1$ for all x and $\sum q(\hat{x}|x)p(x)d(x, \hat{x}) \leq D$.

We can use the method of Lagrange multipliers to find the solution. We set up the functional

$$J(q) = \sum_x \sum_{\hat{x}} p(x)q(\hat{x}|x) \log \frac{q(\hat{x}|x)}{\sum_x p(x)q(\hat{x}|x)}$$

$$+\lambda \sum_x \sum_{\hat{x}} p(x)q(\hat{x}|x)d(x, \hat{x}) \tag{10.116}$$

$$+\sum_x \nu(x) \sum_{\hat{x}} q(\hat{x}|x), \tag{10.117}$$

where the last term corresponds to the constraint that $q(\hat{x}|x)$ is a conditional probability mass function. If we let $q(\hat{x}) = \sum_x p(x)q(\hat{x}|x)$ be the distribution on $\hat{X}$ induced by $q(\hat{x}|x)$, we can rewrite $J(q)$ as

$$J(q) = \sum_x \sum_{\hat{x}} p(x)q(\hat{x}|x) \log \frac{q(\hat{x}|x)}{q(\hat{x})}$$

$$+\lambda \sum_x \sum_{\hat{x}} p(x)q(\hat{x}|x)d(x, \hat{x}) \tag{10.118}$$

$$+\sum_x \nu(x) \sum_{\hat{x}} q(\hat{x}|x). \tag{10.119}$$

Differentiating with respect to $q(\hat{x}|x)$, we have

$$\frac{\partial J}{\partial q(\hat{x}|x)} = p(x) \log \frac{q(\hat{x}|x)}{q(\hat{x})} + p(x) - \sum_{x'} p(x')q(\hat{x}|x') \frac{1}{q(\hat{x})} p(x)$$

$$+\lambda p(x)d(x, \hat{x}) + \nu(x) = 0. \tag{10.120}$$

Setting $\log \mu(x) = \nu(x)/p(x)$, we obtain

$$p(x)\left[\log \frac{q(\hat{x}|x)}{q(\hat{x})} + \lambda d(x, \hat{x}) + \log \mu(x)\right] = 0 \tag{10.121}$$

or

$$q(\hat{x}|x) = \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\mu(x)}. \tag{10.122}$$

Since $\sum_{\hat{x}} q(\hat{x}|x) = 1$, we must have

$$\mu(x) = \sum_{\hat{x}} q(\hat{x})e^{-\lambda d(x,\hat{x})} \tag{10.123}$$

or

$$q(\hat{x}|x) = \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{x}} q(\hat{x})e^{-\lambda d(x,\hat{x})}}. \tag{10.124}$$

Multiplying this by $p(x)$ and summing over all x, we obtain

$$q(\hat{x}) = q(\hat{x}) \sum_{x} \frac{p(x)e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{x}'} q(\hat{x}')e^{-\lambda d(x,\hat{x}')}}. \tag{10.125}$$

If $q(\hat{x}) > 0$, we can divide both sides by $q(\hat{x})$ and obtain

$$\sum_{x} \frac{p(x)e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{x}'} q(\hat{x}')e^{-\lambda d(x,\hat{x}')}} = 1 \tag{10.126}$$

for all $\hat{x} \in \hat{\mathcal{X}}$. We can combine these $|\hat{\mathcal{X}}|$ equations with the equation defining the distortion and calculate λ and the $|\hat{\mathcal{X}}|$ unknowns $q(\hat{x})$. We can use this and (10.124) to find the optimum conditional distribution.

The above analysis is valid if $q(\hat{x})$ is unconstrained (i.e., $q(\hat{x}) > 0$ for all $\hat{x}$). The inequality condition $q(\hat{x}) > 0$ is covered by the Kuhn–Tucker conditions, which reduce to

$$\begin{aligned} \frac{\partial J}{\partial q(\hat{x}|x)} &= 0 \text{ if } q(\hat{x}|x) > 0, \\ &\geq 0 \text{ if } q(\hat{x}|x) = 0. \end{aligned} \tag{10.127}$$

Substituting the value of the derivative, we obtain the conditions for the minimum as

$$\sum_{x} \frac{p(x)e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{x}'} q(\hat{x}')e^{-\lambda d(x,\hat{x}')}} = 1 \quad \text{if } q(\hat{x}) > 0, \tag{10.128}$$

$$\leq 1 \quad \text{if } q(\hat{x}) = 0. \tag{10.129}$$

This characterization will enable us to check if a given $q(\hat{x})$ is a solution to the minimization problem. However, it is not easy to solve for the optimum output distribution from these equations. In the next section we provide an iterative algorithm for computing the rate distortion function. This algorithm is a special case of a general algorithm for finding the minimum relative entropy distance between two convex sets of probability densities.

10.8 COMPUTATION OF CHANNEL CAPACITY AND THE RATE DISTORTION FUNCTION

Consider the following problem: Given two convex sets A and B in $\mathcal{R}^n$ as shown in Figure 10.9, we would like to find the minimum distance between them:

$$d_{\min} = \min_{a \in A, b \in B} d(a, b), \tag{10.130}$$

where $d(a, b)$ is the Euclidean distance between a and b. An intuitively obvious algorithm to do this would be to take any point $x \in A$, and find the $y \in B$ that is closest to it. Then fix this y and find the closest point in A. Repeating this process, it is clear that the distance decreases at each stage. Does it converge to the minimum distance between the two sets? Csiszár and Tusnády [155] have shown that if the sets are convex and if the distance satisfies certain conditions, this alternating minimization algorithm will indeed converge to the minimum. In particular, if the sets are sets of probability distributions and the distance measure is the relative entropy, the algorithm does converge to the minimum relative entropy between the two sets of distributions.

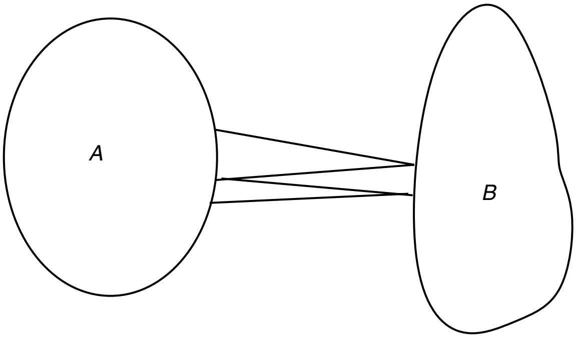


FIGURE 10.9. Distance between convex sets.

To apply this algorithm to rate distortion, we have to rewrite the rate distortion function as a minimum of the relative entropy between two sets. We begin with a simple lemma. A form of this lemma comes up again in theorem 13.1.1, establishing the duality of channel capacity universal data compression.

Lemma 10.8.1 *Let $p(x)p(y|x)$ be a given joint distribution. Then the distribution $r(y)$ that minimizes the relative entropy $D(p(x)p(y|x)||p(x)\, r(y))$ is the marginal distribution $r^*(y)$ corresponding to $p(y|x)$:*

$$D(p(x)p(y|x)||p(x)r^*(y)) = \min_{r(y)} D(p(x)p(y|x)||p(x)r(y)), \tag{10.131}$$

where $r^*(y) = \sum_x p(x)p(y|x)$. Also,

$$\max_{r(x|y)} \sum_{x,y} p(x)p(y|x) \log \frac{r(x|y)}{p(x)} = \sum_{x,y} p(x)p(y|x) \log \frac{r^*(x|y)}{p(x)}, \tag{10.132}$$

where

$$r^*(x|y) = \frac{p(x)p(y|x)}{\sum_x p(x)p(y|x)}. \tag{10.133}$$

Proof

$$D(p(x)p(y|x)||p(x)r(y)) - D(p(x)p(y|x)||p(x)r^*(y))$$

$$= \sum_{x,y} p(x)p(y|x) \log \frac{p(x)p(y|x)}{p(x)r(y)} \tag{10.134}$$

$$- \sum_{x,y} p(x)p(y|x) \log \frac{p(x)p(y|x)}{p(x)r^*(y)} \tag{10.135}$$

$$= \sum_{x,y} p(x)p(y|x) \log \frac{r^*(y)}{r(y)} \tag{10.136}$$

$$= \sum_{y} r^*(y) \log \frac{r^*(y)}{r(y)} \tag{10.137}$$

$$= D(r^*||r) \tag{10.138}$$

$$\geq 0. \tag{10.139}$$

The proof of the second part of the lemma is left as an exercise. □

We can use this lemma to rewrite the minimization in the definition of the rate distortion function as a double minimization,

$$R(D) = \min_{r(\hat{x})} \min_{q(\hat{x}|x):\sum p(x)q(\hat{x}|x)d(x,\hat{x})\leq D} \sum_x \sum_{\hat{x}} p(x)q(\hat{x}|x) \log \frac{q(\hat{x}|x)}{r(\hat{x})}. \tag{10.140}$$

If A is the set of all joint distributions with marginal $p(x)$ that satisfy the distortion constraints and if B the set of product distributions $p(x)r(\hat{x})$ with arbitrary $r(\hat{x})$, we can write

$$R(D) = \min_{q\in B} \min_{p\in A} D(p||q). \tag{10.141}$$

We now apply the process of alternating minimization, which is called the *Blahut–Arimoto algorithm* in this case. We begin with a choice of λ and an initial output distribution $r(\hat{x})$ and calculate the $q(\hat{x}|x)$ that minimizes the mutual information subject to the distortion constraint. We can use the method of Lagrange multipliers for this minimization to obtain

$$q(\hat{x}|x) = \frac{r(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{x}} r(\hat{x})e^{-\lambda d(x,\hat{x})}}. \tag{10.142}$$

For this conditional distribution $q(\hat{x}|x)$, we calculate the output distribution $r(\hat{x})$ that minimizes the mutual information, which by Lemma 10.8.1 is

$$r(\hat{x}) = \sum_x p(x)q(\hat{x}|x). \tag{10.143}$$

We use this output distribution as the starting point of the next iteration. Each step in the iteration, minimizing over $q(\cdot|\cdot)$ and then minimizing over $r(\cdot)$, reduces the right-hand side of (10.140). Thus, there is a limit, and the limit has been shown to be $R(D)$ by Csiszár [139], where the value of D and $R(D)$ depends on λ. Thus, choosing λ appropriately sweeps out the $R(D)$ curve.

A similar procedure can be applied to the calculation of channel capacity. Again we rewrite the definition of channel capacity,

$$C = \max_{r(x)} I(X;Y) = \max_{r(x)} \sum_x \sum_y r(x)p(y|x) \log \frac{r(x)p(y|x)}{r(x)\sum_{x'} r(x')p(y|x')} \tag{10.144}$$

as a double maximization using Lemma 10.8.1,

$$C = \max_{q(x|y)} \max_{r(x)} \sum_x \sum_y r(x)p(y|x) \log \frac{q(x|y)}{r(x)}. \tag{10.145}$$

In this case, the Csiszár–Tusnady algorithm becomes one of alternating maximization—we start with a guess of the maximizing distribution $r(x)$ and find the best conditional distribution, which is, by Lemma 10.8.1,

$$q(x|y) = \frac{r(x)p(y|x)}{\sum_x r(x)p(y|x)}. \tag{10.146}$$

For this conditional distribution, we find the best input distribution $r(x)$ by solving the constrained maximization problem with Lagrange multipliers. The optimum input distribution is

$$r(x) = \frac{\prod_y (q(x|y))^{p(y|x)}}{\sum_x \prod_y (q(x|y))^{p(y|x)}}, \tag{10.147}$$

which we can use as the basis for the next iteration.

These algorithms for the computation of the channel capacity and the rate distortion function were established by Blahut [65] and Arimoto [25] and the convergence for the rate distortion computation was proved by Csiszár [139]. The alternating minimization procedure of Csiszár and Tusnady can be specialized to many other situations as well, including the EM algorithm [166], and the algorithm for finding the log-optimal portfolio for a stock market [123].

SUMMARY

Rate distortion. The rate distortion function for a source $X \sim p(x)$ and distortion measure $d(x, \hat{x})$ is

$$R(D) = \min_{p(\hat{x}|x): \sum_{(x,\hat{x})} p(x)p(\hat{x}|x)d(x,\hat{x}) \le D} I(X; \hat{X}), \tag{10.148}$$

where the minimization is over all conditional distributions $p(\hat{x}|x)$ for which the joint distribution $p(x, \hat{x}) = p(x)p(\hat{x}|x)$ satisfies the expected distortion constraint.

Rate distortion theorem. If $R > R(D)$, there exists a sequence of codes $\hat{X}^n(X^n)$ with the number of codewords $|\hat{X}^n(\cdot)| \le 2^{nR}$ with $Ed(X^n, \hat{X}^n(X^n)) \to D$. If $R < R(D)$, no such codes exist.

Bernoulli source. For a Bernoulli source with Hamming distortion,

$$R(D) = H(p) - H(D). \tag{10.149}$$

Gaussian source. For a Gaussian source with squared-error distortion,

$$R(D) = \frac{1}{2}\log\frac{\sigma^2}{D}. \tag{10.150}$$

Source–channel separation. A source with rate distortion $R(D)$ can be sent over a channel of capacity C and recovered with distortion D if and only if $R(D) < C$.

Multivariate Gaussian source. The rate distortion function for a multivariate normal vector with Euclidean mean-squared-error distortion is given by reverse water-filling on the eigenvalues.

PROBLEMS

10.1 *One-bit quantization of a single Gaussian random variable*. Let $X \sim \mathcal{N}(0, \sigma^2)$ and let the distortion measure be squared error. Here we do not allow block descriptions. Show that the optimum reproduction points for 1-bit quantization are $\pm\sqrt{\frac{2}{\pi}}\sigma$ and that the expected distortion for 1-bit quantization is $\frac{\pi-2}{\pi}\sigma^2$. Compare this with the distortion rate bound $D = \sigma^2 2^{-2R}$ for $R = 1$.

10.2 *Rate distortion function with infinite distortion*. Find the rate distortion function $R(D) = \min I(X; \hat{X})$ for $X \sim$ Bernoulli $(\frac{1}{2})$ and distortion

$$d(x, \hat{x}) = \begin{cases} 0, & x = \hat{x} \\ 1, & x = 1, \hat{x} = 0 \\ \infty, & x = 0, \hat{x} = 1. \end{cases}$$

10.3 *Rate distortion for binary source with asymmetric distortion .* Fix $p(\hat{x}|x)$ and evaluate $I(X;\hat{X})$ and D for

$$X \sim \text{Bernoulli}\left(\frac{1}{2}\right),$$

$$d(x,\hat{x}) = \begin{bmatrix} 0 & a \\ b & 0 \end{bmatrix}.$$

(The rate distortion function cannot be expressed in closed form.)

10.4 *Properties of $R(D)$.* Consider a discrete source $X \in \mathcal{X} = \{1, 2, \ldots, m\}$ with distribution $p_1, p_2, \ldots, p_m$ and a distortion measure $d(i, j)$. Let $R(D)$ be the rate distortion function for this source and distortion measure. Let $d'(i, j) = d(i, j) - w_i$ be a new distortion measure, and let $R'(D)$ be the corresponding rate distortion function. Show that $R'(D) = R(D + \overline{w})$, where $\overline{w} = \sum p_i w_i$, and use this to show that there is no essential loss of generality in assuming that $\min_{\hat{x}} d(i, \hat{x}) = 0$ (i.e., for each $x \in \mathcal{X}$, there is one symbol $\hat{x}$ that reproduces the source with zero distortion). This result is due to Pinkston [420].

10.5 *Rate distortion for uniform source with Hamming distortion.* Consider a source X uniformly distributed on the set $\{1, 2, \ldots, m\}$. Find the rate distortion function for this source with Hamming distortion; that is,

$$d(x,\hat{x}) = \begin{cases} 0 & \text{if } x = \hat{x}, \\ 1 & \text{if } x \neq \hat{x}. \end{cases}$$

10.6 *Shannon lower bound for the rate distortion function.* Consider a source X with a distortion measure $d(x, \hat{x})$ that satisfies the following property: All columns of the distortion matrix are permutations of the set $\{d_1, d_2, \ldots, d_m\}$. Define the function

$$\phi(D) = \max_{\mathbf{p}:\sum_{i=1}^{m} p_i d_i \le D} H(\mathbf{p}). \tag{10.151}$$

The Shannon lower bound on the rate distortion function [485] is proved by the following steps:

(a) Show that $\phi(D)$ is a concave function of D.

(b) Justify the following series of inequalities for $I(X;\hat{X})$ if $Ed(X,\hat{X}) \le D$,

$$I(X;\hat{X}) = H(X) - H(X|\hat{X}) \tag{10.152}$$

$$= H(X) - \sum_{\hat{x}} p(\hat{x}) H(X|\hat{X} = \hat{x}) \tag{10.153}$$

$$\geq H(X) - \sum_{\hat{x}} p(\hat{x}) \phi(D_{\hat{x}}) \tag{10.154}$$

$$\geq H(X) - \phi\left(\sum_{\hat{x}} p(\hat{x}) D_{\hat{x}}\right) \tag{10.155}$$

$$\geq H(X) - \phi(D), \tag{10.156}$$

where $D_{\hat{x}} = \sum_x p(x|\hat{x}) d(x, \hat{x})$.

(c) Argue that

$$R(D) \geq H(X) - \phi(D), \tag{10.157}$$

which is the Shannon lower bound on the rate distortion function.

(d) If, in addition, we assume that the source has a uniform distribution and that the rows of the distortion matrix are permutations of each other, then $R(D) = H(X) - \phi(D)$ (i.e., the lower bound is tight).

10.7 *Erasure distortion.* Consider $X \sim$ Bernoulli $(\frac{1}{2})$, and let the distortion measure be given by the matrix

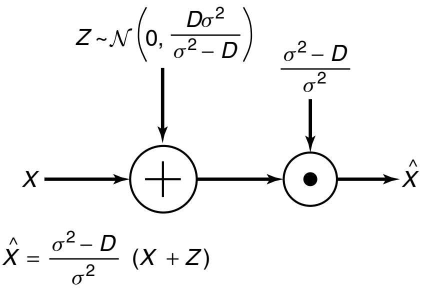
$$d(x, \hat{x}) = \begin{bmatrix} 0 & 1 & \infty \\ \infty & 1 & 0 \end{bmatrix}. \tag{10.158}$$

Calculate the rate distortion function for this source. Can you suggest a simple scheme to achieve any value of the rate distortion function for this source?

10.8 *Bounds on the rate distortion function for squared-error distortion.* For the case of a continuous random variable X with mean zero and variance σ^2 and squared-error distortion, show that

$$h(X) - \frac{1}{2}\log(2\pi e D) \leq R(D) \leq \frac{1}{2}\log\frac{\sigma^2}{D}. \tag{10.159}$$

For the upper bound, consider the following joint distribution:


$Z \sim \mathcal{N}\left(0, \frac{D\sigma^2}{\sigma^2 - D}\right)$
$\frac{\sigma^2 - D}{\sigma^2}$
X
$\hat{X}$
$\hat{X} = \frac{\sigma^2 - D}{\sigma^2}(X + Z)$

Are Gaussian random variables harder or easier to describe than other random variables with the same variance?

10.9 *Properties of optimal rate distortion code*. A good (R, D) rate distortion code with $R \approx R(D)$ puts severe constraints on the relationship of the source X^n and the representations $\hat{X}^n$. Examine the chain of inequalities (10.58–10.71) considering the conditions for equality and interpret as properties of a good code. For example, equality in (10.59) implies that $\hat{X}^n$ is a deterministic function of X^n.

10.10 *Rate distortion*. Find and verify the rate distortion function $R(D)$ for X uniform on $\mathcal{X} = \{1, 2, \ldots, 2m\}$ and

$$d(x, \hat{x}) = \begin{cases} 1 & \text{for } x - \hat{x} \text{ odd,} \\ 0 & \text{for } x - \hat{x} \text{ even,} \end{cases}$$

where $\hat{X}$ is defined on $\hat{\mathcal{X}} = \{1, 2, \ldots, 2m\}$. (You may wish to use the Shannon lower bound in your argument.)

10.11 *Lower bound*. Let

$$X \sim \frac{e^{-x^4}}{\int_{-\infty}^{\infty} e^{-x^4} dx}$$

and

$$\frac{\int x^4 e^{-x^4} dx}{\int e^{-x^4} dx} = c.$$

Define $g(a) = \max h(X)$ over all densities such that $EX^4 \leq a$. Let $R(D)$ be the rate distortion function for X with the density above and with distortion criterion $d(x, \hat{x}) = (x - \hat{x})^4$. Show that $R(D) \geq g(c) - g(D)$.

10.12 *Adding a column to the distortion matrix.* Let $R(D)$ be the rate distortion function for an i.i.d. process with probability mass function $p(x)$ and distortion function $d(x, \hat{x})$, $x \in \mathcal{X}$, $\hat{x} \in \hat{\mathcal{X}}$. Now suppose that we add a new reproduction symbol $\hat{x}_0$ to $\hat{\mathcal{X}}$ with associated distortion $d(x, \hat{x}_0)$, $x \in \mathcal{X}$. Does this increase or decrease $R(D)$, and why?

10.13 *Simplification.* Suppose that $\mathcal{X} = \{1, 2, 3, 4\}$, $\hat{\mathcal{X}} = \{1, 2, 3, 4\}$, $p(i) = \frac{1}{4}$, $i = 1, 2, 3, 4$, and $X_1, X_2, \ldots$ are i.i.d. $\sim p(x)$. The distortion matrix $d(x, \hat{x})$ is given by

	1	2	3	4
1	0	0	1	1
2	0	0	1	1
3	1	1	0	0
4	1	1	0	0

(a) Find $R(0)$, the rate necessary to describe the process with zero distortion.

(b) Find the rate distortion function $R(D)$. There are some irrelevant distinctions in alphabets $\mathcal{X}$ and $\hat{\mathcal{X}}$, which allow the problem to be collapsed.

(c) Suppose that we have a nonuniform distribution $p(i) = p_i$, $i = 1, 2, 3, 4$. What is $R(D)$?

10.14 *Rate distortion for two independent sources.* Can one compress two independent sources simultaneously better than by compressing the sources individually? The following problem addresses this question. Let $\{X_i\}$ be i.i.d. $\sim p(x)$ with distortion $d(x, \hat{x})$ and rate distortion function $R_X(D)$. Similarly, let $\{Y_i\}$ be i.i.d. $\sim p(y)$ with distortion $d(y, \hat{y})$ and rate distortion function $R_Y(D)$. Suppose we now wish to describe the process $\{(X_i, Y_i)\}$ subject to distortions $Ed(X, \hat{X}) \le D_1$ and $Ed(Y, \hat{Y}) \le D_2$. Thus, a rate $R_{X,Y}(D_1, D_2)$ is sufficient, where

$$R_{X,Y}(D_1, D_2) = \min_{p(\hat{x}, \hat{y}|x, y): Ed(X, \hat{X}) \le D_1, Ed(Y, \hat{Y}) \le D_2} I(X, Y; \hat{X}, \hat{Y}).$$

Now suppose that the $\{X_i\}$ process and the $\{Y_i\}$ process are independent of each other.

(a) Show that

$$R_{X,Y}(D_1, D_2) \ge R_X(D_1) + R_Y(D_2).$$

(b) Does equality hold?

Now answer the question.

10.15 *Distortion rate function.* Let

$$D(R) = \min_{p(\hat{x}|x): I(X;\hat{X}) \le R} Ed(X, \hat{X}) \tag{10.160}$$

be the distortion rate function.

(a) Is $D(R)$ increasing or decreasing in R?

(b) Is $D(R)$ convex or concave in R?

(c) Converse for distortion rate functions: We now wish to prove the converse by focusing on $D(R)$. Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim p(x)$. Suppose that one is given a $(2^{nR}, n)$ rate distortion code $X^n \to i(X^n) \to \hat{X}^n(i(X^n))$, with $i(X^n) \in 2^{nR}$, and suppose that the resulting distortion is $D = Ed(X^n, \hat{X}^n(i(X^n)))$. We must show that $D \ge D(R)$. Give reasons for the following steps in the proof:

$$D = Ed(X^n, \hat{X}^n(i(X^n))) \tag{10.161}$$

$$\overset{(a)}{=} E\frac{1}{n}\sum_{i=1}^{n} d(X_i, \hat{X}_i) \tag{10.162}$$

$$\overset{(b)}{=} \frac{1}{n}\sum_{i=1}^{n} \mathrm{Ed}(X_i, \hat{X}_i) \tag{10.163}$$

$$\overset{(c)}{\ge} \frac{1}{n}\sum_{i=1}^{n} D\left(I(X_i; \hat{X}_i)\right) \tag{10.164}$$

$$\overset{(d)}{\ge} D\left(\frac{1}{n}\sum_{i=1}^{n} I(X_i; \hat{X}_i)\right) \tag{10.165}$$

$$\overset{(e)}{\ge} D\left(\frac{1}{n} I(X^n; \hat{X}^n)\right) \tag{10.166}$$

$$\overset{(f)}{\ge} D(R). \tag{10.167}$$

10.16 *Probability of conditionally typical sequences.* In Chapter 7 we calculated the probability that two independently drawn sequences X^n and Y^n are weakly jointly typical. To prove the rate distortion theorem, however, we need to calculate this probability when one of the sequences is fixed and the other is random. The techniques of weak typicality allow us only to calculate the average set size of the conditionally typical set. Using the ideas of strong typicality, on the other hand, provides us with stronger bounds that work for all typical x^n sequences. We outline the proof that $\Pr\{(x^n, Y^n) \in A_\epsilon^{*(n)}\} \approx 2^{-nI(X;Y)}$ for all typical x^n. This approach was introduced by Berger [53] and is fully developed in the book by Csiszár and Körner [149].

Let (X_i, Y_i) be drawn i.i.d. $\sim p(x, y)$. Let the marginals of X and Y be $p(x)$ and $p(y)$, respectively.

(a) Let $A_\epsilon^{*(n)}$ be the strongly typical set for X. Show that

$$|A_\epsilon^{*(n)}| \doteq 2^{nH(X)}. \tag{10.168}$$

(*Hint:* Theorems 11.1.1 and 11.1.3.)

(b) The *joint type* of a pair of sequences (x^n, y^n) is the proportion of times $(x_i, y_i) = (a, b)$ in the pair of sequences:

$$p_{x^n,y^n}(a, b) = \frac{1}{n} N(a, b|x^n, y^n) = \frac{1}{n} \sum_{i=1}^{n} I(x_i = a, y_i = b). \tag{10.169}$$

The *conditional type* of a sequence y^n given x^n is a stochastic matrix that gives the proportion of times a particular element of $\mathcal{Y}$ occurred with each element of $\mathcal{X}$ in the pair of sequences. Specifically, the conditional type $V_{y^n|x^n}(b|a)$ is defined as

$$V_{y^n|x^n}(b|a) = \frac{N(a, b|x^n, y^n)}{N(a|x^n)}. \tag{10.170}$$

Show that the number of conditional types is bounded by $(n + 1)^{|\mathcal{X}||\mathcal{Y}|}$.

(c) The set of sequences $y^n \in \mathcal{Y}^n$ with conditional type V with respect to a sequence x^n is called the *conditional type class* $T_V(x^n)$. Show that

$$\frac{1}{(n + 1)^{|\mathcal{X}||\mathcal{Y}|}} 2^{nH(Y|X)} \le |T_V(x^n)| \le 2^{nH(Y|X)}. \tag{10.171}$$

(d) The sequence $y^n \in \mathcal{Y}^n$ is said to be *ϵ-strongly conditionally typical* with the sequence x^n with respect to the conditional distribution $V(\cdot|\cdot)$ if the conditional type is close to V. The conditional type should satisfy the following two conditions:

(i) For all $(a, b) \in \mathcal{X} \times \mathcal{Y}$ with $V(b|a) > 0$,

$$\frac{1}{n}\left|N(a, b|x^n, y^n) - V(b|a)N(a|x^n)\right| \le \frac{\epsilon}{|\mathcal{Y}| + 1}. \tag{10.172}$$

(ii) $N(a, b|x^n, y^n) = 0$ for all (a, b) such that $V(b|a) = 0$.

The set of such sequences is called the *conditionally typical set* and is denoted $A_\epsilon^{*(n)}(Y|x^n)$. Show that the number of sequences y^n that are conditionally typical with a given $x^n \in \mathcal{X}^n$ is bounded by

$$\frac{1}{(n+1)^{|\mathcal{X}||\mathcal{Y}|}} 2^{n(H(Y|X)-\epsilon_1)} \le |A_\epsilon^{*(n)}(Y|x^n)| \le (n+1)^{|\mathcal{X}||\mathcal{Y}|} 2^{n(H(Y|X)+\epsilon_1)}, \tag{10.173}$$

where $\epsilon_1 \to 0$ as $\epsilon \to 0$.

(e) For a pair of random variables (X, Y) with joint distribution $p(x, y)$, the ϵ-*strongly typical* set $A_\epsilon^{*(n)}$ is the set of sequences $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ satisfying

(i)

$$\left|\frac{1}{n}N(a, b|x^n, y^n) - p(a, b)\right| < \frac{\epsilon}{|\mathcal{X}||\mathcal{Y}|} \tag{10.174}$$

for every pair $(a, b) \in \mathcal{X} \times \mathcal{Y}$ with $p(a, b) > 0$.

(ii) $N(a, b|x^n, y^n) = 0$ for all $(a, b) \in \mathcal{X} \times \mathcal{Y}$ with $p(a, b) = 0$.

The set of ϵ-strongly jointly typical sequences is called the ϵ-*strongly jointly typical set* and is denoted $A_\epsilon^{*(n)}(X, Y)$. Let (X, Y) be drawn i.i.d. $\sim p(x, y)$. For any x^n such that there exists at least one pair $(x^n, y^n) \in A_\epsilon^{*(n)}(X, Y)$, the set of sequences y^n such that $(x^n, y^n) \in A_\epsilon^{*(n)}$ satisfies

$$\frac{1}{(n+1)^{|\mathcal{X}||\mathcal{Y}|}} 2^{n(H(Y|X)-\delta(\epsilon))} \le |\{y^n : (x^n, y^n) \in A_\epsilon^{*(n)}\}| \le (n+1)^{|\mathcal{X}||\mathcal{Y}|} 2^{n(H(Y|X)+\delta(\epsilon))}, \tag{10.175}$$

where $\delta(\epsilon) \to 0$ as $\epsilon \to 0$. In particular, we can write

$$2^{n(H(Y|X)-\epsilon_2)} \le |\{y^n : (x^n, y^n) \in A_\epsilon^{*(n)}\}| \le 2^{n(H(Y|X)+\epsilon_2)}, \tag{10.176}$$

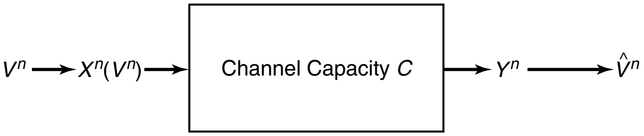
where we can make ϵ_2 arbitrarily small with an appropriate choice of ϵ and n.

(f) Let $Y_1, Y_2, \ldots, Y_n$ be drawn i.i.d. $\sim \prod p(y_i)$. For $x^n \in A_\epsilon^{*(n)}$, the probability that $(x^n, Y^n) \in A_\epsilon^{*(n)}$ is bounded by

$$2^{-n(I(X;Y)+\epsilon_3)} \le \Pr((x^n, Y^n) \in A_\epsilon^{*(n)}) \le 2^{-n(I(X;Y)-\epsilon_3)}, \tag{10.177}$$

where ϵ_3 goes to 0 as $\epsilon \to 0$ and $n \to \infty$.

10.17 *Source–channel separation theorem with distortion.* Let V_1, $V_2, \ldots, V_n$ be a finite alphabet i.i.d. source which is encoded as a sequence of n input symbols X^n of a discrete memoryless channel. The output of the channel Y^n is mapped onto the reconstruction alphabet $\hat{V}^n = g(Y^n)$. Let $D = Ed(V^n, \hat{V}^n) = \frac{1}{n}\sum_{i=1}^n Ed(V_i, \hat{V}_i)$ be the average distortion achieved by this combined source and channel coding scheme.



(a) Show that if $C > R(D)$, where $R(D)$ is the rate distortion function for V, it is possible to find encoders and decoders that achieve a average distortion arbitrarily close to D.

(b) (Converse) Show that if the average distortion is equal to D, the capacity of the channel C must be greater than $R(D)$.

10.18 *Rate distortion.* Let $d(x, \hat{x})$ be a distortion function. We have a source $X \sim p(x)$. Let $R(D)$ be the associated rate distortion function.

(a) Find $\tilde{R}(D)$ in terms of $R(D)$, where $\tilde{R}(D)$ is the rate distortion function associated with the distortion $\tilde{d}(x, \hat{x}) = d(x, \hat{x}) + a$ for some constant $a > 0$. (They are not equal.)

(b) Now suppose that $d(x, \hat{x}) \ge 0$ for all $x, \hat{x}$ and define a new distortion function $d^*(x, \hat{x}) = bd(x, \hat{x})$, where b is some number ≥ 0. Find the associated rate distortion function $R^*(D)$ in terms of $R(D)$.

(c) Let $X \sim N(0, \sigma^2)$ and $d(x, \hat{x}) = 5(x - \hat{x})^2 + 3$. What is $R(D)$?

10.19 *Rate distortion with two constraints*. Let X_i be iid $\sim p(x)$. We are given two distortion functions, $d_1(x, \hat{x})$ and $d_2(x, \hat{x})$. We wish to describe X^n at rate R and reconstruct it with distortions $Ed_1(X^n, \hat{X}_1^n) \le D_1$, and $Ed_2(X^n, \hat{X}_2^n) \le D_2$, as shown here:

$$X^n \longrightarrow i(X^n) \longrightarrow (\hat{X}_1^n(i), \hat{X}_2^n(i))$$

$$\begin{aligned} D_1 &= Ed_1(X_1^n, \hat{X}_1^n) \\ D_2 &= Ed_2(X_1^n, \hat{X}_2^n). \end{aligned}$$

Here $i(\cdot)$ takes on 2^{nR} values. What is the rate distortion function $R(D_1, D_2)$?

10.20 *Rate distortion*. Consider the standard rate distortion problem, X_i i.i.d. $\sim p(x)$, $X^n \to i(X^n) \to \hat{X}^n$, $|i(\cdot)| = 2^{nR}$. Consider two distortion criteria $d_1(x, \hat{x})$ and $d_2(x, \hat{x})$. Suppose that $d_1(x, \hat{x}) \le d_2(x, \hat{x})$ for all $x \in \mathcal{X}, \hat{x} \in \hat{\mathcal{X}}$. Let $R_1(D)$ and $R_2(D)$ be the corresponding rate distortion functions.

(a) Find the inequality relationship between $R_1(D)$ and $R_2(D)$.

(b) Suppose that we must describe the source $\{X_i\}$ at the minimum rate R achieving $d_1(X^n, \hat{X}_1^n) \le D$ and $d_2(X^n, \hat{X}_2^n) \le D$. Thus,

$$X^n \to i(X^n) \to \begin{cases} \hat{X}_1^n(i(X^n)) \\ \hat{X}_2^n(i(X^n)) \end{cases}$$

and $|i(\cdot)| = 2^{nR}$.
Find the minimum rate R.

HISTORICAL NOTES

The idea of rate distortion was introduced by Shannon in his original paper [472]. He returned to it and dealt with it exhaustively in his 1959 paper [485], which proved the first rate distortion theorem. Meanwhile, Kolmogorov and his school in the Soviet Union began to develop rate distortion theory in 1956. Stronger versions of the rate distortion theorem have been proved for more general sources in the comprehensive book by Berger [52].

The inverse water-filling solution for the rate distortion function for parallel Gaussian sources was established by McDonald and Schultheiss

[381]. An iterative algorithm for the calculation of the rate distortion function for a general i.i.d. source and arbitrary distortion measure was described by Blahut [65], Arimoto [25], and Csiszár [139]. This algorithm is a special case of a general alternating minimization algorithm due to Csiszár and Tusnády [155].

CHAPTER 11

INFORMATION THEORY AND STATISTICS

We now explore the relationship between information theory and statistics. We begin by describing the method of types, which is a powerful technique in large deviation theory. We use the method of types to calculate the probability of rare events and to show the existence of universal source codes. We also consider the problem of testing hypotheses and derive the best possible error exponents for such tests (the Chernoff–Stein lemma). Finally, we treat the estimation of the parameters of a distribution and describe the role of Fisher information.

11.1 METHOD OF TYPES

The AEP for discrete random variables (Chapter 3) focuses our attention on a small subset of typical sequences. The method of types is an even more powerful procedure in which we consider sequences that have the same empirical distribution. With this restriction, we can derive strong bounds on the number of sequences with a particular empirical distribution and the probability of each sequence in this set. It is then possible to derive strong error bounds for the channel coding theorem and prove a variety of rate distortion results. The method of types was fully developed by Csiszár and Körner [149], who obtained most of their results from this point of view.

Let $X_1, X_2, \ldots, X_n$ be a sequence of n symbols from an alphabet $\mathcal{X} = \{a_1, a_2, \ldots, a_{|\mathcal{X}|}\}$. We use the notation x^n and $\mathbf{x}$ interchangeably to denote a sequence $x_1, x_2, \ldots, x_n$.

Definition The *type* $P_{\mathbf{x}}$ (or empirical probability distribution) of a sequence $x_1, x_2, \ldots, x_n$ is the relative proportion of occurrences of each

Elements of Information Theory, Second Edition, By Thomas M. Cover and Joy A. Thomas

symbol of $\mathcal{X}$ (i.e., $P_{\mathbf{x}}(a) = N(a|\mathbf{x})/n$ for all $a \in \mathcal{X}$, where $N(a|\mathbf{x})$ is the number of times the symbol a occurs in the sequence $\mathbf{x} \in \mathcal{X}^n$).

The type of a sequence $\mathbf{x}$ is denoted as $P_{\mathbf{x}}$. It is a probability mass function on $\mathcal{X}$. (Note that in this chapter, we will use capital letters to denote types and distributions. We also loosely use the word *distribution* to mean a probability mass function.)

Definition The *probability simplex in $\mathcal{R}^m$* is the set of points $\mathbf{x} = (x_1, x_2, \ldots, x_m) \in \mathcal{R}^m$ such that $x_i \geq 0$, $\sum_{i=1}^m x_i = 1$.

The probability simplex is an $(m-1)$-dimensional manifold in m-dimensional space. When $m = 3$, the probability simplex is the set of points $\{(x_1, x_2, x_3) : x_1 \geq 0, x_2 \geq 0, x_3 \geq 0, x_1 + x_2 + x_3 = 1\}$ (Figure 11.1). Since this is a triangular two-dimensional flat in $\mathcal{R}^3$, we use a triangle to represent the probability simplex in later sections of this chapter.

Definition Let $\mathcal{P}_n$ denote the *set of types with denominator n*.

For example, if $\mathcal{X} = \{0, 1\}$, the set of possible types with denominator n is

$$\mathcal{P}_n = \left\{ (P(0), P(1)) : \left(\frac{0}{n}, \frac{n}{n}\right), \left(\frac{1}{n}, \frac{n-1}{n}\right), \ldots, \left(\frac{n}{n}, \frac{0}{n}\right) \right\}. \quad (11.1)$$

Definition If $P \in \mathcal{P}_n$, the set of sequences of length n and type P is called the *type class* of P, denoted $T(P)$:

$$T(P) = \{\mathbf{x} \in \mathcal{X}^n : P_{\mathbf{x}} = P\}. \quad (11.2)$$

The type class is sometimes called the *composition class* of P.

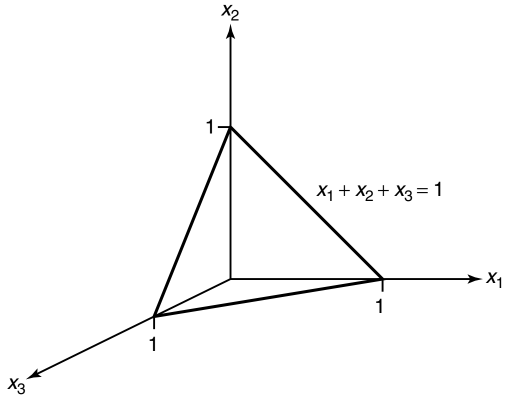


FIGURE 11.1. Probability simplex in $\mathcal{R}^3$.

Example 11.1.1 Let $\mathcal{X} = \{1, 2, 3\}$, a ternary alphabet. Let $\mathbf{x} = 11321$. Then the type $P_{\mathbf{x}}$ is

$$P_{\mathbf{x}}(1) = \frac{3}{5}, \quad P_{\mathbf{x}}(2) = \frac{1}{5}, \quad P_{\mathbf{x}}(3) = \frac{1}{5}. \tag{11.3}$$

The type class of $P_{\mathbf{x}}$ is the set of all sequences of length 5 with three 1's, one 2, and one 3. There are 20 such sequences, and

$$T(P_{\mathbf{x}}) = \{11123, 11132, 11213, \ldots, 32111\}. \tag{11.4}$$

The number of elements in $T(P)$ is

$$|T(P)| = \binom{5}{3, 1, 1} = \frac{5!}{3!\,1!\,1!} = 20. \tag{11.5}$$

The essential power of the method of types arises from the following theorem, which shows that the number of types is at most polynomial in n.

Theorem 11.1.1

$$|\mathcal{P}_n| \le (n+1)^{|\mathcal{X}|}. \tag{11.6}$$

Proof: There are $|\mathcal{X}|$ components in the vector that specifies $P_{\mathbf{x}}$. The numerator in each component can take on only $n+1$ values. So there are at most $(n+1)^{|\mathcal{X}|}$ choices for the type vector. Of course, these choices are not independent (e.g., the last choice is fixed by the others). But this is a sufficiently good upper bound for our needs. □

The crucial point here is that there are only a polynomial number of types of length n. Since the number of sequences is exponential in n, it follows that at least one type has exponentially many sequences in its type class. In fact, the largest type class has essentially the same number of elements as the entire set of sequences, to first order in the exponent.

Now, we assume that the sequence $X_1, X_2, \ldots, X_n$ is drawn i.i.d. according to a distribution $Q(x)$. All sequences with the same type have the same probability, as shown in the following theorem. Let $Q^n(x^n) = \prod_{i=1}^n Q(x_i)$ denote the product distribution associated with Q.

Theorem 11.1.2 *If $X_1, X_2, \ldots, X_n$ are drawn i.i.d. according to $Q(x)$, the probability of $\mathbf{x}$ depends only on its type and is given by*

$$Q^n(\mathbf{x}) = 2^{-n(H(P_{\mathbf{x}})+D(P_{\mathbf{x}}||Q))}. \tag{11.7}$$

Proof

$$
\begin{aligned}
Q^n(\mathbf{x}) &= \prod_{i=1}^{n} Q(x_i) && (11.8)\\
&= \prod_{a\in\mathcal{X}} Q(a)^{N(a|\mathbf{x})} && (11.9)\\
&= \prod_{a\in\mathcal{X}} Q(a)^{nP_{\mathbf{x}}(a)} && (11.10)\\
&= \prod_{a\in\mathcal{X}} 2^{nP_{\mathbf{x}}(a)\log Q(a)} && (11.11)\\
&= \prod_{a\in\mathcal{X}} 2^{n(P_{\mathbf{x}}(a)\log Q(a)-P_{\mathbf{x}}(a)\log P_{\mathbf{x}}(a)+P_{\mathbf{x}}(a)\log P_{\mathbf{x}}(a))} && (11.12)\\
&= 2^{n\sum_{a\in\mathcal{X}}(-P_{\mathbf{x}}(a)\log\frac{P_{\mathbf{x}}(a)}{Q(a)}+P_{\mathbf{x}}(a)\log P_{\mathbf{x}}(a))} && (11.13)\\
&= 2^{n(-D(P_{\mathbf{x}}||Q)-H(P_{\mathbf{x}}))}. \quad \square && (11.14)
\end{aligned}
$$

Corollary *If* $\mathbf{x}$ *is in the type class of* Q*, then*

$$Q^n(\mathbf{x}) = 2^{-nH(Q)}. \tag{11.15}$$

Proof: If $\mathbf{x} \in T(Q)$, then $P_{\mathbf{x}} = Q$, which can be substituted into (11.14). □

Example 11.1.2 The probability that a fair die produces a particular sequence of length n with precisely $n/6$ occurrences of each face (n is a multiple of 6) is $2^{-nH(\frac{1}{6},\frac{1}{6},\ldots,\frac{1}{6})} = 6^{-n}$. This is obvious. However, if the die has a probability mass function $(\frac{1}{3}, \frac{1}{3}, \frac{1}{6}, \frac{1}{12}, \frac{1}{12}, 0)$, the probability of observing a particular sequence with precisely these frequencies is precisely $2^{-nH(\frac{1}{3},\frac{1}{3},\frac{1}{6},\frac{1}{12},\frac{1}{12},0)}$ for n a multiple of 12. This is more interesting.

We now give an estimate of the size of a type class $T(P)$.

Theorem 11.1.3 (*Size of a type class* $T(P)$) *For any type* $P \in \mathcal{P}_n$*,*

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{nH(P)} \le |T(P)| \le 2^{nH(P)}. \tag{11.16}$$

Proof: The exact size of $T(P)$ is easy to calculate. It is a simple combinatorial problem—the number of ways of arranging $nP(a_1), nP(a_2), \ldots,$

$nP(a_{|\mathcal{X}|})$ objects in a sequence, which is

$$|T(P)| = \binom{n}{nP(a_1),\ nP(a_2), \ldots, nP(a_{|\mathcal{X}|})}. \tag{11.17}$$

This value is hard to manipulate, so we derive simple exponential bounds on its value.

We suggest two alternative proofs for the exponential bounds. The first proof uses Stirling's formula [208] to bound the factorial function, and after some algebra, we can obtain the bounds of the theorem. We give an alternative proof. We first prove the upper bound. Since a type class must have probability ≤ 1, we have

$$\begin{aligned} 1 &\geq P^n(T(P)) && (11.18)\\ &= \sum_{\mathbf{x}\in T(P)} P^n(\mathbf{x}) && (11.19)\\ &= \sum_{\mathbf{x}\in T(P)} 2^{-nH(P)} && (11.20)\\ &= |T(P)|2^{-nH(P)}, && (11.21) \end{aligned}$$

using Theorem 11.1.2. Thus,

$$|T(P)| \leq 2^{nH(P)}. \tag{11.22}$$

Now for the lower bound. We first prove that the type class $T(P)$ has the highest probability among all type classes under the probability distribution P:

$$P^n(T(P)) \geq P^n(T(\hat{P})) \quad \text{for all } \hat{P} \in \mathcal{P}_n. \tag{11.23}$$

We lower bound the ratio of probabilities,

$$\begin{aligned} \frac{P^n(T(P))}{P^n(T(\hat{P}))} &= \frac{|T(P)| \prod_{a\in\mathcal{X}} P(a)^{nP(a)}}{|T(\hat{P})| \prod_{a\in\mathcal{X}} P(a)^{n\hat{P}(a)}} && (11.24)\\ &= \frac{\binom{n}{nP(a_1),\, nP(a_2),\ldots,nP(a_{|\mathcal{X}|})} \prod_{a\in\mathcal{X}} P(a)^{nP(a)}}{\binom{n}{n\hat{P}(a_1),\, n\hat{P}(a_2),\ldots,n\hat{P}(a_{|\mathcal{X}|})} \prod_{a\in\mathcal{X}} P(a)^{n\hat{P}(a)}} && (11.25)\\ &= \prod_{a\in\mathcal{X}} \frac{(n\hat{P}(a))!}{(nP(a))!} P(a)^{n(P(a)-\hat{P}(a))}. && (11.26) \end{aligned}$$

Now using the simple bound (easy to prove by separately considering the cases $m \geq n$ and $m < n$)

$$\frac{m!}{n!} \geq n^{m-n}, \tag{11.27}$$

we obtain

$$\begin{aligned}
\frac{P^n(T(P))}{P^n(T(\hat{P}))} &\geq \prod_{a\in\mathcal{X}} (nP(a))^{n\hat{P}(a)-nP(a)} P(a)^{n(P(a)-\hat{P}(a))} && (11.28)\\
&= \prod_{a\in\mathcal{X}} n^{n(\hat{P}(a)-P(a))} && (11.29)\\
&= n^{n\left(\sum_{a\in\mathcal{X}} \hat{P}(a)-\sum_{a\in\mathcal{X}} P(a)\right)} && (11.30)\\
&= n^{n(1-1)} && (11.31)\\
&= 1. && (11.32)
\end{aligned}$$

Hence, $P^n(T(P)) \geq P^n(T(\hat{P}))$. The lower bound now follows easily from this result, since

$$\begin{aligned}
1 &= \sum_{Q\in\mathcal{P}_n} P^n(T(Q)) && (11.33)\\
&\leq \sum_{Q\in\mathcal{P}_n} \max_Q P^n(T(Q)) && (11.34)\\
&= \sum_{Q\in\mathcal{P}_n} P^n(T(P)) && (11.35)\\
&\leq (n+1)^{|\mathcal{X}|} P^n(T(P)) && (11.36)\\
&= (n+1)^{|\mathcal{X}|} \sum_{\mathbf{x}\in T(P)} P^n(\mathbf{x}) && (11.37)\\
&= (n+1)^{|\mathcal{X}|} \sum_{\mathbf{x}\in T(P)} 2^{-nH(P)} && (11.38)\\
&= (n+1)^{|\mathcal{X}|} |T(P)| 2^{-nH(P)}, && (11.39)
\end{aligned}$$

where (11.36) follows from Theorem 11.1.1 and (11.38) follows from Theorem 11.1.2. □

We give a slightly better approximation for the binary case.

Example 11.1.3 (*Binary alphabet*) In this case, the type is defined by the number of 1's in the sequence, and the size of the type class is therefore $\binom{n}{k}$. We show that

$$\frac{1}{n+1}2^{nH\left(\frac{k}{n}\right)} \le \binom{n}{k} \le 2^{nH\left(\frac{k}{n}\right)}. \tag{11.40}$$

These bounds can be proved using Stirling's approximation for the factorial function (Lemma 17.5.1). But we provide a more intuitive proof below.

We first prove the upper bound. From the binomial formula, for any p,

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1. \tag{11.41}$$

Since all the terms of the sum are positive for $0 \le p \le 1$, each of the terms is less than 1. Setting $p = k/n$ and taking the kth term, we get

$$1 \ge \binom{n}{k}\left(\frac{k}{n}\right)^k \left(1-\frac{k}{n}\right)^{n-k} \tag{11.42}$$

$$= \binom{n}{k} 2^{k\log\frac{k}{n} + (n-k)\log\frac{n-k}{n}} \tag{11.43}$$

$$= \binom{n}{k} 2^{n\left(\frac{k}{n}\log\frac{k}{n} + \frac{n-k}{n}\log\frac{n-k}{n}\right)} \tag{11.44}$$

$$= \binom{n}{k} 2^{-nH\left(\frac{k}{n}\right)}. \tag{11.45}$$

Hence,

$$\binom{n}{k} \le 2^{nH\left(\frac{k}{n}\right)}. \tag{11.46}$$

For the lower bound, let S be a random variable with a binomial distribution with parameters n and p. The most likely value of S is $S = \langle np \rangle$. This can easily be verified from the fact that

$$\frac{P(S=i+1)}{P(S=i)} = \frac{n-i}{i+1}\frac{p}{1-p} \tag{11.47}$$

and considering the cases when $i < np$ and when $i > np$. Then, since there are $n+1$ terms in the binomial sum,

$$1 = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \le (n+1) \max_k \binom{n}{k} p^k (1-p)^{n-k} \tag{11.48}$$

$$= (n+1) \binom{n}{\langle np \rangle} p^{\langle np \rangle} (1-p)^{n-\langle np \rangle}. \tag{11.49}$$

Now let $p = k/n$. Then we have

$$1 \le (n+1) \binom{n}{k} \left(\frac{k}{n}\right)^k \left(1 - \frac{k}{n}\right)^{n-k}, \tag{11.50}$$

which by the arguments in (11.45) is equivalent to

$$\frac{1}{n+1} \le \binom{n}{k} 2^{-nH\left(\frac{k}{n}\right)}, \tag{11.51}$$

or

$$\binom{n}{k} \ge \frac{2^{nH\left(\frac{k}{n}\right)}}{n+1}. \tag{11.52}$$

Combining the two results, we see that

$$\binom{n}{k} \doteq 2^{nH\left(\frac{k}{n}\right)}. \tag{11.53}$$

A more precise bound can be found in theorem 17.5.1 when $k \ne 0$ or n.

Theorem 11.1.4 (*Probability of type class*) *for any $P \in \mathcal{P}_n$ and any distribution Q, the probability of the type class $T(P)$ under Q^n is $2^{-nD(P||Q)}$ to first order in the exponent. More precisely,*

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(P||Q)} \le Q^n(T(P)) \le 2^{-nD(P||Q)}. \tag{11.54}$$

Proof: We have

$$Q^n(T(P)) = \sum_{\mathbf{x} \in T(P)} Q^n(\mathbf{x}) \tag{11.55}$$

$$= \sum_{\mathbf{x} \in T(P)} 2^{-n(D(P||Q)+H(P))} \tag{11.56}$$

$$= |T(P)| 2^{-n(D(P||Q)+H(P))}, \tag{11.57}$$

by Theorem 11.1.2. Using the bounds on $|T(P)|$ derived in Theorem 11.1.3, we have

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(P||Q)} \le Q^n(T(P)) \le 2^{-nD(P||Q)}. \qquad \square \tag{11.58}$$

We can summarize the basic theorems concerning types in four equations:

$$|\mathcal{P}_n| \le (n+1)^{|\mathcal{X}|}, \tag{11.59}$$

$$Q^n(\mathbf{x}) = 2^{-n(D(P_{\mathbf{x}}||Q)+H(P_{\mathbf{x}}))}, \tag{11.60}$$

$$|T(P)| \doteq 2^{nH(P)}, \tag{11.61}$$

$$Q^n(T(P)) \doteq 2^{-nD(P||Q)}. \tag{11.62}$$

These equations state that there are only a polynomial number of types and that there are an exponential number of sequences of each type. We also have an exact formula for the probability of any sequence of type P under distribution Q and an approximate formula for the probability of a type class.

These equations allow us to calculate the behavior of long sequences based on the properties of the type of the sequence. For example, for long sequences drawn i.i.d. according to some distribution, the type of the sequence is close to the distribution generating the sequence, and we can use the properties of this distribution to estimate the properties of the sequence. Some of the applications that will be dealt with in the next few sections are as follows:

- The law of large numbers
- Universal source coding
- Sanov's theorem
- The Chernoff–Stein lemma and hypothesis testing
- Conditional probability and limit theorems

11.2 LAW OF LARGE NUMBERS

The concept of type and type classes enables us to give an alternative statement of the law of large numbers. In fact, it can be used as a proof of a version of the weak law in the discrete case. The most important property of types is that there are only a polynomial number of types, and an exponential number of sequences of each type. Since the probability of each type class depends exponentially on the relative entropy distance between the type P and the distribution Q, type classes that are far from the true distribution have exponentially smaller probability.

Given an $\epsilon > 0$, we can define a typical set T_Q^ϵ of sequences for the distribution Q^n as

$$T_Q^\epsilon = \{x^n : D(P_{x^n}||Q) \le \epsilon\}. \tag{11.63}$$

Then the probability that x^n is not typical is

$$\begin{aligned}
1 - Q^n(T_Q^\epsilon) &= \sum_{P:D(P||Q)>\epsilon} Q^n(T(P)) && (11.64)\\
&\le \sum_{P:D(P||Q)>\epsilon} 2^{-nD(P||Q)} \quad \text{(Theorem 11.1.4)} && (11.65)\\
&\le \sum_{P:D(P||Q)>\epsilon} 2^{-n\epsilon} && (11.66)\\
&\le (n+1)^{|\mathcal{X}|} 2^{-n\epsilon} \quad \text{(Theorem 11.1.1)} && (11.67)\\
&= 2^{-n\left(\epsilon - |\mathcal{X}|\frac{\log(n+1)}{n}\right)}, && (11.68)
\end{aligned}$$

which goes to 0 as $n \to \infty$. Hence, the probability of the typical set T_Q^ϵ goes to 1 as $n \to \infty$. This is similar to the AEP proved in Chapter 3, which is a form of the weak law of large numbers. We now prove that the empirical distribution P_{X^n} converges to P.

Theorem 11.2.1 *Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim P(x)$. Then*

$$\Pr\{D(P_{x^n}||P) > \epsilon\} \le 2^{-n(\epsilon - |\mathcal{X}|\frac{\log(n+1)}{n})}, \tag{11.69}$$

and consequently, $D(P_{x^n}||P) \to 0$ with probability 1.

Proof: The inequality (11.69) was proved in (11.68). Summing over n, we find that

$$\sum_{n=1}^{\infty} \Pr\{D(P_{x^n}||P) > \epsilon\} < \infty. \tag{11.70}$$

Thus, the expected number of occurrences of the event $\{D(P_{x^n}||P) > \epsilon\}$ for all n is finite, which implies that the actual number of such occurrences is also finite with probability 1 (Borel–Cantelli lemma). Hence $D(P_{x^n}||P) \to 0$ with probability 1. □

We now define a stronger version of typicality than in Chapter 3.

Definition We define the *strongly typical set* $A_\epsilon^{*(n)}$ to be the set of sequences in $\mathcal{X}^n$ for which the sample frequencies are close to the true values:

$$A_\epsilon^{*(n)} = \left\{ \mathbf{x} \in \mathcal{X}^n : \begin{array}{ll} \left|\frac{1}{n}N(a|\mathbf{x}) - P(a)\right| < \frac{\epsilon}{|\mathcal{X}|}, & \text{if } P(a) > 0 \\ N(a|\mathbf{x}) = 0 & \text{if } P(a) = 0 \end{array} \right\}. \tag{11.71}$$

Hence, the typical set consists of sequences whose type does not differ from the true probabilities by more than $\epsilon/|\mathcal{X}|$ in any component. By the strong law of large numbers, it follows that the probability of the strongly typical set goes to 1 as $n \to \infty$. The additional power afforded by strong typicality is useful in proving stronger results, particularly in universal coding, rate distortion theory, and large deviation theory.

11.3 UNIVERSAL SOURCE CODING

Huffman coding compresses an i.i.d. source with a known distribution $p(x)$ to its entropy limit $H(X)$. However, if the code is designed for some incorrect distribution $q(x)$, a penalty of $D(p||q)$ is incurred. Thus, Huffman coding is sensitive to the assumed distribution.

What compression can be achieved if the true distribution $p(x)$ is unknown? Is there a universal code of rate R, say, that suffices to describe every i.i.d. source with entropy $H(X) < R$? The surprising answer is yes. The idea is based on the method of types. There are $2^{nH(P)}$ sequences of type P. Since there are only a polynomial number of types with denominator n, an enumeration of all sequences x^n with type P_{x^n} such that $H(P_{x^n}) < R$ will require roughly nR bits. Thus, by describing all such sequences, we are prepared to describe any sequence that is likely to arise from any distribution Q having entropy $H(Q) < R$. We begin with a definition.

Definition A *fixed-rate block code* of rate R for a source $X_1, X_2, \ldots, X_n$ which has an unknown distribution Q consists of two mappings: the encoder,

$$f_n : \mathcal{X}^n \to \{1, 2, \ldots, 2^{nR}\}, \tag{11.72}$$

and the decoder,

$$\phi_n : \{1, 2, \ldots, 2^{nR}\} \to \mathcal{X}^n. \tag{11.73}$$

Here R is called the *rate* of the code. The probability of error for the code with respect to the distribution Q is

$$P_e^{(n)} = Q^n(X^n : \phi_n(f_n(X^n)) \neq X^n) \tag{11.74}$$

Definition A rate R block code for a source will be called *universal* if the functions f_n and ϕ_n do not depend on the distribution Q and if $P_e^{(n)} \to 0$ as $n \to \infty$ if $R > H(Q)$.

We now describe one such universal encoding scheme, due to Csiszár and Körner [149], that is based on the fact that the number of sequences of type P increases exponentially with the entropy and the fact that there are only a polynomial number of types.

Theorem 11.3.1 *There exists a sequence of* $(2^{nR}, n)$ *universal source codes such that* $P_e^{(n)} \to 0$ *for every source* Q *such that* $H(Q) < R$.

Proof: Fix the rate R for the code. Let

$$R_n = R - |\mathcal{X}|\frac{\log(n+1)}{n}. \tag{11.75}$$

Consider the set of sequences

$$A = \{\mathbf{x} \in \mathcal{X}^n : H(P_{\mathbf{x}}) \leq R_n\}. \tag{11.76}$$

Then

$$\begin{aligned}
|A| &= \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} |T(P)| && (11.77)\\
&\leq \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} 2^{nH(P)} && (11.78)\\
&\leq \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} 2^{nR_n} && (11.79)\\
&\leq (n+1)^{|\mathcal{X}|} 2^{nR_n} && (11.80)\\
&= 2^{n(R_n + |\mathcal{X}|\frac{\log(n+1)}{n})} && (11.81)\\
&= 2^{nR}. && (11.82)
\end{aligned}$$

By indexing the elements of A, we define the encoding function f_n as

$$f_n(\mathbf{x}) = \begin{cases} \text{index of } \mathbf{x} \text{ in } A & \text{if } x \in A, \\ 0 & \text{otherwise.} \end{cases} \tag{11.83}$$

The decoding function maps each index onto the corresponding element of A. Hence all the elements of A are recovered correctly, and all the remaining sequences result in an error. The set of sequences that are recovered correctly is illustrated in Figure 11.2.

We now show that this encoding scheme is universal. Assume that the distribution of $X_1, X_2, \ldots, X_n$ is Q and $H(Q) < R$. Then the probability of decoding error is given by

$$P_e^{(n)} = 1 - Q^n(A) \tag{11.84}$$

$$= \sum_{P:H(P)>R_n} Q^n(T(P)) \tag{11.85}$$

$$\le (n+1)^{|\mathcal{X}|} \max_{P:H(P)>R_n} Q^n(T(P)) \tag{11.86}$$

$$\le (n+1)^{|\mathcal{X}|} 2^{-n \min_{P:H(P)>R_n} D(P||Q)}. \tag{11.87}$$

Since $R_n \uparrow R$ and $H(Q) < R$, there exists n_0 such that for all $n \ge n_0$, $R_n > H(Q)$. Then for $n \ge n_0$, $\min_{P:H(P)>R_n} D(P||Q)$ must be greater than 0, and the probability of error $P_e^{(n)}$ converges to 0 exponentially fast as $n \to \infty$.

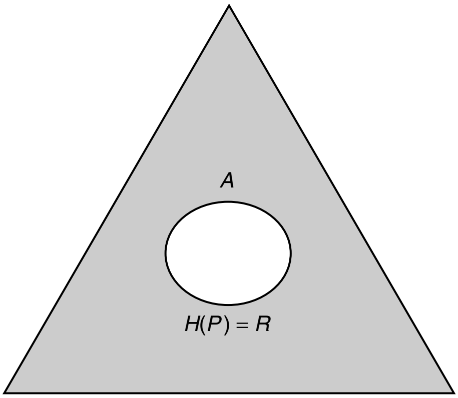


FIGURE 11.2. Universal code and the probability simplex. Each sequence with type that lies outside the circle is encoded by its index. There are fewer than 2^{nR} such sequences. Sequences with types within the circle are encoded by 0.

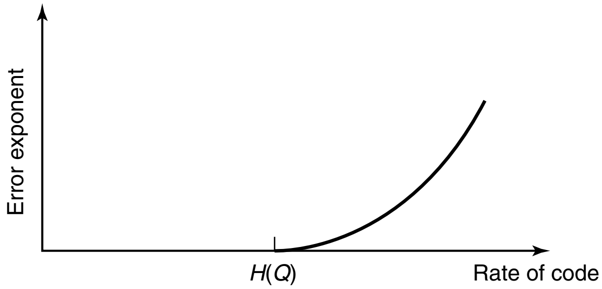


FIGURE 11.3. Error exponent for the universal code.

On the other hand, if the distribution Q is such that the entropy $H(Q)$ is greater than the rate R, then with high probability the sequence will have a type outside the set A. Hence, in such cases the probability of error is close to 1.

The exponent in the probability of error is

$$D^*_{R,Q} = \min_{P:H(P)>R} D(P||Q), \tag{11.88}$$

which is illustrated in Figure 11.3. □

The universal coding scheme described here is only one of many such schemes. It is universal over the set of i.i.d. distributions. There are other schemes, such as the Lempel–Ziv algorithm, which is a variable-rate universal code for all ergodic sources. The Lempel–Ziv algorithm, discussed in Section 13.4, is often used in practice to compress data that cannot be modeled simply, such as English text or computer source code.

One may wonder why it is ever necessary to use Huffman codes, which are specific to a probability distribution. What do we lose in using a universal code? Universal codes need a longer block length to obtain the same performance as a code designed specifically for the probability distribution. We pay the penalty for this increase in block length by the increased complexity of the encoder and decoder. Hence, a distribution specific code is best if one knows the distribution of the source.

11.4 LARGE DEVIATION THEORY

The subject of large deviation theory can be illustrated by an example. What is the probability that $\frac{1}{n}\sum X_i$ is near $\frac{1}{3}$ if $X_1, X_2, \ldots, X_n$ are drawn i.i.d. Bernoulli($\frac{1}{3}$)? This is a small deviation (from the expected outcome) and the probability is near 1. Now what is the probability that $\frac{1}{n}\sum X_i$ is greater than $\frac{3}{4}$ given that $X_1, X_2, \ldots, X_n$ are Bernoulli($\frac{1}{3}$)? This is a large deviation, and the probability is exponentially small. We might estimate the exponent using the central limit theorem, but this is a poor approximation for more than a few standard deviations. We note that $\frac{1}{n}\sum X_i = \frac{3}{4}$ is equivalent to $P_{\mathbf{x}} = (\frac{1}{4}, \frac{3}{4})$. Thus, the probability that $\overline{X}_n$ is near $\frac{3}{4}$ is the probability that type P_X is near $(\frac{3}{4}, \frac{1}{4})$. The probability of this large deviation will turn out to be $\approx 2^{-nD((\frac{3}{4},\frac{1}{4})||(\frac{1}{3},\frac{2}{3}))}$. In this section we estimate the probability of a set of nontypical types.

Let E be a subset of the set of probability mass functions. For example, E may be the set of probability mass functions with mean μ. With a slight abuse of notation, we write

$$Q^n(E) = Q^n(E \cap \mathcal{P}_n) = \sum_{\mathbf{x}: P_{\mathbf{x}} \in E \cap \mathcal{P}_n} Q^n(\mathbf{x}). \tag{11.89}$$

If E contains a relative entropy neighborhood of Q, then by the weak law of large numbers (Theorem 11.2.1), $Q^n(E) \to 1$. On the other hand, if E does not contain Q or a neighborhood of Q, then by the weak law of large numbers, $Q^n(E) \to 0$ exponentially fast. We will use the method of types to calculate the exponent.

Let us first give some examples of the kinds of sets E that we are considering. For example, assume that by observation we find that the sample average of $g(X)$ is greater than or equal to α [i.e., $\frac{1}{n}\sum_i g(x_i) \ge \alpha$]. This event is equivalent to the event $P_{\mathbf{X}} \in E \cap \mathcal{P}_n$, where

$$E = \left\{ P : \sum_{a \in \mathcal{X}} g(a)P(a) \ge \alpha \right\}, \tag{11.90}$$

because

$$\frac{1}{n}\sum_{i=1}^{n} g(x_i) \ge \alpha \Leftrightarrow \sum_{a \in \mathcal{X}} P_{\mathbf{X}}(a)g(a) \ge \alpha \tag{11.91}$$

$$\Leftrightarrow P_{\mathbf{X}} \in E \cap \mathcal{P}_n. \tag{11.92}$$

Thus,

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} g(X_i) \ge \alpha \right) = Q^n(E \cap \mathcal{P}_n) = Q^n(E). \tag{11.93}$$

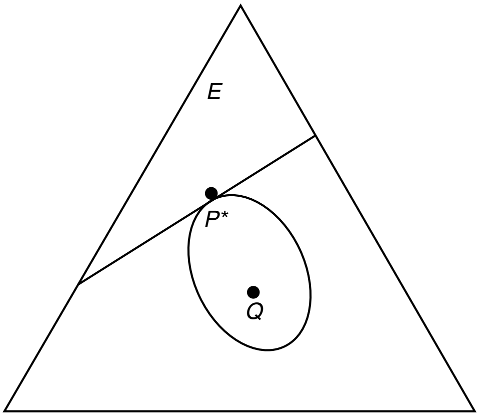


FIGURE 11.4. Probability simplex and Sanov's theorem.

Here E is a half space in the space of probability vectors, as illustrated in Figure 11.4.

Theorem 11.4.1 (*Sanov's theorem*) *Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim Q(x)$. Let $E \subseteq \mathcal{P}$ be a set of probability distributions. Then*

$$Q^n(E) = Q^n(E \cap \mathcal{P}_n) \le (n+1)^{|\mathcal{X}|} 2^{-nD(P^*||Q)}, \tag{11.94}$$

where

$$P^* = \arg\min_{P \in E} D(P||Q) \tag{11.95}$$

is the distribution in E that is closest to Q in relative entropy.

If, in addition, the set E is the closure of its interior, then

$$\frac{1}{n} \log Q^n(E) \to -D(P^*||Q). \tag{11.96}$$

Proof: We first prove the upper bound:

$$Q^n(E) = \sum_{P \in E \cap \mathcal{P}_n} Q^n(T(P)) \tag{11.97}$$

$$\le \sum_{P \in E \cap \mathcal{P}_n} 2^{-nD(P||Q)} \tag{11.98}$$

$$\leq \sum_{P \in E \cap \mathcal{P}_n} \max_{P \in E \cap \mathcal{P}_n} 2^{-nD(P||Q)} \tag{11.99}$$

$$= \sum_{P \in E \cap \mathcal{P}_n} 2^{-n \min_{P \in E \cap \mathcal{P}_n} D(P||Q)} \tag{11.100}$$

$$\leq \sum_{P \in E \cap \mathcal{P}_n} 2^{-n \min_{P \in E} D(P||Q)} \tag{11.101}$$

$$= \sum_{P \in E \cap \mathcal{P}_n} 2^{-nD(P^*||Q)} \tag{11.102}$$

$$\leq (n+1)^{|\mathcal{X}|} 2^{-nD(P^*||Q)}, \tag{11.103}$$

where the last inequality follows from Theorem 11.1.1. Note that P^* need not be a member of $\mathcal{P}_n$. We now come to the lower bound, for which we need a "nice" set E, so that for all large n, we can find a distribution in $E \cap \mathcal{P}_n$ that is close to P^*. If we now assume that E is the closure of its interior (thus, the interior must be nonempty), then since $\cup_n \mathcal{P}_n$ is dense in the set of all distributions, it follows that $E \cap \mathcal{P}_n$ is nonempty for all $n \geq n_0$ for some n_0. We can then find a sequence of distributions P_n such that $P_n \in E \cap \mathcal{P}_n$ and $D(P_n||Q) \to D(P^*||Q)$. For each $n \geq n_0$,

$$Q^n(E) = \sum_{P \in E \cap \mathcal{P}_n} Q^n(T(P)) \tag{11.104}$$

$$\geq Q^n(T(P_n)) \tag{11.105}$$

$$\geq \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(P_n||Q)}. \tag{11.106}$$

Consequently,

$$\liminf \frac{1}{n} \log Q^n(E) \geq \liminf \left(-\frac{|\mathcal{X}| \log(n+1)}{n} - D(P_n||Q) \right)$$
$$= -D(P^*||Q). \tag{11.107}$$

Combining this with the upper bound establishes the theorem. □

This argument can be extended to continuous distributions using quantization.

11.5 EXAMPLES OF SANOV'S THEOREM

Suppose that we wish to find $\Pr\{\frac{1}{n}\sum_{i=1}^{n} g_j(X_i) \geq \alpha_j,\, j = 1, 2, \ldots, k\}$. Then the set E is defined as

$$E = \left\{ P : \sum_a P(a) g_j(a) \geq \alpha_j,\, j = 1, 2, \ldots, k \right\}. \tag{11.108}$$

To find the closest distribution in E to Q, we minimize $D(P||Q)$ subject to the constraints in (11.108). Using Lagrange multipliers, we construct the functional

$$J(P) = \sum_x P(x) \log \frac{P(x)}{Q(x)} + \sum_i \lambda_i \sum_x P(x) g_i(x) + \nu \sum_x P(x). \tag{11.109}$$

We then differentiate and calculate the closest distribution to Q to be of the form

$$P^*(x) = \frac{Q(x) e^{\sum_i \lambda_i g_i(x)}}{\sum_{a \in \mathcal{X}} Q(a) e^{\sum_i \lambda_i g_i(a)}}, \tag{11.110}$$

where the constants λ_i are chosen to satisfy the constraints. Note that if Q is uniform, P^* is the maximum entropy distribution. Verification that P^* is indeed the minimum follows from the same kinds of arguments as given in Chapter 12.

Let us consider some specific examples:

Example 11.5.1 (*Dice*) Suppose that we toss a fair die n times; what is the probability that the average of the throws is greater than or equal to 4? From Sanov's theorem, it follows that

$$Q^n(E) \doteq 2^{-nD(P^*||Q)}, \tag{11.111}$$

where P^* minimizes $D(P||Q)$ over all distributions P that satisfy

$$\sum_{i=1}^{6} i P(i) \geq 4. \tag{11.112}$$

From (11.110), it follows that P^* has the form

$$P^*(x) = \frac{2^{\lambda x}}{\sum_{i=1}^{6} 2^{\lambda i}}, \tag{11.113}$$

with λ chosen so that $\sum i P^*(i) = 4$. Solving numerically, we obtain $\lambda = 0.2519$, $P^* = (0.1031, 0.1227, 0.1461, 0.1740, 0.2072, 0.2468)$, and therefore $D(P^*||Q) = 0.0624$ bit. Thus, the probability that the average of 10000 throws is greater than or equal to 4 is $\approx 2^{-624}$.

Example 11.5.2 (Coins) Suppose that we have a fair coin and want to estimate the probability of observing more than 700 heads in a series of 1000 tosses. The problem is like Example 11.5.1. The probability is

$$P(\overline{X}_n \geq 0.7) \doteq 2^{-nD(P^*||Q)}, \tag{11.114}$$

where P^* is the $(0.7, 0.3)$ distribution and Q is the $(0.5, 0.5)$ distribution. In this case, $D(P^*||Q) = 1 - H(P^*) = 1 - H(0.7) = 0.119$. Thus, the probability of 700 or more heads in 1000 trials is approximately 2^{-119}.

Example 11.5.3 (*Mutual dependence*) Let $Q(x, y)$ be a given joint distribution and let $Q_0(x, y) = Q(x)Q(y)$ be the associated product distribution formed from the marginals of Q. We wish to know the likelihood that a sample drawn according to Q_0 will "appear" to be jointly distributed according to Q. Accordingly, let (X_i, Y_i) be i.i.d. $\sim Q_0(x, y) = Q(x)Q(y)$. We define joint typicality as we did in Section 7.6; that is, (x^n, y^n) is jointly typical with respect to a joint distribution $Q(x, y)$ iff the sample entropies are close to their true values:

$$\left| -\frac{1}{n} \log Q(x^n) - H(X) \right| \leq \epsilon, \tag{11.115}$$

$$\left| -\frac{1}{n} \log Q(y^n) - H(Y) \right| \leq \epsilon, \tag{11.116}$$

and

$$\left| -\frac{1}{n} \log Q(x^n, y^n) - H(X, Y) \right| \leq \epsilon. \tag{11.117}$$

We wish to calculate the probability (under the product distribution) of seeing a pair (x^n, y^n) that looks jointly typical of Q [i.e., (x^n, y^n)

satisfies (11.115)–(11.117)]. Thus, (x^n, y^n) are jointly typical with respect to $Q(x, y)$ if $P_{x^n,y^n} \in E \cap \mathcal{P}_n(X, Y)$, where

$$\begin{aligned} E = \{P(x, y) : &\left| -\sum_{x,y} P(x, y) \log Q(x) - H(X) \right| \le \epsilon, \\ &\left| -\sum_{x,y} P(x, y) \log Q(y) - H(Y) \right| \le \epsilon, \\ &\left| -\sum_{x,y} P(x, y) \log Q(x, y) - H(X, Y) \right| \le \epsilon\}. \end{aligned} \quad (11.118)$$

Using Sanov's theorem, the probability is

$$Q_0^n(E) \doteq 2^{-nD(P^*||Q_0)}, \quad (11.119)$$

where P^* is the distribution satisfying the constraints that is closest to Q_0 in relative entropy. In this case, as $\epsilon \to 0$, it can be verified (Problem 11.10) that P^* is the joint distribution Q, and Q_0 is the product distribution, so that the probability is $2^{-nD(Q(x,y)||Q(x)Q(y))} = 2^{-nI(X;Y)}$, which is the same as the result derived in Chapter 7 for the joint AEP.

In the next section we consider the empirical distribution of the sequence of outcomes given that the type is in a particular set of distributions E. We will show that not only is the probability of the set E essentially determined by $D(P^*||Q)$, the distance of the closest element of E to Q, but also that the conditional type is essentially P^*, so that given that we are in set E, the type is very likely to be close to P^*.

11.6 CONDITIONAL LIMIT THEOREM

It has been shown that the probability of a set of types under a distribution Q is determined essentially by the probability of the closest element of the set to Q; the probability is 2^{-nD^*} to first order in the exponent, where

$$D^* = \min_{P \in E} D(P||Q). \quad (11.120)$$

This follows because the probability of the set of types is the sum of the probabilities of each type, which is bounded by the largest term times the

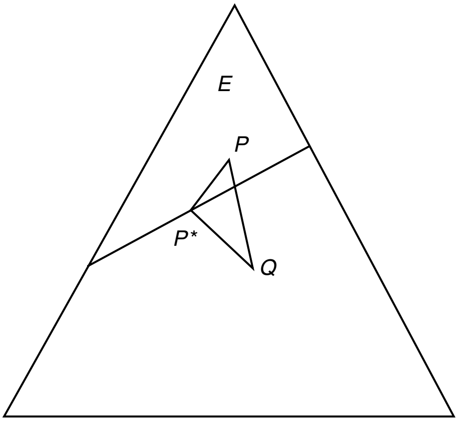


FIGURE 11.5. Pythagorean theorem for relative entropy.

number of terms. Since the number of terms is polynomial in the length of the sequences, the sum is equal to the largest term to first order in the exponent.

We now strengthen the argument to show that not only is the probability of the set E essentially the same as the probability of the closest type P^* but also that the total probability of other types that are far away from P^* is negligible. This implies that with very high probability, the type observed is close to P^*. We call this a *conditional limit theorem.*

Before we prove this result, we prove a "Pythagorean" theorem, which gives some insight into the geometry of $D(P||Q)$. Since $D(P||Q)$ is not a metric, many of the intuitive properties of distance are not valid for $D(P||Q)$. The next theorem shows a sense in which $D(P||Q)$ behaves like the square of the Euclidean metric (Figure 11.5).

Theorem 11.6.1 *For a closed convex set $E \subset \mathcal{P}$ and distribution $Q \notin E$, let $P^* \in E$ be the distribution that achieves the minimum distance to Q; that is,*

$$D(P^*||Q) = \min_{P \in E} D(P||Q). \tag{11.121}$$

Then

$$D(P||Q) \geq D(P||P^*) + D(P^*||Q) \tag{11.122}$$

for all $P \in E$.

Note. The main use of this theorem is as follows: Suppose that we have a sequence $P_n \in E$ that yields $D(P_n||Q) \to D(P^*||Q)$. Then from the Pythagorean theorem, $D(P_n||P^*) \to 0$ as well.

Proof: Consider any $P \in E$. Let

$$P_\lambda = \lambda P + (1-\lambda)P^*. \tag{11.123}$$

Then $P_\lambda \to P^*$ as $\lambda \to 0$. Also, since E is convex, $P_\lambda \in E$ for $0 \le \lambda \le 1$. Since $D(P^*||Q)$ is the minimum of $D(P_\lambda||Q)$ along the path $P^* \to P$, the derivative of $D(P_\lambda||Q)$ as a function of λ is nonnegative at $\lambda = 0$. Now

$$D_\lambda = D(P_\lambda||Q) = \sum P_\lambda(x) \log \frac{P_\lambda(x)}{Q(x)} \tag{11.124}$$

and

$$\frac{dD_\lambda}{d\lambda} = \sum \left((P(x) - P^*(x)) \log \frac{P_\lambda(x)}{Q(x)} + (P(x) - P^*(x)) \right). \tag{11.125}$$

Setting $\lambda = 0$, so that $P_\lambda = P^*$ and using the fact that $\sum P(x) = \sum P^*(x) = 1$, we have

$$0 \le \left(\frac{dD_\lambda}{d\lambda} \right)_{\lambda=0} \tag{11.126}$$

$$= \sum (P(x) - P^*(x)) \log \frac{P^*(x)}{Q(x)} \tag{11.127}$$

$$= \sum P(x) \log \frac{P^*(x)}{Q(x)} - \sum P^*(x) \log \frac{P^*(x)}{Q(x)} \tag{11.128}$$

$$= \sum P(x) \log \frac{P(x)}{Q(x)} \frac{P^*(x)}{P(x)} - \sum P^*(x) \log \frac{P^*(x)}{Q(x)} \tag{11.129}$$

$$= D(P||Q) - D(P||P^*) - D(P^*||Q), \tag{11.130}$$

which proves the theorem. □

Note that the relative entropy $D(P||Q)$ behaves like the square of the Euclidean distance. Suppose that we have a convex set E in $\mathcal{R}^n$. Let A be a point outside the set, B the point in the set closest to A, and C any

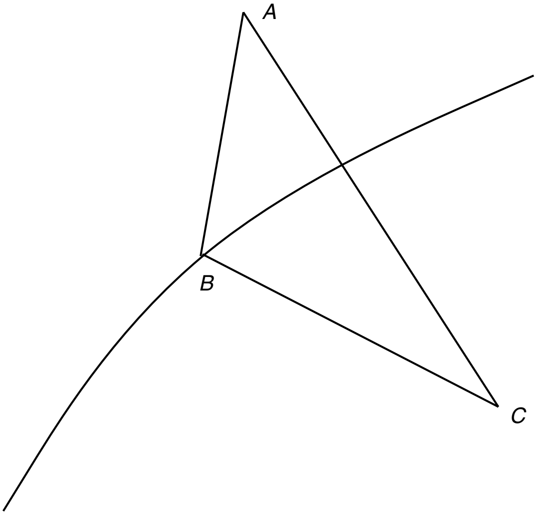


FIGURE 11.6. Triangle inequality for distance squared.

other point in the set. Then the angle between the lines BA and BC must be obtuse, which implies that $l^2_{AC} \geq l^2_{AB} + l^2_{BC}$, which is of the same form as Theorem 11.6.1. This is illustrated in Figure 11.6.

We now prove a useful lemma which shows that convergence in relative entropy implies convergence in the $\mathcal{L}_1$ norm.

Definition The $\mathcal{L}_1$ distance between any two distributions is defined as

$$||P_1 - P_2||_1 = \sum_{a \in \mathcal{X}} |P_1(a) - P_2(a)|. \tag{11.131}$$

Let A be the set on which $P_1(x) > P_2(x)$. Then

$$\begin{aligned} ||P_1 - P_2||_1 &= \sum_{x \in \mathcal{X}} |P_1(x) - P_2(x)| && (11.132) \\ &= \sum_{x \in A} (P_1(x) - P_2(x)) + \sum_{x \in A^c} (P_2(x) - P_1(x)) && (11.133) \\ &= P_1(A) - P_2(A) + P_2(A^c) - P_1(A^c) && (11.134) \\ &= P_1(A) - P_2(A) + 1 - P_2(A) - 1 + P_1(A) && (11.135) \\ &= 2(P_1(A) - P_2(A)). && (11.136) \end{aligned}$$

Also note that

$$\max_{B \subseteq \mathcal{X}} (P_1(B) - P_2(B)) = P_1(A) - P_2(A) = \frac{||P_1 - P_2||_1}{2}. \tag{11.137}$$

The left-hand side of (11.137) is called the *variational distance* between P_1 and P_2.

Lemma 11.6.1

$$D(P_1||P_2) \geq \frac{1}{2\ln 2}||P_1 - P_2||_1^2. \tag{11.138}$$

Proof: We first prove it for the binary case. Consider two binary distributions with parameters p and q with $p \geq q$. We will show that

$$p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} \geq \frac{4}{2\ln 2}(p-q)^2. \tag{11.139}$$

The difference $g(p,q)$ between the two sides is

$$g(p,q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} - \frac{4}{2\ln 2}(p-q)^2. \tag{11.140}$$

Then

$$\begin{aligned} \frac{dg(p,q)}{dq} &= -\frac{p}{q\ln 2} + \frac{1-p}{(1-q)\ln 2} - \frac{4}{2\ln 2}2(q-p) & (11.141) \\ &= \frac{q-p}{q(1-q)\ln 2} - \frac{4}{\ln 2}(q-p) & (11.142) \\ &\leq 0 & (11.143) \end{aligned}$$

since $q(1-q) \leq \frac{1}{4}$ and $q \leq p$. For $q = p$, $g(p,q) = 0$, and hence $g(p,q) \geq 0$ for $q \leq p$, which proves the lemma for the binary case.

For the general case, for any two distributions P_1 and P_2, let

$$A = \{x : P_1(x) > P_2(x)\}. \tag{11.144}$$

Define a new binary random variable $Y = \phi(X)$, the indicator of the set A, and let $\hat{P}_1$ and $\hat{P}_2$ be the distributions of Y. Thus, $\hat{P}_1$ and $\hat{P}_2$ correspond to the quantized versions of P_1 and P_2. Then by the data-processing inequality applied to relative entropies (which is proved in the same way as the data-processing inequality for mutual information), we have

$$\begin{aligned} D(P_1||P_2) &\geq D(\hat{P}_1||\hat{P}_2) && (11.145)\\ &\geq \frac{4}{2\ln 2}(P_1(A) - P_2(A))^2 && (11.146)\\ &= \frac{1}{2\ln 2}||P_1 - P_2||_1^2, && (11.147) \end{aligned}$$

by (11.137), and the lemma is proved. □

We can now begin the proof of the conditional limit theorem. We first outline the method used. As stated at the beginning of the chapter, the essential idea is that the probability of a type under Q depends exponentially on the distance of the type from Q, and hence types that are farther away are exponentially less likely to occur. We divide the set of types in E into two categories: those at about the same distance from Q as P^* and those a distance 2δ farther away. The second set has exponentially less probability than the first, and hence the first set has a conditional probability tending to 1. We then use the Pythagorean theorem to establish that all the elements in the first set are close to P^*, which will establish the theorem.

The following theorem is an important strengthening of the maximum entropy principle.

Theorem 11.6.2 (*Conditional limit theorem*) *Let E be a closed convex subset of $\mathcal{P}$ and let Q be a distribution not in E. Let $X_1, X_2, \ldots, X_n$ be discrete random variables drawn i.i.d. $\sim Q$. Let P^* achieve* $\min_{P\in E}$ *$D(P||Q)$. Then*

$$Pr(X_1 = a|P_{X^n} \in E) \to P^*(a) \tag{11.148}$$

in probability as $n \to \infty$, i.e., the conditional distribution of X_1, given that the type of the sequence is in E, is close to P^* for large n.

Example 11.6.1 If X_i i.i.d. $\sim Q$, then

$$\Pr\left\{X_1 = a \left| \frac{1}{n}\sum X_i^2 \geq \alpha\right.\right\} \to P^*(a), \tag{11.149}$$

where $P^*(a)$ minimizes $D(P||Q)$ over P satisfying $\sum P(a)a^2 \geq \alpha$. This minimization results in

$$P^*(a) = Q(a)\frac{e^{\lambda a^2}}{\sum_a Q(a)e^{\lambda a^2}}, \tag{11.150}$$

where λ is chosen to satisfy $\sum P^*(a)a^2 = \alpha$. Thus, the conditional distribution on X_1 given a constraint on the sum of the squares is a (normalized) product of the original probability mass function and the maximum entropy probability mass function (which in this case is Gaussian).

Proof of Theorem: Define the sets

$$S_t = \{P \in \mathcal{P} : D(P||Q) \leq t\}. \tag{11.151}$$

The set S_t is convex since $D(P||Q)$ is a convex function of P. Let

$$D^* = D(P^*||Q) = \min_{P \in E} D(P||Q). \tag{11.152}$$

Then P^* is unique, since $D(P||Q)$ is strictly convex in P. Now define the set

$$A = S_{D^*+2\delta} \cap E \tag{11.153}$$

and

$$B = E - S_{D^*+2\delta} \cap E. \tag{11.154}$$

Thus, $A \cup B = E$. These sets are illustrated in Figure 11.7. Then

$$Q^n(B) = \sum_{P \in E \cap \mathcal{P}_n : D(P||Q) > D^*+2\delta} Q^n(T(P)) \tag{11.155}$$

$$\leq \sum_{P \in E \cap \mathcal{P}_n : D(P||Q) > D^*+2\delta} 2^{-nD(P||Q)} \tag{11.156}$$

$$\leq \sum_{P \in E \cap \mathcal{P}_n : D(P||Q) > D^*+2\delta} 2^{-n(D^*+2\delta)} \tag{11.157}$$

$$\leq (n+1)^{|\mathcal{X}|} 2^{-n(D^*+2\delta)} \tag{11.158}$$

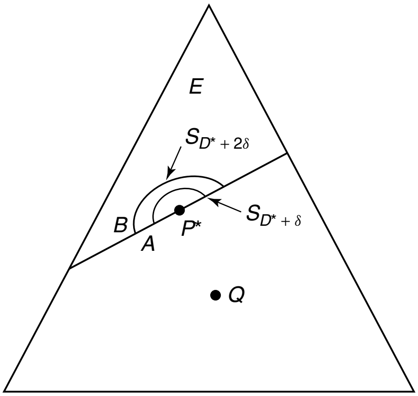


FIGURE 11.7. Conditional limit theorem.

since there are only a polynomial number of types. On the other hand,

$$\begin{aligned} Q^n(A) &\geq Q^n(S_{D^*+\delta} \cap E) && (11.159)\\ &= \sum_{P \in E \cap \mathcal{P}_n : D(P\|Q) \leq D^*+\delta} Q^n(T(P)) && (11.160)\\ &\geq \sum_{P \in E \cap \mathcal{P}_n : D(P\|Q) \leq D^*+\delta} \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(P\|Q)} && (11.161)\\ &\geq \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-n(D^*+\delta)} \quad \text{for } n \text{ sufficiently large,} && (11.162) \end{aligned}$$

since the sum is greater than one of the terms, and for sufficiently large n, there exists at least one type in $S_{D^*+\delta} \cap E \cap \mathcal{P}_n$. Then, for n sufficiently large,

$$\begin{aligned} \Pr(P_{X^n} \in B | P_{X^n} \in E) &= \frac{Q^n(B \cap E)}{Q^n(E)} && (11.163)\\ &\leq \frac{Q^n(B)}{Q^n(A)} && (11.164)\\ &\leq \frac{(n+1)^{|\mathcal{X}|} 2^{-n(D^*+2\delta)}}{\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-n(D^*+\delta)}} && (11.165)\\ &= (n+1)^{2|\mathcal{X}|} 2^{-n\delta}, && (11.166) \end{aligned}$$

which goes to 0 as $n \to \infty$. Hence the conditional probability of B goes to 0 as $n \to \infty$, which implies that the conditional probability of A goes to 1.

We now show that all the members of A are close to P^* in relative entropy. For all members of A,

$$D(P||Q) \le D^* + 2\delta. \tag{11.167}$$

Hence by the "Pythagorean" theorem (Theorem 11.6.1),

$$D(P||P^*) + D(P^*||Q) \le D(P||Q) \le D^* + 2\delta, \tag{11.168}$$

which in turn implies that

$$D(P||P^*) \le 2\delta, \tag{11.169}$$

since $D(P^*||Q) = D^*$. Thus, $P_{\mathbf{x}} \in A$ implies that $D(P_{\mathbf{x}}||Q) \le D^* + 2\delta$, and therefore that $D(P_{\mathbf{x}}||P^*) \le 2\delta$. Consequently, since $\Pr\{P_{X^n} \in A | P_{X^n} \in E\} \to 1$, it follows that

$$\Pr(D(P_{X^n}||P^*) \le 2\delta | P_{X^n} \in E) \to 1 \tag{11.170}$$

as $n \to \infty$. By Lemma 11.6.1, the fact that the relative entropy is small implies that the $\mathcal{L}_1$ distance is small, which in turn implies that $\max_{a\in\mathcal{X}} |P_{X^n}(a) - P^*(a)|$ is small. Thus, $\Pr(|P_{X^n}(a) - P^*(a)| \ge \epsilon | P_{X^n} \in E) \to 0$ as $n \to \infty$. Alternatively, this can be written as

$$\Pr(X_1 = a | P_{X^n} \in E) \to P^*(a) \qquad \text{in probability}, a \in \mathcal{X}. \tag{11.171}$$

In this theorem we have only proved that the marginal distribution goes to P^* as $n \to \infty$. Using a similar argument, we can prove a stronger version of this theorem:

$$\begin{aligned}\Pr(X_1 = a_1, X_2 = a_2, \ldots, X_m \\ = a_m | P_{X^n} \in E) \to \prod_{i=1}^{m} P^*(a_i) \qquad \text{in probability.}\end{aligned} \tag{11.172}$$

This holds for fixed m as $n \to \infty$. The result is not true for $m = n$, since there are end effects; given that the type of the sequence is in E, the last elements of the sequence can be determined from the remaining elements, and the elements are no longer independent. The conditional limit