Elements of Information Theory
Thomas M. Cover, Joy A. Thomas

Print ISBN 0-471-06259-6 Online ISBN 0-471-20061-1

Chapter 13

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 will begin by considering the simple problem of representing a single continuous random variable by a finite number of bits.

13.1 QUANTIZATION

This section on quantization motivates the elegant theory of rate distortion by showing how complicated it is to solve the quantization problem exactly for a single 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 to 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, then 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 codepoints) 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 1 bit to represent X, it is clear that the bit should distinguish whether $X > 0$ or not. To minimize squared error, each reproduced symbol should be at the conditional mean of its region. This is illustrated in Figure 13.1. Thus

$$\hat{X}(x) = \begin{cases} \sqrt{\frac{2}{\pi}}\sigma, & \text{if } x \geq 0, \\ -\sqrt{\frac{2}{\pi}}\sigma, & \text{if } x < 0. \end{cases} \tag{13.1}$$

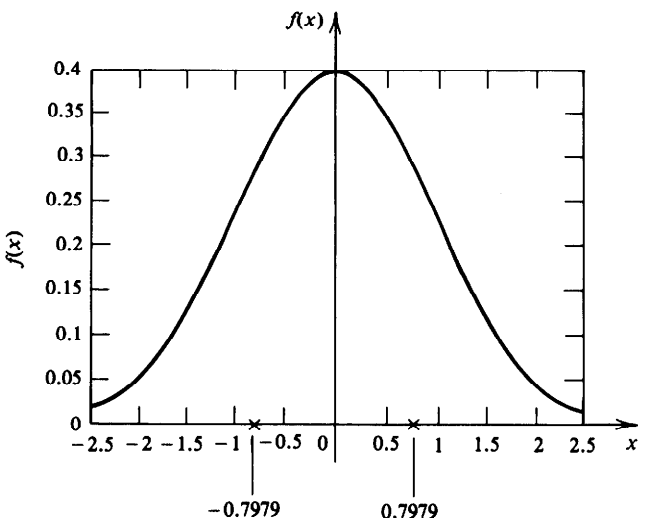

Figure 13.1. One bit quantization of a Gaussian random variable.

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 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 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* [181] (for real-valued random variables) or the *generalized Lloyd algorithm* [80] (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.

13.2 DEFINITIONS

Assume that we have a source that produces a sequence $X_1, X_2, \ldots, X_n$ i.i.d. $\sim p(x), x \in \mathcal{X}$. We will assume that the alphabet is finite for the

proofs in this chapter; 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 $\hat{X}^n \in \hat{\mathcal{X}}^n$, as illustrated in Figure 13.2.

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

$$d: \mathcal{X} \times \hat{\mathcal{X}} \to R^+ \tag{13.2}$$

from the set of source alphabet-reproduction alphabet pairs into the set of non-negative 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, i.e.,

$$d_{\max} \stackrel{\text{def}}{=} \max_{x \in \mathcal{X}, \hat{x} \in \hat{\mathcal{X}}} d(x, \hat{x}) < \infty . \tag{13.3}$$

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

$$d(x, \hat{x}) = \begin{cases} 0 & \text{if } x = \hat{x} \\ 1 & \text{if } x \neq \hat{x} \end{cases}, \tag{13.4}$$

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

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

$$d(x, \hat{x}) = (x - \hat{x})^2 , \tag{13.5}$$

  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

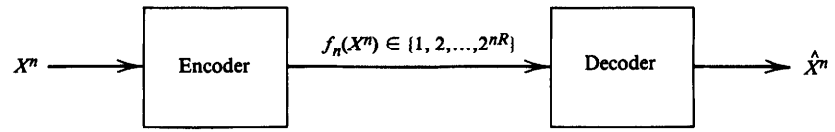

Figure 13.2. Rate distortion encoder and decoder.

speech coding, various authors have pointed out that the mean squared error is not an appropriate measure of distortion as observed by a human observer. 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 very similar 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

$$d(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i) . \tag{13.6}$$

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 distortion between two sequences by the maximum of the per symbol distortions. The theory derived below does not apply directly to this case.

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

$$f_n : \mathcal{X}^n \to \{1, 2, \ldots, 2^{nR}\} , \tag{13.7}$$

and a decoding (reproduction) function,

$$g_n : \{1, 2, \ldots, 2^{nR}\} \to \hat{\mathcal{X}}^n . \tag{13.8}$$

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

$$D = Ed(X^n, g_n(f_n(X^n))) , \tag{13.9}$$

where the expectation is with respect to the probability distribution on X, i.e.,

$$D = \sum_{x^n} p(x^n) d(x^n, g_n(f_n(x^n))) . \tag{13.10}$$

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 $\lim_{n\to\infty} Ed(X^n, g_n(f_n(X^n))) \leq D$.

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, which is the set of achievable rate distortion pairs. We will in general use the rate distortion function rather than the distortion rate function to describe this boundary, though 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

$$R^{(I)}(D) = \min_{p(\hat{x}|x):\, \Sigma_{(x,\hat{x})}\, p(x)p(\hat{x}|x)d(x,\hat{x}) \leq D} I(X; \hat{X}) \tag{13.11}$$

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 8, 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 13.2.1: *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*

$$R(D) = R^{(I)}(D) = \min_{p(\hat{x}|x):\, \Sigma_{(x,\hat{x})}\, p(x)p(\hat{x}|x)d(x,\hat{x}) \leq D} I(X; \hat{X}) \tag{13.12}$$

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.

13.3 CALCULATION OF THE RATE DISTORTION FUNCTION

13.3.1 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 13.3.1: *The rate distortion function for a Bernoulli(p) source with Hamming distortion is given by*

$$R(D) = \begin{cases} H(p) - H(D), & 0 \leq D \leq \min\{p, 1-p\}, \\ 0, & D > \min\{p, 1-p\}. \end{cases} \tag{13.13}$$

Proof: Consider a binary source $X \sim$ 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,

$$R(D) = \min_{p(\hat{x}|x):\, \Sigma_{(x,\hat{x})}\, p(x)p(\hat{x}|x)d(x,\hat{x}) \leq D} I(X; \hat{X}). \tag{13.14}$$

Let $\oplus$ denote modulo 2 addition. Thus $X \oplus \hat{X} = 1$ is equivalent to $X \neq \hat{X}$.

We cannot 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

$$I(X;\hat{X}) = H(X) - H(X|\hat{X}) \tag{13.15}$$

$$= H(p) - H(X \oplus \hat{X}|\hat{X}) \tag{13.16}$$

$$\geq H(p) - H(X \oplus \hat{X}) \tag{13.17}$$

$$\geq H(p) - H(D), \tag{13.18}$$

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

$$R(D) \geq H(p) - H(D). \tag{13.19}$$

We will 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 $0 \leq D \leq p$, we can achieve the value of the rate distortion function in (13.19) by choosing $(X, \hat{X})$ to have the joint distribution given by the binary symmetric channel shown in Figure 13.3.

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

$$r(1-D) + (1-r)D = p, \tag{13.20}$$

or

$$r = \frac{p-D}{1-2D}. \tag{13.21}$$

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

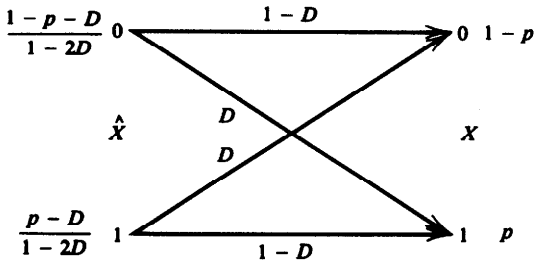

Figure 13.3. Joint distribution for binary source.

$$I(X;\hat{X}) = H(X) - H(X|\hat{X}) = H(p) - H(D), \tag{13.22}$$

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

If $D \geq p$, then 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 $D \geq 1 - p$, we can achieve $R(D) = 0$ by setting $\hat{X} = 1$ with probability 1.

Hence the rate distortion function for a binary source is

$$R(D) = \begin{cases} H(p) - H(D), & 0 \leq D \leq \min\{p, 1-p\}, \\ 0, & D > \min\{p, 1-p\}. \end{cases} \tag{13.23}$$

This function is illustrated in Figure 13.4. □

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 13.2.1.

13.3.2 Gaussian Source

Although Theorem 13.2.1 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 13.3.2: *The rate distortion function for a $\mathcal{N}(0, \sigma^2)$ source with squared error distortion is*

$$R(D) = \begin{cases} \frac{1}{2} \log \frac{\sigma^2}{D}, & 0 \leq D \leq \sigma^2, \\ 0, & D > \sigma^2. \end{cases} \tag{13.24}$$

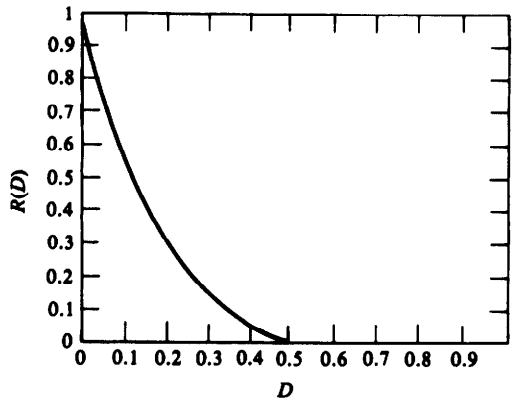

Figure 13.4. Rate distortion function for a binary source.

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

$$R(D) = \min_{f(\hat{x}|x): E(\hat{X}-X)^2 \le D} I(X; \hat{X}). \tag{13.25}$$

As in the previous 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

$$I(X; \hat{X}) = h(X) - h(X|\hat{X}) \tag{13.26}$$

$$= \frac{1}{2}\log(2\pi e)\sigma^2 - h(X - \hat{X}|\hat{X}) \tag{13.27}$$

$$\ge \frac{1}{2}\log(2\pi e)\sigma^2 - h(X - \hat{X}) \tag{13.28}$$

$$\ge \frac{1}{2}\log(2\pi e)\sigma^2 - h(\mathcal{N}(0, E(X - \hat{X})^2)) \tag{13.29}$$

$$= \frac{1}{2}\log(2\pi e)\sigma^2 - \frac{1}{2}\log(2\pi e)E(X - \hat{X})^2 \tag{13.30}$$

$$\ge \frac{1}{2}\log(2\pi e)\sigma^2 - \frac{1}{2}\log(2\pi e)D \tag{13.31}$$

$$= \frac{1}{2}\log\frac{\sigma^2}{D}, \tag{13.32}$$

where (13.28) follows from the fact that conditioning reduces entropy and (13.29) follows from the fact that the normal distribution maximizes the entropy for a given second moment (Theorem 9.6.5). Hence

$$R(D) \ge \frac{1}{2}\log\frac{\sigma^2}{D}. \tag{13.33}$$

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 13.5. If $D \le \sigma^2$, we choose

$$X = \hat{X} + Z, \quad \hat{X} \sim \mathcal{N}(0, \sigma^2 - D), \quad Z \sim \mathcal{N}(0, D), \tag{13.34}$$

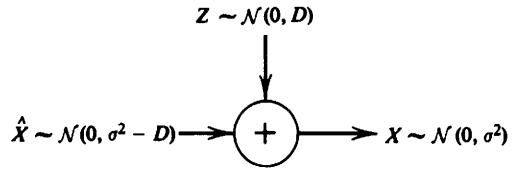

Figure 13.5. Joint distribution for Gaussian source.

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

$$I(X;\hat{X}) = \frac{1}{2}\log\frac{\sigma^2}{D}, \tag{13.35}$$

and $E(X-\hat{X})^2 = D$, thus achieving the bound in (13.33). 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

$$R(D) = \begin{cases} \frac{1}{2}\log\frac{\sigma^2}{D}, & 0 \le D \le \sigma^2, \\ 0, & D > \sigma^2, \end{cases} \tag{13.36}$$

as illustrated in Figure 13.6. □

We can rewrite (13.36) to express the distortion in terms of the rate,

$$D(R) = \sigma^2 2^{-2R}. \tag{13.37}$$

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 13.1. 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 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.

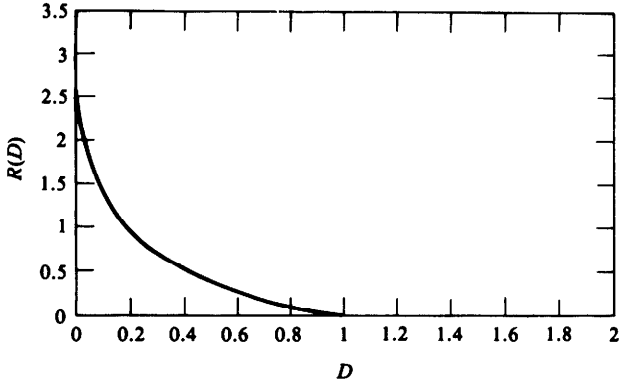

Figure 13.6. Rate distortion function for a Gaussian source.

13.3.3 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 different components to minimize the total distortion. Extending the definition of the information rate distortion function to the vector case, we have

$$R(D) = \min_{f(\hat{x}^m|x^m):\, Ed(X^m, \hat{X}^m) \le D} I(X^m; \hat{X}^m), \tag{13.38}$$

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

$$I(X^m; \hat{X}^m) = h(X^m) - h(X^m|\hat{X}^m) \tag{13.39}$$

$$= \sum_{i=1}^m h(X_i) - \sum_{i=1}^m h(X_i|X^{i-1}, \hat{X}^m) \tag{13.40}$$

$$\ge \sum_{i=1}^m h(X_i) - \sum_{i=1}^m h(X_i|\hat{X}_i) \tag{13.41}$$

$$= \sum_{i=1}^m I(X_i; \hat{X}_i) \tag{13.42}$$

$$\ge \sum_{i=1}^m R(D_i) \tag{13.43}$$

$$= \sum_{i=1}^m \left(\frac{1}{2} \log \frac{\sigma_i^2}{D_i}\right)^+, \tag{13.44}$$

where $D_i = E(X_i - \hat{X}_i)^2$ and (13.41) follows from the fact that conditioning reduces entropy. We can achieve equality in (13.41) by choosing $f(x^m|\hat{x}^m) = \prod_{i=1}^m f(x_i|\hat{x}_i)$ and in (13.43) by choosing the distribution of each $\hat{X}_i \sim \mathcal{N}(0, \sigma_i^2 - D_i)$, as in the previous example. Hence the problem of finding the rate distortion function can be reduced to the following optimization (using nats for convenience):

$$R(D) = \min_{\sum D_i = D} \sum_{i=1}^m \max\left\{\frac{1}{2} \ln \frac{\sigma_i^2}{D_i}, 0\right\}. \tag{13.45}$$

Using Lagrange multipliers, we construct the functional

$$J(D) = \sum_{i=1}^m \frac{1}{2} \ln \frac{\sigma_i^2}{D_i} + \lambda \sum_{i=1}^m D_i, \tag{13.46}$$

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

$$\frac{\partial J}{\partial D_i} = -\frac{1}{2}\frac{1}{D_i} + \lambda = 0, \tag{13.47}$$

or

$$D_i = \lambda'. \tag{13.48}$$

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 (13.48) 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 (13.48) 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 (13.46). In this case the Kuhn-Tucker conditions yield

$$\frac{\partial J}{\partial D_i} = -\frac{1}{2}\frac{1}{D_i} + \lambda, \tag{13.49}$$

where λ is chosen so that

$$\frac{\partial J}{\partial D_i}\begin{cases} = 0, & \text{if } D_i < \sigma_i^2, \\ \leq 0, & \text{if } D_i \geq \sigma_i^2. \end{cases} \tag{13.50}$$

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

Theorem 13.3.3 (*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*

$$R(D) = \sum_{i=1}^{m} \frac{1}{2} \log \frac{\sigma_i^2}{D_i}, \tag{13.51}$$

where

$$D_i = \begin{cases} \lambda, & \text{if } \lambda < \sigma_i^2, \\ \sigma_i^2, & \text{if } \lambda \geq \sigma_i^2, \end{cases} \tag{13.52}$$

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

This gives rise to a kind of reverse "water-filling" as illustrated in Figure 13.7. We choose a constant λ and only describe those random

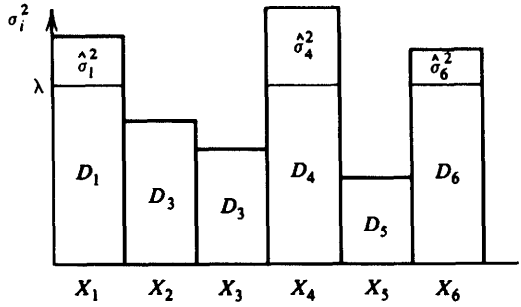

Figure 13.7. Reverse water-filling for independent Gaussian random variables.

variables with variances greater than λ. No bits are used to describe random variables with variance less than λ.

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 different frequency bands. Reverse water-filling on the spectrum yields the rate distortion function.

13.4 CONVERSE TO THE RATE DISTORTION THEOREM

In this section, we prove the converse to Theorem 13.2.1 by showing that we cannot achieve a distortion less than D if we describe X at a rate less than $R(D)$, where

$$R(D) = \min_{p(\hat{x}|x):\ \sum_{(x,\hat{x})} p(x)p(\hat{x}|x)d(x,\hat{x}) \le D} I(X;\hat{X}). \tag{13.53}$$

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 13.4.1 (*Convexity of $R(D)$*): *The rate distortion function $R(D)$ given in (13.53) is a non-increasing 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 non-increasing 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

$$I_{p_\lambda}(X; \hat{X}) \leq \lambda I_{p_1}(X; \hat{X}) + (1 - \lambda) I_{p_2}(X; \hat{X}) . \tag{13.54}$$

Hence by the definition of the rate distortion function,

$$R(D_\lambda) \leq I_{p_\lambda}(X; \hat{X}) \tag{13.55}$$

$$\leq \lambda I_{p_1}(X; \hat{X}) + (1 - \lambda) I_{p_2}(X; \hat{X}) \tag{13.56}$$

$$= \lambda R(D_1) + (1 - \lambda) R(D_2) , \tag{13.57}$$

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

The converse can now be proved.

Proof: (*Converse in Theorem 13.2.1*): 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 $\leq D$, that the rate R of the code satisfies $R \geq R(D)$.

Consider any $(2^{nR}, n)$ rate distortion code defined by functions f_n and g_n. Let $\hat{X}^n = \hat{X}^n(X^n) = g_n(f_n(X^n))$ be the reproduced sequence corresponding to X^n. Then we have the following chain of inequalities:

$$nR \overset{(a)}{\geq} H(\hat{X}^n) \tag{13.58}$$

$$\overset{(b)}{\geq} H(\hat{X}^n) - H(\hat{X}^n|X^n) \tag{13.59}$$

$$\overset{(c)}{=} I(\hat{X}^n; X^n) \tag{13.60}$$

$$= H(X^n) - H(X^n|\hat{X}^n) \tag{13.61}$$

$$\overset{(d)}{=} \sum_{i=1}^{n} H(X_i) - H(X^n|\hat{X}^n) \tag{13.62}$$

$$\overset{(e)}{=} \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}^n, X_{i-1}, \ldots, X_1) \tag{13.63}$$

$$\overset{(f)}{\geq} \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}_i) \tag{13.64}$$

$$= \sum_{i=1}^{n} I(X_i; \hat{X}_i) \tag{13.65}$$

$$\overset{(g)}{\geq} \sum_{i=1}^{n} R(Ed(X_i, \hat{X}_i)) \tag{13.66}$$

$$= n \sum_{i=1}^{n} \frac{1}{n} R(Ed(X_i, \hat{X}_i)) \tag{13.67}$$

$$\overset{(h)}{\geq} nR\left(\frac{1}{n} \sum_{i=1}^{n} Ed(X_i, \hat{X}_i)\right) \tag{13.68}$$

$$\overset{(i)}{=} nR(Ed(X^n, \hat{X}^n)) \tag{13.69}$$

$$= nR(D), \tag{13.70}$$

where

(a) follows from the fact that there are at most 2^{nR} $\hat{X}^n$'s in the range of the encoding function,
(b) from the fact that $\hat{X}^n$ is a function of X^n and thus $H(\hat{X}^n|X^n) = 0$,
(c) from the definition of mutual information,
(d) from the fact that the X_i are independent,
(e) from the chain rule for entropy,
(f) from the fact that conditioning reduces entropy,
(g) from the definition of the rate distortion function,
(h) from the convexity of the rate distortion function (Lemma 13.4.1) and Jensen's inequality, and
(i) from the definition of distortion for blocks of length n.

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. □

13.5 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

$$\left| -\frac{1}{n} \log p(x^n) - H(X) \right| < \epsilon \tag{13.71}$$

$$\left| -\frac{1}{n} \log p(\hat{x}^n) - H(\hat{X}) \right| < \epsilon \tag{13.72}$$

$$\left| -\frac{1}{n} \log p(x^n, \hat{x}^n) - H(X, \hat{X}) \right| < \epsilon \tag{13.73}$$

$$|d(x^n, \hat{x}^n) - Ed(X, \hat{X})| < \epsilon \tag{13.74}$$

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 8.6) 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 $\sim p(x, \hat{x})$, then the distortion between two random sequences

$$d(X^n, \hat{X}^n) = \frac{1}{n} \sum_{i=1}^{n} d(X_i, \hat{X}_i) \tag{13.75}$$

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 13.5.1: *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 13.5.2: *For all* $(x^n, \hat{x}^n) \in A_{d,\epsilon}^{(n)}$,

$$p(\hat{x}^n) \geq p(\hat{x}^n | x^n) 2^{-n(I(X;\hat{X})+3\epsilon)}. \tag{13.76}$$

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

$$p(\hat{x}^n|x^n) = \frac{p(x^n, \hat{x}^n)}{p(x^n)} \tag{13.77}$$

$$= p(\hat{x}^n)\frac{p(x^n, \hat{x}^n)}{p(x^n)p(\hat{x}^n)} \tag{13.78}$$

$$\leq p(\hat{x}^n)\frac{2^{-n(H(X,\hat{X})-\epsilon)}}{2^{-n(H(X)+\epsilon)}2^{-n(H(\hat{X})+\epsilon)}} \tag{13.79}$$

$$= p(\hat{x}^n)2^{n(I(X;\hat{X})+3\epsilon)}, \tag{13.80}$$

and the lemma follows immediately. □

We also need the following interesting inequality.

Lemma 13.5.3: *For* $0 \leq x, y \leq 1$, $n > 0$,

$$(1 - xy)^n \leq 1 - x + e^{-yn}. \tag{13.81}$$

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

$$(1 - y)^n \leq e^{-yn}. \tag{13.82}$$

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

$$(1 - xy)^n = g_y(x) \tag{13.83}$$

$$\leq (1 - x)g_y(0) + xg_y(1) \tag{13.84}$$

$$= (1 - x)1 + x(1 - y)^n \tag{13.85}$$

$$\leq 1 - x + xe^{-yn} \tag{13.86}$$

$$\leq 1 - x + e^{-yn}. \quad \square \tag{13.87}$$

We use this to prove the achievability of Theorem 13.2.1.

Proof (*Achievability in Theorem 13.2.1*): *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 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 (13.53). 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 $\mathscr{C}$ 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 $\mathscr{C}$ as

$$\bar{D} = E_{X^n, \mathscr{C}} d(X^n, \hat{X}^n) \tag{13.88}$$

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

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

$$Ed(X^n, \hat{X}^n(X^n)) \leq D + \epsilon + P_e d_{\max}, \tag{13.89}$$

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, then 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 $\mathscr{C}$ and a randomly chosen source sequence, there is no codeword that is distortion typical with the source sequence. Let $J(\mathscr{C})$ denote the set of source sequences x^n such that at least one codeword in $\mathscr{C}$ is distortion typical with x^n.

Then

$$P_e = \sum_{\mathscr{C}} P(\mathscr{C}) \sum_{x^n : x^n \notin J(\mathscr{C})} p(x^n) . \tag{13.90}$$

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

$$P_e = \sum_{x^n} p(x^n) \sum_{\mathscr{C} : x^n \notin J(\mathscr{C})} p(\mathscr{C}) . \tag{13.91}$$

Let us define

$$K(x^n, \hat{x}^n) = \begin{cases} 1 & \text{if } (x^n, \hat{x}^n) \in A_{d,\epsilon}^{(n)} , \\ 0 & \text{if } (x^n, \hat{x}^n) \notin A_{d,\epsilon}^{(n)} . \end{cases} \tag{13.92}$$

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

$$\Pr((x^n, \hat{X}^n) \notin A_{d,\epsilon}^{(n)}) = \Pr(K(x^n, \hat{X}^n) = 0) = 1 - \sum_{\hat{x}^n} p(\hat{x}^n) K(x^n, \hat{x}^n) , \tag{13.93}$$

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

$$P_e = \sum_{x^n} p(x^n) \sum_{\mathscr{C} : x^n \notin J(\mathscr{C})} p(\mathscr{C}) \tag{13.94}$$

$$= \sum_{x^n} p(x^n) \left[1 - \sum_{\hat{x}^n} p(\hat{x}^n) K(x^n, \hat{x}^n) \right]^{2^{nR}} . \tag{13.95}$$

We now use Lemma 13.5.2 to bound the sum within the brackets. From Lemma 13.5.2, it follows that

$$\sum_{\hat{x}^n} p(\hat{x}^n) K(x^n, \hat{x}^n) \geq \sum_{\hat{x}^n} p(\hat{x}^n | x^n) 2^{-n(I(X;\hat{X})+3\epsilon)} K(x^n, \hat{x}^n) , \tag{13.96}$$

and hence

$$P_e \leq \sum_{x^n} p(x^n)\left(1 - 2^{-n(I(X;\hat{X})+3\epsilon)} \sum_{\hat{x}^n} p(\hat{x}^n|x^n)K(x^n,\hat{x}^n)\right)^{2^{nR}}. \quad (13.97)$$

We now use Lemma 13.5.3 to bound the term on the right hand side of (13.97) and obtain

$$\left(1 - 2^{-n(I(X;\hat{X})+3\epsilon)} \sum_{\hat{x}^n} p(\hat{x}^n|x^n)K(x^n,\hat{x}^n)\right)^{2^{nR}}$$
$$\leq 1 - \sum_{\hat{x}^n} p(\hat{x}^n|x^n)K(x^n,\hat{x}^n) + e^{-(2^{-n(I(X;\hat{X})+3\epsilon)}2^{nR})}. \quad (13.98)$$

Substituting this inequality in (13.97), we obtain

$$P_e \leq 1 - \sum_{x^n} p(x^n)p(\hat{x}^n|x^n)K(x^n,\hat{x}^n) + e^{-2^{-n(I(X;\hat{X})+3\epsilon)}2^{nR}}. \quad (13.99)$$

The last term in the bound is equal to

$$e^{-2^{n(R-I(X;\hat{X})-3\epsilon)}}, \quad (13.100)$$

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 $R > I(X;\hat{X})$ and we can choose ϵ small enough so that the last term in (13.99) goes to 0.

The first two terms in (13.99) 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 13.5.1,

$$1 - \sum_{x^n}\sum_{\hat{x}^n} p(x^n,\hat{x}^n)K(x^n,\hat{x}^n) = \Pr((X^n,\hat{X}^n) \notin A_{d,\epsilon}^{(n)}) \quad (13.101)$$

$$< \epsilon \quad (13.102)$$

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 $\mathscr{C}^*$ 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

$$M \leq \frac{(\sqrt{n(P+N)})^n}{(\sqrt{nN})^n} = \left(\frac{P+N}{N}\right)^{n/2}. \tag{13.103}$$

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

$$2^{nC} = \left(\frac{P+N}{N}\right)^{n/2} \tag{13.104}$$

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

$$2^{nR(D)} = \left(\frac{\sigma^2}{D}\right)^{n/2}. \tag{13.105}$$

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 which have a large minimum distance between codewords, while 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.

13.6 STRONGLY TYPICAL SEQUENCES AND RATE DISTORTION

In the last section, we proved the existence of a rate distortion code of rate $R(D)$ with average distortion close to D. But a stronger statement is true—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 stronger result is more involved; we will only give an outline of the proof. The method of proof is similar to the proof in the previous section; the main difference is that we will use strongly typical sequences rather than weakly typical sequences. This will enable us to give a lower bound to the probability that a typical source sequence is not well represented by a randomly chosen codeword in (13.93). This will give a more intuitive proof of the rate distortion theorem.

We will 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 [28] and were explored in detail in the book by Csiszár and Körner [83]. We will define strong typicality (as in Chapter 12) and state a fundamental lemma. The proof of the lemma will be left as a problem at the end of the chapter.

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

$$\left| \frac{1}{n} N(a|x^n) - p(a) \right| < \frac{\epsilon}{|\mathcal{X}|}, \tag{13.106}$$

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

$$\left| \frac{1}{n} N(a, b|x^n, y^n) - p(a, b) \right| < \frac{\epsilon}{|\mathcal{X}||\mathcal{Y}|}, \tag{13.107}$$

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 13.6.1: *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 8.6.1 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 13.6.2: *Let* $Y_1, Y_2, \ldots, Y_n$ *be drawn i.i.d.* $\sim \prod p(y)$. *For* $x^n \in A_\epsilon^{*(n)}$, *the probability that* $(x^n, Y^n) \in A_\epsilon^{*(n)}$ *is bounded by*

$$2^{-n(I(X;Y)+\epsilon_1)} \leq \Pr((x^n, Y^n) \in A_\epsilon^{*(n)}) \leq 2^{-n(I(X;Y)-\epsilon_1)}, \tag{13.108}$$

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 a problem 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 the last section.

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 $\mathscr{C}$ 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 the last section, we calculate the expected distortion over the random choice of codebook as

$$D = E_{X^n, \mathscr{C}} d(X^n, \hat{X}^n) \tag{13.109}$$

$$= E_{\mathscr{C}} \sum_{x^n} p(x^n) d(x^n, \hat{X}^n(x^n)) \tag{13.110}$$

$$= \sum_{x^n} p(x^n) E_{\mathscr{C}} d(x^n, \hat{X}^n), \tag{13.111}$$

where the expectation is over the random choice of codebook.

For a fixed codebook $\mathscr{C}$, we divide the sequences $x^n \in \mathscr{X}^n$ into three categories as shown in Figure 13.8.

- *The non-typical 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 non-typical 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* $\hat{X}^n$ *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

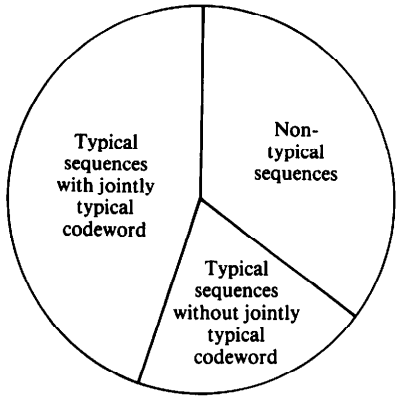

Figure 13.8. Classes of source sequences in rate distortion theorem.

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 (13.93), we can expand the probability of error as

$$P_e = \sum_{x^n \in A_\epsilon^{*(n)}} p(x^n)[1 - \Pr((x^n, \hat{X}^n) \in A_\epsilon^{*(n)})]^{2^{nR}}. \tag{13.112}$$

From Lemma 13.6.2, we have

$$\Pr((x^n, \hat{X}^n) \in A_\epsilon^{*(n)}) \geq 2^{-n(I(X;\hat{X})+\epsilon_1)}. \tag{13.113}$$

Substituting this in (13.112) and using the inequality $(1-x)^n \leq e^{-nx}$, we have

$$P_e \leq e^{-(2^{nR}2^{-n(I(X;\hat{X})+\epsilon_1)})} \tag{13.114}$$

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 $D + \delta$. □

13.7 CHARACTERIZATION OF THE RATE DISTORTION FUNCTION

We have defined the information rate distortion function as

$$R(D) = \min_{q(\hat{x}|x):\, \Sigma_{(x,\hat{x})} p(x)q(\hat{x}|x)d(x,\hat{x}) \leq D} I(X;\hat{X}), \tag{13.115}$$

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 $\Sigma_{\hat{x}} q(\hat{x}|x) = 1$ for all x and $\Sigma q(\hat{x}|x)p(x)d(x,\hat{x}) \leq D$.

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

$$J(q) = \sum_x \sum_{\hat{x}} p(x)q(\hat{x}|x) \log \frac{q(\hat{x}|x)}{\Sigma_x p(x)q(\hat{x}|x)} + \lambda \sum_x \sum_{\hat{x}} p(x)q(\hat{x}|x)d(x,\hat{x}) \tag{13.116}$$

$$+ \sum_x \nu(x) \sum_{\hat{x}} q(\hat{x}|x), \tag{13.117}$$

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

$$J(q) = \sum_x \sum_{\hat{x}} p(x)q(\hat{x}|x) \log \frac{q(\hat{x}|x)}{q(\hat{x})} + \lambda \sum_x \sum_{\hat{x}} p(x)q(\hat{x}|x)d(x,\hat{x}) \quad (13.118)$$

$$+ \sum_x \nu(x) \sum_{\hat{x}} q(\hat{x}|x) . \quad (13.119)$$

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

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

$$+ \nu(x) = 0 . \quad (13.120)$$

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

$$p(x) \left[\log \frac{q(\hat{x}|x)}{q(\hat{x})} + \lambda d(x,\hat{x}) + \log \mu(x) \right] = 0 \quad (13.121)$$

or

$$q(\hat{x}|x) = \frac{q(\hat{x}) e^{-\lambda d(x,\hat{x})}}{\mu(x)} . \quad (13.122)$$

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

$$\mu(x) = \sum_{\hat{x}} q(\hat{x}) e^{-\lambda d(x,\hat{x})} \quad (13.123)$$

or

$$q(\hat{x}|x) = \frac{q(\hat{x}) e^{-\lambda d(x,\hat{x})}}{\Sigma_{\hat{x}} q(\hat{x}) e^{-\lambda d(x,\hat{x})}} . \quad (13.124)$$

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

$$q(\hat{x}) = q(\hat{x}) \sum_x \frac{p(x) e^{-\lambda d(x,\hat{x})}}{\Sigma_{\hat{x}'} q(\hat{x}') e^{-\lambda d(x,\hat{x}')}} . \quad (13.125)$$

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

$$\sum_x \frac{p(x) e^{-\lambda d(x,\hat{x})}}{\Sigma_{\hat{x}'} q(\hat{x}') e^{-\lambda d(x,\hat{x}')}} = 1 \quad (13.126)$$

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 $|\hat{\mathscr{X}}|$ unknowns $q(\hat{x})$. We can use this and (13.124) to find the optimum conditional distribution.

The above analysis is valid if all the output symbols are active, i.e., $q(\hat{x}) > 0$ for all $\hat{x}$. But this is not necessarily the case. We would then have to apply the Kuhn-Tucker conditions to characterize the minimum. The inequality condition $q(\hat{x}) > 0$ is covered by the Kuhn-Tucker conditions, which reduce to

$$\frac{\partial J}{\partial q(\hat{x}|x)} \begin{cases} = 0 & \text{if } q(\hat{x}|x) > 0\,, \\ \geq 0 & \text{if } q(\hat{x}|x) = 0\,. \end{cases} \tag{13.127}$$

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

$$\sum_x \frac{p(x)e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{x}'} q(\hat{x}')e^{-\lambda d(x,\hat{x}')}} = 1 \quad \text{if } q(\hat{x}) > 0\,, \tag{13.128}$$

$$\sum_x \frac{p(x)e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{x}'} q(\hat{x}')e^{-\lambda d(x,\hat{x}')}} \leq 1 \quad \text{if } q(\hat{x}) = 0\,. \tag{13.129}$$

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.

13.8 COMPUTATION OF CHANNEL CAPACITY AND THE RATE DISTORTION FUNCTION

Consider the following problem: Given two convex sets A and B in $\mathscr{R}^n$ as shown in Figure 13.9, we would like to the find the minimum distance between them

$$d_{\min} = \min_{a \in A,\, b \in B} d(a, b)\,, \tag{13.130}$$

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 [85] have shown that if the sets are convex and if the distance satisfies certain conditions, then this alternating minimiza-

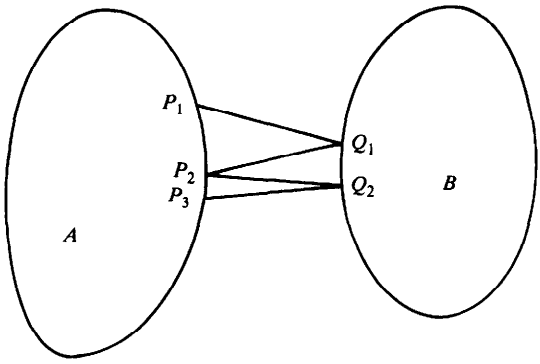

Figure 13.9. Distance between convex sets.

tion 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, then the algorithm does converge to the the minimum relative entropy between the two sets of distributions.

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:

Lemma 13.8.1: *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 $p(y|x)$, i.e.,*

$$D(p(x)p(y|x)\|p(x)r^*(y)) = \min_{r(y)} D(p(x)p(y|x)\|p(x)r(y)), \tag{13.131}$$

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

$$\max_{r(x|y)} \sum_{x,y} p(x)p(y|x) \log \frac{r(x|y)}{p(x)} = \sum_{x,y} p(x)p(y|x) \log \frac{r^*(x|y)}{p(x)}, \tag{13.132}$$

where

$$r^*(x|y) = \frac{p(x)p(y|x)}{\Sigma_x p(x)p(y|x)}. \tag{13.133}$$

Proof:

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

$$= \sum_{x,y} p(x)p(y|x) \log \frac{p(x)p(y|x)}{p(x)r(y)} \tag{13.134}$$

$$-\sum_{x,y} p(x)p(y|x)\log\frac{p(x)p(y|x)}{p(x)r^*(y)} \tag{13.135}$$

$$=\sum_{x,y} p(x)p(y|x)\log\frac{r^*(y)}{r(y)} \tag{13.136}$$

$$=\sum_{y} r^*(y)\log\frac{r^*(y)}{r(y)} \tag{13.137}$$

$$=D(r^*\|r) \tag{13.138}$$

$$\geq 0\,. \tag{13.139}$$

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,

$$R(D)=\min_{r(\hat{x})}\ \min_{q(\hat{x}|x):\,\Sigma\, p(x)q(\hat{x}|x)d(x,\hat{x})\leq D}\ \sum_{x}\sum_{\hat{x}} p(x)q(\hat{x}|x)\log\frac{q(\hat{x}|x)}{r(\hat{x})}\,. \tag{13.140}$$

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 $r(\hat{x})$, then we can write

$$R(D)=\min_{q\in B}\min_{p\in A} D(p\|q)\,. \tag{13.141}$$

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 a distortion constraint. We can use the method of Lagrange multipliers for this minimization to obtain

$$q(\hat{x}|x)=\frac{r(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{x}} r(\hat{x})e^{-\lambda d(x,\hat{x})}}\,. \tag{13.142}$$

For this conditional distribution $q(\hat{x}|x)$, we calculate the output distribution $r(\hat{x})$ that minimizes the mutual information, which by Lemma 13.8.1 is

$$r(\hat{x})=\sum_{x} p(x)q(\hat{x}|x)\,. \tag{13.143}$$

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 (13.140). Thus there is a limit, and the limit has been shown to be $R(D)$ by Csiszár [79],

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,

$$C = \max_{r(x)} I(X; Y) = \max_{r(x)} \sum_x \sum_y r(x)p(y|x) \log \frac{r(x)p(y|x)}{r(x) \Sigma_{x'} r(x')p(y|x')} \tag{13.144}$$

as a double maximization using Lemma 13.8.1,

$$C = \max_{q(x|y)} \max_{r(x)} \sum_x \sum_y r(x)p(y|x) \log \frac{q(x|y)}{r(x)} . \tag{13.145}$$

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 13.8.1,

$$q(x|y) = \frac{r(x)p(y|x)}{\Sigma_x r(x)p(y|x)} . \tag{13.146}$$

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

$$r(x) = \frac{\Pi_y (q(x|y))^{p(y|x)}}{\Sigma_x \Pi_y (q(x|y))^{p(y|x)}} , \tag{13.147}$$

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 [37] and Arimoto [11] and the convergence for the rate distortion computation was proved by Csiszár [79]. The alternating minimization procedure of Csiszár and Tusnady can be specialized to many other situations as well, including the EM algorithm [88], and the algorithm for finding the log-optimal portfolio for a stock market [64].

SUMMARY OF CHAPTER 13

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

$$R(D) = \min_{p(\hat{x}|x) : \Sigma_{(x,\hat{x})} p(x)p(\hat{x}|x)d(x,\hat{x}) \le D} I(X; \hat{X}) , \tag{13.148}$$

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 number of codewords $|\hat{X}^n(\cdot)| \leq 2^{nR}$ 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,

$$R(D) = H(p) - H(D). \tag{13.149}$$

Gaussian source: For a Gaussian source with squared error distortion,

$$R(D) = \frac{1}{2} \log \frac{\sigma^2}{D}. \tag{13.150}$$

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 FOR CHAPTER 13

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 $R = 1$.

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

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

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

$$X \sim \text{Bern}(1/2),$$
$$d(x, \hat{x}) = \begin{bmatrix} 0 & a \\ b & 0 \end{bmatrix}.$$

   ($R(D)$ cannot be expressed in closed form.)

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 $R'(D) = R(D + \bar{w})$, where $\bar{w} = \Sigma\, p_i w_i$, 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 $\hat{x}$ which reproduces the source with zero distortion.

This result is due to Pinkston [209].

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, i.e.,

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

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

$$\phi(D) = \max_{\mathbf{p}:\, \sum_{i=1}^{m} p_i d_i \leq D} H(\mathbf{p}). \tag{13.151}$$

The Shannon lower bound on the rate distortion function [245] 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 $Ed(X, \hat{X}) \leq D$,

$$I(X; \hat{X}) = H(X) - H(X|\hat{X}) \tag{13.152}$$

$$= H(X) - \sum_{\hat{x}} p(\hat{x}) H(X|\hat{X} = \hat{x}) \tag{13.153}$$

$$\geq H(X) - \sum_{\hat{x}} p(\hat{x}) \phi(D_{\hat{x}}) \tag{13.154}$$

$$\geq H(X) - \phi\left(\sum_{\hat{x}} p(\hat{x}) D_{\hat{x}}\right) \tag{13.155}$$

$$\geq H(X) - \phi(D), \tag{13.156}$$

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

(c) Argue that

$$R(D) \geq H(X) - \phi(D), \tag{13.157}$$

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.

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

$$d(x, \hat{x}) = \begin{bmatrix} 0 & 1 & \infty \\ \infty & 1 & 0 \end{bmatrix}. \tag{13.158}$$

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?

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

$$h(X) - \frac{1}{2}\log(2\pi e)D \le R(D) \le \frac{1}{2}\log\frac{\sigma^2}{D}. \tag{13.159}$$

For the upper bound, consider the joint distribution shown in Figure 13.10. Are Gaussian random variables harder or easier to describe than other random variables with the same variance?

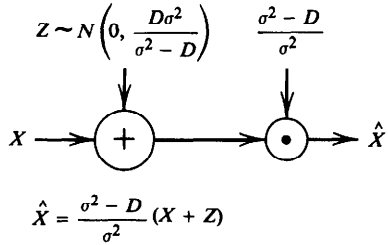

Figure 13.10. Joint distribution for upper bound on rate distortion function.

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 (13.58–13.70) considering the conditions for equality and interpret as properties of a good code. For example, equality in (13.59) implies that $\hat{X}^n$ is a deterministic function of X^n.

10. *Probability of conditionally typical sequences.* In Chapter 8, we calculated the probability that two independently drawn sequences X^n and Y^n will be 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 which work for all typical x^n sequences. We will 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 [28] and is fully developed in the book by Csiszár and Körner [83].

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

$$|A_\epsilon^{*(n)}| \doteq 2^{nH(X)} \tag{13.160}$$

Hint: Theorem 12.1.1 and 12.1.3.

(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, i.e.,

$$p_{x^n, y^n}(a, b) = \frac{1}{n} N(a, b|x^n, y^n) = \frac{1}{n} \sum_{i=1}^{n} I(x_i = a, y_i = b). \tag{13.161}$$

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

$$V_{y^n|x^n}(b|a) = \frac{N(a, b|x^n, y^n)}{N(a|x^n)}. \tag{13.162}$$

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

$$\frac{1}{(n+1)^{|\mathcal{X}||\mathcal{Y}|}} 2^{nH(Y|X)} \le |T_V(x^n)| \le 2^{nH(Y|X)}. \tag{13.163}$$

(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 $V(b|a) > 0$,

$$\frac{1}{n} \left| N(a, b|x^n, y^n) - V(b|a) N(a|x^n) \right| \le \frac{\epsilon}{|\mathcal{Y}| + 1}. \tag{13.164}$$

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

$$\frac{1}{(n+1)^{|\mathcal{X}||\mathcal{Y}|}} 2^{n(H(Y|X) - \epsilon_1)} \le |A_\epsilon^{*(n)}(Y|x^n)| \le (n+1)^{|\mathcal{X}||\mathcal{Y}|} 2^{n(H(Y|X) + \epsilon_1)}, \tag{13.165}$$

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.

$$\left| \frac{1}{n} N(a, b|x^n, y^n) - p(a, b) \right| < \frac{\epsilon}{|\mathcal{X}||\mathcal{Y}|} \tag{13.166}$$

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

$$\frac{1}{(n+1)^{|\mathcal{X}||\mathcal{Y}|}} 2^{n(H(Y|X)-\delta(\epsilon))} \le \left| \{y^n : (x^n, y^n) \in A_\epsilon^{*(n)}\} \right| \le (n+1)^{|\mathcal{X}||\mathcal{Y}|} 2^{n(H(Y|X)+\delta(\epsilon))}, \tag{13.167}$$

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

$$2^{n(H(Y|X)-\epsilon_2)} \le \left| \{y^n : (x^n, y^n) \in A_\epsilon^{*(n)}\} \right| \le 2^{n(H(Y|X)+\epsilon_2)}, \tag{13.168}$$

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

$$2^{-n(I(X;Y)+\epsilon_3)} \le \Pr((x^n, Y^n) \in A_\epsilon^{*(n)}) \le 2^{-n(I(X;Y)-\epsilon_3)}, \tag{13.169}$$

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

HISTORICAL NOTES

The idea of rate distortion was introduced by Shannon in his original paper [238]. He returned to it and dealt with it exhaustively in his 1959 paper [245], 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 [27].

The inverse water-filling solution for the rate-distortion function for parallel Gaussian sources was established by McDonald and Schultheiss [190]. 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 [37] and Arimoto [11] and Csiszár [79]. This algorithm is a special case of general alternating minimization algorithm due to Csiszár and Tusnady [85].