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

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

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 $T(P)$:

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

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

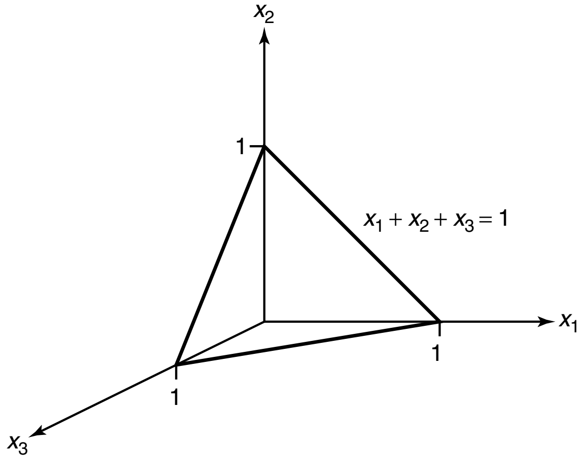

FIGURE 11.1. Probability simplex in $\mathcal{R}^3$.

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* **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

$$1 \ge P^n(T(P)) \tag{11.18}$$

$$= \sum_{\mathbf{x}\in T(P)} P^n(\mathbf{x}) \tag{11.19}$$

$$= \sum_{\mathbf{x}\in T(P)} 2^{-nH(P)} \tag{11.20}$$

$$= |T(P)|2^{-nH(P)}, \tag{11.21}$$

using Theorem 11.1.2. Thus,

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

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:

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

We lower bound the ratio of probabilities,

$$\frac{P^n(T(P))}{P^n(T(\hat{P}))} = \frac{|T(P)| \prod_{a\in\mathcal{X}} P(a)^{nP(a)}}{|T(\hat{P})| \prod_{a\in\mathcal{X}} P(a)^{n\hat{P}(a)}} \tag{11.24}$$

$$= \frac{\binom{n}{nP(a_1),\, nP(a_2),\ldots,nP(a_{|\mathcal{X}|})} \prod_{a\in\mathcal{X}} P(a)^{nP(a)}}{\binom{n}{n\hat{P}(a_1),\, n\hat{P}(a_2),\ldots,n\hat{P}(a_{|\mathcal{X}|})} \prod_{a\in\mathcal{X}} P(a)^{n\hat{P}(a)}} \tag{11.25}$$

$$= \prod_{a\in\mathcal{X}} \frac{(n\hat{P}(a))!}{(nP(a))!} P(a)^{n(P(a)-\hat{P}(a))}. \tag{11.26}$$

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

$$\begin{aligned} 1 &\ge \binom{n}{k} \left(\frac{k}{n}\right)^k \left(1 - \frac{k}{n}\right)^{n-k} && (11.42)\\ &= \binom{n}{k} 2^{k \log \frac{k}{n} + (n-k) \log \frac{n-k}{n}} && (11.43)\\ &= \binom{n}{k} 2^{n\left(\frac{k}{n} \log \frac{k}{n} + \frac{n-k}{n} \log \frac{n-k}{n}\right)} && (11.44)\\ &= \binom{n}{k} 2^{-nH\left(\frac{k}{n}\right)}. && (11.45) \end{aligned}$$

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,

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

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

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

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

which by the arguments in (11.45) is equivalent to

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

or

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

Combining the two results, we see that

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

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,*

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

Proof: We have

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

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

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

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

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

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

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

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

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

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

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

$$P_e^{(n)} = 1 - Q^n(A) \tag{11.84}$$

$$= \sum_{P:H(P)>R_n} Q^n(T(P)) \tag{11.85}$$

$$\leq (n+1)^{|\mathcal{X}|} \max_{P:H(P)>R_n} Q^n(T(P)) \tag{11.86}$$

$$\leq (n+1)^{|\mathcal{X}|} 2^{-n \min_{P:H(P)>R_n} D(P||Q)}. \tag{11.87}$$

Since $R_n \uparrow R$ and $H(Q) < R$, there exists n_0 such that for all $n \geq n_0$, $R_n > H(Q)$. Then for $n \geq n_0$, $\min_{P:H(P)>R_n} D(P||Q)$ must be greater than 0, and the probability of error $P_e^{(n)}$ converges to 0 exponentially fast as $n \to \infty$.

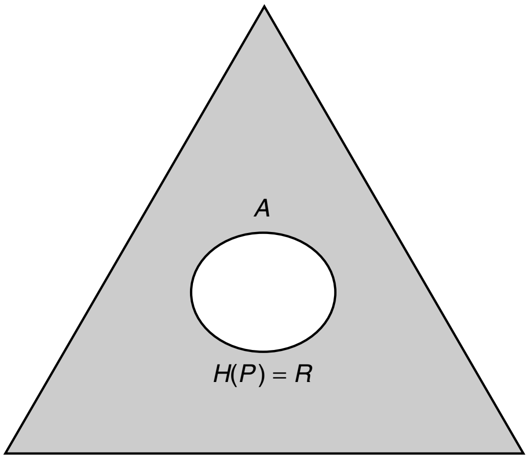

FIGURE 11.2. Universal code and the probability simplex. Each sequence with type that lies outside the circle is encoded by its index. There are fewer than 2^{nR} such sequences. Sequences with types within the circle are encoded by 0.

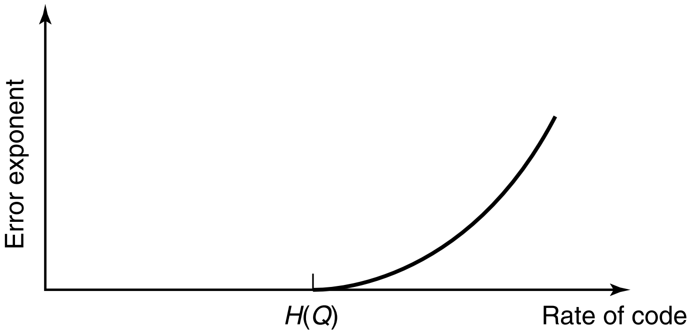

FIGURE 11.3. Error exponent for the universal code.

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., $\frac{1}{n}\sum_i g(x_i) \geq \alpha$]. This event is equivalent to the event $P_{\mathbf{X}} \in E \cap \mathcal{P}_n$, where

$$E = \left\{ P : \sum_{a \in \mathcal{X}} g(a)P(a) \geq \alpha \right\}, \tag{11.90}$$

because

$$\frac{1}{n}\sum_{i=1}^{n} g(x_i) \geq \alpha \Leftrightarrow \sum_{a \in \mathcal{X}} P_{\mathbf{X}}(a)g(a) \geq \alpha \tag{11.91}$$

$$\Leftrightarrow P_{\mathbf{X}} \in E \cap \mathcal{P}_n. \tag{11.92}$$

Thus,

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} g(X_i) \geq \alpha\right) = Q^n(E \cap \mathcal{P}_n) = Q^n(E). \tag{11.93}$$

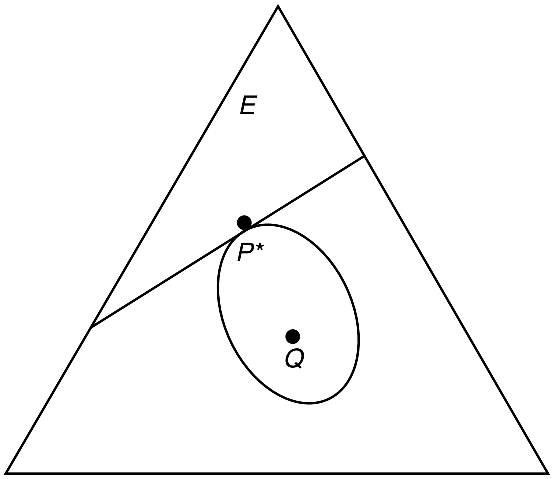

FIGURE 11.4. Probability simplex and Sanov's theorem.

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*

$$Q^n(E) = Q^n(E \cap \mathcal{P}_n) \leq (n+1)^{|\mathcal{X}|} 2^{-nD(P^*||Q)}, \tag{11.94}$$

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:

$$\begin{aligned} Q^n(E) &= \sum_{P \in E \cap \mathcal{P}_n} Q^n(T(P)) && (11.97)\\ &\leq \sum_{P \in E \cap \mathcal{P}_n} 2^{-nD(P||Q)} && (11.98) \end{aligned}$$

$$\leq \sum_{P\in E\cap\mathcal{P}_n} \max_{P\in E\cap\mathcal{P}_n} 2^{-nD(P||Q)} \tag{11.99}$$

$$= \sum_{P\in E\cap\mathcal{P}_n} 2^{-n\min_{P\in E\cap\mathcal{P}_n} D(P||Q)} \tag{11.100}$$

$$\leq \sum_{P\in E\cap\mathcal{P}_n} 2^{-n\min_{P\in E} D(P||Q)} \tag{11.101}$$

$$= \sum_{P\in E\cap\mathcal{P}_n} 2^{-nD(P^*||Q)} \tag{11.102}$$

$$\leq (n+1)^{|\mathcal{X}|}2^{-nD(P^*||Q)}, \tag{11.103}$$

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 $\Pr\{\frac{1}{n}\sum_{i=1}^n g_j(X_i) \ge \alpha_j, j = 1, 2, \ldots, k\}$. Then the set E is defined as

$$E = \left\{ P : \sum_a P(a) g_j(a) \ge \alpha_j,\ j = 1, 2, \ldots, k \right\}. \tag{11.108}$$

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

$$\sum_{i=1}^{6} i P(i) \ge 4. \tag{11.112}$$

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

$$\begin{aligned} E = \{P(x, y) : &\left| -\sum_{x,y} P(x, y) \log Q(x) - H(X) \right| \le \epsilon, \\ &\left| -\sum_{x,y} P(x, y) \log Q(y) - H(Y) \right| \le \epsilon, \\ &\left| -\sum_{x,y} P(x, y) \log Q(x, y) - H(X, Y) \right| \le \epsilon\}. \end{aligned} \tag{11.118}$$

Using Sanov's theorem, the probability is

$$Q_0^n(E) \doteq 2^{-nD(P^*||Q_0)}, \tag{11.119}$$

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

$$D^* = \min_{P \in E} D(P||Q). \tag{11.120}$$

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

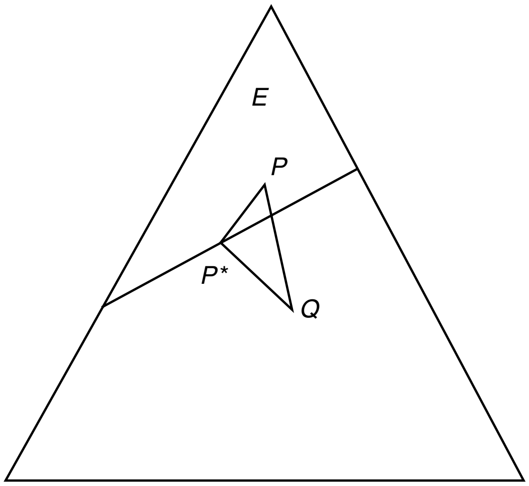

FIGURE 11.5. Pythagorean theorem for relative entropy.

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

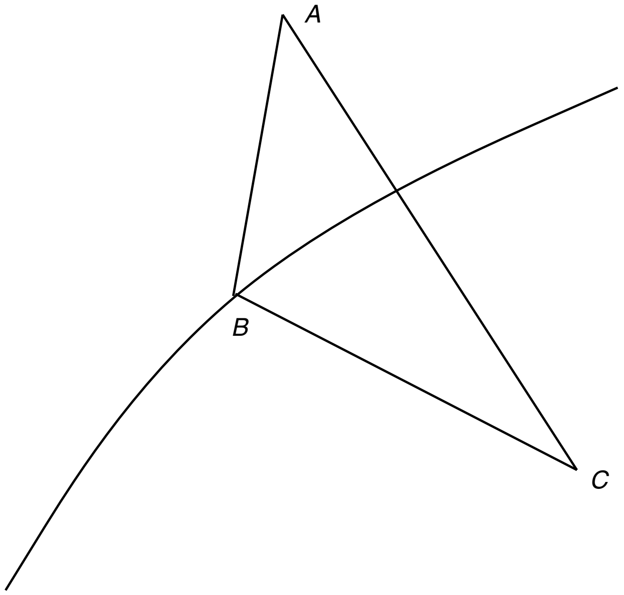

FIGURE 11.6. Triangle inequality for distance squared.

other point in the set. Then the angle between the lines BA and BC must be obtuse, which implies that $l_{AC}^2 \geq l_{AB}^2 + l_{BC}^2$, 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

$$D(P_1||P_2) \geq D(\hat{P}_1||\hat{P}_2) \tag{11.145}$$

$$\geq \frac{4}{2\ln 2}(P_1(A) - P_2(A))^2 \tag{11.146}$$

$$= \frac{1}{2\ln 2}||P_1 - P_2||_1^2, \tag{11.147}$$

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

$$\begin{aligned} Q^n(B) &= \sum_{P \in E \cap \mathcal{P}_n : D(P||Q) > D^*+2\delta} Q^n(T(P)) && (11.155)\\ &\leq \sum_{P \in E \cap \mathcal{P}_n : D(P||Q) > D^*+2\delta} 2^{-nD(P||Q)} && (11.156)\\ &\leq \sum_{P \in E \cap \mathcal{P}_n : D(P||Q) > D^*+2\delta} 2^{-n(D^*+2\delta)} && (11.157)\\ &\leq (n+1)^{|\mathcal{X}|} 2^{-n(D^*+2\delta)} && (11.158) \end{aligned}$$

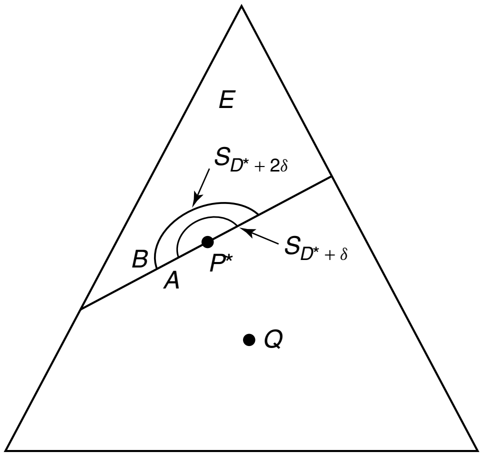

FIGURE 11.7. Conditional limit theorem.

since there are only a polynomial number of types. On the other hand,

$$Q^n(A) \geq Q^n(S_{D^*+\delta} \cap E) \tag{11.159}$$

$$= \sum_{P \in E \cap \mathcal{P}_n : D(P||Q) \leq D^*+\delta} Q^n(T(P)) \tag{11.160}$$

$$\geq \sum_{P \in E \cap \mathcal{P}_n : D(P||Q) \leq D^*+\delta} \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(P||Q)} \tag{11.161}$$

$$\geq \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-n(D^*+\delta)} \quad \text{for } n \text{ sufficiently large,} \tag{11.162}$$

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,

$$\Pr(P_{X^n} \in B | P_{X^n} \in E) = \frac{Q^n(B \cap E)}{Q^n(E)} \tag{11.163}$$

$$\leq \frac{Q^n(B)}{Q^n(A)} \tag{11.164}$$

$$\leq \frac{(n+1)^{|\mathcal{X}|} 2^{-n(D^*+2\delta)}}{\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-n(D^*+\delta)}} \tag{11.165}$$

$$= (n+1)^{2|\mathcal{X}|} 2^{-n\delta}, \tag{11.166}$$

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 theorem states that the first few elements are asymptotically independent with common distribution P^*.

Example 11.6.2 As an example of the conditional limit theorem, let us consider the case when n fair dice are rolled. Suppose that the sum of the outcomes exceeds $4n$. Then by the conditional limit theorem, the probability that the first die shows a number $a \in \{1, 2, \ldots, 6\}$ is approximately $P^*(a)$, where $P^*(a)$ is the distribution in E that is closest to the uniform distribution, where $E = \{P : \sum P(a)a \geq 4\}$. This is the maximum entropy distribution given by

$$P^*(x) = \frac{2^{\lambda x}}{\sum_{i=1}^{6} 2^{\lambda i}}, \tag{11.173}$$

with λ chosen so that $\sum i P^*(i) = 4$ (see Chapter 12). Here P^* is the conditional distribution on the first (or any other) die. Apparently, the first few dice inspected will behave as if they are drawn independently according to an exponential distribution.

11.7 HYPOTHESIS TESTING

One of the standard problems in statistics is to decide between two alternative explanations for the data observed. For example, in medical testing, one may wish to test whether or not a new drug is effective. Similarly, a sequence of coin tosses may reveal whether or not the coin is biased.

These problems are examples of the general hypothesis-testing problem. In the simplest case, we have to decide between two i.i.d. distributions. The general problem can be stated as follows:

Problem 11.7.1 Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim Q(x)$. We consider two hypotheses:

- H_1: $Q = P_1$.
- H_2: $Q = P_2$.

Consider the general decision function $g(x_1, x_2, \ldots, x_n)$, where $g(x_1, x_2, \ldots, x_n) = 1$ means that H_1 is accepted and $g(x_1, x_2, \ldots, x_n) = 2$ means that H_2 is accepted. Since the function takes on only two values, the test can also be specified by specifying the set A over which $g(x_1, x_2, \ldots, x_n)$ is 1; the complement of this set is the set where $g(x_1, x_2, \ldots, x_n)$ has the value 2. We define the two probabilities of error:

$$\alpha = \Pr(g(X_1, X_2, \ldots, X_n) = 2 | H_1 \text{ true}) = P_1^n(A^c) \tag{11.174}$$

and

$$\beta = \Pr(g(X_1, X_2, \ldots, X_n) = 1 | H_2 \text{ true}) = P_2^n(A). \tag{11.175}$$

In general, we wish to minimize both probabilities, but there is a tradeoff. Thus, we minimize one of the probabilities of error subject to a constraint on the other probability of error. The best achievable error exponent in the probability of error for this problem is given by the Chernoff–Stein lemma.

We first prove the Neyman–Pearson lemma, which derives the form of the optimum test between two hypotheses. We derive the result for discrete distributions; the same results can be derived for continuous distributions as well.

Theorem 11.7.1 (*Neyman–Pearson lemma*) *Let $X_1, X_2, \ldots, X_n$ be drawn i.i.d. according to probability mass function Q. Consider the decision problem corresponding to hypotheses $Q = P_1$ vs. $Q = P_2$. For $T \geq 0$, define a region*

$$A_n(T) = \left\{ x^n : \frac{P_1(x_1, x_2, \ldots, x_n)}{P_2(x_1, x_2, \ldots, x_n)} > T \right\}. \tag{11.176}$$

Let

$$\alpha^* = P_1^n(A_n^c(T)), \qquad \beta^* = P_2^n(A_n(T)) \tag{11.177}$$

be the corresponding probabilities of error corresponding to decision region A_n. Let B_n be any other decision region with associated probabilities of error α and β. If $\alpha \leq \alpha^*$, then $\beta \geq \beta^*$.

Proof: Let $A = A_n(T)$ be the region defined in (11.176) and let $B \subseteq \mathcal{X}^n$ be any other acceptance region. Let ϕ_A and ϕ_B be the indicator functions of the decision regions A and B, respectively. Then for all $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$,

$$(\phi_A(\mathbf{x}) - \phi_B(\mathbf{x}))(P_1(\mathbf{x}) - TP_2(\mathbf{x})) \geq 0. \tag{11.178}$$

This can be seen by considering separately the cases $\mathbf{x} \in A$ and $\mathbf{x} \notin A$. Multiplying out and summing this over the entire space, we obtain

$$0 \leq \sum (\phi_A P_1 - T\phi_A P_2 - P_1\phi_B + TP_2\phi_B) \tag{11.179}$$

$$= \sum_A (P_1 - TP_2) - \sum_B (P_1 - TP_2) \tag{11.180}$$

$$= (1-\alpha^*) - T\beta^* - (1-\alpha) + T\beta \tag{11.181}$$

$$= T(\beta - \beta^*) - (\alpha^* - \alpha). \tag{11.182}$$

Since $T \ge 0$, we have proved the theorem. □

The Neyman–Pearson lemma indicates that the optimum test for two hypotheses is of the form

$$\frac{P_1(X_1, X_2, \ldots, X_n)}{P_2(X_1, X_2, \ldots, X_n)} > T. \tag{11.183}$$

This is the likelihood ratio test and the quantity $\frac{P_1(X_1,X_2,\ldots,X_n)}{P_2(X_1,X_2,\ldots,X_n)}$ is called the *likelihood ratio*. For example, in a test between two Gaussian distributions [i.e., between $f_1 = \mathcal{N}(1, \sigma^2)$ and $f_2 = \mathcal{N}(-1, \sigma^2)$], the likelihood ratio becomes

$$\frac{f_1(X_1, X_2, \ldots, X_n)}{f_2(X_1, X_2, \ldots, X_n)} = \frac{\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X_i-1)^2}{2\sigma^2}}}{\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X_i+1)^2}{2\sigma^2}}} \tag{11.184}$$

$$= e^{+\frac{2\sum_{i=1}^n X_i}{\sigma^2}} \tag{11.185}$$

$$= e^{+\frac{2n\overline{X}_n}{\sigma^2}}. \tag{11.186}$$

Hence, the likelihood ratio test consists of comparing the sample mean $\overline{X}_n$ with a threshold. If we want the two probabilities of error to be equal, we should set $T = 1$. This is illustrated in Figure 11.8.

In Theorem 11.7.1 we have shown that the optimum test is a likelihood ratio test. We can rewrite the log-likelihood ratio as

$$L(X_1, X_2, \ldots, X_n) = \log \frac{P_1(X_1, X_2, \ldots, X_n)}{P_2(X_1, X_2, \ldots, X_n)} \tag{11.187}$$

$$= \sum_{i=1}^n \log \frac{P_1(X_i)}{P_2(X_i)} \tag{11.188}$$

$$= \sum_{a \in \mathcal{X}} nP_{X^n}(a) \log \frac{P_1(a)}{P_2(a)} \tag{11.189}$$

$$= \sum_{a \in \mathcal{X}} nP_{X^n}(a) \log \frac{P_1(a)}{P_2(a)} \frac{P_{X^n}(a)}{P_{X^n}(a)} \tag{11.190}$$

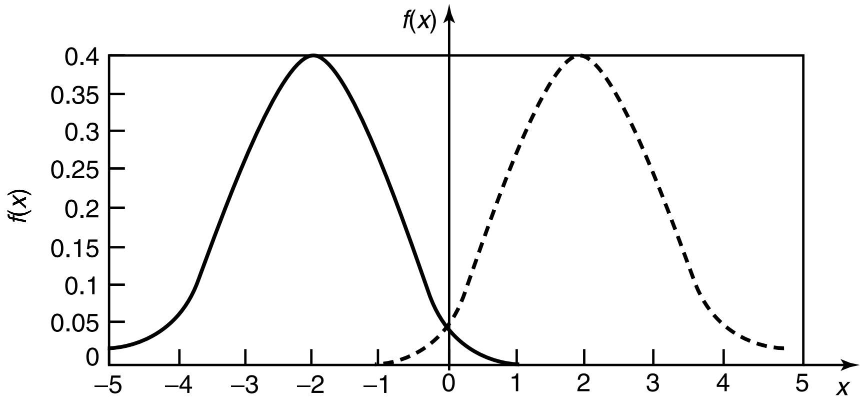

FIGURE 11.8. Testing between two Gaussian distributions.

$$= \sum_{a\in\mathcal{X}} nP_{X^n}(a) \log \frac{P_{X^n}(a)}{P_2(a)}$$

$$- \sum_{a\in\mathcal{X}} nP_{X^n}(a) \log \frac{P_{X^n}(a)}{P_1(a)} \tag{11.191}$$

$$= nD(P_{X^n}||P_2) - nD(P_{X^n}||P_1), \tag{11.192}$$

the difference between the relative entropy distances of the sample type to each of the two distributions. Hence, the likelihood ratio test

$$\frac{P_1(X_1, X_2, \ldots, X_n)}{P_2(X_1, X_2, \ldots, X_n)} > T \tag{11.193}$$

is equivalent to

$$D(P_{X^n}||P_2) - D(P_{X^n}||P_1) > \frac{1}{n}\log T. \tag{11.194}$$

We can consider the test to be equivalent to specifying a region of the simplex of types that corresponds to choosing hypothesis H_1. The optimum region is of the form (11.194), for which the boundary of the region is the set of types for which the difference between the distances is a constant. This boundary is the analog of the perpendicular bisector in Euclidean geometry. The test is illustrated in Figure 11.9.

We now offer some informal arguments based on Sanov's theorem to show how to choose the threshold to obtain different probabilities of error. Let B denote the set on which hypothesis 1 is accepted. The probability

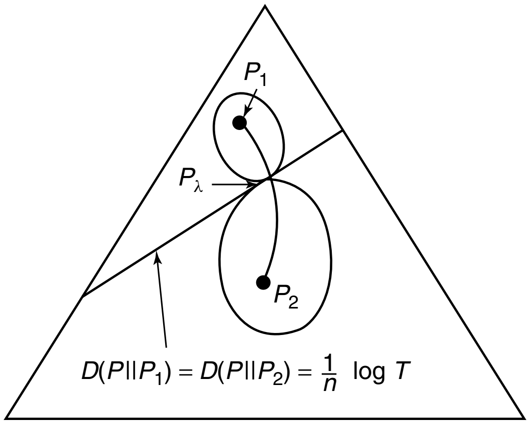

FIGURE 11.9. Likelihood ratio test on the probability simplex.

of error of the first kind is

$$\alpha_n = P_1^n(P_{X^n} \in B^c). \tag{11.195}$$

Since the set B^c is convex, we can use Sanov's theorem to show that the probability of error is determined essentially by the relative entropy of the closest member of B^c to P_1. Therefore,

$$\alpha_n \doteq 2^{-nD(P_1^*||P_1)}, \tag{11.196}$$

where P_1^* is the closest element of B^c to distribution P_1. Similarly,

$$\beta_n \doteq 2^{-nD(P_2^*||P_2)}, \tag{11.197}$$

where P_2^* is the closest element in B to the distribution P_2.

Now minimizing $D(P||P_2)$ subject to the constraint $D(P||P_2) - D(P||P_1) \geq \frac{1}{n}\log T$ will yield the type in B that is closest to P_2. Setting up the minimization of $D(P||P_2)$ subject to $D(P||P_2) - D(P||P_1) = \frac{1}{n}\log T$ using Lagrange multipliers, we have

$$J(P) = \sum P(x)\log\frac{P(x)}{P_2(x)} + \lambda\sum P(x)\log\frac{P_1(x)}{P_2(x)} + \nu\sum P(x). \tag{11.198}$$

Differentiating with respect to $P(x)$ and setting to 0, we have

$$\log\frac{P(x)}{P_2(x)} + 1 + \lambda\log\frac{P_1(x)}{P_2(x)} + \nu = 0. \tag{11.199}$$

Solving this set of equations, we obtain the minimizing P of the form

$$P_2^* = P_{\lambda^*} = \frac{P_1^\lambda(x) P_2^{1-\lambda}(x)}{\sum_{a \in \mathcal{X}} P_1^\lambda(a) P_2^{1-\lambda}(a)}, \tag{11.200}$$

where λ is chosen so that $D(P_{\lambda^*}||P_1) - D(P_{\lambda^*}||P_2) = \frac{1}{n}\log T$.

From the symmetry of expression (11.200), it is clear that $P_1^* = P_2^*$ and that the probabilities of error behave exponentially with exponents given by the relative entropies $D(P^*||P_1)$ and $D(P^*||P_2)$. Also note from the equation that as $\lambda \to 1$, $P_\lambda \to P_1$ and as $\lambda \to 0$, $P_\lambda \to P_2$. The curve that P_λ traces out as λ varies is a geodesic in the simplex. Here P_λ is a normalized convex combination, where the combination is in the exponent (Figure 11.9).

In the next section we calculate the best error exponent when one of the two types of error goes to zero arbitrarily slowly (the Chernoff–Stein lemma). We will also minimize the weighted sum of the two probabilities of error and obtain the Chernoff information bound.

11.8 CHERNOFF–STEIN LEMMA

We consider hypothesis testing in the case when one of the probabilities of error is held fixed and the other is made as small as possible. We will show that the other probability of error is exponentially small, with an exponential rate equal to the relative entropy between the two distributions. The method of proof uses a relative entropy version of the AEP.

Theorem 11.8.1 (*AEP for relative entropy*) *Let $X_1, X_2, \ldots, X_n$ be a sequence of random variables drawn i.i.d. according to $P_1(x)$, and let $P_2(x)$ be any other distribution on $\mathcal{X}$. Then*

$$\frac{1}{n}\log\frac{P_1(X_1, X_2, \ldots, X_n)}{P_2(X_1, X_2, \ldots, X_n)} \to D(P_1||P_2) \qquad \textit{in probability.} \tag{11.201}$$

Proof: This follows directly from the weak law of large numbers.

$$\frac{1}{n}\log\frac{P_1(X_1, X_2, \ldots, X_n)}{P_2(X_1, X_2, \ldots, X_n)} = \frac{1}{n}\log\frac{\prod_{i=1}^n P_1(X_i)}{\prod_{i=1}^n P_2(X_i)} \tag{11.202}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \log \frac{P_1(X_i)}{P_2(X_i)} \tag{11.203}$$

$$\to E_{P_1} \log \frac{P_1(X)}{P_2(X)} \text{ in probability} \tag{11.204}$$

$$= D(P_1||P_2). \qquad \square \tag{11.205}$$

Just as for the regular AEP, we can define a relative entropy typical sequence as one for which the empirical relative entropy is close to its expected value.

Definition For a fixed n and $\epsilon > 0$, a sequence $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ is said to be *relative entropy typical* if and only if

$$D(P_1||P_2) - \epsilon \le \frac{1}{n} \log \frac{P_1(x_1, x_2, \ldots, x_n)}{P_2(x_1, x_2, \ldots, x_n)} \le D(P_1||P_2) + \epsilon. \tag{11.206}$$

The set of relative entropy typical sequences is called the *relative entropy typical set* $A_\epsilon^{(n)}(P_1||P_2)$.

As a consequence of the relative entropy AEP, we can show that the relative entropy typical set satisfies the following properties:

Theorem 11.8.2

1. *For* $(x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)}(P_1||P_2)$,

$$\begin{aligned} &P_1(x_1, x_2, \ldots, x_n)2^{-n(D(P_1||P_2)+\epsilon)} \\ &\quad \le P_2(x_1, x_2, \ldots, x_n) \\ &\quad \le P_1(x_1, x_2, \ldots, x_n)2^{-n(D(P_1||P_2)-\epsilon)}. \end{aligned} \tag{11.207}$$

2. $P_1(A_\epsilon^{(n)}(P_1||P_2)) > 1 - \epsilon$, *for n sufficiently large.*
3. $P_2(A_\epsilon^{(n)}(P_1||P_2)) < 2^{-n(D(P_1||P_2)-\epsilon)}$.
4. $P_2(A_\epsilon^{(n)}(P_1||P_2)) > (1 - \epsilon)2^{-n(D(P_1||P_2)+\epsilon)}$, *for n sufficiently large.*

Proof: The proof follows the same lines as the proof of Theorem 3.1.2, with the counting measure replaced by probability measure P_2. The proof of property 1 follows directly from the definition of the relative entropy typical set. The second property follows from the AEP for relative entropy (Theorem 11.8.1). To prove the third property, we write

$$P_2(A_\epsilon^{(n)}(P_1||P_2)) = \sum_{x^n \in A_\epsilon^{(n)}(P_1||P_2)} P_2(x_1, x_2, \ldots, x_n) \tag{11.208}$$

$$\leq \sum_{x^n \in A_\epsilon^{(n)}(P_1||P_2)} P_1(x_1, x_2, \ldots, x_n) 2^{-n(D(P_1||P_2)-\epsilon)} \tag{11.209}$$

$$= 2^{-n(D(P_1||P_2)-\epsilon)} \sum_{x^n \in A_\epsilon^{(n)}(P_1||P_2)} P_1(x_1, x_2, \ldots, x_n) \tag{11.210}$$

$$= 2^{-n(D(P_1||P_2)-\epsilon)} P_1(A_\epsilon^{(n)}(P_1||P_2)) \tag{11.211}$$

$$\leq 2^{-n(D(P_1||P_2)-\epsilon)}, \tag{11.212}$$

where the first inequality follows from property 1, and the second inequality follows from the fact that the probability of any set under P_1 is less than 1.

To prove the lower bound on the probability of the relative entropy typical set, we use a parallel argument with a lower bound on the probability:

$$P_2(A_\epsilon^{(n)}(P_1||P_2)) = \sum_{x^n \in A_\epsilon^{(n)}(P_1||P_2)} P_2(x_1, x_2, \ldots, x_n) \tag{11.213}$$

$$\geq \sum_{x^n \in A_\epsilon^{(n)}(P_1||P_2)} P_1(x_1, x_2, \ldots, x_n) 2^{-n(D(P_1||P_2)+\epsilon)} \tag{11.214}$$

$$= 2^{-n(D(P_1||P_2)+\epsilon)} \sum_{x^n \in A_\epsilon^{(n)}(P_1||P_2)} P_1(x_1, x_2, \ldots, x_n) \tag{11.215}$$

$$= 2^{-n(D(P_1||P_2)+\epsilon)} P_1(A_\epsilon^{(n)}(P_1||P_2)) \tag{11.216}$$

$$\geq (1-\epsilon) 2^{-n(D(P_1||P_2)+\epsilon)}, \tag{11.217}$$

where the second inequality follows from the second property of $A_\epsilon^{(n)}(P_1||P_2)$. □

With the standard AEP in Chapter 3, we also showed that any set that has a high probability has a high intersection with the typical set, and therefore has about 2^{nH} elements. We now prove the corresponding result for relative entropy.

Lemma 11.8.1 *Let $B_n \subset \mathcal{X}^n$ be any set of sequences $x_1, x_2, \ldots, x_n$ such that $P_1(B_n) > 1 - \epsilon$. Let P_2 be any other distribution such that $D(P_1||P_2) < \infty$. Then $P_2(B_n) > (1 - 2\epsilon)2^{-n(D(P_1||P_2)+\epsilon)}$.*

Proof: For simplicity, we will denote $A_\epsilon^{(n)}(P_1||P_2)$ by A_n. Since $P_1(B_n) > 1 - \epsilon$ and $P(A_n) > 1 - \epsilon$ (Theorem 11.8.2), we have, by the union of events bound, $P_1(A_n^c \cup B_n^c) < 2\epsilon$, or equivalently, $P_1(A_n \cap B_n) > 1 - 2\epsilon$. Thus,

$$P_2(B_n) \geq P_2(A_n \cap B_n) \tag{11.218}$$

$$= \sum_{x^n \in A_n \cap B_n} P_2(x^n) \tag{11.219}$$

$$\geq \sum_{x^n \in A_n \cap B_n} P_1(x^n) 2^{-n(D(P_1||P_2)+\epsilon)} \tag{11.220}$$

$$= 2^{-n(D(P_1||P_2)+\epsilon)} \sum_{x^n \in A_n \cap B_n} P_1(x^n) \tag{11.221}$$

$$= 2^{-n(D(P_1||P_2)+\epsilon)} P_1(A_n \cap B_n) \tag{11.222}$$

$$\geq 2^{-n(D(P_1||P_2)+\epsilon)}(1 - 2\epsilon), \tag{11.223}$$

where the second inequality follows from the properties of the relative entropy typical sequences (Theorem 11.8.2) and the last inequality follows from the union bound above. □

We now consider the problem of testing two hypotheses, P_1 vs. P_2. We hold one of the probabilities of error fixed and attempt to minimize the other probability of error. We show that the relative entropy is the best exponent in probability of error.

Theorem 11.8.3 (*Chernoff–Stein Lemma*) *Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim Q$. Consider the hypothesis test between two alternatives, $Q = P_1$ and $Q = P_2$, where $D(P_1||P_2) < \infty$. Let $A_n \subseteq \mathcal{X}^n$ be an acceptance region for hypothesis H_1. Let the probabilities of error be*

$$\alpha_n = P_1^n(A_n^c), \qquad \beta_n = P_2^n(A_n). \tag{11.224}$$

and for $0 < \epsilon < \frac{1}{2}$, define

$$\beta_n^\epsilon = \min_{\substack{A_n \subseteq \mathcal{X}^n \\ \alpha_n < \epsilon}} \beta_n. \tag{11.225}$$

Then

$$\lim_{n\to\infty} \frac{1}{n} \log \beta_n^{\epsilon} = -D(P_1||P_2). \tag{11.226}$$

Proof: We prove this theorem in two parts. In the first part we exhibit a sequence of sets A_n for which the probability of error β_n goes exponentially to zero as $D(P_1||P_2)$. In the second part we show that no other sequence of sets can have a lower exponent in the probability of error.

For the first part, we choose as the sets $A_n = A_\epsilon^{(n)}(P_1||P_2)$. As proved in Theorem 11.8.2, this sequence of sets has $P_1(A_n^c) < \epsilon$ for n large enough. Also,

$$\lim_{n\to\infty} \frac{1}{n} \log P_2(A_n) \le -(D(P_1||P_2) - \epsilon) \tag{11.227}$$

from property 3 of Theorem 11.8.2. Thus, the relative entropy typical set satisfies the bounds of the lemma.

To show that no other sequence of sets can to better, consider any sequence of sets B_n with $P_1(B_n) > 1 - \epsilon$. By Lemma 11.8.1, we have $P_2(B_n) > (1 - 2\epsilon)2^{-n(D(P_1||P_2)+\epsilon)}$, and therefore

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \log P_2(B_n) &> -(D(P_1||P_2) + \epsilon) + \lim_{n\to\infty} \frac{1}{n} \log(1 - 2\epsilon) \\ &= -(D(P_1||P_2) + \epsilon). \end{aligned} \tag{11.228}$$

Thus, no other sequence of sets has a probability of error exponent better than $D(P_1||P_2)$. Thus, the set sequence $A_n = A_\epsilon^{(n)}(P_1||P_2)$ is asymptotically optimal in terms of the exponent in the probability. □

Not that the relative entropy typical set, although asymptotically optimal (i.e., achieving the best asymptotic rate), is not the optimal set for any fixed hypothesis-testing problem. The optimal set that minimizes the probabilities of error is that given by the Neyman–Pearson lemma.

11.9 CHERNOFF INFORMATION

We have considered the problem of hypothesis testing in the classical setting, in which we treat the two probabilities of error separately. In the derivation of the Chernoff–Stein lemma, we set $\alpha_n \le \epsilon$ and achieved $\beta_n \doteq 2^{-nD}$. But this approach lacks symmetry. Instead, we can follow a Bayesian approach, in which we assign prior probabilities to both hypotheses. In this case we wish to minimize the overall probability of error given by the weighted sum of the individual probabilities of error. The resulting error exponent is the *Chernoff information*.

The setup is as follows: $X_1, X_2, \ldots, X_n$ i.i.d. $\sim Q$. We have two hypotheses: $Q = P_1$ with prior probability π_1 and $Q = P_2$ with prior probability π_2. The overall probability of error is

$$P_e^{(n)} = \pi_1 \alpha_n + \pi_2 \beta_n. \tag{11.229}$$

Let

$$D^* = \lim_{n\to\infty} -\frac{1}{n} \log \min_{A_n \subseteq \mathcal{X}^n} P_e^{(n)}. \tag{11.230}$$

Theorem 11.9.1 (*Chernoff*) *The best achievable exponent in the Bayesian probability of error is D^*, where*

$$D^* = D(P_{\lambda^*}||P_1) = D(P_{\lambda^*}||P_2), \tag{11.231}$$

with

$$P_\lambda = \frac{P_1^\lambda(x) P_2^{1-\lambda}(x)}{\sum_{a\in\mathcal{X}} P_1^\lambda(a) P_2^{1-\lambda}(a)}, \tag{11.232}$$

and λ^* the value of λ such that

$$D(P_{\lambda^*}||P_1) = D(P_{\lambda^*}||P_2). \tag{11.233}$$

Proof: The basic details of the proof were given in Section 11.8. We have shown that the optimum test is a likelihood ratio test, which can be considered to be of the form

$$D(P_{X^n}||P_2) - D(P_{X^n}||P_1) > \frac{1}{n} \log T. \tag{11.234}$$

The test divides the probability simplex into regions corresponding to hypothesis 1 and hypothesis 2, respectively. This is illustrated in Figure 11.10.

Let A be the set of types associated with hypothesis 1. From the discussion preceding (11.200), it follows that the closest point in the set A^c to P_1 is on the boundary of A and is of the form given by (11.232). Then from the discussion in Section 11.8, it is clear that P_λ is the distribution

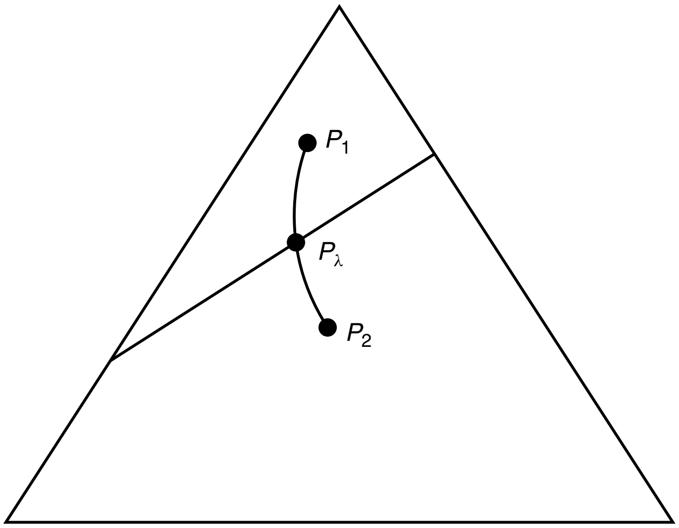

FIGURE 11.10. Probability simplex and Chernoff information.

in A that is closest to P_2; it is also the distribution in A^c that is closest to P_1. By Sanov's theorem, we can calculate the associated probabilities of error,

$$\alpha_n = P_1^n(A^c) \doteq 2^{-nD(P_{\lambda^*}||P_1)} \tag{11.235}$$

and

$$\beta_n = P_2^n(A) \doteq 2^{-nD(P_{\lambda^*}||P_2)}. \tag{11.236}$$

In the Bayesian case, the overall probability of error is the weighted sum of the two probabilities of error,

$$P_e \doteq \pi_1 2^{-nD(P_\lambda||P_1)} + \pi_2 2^{-nD(P_\lambda||P_2)} \doteq 2^{-n\min\{D(P_\lambda||P_1), D(P_\lambda||P_2)\}}, \tag{11.237}$$

since the exponential rate is determined by the worst exponent. Since $D(P_\lambda||P_1)$ increases with λ and $D(P_\lambda||P_2)$ decreases with λ, the maximum value of the minimum of $\{D(P_\lambda||P_1), D(P_\lambda||P_2)\}$ is attained when they are equal. This is illustrated in Figure 11.11. Hence, we choose λ so that

$$D(P_\lambda||P_1) = D(P_\lambda||P_2). \tag{11.238}$$

Thus, $C(P_1, P_2)$ is the highest achievable exponent for the probability of error and is called the Chernoff information. □

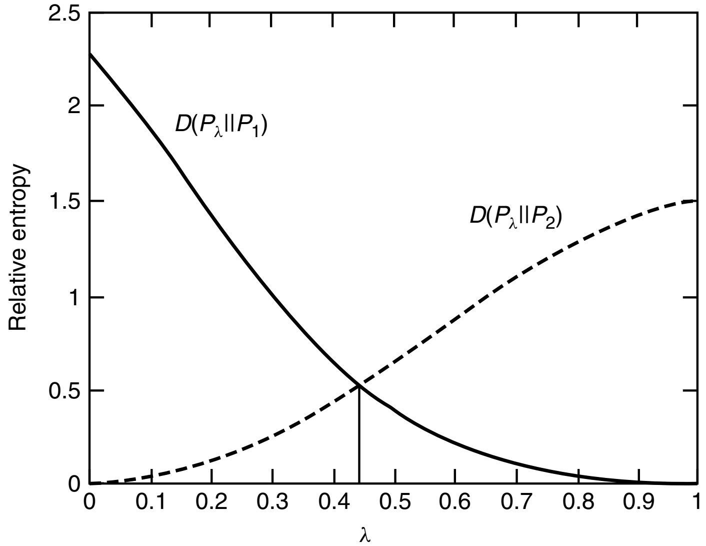

FIGURE 11.11. Relative entropy $D(P_\lambda||P_1)$ and $D(P_\lambda||P_2)$ as a function of λ.

The definition $D^* = D(P_{\lambda^*}||P_1) = D(P_{\lambda^*}||P_2)$ is equivalent to the standard definition of *Chernoff information*,

$$C(P_1, P_2) \overset{\triangle}{=} -\min_{0\le\lambda\le 1} \log\left(\sum_x P_1^\lambda(x) P_2^{1-\lambda}(x)\right). \tag{11.239}$$

It is left as an exercise to the reader to show the equivalence of (11.231) and (11.239).

We outline briefly the usual derivation of the Chernoff information bound. The maximum a posteriori probability decision rule minimizes the Bayesian probability of error. The decision region A for hypothesis H_1 for the maximum a posteriori rule is

$$A = \left\{\mathbf{x} : \frac{\pi_1 P_1(\mathbf{x})}{\pi_2 P_2(\mathbf{x})} > 1\right\}, \tag{11.240}$$

the set of outcomes where the a posteriori probability of hypothesis H_1 is greater than the a posteriori probability of hypothesis H_2. The probability of error for this rule is

$$P_e = \pi_1 \alpha_n + \pi_2 \beta_n \tag{11.241}$$

$$= \sum_{A^c} \pi_1 P_1 + \sum_A \pi_2 P_2 \tag{11.242}$$

$$= \sum \min\{\pi_1 P_1, \pi_2 P_2\}. \tag{11.243}$$

Now for any two positive numbers a and b, we have

$$\min\{a, b\} \le a^\lambda b^{1-\lambda} \quad \text{for all } 0 \le \lambda \le 1. \tag{11.244}$$

Using this to continue the chain, we have

$$\begin{aligned} P_e &= \sum \min\{\pi_1 P_1, \pi_2 P_2\} && (11.245)\\ &\le \sum (\pi_1 P_1)^\lambda (\pi_2 P_2)^{1-\lambda} && (11.246)\\ &\le \sum P_1^\lambda P_2^{1-\lambda}. && (11.247) \end{aligned}$$

For a sequence of i.i.d. observations, $P_k(\mathbf{x}) = \prod_{i=1}^n P_k(x_i)$, and

$$\begin{aligned} P_e^{(n)} &\le \sum \pi_1^\lambda \pi_2^{1-\lambda} \prod_i P_1^\lambda(x_i) P_2^{1-\lambda}(x_i) && (11.248)\\ &= \pi_1^\lambda \pi_2^{1-\lambda} \prod_i \sum P_1^\lambda(x_i) P_2^{1-\lambda}(x_i) && (11.249)\\ &\le \prod_{x_i} \sum P_1^\lambda P_2^{1-\lambda} && (11.250)\\ &= \left(\sum_x P_1^\lambda P_2^{1-\lambda} \right)^n, && (11.251) \end{aligned}$$

where (11.250) follows since $\pi_1 \le 1, \pi_2 \le 1$. Hence, we have

$$\frac{1}{n} \log P_e^{(n)} \le \log \sum P_1^\lambda(x) P_2^{1-\lambda}(x). \tag{11.252}$$

Since this is true for all λ, we can take the minimum over $0 \le \lambda \le 1$, resulting in the Chernoff information bound. This proves that the exponent is no better than $C(P_1, P_2)$. Achievability follows from Theorem 11.9.1.

Note that the Bayesian error exponent does not depend on the actual value of π_1 and π_2, as long as they are nonzero. Essentially, the effect of the prior is washed out for large sample sizes. The optimum decision rule is to choose the hypothesis with the maximum a posteriori probability, which corresponds to the test

$$\frac{\pi_1 P_1(X_1, X_2, \ldots, X_n)}{\pi_2 P_2(X_1, X_2, \ldots, X_n)} \gtrless 1. \tag{11.253}$$

Taking the log and dividing by n, this test can be rewritten as

$$\frac{1}{n}\log\frac{\pi_1}{\pi_2}+\frac{1}{n}\sum_i \log\frac{P_1(X_i)}{P_2(X_i)} \lessgtr 0, \tag{11.254}$$

where the second term tends to $D(P_1||P_2)$ or $-D(P_2||P_1)$ accordingly as P_1 or P_2 is the true distribution. The first term tends to 0, and the effect of the prior distribution washes out.

Finally, to round off our discussion of large deviation theory and hypothesis testing, we consider an example of the conditional limit theorem.

Example 11.9.1 Suppose that major league baseball players have a batting average of 260 with a standard deviation of 15 and suppose that minor league ballplayers have a batting average of 240 with a standard deviation of 15. A group of 100 ballplayers from one of the leagues (the league is chosen at random) are found to have a group batting average greater than 250 and are therefore judged to be major leaguers. We are now told that we are mistaken; these players are minor leaguers. What can we say about the distribution of batting averages among these 100 players? The conditional limit theorem can be used to show that the distribution of batting averages among these players will have a mean of 250 and a standard deviation of 15. To see this, we abstract the problem as follows.

Let us consider an example of testing between two Gaussian distributions, $f_1 = \mathcal{N}(1, \sigma^2)$ and $f_2 = \mathcal{N}(-1, \sigma^2)$, with different means and the same variance. As discussed in Section 11.8, the likelihood ratio test in this case is equivalent to comparing the sample mean with a threshold. The Bayes test is "Accept the hypothesis $f = f_1$ if $\frac{1}{n}\sum_{i=1}^n X_i > 0$." Now assume that we make an error of the first kind (we say that $f = f_1$ when indeed $f = f_2$) in this test. What is the conditional distribution of the samples given that we have made an error?

We might guess at various possibilities:

- The sample will look like a $(\frac{1}{2}, \frac{1}{2})$ mix of the two normal distributions. Plausible as this is, it is incorrect.
- $X_i \approx 0$ for all i. This is quite clearly very unlikely, although it is conditionally likely that $\overline{X}_n$ is close to 0.
- The correct answer is given by the conditional limit theorem. If the true distribution is f_2 and the sample type is in the set A, the conditional distribution is close to f^*, the distribution in A that is closest to f_2. By symmetry, this corresponds to $\lambda = \frac{1}{2}$ in (11.232). Calculating

the distribution, we get

$$f^*(x) = \frac{\left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-1)^2}{2\sigma^2}}\right)^{\frac{1}{2}}\left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x+1)^2}{2\sigma^2}}\right)^{\frac{1}{2}}}{\int\left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-1)^2}{2\sigma^2}}\right)^{\frac{1}{2}}\left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x+1)^2}{2\sigma^2}}\right)^{\frac{1}{2}}dx} \tag{11.255}$$

$$= \frac{\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x^2+1)}{2\sigma^2}}}{\int\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x^2+1)}{2\sigma^2}}dx} \tag{11.256}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{x^2}{2\sigma^2}} \tag{11.257}$$

$$= \mathcal{N}(0,\sigma^2). \tag{11.258}$$

It is interesting to note that the conditional distribution is normal with mean 0 and with the same variance as the original distributions. This is strange but true; if we mistake a normal population for another, the "shape" of this population still looks normal with the same variance and a different mean. Apparently, this rare event does not result from bizarre-looking data.

Example 11.9.2 (*Large deviation theory and football*) Consider a very simple version of football in which the score is directly related to the number of yards gained. Assume that the coach has a choice between two strategies: running or passing. Associated with each strategy is a distribution on the number of yards gained. For example, in general, running

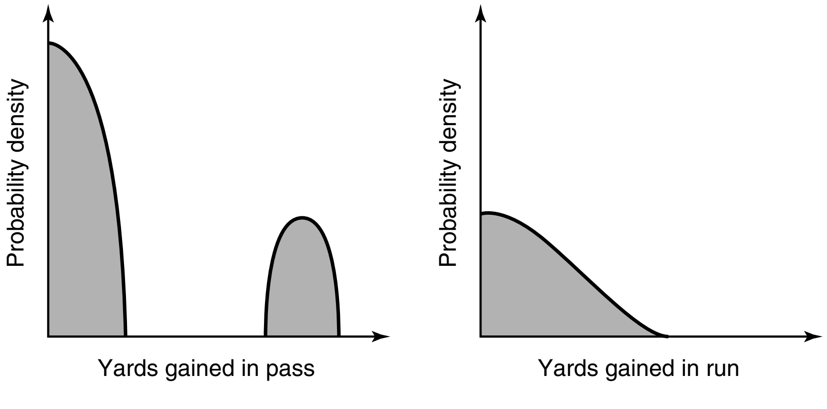

FIGURE 11.12. Distribution of yards gained in a run or a pass play.

results in a gain of a few yards with very high probability, whereas passing results in huge gains with low probability. Examples of the distributions are illustrated in Figure 11.12.

At the beginning of the game, the coach uses the strategy that promises the greatest expected gain. Now assume that we are in the closing minutes of the game and one of the teams is leading by a large margin. (Let us ignore first downs and adaptable defenses.) So the trailing team will win only if it is very lucky. If luck is required to win, we might as well assume that we will be lucky and play accordingly. What is the appropriate strategy?

Assume that the team has only n plays left and it must gain l yards, where l is much larger than n times the expected gain under each play. The probability that the team succeeds in achieving l yards is exponentially small; hence, we can use the large deviation results and Sanov's theorem to calculate the probability of this event. To be precise, we wish to calculate the probability that $\sum_{i=1}^{n} Z_i \geq n\alpha$, where Z_i are independent random variables and Z_i has a distribution corresponding to the strategy chosen.

The situation is illustrated in Figure 11.13. Let E be the set of types corresponding to the constraint,

$$E = \left\{ P : \sum_{a \in \mathcal{X}} P(a)a \geq \alpha \right\}. \tag{11.259}$$

If P_1 is the distribution corresponding to passing all the time, the probability of winning is the probability that the sample type is in E, which by Sanov's theorem is $2^{-nD(P_1^*||P_1)}$, where P_1^* is the distribution in E that is closest to P_1. Similarly, if the coach uses the running game all the time,

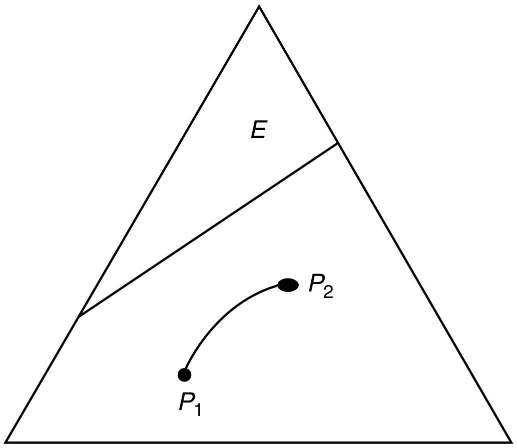

FIGURE 11.13. Probability simplex for a football game.

the probability of winning is $2^{-nD(P_2^*||P_2)}$. What if he uses a mixture of strategies? Is it possible that $2^{-nD(P_\lambda^*||P_\lambda)}$, the probability of winning with a mixed strategy, $P_\lambda = \lambda P_1 + (1-\lambda)P_2$, is better than the probability of winning with either pure passing or pure running? The somewhat surprising answer is yes, as can be shown by example. This provides a reason to use a mixed strategy other than the fact that it confuses the defense.

We end this section with another inequality due to Chernoff, which is a special version of Markov's inequality. This inequality is called the *Chernoff bound.*

Lemma 11.9.1 *Let Y be any random variable and let $\psi(s)$ be the moment generating function of Y,*

$$\psi(s) = Ee^{sY}. \tag{11.260}$$

Then for all $s \geq 0$,

$$\Pr(Y \geq a) \leq e^{-sa}\psi(s), \tag{11.261}$$

and thus

$$\Pr(Y \geq a) \leq \min_{s\geq 0} e^{-sa}\psi(s). \tag{11.262}$$

Proof: Apply Markov's inequality to the nonnegative random variable e^{sY}. □

11.10 FISHER INFORMATION AND THE CRAMÉR–RAO INEQUALITY

A standard problem in statistical estimation is to determine the parameters of a distribution from a sample of data drawn from that distribution. For example, let $X_1, X_2, \ldots, X_n$ be drawn i.i.d. $\sim \mathcal{N}(\theta, 1)$. Suppose that we wish to estimate θ from a sample of size n. There are a number of functions of the data that we can use to estimate θ. For example, we can use the first sample X_1. Although the expected value of X_1 is θ, it is clear that we can do better by using more of the data. We guess that the best estimate of θ is the sample mean $\overline{X}_n = \frac{1}{n}\sum X_i$. Indeed, it can be shown that $\overline{X}_n$ is the minimum mean-squared-error unbiased estimator.

We begin with a few definitions. Let $\{f(x;\theta)\}$, $\theta \in \Theta$, denote an indexed family of densities, $f(x;\theta) \geq 0$, $\int f(x;\theta)\,dx = 1$ for all $\theta \in \Theta$. Here Θ is called the *parameter set.*

Definition An *estimator* for θ for sample size n is a function $T : \mathcal{X}^n \to \Theta$.

An estimator is meant to approximate the value of the parameter. It is therefore desirable to have some idea of the goodness of the approximation. We will call the difference $T - \theta$ the *error* of the estimator. The error is a random variable.

Definition The *bias* of an estimator $T(X_1, X_2, \ldots, X_n)$ for the parameter θ is the expected value of the error of the estimator [i.e., the bias is $E_\theta T(x_1, x_2, \ldots, x_n) - \theta$]. The subscript θ means that the expectation is with respect to the density $f(\cdot; \theta)$. The estimator is said to be *unbiased* if the bias is zero for all $\theta \in \Theta$ (i.e., the expected value of the estimator is equal to the parameter).

Example 11.10.1 Let $X_1, X_2, \ldots, X_n$ drawn i.i.d. $\sim f(x) = (1/\lambda)$ $e^{-x/\lambda}$, $x \geq 0$ be a sequence of exponentially distributed random variables. Estimators of λ include X_1 and $\overline{X}_n$. Both estimators are unbiased.

The bias is the expected value of the error, and the fact that it is zero does not guarantee that the error is low with high probability. We need to look at some loss function of the error; the most commonly chosen loss function is the expected square of the error. A good estimator should have a low expected squared error and should have an error that approaches 0 as the sample size goes to infinity. This motivates the following definition:

Definition An estimator $T(X_1, X_2, \ldots, X_n)$ for θ is said to be *consistent in probability* if
$T(X_1, X_2, \ldots, X_n) \to \theta$ in probability as $n \to \infty$.

Consistency is a desirable asymptotic property, but we are interested in the behavior for small sample sizes as well. We can then rank estimators on the basis of their mean-squared error.

Definition An estimator $T_1(X_1, X_2, \ldots, X_n)$ is said to *dominate* another estimator $T_2(X_1, X_2, \ldots, X_n)$ if, for all θ,

$$E\left(T_1(X_1, X_2, \ldots, X_n) - \theta\right)^2 \leq E\left(T_2(X_1, X_2, \ldots, X_n) - \theta\right)^2. \tag{11.263}$$

This raises a natural question: Is there a best estimator of θ that dominates every other estimator? To answer this question, we derive the Cramér–Rao lower bound on the mean-squared error of any estimator. We first define the score function of the distribution $f(x; \theta)$. We then use the Cauchy–Schwarz inequality to prove the Cramér–Rao lower bound on the variance of all unbiased estimators.

Definition The *score* V is a random variable defined by

$$V = \frac{\partial}{\partial\theta} \ln f(X;\theta) = \frac{\frac{\partial}{\partial\theta} f(X;\theta)}{f(X;\theta)}, \tag{11.264}$$

where $X \sim f(x;\theta)$.

The mean value of the score is

$$\begin{aligned} EV &= \int \frac{\frac{\partial}{\partial\theta} f(x;\theta)}{f(x;\theta)} f(x;\theta)\, dx && (11.265)\\ &= \int \frac{\partial}{\partial\theta} f(x;\theta)\, dx && (11.266)\\ &= \frac{\partial}{\partial\theta} \int f(x;\theta)\, dx && (11.267)\\ &= \frac{\partial}{\partial\theta} 1 && (11.268)\\ &= 0, && (11.269)\end{aligned}$$

and therefore $EV^2 = \operatorname{var}(V)$. The variance of the score has a special significance.

Definition The *Fisher information* $J(\theta)$ is the variance of the score:

$$J(\theta) = E_\theta \left[\frac{\partial}{\partial\theta} \ln f(X;\theta) \right]^2. \tag{11.270}$$

If we consider a sample of n random variables $X_1, X_2, \ldots, X_n$ drawn i.i.d. $\sim f(x;\theta)$, we have

$$f(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i;\theta), \tag{11.271}$$

and the score function is the sum of the individual score functions,

$$\begin{aligned} V(X_1, X_2, \ldots, X_n) &= \frac{\partial}{\partial\theta} \ln f(X_1, X_2, \ldots, X_n; \theta) && (11.272)\\ &= \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \ln f(X_i;\theta) && (11.273)\\ &= \sum_{i=1}^{n} V(X_i), && (11.274)\end{aligned}$$

where the $V(X_i)$ are independent, identically distributed with zero mean. Hence, the n-sample Fisher information is

$$J_n(\theta) = E_\theta\left[\frac{\partial}{\partial\theta}\ln f(X_1, X_2, \ldots, X_n; \theta)\right]^2 \tag{11.275}$$

$$= E_\theta V^2(X_1, X_2, \ldots, X_n) \tag{11.276}$$

$$= E_\theta\left(\sum_{i=1}^{n} V(X_i)\right)^2 \tag{11.277}$$

$$= \sum_{i=1}^{n} E_\theta V^2(X_i) \tag{11.278}$$

$$= nJ(\theta). \tag{11.279}$$

Consequently, the Fisher information for n i.i.d. samples is n times the individual Fisher information. The significance of the Fisher information is shown in the following theorem.

Theorem 11.10.1 (*Cramér–Rao inequality*) *The mean-squared error of any unbiased estimator $T(X)$ of the parameter θ is lower bounded by the reciprocal of the Fisher information:*

$$\mathrm{var}(T) \ge \frac{1}{J(\theta)}. \tag{11.280}$$

Proof: Let V be the score function and T be the estimator. By the Cauchy–Schwarz inequality, we have

$$(E_\theta[(V - E_\theta V)(T - E_\theta T)])^2 \le E_\theta(V - E_\theta V)^2 E_\theta(T - E_\theta T)^2. \tag{11.281}$$

Since T is unbiased, $E_\theta T = \theta$ for all θ. By (11.269), $E_\theta V = 0$ and hence $E_\theta(V - E_\theta V)(T - E_\theta T) = E_\theta(VT)$. Also, by definition, $\mathrm{var}(V) = J(\theta)$. Substituting these conditions in (11.281), we have

$$[E_\theta(VT)]^2 \le J(\theta)\mathrm{var}(T). \tag{11.282}$$

Now,

$$E_\theta(VT) = \int \frac{\frac{\partial}{\partial\theta} f(x;\theta)}{f(x;\theta)} T(x) f(x;\theta)\, dx \tag{11.283}$$

$$= \int \frac{\partial}{\partial \theta} f(x;\theta) T(x)\, dx \tag{11.284}$$

$$= \frac{\partial}{\partial \theta} \int f(x;\theta) T(x)\, dx \tag{11.285}$$

$$= \frac{\partial}{\partial \theta} E_\theta T \tag{11.286}$$

$$= \frac{\partial}{\partial \theta} \theta \tag{11.287}$$

$$= 1, \tag{11.288}$$

where the interchange of differentiation and integration in (11.285) can be justified using the bounded convergence theorem for appropriately well behaved $f(x;\theta)$, and (11.287) follows from the fact that the estimator T is unbiased. Substituting this in (11.282), we obtain

$$\mathrm{var}(T) \ge \frac{1}{J(\theta)}, \tag{11.289}$$

which is the Cramér–Rao inequality for unbiased estimators. □

By essentially the same arguments, we can show that for any estimator

$$E(T-\theta)^2 \ge \frac{(1+b_T'(\theta))^2}{J(\theta)} + b_T^2(\theta), \tag{11.290}$$

where $b_T(\theta) = E_\theta T - \theta$ and $b_T'(\theta)$ is the derivative of $b_T(\theta)$ with respect to θ. The proof of this is left as a problem at the end of the chapter.

Example 11.10.2 Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim \mathcal{N}(\theta, \sigma^2)$, σ^2 known. Here $J(\theta) = n/\sigma^2$. Let $T(X_1, X_2, \ldots, X_n) = \overline{X}_n = \frac{1}{n}\sum X_i$. Then $E_\theta(\overline{X}_n - \theta)^2 = \sigma^2/n = 1/J(\theta)$. Thus, $\overline{X}_n$ is the minimum variance unbiased estimator of θ, since it achieves the Cramér–Rao lower bound.

The Cramér–Rao inequality gives us a lower bound on the variance for all unbiased estimators. When this bound is achieved, we call the estimator efficient.

Definition An unbiased estimator T is said to be *efficient* if it meets the Cramér–Rao bound with equality [i.e., if $\mathrm{var}(T) = \frac{1}{J(\theta)}$].

The Fisher information is therefore a measure of the amount of "information" about θ that is present in the data. It gives a lower bound on the error in estimating θ from the data. However, it is possible that there does not exist an estimator meeting this lower bound.

We can generalize the concept of Fisher information to the multiparameter case, in which case we define the Fisher information matrix $J(\theta)$ with elements

$$J_{ij}(\theta) = \int f(x;\theta) \frac{\partial}{\partial\theta_i} \ln f(x;\theta) \frac{\partial}{\partial\theta_j} \ln f(x;\theta) \; dx. \tag{11.291}$$

The Cramér–Rao inequality becomes the matrix inequality

$$\Sigma \geq J^{-1}(\theta), \tag{11.292}$$

where Σ is the covariance matrix of a set of unbiased estimators for the parameters θ and $\Sigma \geq J^{-1}(\theta)$ in the sense that the difference $\Sigma - J^{-1}$ is a nonnegative definite matrix. We will not go into the details of the proof for multiple parameters; the basic ideas are similar.

Is there a relationship between the Fisher information $J(\theta)$ and quantities such as entropy defined earlier? Note that Fisher information is defined with respect to a family of parametric distributions, unlike entropy, which is defined for all distributions. But we can parametrize any distribution $f(x)$ by a location parameter θ and define Fisher information with respect to the family of densities $f(x-\theta)$ under translation. We explore the relationship in greater detail in Section 17.8, where we show that while entropy is related to the volume of the typical set, the Fisher information is related to the surface area of the typical set. Further relationships of Fisher information to relative entropy are developed in the problems.

SUMMARY

Basic identities

$$Q^n(\mathbf{x}) = 2^{-n(D(P_{\mathbf{x}}||Q)+H(P_{\mathbf{x}}))}, \tag{11.293}$$

$$|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}|}, \tag{11.294}$$

$$|T(P)| \doteq 2^{nH(P)}, \tag{11.295}$$

$$Q^n(T(P)) \doteq 2^{-nD(P||Q)}. \tag{11.296}$$

Universal data compression

$$P_e^{(n)} \le 2^{-nD(P_R^*||Q)} \quad \text{for all } Q, \tag{11.297}$$

where

$$D(P_R^*||Q) = \min_{P:H(P)\ge R} D(P||Q). \tag{11.298}$$

Large deviations (Sanov's theorem)

$$Q^n(E) = Q^n(E \cap \mathcal{P}_n) \le (n+1)^{|\mathcal{X}|} 2^{-nD(P^*||Q)}, \tag{11.299}$$

$$D(P^*||Q) = \min_{P\in E} D(P||Q). \tag{11.300}$$

If E is the closure of its interior, then

$$Q^n(E) \doteq 2^{-nD(P^*||Q)}. \tag{11.301}$$

$\mathcal{L}_1$ bound on relative entropy

$$D(P_1||P_2) \ge \frac{1}{2\ln 2}||P_1 - P_2||_1^2. \tag{11.302}$$

Pythagorean theorem. If E is a convex set of types, distribution $Q \notin E$, and P^* achieves $D(P^*||Q) = \min_{P\in E} D(P||Q)$, we have

$$D(P||Q) \ge D(P||P^*) + D(P^*||Q) \tag{11.303}$$

for all $P \in E$.

Conditional limit theorem. If $X_1, X_2, \ldots, X_n$ i.i.d. $\sim Q$, then

$$\Pr(X_1 = a|P_{X^n} \in E) \to P^*(a) \quad \text{in probability}, \tag{11.304}$$

where P^* minimizes $D(P||Q)$ over $P \in E$. In particular,

$$\Pr\left\{X_1 = a \left| \frac{1}{n}\sum_{i=1}^n X_i \ge \alpha \right.\right\} \to \frac{Q(a)e^{\lambda a}}{\sum_x Q(x)e^{\lambda x}}. \tag{11.305}$$

Neyman–Pearson lemma. The optimum test between two densities P_1 and P_2 has a decision region of the form "accept $P = P_1$ if $\frac{P_1(x_1,x_2,\ldots,x_n)}{P_2(x_1,x_2,\ldots,x_n)} > T$."

Chernoff–Stein lemma. The best achievable error exponent β_n^ϵ if $\alpha_n \le \epsilon$:

$$\beta_n^\epsilon = \min_{\substack{A_n \subseteq \mathcal{X}^n \\ \alpha_n < \epsilon}} \beta_n, \tag{11.306}$$

$$\lim_{n\to\infty} \frac{1}{n} \log \beta_n^\epsilon = -D(P_1||P_2). \tag{11.307}$$

Chernoff information. The best achievable exponent for a Bayesian probability of error is

$$D^* = D(P_{\lambda^*}||P_1) = D(P_{\lambda^*}||P_2), \tag{11.308}$$

where

$$P_\lambda = \frac{P_1^\lambda(x) P_2^{1-\lambda}(x)}{\sum_{a\in\mathcal{X}} P_1^\lambda(a) P_2^{1-\lambda}(a)} \tag{11.309}$$

with $\lambda = \lambda^*$ chosen so that

$$D(P_\lambda||P_1) = D(P_\lambda||P_2). \tag{11.310}$$

Fisher information

$$J(\theta) = E_\theta \left[\frac{\partial}{\partial\theta} \ln f(x;\theta) \right]^2. \tag{11.311}$$

Cramér–Rao inequality. For any unbiased estimator T of θ,

$$E_\theta (T(X) - \theta)^2 = \text{var}(T) \ge \frac{1}{J(\theta)}. \tag{11.312}$$

PROBLEMS

11.1 *Chernoff–Stein lemma.* Consider the two-hypothesis test

$$H_1 : f = f_1 \quad \text{vs.} \quad H_2 : f = f_2.$$

Find $D(f_1 \parallel f_2)$ if

(a) $f_i(x) = N(0, \sigma_i^2), i = 1, 2.$

(b) $f_i(x) = \lambda_i e^{-\lambda_i x}, x \geq 0, i = 1, 2.$

(c) $f_1(x)$ is the uniform density over the interval [0, 1] and $f_2(x)$ is the uniform density over $[a, a + 1]$. Assume that $0 < a < 1$.

(d) f_1 corresponds to a fair coin and f_2 corresponds to a two-headed coin.

11.2 *Relation between $D(P \parallel Q)$ and chi-square.* Show that the χ^2 statistic

$$\chi^2 = \Sigma_x \frac{(P(x) - Q(x))^2}{Q(x)}$$

is (twice) the first term in the Taylor series expansion of $D(P \parallel Q)$ about Q. Thus, $D(P \parallel Q) = \frac{1}{2}\chi^2 + \cdots$. [*Suggestion:* Write $\frac{P}{Q} = 1 + \frac{P-Q}{Q}$ and expand the log.]

11.3 *Error exponent for universal codes.* A universal source code of rate R achieves a probability of error $P_e^{(n)} \doteq e^{-nD(P^* \| Q)}$, where Q is the true distribution and P^* achieves $\min D(P \parallel Q)$ over all P such that $H(P) \geq R$.

(a) Find P^* in terms of Q and R.

(b) Now let X be binary. Find the region of source probabilities $Q(x)$, $x \in \{0, 1\}$, for which rate R is sufficient for the universal source code to achieve $P_e^{(n)} \rightarrow 0$.

11.4 *Sequential projection.* We wish to show that projecting Q onto P_1 and then projecting the projection $\hat{Q}$ onto $P_1 \bigcap P_2$ is the same as projecting Q directly onto $P_1 \bigcap P_2$. Let $\mathcal{P}_1$ be the set of probability mass functions on $\mathcal{X}$ satisfying

$$\sum_x p(x) = 1, \tag{11.313}$$

$$\sum_x p(x) h_i(x) \geq \alpha_i, \qquad i = 1, 2, \ldots, r. \tag{11.314}$$

Let $\mathcal{P}_2$ be the set of probability mass functions on $\mathcal{X}$ satisfying

$$\sum_x p(x) = 1, \tag{11.315}$$

$$\sum_x p(x) g_j(x) \geq \beta_j, \qquad j = 1, 2, \ldots, s. \tag{11.316}$$

Suppose that $Q \notin P_1 \bigcup P_2$. Let P^* minimize $D(P \parallel Q)$ over all $P \in \mathcal{P}_1$. Let R^* minimize $D(R \parallel Q)$ over all $R \in \mathcal{P}_1 \bigcap \mathcal{P}_2$. Argue that R^* minimizes $D(R \parallel P^*)$ over all $R \in P_1 \bigcap P_2$.

11.5 *Counting*. Let $\mathcal{X} = \{1, 2, \ldots, m\}$. Show that the number of sequences $x^n \in \mathcal{X}^n$ satisfying $\frac{1}{n}\sum_{i=1}^n g(x_i) \geq \alpha$ is approximately equal to 2^{nH^*}, to first order in the exponent, for n sufficiently large, where

$$H^* = \max_{P:\sum_{i=1}^m P(i)g(i) \geq \alpha} H(P). \tag{11.317}$$

11.6 *Biased estimates may be better*. Consider the problem of estimating μ and σ^2 from n samples of data drawn i.i.d. from a $\mathcal{N}(\mu, \sigma^2)$ distribution.

(a) Show that $\overline{X}_n$ is an unbiased estimator of μ.

(b) Show that the estimator

$$S_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2 \tag{11.318}$$

is a biased estimator of σ^2 and the estimator

$$S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2 \tag{11.319}$$

is unbiased.

(c) Show that S_n^2 has a lower mean-squared error than that of S_{n-1}^2. This illustrates the idea that a biased estimator may be "better" than an unbiased estimator.

11.7 *Fisher information and relative entropy*. Show for a parametric family $\{p_\theta(x)\}$ that

$$\lim_{\theta' \to \theta} \frac{1}{(\theta - \theta')^2} D(p_\theta || p_{\theta'}) = \frac{1}{\ln 4} J(\theta). \tag{11.320}$$

11.8 *Examples of Fisher information*. The Fisher information $J(\Theta)$ for the family $f_\theta(x), \theta \in \mathbf{R}$ is defined by

$$J(\theta) = E_\theta\left(\frac{\partial f_\theta(X)/\partial\theta}{f_\theta(X)}\right)^2 = \int \frac{(f_\theta^{'})^2}{f_\theta}.$$

Find the Fisher information for the following families:

(a) $f_\theta(x) = N(0,\theta) = \frac{1}{\sqrt{2\pi\theta}} e^{-\frac{x^2}{2\theta}}$

(b) $f_\theta(x) = \theta e^{-\theta x}, x \geq 0$

(c) What is the Cramèr–Rao lower bound on $E_\theta(\hat{\theta}(X) - \theta)^2$, where $\hat{\theta}(X)$ is an unbiased estimator of θ for parts (a) and (b)?

11.9 *Two conditionally independent looks double the Fisher information.* Let $g_\theta(x_1, x_2) = f_\theta(x_1) f_\theta(x_2)$. Show that $J_g(\theta) = 2J_f(\theta)$.

11.10 *Joint distributions and product distributions.* Consider a joint distribution $Q(x, y)$ with marginals $Q(x)$ and $Q(y)$. Let E be the set of types that look jointly typical with respect to Q:

$$\begin{aligned} E = \{P(x, y) : &-\sum_{x,y} P(x, y) \log Q(x) - H(X) = 0, \\ &-\sum_{x,y} P(x, y) \log Q(y) - H(Y) = 0, \\ &-\sum_{x,y} P(x, y) \log Q(x, y) \\ &-H(X, Y) = 0\}. \end{aligned} \tag{11.321}$$

(a) Let $Q_0(x, y)$ be another distribution on $\mathcal{X} \times \mathcal{Y}$. Argue that the distribution P^* in E that is closest to Q_0 is of the form

$$P^*(x, y) = Q_0(x, y) e^{\lambda_0 + \lambda_1 \log Q(x) + \lambda_2 \log Q(y) + \lambda_3 \log Q(x,y)}, \tag{11.322}$$

where $\lambda_0,\ \lambda_1,\ \lambda_2$, and λ_3 are chosen to satisfy the constraints. Argue that this distribution is unique.

(b) Now let $Q_0(x, y) = Q(x)Q(y)$. Verify that $Q(x, y)$ is of the form (11.322) and satisfies the constraints. Thus, $P^*(x, y) = Q(x, y)$ (i.e., the distribution in E closest to the product distribution is the joint distribution).

11.11 *Cramér–Rao inequality with a bias term.* Let $X \sim f(x; \theta)$ and let $T(X)$ be an estimator for θ. Let $b_T(\theta) = E_\theta T - \theta$ be the bias of the estimator. Show that

$$E(T - \theta)^2 \geq \frac{[1 + b_T'(\theta)]^2}{J(\theta)} + b_T^2(\theta). \tag{11.323}$$

11.12 *Hypothesis testing*. Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim p(x)$. Consider the hypothesis test $H_1 : p = p_1$ vs. $H_2 : p = p_2$. Let

$$p_1(x) = \begin{cases} \frac{1}{2}, & x = -1 \\ \frac{1}{4}, & x = 0 \\ \frac{1}{4}, & x = 1 \end{cases}$$

and

$$p_2(x) = \begin{cases} \frac{1}{4}, & x = -1 \\ \frac{1}{4}, & x = 0 \\ \frac{1}{2}, & x = 1. \end{cases}$$

Find the error exponent for Pr{Decide $H_2|H_1$ true} in the best hypothesis test of H_1 vs. H_2 subject to Pr{Decide $H_1|H_2$ true} $\leq \frac{1}{2}$.

11.13 *Sanov's theorem*. Prove a simple version of Sanov's theorem for Bernoulli(q) random variables.
Let the proportion of 1's in the sequence $X_1, X_2, \ldots, X_n$ be

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{11.324}$$

By the law of large numbers, we would expect $\overline{X}_n$ to be close to q for large n. Sanov's theorem deals with the probability that p_{X^n} is far away from q. In particular, for concreteness, if we take $p > q > \frac{1}{2}$, Sanov's theorem states that

$$\begin{aligned} &-\frac{1}{n}\log \Pr\left\{(X_1, X_2, \ldots, X_n) : \overline{X}_n \geq p\right\} \\ &\to p \log \frac{p}{q} + (1-p)\log\frac{1-p}{1-q} \\ &= D((p, 1-p)||(q, 1-q)). \end{aligned} \tag{11.325}$$

Justify the following steps:

- $\Pr\left\{(X_1, X_2, \ldots, X_n) : \overline{X}_n \geq p\right\} \leq \sum_{i=\lfloor np \rfloor}^{n} \binom{n}{i} q^i (1-q)^{n-i}.$ (11.326)

- Argue that the term corresponding to $i = \lfloor np \rfloor$ is the largest term in the sum on the right-hand side of the last equation.
- Show that this term is approximately 2^{-nD}.
- Prove an upper bound on the probability in Sanov's theorem using the steps above. Use similar arguments to prove a lower bound and complete the proof of Sanov's theorem.

11.14 *Sanov.* Let X_i be i.i.d. $\sim N(0, \sigma^2)$.

(a) Find the exponent in the behavior of $\Pr\{\frac{1}{n}\sum_{i=1}^{n} X_i^2 \geq \alpha^2\}$. This can be done from first principles (since the normal distribution is nice) or by using Sanov's theorem.

(b) What do the data look like if $\frac{1}{n}\sum_{i=1}^{n} X_i^2 \geq \alpha$? That is, what is the P^* that minimizes $D(P \parallel Q)$?

11.15 *Counting states.* Suppose that an atom is equally likely to be in each of six states, $X \in \{s_1, s_2, s_3, \ldots, s_6\}$. One observes n atoms $X_1, X_2, \ldots, X_n$ independently drawn according to this uniform distribution. It is observed that the frequency of occurrence of state s_1 is twice the frequency of occurrence of state s_2.

(a) To first order in the exponent, what is the probability of observing this event?

(b) Assuming n large, find the conditional distribution of the state of the first atom X_1, given this observation.

11.16 *Hypothesis testing.* Let $\{X_i\}$ be i.i.d. $\sim p(x)$, $x \in \{1, 2, \ldots\}$. Consider two hypotheses, $H_0 : p(x) = p_0(x)$ vs. $H_1 : p(x) = p_1(x)$, where $p_0(x) = \left(\frac{1}{2}\right)^x$ and $p_1(x) = qp^{x-1}$, $x = 1, 2, 3, \ldots$.

(a) Find $D(p_0 \parallel p_1)$.

(b) Let $\Pr\{H_0\} = \frac{1}{2}$. Find the minimal probability of error test for H_0 vs. H_1 given data $X_1, X_2, \ldots, X_n \sim p(x)$.

11.17 *Maximum likelihood estimation.* Let $\{f_\theta(x)\}$ denote a parametric family of densities with parameter $\theta \epsilon \mathcal{R}$. Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim f_\theta(x)$. The function

$$l_\theta(x^n) = \ln\left(\prod_{i=1}^{n} f_\theta(x_i)\right)$$

is known as the *log likelihood function*. Let θ_0 denote the true parameter value.

(a) Let the expected log likelihood be

$$E_{\theta_0} l_\theta(X^n) = \int \left(\ln \prod_{i=1}^{n} f_\theta(x_i) \right) \prod_{i=1}^{n} f_{\theta_0}(x_i) dx^n,$$

and show that

$$E_{\theta_0}(l(X^n)) = (-h(f_{\theta_0}) - D(f_{\theta_0}||f_\theta))n.$$

(b) Show that the maximum over θ of the expected log likelihood is achieved by $\theta = \theta_0$.

11.18 *Large deviations*. Let $X_1, X_2, \ldots$ be i.i.d. random variables drawn according to the geometric distribution

$$\Pr\{X = k\} = p^{k-1}(1-p), \qquad k = 1, 2, \ldots .$$

Find good estimates (to first order in the exponent) of:

(a) $\Pr\{\frac{1}{n} \sum_{i=1}^{n} X_i \geq \alpha\}$.

(b) $\Pr\{X_1 = k | \frac{1}{n} \sum_{i=1}^{n} X_i \geq \alpha\}$.

(c) Evaluate parts (a) and (b) for $p = \frac{1}{2}, \alpha = 4$.

11.19 *Another expression for Fisher information*. Use integration by parts to show that

$$J(\theta) = -E \frac{\partial^2 \ln f_\theta(x)}{\partial \theta^2}.$$

11.20 *Stirling's approximation*. Derive a weak form of Stirling's approximation for factorials; that is, show that

$$\left(\frac{n}{e}\right)^n \leq n! \leq n \left(\frac{n}{e}\right)^n \tag{11.327}$$

using the approximation of integrals by sums. Justify the following steps:

$$\ln(n!) = \sum_{i=2}^{n-1} \ln(i) + \ln(n) \leq \int_2^{n-1} \ln x \, dx + \ln n = \cdots \tag{11.328}$$

and

$$\ln(n!) = \sum_{i=1}^{n} \ln(i) \ge \int_0^n \ln x \, dx = \cdots . \tag{11.329}$$

11.21 *Asymptotic value of* $\binom{n}{k}$. Use the simple approximation of Problem 11.20 to show that if $0 \le p \le 1$, and $k = \lfloor np \rfloor$ (i.e., k is the largest integer less than or equal to np), then

$$\lim_{n\to\infty} \frac{1}{n} \log \binom{n}{k} = -p \log p - (1-p)\log(1-p) = H(p). \tag{11.330}$$

Now let p_i, $i = 1, \ldots, m$ be a probability distribution on m symbols (i.e., $p_i \ge 0$ and $\sum_i p_i = 1$). What is the limiting value of

$$\begin{aligned} &\frac{1}{n} \log \begin{pmatrix} n \\ \lfloor np_1 \rfloor \; \lfloor np_2 \rfloor \; \ldots \lfloor np_{m-1} \rfloor \; n - \sum_{j=0}^{m-1} \lfloor np_j \rfloor \end{pmatrix} \\ &\quad = \frac{1}{n} \log \frac{n!}{\lfloor np_1 \rfloor ! \; \lfloor np_2 \rfloor ! \; \ldots \lfloor np_{m-1} \rfloor ! \; (n - \sum_{j=0}^{m-1} \lfloor np_j \rfloor)!} ? \end{aligned} \tag{11.331}$$

11.22 *Running difference*. Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim Q_1(x)$, and $Y_1, Y_2, \ldots, Y_n$ be i.i.d. $\sim Q_2(y)$. Let X^n and Y^n be independent. Find an expression for $\Pr\{\sum_{i=1}^n X_i - \sum_{i=1}^n Y_i \ge nt\}$ good to first order in the exponent. Again, this answer can be left in parametric form.

11.23 *Large likelihoods*. Let $X_1, X_2, \ldots$ be i.i.d. $\sim Q(x)$, $x \in \{1, 2, \ldots, m\}$. Let $P(x)$ be some other probability mass function. We form the log likelihood ratio

$$\frac{1}{n} \log \frac{P^n(X_1, X_2, \ldots, X_n)}{Q^n(X_1, X_2, \ldots, X_n)} = \frac{1}{n} \sum_{i=1}^{n} \log \frac{P(X_i)}{Q(X_i)}$$

of the sequence X^n and ask for the probability that it exceeds a certain threshold. Specifically, find (to first order in the exponent)

$$Q^n\left(\frac{1}{n} \log \frac{P(X_1, X_2, \ldots, X_n)}{Q(X_1, X_2, \ldots, X_n)} > 0\right).$$

There may be an undetermined parameter in the answer.

11.24 *Fisher information for mixtures.* Let $f_1(x)$ and $f_0(x)$ be two given probability densities. Let Z be Bernoulli(θ), where θ is unknown. Let $X \sim f_1(x)$ if $Z = 1$ and $X \sim f_0(x)$ if $Z = 0$.

(a) Find the density $f_\theta(x)$ of the observed X.

(b) Find the Fisher information $J(\theta)$.

(c) What is the Cramér–Rao lower bound on the mean-squared error of an unbiased estimate of θ?

(d) Can you exhibit an unbiased estimator of θ?

11.25 *Bent coins.* Let $\{X_i\}$ be iid $\sim Q$, where

$$Q(k) = \Pr(X_i = k) = \binom{m}{k} q^k (1-q)^{m-k} \quad \text{for } k = 0, 1, 2, \ldots, m.$$

Thus, the X_i's are iid $\sim$ Binomial(m, q). Show that as $n \to \infty$,

$$\Pr\left(X_1 = k \left| \frac{1}{n} \sum_{i=1}^{n} X_i \geq \alpha \right.\right) \to P^*(k),$$

where P^* is Binomial(m, λ) (i.e., $P^*(k) = \binom{m}{k} \lambda^k (1-\lambda)^{m-k}$ for some $\lambda \in [0, 1]$). Find λ.

11.26 *Conditional limiting distribution*

(a) Find the exact value of

$$\Pr\left\{X_1 = 1 \left| \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{1}{4} \right.\right\} \tag{11.332}$$

if $X_1, X_2, \ldots,$ are Bernoulli($\frac{2}{3}$) and n is a multiple of 4.

(b) Now let $X_i \epsilon \{-1, 0, 1\}$ and let $X_1, X_2 \ldots$ be i.i.d. uniform over $\{-1, 0, +1\}$. Find the limit of

$$\Pr\left\{X_1 = +1 \left| \frac{1}{n} \sum_{i=1}^{n} X_i^2 = \frac{1}{2} \right.\right\} \tag{11.333}$$

for $n = 2k, k \to \infty$.

11.27 *Variational inequality*. Verify for positive random variables X that

$$\log E_P(X) = \sup_Q \left[E_Q(\log X) - D(Q||P) \right], \tag{11.334}$$

where $E_P(X) = \sum_x x P(x)$ and $D(Q||P) = \sum_x Q(x) \log \frac{Q(x)}{P(x)}$ and the supremum is over all $Q(x) \geq 0$, $\sum Q(x) = 1$. It is enough to extremize $J(Q) = E_Q \ln X - D(Q||P) + \lambda(\sum Q(x) - 1)$.

11.28 *Type constraints*

(a) Find constraints on the type P_{X^n} such that the sample variance $\overline{X_n^2} - (\overline{X}_n)^2 \leq \alpha$, where $\overline{X_n^2} = \frac{1}{n} \sum_{i=1}^n X_i^2$ and $\overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$.

(b) Find the exponent in the probability $Q^n(\overline{X_n^2} - (\overline{X}_n)^2 \leq \alpha)$. You can leave the answer in parametric form.

11.29 *Uniform distribution on the simplex*. Which of these methods will generate a sample from the uniform distribution on the simplex $\{x \in R^n : x_i \geq 0, \quad \sum_{i=1}^n x_i = 1\}$?

(a) Let Y_i be i.i.d. uniform $[0, 1]$ with $X_i = Y_i / \sum_{j=1}^n Y_j$.

(b) Let Y_i be i.i.d. exponentially distributed $\sim \lambda e^{-\lambda y}$, $y \geq 0$, with $X_i = Y_i / \sum_{j=1}^n Y_j$.

(c) (*Break stick into n parts*) Let $Y_1, Y_2, \ldots, Y_{n-1}$ be i.i.d. uniform $[0, 1]$, and let X_i be the length of the ith interval.

HISTORICAL NOTES

The method of types evolved from notions of strong typicality; some of the ideas were used by Wolfowitz [566] to prove channel capacity theorems. The method was fully developed by Csiszár and Körner [149], who derived the main theorems of information theory from this viewpoint. The method of types described in Section 11.1 follows the development in Csiszár and Körner. The $\mathcal{L}_1$ lower bound on relative entropy is due to Csiszár [138], Kullback [336], and Kemperman [309]. Sanov's theorem [455] was generalized by Csiszár [141] using the method of types.