23 The Method of Types and Sanov's Theorem

Our analysis of large deviations thus far has been fairly general, and thus applicable to a range of problems involving both continuous and discrete distributions. However, when the distribution is discrete and over a finite alphabet $\mathcal{Y}$, a framework of analysis referred to as the *method of types* can be applied instead.

This framework leads to Sanov's theorem, which is both more and less general than Cramér's theorem. It is more general in that it characterizes the behavior of a greater variety of large deviation events, but it is less general in that it only applies to distributions over finite alphabets. In addition, the analysis is at a relatively coarse scale, with the associated bounds depending strongly on the cardinality of the alphabet.

Ultimately, perhaps the greatest value of the analysis via the method of types is the insight it provides. In particular, this approach leads to valuable information geometric perspectives, which help to significant refine our intuition.

Without loss of generality, in this section we let $\mathcal{Y} = \{1, 2, \ldots, M\}$. And unless otherwise indicated, all logarithms continue to be base-2.

23.1 Types and Type Classes

Definition 1. *Let* $\mathbf{y} = [y_1, \ldots, y_N]^{\mathrm{T}}$ *be a sequence drawn from a finite alphabet* $\mathcal{Y}$*. The* empirical distribution*, or* type*, of* $\mathbf{y}$ *is a probability distribution over the same alphabet defined as follows:*

$$\hat{p}(b; \mathbf{y}) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}_b(y_n) = \frac{N_b(\mathbf{y})}{N}, \tag{1}$$

where $N_b(\mathbf{y})$ *is the number of times the value* b *appears in the sequence* $\mathbf{y}$*.*

A type $\hat{p}(\cdot; \mathbf{y})$ evidently describes the relative frequencies of the elements of $\mathcal{Y}$ in $\mathbf{y}$.

Example 1. Suppose $\mathbf{y} = \begin{bmatrix} 1 & 0 & 1 \end{bmatrix}^{\mathrm{T}}$. Then

$$\hat{p}_y(b; \mathbf{y}) = \begin{cases} 1/3 & b = 0 \\ 2/3 & b = 1 \end{cases}.$$

The type can be viewed as an estimate of the true distribution that generated the sequence $\mathbf{y}$. Indeed, since

$$\mathbb{E}\left[\mathbb{1}_b(\mathsf{y})\right] = 1 \cdot \mathbb{P}\left(\mathsf{y} = b\right) + 0 \cdot \mathbb{P}\left(\mathsf{y} \neq b\right) = p(b), \tag{2}$$

it follows from the weak law of large numbers that for each $b \in \mathcal{Y}$,

$$\lim_{N\to\infty} \mathbb{P}\left(|\hat{p}(b;\mathbf{y}) - p(b)| > \epsilon\right) = 0, \tag{3}$$

i.e., the type converges in probability to the true distribution, which we express using the notation $\hat{p}(b;\mathbf{y}) \xrightarrow{\text{p}} p(b)$.[1]

Example 2. If $\mathcal{Y} = \{0,1\}$ and

$$p(b) = \begin{cases} 1/3 & b = 0 \\ 2/3 & b = 1, \end{cases}$$

then (3) implies that in a long sequence of samples from p, roughly $1/3$ of the elements will be 0's.

In fact, as we will discuss later in these notes, (3) implies yet another way to define a typical set, with connections to our analysis. However, we do not yet need such insights.

Definition 2. *The* set of types $\mathcal{P}_N^{\mathcal{Y}}$ *is the set of all possible types for sequences of length* N *generated from an alphabet* $\mathcal{Y}$*.*

When there is no risk of confusion, we may omit the superscript, letting the alphabet be implicit.

Example 3. Suppse $\mathcal{Y} = \{0,1\}$. Then there are $N+1$ possible types:

$$\mathcal{P}_N(\mathcal{Y}) = \left\{(0,1), \left(\frac{1}{N}, \frac{N-1}{N}\right), \left(\frac{2}{N}, \frac{N-2}{N}\right), \ldots, (1,0)\right\}. \tag{4}$$

Definition 3. *Let* p *be a type defined over a finite alphabet* $\mathcal{Y}$ *for sequences of length* N*, i.e.,* $p \in \mathcal{P}_N^{\mathcal{Y}}$*. The set of all sequences of length* N *whose type is equal to* p

$$\mathcal{T}_N^{\mathcal{Y}}(p) = \left\{\mathbf{y} \in \mathcal{Y}^N \colon \hat{p}(\cdot;\mathbf{y}) \equiv p(\cdot)\right\} \tag{5}$$

is called the type class*.*

Again, when there is no risk of confusion, we may omit the superscript, letting the alphabet be implicit.

[1] More generally, we write $z_N \xrightarrow{\text{p}} a$ when for any $\epsilon > 0$,

$$\lim_{N\to\infty} \mathbb{P}\left(|z_N - a| > \epsilon\right) = 0.$$

Example 4. Suppose $\mathcal{Y} = \{0, 1\}$, let

$$p(y) = \begin{cases} 1/3 & y = 0 \\ 2/3 & y = 1. \end{cases}$$

Then

$$T_3(p) = \left\{ \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \right\}. \tag{6}$$

More generally, when N is a multiple of 3, the number of sequences in the type class for p is

$$|T_N(p)| = \binom{N}{N/3}.$$

Finally, we note that many calculations involving types are simplified by replacing sample geometric means with symbol geometric means.

Identity 1. *For an arbitrary choice of $g(\cdot)$,*

$$\left(\prod_{n=1}^{N} g(y_n) \right)^{1/N} = \prod_{b=1}^{M} g(b)^{\hat{p}(b;\mathbf{y})}. \tag{7}$$

Moreover, for the choice $g(\cdot) = 2^{t(\cdot)}$, (7) becomes, after taking logarithms,

$$\frac{1}{N} \sum_{n=1}^{N} t(y_n) = \sum_{b=1}^{M} \hat{p}(b; \mathbf{y})\, t(b), \tag{8}$$

which relates sample arithmetic means to symbol-wise ones.

Proof. Using the (almost trivial) identity

$$g(y) = \prod_{b=1}^{M} g(b)^{\mathbb{1}_b(y)} \tag{9}$$

we obtain

$$\begin{aligned} \left(\prod_{n=1}^{N} g(y_n) \right)^{1/N} &= \left(\prod_{n=1}^{N} \prod_{b=1}^{M} g(b)^{\mathbb{1}_b(y_n)} \right)^{1/N} \\ &= \prod_{b=1}^{M} \prod_{n=1}^{N} g(b)^{\mathbb{1}_b(y_n)/N} \\ &= \prod_{b=1}^{M} g(b)^{(1/N)\sum_{n=1}^{N} \mathbb{1}_b(y_n)} \\ &= \prod_{b=1}^{M} g(b)^{\hat{p}(b;\mathbf{y})}. \end{aligned}$$

□

23.2 Exponential Rate Notation

It is convenient to introduce the following notation for describing the asymptotic rate of exponential growth or decay of functions. Specifically, we use $f(N) \doteq 2^{N\alpha}$ to denote

$$\lim_{N\to\infty} \frac{\log f(N)}{N} = \alpha, \tag{10}$$

i.e., the growth rate of $f(N)$ is equal to α. Phrased differently, (10) expresses that the approximation $f(N) \cong 2^{N\alpha}$ is accurate to first-order in the exponent. Clearly, we can also use a different base—e.g., e—in the exponential and the logarithm without changing the meaning of the notation. We similarly define the notation $\dot{<}$, $\dot{>}$, $\dot{\leq}$, and $\dot{\geq}$.

Example 5. The following examples are readily obtained directly from the definition (10). First, we have

$$f(N) = 2^{N\alpha+\sqrt{N}\beta} \doteq 2^{N\alpha},$$

which emphasizes that the notion of approximation is sufficiently coarse-grained that sublinear terms in exponent are not distinguished by the notation. Second, we have

$$f(N) = N^{\beta}2^{N\alpha} = 2^{N\alpha+\beta\log N} \doteq 2^{N\alpha},$$

which emphasizes that the polynomial growth factors are also not distinguished by the notation since they also correspond to sublinear (in this case, logarithmic) terms in the exponent. This of course includes constant factors.

It is important to correctly interpret limiting cases of this notation. In particular,

(i) $f(N) \doteq \infty \doteq 2^{N\infty}$ means that $f(\cdot)$ grows superexponentially; e.g., $f(N) = 2^{N^2}$.

(ii) $f(N) \doteq 0 \doteq 2^{-N\infty}$ means that $f(\cdot)$ decays superexponentially; e.g., $f(N) = 2^{-N^2}$.

(iii) $f(N) \doteq 1 \doteq 2^{N\cdot 0}$ means that $f(\cdot)$ grows or decays subexponentially; e.g., $f(N) = N^2$ or $f(N) = 1 + N$ or $f(N) = 1/\sqrt{N}$ or $f(N) = 2 - 1/N$. Note that although such notation includes the possibility that $f(N)$ converges to a constant, the coarseness of approximation inherent in the notation does not allow us to distinguish this possibility from the others.

Finally, we can use the definition (10) to deduce the behavior of, e.g., addition and multiplication under this notation. For instance, if $f(N) \doteq 2^{N\alpha}$ and $g(N) \doteq 2^{N\beta}$, then

$$f(N) + g(N) \doteq 2^{N\max(\alpha,\beta)}$$

and

$$f(N) \cdot g(N) \doteq 2^{N(\alpha+\beta)}$$

23.3 Properties of Types

Now we are ready to establish the properties of types needed for our large deviation analysis. Of particular relevance is how likely it is that sampling from some distribution q will generate a sequence whose type is p. As intuition might suggest based on our related analysis of typical sets earlier in the notes, in fact

$$Q\left\{\mathcal{T}_N^{\mathcal{Y}}(p)\right\} \cong 2^{-ND(p\|q)}.$$

In this section, we develop a suitably precise statement of this result.

As our first step, we establish that relative to the number of possible sequences (which is exponential in N), there aren't very many types.

Lemma 1. *For any finite alphabet $\mathcal{Y}$,*

$$\left|\mathcal{P}_N^{\mathcal{Y}}\right| \leq (N+1)^{|\mathcal{Y}|}, \tag{11}$$

i.e., the number of types is polynomial in the sequence length.

Proof. Each type in $\mathcal{P}_N^{\mathcal{Y}}$ can be represented as a string of length $|\mathcal{Y}|$ whose elements correspond to frequency counts and take integer values between 0 and N. This immediately implies the statement of the theorem. □

While this bound is obviously not tight—indeed, Example 3 shows that there are $N+1$ types when $M=2$, not $(N+1)^2$—it is sufficient for our needs. Indeed, ultimately we need only that the number of types is subexponential.

One obvious implication of this result is that at least one of the type classes must contain exponentially many sequences. In fact, we will show that essentially *all* the type classes are exponentially large.

As a first step note that when sequences are generated i.i.d. according to some distribution, sequences that are permutations of one another must have the same probability. Since the largest collection of sequences that are permutations of one another is a type class, this means that the probability of a sequence depends only on its type.[2] The following theorem quantifies the exponentially small probability of a sequence as a function of its type.

Lemma 2. *Let $\mathbf{y} = [y_1, \ldots y_N]^{\mathrm{T}}$ be a sequence drawn from a finite alphabet $\mathcal{Y}$, i.e., $\mathbf{y} \in \mathcal{Y}^N$, and let $q \in \mathcal{P}^{\mathcal{Y}}$ be any distribution defined over the same alphabet. Then*

$$q^N(\mathbf{y}) = 2^{-N(D(\hat{p}(\cdot;\mathbf{y})\|q)+H(\hat{p}(\cdot;\mathbf{y})))}. \tag{12}$$

[2]This means a sequence generated in an i.i.d. manner can be equivalently represented by its type for the purposes of inference, i.e., the type is a sufficient statistic in the sense developed earlier in the notes.

Proof. Using (8) with $t(\cdot) = \log q(\cdot)$, we have

$$\begin{aligned}\frac{1}{N}\log q^N(\mathbf{y}) &= \frac{1}{N}\sum_{n=1}^{N}\log q(y_n)\\ &= \sum_{b\in\mathcal{Y}}\hat{p}(b;\mathbf{y})\log q(b)\\ &= \mathbb{E}_{\hat{p}(\cdot;\mathbf{y})}\left[\log q(Y)\right]\\ &= -D(\hat{p}(\cdot;\mathbf{y})\|q) - H(\hat{p}(\cdot;\mathbf{y})).\end{aligned}$$

□

As special cases, we have that

$$q^N(\mathbf{y}) = 2^{-N(D(p\|q)+H(p))} \quad \text{for } \mathbf{y}\in\mathcal{T}_N^{\mathcal{Y}}(p), \tag{13}$$

which follows from recognizing that by definition $\hat{p}(\cdot;\mathbf{y}) = p$ for $\mathbf{y}\in\mathcal{T}_N^{\mathcal{Y}}(p)$, and

$$p^N(\mathbf{y}) = 2^{-NH(p)} \quad \text{for } \mathbf{y}\in\mathcal{T}_N^{\mathcal{Y}}(p), \tag{14}$$

which follows from setting $q = p$ in (13).

Next we establish that every (nondegenerate) type class contains exponentially many sequences.

Lemma 3. *Let $p\in\mathcal{P}_N^{\mathcal{Y}}$ for a finite alphabet $\mathcal{Y}$. Then $\left|\mathcal{T}_N^{\mathcal{Y}}(p)\right| \doteq 2^{NH(p)}$. Specifically,*

$$cN^{-|\mathcal{Y}|}2^{NH(p)} \le \left|\mathcal{T}_N^{\mathcal{Y}}(p)\right| \le 2^{NH(p)}, \tag{15}$$

where c is a constant that may depend on $|\mathcal{Y}|$ and p but does not depend on N.

Part of our proof makes use of the following fact, which can be viewed as a coarse version of Stirling's approximation.[3]

Fact 1. *For $n\ge 1$ we have*[4]

$$e\left(\frac{n}{e}\right)^n \le n! \le ne\left(\frac{n}{e}\right)^n \tag{16}$$

[3]A finer-grained version of Stirling's approximation is $n!\cong\sqrt{2\pi n}(n/e)^n$, which is accurate in the sense that

$$\lim_{n\to\infty}\frac{n!}{\sqrt{2\pi n}(n/e)^n} = 1.$$

[4]To verify (16), it suffices to write $n! = \exp\left(\sum_{i=1}^{n}\ln i\right)$, then upper and lower bound the sum in the exponent by suitable integrals; specifically,

$$n\ln n - n + 1 = \int_1^n \ln x\,\mathrm{d}x \le \sum_{i=1}^{n}\ln i \le \ln n + \int_1^n \ln x\,\mathrm{d}x = (n+1)\ln n - n + 1,$$

Proof. To obtain the upper bound, we write

$$\begin{aligned}\left|\mathcal{T}_N^{\mathcal{Y}}(p)\right| &= \sum_{\mathbf{y}\in\mathcal{T}_N^{\mathcal{Y}}(p)} 1 \\ &= 2^{NH(p)} \sum_{\mathbf{y}\in\mathcal{T}_N^{\mathcal{Y}}(p)} 2^{-NH(p)} \\ &= 2^{NH(p)} \sum_{\mathbf{y}\in\mathcal{T}_N^{\mathcal{Y}}(p)} p^N(\mathbf{y}) \qquad (17) \\ &\leq 2^{NH(p)},\end{aligned}$$

where to obtain (17) we have used (14).

To obtain the lower bound, we note that the number of sequences in the type class corresponds to the number of different ways of choosing $N_1 = Np(1)$ elements with value 1, $N_2 = Np(2)$ elements with value 2, and so on, which is expressed by the associated multinomial coefficient, i.e.,

$$\left|\mathcal{T}_N^{\mathcal{Y}}(p)\right| = \binom{N}{N_1, N_2, \ldots, N_M} = \frac{N!}{N_1!\, N_2! \cdots N_M!} \qquad (18)$$

We can therefore lower bound (18) using (16), yielding

$$\begin{aligned}\left|\mathcal{T}_N^{\mathcal{Y}}(p)\right| &\geq \frac{e(N/e)^N}{e^M \left(\prod_{b=1}^M N_b\right) \prod_{b=1}^M (N_b/e)^{N_b}} \\ &= \frac{e^{1-N}\, \tilde{c}(p)\, 2^{N\log N}}{(Ne)^M e^{-N} \prod_{b=1}^M 2^{N_b \log N_b}} \\ &= \frac{e\, \tilde{c}(p)\, 2^{\sum_{b=1}^M N_b \log N}}{(Ne)^M 2^{\sum_{b=1}^M N_b \log N_b}} \\ &= \frac{e\, \tilde{c}(p)\, 2^{-N\sum_{b=1}^M (N_b/N)\log(N_b/N)}}{(Ne)^M} \\ &= e^{1-M}\, \tilde{c}(p)\, N^{-M}\, 2^{-NH(p)}\end{aligned}$$

where

$$\tilde{c}(p) \triangleq \left(\prod_{b=1}^M p(b)\right)^{-1}.$$

□

We are now ready to combine Lemmas 2 and 3 to establish our main result—that there is an exponentially small probability that a sequence obtained by sampling from p will have a type q, when $q \neq p$.

Theorem 1. *Let $p \in \mathcal{P}_N^{\mathcal{Y}}$, and let $q \in \mathcal{P}^{\mathcal{Y}}$ be a distribution over the same alphabet. Then*

$$c\, N^{-|\mathcal{Y}|}\, 2^{-ND(p\|q)} \leq Q\left\{\mathcal{T}_N^{\mathcal{Y}}(p)\right\} \leq 2^{-ND(p\|q)}, \tag{19}$$

where c is a constant that does not depend on N, but may depend on $\mathcal{Y}$ and p, whence[5]

$$Q\left\{\mathcal{T}_N^{\mathcal{Y}}(p)\right\} \doteq 2^{-ND(p\|q)}. \tag{20}$$

Proof. We have

$$\begin{aligned} Q\left\{\mathcal{T}_N^{\mathcal{Y}}(p)\right\} &= \sum_{\mathbf{y}\in\mathcal{T}_N^{\mathcal{Y}}(p)} q^N(\mathbf{y}) \\ &= \sum_{\mathbf{y}\in\mathcal{T}_N^{\mathcal{Y}}(p)} 2^{-N(D(p\|q)+H(p))} \qquad (21) \\ &= \left|\mathcal{T}_N^{\mathcal{Y}}(p)\right| 2^{-N(D(p\|q)+H(p))}, \qquad (22) \end{aligned}$$

where to obtain (21) we have used (13), In turn, applying (15) to (22) we obtain (19). □

Note that for the special case $q = p$, Theorem 1 establishes that $P\left\{\mathcal{T}_N^{\mathcal{Y}}(p)\right\} \doteq 1$, i.e., the probability a sequence will have type p if generated i.i.d. from p is not exponentially small. As such, it is the most likely type to occur.

As an aside, stronger statements are possible using a finer grained analysis of the size of the type class based on a more accurate Stirling's approximation, such as that mentioned earlier. For example, it can be shown that using such an approximation, $\mathcal{T}_N^{\mathcal{Y}}(p) = O(N^{-|\mathcal{Y}|/2})2^{-NH(p)}$, so $P\left\{\mathcal{T}_N^{\mathcal{Y}}(p)\right\} = O(N^{-|\mathcal{Y}|/2})$, i.e., the probability decays polynomially. Hence, with high probability, a sequence generated by p will have a type other than p. However, while the sequence does not have type p with high probability, from (3) we know that for any small fixed $\epsilon > 0$, the type of the sequence is within ϵ of p with high probability. Moreover, of the $O(N)$ types in this neighborhood, we emphasize that p is the most likely.

23.4 Large Deviation Analysis via Types

The method of types provides a convenient framework for analyzing the behavior of any of a wide range of possible atypical events based on Theorem 1, as we now develop.

The key to applying the method of types is exploiting that for sequences generated i.i.d. from a distribution, events involving sequences can be expressed as events

[5]The astute reader will notice a subtlety in this expression. In particular, (20) describes a limit in N, but the assumption $p \in \mathcal{P}_N^{\mathcal{Y}}$ cannot hold for all N sufficiently large. However, if $p \in \mathcal{P}_{N_0}^{\mathcal{Y}}$ for some N_0, then $p \in \mathcal{P}_N^{\mathcal{Y}}$ for $N = kN_0$ for all $k \geq 1$, so the limit can be taken with respect to such a subsequence.

involving their corresponding types. From this perspective, any event of interest can be expressed in the form

$$\mathcal{R} = \left\{ \mathbf{y} \in \mathcal{Y}^N \colon \hat{p}(\cdot; \mathbf{y}) \in \mathcal{S} \cap \mathcal{P}_N^{\mathcal{Y}} \right\},$$

where $\mathcal{S} \subset \mathcal{P}^{\mathcal{Y}}$.

We now turn to evaluating the probability of the event $\mathcal{R}$. In our development, we will adopt the following convenient abuse of notation

$$\mathbb{P}_q \left[\hat{p}(\cdot; \mathbf{y}) \in \mathcal{S} \right] \triangleq Q \left\{ \mathcal{S} \right\} = Q \left\{ \mathcal{S} \cap \mathcal{P}_N^{\mathcal{Y}} \right\} = Q \left\{ \mathcal{R} \right\} \triangleq \mathbb{P}_q \left[\mathbf{y} \in \mathcal{R} \right], \tag{23}$$

since the event $\mathbf{y} \in \mathcal{R}$ is the event $\hat{p}(\cdot; \mathbf{y}) \in \mathcal{S} \cap \mathcal{P}_N^{\mathcal{Y}}$.

If the samples are generated by a probability distribution q, then according to Theorem 1 the probability of any type class $\mathcal{T}_N^{\mathcal{Y}}(p)$ decays exponentially, i.e., $Q \left\{ \mathcal{T}_N^{\mathcal{Y}}(p) \right\} \doteq 2^{-ND(p\|q)}$. Since by Lemma 1 there are at most a polynomial number of types, the exponential that corresponds to the type "closest" to q dominates the sum. The result is express in the form of the following theorem, and the proof formalizes the associated reasoning.

Theorem 2 (Sanov's Theorem)**.** *Let $\mathcal{S} \subset \mathcal{P}^{\mathcal{Y}}$ be a closed set, and let $q \in \mathcal{P}^{\mathcal{Y}}$ be arbitrary. Then*

$$Q \left\{ \mathcal{S} \cap \mathcal{P}_N^{\mathcal{Y}} \right\} \leq (N+1)^{|\mathcal{Y}|} \, 2^{-ND(p_* \| q)}, \tag{24}$$

whence,

$$Q \left\{ \mathcal{S} \cap \mathcal{P}_N^{\mathcal{Y}} \right\} \dot{\leq} 2^{-ND(p_* \| q)}, \tag{25}$$

where

$$p_* = \arg\min_{p \in \mathcal{S}} D(p \| q) \tag{26}$$

is the I-projection of q onto $\mathcal{S}$. If $\mathcal{S}$ is also the closure of its interior,[6] *then*

$$Q \left\{ \mathcal{S} \cap \mathcal{P}_N^{\mathcal{Y}} \right\} \doteq 2^{-ND(p_* \| q)}. \tag{27}$$

The geometry of Sanov's theorem is depicted in Fig. 1. As the theorem expresses, the probability of a large deviation is characterized by the distance (as measure by information divergence) of the set $\mathcal{S}$ from the generating distribution q.

It should be emphasized that Sanov's thereom is a generalization of Cramér's Theorem we developed in the last installment of the notes, and accordingly we refer to

$$D_* = D(p_* \| q) = \min_{p \in \mathcal{S}} D(p \| q) \tag{28}$$

as the Chernoff exponent of the large deviation event associated with $\mathcal{S}$. In the next section will will discuss the specialization of Sanov's Theorem to the scenario of Cramér's Theorem, and the associated geometry.

[6]Note that this is a stronger condition than just requiring $\mathcal{S}$ be closed. For example, for $|\mathcal{Y}| = 2$, the discrete set $\mathcal{S} = \{1/3, 2/3\} \cup \{1/2, 1/2\}$ is closed but the closure of its interior is empty. Closed convex sets are examples of sets that are closures of their interiors.

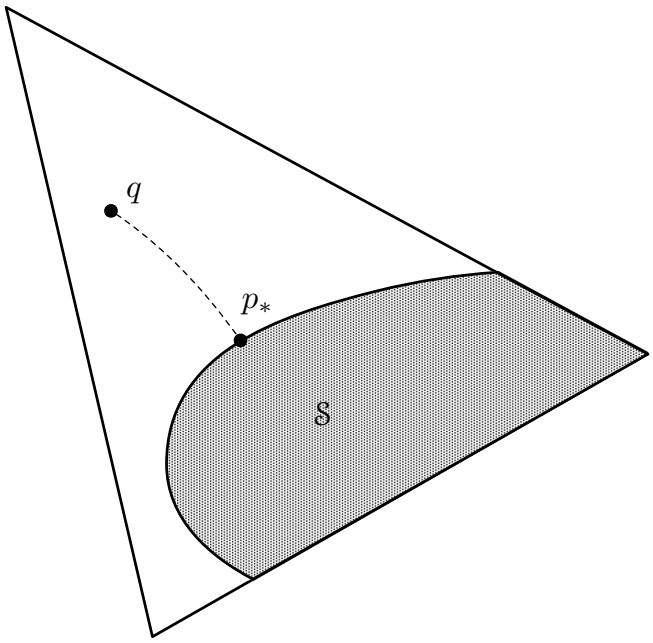

Figure 1: The geometry of Sanov's theorem. The probability that a sequence generated i.i.d. according to q will have a type in the set $\mathcal{S}$ is exponentially small with Chernoff exponent $D(p_*\|q)$, where p_* is the I-projection of q onto $\mathcal{S}$.

Proof. To obtain (24) we write

$$\begin{aligned} Q\left\{\mathcal{S}\cap\mathcal{P}_N^{\mathcal{Y}}\right\} &= \sum_{p\in\mathcal{S}\cap\mathcal{P}_N^{\mathcal{Y}}} Q\left\{\mathcal{T}_N^{\mathcal{Y}}(p)\right\} \\ &\le \sum_{p\in\mathcal{S}\cap\mathcal{P}_N^{\mathcal{Y}}} 2^{-ND(p\|q)} && (29)\\ &\le \sum_{p\in\mathcal{S}\cap\mathcal{P}_N^{\mathcal{Y}}} 2^{-ND(p_*\|q)} && (30)\\ &\le |\mathcal{P}_N^{\mathcal{Y}}|\, 2^{-ND(p_*\|q)} \\ &\le (N+1)^{|\mathcal{Y}|}\, 2^{-ND(p_*\|q)}, && (31)\end{aligned}$$

where to obtain (29) we have used the upper bound in (19), where to obtain (30) we have used that, by construction, $D(p\|q) \ge D(p_*\|q)$ for any $p \in \mathcal{S}$, and where to obtain (31) we have used (11).

Note that in general p_* will not be in the set of types $\mathcal{P}_N^{\mathcal{Y}}$. However, if $\mathcal{S}$ is a closure of its interior, as $N \to \infty$ the set $\mathcal{S} \cap \mathcal{P}_N^{\mathcal{Y}}$ contains distributions that are arbitrarily close to p_*. Formally, there exists a sequence of distributions $\{p_N \in \mathcal{S} \cap \mathcal{P}_N^{\mathcal{Y}}\}$ for $N = 1, 2, \ldots$ such that

$$\lim_{N\to\infty} D(p_N\|q) = D(p_*\|q).$$

We use this sequence to obtain (27). In particular, we have

$$\begin{aligned} Q\left\{\mathcal{S}\cap\mathcal{P}_N^{\mathcal{Y}}\right\} &= \sum_{p\in\mathcal{S}\cap\mathcal{P}_N^{\mathcal{Y}}} Q\left\{\mathcal{T}_N^{\mathcal{Y}}(p)\right\} && (32)\\ &\ge Q\left\{\mathcal{T}_N^{\mathcal{Y}}(p_N)\right\} && (33)\\ &\ge c\, N^{-|\mathcal{Y}|} 2^{-ND(p_N\|q)}, && (34)\end{aligned}$$

where to obtain (33) we have selected one term in the summation of (32), and where to obtain (34) we have used the lower bound in (19).

Rearranging the left- and right-hand sides of (34) and taking limits, we obtain

$$\begin{aligned} \frac{1}{N}\log Q\left\{\mathcal{S}\cap\mathcal{P}_N^{\mathcal{Y}}\right\} &\ge \frac{\log c - |\mathcal{Y}|\log N}{N} - D(p_N\|q) && (35)\\ &\to -D(p_*\|q) \qquad \text{as } N\to\infty, && (36)\end{aligned}$$

whence

$$Q\left\{\mathcal{S}\cap\mathcal{P}_N^{\mathcal{Y}}\right\} \doteq 2^{-ND(p_*\|q)}. \tag{37}$$

Combining (37) with (25) we obtain (27). □

23.5 Example: Large Deviations of Sample Averages

Reviting the large deviation scenario consider in the last installment of the notes, it is straightforward to apply Theorem 2 to evaluate the probability of the event

$$\frac{1}{N}\sum_{n=1}^{N} t(y_n) \geq \gamma,$$

where the data are generated i.i.d. according to a distribution q with mean μ, where $t(\cdot)$ be a real-valued statistic, and where $\gamma > \mu$. In particular, using the identity (8) we obtain that

$$\begin{aligned}
\mathcal{R} &= \left\{ \mathbf{y} \in \mathcal{Y}^N \colon \frac{1}{N}\sum_{n=1}^{N} t(y_n) \geq \gamma \right\} \\
&= \left\{ \mathbf{y} \in \mathcal{Y}^N \colon \sum_{b \in \mathcal{Y}} \hat{p}(b;\mathbf{y})\, t(b) \geq \gamma \right\} \\
&= \left\{ \mathbf{y} \in \mathcal{Y}^N \colon \mathbb{E}_{\hat{p}(\cdot;\mathbf{y})}\left[t(\mathsf{y})\right] \geq \gamma \right\} \\
&= \left\{ \mathbf{y} \in \mathcal{Y}^N \colon \hat{p}(\cdot;\mathbf{y}) \in \mathcal{S} \cap \mathcal{P}_N^{\mathcal{Y}} \right\}, \qquad (38)
\end{aligned}$$

where

$$\mathcal{S} = \left\{ p \in \mathcal{P}^{\mathcal{Y}} \colon \mathbb{E}_p\left[t(\mathsf{y})\right] \geq \gamma \right\}. \qquad (39)$$

Recall from our earlier development of information geometry that the constraint $\mathbb{E}_p\left[t(\mathsf{y})\right] = \gamma$ defines a linear family $\mathcal{L}$ within the simplex $\mathcal{P}^{\mathcal{Y}}$, so the boundary of $\mathcal{S}$ is an affine subspace restricted to the simplex. Moreover, the I-projection of q onto this linear family lies along the one-parameter linear exponential family

$$p_{\mathsf{y}}(y;x) = q(y)\, e^{xt(y)-\alpha(x)},$$

which is "orthogonal" to $\mathcal{L}$. The choice of x in Cramér's Theorem corresponds precisely to the intersection between this exponential family and $\mathcal{L}$, which represents the closest type in $\mathcal{S}$ to q in a divergence sense. This closest type provides dominant contribution to the large deviation probability since it corresponds to the smallest exponent, and thus the associated divergence is the Chernoff exponent.

It is worth noting that while Theorem 2 is more powerful than the scalar version of Cramér's Theorem, it is less powerful than its vector generalization. To see this, consider any distribution q and set $\mathcal{S}$ in Theorem 2, and consider a special case of the vector Cramér's Theorem in which, with some abuse of notation,

$$\mathbf{y}_n \triangleq \begin{bmatrix} t_1(\mathsf{y}_n) \\ t_2(\mathsf{y}_n) \\ \vdots \\ t_M(\mathsf{y}_n) \end{bmatrix},$$

with y_n distributed i.i.d. according to q, and with $t_1, t_2, \ldots t_M$ chosen such that

$$\mathbf{T} \triangleq \begin{bmatrix} t_1(1) & t_1(2) & \cdots & t_1(M) \\ t_2(1) & t_2(2) & \cdots & t_2(M) \\ \vdots & \vdots & \ddots & \vdots \\ t_M(1) & t_M(2) & \cdots & t_M(M) \end{bmatrix}$$

is nonsingular. Then since

$$\frac{1}{N}\sum_{n=1}^{N} \mathbf{y}_n = \sum_{b=1}^{M} \hat{p}(b;\mathbf{y}) \begin{bmatrix} t_1(b) \\ t_2(b) \\ \vdots \\ t_M(b) \end{bmatrix} = \mathbf{T} \begin{bmatrix} \hat{p}(1;\mathbf{y}) \\ \hat{p}(2;\mathbf{y}) \\ \vdots \\ \hat{p}(M;\mathbf{y}) \end{bmatrix} \triangleq t(\hat{p}(\cdot;\mathbf{y})), \tag{40}$$

we see that the (linear) mapping t is one-to-one. Thus, the large deviation event $\hat{p}(\cdot;\mathbf{y}) \in \mathcal{S}$ in Theorem 2 can be equivalently expressed as

$$\frac{1}{N}\sum_{n=1}^{N} \mathbf{y}_n \in \mathcal{F}, \quad \text{with } \mathcal{F} \triangleq t(\mathcal{S}),$$

in Cramér's Theorem.

A simple example of a large deviation event that cannot be directly analyzed by the scalar version of Cramér's Theorem, but can by Theorem 2 is

$$\mathbb{P}\left(\frac{1}{N}\sum_{n=1}^{N} y_n \geq \gamma_1 \quad \text{and} \quad \frac{1}{N}\sum_{n=1}^{N} s(y_n) \geq \gamma_2 \right),$$

for some s. However, we emphasize that applying, instead, the vector version of Cramér's Theorem to the analysis of such events has two notable advantages: 1) it is not limited to distributions over finite alphabets and, in particular, can be used for continuous distributions; and 2) it provides better bounds on the large deviation probability since they are not dependent on the cardinality of the alphabet (though our simple statement of the vector version does not reveal them).

23.6 Dominant Atypical Behavior

While Sanov's theorem tells us how unlikely large deviation events are in an exponential sense, in many applications we also want to know *how* they happen. In particular, we want to know the type of the data. Although Sanov's theorem *suggests* that when event happens the type will be p_*, it doesn't say something quite that strong. However, with only a little more analysis, we can indeed verify that with high probability the observed type will be close to p_*, provided we restrain the class of events a little further.

A formal statement of the result is as follows.

Theorem 3 (Conditional Limit Theorem)**.** *Let $\mathcal{S} \subset \mathcal{P}^{\mathcal{Y}}$ be a (nonempty) closed and convex set, and let $\mathbf{y} \in \mathcal{Y}^N$. If the elements of $\mathbf{y}$ are i.i.d. according to distribution $q \in \mathcal{P}^{\mathcal{Y}}$ where $q \notin \mathcal{S}$, then as $N \to \infty$,*

$$\lim_{N\to\infty} \mathbb{P}\left(|\hat{p}(b;\mathbf{y}) - p_*(b)| > \epsilon \mid \hat{p}(\cdot;\mathbf{y}) \in \mathcal{S}\right) = 0, \quad \textit{for all } b \in \mathcal{Y}, \tag{41}$$

where p_* is the I-projection defined in (26), i.e., the type of the atypical sequence satisfies $\hat{p}(b;\mathbf{y}) \xrightarrow{\text{p}} p_*(b)$ as $N \to \infty$.

There are three steps to establishing the theorem. First, we show that the set of all the distributions in $\mathcal{S}$ that are close to within divergence $D(p_*\|q)$ of q occurs with high probability. Second, we argue via information geometry that all distributions in this set must also be close in divergence from p_*. Finally, we use that when distributions are close in divergence, their constitutent probabilities are close in absolute difference as well. The latter result is expressed via the following result, whose proof we leave as an exercise.[7]

Lemma 4 (Pinsker's Inequality)**.** *For any two distributions p and q*

$$\|p - q\|_1 \triangleq \sum_{b\in\mathcal{Y}} |p(b) - q(b)| \leq \sqrt{2 \ln 2\, D(p\|q)}. \tag{42}$$

The L_1 distance on the left-hand side of (42) is referred to as the *variational distance* between the distributions p and q. Note, too, that since

$$\max_{b\in\mathcal{Y}} |p(b) - q(b)| \leq \|p - q\|_1,$$

the worst-case difference in symbol probabilities is bounded via (42) as well.

Proof of Theorem 3. First, let us define a neighborhood of types in $\mathcal{S}$ that are closest in divergence to q. In particular,

$$\mathcal{S}_*(\epsilon) = \{p \in \mathcal{P}\colon D(p\|q) \leq D_* + \epsilon\} \cap \mathcal{S}$$

where

$$D_* \triangleq D(p_*\|q) = \arg\min_{p\in\mathcal{S}} D(p\|q).$$

Note that the convexity of $D(p\|q)$ in p ensures D_* is unique.

Let us now examine the relative probabilities of the sets $\mathcal{S}_*(2\delta)$ and $\mathcal{S}\backslash\mathcal{S}_*(2\delta)$ for $\delta > 0$ and show that the latter probability is negligible by comparison.

[7]Note that Pinsker's Inequality ultimately implies that convergence in divergence of a sequence of distributions $p_1, p_2, \ldots$ is stronger than L_1 convergence of the sequence.

Proceeding, we have

$$\begin{aligned} Q\left\{\mathcal{S}\backslash\mathcal{S}_*(2\delta)\right\} &= \sum_{p\in\mathcal{S}\cap\mathcal{P}_N:\ D(p\|q)>D_*+2\delta} Q\left\{\mathcal{T}_N^{\mathcal{Y}}(p)\right\} \\ &\le \sum_{p\in\mathcal{S}\cap\mathcal{P}_N:\ D(p\|q)>D_*+2\delta} 2^{-ND(p\|q)} \qquad (43) \\ &\le \sum_{p\in\mathcal{S}\cap\mathcal{P}_N:\ D(p\|q)>D_*+2\delta} 2^{-N(D_*+2\delta)} \\ &\le (N+1)^{|\mathcal{Y}|}2^{-N(D_*+2\delta)}, \qquad (44) \end{aligned}$$

where to obtain (43) we have used the upper bound in (19), and where to obtain (44) we have used (11).

Next, we have

$$\begin{aligned} Q\left\{\mathcal{S}_*(\delta)\right\} &\ge \sum_{p\in\mathcal{S}\cap\mathcal{P}_N:\ D(p\|q)\le D_*+\delta} Q\left\{\mathcal{T}_N^{\mathcal{Y}}(p)\right\} \\ &\ge \sum_{p\in\mathcal{S}\cap\mathcal{P}_N:\ D(p\|q)\le D_*+\delta} c\,N^{-|\mathcal{Y}|}\,2^{-ND(p\|q)} \qquad (45) \\ &\ge c\,N^{-|\mathcal{Y}|}\,2^{-N(D_*+\delta)} \quad \text{for } N \text{ sufficiently large,} \qquad (46) \end{aligned}$$

where to obtain (45) we have used the lower bound in (19), and where to obtain (46) we have used that for N sufficiently large there is at least one type in $\mathcal{S}_*(\delta)$.

Using (44) and (46) we then see that

$$\begin{aligned} \mathbb{P}_q\left[\hat{p}(\cdot;\mathbf{y})\in\mathcal{S}\backslash\mathcal{S}_*(2\delta)\mid\hat{p}(\cdot;\mathbf{y})\in\mathcal{S}\right] &= \frac{Q\left\{\mathcal{S}\backslash\mathcal{S}_*(2\delta)\right\}}{Q\left\{\mathcal{S}\right\}} \\ &\le \frac{Q\left\{\mathcal{S}\backslash\mathcal{S}_*(2\delta)\right\}}{Q\left\{\mathcal{S}_*(\delta)\right\}} \\ &\le \frac{(N+1)^{|\mathcal{Y}|}2^{-N(D_*+2\delta)}}{c\,N^{-|\mathcal{Y}|}\,2^{-N(D_*+\delta)}} \quad \text{for } N \text{ sufficiently large} \\ &= \frac{(N(N+1))^{|\mathcal{Y}|}}{c}2^{-N\delta}, \end{aligned}$$

so

$$\begin{aligned} \mathbb{P}_q\left[\hat{p}(\cdot;\mathbf{y})\in\mathcal{S}_*(2\delta)\mid\hat{p}(\cdot;\mathbf{y})\in\mathcal{S}\right] &\ge 1-\frac{(N(N+1))^{|\mathcal{Y}|}}{c}2^{-N\delta} \\ &\to 1, \quad \text{as } N\to\infty. \qquad (47) \end{aligned}$$

Hence, with high probability, the observed type will be in $\mathcal{S}_*(2\delta)$.

We next show that all types in this set are close in divergence sense to p_*. In particular, for $p\in\mathcal{S}_*(2\delta)$ we have

$$D(p\|p_*)+D(p_*\|q)\le D(p\|q)\le D_*+2\delta \qquad (48)$$

where the first inequality follows from the information version of Pythagoras' Theorem since $\mathcal{S}_*(2\delta)$ is a closed convex set. In turn, recognizing the second term on the left-hand side of (48) as D_*, we obtain

$$D(p\|p_*) \leq 2\delta. \tag{49}$$

Thus, with high probability, the observed type will be close in divergence to p_*.

It remains only to use that closeness in divergence implies closeness in variational distance, as Lemma 4 establishes. In particular, for any $b \in \mathcal{Y}$ we have

$$|p(b) - p_*(b)| \leq \sum_{b' \in \mathcal{Y}} |p(b') - p_*(b')| \leq \sqrt{2 \ln 2 \, D(p\|p_*)} \leq \sqrt{4\delta \ln 2} \triangleq \epsilon \tag{50}$$

Since ϵ can be made as small as desired by choice of δ, (41) follows. □

23.7 Typical Sets Revisited

One way to connect the method of types to our earlier typical set analysis is through the notion of strong typicality, as we'll now describe. In particular, motivated by (3), consider the following alternative notion of a typical set.

Definition 4 (Strongly Typical Set)**.** *Let* $\mathbf{y} = [y_1, \ldots, y_N]^{\mathrm{T}}$ *be a sequence of* N *elements, each taking a value from an alphabet* $\mathcal{Y}$*, i.e.,* $\mathbf{y} \in \mathcal{Y}^N$*, and let* $\epsilon > 0$ *be a (small) positive constant. The sequence* $\mathbf{y}$ *is called* strongly ϵ-typical *with respect to the probability distribution* p *if for all* $b \in \mathcal{Y}$

$$|\hat{p}(b; \mathbf{y}) - p(b)| \leq \epsilon. \tag{51}$$

A set $\mathcal{T}_\epsilon^{\mathrm{s}}(p; N)$ *of all strongly* ϵ*-typical sequences of length* N *is called the* strongly ϵ-typical set *with respect to the probability distribution* p*.*

With this definition, (3) can be interpreted as telling us that the strongly typical set occurs with high probability.

The strongly typical set is so-named not because it uses the Strong Law of Large Numbers (which it does not), but because strong typicality implies weaker forms of

typicality such as that previously discussed. In particular, since

$$\begin{aligned} L_{p|q}(\mathbf{y}) &= \frac{1}{N} \log \frac{p^N(\mathbf{y})}{q^N(\mathbf{y})} \\ &= \frac{1}{N} \sum_{n=1}^{N} \log \frac{p(y_n)}{q(y_n)} \\ &= \frac{1}{N} \sum_{b \in \mathcal{Y}} N_b(\mathbf{y}) \log \frac{p(b)}{q(b)} \\ &= \sum_{b \in \mathcal{Y}} \hat{p}(b; \mathbf{y}) \log \frac{p(b)}{q(b)} \\ &= \mathbb{E}_{\hat{p}(\cdot;\mathbf{y})} \left[\log \frac{p(y)}{q(y)} \right] \\ &= D(\hat{p}(\cdot;\mathbf{y}) \| q) - D(\hat{p}(\cdot;\mathbf{y}) \| p), \end{aligned} \tag{52}$$

the divergence ϵ-typical sequences $\mathcal{T}_\epsilon(p|q;N)$ are those for which

$$|D(\hat{p}(\cdot;\mathbf{y}) \| q) - D(\hat{p}(\cdot;\mathbf{y}) \| p) - D(p \| q)| \leq \epsilon. \tag{53}$$

Note that our original typical set was just an instance of a divergence typical set corresponding to choosing for q the uniform distribution $\mathtt{U}$, and thus doesn't require a separate discussion. Indeed, $D(p\|\mathtt{U}) = -H(p)$, so $\mathcal{T}_\epsilon(p;N) = \mathcal{T}_\epsilon(p|\mathtt{U};N)$, from which it follows immediately from (53) that those original typical sequences satisfy

$$|H(\hat{p}(\cdot;\mathbf{y})) - H(p) - D(\hat{p}(\cdot;\mathbf{y}) \| p)| \leq \epsilon. \tag{54}$$

While it is sufficient for $\hat{p}(b;\mathbf{y})$ to be close to $p(b)$ for all $b \in \mathcal{Y}$ for a sequence to be typical in the sense of (53), it is not necessary. As a result, (53) characterizes what is referred to as *weakly* ϵ-typical set.

Thus, strong typicality implies (weak) divergence typicality. Specifically, given N and $\epsilon > 0$, then there is an $\epsilon' > 0$ such that $\mathcal{T}^{\mathrm{s}}_{\epsilon'}(p;N) \subset \mathcal{T}_\epsilon(p;N)$. And, further, given q, there is an $\epsilon'' > 0$ such that $\mathcal{T}^{\mathrm{s}}_{\epsilon''}(p;N) \subset \mathcal{T}_\epsilon(p|q;N)$.

The geometry of the typical sets in the probability simplex is useful to visualize. In particular, as depicted in Fig. 2, the strongly typical set is a ball of types centered at p, while the divergence typical set are the types that lie between two parallel affine subspaces straddling p. That this is the geometry of the divergence typical set follows from the fact that since

$$\frac{1}{N} \sum_{n=1}^{N} t(y_n) = \sum_{b=1}^{M} \hat{p}(b;\mathbf{y})\, t(b),$$

for any statistic t, the event

$$\frac{1}{N} \sum_{n=1}^{N} t(y_n) = \mathbb{E}_p\left[y\right]$$

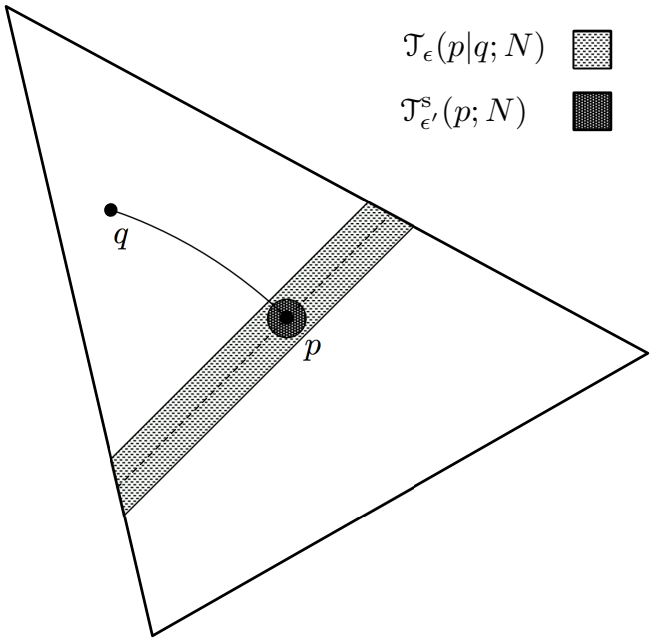

Figure 2: The geometry of the typical sets. The strongly typical set is the darkly shaded ball whose diameter is determined by ϵ'. The divergence typical set is the lightly shaded strip whose width is determined by ϵ.

corresponds to the observed type lying on an affine subspace through p, i.e., the type lying in a linear family, where the orientiation of the family is controlled by the choice of t. The choice

$$t(\cdot) = \log \frac{p(\cdot)}{q(\cdot)}$$

has this orientation controlled by q since p is fixed. More specifically, the curve from q to p represents the one-dimensional exponential family connecting them, and, in turn, the parallel subspace through p defines the corresponding orthogonal linear family.

From the above perspective, in our development of Cramér's Theorem in the last installment of the notes, q was fixed, but $p = p(\cdot; x)$ was chosen according to a canonical exponential family. As a result, the associated orthogonal linear family aligns exactly with the subspace forming the boundary of the large deviation event type set $\mathcal{S}$, and the divergence typical set lies along this boundary just inside $\mathcal{S}$.

Finally, note that in principle the divergence typical set can be defined for continuous distributions since divergence remains well-defined. For the original special case, this requires replacing entropy $H(p)$ with differential entropy $h(p)$. However, the strongly typical set for continuous distributions is less straightforward and generally less useful.

23.8 Further Reading

T. Cover and J. Thomas, *Elements of Information Theory*, provides additional discussion and development of the method of types.

I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, contains a more extensive development of the method of types and its applications.

I. Csiszár, "Information Theory and Statistics: A Tutorial," *Foundations and Trends in Information Theory*, further develops the method of types for applications in statistics.