Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
6.437 Inference and Information
Spring 2015

22 Typical Sequences and Large Deviations

Over the next few installments of the notes, we develop asymptotic analysis tools for our key problems of interest. As we will see, in the regime where the number of observations is relatively large, the behavior of likelihood models—and thus, in turn, hypothesis testing, parameter estimation and inference—often simplifies in important ways, which facilitates efficient inference.

The key insights can all be developed from analysis with comparatively simple models whose observations take the form of i.i.d. samples. Accordingly, we restrict our attention to these models, but note that such analysis can be generalized to include richer classes of models capable of capturing an even broader range of phenomena.

We begin by developing the foundations of such asymptotic analysis. In particular, we focus on understanding what sample sequences are typical of a distribution, and which are atypical. We then turn to quantifying how unlikely atypical sequences are. As a prototype application, we analyze the behavior of sample averages.

We begin with the law of large numbers, which is central to asymptotic analysis.[1]

22.1 The Weak Law of Large Numbers

We will make substantial use of the Weak Law of Large Numbers (WLLN) in our derivations, which expresses how the average of a collection of i.i.d. samples concentrates around the mean of the corresponding distribution.

Theorem 1 (Weak Law of Large Numbers)**.** *Let $\mathsf{w}_1, \dots, \mathsf{w}_N$ be a set of i.i.d. random variables with a mean $\mu < \infty$. Then for any $\epsilon > 0$,*

$$\lim_{N\to\infty} \mathbb{P}\left(\left| \frac{1}{N} \sum_{n=1}^{N} \mathsf{w}_n - \mu \right| > \epsilon \right) = 0, \tag{1}$$

i.e., the sample average converges in probability to μ.

The simplest proof of this result is based on a straightforward application of the Chebyshev inequality, and applies when the underlying distribution has finite variance. To obtain the Chebyshev inequality, we start from the Markov inequality: if v is a nonnegative random variable, then for any $\delta > 0$,[2]

$$\mathbb{P}(\mathsf{v} \ge \delta) \le \frac{\mathbb{E}[\mathsf{v}]}{\delta}. \tag{2}$$

[1] In fact, more generally, one could argue that the law of large numbers is what ultimately makes probability theory itself both meaningful and useful.

[2] Eq. (2) follows from a simple application of chain rule:

$$\mathbb{E}[\mathsf{u}] = \mathbb{P}(\mathsf{u} \ge \delta)\, \mathbb{E}[\mathsf{u} | \mathsf{u} \ge \delta] + \mathbb{P}(\mathsf{u} < \delta)\, \mathbb{E}[\mathsf{u} | \mathsf{u} < \delta] \ge \delta\, \mathbb{P}(\mathsf{u} \ge \delta).$$

The Chebyshev inequality, in turn, follows from the choices $\delta = \epsilon^2$ and $\mathsf{v} = (\mathsf{v} - \mathbb{E}[\mathsf{v}])^2$ in (2), yielding

$$\mathbb{P}(|\mathsf{u} - \mathbb{E}[\mathsf{u}]| > \epsilon) \leq \frac{\mathrm{var}[\mathsf{u}]}{\epsilon^2} \tag{3}$$

for all $\epsilon > 0$. Eq. (1), in turn, follows from making the choice

$$\mathsf{u} = \frac{1}{N} \sum_{n=1}^{N} \mathsf{w}_n$$

and noting that $\mathbb{E}[\mathsf{u}] = \mu$ and $\mathrm{var}[\mathsf{u}] = \mathrm{var}[\mathsf{w}]/N$. It is worth emphasizing, however, that more elaborate proofs exist that do not require a finite variance assumption.

The WLLN does not say anything about the rate at which the sample average converges to its expectation. Use of the Chebyshev inequality as above establishes that the rate is at least as fast as $1/N$. In fact, as we will develop later in this installment of the notes, the convergence is much faster than this—in fact, the rate is exponential in the number of samples N.

22.2 Typical Sequences

The law of large numbers can be used to deduce other properties of i.i.d. samples from a distribution. In particular, as we now develop, it tells us about which sequences are typical of the distribution. Typical sequences are the ones that appear with high probability as we generate i.i.d. samples. In the sequel, we will emphasize the case of discrete distributions. However, the basic concepts also apply to continuous distributions, as we'll describe at the end of the section.

To gain intuition about typicality, consider a sequence $\mathbf{y} = [y_1, \ldots y_N]^{\mathrm{T}}$ generated i.i.d. from a distribution p. The corresponding normalized log-likelihood is given by[3]

$$L_p(\mathbf{y}) = \frac{1}{N} \log p_{\mathbf{y}}(\mathbf{y}) = \frac{1}{N} \log \prod_{n=1}^{N} p(y_n) = \frac{1}{N} \sum_{n=1}^{N} \log p(y_n). \tag{4}$$

The sequence of log-likelihood values $\{\log p(y_n)\}_{n=1}^{N}$ can itself be viewed as a collection of i.i.d. samples of a new random variable $\mathsf{w} = \log p(\mathsf{y})$. Moreover, since

$$\mathbb{E}[\mathsf{w}] = \mathbb{E}_p[\log p(\mathsf{y})] = -H(p), \tag{5}$$

with $H(p)$ denoting, as usual, the entropy of the distribution p, it follows from the WLLN that for any $\epsilon > 0$,

$$\lim_{N \to \infty} \mathbb{P}(|L_p(\mathbf{y}) + H(p)| > \epsilon) = 0, \tag{6}$$

i.e., the distribution of the average log-likelihood values concentrates around $-H(\mathsf{y})$ as the sequence length increases. We now formalize the notion of a typical sequence.

[3] Throughout our development of asymptotic analysis, all logarithms will be base 2, for convenience.

Definition 1 (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* ϵ-typical *with respect to the probability distribution* p *if*

$$|L_p(\mathbf{y}) + H(p)| \leq \epsilon. \tag{7}$$

A set of all ϵ*-typical sequences of length* N*, i.e.,*

$$\mathcal{T}_\epsilon(p; N) = \left\{ \mathbf{y} \in \mathcal{Y}^N \colon \ |L_p(\mathbf{y}) + H(p)| \leq \epsilon \right\}, \tag{8}$$

is called the ϵ-typical set *with respect to the probability distribution* p*.*

The typical set has some interesting properties. First, (6) establishes directly that a sequence from the typical set is generated with high probability; specifically,

$$\mathbb{P}\left(\mathbf{y} \in \mathcal{T}_\epsilon(p; N)\right) \cong 1 \qquad \text{for large } N. \tag{9}$$

Second, rearranging (4) and using that by definition the typical sequences $\mathbf{y}$ satisfy $L_p(\mathbf{y}) \cong -H(p)$, we obtain

$$p_{\mathsf{y}}(\mathbf{y}) \cong 2^{-NH(p)}, \tag{10}$$

i.e., the typical sequences are about equiprobable and each occurs with a probability of roughly $2^{-NH(p)}$. But then (9) and (10) together imply that there must be about $2^{NH(p)}$ typical sequences, i.e.,

$$|\mathcal{T}_\epsilon(p; N)| \cong 2^{NH(p)}. \tag{11}$$

Now for a distribution over an alphabet $\mathcal{Y}$ of size $M = |\mathcal{Y}|$ there are $M^N = 2^{N \log M}$ possible sequences. Thus whenever $H(p) < \log M$, only a vanishingly small (specifically, exponentially small) fraction of the possible sequences occur with non-negligible probability.

These results have substantial implications, as we'll develop. For the moment, however, note that they provide a new and important operational interpretation of entropy as characterizing the (exponential) size of the typical set.

Before developing the broader implications, we first need to more precisely characterize these approximations. To do so, it will be convenient to introduce a bit of additional notation that simplifies our derivations. First, we let $\mathcal{P}^{\mathcal{Y}}$ denote the probability simplex (i.e., all distributions) associated with the alphabet $\mathcal{Y}$. When there is no risk of confusion, we omit the superscript and let the alphabet be implicit.

Second, we use p^N to denote the distribution for a sequence of N i.i.d. variables distributed according to p, i.e., with $\mathbf{y} = [y_1, \ldots, y_N]^{\mathrm{T}}$,

$$p^N(\mathbf{y}) = \prod_{n=1}^{N} p(y_n), \tag{12}$$

and similarly for q^N.

Third, we use $P\left\{\mathcal{T}\right\}$ to denote the total probability assigned under $p \in \mathcal{P}^{\mathcal{Y}}$ to sequences in a set $\mathcal{A}$:

$$P\left\{\mathcal{A}\right\} \triangleq \sum_{\mathbf{y}\in\mathcal{A}} p^N(\mathbf{y}). \tag{13}$$

In other words $P\left\{\mathcal{A}\right\}$ denotes the probability of generating a sequence from the set $\mathcal{A}$ if we generate samples using the distribution p, i.e., $P\left\{\mathcal{A}\right\} = \mathbb{P}_p\left[\mathbf{y} \in \mathcal{A}\right]$, where the subscript of $\mathbb{P}$ emphasizes the dependence on p. Similarly, for an arbitrary distribution $q \in \mathcal{P}^{\mathcal{Y}}$,

$$Q\left\{\mathcal{A}\right\} \triangleq \sum_{\mathbf{y}\in\mathcal{A}} q^N(\mathbf{y}). \tag{14}$$

Equipped with this notation, the following theorem expresses the relevant properties of typical sequences with suitable precision for our needs.

Theorem 2 (Asymptotic Equipartition Property (AEP))**.** *Let $\mathcal{T}_\epsilon(p;N)$ be the ϵ-typical set with respect to the probability distribution p for a given $\epsilon > 0$. Then*

$$\begin{aligned} \lim_{N\to\infty} P\left\{\mathcal{T}_\epsilon(p;N)\right\} &= 1 && \text{(AEP-1)} \\ 2^{-N(H(p)+\epsilon)} \le p^N(\mathbf{y}) &\le 2^{-N(H(p)-\epsilon)} \quad \textit{for all } \mathbf{y} \in \mathcal{T}_\epsilon(p;N) && \text{(AEP-2)} \\ (1-\epsilon)2^{N(H(p)-\epsilon)} \le \left|\mathcal{T}_\epsilon(p;N)\right| &\le 2^{N(H(p)+\epsilon)} \quad \textit{for a sufficiently large } N. && \text{(AEP-3)} \end{aligned}$$

Proof. First, (AEP-1) follows immediately from (9).

Next, consider an ϵ-typical sequence $\mathbf{y} \in \mathcal{T}_\epsilon(p;N)$ and use (7) to obtain

$$-H(p) - \epsilon \le L_p(\mathbf{y}) \le -H(p) + \epsilon,$$

which when we substitute for $L_p(\mathbf{y})$ via (4) yields

$$-H(p) - \epsilon \le \frac{1}{N}\log p^N(\mathbf{y}) \le -H(p) + \epsilon. \tag{15}$$

Finally, scaling and exponentiating (15) we obtain (AEP-2).

To obtain the upper bound in (AEP-3), we write

$$\begin{aligned} \left|\mathcal{T}_\epsilon(p;N)\right| &= \sum_{\mathbf{y}\in\mathcal{T}_\epsilon(p;N)} 1 \\ &= 2^{N(H(p)+\epsilon)} \sum_{\mathbf{y}\in\mathcal{T}_\epsilon(p;N)} 2^{-N(H(p)+\epsilon)} \\ &\le 2^{N(H(p)+\epsilon)} \sum_{\mathbf{y}\in\mathcal{T}_\epsilon(p;N)} p^N(\mathbf{y}) \\ &= 2^{N(H(p)+\epsilon)}\, P\left\{\mathcal{T}_\epsilon(p;N)\right\} \\ &\le 2^{N(H(p)+\epsilon)}, \end{aligned}$$

where to obtain the first inequality we used the lower bound in (AEP-2).

To obtain the lower bound in (AEP-3), we write

$$
\begin{aligned}
|\mathcal{T}_\epsilon(p;N)| &= \sum_{\mathbf{y}\in\mathcal{T}_\epsilon(p;N)} 1 \\
&= 2^{N(H(p)-\epsilon)} \sum_{\mathbf{y}\in\mathcal{T}_\epsilon(p;N)} 2^{-N(H(p)-\epsilon)} \\
&\geq 2^{N(H(p)-\epsilon)} \sum_{\mathbf{y}\in\mathcal{T}_\epsilon(p;N)} p^N(\mathbf{y}) \\
&= 2^{N(H(p)-\epsilon)} P\left\{\mathcal{T}_\epsilon(p;N)\right\} \\
&\geq (1-\epsilon) 2^{N(H(p)-\epsilon)},
\end{aligned}
$$

where to obtain the first inequality we have used the upper bound in (AEP-2), and where to obtain the last inequality we have used that (AEP-1) implies that $P\left\{\mathcal{T}_\epsilon(p;N)\right\} \geq 1-\epsilon$ for sufficiently large N. □

22.2.1 Continuous Distributions

The concept of typical sequences extends naturally to samples from continuous distributions. In such cases, since (5) becomes

$$\mathbb{E}\left[\mathsf{w}\right] = \mathbb{E}_p\left[\log p(\mathsf{y})\right] = -h(p),$$

with $h(p)$ denoting the differential entropy of p, it follows that the typical sequences are those whose normalized log-likelihoods are close to $-h(p)$.

In turn, the typical set is similarly well defined. And although, of course, its cardinality is infinite in this case, it has a well defined *volume.* In particular, [cf. (11)], the volume of the typical set satisfies

$$\text{vol}[\mathcal{T}_\epsilon(p;N)] \cong 2^{Nh(p)}. \tag{16}$$

since $p^N(\mathbf{y}) \cong 2^{-Nh(p)}$. The continuous version of the AEP (Theorem 2) is thus obtained by replacing $|\cdot|$ with $\text{vol}[\cdot]$ and $H(p)$ with $h(p)$, and the proof requires simply replacing summation with integration in the appropriate steps.

22.3 Atypical Sequences and Large Deviations Analysis

The sequences not in the typical set can be viewed as “atypical” ones, and while they occur rarely, when they do they give rise to anomolous behavior. Such phenomena are characterized using *large deviations analysis,* which now develop. Among other results, we will show that atypical events are exponentially unlikely, and characterize the associated exponent.

To develop our main results, let us develop an alternative characterization of the typical set. Our develope applies equally well to both discrete and continuous distributions. In particular, consider a sequence $\mathbf{y} = [y_1, \dots y_N]^{\mathrm{T}}$ generated i.i.d. according to a distribution p over an alphabet $\mathcal{Y}$. But now also consider a second reference distribution q, and form a normalized log-likelihood ratio

$$L_{p|q}(\mathbf{y}) = \frac{1}{N} \log \frac{p^N(\mathbf{y})}{q^N(\mathbf{y})} = \frac{1}{N} \sum_{n=1}^{N} \log \frac{p(y_n)}{q(y_n)}. \tag{17}$$

As before, (17) the average of a sequence of N i.i.d. samples of a random variable $\mathsf{w} = \log(p(\mathsf{y})/q(\mathsf{y}))$ whose mean is $\mathbb{E}_p[\mathsf{w}] = D(p\|q)$, where we add the expectation subscript to remind ourselves that p is the distribution generating the data. Applying the WLLN in this case yields

$$\lim_{N\to\infty} \mathbb{P}\left(\left|L_{p|q}(\mathbf{y}) - D(p\|q)\right| > \epsilon\right) = 0, \tag{18}$$

which leads to the following essentially equivalent description of the typical set.

Definition 2 (Typical Set, Another Characterization). *Let* $\mathbf{y} = [y_1, \dots y_N]^{\mathrm{T}}$ *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* divergence ϵ-typical *with respect to a distribution* $p \in \mathcal{P}^{\mathcal{Y}}$ *relative to a reference distribution* $q \in \mathcal{P}^{\mathcal{Y}}$ *if*

$$\left|L_{p|q}(\mathbf{y}) - D(p\|q)\right| \le \epsilon. \tag{19}$$

The set of all divergence ϵ*-typical sequences of length* N

$$\mathcal{T}_\epsilon(p|q;N) = \left\{\mathbf{y} \in \mathcal{Y}^N \colon \left|L_{p|q}(\mathbf{y}) - D(p\|q)\right| \le \epsilon\right\} \tag{20}$$

is called a divergence ϵ-typical set *with respect to* p *relative to* q*.*

It should be emphasized that the two notions of typical set are not strictly equivalent. Since

$$L_{p|q}(\mathbf{y}) - D(p\|q) = (L_p(\mathbf{y}) + H(p)) - \left(\frac{1}{N}\sum_{n=1}^{N} \log q(y_n) - \mathbb{E}_p[\log q(\mathsf{y})]\right),$$

it is evident that no matter how ϵ and ϵ' are chosen such that $\mathcal{T}_\epsilon(p;N) \ne \mathcal{Y}^N$ and $\mathcal{T}'_\epsilon(p|q;N) \ne \mathcal{Y}^N$, neither typical set is a subset of the other. However, the two notions of typicality are essentially equivalent in one important sense. In particular, first note that (18) means that sequences of samples from p will also fall in the divergence typical set with high probability. Then note that if we have two sets whose probabilities are close to one, their intersection must also occur with probability close to one.[4] This means that with high probability, a sequence of samples from p will fall into both typical sets. Hence, the sequences that the two sets do not have in common occur with negligible probability under p, even though there may be many of them. It is this equivalent behavior of the two sets under p that is more relevant to our analysis.

The value of this alternative characterization of typicality is that we can readily characterize how likely it is that sampling a different distribution, say q, will produce sequences typical with respect to p. In particular, rearranging (17) and using that sequences $\mathbf{y}$ that are typical with respect to p satisfy $L_{p|q}(\mathbf{y}) \cong D(p\|q)$, we obtain

$$q^N(\mathbf{y}) \cong p^N(\mathbf{y}) 2^{-ND(p\|q)},$$

which when we sum over all $\mathbf{y}$ in the typical set yields

$$Q\left\{\mathcal{T}_\epsilon(p|q;N)\right\} \cong 2^{-ND(p\|q)}. \tag{21}$$

The implication of (21) is worth emphasizing: the probability that sampling from a distribution q will produce *any* sequence in the typical set for p is exponentially small, with the rate given by the information divergence of q from p.

Evidently, our atypical sequence analysis yields a new operational interpretation of information divergence.

If we visual the total collection of $2^{N\log\mathcal{Y}}$ possible sequences, our results thus establish that: i) the typical set for every (nonunform) distribution includes only a vanishing subset of the collection; and ii) the typical sets for different (nonuniform) distributions are essentially disjoint.[5]

As before, it is useful to make the approximate reasoning described above more precise, which the following theorem accomplishes.

Theorem 3. *If $\mathcal{T}_\epsilon(p|q;N)$ be the divergence ϵ-typical set with respect to p relative to q, then*

$$(1-\epsilon)2^{-N(D(p\|q)+\epsilon)} \le Q\left\{\mathcal{T}_\epsilon(p|q;N)\right\} \le 2^{-N(D(p\|q)-\epsilon)} \quad \textit{for a sufficiently large } N. \tag{22}$$

Proof. From (19) we obtain that every $\mathbf{y} \in \mathcal{T}_\epsilon(p|q;N)$ satisfies

$$D(p\|q) - \epsilon \le L_{p|q}(\mathbf{y}) \le D(p\|q) + \epsilon$$

[4] Formally, suppose we have sets $\mathcal{A}$ and $\mathcal{B}$ such that $\mathbb{P}(\mathcal{A}) \ge 1-\delta$ and $\mathbb{P}(\mathcal{B}) \ge 1-\delta$ for some small $\delta > 0$. Then

$$\mathbb{P}(\mathcal{A}\cap\mathcal{B}) = \mathbb{P}(\mathcal{A}) + \mathbb{P}(\mathcal{B}) - \mathbb{P}(\mathcal{A}\cup\mathcal{B}) \ge (1-\delta) + (1-\delta) - 1 = 1-2\delta.$$

[5] More accurately, any collection of typical sets is effectively disjoint if N is chosen sufficiently large.

which when we substitute for $L_{p|q}(\mathbf{y})$ using (17) we obtain, after some rearrangement,

$$p^N(\mathbf{y})2^{-N(D(p\|q)+\epsilon)} \leq q^N(\mathbf{y}) \leq p^N(\mathbf{y})2^{-N(D(p\|q)-\epsilon)}. \tag{23}$$

Summing[6] (23) over all sequences $\mathbf{y} \in \mathcal{T}_\epsilon(p|q;N)$, we obtain

$$P\left\{\mathcal{T}_\epsilon(p|q;N)\right\}2^{-N(D(p\|q)+\epsilon)} \leq Q\left\{\mathcal{T}_\epsilon(p|q;N)\right\} \leq P\left\{\mathcal{T}_\epsilon(p|q;N)\right\}2^{-N(D(p\|q)-\epsilon)}. \tag{24}$$

But (18) implies that $P\left\{\mathcal{T}_\epsilon(p|q;N)\right\} \geq 1-\epsilon$ for sufficiently large N, allowing us to further bound (24), yielding (22). □

Let's develop a simple but useful application of Theorem 3 before turning to our main results in the next section.

Example 1. Suppose $p_{\mathsf{x},\mathsf{y}}$ is a (joint) distribution over $\mathcal{X} \times \mathcal{Y}$, with marginals p_x and p_y. And suppose we draw samples x_n according to the marginal p_x, and independently draw samples y_n according to p_y, i.e., $(\mathsf{x}_n, \mathsf{y}_n)$ are drawn according to the product distribution $q_{\mathsf{x},\mathsf{y}}(x,y) = p_\mathsf{x}(x)\,p_\mathsf{y}(y)$. Then using (22), the probability that these samples will be typical of $p_{\mathsf{x},\mathsf{y}}$ is, for sufficiently large N, roughly

$$Q\left\{\mathcal{T}_\epsilon(p|q;N)\right\} \cong 2^{-N(D(p\|q))}$$

where

$$D(p\|q) = D(p_{\mathsf{x},\mathsf{y}} \parallel p_\mathsf{x}\,p_\mathsf{y}) = I(\mathsf{x};\mathsf{y}),$$

where $I(\mathsf{x};\mathsf{y})$ denotes, as usual, mutual information. As such, this example provides a new and useful operational interpretation of mutual information.

22.4 Large Deviations of Sample Averages

As an important example application of our atypical sequence development, here we characterize *how* sample averages converge to the mean. Specifically, an immediate consequence of the WLLN is that

$$\lim_{N\to\infty} \mathbb{P}\left(\left|\frac{1}{N}\sum_{n=1}^{N} \mathsf{y}_n - \mu\right| \leq \epsilon\right) = 1. \tag{25}$$

when $\mathbf{\mathsf{y}} = [\mathsf{y}_1, \ldots, \mathsf{y}_N]^\mathrm{T}$ is a sequence of N i.i.d. random variables with mean μ.

To understand the associated rate of convergence, we characterize the *large deviation probability*

$$\mathbb{P}\left(\frac{1}{N}\sum_{n=1}^{N} \mathsf{y}_n \geq \gamma\right) \tag{26}$$

[6] In the case of continuous distributions, summation is replaced with integration.

for an arbitrary $\gamma > \mu$, when N is large. The key to such analysis is recognizing that since the sample average is close to μ when the samples are typical of q, when the sample average is different from μ, the samples must be atypical of q and thus typical of a different distribution p. Our analysis will reveal this exponentially unlikely p.

Interestingly, although there are many atypical distributions that can generate the large deviation event, one dominates all the others. In particular, it is one that simultaneously is an element of the one-dimensional canonical exponential family that includes q and has a mean arbitrarily close to γ. Among other implications, this means that although the large deviation event (26) corresponds to the sample average taking on one of an infinite number of possible values (corresponding to the range $[\gamma, \infty)$), the dominant atypical event is

$$\frac{1}{N}\sum_{n=1}^{N} y_n \cong \gamma.$$

The results of such an analysis are expressed by the following theorem, whose proof we emphasize in advance applies to both discrete and continuous distributions.

Theorem 4 (Cramér's Theorem)**.** *If $\mathsf{y} = [\mathsf{y}_1, \ldots, \mathsf{y}_N]^{\mathrm{T}}$ is a sequence of N i.i.d. random variables generated from a distribution q with mean $\mu < \infty$, then for any $\gamma > \mu$,*[7]

$$\lim_{N\to\infty} -\frac{1}{N}\log \mathbb{P}\left(\frac{1}{N}\sum_{n=1}^{N} \mathsf{y}_n \geq \gamma\right) = E_{\mathrm{C}}(\gamma), \tag{27}$$

where $E_{\mathrm{C}}(\gamma)$ is referred to as the Chernoff *exponent and is defined via*

$$E_{\mathrm{C}}(\gamma) \triangleq D(p(\cdot;x)\|q) \tag{28a}$$

with

$$p(y;x) = q(y)\, e^{xy-\alpha(x)}, \tag{28b}$$

and with $x > 0$ chosen such that

$$\mathbb{E}_{p(\cdot;x)}\left[\mathsf{y}\right] = \gamma. \tag{28c}$$

Before proceeding to a proof, it is worth emphasizing that this theorem expresses that the following large N approximation

$$\mathbb{P}\left(\frac{1}{N}\sum_{n=1}^{N} \mathsf{y}_n \geq \gamma\right) \cong 2^{-NE_{\mathrm{C}}(\gamma)}.$$

[7] Since the risk of confusion is small, We omit the subscript q in using the notation $\mathbb{P}(\cdot)$.

is accurate to first order in the exponential scaling with N.[8] In fact, for any $\epsilon > 0$ and all N sufficiently large,

$$2^{-N(E_{\mathrm{C}}(\gamma)+\epsilon)} \leq \mathbb{P}\left(\frac{1}{N}\sum_{n=1}^{N} \mathsf{y}_n \geq \gamma\right) \leq 2^{-NE_{\mathrm{C}}(\gamma)},$$

the upper bound of which is established in our proof. The lower bound requires slightly more refined analysis than we have included in our proof.

Proof. We start by deriving an upper bound on the large deviation probability. In particular, note for any constant $x > 0$ we have

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{N}\sum_{n=1}^{N} \mathsf{y}_n \geq \gamma\right) &= \mathbb{P}\left(e^{x\sum_{n=1}^{N}\mathsf{y}_n} \geq e^{Nx\gamma}\right) \\
&\leq e^{-Nx\gamma}\,\mathbb{E}\left[e^{x\sum_{n=1}^{N}\mathsf{y}_n}\right] \qquad (29)\\
&= e^{-Nx\gamma}(\mathbb{E}\left[e^{x\mathsf{y}}\right])^N \qquad (30)\\
&= e^{-N(x\gamma-\alpha(x))}, \qquad (31)\\
&\leq e^{-N(x_*\gamma-\alpha(x_*))}, \qquad (32)
\end{aligned}$$

where to obtain (29) we have used the Markov inequality (2), where to obtain (30) we have used that the y_n are i.i.d. variables, where in (31)

$$\alpha(x) \triangleq \ln \mathbb{E}\left[e^{x\mathsf{y}}\right], \qquad (33)$$

and where in (32)

$$x_* = \underset{x>0}{\arg\max}\,\varphi(x), \qquad \varphi(x) \triangleq x\gamma - \alpha(x).$$

At this point, we recognize (33) as the log-partition function of the canonical exponential family (28b), where we note that $x > 0$ "tilts" the distribution q so that the mean of $p(\cdot; x)$ is larger that of q, i.e., μ.

In fact, (32) corresponds to tilting the distribution so the new mean is γ. To see this, note first that

$$\frac{\mathrm{d}}{\mathrm{d}x}\varphi(x) = \gamma - \dot{\alpha}(x) \qquad (34)$$

and

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\varphi(x) = -\ddot{\alpha}(x), \qquad (35)$$

[8] We will revisit this notion of approximation in more detail in a subsequent installment of the notes.

Then recall that for the distributions defined via (28b) we have $\lim_{x\to 0} \mathbb{E}_{p(\cdot;x)}[\mathsf{y}] = \mu$ and, from our earlier development of exponential families,

$$\mathbb{E}_{p(\cdot;x)}[\mathsf{y}] = \dot{\alpha}(x), \tag{36}$$

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{E}_{p(\cdot;x)}[\mathsf{y}] = \ddot{\alpha}(x) = \mathrm{var}_{p(\cdot;x)}(\mathsf{y}) > 0 \tag{37}$$

for any distribution q that does not assign all probability to a single value.

In turn, (37) implies that $\varphi(x)$ is concave, and thus maximized either in the limits $x \to 0$ or $x \to \infty$, or at the unique stationary point corresponding to setting (34) to zero, i.e., $\dot{\alpha}(x) = \gamma$. But from (37) we see that $\mathbb{E}_{p(\cdot;x)}[\mathsf{y}]$ is a strictly increasing function of $x > 0$, and, moreover, a one-to-one (i.e., invertible) function of x. Hence, for $\gamma > \mu$ it must be that x_* is the unique positive value such that

$$\mathbb{E}_{p(\cdot;x_*)}[\mathsf{y}] = \dot{\alpha}(x_*) = \gamma. \tag{38}$$

Next, we note that the resulting exponent in (32) can be expressed in the form of a divergence. In particular, since

$$\frac{D(p(\cdot;x)\|q)}{\log(e)} = \mathbb{E}_{p(\cdot;x)}\left[\ln\frac{p(\mathsf{y};x)}{q(\mathsf{y})}\right] = x\,\mathbb{E}_{p(\cdot;x)}[\mathsf{y}] - \alpha(x) = x\dot{\alpha}(x) - \alpha(x), \tag{39}$$

using (38) we get

$$x_*\gamma - \alpha(x_*) = x_*\dot{\alpha}(x_*) - \alpha(x_*) = \frac{D(p(\cdot;x_*)\|q)}{\log(e)}, \tag{40}$$

and thus (32) can be equivalently expressed as

$$\mathbb{P}\left(\frac{1}{N}\sum_{n=1}^{N}\mathsf{y}_n \geq \gamma\right) \leq 2^{-NE_{\mathrm{C}}(\gamma)}. \tag{41}$$

where, in accordance with (28),

$$E_{\mathrm{C}}(\gamma) = D(p(\cdot;x_*)\|q).$$

Next, we turn to lower bounding the large deviation probabilty. Here we make use of our results on atypical sequences. In particular, one of the (many) ways the sample average can end up being at least γ is when the generated sequence is divergence $x_\epsilon\epsilon$-typical of $p(\cdot;x_\epsilon)$ with x_ϵ chosen such that[9]

$$\mathbb{E}_{p(\cdot;x_\epsilon)}[\mathsf{y}] = \dot{\alpha}(x_\epsilon) = \gamma + \epsilon. \tag{42}$$

[9] It should be emphasized that there are many other choices of distribution with this mean what also represent ways of obtaining such sample average behavior, but would ultimately lead to looser lower bounds; i.e., our choice turns out to be special.

To see this, note that for any sequence in the set $\mathbf{y} \in \mathcal{T}_{x_\epsilon \epsilon}(p(\cdot; x_\epsilon)|q; N)$ we have, using that

$$L_{p(\cdot;x)|q}(\mathbf{y}) = \frac{1}{N} \sum_{n=1}^{N} \log \frac{p(y_n)}{q(y_n)} = x \left(\frac{1}{N} \sum_{n=1}^{N} y_n \right) - \alpha(x) \tag{43}$$

in (19),

$$-x_\epsilon \epsilon \leq x_\epsilon \frac{1}{N} \sum_{n=1}^{N} y_n - x_\epsilon \, \mathbb{E}_{p(\cdot;x_\epsilon)} [y] \leq x_\epsilon \epsilon.$$

which using (42) yields, after some rearrangement,

$$\gamma \leq \frac{1}{N} \sum_{n=1}^{N} y_n \leq \gamma + 2\epsilon. \tag{44}$$

Hence,

$$\begin{aligned} &\mathbb{P} \left(\frac{1}{N} \sum_{n=1}^{N} y_n \geq \gamma \right) \\ &\qquad \geq \mathbb{P}_q \left[\frac{1}{N} \sum_{n=1}^{N} y_n \geq \gamma \;\middle|\; \mathbf{y} \in \mathcal{T}_{x_\epsilon \epsilon}(p(\cdot; x_\epsilon)|q; N) \right] Q \left\{ \mathcal{T}_{x_\epsilon \epsilon}(p(\cdot; x_\epsilon)|q; N) \right\} \\ &\qquad = Q \left\{ \mathcal{T}_{x_\epsilon \epsilon}(p(\cdot; x_\epsilon)|q; N) \right\}. \end{aligned} \tag{45}$$

Thus, it remains only to lower bound (45) using Theorem 3. Proceeding, we have

$$\begin{aligned} Q \left\{ \mathcal{T}_{x_\epsilon \epsilon}(p(\cdot; x_\epsilon)|q; N) \right\} &\geq (1 - x_\epsilon \epsilon) 2^{-N(D(p(\cdot;x_\epsilon)\|q) + x_\epsilon \epsilon)} && (46) \\ &= (1 - x_\epsilon \epsilon) 2^{-N(E_{\mathrm{C}}(\gamma+\epsilon) + x_\epsilon \epsilon)}, && (47) \end{aligned}$$

where (46) follows from (22), and where (47) follows from the fact that in accordance with (28), (42) implies

$$D(p(\cdot; x_\epsilon)\|q) = E_{\mathrm{C}}(\gamma + \epsilon).$$

Using (47) with (45), we obtain

$$\begin{aligned} \lim_{N\to\infty} -\frac{1}{N} \log \mathbb{P} \left(\frac{1}{N} \sum_{n=1}^{N} y_n \geq \gamma \right) &\leq \lim_{N\to\infty} -\frac{1}{N} \log(1 - x_\epsilon \epsilon) + E_{\mathrm{C}}(\gamma + \epsilon) + x_\epsilon \epsilon \\ &= E_{\mathrm{C}}(\gamma + \epsilon) + x_\epsilon \epsilon \end{aligned}$$

which can be made arbitrarily close to $E_{\mathrm{C}}(\gamma)$ by choosing ϵ small enough.

Finally, since our upper and lower bounds coincide, we obtain (27).

□

Note that since the large deviation events for progressively larger γ are nested, we would expect that $E_{\mathrm{C}}(\gamma)$ to be a nondecreasing function of γ. This is indeed true, but we can make a stronger statement. In particular, note that the divergence of q from $p(\cdot;x)$ is a strictly monotonic and one-to-one function of x. Indeed, we have, via (39),

$$\frac{\mathrm{d}}{\mathrm{d}x} D(p(\cdot;x)\|q)/\log(e) = \dot{\alpha}(x) + x\ddot{\alpha}(x) - \dot{\alpha}(x) = x\ddot{\alpha}(x) > 0 \quad \text{for all } x > 0.$$

In addition, as developed in the course of the proof of Theorem 4, x as defined by (28c) is a strictly monotonic and one-to-one function of γ. Thus, the composition $E_{\mathrm{C}}(\cdot)$ must be monotonic, one-to-one function of its argument, and we deduce that the Chernoff exponent strictly and smoothly increases with γ.

As a final comment, note that Theorem 4 suggests that (28) describes how the large deviation event happens when it does. In particular, it suggests that when the large deviation event happens, the empirical distribution of the data is given by (28b) with x given by (28c). In fact, although under the right conditions this is true in an appropriate sense, Theorem 4 doesn't actually establish this. To do so requires further work, which takes the form of what is referred to as a conditional limit theorem. We will discuss such a theorem in the next installment of the notes when we revisit large deviation analysis using different tools.

We conclude the section with some simple examples.

Example 2. Suppose $\mathsf{y}_1, \mathsf{y}_2, \ldots$ are i.i.d. according to q, where q is the unit Gaussian distribution $\mathtt{N}(0,1)$, i.e.,

$$q(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2},$$

and we want to estimate the large deviation probability with $\gamma > 0$ using Theorem 4. Applying (28b), note that

$$p(y;x) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2} e^{xy} e^{-\alpha(x)}, \tag{48}$$

which is an exponentiated quadratic and therefore Gaussian. Completing the square in the exponent of (48), we get

$$p(y;x) = \frac{1}{\sqrt{2\pi}} e^{-(y-x)^2/2},$$

where we note $\alpha(x) = \gamma^2/2$ is the required normalization; hence, p is $\mathtt{N}(x,1)$. The constraint (28c) establishes that $x = \gamma$.

Using (40), we obtain the Chernoff exponent as

$$E_{\mathrm{C}}(\gamma) = D(\mathtt{N}(\gamma,1)\|\mathtt{N}(0,1)) = x\gamma - \alpha(x) = \gamma^2 - \frac{1}{2}\gamma^2 = \frac{1}{2}\gamma^2. \tag{49}$$

Hence, the probability that the sample average of N unit Gaussian random variables will be at least γ is

$$\mathbb{P}\left(\frac{1}{N}\sum_{n=1}^{N} \mathsf{y}_n \geq \gamma\right) \cong 2^{-N\gamma^2/2},$$

and, furthermore, when this large deviation event happens, Theorem 4 suggests that the overwhelmingly most likely way is when the samples are still typical of a Gaussian distribution, but one with a mean of γ instead of 0.

Example 3. Let's repeat Example 2 with q now an exponential distribution with unit-mean, i.e.,

$$q(y) = e^{-y}, \quad y \geq 0,$$

and consider the large deviation probability for $\gamma > 1$. Then

$$p(y; x) = e^{-y}\, e^{xy}\, e^{-\alpha(x)}$$

is evidently also exponential, but now with mean $1/(1-x)$, so $x = 1 - 1/\gamma$, and $\alpha(x) = -\ln(1-x)$ is the corresponding normalization. Hence,

$$E_{\mathrm{C}}(\gamma) = (\gamma - 1) + \ln\gamma.$$

so

$$\mathbb{P}\left(\frac{1}{N}\sum_{n=1}^{N} \mathsf{y}_n \geq \gamma\right) \cong 2^{-N[(\gamma-1)+\ln(\gamma)]},$$

Moreover, when this large deviation event happens, Theorem 4 suggests that the dominant way is when the samples are still typical of an exponential distribution, but one with a mean of *gamma* instead of 1.

Example 4. Now suppose the distribution is uniform over the interval $[0, 1]$, and consider the large deviation probability for $\gamma > 1/2$. Then

$$p(y; x) = \begin{cases} e^{xy}\, e^{-\alpha(x)}, & y \in [0, 1] \\ 0, & \text{otherwise}, \end{cases}$$

which is not uniform. The normalization is straightforward to calculate as

$$\alpha(x) = \ln(e^x - 1) - \ln x$$

and the mean is

$$\mathbb{E}_{p(\cdot;x)}[\mathsf{y}] = 1 + \frac{1}{e^x - 1} - \frac{1}{x},$$

which can be solved numerically for x. In turn, the associated Chernoff exponent is, in terms of x,

$$E_{\mathrm{C}}(\gamma) = x\gamma - \alpha(x) = x + \frac{x}{e^x - 1} - 1 - \ln(e^x - 1) + \ln x.$$

As an alternative to numerical solutions for x, we can use simple but fairly good approximations in different regimes. For example, the approximation

$$x \cong 12(\gamma - 1/2) \tag{50}$$

is correct in the limit $\gamma \to 1/2$, and more generally accurate for $1/2 < \gamma \lesssim 2/3$. Using (50) we get the following Chernoff exponent approximation

$$E_{\mathrm{C}}(\gamma) \cong 12\gamma \left(\gamma - \frac{1}{2}\right)^2$$

for this regime. Evidently, if N is only moderately large, observing sample means close to but greater than 1/2 is not quite so unlikely, since the exponent grows slowly from 0 with increasing γ.

By contrast, the approximation

$$x \cong \frac{1}{1-\gamma} \tag{51}$$

is correct in the limit $\gamma \to 1$, and more generally accurate for $4/5 \lesssim \gamma < 1$. Using (51) we get the following Chernoff exponent approximation

$$E_{\mathrm{C}}(\gamma) \cong -\ln(1-\gamma) - 1$$

for this regime. Evidently, it is overwhelmingly unlikely that we will observe a sample mean greater than or equal to 1, since the associated exponent becomes infinite as γ approaches 1.

22.5 Extensions

The above analysis of large deviations of samples averages can be extended in some key ways that make it useful to a broader range of applications. First, it can be essentially trivially extended to averages of functions of the i.i.d. samples. In particular, we can similarly derive the rate at which

$$\frac{1}{N}\sum_{n=1}^{N} t(\mathsf{y}_n) \geq \gamma' \tag{52}$$

decays exponentially in N. Conceptually, it suffices to apply Theorem 4 to the random variables $\mathsf{t}_n = t(\mathsf{y}_n)$ whose distribution is derived from q. Equivalently, one can rederive the theorem generally as before, but now with the linear exponential family

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

as the relevant tilted distribution.

A more powerful generalization is the vector version of Cramér's Theorem. In this version, we consider sequences of N i.i.d. vectors $\mathbf{y}_1, \dots \mathbf{y}_N$, each of dimension k, generated according to a multivariate distribution q. Then when a set $\mathcal{F} \subset \mathbb{R}^k$ is the closure of its interior, we have

$$\lim_{N \to \infty} -\frac{1}{N} \log \mathbb{P} \left(\frac{1}{N} \sum_{n=1}^{N} \mathbf{y}_n \in \mathcal{F} \right) = E_{\mathrm{C}}(\mathcal{F}), \tag{53}$$

where the Chernoff exponent $E_{\mathrm{C}}(\mathcal{F})$ is defined via

$$E_{\mathrm{C}}(\mathcal{F}) \triangleq \min_{\{\mathbf{x}:\ \mathbb{E}_{p(\cdot;\mathbf{x})}[\mathbf{y}] \in \mathcal{F}\}} D(p(\cdot;\mathbf{x}) \| q), \tag{54}$$

with

$$p(\mathbf{y};\mathbf{x}) = q(\mathbf{y})\, e^{\mathbf{x}^{\mathrm{T}}\mathbf{y} - \alpha(\mathbf{x})}. \tag{55}$$

Finally, the results can be extended beyond the i.i.d. case to dependent samples—in particular, when the samples come from a sufficiently ergodic random process, meaning that, e.g., the WLLN applies. The Gärtner-Ellis Theorem is the natural generalization of Cramér's Theorem in this case, and can be used to analyze the large deviation behavior of finite-state Markov chains, for example.

In the next installment of the notes, via a different framework we will develop an information geometric view of our large deviation results, which leads to important additional insights. In contrast to the preceding development, the framework we pursue next is specific to discrete distributions over finite alphabets.

22.6 Further Reading

T. Cover and J. Thomas, *Elements of Information Theory*, provides develops additional aspects of typical sets and their applications, including for large deviation analysis.

A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, provides a more advanced treatment of large deviation analysis.