Statistical Inference I: Project 2
Large Deviation Theory

Chi-Ning Chou
January 14, 2016

Abstract

Large deviation theory developed in late 20th century concerns the rate of asymptotic behavior of a sequence of random variables. Specifically, we would like to analyze the tail probability/rare event probability as the sample size n goes to infinity. In general, we consider $\mathbb{P}[X_n \in A]$ in a hope that we can characterize it with rate function $I(A) = \lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}[X_n \in A]$. With rate function, we can have a feeling about the speed of convergence, which is very important in lots of application involve asymptotic analysis such as Information theory, statistical mechanics, etc.

In this project, we surveyed [BZ79] which mainly focused on developing a complete theory for applying large deviation results in general settings. Their important intuitions and results will be discussed in Section 2. In addition to technical parts, we presented some applications of large deviation theory in Information theory in Section 3.

Keywords: Large deviation, Cramer's theorem, Sanov's theorem

1 Introduction

Large deviation theory deals with the probability sequence $\mathbb{P}[X_n \in A]$ which follows the so called *Large Deviation Principle (LDP)*. Namely, there exists a rate function $I(A)$ such that $\mathbb{P}[X_n \in A] = e^{-nI(A)\pm o(n)}$, or asymptotically $\lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}[X_n \in A] = I(A)$. There are several interesting points here about the results of large deviation theory. First, it reveals an exponential convergent rate, which is asymptotically tight and optimal. Second, the limiting behavior with respect to a set A can be simply quantified with a single rate function I. Finally, the setting of large deviation is quite general, lots of asymptotic behaviors of random sequence can be characterized using a general form, which will be introduced in Section 2.

As a result, there might be some questions we would like to ask:

- How do we know a random sequence enjoys LDP?
- Why exponential convergent rate?
- How to find the rate function?

To answer these questions, let's have a look at the following example for some intuitions.

Example 1 (Sample mean of standard normal) *Consider the n iid standard normals and their sample mean $\bar{X}_n \sim N(0, \frac{1}{n})$. Let $A_t = [t, \infty)$, we have*

$$\mathbb{P}[\bar{X}_n \in A_t] = \int_t^\infty \sqrt{\frac{1}{2\pi/n}} e^{-x^2/(2/n)} dx \xrightarrow{n\to\infty} e^{-nt^2/2}$$

The sample mean of n iid standard normals follows LDP with rate function $I(A_t) = \frac{t^2}{2}$.

From Example 1, we can see that the first step is to write the probability we concern into an integration form $\mathbb{P}[X_n \in A] = \int_A e^{-ng(x)+o(n)} dx$. Then, as n goes to infinity, by the *Laplace's principle*, the integration will be dominated by the supremum, *i.e.*, $\mathbb{P}[X_n \in A] \approx \sup_{x\in A} e^{-ng(x)} = exp(-n \inf_{x\in A} g(x))$ as $n \to \infty$. Thus, we yield the rate function $I(A) = \inf_{x\in A} g(x)$. Let's see another example about Chernoff bound.

Example 2 (Chernoff's bound) *Consider n iid random variables $X_1, ..., X_n$ with mean $\mu = \mathbb{E}[X_1]$, moment generating function $M(\theta) = \mathbb{E}[e^{\theta X_1}]$, and sample mean $S_n = \frac{1}{n}\sum_{i=1}^n X_i$.*
If $a > \mu$, $\exists \theta > 0$ such that $\frac{M(\theta)}{e^{\theta a}} < 1$ and

$$\mathbb{P}[S_n > a] \leq (\frac{M(\theta)}{e^{\theta a}})^n$$

If $a < \mu$, $\exists \theta < 0$ such that $\frac{M(\theta)}{e^{\theta a}} < 1$ and

$$\mathbb{P}[S_n < a] \leq (\frac{M(\theta)}{e^{\theta a}})^n$$

Once we rewrite the upper bound as

$$e^{-n(\theta a - \log M(\theta))}$$

we can see that take $I(a) = \sup_\theta(\theta a - \log M(\theta))$ gives an upper bound on the decaying rate of $\mathbb{P}[S_n > a]$ or $\mathbb{P}[S_n < a]$.

The discussion here is very rough and informal, however, we can get some intuitions that the development of large deviation theory is actually based on two key observations: the rewrite of probability integration and the dominance of single atom in the event set.

- (**Rewrite of probability integration**) In Example 1, we simply write down the tail probability as an integration over an exponential function. and from Example 2, we can see that a general rate function $I(a) = \sup_\theta(\theta a - M(\theta))$ provides an upper bound on the decaying rate of tail probability. However, in general setting, it might not be easy to do so as the rare event set could be non-trivial. To extend the results, several techniques such as cumulant generating function can help us rewriting the probability in a desired exponential integration form. Then, we can upper bound and lower bound the probability with exponential term and derive the rate function.
- (**Dominance of single atom**) Laplace's transformation gave us a feeling about the dominance of probability density of single atom in an exponential integration. In a more complicated setting, we would like to to derive similar results. Basically, our goal is to upper and lower bound the probability with a function of single atom. In Section 2, we will see that Varadham, Bahadur, Cramer etc. presented some nice theorems with the help of entropy function and Legendre-Fenchel transformation.

In our future discussion, we will see how people use different method to elegantly rewrite and bound the probability of rare event.

2 Large deviation principle

From the examples in the previous section, it's tempting to generalize the results in two directions:

- Extend to general rare event instead of only for tail event.
- Provide a lower bound on the rare event probability.

In this section, we will first state the famous Cramer's theorem to show how we can answer the above two questions by simply extending the idea of Chernoff's bound to open/closed set. Then, the results for general settings discussed in the works of Bahadur and Zabell [BZ79] will be presented. From these two theorems, we can get some feelings about the underlying philosophy of large deviation principle.

2.1 Cramer's theorem

The goal of Cramer's theorem was to derive rate function for open/closed interval. Start from considering the probability of sample mean lying in $[a^+, \infty)$ or $(-\infty, a^-]$, we can use Chernoff's bound to get an upper bound for the rare event probability, *i.e.,* a lower bound for rate function. The main difficulty is on handling the lower bound and extend to general open/closed set. After resolving those mathematical issues, we will have the following theorem.

Theorem 3 (Cramer's theorem) *Consider n iid random variables $X_1, ..., X_n$ with the same settings as in Example 2, the following holds.*
For any closed set $F \subseteq \mathbb{R}$

$$\lim \inf_{n\to\infty} -\frac{1}{n}\mathbb{P}[S_n \in F] \geq \sup_{x\in F} I(x)$$

For any open set $U \subseteq \mathbb{R}$

$$\lim \sup_{n\to\infty} -\frac{1}{n}\mathbb{P}[S_n \in U] \leq \sup_{x\in U} I(x)$$

,where $I(a) = \sup_\theta(\theta a - M(\theta))$.

Proof: Basically, the proof of Cramer's theorem follows the philosophy is based on the idea of moment generating function used in Chernoff bound plus some partition arguments. For more details, one can refer to a nice tutorial by Morters. [Mör08] ■

That is, Cramer's theorem generalized the results in Example 2 to a general rare event setting and provide a lower bound. Thus, we can use the rate function $I(a)$ to fully characterize the asymptotic behaviors of rare events defined on open/closed set.

Cramer's theorem provides a nice results for finding the rate function for an open/closed rare event set. However, what if the rare event set is not that beautiful to characterize? Is there any large deviation theory for general cases?

2.2 Bahadur and Zabell's extension

In the section 3 of [BZ79], Bahadur and Zabell considered general rare event set characterized by a functional defined on a vector space. Then, they used cumulant generating function, which is a extended variation of moment generating function, and derive analogous results. Now, let's take a look at how they formalize the idea.

The settings and notations are basically following the previous examples and theorems. The only difference is that now we consider a general topological vector space V instead of simply $\mathbb{R}$. And the rare event set we consider is induced from a point $v \in V$ and a functional $f \in V^*$ defined as follow.

$$H_f(v) = \{w \in V : f(w) \geq f(v)\}$$

Intuitively, the rare event set collects all the point having larger function value than a point v. As we can transform sample mean $\bar{X}_n$ into $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^n f(X_i)$, now we can discuss the rare event probability

$$\mathbb{P}[\bar{Y}_n \in H_f(v)] \approx e^{-nI(f,v)}$$

Immediately, we have a lemma to quantify the rate function $I(f, v)$.

Lemma 4 *For each $v \in V$ and $f \in V^*$, we have*

$$I(f, v) = \sup_{\theta \geq 0} \theta f(v) - \log \mathbb{E}[e^{\theta f(X)}]$$

Note that when taking $f(x) = x$, the above lemma will be exactly the same as Cramer's theorem.

The $\mathbb{E}[e^{\theta f(X)}]$ is the so called cumulant transformation and $I(f, v)$ is simply the Fenchel transformation of it. Basically, here we just first map the element in a vector space V back to $\mathbb{R}$ and do the same procedure as we did before to construct upper bound and lower bound for the rare event probability. To illustrate why the above lemma holds, one can refer to a extension of Laplace principle: *Varadhan's theorem.* [Var66] In one sentence, Varadhan's theorem told us that we can approximate the cumulant generating function with $\mathbb{E}[e^{nf(X)}] \approx e^{n \sup_\theta f(\theta) - I(\theta)}$.

Now, we can finally derive general large deviation theory with the above line of works.

Theorem 5 (Bahadur, Zabell: Theorem 3.1,3.2) *If V satisfies some regular condition, then for any J be a Borel measurable subset of V, we have*

$$\lim_{n \to \infty} -\frac{1}{n} \mathbb{P}[\bar{X}_n \in J] = \sup_{v \in J, f \in V^*} I(f, v)$$

Intuitively, the rate function of sample mean lying in some Borel measurable rare event set can quantified with the supremum of Fenchel-Legendre transform over any functionals. As a result, with these theorem, we can compute the rate function accordingly.

2.3 Remark

There are two remarks I want to emphasize.

1. The result presented in Theorem 5 is not very mechanical. That is to say, for an ugly rare event set J, me might have no idea how to compute its rate function even we have known the above closed form. Thus, there are other large deviation theorems was developed for computational convenience. For example, the Sanov's theorem, which will be introduced in Section 3 with its application in Information theory.
2. I omitted the discussion about point entropy in this section for the elegance of illustration. Basically, we can further find out that the probability will actually be dominated by single atom in the rare event set. This is the reason why there will be a supremum term over the point in J in Theorem 5. For intuition, the reason why the phenomenon will happen is similar to what happen in Laplace's principle and Varadhan's theorem.

3 Applications in Information theory

Information theory discusses the fundamental limit in a communication system. Started from Shannon's 1948 paper: *A Mathematical Theory of Communication* [Sha01], people have derived and presented many great results which are widely used in practice. Specifically, Information theory focus on the asymptotic behaviors of a communication system. For example, in source coding, the goal would be finding the optimal data compressing rate, and in channel coding, the goal would become achieving the optimal capacity. For a complete introduction to Information theory, please refer to the nice book by Cover and Thomas. [CT12].

Concretely, there are two aspects in each subfield of Information theory: *Achievability* and *converse* (lower bound). The analysis of these two aspects are sometimes non-trivial. In some cases, with the help of large deviation theory, we can easily derive the results and even get more intuitive interpretation.

In this section, we would first present the Sanov's theorem in large deviation theory and then show some related applications.

3.1 Sanov's theorem

As all the theorems in large deviation theory, Sanov's theorem provides a way to compute the exponent of rare event. The reason why it is important in Information theory is that the exponent it provides is in the form of *Kullback-Leibler divergence*, which has strong connection to the important information measure, Shannon entropy, in Information theory.

The scenario we concern is as follow: Consider an i.i.d. discrete random variables $X_1, ..., X_n \sim P$ on a finite support $\mathcal{X}$ and a set $\mathcal{F}$ consisting the probability distributions that cause rare event. Our goal is to analyze the probability that the empirical distribution of $X_1, ..., X_n$: $\Pi_{X_1,...,X_n}$ falling into $\mathcal{F}$. Sanov's theorem gave us an upper bound for the probability as follow.

Theorem 6 (Sanov) *With the above setting, we have*

$$P^n(\Pi_{X_1,...,X_n} \in \mathcal{F}) \leq (n+1)^{|\mathcal{X}|} 2^{-nD_{KL}(P^*||P)}$$

, where $P^* = \arg\inf_{Q\in\mathcal{F}} D_{KL}(Q||P)$*. Moreover, under mild regular condition, we have*

$$\lim_{n\to\infty} -\frac{1}{n} \log P^n(\Pi_{X_1,...,X_n} \in \mathcal{F}) = D_{KL}(P^*||P)$$

Proof: Both upper and lower bounds can be elegantly proved in *method of types* [Csi98], which is a beautiful techniques in Information theory. Here, we omitted the details. ■

The message from Sanov's theorem is a vividly geometrical interpretation as illustrated in Figure 1. One can see that the exponent of the probability of rare event $\mathcal{F}$ is exactly the minimum KL-divergence between the underlying distribution and the probability distribution in $\mathcal{F}$.

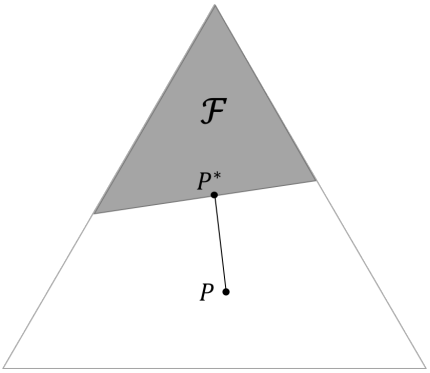

Figure 1: Geometrical interpretation of Sanov's theorem.

With the help of Sanov's theorem, one can easily calculate the error exponent via minimizing the KL-divergence under certain constraint as long as the rare event set $\mathcal{F}$ is good. (e.g. convex)

3.2 Example: Performance of a channel

Now, we would like to show an example to convince you the power of Sanov's theorem in Information theory. Consider a channel transmitting some messages from a finite alphabet set $\mathcal{X}$ with underlying distribution P. We use these messages to encode information, however, there will be some costs during the transmission. The cost function $g : \mathcal{X} \to \mathbb{R}^+$ records the cost of each symbol and the average cost of a message $x_1, ..., x_n$ will be $\frac{1}{n}\sum_{i=1}^{n} g(x_i)$. Our goal here is to analyze the probability of the expectation cost being greater than a certain level c. That is, the probability of empirical distribution falling into rare event set $\mathcal{F} = \{Q : \mathbb{E}_Q[g(X) \geq c]\}$. After knowing the error probability, one can derive the rate function and find the fundamental limit of this channel.

By Sanov's theorem, the error exponent can be computed as follow

$$\lim_{n\to\infty} -\frac{1}{n} P^n(\Pi_{X_1,...,X_n} \in \mathcal{F}) = \inf_{Q\in\mathcal{F}} D_{KL}(Q||P)$$

To be more precise, if the cost function g is convex, we can compute the error exponent by solving a convex optimization problem.

$$\begin{array}{rl} \text{minimize} & \sum_{a\in\mathcal{X}} Q(a) \log \frac{Q(a)}{P(a)} \\ \text{subject to} & c - \sum_{a\in\mathcal{X}} Q(a) \leq 0 \\ & -Q(a) \leq 0,\ \forall a \in \mathcal{X} \\ & \sum_{a\in\mathcal{X}} Q(a) = 1 \end{array}$$

We can use the KKT conditions and find the error exponent

$$\lim_{n\to\infty} -\frac{1}{n} P^n(\Pi_{X_1,\ldots,X_n} \in \mathcal{F}) = \sum_{a\in\mathcal{X}} \frac{p(a)e^{\mu g(a)}}{\sum_{b\in\mathcal{X}} p(b)e^{\mu g(b)}} \log \frac{e^{\mu g(a)}}{\sum_{b\in\mathcal{X}} p(b)e^{\mu g(b)}}$$

, where μ makes $\sum_{a\in\mathcal{X}} g(a)p(a)e^{\mu g(a)} = c$.

Although the result above is looked complicated, it has a close form for us to analyze. From this simple example, we can taste the power of Sanov's theorem.

4 Conclusion

In this survey, we have a overview study for large deviation theory. From the examples and applications, we can see that knowing the rate function of rare event is quite useful and powerful in many fields involving asymptotic analysis. However, how to efficiently and correctly derive the rate function is a non-trivial question, especially when the rare event set is ugly. In Section 2, we see two theorems: Cramer's theorem and Bahadur-Zabell's theorem that characterized the rate function based on the observation in Laplace's principle. Finally, we know that the rate function is simply the supremum of the Fenchel-Legendre transform of cumulant generating function. In the end, we use Figure 2 to visualize the historical development of large deviation theory.

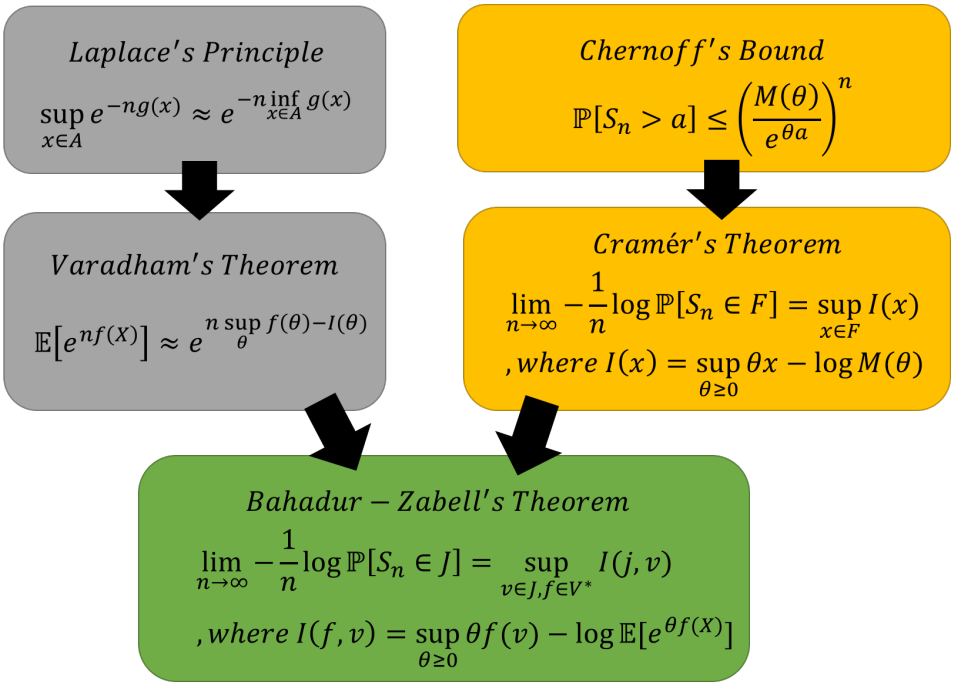

Figure 2: Historical flow of large deviation theory.

References

BZ79. RR Bahadur and SL Zabell. Large deviations of the sample mean in general vector spaces. *Annals of Probability*, 7(4):587–621, 1979.

Csi98. Imre Csiszár. The method of types [information theory]. *Information Theory, IEEE Transactions on*, 44(6):2505–2523, 1998.

CT12. Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

Mör08. Peter Mörters. Large deviation theory and applications, 2008.

Sha01. Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

Var66. SR Srinivasa Varadhan. Asymptotic probabilities and differential equations. *Communications on Pure and Applied Mathematics*, 19(3):261–286, 1966.