Lecture 5. Limit theorems

Tiejun Li[1,2]

[1]School of Mathematical Sciences (SMS),
&
[2]Center for Machine Learning Research (CMLR),
Peking University,
Beijing 100871,
P.R. China
tieli@pku.edu.cn

Office: No. 1 Science Building, Room 1376E

Table of Contents

Law of Large Numbers

- Let $\{X_j\}_{j=1}^{\infty}$ be a sequence of independently and identically distributed (abbreviated as i.i.d. in the later text) random variables. Let $\eta = \mathbb{E}X_1$ and S_n the partial sum of X_j from 1 to n.

Law of Large Numbers

- Let $\{X_j\}_{j=1}^{\infty}$ be a sequence of independently and identically distributed (abbreviated as i.i.d. in the later text) random variables. Let $\eta = \mathbb{E}X_1$ and S_n the partial sum of X_j from 1 to n.
- The well-known law of large numbers validates the intuitive characterization of the mathematical expectation: it is the limit of empirical average when the sample size n goes to infinity. It is also the theoretical basis of the Monte Carlo methods.

Weak Law of Large Numbers

Theorem (Weak law of large numbers (WLLN))

For i.i.d. random variables $\{X_j\}_{j=1}^{\infty}$ *with* $\mathbb{E}|X_j| < \infty$, *we have*

$$\frac{S_n}{n} \to \eta \qquad \textit{in probability.}$$

Proving the result under the stated assumption is quite involved. We will give a proof of the WLLN under the stronger assumption that $\mathbb{E}|X_j|^2 < \infty$.

Weak Law of Large Numbers: Proof

Proof.

Without loss of generality, we can assume $\eta = 0$. Using Chebyshev's inequality, we have

$$\mathbb{P}\left\{\left|\frac{S_n}{n}\right| > \epsilon\right\} \leq \frac{1}{\epsilon^2}\mathbb{E}\left|\frac{S_n}{n}\right|^2$$

for any $\epsilon > 0$. Using independence, we have

$$\mathbb{E}|S_n|^2 = \sum_{i,j=1}^{n} \mathbb{E}(X_i X_j) = n\mathbb{E}|X_1|^2.$$

Hence

$$\mathbb{P}\left\{\left|\frac{S_n}{n}\right| > \epsilon\right\} \leq \frac{1}{n\epsilon^2}\mathbb{E}|X_1|^2 \to 0,$$

as $n \to \infty$. □

Strong Law of Large Numbers

Theorem (Strong law of large numbers (SLLN))

For i.i.d. random variables $\{X_j\}_{j=1}^{\infty}$ *we have*

$$\frac{S_n}{n} \to \eta \quad \textit{a.s.}$$

if and only if $\mathbb{E}|X_j| < \infty$.

Strong Law of Large Numbers: Proof

Proof.

We will only prove SLLN under the stronger assumption that $\mathbb{E}|X_j|^4 < \infty$. More general case may be referred to [1].
Without loss of generality, we can assume $\eta = 0$. Using Chebyshev's inequality, we obtain

$$\mathbb{P}\left\{\left|\frac{S_n}{n}\right| > \epsilon\right\} \leq \frac{1}{\epsilon^4}\mathbb{E}\left|\frac{S_n}{n}\right|^4.$$

Using independence, we get

$$\mathbb{E}|S_n|^4 = \sum_{i,j,k,l=1}^{n} \mathbb{E}(X_i X_j X_k X_l) = n\mathbb{E}|X_j|^4 + 3n(n-1)(\mathbb{E}|X_j|^2)^2.$$

[1] K.L. Chung. A course in probability theory. Academic Press, third edition, 2001.

Strong Law of Large Numbers: Proof

We have $(\mathbb{E}|X_j|^2)^2 \leq \mathbb{E}|X_j|^4 < \infty$ by Hölder inequality. Hence

$$\mathbb{P}\left\{\left|\frac{S_n}{n}\right| > \epsilon\right\} \leq \frac{C}{n^2\epsilon^4}.$$

Since the series $1/n^2$ is summable we get

$$\mathbb{P}\left\{\left|\frac{S_n}{n}\right| > \epsilon, \text{ i.o.}\right\} = 0$$

by Borel-Cantelli lemma. This implies that

$$\frac{S_n}{n} \to 0 \qquad \text{a.s.}$$

and we are done. □

Law of Large Numbers: Counter Example

Example (Cauchy distribution)

The following example shows that the law of large numbers does not hold if the assumed condition is not satisfied. Consider the i.i.d. random variables $\{X_j\}_{j=1}^{\infty}$ with Cauchy distribution having probability density function

$$\frac{1}{\pi(1+x^2)}, \quad x \in \mathbb{R}.$$

We have $\mathbb{E}X_j = 0$ by symmetry and $\mathbb{E}|X_j| = \infty, \mathbb{E}|X_j|^2 = \infty$. In this case, we can prove S_n/n always has the same distribution as X_1. Thus the weak and strong law of large numbers are both violated.

Table of Contents

Central Limit Theorem

The following central limit theorem explains why the normal or normal-like distributions are so widely observed in the nature.

Theorem (Lindeberg-Lévy central limit theorem (CLT))

Let $\{X_j\}_{j=1}^{\infty}$ *be a sequence of i.i.d. random variables. Assume that* $\mathbb{E}X_j^2 < \infty$ *and let* $\sigma^2 = \mathrm{var}(X_j)$. *Then*

$$\frac{S_n - n\eta}{\sqrt{n\sigma^2}} \to N(0,1)$$

in the sense of distribution.

Central Limit Theorem: Proof

Proof.

Assume without loss of generality $\eta = 0$ and $\sigma = 1$, otherwise we can shift and rescale X_j. Let f be the characteristic function of X_1 and let g_n be the characteristic function of $S_n/\sqrt{n}$. Then

$$g_n(\xi) = \mathbb{E}e^{i\xi S_n/\sqrt{n}} = \prod_{j=1}^{n} \mathbb{E}e^{i\xi X_j/\sqrt{n\sigma^2}} = \prod_{j=1}^{n} f\left(\frac{\xi}{\sqrt{n}}\right) = f^n\left(\frac{\xi}{\sqrt{n}}\right).$$

Using Taylor expansion and the properties of characteristic functions we obtain

$$\begin{aligned} f\left(\frac{\xi}{\sqrt{n}}\right) &= f(0) + \frac{\xi}{\sqrt{n}} f'(0) + \frac{1}{2}\left(\frac{\xi}{\sqrt{n}}\right)^2 f''(0) + o\left(\frac{1}{n}\right) \\ &= 1 - \frac{\xi^2}{2n} + o\left(\frac{1}{n}\right) \end{aligned}$$

Central Limit Theorem: Proof

Hence

$$g_n(\xi) = f\left(\xi/\sqrt{n}\right)^n = \left(1 - \frac{\xi^2}{2n} + o\left(\frac{1}{n}\right)\right)^n \to e^{-\frac{1}{2}\xi^2} \qquad \text{as } n \to \infty$$

for every $\xi \in \mathbb{R}^1$. This completes the proof by using Levy's continuity theorem. □

CLT: Implication on MC

The central limit theorem is the theoretical basis for the assumption that additive noise can be modeled by Gaussian noises. It also gives an estimate for the rate of convergence in the law of large numbers. Since by CLT we have

$$\frac{S_n}{n} - \eta \sim \frac{\sigma}{\sqrt{n}}.$$

The rate of convergence of S_n/n to η is $O(n^{-\frac{1}{2}})$. This is the reason why most Monte Carlo methods has a rate of convergence of $O(n^{-\frac{1}{2}})$ where n is the sample size.

CLT: Application in polymer physics

The central limit theorem is fundamental to understand the end-to-end statistics for a polymer. The simplest model for flexible polymers is called the freely jointed chain, in which a polymer consists of K units, each of length b_0 and able to point in any direction independently of each other.

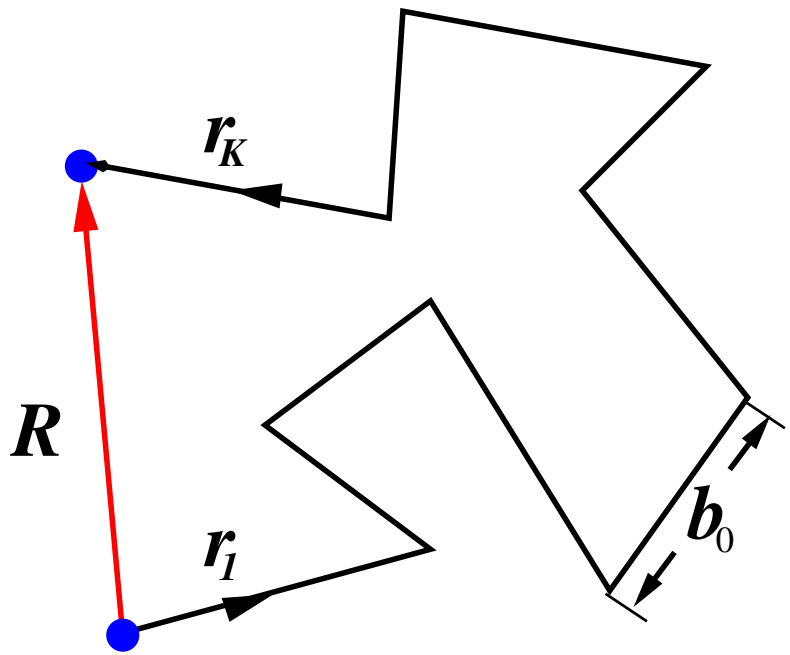

Figure: Schematics of the freely jointed chain.

CLT: Application in polymer physics

Denote the bond vectors as $\boldsymbol{r}_k$ $(k = 1, \ldots, K)$, which has i.i.d. distribution density

$$p(\boldsymbol{r}) = \frac{1}{4\pi b_0^2}\delta(\boldsymbol{r} - b_0).$$

The end-to-end vector

$$\boldsymbol{R} = \sum_{k=1}^{K} \boldsymbol{r}_k.$$

From the central limit theorem, we have asymptotically

$$\boldsymbol{R} \sim N(0, Kb_0^2\boldsymbol{I}), \quad K \gg 1.$$

Note that this Gaussian type approximation as $K \gg 1$ is independent of the choice of the bond vector distribution. This model is called *Gaussian chain* in polymer physics.

Table of Contents

Laplace asymptotics

- Laplace method is the basis of large deviation theory. It is widely used in many fields of applied mathematics. We will only introduce the one-dimensional version of Laplace asymptotics in this section. For more details, see [2].

[2] C.M. Bender and S.A. Orszag. Advanced Mathematical Methods for Scientists and Engineers: Asymptotic Methods and Perturbation Theory.

Laplace asymptotics

- Laplace method is the basis of large deviation theory. It is widely used in many fields of applied mathematics. We will only introduce the one-dimensional version of Laplace asymptotics in this section. For more details, see [2].
- Let us consider the Laplace integral

$$F(t) = \int_{\mathbb{R}} e^{th(x)} dx, \qquad t \gg 1$$

where $h(x) \in C^2(\mathbb{R})$, $h(0) = 0$ is the only global maximum such that

$$h(x) \leq -b \qquad \text{if} \qquad |x| \geq c$$

for positive reals b, c. Suppose $h(x) \to -\infty$ fast enough as $x \to \infty$ to ensure the convergence of F for $t = 1$ and assume $h''(0) < 0$, then the Laplace Lemma holds.

[2] C.M. Bender and S.A. Orszag. Advanced Mathematical Methods for Scientists and Engineers: Asymptotic Methods and Perturbation Theory.

Laplace asymptotics

Lemma (Laplace Lemma)

As $t \to \infty$, to leading order

$$F(t) \sim \sqrt{2\pi}(-th''(0))^{-\frac{1}{2}}.$$

Laplace asymptotics

Lemma (Laplace Lemma)

As $t \to \infty$, to leading order

$$F(t) \sim \sqrt{2\pi}(-th^{''}(0))^{-\frac{1}{2}}.$$

Proof.

If $h(x) = h''(0)x^2/2, h^{''}(0) < 0$, then

$$\int_{\mathbb{R}} e^{th(x)}dx = \sqrt{2\pi}(-th^{''}(0))^{-\frac{1}{2}}.$$

In general, for any $\epsilon > 0$, there exists $\delta > 0$ such that for any $|x| \leq \delta$,

$$|h(x) - \frac{h^{''}(0)}{2}x^2| \leq \epsilon x^2.$$

Laplace asymptotics: Proof

It follows that

$$\int_{[-\delta,\delta]} \exp\Big(\frac{tx^2}{2}(h''(0)-2\epsilon)\Big)dx \le \int_{[-\delta,\delta]} \exp\Big(th(x)\Big)dx \le \int_{[-\delta,\delta]} \exp\Big(\frac{tx^2}{2}(h''(0)+2\epsilon)\Big)dx.$$

For this $\delta > 0$, there exists $\eta > 0$ by assumptions such that

$$h(x) \le -\eta \quad \text{if} \quad |x| \ge \delta,$$

thus

$$\int_{|x|\ge\delta} \exp\Big(th(x)\Big)dx \le e^{-(t-1)\eta} \int_{\mathbb{R}} e^{h(x)}dx \sim \mathcal{O}(e^{-\alpha t}),\ \alpha > 0, \quad \text{for} \quad t > 1.$$

Laplace asymptotics: Proof

First consider the upper bound, we have

$$\begin{aligned}
&\int_{\mathbb{R}} \exp\Big(th(x)\Big)dx \\
\leq &\int_{\mathbb{R}} \exp\Big(\frac{tx^2}{2}(h''(0)+2\epsilon)\Big)dx - \int_{|x|\geq\delta} \exp\Big(\frac{tx^2}{2}(h''(0)+2\epsilon)\Big)dx + \mathcal{O}(e^{-\alpha t}) \\
= &\sqrt{2\pi}\Big[t(-h^{''}(0)-2\epsilon)\Big]^{-\frac{1}{2}} + \mathcal{O}(e^{-\beta t})
\end{aligned}$$

where $\beta > 0$. In fact, we ask $\epsilon < -h''(0)/2$ here.
The proof of lower bound is similar. By the arbitrary smallness of ϵ, we have

$$\lim_{t\to\infty} F(t)/\sqrt{2\pi}(-th^{''}(0))^{-\frac{1}{2}} = 1,$$

which completes the proof. □

Laplace asymptotics: LDT form

The result is easily extended to the case where $h(0) \neq 0$. The term $e^{th(0)}$ will appear in the leading order and another commonly used form ignoring the prefactor is the so-called saddle point approximation

$$\lim_{t\to\infty} \frac{1}{t} \log F(t) = \sup_{x\in\mathbb{R}} h(x),$$

which is the typical form in large deviation theory and widely used in physics literature.

Table of Contents

Cramér's Theorem for Large Deviations

- Let $\{X_j\}_{j=1}^n$ be a sequence of i.i.d. random variables and let $\eta = \mathbb{E}X_j$.

Cramér's Theorem for Large Deviations

- Let $\{X_j\}_{j=1}^n$ be a sequence of i.i.d. random variables and let $\eta = \mathbb{E}X_j$.
- The laws of large numbers says that for any $\epsilon > 0$, with probability close to 1, $|S_n/n - \eta| < \epsilon$ for large enough n; conversely if $y \neq \eta$, then the probability that S_n/n is close to y goes to zero as $n \to \infty$.

Cramér's Theorem for Large Deviations

- Let $\{X_j\}_{j=1}^n$ be a sequence of i.i.d. random variables and let $\eta = \mathbb{E}X_j$.
- The laws of large numbers says that for any $\epsilon > 0$, with probability close to 1, $|S_n/n - \eta| < \epsilon$ for large enough n; conversely if $y \neq \eta$, then the probability that S_n/n is close to y goes to zero as $n \to \infty$.
- Events of this type, i.e. $\{|S_n/n - y| < \epsilon\}$, are called *large deviation events* compared with the *small deviation events* from the mean like the set $\{|S_n/n - \eta| \leq c/\sqrt{n}\}$ in central limit theorem.

Rate function

To estimate the precise rate at which $\mathbb{P}\{|S_n/n - y| < \epsilon\}$ goes to zero, we assume here that the distribution μ of the X_j's have finite exponential moments. Let us define the moment generating function

$$M(\lambda) = \mathbb{E}e^{\lambda X_j} = \int_{\mathbb{R}} e^{\lambda x} d\mu(x) < \infty, \quad \lambda \in \mathbb{R},$$

the cumulant generating function

$$\Lambda(\lambda) = \log M(\lambda)$$

and the Legendre-Fenchel transform of $\Lambda(\lambda)$ as

$$I(x) = \sup_{\lambda}\{x\lambda - \Lambda(\lambda)\}.$$

Then we have the large deviation type theorem for the i.i.d. sums.

Cramér's Theorem

Theorem (Cramér's Theorem)

The distribution of the empirical average μ_n defined by

$$\mu_n(\Gamma) = \mathbb{P}\{S_n/n \in \Gamma\}$$

satisfies the large deviation principle:

Cramér's Theorem

Theorem (Cramér's Theorem)

The distribution of the empirical average μ_n defined by

$$\mu_n(\Gamma) = \mathbb{P}\{S_n/n \in \Gamma\}$$

satisfies the large deviation principle:

(i) *For any closed set $F \in \mathcal{B}$*

$$\overline{\lim_{n\to\infty}} \frac{1}{n} \log \mu_n(F) \leq -\inf_{x\in F} I(x).$$

Cramér's Theorem

Theorem (Cramér's Theorem)

The distribution of the empirical average μ_n defined by

$$\mu_n(\Gamma) = \mathbb{P}\{S_n/n \in \Gamma\}$$

satisfies the large deviation principle:

(i) *For any closed set $F \in \mathcal{B}$*

$$\overline{\lim_{n\to\infty}} \frac{1}{n} \log \mu_n(F) \leq - \inf_{x\in F} I(x).$$

(ii) *For any open set $G \in \mathcal{B}$*

$$\underline{\lim_{n\to\infty}} \frac{1}{n} \log \mu_n(G) \geq - \inf_{x\in G} I(x).$$

Cramér's Theorem

Theorem (Cramér's Theorem)

The distribution of the empirical average μ_n defined by

$$\mu_n(\Gamma) = \mathbb{P}\{S_n/n \in \Gamma\}$$

satisfies the large deviation principle:

(i) *For any closed set $F \in \mathcal{B}$*

$$\overline{\lim_{n\to\infty}} \frac{1}{n} \log \mu_n(F) \leq -\inf_{x\in F} I(x).$$

(ii) *For any open set $G \in \mathcal{B}$*

$$\underline{\lim_{n\to\infty}} \frac{1}{n} \log \mu_n(G) \geq -\inf_{x\in G} I(x).$$

$I(x)$ is called the rate function.

Cramér's Theorem: Intuition

For the so-called I-continuity set Γ, i.e.
$\inf_{x\in\Gamma^\circ} I(x) = \inf_{x\in\bar{\Gamma}} I(x)$, this theorem suggests that roughly

$$\mu_n(\Gamma) \asymp \exp\left(-n \inf_{x\in\Gamma} I(x)\right).$$

Here we use the notation "$\asymp$" instead of "$\approx$" since the equivalence is in the logarithmic scale, i.e. $a_n \asymp b_n$ if $\log a_n \sim \log b_n$.

Cramér's Theorem: Intuition

For the so-called I-continuity set Γ, i.e. $\inf_{x\in\Gamma^\circ} I(x) = \inf_{x\in\bar{\Gamma}} I(x)$, this theorem suggests that roughly

$$\mu_n(\Gamma) \asymp \exp\left(-n \inf_{x\in\Gamma} I(x)\right).$$

Here we use the notation "$\asymp$" instead of "$\approx$" since the equivalence is in the logarithmic scale, i.e. $a_n \asymp b_n$ if $\log a_n \sim \log b_n$.

Before the proof, we need some results on the Legendre-Fenchel transform and some elementary properties of $I(x)$.

Legendre-Fenchel Transform

Lemma (Legendre-Fenchel Transform)

Suppose $f(x) : \mathbb{R}^d \to \bar{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$ *is a lower semicontinuous convex function. The conjugate function* $F(y)$ *of* $f(x)$ *(Legendre-Fenchel transform) defined as*

$$F(y) = \sup_x \{(x, y) - f(x)\}$$

has the following properties:

Legendre-Fenchel Transform

Lemma (Legendre-Fenchel Transform)

Suppose $f(x) : \mathbb{R}^d \to \bar{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$ *is a lower semicontinuous convex function. The conjugate function* $F(y)$ *of* $f(x)$ *(Legendre-Fenchel transform) defined as*

$$F(y) = \sup_x \{(x, y) - f(x)\}$$

has the following properties:

(i) F *is also a lower semicontinuous convex function.*

Legendre-Fenchel Transform

Lemma (Legendre-Fenchel Transform)

Suppose $f(x) : \mathbb{R}^d \to \bar{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$ *is a lower semicontinuous convex function. The conjugate function* $F(y)$ *of* $f(x)$ *(Legendre-Fenchel transform) defined as*

$$F(y) = \sup_x \{(x, y) - f(x)\}$$

has the following properties:

(i) F *is also a lower semicontinuous convex function.*

(ii) *Fenchel inequality holds*

$$(x, y) \leq f(x) + F(y).$$

Legendre-Fenchel Transform

Lemma (Legendre-Fenchel Transform)

Suppose $f(x) : \mathbb{R}^d \to \bar{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$ *is a lower semicontinuous convex function. The conjugate function* $F(y)$ *of* $f(x)$ *(Legendre-Fenchel transform) defined as*

$$F(y) = \sup_x \{(x, y) - f(x)\}$$

has the following properties:

(i) F *is also a lower semicontinuous convex function.*

(ii) *Fenchel inequality holds*

$$(x, y) \leq f(x) + F(y).$$

(iii) *The conjugacy relation holds:*

$$f(x) = \sup_y \{(x, y) - F(y)\}.$$

Legendre-Fenchel Transform

where we utilize the rule

$$\alpha + \infty = \infty,\ \alpha - \infty = -\infty \quad \textit{for } \alpha \textit{ finite}$$
$$\alpha \cdot \infty = \infty,\ \alpha \cdot (-\infty) = -\infty, \quad \textit{for } \alpha > 0$$
$$0 \cdot \infty = 0 \cdot (-\infty) = 0, \quad \inf \emptyset = \infty, \sup \emptyset = -\infty$$

The readers may be referred to [3] [4] for proof details.

[3] R.T. Rockafellar. Convex analysis. Princeton University Press, Princeton, 1970.

[4] R.T. Rockafellar and R. J-B Wets. Variational Analysis. Springer-Verlag, Berlin and Heidelberg, 2009.

Heuristic derivation of the rate function

Now we apply the Laplace asymptotics to explain heuristically why the rate function takes the interesting form

$$I(x) = \sup_{\lambda} \{x\lambda - \Lambda(\lambda)\}.$$

Heuristic derivation of the rate function

Now we apply the Laplace asymptotics to explain heuristically why the rate function takes the interesting form

$$I(x) = \sup_{\lambda}\{x\lambda - \Lambda(\lambda)\}.$$

Suppose the Cramér's theorem is already correct, then roughly we have

$$\mu_n(dx) \propto \exp(-nI(x))dx$$

and thus by Laplace asymptotics

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}^{\mu_n}(\exp(n\Phi(x))) &:= \lim_{n\to\infty} \frac{1}{n} \log \int_{\mathbb{R}} \exp(n\Phi(x))\mu_n(dx) \\ &= \sup_{x}\{\Phi(x) - I(x)\}.\end{aligned}$$

Heuristic derivation of the rate function

Now we take $\Phi(x) = \lambda x$ then

$$\mathbb{E}^{\mu_n}(\exp(n\lambda x)) = \mathbb{E}\exp(\lambda \sum_{j=1}^{n} X_j) = [\mathbb{E}\exp(\lambda X_j]^n = \Big(M(\lambda)\Big)^n.$$

The equation leads to

$$\Lambda(\lambda) = \sup_x \{\lambda x - I(x)\}.$$

By the conjugacy relation of Legendre-Fenchel transform, we obtain the rate function $I(x)$.

Properties of Rate Function

Lemma (Properties of rate function)

The rate function $I(x)$ has the following properties:

Properties of Rate Function

Lemma (Properties of rate function)

The rate function $I(x)$ has the following properties:

(i) *$I(x)$ is convex and lower semicontinuous.*

Properties of Rate Function

Lemma (Properties of rate function)

The rate function $I(x)$ has the following properties:

(i) $I(x)$ *is convex and lower semicontinuous.*

(ii) $I(x)$ *is non-negative and* $I(\eta) = 0$.

Properties of Rate Function

Lemma (Properties of rate function)

The rate function $I(x)$ has the following properties:

(i) $I(x)$ *is convex and lower semicontinuous.*

(ii) $I(x)$ *is non-negative and* $I(\eta) = 0$.

(iii) $I(x)$ *is non-decreasing in* $[\eta, \infty)$ *and non-increasing in* $(-\infty, \eta]$.

Properties of Rate Function

Lemma (Properties of rate function)

The rate function $I(x)$ has the following properties:

(i) $I(x)$ *is convex and lower semicontinuous.*

(ii) $I(x)$ *is non-negative and* $I(\eta) = 0$.

(iii) $I(x)$ *is non-decreasing in* $[\eta, \infty)$ *and non-increasing in* $(-\infty, \eta]$.

(iv) *If* $x > \eta$, $I(x) = \sup_{\lambda > 0}\{\lambda x - \Lambda(\lambda)\}$; *If* $x < \eta$, $I(x) = \sup_{\lambda < 0}\{\lambda x - \Lambda(\lambda)\}$.

Properties of Rate Function: Proof

Proof.

(i) The convexity of $\Lambda(\lambda)$ follows by Hölder's inequality. For any $0 \leq \theta \leq 1$,

$$\begin{aligned}
\Lambda(\theta\lambda_1 + (1-\theta)\lambda_2) &= \log \mathbb{E}\Big(\exp(\theta\lambda_1 X_j) \exp((1-\theta)\lambda_2 X_j) \Big) \\
&\leq \log \Big(\big(\mathbb{E} \exp\left(\lambda_1 X_j\right)\big)^{\theta} \big(\mathbb{E} \exp(\lambda_2 X_j)\big)^{(1-\theta)} \Big) \\
&= \theta\Lambda(\lambda_1) + (1-\theta)\Lambda(\lambda_2)
\end{aligned}$$

Thus $\Lambda(\lambda)$ is a convex function. The rest is a direct application of Lemma.

Properties of Rate Function: Proof

Proof.

(i) The convexity of $\Lambda(\lambda)$ follows by Hölder's inequality. For any $0 \leq \theta \leq 1$,

$$\begin{aligned}\Lambda(\theta\lambda_1 + (1-\theta)\lambda_2) &= \log \mathbb{E}\Big(\exp(\theta\lambda_1 X_j) \exp((1-\theta)\lambda_2 X_j)\Big) \\ &\leq \log \Big(\left(\mathbb{E} \exp\left(\lambda_1 X_j\right)\right)^{\theta} \left(\mathbb{E}\exp(\lambda_2 X_j)\right)^{(1-\theta)}\Big) \\ &= \theta\Lambda(\lambda_1) + (1-\theta)\Lambda(\lambda_2)\end{aligned}$$

Thus $\Lambda(\lambda)$ is a convex function. The rest is a direct application of Lemma.

(ii) Taking $\lambda = 0$, we obtain $x \cdot 0 - \Lambda(0) = 0$. Thus $I(x) \geq 0$. On the other hand, we have

$$\Lambda(\lambda) = \log \mathbb{E}\exp(\lambda X_j) \geq \log \exp(\lambda\eta) = \lambda\eta$$

by Jensen's inequality. This gives $I(\eta) \leq 0$. Combing with $I(x) \geq 0$ we get the result.

Properties of Rate Function: Proof

(iii) From the convexity of $I(x)$ and it achieves minimum at $x = \eta$, we immediately obtain the desired monotone property in $(-\infty, \eta]$ and $[\eta, \infty)$.

Properties of Rate Function: Proof

(iii) From the convexity of $I(x)$ and it achieves minimum at $x = \eta$, we immediately obtain the desired monotone property in $(-\infty, \eta]$ and $[\eta, \infty)$.
(iv) If $x > \eta$, then when $\lambda \leq 0$

$$\lambda x - \Lambda(\lambda) \leq \lambda\eta - \Lambda(\lambda) \leq 0,$$

Thus the supremum is only achieved when $\lambda > 0$ by the non-negativity of $I(x)$. Similar proof can be applied to the case $x < \eta$. □

Proof of Cramér's Theorem: Upper Bound

Proof of Cramér's Theorem. Without loss of generality, we assume $\eta = 0$. (i) *Upper bound.* Suppose $x > 0$, $J_x := [x, \infty)$. For $\lambda > 0$,

$$\begin{aligned}\mu_n(J_x) &= \int_x^\infty \mu_n(dy) \le e^{-\lambda x} \int_x^\infty e^{\lambda y} \mu_n(dy) \\ &\le e^{-\lambda x} \int_{-\infty}^\infty e^{\lambda y} \mu_n(dy) = e^{-\lambda x} \Big[M(\frac{\lambda}{n}) \Big]^n.\end{aligned}$$

Taking $n\lambda$ instead of λ in the above equation, we obtain

$$\frac{1}{n} \log \mu_n(J_x) \le -(\lambda x - \Lambda(\lambda))$$

and accordingly

$$\frac{1}{n} \log \mu_n(J_x) \le -\sup_{\lambda > 0} \{\lambda x - \Lambda(\lambda)\} = -I(x).$$

Proof of Cramér's Theorem: Upper Bound

If $x < 0$, we can define $\tilde{J}_x = (-\infty, x]$. Similarly as above we get

$$\frac{1}{n} \log \mu_n(\tilde{J}_x) \leq -I(x).$$

For a closed set $F \in \mathcal{B}$, if $0 \in F$, $\inf_{x \in F} I(x) = 0$. Then the upper bound holds obviously. Otherwise, let (x_1, x_2) is the maximal interval satisfying the condition $(x_1, x_2) \cap F = \emptyset$ and $0 \in (x_1, x_2)$. So $x_1, x_2 \in F$, $F \subset \tilde{J}_{x_1} \cup J_{x_2}$. From monotonicity of $I(x)$ in $(-\infty, 0]$ and $[0, \infty)$, we obtain

$$\begin{aligned} \overline{\lim_{n\to\infty}} \frac{1}{n} \log \mu_n(F) &\leq \max\left(\overline{\lim_{n\to\infty}} \frac{1}{n} \log \mu_n(\tilde{J}_{x_1}), \overline{\lim_{n\to\infty}} \frac{1}{n} \log \mu_n(J_{x_2}) \right) \\ &\leq -\min(I(x_1), I(x_2)) = -\inf_{x \in F} I(x). \end{aligned}$$

Proof of Cramér's Theorem: Lower Bound

(ii) *Lower bound.* For any nonempty open set G, it is sufficient to prove that for any $x \in G$

$$\varliminf_{n\to\infty} \frac{1}{n} \log \mu_n(G) \geq -I(x).$$

Now fix x and assume $I(x) < \infty$.

Proof of Cramér's Theorem: Lower Bound

(ii) *Lower bound.* For any nonempty open set G, it is sufficient to prove that for any $x \in G$

$$\varliminf_{n\to\infty} \frac{1}{n} \log \mu_n(G) \geq -I(x).$$

Now fix x and assume $I(x) < \infty$.
Case 1. If the supremum

$$I(x) = \sup_{\lambda}\{\lambda x - \Lambda(\lambda)\}$$

can not be achieved, then $x \neq 0$. Suppose $x > 0$ and there exists $\lambda_n \to \infty$ such that

$$I(x) = \lim_{n\to\infty} (\lambda_n x - \Lambda(\lambda_n)).$$

Proof of Cramér's Theorem: Lower Bound

We have

$$\int_{-\infty}^{x-0} \exp(\lambda_n(y-x)\mu(dy) \to 0 \text{ as } n \to \infty.$$

by dominated convergence theorem. On the other hand

$$\begin{aligned}\lim_{n\to\infty} \int_x^{\infty} \exp(\lambda_n(y-x)\mu(dy) &= \lim_{n\to\infty} \int_{-\infty}^{\infty} \exp(\lambda_n(y-x)\mu(dy) \\ &= \exp(-I(x)) < \infty.\end{aligned}$$

Thus $\mu((x,\infty)) = 0$ and

$$\exp(-I(x)) = \lim_{n\to\infty} \int_x^{\infty} \exp(\lambda_n(y-x)\mu(dy) = \mu(\{x\}).$$

Proof of Cramér's Theorem: Lower Bound

We have

$$\mu_n(G) \geq \mu_n(\{x\}) \geq (\mu(\{x\}))^n = \exp(-nI(x))$$

and thus

$$\frac{1}{n} \log \mu_n(G) \geq -I(x).$$

Similar proof can be applied to the case $x < a$.

Proof of Cramér's Theorem: Lower Bound

We have

$$\mu_n(G) \geq \mu_n(\{x\}) \geq (\mu(\{x\}))^n = \exp(-nI(x))$$

and thus

$$\frac{1}{n} \log \mu_n(G) \geq -I(x).$$

Similar proof can be applied to the case $x < a$.
Case 2. Suppose that the supremum is attained at λ_0 such that

$$I(x) = \lambda_0 x - \Lambda(\lambda_0).$$

Then $x = \Lambda'(\lambda_0) = M'(\lambda_0)/M(\lambda_0)$. Define a new probability measure as

$$\tilde{\mu}(dy) = \frac{1}{M(\lambda_0)} \exp(\lambda_0 y)\mu(dy).$$

Proof of Cramér's Theorem: Lower Bound

It has the expectation

$$\int_{\mathbb{R}} y\tilde{\mu}(dy) = \frac{1}{M(\lambda_0)} \int_{\mathbb{R}} y \exp(\lambda_0 y)\mu(dy) = \frac{M'(\lambda_0)}{M(\lambda_0)} = x.$$

If $x \geq 0$, then $\lambda_0 \geq 0$. For sufficiently small $\delta > 0$, we have $(x - \delta, x + \delta) \subset G$,

$$\begin{aligned}
\mu_n(G) &\geq \mu_n(x - \delta, x + \delta) \\
&= \int_{\left\{\left|\frac{1}{n}\sum_{j=1}^{n} y_j - x\right| < \delta\right\}} \mu(dy_1) \cdots \mu(dy_n) \\
&\geq \exp(-n\lambda_0(x + \delta)) \int_{\left\{\left|\frac{1}{n}\sum_{j=1}^{n} y_j - x\right| < \delta\right\}} \exp(\lambda_0 y_1) \cdots \exp(\lambda_0 y_n)\mu(dy_1) \cdots \mu(dy_n) \\
&= \exp(-n\lambda_0(x + \delta))M(\lambda_0)^n \int_{\left\{\left|\frac{1}{n}\sum_{j=1}^{n} y_j - x\right| < \delta\right\}} \tilde{\mu}(dy_1) \cdots \tilde{\mu}(dy_n).
\end{aligned}$$

Proof of Cramér's Theorem

By the WLLN, we have

$$\int_{\left\{\left|\frac{1}{n}\sum_{j=1}^{n} y_j - x\right| < \delta\right\}} \tilde{\mu}(dy_1)\cdots\tilde{\mu}(dy_n) \to 1 \text{ as } n \to \infty.$$

Thus

$$\varliminf_{n\to\infty} \frac{1}{n}\log \mu_n(G) \geq -\lambda_0(x+\delta)+\Lambda(\lambda_0) = -I(x)-\lambda_0\delta \text{ for all } 0 < \delta \ll 1.$$

Similar proof can be applied to the case $x < a$. □

Example of Cramér's Theorem

Example (Cramér's theorem applied to Ber(p))

We have $\Lambda(\lambda) = \ln(pe^{\lambda} + q)$ where $q = 1 - p$. The rate function

$$I(x) = \begin{cases} x \log \frac{x}{p} + (1-x) \ln \frac{1-x}{q}, & x \in [0,1], \\ \infty, & otherwise. \end{cases}$$

Here we take the convention $0 \log 0 = 0$. It is obvious that $I(x) \geq 0$, and $I(x)$ achieves its global minimum 0 at $x^* = p$.

Example of Cramér's Theorem

Example (Cramér's theorem applied to Ber(p))

We have $\Lambda(\lambda) = \ln(pe^{\lambda} + q)$ where $q = 1 - p$. The rate function

$$I(x) = \begin{cases} x \log \frac{x}{p} + (1-x) \ln \frac{1-x}{q}, & x \in [0,1], \\ \infty, & otherwise. \end{cases}$$

Here we take the convention $0 \log 0 = 0$. It is obvious that $I(x) \geq 0$, and $I(x)$ achieves its global minimum 0 at $x^* = p$.

- $I(x)$ has important background in information theory. It is called relative entropy, or Kullback-Leibler distance between two distributions μ and ν defined as follows

$$D(\mu||\nu) = \sum_{i=1}^{r} \mu_i \log \frac{\mu_i}{\nu_i},$$

where $\mu = (\mu_1, \mu_2, \ldots, \mu_r)$, $\nu = (\nu_1, \nu_2, \ldots, \nu_r)$. In the previous case, we have $r = 2$, $\mu = (x, 1-x)$ and $\nu = (p, q)$, the underlying Bernoulli distribution.

Connections with statistical mechanics

There are intimate relations between the large deviation theory and equilibrium statistical mechanics [5]. Now let us only consider the simplest case here. For the Bernoulli trials with parameter p, we can obtain the rate function as

$$I(x) = x \ln \frac{x}{p} + (1-x) \ln \frac{1-x}{q}, \ x \in [0,1]$$

which is also called the relative entropy. When $p = 1/2$ we have

$$I(x) = x \ln x + (1-x) \ln(1-x) + \ln 2, \ x \in [0,1].$$

In this case, the rate function is exactly the negative Shannon entropy up to a constant $\ln 2$. Below we will show that it has direct connection to Boltzmann entropy in statistical mechanics.

[5] R.S. Ellis. Entropy, Large deviations, and statistical mechanics. Springer-Verlag, 1st edition edition, 1985.

Connections with statistical mechanics: Entropy

- Consider a system with n independent spins being up or down with equal probability $1/2$. If it is up, we label it as 1, and 0 otherwise. We define the set of microstates as

$$\Omega = \{\omega : \omega = (s_1, s_2, \ldots, s_n), s_i = 1 \text{ or } 0\}.$$

For each microstate ω, we define its mean energy as

$$h_n(\omega) = \frac{1}{n} \sum_{i=1}^{n} s_i.$$

Connections with statistical mechanics: Entropy

- Consider a system with n independent spins being up or down with equal probability $1/2$. If it is up, we label it as 1, and 0 otherwise. We define the set of microstates as

$$\Omega = \{\omega : \omega = (s_1, s_2, \ldots, s_n), s_i = 1 \text{ or } 0\}.$$

  For each microstate ω, we define its mean energy as

$$h_n(\omega) = \frac{1}{n} \sum_{i=1}^{n} s_i.$$

- In thermodynamics, the entropy is a function of the macrostate energy. In statistical mechanics, Boltzmann gives a clear mathematical definition of the entropy

$$S = k_B \ln W$$

  in the micro-canonical ensemble (the number of spins n and total energy $h_n = E$ are fixed in this set-up), where k_B is the Boltzmann constant, W is the number of the microstates corresponding to the fixed energy E. Actually this formula is carved in Boltzmann's tombstone.

Connections with statistical mechanics: LDT

- From large deviation theory we have

$$I(E) = \lim_{n\to\infty} -\frac{1}{n} \ln \mathbb{P}(h_n \in [E, E + dE]),$$

where dE is an infinitesimal quantity and

$$\begin{aligned} I(E) &= \lim_{n\to\infty} -\frac{1}{n} \ln \frac{W(h_n \in [E, E + dE])}{2^n} \\ &= \ln 2 - \frac{1}{k_B} \lim_{n\to\infty} \frac{1}{n} S_n(E). \end{aligned}$$

Connections with statistical mechanics: LDT

- From large deviation theory we have

$$I(E) = \lim_{n\to\infty} -\frac{1}{n} \ln \mathbb{P}(h_n \in [E, E + dE]),$$

where dE is an infinitesimal quantity and

$$\begin{aligned} I(E) &= \lim_{n\to\infty} -\frac{1}{n} \ln \frac{W(h_n \in [E, E + dE])}{2^n} \\ &= \ln 2 - \frac{1}{k_B} \lim_{n\to\infty} \frac{1}{n} S_n(E). \end{aligned}$$

- Taking the normalization of S with $1/n$ in the $n \to \infty$ limit, we obtain

$$k_B I(E) = k_B \ln 2 - S(E),$$

where $S(E)$ is the Boltzmann entropy in statistical mechanics. So we have that the rate function is the negative entropy (with factor $1/k_B$) up to an additive constant. This is a general statement.

Connections with statistical mechanics: Free Energy

In the canonical ensemble in statistical mechanics (the number of spins n and the temperature T are fixed in this set-up), let us investigate the physical meaning of Λ. The logarithmic moment generating function of $H_n(\omega) = nh_n(\omega)$ with normalization $1/n$ is

$$\Lambda(\lambda) = \lim_{n\to\infty} \frac{1}{n} \ln \mathbb{E} e^{\lambda H_n},$$

where we take H_n instead of a single R.V. s_i since it admits more general interpretation. Take $\lambda = -\beta = -(k_B T)^{-1}$, we have

$$\Lambda(-\beta) = \lim_{n\to\infty} \frac{1}{n} \ln(\sum_{\omega} e^{-\beta H_n(\omega)}) - \ln 2.$$

Thus the free energy $F(\beta)$ is the negative logarithmic moment generating function up to a constant.

Connections with statistical mechanics

Define the partition function

$$Z_n(\beta) = \sum_{\omega} e^{-\beta H_n(\omega)}$$

and free energy

$$F_n(\beta) = -\beta^{-1} \ln Z_n(\beta),$$

we have

$$\Lambda(-\beta) = -\beta \lim_{n\to\infty} \frac{1}{n} F_n(\beta) - \ln 2 = -\beta F(\beta) - \ln 2.$$

Connections with statistical mechanics

According to the large deviation theory we have

$$-\beta F(\beta) - \ln 2 = \sup_E \{-\beta E - \ln 2 + k_B^{-1} S(E)\},$$

i.e.

$$F(\beta) = \inf_E \{E - TS(E)\}.$$

The infimum is achieved at the critical point E^* such that

$$\frac{\partial S(E)}{\partial E}\Big|_{E=E^*} = \frac{1}{T},$$

which is exactly a thermodynamic relation between S and T. Here E^* is essentially the internal energy U.