Introductory examples and definitions. Cramér's theorem

Elena Kosygina

July 22, 2018

In this lecture we introduce the large deviation principle (LDP), first by considering empirical means of Bernoulli sequences and then discussing the general definition. The main focus of the lecture and the first problem set is the classical Cramér Theorem on $\mathbb{R}$ and its applications. Mastering the Legendre-Fenchel transform is also one of the goals. This transform and related convex analysis notions play an important role in other areas of mathematics (Hamiltonian mechanics, statistical physics, optimal control, etc.).

First Examples and the Large Deviation Principle

Notation: $\mathcal{B}(\mathbb{R})$ denotes the Borel σ-algegra on $\mathbb{R}$.

Let $(X_i)_{i\in\mathbb{N}}$ be an i.i.d. sequence of random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with a common distribution μ. Denote by μ_n be the distribution of *empirical means*, $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, that is for every Borel set B in $\mathbb{R}$

$$\mu_n(B) = \mathbb{P}(\overline{X}_n \in B). \tag{1}$$

The following limit theorems for empirical means of i.i.d. random variables are familiar from a standard probability course. Probably, the very first one that you learn about is the strong law of large numbers.

Theorem 1 (Strong law of large numbers)**.** *Assume that* $\mathbb{E}|X_i| < \infty$ *and denote* $\mathbb{E}[X_i]$ *by* $\overline{x}$*. Then*

$$\mathbb{P}\left(\omega : \lim_{n\to\infty} \overline{X}_n(\omega) = \overline{x}\right) = 1.$$

Equivalently, we shall also say that $\overline{X}_n \to \overline{x}$ *almost surely (a.s.) or with probability 1.*

We shall be mostly using the weak law of large numbers.

Theorem 2 (Weak law of large numbers)**.** *Assume that* $\mathbb{E}|X_i| < \infty$ *and denote* $\mathbb{E}[X_i]$ *by* $\overline{x}$*. Then the sequence* $(\overline{X}_n)_{n\in\mathbb{N}}$ *converges to* $\overline{x}$ *in probability, i.e. for every* $\epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(|\overline{X}_n - \overline{x}| \geq \epsilon) = 0.$$

The proof of the weak law of large numbers is a simple application of Chebyshev inequality if we assume that $\mathbb{E}[X_i^2] < \infty$. Below this will always be the case. The proofs of the weak and strong laws of large numbers under the stated assumptions can be found, for example, in R. Durrett's book *Probability: Theory and Examples*, Theorems 2.2.9 and 2.4.1 respectively.

Informally, both the weak and the strong laws of large numbers assert that as $n \to \infty$ the empirical means $\overline{X}_n$ approach the mean $\overline{x}$. The central limit theorem tells us that typically the deviations of $\overline{X}_n$ from $\overline{x}$ are of order $1/\sqrt{n}$. For example, there is less than 1/2% chance that these deviations exceed $3\sigma/\sqrt{n}$ where σ is the standard deviation of X_i.

Theorem 3 (Central limit theorem)**.** *Assume that* $\mathbb{E}[X_i^2] < \infty$*. Denote* $\mathbb{E}[X_i]$ *by* $\overline{x}$ *and* $Var(X_i)$ *by* σ^2 *and suppose that* $\sigma \neq 0$*. Then for all* $a, b \in \mathbb{R}$*,* $a < b$*,*

$$\mathbb{P}\left(\frac{a\sigma}{\sqrt{n}} < \overline{X}_n - \overline{x} \leq \frac{b\sigma}{\sqrt{n}}\right) \to \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\,dx \quad \text{as } n \to \infty.$$

Note that the weak law of large numbers and the central limit theorem can be equivalently written in terms of the measures μ_n, $n \in \mathbb{N}$.

The weak law of large numbers: for every $\epsilon > 0$

$$\lim_{n\to\infty} \mu_n((\overline{x}-\epsilon, \overline{x}+\epsilon)) = 1.$$

Denoting by δ_x a Dirac measure at x, we can easily show that the weak law of large numbers is equivalent to the following statement: the sequence of measures $(\mu_n)_{n\in\mathbb{N}}$ *converges weakly* to $\delta_{\overline{x}}$ as $n \to \infty$, i.e. $\mu_n \Rightarrow \delta_{\overline{x}}$ as $n \to \infty$.

Definition: Let μ_n, $n \in \mathbb{N}$, and μ be probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. We say that the sequence $(\mu_n)_{n\in\mathbb{N}}$ converges weakly to μ and write $\mu_n \Rightarrow \mu$ if

$$\int_{\mathbb{R}} f\,d\mu_n \to \int_{\mathbb{R}} f\,d\mu \quad \text{as } n \to \infty$$

for every continuous bounded function $f : \mathbb{R} \to \mathbb{R}$.

Exercise. Show that a sequence $(Y_n)_{n\in\mathbb{N}}$ of random variables converges to a constant $c \in \mathbb{R}$ in probability iff the sequence $(\mu_n)_{n\in\mathbb{N}}$ of distributions of Y_n, $n \in \mathbb{N}$, converges weakly to δ_c.

In our case, $Y_n = \overline{X}_n$, $n \in \mathbb{N}$, and $c = \overline{x}$.

The central limit theorem: for all $a, b \in \mathbb{R}$, $a < b$,

$$\lim_{n\to\infty} \mu_n((\overline{x} + a\sigma/\sqrt{n}, \overline{x} + b\sigma/\sqrt{n}]) = \chi((a,b]) := \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\,dx.$$

Equivalently, the distributions of $\frac{\overline{X}_n - \overline{x}}{\sigma/\sqrt{n}}$ converge weakly to χ.

Deviations of order 1 as opposed to "normal" deviations of order $1/\sqrt{n}$ are said to be large. We shall see below that for many families of distributions the probabilities of large deviations of empirical means from the mean decay exponentially fast with n. The goal is to find appropriate conditions and quantify the rate of decay.

Bernoulli Sequences

Suppose that $\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = 0) = 1/2$. Then μ_n is supported on $A_n = \{i/n,\ i = 0, 1, \ldots, n\} \subset [0,1]$ and assigns a binomial weight to every $i/n \in A_n$:

Exercise: work out the asymmetric case, $P(X_i = 1) = 1 - P(X_i = 0) = p \neq 1/2$.

$$\mu_n\left(\left\{\frac{i}{n}\right\}\right) = \mathbb{P}\left(\overline{X}_n = \frac{i}{n}\right) = \frac{1}{2^n}\binom{n}{i}, \quad i \in \{0, 1, \ldots, n\}.$$

We would like to compute for every $a \in (1/2, 1]$

$$\lim_{n\to\infty} \frac{1}{n} \ln \mu_n\left([a,\infty)\right) = \lim_{n\to\infty} \frac{1}{n} \ln \mu_n\left([a,1]\right).$$

Question: If $a \in [0, 1/2]$, then what is the value of this limit?
Answer: 0.

Denoting by i_n the smallest integer in $[an, n]$, noticing that

$$\max_{i\in[an,n]} \binom{n}{i} = \binom{n}{i_n},$$

The easiest way to check this is to show that $\binom{n}{i}/\binom{n}{i+1} \geq 1$ for all $i \geq n/2$.

and using Lemma 4 and Stirling's formula[1]

$$n! = c_n\sqrt{n}\left(\frac{n}{e}\right)^n, \quad \text{where } c_n \to \sqrt{2\pi} \text{ as } n \to \infty,$$

we calculate

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n} \ln \mu_n([a,1]) &= -\ln 2 + \lim_{n\to\infty} \frac{1}{n} \ln \sum_{i=i_n}^{n} \binom{n}{i} \\ &= -\ln 2 + \lim_{n\to\infty} \frac{1}{n} \ln \max_{i_n\leq i\leq n} \binom{n}{i} \\ &= -\ln 2 + \lim_{n\to\infty} \frac{1}{n} \ln \binom{n}{i_n} \\ &= -\ln 2 - \lim_{n\to\infty} \left[\frac{i_n}{n} \ln \frac{i_n}{n} + \left(1 - \frac{i_n}{n}\right) \ln\left(1 - \frac{i_n}{n}\right)\right].\end{aligned}$$

Lemma 4 (Swapping $\ln\sum$ for $\ln\max$)**.** *Let $a_i^{(n)}$, $1 \leq i \leq N_n$, be arrays of positive numbers and $m_n = \max_{1\leq i\leq N_n} a_i^{(n)}$.*

If $\lim_{n\to\infty} \frac{1}{n} \ln N_n = 0$ then

$$\lim_{n\to\infty} \frac{1}{n} \ln\left(\frac{1}{m_n} \sum_{i=1}^{N_n} a_i^{(n)}\right) = 0.$$

Proof. The statement immediately follows from the inequalities

$$1 \leq \frac{1}{m_n} \sum_{i=1}^{N_n} a_i^{(n)} \leq N_n.$$

□

By the definition of i_n, $i_n/n \to a$ as $n \to \infty$, and we conclude that

$$\lim_{n\to\infty} \frac{1}{n} \ln \mu_n([a,1]) = -\left(\ln 2 + a\ln a + (1-a)\ln(1-a)\right).$$

A symmetric computation shows that for $a \in [0, 1/2)$

$$\lim_{n\to\infty} \frac{1}{n} \ln \mu_n([0,a]) = -\left(\ln 2 + a\ln a + (1-a)\ln(1-a)\right).$$

If we put

$$\mathcal{I}(x) = \begin{cases} \ln 2 + x\ln x + (1-x)\ln(1-x), & \text{if } x \in (0,1); \\ \infty, & \text{if } x \notin (0,1), \end{cases}$$

then upon a moment's reflection we shall agree that, in fact, for any interval $I \subset \mathbb{R}$ of positive length

$$\lim_{n\to\infty} \frac{1}{n} \ln \mu_n(I) = -\inf_{x\in I} \mathcal{I}(x). \tag{2}$$

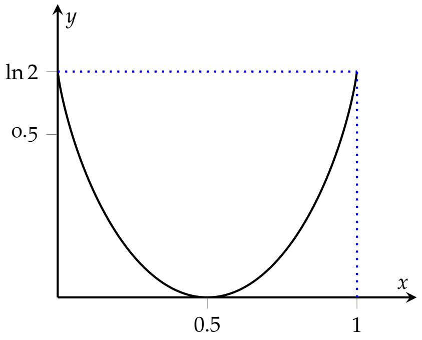

Figure 1: Graph of $\mathcal{I}(x)$.

The infimum over the empty set is conveniently defined to be ∞.

Can we replace I with an arbitrary Borel set B? To see that the limit, in general, need not exist take $I = [1/2, 1/2]$. Then $\mu_n(I) = 0$ for odd n, $\mu_n(I) = \binom{n}{n/2}\frac{1}{2^n}$ for even n, and

$$-\infty = \liminf_{n\to\infty} \frac{1}{n} \ln \mu_n(I) < \limsup_{n\to\infty} \frac{1}{n} \ln \mu_n(I) = 0 = -\mathcal{I}\left(\frac{1}{2}\right).$$

The point $1/2 = E[X_i]$ is special but consider also replacing I with $B = \{1/4, 1/3, 1/\sqrt{5}\}$.

What is the right statement then?

[1] It often suffices to know only that $\ln(n!) = n\ln n - n + O(\ln n)$ as $n \to \infty$.

Large deviation principle

The statement we are looking for is called the large deviation principle (LDP).

Definition 1 (LDP). *A sequence of probability measures on a Polish space $\mathbb{X}$ equipped with the Borel σ-algebra $\mathcal{X}$ satisfies a LDP with a rate function $\mathcal{I} : \mathbb{X} \to [0, \infty]$ if*

(i) $\mathcal{I}$ has compact sub-level sets $\{x \in \mathbb{X} : \mathcal{I}(x) \le \ell\}$ for all $\ell \in [0, \infty)$;

(ii) for every closed set $C \subset \mathbb{X}$

$$\limsup_{n\to\infty} \frac{1}{n} \ln \mu_n(C) \le -\inf_{x\in C} \mathcal{I}(x); \tag{3}$$

(iii) for every open set $O \subset \mathbb{X}$

$$\liminf_{n\to\infty} \frac{1}{n} \ln \mu_n(O) \ge -\inf_{x\in O} \mathcal{I}(x). \tag{4}$$

Sometimes $\mathcal{I}$ satisfying *(i)* is called a "good rate function". A general rate function is supposed to be lower-semi-continuous, i.e. be such that all the sub-level sets in *(i)* are closed. We shall only work with good rate functions.

Since $\mathbb{X}$ is closed and $\mu_n(\mathbb{X}) = 1$, *(ii)* implies that $\inf_{x\in\mathbb{X}} \mathcal{I}(x) = 0$ and, in particular, $\mathcal{I} \not\equiv \infty$.

Putting this differently, if $\mathcal{I}$ satisfies *(i)* then a LDP holds iff for every $B \in \mathcal{X}$

$$-\inf_{x\in B^o} \mathcal{I}(x) \le \liminf_{n\to\infty} \frac{1}{n} \ln \mu_n(B) \le \limsup_{n\to\infty} \frac{1}{n} \ln \mu_n(B) \le -\inf_{x\in\overline{B}} \mathcal{I}(x).$$

Notation: for a Borel set B we denote by $\overline{B}$ the closure of B (intersection of all closed sets containing B), by B^o the interior of B (the union of all open sets contained in B), and by ∂B the boundary of A: $\partial B = \overline{B} \setminus B^o$.

Normal Random variables and other examples

Let $(X_i)_{i\in\mathbb{N}}$ be an i.i.d. sequence of standard normal random variables. Then $\overline{X}_n$ is normal with mean 0 and variance $1/n$, $\mu_n \Rightarrow \delta_0$, and for any interval $I \subset \mathbb{R}$ with endpoints $a < b$

$$\lim_{n\to\infty} \frac{1}{n} \ln \mu_n(I) = \lim_{n\to\infty} \frac{1}{n} \ln \left(\frac{1}{\sqrt{2\pi}} \int_{a\sqrt{n}}^{b\sqrt{n}} e^{-\frac{x^2}{2}}\, dx \right) = -\inf_{x\in I} \frac{x^2}{2}.$$

Exercise: provide the details of this computation.

This calculation gives us a candidate for the rate function: $\mathcal{I}(x) = x^2/2$.

Exercise 1. For each sequence $(\mu_n)_{n\in\mathbb{N}}$ determine if it satisfies a LDP.

(a) $(\mathbb{X}, \mathcal{X}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and μ_n is a uniform measure on $[-n, n]$;

(b) $(\mathbb{X}, \mathcal{X}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and μ_n is a uniform measure on $[-n^{-1}, n^{-1}]$;

(c) $(\mathbb{X}, \mathcal{X}) = ([-1, 1], \mathcal{B}([-1, 1]))$ and μ_n is a uniform measure on $[-1, 1]$.

This is Exercise III.9 on p. 30 in

Frank den Hollander. *Large deviations*, volume 14 of *Fields Institute Monographs*. American Mathematical Society, Providence, RI, 2000

Exercise 2. Find a candidate for the LD rate function for μ_n if $(X_i)_{i\in\mathbb{N}}$ are i.i.d.

(a) exponential;

(b) Poisson

random variables with parameter $\lambda > 0$. Is there a relationship between the two rate functions (associated to the same λ)? *Hint:* If $(N_t)_{t\geq 0}$ is a Poisson process with rate λ and τ_n is the time of the n-th arrival then $N_t = \sup\{n \geq 0 : \tau_n \leq t\}$.
Answer:

$$\text{(a)}\quad \mathcal{I}_{\text{Exp}}(x) = \begin{cases} \lambda x - \ln(\lambda x) - 1, & \text{if } x > 0; \\ \infty, & \text{if } x \leq 0; \end{cases}$$

$$\text{(b)}\quad \mathcal{I}_{\text{Poi}}(x) = \begin{cases} x\ln(x/\lambda) - (x-\lambda), & \text{if } x > 0; \\ \lambda, & \text{if } x = 0; \\ \infty, & \text{if } x < 0, \end{cases}$$

and $\mathcal{I}_{\text{Poi}}(x) = |x|\mathcal{I}_{\text{Exp}}\left(\frac{1}{x}\right)$.

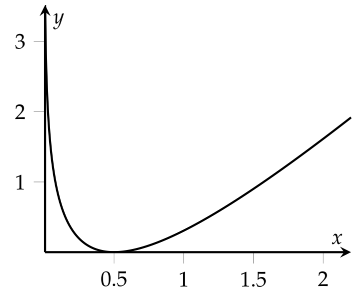

Figure 2: Graph of $\mathcal{I}_{\text{Exp}}(x)$ with $\lambda = 2$.

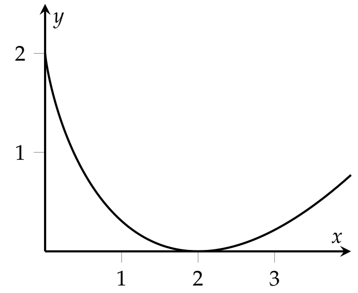

Figure 3: Graph of $\mathcal{I}_{\text{Poi}}(x)$ with $\lambda = 2$.

Cramér's Theorem

We have worked out several examples and could now really appreciate a general theorem which would tell us when distributions μ_n of $\overline{X}_n$ (empirical means of an i.i.d. sequence of random variables) satisfy a LDP and how to compute the rate function.

Let $M : \mathbb{R} \to (0, \infty]$ be the moment generating function (MGF) associated to the distribution measure μ of a random variable X:

$$M(t) = \int_{\mathbb{R}} e^{tx}\, d\mu = \mathbb{E}\left[e^{tX}\right], \quad t \in \mathbb{R}.$$

Its logarithm, $\Lambda(t) := \ln M(t)$, is called the *logarithmic moment generating function* or *cumulant generating function* associated to μ. We shall denote the domain of Λ, $\{t \in \mathbb{R} : \Lambda(t) < \infty\}$, by D_Λ. We note that D_Λ is always an interval and it contains 0. The next lemma summarizes basic properties of Λ.

Harald Cramér (1893 - 1985) was a Swedish mathematician, actuary, and statistician. His paper *Sur un nouveau théorème-limite de la théorie des probabilités*, Actualités scientfiques et industrielles: 736, pp. 2-23, 1938, is considered to be the first on the subject. See `https://arxiv.org/abs/1802.05988` for the original text alongside with English translation by H. Touchette.

Lemma 5. *Let Λ be a logarithmic moment generating function associated to a non-degenerate probability measure μ. Then* *(a)* $\Lambda \in C^\infty(D^o_\Lambda)$ *(infinitely many times differentiable on the interior of its domain) and if* X *is a random variable with distribution* μ *then* $\forall t \in D^o_\Lambda$

$$\Lambda'(t) = \frac{M'(t)}{M(t)} = \frac{\mathbb{E}\left[Xe^{tX}\right]}{\mathbb{E}\left[e^{tX}\right]},$$

$$\Lambda''(t) = \frac{M''(t)}{M(t)} - \left(\frac{M'(t)}{M(t)}\right)^2 = \frac{\mathbb{E}\left[X^2e^{tX}\right]}{\mathbb{E}\left[e^{tX}\right]} - \left(\frac{\mathbb{E}\left[Xe^{tX}\right]}{\mathbb{E}\left[e^{tX}\right]}\right)^2.$$

In particuar, if $0 \in D^o_\Lambda$ *then* $\Lambda'(0) = \mathbb{E}[X]$ *and* $\Lambda''(0) = Var(X)$*.*

(b) $\Lambda(t)$ *is strictly convex.*

Proof. (a) Since $M > 0$ and $D_M = D_\Lambda$, it is enough to prove that $M \in C^\infty(D^o_M)$. Let $t \in D^o_\Lambda$. Then there is a $\delta > 0$ such that $(t - 2\delta, t + 2\delta) \subset D^o_M$. We look at the difference quotients

$$\frac{M(t+h) - M(t)}{h} = \int_{\mathbb{R}} e^{tx}\,\frac{e^{hx} - 1}{h}\,d\mu$$

and note that

$$\lim_{h\to 0} e^{tx}\,\frac{e^{hx} - 1}{h} = xe^{tx}.$$

Moreover, for all h such that $0 < |h| \le \delta$

$$\left| e^{tx}\frac{e^{hx} - 1}{h}\right| \le e^{tx}\frac{e^{\delta|x|} - 1}{\delta} \le \frac{1}{\delta}\left(e^{(t-\delta)x} + e^{(t+\delta)x}\right).$$

The right hand side of the last inequality is integrable with respect to μ by our choice of δ. By the dominated convergence theorem we conclude that

$$\lim_{h\to 0}\frac{M(t+h) - M(t)}{h} = \lim_{h\to 0}\int_{\mathbb{R}} e^{tx}\,\frac{e^{hx} - 1}{h}\,d\mu = \int_{\mathbb{R}} xe^{tx}\,d\mu = \mathbb{E}\left[Xe^{tX}\right].$$

The proof for the second and higher derivatives is very similar. The formulas for Λ' and Λ'' follow simply by the chain and quotient rules.

(b) Let $t \in D^o_\Lambda$. For this t define a new probability measure ν_t by

$$\nu_t(B) = \frac{1}{M(t)}\int_B e^{tx}\,d\mu \quad \forall B \in \mathcal{B}(\mathbb{R}).$$

Let Y_t be a random variable with distribution ν_t. Since $t \in D^o_\Lambda = D^o_M$, the MGF of Y_t, M_{Y_t}, is finite in some neighborhood of the origin: there is a $\delta > 0$ such that for all $|s| < \delta$

$$\mathbb{E}\left[e^{sY_t}\right] = \int_{\mathbb{R}} e^{sy}\,d\nu_t = \frac{1}{M(t)}\int_{\mathbb{R}} e^{(t+s)y}\,d\mu = \frac{M(t+s)}{M(t)} < \infty.$$

Another proof of (b) can be given using Hölder inequality. Take $\alpha \in (0,1)$ and $t_1, t_2 \in D_\Lambda$. Then Hölder inequality with $f(x) = e^{\alpha t_1 x}$, $g(x) = e^{(1-\alpha)t_2 x}$, $p = 1/\alpha$, $q = 1/(1-\alpha)$ gives

$$\begin{aligned} M(\alpha t_1 + (1-\alpha)t_2)) &= \int_{\mathbb{R}} e^{\alpha t_1 x + (1-\alpha)t_2 x}\,d\mu \\ &= \int_{\mathbb{R}} \left(e^{t_1 x}\right)^\alpha \left(e^{t_2 x}\right)^{1-\alpha}\,d\mu \\ &\le \left(\int_{\mathbb{R}} e^{t_1 x}\,d\mu\right)^\alpha \left(\int_{\mathbb{R}} e^{t_2 x}\,d\mu\right)^{1-\alpha} \\ &= (M(t_1))^\alpha (M(t_2))^{1-\alpha}. \end{aligned}$$

Taking the logarithm of both sides we get the convexity of Λ. To show that for $t_1 \ne t_2$ and $\alpha \in (0,1)$ the inequality is strict we recall that for given $p, q > 1$, $p^{-1} + q^{-1} = 1$

$$\|fg\|_{L^1} = \|f\|^{1/p}_{L^p}\|g\|^{1/q}_{L^q}$$

if and only if

$$\frac{|f|^p}{\|f\|^p_{L_p}} = \frac{|g|^q}{\|g\|^q_{L_q}} \quad \mu\text{-a.s..}$$

If for for some $t_1 \ne t_2, \alpha \in (0,1)$ we had an equality then we would have that $e^{t_1 x}/M(t_1) = e^{t_2 x}/M(t_2)$ μ-a.s.. Since we assumed that $t_1 \ne t_2$, μ has to be degenerate. This contradicts our assumption.

By part (a) we conclude that the logarithmic MGF of Y_t, Λ_{Y_t}, is smooth near the origin and $\Lambda''_{Y_t}(0) = \mathrm{Var}(Y_t) > 0$ (μ and, hence, ν are non-degenerate). But note that

$$\Lambda''(t) = \Lambda''_{Y_t}(s)\Big|_{s=0} = \mathrm{Var}(Y_t).$$

Thus, we have shown that for every $t \in D^o_\Lambda$ the second derivative of Λ at t is strictly positive. This proves strict convexity of Λ. □

Theorem 6 (Cramér's Theorem)**.** *Suppose that* $0 \in D^o_\Lambda$. *Then measures* μ_n *satisfy a LDP with the rate function*

$$\mathcal{I}(x) = \sup_{t\in\mathbb{R}}(xt - \Lambda(t)), \quad x \in \mathbb{R}. \tag{5}$$

Exercise. Show that under the conditions of Cramér's theorem
(a) $\mathbb{E}[|X_i|] < \infty$;
(b) $\mathcal{I}(\overline{x}) = 0$.

First of all, let's see why under the conditions of the theorem

$$\mathcal{I}(x) = \sup_{t\in\mathbb{R}}(xt - \Lambda(t))$$

satisfies the conditions of Definition 1. Note that

$$\mathcal{I}(x) \geq xt - \Lambda(t), \quad \forall t \in \mathbb{R}. \tag{6}$$

- Setting $t = 0$ in (6) we see that $\mathcal{I}(x) \geq 0$.
- If $\mathcal{I}(x_k) \leq c$ for all $k \in \mathbb{N}$ and $x_k \to x$ then for all $t \in \mathbb{R}$

  $$c \geq \liminf_{k\to\infty} \mathcal{I}(x_k) \geq \liminf_{k\to\infty}(x_k t - \Lambda(t)) = xt - \Lambda(t).$$

  Taking the supremum of the right hand side over $t \in \mathbb{R}$ we conclude that $I(x) \leq c$. This shows that the sub-level sets of $\mathcal{I}$ are closed, i.e. $\mathcal{I}$ is lower semi-continuous.
- Finally we note that by (6) for every $t \in \mathbb{R}$ and $x \neq 0$

  $$\frac{\mathcal{I}(x)}{|x|} \geq t \operatorname{sign} x - \frac{\Lambda(t)}{|x|}.$$

  Since $\Lambda(t) < \infty$ for all $|t| < \delta$, we can choose $t = \frac{1}{2}\delta \operatorname{sign} x$ and get that

  $$\liminf_{|x|\to\infty} \frac{\mathcal{I}(x)}{|x|} \geq \delta > 0.$$

  Thus, $\mathcal{I}(x) \to \infty$ as $|x| \to \infty$. This implies that sub-level sets of $\mathcal{I}$ are bounded. Since every closed and bounded subset of $\mathbb{R}$ is compact, part *(i)* of the Definition 1 is also satisfied, and $\mathcal{I}$ is a rate function.

A proof of Cramér's theorem under slightly varying assumptions can be found in virtually every book on large deviations. We shall concentrate on two main ideas of the proof. These ideas are very useful in their own right.

Idea I: Chernoff bound, a.k.a. exponential Chebyshev inequality

The upper bound is about closed sets. Since $0 \in D^o_\Lambda$, there is a $\delta > 0$ such that $M(t) < \infty$ for all t with $|t| < \delta$. This implies, in particular, that all moments of X_i are finite.

Let us consider again an interval $[x, \infty)$ for some $x > \overline{x} := E[X_i]$, and try to estimate $\mu_n([x,\infty))$. For for all $t > 0$

$$\mu_n([x,\infty)) = \mathbb{P}(\overline{X}_n \geq x) = \mathbb{P}\left(e^{tn\overline{X}_n} \geq e^{tnx}\right).$$

A combination of exponentiation and Markov inequality is widely known as "Chernoff bound" as well as "exponential Chebyshev inequality". According to Herman Chernoff, the bound was suggested to him by Herman Rubin (see p. 340 of J. Bather. *A Conversation with Herman Chernoff* at `http://www.jstor.org/stable/2246029`). Wouldn't it be appropriate then to rename it to "Herman bound"? :-)

Applying Markov inequality and using independence we get

$$\mu_n([x,\infty)) \leq e^{-tnx}\mathbb{E}\left[e^{tn\overline{X}_n}\right] = e^{-tnx}\left(M(t)\right)^n = e^{-n(xt-\Lambda(t))}.$$

For some t the right hand side can be infinite. But we know that for $t \in (0, \delta)$ it is finite. Minimizing the right hand side over $t \geq 0$ (for $t = 0$ the inequality holds trivially) we see that for all $n \in \mathbb{N}$

$$\frac{1}{n}\ln \mu_n([x,\infty)) \leq \inf_{t\geq 0}(-xt + \Lambda(t)) = -\sup_{t\geq 0}(xt - \Lambda(t)).$$

In fact, for $x > \overline{x}$

$$\sup_{t\geq 0}(xt - \Lambda(t)) = \sup_{x\in\mathbb{R}}(xt - \Lambda(t)) = \mathcal{I}(x). \tag{7}$$

Indeed,

- by the concavity of the logarithm, for $t < 0$ and $x > \overline{x}$ we have

  $$xt - \ln \mathbb{E}\left[e^{tX_i}\right] < \overline{x}t - t\mathbb{E}[X_i] = 0,$$

  and, thus, $\sup_{t<0}(xt - \Lambda(t)) \leq 0$;

- on the other hand,

  $$\sup_{t\geq 0}(xt - \Lambda(t)) \geq xt - \Lambda(t)\Big|_{t=0} = 0.$$

We conclude that

$$\frac{1}{n}\ln \mu_n([x,\infty)) \leq -\mathcal{I}(x).$$

A similar argument shows that for $x < \overline{x}$

$$\frac{1}{n}\ln \mu_n((-\infty, x]) \leq -\mathcal{I}(x).$$

Now let C be an arbitrary closed set in $\mathbb{R}$. If $\inf_{x\in C}\mathcal{I}(x) = 0$ then the LD upper bound trivially holds. Therefore, it is enough to consider the case when $\inf_{x\in C}\mathcal{I}(x) > 0$. Since $\mathcal{I}(\overline{x}) = 0$, we conclude that $\overline{x} \in C^c$, which is an open set. Take the union of all open intervals in C^c which contain $\overline{x}$. It is an interval $(x_\ell, x_r) \subset C^c$ so that $C \subset (x_\ell, x_r)^c$. Observe also that at least one of the endpoints of (x_ℓ, x_r) is finite as C is non-empty.

Every open set in $\mathbb{R}$ is a countable disjoint union of open intervals.

If $x_\ell > -\infty$ then $x_\ell \in C$ and $\mathcal{I}(x_\ell) \geq \inf_{x\in C}\mathcal{I}(x)$. In the same way, if $x_r < \infty$ then $\mathcal{I}(x_r) \geq \inf_{x\in C}\mathcal{I}(x)$. Applying upper bounds obtained earlier for $(-\infty, x_\ell]$ and $[x_r, \infty)$ we get

$$\mu_n(C) \leq \mu_n((-\infty, x_\ell]) + \mu_n([x_r, \infty)) \leq 2e^{-n\inf_{x\in C}\mathcal{I}(x)}.$$

This proves the required LD upper bound.

Idea II: change of measure, a.k.a. exponential "tilting"

Lower bound is about open sets. We shall not give all details here but restrict ourselves to a typical case in which the idea of the proof is most transparent.

Let O be an open set. If $\overline{x} \in O$ then there is a $\delta > 0$ such that $(\overline{x}-\delta, \overline{x}+\delta) \subset O$. By the weak law of large numbers,

$$\mu_n((\overline{x}-\delta, \overline{x}+\delta)) = \mathbb{P}(|\overline{X}_n - \overline{x}| < \delta) \to 1 \quad \text{as } n \to \infty,$$

and, thus,

$$\liminf_{n\to\infty} \frac{1}{n} \ln \mu_n(O) \geq \liminf_{n\to\infty} \frac{1}{n} \ln \mu_n((\overline{x}-\delta, \overline{x}+\delta)) = 0 = -\inf_{x\in O}\mathcal{I}(x).$$

This is the simplest case, and it gives us two ideas. First, we note that it is enough to show that for each $x_0 \in O$ and all $\delta > 0$

$$\liminf_{n\to\infty} \frac{1}{n} \ln \mu_n((x_0-\delta, x_0+\delta)) \geq -\mathcal{I}(x_0). \tag{8}$$

The second idea is, for a given $x_0 \neq \overline{x}$, to introduce a new "tilted" measure $\widetilde{\mu}$ with the mean equal to x_0, and perform a change of measure from μ to $\widetilde{\mu}$, and, thus, from μ_n to $\widetilde{\mu}_n$. Then the event

$$\{|\overline{X}_n - x_0| < \delta\}$$

will be a typical event for the new measure, i.e.

$$\widetilde{\mu}_n((x_0-\delta, x_0+\delta)) \to 1, \quad \text{as } n \to \infty,$$

and a lower bound will be obtained simply by estimating "the cost" of this change of measure.

Here are the details. Assume that $\mathcal{I}(x_0) < \infty$.[2] In addition, suppose that the supremum in (5) is attained at some $t_0 \in \mathbb{R}$ so that[3]

[2] If $\mathcal{I}(x_0) = \infty$ then (8) holds trivially.

[3] Our assumptions imply, in particular, that $t_0 \in D_\Lambda^o$, i.e. $\Lambda(t_0) < \infty$.

$$\mathcal{I}(x_0) = x_0 t_0 - \Lambda(t_0),$$

where

$$x_0 = \Lambda'(t) = \frac{M'(t_0)}{M(t_0)} = \frac{1}{M(t_0)} \int_{\mathbb{R}} x e^{x t_0} d\mu =: \int_{\mathbb{R}} x \, d\widetilde{\mu},$$

and we defined for every $A \in \mathcal{B}$

$$\widetilde{\mu}(A) := \frac{1}{M(t_0)} \int_A e^{x t_0} \, d\mu \quad \text{(in short, } d\mu = \frac{e^{x t_0}}{M(t_0)} \, d\widetilde{\mu}).$$

This is exactly the measure we are looking for: by the weak law of large numbers for every $\epsilon > 0$

$$\widetilde{\mu}_n((x_0 - \epsilon, x_0 + \epsilon)) \to 1 \quad \text{as } n \to \infty.$$

Moreover, for every $\epsilon \in (0, \delta)$

$$\begin{aligned}
\mu_n((x_0 - \delta, x_0 + \delta)) \geq & \mu_n((x_0 - \epsilon, x_0 + \epsilon)) = \mathbb{P}\left(|\overline{X}_n - x_0| < \epsilon\right) \\
&= \underset{\left|\sum_{i=1}^n x_i - x_0 n\right| < \epsilon n}{\int \cdots \int} d\mu(x_1) \ldots d\mu(x_n) \\
= (M(t_0))^n & \underset{\left|\sum_{i=1}^n x_i - x_0 n\right| < \epsilon n}{\int \cdots \int} e^{-t_0 \sum_{i=1}^n x_i} d\widetilde{\mu}(x_1) \ldots d\widetilde{\mu}(x_n) \\
= e^{n(\Lambda(t_0) - x_0 t_0)} & \underset{\left|\sum_{i=1}^n x_i - x_0 n\right| < \epsilon n}{\int \cdots \int} e^{-t_0 \left(\sum_{i=1}^n x_i - n x_0\right)} d\widetilde{\mu}(x_1) \ldots d\widetilde{\mu}(x_n) \\
\geq e^{n(\Lambda(t_0) - x_0 t_0 - \epsilon |t_0|)} & \underset{\left|\sum_{i=1}^n x_i - x_0 n\right| < \epsilon n}{\int \cdots \int} d\widetilde{\mu}(x_1) \ldots d\widetilde{\mu}(x_n) \\
&= e^{n(\Lambda(t_0) - x_0 t_0 - \epsilon |t_0|)} \widetilde{\mu}_n((x_0 - \epsilon, x_0 + \epsilon)).
\end{aligned}$$

Taking the logarithm, dividing by n, passing to the $\liminf$ as $n \to \infty$, and finally letting $\epsilon \downarrow 0$ we get (8).

The assumption in Cramér's theorem is weaker, it does not guarantee, in general, the existence of such $t_0 \in \mathbb{R}$ for every x_0 with $\mathcal{I}(x_0) < \infty$. Thus, a more careful analysis is needed. For details and an even more general version of Cramér's theorem on $\mathbb{R}$ we refer the reader to [4]Theorem 2.2.3.

Consider, for example, the case of Poisson distribution and take $x_0 = 0$. For this x_0 the supremum in (5) is "attained" at $t_0 = -\infty$.

Properties of the Legendre-Fenchel transform

We discuss the transformation defined by the right hand side of (5) in more detail. Recall from Lemma 5 that a logarithmic MGF Λ is always C^∞ and strictly convex in D_Λ^o. Note that if $D_\Lambda \neq \{0\}$ then D_Λ^o is a non-empty interval which either contains the origin or has it as one of its end points. The strict convexity of Λ on D_Λ^o implies that

Exercise. Show that if $D_\Lambda \neq \{0\}$ then D_Λ^o is a non-empty open interval which either contains the origin or has it as one of its end points.

[4] Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications*, volume 38 of *Applications of Mathematics*. Springer-Verlag, New York, second edition, 1998

$\Lambda'(t)$ is increasing on D^o_Λ. If we now define

$$x^- = \inf_{t \in D^o_\Lambda} \Lambda'(t) \quad \text{and} \quad x^+ = \sup_{t \in D^o_\Lambda} \Lambda'(t), \tag{9}$$

then Λ' is a bijection from D^o_Λ onto (x^-, x^+). Thus, for each $x \in (x^-, x^+)$ the equation

$$x = \Lambda'(t) \ \text{ has a unique solution } \ t_x \in D^o_\Lambda. \tag{10}$$

Lemma 7. *Let Λ be a logarithmic MGF of a distribution μ of a random variable X and define*

$$\Lambda^*(x) = \sup_{t \in \mathbb{R}} (tx - \Lambda(t)) = \sup_{t \in D_\Lambda} (tx - \Lambda(t)). \tag{11}$$

Λ^* has the following properties.

(a) *Λ^* is a convex function on $\mathbb{R}$.*

(b) *If $D_\Lambda = \{0\}$ then $\Lambda^* \equiv 0$. If $D_\Lambda \neq \{0\}$ then (x^-, x^+) defined in* (9) *is non-empty, $\Lambda^* \in C^\infty((x^-, x^+))$, and for every $x \in (x^-, x^+)$ and t_x as in* (10)

$$\Lambda^*(x) = xt_x - \Lambda(t_x).$$

(c) $\inf_{x \in \mathbb{R}} \Lambda^*(x) = 0$. *If $\mathbb{E}|X| < \infty$ then $\Lambda^*(\mathbb{E}[X]) = 0$, i.e. the infimum is attained at $\mathbb{E}[X]$.*

Parts of the proof of this lemma follow the proof of Lemma 2.2.5 in .

Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications*, volume 38 of *Applications of Mathematics*. Springer-Verlag, New York, second edition, 1998

Proof. Statement (a) follows from the definition of Λ^*: for all $x, y \in \mathbb{R}$ and $\alpha \in [0,1]$

$$\begin{aligned}\Lambda^*(\alpha x + (1-\alpha)y) &= \sup_{t \in \mathbb{R}} (t(\alpha x + (1-\alpha)y) - \Lambda(t)) = \sup_{t \in \mathbb{R}} (\alpha(tx - \Lambda(t)) + (1-\alpha)(ty - \Lambda(t))) \\ &\leq \sup_{t \in \mathbb{R}} (\alpha(tx - \Lambda(t)) + \sup_{t \in \mathbb{R}} ((1-\alpha)(ty - \Lambda(t)) = \alpha\Lambda^*(x) + (1-\alpha)\Lambda^*(y).\end{aligned}$$

The first statement in (b) is trivial. Assume now that $D_\Lambda \neq \{0\}$. As we pointed out at the beginning of this subsection, this means that D^o_Λ is a non-empty open interval. Hence, $(x^-, x^+) \neq \emptyset$ and for every $x \in (x^-, x^+)$ there is a unique solution t_x of $x = \Lambda'(t)$ (see (9) and (10)). Since the function $t \mapsto tx - \Lambda(t)$ is concave, its supremum is attained at t_x, and we have $\Lambda^*(x) = xt_x - \Lambda(t_x)$. What is left to show is that $\Lambda^* \in C^\infty((x^-, x^+))$. By Lemma 5, $\Lambda' \in C^\infty(D^o_\Lambda)$ and is increasing on D^o_Λ. The implicit function theorem implies that the function $x \mapsto t_x$ defined on (x^-, x^+) is $C^\infty((x^-, x^+))$. Recalling that $\Lambda^*(x) = xt_x - \Lambda(t_x)$ on (x^-, x^+) we conclude that $\Lambda^* \in C^\infty((x^-, x^+))$ as claimed.

(c) If $D_\Lambda = \{0\}$ then the statement is obvious from (b). Hence, assume that $D_\Lambda \neq \{0\}$. As before, D^o_Λ is then a non-empty interval. If it contains the origin then for $\overline{x} = \Lambda'(0) = \mathbb{E}[X]$ (see Lemma 5) by part (b) we have that $\Lambda^*(\overline{x}) = 0$, and we are done, since Λ^* is non-negative by the definition (put $t = 0$ in the formula for Λ^*).

Suppose now that 0 is the left endpoint of D^o_λ. This implies that $-\infty \leq \overline{x} < \infty$. Indeed, take any $t \in D^o_\Lambda$. Then $t > 0$ and

$$\overline{x} = \int_{\mathbb{R}} x\, d\mu \leq \int_0^\infty x\, d\mu \leq \frac{1}{t}\int_0^\infty tx + 1\, d\mu \leq \frac{1}{t}\int_0^\infty e^{tx}\, d\mu < \infty.$$

By Jensen inequality,

$$\Lambda(t) = \ln \mathbb{E}\left[e^{tX}\right] \geq \mathbb{E}\left[\ln e^{tX}\right] = t\overline{x}.$$

If $\overline{x} \in \mathbb{R}$ then

$$0 \leq \Lambda^*(\overline{x}) = \sup_{t\in\mathbb{R}}(t\overline{x} - \Lambda(t)) \leq \sup_{t\in\mathbb{R}}(t\overline{x} - t\overline{x}) = 0.$$

If $\overline{x} = -\infty$ then $\Lambda(t) = \infty$ for all $t < 0$ and $\Lambda^*(x) = \sup_{t\geq 0}(tx - \Lambda(t))$. But this implies that Λ^* is non-decreasing on $\mathbb{R}$ and $\inf_{x\in\mathbb{R}} \Lambda^*(x) = \lim_{x\to-\infty} \Lambda^*(x)$. Moreover, by Chernoff bound,

$$\ln \mu([x,\infty)) \leq \inf_{t\geq 0} \ln \mathbb{E}\left[e^{t(X-x)}\right] = \inf_{t\geq 0}(-tx + \Lambda(t)) = -\Lambda^*(x).$$

Therefore,

$$0 \leq \inf_{x\in\mathbb{R}} \Lambda^*(x) = \lim_{x\to-\infty} \Lambda^*(x) \leq -\lim_{x\to-\infty} \ln \mu([x,\infty)) = 0.$$

The case when 0 is the right endpoint of D^o_Λ can be reduced to the already considered case by considering the logarithmic MGF of $-X$.

□

The transformation $\Lambda \mapsto \Lambda^*$ is called a *Legendre-Fenchel transform* of Λ. In fact, Legendre-Fenchel transform can be defined for an arbitrary function $f : \mathbb{R} \to [-\infty, \infty]$,

$$f^*(x) = \sup_{t\in\mathbb{R}}(tx - f(t)).$$

The function f^* is convex and is called a convex conjugate of f. In this very general setting we always have that $f \leq (f^*)^*$. An important result is Fenchel-Moreau Theorem.[5]

Theorem 8. *Assume that $f : \mathbb{R} \to (-\infty, \infty]$ and $f \not\equiv \infty$. Then $f = (f^*)^*$ if and only if f is convex and lower semi-continuous.*

[5] see, for example, p. 53 of Firas Rassoul-Agha and Timo Seppäläinen. *A course on large deviations with an introduction to Gibbs measures*, volume 162 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2015

References

Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications*, volume 38 of *Applications of Mathematics*. Springer-Verlag, New York, second edition, 1998.

Frank den Hollander. *Large deviations*, volume 14 of *Fields Institute Monographs*. American Mathematical Society, Providence, RI, 2000.

Firas Rassoul-Agha and Timo Seppäläinen. *A course on large deviations with an introduction to Gibbs measures*, volume 162 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2015.