Advanced Probability Theory
Supplementary Materials 2

Xinyi Li

BICMR, Peking University

December 29, 2020

Based on lecture materials of Yuan Zhang from Fall 2018

Caveat

Remark: Similar results does not hold for odd order derivatives:

- Recall that if a random variable X has moments up to k-th order, then the characteristic function is k times continuously differentiable on the entire real line.
- However, if a characteristic function φ_X has a k-th derivative at zero, then the random variable X has all moments up to k if k is even, but only up to $k-1$ if k is odd.

When is a function the ch.f. of a probability measure?

Definition

Given φ *and* $x_1, \ldots, x_n \in \mathbb{R}$, *we can consider the matrix with* (i, j) *entry given by* $\varphi(x_i - x_j)$. *Call* φ *positive definite if this matrix is always positive semi-definite Hermitian.*

Theorem (Bochner's theorem (not required))

A function from $\mathbb{R}$ *to* $\mathbb{C}$ *which is continuous at origin with* $\varphi(0) = 1$ *is a ch.f. of some probability measure on* $\mathbb{R}$ *if and only if it is positive definite.*

Deviations, typical and atypical

- Let $X_1, X_2, \cdots$ be i.i.d. with $E[X_1] = \mu$ and let
$$S_n = X_1 + X_2 + \cdots + X_n.$$
  - According to CLT, the **typical** value of $S_n - n\mu$ is $O(\sqrt{n})$.
  - What about **atypical** deviations of $S_n - n\mu$?
  - According to (W)LLN, we know that for any $a > \mu$,
$$P(S_n > na) \to 0.$$
- In this lecture, we consider the **large deviation problem**, i.e., the speed that $P(S_n > na)$ goes to 0 as $n \to \infty$.
- Large deviation is in some sense a generalization of WLLN. People are also interested in deviations in between the scale Researches in this direction is called **moderate deviation**.

Theorem (Moderate deviations (not required))

Let X_i's be absolutely integrable i.i.d. with mean μ, variance $\sigma^2 > 0$ and finite M.G.F. $E[e^{\theta X_1}] < \infty$ for all $\theta \in \mathbb{R}$. Then

$$\lim_{n\to\infty} \frac{1}{n^{2\alpha-1}} \log P[S_n - \rho n \geq y n^{\alpha}] = -\frac{y^2}{2\sigma^2} \quad (y > 0,\ \frac{1}{2} < \alpha < 1).$$

Moment generating function and large deviation

- We will show that if the **moment generating function**

$$\psi(\theta) = E[\exp(\theta X_1)] < \infty$$

for some $\theta > 0$, then $P(S_n > na) \to 0$ exponentially fast.

- A crude observation: S_n has M.G.F. $\psi^n(\theta)$ and hence for some $\theta > 0$,

$$P(S_n \geq na)e^{\theta na} \leq E[e^{\theta S_n}] = \psi^n(\theta) < \infty.$$

In other words, $P(S_n \geq na) \leq \exp\Big[n\big(\log\psi(\theta) - \theta a\big)\Big]$, and when "$a$ is big", we see there is an exponential decay in n.

Question: Recall that $P(S_n \geq na) \leq \exp\Big[n\big(\log\psi(\theta) - \theta a\big)\Big]$. Do we really have exponential decay? Do we have a lower bound on the probability (= upper bound on the rate)?

Supermultiplicativity

- Let $\pi_n = P(S_n \geq na)$. Then
$$\pi_{n+m} \geq P(S_n \geq na,\ S_{n+m} - S_n \geq ma) = \pi_n \pi_m.$$
- So we let $\gamma_n = \log \pi_n$, $\gamma_{n+m} \geq \gamma_n + \gamma_m$.

Lemma

As $n \to \infty$ the limit of γ_n exists and
$$\lim_{n\to\infty} \frac{\gamma_n}{n} = \sup_{n \in \mathbb{Z}^+} \frac{\gamma_n}{n}.$$

Proof: It suffices to show that for any m, and $\epsilon > 0$, $\gamma_n/n \geq \gamma_m/m - \epsilon$ for all sufficiently large n. □

Question: What if for some δ_n, we have $\delta_{n+m} \leq \delta_n \delta_m$?

Large Deviations

- We define $\gamma(a) = \lim_{n\to\infty} \gamma_n/n \leq 0$. Then for any distribution and any n and a,

$$P(S_n \geq na) \leq e^{n\gamma(a)}.$$

- So we now want to prove $\gamma(a) < 0$ for $a > \mu$.
- In the rest of this lecture, we **assume**

$$\psi(\theta_0) = E[\exp(\theta_0 X_1)] < \infty$$

for some $\theta_0 > 0$.

Large Deviations

Recall that for some $\theta > 0$, letting $\kappa(\theta) = \log \psi(\theta)$, we have

$$P(S_n \geq na) \leq \exp\Big(-n[a\theta - \kappa(\theta)]\Big).$$

Lemma

If $a > \mu$, then $a\theta - \kappa(\theta) > 0$ for all sufficiently small θ.

Proof: Note that $\kappa(0) = 0$. It suffices to prove

1. $\psi(\theta)$ is right continuous at 0.
2. $\psi(\theta)$ is differentiable on $(0, \theta_0)$.
3. $\kappa'(\theta) \to \mu$ as $\theta \to 0$.

To check (iii), observe that $\kappa'(\theta) = \psi'(\theta)/\psi(\theta)$. □

Large Deviations

- With the lemma above, we will further strengthen our upper bounds by finding the the maximum of
$$\lambda(\theta) = a\theta - \kappa(\theta).$$
- Let
$$\theta_+ = \sup\{\theta : \ \psi(\theta) < \infty\}, \quad \theta_- = \inf\{\theta : \ \psi(\theta) < \infty\}.$$
- Note that $\psi(\theta) < \infty$ for all $\theta \in (\theta_-, \theta_+)$.
- Now since that $\psi(\theta) \in C^\infty$ within (θ_-, θ_+), we have
$$\lambda'(\theta) = a - \frac{\psi'(\theta)}{\psi(\theta)}.$$
So the maximal point of θ must satisfy $\psi'(\theta)/\psi(\theta) = a$.
- For the existence and uniqueness of such point(s), we introduce a new distribution, and use a trick named "tilting".

Large Deviations

- We now introduce the distribution F_θ by "reweighting F":

$$F_\theta(x) = \frac{1}{\psi(\theta)} \int_{-\infty}^{x} e^{y\theta} dF(y).$$

Then it is easy to compute its mean:

$$\int x dF_\theta(x) = \frac{1}{\psi(\theta)} \int_{-\infty}^{\infty} y e^{\theta y} dF(y) = \frac{\psi'(\theta)}{\psi(\theta)}.$$

- Take a another derivative, we have

$$\psi''(\theta) = \int_{-\infty}^{\infty} x^2 e^{\theta x} dF(x).$$

- So we have

$$\frac{d}{d\theta} \frac{\psi'(\theta)}{\psi(\theta)} = \frac{\psi''(\theta)}{\psi(\theta)} - \left(\frac{\psi'(\theta)}{\psi(\theta)} \right)^2 = \int x^2 F_\theta(x) - \left(\int x F_\theta(x) \right)^2$$

which is the variance of distribution F_θ and ≥ 0.

Large Deviations

- Now we further assume that X_1 is **not** a.s. a constant. Then F_θ is not a Dirac-mass either.
- By the calculation from the last slide, $\psi'(\theta)/\psi(\theta)$ is strictly increasing.
- So $\lambda(\theta)$ is a concave function.
- And note that $a > \mu$, which implies $\lambda'(\theta) > 0$ for all sufficiently small $\theta > 0$. This implies that the θ satisfying $\psi'(\theta)/\psi(\theta) = a$ has to be unique.
- We denote such point (if exists) by θ_a.

Large Deviations

If such θ_a exists, we can show the upper bound at θ_a is asymptotically sharp:

Theorem

For non-constant X_i's such that

$$\psi(\theta_0) = E[\exp(\theta_0 X_1)] < \infty$$

for some $\theta_0 > 0$, suppose there further exists some θ_a such that $\psi'(\theta_a)/\psi(\theta_a) = a$. Then as $n \to \infty$,

$$n^{-1} \log[P(S_n \geq na)] \to -\lambda(\theta_a).$$

Proof: By earlier discussions, we have already proved that

$$\limsup_{n\to\infty} n^{-1} \log[P(S_n \geq na)] \leq -\lambda(\theta_a).$$

So it suffices to concentrate on proving the other direction.

Large Deviations

We now prove

$$\liminf_{n\to\infty} n^{-1} \log[P(S_n \geq na)] \geq -\lambda(\theta_a).$$

- Pick $\theta \in (\theta_a, \theta_+)$ (we will see later why we require $\theta > \theta_a$);
- Let $X_1^\theta, X_2^\theta, \ldots$ be i.i.d. with distribution F_θ;
- write $S_n^\theta = X_1^\theta + \cdots + X_n^\theta$.

Recall that

$$F_\theta(x) = \frac{1}{\psi(\theta)} \int_{-\infty}^{x} e^{\theta y} dF(y).$$

For each $A \in \mathcal{R}$,

$$\int_A e^{\theta y} dF(y) = 0 \implies \int_A dF(y) = 0$$

Thus F is absolutely continuous with respect to F_θ (denoted $F \ll F_\theta$).

Radon-Nikodym Derivative

Theorem (Radon-Nikodym Theorem)

Let μ and ν be σ-finite measures on $(\Omega, \mathcal{F})$. If $\nu \ll \mu$, there is a function $f \in \mathcal{F}$ so that for all $A \in \mathcal{F}$,

$$\int_A f \, d\mu = \nu(A).$$

This f is usually denoted $d\nu/d\mu$ and called the **Radon-Nikodym derivative.**

Since $F \ll F_\theta$, we see that the Radon-Nikodym derivative exists and is given by

$$\frac{dF}{dF_\theta}(x) = e^{-\theta x}\psi(\theta).$$

Large Deviations

Now let F_θ^n and F^n be the distributions of S_n^θ and S_n, we have

Lemma

$$F^n \ll F_\theta^n \quad \textit{and} \quad dF^n/dF_\theta^n = e^{-\theta x}\psi(\theta)^n.$$

Proof: This lemma follows from induction and the convolution formula. Suppose the result holds for all $k \leq n-1$, now

$$\begin{aligned} F^n(z) = F^{n-1} * F(z) &= \int_{-\infty}^{\infty} dF^{n-1}(x) \int_{-\infty}^{z-x} dF(y) \\ &= \int_{-\infty}^{\infty} e^{-\theta x}\psi(\theta)^{n-1} dF_\theta^{n-1}(x) \int_{-\infty}^{z-x} e^{-\theta y}\psi(\theta) dF_\theta(y) \\ &= \psi(\theta)^n \int_{\mathbb{R}^2} e^{-\theta(x+y)} 1_{x+y\leq z} dF_\theta^{n-1}(x) dF_\theta(y) \\ &= \psi(\theta)^n \int_{-\infty}^{z} e^{-\theta t} dF_\theta^n(t). \end{aligned}$$

Hence we can readily read the R-N derivative from above. □

Intuition behind the tilting

Question: Why do we want to introduce the measure F_θ?

- Intuitively, the new measure is like a "distorting mirror" – it "distorts" our view on how each event is likely to happen.
- So, when we want to estimate a **rare** event A under P, suppose
  - we can construct a new measure Q such that $Q[A]$ is easily calculable, e.g., $Q[A] \approx 1$;
  - we have a uniform lower bound of the R-N derivative $dP/dQ \geq c$ **on** A,

  then we can conclude that $P[A] = \int_A \frac{dP}{dQ} dQ \geq cQ[A]$.
- In addition, if $Q[A] \approx 1$, we can say that A is "realized" when the randomness behind A behaves **as if** it were governed by Q.

Example (A toy example of Boltzmann-Sanov Theorem)

Let X_i i.i.d. with $P(X_1 = 1) = P(X_1 = -1) = 1/2$ and consider $A = \{S_n = n\}$. Then $P^n(A) = 2^{-n}$. Introduce a new measure Q s.t. $Q[X_1 = 1] = 1$. Then $Q^n[A] = 1$ and $dP^n/dQ^n = 2^{-n}$ on A.

Large Deviations

Back to the proof of our theorem, for any $\nu > a$,

$$\begin{aligned} P(S_n \geq na) &\geq \int_{na}^{n\nu} \psi(\theta)^n e^{-\theta t} dF_\theta^n(t) \\ &\geq \psi(\theta)^n e^{-\theta n\nu}[F_\theta^n(n\nu) - F_\theta^n(na)]. \end{aligned} \tag{1}$$

- Recall that X_i^θ has mean $\psi'(\theta)/\psi(\theta)$.
- Since we have picked $\theta > \theta_a$, $a < \psi'(\theta)/\psi(\theta)$.
- Now pick ν such that $\psi'(\theta)/\psi(\theta) < \nu$. By (weak) law of large numbers,

  $$F_\theta^n(n\nu) - F_\theta^n(na) \to 1,$$

  which implies

  $$\liminf_{n\to\infty} n^{-1} \log P(S_n > na) \geq -\theta\nu + \log(\psi(\theta)).$$

Now since we can choose $\psi'(\theta)/\psi(\theta)$ and ν arbitrarily close to a, the proof is complete. □

Large Deviations

Example (Standard Normal Distribution)

$$\psi(\theta) = E[\exp(\theta X)] = \frac{1}{\sqrt{2\pi}} \int e^{\theta x - x^2/2} dx = \exp(\theta^2/2).$$

Thus we have $\psi'(\theta) = \theta \exp(\theta^2/2)$, *and* $\lambda'(\theta) = a - \theta$, *with* $\theta_a = a$, $\gamma(a) = -a^2/2$.

Remark 1: Recall the following calculation: for $x > 0$,

$$(x^{-1} - x^{-3}) \exp(-x^2/2) \leq \int_x^\infty \exp(-y^2/2) dy \leq x^{-1} \exp(-x^2/2).$$

We can give a direct estimate (note that $S_n \sim \mathcal{N}(0, n)$):

$$P(S_n \geq na) \sim (\sqrt{2\pi n}a)^{-1} e^{-na^2/2},$$

which implies

$$n^{-1} \log P(S_n \geq na) \to -a^2/2.$$

Large Deviations

Example (Standard Normal Distribution)

$$\psi(\theta) = E[\exp(\theta X)] = \frac{1}{\sqrt{2\pi}} \int e^{\theta x - x^2/2} dx = \exp(\theta^2/2).$$

Thus we have $\psi'(\theta) = \theta \exp(\theta^2/2)$, *and* $\lambda'(\theta) = a - \theta$, *with* $\theta_a = a$, $\gamma(a) = -a^2/2$.

Remark 2: Intuitively, the large deviation event $\{S_n \geq na\}$ is realized by letting each X_i deviate to (a neighborhood of) a. To have a more precise picture, consider the tilted measure $F_{\theta_a} = F_a$:

$$F_a(x) = (\sqrt{2\pi})^{-1} e^{a^2/2} \int_{-\infty}^{x} e^{ay} e^{-y^2/2} dy.$$

This is **exactly** the distribution of $\mathcal{N}(a, 1)$. In other words, we can **roughly** say that the large deviation event is realized when each X_i is (independently) tilted to $\mathcal{N}(a, 1)$.

Exponential distributions

Example (Exponential distribution with mean 1)

The M.G.F. is finite for all $\theta < 1$*.*

$$\psi(\theta) = E[\exp(\theta X)] = \int e^{(\theta-1)x} dx = \frac{1}{1-\theta}.$$

Hence $\kappa(\theta) = -\log(1-\theta)$, $\kappa'(\theta) = 1/(1-\theta)$*. For* $a > 1$*,* $\theta_a = 1 - a^{-1}$*. And*

$$\gamma(a) = -a\theta_a + \kappa(\theta_a) = -a + 1 + \log(a).$$

In this case,

$$F_{\theta_a}(x) = \frac{1}{1-\theta_a} \int_0^x e^{(\theta_a - 1)y} dy = a \int_0^x e^{y/a} dy.$$

I.e., $F_{\theta_a}(x)$ *is the exponential distribution with mean* a*.*

Remark: What can we say now?

Large Deviations

We know that $k'(\theta) = \psi'(\theta)/\psi(\theta)$ is increasing and $k'(0) = \mu$. We have treated the case where we can indeed find a solution for $\psi'(\theta)/\psi(\theta) = a$. We now treat the cases where we cannot find one.

Theorem (2.7.10 of Durrett)

Suppose $x_o = \sup\{x : F(x) < 1\} = \infty$, $\theta_+ < \infty$, *and* $\psi'(\theta)/\psi(\theta)$ *increases to a finite limit* a_0 *as* $\theta \uparrow \theta_+$. *If* $a_0 \le a < \infty$

$$n^{-1} \log P(S_n \ge na) \to -a\theta_+ + \log \psi(\theta_+)$$

i.e., $\gamma(a)$ *is linear for* $a \ge a_0$.

Remark: We will omit the proof (which can be found on p.110, of Durrett), and only comment that θ_+ is the supremum of θ such that $\psi(\theta) < \infty$, this is as far as we can tilt.

LDP for bounded r.v.'s

There is also one possibility: $x_o = \sup\{x : \; F(x) < 1\} < \infty$.

Theorem

Suppose $x_o = \sup\{x : \; F(x) < 1\} < \infty$ and F has no mass at x_o. Then $\psi(\theta) < \infty$ for all $\theta > 0$ and $\psi'(\theta)/\psi(\theta) \to x_o$ as $\theta \to \infty$.

Proof: Note that $P(X \leq x_o) = 1$. So for any $0 \leq \theta < \infty$,

$$\psi(\theta) = \int_{-\infty}^{x_o} e^{\theta x} dF(x) \leq e^{\theta x_o} < \infty.$$

And for sufficiently large θ, almost all mass of $\psi'(\theta)$ and $\psi(\theta)$ are concentrated around x_o. □

Remark: Using the same argument, when $x_o = \infty$ and $\theta_+ = \infty$, $\psi'(\theta)/\psi(\theta) \to \infty$ as $\theta \to \infty$. This is the case where we do have a solution.

Large Deviations

Finally we deal with the case where F admits some mass at x_o.

Theorem

Suppose $x_o = \sup\{x : \ F(x) < 1\} < \infty$. *Then,*

$$n^{-1} \log P(S_n \geq n x_o) = \log P(X_i = x_o).$$

Now, we have shown the decaying asymptotic for **ALL** possible situations:

- If $x_o < \infty$:
  - When $a < x_o$: exponential, rate=θ_a.
  - When $a = x_o$: exponential if $P(X_1 = x_o) > 0$, 0 otherwise.
  - When $a > x_o$: 0.
- If $x_o = \infty$:
  - If $\theta_+ = \infty$, exponential, rate=θ_a.
  - If $\theta_+ < \infty$
    - If $\psi'(\theta)/\psi(\theta) \to \infty$ as $\theta \to \theta_+$, exponential,rate=θ_a.
    - If $\psi'(\theta)/\psi(\theta) \to a_0$ as $\theta \to \theta_+$
      - When $a < a_0$: exponential, rate=θ_a.
      - When $a \geq a_0$: exponential, rate=θ_+.

Coin flips

We now see an example with $x_o < \infty$.

Example

Consider $P(X_i = 1) = P(X_i = -1) = 1/2$. *Then*

$$\psi(\theta) = E[\exp(\theta X)] = (e^\theta + e^{-\theta})/2, \quad \frac{\psi'(\theta)}{\psi(\theta)} = \frac{e^\theta - e^{-\theta}}{e^\theta + e^{-\theta}}.$$

- *When* $a > 1$, θ_a *does not exist; actually* $P(S_n \geq na) = 0$.
- *When* $a \leq 1$, $\theta_a = \frac{1}{2}\big(\log(a+1) - \log(1-a)\big)$, *and*

$$\gamma(a) = -\frac{(1+a)\log(1+a) + (1-a)\log(1-a)}{2}.$$

Cramér's theorem

Let us restate our theorem in a form which usually appears in other sources. Let $X_1, X_2, \ldots$ be a sequence of i.i.d. r.v.'s, with M.G.F.

$$\psi(\theta) = E[\exp(\theta X_1)].$$

Let $I(a)$ be the **Legendre transform** of $\log \psi(\cdot)$ (sometimes called the **free energy**):

$$I(a) := \sup_{\theta \in \mathbb{R}} \Big(\theta a - \log \psi(\theta) \Big).$$

Theorem (Cramér's theorem, baby form)

$$\lim_{n \geq \infty} \frac{1}{n} \log P_n(\{S_n \geq na\}) = I(a).$$

We say that X_n satisfies a **large deviation principle** with **rate function** I.

Cramér's theorem

We want to go beyond events like $S_n/n \in [a, \infty)$, and claim something like for any Borel A,

$$\lim_{n\to\infty} \frac{1}{n} \log P(S_n/n \in A) = -\inf_{x\in A} I(x).$$

However this is not true:

Example

Let $X_i \in \mathbb{Z}$ be i.i.d., and $A := \{m/p : m \in \mathbb{Z}, p \text{ odd prime}\}$. Then $P(S_n/n \in A) = 1$ for n odd prime and $= 0$ for $n = 2^k$.

Theorem (Cramér's theorem, general form)

With the same notations, one has for any **closed** *F,*

$$\limsup_{n\to\infty} n^{-1} \log P(S_n/n \in F) \leq -\inf_{x\in F} I(x);$$

while for any **open** *G,*

$$\liminf_{n\to\infty} n^{-1} \log P_n(S_n/n \in G) \geq -\inf_{x\in G} I(x).$$

Chernoff bound and Hoeffding's inequality

M.G.F. and so-called "moment method" is very powerful. The following bounds is a very common result for (not i.d.) independent r.v.'s.

Lemma (Chernoff bound)

Let X_i be independent bernoulli r.v's. Write $S_n = X_1 + \cdots + X_n$ and let $\mu = E[S_n]$. Then for $\delta > 0$,

$$P\Big(S_n > (1+\delta)\mu\Big) \le e^{-\frac{\delta^2\mu}{2+\delta}}; \quad P\Big(S_n < (1-\delta)\mu\Big) \le e^{-\frac{\delta^2\mu}{2}}.$$

Proof: Let $p_i = P(X_i = 1)$ and write M_i for the M.G.F. for X_i. Write M for the M.G.F. of S_n. We only work out details for the upper tail. Then

$$M_i(\theta) = 1 + p_i(e^\theta - 1) \le e^{p_i(e^\theta - 1)}, \text{ hence } M(\theta) = \prod_{i=1}^n M_i \le e^{\mu(e^\theta - 1)}.$$

Chernoff bound and Hoeffding's inequality

Hence by Markov inequality, for any $\theta > 0$,

$$P\Big(S_n > (1+\delta)\mu\Big) \le e^{-\theta(1+\delta)\mu} M(\theta) \le e^{\mu(e^\theta - 1 - \theta(1+\delta))}.$$

Optimizing our choice of θ (minimal value of RHS attained at $\theta = \log(1+\delta)$), hence we have

$$\text{RHS} \le e^{-\frac{\delta^2 \mu}{2+\delta}}, \text{ or for } \delta \in (0,1), \text{ RHS} \le e^{-\frac{\delta^2 \mu}{3}}.$$

The lower tail is complete analogous. The optimal θ is at $\log(1-\delta)$. □

Theorem (Hoeffding's inequality for bounded r.v.'s)

Let X_i be independent r.v.'s such that $X_i \in [a_i, b_i]$ a.s. Write $S_n = X_1 + \cdots + X_n$ and let $\mu = E[S_n]$. Then for $\delta > 0$,

$$P\Big(|S_n - \mu| \ge \delta)\Big) \le 2\exp\left(-\frac{2n^2\delta^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Generalization of Hoeffding's inequality

- For the case of martigales, **Azuma-Hoeffding inequality** (not required) holds.
- Alternatively, consider r.v.'s that are not bounded.

A random variable is **sub-Gaussian**, if and only if for some $C < \infty$ and $c > 0$,

$$P(|X| \geq t) \leq Ce^{-ct^2}.$$

Equivalently, one can also define sub-Gaussianity through the finiteness of ψ_2-norm in Birnbaum-Orlicz space:

$$\|X\|_{\psi_2} = \inf\{c \geq 0 : E[e^{X^2/c^2}] \leq 2\}.$$

Theorem (Hoeffding's inequality for sub-Gaussian r.v.'s)

Let X_i be independent zero-mean sub-Gaussian r.v.'s.Write $S_n = X_1 + \cdots + X_n$. Then there exists some $c > 0$ such that for any $\delta > 0$,

$$P\Big(|S_n| \geq \delta)\Big) \leq 2\exp\left(-c\delta^2 \Big/ \textstyle\sum_{i=1}^n \|X\|_{\psi_2}\right).$$

Concentration inequalities

In probability theory, **concentration inequalities** refer to various bounds on how a random variable deviates from some typical value (usually its mean). Examples include:

- Markov inequality and Chebyshev inequality (and "simple" corollaries of these inequalities);
- LLN, CLT, moderate and large deviations;
- Chernoff bound and Hoeffding's inequality;
- Paley-Zygmund inequality: if $Z \geq 0$ is a r.v. with finite variance, then for $0 \leq \theta \leq 1$,
$$P(Z > \theta E[Z]) \geq (1-\theta)^2 E[Z]^2 / E[Z^2].$$
- Efron-Stein inequality: if $X_1, \ldots, X_n$, $X_1', \ldots, X_n'$'s are independent such that $X_i \overset{\mathrm{d}}{=} X_i'$. Let $Y = (X_1, \ldots, X_n)$ and $Y_i = (X_1, \ldots, X_i', \ldots, X_n)$. Then for a nice function f,
$$\mathrm{var}(f(Y)) \leq \tfrac{1}{2} \textstyle\sum_{i=1}^n E[(f(Y) - f(Y_i)^2].$$
- For specific distributions/models/questions, more specific inequalities apply.

Wasserstein metric

Remark: The last statement is not true if we are working in the continuum: in general, convergence in total variation distance is stronger than weak convergence.

- In various applications, we sometimes consider the **Wasserstein distance** between measures, which comes from **transport theory**.
- The p-th Wasserstein distance between two probability measures μ and ν on M with p-th moment is defined as

$$W_p(\mu, \nu) := \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{M \times M} d(x, y)^p d\gamma(x, y) \right)^{1/p},$$

  where $\Gamma(\mu, \nu)$ is the set of all **couplings** of μ and ν.
- When $p = 1$, it is also called the **earth mover's distance**.
- One can show that W_p defines a metric and convergence under W_p-metric is equivalent to weak convergence plus convergence of the first p-th moments.

Poisson Process

We can also define the process in a more abstract way.

Let $\Omega = (\mathbb{Z}^+ \cup \{0\})^{[0,\infty)} = \{\omega : [0,\infty) \to \mathbb{Z}^+ \cup \{0\}\}$, and $\mathcal{F}_o$ be the σ-algebra generated by all finite-dimensional columns, i.e. events in the following form:

$$\{N_{t_i} \in S_i,\ i = 1,\dots,k\} 0 \le t_0 < t_1 < t_2 < \cdots < t_k,$$

with $S_i \subseteq \mathbb{Z}^+ \cup \{0\}$, $i = 1,\dots,k$.

We can always check the **consistency** of the distributions and use a generalized form of Kolgomorov's extension theorem to construct a probability measure on the space above. But...

Question: Are the following sets

$$\{\omega :\ \text{continuous}\}, \quad \{\omega :\ \text{monotone}\}$$

measurable with respect to $\mathcal{F}_o$? (Exercise 7.1.4)

Poisson Process

The potential problem of using extension theorems directly is we may need to introduce uncountable unions of events to discuss properties w.r.t. the trajectory.

Solution:

- Construct the process on $\mathbb{Q}$, and let

$$N(t) = \lim_{s \to t^+,\ s \in \mathbb{Q}} N(s).$$

  See 7.1 for a similar definition for Brownian Motion.

- Construction through Markov semi-group. We consider a family of operators $S^\lambda(\cdot)$ from bounded maps $\mathbb{Z}^+ \to \mathbb{R}$ to itself:

$$S^\lambda(t) f(\cdot) := E[f(\cdot + N_t)], \text{ where } N_t \sim \text{Poisson}(\lambda t).$$

Poisson Process

One can see that

$$S^{\lambda}(t)f(n) = \sum_{k=0}^{\infty} \frac{(\lambda t)^k e^{-\lambda t}}{k!} f(n+k),$$

and has the following properties:

1. $S^{\lambda}(0) = \mathrm{Id}$;
2. $t \to S^{\lambda}(t)f(n)$ is right continuous for all bounded f;
3. $S^{\lambda}(t+s) = S^{\lambda}(t)S^{\lambda}(s)$ (Markov property);
4. $S^{\lambda}(t)1 = 1$;
5. If $f \geq 0$, $S^{\lambda}(t)f \geq 0$.

Poisson Process

Definition

A family of operators $S(t),\ t \geq 0$ *satisfying* (I) − (V) *is call a* **Markov Semigroup**.

Theorem (Blumenthal and Getoor 1968, Gihman and Skorohod 1975, Liggett, 1985)

If $S(t)$ *is a Markov semigroup. Then there exists a unique Markov process* $\eta_{t\geq 0}$ *governed by* E^η *s.t.*

$$S(t)f(\eta_0) = E^\eta[f(\eta_t)].$$

Remark: Here the Markov process is defined on the path space $D[0, \infty)$, the family of all functions that are right continuous and have left limits (abbreviated as RCLL or càdlàg (=*continue à droite, limite à gauche* in French) functions).

Poisson Thinning

One can also generalize the thinning arguments to define inhomogenous Poisson process with **finite** transition rate function:

Theorem

Suppose that in a Poisson process with rate λ, for a point that lands at time s, we keep it with probability $p(s)$. Then the result is an **inhomogenous Poisson process** *with rate $\lambda p(s)$.*

Definition (Inhomogenous Poisson process as time change of Poisson process)

For $p(t)$, and the standard Poisson process N_t with rate λ, we call

$$\hat{N}(t) = N\left(\int_0^t \lambda p(s) ds\right)$$

be the inhomogenous Poisson process with transition rate function $\lambda(t) = \lambda p(t)$.

Poisson Point Process

Question: Are there any other way(s) to define/ construct/ simulate a nonhomogeneous Poisson process?

- Let $(S, \mathcal{S}, \lambda)$ be a measure space. A **Poisson Point Process** (abbr. as PPP) is a **random mapping** from $\mathcal{S}$ to $\{0, 1, 2, \cdots, \infty\}$.
- Though not mentioned in this book, such mappings are usually required to additive.
- In this case, ω can be written as a finite or countably infinite point measure on S,

$$\omega = \sum_{i=1}^{\infty} \delta_{s_i}.$$

  Here we use the convention $\delta_{\emptyset} \equiv 0$.
- Let $\Omega(S)$ be the space of all such point measures, and $\mathcal{A}(S)$ be the sigma algebra on $\Omega(S)$ generated by $\omega(A),\ A \in \mathcal{S}$.

Poisson Point Process

Now we construct a random mapping which is actually governed by a probability measure on $(\Omega(S), \mathcal{A}(S))$ associated with λ.

Definition

Suppose λ is sigma-finite, we say a random measure μ is a Poisson Point Process/ Poisson random measure with **intensity measure** λ *if*

- *For all $B \in \mathcal{S}$, $\mu(B) \sim Poisson(\lambda(B))$.*
- *If $B_1, \cdots, B_n$ are disjoint sets in S, then the random variables $\mu(B_1), \cdots, \mu(B_n)$ are also independent.*

Theorem

Such random measure μ exists and is unique.

Poisson Point Process

Ideas of proof: Uniqueness: $\pi - \lambda$ argument.

Existence: Suppose λ is a finite measure. Then $\nu = \lambda(\cdot)/\lambda(S)$ is a probability measure. Let $X_1, X_2, \cdots$ be i.i.d. with distribution ν, and $N \sim \text{Poisson}(\lambda(S))$. Then $\mu := \sum_{i=1}^{N} \delta_{X_i}$ is what we are looking for.

Suppose λ is an infinite but sigma-finite measure. We can repeat the construction above on its partition. □

Remark 1: Suppose λ is an infinite but sigma-finite measure. The random measure μ is with probability one an infinite point measure.

Remark 2: The regular Poisson process is a PPP on $[0, \infty)$ with intensity measure $\lambda *$Lebesgue measure.

Remark 3: The inhomogeneous Poisson process is a PPP on $[0, \infty)$ where the Radon-Nikodym derivative of the intensity measure with respect to Lebesgue measure is $\lambda p(t)$.

Central Limit Theorem: Berry-Esseen Bounds

Theorem (Berry-Esseen Bounds)

Let $X_1, X_2, \cdots$ be i.i.d. with $E[X_i] = 0$, $E[X_i^2] = \sigma^2$, and $E[|X_i|^3] = \rho < \infty$. Let $\mathcal{N}(x)$ is the distribution of the standard normal distribution, then for all $n \geq 1$ and $x \in \mathbb{R}$

$$|F_n(x) - \mathcal{N}(x)| \leq 3\rho/\sigma^3\sqrt{n}.$$

Remark: The third moment bound ($E[|X_i|^3] = \rho < \infty$) is necessary for the $O(n^{-1/2})$ error.)

Recall **Pólya's distribution**:

- Density $h_L(x) = (1 - \cos(Lx))/(\pi L x^2)$
- Ch.f. $\varphi_L(t) = (1 - |t/L|)^+$.

We will write H_L for its distribution function. Note that $\varphi_L(t)$ has compact support.

Central Limit Theorem: Berry-Esseen Bounds

Proof Idea: We will convolve F_n and $\mathcal{N}(x)$ with Pólya's distribution respectively, and

1. show that the difference between the distributions are bounded by the difference of the convolved distributions;
2. observe that the ch.f. of the convolved distributions also has compact support;
3. bound the difference between the convolved distributions.

Remark: In retrospect, looking at the calculation above from the distributional side, we can say that vaguely speaking, convolution with a smooth kernel improves regularity. While from the ch.f. side, it is simply a story of truncation.

LCLT for lattice r.v.'s

Theorem

Let X_i be i.i.d. r.v.'s with $E[X_i] = 0$, $E[X_i^2] = \sigma^2 \in (0, \infty)$. Suppose in addition $P(X_i \in b + h\mathbb{Z}) = 1$, i.e. X_i are lattice with span h. Let

$$p_n(x) = P(S_n/\sqrt{n} = x) \text{ for } x \in \mathcal{L}_n = \{(nb + h\mathbb{Z})/\sqrt{n}\},$$

and $n(x)$ be the density of $\mathcal{N}(0, \sigma^2)$. Then

$$\lim_{n\to\infty} \sup_{x\in\mathcal{L}_n} \left| \frac{\sqrt{n}}{h} p_n(x) - n(x) \right| = 0.$$

Remark: Actually, we can even have explicit bounds on the convergence rate. E.g., if $|x| = O(\sqrt{n})$, then we have

$$p_n(x) = n^{-1/2} h n(x)\big(1 + O(n^{-1})\big).$$

For more details, see Chap. 2 of "Random walk, a modern introduction" by Lawler and Limic.

Ch.f. for lattice r.v.'s, an example

Theorem

Let $p_n^{(d)}(\cdot)$ stand for the n-step transition probability for d-dimensional simple random walk. Prove that $p_{2n}^{(d)}(0)$ is monotone decreasing in d.

Proof: Let S_n be the location of a d-dimensional SRW started from 0 at time n. Then

$$\varphi_{S_{2n}}(t_1, \ldots, t_d) = d^{-2n} \left(\sum_{i=1}^{d} \cos t_i \right)^{2n} .$$

Hence

$$p_{2n}^{(d)}(0) = (2\pi)^{-d} \int_{[-\pi,\pi]^d} \left(\frac{1}{d} \sum_{i=1}^{d} \cos t_i \right)^{2n} dt_1 dt_2 \cdots dt_n .$$

Ch.f. for lattice r.v.'s, an example

By Jensen's inequality and the convexity of $f(x) = x^{2n}$,

$$\left(\frac{1}{d} \sum_{i=1}^{d} \cos t_i \right)^{2n} \le \frac{1}{d} \sum_{k=1}^{d} \left(\frac{1}{d-1} \sum_{i=1,\, i \neq k}^{d} \cos t_i \right)^{2n} .$$

The claim follows from the the observation that $\forall 1 \le k \le d$,

$$p_{2n}^{(d-1)}(0) = \frac{1}{(2\pi)^{d-1}} \int_{[-\pi,\pi]^{d-1}} \left(\frac{1}{d-1} \sum_{i=1,\, i \neq k}^{d} \cos t_i \right)^{2n} dt_1 \cdot\cdot \widehat{dt_k} \cdot\cdot dt_n,$$

where $\widehat{dt_k}$ means dt_k does not appear in the integral. $\square$

LCLT for nonlattice r.v.'s

Remark: For nonlattice r.v.'s, one does not have a good control over its ch.f. in general. E.g., let X be a nonlattice r.v. s.t.

$$P(X = 1/n) = P(X = -1/n) = 2^{-n-1}.$$

Then, letting $t_n = \mathrm{lcm}(1, \ldots, n)$, one can see that

$$\varphi_X(t_n) = \sum_{j=1}^{\infty} 2^{-j} \cos(2\pi t_n / j) \geq 1 - 2^{-n+1}.$$

In other words, $\limsup_{t \to \infty} |\varphi_X(t)| = 1$.

Stein's Lemma, full version

Lemma

If $Z \sim \mathcal{N}(\mu, \sigma^2)$, then for all absolutely continuous functions f with $E[f'(Z)] < \infty$,

$$E[(Z - \mu) f(Z)] = \sigma^2 E[f'(Z)].$$

If the identity above holds for all absolutely continuous functions s.t. $\|f'\| < \infty$, then $Z \sim \mathcal{N}(\mu, \sigma^2)$.

Remark: Let's think one step forward. For a random variable W with $E[W] = 0$, $\mathrm{var}(W) = 1$, if $E[f'(W)] - E[Wf(W)]$ is close to zero for "many" functions f, then W should be close to Z in distribution.

Proof: The first claim follows from partial integration.

Proof of Stein's Lemma

For the second claim, WLOG assume $E[Z] = 0$, $\text{var}(Z) = 1$. Observe that solution of the following differential equation

$$f'(x) - xf(x) = g(x)$$

for some f such that $\lim_{x \to -\infty} f(x)e^{-x^2/2} = 0$ is given by

$$f(y) = e^{y^2/2} \int_{-\infty}^{y} g(x)e^{-x^2} dx.$$

Take $g(x) = g_{x_0}(x) = 1_{x \le x_0} - \Phi(x_0)$, then

$$0 = E[f'_{x_0}(Z) - Zf_{x_0}(Z)] = P(Z \le x_0) - \Phi(x_0).$$

So $Z \sim \mathcal{N}(0, 1)$. □

Remark: The form on the RHS of the last identity inspires us to realize that this method can yield Berry-Esseen type bounds.

The crux of Stein's method

- In other words, the stein equation can give some information on how a random variable deviates from standard normal. More precisely, recall for $g_{x_0}(x) = 1_{x \le x_0} - \Phi(x_0)$, and for random variables $X \sim P$, $Z \sim \mathcal{N}(0,1)$,
$$|P(X \le x_0) - P(Z \le x_0)| = |E[f'_{x_0}(X) - X f_{x_0}(X)]|.$$
- Now consider some metric between the distributions of X and Z defined through
$$d_{\mathcal{H}}(X,Z) := \sup_{h \in \mathcal{H}} |E[h(X)] - E[h(Z)]|,$$
where $\mathcal{H}$ is a certain collection of test functions. For any $h \in \mathcal{H}$, let f_h solve
$$f'_h(w) - w f_h(w) = h(w) - E[h(Z)].$$
Then
$$d_{\mathcal{H}}(X,Z) = \sup_{h \in \mathcal{H}} |E[f'_h(X) - X f_h(X)]|.$$

More discussions on Wasserstein metric

Recall the Wasserstein-1 distance (earth mover's distance) on $\mathbb{R}$:

- For two random variables $X \sim \mu$ and $Y \sim \nu$, with finite first moment:

$$W_1(\mu, \nu) := \inf E[|X - Y|],$$

where inf runs over all **couplings** of X and Y.

- Equivalently, let $\mathcal{L}$ be the set of Lipschitz-continuous functions with Lipschitz constant 1:

$$\mathcal{L} := \{g : \mathbb{R} \to \mathbb{R}; |g(y) - g(x)| \leq |y - x|\},$$

then

$$W_1(\mu, \nu) = \sup_{g \in \mathcal{L}} |E[g(Y)] - E[g(X)]|.$$

- **Question**: Why are the two definitions equivalent?
- **Hint**: In the special case of $d = 1$,

$$W_1(\mu, \nu) = \int_0^1 \left| F^{-1}(z) - G^{-1}(z) \right| dz.$$

Sum of independent r.v.'s

Theorem

Let X_i be independent r.v.'s with $E[X_i] = 0$, $var(X_i) = 1$ and $E[|X_i^4|] < \infty$. Let $W = S_n/\sqrt{n}$ and $Z \sim \mathcal{N}(0,1)$. Then

$$W_1(W,Z) \leq \frac{1}{n^{3/2}} \sum_{i=1}^n E[|X_i|^3] + \frac{C}{n} \sqrt{\sum_{i=1}^n E[|X_i|^4]}.$$

Remark: In the i.i.d. case, one has

$$W_1(W,Z) \leq \frac{1}{n^{1/2}} \left(E[|X_i|^3] + C\sqrt{E[|X_i|^4]} \right).$$

Proof: By the discussion above, we just need to bound $\sup_{h \in \mathcal{L}} |E[f_h'(X) - X f_h(X)]|$. For a test function $f = f_h$, we claim that

$$E[f'(W) - W f(W)] \leq \frac{\|f''\|}{2n^{3/2}} \sum_{i=1}^n E[|X_i|^3] + \frac{\|f'\|}{n} \sqrt{\sum_{i=1}^n E[|X_i|^4]}.$$

Proof of CLT via Stein's method

To see the claim

$$E[f'(W) - Wf(W)] \leq \frac{\|f''\|}{2n^{3/2}} \sum_{i=1}^{n} E[|X_i|^3] + \frac{\|f'\|}{n} \sqrt{\sum_{i=1}^{n} E[|X_i|^4]}.$$

implies the theorem, note that the explicit solution

$$f_h(y) = e^{y^2/2} \int_{-\infty}^{y} \big(h(x) - E[h(Z)]\big) e^{-x^2/2} dx$$

gives us the following bounds (after some calculation):

$$\|f_h\| \leq 2\|h\|, \quad \|f_h'\| \leq \sqrt{2/\pi}\|h'\|, \quad \text{and } \|f_h''\| \leq 2\|h'\|,$$

and that $h \in \mathcal{L}$ implies $\|h'\| \leq 1$.

Proof of CLT via Stein's method

To prove the claim, let $W_i = W - X_i/\sqrt{n}$. One has

$$E[Wf(W)] = E\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\big(X_i f(W) - X_i f(W_i)\big)\right].$$

Hence,

$$\begin{aligned} E[Wf(W)] &= E\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i\big(f(W) - f(W_i) - (W - W_i)f'(W)\big)\right] \\ &\quad + E\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i(W - W_i)f'(W)\right] \\ &= \mathrm{I} + \mathrm{II}. \end{aligned}$$

So

$$|E[f'(W) - Wf(W)]| \le |\mathrm{I}| + \left|E\left[f'(W)\left(1 - \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i(W - W_i)\right)\right]\right|.$$

Denouement

On the one hand, by Taylor expansion, we have

$$\mathrm{I} \leq \frac{\|f''\|}{2n^{1/2}} \sum_{i=1}^{n} E[|X_i(W - W_i)^2] = \frac{\|f''\|}{2n^{3/2}} \sum_{i=1}^{n} E[|X_i|^3].$$

On the other hand, to bound

$$\mathrm{II}' := \left| E\left[f'(W) \left(1 - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i(W - W_i) \right) \right] \right|,$$

note that

$$\mathrm{II}' \leq \frac{\|f'\|}{n} E\left[\left| \sum_{i=1}^{n} (1 - X_i^2) \right| \right] \leq \frac{\|f'\|}{n} \sqrt{\operatorname{var}\left(\sum_{i=1}^{n} X_i^2 \right)},$$

where in the last step we have used Cauchy-Schwarz. Finally, note that the RHS$\leq \frac{\|f'\|}{n} \sqrt{\sum_{i=1}^{n} E[|X_i|^4]}$. □

Stein's Method for Poisson r.v.'s

- **Fact**: $Z \sim \text{Poisson}(\lambda) \Leftrightarrow$ for all $f : \mathbb{Z}^+ \to \mathbb{R}$ with $E[f(Z)] = 0$, one has

$$E[\lambda f(Z+1) - Zf(Z)] = 0.$$

- Hence for test indicator functions $I_A(\cdot) = 1_{\cdot \in A}$ for $A \in \mathbb{Z}^+$, consider the Stein equation

$$\lambda g(j+1) - jg(j) = I_A(j) - P(Z \in A).$$

- For some r.v. W we hence have

$$\lambda g(W) - Wg(W) = I_A(W) - P(Z \in A).$$

Taking expectation, one has

$$d_{\text{TV}}(W, Z) \leq \sup_g |E[\lambda g(W+1) - Wg(W)]|,$$

where sup is over all functions g which solve the Stein equation.

Summary: general form of Stein's method

- Start with a target distribution μ.
- Find characterization: some operator $\mathcal{A}$ such that $X \sim \mu$ if and only if for all test functions f,
$$E[\mathcal{A}f(X)] = 0.$$
- For each function h find solution $f = f_h$ of the Stein equation:
$$h(x) - \int h d\mu = \mathcal{A}f(x).$$
- Then for any variable W,
$$E[h(W)] - \int h d\mu = E[\mathcal{A}f(W)].$$
- Finally, according to the type of convergence we want, give bounds on the test function f (or its derivatives).

Review: Characteristic functions and moments

Theorem

If $\int |x|^n \mu(dx) < \infty$, then its ch.f. φ has a continuous derivative of order n given by

$$\varphi^{(n)}(t) = \int (ix)^n e^{itx} \mu(dx).$$

In particular,

$$\varphi^{(n)}(0) = E[(iX)^n].$$

Lemma (See Lemma 3.3.19, p. 134 of Durrett)

$$\left| e^{ix} - \sum_{m=0}^{n} \frac{(ix)^m}{m!} \right| \leq \min \left(\frac{|x|^{n+1}}{(n+1)!}, \ \frac{2|x|^n}{n!} \right).$$

Review: Characteristic functions and moments

When $E[|X|^n] < \infty$, take expectation, and use Jensen's ineq.,

$$\left|\varphi(t) - \sum_{m=0}^{n} \frac{E[(iX)^m]t^m}{m!}\right| \le E\left|e^{itx} - \sum_{m=0}^{n} \frac{(itX)^m}{m!}\right|$$
$$\le CE\left[\min(|tX|^{n+1},\ 2|tX|^n)\right].$$

In particular, we can conclude with the following theorem:

Theorem

If $E|X|^2 < \infty$ *then*

$$\varphi(t) = 1 + itE[X] - t^2E[X^2]/2 + o(t^2).$$

Theorem

If

$$\limsup_{h\downarrow 0} \frac{\varphi(h) - 2\varphi(0) + \varphi(-h)}{h^2} > -\infty,$$

then $E[X^2] < \infty$.

Caveat

Remark: Similar results does not hold for odd order derivatives:

- Recall that if a random variable X has moments up to k-th order, then the characteristic function is k times continuously differentiable on the entire real line.
- However, if a characteristic function φ_X has a k-th derivative at zero, then the random variable X has all moments up to k if k is even, but only up to $k-1$ if k is odd.

The moment problem (not required)

Consider a sequence of distributions F_n. Suppose there exist $(\mu_k)_{k\in\mathbb{Z}^+}$ s.t.

$$\int x^k dF_n(x) \to \mu_k, \quad \forall k \in \mathbb{Z}^+. \tag{2}$$

The sequence of distributions is tight and hence every subsequential limit has the same moments.

- It is easy to see that this is true if F is concentrated on a finite interval $[-M, M]$ since every continuous function can be approximated uniformly on $[-M, M]$ by polynomials.
- However, the result is false in general.

Theorem (not required)

If $\limsup_{k\to\infty} \mu_{2k}^{1/2k}/2k < \infty$, *then there is at most one distribution* F *with moments* μ_k. *Suppose in addition there exist distributions* F_n *satisfying* (2), *then* $F_n \Rightarrow F$.

Example: Random matrices and semicircle law

Example

Let $Z_{i,j}$, $1 \leq i < j$ be i.i.d. $\mathcal{N}(0,1)$ and Y_i, $1 \leq i$ be i.i.d. $\mathcal{N}(0,2)$, independent of each other. Consider the (symmetric) $N \times N$ matrix X_N with entries

$$X_N(i,j) = X_N(j,i) = \begin{cases} Z_{i,j}/\sqrt{N} & \text{if } i < j; \\ Y_i/\sqrt{N} & \text{if } i = j. \end{cases}$$

These matrices are referred to as **Gaussian orthogonal ensemble** *(GOE) matrices.*

Let λ_i^N denote the order statistics of the (necessarily real) eigenvalues of X_N , with $\lambda_1^N \leq \lambda_2^N \leq \cdots \leq \lambda_N^N$, define the **empirical distribution** of the eigenvalues as the (random) probability measure on $\mathbb{R}$ defined by

$$L_N = \frac{1}{N} \sum_{i=1}^{N} \delta_{\lambda_i^N}.$$

Semicircle law

Let $\sigma(x) = \frac{1}{2\pi}\sqrt{4-x^2}1_{|x|\leq 2}$ be the density of the **semicircle distribution**.

Theorem

The empirical distribution of eigenvalues L_N converges weakly a.s. to the semicircle distribution.

Remark: Note that here L_N is a **random measure**; i.e., a r.v. in $M_1(\mathbb{R})$, the space of probability measures on the real line, and the topology (or notion of continuity) is provided by weak convergence.

- $L_N \stackrel{\mathrm{P}}{\to} \sigma \in M_1(\mathbb{R})$ if, for every test function $f \in C_b(\mathbb{R})$,
$$\int f dL_N \stackrel{\mathrm{P}}{\to} \int f d\sigma;$$
- $L_N \stackrel{\text{a.s.}}{\to} \sigma$ if, for every $f \in C_b(\mathbb{R})$,
$$\int f dL_N \stackrel{\text{a.s.}}{\to} \int f d\sigma.$$

Semicircle law

Let $\sigma(x) = \frac{1}{2\pi}\sqrt{4-x^2}1_{|x|\leq 2}$ be the density of the **semicircle distribution**.

Theorem

$$L_N \overset{\text{a.s.}}{\to} \sigma.$$

Remark 1: This result is quite "universal" in the sense that it holds for many models of random matrices under various setups.

Remark 2: This theorem has a number of proofs. We are only going to see one of them, using the moment method.

Theorem (Eigenvalue interlacing inequality)

For any $n \times n$ Hermitian matrix A_n, let A_{n-1} be its top-left $(n-1)\times(n-1)$ minor. Then, one has

$$\lambda_i(A_n) \leq \lambda_i(A_{n-1}) \leq \lambda_{i+1}(A_n)$$

for all $1 \leq i < n$.

Moment calculation

Define the moments $m_k := \langle \sigma, x^k \rangle$ for the semicircle distribution. One can check

$$m_{2k} = C_k, \quad m_{2k+1} = 0,$$

where $C_k = (2k)!/(k+1)!k!$ is the **Catalan number**.

Remark: As $\limsup_{k\to\infty} m_{2k}^{1/2k}/2k < \infty$, we see that σ is completely characterized by its moments.

Note that if $U_N \sim L_N$, then

$$E[U_N^k] = \frac{1}{N}\sum_{i=1}^{N}(\lambda_i^N)^k = \frac{1}{N}E[\mathrm{tr}(X_N^k)].$$

Note that for k odd, $E[U_N^k] = 0$. By some calculation from combinatorics,

$$E[\mathrm{tr}(X_N^{2k})] = C_k n(n-1)\cdots(n-k)/n^k = \big(C_k + o_k(1)\big)n.$$

Discussion

- The calculations from the last slide already shows weak convergence.
- To upgrade this convergence to a.s. convergence, recall the eigenvalue interlacing inequality, it suffices to show it for any sub-sequence n_j s.t. $n_{j+1}/n_j \geq c > 1$ for any j.
- Finally, it suffices to obtain some second-moment bound, e.g.

$$\mathrm{var}(U_N^k) = O_k(N^{-ck})$$

  for some $c_k > 0$. This bound again comes from (very careful) combinatorial calculations.
- However, if we only want to show $L_N \xrightarrow{\text{P}} \sigma$, it suffices to prove the following: for all $k \in \mathbb{N}$ and $\epsilon > 0$,

$$\lim_{N\to\infty} P\left(\left|\langle L_N, x^k\rangle - E[U_N^k]\right| > \epsilon\right) = 0.$$

- In the process of proving this, one still need to show $\mathrm{var}(U_N^k) = o_k(1)$, which is slightly easier.

References

- An Introduction to Random Matrices by Anderson, Guionnet and Zeitouni.
- Topics in Random Matrix Theory by Tao.
- And many more...