

LEC018 Review of Probability II

VG441 SS2020

Cong Shi
Industrial & Operations Engineering
University of Michigan

之后会讲

online learning.

- WLLN

Let X_1, \dots, X_n be i.i.d. having $\mathbb{E}[X] = \mu$ and variance σ^2 , then

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{and} \quad \bar{X}_n - \mu \xrightarrow{i.p.} 0 \quad \text{as } n \rightarrow \infty.$$

- CLT

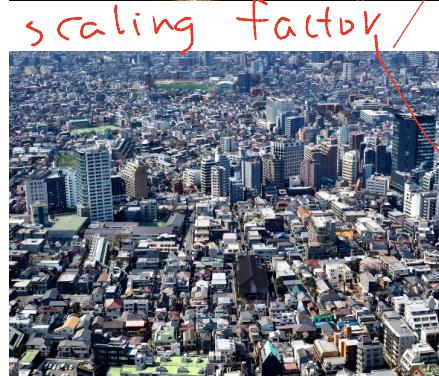
Let X_1, \dots, X_n be i.i.d. having $\mathbb{E}[X_1] = \mu$ and variance σ^2 , then

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{and} \quad \sqrt{n} (\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2) \quad \text{as } n \rightarrow \infty.$$



Limit Theorems

- Statistical snapshot at different levels



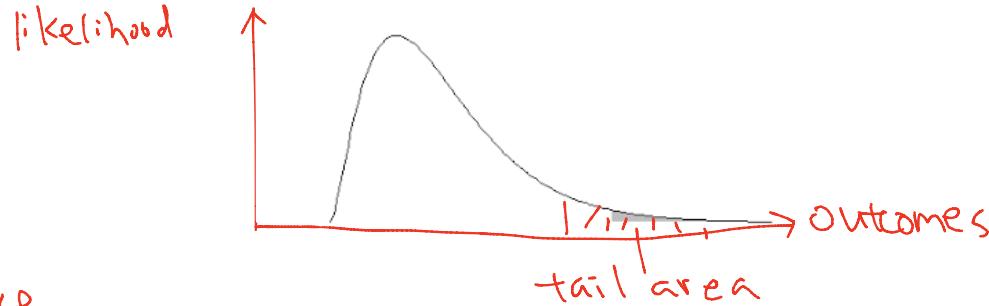
$$\frac{\sum}{n} \rightarrow 0$$

$$\frac{\sum}{\sqrt{n}} \rightarrow N(0, \sigma^2)$$

$$\frac{\sum}{1}$$

→ noisy

Tail Approximations



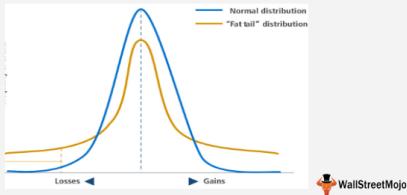
为了避免
worst case.



Disney land
queue

Tail Risk

Tail Risk relates to the risk of a portfolio moving its normal deviation from the current price three times.



finance
pricing

risk management



rare event / SCM disruption

Tail Approximations

SLLN + CLT tells us $S_n \approx n\mu + \sqrt{n}\sigma N(0, 1)$. So CLT handles deviation of size \sqrt{n} : $\mathbb{P}(S_n > n\mu + \delta\sqrt{n}) \approx \mathbb{P}(Z > \delta/\sigma)$ for moderate or large n .

$$Z = N(0, 1)$$

A simple example:

Consider i.i.d. r.v.'s X_1, \dots, X_{16} where $X_i \sim U[0, 1]$ for all $i = 1, \dots, 16$. We want to bound

$$\mathbb{P}\left(\sum_{i=1}^{16} X_i \geq 10\right). \quad \text{Uniformly}$$

Let $S_{16} = \sum_{i=1}^{100} X_i$ and then $\mathbb{E}(S_{16}) = 16\mathbb{E}(X_1) = 8$. The true distribution is the Irwin-Hall distribution with n .



- Using Markov's inequality,

$$\frac{\mathbb{E}[S_{16}]}{16}$$

$$\mathbb{P}(S_{16} \geq 10) \leq \frac{\mathbb{E}(S_{16}/10)}{10} = 8/10 = 0.80. \quad \text{太弱了.}$$

也可以用蒙特卡洛
simulation
几百万次

- Using Chebyshev's inequality,

$$\because \text{Var}(U[0,1]) = 1/12$$

$$\mathbb{P}(S_{16} \geq 10) = \frac{1}{2} \mathbb{P}(|S_{16} - 8| \geq 2) \leq \frac{1}{2} \left(\frac{1}{2^2}\right) \sigma_{S_{16}}^2 = \frac{1}{8} \left(\frac{16}{12}\right) = 0.17.$$

兩端的↑

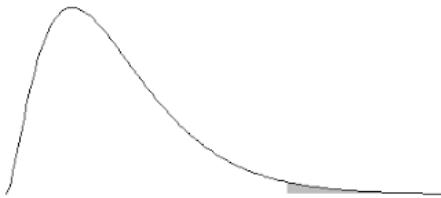
- Using CLT, we have

$$\mathbb{P}(S_{16} \geq 10) = \mathbb{P}(S_{16} \geq 8 + 0.5(4)) \approx \mathbb{P}\left(Z \geq \frac{0.5}{\sqrt{1/12}}\right) = 1 - \Phi(1.732) = 0.042. \quad \leftarrow \text{这个取值}$$

CDF table

or 计算器

Big Question



How many samples/experiments n do you need for the performance to be robust (having a tail that is small, preferably exponentially smaller as n increases)?

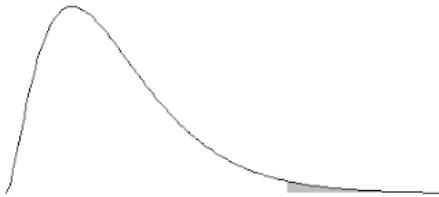
Goal:

exponential decay

on tail in the

amount of samples.

Big Question



How many samples/experiments n do you need for the performance to be robust (having a tail that is small, preferably exponentially smaller as n increases)?

CLT “roughly” handles deviation of size \sqrt{n} :
It takes time to bb approximation

$$\mathbb{P}(S_n > n\mu + \delta\sqrt{n}) \approx \mathbb{P}(Z > \delta/\sigma) \quad \text{for moderate or large } n.$$

Large deviation “exactly” handles deviation of size n :

$$\mathbb{P}(S_n > n\mu + n\delta).$$

CLT is insufficient to bound the above in a sense that

$$\mathbb{P}\left(S_n > n(\underbrace{\mu + \delta}_a)\right) = \mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} > \underbrace{\frac{\sqrt{n}(a - \mu)}{\sigma}}_Z\right), \quad \text{but} \quad \sqrt{n}\left(\frac{a - \mu}{\sigma}\right) \rightarrow \infty!$$

blow up

P(Z > ∞) is meaningless 7

Concentration inequalities

major branch
in statistic / ML.

Consider an i.i.d. sequence X_1, X_2, \dots . Fix a value $a > \mu$ and fix a positive parameter $\theta > 0$. We have

$$\begin{aligned} \mathbb{P}\left(\frac{\sum_{1 \leq i \leq n} X_i}{n} > a\right) &= \mathbb{P}\left(\sum_{1 \leq i \leq n} X_i > na\right) = \mathbb{P}\left(e^{\theta \sum_{1 \leq i \leq n} X_i} > e^{\theta na}\right) \\ &\stackrel{\text{exponentiate.}}{\rightarrow} \mathbb{E}[e^{\theta \sum_{1 \leq i \leq n} X_i}] = \frac{\mathbb{E}[e^{\theta X_1} \dots e^{\theta X_n}]}{(e^{\theta a})^n} = \frac{[\mathbb{M}(\theta)]^n}{(e^{\theta a})^n} \end{aligned}$$

[由式子的利用]
moment

empirical average.
用Markov

But recall that X_i 's are i.i.d. Therefore $\mathbb{E}[e^{\theta X_1} \dots e^{\theta X_n}] = (\mathbb{E}[e^{\theta X_1}])^n$. Thus, we obtain an upper bound

$$\mathbb{P}\left(\frac{\sum_{1 \leq i \leq n} X_i}{n} > a\right) \leq \left(\frac{\mathbb{E}[e^{\theta X_1}]}{e^{\theta a}}\right)^n.$$

(1) $\mathbb{M}(\theta)$
moment generating function of X_1

First assume for a moment that $\mathbb{E}(\theta X_1)$ is finite for all θ in some interval $[0, \theta_0]$.

Note that when $\theta = 0$ the ratio $\frac{\mathbb{E}[e^{\theta X_1}]}{e^{\theta a}} = 1$. Now differentiate this ratio with respect to θ at $\theta = 0$:

$$\frac{d}{d\theta} \frac{\mathbb{E}[e^{\theta X_1}]}{e^{\theta a}} \Big|_{\theta=0} = \frac{\mathbb{E}[X_1 e^{\theta X_1}] e^{\theta a} - a e^{\theta a} \mathbb{E}[e^{\theta X_1}]}{e^{2\theta a}} \Big|_{\theta=0} = \mathbb{E}[X_1] - a = \mu - a < 0.$$

Therefore, for sufficiently small θ the ratio $\frac{\mathbb{E}[e^{\theta X_1}]}{e^{\theta a}}$ is less than unity!

Concentration inequalities

Given an i.i.d. sequence X_1, \dots, X_n suppose $\mathbb{E}[e^{\theta X_1}]$ is finite for all θ in some interval $[0, \theta_0)$. Let $a > \mu = \mathbb{E}[X_1]$. Then for some sufficiently small $\theta > 0$ there holds $\frac{\mathbb{E}[e^{\theta X_1}]}{e^{\theta a}} < 1$ and, moreover,

$$\mathbb{P}\left(\frac{\sum_{1 \leq i \leq n} X_i}{n} > a\right) \leq \left(\frac{\mathbb{E}[e^{\theta X_1}]}{e^{\theta a}}\right)^n.$$

In other words, the large deviation probability is exponentially small.

我们可以 choose θ to minimize α

One degree of freedom (that we can leverage)...

How small can we make this ratio? We have some freedom in choosing θ as long as $\mathbb{E}[e^{\theta X_1}]$ is finite. So we could try to find θ which minimizes the ratio $\frac{\mathbb{E}[e^{\theta X_1}]}{e^{\theta a}}$. This is what we will do. The surprising conclusion of the large deviations theory is that such a minimizing value θ^* exists and is tight. Namely it provides the *correct decay rate*!

$$\begin{aligned} P\left(\frac{\sum_{n>a} X_i}{n} \geq a\right) &\leq \left(\frac{\mathbb{E}[e^{\theta X_1}]}{e^{\theta a}}\right)^n = e^{(\log \mathbb{E}[e^{\theta X_1}]) - \theta a)n} \\ &= e^{-n(\theta a - \log \mathbb{E}[e^{\theta X_1}])} \end{aligned}$$

factor (要 maximize)

Chernoff Bound

Arguably the most useful inequality in probability theory:

A Legendre transform of a r.v. X is the function $l(a) \triangleq \sup_{\theta}(\theta a - \log M(\theta))$.

We have established an upper bound on the probability of large deviations

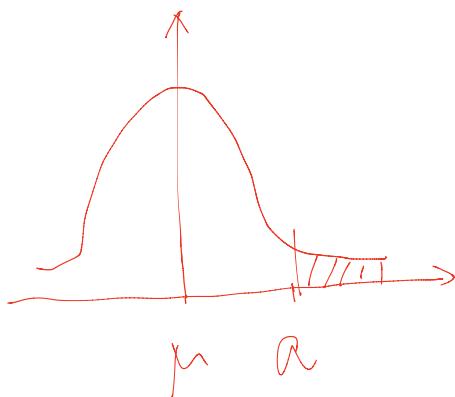
$$\text{if } a > E[X_1] = \mu \quad \mathbb{P}\left(\frac{\sum_{1 \leq i \leq n} X_i}{n} > a\right) \leq e^{-l(a)n}, \quad \text{tightest possible.}$$

where $l(a)$ is the Legendre transform corresponding to the distribution of random variable X_1 . This upper bound is **tight!**

(up to linearity of n)

$$e^{-l(a)n + o(n)}$$

$$\begin{aligned} M(\theta) &= E[e^{\theta X}] \\ &= e^{\theta \mu + \theta^2/2} \end{aligned}$$



Exponential Distribution Example

$$\text{pdf is } f(x) = \lambda \cdot e^{-\lambda x} \quad M(\theta) = E[e^{\theta x}] = \int e^{\theta x} \cdot f(x) dx$$

Exponential distribution with parameter λ . Recall that $M(\theta) = \lambda / (\lambda - \theta)$ when $\theta < \lambda$ and $M(\theta) = \infty$ otherwise. Therefore when $\theta < \lambda$,

$$l(a) = \sup_{\theta} \left(a\theta - \log \frac{\lambda}{\lambda - \theta} \right) = \sup_{\theta} (a\theta - \log \lambda + \log(\lambda - \theta)).$$

Setting the derivative of $g(\theta) = a\theta - \log \lambda + \log(\lambda - \theta)$ equal to zero we obtain the equation $a - 1/(\lambda - \theta) = 0$ which has the unique solution $\theta^* = \lambda - 1/a$. Therefore,

$$l(a) = a(\lambda - 1/a) - \log \lambda + \log(\lambda - 1/a) = a\lambda - 1 - \log \lambda - \log a.$$

The large deviations bound then tells us that when $a > 1/\lambda$,

$$\mathbb{P}\left(\frac{\sum_{1 \leq i \leq n} X_i}{n} > a\right) \approx e^{-(a\lambda - 1 - \log \lambda - \log a)n}. \rightarrow \text{more useful bound.}$$

Say $\lambda = 1$ and $a = 1.2$. This approximation gives $\approx e^{-(0.2 - \log 1.2)n}$. Recall that the process $X_1, X_1 + X_2, \dots, X_1 + \dots + X_n, \dots$ is a Poisson process with $\lambda = 1$. We can compute the probability $\mathbb{P}(\sum_{1 \leq i \leq n} X_i > 1.2n)$ exactly: it is the probability that the Poisson process has at most $n-1$ events before time $1.2n$. Thus,

$$\mathbb{P}\left(\frac{\sum_{1 \leq i \leq n} X_i}{n} > 1.2\right) = \mathbb{P}\left(\sum_{1 \leq i \leq n} X_i > 1.2n\right) = \sum_{0 \leq k \leq n-1} \frac{(1.2n)^k}{k!} e^{-1.2n}. \rightarrow \sum \text{poisson masses}$$

It is not at all clear how revealing this expression is. In hindsight, we know that it is approximately $e^{-(0.2 - \log 1.2)n}$. \rightarrow "useless"

Normal Distribution Example

Standard normal distribution. Recall that $M(\theta) = e^{\frac{\theta^2}{2}}$ when X_1 has the standard Normal distribution. The expected value $\mu = 0$. Thus we fix $a > 0$ and obtain

$$l(a) = \sup_{\theta} (a\theta - \theta^2/2) = a^2/2,$$

achieved at $\theta^* = a$. Again we see that $l(a)$ is (as it should be) a convex function of a . Thus for $a > 0$, the large deviations theory predicts that

$$\mathbb{P}\left(\frac{\sum_{1 \leq i \leq n} X_i}{n} > a\right) \approx e^{-\frac{a^2}{2}n}. \quad \text{more useful}$$

Again we could compute this probability directly. We know that $\frac{\sum_{1 \leq i \leq n} X_i}{n}$ is distributed as a Normal random variable with mean zero and variance $1/n$. Thus

$$\mathbb{P}\left(\frac{\sum_{1 \leq i \leq n} X_i}{n} > a\right) = \frac{\sqrt{n}}{\sqrt{2\pi}} \int_a^\infty e^{-\frac{t^2}{2}} dt. \quad \text{less useful}$$

One could show that this integral is “dominated” by its part around a , namely, $\frac{\sqrt{n}}{\sqrt{2\pi}} e^{-\frac{a^2}{2}n}$. This is consistent with the large deviations theory. The lower order term $\frac{\sqrt{n}}{\sqrt{2\pi}}$ disappears in the approximation on the log scale.

$$\begin{aligned} \frac{\sum X_i}{n} &\sim Z(0, \frac{1}{n}) \\ P(Z(0, \frac{1}{n}) > a) &= 1 - \text{CDF} \end{aligned}$$

Poisson Distribution Example

Poisson distribution. Suppose X has a Poisson distribution with parameter λ . Recall that $M(\theta) = e^{e^{\theta\lambda} - \lambda}$. Then

$$l(a) = \sup_{\theta} (a\theta - (e^{\theta\lambda} - \lambda)).$$

$$e^{-\text{train}} = e^{-[\dots]n}$$

Useful

Setting derivative to zero we obtain $\theta^* = \log(a/\lambda)$ and $l(a) = a \log(a/\lambda) - (a - \lambda)$.

In this case as well we can compute the large deviations probability explicitly. The sum $X_1 + \dots + X_n$ of Poisson random variables is also a Poisson random variable with parameter λn . Therefore

$$\mathbb{P}\left(\sum_{1 \leq i \leq n} X_i > an\right) = \underbrace{\sum_{m>an} \frac{(\lambda n)^m}{m!} e^{-\lambda n}}_{\sum \text{Poisson masses}}$$

$X_i \sim \text{Poisson}(\lambda)$

$X_1 + \dots + X_n \sim \text{Poisson}(n\lambda)$

But again it is hard to infer a more explicit rate of decay using this expression.

useless

similar to
chernoff.

Hoeffding Inequality

$$\forall i, a \leq x_i \leq b$$

Let X_1, \dots, X_n be i.i.d. random variables on a bounded support $[a, b]$. Let $\mathbb{E}(X_1) = \mu$. Then, for any $\epsilon > 0$,

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) \leq 2e^{-\frac{2n\epsilon^2}{(b-a)^2}}.$$

\downarrow \downarrow
empirical mean true mean

(两tail)

Example:

$$a=0, b=1$$

If $X_1, \dots, X_n \sim \text{Bernoulli}(p)$, then

$$\mathbb{P}(|\bar{X}_n - p| > \epsilon) \leq 2e^{-2n\epsilon^2}.$$

* 不需要知道 $M_X(\theta)$

* 这个 bound really tight

背这个公式 !!!

Hoeffding Inequality

Let X_1, \dots, X_n be i.i.d. random variables on a bounded support $[a, b]$. Let $\mathbb{E}(X_1) = \mu$. Then, for any $\epsilon > 0$,

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) \leq 2e^{-\frac{2n\epsilon^2}{(b-a)^2}}$$

由 subgaussian
property
(light-tailed)

Proof. WLOG assume $\mu = 0$. (if $\mu \neq 0$, $\tilde{X}_i = X_i - \mu$ 即可)

$$\mathbb{P}(X_1 + \dots + X_n \geq n\epsilon) = \mathbb{P}(e^{t(X_1 + \dots + X_n)} \geq e^{tn\epsilon}) \leq \frac{\mathbb{E}[e^{t(X_1 + \dots + X_n)}]}{e^{tn\epsilon}} = \frac{(\mathbb{E}e^{tX_1})^n}{e^{tn\epsilon}}.$$

We bound the MGF of X_1 . Below is a very useful inequality: for any X with zero mean and bounded support $[a, b]$,

$$\mathbb{E}e^{tX} \leq e^{t^2(b-a)^2/8}. \quad (1)$$

We will show this later. Now we have

$$\mathbb{P}(X_1 + \dots + X_n \geq n\epsilon) \leq \inf_t \frac{(\mathbb{E}e^{tX_1})^n}{e^{tn\epsilon}} \leq \inf_t \left(e^{nt^2(b-a)^2/8 - tn\epsilon} \right).$$

Choose $t = 4\epsilon/(b-a)^2$ such that the exponent is minimized. This completes the proof. \square

对 t 求导 $g'(t) = 0$

P18 用到的 lemma → Bound the MGF

We then prove that our claim is true, i.e., for any X with zero mean and bounded support $[a, b]$,

$$\mathbb{E}e^{tX} \leq e^{t^2(b-a)^2/8}. \quad (1)$$

good tight bound
on MGF with range $[a, b]$

We first write $X = \frac{b-X}{b-a}a + \frac{X-a}{b-a}b$. By convexity, $(\because E[X] = M = 0)$

$$\begin{aligned} \mathbb{E}e^{tX} &\leq \mathbb{E}\left(\frac{b-X}{b-a}e^{ta}\right) + \mathbb{E}\left(\frac{X-a}{b-a}e^{tb}\right) = \frac{b}{b-a}(e^{ta}) - \frac{a}{b-a}(e^{tb}) \\ &= e^{at} + e^{at}\left(\frac{a}{b-a}\right) - e^{bt}\left(\frac{a}{b-a}\right) = e^{at}\left(1 + \left(\frac{a}{b-a}\right) - \left(\frac{a}{b-a}\right)e^{t(b-a)}\right) = e^{g(u)}, \end{aligned}$$

where

$$g(u) = -\gamma u + \log(1 - \gamma + \gamma e^u), \quad \gamma = -\frac{a}{b-a}, \quad u = t(b-a).$$

$\nabla g(u) = 0$

Note that $g(0) = g'(0) = 0$ and $g''(x) \leq 1/4$ for all $x > 0$. Using Taylor's Theorem, we have for some $\xi \in [0, u]$,

Taylor \rightarrow

$$g(u) = g(0) + ug'(0) + \frac{u^2}{2}g''(\xi) \leq u^2/8 = t^2(b-a)^2/8.$$

展开
(RVv186)

$$g''(u) = \beta(1-\beta) \text{ 形式}$$

$$\therefore g''(u) \leq 1/4$$

Hoeffding Inequality

Let X_1, \dots, X_n be i.i.d. random variables on a bounded support $[a, b]$. Let $\mathbb{E}(X_1) = \mu$. Then, for any $\epsilon > 0$,

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) \leq 2e^{-\frac{2n\epsilon^2}{(b-a)^2}}.$$

Corollary:

exercise

Let X_1, \dots, X_n be i.i.d. random variables on a bounded support $[a, b]$. Let $\mathbb{E}(X_1) = \mu$. Then

$$|\bar{X}_n - \mu| \leq \sqrt{\frac{(b-a)^2}{2n} \log\left(\frac{2}{\delta}\right)}, \quad \text{with probability at least } 1 - \delta.$$

$$\mathbb{P}(|\bar{X}_n - \mu| \leq \sqrt{\frac{(b-a)^2}{2n} \log \frac{2}{\delta}}) \geq 1 - \delta$$

high-probability bound

$$\text{e.g. } \delta = 1\%, \text{ then } P \rightarrow 99\%$$