IOE 516

Stochastic Processes II

Winter Term, 2024

Prof. Xiuli Chao
Email: xchao@umich.edu

Recap

- Last week we discussed three fundamental limit theorems: WLLN, SLLN, and CLT.

- These results are useful. For example, CLT can help estimate tail probabilities.

Tail approximation of sum

- One application of CLT is in computing tail probabilities of sum of random variables. Let Φ be the cdf of standard normal.

- Let $X_1, X_2, \ldots$ be i.i.d. with mean μ and variance σ^2, and $S_n = \sum_{i=1}^n X_i$. By CLT, $(S_n - n\mu)/(\sigma\sqrt{n})$ is approximately standard normal when n is large. Thus, we can estimate the probability for $S_n > x$ as follows:

$$
\begin{aligned}
P(S_n > x) &= P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} > \frac{x - n\mu}{\sigma\sqrt{n}}\right) \\
&\approx 1 - \Phi\left(\frac{x - n\mu}{\sigma\sqrt{n}}\right)
\end{aligned}
$$

Example

- A car insurance firm has 10,000 customers. The annual premium is $\$800$. The claim for each customer over the year is random and historical data show that the mean is $\$700$ and standard deviation is $\$500$. What is the probability that the firm makes a profit of $\$1$ MM from these customers?

- Let X_i denote the net contribution of customer i which is the premium minus claim, then it has mean $\mu = 100$ and standard deviation $\$500$.

- The total profit is $S = \sum_{i=1}^{10000} X_i$. By the result from previous page,

$$
\begin{aligned}
P(S > 1,000,000) &= 1 - \Phi\left(\frac{x - n\mu}{\sigma\sqrt{n}}\right) \\
&= 1 - \Phi\left(\frac{1,000,000 - 10,000 \times 100}{500 \times 100}\right) = 0.5.
\end{aligned}
$$

Ordinary deviation

- CLT implies that, for large n we can estimate $P(S_n - n\mu > \sqrt{n}a)$ as follows:

$$P(S_n - n\mu > \sqrt{n}a) \approx P(Z > a/\sigma) = 1 - \Phi(a/\sigma).$$

- Deviations of S_n from its mean by the order of $\sqrt{n}$ is known as **ordinary deviation**.

- Thus, the probability for ordinary deviation is can be easily estimated.

However,

- Ordinary deviation is useful, but not enough for operations applications.

- Today we discuss more deviation results, known as concentration inequalities.

- First we discuss Large Deviation. Since Large Deviation is based MGF, let us elaborate further on MGF.

Some property of MGF

- We need some property of GMF $M_X(\theta)$. We focus on $\theta \geq 0$. Recall that $M_X(0) = 1$ for any X.

- We already know that, for some r.v.'s (e.g., lognormal), their MGF $M_X(\theta)$ is infinity for any $\theta > 0$.

- It can be argued that, if there exists $\theta_0 > 0$ such that $M_X(\theta_0)$ if finite, then $M_X(\theta) < \infty$ for all $0 \leq \theta < \theta_0$. This shows that the range of parameter θ sudh that $M_X(\theta) < \infty$ is an interval containing 0.

Large deviation theory

- Recall that $S_n = \sum_{i=1}^n X_i$. Large deviation is concerned with the event that S_n deviates from its mean by the order of n.

- That is, what is $P(S_n - n\mu > na)$ for $a > 0$.

- This probability should be small for moderate or large n, but how small and how to evaluate that?

- There is an entire subject area in probability, known as **Large Deviation Theory**. There are several good books are on this subject.

First, some preliminaries

- To present large deviation theory, we need some preparation. The first one is Chernoff bound.

- Let $X_1, X_2, \ldots$ i.i.d. with mean μ and variance σ^2.

- MGF $M(\theta) = E[e^{\theta X_1}]$, assuming finite near 0.

- **Properties of** $M(\theta)$: (i) $M(0) = 1$, and (ii) $M'(0) = \mu$.

Chernoff bound

- Let $s = \mu + a$. For $\theta > 0$, we have

$$P(S_n \geq ns) = P\left(e^{\theta S_n} > e^{\theta ns}\right) \leq \frac{E[e^{\theta S_n}]}{e^{\theta na}}$$
$$= \; e^{-\theta ns}\Big(E[e^{\theta X_1}]\Big)^n = e^{-n(\theta s - \log M(\theta))}$$

- $\theta s - \log M(\theta)$ is 0 when $\theta = 0$. Assuming finite on $[0, \theta_0)$, then

$$(\theta s - \log M(\theta))'|_{\theta-} = s - \frac{M'(0)}{M(0)} = s - \mu > 0.$$

- This confirms that there exists $\theta > 0$ such that $\theta s - \log M(\theta) > 0$. Thus, $P(S_n \geq na)$ goes to zero exponentially fast!

Remark

- $P(S_n \geq na)$ goes to zero at exponential rate $\theta s - \log M(\theta)$ when it is greater than 0.

- What is the fastest rate? It is clearly the largest $\theta s - \log M(\theta)$.

- That is, we need to identify that θ that maximizes $\theta s - \log M(\theta)$.

Fenchel-Legendre transform

- The Legendre transform of a r.v. is defined as

$$\Lambda^*(s) = \sup_{\theta \geq 0} \Big(\theta s - \log M(\theta) \Big) > 0$$

- **Properties of** $\Lambda(\theta) = \log M(\theta)$:

  (i) $\Lambda(0) = 0$,

  (ii) $\Lambda'(0) = \mu$, and

  (iii) $\Lambda(\theta)$ is convex.

- Thus, $\theta s - \Lambda(\theta)$ is a concave function. Let its maximizer be θ^*.

Cramer-Chernoff Theorem

- By Chernoff bound,

$$P(S_n \geq sn) \leq e^{-\Lambda^*(s)n}.$$

- This implies

$$\frac{1}{n} \cdot \log P(S_n \geq sn) \leq -\Lambda^*(s).$$

- **Large deviation theory**, known as Cramer-Chernoff Theorem, shows that the upper bound is actually tight. That is

$$\frac{1}{n} \cdot \log P(S_n \geq sn) \to -\Lambda^*(s) \quad \text{as } n \to \infty.$$

Example

- We use Poisson with parameter λ to illustrate.

- GMF is

$$E[e^{\theta X}] = \sum_{n=0}^{\infty} e^{\theta n} e^{-\lambda} \frac{\lambda^n}{n!} = e^{\lambda(e^{\theta}-1)}$$

- $\Lambda^*(s) = \sup_{\theta \geq 0}\{\theta s - \lambda(e^{\theta} - 1)\} = s \log(s/\lambda) - s + \lambda$ with $\theta^* = \log(s/\lambda)$.

- By LDT, $P(S_n > sn) \approx e^{-(s \log(s/\lambda) - s + \lambda)n}$. The LHS is

$$P(S_n > ns) = \sum_{k \geq sn} e^{-\lambda n} \frac{(\lambda n)^k}{k!}.$$

- This decay rate is hard to assess but LDT has provided the solution.

Remark

- How strong is "MGF is finite near 0"?

- It is basically those random variables whose tail function decays at least exponentially fast.

- This shows that, large deviation result is quite natural and expected.

- Let us argue it using non-negative random variables.

Exponential decay

- **Claim.** MGF is finite near 0 iff its tail is exponentially bounded, i.e., there exist $\mu > 0, \lambda > 0$, such that

$$P(X > t) \leq \mu e^{-\lambda t}.$$

- **Why?** If there exists $\theta_0 > 0$ such that $E[e^{\theta_0 X}] = \mu$ is finite, then by Markov inequality,

$$P(X > t) = P(e^{\theta_0 X} > e^{\theta_0 t}) \leq \frac{E[e^{\theta_0}]}{e^{\theta_0 t}} = \mu e^{-\theta_0 t}.$$

- On the other hand, if $P(X > t) \leq \mu e^{-\lambda t}$, then for any $\theta_0 < \lambda$, we have

$$E[e^{\theta_0 X}] = -\int_0^\infty e^{\theta_0 t} d\bar{F}(t) = -e^{\theta_0 t}\bar{F}(t)|_0^\infty + \theta_0 \int_0^\infty \bar{F}(t) e^{\theta_0 t} dt = \cdots$$

Medium deviation

- LDT considers the probability that S_n deviates from its mean by an. How about deviating by λ_n, say $\lambda_n = n^{2/3}, \sqrt{n \log n}, \sqrt{n} \log n$, etc.? These are particularly important for studying algorithms of operations problems.

- When λ_n lies in between n (large deviation) and $\sqrt{n}$ (ordinary deviation), it is called **medium deviation**.

- In the following we assume the mean μ of X_i is zero. If not, the formula should replace S_n by $S_n - n\mu$.

Medium deviation theory

- If λ_n grows faster than slower than $n^{2/3}$ but faster than $n^{1/2}$, then under the condition of finite MGF near 0, we have

$$P(S_n > \lambda_n) \approx \frac{\sigma}{\lambda_n}\sqrt{\frac{n}{2\pi}} \cdot e^{-\frac{\lambda_n^2}{2\sigma^2 n}}.$$

Important cases

- Of particular importance is the special case that $\lambda_n = c\sqrt{\sigma n \log n}$:

$$P(S_n > c\sqrt{\sigma n \log n}) \approx \frac{1}{c\sqrt{2\pi \log n}} \cdot n^{-c^2/2}.$$

- **Examples.** For $c = \sqrt{2}$ and 2, we obtain

$$\begin{aligned} P(S_n > \sqrt{2\sigma n \log n}) &\approx \frac{1}{2\sqrt{\pi \log n}} \cdot \frac{1}{n}, \\ P(S_n > 2\sqrt{\sigma n \log n}) &\approx \frac{1}{2\sqrt{2\pi \log n}} \cdot \frac{1}{n^2}. \end{aligned}$$

A very useful result

- In general, we have for any $\alpha > 0$,

$$P(S_n > \sqrt{2\alpha\sigma n \log n}) \approx \frac{1}{2\sqrt{\alpha\pi \log n}} \cdot \frac{1}{n^\alpha}.$$

- **Remark.** The result above does not require MGF be finite near zero. A sufficient condition is $E[|X_1|^{2(\alpha+1)+\delta}] < \infty$ for small $\delta > 0$. This condition is much weaker than those imposed in the operations literature.

Concentration inequalities

- A random variable concentrates its mass around the mean. How fast does it spread out, especially when the random variable is sum of random variables?

- Ordinary deviation, medium deviation, and large deviation are all concentration inequalities. But there are many more.

- Together, they play a central role in analyzing learning and approximation algorithms in the operations literature.

Discussion

- Concentration inequalties take the form

$$P(|X - E[X]| \geq \epsilon)$$

  for any given $\epsilon > 0$. It usually suffices to study $P(X - E[X] > \epsilon)$, and it is WLOG to assume $E[X] = 0$.

- Chebshev inequality is the simplest general result on concentration inequality.

Discussion

- Hot to get the result for $P(|X - E[X]| > \epsilon)$ from $P(X - E[X] > \epsilon)$?

- In all our results, the upper bounds for $P(|X - E[X]| > \epsilon)$ is twice of $P(X - E[X] > \epsilon)$.

Example: Normal

- X is normal N(0,1). Find upper bound for $P(X > \epsilon)$ for $\epsilon > 0$. What do you get if you apply Chebshev inequality?

- **Sharper result.** $P(X > \epsilon) \le \frac{e^{-\epsilon^2/2}}{\epsilon}$

- **Why?** Here is the argument:

$$P(X > \epsilon) = \int_\epsilon^\infty \phi(s)ds \le \frac{1}{\epsilon}\int_\epsilon^\infty s\phi(s)ds = \frac{e^{-\epsilon^2/2}}{\sqrt{2\pi}\epsilon}$$

Example:
Sample mean of normal

- X_i is normal $N(0,1)$ and $S_n = \sum_{i=1}^{n} X_i$. Find upper bound for $P(S_n > \epsilon)$ for $\epsilon > 0$. What do you get if you apply Chebshev inequality?

- **Sharpger result.** $P(S_n > \epsilon) \leq \frac{\sqrt{n}}{\sqrt{2\pi}\epsilon} e^{-\epsilon^2/(2n)}$

- **Why?**

Comparisons with large and medium deviations theory

- How do you compare the results with large and medium deviation theory results?

- **Homework**: Substitute ϵ by an and $c\sqrt{2n \log n}$, respectively, and compare the results with LDT and MDT.

Hoeffding inequality

- The most classic concentration inequality is Hoeffding inequality for bounded random variables. Given the role it played in the development, we discuss it in detail.

- **Hoeffding inequality**. Suppose $X_1, X_2, \ldots$ are i.i.d. on a bounded support $[a, b]$ and $E[X_1] = \mu$. Then, for $\epsilon > 0$,

$$P(S_n - n\mu > \epsilon) \leq e^{\frac{-2\epsilon^2}{n(b-a)^2}}.$$

- Important special case: **Bernoulli** with mean p, then

$$P(S_n - np > \epsilon) \leq e^{-2\epsilon^2/n}.$$

Remark

- Substituting ϵ by an and $\sqrt{\alpha n \log n}$ leads to similar result of Large deviation and medium deviation.

Why?

- WLOG we assume $\mu = 0$. By Chernoff bound: For any $\theta > 0$,
$$P(S_n > \epsilon) = P(e^{\theta S_n} > e^{\theta\epsilon}) \le e^{-\theta\epsilon}(E[e^{\theta X_1}])^n = e^{-\theta + n \log E[e^{\theta X_1}]}. \quad (1)$$

- Now, since X_1 has support on $[a, b]$, we can say more about $E[e^{\theta X_1}]$. We can show the following:
$$E[e^{\theta X_1}] \le e^{\theta^2 (b-a)^2/8}.$$

- How to prove? First, note that, by $X \in [a, b]$,
$$X = \frac{b - X}{b - a} a + \frac{X - a}{b - a} b,$$

- Since $e^{\theta X}$ is convex in X, we have
$$e^{\theta X} \le \frac{b - X}{b - a} e^{\theta a} + \frac{X - a}{b - a} e^{\theta b}$$

- Thus

$$
\begin{aligned}
E[e^{\theta X}] &\le \frac{b}{b-a}e^{\theta a} - \frac{a}{b-a}e^{\theta b} \\
&= e^{\theta a}\left(\frac{b}{b-a} - \frac{a}{b-a}e^{\theta(b-a)}\right) \\
&= e^{\theta a + \log\left(1+\frac{a}{b-a}-\frac{a}{b-a}e^{\theta(b-a)}\right)} \\
&= e^{g(u)},
\end{aligned}
$$

  where

$$
\begin{aligned}
g(u) &= -\gamma u + \log(1-\gamma+\gamma e^{u}), \\
\gamma &= -\frac{a}{b-a}, \\
u &= \theta(b-a).
\end{aligned}
$$

- It is an exercise for you to show that $g(0) = g'(0) = 0$ and $g''(x) \le 1/4$ for all $x \ge 0$. By Taylor's theorem,

$$
g(u) = g(0) + g'(0)u + g''(\xi)\frac{u^2}{2} \le u^2/8 = (b-a)^2\theta^2/8.
$$

- Substituting to (1) yields

$$P(S_n \geq \epsilon) \leq \inf_{\theta} e^{n(b-a)^2\theta^2/8-\epsilon\theta}$$

- The RHS is minimized when $\theta = 4\epsilon/n(b-a)^2$, to obtain

$$P(S_n \geq \epsilon) \leq e^{-\frac{2\epsilon^2}{n(b-a)^2}}.$$