IE 498: Online Learning and Decision Making, Fall 2019 Aug 27th, Aug 29th

Lecture 01 & 02: the Central Limit Theorem and Tail Bounds

Lecturer: Yuan Zhou *Scribe: Yuan Zhou*

1 Central Limit Theorem for *i.i.d.* random variables

Let us say that we want to analyze the total sum of a certain kind of result in a series of repeated independent random experiments each of which has a well-defined expected value and finite variance. In other words, a certain kind of result (e.g. whether the experiment is a "success") has some probability to be produced in each experiment. We would like to repeat the experiment many times independently and understand the total sum of the results.

1.1 Bernoulli variables

We first consider the sum of a bunch of Bernoulli variables.

Specifically, let X_1, X_2, ..., X_n be i.i.d. random variables with

$$\Pr[X_i = 1] = p, \qquad \Pr[X_i = 0] = 1 - p.$$

Let $S = S_n = X_1 + X_2 + ... + X_n$ and we want to understand S.

According to the linearity of expectation, we have

$$\mathbb{E}[S] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + ... + \mathbb{E}[X_n] = np.$$

Since $X_1, X_2, ...X_n$ are independent, we have $\mathrm{Var}[S] = np(1-p)$.

Now let us use a linear transformation to make S mean 0 and variance 1. I.e., let us introduce Z_n, a linear function of S_n, to be

$$Z_n = \frac{S_n - np}{\sqrt{np(1-p)}}.$$

Using $\mu = np$ and $\sigma = \sqrt{np(1-p)}$, we have

$$Z_n = \frac{S_n - \mu}{\sigma}.$$

Via this transformation, we do not lose any information about $S = S_n$. Specifically, for any u, we have

$$\Pr[S_n \le u] = \Pr[\sigma Z_n + \mu \le u] = \Pr\left[Z_n \le \frac{u-\mu}{\sigma}\right].$$

Therefore, we proceed to study the distribution of Z_n.

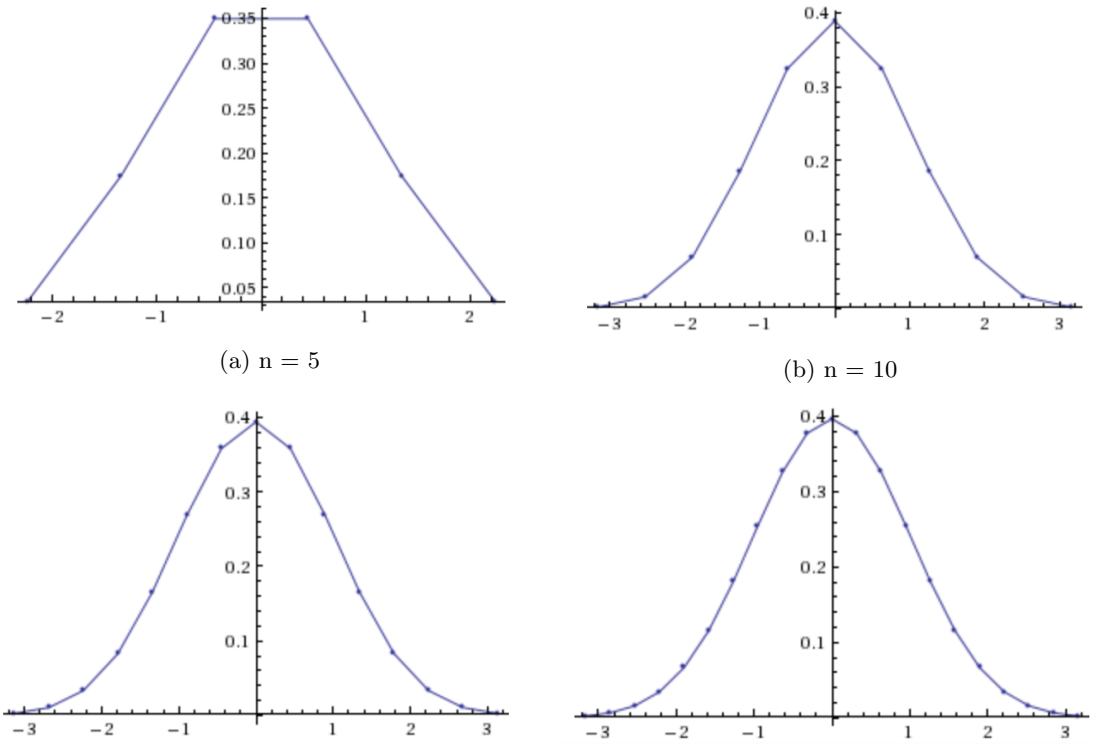
As a special instance, let us temporarily set $p = \frac{1}{2}$ so that X_i's become unbiased coin flips. In such case, we have

$$Z_n = \frac{X_1 + X_2 + ... + X_n - \frac{1}{2}n}{\frac{1}{2}\sqrt{n}} = \frac{1}{\sqrt{n}}((2X_1 - 1) + (2X_2 - 1) + ... + (2X_n - 1)).$$

For each integer $a \in [0, n]$, we have

$$\Pr\left[Z_n = \frac{2a - n}{\sqrt{n}}\right] = \frac{\binom{n}{a}}{2^n}.$$

Therefore, we can easily plot the probability density curve of Z_n. In Figure 1, we plot the density curve for a few values of n.



(a) n = 5

(b) n = 10

(c) n = 20

(d) n = 40

Figure 1: Probability density curves of Z_n for a few values of n

We can see that as $n \to \infty$, the probability density curve converges to a fixed continuous curve as illustrated in Figure 2.

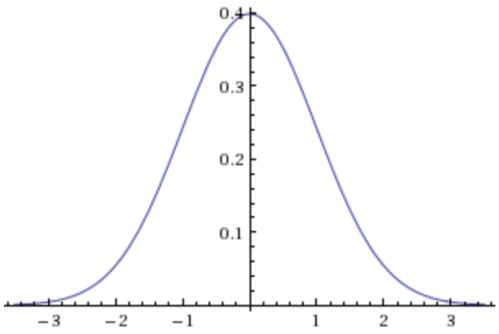


Figure 2: The famous "Bell curve" – the probability density function of a standard Gaussian variable

Indeed, even when $p = \Pr[X_i = 1]$ is a constant in $(0, 1)$ other than $\frac{1}{2}$, the probability density curve of Z_n still converges to the same curve as $n \to \infty$. We call the probability distribution using such curve as pdf the *Gaussian distribution* (or *Normal distribution*).

1.2 The Central Limit Theorem

The Central Limit Theorem (CLT) for i.i.d. random variables can be stated as follows.

Theorem 1 (the Central Limit Theorem)**.** *Let Z be a standard Gaussian. For any i.i.d $X_1, X_2, ..., X_n$ (not necessarily binary valued), as $n \to \infty$, we have $Z_n \to Z$ in the sense that $\forall u \in \mathbb{R}, \Pr[Z_n \leq u] \to \Pr[Z \leq u]$.*

More specifically, for each $\epsilon > 0$, there exists $N \in \mathbb{N}$ so that for every $n > N$ and every $u \in \mathbb{R}$, we have

$$|\Pr[Z_n \leq u] - \Pr[Z \leq u]| < \epsilon.$$

Definition 2. *We use $Z \sim \mathcal{N}(0, 1)$ to denote that Z is a standard Gaussian variable. More specifically, Z is a continuous random variable with probability density function*

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

We also use $Y \sim \mathcal{N}(\mu, \sigma)$ to denote that Y is a Gaussian variable with mean μ and variance σ^2, i.e. $Y = \sigma Z + \mu$ where Z is a standard Gaussian.

Now we introduce a few facts about Gaussian variables.

Theorem 3. *Let $\vec{Z} = (Z_1, Z_2, ..., Z_d) \in \mathbb{R}^d$, where $Z_1, Z_2, ..., Z_d$ are i.i.d. standard Gaussians. Then the distribution of $\vec{Z}$ is rotationally symmetric. I.e., the probability density will be the same for $\vec{z_1}$ and $\vec{z_2}$ when $\|\vec{z_1}\| = \|\vec{z_2}\|$.*

Proof. The probability density function of $\vec{Z}$ at $\vec{z} = (z_1, z_2, ..., z_d)$ is

$$\phi(z_1)\phi(z_2)...\phi(z_d) = \left(\frac{1}{\sqrt{2\pi}}\right)^d e^{-(z_1^2+z_2^2+...+z_d^2)/2} = \left(\frac{1}{\sqrt{2\pi}}\right)^d e^{-\|\vec{z}\|^2},$$

which only depends on $\|\vec{z}\|$. □

The following corollary says that the function $\phi(\cdot)$ is indeed a probability density function.

Corollary 4. $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \mathrm{d}z = 1$

Corollary 5. *Linear combination of independent gaussians is still gaussian.*

2 The Berry-Esseen Theorem (CLT with error bounds)

When designing and analyzing algorithms, we usually need to know the convergence rate in order to derive a guarantee on the performance (e.g. time/space complexity) of the algorithm. In this sense, the Central Limit Theorem (Theorem 1) may not be practically useful. The following Berry-Esseen theorem strengthens the CLT with concrete error bounds.

Theorem 6 (the Berry-Esseen Theorem)**.** *Let $X_1, X_2, ..., X_n$ be independent. Assume w.l.o.g. that $\mathbb{E}(X_i) = 0$ and $\mathrm{Var}(X_i) = \sigma_i^2$ and $\sum_{i=1}^n \sigma_i^2 = 1$. Let $Z = X_1 + X_2 + ... + X_n$. (Note that $\mathbb{E}[Z] = 1, \mathrm{Var}[Z] = 1$.) Then $\forall u \in \mathbb{R}$, we have*

$$\left| \Pr[S \le u] - \Pr_{Z \sim \mathcal{N}(0,1)}[Z \le u] \right| \le O(1) \cdot \beta,$$

where $\beta = \sum_{i=1}^{n} \mathbb{E}\,|X_i|^3$.

Remark 1. *The hidden constant in the upperbound of the theorem can be as good as .5514 by [She13].*

Remark 2. *The Berry-Esseen theorem does not need X_i's to be identical. Independence among variables is still essential.*

We still use the unbiased coin flips example to see how this bound works.

Let

$$X_i = \begin{cases} +\frac{1}{\sqrt{N}}, & w.p. \frac{1}{2} \\ -\frac{1}{\sqrt{N}}, & w.p. \frac{1}{2} \end{cases}$$

be independent random variables.

We can check that $\mathbb{E}[X_i] = 0$ and $\mathrm{Var}(X_i) = \frac{1}{n}, \sum \sigma_i^2 = 1$ satisfy the requirement in the Berry-Esseen theroem. We can also compute that $\mathbb{E}\,|X_i|^3 = \frac{1}{n^{\frac{3}{2}}}$, and therefore $\beta = \frac{1}{\sqrt{n}}$.

According to the Berry-Esseen theorem, we have

$$\forall u \in \mathbb{R}, \left| \Pr[S \le u] - \Pr_{Z \sim \mathcal{N}(0,1)}[Z \le u] \right| \le \frac{.56}{\sqrt{n}}. \tag{1}$$

The right-hand side ($\frac{.56}{\sqrt{n}}$) gives a concrete convergence rate.

Now let us investigate whether the $O\left(\frac{1}{\sqrt{n}}\right)$ upper bound can be improved. Say n is even, then $S = \frac{\#\text{Heads} - \#\text{Tails}}{\sqrt{n}}$. Then $S = 0 \Leftrightarrow \#H = \#T = \frac{n}{2}$. Now let us estimate this probability using (1). For $\epsilon > 0$, we have

$$\begin{aligned} \Pr[\#H = \#T] &= \Pr[S = 0] = \Pr[S \le 0] - \Pr[S \le -\epsilon] \\ &= (\Pr[S \le 0] - \Pr[Z \le 0]) - (\Pr[S \le -\epsilon] - \Pr[Z \le -\epsilon]) + (\Pr[Z \le 0] - \Pr[Z \le -\epsilon]) \\ &\le |\Pr[S \le 0] - \Pr[Z \le 0]| - |\Pr[S \le -\epsilon] - \Pr[Z \le -\epsilon]| + \Pr[-\epsilon < Z \le 0]. \end{aligned}$$

Taking $\epsilon \to 0^+$, we have

$$\Pr[\#H = \#T] \le |\Pr[S \le 0] - \Pr[Z \le 0]| - |\Pr[S \le -\epsilon] - \Pr[Z \le -\epsilon]| \le \frac{.56}{\sqrt{n}} + \frac{.56}{\sqrt{n}} = \frac{1.12}{\sqrt{n}}, \quad (2)$$

where the last inequality is because of (1).

On the other hand, it is easy to see that

$$Pr[\#H = \#T] = \frac{\binom{n}{\frac{n}{2}}}{2^n}.$$

Using Sterling's approximation, when $n \to \infty$, we have

$$\Pr[\#H = \#T] \to \frac{\sqrt{2\pi n}(\frac{n}{e})^n}{2\pi \cdot \frac{n}{2} \cdot (\frac{n}{2e})^n \cdot 2^n} = \frac{\sqrt{2}}{\sqrt{\pi n}} \approx \frac{.798}{\sqrt{n}}. \quad (3)$$

If we had a essentially better upper bound (say $o\left(\frac{1}{\sqrt{n}}\right)$) in (1), we would get an upper bound of $o\left(\frac{1}{\sqrt{n}}\right)$ in (2). This would contradict (3). Therefore the upper bound in (1) given by the Berry-Esseen theorem is asymptotically tight.

3 Tail Bounds

Let us consider the unbiased coin flips again. I.e., let the outcome of the i-th coin toss to be a random variables

$$X_i = \begin{cases} +1, & w.p.\ \frac{1}{2} \\ -1, & w.p.\ \frac{1}{2} \end{cases}.$$

We assume all coin tosses are independent and let would like to study the sum S of the first n coin tosses,

$$S = \sum_{i=1}^{n} X_i.$$

In this lecture, we would like to study the probability that S greatly deviates from its mean $\mathbb{E}\, S = 0$. Specifically, for some parameter t, we would like to estimate the probability $\Pr[S > t]$. Intuitively, we know that such probability should be "small" for large enough t. The goal of this lecture (and part of the next one) is to derive qualitative upper bounds on the tail probability mass parameterized by t.

As we did in the previous lecture, using the Berry-Esseen theorem, we know that

$$\left|\Pr[S \ge \sqrt{n} \cdot t] - \Pr[G \ge t]\right| \le \frac{O(1)}{\sqrt{n}},$$

where $G \sim \mathcal{N}(0, 1)$ is a standard Gaussian. For convenience, we may also use the following informal notation

$$\Pr[S \ge \sqrt{n} \cdot t] = \Pr[G \ge t] \pm \frac{O(1)}{\sqrt{n}}. \quad (4)$$

Using basic calculus, we can estimate that

$$\Pr[G \geq t] = \int_t^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \mathrm{d}t \leq O(1) \cdot e^{\frac{-t^2}{2}}. \tag{5}$$

Now let us fix the parameter $t = 10\sqrt{\ln n}$. Combining (4) and (5), we have

$$\begin{aligned}\Pr[S \geq \sqrt{n} \cdot t] \leq \Pr[G \geq t] + O\left(\frac{1}{\sqrt{n}}\right) &= O\left(\exp\left(-\frac{(10\sqrt{\ln n})^2}{2}\right)\right) + O\left(\frac{1}{\sqrt{n}}\right) \\ &= O\left(\frac{1}{n^{50}}\right) + O\left(\frac{1}{\sqrt{n}}\right) = O\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

We see that the tail mass of the standard Gaussian is only $O\left(\frac{1}{n^{50}}\right)$. However, the error term $O\left(\frac{1}{\sqrt{n}}\right)$ introduced by the Berry-Esseen theorem is much greater. This error term is the main reason that we can't get better results. In the following part of this lecture, we will try various other methods to improve the upper bound.

4 Markov inequality

When we only know the mean of a nonnegative random variable, Markov inequality gives a simple upper bound on the probability that it deviates from its mean.

Theorem 7 (Markov inequality)**.** *Given a random variable X, assume $X \geq 0$. For each parameter $t \geq 1$, we have* $\Pr[X \geq t \cdot \mathbb{E}[X]] \leq \frac{1}{t}$.

Proof. For each $\alpha > 0$, we have

$$\begin{aligned}\mathbb{E}[X] = \Pr[X \geq \alpha] \cdot \mathbb{E}[X|X \geq \alpha] + \Pr[X < \alpha] \cdot \mathbb{E}[X|X < \alpha] \\ \geq \Pr[X \geq \alpha] \cdot \alpha + \Pr[X < \alpha] \cdot 0 = \Pr[X \geq \alpha] \cdot \alpha.\end{aligned}$$

Dividing both sides of the inequality by $\alpha > 0$

$$\frac{\mathbb{E}[X]}{\alpha} \geq \Pr[X \geq \alpha]$$

Taking $\alpha = \mathbb{E}[X] \cdot t$ we get the desired bound. □

Now let us try to apply the Markov inequality to bounding the tail mass of S. Since S is not a nonnegative random variable, we cannot directly apply the inequality. However, note that $S \geq -n$ always holds. We apply the inequality to $T = S + n$ where $\mathbb{E}\,T = \mathbb{E}\,S + n = n$. Let $t = 10\sqrt{n \ln n}$. We have

$$\Pr[S \geq t] = \Pr[T \geq t + n] = \Pr\left[T \geq (\mathbb{E}\,T) \cdot \frac{t+n}{n}\right] \leq \frac{n}{t+n} = \frac{n}{n + 10\sqrt{n \ln n}}.$$

This is a very bad bound – it does not even converge to 0 as n grows!

5 Chebyshev inequality

The Chebyshev inequality not only uses the mean of the random variable, but also needs the variance (or the second moment). Since we have more information about the random variable, we may potentially get better bounds.

Theorem 8 (Chebyshev inequality)**.** *Asumme that* $\mathbb{E}[X] = \mu$ *and* $\mathrm{Var}[X] = \sigma^2 > 0$*. For every parameter* $t > 0$*, we have* $\Pr[|X - \mu| \geq t \cdot \sigma] \leq \frac{1}{t^2}$.

Proof. Let $Y = (X - \mu)^2$. We can check that $E[Y] = \sigma^2$ and $Y \geq 0$. Applying Markov inequality, we have

$$\Pr[|X - \mu| \geq t \cdot \sigma] = \Pr[(X - \mu)^2 \geq t^2 \cdot \sigma^2] = \Pr[Y \geq t^2 \cdot E[Y]] \leq \frac{1}{t^2}.$$

□

Now let us go back to the scenario discussed at the beginning of this lecture. We compute that

$$\mu = \mathbb{E}\, S = 0 \qquad \text{and} \qquad \sigma = \sqrt{\mathrm{Var}[S]} = \sqrt{\mathbb{E}\, S^2} = \sqrt{n}.$$

Therefore

$$\begin{aligned}\Pr[S \geq 10\sqrt{n \ln n}] &\leq \Pr[|S| \geq 10\sqrt{n \ln n}] \\ &= \Pr\left[|S| \geq \sigma \cdot \frac{10\sqrt{n \ln n}}{\sigma}\right] \leq \frac{\sigma^2}{(10\sqrt{n \ln n})^2} = \frac{1}{100 \ln n}.\end{aligned}$$

This bound is still not as good as expected. However, at least it converges to 0 as $n \to \infty$.

Remark 3. *Note that Chebyshev inequality only needs pairwise independence among* X_i*'s. Specifically, when computing the variance of* S*, we have*

$$\begin{aligned}\mathrm{Var}[S] &= \mathrm{Var}[X_1 + X_2 + ... + X_n] \\ &= \mathbb{E}[(X_1 + X_2 + ... + X_n)^2] - \mathbb{E}[(X_1 + X_2 + ... + X_n)]^2 = \mathbb{E}[(X_1 + X_2 + ... + X_n)^2] \\ &= \sum_{i=1}^{n} \mathbb{E}[X_i^2] + \sum_{i \neq j} \mathbb{E}[X_i X_j] = \sum_{i=1}^{n} \mathbb{E}[X_i^2] = n.\end{aligned}$$

In the penultimate equality, we used the fact that X_i *is independent from* X_j *for* $i \neq j$*.*

6 The fourth moment method

Using the first two moments, we had better bounds then only using the mean of the random variable. Now let us try to extend this method to the fourth moment.

Let us consider $S^4 \geq 0$. By Markov inequality, we have

$$\Pr[S \geq 10\sqrt{n \ln n}] \leq \Pr\left[S^4 \geq (10\sqrt{n \ln n})^4\right] \leq \frac{\mathbb{E}\, S^4}{10000 n^2 \ln^2 n}. \tag{6}$$

Now let us estimate

$$\begin{aligned}\mathbb{E}[S^4] &= \mathbb{E}\left[\left(\sum_i X_i\right)^4\right] \\ &= \sum_i \mathbb{E}\, X_i^4 + \sum_i \sum_{j\neq i} \mathbb{E}\, X_i^2 X_j^2 \cdot \frac{1}{2}\binom{4}{2} + \sum_i \sum_{j\neq i} \mathbb{E}\, X_i X_j^3 \cdot \binom{4}{1} + \sum_i \sum_{j\neq i} \sum_{k:k\neq i,k\neq j} \mathbb{E}\, X_i X_j X_k^2 \cdot \binom{4}{2} \\ &\qquad + \sum_i \sum_{j\neq i} \sum_{k:k\neq i,k\neq j} \sum_{q:q\neq i,q\neq j,q\neq k} \mathbb{E}\, X_i X_j X_k X_q. \quad (7)\end{aligned}$$

Fortunately because of independence and $\mathbb{E}\, X_i = 0$, we have that $\mathbb{E}\, X_i X_j^3 = \mathbb{E}\, X_i X_j X_k^2 = \mathbb{E}\, X_i X_j X_k X_q = 0$. Therefore we can simplify (7) by

$$\mathbb{E}[S^4] = \sum_i \mathbb{E}\, X_i^4 + \sum_i \sum_{j\neq i} \mathbb{E}\, X_i^2 X_j^2 \cdot 3 = n + 3n(n-1) \leq 3n^2. \quad (8)$$

Combining (6) and (8), we get

$$\Pr[S \geq 10\sqrt{n \ln n}] \leq \frac{3n^2}{10000 n^2 \ln^2 n} = \frac{3}{10000 \ln^2 n}.$$

This is a better bound than what we get from Chebyshev.

Remark 4. *When using the fourth moment method, we only used the independence among every quadruple of random variables. Therefore the bound works for* 4*-wise independent random variables too.*

Remark 5. *We can extend this method to considering S^{2k} for positive integer k's, and picking the k that optimizes the upper bound. However, this plan would lead to the painful estimation of $\mathbb{E}\, S^{2k}$. We will use a slightly different method to get better upper bounds.*

7 The "Chernoff method"

Instead of S^{2k}, let us consider the function $e^{\lambda S}$ for some positive parameter λ.

Since e^x is a monotonically increasing function, we have

$$\Pr[S \geq 10\sqrt{n \ln n}] = \Pr[\lambda S \geq 10\lambda\sqrt{n \ln n}] \leq \Pr[e^{\lambda S} \geq e^{10\lambda\sqrt{n \ln n}}].$$

By Markov inequality (also checking that $e^{\lambda S} > 0$, we have

$$\Pr[e^{\lambda S} \geq e^{10\lambda\sqrt{n \ln n}}] \leq \frac{\mathbb{E}\, e^{\lambda S}}{e^{10\lambda\sqrt{n \ln n}}}. \quad (9)$$

Now it remains to upper bound $\mathbb{E}\, e^{\lambda S}$. We have

$$\mathbb{E}\, e^{\lambda S} = \mathbb{E}\, e^{\lambda \sum_i X_i}] = \mathbb{E} \prod_i e^{\lambda X_i} = \prod_i \mathbb{E}\, e^{\lambda X_i}. \quad (10)$$

Note that in the last equality, we used the full independence among all X_i's.

On the other hand, by the distribution of X_i, we have

$$\begin{aligned}\mathbb{E}\, e^{\lambda X_i} &= \frac{1}{2}e^{\lambda} + \frac{1}{2}e^{-\lambda} \\ &= \frac{1}{2}\left(1+\lambda+\frac{\lambda^2}{2!}+\frac{\lambda^3}{3!}+\frac{\lambda^4}{4!}+\dots\right)+\frac{1}{2}\left(1-\lambda+\frac{\lambda^2}{2!}-\frac{\lambda^3}{3!}+\frac{\lambda^4}{4!}-\dots\right) \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Taylor expansion)} \\ &= 1+\frac{\lambda^2}{2!}+\frac{\lambda^4}{4!}+\frac{\lambda^6}{6!}+\ldots \le e^{\lambda^2/2}.\end{aligned}$$

Getting back to (10), we have

$$\mathbb{E}\, e^{\lambda S} \le \prod_i e^{\lambda^2/2} = e^{n\lambda^2/2}.$$

Combining this with (9), we have

$$\Pr[e^{\lambda S} \ge e^{10\lambda\sqrt{n\ln n}}] \le e^{n\lambda^2/2 - 10\lambda\sqrt{n\ln n}}. \tag{11}$$

Picking $\lambda = 10\sqrt{\frac{\ln n}{n}}$, we minimize the right-hand side of (11) and get our desired upper bound

$$\Pr[e^{\lambda S} \ge e^{10\lambda\sqrt{n\ln n}}] \le e^{50\ln n - 100\ln n} = \frac{1}{n^{50}}.$$

In the beginning of the next lecture, we are going to extend this method to more general random variables and general thresholds, and go through the proof the famous Chernoff bound.

8 The Chernoff Bound

Using the same idea, we prove the following Chernoff Bound.

Theorem 9. *Let $X_1, X_2, \dots, X_n$ be independent Bernoulli random variables. Assume $\mathbb{E}\, X_i = p_i$ for every $i \in [n]$. Let $X = \sum_{i=1}^n X_i$, $\mu = \mathbb{E}\, X$. Then for any $\delta > 0$, it holds that*

1. $\Pr[X \ge (1+\delta)\mu] \le \left[\dfrac{e^{\delta}}{(1+\delta)^{1+\delta}}\right]^{\mu}$;
2. $\Pr[X \ge (1-\delta)\mu] \le \left[\dfrac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\mu}$.

Remark 6.

1. *The Chernoff Bound also holds when $X_i \in [0,1]$ instead of being Bernoulli.*
2. *When $\delta \in (0,1]$, the Chernoff Bound also implies the following inequalities which might be easier to use.*

$$\Pr[X \ge (1+\delta)\mu] \le \exp\left(-\delta^2\mu/3\right), \quad \textit{and} \quad \Pr[X \ge (1-\delta)\mu] \le \exp\left(-\delta^2\mu/3\right).$$

Proof of Theorem 9. We only prove the upper tail, and the lower tail can be shown in the similar way.

For any $\lambda > 0$, we have

$$\Pr[X \ge (1+\delta)\mu] = \Pr\left[e^{\lambda X} \ge e^{\lambda(1+\delta)\mu}\right] \le \frac{\mathbb{E}\, e^{\lambda X}}{e^{\lambda(1+\delta)\mu}}, \tag{12}$$

where the inequality is due to Markov Inequality.

Note that

$$\mathbb{E}\, e^{\lambda X_i} = p_i e^{\lambda} + (1 - p_i) = 1 + p_i(e^{\lambda} - 1) \leq \exp\left(p_i(e^{\lambda} - 1)\right),$$

where the last inequality is because of $1 + x \leq e^x$ for all real value x. Therefore we have

$$\mathbb{E}\, e^{\lambda X} = \prod_{i=1}^{n} \mathbb{E}\, e^{\lambda X_i} = \exp\left(\sum_{i=1}^{n} p_i(e^{\lambda} - 1)\right) = \exp\left(\mu(e^{\lambda} - 1)\right).$$

Combining with (12), we have

$$\Pr\left[X \geq (1+\delta)\mu\right] \leq \exp\left(\mu(e^{\lambda} - 1) - \lambda(1+\delta)\mu\right).$$

Setting $\lambda = \ln(1+\delta) > 0$, we prove the desired upper bound. □

9 Other useful tail bounds

In class we also mentioned the following concentration inequalities which will be useful in later lectures.

Theorem 10 (Hoeffding's Inequality)**.** *Let $X_1, X_2, \ldots, X_n$ be independent random variables, Suppose $X_i \in [a_i, b_i]$ for every $i \in [n]$. Let $X = \sum_{i=1}^{n} X_i$. For any $t > 0$, it holds that*

$$\Pr[X - \mathbb{E}\, X \geq t] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Theorem 11 (Bernstein Inequality)**.** *Let $X_1, X_2, \ldots, X_n$ be independent random variables, Suppose $\mathbb{E}\, X_i = 0$ and $|X_i| \leq M$ for every $i \in [n]$. Let $X = \sum_{i=1}^{n} X_i$. For any $t > 0$, it holds that*

$$\Pr[X \geq t] \leq \exp\left(-\frac{\frac{1}{2}t^2}{\sum_{i=1}^{n} \mathbb{E}\, X_i^2 + \frac{1}{3}Mt}\right).$$

References

[She13] I. G. Shevtsova. On the absolute constants in the Berry–Esseen inequality and its structural and nonuniform improvements. *Inform. Primen.*, **7**(1):124–125, 2013.