

Lecture 14: Concentration Inequalities

Lecturer: Yiqiao Zhong

Last update: 15:08, 2024-12-09

Disclaimer: materials in this pdf are used for STAT 709 “Mathematical Statistics” at the University of Wisconsin–Madison, but they are not guaranteed to be accurate. Consult textbooks whenever possible.

In this lecture, we will introduce some basics for understanding high-dimensional phenomena.

What are concentration inequalities? From the LLN, we know that an empirical mean $\hat{\mu}$ “concentrates” around the population mean μ . This is characterized by the asymptotic notion of convergence in probability. When we zoom in on the difference $\hat{\mu} - \mu$ by multiplying it with a factor of $\sqrt{n}$, then we know that the $\sqrt{n}(\hat{\mu} - \mu)$ has a normal distribution asymptotically.

Concentration inequalities provide non-asymptotic bounds on the difference as an alternative way of characterizing the concentration. Usually we will see statements like “With very high probability, a certain event (e.g. inequality) happens.” Often the exception probability decays exponentially in the number of variables. Compared with classical asymptotic theory, concentration inequalities often give us refined quantitative control of extreme/exception probabilities.

In this lecture, we use C and c to refer to a generic constant.

1 Motivation for concentration inequalities

Recall the central limit theorem: if $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} X$ satisfy $\mathbb{E}X = 0$ and $\mathbb{E}X^2 = 1$, then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \xrightarrow{d} \mathcal{N}(0, 1).$$

Recall the Gaussian tail probability bounds from Lecture 1.

Proposition 1 (Gaussian tail probability). *Let $G \sim \mathcal{N}(0, 1)$, then for any $t > 0$,*

$$\left(\frac{1}{t} - \frac{1}{t^3}\right) \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \leq \mathbb{P}(G \geq t) \leq \frac{1}{t} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}.$$

Let F and P be the Gaussian CDF and Gaussian pdf respectively. From the above proposition, if $t \rightarrow \infty$, then

$$1 - F(t) = (1 + o(1)) \frac{1}{t} P(t).$$

If we pretend that the following approximation is valid,

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n X_i > t\right) = \mathbb{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i > \sqrt{n}t\right) \approx \mathbb{P}(G > \sqrt{n}t),$$

then the tail probability of empirical mean of n i.i.d. random variables is $O\left(e^{-nt^2/2}\right)$. This argument is wrong, but the conclusion about an exponentially decaying tail probability holds in many cases.

Proposition 2 (Berry-Esseen Inequality). *Suppose that $X_1, X_2, \dots, X_n$ are i.i.d. random variables with zero mean and finite third moments. Then, for all $t \in \mathbb{R}$,*

$$\left| \mathbb{P} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \geq t \right) - \mathbb{P}(G \geq t) \right| \leq \frac{C \mathbb{E}|X|^3}{\sqrt{n}},$$

where C is a positive constant.

The error term is $O\left(\frac{1}{\sqrt{n}}\right)$, much larger than $O\left(e^{-nt^2/2}\right)$ if $t \gg \frac{1}{\sqrt{n}}$. Moreover, the error term cannot be made smaller in general due to the next simple example.

Thus, unfortunately, there is no hope of deriving a non-asymptotic bound on the tail probability of $\mathbb{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \geq t\right)$ by combining the Gaussian tail probability bound and the Berry-Esseen Inequality.

Example 1. Let $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Bernoulli}(\frac{1}{2})$, where n is an even integer, then

$$\mathbb{P} \left(\sum_{i=1}^n Y_i = \frac{n}{2} \right) = 2^{-n} \binom{n}{n/2} = 2^{-n} \frac{n!}{[(n/2)!]^2}.$$

Applying Stirling's approximation:

$$n! = (1 + o(1)) \sqrt{2\pi n} \left(\frac{n}{e}\right)^n,$$

we have

$$\mathbb{P} \left(\sum_{i=1}^n Y_i = \frac{n}{2} \right) = (1 + o(1)) \sqrt{\frac{2}{\pi}} \frac{1}{\sqrt{n}}$$

such that

$$\mathbb{P} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n (Y_i - \mathbb{E}Y_i) = 0 \right) \xrightarrow{n \rightarrow \infty} \sqrt{\frac{2}{\pi}} \frac{1}{\sqrt{n}}.$$

Note that $\mathbb{P}(G = 0) = 0$. In this example, the gap between the two CDFs is $O(\frac{1}{\sqrt{n}})$.

2 Subgaussian variables and Hoeffding's Inequality

First, we introduce an important definition.

Theorem 1 (Subgaussian random variables). *Let X be a random variable with $\mathbb{E}X = 0$. Consider three conditions.*

$$(i) \quad \mathbb{P}(|X| \geq t) \leq 2e^{-t^2/K_1^2}, \quad \forall t \geq 0.$$

$$(ii) \quad \sup_{p \geq 1} \frac{1}{\sqrt{p}} (\mathbb{E}|X|^p)^{1/p} \leq K_2.$$

$$(iii) \quad \mathbb{E}[e^{\lambda X}] \leq e^{K_3^2 \lambda^2}, \quad \forall \lambda \in \mathbb{R}.^1$$

Conditions (i), (ii), (iii) are equivalent in the sense that $K_i \leq CK_j$, $\forall i \neq j$ for certain absolute constant C . We say X is a subgaussian random variable if any of the three conditions hold.

¹Note: $\mathbb{E}[e^{\lambda x}]$ is called moment generating function (m.g.f.) of X .

Proof. (i) $\Rightarrow$ (ii): Without loss of generality, assume $K_1 = 1$. Otherwise set $X := X/K_1$. Observe

$$\begin{aligned}
\mathbb{E}|X|^p &= \int_0^\infty \mathbb{P}(|X| \geq t) p t^{p-1} dt \quad (\text{see Lecture 4, or use Fubini's theorem}) \\
&\leq \int_0^\infty 2 p t^{p-1} e^{-t^2} dt \\
&= p \int_0^\infty u^{(p-2)/2} e^{-u} du \quad (\text{change of variables } u = t^2) \\
&= p \Gamma(p/2) \quad (\text{Gamma function}) \\
&\leq 3p \left(\frac{p}{2}\right)^{p/2} \quad (\text{by property of Gamma functions}).
\end{aligned}$$

(ii) $\Rightarrow$ (iii): Without loss of generality, assume $K_2 = 1$ for simplicity. Let $\tilde{X} \stackrel{d}{=} X$ be an independent copy. Note that $X - \tilde{X}$ has a symmetric distribution, i.e., $X - \tilde{X} \stackrel{d}{=} \tilde{X} - X$ ([exercise](#)). Recall Jensen's inequality: If Z is a random variable, $h : \mathbb{R} \rightarrow \mathbb{R}$ is a convex function, then $\mathbb{E}g(Z) \geq g(\mathbb{E}Z)$. Therefore,

$$\mathbb{E}e^{-\lambda\tilde{X}} \geq e^{-\lambda\mathbb{E}\tilde{X}} = 1.$$

So,

$$\begin{aligned}
\mathbb{E}e^{\lambda X} &\leq \mathbb{E}e^{\lambda X} \mathbb{E}e^{-\lambda\tilde{X}} = \mathbb{E}e^{\lambda(X-\tilde{X})} \quad (\text{by independence between } X \text{ and } \tilde{X}) \\
&= \mathbb{E} \left[\sum_{k=0}^\infty \frac{[\lambda(X-\tilde{X})]^k}{k!} \right] \\
&= \sum_{k=0}^\infty \frac{\lambda^k \mathbb{E}[(X-\tilde{X})^k]}{k!} \quad (\text{by Fubini's theorem}) \\
&= 1 + \sum_{k=2, k \text{ even}}^\infty \frac{\lambda^k \mathbb{E}[(X-\tilde{X})^k]}{k!} \quad (\text{by symmetry}) \\
&\leq 1 + \sum_{k=2, k \text{ even}}^\infty C_1 \frac{(2\lambda)^k k^{k/2}}{\sqrt{2\pi k} (k/e)^k} \quad (\text{by Stirling's formula}) \\
&\leq 1 + \sum_{k=2, k \text{ even}}^\infty C_2 \frac{(2\sqrt{e}\lambda)^k}{\sqrt{2\pi k/2} (k/(2e))^{k/2}} \\
&\leq 1 + \sum_{m=2}^\infty C_2 \frac{(4e\lambda^2)^m}{m!} \quad (\text{letting } m := k/2.) \\
&\leq 1 + \sum_{m=2}^\infty \frac{(4C_2 e\lambda^2)^m}{m!} \quad (\text{by Stirling's formula}) \\
&= e^{K_2 \lambda^2} \quad (\text{Choose } K_2 = 4C_2 e).
\end{aligned}$$

(iii) $\Rightarrow$ (i): Without loss of generality, assume $K_3 = 1$. For any $\lambda > 0$, $t > 0$,

$$\begin{aligned}
\mathbb{P}(X \geq t) &\leq \mathbb{P}(e^{\lambda X} \geq e^{\lambda t}) \\
&\leq e^{-\lambda t} \mathbb{E}e^{\lambda X} \quad (\text{by Markov's inequality}) \\
&\leq e^{-\lambda t + \lambda^2} \\
&\leq e^{-t^2/4} \quad (\text{choose } \lambda = t/2 \text{ to minimize } -\lambda t + \lambda^2).
\end{aligned}$$

Repeating the same for $-X$, we have

$$\mathbb{P}(-X \geq t) \leq e^{-t^2/4}.$$

Therefore,

$$\mathbb{P}(|X| \geq t) \leq 2e^{-t^2/4}, \quad \forall t > 0.$$

□

The smallest K_2 is called subgaussian norm of X , denoted by $\|X\|_{\psi_2}$. By Minkowski inequality, it is easy to check that it satisfies the requirements of a norm ([exercise](#)). Like the variance, the subgaussian norm measures the spread of a zero-mean random variable, but it requires the probability tail decay as fast as a Gaussian tail in order for the norm to be finite.

Example 2. If $X \sim \mathcal{N}(0, \sigma^2)$, then $\|X\|_{\psi_2} \leq C\sigma$.

Example 3. If $|X| \leq M$ a.s., then $\|X\|_{\psi_2} \leq CM$.

For a sum of independent subgaussian random variables, we have the desired Gaussian tail bound.

Theorem 2 (Hoeffding's Inequality). Let $X_1, X_2, \dots, X_n$ be independent random variables, such that $\mathbb{E}X_i = 0$ and each X_i is subgaussian. Let $a = (a_1, \dots, a_n)$ be a nonrandom vector. Denote $K := \max_{1 \leq i \leq n} \|X_i\|_{\psi_2}$. Then

$$\mathbb{P}\left(\left|\sum_{i=1}^n a_i X_i\right| \geq t\right) \leq 2 \exp\left(-\frac{ct^2}{K^2 \|a\|^2}\right) \quad (1)$$

holds for certain absolute constant $c > 0$.

In particular, if we choose $a_i = 1/n$ for all i then Hoeffding's inequality says that we have the Gaussian tail bound.

Apart from the constant, this bound cannot be improved. In the case $X_i \stackrel{\text{i.i.d.}}{\sim} N(0, 1)$, we have $\sum_{i=1}^n a_i X_i \sim N(0, \|a\|^2)$, so by Proposition 1 the tail probability is roughly characterized by $\exp(-t^2/(2\|a\|^2))$.

Proof of Theorem 2.

$$\begin{aligned} \mathbb{E} \exp\left(\lambda \sum_{i=1}^n a_i X_i\right) &= \prod_{i=1}^n \mathbb{E} \exp(\lambda a_i X_i) \quad (\text{by independence}) \\ &\leq \prod_{i=1}^n \exp(C\lambda^2 a_i^2 \|X_i\|_{\psi_2}^2) \quad (\text{by Theorem 1 (iii)}) \\ &\leq \exp(C\lambda^2 \sum_{i=1}^n a_i^2 K^2) \\ &= \exp(C\lambda^2 \|a\|^2 K^2). \end{aligned}$$

Then use Theorem 1 (i). □

3 Subexponential variables and Bernstein's Inequality

We introduce another notion that characterizes fast decay of the tail probability, though not as fast as subgaussian tails.

Theorem 3 (Subexponential random variables). Let X be a random variable with $\mathbb{E}X = 0$.

$$(i) \quad \mathbb{P}(|X| \geq t) \leq 2e^{-t/K_1}, \quad \forall t \geq 0.$$

$$(ii) \sup_{p \geq 1} \frac{1}{p} (\mathbb{E}|X|^p)^{1/p} \leq K_2.$$

$$(iii) \mathbb{E}[e^{\lambda X}] \leq e^{K_3 \lambda^2}, \quad \forall |\lambda| \leq 1/K_3.$$

Conditions (a), (b), (c) are equivalent in the sense that $K_i \leq CK_j$, $\forall i \neq j$ for certain absolute constant C .

The smallest K_2 is called subexponential norm of X , denoted by $\|X\|_{\psi_1}$. By definition, subgaussian random variables are necessarily subexponential random variables with $\|X\|_{\psi_1} \leq \|X\|_{\psi_2}$; this is because

$$\frac{1}{p} (\mathbb{E}|X|^p)^{1/p} \geq \frac{1}{\sqrt{p}} (\mathbb{E}|X|^p)^{1/p}, \quad \text{for any } p \geq 1$$

so taking $\sup_{p \geq 1}$ yields $\|X\|_{\psi_1} \leq \|X\|_{\psi_2}$. In fact, intuitively, a subexponential random variable can be viewed as a square of a subgaussian random variable. This intuition is captured by the following lemma.

Centering. For random variables with nonzero mean, conditions (i) and (ii) in Theorem 3 (and similarly in Theorem 3) are still equivalent. Moreover, we have bound

$$\|X - \mathbb{E}X\|_{\psi_2} \leq C\|X\|_{\psi_2}.$$

See Lemma 2.6.8 of the book [1].

Lemma 1. Suppose that X, Y are subgaussian (not necessarily independent). Then XY is subexponential with

$$\|XY\|_{\psi_1} \leq \|X\|_{\psi_2} \|Y\|_{\psi_2}.$$

Proof. We use the second condition of subgaussian and subexponential random variables.

$$\begin{aligned} [\mathbb{E}(|X|^p |Y|^p)]^{1/p} &\leq [\mathbb{E}|X|^{2p} \mathbb{E}|Y|^{2p}]^{1/(2p)} \quad (\text{Cauchy-Schwarz Inequality}) \\ &\leq C\|X\|_{\psi_2} \sqrt{p} \|Y\|_{\psi_2} \sqrt{p}. \end{aligned}$$

□

Example 4. Let $G \sim \mathcal{N}(0, 1)$, then G^2 (namely chi-square χ_1^2) is subexponential, $\|G^2\|_{\psi_1} \leq C$.

Example 5. Let $X \sim \text{Exp}(1)$ be an exponential random variable, then $\|X\|_{\psi_1} \leq C$.

Example 6. Let X be a random variable such that $|X| \leq M$ is bounded for $M \geq 0$. Then, $\|X\|_{\psi_1} \leq CM$.

Now we are in the position to state the concentration inequality for a sum of subexponential random variables.

Theorem 4 (Bernstein's Inequality). Let $X_1, \dots, X_n$ be independent random variables, such that $\mathbb{E}X_i = 0$ and each X_i is subexponential. Let $a = (a_1, \dots, a_n)$ be a nonrandom vector. Denote $K := \max_{1 \leq i \leq n} \|X_i\|_{\psi_1}$. Then

$$\mathbb{P}\left(\left|\sum_{i=1}^n a_i X_i\right| \geq t\right) \leq 2 \exp\left(-c \min\left(\frac{t^2}{K^2 \|a\|^2}, \frac{t}{K \|a\|_\infty}\right)\right), \quad (2)$$

where $c > 0$ is a constant.

Subexponential random variables have “heavier” tails compared with subgaussian random variables. Expectedly, the upper bound in Eqn. 2 is worse than the bound in Hoeffding's theorem, since it involves additional term $t/(K\|a\|_\infty)$ in the exponent.

Proof. We apply a similar strategy as in the proof for Hoeffding's Inequality.

$$\mathbb{P}\left(\left|\sum_{i=1}^n a_i X_i\right| \geq t\right) \leq e^{-\lambda t} \prod_{i=1}^n \mathbb{E} e^{a_i \lambda X_i}.$$

If $|\lambda| \leq \frac{C}{\|a\|_\infty K}$, then by (c),

$$\mathbb{P}\left(\left|\sum_{i=1}^n a_i X_i\right| \geq t\right) \leq 2 \exp(-\lambda t + C\lambda^2 K^2 \|a\|^2).$$

We only need to optimize $-\lambda t + C\lambda^2 K^2 \|a\|^2$ over $|\lambda| \leq \frac{C}{\|a\|_\infty K}$. Some calculation yields the desired upper bound. $\square$

Theorem 5 (Bernstein's Inequality for bounded random variables). *Let $X_1, \dots, X_n$ be independent random variables, such that $\mathbb{E}X_i = 0$ and $\max_{1 \leq i \leq n} |X_i| \leq M$ a.s.. Then*

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| \geq t\right) \leq 2 \exp\left(-\frac{t^2/2}{\sigma^2 + Mt/3}\right),$$

where $\sigma^2 = \sum_{i=1}^n \mathbb{E}X_i^2$ is total variance.

Note that

$$\max\{\sigma^2, Mt/3\} \leq \sigma^2 + Mt/3 \leq 2 \max\{\sigma^2, Mt/3\},$$

so the exponent in the upper bound of Theorem 5 is roughly (up to a constant²)

$$-\frac{t^2}{\max\{\sigma^2, Mt\}} = -\min\left\{\frac{t^2}{\sigma^2}, \frac{t}{M}\right\}$$

This is similar to the exponent in the bound in (2)

Remark 1. *There is a popular variant of Bernstein's inequality.*

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| \geq \frac{C(\sigma + M)}{\delta}\right) \leq 2 \exp\left(-\min\left(\frac{1}{\delta}, \frac{1}{\delta^2}\right)\right), \quad \text{for any } \delta > 0, \quad (3)$$

where C is a certain constant. This inequality is a consequence of Theorem 5. In fact, note that

$$\frac{-t^2/2}{\sigma + Mt/3} \leq \frac{-t^2}{2 \max\{\sigma^2, Mt/3\}} = -\min\left\{\frac{t^2}{4\sigma^2}, \frac{3t}{4M}\right\}.$$

So choosing $t = \frac{C(\sigma+M)}{\delta}$ with a sufficiently large C leads to

$$\frac{t^2}{4\sigma^2} \geq \frac{1}{\delta^2}, \quad \frac{3t}{4M} \geq \frac{1}{\delta}$$

and therefore leads to the upper bound in Eqn. 3.

4 Understanding the two tail probabilities

Let $X_1, X_2, \dots, X_n$ be independent random variables, and $a = (a_1, \dots, a_n)$ a nonrandom vector. We are interested in the tail probability of the sum $S_n = \sum_{i=1}^n a_i X_i$. The variance of S_n is $\text{Var}(S_n) = \|a\|^2$.

²When we say two quantities s and t are up to a constant, we mean $ct \leq s \leq Ct$ for absolute constants $c, C > 0$.

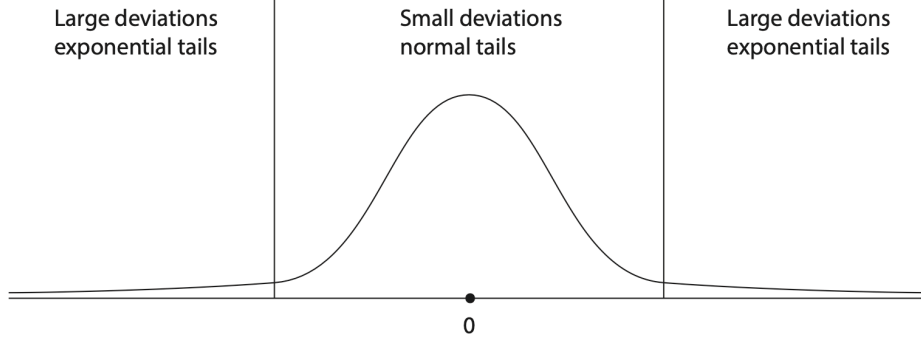


Figure 2.3 Bernstein's inequality for a sum of sub-exponential random variables gives a mixture of two tails: sub-gaussian for small deviations and sub-exponential for large deviations.

1. **Gaussian distribution.** In the canonical case $X_i \stackrel{d}{=} N(0, 1)$, we have $S_n \stackrel{d}{=} N(0, \|a\|^2)$, so the distribution of S_n is completely determined by its variance. According to Proposition 1, the tail probability is mainly determined by $e^{-t^2/(2\|a\|^2)}$ (ignoring poly-factor).
2. **Asymptotic normal distribution.** Under certain conditions we expect the asymptotic normality $S_n/\|a\| \xrightarrow{d} N(0, 1)$, so approximately $S_n \stackrel{d}{\approx} N(0, \|a\|^2)$. However, as we explained in Section 1, CLT is informative about “bulk” probability but not tail probability.
3. **Subgaussian distribution.** If the tail probability of each individual X_i is “lighter” (in the sense of Theorem 1(i)) than the standard Gaussian tail, then we still get a “light” tail controlled by $e^{-ct^2/(2\|a\|^2)}$ (up to constant c in the exponent).
4. **Subexponential distribution.** If the tail of X_i is slightly heavier than Gaussian, then S_n shows a mixture of two tails in Eqn. 2—one is Gaussian $e^{-ct^2/(2\|a\|^2)}$ and another is exponential $e^{-ct/\|a\|_\infty}$. The tail probability is controlled by the Gaussian tail when t is moderately large and by the exponential tail when t is very large. We cannot control the exponential tail: consider $a = (1, 0, \dots, 0)$ and $X_1 \stackrel{d}{=} \text{Exp}(1)$, where the tail probability is purely exponential.

The discussion about subexponential distribution suggests two scales of S_n : (i) one is the standard deviation $\|a\|$ of S_n which is already expected from CLT, (ii) another is $\|a\|_\infty$ which cannot be controlled by Gaussian tail due to the heavier tail of individual X_i . Usually, the first tail plays the dominant role at small deviation (moderate t), and the second at large deviation (large t). See Figure 2.3.

Theorem 5 controls the tail probability from a slightly different way: first using a variance quantity σ^2 similar to $\|a\|^2$ above, second using a crude bound M on X_i (trivially leads to subexponential norm M) to control the potentially “heavier” tail of X_i .

High-probability bound. Finally, to get a high-probability bound on S_n , we conclude that $|S_n|$ is bounded by the following quantity with probability at least $1 - \delta$.

1. Gaussian or subgaussian case: $O(\sqrt{\log(1/\delta)} \sigma)$ where σ is the variance of S_n .
2. Subexponential case: $O(\log(1/\delta)(\sigma + M))$ by Eqn. 3.

5 Further reading

The recent book [1] on high-dimensional probability provides an excellent introduction to this topic. The materials (esp. Figure 2.3) of this lecture are mostly taken or adapted from Section 2 of this book.

References

- [1] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.