

Lecture 13: Concentration Inequalities

February 21, 2025

Lecturer: Lei Wu

Scribe: Lei Wu

Abstract

Concentration phenomena and inequalities play a pivotal role in high-dimensional probability, as they provide powerful tools for understanding the behavior of random variables and their aggregates in high-dimensional spaces. These tools allow us to bound the deviations of random variables from their expectations, ensuring that the aggregate behavior remains predictable even in complex systems. This is particularly critical in modern applications such as machine learning, statistical inference, and optimization, where datasets and models often involve thousands or millions of dimensions. Great textbooks on this topic include:

- [High-Dimensional Probability](#) by Roman Vershynin.
- [Probability in High Dimension](#) by Roman van Handel.

For more details related to the materials covered in this lecture, we specifically refer to [[Vershynin, 2018](#), Section 2].

1 Introduction

Let X_1, \dots, X_n be i.i.d. random variables with expectation μ . Then,

$$\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n X_i \right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] = \mu.$$

We are interested in when *the empirical mean $\frac{1}{n} \sum_{i=1}^n X_i$ will concentrate around the population mean μ .*

- What conditions are required for the random variable X_i ?
- What does the “concentration” mean?

Let first review two classical results in standard probability theory textbook.

Theorem 1.1 (Strong law of large numbers (LLN)). *Let X_1, \dots, X_n be a sequence of i.i.d. random variables with expectation μ . Then,*

$$\frac{1}{n} \sum_{i=1}^n X_i \rightarrow \mu \quad \text{almost surely.}$$

LLN shows that as long as the expectation μ is finite, the empirical mean will converge to μ . In other words, as long as we have sufficient samples, $\frac{1}{n} \sum_{i=1}^n X_i$ will always concentrate at μ . Unfortunately, the rate of “concentration” in LLN can be arbitrarily slow. The next theorem, the central limit theorem, makes one step further shows that if the second moment is finite, the deviation should be on the order of $O(1/\sqrt{n})$.

Theorem 1.2 (Central limit theorem (CLT)). *Let X_1, \dots, X_n be a sequence of i.i.d. random variables with mean μ and variance σ^2 . Then,*

$$\sqrt{n} \left(\frac{X_1 + X_2 + \dots + X_n}{n} - \mu \right) \rightarrow \mathcal{N}(0, \sigma^2) \quad \text{in distribution.}$$

CLT implies that $\frac{1}{n} \sum_{i=1}^n X_i \approx \mu + \frac{\sigma}{\sqrt{n}} Z$, where Z is the standard normal random variable. Thus, the CLT provides a precise characterization of how the empirical mean deviates from the population mean μ when the deviation is on the order of $1/\sqrt{n}$. In other words, it delivers a strong, yet asymptotic, understanding of the distribution of small deviations. However, if we are interested in large deviations—those whose magnitude does not diminish as n grows—or in non-asymptotic descriptions, the CLT alone is not sufficient.

2 Linear Concentration

Moment-based concentration. By Chebyshev's inequality,

$$\mathbb{P} \left\{ \left| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right| \geq t \right\} = \mathbb{P} \left\{ \left| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right|^2 \geq t^2 \right\} \leq \frac{\mathbb{E} \left[\left| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right|^2 \right]}{t^2} \leq \frac{\sigma^2}{nt^2}.$$

This probability of having large deviations is in the order of $O(1/n)$.

However, on the other hand, from CLT, we “anticipate” that

$$\begin{aligned} \mathbb{P} \left\{ \left| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right| \geq t \right\} &\approx \mathbb{P} \left\{ \left| \frac{\sigma Z}{\sqrt{n}} \right| \geq t \right\} = 2\mathbb{P} \left\{ Z \geq \frac{\sqrt{nt}}{\sigma} \right\} \\ &= \sqrt{\frac{2}{\pi}} \int_{\frac{\sqrt{nt}}{\sigma}}^{\infty} e^{-\frac{x^2}{2}} dx \lesssim e^{-\frac{1}{2} \left(\frac{\sqrt{nt}}{\sigma} \right)^2} = e^{-\frac{nt}{2\sigma^2}}. \end{aligned} \quad (1)$$

This suggests that the tail can decay exponentially fast, which is much stronger than the one provided by Chebyshev's inequality. Unfortunately, this calculation is not correct since $\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i - \mu \rightarrow \sigma Z$ can be arbitrarily slow. Therefore, we need to control somethings stronger than the second-order moments.

Let us first look at a simple example.

Theorem 2.1 (Hoeffding's inequality). *Let X_1, \dots, X_n be i.i.d. symmetric Bernoulli random variable, i.e., $\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = \frac{1}{2}$. Then,*

$$\mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n X_i \geq t \right\} \lesssim e^{-\frac{nt}{2}}.$$

Proof. We have

$$\begin{aligned} \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n X_i \geq t \right\} &= \mathbb{P} \left\{ e^{\lambda \sum_{i=1}^n X_i} \geq e^{n\lambda t} \right\} \leq \frac{\mathbb{E}[e^{\lambda \sum_{i=1}^n X_i}]}{e^{n\lambda t}} \\ &= e^{-n\lambda t} \prod_{i=1}^n \mathbb{E}[e^{\lambda X_i}] = e^{-n\lambda t + n\psi(\lambda)}, \end{aligned} \quad (2)$$

where

$$\psi(\lambda) = \log \mathbb{E}[e^{\lambda X}] = \log\left(\frac{e^\lambda + e^{-\lambda}}{2}\right) \leq \lambda^2/2. \quad (3)$$

Plugging it into (2), we have

$$\mathbb{P}\left\{\frac{1}{n} \sum_{i=1}^n X_i \geq t\right\} \leq \inf_{\lambda > 0} e^{-n\lambda t + n\psi(\lambda)} = \inf_{\lambda} e^{-n(\lambda t - \lambda^2/2)} = e^{-nt^2/2}.$$

□

Remark 2.2. The above approach is often referred as the *Chernoff-Cramer method*.

From the proof, we can see that the key ingredient is the log-moment generating function (log-MGF):

$$\psi(\lambda) = \log \mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}] \quad (4)$$

and the Legendre conjugate of the log-MGF:

$$\psi^*(t) = \sup_{\lambda > 0} \{\lambda t - \psi(\lambda)\}. \quad (5)$$

Lemma 2.3. *If X has a log-MGF ψ with the Legendre dual ψ^* , then*

$$\mathbb{P}\{X - \mathbb{E}[X] \geq t\} \leq e^{-\psi^*(t)}.$$

Let X_1, \dots, X_n be i.i.d. random variable. Then,

$$\mathbb{P}\left\{\left|\frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}[X]\right| \geq t\right\} \leq 2e^{-n\psi^*(t)}.$$

The above lemma implies that $\psi^*(t)$ controls the rate of concentration.

Definition 2.4 (sub-Gaussian). A random variable X is said to be sub-Gaussian with variance proxy σ^2 if $\psi(\lambda) \leq \frac{\lambda^2 \sigma^2}{2}$.

The sub-Gaussian assumption implies that

$$\psi^*(t) = \sup_{\lambda > 0} \{\lambda t - \psi(\lambda)\} \geq \sup_{\lambda > 0} \left\{ \lambda t - \frac{\lambda^2 \sigma^2}{2} \right\} = \frac{t^2}{2\sigma^2}.$$

By Lemma 2.3, the tail of X satisfies

$$\mathbb{P}\{|X - \mathbb{E}[X]| \geq t\} \leq 2e^{-\frac{t^2}{2\sigma^2}}, \quad (6)$$

which is similar to the tail of Gaussian. In fact, the tail estimate (6) is often used as the equivalent definition of the sub-Gaussian class.

Lemma 2.5. *If the tail behavior of X satisfies*

$$\mathbb{P}\{|X| \geq t\} \leq C_1 e^{-C_2 t^2} \text{ for all } t \geq 0. \quad (7)$$

Then, $\varphi(\lambda) \leq K_1 \lambda^2$ for some constant K_1 .

Proof. With loss of generality, we consider only the case of $\lambda \geq 0$. Then, we have

$$\begin{aligned}
\mathbb{E}[e^{\lambda X}] &= \int_0^\infty \mathbb{P}\{e^{\lambda X} \geq t\} dt \\
&= \int_{-\infty}^\infty \mathbb{P}\{e^{\lambda X} \geq e^{\lambda s}\} \lambda e^{\lambda s} ds \quad (t = e^{\lambda s}) \\
&= \lambda \left(\int_{-\infty}^0 \mathbb{P}\{X \geq s\} e^{\lambda s} ds + \int_0^\infty \mathbb{P}\{X \geq s\} e^{\lambda s} ds \right) \\
&\leq \lambda \left(\int_{-\infty}^0 e^{\lambda s} ds + 2 \int_0^\infty e^{-C_1 t^2 + \lambda s} ds \right) \\
&\leq 1 + C_1 \lambda e^{K\lambda^2} \\
&\leq e^{K_1 \lambda^2},
\end{aligned}$$

where C, K, K_1 are some absolute positive constants. □

Corollary 2.6 (Chernoff's bound). *Let X_1, \dots, X_n be i.i.d. sub-Gaussian random variables with mean μ and variance proxy σ^2 . Then*

$$\mathbb{P}\left\{\left|\frac{1}{n} \sum_{i=1}^n X_i - \mu\right| \geq t\right\} \leq 2e^{-\frac{nt^2}{2\sigma^2}}.$$

By Lemma 6, we can conclude that as long as each random variable has a sub-Gaussian tail, we have $P(|\frac{1}{n} \sum_i X_i - \mu| \geq t) \leq 2e^{-K_1 nt^2}$ for some constant K_1 .

Examples:

- **Gaussian RV:** For $g \sim \mathcal{N}(0, 1)$, its tail behavior satisfies [Vershynin, 2018, Proposition 2.1.2]

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

- **Bounded RV:** Bounded random variables obviously satisfy the tail behavior (7). Specifically, the following lemma provides a tight estimate of the variance proxy.

Lemma 2.7 (Hoeffding's lemma). *Assume $a \leq X \leq b$. Then, $\psi(\lambda) \leq \lambda^2(b-a)^2/8$.*

Proof. WLOG, assume that $\mathbb{E}[X] = 0$. Recall that $\psi(\lambda) = \log \mathbb{E}[e^{\lambda X}]$. Then,

$$\psi'(\lambda) = \frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]}, \quad \psi''(\lambda) = \frac{\mathbb{E}[X^2 e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} - \left(\frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]}\right)^2.$$

Let \mathbb{Q} denote the distribution with $\frac{d\mathbb{Q}}{d\mathbb{P}} = e^{\lambda X} / \mathbb{E}[e^{\lambda X}]$. Then, we can rewrite the second-order derivative as $\text{Var}_{\mathbb{Q}}[X]$. Since $X \in [a, b]$, we have

$$\text{Var}_{\mathbb{Q}}[X] = \mathbb{E}_{\mathbb{Q}}[|X - \mathbb{E}_{\mathbb{Q}}[X]|^2] \leq \mathbb{E}_{\mathbb{Q}}\left[\left|X - \frac{b+a}{2}\right|^2\right] \leq \mathbb{E}_{\mathbb{Q}}\left[\left|\frac{b-a}{2}\right|^2\right] = \frac{(b-a)^2}{4},$$

where the second step use the fact that $\mathbb{E}[Z] = \operatorname{argmin}_{\mu} \mathbb{E}[(Z - \mu)^2]$ for any random variable Z . Hence,

$$\psi(0) = 0, \quad \psi'(0) = \mathbb{E}[X] = 0, \quad \psi''(\lambda) \leq \frac{(b-a)^2}{4},$$

which implies

$$\begin{aligned} \psi(\lambda) &= \psi(0) + \int_0^\lambda \psi'(t) dt \\ &= \psi(0) + \int_0^\lambda \left(\psi'(0) + \int_0^t \psi''(s) ds \right) dt \\ &= \psi(0) + \psi'(0)\lambda + \int_0^\lambda dt \int_0^s \psi''(s) ds \leq \frac{(b-a)^2 \lambda^2}{8}. \end{aligned}$$

□

Remark 2.8. The Hoeffding's lemma is sharp when X is the symmetric Bernoulli distribution, i.e., $\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = 1/2$. See Eq. (3).

Corollary 2.9 (Hoeffding's inequality). *Let X_1, \dots, X_n be i.i.d. random variables. If $a \leq X_i \leq b$, then,*

$$\mathbb{P} \left\{ \left| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right| \geq t \right\} \leq 2e^{-\frac{2nt^2}{(b-a)^2}}.$$

3 Nonlinear Concentration

Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be a (nonlinear) function and consider the following concentration:

$$f(X_1, \dots, X_n) \approx \mathbb{E}[f(X_1, \dots, X_n)] \quad \text{with high probability?}$$

The preceding results correspond to $f(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n x_i$. Can we extend it to nonlinear functions?

- If f only depends on one coordinate, we can not anticipate any concentration.
- If f is equally robust to small changes for all coordinates, we anticipate that this case will behave like the empirical mean.

Theorem 3.1 (McDiarmid's inequality). *Let f be a function satisfying the Bounded Difference Property: there exist constants $\{L_i\}_{i=1}^n$ such that the following holds for all $i \in [n]$*

$$\sup_{x_1, \dots, x_n, \tilde{x}_i} |f(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n) - f(x_1, \dots, x_{i-1}, \tilde{x}_i, x_{i+1}, \dots, x_n)| \leq L_i$$

Let $\sigma^2 := \frac{1}{4} \sum_{i=1}^n L_i^2$ and X_1, \dots, X_n are i.i.d.. Then, $Z = f(X_1, \dots, X_n)$ is sub-Gaussian with variance proxy σ :

$$\mathbb{P}\{|Z - \mathbb{E}[Z]| \geq t\} \leq 2e^{-\frac{t^2}{2\sigma^2}}.$$

One can think L_i as a measure of the sensitivity of f to the i -th coordinates. For the case of empirical mean, $D_i f(x) = O(1/n)$ for every i . This recovers the Hoeffding's inequality (Corollary 2.9). Thus, we can view McDiarmid's inequality as a nonlinear version of Hoeffding's inequality. Question: Is there a similar nonlinear Chernoff's inequality?

The proof needs following lemmas.

Lemma 3.2 (Azuma's lemma). *Let $\{\mathcal{F}_i\}_{i=1}^n$ be a filtration. Assume σ_i to be positive constants and $\{\Delta_i\}$ random variables such that*

1. $\mathbb{E}[\Delta_i | \mathcal{F}_{i-1}] = 0$ (Martingale difference property).
2. $\log \mathbb{E}[e^{\lambda \Delta_i} | \mathcal{F}_{i-1}] \leq \frac{\lambda^2 \sigma_i^2}{2}$ (Conditional sub-Gaussian property).

Then, $\sum_{i=1}^n \Delta_i$ is sub-Gaussian with the proxy variance $\sum_{i=1}^n \sigma_i^2$.

Proof. This time, we do not have the independence. Instead, we can exploit the conditional independence, i.e., the martingale property. Consider the condition on the filtration

$$\begin{aligned} \mathbb{E} \left[e^{\lambda \sum_{i=1}^n \Delta_i} \right] &= \mathbb{E} \left[\mathbb{E} \left[e^{\lambda \sum_{i=1}^n \Delta_i} | \mathcal{F}_{n-1} \right] \right] \\ &= \mathbb{E} \left[e^{\lambda \sum_{i=1}^{n-1} \Delta_i} \mathbb{E} \left[e^{\lambda \Delta_n} | \mathcal{F}_{n-1} \right] \right] \leq e^{\frac{\lambda^2 \sigma_n^2}{2}} \mathbb{E} \left[e^{\lambda \sum_{i=1}^{n-1} \Delta_i} \right] \end{aligned}$$

By induction, we conclude that

$$\mathbb{E} \left[e^{\lambda \sum_{i=1}^n \Delta_i} \right] \leq e^{\frac{\lambda^2 \sum_{i=1}^n \sigma_i^2}{2}}.$$

This means $\sum_{i=1}^n \Delta_i$ is sub-Gaussian with the proxy variance $\sum_{i=1}^n \sigma_i^2$. \square

Lemma 3.3 (Azuma-Hoeffding's inequality). *Under the assumption of Lemma 3.2, assume $A_i \leq \Delta_i \leq B_i$ almost surely and A_i, B_i are \mathcal{F}_{i-1} -measurable. Then, $\sum_{i=1}^n \Delta_i$ is sub-Gaussian with the proxy variance $\sigma^2 = \frac{1}{4} \sum_{i=1}^n \|B_i - A_i\|_{L^\infty}^2$. In particular,*

$$\mathbb{P} \left\{ \left| \sum_{i=1}^n \Delta_i \right| \geq t \right\} \leq 2e^{-\frac{t^2}{2\sigma^2}}.$$

Proof. Combining Lemma 2.3, 2.7 and 3.2, we complete the proof. \square

Proof of McDiarmid's inequality. To analyze the behavior of $f(X_1, \dots, X_n)$, consider the following decomposition

$$\begin{aligned} f(X) - \mathbb{E}[f(X)] &= f(X) - \mathbb{E}[f(X) | X_1, \dots, X_{n-1}] \\ &\quad + \mathbb{E}[f(X) | X_1, \dots, X_{n-1}] - \mathbb{E}[f(X) | X_1, \dots, X_{n-2}] \\ &\quad + \dots + \mathbb{E}[f(X) | X_1] - \mathbb{E}[f(X)] \\ &= \sum_{i=1}^n \Delta_i, \end{aligned} \tag{8}$$

where $\Delta_i = \mathbb{E}[f(X) | X_1, \dots, X_i] - \mathbb{E}[f(X) | X_1, \dots, X_{i-1}]$. Let $\mathcal{F}_i = \sigma(X_1, \dots, X_i)$. Then, $\mathbb{E}[\Delta_i | \mathcal{F}_{i-1}] = 0$ and

$$\Delta_i = \mathbb{E} \left[\mathbb{E}[f(X_1, \dots, X_i, \dots, X_n) | X_i] - f(X) | X_1, \dots, X_{i-1} \right].$$

Let

$$A_i = \mathbb{E}[\inf_{\alpha} f(X_1, \dots, X_{i-1}, \alpha, X_{i+1}, \dots, X_n) - f(X_1, \dots, X_n) | X_1, \dots, X_{i-1}]$$

$$B_i = \mathbb{E}[\sup_{\alpha} f(X_1, \dots, X_{i-1}, \alpha, X_{i+1}, \dots, X_n) - f(X_1, \dots, X_n) | X_1, \dots, X_{i-1}]$$

By the assumption of f , it is easy to verify that

$$A_i \leq \Delta_i \leq B_i, \quad |B_i - A_i| \leq \|D_i f\|_{L^\infty}.$$

Using the Azuma-Hoeffding lemma, $f(X) - \mathbb{E}[f(X)]$ is a sub-Gaussian with the variance proxy $\sigma^2 = \frac{1}{4} \sum_{i=1}^n \|D_i f\|_{L^\infty}^2$. This directly implies that

$$\mathbb{P}\{|f(X) - \mathbb{E}[f(X)]| \geq t\} \leq 2e^{-\frac{2}{\sum_{i=1}^n \|D_i f\|_{L^\infty}^2} t^2}.$$

Thus, we complete the proof. \square

4 Maximal Inequality

Lemma 4.1 (Maximal inequality). *Assume that X_1, \dots, X_n be n sub-Gaussian random variables with zero mean and the variance proxy σ^2 . Then,*

$$\mathbb{E}[\max_{i \in [n]} X_i] \leq \sigma \sqrt{2 \log n}.$$

Proof. Recalling the LogSumExp trick we introduced in Lecture 3, we have for any $\beta > 0$:

$$\max_{i \in [n]} X_i \leq \frac{1}{\lambda} \log \sum_{i=1}^n e^{\lambda X_i}.$$

For any $\lambda > 0$,

$$\begin{aligned} \mathbb{E}[\max_{i \in [n]} X_i] &\leq \frac{1}{\lambda} \mathbb{E}[\log \sum_{i=1}^n e^{\lambda X_i}] \\ &\leq \frac{1}{\lambda} \log \sum_{i=1}^n \mathbb{E}[e^{\lambda X_i}] \quad (\text{Jensen's inequality}) \\ &\leq \frac{1}{\lambda} \log \sum_{i=1}^n e^{\frac{\sigma^2 \lambda^2}{2}} = \frac{\log n}{\lambda} + \frac{\sigma^2 \lambda}{2}. \end{aligned}$$

Taking $\lambda = \sqrt{2 \log(n)/\sigma^2}$ completes the proof. \square

Note that in the maximal inequality, we do not assume that X_1, \dots, X_n are independent. In fact, the bound in Lemma 4.1 is sharp.

Lemma 4.2. *Let X_1, \dots, X_n be independent $\mathcal{N}(0, 1)$ random variables. Then,*

$$\mathbb{E} \max_{i \in [n]} X_i \geq c \sqrt{\log n}.$$

Next, we turn to provide high-probability bound.

Lemma 4.3. *Let X_1, \dots, X_n be sub-Gaussian random variables with zero mean and variance proxy σ^2 . Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, it holds that*

$$\max_{i \in [n]} X_i \leq \sigma \sqrt{2 \log\left(\frac{2n}{\delta}\right)}.$$

Proof. By union bound, we have

$$\mathbb{P}\{\max_i X_i \geq t\} \leq \sum_{i=1}^n \mathbb{P}\{|X_i| \geq t\} \leq 2ne^{-t/(2\sigma^2)}.$$

For this probability to be smaller than δ , we require $2ne^{-t/(2\sigma^2)} \leq \delta$. It gives $t \geq \sqrt{2\sigma^2 \log(2n/\delta)}$. \square

References

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