**Topics in Deep Learning Theory (Spring 2025)** 

### Lecture 0: Concentration Inequalities

Instructor: Lei Wu

Date: May 16, 2025

#### 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  $X_1, \ldots, X_n$  be i.i.d. random variables with expectation  $\mu$ . 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  $\mu$ .

- 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  $X_1, \ldots, X_n$  be a sequence of *i.i.d.* random variables with expectation  $\mu$ . Then,

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

LLN shows that as long as the expectation  $\mu$  is finite, the empirical mean will converge to  $\mu$ . In other words, as long as we have sufficient samples,  $\frac{1}{n} \sum_{i=1}^{n} X_i$  will always concentrate at  $\mu$ . 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  $X_1, \ldots, X_n$  be a sequence of i.i.d. random variables with mean  $\mu$  and variance  $\sigma^2$ . Then,

$$\sqrt{n}\left(\frac{X_1+X_2+\cdots+X_n}{n}-\mu\right) \to \mathcal{N}(0,\sigma^2)$$
 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  $\mu$  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

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}X_{i}-\mu\right|\geq t\right\}\approx\mathbb{P}\left\{\left|\frac{\sigma Z}{\sqrt{n}}\right|\geq t\right\}=2\mathbb{P}\left\{Z\geq\frac{\sqrt{n}t}{\sigma}\right\}$$
$$=\sqrt{\frac{2}{\pi}}\int_{\frac{\sqrt{n}t}{\sigma}}^{\infty}e^{-\frac{x^{2}}{2}}\,\mathrm{d}x\lesssim e^{-\frac{1}{2}(\frac{\sqrt{n}t}{\sigma})^{2}}=e^{-\frac{nt^{2}}{2\sigma^{2}}}.$$
(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  $X_1, \ldots, X_n$  be i.i.d. symmetric Bernoulli random variable, i.e.,  $\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = \frac{1}{2}$ . Then,

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

Proof. We have

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

where

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

Plugging it into (2), we have

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

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):

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

and the Legendre conjugate of the log-MGF:

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

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

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

Let  $X_1, \ldots, X_n$  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  $\sigma^2$  if  $\psi(\lambda) \leq \frac{\lambda^2 \sigma^2}{2}$ .

The sub-Gaussian assumption implies that

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

By Lemma 2.3, the tail of X satisfies

$$\mathbb{P}\{|X - \mathbb{E}[X]| \ge t\} \le 2e^{-\frac{t^2}{2\sigma^2}},\tag{6}$$

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

$$\mathbb{P}\left\{|X| \ge t\right\} \le C_1 e^{-C_2 t^2} \text{ for all } t \ge 0.$$

$$\tag{7}$$

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

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

$$\begin{split} \mathbb{E}[e^{\lambda X}] &= \int_0^\infty \mathbb{P}\left\{e^{\lambda X} \ge t\right\} \mathrm{d}t \\ &= \int_{-\infty}^\infty \mathbb{P}\left\{e^{\lambda X} \ge e^{\lambda s}\right\} \lambda e^{\lambda s} \mathrm{d}s \qquad (t = e^{\lambda s}) \\ &= \lambda \left(\int_{-\infty}^0 \mathbb{P}\left\{X \ge s\right\} e^{\lambda s} \mathrm{d}s + \int_0^\infty \mathbb{P}\left\{X \ge s\right\} e^{\lambda s} \mathrm{d}s\right) \\ &\leq \lambda \left(\int_{-\infty}^0 e^{\lambda s} \mathrm{d}s + 2\int_0^\infty e^{-C_1 t^2 + \lambda s} \mathrm{d}s\right) \\ &\leq 1 + C_1 \lambda e^{K\lambda^2} \\ &< e^{K_1\lambda^2}, \end{split}$$

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

**Corollary 2.6** (Chernoff's bound). Let  $X_1, \ldots, X_n$  be i.i.d. sub-Gaussian random variables with mean  $\mu$  and variance proxy  $\sigma^2$ . 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]

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

• **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  $a \le X \le b$ . Then,  $\psi(\lambda) \le \lambda^2 (b-a)^2/8$ .

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

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

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  $\operatorname{Var}_Q[X]$ . Since  $X \in [a, b]$ , we have

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

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,

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

which implies

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

**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  $X_1, \ldots, X_n$  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:

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

The preceding results correspond to  $f(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^n x_i$ . 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  $i \in [n]$ 

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

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

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

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 viewed 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  $\sigma_i$  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

$$\mathbb{E}\left[e^{\lambda\sum_{i=1}^{n}\Delta_{i}}\right] = \mathbb{E}\left[\mathbb{E}\left[e^{\lambda\sum_{i=1}^{n}\Delta_{i}}|\mathcal{F}_{n-1}\right]\right]$$
$$\mathbb{E}\left[e^{\lambda\sum_{i=1}^{n-1}\Delta_{i}}\mathbb{E}\left[e^{\lambda\Delta_{n}}|\mathcal{F}_{n-1}\right]\right] \le e^{\frac{\lambda^{2}\sigma_{n}^{2}}{2}}\mathbb{E}\left[e^{\lambda\sum_{i=1}^{n-1}\Delta_{i}}\right]$$

By induction, we conclude that

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

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

**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  $\sigma^2 = \frac{1}{4} \sum_{i=1}^n \|B_i - A_i\|_{L^{\infty}}$ . 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.

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

$$f(X) - \mathbb{E}[f(X)] = f(X) - \mathbb{E}[f(X)|X_1, \dots, X_{n-1}] + \mathbb{E}[f(X)|X_1, \dots, X_{n-1}] - \mathbb{E}[f(X)|X_1, \dots, X_{n-2}] + \dots + \mathbb{E}[f(X)|X_1] - \mathbb{E}[f(X)] = \sum_{i=1}^n \Delta_i,$$
(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

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

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

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

Thus, we complete the proof.

# **4** Maximal Inequality

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

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

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

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

For any  $\lambda > 0$ ,

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

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

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

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

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

Next, we turn to provide high-probability bound.

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

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

Proof. By union bound, we have

$$\mathbb{P}\{\max_{i} X_i \ge t\} \le \sum_{i=1}^{n} \mathbb{P}\{|X_i| \ge t\} \le 2ne^{-t/(2\sigma^2)}.$$

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

# References

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