Mathematical Introduction to Machine Learning

Lecture 13: Concentration inequalties

December 7, 2023

Lecturer: Lei Wu *Scribe: Lei Wu*

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

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

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

- What conditions are required for the random variable X_i?
- What does the "concentration" means?

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

Theorem 0.1 (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,*

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

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 0.2 (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,*

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

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, it provides a precise characterization how the empirical mean deviates from the population mean μ when the deviation is in the order of $1/\sqrt{n}$. CLT is strong in the sense that it provide a precise (bu asymptotic) characterization of the whole distribution of (small) deviations However, it is also not sufficient if we are interested in "large deviations", whose magnitudes do not depend on n?

1 Linear Concentration

By Chebyshev's inequality,

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

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

On the other hand, from CLT, we "anticipate" that

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

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 $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i - \mu \to \sigma Z$ 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 1.1 (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,*

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

Proof. We have

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

where

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

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 1.2. 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:

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

Lemma 1.3. *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 $X_1, \ldots, X_n$ be i.i.d. random variable. Then,

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

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

Definition 1.4 (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

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

By Lemma 1.3, the tail of X satisfies

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

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 1.5. *If the tail behavior of X satisfies*

$$\mathbb{P}\left\{|X| \geq t\right\} \leq 2e^{-C_1 t} \textit{ for all } t \geq 0. \tag{7}$$

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

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

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

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

□

Corollary 1.6 (Chernoff bound)**.** *Let $X_1, \dots, X_n$ be i.i.d. sub-Gaussian random variables with mean μ and variance proxy σ^2. Then*

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

By Lemma 6, we can conclude that as long as each random variable has a sub-Gaussian tail, we have $P(|\frac{1}{n}\sum_i X_i - \mu| \geq t) \leq 2e^{-K_1 t^2}$ 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 1.7 (Hoffding'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,

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

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

$$\mathrm{Var}_{\mathbb{Q}}[X] = \mathbb{E}_{\mathbb{Q}}[|X - \mathbb{E}_X|^2] \leq \mathbb{E}_{\mathbb{Q}}[|X - \frac{b-a}{2}|^2] \leq \mathbb{E}_{\mathbb{Q}}[|\frac{b-a}{2}|^2] = \frac{(b-a)^2}{4}.$$

Hence,

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

which implies

$$\psi(\lambda) = \psi(0) + \int_0^{\lambda} \int_0^{s} \psi''(s)\,\mathrm{d}s \leq \frac{(b-a)^2 \lambda^2}{8}.$$

□

Remark 1.8. 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 1.9 (Hoeffding's inequality)**.** *Let $X_1, \dots, X_n$ be i.i.d. random variables. If $a \leq X_i \leq b$, then,*

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

2 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 2.1 (McDiarmid's inequality)**.** *Define*

$$D_i f(x) = \sup_\alpha f(x_1, \dots, x_{i-1}, \alpha, x_{i+1}, \dots, x_n) - \inf_\alpha f(x_1, \dots, x_{i-1}, \alpha, x_{i+1}, \dots, x_n).$$

Assume that D_i is bounded for all i and let $\sigma^2 := \frac{1}{4} \sum_{i=1}^n \|D_i f\|_{L^\infty}^2$. Then,

$$\mathbb{P}\{|f(X_1, \dots, X_n) - \mathbb{E}[f]| \geq t\} \leq 2e^{-\frac{t^2}{2\sigma^2}}.$$

One can think $D_i f(x)$ 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 1.9). 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 2.2 (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

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

By induction, we conclude that

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

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

Lemma 2.3 (Azuma-Hoeffding's inequality)**.** *Under the assumption of Lemma 2.2, assume $A_i \le \Delta_i \le B_i$ 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,*

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

Proof. Combining Lemma 1.3, 1.7 and 2.2, we complete the proof. □

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

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

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

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

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. □

3 Maximal Inequality

Lemma 3.1 (Maximal inequality)**.** *Assume that $X_1, \dots, X_n$ be n sub-Gaussian random variables with the variance proxy σ^2. Then,*

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

Proof. Recall the LogSumExp trick we introduced in Lecture 3:

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

For any $\lambda > 0$,

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

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

Note that in the maximal inequality, we do not assume that $X_1, \ldots, X_n$ are independent. In fact, the bound in Lemma 3.1 is sharp.

Lemma 3.2. *Let* $\boldsymbol{X}_1, \ldots, X_n$ *be independent* $\mathcal{N}(0,1)$ *random variables. Then,*

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

References

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