Theoretical Deep Learning

Lecture 2: Concentration Inequalities

July 25, 2021

Lecturer: Lei Wu *Scribe: Lei Wu*

If $X_1, \dots, X_n$ are 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. Specifically:

- 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 mean is finite, the empirical mean will converge. 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 convergence is guaranteed with the rate 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,*

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

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 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 Concentration Inequalities

By Chebysheff's inequality,

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

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

On the other hand, from CLT, we expect that

$$\mathbb{P}\{|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu| \geq t\} \approx \mathbb{P}\{|\frac{\sigma Z}{\sqrt{n}}| \geq t\} = 2\mathbb{P}\{Z \geq \frac{\sqrt{n}t}{\sigma}\}$$

$$= \sqrt{\frac{2}{\pi}} \int_{\frac{\sqrt{n}t}{\sigma}}^{\infty} e^{-\frac{x^2}{2}}\,\mathrm{d}x \lesssim e^{-\frac{1}{2}(\frac{\sqrt{n}t}{\sigma})^2} = e^{-\frac{nt^2}{2\sigma^2}}. \tag{1.1}$$

This suggests that the tail can decay exponentially fast, which is much stronger than the one provided by Chebysheff'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 variance/second moments.

Let us first look at a sample 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,*

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

Proof. We have

$$\mathbb{P}\{\frac{1}{n}\sum_{i=1}^{n} X_i \geq t\} = \mathbb{P}\{e^{\lambda\sum_{i=1}^{n}} \geq e^{n\lambda t}\} \leq \frac{\mathbb{E}[e^{\lambda\frac{1}{n}\sum_{i=1}^{n} X_i}]}{e^{n\lambda t}}$$

$$= e^{-n\lambda t}\prod_{i=1}^{n}\mathbb{E}[e^{\lambda X_i}] = e^{-n\lambda t + n\psi(\lambda)}, \tag{1.2}$$

where

$$\psi(\lambda) = \log \mathbb{E}[e^{\lambda X}] = \log(\frac{e^{\lambda} + e^{-\lambda}}{2}) \leq \lambda^2/2. \tag{1.3}$$

Plugging it into (1.2), we have

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

□

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:

$$\psi(\lambda) = \log \mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}] \tag{1.4}$$

and the Legendre dual of the log-moment generating function:

$$\psi^*(t) = \sup_{\lambda>0}\{\lambda t - \psi(\lambda)\}. \tag{1.5}$$

Lemma 1.3. *If X has a log-moment generating function ψ with 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,

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

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

$$\mathbb{P}\{|X - \mathbb{E}[X]| \geq t\} \leq 2e^{-\frac{t^2}{2\sigma^2}}, \tag{1.6}$$

which is similar to the tail of Gaussian. In fact, the tail estimate (1.6) is often used as the equivalent definition of the sub-Gaussian class.

Corollary 1.5 (Chernoff bound)**.** *Let $X_1, \ldots, X_n$ be i.i.d. sub-Gaussian random variables with mean μ and variance proxy σ^2. Then*

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

The following Hoeffding's lemma implies that all the bounded random variables are sub-Gaussian.

Lemma 1.6 (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,

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

Let $\mathbb{Q}$ denote the distribution with $\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} = e^{\lambda X} / \mathbb{E}[e^{\lambda X}]$. 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

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

$\square$

Remark 1.7. 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. (1.3).

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

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

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 expect a lot of oscillations.
- If f is equally robust to small changes for all coordinates, we anticipate that this case will behave like the empirical mean.

Theorem 1.9 (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,

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

One can think $D_i f(x)$ as a measure of the sensitivity of f to the i-th coordinates. Considering the case of empirical mean, $D_i f(x) = O(1/n)$ for every i. This recovers the Hoeffding's inequality (Corollary 1.8).

The proof needs following lemmas.

Lemma 1.10 (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. $\log \mathbb{E}[e^{\lambda \Delta_i} | \mathcal{F}_{i-1}] \le \frac{\lambda^2 \sigma_i^2}{2}$ *(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

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

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 1.11 (Azuma-Hoeffding's inequality)**.** *Under the assumption of Lemma 1.10, 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,*

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

Proof. Combining Lemma 1.3, 1.6 and 1.10, we complete the proof. □

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

$$\begin{aligned} f(X) - \mathbb{E}[f(X)] &= f(X) - \mathbb{E}[f(X)|X_1, \dots, X_{n-1}] \\ &\quad + \mathbb{E}[f(X)|X_1, \dots, X_{n-1}] - \mathbb{E}[f(X)|X_1, \dots, X_{n-2}] \\ &\quad + \cdots + \mathbb{E}[f(X)|X_1] - \mathbb{E}[f(X)] \\ &= \sum_{i=1}^n \Delta_i, \end{aligned} \tag{1.7}$$

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. □

Lemma 1.12 (Maximal inequality)**.** *Assume $X_1, \dots, X_n$ be i.i.d. sub-Gaussian random variable with the variance proxy σ^2. Then,*

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

Proof. For any $\lambda > 0$,

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

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

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