COMP 7070 Advanced Topics in AI and ML February 06, 2024

Lecture 5: SubGaussian Random Variables and Concentration Inequalities

Instructor: Yifan Chen *Scribes: Xiong Peng, Riwei Lai* *Proof reader: Zhanke Zhou*

5.1 Basic Inequalities

We present several fundamental inequalities used in probability theory.

① **Markov's Inequality**: For a non-negative random variable X, the probability that X is at least t is bounded by the expected value of X over t:

$$P(X \geq t) \leq \frac{\mathbb{E}[X]}{t}, \quad \text{for } X \geq 0.$$

Proof. The expected value of X is:

$$\mathbb{E}[X] = \int_0^\infty x \cdot p(x)\, dx.$$

This can be split as:

$$\int_0^t x \cdot p(x)\, dx + \int_t^\infty x \cdot p(x)\, dx.$$

Since $x \geq t$ for the second integral, we have:

$$\mathbb{E}[X] \geq \int_t^\infty t \cdot p(x)\, dx = t \cdot P(X \geq t).$$

□

② **Chebyshev's Inequality**: For a random variable X with mean μ and variance $\text{Var}(X)$, the probability that the deviation of X from μ is at least t is bounded by the variance over t^2:

$$P(|X - \mu| \geq t) \leq \frac{\text{Var}(X)}{t^2}.$$

Proof. Apply Markov's inequality to the non-negative random variable $|X-\mu|^2$, we have

$$LHS = P(|X-\mu|^2 \geq t^2) \leq \frac{1}{t^2}\mathbb{E}[|X-\mu|^2] = RHS.$$

□

③ **Chernoff Bound**: The Chernoff bound combines the moment generating function with Markov's inequality to provide an exponential bound on the tail probabilities.

$$P(X - \mu \geq t) = P(\exp(\lambda(X-\mu)) \geq \exp(\lambda t)), \quad \forall \lambda > 0.$$

$$\leq \exp(-\lambda t) \cdot \mathbb{E}[\exp(\lambda(X-\mu))], \quad \lambda \in [-b, b].$$

Let $\phi(\lambda) \equiv \mathbb{E}[\exp(\lambda(X-\mu))]$, this leads to:

$$P(X - \mu \geq t) \leq \exp(-\lambda t) \cdot \phi(\lambda), \quad \forall \lambda \in [0, b].$$

$$\implies P(X - \mu \geq t) \leq \inf_{\lambda \in [0,b]} \exp(-\lambda t) \cdot \phi(\lambda).$$

5.2 Subgaussian

5.2.1 Definition

A random variable X, subject to $\mathbb{E}[\exp(\lambda(X-\mu))] \leq \exp(\frac{\lambda^2\sigma^2}{2})$ for all $\lambda \in \mathbb{R}$.

① Subgaussian with Chernoff bound.

$$P(X-\mu \geq t) \leq \inf_{\lambda>0} \exp(-\lambda t)\Phi(\lambda) \leq \inf_{\lambda>0} \exp(\frac{1}{2}\sigma^2\lambda^2 - \lambda t).$$

where $\lambda = \frac{t}{\sigma^2}$, then we have

$$P(X-\mu \geq t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

If X is subgaussian, then $-X$ is also subgaussian.

$$P(-X-(-\mu) \geq t) = \mathbb{P}(X-\mu \leq -t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Therefore,

$$P(|X-\mu| \geq t) \leq 2\cdot\exp\left(-\frac{t^2}{2\sigma^2}\right).$$

② Any bounded random variable is subgaussian.

Proof. Let $X \in [a,b]$ almost surely. Then

$$\begin{aligned}\mathbb{E}[\exp(\lambda(X-\mu))] &= \mathbb{E}_X \exp(\lambda(X-\mathbb{E}[X'])),\\ &\leq \mathbb{E}_X\mathbb{E}'_X \exp(\lambda(X-X')),\\ &= \mathbb{E}_X\mathbb{E}'_X\left[\mathbb{E}_\epsilon \exp(\lambda(X-X')\cdot\epsilon)\right].\end{aligned}$$

ϵ is a Rademacher random variable, meaning $\epsilon = \{1 \text{ with probability } \frac{1}{2}, -1 \text{ with probability } \frac{1}{2}\}$, thus we have

$$\begin{aligned}\mathbb{E}_\epsilon \exp(\lambda\cdot(X-X')\cdot\epsilon) &= \frac{1}{2}\exp(\lambda\cdot(X-X')) + \frac{1}{2}\exp(\lambda\cdot(X'-X)),\\ &\leq \exp\left(\frac{1}{2}\lambda^2\cdot(X-X')^2\right).\end{aligned}$$

Thus we have

$$\begin{aligned}\mathbb{E}[\exp(\lambda(X-\mu))] &\leq \mathbb{E}_X\mathbb{E}'_X \exp\left(\frac{1}{2}\lambda^2(X-X')^2\right),\\ &\leq \exp\left(\frac{1}{2}\lambda^2(b-a)^2\right).\end{aligned}$$

Hence X is subgaussian with $\sigma^2 = (b-a)^2$. □

③ Additivity of Subgaussian.

Let X_i be subgaussian, i.e. $X_i \sim \text{SubG}(\sigma_i^2)$, then $\sum X_i$ is also subgaussian given X_i's are independent, and $\sum X_i \sim \text{SubG}(\sum \sigma_i^2)$.

We can further derive the Hoeffding bound:

$$P\left(\sum(X_i-\mu) \geq t\right) \leq \exp\left(-\frac{t^2}{2\sum\sigma_i^2}\right).$$

④ If we know $P(|X| > t)$, then we can have:

$$\mathbb{E}[|X|^k] = \int_0^{\infty} P(|X|^k > t)\, dt \leq \int_0^{\infty} 2 \cdot \exp\left(-\frac{t^{2/k}}{2\sigma^2}\right) dt.$$

by using the bound for $P(|X| > t)$:

$$P(|X| > t) \leq 2 \cdot \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

We can get:

$$\mathbb{E}[|X|^k] \approx \left(2\sigma^2\right)^{\frac{k}{2}} \cdot k \cdot \Gamma\left(\frac{k}{2}\right) = \mathcal{O}\left(\sigma^k\right).$$

5.2.2 $f(X) - \mathbb{E}f(X)$

Let $f(X) \equiv f(X_1, X_2, \ldots, X_n)$, if f has a bounded difference, $f(X) - \mathbb{E}f(X)$ will be subgaussian.

① Doob construction. Construct a martingale with $f(X)$ and $X_{1:n}$.

$$Y_k = \mathbb{E}[f(X)|F_k], \quad \mathcal{F}_k = \sigma(X_1, \ldots, X_k).$$

Definition of martingale:

$$\mathbb{E}[Y_{k+1}|\mathcal{F}_k] = Y_k.$$

Which can be derived as follows:

$$\mathbb{E}[Y_{k+1}|\mathcal{F}_k] = \mathbb{E}[\mathbb{E}[f(X)|\mathcal{F}_{k+1}]|\mathcal{F}_k] \overset{\text{Tower property}}{=} \mathbb{E}[f(X)|\mathcal{F}_k] \equiv Y_k.$$

Let

$$D_k = Y_k - Y_{k-1},$$

then

$$\mathbb{E}[D_{k+1}|\mathcal{F}_k] = \mathbb{E}[Y_{k+1} - Y_k|\mathcal{F}_k] = 0,$$

finally,

$$Y_n - Y_0 = f(X) - \mathbb{E}[f(X)] = \sum_{i=1}^{n} D_k.$$

② Azuma-Hoeffding. For $D_k \in [a_k, b_k]$, $\sum_{k=1}^{n} D_k$ is subG.

Proof.

$$\mathbb{E}[\exp(\lambda \sum_{k=1}^{n} \cdot D_k) = \mathbb{E}[\mathbb{E}[\exp(\lambda \cdot \sum_{k=1}^{n} D_k) \cdot \exp(\lambda D_n|\mathcal{F}_{n-1})]].$$

and we have

$$\mathbb{E}[\exp(\lambda \sum_{k=1}^{n} \cdot D_k) = \mathbb{E}[\exp(\lambda \cdot \sum_{k=1}^{n-1} D_K)] \cdot \mathbb{E}[\exp(\lambda D_n)|\mathcal{F}_{n-1}].$$

since $D_k|\mathcal{F}_{k-1}$ bdd is subG, we have

$$\mathbb{E}[\exp(\lambda D_k)|\mathcal{F}_{k-1}] \leq \exp(\frac{\lambda^2(b_k - a_k)^2}{8}).$$

then

$$\mathbb{E}[\exp(\lambda \sum_{k=1}^{n} \cdot D_k) \leq \mathbb{E}[\exp(\lambda \cdot \sum_{k=1}^{n-1} D_K)] \cdot \exp(\frac{\lambda^2(b_k - a_k)^2}{8}) \leq \exp(\frac{\lambda^2}{8} \sum_{k=1}^{n} (b_k - a_k)^2).$$

thus $\sum D_k$ is subG with $\sigma^2 = \frac{1}{4} \sum_{k=1}^{n} (b_k - a_k)^2$.

□

③ Bounded Difference Inequality.

$$\forall x, x_k^{'}$$

if

$$|f(\mathbf{x}) - f(\mathbf{x}_k^{'})| \leq L_k.$$

Here

$$\mathbf{x}_k^{'} = \begin{cases} x_k^{'}, & \text{if } x_k = x_k', \\ x_j, & \text{if } x_k \neq x_k', \end{cases}$$

Define $\sum D_k = f(\mathbf{x}) - \mathbb{E}f(\mathbf{x})$, we have $\sum D_k$ is subG.

Proof. Using Azuma-Hoeffding inequality to show D_k is bounded:
Let

$$\begin{aligned} D_k &= Y_k - Y_{k-1}, \\ A_k &= \inf_x \mathbb{E}[f(\mathbf{x}) \mid \mathbf{X}_{1\sim k-1}, \mathbf{X}_k = x] - Y_{k-1}, \\ B_k &= \sup_x \mathbb{E}[f(\mathbf{x}) \mid \mathbf{X}_{1\sim k-1}, \mathbf{X}_k = x] - Y_{k-1}. \end{aligned}$$

Then we have

$$A_k \leq D_k \leq B_k,$$

$$B_k - A_k \leq \sup_{x,y} \mathbb{E}[f(\mathbf{X})_{1\sim k-1}, x, \mathbf{X}_{k+1\sim n})] - \mathbb{E}[f(\mathbf{X})_{1\sim k-1}, y, \mathbf{X}_{k+1\sim n})] \leq \sup_{x,y} L_k = L_k.$$

So that D_k is bdd.
By Azuma-Hoeffding inequality, we have $\sum D_k$ is subG, which completes the proof. □

④ Rademacher complexity: the complexity of a vector collection $\mathcal{A}$:

$$\left\{ \begin{bmatrix} cf(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}, \begin{bmatrix} cf'(X_1) \\ \vdots \\ f'(X_n) \end{bmatrix}, \dots \right\}, \text{where} f \in \mathcal{F} \Rightarrow \text{all the models.}$$

Assume that ε is a Rademacher vector, we have

$$\mathbb{E}_\varepsilon Z(\mathcal{A}) = \mathbb{E} \sup_{a \in \mathcal{A}} \langle a, \varepsilon \rangle.$$

Define $\varepsilon \to \varepsilon^{'k}$ as the k-th element of $\varepsilon^{'k} \neq \varepsilon_k$ and $f(\varepsilon)$ as $Z(\mathcal{A})$, we have $f(\varepsilon) - f(\varepsilon^{'k})$ has bounded difference.

Proof. Since

$$f(\varepsilon^{'k}) = \sup_{a \in \mathcal{A}} \langle a, \varepsilon^{'k} \rangle \geq \langle a, \varepsilon^{'k} \rangle, \forall a \in \mathcal{A}.$$

Which can be transferred to:

$$\langle a, \varepsilon \rangle - f(\varepsilon^{'k}) \leq \langle a, \varepsilon - \varepsilon^{'k} \rangle, \forall a \in \mathcal{A}.$$

And we have

$$\sup_a \langle a, \varepsilon \rangle - f(\varepsilon^{'k}) \leq \sup_a \langle a, \varepsilon - \varepsilon^{'k} \rangle.$$

So, finally, we have

$$f(\varepsilon) - f(\varepsilon^{'k}) \leq \sup_a \; 2 \cdot |a_k| =: L_k.$$

which completes the proof. □

⑤ Maximal Inequality: (worst case won't happen w.h.p.)

$$\frac{1}{n}\sum z_i \to \infty \Rightarrow \text{w.h.p } \left|\frac{1}{n}\sum z_i\right| \le t.$$

Given $X_{i\sim N}$ not i.i.d. but $\mathbb{E}\left[\max_i X_i\right]$ is sub-G(δ^2)

$$\begin{aligned}
(1) \quad \mathbb{E}[\max_i X_i] &= \frac{1}{s}\mathbb{E}\left[\log\left(\exp\left(s\cdot\max_i X_i\right)\right)\right], \quad \forall s>0 \\
&\le \frac{1}{s}\log\left(\mathbb{E}\left[\exp\left(s\cdot\max_i X_i\right)\right]\right) \\
&= \frac{1}{s}\log\left(\mathbb{E}\left[\max_i \ \exp\left(s\cdot X_i\right)\right]\right) \\
&\le \frac{1}{s}\log\left(\mathbb{E}\left[\sum_i \ \exp\left(s\cdot X_i\right)\right]\right) \\
&= \frac{1}{s}\log\left(\sum_i \exp\left(\frac{\delta^2 s^2}{2}\right)\right) \\
&= \frac{1}{s}\log N + \frac{\delta^2}{2}S, \quad \forall s>0 \\
\Rightarrow LHS &\le \inf_{s>0} RHS = \delta\cdot\sqrt{2\log N}.
\end{aligned}$$

$$\begin{aligned}
(2) \quad \mathbf{P}(\max_i X_i > t) &= \mathbf{P}\left(\bigcup_i (X_i > t)\right) \\
&\le \sum_i \mathbf{P}(X_i > t) = N\cdot\exp\left(-\frac{t^2}{2\delta^2}\right).
\end{aligned}$$

$$\Rightarrow N\cdot\exp\left(-\frac{t^2}{2\delta^2}\right) \le \varepsilon \Rightarrow t = O\left(\delta\cdot\sqrt{\log\frac{N}{\varepsilon}}\right).$$

$$(3) \quad \mathbb{E}\left[\max_i |X_i|\right] = \mathbb{E}\left[\max_{i\in[N]}\max\{X_i, -X_i\}\right] \le \delta\cdot\sqrt{2\log(2N)}.$$

$$(4) \quad \mathbf{P}(\max_i |X_i| > t) \le 2N\cdot\exp\left(-\frac{t^2}{2\delta^2}\right).$$

⑥ HW: Proof Thm 1.9. of Chapter 1.

$$\mathbb{E}\left[\max_{\Theta\in\mathcal{B}_2}\Theta X\right] \le 4\delta\sqrt{d}.$$