COMP 7070 Advanced Topics in AI and ML Feb 19, 2025

Lecture 5: SubGaussian Random Variables and Concentration Inequalities

Instructor: Yifan Chen *Scribes: Yurui Lai, Xiaoyang Lin* *Proof reader: Xiong Peng*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

5.1 Motivation

In many practical applications,drawing samples from a population is essential for inference and decision making. Understanding basic inequalities helps us assess the deviation between the means of samples and the population, thereby making our **sampling** methods accurate. Moreover, **random projection** plays a key role in data compression. By studying these inequalities, we can better understand the impact of random projections on data distribution, ensuring the minimization of information loss of the random projection. In the fields of finance, engineering, and science, **risk analysis** involves assessing uncertainty and potential losses. These inequalities provide tools for quantifying risks, ensuring the robustness of systems and decision-making processes.

5.2 Basic Inequalities

We here give the definitions and proofs to the basic inequalities, including Markov's Inequality, Chebyshev's Inequality and Chernoff Bound.

Lemma 5.2.1 (Markov's Inequality) *For any random variable $X \geq 0$ and for all $t > 0$, it holds that:*

$$P(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

Proof: By definition,

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

where $p(x)$ is the probability density function of X. Split the integral into two parts:

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

In the second integral, since $x \geq t$, we have

$$\int_t^\infty x\, p(x)\, dx \geq \int_t^\infty t\, p(x)\, dx = t \cdot P(X \geq t).$$

Thus,

$$\mathbb{E}[X] \geq t \cdot P(X \geq t),$$

which implies

$$P(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

■

Markov's Inequality provides an upper bound on the tail probability of a non-negative random variable. It serves as a building block for other important inequalities, such as Chebyshev's Inequality and the Chernoff Bound.

Lemma 5.2.2 (Chebyshev's Inequality) *For any random variable X with expected value μ and variance σ^2, and for all $t > 0$,*

$$P(|X - \mu| \geq t\sigma) \leq \frac{1}{t^2}.$$

Proof: Let $Y = (X - \mu)^2$, so that $Y \geq 0$. By Markov's Inequality, for any $k > 0$,

$$P(Y \geq k) \leq \frac{\mathbb{E}[Y]}{k}.$$

Choosing $k = t^2\sigma^2$ and noting that $\mathbb{E}[Y] = \sigma^2$ gives:

$$P\big((X - \mu)^2 \geq t^2\sigma^2\big) \leq \frac{\sigma^2}{t^2\sigma^2} = \frac{1}{t^2}.$$

Since $(X - \mu)^2 \geq t^2\sigma^2$ is equivalent to $|X - \mu| \geq t\sigma$, the result follows. ■

Chebyshev's Inequality utilizes the variance of a random variable, providing a tighter bound than Markov's Inequality on the probability of deviation from the mean. This makes it applicable to any distribution with finite variance, not just non-negative variables.

Lemma 5.2.3 (Chernoff Bound) *For any random variable X with mean μ and for all $t > 0$, the following holds:*

$$P(X - \mu \geq t) \leq \inf_{\lambda \in (0,b]} \exp(-\lambda t) \cdot \mathbb{E}\left[\exp(\lambda(X - \mu))\right],$$

where λ is a tuning parameter and b is a positive constant such that the moment generating function is finite for $\lambda \in [-b, b]$.

Proof: For any $\lambda > 0$, note that

$$P(X - \mu \geq t) = P\left(\exp(\lambda(X - \mu)) \geq \exp(\lambda t)\right).$$

By Markov's Inequality,

$$P\left(\exp(\lambda(X - \mu)) \geq \exp(\lambda t)\right) \leq \exp(-\lambda t) \cdot \mathbb{E}\left[\exp(\lambda(X - \mu))\right].$$

Defining the moment generating function as $\varphi(\lambda) = \mathbb{E}\left[\exp(\lambda(X - \mu))\right]$ and optimizing over λ in the interval $(0, b]$ yields the stated bound. ■

The Chernoff bound shows that the probability of a significant deviation from the mean decays exponentially with the size of the deviation. This is much stronger than bounds provided by Markov's inequality or Chebyshev's inequality, which only give polynomial decay.

5.3 Subgaussian

Subgaussian random variables are important in probability and statistics because they have exponentially decaying tail probabilities and bounded moment generating functions. These properties make them useful for deriving concentration inequalities and controlling the deviation of random variables from their means.

Definition 5.3.1 (Subgaussian Random Variable) *A real-valued random variable x is said to be sub-Gaussian if $\exists \sigma > 0$ such that $\forall \lambda \in \mathbb{R}$:*

$$\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\sigma^2\lambda^2}{2}}$$

Theorem 5.3.2 (Subgaussian Chernoff Bound) *For any subgaussian random variable X with parameter σ, $\forall t > 0$:*

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

Proof: Recall the Chernoff bound, we have:

$$P(X - \mu \geq t) \leq \inf_{\lambda > 0} \exp(-\lambda t) \cdot \mathbb{E}[\exp(\lambda(X - \mu))].$$

Then we use the property of subgaussian:

$$\mathbb{E}[\exp(\lambda(X - \mu))] \leq \exp\left(\frac{\lambda^2 \sigma^2}{2}\right).$$

Then we combine above together:

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

We set $\lambda = \frac{t}{\sigma^2}$ and get:

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

Since $-X$ is also subgaussian,

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

Finally, we obtain Subgaussian Chernoff Bound,

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

■

Theorem 5.3.3 (Bounded Random Variable is Subgaussian) *Any bounded random variable $X \in [a, b]$ is subgaussian with $\sigma^2 = (a - b)^2$.*

Proof: Firstly, let $\epsilon = \{1 \quad probability = 0.5, -1 \quad probability = 0.5\}$, we have:

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

The right-hand-side can be rewritten and bounded by:

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

Then we combine the above inequalities together:

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

Finally we get that X is subgaussian with $\sigma^2 = (a - b)^2$. ■

Theorem 5.3.4 (Additivity of Subgaussian) *Let $X_i \sim SubG(\sigma_i^2)$, then $\sum X_i$ is also subgaussian given X_i's are independent, and $\sum X_i \sim SubG(\sum \sigma_i^2)$. And the Hoeffding bound is:*

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

Proof: The proof follows from the properties of subgaussian random variables and the Hoeffding bound. Since each X_i is subgaussian, their sum is also subgaussian with variance parameter $\sum \sigma_i^2$.

The Hoeffding bound then follows from the Chernoff bound applied to the sum of subgaussian random variables.We firstly give the Chernoff bound for the sum of subgaussian random variables:

$$P\left(\sum (X_i - \mu) \geq t\right) \leq \inf_{\lambda > 0} \exp(-\lambda t) \cdot \mathbb{E}\left[\exp\left(\lambda \sum (X_i - \mu)\right)\right].$$

We use the subgaussian property for each X_i:

$$\mathbb{E}\left[\exp\left(\lambda(X_i - \mu)\right)\right] \leq \exp\left(\frac{\lambda^2 \sigma_i^2}{2}\right).$$

Since the X_i's are independent, the moment generating function of the sum is the product of the individual moment generating functions:

$$\mathbb{E}\left[\exp\left(\lambda \sum (X_i - \mu)\right)\right] = \prod_{i=1}^{n} \mathbb{E}\left[\exp\left(\lambda(X_i - \mu)\right)\right] \leq \prod_{i=1}^{n} \exp\left(\frac{\lambda^2 \sigma_i^2}{2}\right) = \exp\left(\frac{\lambda^2}{2} \sum \sigma_i^2\right).$$

Then we substitute this back into the Chernoff bound:

$$P\left(\sum (X_i - \mu) \geq t\right) \leq \inf_{\lambda > 0} \exp\left(-\lambda t + \frac{\lambda^2}{2} \sum \sigma_i^2\right).$$

To find the optimal λ, we take the derivative of the exponent with respect to λ and make it zero:

$$-\frac{d}{d\lambda}\left(-\lambda t + \frac{\lambda^2}{2} \sum \sigma_i^2\right) = -t + \lambda \sum \sigma_i^2 = 0 \implies \lambda = \frac{t}{\sum \sigma_i^2}.$$

We also take $\lambda = \frac{t}{\sum \sigma_i^2}$ back the above equation:

$$-\lambda t + \frac{\lambda^2}{2} \sum \sigma_i^2 = -\frac{t^2}{\sum \sigma_i^2} + \frac{t^2}{2(\sum \sigma_i^2)^2} \sum \sigma_i^2 = -\frac{t^2}{2 \sum \sigma_i^2}.$$

Finally, we get the Hoeffding bound:

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

■

Theorem 5.3.5 (Bounding Moments Using Tail Probabilities) *If $P(|X| > t)$, then the following holds:*

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

Proof: We firstly give the definition of the expected value of $|X|^k$:

$$\mathbb{E}[|X|^k] = \int_0^\infty P(|X|^k > t)\, dt$$

We use Chernoff bound for $P(|X| > t)$:

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

Then we have:

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

Because $u = \frac{t^{2/k}}{2\sigma^2}$, we get:

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

■

5.3.1 $f(x) - \mathbb{E}f(x)$

Theorem 5.3.6 (Bounded Difference and Subgaussian Property) *Let $f(X) \equiv f(X_1, X_2, \ldots, X_n)$. If f has a bounded difference, then $f(X) - \mathbb{E}f(X)$ is subgaussian.*

Proof: We use the Doob construction to construct a martingale with $f(X)$ and $X_{1:n}$.Let Y_k be:

$$Y_k = \mathbb{E}[f(X)|\mathcal{F}_k]$$

where $\mathcal{F}_k = \sigma(X_1, \ldots, X_k)$. Then a martingale is defined as:

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

This 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 we have:

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

Since f has a bounded difference, each D_k is bounded. Therefore, $f(X) - \mathbb{E}f(X)$ is a sum of bounded random variables, which implies it is subgaussian. ■

Theorem 5.3.7 (Azuma-Hoeffding Inequality) *For $D_k \in [a_k, b_k]$, $\sum_{k=1}^{n} D_k$ is subgaussian.*

Proof:

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

and we have

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

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

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

then

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

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

■

When contextualized for functions with per-variable output sensitivity bounds, the Azuma-Hoeffding result crystallizes into the Bounded Difference Inequality.

Theorem 5.3.8 (Bounded Difference Inequality) $\forall x, x'_k$

$$\textit{if} \quad \left|f(x) - f(x'_k)\right| \leq L_k$$

Here

$$x'_k = \begin{cases} x'_k & \textit{if } x_k = x'_k \\ x_j & \textit{if } x_k \neq x'_k \end{cases}$$

Define $\sum D_k = f(x) - Ef(x)$, we have $\sum D_k$ is subG.

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

Let

$$D_k = Y_k - Y_{k-1},$$
$$A_k = \inf_x \mathbb{E}\left[f(x) \mid X_{1\sim k-1}, X_k = x\right] - Y_{k-1},$$
$$B_k = \sup_x \mathbb{E}\left[f(x) \mid X_{1\sim k-1}, X_k = x\right] - Y_{k-1}$$

Then we have

$$A_k \le D_k \le B_k,$$
$$B_k - A_k \le \sup_{x,y} \left(\mathbb{E}\left[f(X)_{1\sim k-1}, x, X_{k+1\sim n}\right] - \mathbb{E}\left[f(X)_{1\sim k-1}, y, X_{k+1\sim n}\right]\right) \le \sup_{x,y} L_k = L_k$$

So that D_k is bdd. By theorem. 5.3.7, $\sum D_k$ is subG. ■

Theorem 5.3.9 (Rademacher complexity) *The complexity of a vector collection $\mathcal{A}$:*

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

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

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

Which can be transferred to:

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

And we have

$$\sup_a \Big(\langle a, \varepsilon\rangle - f(\varepsilon'^k)\Big) \le \sup_a \langle a, \varepsilon - \varepsilon'^k\rangle$$

Finally,

$$f(\varepsilon) - f(\varepsilon'^k) \le \sup_a 2 \cdot |a_k| =: L_k$$

which completes the proof. ■

5.3.2 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)

Theorem 5.3.10 (Expectation bound) *for $i \in [1, N]$, $\mathbb{E}\left[\max_i X_i\right] \le \delta \cdot \sqrt{2\log N}$*

Proof:

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

■

Theorem 5.3.11 (Tail Probability Bound) *for* $i\in[1,N]$, $P\left(\max_i X_i > t\right) \le N\cdot\exp\left(-\frac{t^2}{2\delta^2}\right)$

Proof:

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

■

From above proof, we can also calculate the bound t:

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

According to theorem 5.3.10 and 5.3.11, we can get the bounds of $\max_i |X_i|$ are as follows:

$$\begin{aligned}
\mathbb{E}\left[\max_i |X_i|\right] &\le \delta\cdot\sqrt{2\log(2N)}\\
\mathbf{P}\left(\max_i |X_i| > t\right) &\le 2N\cdot\exp\left(-\frac{t^2}{2\delta^2}\right)
\end{aligned}$$