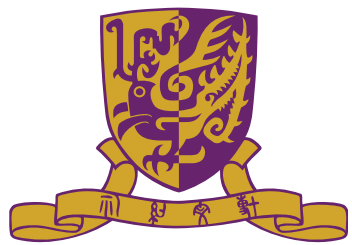
DDA5001 Machine Learning

Proofs of Porbabilistic Inequalities

Xiao Li

School of Data Science
The Chinese University of Hong Kong, Shenzhen



Law of Large Numbers

Typically, if we have $X, X_1, X_2, \cdots, X_n$ are i.i.d., then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mathbb{E}[X] \text{ as } n \to \infty$$

Law of large numbers: If $\mathbb{E}[|X|] < \infty$, then

$$\Pr\left[\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i \neq \mathbb{E}[X]\right] = 0$$

This is an asymptotic (infinite n) result. We need non-asymptotic (finite n) concentration bounds.

Concentration Using Moments

Markov's Inequality

Basic method: Control tail probability $\Pr[X \geq t]$ by controlling the moments of the random variable X. For example:

- Markov's inequality (requires only existence of the first moment)
- $\vdots$
- Chernoff bound (requires existence of the moment generating function)

Most elementary one:

Lemma: Markov's inequality

Given a non-negative random variable X with finite mean, we have

$$\Pr[X \geq t] \leq \frac{\mathbb{E}[X]}{t} \quad \forall t > 0$$

- Markov's inequality is tight, i.e., non-improvable in general.
- Smaller $\mathbb{E}[X]$ and/or larger t implies a smaller probability

Proof

Note that

$$\Pr[X \geq t] = \int_t^{+\infty} p(x)dx = \mathbb{E}[1_{X \geq t}]$$

Case I: $X \geq t$, then

$$X/t \geq 1 \geq 1_{X \geq t}$$

Case II: $X < t$, then

$$X/t \geq 0 = 1_{X \geq t}$$

Combing the above, we have

$$\Pr[X \geq t] = \mathbb{E}[1_{X \geq t}] \leq \frac{E[X]}{t}$$

More Extensions

k-th moments:

$$\Pr[|X - \mathbb{E}[X]| \geq t] \leq \frac{\mathbb{E}[|X - \mathbb{E}[X]|^k]}{t^k} \quad \forall t > 0$$

More generally:

- For any strictly increasing nonnegative function ϕ, we have

$$\Pr[\phi(X) \geq \phi(t)] \leq \frac{\mathbb{E}[\phi(X)]}{\phi(t)} \quad \forall t > 0$$

Choosing a specified ϕ can lead to much sharper bounds.

Concentration for Sub-Gaussian

Moment Generating Function

Definition: For a random variable X, the moment generating function (MGF) is defined as

$$M_X(\lambda) = \mathbb{E}[\exp(\lambda X)]$$

Example:

- Normal distribution $X \sim \mathcal{N}(0, \sigma^2)$,

$$M_X(\lambda) = \exp\left(\frac{\lambda^2 \sigma^2}{2}\right)$$

- Rademacher random variable: $\Pr[X = 1] = \frac{1}{2}$ and $\Pr[X = -1] = \frac{1}{2}$

$$M_X(\lambda) \leq \exp\left(\frac{\lambda^2}{2}\right)$$

Chernoff Bounds

Prop: Chernoff bounds

For any random variable X and any $t > 0$, we have

$$\Pr[X - \mathbb{E}[X] \geq t] \leq \min_{\lambda \geq 0} \mathbb{E}\left[e^{\lambda(X - \mathbb{E}[X])}\right] e^{-\lambda t}$$

and

$$\Pr[X - \mathbb{E}[X] \leq -t] \leq \min_{\lambda \geq 0} \mathbb{E}\left[e^{\lambda(\mathbb{E}[X] - X)}\right] e^{-\lambda t}$$

Proof: We first note that

$$X - \mathbb{E}[X] \geq t \iff e^{\lambda(X - \mathbb{E}[X])} \geq e^{\lambda t}$$

Applying the Markov's inequality, we have

$$\begin{aligned}\Pr[X - \mathbb{E}[X] \geq t] &= \Pr[e^{\lambda(X - \mathbb{E}[X])} \geq e^{\lambda t}] \\ &\leq \frac{\mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}]}{e^{\lambda t}}\end{aligned}$$

Optimizing over λ gives the desired result.

Sub-Gaussian Random Variable

Definition: A random variable X with mean $\mu = \mathbb{E}[X]$ is called sub-Gaussian, if there exists a positive number σ such that

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

$\sigma > 0$ is the sub-Gaussian parameter.

Example:

- Gaussian distribution $X \sim \mathcal{N}(\mu, \sigma^2)$ (equality holds).
- Rademacher random variable: $\Pr[X = 1] = \frac{1}{2}$ and $\Pr[X = -1] = \frac{1}{2}$

$$\mathbb{E}[e^{\lambda(X-0)}] \leq \exp\left(\frac{\lambda^2}{2}\right)$$

- Any bounded random variable on $[a, b]$, we have $\sigma \leq \frac{b-a}{2}$.

$\rightsquigarrow$ See [1, Subsection 2.1.2].

[1] Wainwright, M. J. (2019). High-dimensional statistics: A non-asymptotic viewpoint. Cambridge University Press.

Recall Concentration Inequality for Sub-Gaussian

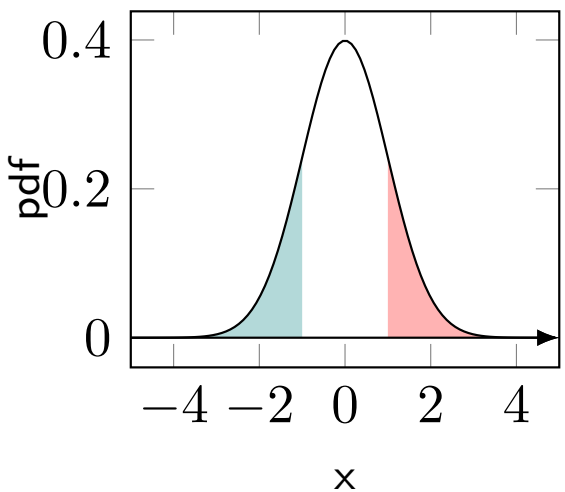
Theorem: Sub-Gaussian Concentration

Suppose X is a sub-Gaussian random variable with mean μ and parameter σ, then for any $t > 0$, we have

$$\Pr[|X - \mu| \geq t] \leq 2e^{-\frac{t^2}{2\sigma^2}}$$

- This bound is much sharper.
- Exponential decay with respect to t.
- Equivalently,

$$\Pr[|X - \mu| \leq t] \geq 1 - 2e^{-\frac{t^2}{2\sigma^2}}$$



Proof

Recall the Chernoff bounds,

$$\Pr[X - \mu \geq t] \leq \min_{\lambda \geq 0} \mathbb{E}\left[e^{\lambda(X-\mu)}\right] e^{-\lambda t}$$

By the definition of sub-Gaussian random variable, we have

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

Combing the above two inequalities yields

$$\Pr[X - \mu \geq t] \leq \min_{\lambda \geq 0} e^{\left(\frac{\lambda^2\sigma^2}{2} - \lambda t\right)}.$$

Note that the quadratic function $q(\lambda) = \frac{\lambda^2\sigma^2}{2} - \lambda t$ attains its minimum at $\lambda = \frac{t}{\sigma^2}$. Optimizing the RHS over λ provides

$$\Pr[X - \mu \geq t] \leq e^{\left(-\frac{t^2}{2\sigma^2}\right)}.$$

By a symmetric argument, one can deduce the other side.

Consequence: Hoeffding's Inequality

Proof of Hoeffding's Inequality

Theorem: Hoeffding's Inequality

Suppose X_i are independent sub-Gaussian random variables with mean μ_i and sub-Gaussian parameter σ_i for $i = 1, \cdots, n$, then for any $t > 0$, we have

$$\Pr\left[\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right] \leq e^{-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2}}$$

Proof:

$$\Pr\left[\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right] \leq \min_{\lambda \geq 0} \mathbb{E}\left[e^{\lambda \sum_{i=1}^{n}(X_i - \mu_i)}\right] e^{-\lambda t} \quad \text{Chernoff bound}$$

$$= \min_{\lambda \geq 0} \mathbb{E}\left[\prod_{i=1}^{n} e^{\lambda(X_i - \mu_i)}\right] e^{-\lambda t} \leq \min_{\lambda \geq 0} \prod_{i=1}^{n} e^{\lambda^2 \sigma_i^2/2 - \lambda t}$$

$$= \min_{\lambda \geq 0} e^{\lambda^2 \sum_{i=1}^{n}\sigma_i^2/2 - \lambda t} = e^{-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2}}$$

- The Hoeffding's Inequality used in our lecture can be seen as a corollary of this more general Hoeffding's Inequality.