CSE 392: Matrix and Tensor Algorithms for Data — Spring 2024

Lecture 2 — Jan. 22nd

Instructor: Shashanka Ubaru — *Scribe: Yifan Zhang*

1 Probability Review

Here are some basic facts about the probability theory.

1. If X is a random variable on $\mathbb{R}$ with density $p(x)$, then $\mathbb{E}\,X = \int xp(x)dx$.
2. If X is discrete with probability mass function q supported on $S \subseteq \mathbb{R}$, then $\mathbb{E}\,X = \sum_{s\in S} sq(s)$.
3. $\operatorname{Var} X = \mathbb{E}(X - \mathbb{E}\,X)^2 = \mathbb{E}\,X^2 - (\mathbb{E}\,X)^2$.
4. For a scalar α, $\mathbb{E}(\alpha X) = \alpha\,\mathbb{E}\,X$ and $\operatorname{Var}(\alpha X) = \alpha^2 \operatorname{Var} X$.
5. For constants α, β, $\mathbb{E}(\alpha X + \beta Y) = \alpha\,\mathbb{E}\,X + \beta\,\mathbb{E}\,Y$.
6. For disjoint events $\{A_i\}_i$, $\mathbb{E}\,X = \sum_i \mathbb{E}(X|A_i)\,\mathbb{P}(A_i)$.
7. If X and Y are independent, then $\mathbb{E}\,XY = \mathbb{E}\,X\,\mathbb{E}\,Y$ and $\operatorname{Var}(X+Y) = \operatorname{Var} X + \operatorname{Var} Y$.
8. For two events A and B, $\mathbb{P}(A\cap B) = \mathbb{P}(A)\,\mathbb{P}(A|B) = \mathbb{P}(B)\,\mathbb{P}(B|A)$.
9. A and B are independent if and only if $\mathbb{P}(A\cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$.
10. A and B are called mutually exclusive if $\mathbb{P}(A\cap B) = 0$.
11. $\|X\|_p = (\mathbb{E}\,|X|^p)^{1/p}$ defines a norm on random variables for all $1 \leqslant p < \infty$.

2 Concentration Inequalities

Proposition 2.1 (Markov's inequality)**.** *Let X be a non-negative random variable. Then for any $t > 0$,*

$$\mathbb{P}(X \geqslant t) \leqslant \frac{\mathbb{E}\,X}{t}.$$

Proof. Let the distribution of X be μ. If X does not have finite expectation, the inequality trivially holds. Assume X is integrable, then $\mathbb{P}(X \geqslant t) = \int_t^{+\infty} d\mu = t^{-1}\int_t^{+\infty} td\mu \leqslant t^{-1}\int_t^{+\infty} xd\mu \leqslant t^{-1}\,\mathbb{E}\,X$. ■

Proposition 2.2 (Chebyshev's inequality)**.** *Let X be a random variable with finite expectation, then for any $k > 0$,*

$$\mathbb{P}(|X - \mathbb{E}\,X| \geqslant k) \leqslant \frac{\operatorname{Var} X}{k^2}.$$

Proof. Apply Markov's inequality to $Y = (X - \mathbb{E}\, X)^2$. ■

The following sub-additivity property of probability measures can be useful. It is often called the union bound.

Proposition 2.3 (Union bound). *For countably many events* $\{A_i\}_i$,

$$\mathbb{P}\Big(\bigcup_i A_i\Big) \leqslant \sum_i \mathbb{P}(A_i).$$

In particular,

$$\mathbb{P}\Big(\bigcap_i A_i\Big) \geqslant 1 - \sum_i \mathbb{P}(A_i^c).$$

The second bound is proved by applying the first bound to A_i^c. This is useful when we want to lower bound the probability of the good event in which all conditions A_i are satisfied.

Next, we define two important classes of random variables called respectively the sub-Gaussian and sub-exponential random variables.

Definition 2.4 (sub-Gaussian). *A random variable X is called sub-Gaussian if there is some constant $C < +\infty$ such that $\|X\|_p \leqslant C\sqrt{p}$ for all $p \geqslant 1$. The infimum of all possible choices of C is called the sub-Gaussian norm of X, denoted as $\|X\|_{\psi_2}$.*

Definition 2.5 (sub-exponential). *A random variable X is called sub-exponential if there is some constant $C < +\infty$ such that $\|X\|_p \leqslant Cp$ for all $p \geqslant 1$. The infimum of all possible choices of C is called the sub-exponential norm of X, denoted as $\|X\|_{\psi_1}$.*

Proposition 2.6. *Sub-Gaussian and sub-exponential random variables respectively form two vector spaces, and $\|\cdot\|_{\psi_2}$, $\|\cdot\|_{\psi_1}$ are valid norms on the said spaces, respectively.*

Proposition 2.7. *Normal random variables are sub-Gaussian. Gamma and exponential random variables are sub-exponential.*

We can control the growth rate of the moment generating function of these classes of random variables. Applying the Markov's inequality to random variables $e^{\lambda X}$ for some carefully chosen $\lambda > 0$ can then lead to the so-called Cramer-Chernoff bound. Below are some examples.

Proposition 2.8 (concentration for sub-Gaussian rvs). *Let X be a sub-Gaussian random variable. Then for any $t \geqslant 0$,*

$$\mathbb{P}(|X - \mathbb{E}\, X| \geqslant t) \leqslant 2e^{-ct^2/\|X\|_{\psi_2}^2},$$

where c is an absolute constant.

Proposition 2.9 (Chernoff bounds for Bernoulli). *Let X_i, $i = 1, \ldots, n$ be independent Bernoulli random variables with success rate p_i. Let $S = \sum_{i=1}^n X_i$. Then for all $\delta > 0$,*

$$\mathbb{P}(S \geqslant (1+\delta)\,\mathbb{E}\, S) \leqslant e^{-\frac{\delta^2}{2+\delta}\cdot \mathbb{E}\, S},$$

and for all $0 < \delta < 1$,

$$\mathbb{P}(S \leqslant (1-\delta)\,\mathbb{E}\, S) \leqslant e^{-\frac{\delta^2}{2}\cdot \mathbb{E}\, S},$$

Proposition 2.10 (Bernstein's inequality for sub-exponential rvs)**.** *Let X_i, $i = 1, \dots, n$ be independent sub-exponential random variables. Let $S = \sum_i X_i$ and $\mu = \mathbb{E}\, S$. Then for some absolute constant c,*

$$\mathbb{P}(|S - \mu| \geqslant t) \leqslant 2 \exp\left(-c \min\left\{\frac{t^2}{\sum_{i=1}^n \|X_i\|_{\psi_1}^2}, \frac{t}{\max_i \|X_i\|_{\psi_1}}\right\}\right).$$

Proposition 2.11 (Hoeffding's inequality for bounded rvs)**.** *Let X_i, $i = 1, \dots, n$ be independent random variables. Let $S = \sum_{i=1}^n X_i$. Then for all $t > 0$,*

$$\mathbb{P}(|S - \mathbb{E}\, S| \geqslant t) \leqslant 2e^{-2t^2/\sum_i (b_i - a_i)^2}.$$

Example 2.12. *Consider flipping a biased coin which lands on heads with probability p. We want to find k such that after k flips we ensure*

$$\mathbb{P}(|\#heads - pk| \geqslant \varepsilon k) \leqslant \delta.$$

To this end, let X_i be a Bernoulli random variable taking value 1 if the ith flip is a head, and $S = \sum_i X_i$. Using Hoeffding (Proposition 2.11),

$$\mathbb{P}(|S - pk| \geqslant \varepsilon k) \leqslant 2e^{-\frac{2(\varepsilon k)^2}{k}} = 2e^{-2\varepsilon^2 k}.$$

To ensure that the right-hand side is bounded by δ, the desired k is

$$k_{Hoeff} = \mathcal{O}(\varepsilon^{-2} \log(1/\delta)).$$

Using Chernoff (Proposition 2.9), for $\varepsilon < p$ we have

$$\mathbb{P}(|S - pk| \geqslant \varepsilon k) \leqslant 2e^{-\frac{(\varepsilon/p)^2}{3} pk} = 2e^{-\frac{\varepsilon^2}{3p} k}.$$

Consequently the desired k is

$$k_{Chern} = \mathcal{O}(p\varepsilon^{-2} \log(1/\delta)).$$

Using a naive bound like Chebyshev, we have

$$\mathbb{P}(|S - pk| \geqslant \varepsilon k) \leqslant \frac{kp(1-p)}{\varepsilon^2 k^2} = \frac{p(1-p)}{\varepsilon^2} k^{-1}$$

The desired k is then

$$k_{Cheby} = \mathcal{O}(p\varepsilon^{-2}\delta^{-1}).$$