Solution to "Chapter 2: Basic tail and concentration bounds"

Jiaxin Hu

July 16, 2020

1 Summary

Theorem 1.1 (Markov's inequality). *Suppose $X \geq 0$ is a random variable with finite mean, we have*

$$\mathbb{P}(X \geq t) \leq \frac{E[X]}{t}, \quad \text{for all } t > 0. \tag{1}$$

Theorem 1.2 (Chebyshev's inequality). *Suppose $X \geq 0$ is a random variable with finite mean μ and finite variance, we have*

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{var(X)}{t^2}, \quad \text{for all } t > 0. \tag{2}$$

Theorem 1.3 (Markov's inequality for polynomial moments). *Suppose the random variable X has a central moment of order k. Applying Markov's inequality to the random variable $|X-\mu|^k$ yields*

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\mathbb{E}[|X-\mu|^k]}{t^k}, \quad \text{for all } t > 0.$$

Theorem 1.4 (Chernoff bound). *Suppose the random variable X has a moment generating function in the neighborhood of 0; i.e. $\varphi_X(\lambda) = \mathbb{E}[e^{\lambda X}] < +\infty$, for all $\lambda \in (-b, b)$ with some $b > 0$. Applying Markov's inequality to the random variable $Y = e^{\lambda(X-\mu)}$ yields*

$$\mathbb{P}((X - \mu) \geq t) \leq \frac{\mathbb{E}[e^{\lambda(X-\mu)}]}{e^{\lambda t}}, \quad \text{for all } \lambda \in (-b, b).$$

Optimizing the choice of λ for the tightest bound yields the Chernoff bound

$$\mathbb{P}((X - \mu) \geq t) \leq \inf_{\lambda \in [0,b)} \frac{\mathbb{E}[e^{\lambda(X-\mu)}]}{e^{\lambda t}}.$$

Theorem 1.5 (Hoeffding bound for bounded variable). *Consider a random variable X with mean $\mu = \mathbb{E}(X)$. Assume that X is bounded and $X \in [a, b]$ almost surely, where a, b are two constants. Then, for any $\lambda \in \mathbb{R}$, we have*

$$\mathbb{E}[e^{\lambda X}] \leq e^{\frac{s(b-a)^2}{8}}.$$

Particularly, the variable $X \sim subG(\frac{(b-a)^2}{4})$.

Proof. See Exercise 2.4. □

Theorem 1.6 (Moment of sub-Gaussian variable)**.** *Let $X \sim sG(\sigma^2)$. For all integer $k \geq 1$, we have*

$$\mathbb{E}[|X|^k] \leq k2^{k/2}\sigma^k\Gamma(\frac{k}{2}), \tag{3}$$

where the Gamma function is defined as $\Gamma(x) = \int_0^{+\infty} t^{x-1}e^{-t}dt$.

Theorem 1.7 (One-sided Bernstein's inequality)**.** *Let X be a random variable. If $X \leq b$ almost surely, then*

$$\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \leq exp\left\{\frac{\lambda^2\mathbb{E}[X^2]/2}{1-b\lambda/3}\right\}, \quad \textit{for all } \lambda \in [0, 3/b).$$

Consequently, suppose there are n independent variables $X_i \leq b$ almost surely, for all $i \in [n]$. We have

$$\mathbb{P}[\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \geq n\delta] \leq exp\left\{-\frac{n\delta^2}{\sum_{i=1}^n \mathbb{E}[X_i^2]/n + b\delta/3}\right\}, \quad \textit{for all } \delta \geq 0. \tag{4}$$

Particularly, suppose there are n independent non-negative variables $Y_i \geq 0$, for all $i \in [n]$. The equation (4) becomes

$$\mathbb{P}[\sum_{i=1}^n (Y_i - \mathbb{E}[Y_i]) \leq n\delta] \leq exp\left\{-\frac{n\delta^2}{\sum_{i=1}^n \mathbb{E}[Y_i^2]/n}\right\}, \quad \textit{for all } \delta \geq 0. \tag{5}$$

2 Exercises

2.1 Exercise 2.1

(Tightness of inequalities.) The Markov's and Chebyshev's inequalities can not be improved in general.

(a) Provide a random variable $X \geq 0$ that attains the equality in Markov's inequality (1).

(b) Provide a random variable Y that attains the equality in Chebyshev's inequality (2).

Solution:

(a) For a given constant $t > 0$, we define a variable $Y_t = X - t\boldsymbol{I}_{[X \geq t]}$, where $\boldsymbol{I}$ is the indicator function. Noted that Y_t is a nonnegative variable, the Markov's inequality follows by taking expectation to Y,

$$\mathbb{E}[Y_t] = \mathbb{E}[X] - t\mathbb{P}[X \geq t] \geq 0.$$

Therefore, Markov's inequality meets the equality if and only if the expectation $\mathbb{E}[Y_t] = 0$. Since Y_t is nonnegative, we have $\mathbb{P}(Y_t = 0) = 1$. Note that $Y_t = 0$ if and only if $X = 0$ or $X = t$.

Hence, for the given constant $t > 0$, the nonnegative variable X with distribution $\mathbb{P}(X \in \{0, t\}) = 1$ attains the equality of Markov's inequality.

(b) Chebyshev's inequality follows by applying Markov's inequality to the non-negative random variable $Z = (X - \mathbb{E}[X])^2$. Similarly as part (a), given a constant $t > 0$, the variable $Z = (X - \mathbb{E}[X])^2$ with distribution $\mathbb{P}(Z \in \{0, t^2\}) = 1$ attains the equality of the Markov's inequality for Z and the Chebyshev's inequality for X. By transformation, the distribution of X satisfies the followings formula,

$$\mathbb{P}(X = x) = \begin{cases} p & \text{if } x = c, \\ \frac{1-p}{2} & \text{if } x = c - t \text{ or } x = c + t, \\ 0 & \text{ otherwise }, \end{cases}$$

where $c \in \mathbb{R}$ is a constant and $p \in [0, 1]$.

Remark 1 (Tightness of Markov's inequality)**.** Only a few variables can attain the equalities in Markov's and Chebyshev's inequalities. In research, we should pay attention to the concentration bounds tighter than Markov's inequality.

2.2 Exercise 2.2

Lemma 1 (Standard normal distribution)**.** *Let $\phi(z)$ be the density function of a standard normal variable $Z \sim N(0, 1)$. Then,*

$$\phi'(z) + z\phi(z) = 0, \tag{6}$$

and

$$\phi(z)(\frac{1}{z} - \frac{1}{z^3}) \leq \mathbb{P}(Z \geq z) \leq \phi(z)(\frac{1}{z} - \frac{1}{z^3} + \frac{3}{z^5}), \quad \textit{for all } z > 0. \tag{7}$$

Proof. First, we prove the equation (6).
The pdf of standard normal distribution is

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{z^2}{2}).$$

The equation (6) follows by taking the derivative of $\phi(z)$.

$$\phi'(z) = -z\frac{1}{\sqrt{2\pi}} \exp(-\frac{z^2}{2}) = -z\phi(z).$$

Next, we prove the equation (7).
We write the upper tail probability of the standard normal variable as

$$\mathbb{P}(Z \geq z) = \int_z^{+\infty} \phi(t)dt = \int_z^{+\infty} -\frac{1}{t}\phi'(t)dt = \frac{1}{z}\phi(z) - \int_z^{+\infty} \frac{1}{t^2}\phi(t)dt, \tag{8}$$

where the second equality follows by the equation (7). Applying the equation (7) to the last term in equation (8) yields

$$\int_z^{+\infty} \frac{1}{t^2}\phi(t)dt = \int_z^{+\infty} \frac{1}{t^3}\phi'(t)dt = -\frac{1}{z^3}\phi(z) + \int_z^{+\infty} \frac{3}{t^4}\phi(t)dt \geq -\frac{1}{z^3}\phi(z) \tag{9}$$

Plugging the equation (9) into the equation (8) , we have $\mathbb{P}(Z \geq z) \geq \phi(z)(\frac{1}{z} - \frac{1}{z^3})$. On the other hand, applying the equation (7) again to the equation (9) yields

$$\int_z^{+\infty} \frac{3}{t^4}\phi(t)dt = \int_z^{+\infty} -\frac{3}{t^5}\phi'(t)dt = \frac{3}{z^5}\phi(z) - \int_z^{+\infty} \frac{15}{t^6}\phi(t)dt \leq \frac{3}{z^5}\phi(z). \tag{10}$$

Combing equations (8), (9), and (10), we have $\mathbb{P}(Z \geq z) \leq \phi(z)(\frac{1}{z} - \frac{1}{z^3} + \frac{3}{z^5})$. □

Remark 2. Direct calculation of tail probability for a univariate normal variable is hard. Equation (7) provides a numerical approximation to the tail probability. Particularly, the tail probability decays at the rate of $z^{-1}e^{-z^2/2}$ as $z \to +\infty$. The decay rate is faster than polynomial rate $\mathcal{O}(z^{-\alpha})$, for any $\alpha \geq 1$.

2.3 Exercise 2.3

Lemma 2 (Polynomial bound and Chernoff bound)**.** *Let $X \geq 0$ be a non-negative variable. Suppose that the moment generating function of X, denoted $\varphi_X(\lambda)$, exists in the neighborhood of $\lambda = 0$. Given some $\delta > 0$, we have*

$$\inf_{k \in \mathbb{Z}_+} \frac{\mathbb{E}[|X|^k]}{\delta^k} \leq \inf_{\lambda > 0} \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda \delta}}. \tag{11}$$

Consequently, an optimized bound based on polynomial moments is always at least as good as the Chernoff upper bound.

Proof. By power series, we have

$$e^{\lambda X} = \sum_{k=0}^{+\infty} \frac{X^k \lambda^k}{k!}, \qquad \text{for all } \lambda \in \mathbb{R}. \tag{12}$$

Since the moment generating function $\varphi_X(\lambda)$ exists in the neighborhood of $\lambda = 0$, there exists a constant $b > 0$ such that

$$\mathbb{E}[e^{\lambda X}] = \sum_{k=0}^{+\infty} \frac{\mathbb{E}[|X|^k] \lambda^k}{k!} < +\infty, \qquad \text{for all } \lambda \in (0, b).$$

Therefore, the moment $\mathbb{E}[|X|^k]$ exists for all $k \in \mathbb{Z}_+$. Applying power series (12) to the right hand side of equation (11) yields

$$\inf_{\lambda > 0} \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda \delta}} = \frac{\sum_{k=0}^{+\infty} \frac{\mathbb{E}[|X|^k] \lambda^k}{k!}}{\sum_{k=0}^{+\infty} \frac{\lambda^k \delta^k}{k!}}. \tag{13}$$

By Cauchy's third inequality, we have

$$\frac{\sum_{k=0}^{+\infty} \frac{\mathbb{E}[|X|^k] \lambda^k}{k!}}{\sum_{k=0}^{+\infty} \frac{\lambda^k \delta^k}{k!}} \geq \inf_{k \in \mathbb{Z}_+} \frac{\mathbb{E}[|X|^k]}{\delta^k} \tag{14}$$

Therefore, the equation (11) follows by combining the equation (13) with equation (14). □

Remark 3. Applying different functions $g(X)$ to the Markov's inequality leads to different bounds for the tail probability of variable X. Equation (11) implies that optimized polynomial bound is tighter than the Chernoff bound, provided that the moment generating function of X exsits.

2.4 Exercise 2.4

In Exercise 2.4, we prove the Theorem 1.5, the Hoeffding bound for a bounded variable.

Proof. Suppose X is a bounded random variable, where $X \in [a, b]$ almost surely, and $a \leq b \in \mathbb{R}$ are two constants. Let $\mu = \mathbb{E}[X]$. Define the function

$$g(\lambda) = \log \mathbb{E}[e^{\lambda X}], \qquad \text{for all } \lambda \in \mathbb{R}.$$

Applying Taylor Expansion to $g(\lambda)$ at 0, we have

$$g(\lambda) = g(0) + g'(0)\lambda + \frac{g''(\lambda_0)}{2}\lambda^2, \quad \text{where } \lambda_0 = t\lambda, \text{ for some } t \in [0, 1]. \tag{15}$$

In equation (15), the term $g(0) = \log \mathbb{E}[e^0] = 0$. By power series (12), we calculate the first derivative $g'(\lambda)$ as following,

$$\begin{aligned} g'(\lambda) &= \left(\log \sum_{k=0}^{n} \frac{\lambda^k}{k!}\mathbb{E}[X^k]\right)' \\ &= \sum_{k=0}^{n} \frac{\lambda^k}{k!}\mathbb{E}[X^{(k+1)}] \bigg/ \sum_{k=0}^{n} \frac{\lambda^k}{k!}\mathbb{E}[X^k] \\ &= \frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]}. \end{aligned} \tag{16}$$

Therefore, $g'(0) = \mathbb{E}[X] = \mu$. Taking the derivative to equation (16), we calculate the second-order derivative $g''(\lambda)$ as following,

$$\begin{aligned} g''(\lambda) &= \sum_{k=0}^{n} \frac{\lambda^k}{k!}\mathbb{E}[X^{(k+2)}] \bigg/ \sum_{k=0}^{n} \frac{\lambda^k}{k!}\mathbb{E}[X^k] - \left(\sum_{k=0}^{n} \frac{\lambda^k}{k!}\mathbb{E}[X^{(k+1)}] \bigg/ \sum_{k=0}^{n} \frac{\lambda^k}{k!}\mathbb{E}[X^k]\right)^2 \\ &= \frac{\mathbb{E}[X^2e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} - \left(\frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]}\right)^2. \end{aligned}$$

We interpret the second-order derivative $g''(\lambda)$ interpreted as the variance of X with the re-weighted distribution $dP' = e^{\lambda X}/\mathbb{E}[e^{\lambda X}]dP_X$, where P_X is the distribution of X. Taking integral of 1 with respect to dP', we have

$$\int_{-\infty}^{+\infty} dP' = \int_{-\infty}^{+\infty} \frac{e^{\lambda X}}{\mathbb{E}[e^{\lambda X}]} dP_X = 1,$$

which implies that the function P' is indeed a distribution. Under all possible re-weighted distributions, the variance of X is upper bounded as following,

$$var(X) = var(X - \frac{a+b}{2}) \leq \mathbb{E}[(X - \frac{a+b}{2})^2] \leq \frac{(b-a)^2}{4},$$

where the term $\frac{(b-a)^2}{4}$ follows by letting X distribute only on the boundary a and b. Hence, the second-order derivative $g''(\lambda) \leq \frac{(b-a)^2}{4}$. We plug the results of g' and g'' into the equation (15). Then,

$$g(\lambda) = g(0) + g'(0)\lambda + \frac{g''(\lambda_0)}{2}\lambda^2 \leq 0 + \lambda\mu + \frac{(b-a)^2}{8}\lambda^2. \tag{17}$$

Taking the exponential on both sides of the inequality (17), we have

$$\mathbb{E}[e^{\lambda X}] = \exp(g(\lambda)) \leq e^{\mu\lambda + \frac{(b-a)^2}{8}\lambda^2}. \tag{18}$$

The equation (18) implies that X is a sub-Gaussian variable with at most $\sigma = \frac{(b-a)}{2}$. □

Remark 4. For any bounded random variable X supported on $[a, b]$, X is a sub-gaussian variable with parameter at most $\sigma^2 = (b-a)^2/4$. All the properties for sub-Gaussian variables apply to the bounded variables.

2.5 Exercise 2.5

Lemma 3 (Sub-Gaussian bounds and means/variance)**.** *Let X be a random variable such that*

$$\mathbb{E}[e^{\lambda X}] \le e^{\frac{\lambda^2\sigma^2}{2}+\mu\lambda}, \quad \text{for all } \lambda \in \mathbb{R}. \tag{19}$$

Then, $\mathbb{E}[X] = \mu$ and $var(X) \le \sigma^2$.

Proof. By equation (19), the moment generating function of X, denoted $\varphi_X(\lambda)$, exists in the neighborhood of $\lambda = 0$. Hence, the mean and variance of X exist. For all λ in the neighborhood of $\lambda = 0$, applying power series on both sides of equation (19) yields

$$\lambda\mathbb{E}[X] + \frac{\lambda^2}{2}\mathbb{E}[X^2] + o(\lambda^2) \le \mu\lambda + \frac{\lambda^2\sigma^2 + \lambda^2\mu^2}{2} + o(\lambda^2). \tag{20}$$

Dividing by $\lambda > 0$ on both sides of equation (20) and letting $\lambda \to 0^+$, we have $\mathbb{E}(X) \le \mu$. Dividing by $\lambda < 0$ on both sides of equation (20) and letting $\lambda \to 0^-$, we have $\mathbb{E}(X) \ge \mu$. Therefore, the mean $\mathbb{E}[X] = \mu$. Then, we divide $2/\lambda^2$ on both sides of equation (20). The term $\mathbb{E}[X]\lambda$ and $\mu\lambda$ are cancelled. We have $\mathbb{E}[X^2] \le \sigma^2 + \mu^2$, and thus $var(X) \le \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \sigma^2$. □

Question: Let σ^2_{min} denote the smallest possible σ satisfying the inequality (19). Is it true that $var(X) = \sigma^2_{min}$?

Solution: The statement that $var(X) = \sigma^2_{min}$ is not necessarily true. Recall the function $g(\lambda)$ in exercise 2.4. By the results in exercise 2.4, the equation (19) is equal to

$$g''(\lambda) \le \sigma^2, \quad \text{for all } \lambda \in \mathbb{R},$$

where $g''(\lambda)$ is the variance of X with with the re-weighted distribution $dP' = e^{\lambda X}/\mathbb{E}[e^{\lambda X}]dP_X$, where P_X is the distribution of X and $g''(0) = var(X)$. Therefore, $\max_\lambda g''(\lambda) = \sigma^2_{min}$. To let the equality $var(X) = \sigma^2_{min}$ hold, we need to show that $\max_\lambda g''(\lambda) = g''(0)$ holds for X.

However, the statement $\max_\lambda g''(\lambda) = g''(0)$ is not always true. A counter example is below. Consider a random variable $Y \sim Ber(1/3)$. The variance of Y is $var(Y) = 2/9$. Let $\lambda = 1$. The re-weighted distribution dP' is

$$P'(Y=0) = \frac{2}{3\mathbb{E}[e^Y]} \quad \text{and} \quad P'(Y=1) = \frac{e}{3\mathbb{E}[e^Y]}, \quad \text{where } \mathbb{E}[e^Y] = \frac{2}{3} + \frac{e}{3}.$$

The variance of Y with dP' is $2/3\mathbb{E}[e^Y] \times e/3\mathbb{E}[e^Y] = 0.2442 > 2/9$. Therefore, for this variable Y, we have $var(Y) < g''(1) \le \max_\lambda g''(\lambda) = \sigma^2_{min}$.

Remark 5. Parameters of a sub-Gaussian distribution provide the exact value of the mean, $\mathbb{E}[X] = \mu$, and an upper bound of the variance, $var(X) \le \sigma^2$. For any variable X whose moment generating function exists, the tail distribution of X can be bounded by a sub-Gaussian distribution with a proper choice of σ^2.

2.6 Exercise 2.6

Lemma 4 (Lower bounds on squared sub-Gaussians)**.** *Let $\{X_i\}_{i=1}^n$ be an i.i.d. sequence of zero-mean sub-Gaussian variables with parameter σ. The normalized summation $Z_n = \frac{1}{n}\sum_{i=1}^n X_i^2$ satisfies*

$$\mathbb{P}[Z_n - \mathbb{E}[Z_n] \le \sigma^2\delta] \le e^{-n\delta^2/16}, \quad \text{for all } \delta \ge 0. \tag{21}$$

The equation (21) implies that the lower tail of the summation of squared sub-Gaussian variables behave in a sub-Gaussian way.

Proof. Since X_i^2 are i.i.d. nonnegative variables, we apply the equation (5) in Theorem 1.7 to the variables X_i^2, where $i \in [n]$. Then,

$$\mathbb{P}[\sum_{i=1}^{n}(X_i^2 - \mathbb{E}[X_i^2]) \leq n\sigma^2\delta] \leq exp\left\{-\frac{n\delta^2\sigma^4}{\mathbb{E}[X_1^4]}\right\}, \quad \text{for all } \delta \geq 0. \tag{22}$$

By the equation (3) in Theorem 1.6, we have

$$\mathbb{E}[X_1^4] \leq 16\sigma^4. \tag{23}$$

Combing equations (22), (23), and the definition of Z_n, we have

$$\mathbb{P}[Z_n - \mathbb{E}[Z_n] \leq \sigma^2\delta] \leq exp\left\{-\frac{n\delta^2}{16}\right\}, \quad \text{for all } \delta \geq 0.$$

□

Remark 6. Equation (21) implies that the lower tail of the summation of squared sub-Gaussian variables behave in a sub-Gaussian way. In following sections, we will show that the variable $Z_n - \mathbb{E}[Z_n]$ in Lemma 4 is a sub-exponential variable.