

Solution to “Chapter 2: Basic tail and concentration bounds”

Jiixin Hu

July 23, 2020

1 Summary

Theorem 1.1 (Markov’s inequality). *Let $X \geq 0$ be a random variable with a finite mean. We have*

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

Theorem 1.2 (Chebyshev’s inequality). *Let $X \geq 0$ be a random variable with a finite mean μ and a finite variance. We have*

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

Theorem 1.3 (Markov’s inequality for polynomial moments). *Let X be a random variable. Suppose that the order k central moment of X exists. 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). *Let X be a random variable. Suppose that the moment generating function of X , denoted $\varphi_X(\lambda)$, exists 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, we obtain 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). *Let X be a random variable with $\mu = \mathbb{E}(X)$. Suppose that $X \in [a, b]$ almost surely, where $a \leq b \in \mathbb{R}$ are two constants. Then, we have*

$$\mathbb{E}[e^{\lambda X}] \leq e^{\frac{s(b-a)^2}{8}}, \quad \text{for all } \lambda \in \mathbb{R}.$$

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

Proof. See Exercise 2.4. □

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

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

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. Suppose $X \leq b$ almost surely. We have*

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

Consequently, let X_i be independent variables, and $X_i \leq b$ almost surely, for all $i \in [n]$. We have

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

Particularly, let X_i be independent nonnegative variables, for all $i \in [n]$. The equation (4) becomes

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

Definition 1 (Bernstein's condition). Let X be a random variable with mean $\mu = \mathbb{E}[X]$ and variance $\sigma^2 = \text{var}(X)$. We say X satisfies the Bernstein's condition with parameter b if

$$\left| \mathbb{E}[(X - \mu)^k] \right| \leq \frac{1}{2} k! \sigma^2 b^{k-2}, \quad \text{for } k = 3, 4, \dots$$

Note that bounded random variables satisfy the Bernstein's condition.

Theorem 1.8 (Bernstein-type bound). *For any variable X satisfying the Bernstein's condition, we have*

$$\mathbb{E} \left[e^{\lambda(X - \mu)} \right] \leq \exp \left\{ \frac{\lambda^2 \sigma^2}{2(1 - b|\lambda|)} \right\}, \quad \text{for all } |\lambda| \leq \frac{1}{b},$$

and the concentration inequality

$$\mathbb{P} [|X - \mu| \geq t] \leq 2 \exp \left\{ -\frac{t^2}{2(\sigma^2 + bt)} \right\}, \quad \text{for all } t \geq 0.$$

2 Exercises

2.1 Exercise 2.1

(Tightness of inequalities.) The Markov's and Chebyshev's inequalities are not able to 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 $Y_t = X - t\mathbf{1}[X \geq t]$, where $\mathbf{1}$ is the indicator function. Note that Y_t is a nonnegative variable. The Markov's inequality follows by taking the expectation to Y_t ,

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

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 nonnegative random variable $Z = (X - \mathbb{E}[X])^2$. Simialrly as in 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 . Consequently, the variable X attains the equality of 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 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

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

Proof. First, we prove the equation (6).

The pdf of the standard normal distribution is

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

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

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

Next, we prove the equation (7).

We write the upper tail probability of the standard normal variable as

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

where the second equality follows by the equation (6). Applying the equation (6) to the last term in equation (8) yields

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

Plugging the equation (9) into the equation (8), we obtain $\mathbb{P}(Z \geq z) \geq \phi(z) \left(\frac{1}{z} - \frac{1}{z^3} \right)$. Applying the equation (6) again to the equation (9) yields

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

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

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 $z \rightarrow +\infty$. 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 nonnegative 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*

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

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

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

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

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

Hence, 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

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

By Cauchy's third inequality, we have

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

Therefore, we obtain the equation (11) by combining the equation (13) with equation (14). \square

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 the optimized polynomial bound is at least as tight as the Chernoff bound, provided that the moment generating function of X exists in the neighborhood of 0.

2.4 Exercise 2.4

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

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

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

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

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

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

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

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

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

We interpret the second-order derivative $g''(\lambda)$ 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 the 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 a valid probability distribution. Under all possible re-weighted distributions, the variance of X is upper bounded as follows,

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

where the term $\frac{(b-a)^2}{4}$ follows by letting X supported on the boundaries a and b only. 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,

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

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

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

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

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*

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

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

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

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

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

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

Solution: The statement that $\text{var}(X) = \sigma_{\min}^2$ 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

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

where $g''(\lambda)$ is the variance of X with the re-weighted distribution defined in Exercise 2.4. Therefore, we have $\max_{\lambda} g''(\lambda) = \sigma_{\min}^2$. Note that $g''(0) = \text{var}(X)$. To let the equality $\text{var}(X) = \sigma_{\min}^2$ 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 necessarily true. A counter example is below. Consider a random variable $Y \sim \text{Ber}(1/3)$. The variance of Y is $\text{var}(Y) = 2/9$. 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, we have $\text{var}(Y) < g''(1) \leq \max_{\lambda} g''(\lambda) = \sigma_{\min}^2$. The statement $\max_{\lambda} g''(\lambda) = g''(0)$ is not true for this variable Y .

Remark 5. Parameters of a sub-Gaussian distribution provide the exact value of the mean and an upper bound of the variance; i.e., $\mathbb{E}[X] = \mu$ and $\text{var}(X) \leq \sigma^2$. Suppose the moment generating function of variable X exists over the entire real interval. Then, the tail distribution of X is 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 sum $Z_n = \frac{1}{n} \sum_{i=1}^n X_i^2$ satisfies*

$$\mathbb{P}[Z_n - \mathbb{E}[Z_n] \leq \sigma^2 \delta] \leq e^{-n\delta^2/16}, \quad \text{for all } \delta \geq 0. \quad (21)$$

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

Proof. Since X_i^2 are i.i.d. nonnegative variables, we apply the equation (5) to the variables $\{X_i^2\}_{i=1}^n$. Then, we have

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

By equation (3), we have

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

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

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

□

Remark 6. Equation (21) implies that the lower tail of the sum of squared sub-Gaussian variables behaves 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.

2.7 Exercise 2.7

Lemma 5 (Bennett's inequality). *Let X_1, \dots, X_n be a sequence of independent zero-mean random variables with $|X_i| \leq b$ and $\text{var}(X_i) = \sigma_i^2$, for all $i \in [n]$. Then, we have the Bennett's inequality*

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

where $\sigma^2 = \frac{1}{n} \sum_{i=1}^n \sigma_i^2$ and $h(t) := (1+t) \log(1+t) - t$ for $t \geq 0$.

Proof. First, we consider the moment generating function of X_i , for all $i \in [n]$.

By power series, for all $i \in [n]$, we have

$$\mathbb{E} \left[e^{\lambda X_i} \right] = \sum_{k=0}^{+\infty} \frac{\lambda^k \mathbb{E}[X_i^k]}{k!} = 1 + 0 + \sum_{k=2}^{+\infty} \frac{\lambda^k \mathbb{E}[X_i^k]}{k!} \leq \exp \left\{ \sum_{k=2}^{+\infty} \frac{\lambda^k \mathbb{E}[X_i^k]}{k!} \right\}, \quad (24)$$

where the 0 comes from the fact that $\mathbb{E}[X_i] = 0$, and the last inequality follows from $1 + x \leq e^x$. By $|X_i| < b$, we bound the last term in equation (24) as follows

$$\sum_{k=2}^{+\infty} \frac{\lambda^k \mathbb{E}[X_i^k]}{k!} \leq \sum_{k=2}^{+\infty} \frac{\lambda^k \mathbb{E}[X_i^2 |X_i|^{k-2}]}{k!} \leq \sum_{k=2}^{+\infty} \frac{\lambda^k \sigma_i^2 b^{k-2}}{k!} = \sigma_i^2 \left(\frac{e^{\lambda b} - 1 - \lambda b}{b^2} \right). \quad (25)$$

Combing the equation (24) with equation (25), we obtain the following upper bound of the moment generating function of $\sum_{i=1}^n X_i$.

$$\mathbb{E} \left[e^{\lambda \sum_{i=1}^n X_i} \right] = \prod_{i=1}^n \mathbb{E} \left[e^{\lambda X_i} \right] \leq \exp \left\{ n\sigma^2 \left(\frac{e^{\lambda b} - 1 - \lambda b}{b^2} \right) \right\}, \quad (26)$$

where $\sigma^2 = \frac{1}{n} \sum_{i=1}^n \sigma_i^2$. Combing the Chernoff bound with equation (26), the upper tail of $\sum_{i=1}^n X_i$ follows

$$\begin{aligned} \mathbb{P} \left[\sum_{i=1}^n X_i \geq n\delta \right] &\leq \exp \left\{ n\sigma^2 \left(\frac{e^{\lambda b} - 1 - \lambda b}{b^2} \right) - \lambda n\delta \right\} \\ &= \exp \left\{ \frac{n\sigma^2}{b^2} \left(e^{\lambda b} - \lambda b - \lambda \frac{\delta b^2}{\sigma^2} - 1 \right) \right\}, \quad \text{for all } \delta \geq 0. \end{aligned} \quad (27)$$

The upper bound (27) achieves the minimum when $\lambda = b^{-1} \log \left(1 + \frac{\delta b}{\sigma^2} \right)$ by the first-order condition of minimization. Plugging $\lambda = b^{-1} \log \left(1 + \frac{\delta b}{\sigma^2} \right)$ into the equation (27), we obtain the Bennett's inequality

$$\mathbb{P} \left[\sum_{i=1}^n X_i \geq n\delta \right] \leq \exp \left\{ -\frac{n\sigma^2}{b^2} h \left(\frac{b\delta}{\sigma^2} \right) \right\}, \quad \text{for all } \delta \geq 0, \quad (28)$$

where $h(t) := (1+t) \log(1+t) - t$ for $t \geq 0$.

Further, we show that the Bennett's inequality is at least as good as the Bernstein's inequality.

The Bernstein's inequality for $\sum_{i=1}^n X_i$ is

$$\mathbb{P} \left[\sum_{i=1}^n X_i \geq n\delta \right] \leq \exp \left\{ \frac{-3n\delta^2}{(2b\delta + 6\sigma^2)} \right\} = \exp \left\{ -\frac{n\sigma^2}{b^2} g \left(\frac{b\delta}{\sigma^2} \right) \right\}, \quad \text{for all } \delta \geq 0, \quad (29)$$

where $g(t) := \frac{3t^2}{2t+6}$ for $t \geq 0$. Since $g(t) \leq h(t)$ holds for all $t \geq 0$, we conclude that the Bennett's inequality (28) is at least as good as Bernstein's inequality (29). \square

Remark 7. So far, we have three inequalities controlling the tail of bounded variables: Hoeffding's inequality, Bernstein's inequality, and Bennett's inequality. Particularly, Hoeffding's inequality implies the sub-Gaussianity of bounded variables. As the proof for Lemma 5 shows, Bennett's inequality is at least as good as the Bernstein's inequality, for bounded random variables.