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

Jiaxin Hu

August 7, 2020

1 Summary

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

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

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*

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

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*

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

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*

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

Optimizing the choice of λ for the tightest bound, we obtain the Chernoff bound

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

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 $X \sim \operatorname{subG}\left(\frac{(b-a)^2}{4}\right)$.

Proof. See Exercise 2.4. $\square$

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

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

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

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

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

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

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

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

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

$$\left|\mathbb{E}[(X-\mu)^k]\right| \leq \frac{1}{2} k! \sigma^2 b^{k-2}, \quad \text{for } k = 3, 4, ... \tag{6}$$

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*

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

and the concentration inequality

$$\mathbb{P}\left[|X-\mu| \geq t\right] \leq 2\exp\left\{-\frac{t^2}{2(\sigma^2 + bt)}\right\}, \quad \textit{for all } t \geq 0. \tag{7}$$

Definition 2 (Bounded difference property)**.** Let $f \colon \mathbb{R}^n \mapsto \mathbb{R}$ be a function. The function f satisfies the bounded difference property with parameter $(L_1, ..., L_n)$ if we have

$$\left|f(x^{(k)}) - f(x^{'(k)})\right| \leq L_k, \tag{8}$$

for all $k \in [n]$ and for all $x^{(k)} = (x_1, ..., x_k, ..., x_n), x^{'(k)} = (x_1, ..., x_k', ..., x_n) \in \mathbb{R}^n$.

Theorem 1.9 (Bounded differences inequality)**.** *Let $f \colon \mathbb{R}^n \mapsto \mathbb{R}$ be a function satisfies the bounded difference property (8), and the random variable $X = (X_1, ..., X_n)$ has independent components. Then,*

$$\mathbb{P}\left[|f(X) - \mathbb{E}[f(X)]| \geq t\right] \leq 2e^{-\frac{2t^2}{\sum_{i=1}^n L_i^2}}, \quad \textit{for all } t \geq 0.$$

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 $Y_t = X - t\mathbb{1}[X \geq t]$, where $\mathbb{1}$ 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,*

$$\phi'(z) + z\phi(z) = 0, \tag{9}$$

and

$$\phi(z)\left(\frac{1}{z} - \frac{1}{z^3}\right) \leq \mathbb{P}(Z \geq z) \leq \phi(z)\left(\frac{1}{z} - \frac{1}{z^3} + \frac{3}{z^5}\right), \quad \textit{for all } z > 0. \tag{10}$$

Proof. First, we prove the equation (9).
The pdf of the standard normal distribution is

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

The equation (9) 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 (10).
We write the upper tail probability of the standard normal variable as

$$\mathbb{P}(Z \geq z) = \int_z^{+\infty} \phi(t)dt = \int_z^{+\infty} -\frac{1}{t}\phi'(t)dt = \frac{1}{z}\phi(z) - \int_z^{+\infty} \frac{1}{t^2}\phi(t)dt, \tag{11}$$

where the second equality follows by the equation (9). Applying the equation (9) to the last term in equation (11) yields

$$\int_z^{+\infty} \frac{1}{t^2}\phi(t)dt = \int_z^{+\infty} \frac{1}{t^3}\phi'(t)dt = -\frac{1}{z^3}\phi(z) + \int_z^{+\infty} \frac{3}{t^4}\phi(t)dt \geq -\frac{1}{z^3}\phi(z) \tag{12}$$

Plugging the equation (12) into the equation (11), we obtain $\mathbb{P}(Z \geq z) \geq \phi(z)\left(\frac{1}{z} - \frac{1}{z^3}\right)$. Applying the equation (9) again to the equation (12) yields

$$\int_z^{+\infty} \frac{3}{t^4}\phi(t)dt = \int_z^{+\infty} -\frac{3}{t^5}\phi'(t)dt = \frac{3}{z^5}\phi(z) - \int_z^{+\infty} \frac{15}{t^6}\phi(t)dt \leq \frac{3}{z^5}\phi(z). \tag{13}$$

Combing equations (11), (12) and (13), we obtain $\mathbb{P}(Z \geq z) \leq \phi(z)\left(\frac{1}{z} - \frac{1}{z^3} + \frac{3}{z^5}\right)$. □

Remark 2. Direct calculation of tail probability for a univariate normal variable is hard. Equation (10) 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 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*

$$\inf_{k\in\mathbb{Z}_+} \frac{\mathbb{E}[|X|^k]}{\delta^k} \leq \inf_{\lambda>0} \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda\delta}}. \tag{14}$$

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

$$e^{\lambda X} = \sum_{k=0}^{+\infty} \frac{X^k\lambda^k}{k!}, \qquad \text{for all } \lambda \in \mathbb{R}. \tag{15}$$

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

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

Hence, the moment $\mathbb{E}[|X|^k]$ exists, for all $k \in \mathbb{Z}_+$. Applying power series (15) to the right hand side of equation (14) yields

$$\inf_{\lambda>0} \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda\delta}} = \frac{\sum_{k=0}^{+\infty} \frac{\mathbb{E}[|X|^k]\lambda^k}{k!}}{\sum_{k=0}^{+\infty} \frac{\lambda^k\delta^k}{k!}}. \tag{16}$$

By Cauchy's third inequality, we have

$$\frac{\sum_{k=0}^{+\infty} \frac{\mathbb{E}[|X|^k]\lambda^k}{k!}}{\sum_{k=0}^{+\infty} \frac{\lambda^k\delta^k}{k!}} \geq \inf_{k\in\mathbb{Z}_+} \frac{\mathbb{E}[|X|^k]}{\delta^k} \tag{17}$$

Therefore, we obtain the equation (14) by combining the equation (16) with equation (17). □

Remark 3. Applying different functions $g(X)$ to the Markov's inequality leads to different bounds for the tail probability of variable X. Equation (14) implies that the optimized polynomial bound is at least as tight as the Chernoff bound, provided that the moment generating function of X exsits 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

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

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

$$g(\lambda) = g(0) + g'(0)\lambda + \frac{g''(\lambda_0)}{2}\lambda^2, \quad \text{where } \lambda_0 = t\lambda, \text{ for some } t \in [0, 1]. \tag{18}$$

In equation (18), the term $g(0) = \log \mathbb{E}[e^0] = 0$. By power series (15), we obtain the first derivative $g'(\lambda)$ as follows,

$$\begin{aligned} g'(\lambda) &= \left(\log \sum_{k=0}^{n} \frac{\lambda^k}{k!}\mathbb{E}[X^k]\right)' \\ &= \sum_{k=0}^{n} \frac{\lambda^k}{k!}\mathbb{E}[X^{(k+1)}] \Bigg/ \sum_{k=0}^{n} \frac{\lambda^k}{k!}\mathbb{E}[X^k] \\ &= \frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]}. \end{aligned} \tag{19}$$

Therefore, $g'(0) = \mathbb{E}[X] = \mu$. Taking the derivative to equation (19), we obtain the second-order derivative $g''(\lambda)$ as follows,

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

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,

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

where the term $\frac{(b-a)^2}{4}$ follows by letting X 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 (18). Then,

$$g(\lambda) = g(0) + g'(0)\lambda + \frac{g''(\lambda_0)}{2}\lambda^2 \leq 0 + \lambda\mu + \frac{(b-a)^2}{8}\lambda^2. \tag{20}$$

Taking the exponentiation on both sides of the inequality (20), we have

$$\mathbb{E}[e^{\lambda X}] = \exp(g(\lambda)) \leq e^{\mu\lambda + \frac{(b-a)^2}{8}\lambda^2}. \tag{21}$$

The equation (21) 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*

$$\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\lambda^2\sigma^2}{2} + \mu\lambda}, \quad \textit{for all } \lambda \in \mathbb{R}. \tag{22}$$

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

Proof. By equation (22), 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 (22) yields

$$\lambda\mathbb{E}[X] + \frac{\lambda^2}{2}\mathbb{E}[X^2] + o(\lambda^2) \leq \mu\lambda + \frac{\lambda^2\sigma^2 + \lambda^2\mu^2}{2} + o(\lambda^2). \tag{23}$$

Dividing by $\lambda > 0$ on both sides of equation (23) and letting $\lambda \to 0^+$, we have $\mathbb{E}(X) \le \mu$. Dividing by $\lambda < 0$ on both sides of equation (23) and letting $\lambda \to 0^-$, we have $\mathbb{E}(X) \ge \mu$. Therefore, we obtain the mean $\mathbb{E}[X] = \mu$. Then, we divide $2/\lambda^2$ on both sides of equation (23), for $\lambda \ne 0$. The term $\mathbb{E}[X]\lambda$ and $\mu\lambda$ are cancelled. We have $\mathbb{E}[X^2] \le \sigma^2 + \mu^2$, and thus the $\mathrm{var}(X) \le \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \sigma^2$. □

Question: Let σ^2_{min} denote the smallest possible σ satisfying the inequality (22). Is it true that $\mathrm{var}(X) = \sigma^2_{min}$?

Solution: The statement that $\mathrm{var}(X) = \sigma^2_{min}$ is not necessarily true. Recall the function $g(\lambda)$ in Exercise 2.4. By the results in Exercise 2.4, the equation (22) 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 the re-weighted distribution defined in Exercise 2.4. Therefore, we have $\max_\lambda g''(\lambda) = \sigma^2_{min}$. Note that $g''(0) = \mathrm{var}(X)$. To let the equality $\mathrm{var}(X) = \sigma^2_{min}$ 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 necessarily true. A counter example is below. Consider a random variable $Y \sim Ber(1/3)$. The variance of Y is $\mathrm{var}(Y) = 2/9$. 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 $\mathrm{var}(Y) < g''(1) \le \max_\lambda g''(\lambda) = \sigma^2_{min}$. 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 $\mathrm{var}(X) \le \sigma^2$. 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*

$$\mathbb{P}[Z_n - \mathbb{E}[Z_n] \le \sigma^2\delta] \le e^{-n\delta^2/16}, \quad \textit{for all } \delta \ge 0. \tag{24}$$

The equation (24) 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

$$\mathbb{P}\left[\sum_{i=1}^n (X_i^2 - \mathbb{E}[X_i^2]) \le n\sigma^2\delta\right] \le \exp\left\{-\frac{n\delta^2\sigma^4}{\mathbb{E}[X_1^4]}\right\}, \quad \text{for all } \delta \ge 0. \tag{25}$$

By equation (3), we have

$$\mathbb{E}[X_1^4] \le 16\sigma^4. \tag{26}$$

Combing equations (25), (26) 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 (24) 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 $X_1, ..., X_n$ 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*

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

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

$$\mathbb{E}\left[e^{\lambda X_i}\right] = \sum_{k=0}^{+\infty} \frac{\lambda^k \mathbb{E}[X_i^k]}{k!} = 1 + 0 + \sum_{k=2}^{+\infty} \frac{\lambda^k \mathbb{E}[X_i^k]}{k!} \leq \exp\left\{\sum_{k=2}^{+\infty} \frac{\lambda^k \mathbb{E}[X_i^k]}{k!}\right\}, \tag{27}$$

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 (27) as follows

$$\sum_{k=2}^{+\infty} \frac{\lambda^k \mathbb{E}[X_i^k]}{k!} \leq \sum_{k=2}^{+\infty} \frac{\lambda^k \mathbb{E}[X_i^2 |X_i|^{k-2}]}{k!} \leq \sum_{k=2}^{+\infty} \frac{\lambda^k \sigma_i^2 b^{k-2}}{k!} = \sigma_i^2\left(\frac{e^{\lambda b} - 1 - \lambda b}{b^2}\right). \tag{28}$$

Combing the equation (27) with equation (28), we obtain the following upper bound of the moment generating function of $\sum_{i=1}^{n} X_i$.

$$\mathbb{E}\left[e^{\lambda \sum_{i=1}^{n} X_i}\right] = \prod_{i=1}^{n} \mathbb{E}\left[e^{\lambda X_i}\right] \leq \exp\left\{n\sigma^2\left(\frac{e^{\lambda b} - 1 - \lambda b}{b^2}\right)\right\}, \tag{29}$$

where $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2$. Combing the Chernoff bound with equation (29), the upper tail of $\sum_{i=1}^{n} X_i$ follows

$$\begin{aligned}\mathbb{P}\left[\sum_{i=1}^{n} X_i \geq n\delta\right] &\leq \exp\left\{n\sigma^2\left(\frac{e^{\lambda b} - 1 - \lambda b}{b^2}\right) - \lambda n\delta\right\} \\ &= \exp\left\{\frac{n\sigma^2}{b^2}\left(e^{\lambda b} - \lambda b - \lambda\frac{\delta b^2}{\sigma^2} - 1\right)\right\}, \quad \text{for all } \delta \geq 0. \end{aligned} \tag{30}$$

The upper bound (30) 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 (30), we obtain the Bennett's inequality

$$\mathbb{P}\left[\sum_{i=1}^{n} X_i \geq n\delta\right] \leq \exp\left\{-\frac{n\sigma^2}{b^2} h\left(\frac{b\delta}{\sigma^2}\right)\right\}, \quad \text{for all } \delta \geq 0, \tag{31}$$

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

$$\mathbb{P}\left[\sum_{i=1}^{n} X_i \geq n\delta\right] \leq \exp\left\{\frac{-3n\delta^2}{(2b\delta + 6\sigma^2)}\right\} = \exp\left\{-\frac{n\sigma^2}{b^2} g\left(\frac{b\delta}{\sigma^2}\right)\right\}, \quad \text{for all } \delta \geq 0, \tag{32}$$

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 (31) is at least as good as Bernstein's inequality (32). $\square$

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.

2.8 Exercise 2.8

Lemma 6 (Bernstein and expectation)**.** *Let Z be a nonnegative random variable satisfying the following concentration inequality*

$$\mathbb{P}[Z \geq t] \leq C e^{-\frac{t^2}{2(\nu^2 + Bt)}}, \quad \textit{for all } t \geq 0, \tag{33}$$

where (ν, B) are two positive constants and $C \geq 1$. Then, the expectation of Z satisfies

$$\mathbb{E}[Z] \leq 2\nu(\sqrt{\pi} + \sqrt{\log C}) + 4B(1 + \log C). \tag{34}$$

Further, let $\{X_i\}_{i=1}^{n}$ be a sequence of i.i.d. zero-mean variables satisfying the Bernstein condition (6). The sample mean of $\{X_i\}_{i=1}^{n}$ satisfies

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right|\right] \leq 2\sigma(\sqrt{\pi} + \sqrt{\log 2}) + 4b(1 + \log 2). \tag{35}$$

Proof. First, we prove the equation (34).

By equation (33), we have

$$\begin{aligned}\mathbb{P}[Z \geq t] &\leq C \max\left\{\exp\left(-\frac{t^2}{4\nu^2}\right), \exp\left(-\frac{t^2}{4Bt}\right)\right\} \\ &\leq C\exp\left(-\frac{t^2}{4\nu^2}\right) + C\exp\left(-\frac{t^2}{4Bt}\right).\end{aligned} \tag{36}$$

Plugging the inequality (36) to $\mathbb{E}[Z] = \int_0^{+\infty} \mathbb{P}[Z \geq t]dt$, we have

$$\mathbb{E}[Z] = \int_0^{+\infty} \min\left\{1, C\exp\left(-\frac{t^2}{4\nu^2}\right)\right\} dt + \int_0^{+\infty} \min\left\{1, C\exp\left(-\frac{t^2}{4Bt}\right)\right\} dt := I_1 + I_2.$$

To evaluate I_1, we spilt the integral to avoid the minimization. Solving $1 = C\exp\left(-\frac{t^2}{4\nu^2}\right)$, the minimization term becomes

$$\min\left\{1, C\exp\left(-\frac{t^2}{4\nu^2}\right)\right\} = \begin{cases} 1 & \text{when } t < 2\nu\sqrt{\log C}, \\ C\exp\left(-\frac{t^2}{4\nu^2}\right) & \text{when } t \geq 2\nu\sqrt{\log C}. \end{cases}$$

Therefore, we evaluate I_1 as follows,

$$\begin{aligned} I_1 &= \int_0^{2\nu\sqrt{\log C}} 1dt + \int_{2\nu\sqrt{\log C}}^{+\infty} C\exp\left(-\frac{t^2}{4\nu^2}\right) dt \\ \left(\text{let} \quad y = \frac{t}{2\nu} - \sqrt{\log C}\right) \quad &= 2\nu\sqrt{\log C} + 2\nu \int_0^{+\infty} \exp\left(-y^2 - 2y\sqrt{\log C}\right) dy \\ &\leq 2\nu\sqrt{\log C} + 2\nu \int_0^{+\infty} \exp\left(-y^2\right) dy. \end{aligned}$$

By Gaussian integral $\int_0^{+\infty} e^{-x^2} dx = \frac{\sqrt{\pi}}{2}$, we obtain $I_1 \leq 2\nu(\sqrt{\pi} + \sqrt{\log C})$. Similarly, we evaluate I_2 as follows

$$\begin{aligned} I_2 &= \int_0^{4B\log C} 1dt + \int_{4B\log C}^{+\infty} C\exp\left(-\frac{t}{4B}\right) dt \\ &= 4B(\log C + 1). \end{aligned}$$

Hence, we obtain the expectation of Z,

$$\mathbb{E}[Z] = I_1 + I_2 \leq 2\nu(\sqrt{\pi} + \sqrt{\log C}) + 4B(\log C + 1).$$

Next, we prove the equation (35).

For all $i \in [n]$, since X_i satisfies the Bernstein condition with parameter (σ, b), the variable X_i satisfies the concentration bound (7),

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

By equation (34), we have

$$\mathbb{E}[|X_i|] \leq 2\sigma(\sqrt{\pi} + \sqrt{\log 2}) + 4b(1 + \log 2).$$

Therefore, the expectation of the sample mean satisfies

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n X_i\right|\right] \leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}[|X_i|] \leq 2\sigma(\sqrt{\pi} + \sqrt{\log 2}) + 4b(1 + \log 2).$$

□

Remark 8. For a nonnegative random variables satisfying the Bernstein-type inequality (33), the expectation of the variable is upper bounded by a function of the parameter (ν, B, C). Particularly, for a zero-mean random variable X satisfying the Bernstein condition with parameter b, $|X|$ satisfies the inequality (33) with parameters $(\sigma, b, 2)$, where $\sigma^2 = \mathrm{var}(X)$. Then, the expectation of the absolute variable $|X|$ is upper bounded by a function of (σ^2, b).

2.9 Exercise 2.9

Lemma 7 (Sharp upper bounds on binomial tails)**.** *Let* $\{X_i\}_{i=1}^n$ *be a sequence of i.i.d. Bernoulli variables with parameter* $\alpha \in (0, 1/2]$*. Consider the binomial random variable* $Z_n = \sum_{i=1}^n X_i$*. The tail probability of* Z_n *is upper bounded as*

$$\mathbb{P}\left[Z_n \leq \delta n\right] \leq e^{-nD(\delta\|\alpha)}, \quad \textit{for all } \delta \in (0, \alpha), \tag{37}$$

where the quantity

$$D(\delta\|\alpha) := \delta \log \frac{\delta}{\alpha} + (1-\delta) \log \frac{1-\delta}{1-\alpha} \tag{38}$$

is the KL divergence between the Bernoulli distributions with parameters δ *and* α*, respectively. Further, the upper bound (37) is strictly tighter than the Hoeffding bound for all* $\delta \in (0, \alpha)$*.*

Proof. Applying the Chernoff inequality to the random variable Z_n, we have

$$\mathbb{P}[Z_n \leq \delta n] \leq \frac{\mathbb{E}\left[e^{\lambda Z_n}\right]}{e^{\lambda \delta n}} = \left\{\exp\left\{\log\left(1-\alpha+\alpha e^{\lambda}\right)\right) - \lambda\delta\right\}\right\}^n, \quad \text{for all } \lambda \in \mathbb{R}, \tag{39}$$

where the second equality follows from the moment generating function of Z_n,

$$\mathbb{E}\left[e^{\lambda Z_n}\right] = \left(1-\alpha+\alpha e^{\lambda}\right)^n.$$

The upper bound (39) achieves the minimum when $\lambda = \log \frac{\delta(1-\alpha)}{\alpha(1-\delta)}$ by the first-order condition. Plugging the minimizer into equation (39), we obtain the result

$$\mathbb{P}\left[Z_n \leq \delta n\right] \leq e^{-nD(\delta\|\alpha)}, \quad \text{for all } \delta \in (0, \alpha).$$

Next, we show that equation (37) is strictly tighter than the Hoeffding bound of Z_n, for all $\delta \in (0, \alpha)$.

Hoeffding bound of random variable Z_n is

$$\mathbb{P}\left[Z_n \leq \delta n\right] \leq e^{-2n(\delta-\alpha)^2}, \quad \text{for all } \delta \in (0, \alpha). \tag{40}$$

Note that parameter α is fixed. Consider the function $g(\delta) = 2(\delta-\alpha)^2 - D\left(\delta\|\alpha\right)$ for $\delta \in (0, \alpha)$. The first-order derivative and second-order derivative of g are

$$g'(\delta) = 4(\delta - \alpha) - \log\frac{\delta}{\alpha} + \log\frac{1-\delta}{1-\alpha} \quad \text{and} \quad g''(\delta) = 4 - \frac{1}{\delta} - \frac{1}{1-\delta}.$$

Note that $g(\alpha) = g'(\alpha) = 0$, and $g''(\delta) < 0$ for all $0 < \delta < 1/2$. Then, we have $g(\delta) < 0$, for all $\delta \in (0, \alpha)$. Hence, the upper bound (37) is strictly tighter than Hoeffding bound (40). $\square$

Remark 9. The upper bound on a binomial tail is a function of the KL divergence between Bernoulli distributions with distinct parameters. The bound using KL divergence is strictly tighter than the Hoeffding bound. The underperformance of Hoeffding bound may attribute to the utilization of the sub-Gaussian parameter σ^2. The sub-Gaussian parameter σ^2 is not necessarily the optimal choice to describe the tail performance of a random variable with good properties.

2.10 Exercise 2.10

Lemma 8 (Lower bounds on binomial tails)**.** *Let $\{X_i\}_{i=1}^n$ be a sequence of i.i.d. Bernoulli variables with parameter $\alpha \in (0, 1/2]$. Consider the binomial random variable $Z_n := \sum_{i=1}^n X_i$. For some fixed $\delta \in (0, \alpha)$, let $m = \lfloor n\delta \rfloor$; i.e., m is the largest integer less or equal to $n\delta$. Let $\tilde{\delta} = m/n$. We have*

$$\frac{1}{n} \log \mathbb{P}\left[Z_n \leq n\delta\right] \geq \frac{1}{n} \log \binom{n}{m} + \tilde{\delta} \log \alpha + (1 - \tilde{\delta}) \log(1 - \alpha). \tag{41}$$

Further, the binomial coefficient satisfies

$$\frac{1}{n} \log \binom{n}{m} \geq \phi(\tilde{\delta}) - \frac{\log(n+1)}{n}. \tag{42}$$

Consequently, the lower tail of binomial variable Z_n satisfies

$$\mathbb{P}[Z_n \leq n\delta] \geq \frac{1}{n+1} e^{-nD(\delta \| \alpha)}, \tag{43}$$

where $D(\delta\|\alpha)$ is the KL divergence defined in equation (38).

Proof. First, we prove equation (41).

Since Z_n is a binomial variable with size n and probability α, we have

$$\begin{aligned}
\mathbb{P}[Z_n \leq n\delta] &= \sum_{k=1}^{m} \binom{n}{k} \alpha^k (1-\alpha)^{n-k} \\
&\geq \binom{n}{m} \alpha^m (1-\alpha)^{n-m}.
\end{aligned} \tag{44}$$

Taking the log and dividing by n on the both sides of the inequality (44) , we obtain

$$\frac{1}{n} \log \mathbb{P}[Z_n \leq n\delta] \geq \frac{1}{n} \log \binom{n}{m} + \tilde{\delta} \log \alpha + (1 - \tilde{\delta}) \log(1 - \alpha).$$

Next, we prove equation (42).

Consider a binomial variable $Y \sim Bin(n, \tilde{\delta})$. For all $k = 0, 1, ..., (n-1)$, the ratio between $\mathbb{P}[Y = k]$ and $\mathbb{P}[Y = k+1]$ is

$$\frac{\mathbb{P}[Y = k+1]}{\mathbb{P}[Y = k]} = \frac{\binom{n}{k+1} \tilde{\delta}^{k+1} (1-\tilde{\delta})^{n-k-1}}{\binom{n}{k} \tilde{\delta}^{k} (1-\tilde{\delta})^{n-k}} = \frac{(n-k)\tilde{\delta}}{(k+1)(1-\tilde{\delta})}.$$

To let the ratio $\frac{\mathbb{P}[Y=k+1]}{\mathbb{P}[Y=k]} \geq 1$, we need $(k+1) \leq (n+1)\tilde{\delta}$. Thus, the probability $\mathbb{P}[Y = l]$ achieves the maximum when $l = n\tilde{\delta}$. Consequently, we have

$$(n+1)\mathbb{P}[Y = n\tilde{\delta}] \geq 1 \quad \Leftrightarrow \quad \mathbb{P}[Y = n\tilde{\delta}] \geq \frac{1}{n+1}. \tag{45}$$

Taking the log and dividing by n on the both sides of the inequality (45), we obtain

$$\frac{1}{n} \log \binom{n}{m} \geq \tilde{\delta} \log(\alpha) - (1 - \tilde{\delta}) \log(1 - \alpha) - \frac{\log(n+1)}{n} = \phi(\tilde{\delta}) - \frac{\log(n+1)}{n}.$$

Last, we prove the lower bound equation (43).

Plugging equation (42) into equation (41), we have

$$\frac{1}{n}\log\mathbb{P}\left[Z_n \le n\delta\right] \ge \phi(\tilde{\delta}) + \tilde{\delta}\log\alpha + (1-\tilde{\delta})\log(1-\alpha) - \frac{\log(n+1)}{n} \ge -D(\delta\|\alpha) - \frac{\log(n+1)}{n}. \quad (46)$$

Multiplying n and exponentiating on the both sides of the inequality (46), we obtain the result

$$\mathbb{P}\left[Z_n \le n\delta\right] \ge \frac{1}{n+1}e^{-nD(\delta\|\alpha)}.$$

□

Remark 10. The lower bound on a binomial tail is also a function of the KL divergence between Bernoulli distributions with distinct parameters. Combining the upper bound (37) and lower bound (43), we conclude that the binomial tail is (upper and lower) bounded by the functions of KL divergence between Bernoulli distributions.

2.11 Exercise 2.11

Lemma 9 (Gaussian maxima)**.** *Let $\{X_i\}_{i=1}^n$ be a sequence of i.i.d. normal random variables following $N(0,\sigma^2)$. Consider the random variable $Z_n := \max_{i\in[n]}|X_i|$. Since the tail bound $\mathbb{P}[U \ge t] \le \sqrt{\frac{2}{\pi}}\frac{1}{t}e^{-t^2/2}$ holds for all standard normal random variable U, the expectation of Z_n is upper bounded as follows*

$$\mathbb{E}[Z_n] \le \sqrt{2\sigma^2\log n} + \frac{4\sigma}{\sqrt{2\log n}}, \quad \textit{for all } n \ge 2.$$

Proof. Consider the tail probability of Z_n. We have

$$\mathbb{P}[Z_n \ge t] = \mathbb{P}[\max_{i\in[n]}|X_i| \ge t] = 1 - \mathbb{P}[\max_{i\in[n]}|X_i| < t] = 1 - (1 - \mathbb{P}[|X_1| \ge t])^n,$$

where the last equality follows from the independence of $\{X_i\}_{i=1}^n$. By Bernoulli's inequality, we have

$$(1 - \mathbb{P}[|X_1| \ge t])^n \ge 1 - n\mathbb{P}[|X_1| \ge t], \quad \text{for all } t > 0.$$

Therefore, the expectation of Z_n follows

$$\mathbb{E}[Z_n] = \int_0^{+\infty}\mathbb{P}[Z_n \ge t]dt \le c + \int_c^{+\infty} n\mathbb{P}[|X_1| \ge t]dt, \quad \text{for all } c > 0. \quad (47)$$

Since the tail bound $\mathbb{P}[U \ge t] \le \sqrt{\frac{2}{\pi}}\frac{1}{t}e^{-t^2/2}$ holds for all standard normal random variable U, the tail bound of $|X_1|$ satisfies

$$\mathbb{P}[|X_1| \ge t] \le 2\sqrt{\frac{2}{\pi}}\frac{\sigma}{t}e^{-\frac{t^2}{2\sigma^2}}.$$

Hence, the last integral in equation (47) follows

$$\begin{aligned} \int_c^{+\infty} n\mathbb{P}[|X_1| \ge t]dt &\le \frac{2n\sigma}{c}\sqrt{\frac{2}{\pi}}\int_c^{+\infty} e^{-\frac{t^2}{2\sigma^2}}dt \\ (\text{let } u = \frac{t}{\sigma}) \quad &\le \frac{2n\sigma^2}{c}\sqrt{\frac{2}{\pi}}\int_{\frac{c}{\sigma}}^{+\infty} ue^{-\frac{u^2}{2}}dt \\ &= \frac{2n\sigma^2}{c}\sqrt{\frac{2}{\pi}}e^{-\frac{c^2}{2\sigma^2}}. \end{aligned} \quad (48)$$

For all $n \geq 2$, let $c = \sqrt{2\sigma^2 \log n}$ and plug the c into equations (47) and (48). Then, we obtain the result

$$\mathbb{E}[Z_n] \leq \sqrt{2\sigma^2 \log n} + \frac{2\sigma}{\sqrt{2\log n}}\sqrt{\frac{2}{\pi}} \leq \sqrt{2\sigma^2 \log n} + \frac{4\sigma}{\sqrt{2\log n}}, \quad \text{for all } n \geq 2.$$

□

2.12 Exercise 2.12

Lemma 10 (Sharp upper bounds for sub-Gaussian maxima)**.** *Let $\{X_i\}_{i=1}^n$ be a sequence of zero-mean sub-Gaussian variables with parameter σ. Then, we have*

$$\mathbb{E}\left[\max_{i\in[n]} X_i\right] \leq \sqrt{2\sigma^2 \log n}, \quad \textit{for all } n \geq 1.$$

Note that independence assumptions are unnecessary.

Proof. For all $t > 0$, the function $f(x) = \exp(tx)$ is a convex function. Apply the function $f(x)$ to $\mathbb{E}\left[\max_{i\in[n]} X_i\right]$. By Jensen's inequality, we have

$$\exp\left\{t\mathbb{E}\left[\max_{i\in[n]} X_i\right]\right\} \leq \mathbb{E}\left[\exp\left(t\max_{i\in[n]} X_i\right)\right] \leq \sum_{i=1}^{n} \mathbb{E}\left[\exp\left(tX_i\right)\right] \leq ne^{\frac{t^2\sigma^2}{2}}, \quad \text{for all } t > 0, \tag{49}$$

where the last inequality follows from the sub-Gaussianity of X_is. Take the log of both sides of equation (49). We have

$$\mathbb{E}\left[\max_{i\in[n]} X_i\right] \leq \frac{\log n}{t} + \frac{t\sigma^2}{2}. \tag{50}$$

Plugging $t = \frac{\sqrt{2\log n}}{\sigma}$ into the equation (50), we obtain

$$\mathbb{E}\left[\max_{i\in[n]} X_i\right] \leq \sqrt{2\sigma^2 \log n}.$$

□

2.13 Exercise 2.13

Lemma 11 (Operations on sub-Gaussian variables)**.** *Let X_1 and X_2 be two zero-mean sub-Gaussian variables with parameters σ_1 and σ_2 respectively.*

(a). If X_1 and X_2 are independent, then random variable $X_1 + X_2$ is sub-Gaussian with parameter $\sqrt{\sigma_1^2 + \sigma_2^2}$.

(b). Without independence assumptions, then random variable $X_1 + X_2$ is sub-Gaussian with parameter at most $\sigma_1 + \sigma_2$.

(c). If X_1 and X_2 are independent, then random variable $X_1 X_2$ is sub-exponential with parameter $(\nu, b) = \left(\sqrt{2}\sigma_1\sigma_2, \frac{1}{\sqrt{2}\sigma_1\sigma_2}\right)$.

Proof. First, we prove (a).

Since $X_1 \sim \mathrm{subG}(\sigma_1)$ and $X_2 \sim \mathrm{subG}(\sigma_2)$ are independent, we have

$$\mathbb{E}\left[e^{t(X_1+X_2)}\right] = \mathbb{E}\left[e^{tX_1}\right]\mathbb{E}\left[e^{tX_2}\right] \leq e^{\frac{t^2(\sigma_1^2+\sigma_2^2)}{2}}, \quad \text{for all } t \in \mathbb{R}.$$

Therefore, $X_1 + X_2$ is sub-Gaussian with parameter $\sqrt{\sigma_1^2 + \sigma_2^2}$.

Next, we prove (b).

For all $t \in \mathbb{R}$, we have

$$\mathbb{E}\left[e^{t(X_1+X_2)}\right] = \mathbb{E}\left[e^{tX_1}e^{tX_2}\right]. \tag{51}$$

Introduce $p, q \geq 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$. By Holder's inequality, the equation (51) becomes

$$\mathbb{E}\left[e^{t(X_1+X_2)}\right] \leq \left(\mathbb{E}\left[e^{ptX_1}\right]\right)^{\frac{1}{p}}\left(\mathbb{E}\left[e^{ptX_2}\right]\right)^{\frac{1}{q}} \leq e^{\frac{pt^2\sigma_1^2+qt^2\sigma_2^2}{2}}, \tag{52}$$

where the last inequality follows from the sub-Gaussianity of X_1 and X_2. The inequality (52) achieves its minimum when $p = \sigma_2/\sigma_1 + 1$. Plugging the minimizer to inequality (52), we obtain

$$\mathbb{E}\left[e^{t(X_1+X_2)}\right] \leq e^{\frac{t^2(\sigma_1+\sigma_2)^2}{2}}.$$

Therefore, the random variable $X_1 + X_2$ is sub-Gaussian with parameter at most $\sigma_1 + \sigma_2$.

Last, we prove (c).

□

2.14 Exercise 2.14

Lemma 12. *Let X be a scalar random variable. Suppose there exist two positive constant c_1, c_2 such that*

$$\mathbb{P}\left[|X - \mathbb{E}[X]| \geq t\right] \leq c_1 e^{-c_2 t^2}, \quad \textit{for all } t \geq 0. \tag{53}$$

Then, the variance $\mathrm{var}(X) \leq c_1/c_2$.

Let m_X be the median of X; i.e., $\mathbb{P}[X \geq m_X] \geq 1/2$ and $\mathbb{P}[X \leq m_X] \geq 1/2$. Note that the median m_X is not necessarily unique. Given mean concentration (53), for all median m_X, we have

$$\mathbb{P}[|X - m_X| \geq t] \leq c_3 e^{-c_4 t^2}, \quad \textit{for all } t \geq 0, \tag{54}$$

where $c_3 := 4c_1$ and $c_4 := c_2/8$.
Conversely, if the equation (54) holds, the mean concentration (53) also holds with parameter $c_1 = 2c_3$ and $c_2 = c_4/4$.

Proof. First, we show the variance $\mathrm{var}(X) \leq c_1/c_2$.

By the definition of variance, we have

$$\mathrm{var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \int_0^{+\infty} \mathbb{P}\left[|X - \mathbb{E}[X]| \geq \sqrt{t}\right] dt. \tag{55}$$

Since X satisfies the mean concentration equation (53), for all $t > 0$, we have

$$\mathbb{P}\left[|X - \mathbb{E}[X]| \geq \sqrt{t}\right] \leq c_1 e^{-c_2 t}. \tag{56}$$

Plugging the equation (56) into equation (55), we obtain

$$\mathrm{var}(X) \le \frac{c_1}{c_2}.$$

Next, we show by an example that the median of X, m_X, is not necessarily unique.

Consider a random variable $Y \sim Bin(5, 1/2)$. Then, we have $\mathbb{P}[Y \ge 3] = 1/2$ and $\mathbb{P}[Y \le 3] > 1/2$. Meanwhile, we also have $\mathbb{P}[Y \ge 2] > 1/2$ and $\mathbb{P}[Y \le 2] = 1/2$. By the definition of median, the number 2 and 3 both are the median of Y. Therefore, the median of a random variable is not necessarily unique.

Then, we prove the median concentration (54), provided that random variable X satisfies the mean concentration (53).

Let $\Delta = |\mathbb{E}[X] - m_X|$. Consider the following two cases.

1. Case 1: Suppose $t \ge 2\Delta$.

   Note the triangle inequality $|X - \mathbb{E}[X]| \ge |X - m_X| - \Delta$ and the assumption $\frac{t}{2} \ge \Delta$. We have
   $$\mathbb{P}[|X - m_X| \ge t] \le \mathbb{P}\left[|X - m_X| \ge \frac{t}{2} + \Delta\right] \le \mathbb{P}\left[|X - \mathbb{E}[X]| \ge \frac{t}{2}\right].$$
   By equation (53), we obtain
   $$\mathbb{P}[|X - m_X| \ge t] \le c_1 e^{-\frac{c_2 t^2}{4}}. \tag{57}$$

2. Case 2: Suppose $t < 2\Delta$.

   By the definition of median, we have
   $$\mathbb{P}[|X - \mathbb{E}[X]| \ge \Delta] \ge \mathbb{P}[X \ge m_X] \ge \frac{1}{2}. \tag{58}$$
   Meanwhile, by equation (53), we have
   $$\mathbb{P}[|X - \mathbb{E}[X]| \ge \Delta] \le c_1 e^{-c_2 \Delta^2}. \tag{59}$$
   Combing the assumption $\Delta > \frac{t}{2}$ and inequalities (58) (59), we obtain
   $$\mathbb{P}[|X - m_X| \ge t] \le 1 \le 2c_1 e^{-c_2 \Delta^2} \le 2c_1 e^{-\frac{c_2 t^2}{4}}. \tag{60}$$

Hence, combing the inequality (57) and inequality (60), we conclude that

$$\mathbb{P}[|X - m_X| \ge t] \le c_3 e^{-c_4 t^2},$$

where $c_3 = 2c_1$ and $c_4 = c_2/4$. The upper bound also holds when $c_3 = 4c_1$ and $c_4 = c_2/8$.

Last, we prove the mean concentration (53), provided that the random variable X satisfies the median concentration (54).

As the proof for median concentration (54), we also consider the following two cases.

1. Case 1: Suppose $t \geq 2\Delta$.

   Note the triangle inequality $|X - m_X| + \Delta \geq |X - \mathbb{E}[X]|$ and the assumption $\frac{t}{2} \geq \Delta$. We have
$$\mathbb{P}[|X - \mathbb{E}[X]| \geq t] \leq \mathbb{P}\left[|X - m_X| \geq \frac{t}{2}\right] \leq c_3 e^{-\frac{c_4 t^2}{4}},$$
   where the last inequality follows from the median concentration (54).

2. Case 2: Suppose $t < 2\Delta$.

□

2.15 Exercise 2.15

Lemma 13. *Let $\{X_i\}_{i=1}^n$ be i.i.d. sample of random variables with density f on the real line. A standard estimate of f is the kernel density estimate*

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right),$$

where $K \colon \mathbb{R} \mapsto [0, +\infty)$ is a kernel function satisfying $\int_{-\infty}^{+\infty} K(x)dx = 1$, and $h > 0$ is a bandwidth parameter. Suppose we assess the estimate $\hat{f}_n(x)$ using the L^1 norm; i.e.,

$$\left\|\hat{f}_n - f\right\|_1 = \int_{-\infty}^{+\infty} |\hat{f}_n(t) - f(t)| dt.$$

The upper tail probability of the L^1 norm satisfies

$$\mathbb{P}\left[\left\|\hat{f}_n - f\right\|_1 - \mathbb{E}\left[\left\|\hat{f}_n - f\right\|_1\right] \geq \delta\right] \leq e^{-\frac{n\delta^2}{8}}, \quad \textit{for all } \delta \geq 0.$$

Proof. Note that $\left\|\hat{f}_n - f\right\|_1$ is a function of $X = (X_1, ..., X_n)$, denoted $g(X) = g(X_1, ..., X_n) = \left\|\hat{f}_n - f\right\|_1$ and $g \colon \mathbb{R}^n \mapsto \mathbb{R}$. Let $x^{(k)} = (x_1, ..., x_k, ..., x_n) \in \mathbb{R}^n$ and $x'^{(k)} = (x_1, ..., x'_k, ..., x_n) \in \mathbb{R}^n$ be two vectors. The absolute difference between the $g(x^{(k)})$ and $g(x'^{(k)})$ is

$$\begin{aligned}
|g(x^{(k)}) - g(x'^{(k)})| &= \left|\int_{-\infty}^{+\infty} \left|\frac{1}{nh} \sum_{i=1}^n K\left(\frac{t - x_i}{h}\right) - f(t)\right| - \left|\frac{1}{nh} \sum_{i \neq k}^n K\left(\frac{t - x_i}{h}\right) - f(t) + \frac{1}{nh} K\left(\frac{t - x'_k}{h}\right)\right| dt\right| \\
&\leq \int_{-\infty}^{+\infty} \left|\frac{1}{nh}\left(K\left(\frac{t - x_k}{h}\right) - K\left(\frac{t - x'_k}{h}\right)\right)\right| dt \\
&\leq \frac{1}{nh}\left(\int_{-\infty}^{+\infty} K\left(\frac{t - x_k}{h}\right) dt + \int_{-\infty}^{+\infty} K\left(\frac{t - x'_k}{h}\right) dt\right). \qquad (61)
\end{aligned}$$

Let $t_1 = \frac{t - x_k}{h}$. By the definition of K, the first integral in equation (61) are

$$\int_{-\infty}^{+\infty} K\left(\frac{t - x_k}{h}\right) dt = h \int_{-\infty}^{+\infty} K(t_1) dt_1 = h.$$

Similarly, the second integral $\int_{-\infty}^{+\infty} K\left(\frac{t-x'_k}{h}\right) dt = h$. Then, we conclude the function g satisfies

$$|g(x^{(k)}) - g(x^{'(k)})| \leq \frac{2}{n}. \tag{62}$$

The inequality (62) also holds for all $k \in [n]$ and for all $x^{(k)}, x^{'(k)} \in \mathbb{R}^n$. Therefore, $g(X)$ satisfies the bounded difference property (8) with parameters $(\frac{2}{n}, ..., \frac{2}{n})$. By Theorem 1.9, we obtain

$$\mathbb{P}\left[\left\|\hat{f}_n - f\right\|_1 - \mathbb{E}\left[\left\|\hat{f}_n - f\right\|_1\right] \geq \delta\right] \leq e^{-\frac{n\delta^2}{2}} \leq e^{-\frac{n\delta^2}{8}}.$$

□

2.16 Exercise 2.16

Lemma 14. *Let $\{X_i\}_{i=1}^n$ be a sequence of independent variables taking values in a Hilbert space $\mathbb{H}$. Suppose that $\|X_i\|_{\mathbb{H}} \leq b_i$ almost surely, for all $i \in [n]$. Consider the real valued random variable $S_n = \|\sum_{i=1}^n X_i\|_{\mathbb{H}}$. The concentration bound of S_n is*

$$\mathbb{P}\left[|S_n - \mathbb{E}[S_n]| \geq n\delta\right] \leq 2e^{-\frac{n\delta^2}{8b^2}}, \quad \text{for all } \delta \geq 0, \tag{63}$$

where $b^2 = \frac{1}{n}\sum_{i=1}^n b_i^2$. The upper tail probability bound of S_n is

$$\mathbb{P}\left[\frac{S_n}{n} \geq a + \delta\right] \leq e^{-\frac{n\delta^2}{8b^2}}, \quad \text{for all } \delta \geq 0, \tag{64}$$

where $a = \sqrt{\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\|X_i\|_{\mathbb{H}}^2\right]}$.

Proof. First, we prove the concentration bound (63).

Note that S_n is a function of $X = (X_1, ..., X_n)$, denoted $S_n(X) = S_n(X_1, ..., X_n)$, and $S_n(X) \colon \mathbb{R}^n \mapsto \mathbb{R}$. Let $x^{(k)} = (x_1, ..., x_k, ..., x_n) \in \mathbb{R}^n$ and $x^{'(k)} = (x_1, ..., x'_k, ..., x_n) \in \mathbb{R}^n$ be two vectors. The absolute difference between $S_n(x^{(k)})$ and $S_n(x^{'(k)})$ is

$$\begin{aligned}
\left|S_n(x^{(k)}) - S_n(x^{'(k)})\right| &= \left|\|x_1 + \cdots + x_k + \cdots + x_n\|_{\mathbb{H}} - \|x_1 + \cdots + x'_k + \cdots + x_n\|_{\mathbb{H}}\right| \\
&\leq \|x_k - x'_k\|_{\mathbb{H}} \\
&\leq 2b_k,
\end{aligned} \tag{65}$$

where the last inequality follows by the boundedness $\|X_k\|_{\mathbb{H}} \leq b_k$, for all $k \in [n]$. The inequality (65) also holds for all $k \in [n]$ and for all $x^{(k)}, x^{'(k)} \in \mathbb{R}^n$. Therefore, $S_n(X)$ satisfies the bounded property (8) with parameters $(2b_1, ..., 2b_n)$. By Theorem 1.9, we obtain

$$\mathbb{P}\left[|S_n - \mathbb{E}[S_n]| \geq n\delta\right] \leq 2e^{-\frac{n\delta^2}{2b^2}} \leq 2e^{-\frac{n\delta^2}{8b^2}}, \quad \text{for all } \delta \geq 0.$$

Next, we prove the upper tail probability bound (64).

The expectation of $S_n(X) = \mathbb{E}\left[\sum_{i=1}^n \|X_i\|_{\mathbb{H}}\right]$ satisfies

$$\mathbb{E}\left[\left\|\sum_{i=1}^n X_i\right\|_{\mathbb{H}}\right] \leq \sqrt{\mathbb{E}\left[\left\|\sum_{i=1}^n X_i\right\|_{\mathbb{H}}^2\right]} = \sqrt{\sum_{i=1}^n \mathbb{E}\left[\|X_i\|_{\mathbb{H}}^2\right]} = na, \tag{66}$$

where the first inequality follows by the fact that $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$ and the second equation follows by the independence of $\{X_i\}_{i=1}^n$. Therefore, combing the inequality (66) with the concentration bound (63), we obtain

$$\mathbb{P}\left[\frac{S_n}{n} \geq a + \delta\right] = \mathbb{P}\left[S_n - na \geq n\delta\right] \leq \mathbb{P}\left[S_n - \mathbb{E}[S_n] \geq n\delta\right] \leq e^{-\frac{n\delta^2}{8b^2}}.$$

□

2.17 Exercise 2.17

Lemma 15 (Hanson-Wright Inequality)**.** *Let $\boldsymbol{Q} = [\![\boldsymbol{Q}_{ij}]\!] \in \mathbb{R}^{n\times n}$ be a positive semi-definite matrix, and $\{X_i\}_{i=1}^n$ be i.i.d. random variables with mean zero, variance 1 and $\sigma-$sub-Gaussian. Consider the random variable $Z = \sum_{i=1}^n \sum_{j=1}^n \boldsymbol{Q}_{ij} X_i X_j$. The Hanson-Wright inequality for Z is*

$$\mathbb{P}\left[Z \geq \text{Trace}(\boldsymbol{Q}) + \sigma t\right] \leq 2e^{-\min\left\{\frac{c_1 t}{\|\boldsymbol{Q}\|_{op}}, \frac{c_2 t^2}{\|\boldsymbol{Q}\|_F^2}\right\}}, \quad \textit{for all } t \geq 0,$$

where $\|\boldsymbol{Q}\|_{op}$ is the operator norm of matrix $\boldsymbol{Q}$, $\|\boldsymbol{Q}\|_F$ is the Frobenius norm of matrix $\boldsymbol{Q}$, and c_1, c_2 are two universal constants independent with n and t.

The proof for Hanson-Wright Inequality under the special case that $X_i \overset{i.i.d.}{\sim} N(0,1)$ is below. Let $\sigma = 1$.

Proof. Since the random variables $X_{i\,i=1}^{\,n}$ are independent and $\text{var}(X_i) = 1$ for all $i \in [n]$, the expectation $\mathbb{E}[Z] = \sum_{i=1}^n \boldsymbol{Q}_{ii} = \text{Trace}(\boldsymbol{Q})$. Then, the upper tail probability of Z becomes

$$\mathbb{P}\left[Z \geq \text{Trace}(\boldsymbol{Q}) + t\right] = \mathbb{P}\left[\left(\sum_{i=1}^n \boldsymbol{Q}_{ii}(X_i^2 - 1) + \sum_{i\neq j} \boldsymbol{Q}_{ij} X_i X_j\right) \geq t\right]$$

For two random variable X, Y and constant t, we have the inequality $\mathbb{P}[X + Y \geq t] \leq \mathbb{P}[X \geq t/2] + \mathbb{P}[Y \geq t/2]$. Hence, the tail probability is upper bounded as

$$\mathbb{P}\left[Z \geq \text{Trace}(\boldsymbol{Q}) + t\right] \leq \mathbb{P}\left[\sum_{i=1}^n \boldsymbol{Q}_{ii}(X_i^2 - 1) \geq \frac{t}{2}\right] + \mathbb{P}\left[\sum_{i\neq j} \boldsymbol{Q}_{ii} X_i X_j \geq \frac{t}{2}\right] =: p_1 + p_2.$$

Consider p_1. For a standard normal variable X, the squared variable $X^2 \sim \chi_1^2$ and X^2 is sub-exponential with parameter $(\nu, b) = (2, 4)$. Given a constant c, cX^2 is still sub-exponential with parameters $(2c, 4c^2)$. Hence, for all $i \in [n]$, the variable $\boldsymbol{Q}_{ii} X_i^2$ is sub-exponential with parameters $(2\boldsymbol{Q}_{ii}, 4\boldsymbol{Q}_{ii}^2)$. By the Bernstein-type inequalities(non-asymptotic 5.16), the probability p_1 is upper bounded as following,

$$p_1 = \mathbb{P}\left[\sum_{i=1}^n \left(\boldsymbol{Q}_{ii} X_i^2 - \boldsymbol{Q}_{ii}\mathbb{E}[X_i^2]\right) \geq \frac{t}{2}\right] \leq \exp\left(-\min\left\{\frac{ct^2}{\sum_i^n \boldsymbol{Q}_{ii}^2}, \frac{c' t}{\max_{i\in[n]} |\boldsymbol{Q}_{ii}|}\right\}\right),$$

where c and c' are two constants independent with n and t.

Consider p_2. For two i.i.d. standard normal variables X, Y, the product $XY = \frac{1}{4}(X+Y)^2 - \frac{1}{4}(X-Y)^2$, where $\frac{1}{2}(X+Y)^2$ and $\frac{1}{2}(X-Y)^2$ follow the chi-square distribution χ_1^2 independently.

Then, XY is the subtraction of two independent sub-exponential variables. Let D_1 and D_2 be two identical and independent sub-exponential variables with parameters (ν, b). We have

$$\mathbb{E}[e^{\lambda(D_1-D_2)}] = \mathbb{E}[e^{\lambda D_1}]\mathbb{E}[e^{-\lambda D_2}] \le e^{v^2\lambda^2}, \quad \text{for all } |\lambda| < \frac{1}{b}.$$

Therefore, the subtraction of two identical and independent sub-exponential variables $D_1 - D_2$ is still sub-exponential with parameters $(\sqrt{2}\nu, b)$.

Back to $\boldsymbol{Q}_{ij}X_iX_j$, where $i \ne j$. By the discussion above, the variable $\boldsymbol{Q}_{ij}X_iX_j$ is sub-exp onential with parameters $(\sqrt{2}\boldsymbol{Q}_{ij}, \boldsymbol{Q}_{ij}^2)$. By the Bernstein-type inequalities, the probability p_2 is upper bounded as following,

$$p_2 = \mathbb{P}\left[\sum_{i\ne j}\boldsymbol{Q}_{ii}X_iX_j \ge \frac{t}{2}\right] \le \exp\left(-\min\left\{\frac{c''t^2}{\sum_{i\ne j}\boldsymbol{Q}_{ij}^2}, \frac{c'''t}{\max_{i\ne j}|\boldsymbol{Q}_{ij}|}\right\}\right),$$

where c'' and c''' are two constants independent with n and t.

Therefore, by the definition of Frobenius norm and the fact that $\max_{i,j\in[n]}|\boldsymbol{Q}_{ij}| \le \|\boldsymbol{Q}\|_{op}$, we obtain the goal inequality

$$\mathbb{P}\left[Z \ge \text{Trace}(\boldsymbol{Q}) + t\right] \le p_1 + p_2 \le 2\exp\left(-\min\left\{\frac{c_2t^2}{\|\boldsymbol{Q}\|_F^2}, \frac{c_1t}{\|\boldsymbol{Q}\|_{op}}\right\}\right),$$

for some universal constants c_1, c_2 independent with n and t. $\square$