Chapter 2: *Convexity*

Xuzhi Yang and Zetai Cen

Last updated: June 7, 2024

Contents

1 Introduction

Keywords: log moment generating functions; Hoeffding's inequality; Bennett's inequality; large deviation bounds; Orlicz norms; subgaussian distributions.

1. Chernoff's inequality;

2. Sub-Gaussian, sub-exponential **definition** and **concentration inequalities**, i.e. Hoeffding, Bennett, Bernstein inequalities;

3. Introduce the notion of Orlicz norm, which incorporates sub-Gaussian and sub-exponential as special cases. One example of random projection;

2 The Cramér-Chernoff method

Definition 2.1. *(log-moment generating function) For all $\lambda \in \mathbb{R}_+$, we define the log-moment generating function of a random variable X as*

$$\psi_X(\lambda) = \log \mathbb{E}[\exp(\lambda X)].$$

Definition 2.2. *(Cramér transform) The Cramér transform of random variable X is defined for on $t \in \mathbb{R}$ as*

$$\psi_X^*(t) = \sup_{\lambda \in \mathbb{R}_+} \{\lambda t - \psi_X(\lambda)\}.$$

Theorem 2.1. *(Chernoff's inequality) Given a zero mean random variable X and any real number t, it holds that*

$$\mathbb{P}(X \geq t) \leq \exp\{-\psi_X^*(t)\}. \tag{1}$$

Proof. For any $\lambda \in \mathbb{R}_+$, we have by Markov's inequality that

$$\begin{aligned}\mathbb{P}(X \geq t) &= \mathbb{P}(\exp(\lambda X) \geq \exp(\lambda t)) \\ &\leq e^{-\lambda t}\, \mathbb{E}[e^{\lambda X}] = \exp\big\{ - [\lambda t - \psi_X(\lambda)]\big\},\end{aligned}$$

which concludes the proof by maximising $\lambda t - \psi_X(\lambda)$ over λ. $\square$

Observe trivially $\psi_X(0) = 0$ and hence $\psi_X^*(t) \geq 0$. By Jensen's inequality we also have $\psi_X(\lambda) \geq \lambda \mathbb{E}[X]$, so that $\lambda t - \psi_X(\lambda) \leq 0$ for any $\lambda < 0$ if $t \geq \mathbb{E}[X]$. This is useful since we are interested in the tail probability, i.e. $t > 0$ in Theorem 2.1. In this case, we can rewrite the Cramér transform as

$$\psi_X^*(t) = \sup_{\lambda \in \mathbb{R}} \{\lambda t - \psi_X(\lambda)\},$$

which is actually the Legendre-Fenchel transform of $\psi_X(\lambda)$.

3 Large deviation inequalities

Before introducing the sub-Gaussian family, we decide to start with the following claim which describes different but equivalent ways to characterise sub-Gaussianity, see Appendix A for the proof.

3.1 Sub-Gaussian and Hoeffding's inequality

Claim 3.1. *(sub-Gaussian properties) Given a zero mean random variable X, the following properties are equivalent:*

(I) (MGF of X) There is a constant $\sigma \geq 0$ such that

$$\mathbb{E}[\exp(\lambda X)] \leq \exp(\sigma^2 \lambda^2 / 2) \quad \forall \lambda \in \mathbb{R}. \tag{2}$$

(II) (tails of X majorized by normal distribution) There is a constant $c \geq 0$ and Gaussian random variable $Z \sim N(0, \tau^2)$ such that

$$\mathbb{P}(|X| \geq s) \leq c\,\mathbb{P}(|Z| \geq s) \quad \forall s \geq 0. \tag{3}$$

(III) (moments of X) There is a constant $\theta \geq 0$ such that

$$\mathbb{E}[X^{2k}] \leq \frac{(2k)!}{2^k k!}\,\theta^{2k} \quad \forall k = 1, 2, 3, \ldots. \tag{4}$$

(IV) (MGF of X^2) There is a constant $\sigma \geq 0$ such that

$$\mathbb{E}[\exp(\lambda x^2 / 2\sigma^2)] \leq 1/\sqrt{1-\lambda} \quad \forall \lambda \in [0, 1). \tag{5}$$

Definition 3.1. *(sub-Gaussian distribution) A random variable X is sub-Gaussian if $X - \mathbb{E}[X]$ satisfies any property in Claim 3.1. In particular, we say X is sub-Gaussian with parameter σ if $X - \mathbb{E}[X]$ satisfies (I) with the constant σ.*

Clearly and without surprise, normal random variable is a member of such sub-Gaussian family. In fact, any bounded distribution is sub-Gaussian. This can be seen by letting X be the demeaned bounded distribution with support $[a, b]$, X' its independent copy, and ε an independent Rademacher variable, then by Jensen's inequality,

$$\begin{aligned}
\mathbb{E}_X[\exp(\lambda X)] &= \mathbb{E}_X\Big[\exp\Big(\lambda(X - \mathbb{E}_{X'}[X'])\Big)\Big] \leq \mathbb{E}_{X,X'}\Big[\exp\Big(\lambda(X - X')\Big)\Big] \\
&= \mathbb{E}_{X,X'}\Big[\mathbb{E}_\varepsilon\Big\{\exp\Big(\lambda\varepsilon(X - X')\Big)\Big\}\Big] = \mathbb{E}_{X,X'}\Big[\frac{1}{2}\Big(e^{\lambda(X-X')} + e^{-\lambda(X-X')}\Big)\Big] \\
&\leq \mathbb{E}_{X,X'}\Big[e^{\lambda^2(X-X')^2/2}\Big] \leq e^{\lambda^2(b-a)^2/2},
\end{aligned}$$

where the second last inequality used $e^x + e^{-x} \leq 2e^{x^2/2}$ for all $x \in \mathbb{R}$, and the last used $|X - X'| \leq b - a$.

Remark 3.1. *Intuitively, we want the squared sub-Gaussian parameter σ^2 to look like the variance, and obtain related deviation bounds. Sometimes people call σ^2 as the variance proxy. It is pointed out that the above variance proxy for bounded variable is not tight and hence this proxy plugged in Theorem 3.1 is not as strong as Theorem 6.1 in Appendix A.*

With sub-Gaussian random variables, we may easily obtain an (upper) deviation inequality using Chernoff's inequality,

$$\mathbb{P}(X - \mu \geq t) \leq e^{-\lambda t}\, \mathbb{E}[e^{\lambda X}] \leq \inf_{\lambda \geq 0} e^{-\lambda t} e^{\lambda^2 \sigma^2/2} = e^{-t^2/2\sigma^2}. \tag{6}$$

As we are more interested in bounding the tails of a sum of random variables, i.e. we want

$$\mathbb{P}\Big(\sum_{i=1}^{n}(X_i - \mu_i) \geq t\Big) \leq \text{something small},$$

for $t \geq 0$, we now consider $X_1, X_2, \ldots, X_n$ being independent sub-Gaussian random variables. To handle this, we have a good thing on sub-Gaussian family stated in the following claim.

Claim 3.2. *If X, Y are independent sub-Gaussian with parameters σ_X, σ_Y, $X + Y$ is also sub-Gaussian with parameter $(\sigma_X^2 + \sigma_Y^2)$.*

Proof.

$$\mathbb{E}[e^{\lambda(X+Y)}] = \mathbb{E}[e^{\lambda X}] \cdot \mathbb{E}[e^{\lambda Y}] \leq e^{\lambda^2 \sigma_X^2/2 + \lambda^2 \sigma_Y^2/2} = e^{\lambda^2(\sigma_X^2/2 + \sigma_Y^2)/2}.$$

□

We may introduce the following inequality, the proof is omitted as it should be straightforward using (6) and Claim 3.2.

Theorem 3.1. *(Hoeffding's inequality) Let X_i with mean μ_i be sub-Gaussian with parameter σ_i, $i = 1, 2, \ldots, n$. For all $t \geq 0$,*

$$\mathbb{P}\Big(\sum_{i=1}^{n}(X_i - \mu_i) \geq t\Big) \leq \exp\Big\{-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2}\Big\}.$$

Many variants of Hoeffding's inequality may be written given the above, such as for bounded random variables, normal distributions, etc.

3.2 Sub-exponential distributions and Bernstein's inequality

Similar to sub-Gaussian distributions, we may define a family general enough yet useful to construct some deviation inequalities. We start with some properties in the following claim, and refer to Section 2.5 of Wainwright (2019) for the proof.

Claim 3.3. *(sub-exponential properties) Given a zero mean random variable X, the following properties are equivalent:*

(I) (MGF of X) There are non-negative constants (ν, α) such that

$$\mathbb{E}[\exp(\lambda X)] \leq \exp(\nu^2\lambda^2/2) \quad \forall|\lambda| \leq \frac{1}{\alpha}. \tag{7}$$

(II) (tails of X majorized by exponential distribution) There are constants $c_1, c_2 > 0$ such that

$$\mathbb{P}(|X| \geq s) \leq c_1 \, e^{-c_2 s} \quad \forall s > 0. \tag{8}$$

(III) (moments of X) The quantity defined below is finite,

$$\gamma := \sup_{k \geq 2} \left\{ \frac{\mathbb{E}[X^k]}{k!} \right\}^{1/k}. \tag{9}$$

(IV) (finite MGF of X on a neighbourhood of zero) There is a positive constant c_3 such that

$$\mathbb{E}[\exp(\lambda X)] < \infty \quad \forall|\lambda| \leq c_3. \tag{10}$$

Definition 3.2. *(sub-exponential distribution) A random variable X is sub-exponential if $X - \mathbb{E}[X]$ satisfies any property in Claim 3.3. In particular, we say X is sub-exponential with parameters (ν, α) if $X - \mathbb{E}[X]$ satisfies (I) with the constants (ν, α).*

With the definition above, we immediately see that sub-Gaussian with σ is sub-exponential with $(\sigma, 0)$. In fact, sub-exponential distribution behaves similarly to sub-Gaussian at a neighbourhood of zero and similarly to exponential distribution further. In other words, we have the following claim.

Claim 3.4. *Given X is sub-exponential with parameters (ν, α), for any $t \geq 0$ we have*

$$\mathbb{P}(X - \mu \geq t) \leq \begin{cases} e^{-t^2/2\nu^2} & 0 \leq t \leq \frac{\nu^2}{\alpha}, \\ e^{-t/2\alpha} & t > \frac{\nu^2}{\alpha}. \end{cases} \tag{11}$$

Proof. Using Chernoff's inequality and the sub-exponential tail, we may arrive at the below for all $\lambda \in [0, 1/\alpha]$,

$$\mathbb{P}(X - \mu \geq t) \leq \exp\Big(- \lambda t + \frac{\lambda^2 \nu^2}{2} \Big) =: \exp(g(\lambda, t)).$$

The value of $\arg\min_\lambda g(\lambda, t)$ without the constraint $\lambda \in [0, 1/\alpha]$ is $\lambda^* := t/\nu^2$, so if $t \leq \frac{\nu^2}{\alpha}$,

$$\arg\min_{\lambda \in [0,1/\alpha]} g(\lambda, t) = \lambda^*.$$

On the other hand, we may assume $t > \frac{\nu^2}{\alpha}$, implying $\lambda^* > 1/\alpha$, so

$$\arg\min_{\lambda \in [0,1/\alpha]} g(\lambda, t) = 1/\alpha.$$

Plugging in the two cases, we obtain our results. □

Checking a random variable is sub-exponential distribution might not be straightforward if we cannot practically compute/bound the MGF, so alternatively we can look at its polynomial moments. Let X be a random variable with mean μ and variance σ^2, we say that **Bernstein's condition** with parameter b holds if

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

When X satisfies Bernstein's condition with parameter b, we can actually show it is sub-exponential with parameters $(\sqrt{2}\sigma, 2b)$ (proof omitted here). With Bernstein's condition, we have the following Bernstein-type inequality.

Theorem 3.2. *(Bernstein's inequality) If X satisfies the Bernstein's condition with parameter b, we have*

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq \exp\Big\{ \frac{\lambda^2 \sigma^2 / 2}{1 - b|\lambda|} \Big\}, \quad \forall |\lambda| \leq \frac{1}{b}.$$

Moreover, for all $t \geq 0$,

$$\mathbb{P}\Big(X - \mu \geq t \Big) \leq \exp\Big\{ - \frac{t^2}{2(\sigma^2 + bt)} \Big\}.$$

TBD: sub-exponential sum is also sub-exponential.

4 Mcdiarmid's inequality

Given vectors $x = (x_j)_{j=1}^n, x' = (x'_j)_{j=1}^n \in \mathbb{R}^n$ and an index $k \in [n]$, we define $x^{\setminus k} := (x_j^{\setminus k})_{j=1}^n$ such that

$$x_j^{\setminus k} = \begin{cases} x_j, & \text{if } j \neq k, \\ x'_k, & \text{if } j = k. \end{cases}$$

Definition 4.1 (Bounded difference property)**.** *We say that $f\mathbb{R}^n \to \mathbb{R}$ satisfies the bounded difference property with parameters $(L_1, \ldots, L_n)$ if*

$$|f(x) - f(x^{\setminus k})| \leq L_k, \quad \textit{for all } x, x' \in \mathbb{R}^n.$$

Theorem 4.1. *Suppose that f satisfies the bounded difference property* (12) *with parameter $(L_1, \ldots, L_n)$ and $X = (X_1, \ldots, X_n)$ has independent components. Then*

$$\mathbb{P}(|f(X) - \mathbb{E}f(X)| \geq t) \leq 2\exp(-\frac{2t^2}{\sum_{k=1}^n L_k^2}).$$

5 Lipschitz functions

6 Appendix A

6.1 Proof of Claim 3.1

Proof. We establish the proof by the cycle (I) $\Rightarrow$ (II) $\Rightarrow$ (III) $\Rightarrow$ (I), and then (I) $\Leftrightarrow$ (IV).

From (I) to (II), it suffices to work on the one-sided result in (3). We show below that given $Z \sim N(0, 2\sigma^2)$, for all $s \geq 0$,[1]

$$\frac{\mathbb{P}(X \geq s)}{\mathbb{P}(Z \geq s)} \leq \sqrt{2\pi}\sqrt{8}e.$$

First using Chernoff's inequality, we have $\mathbb{P}(X \geq s) \leq \exp(-s^2/2\sigma^2)$. Using Mills ratio, we also have

$$\mathbb{P}(Z \geq s) = \mathbb{P}(Z/\sqrt{2\sigma^2} \geq s/\sqrt{2\sigma^2}) \geq \frac{1}{\sqrt{2\pi}}\left\{\frac{\sqrt{2}\sigma}{s} - \frac{(\sqrt{2}\sigma)^3}{s^3}\right\}\exp(-s^2/4\sigma^2).$$

Consider $s \in [0, 2\sigma]$, we have the above decreasing and hence

$$\frac{\mathbb{P}(X \geq s)}{\mathbb{P}(Z \geq s)} \leq \frac{1}{\mathbb{P}(Z \geq 2\sigma)} \leq \sqrt{2\pi}\sqrt{8}e.$$

On the other hand, consider $s \in (2\sigma, \infty)$. By sub-Gaussian tail above, we have

$$\begin{aligned}\frac{\mathbb{P}(X \geq s)}{\mathbb{P}(Z \geq s)} &\leq \sup_{s \geq 2\sigma}\left\{\sqrt{2\pi}\,\frac{s^3}{s^2\sqrt{2}\sigma - (\sqrt{2}\sigma)^3}\,\exp(-s^2/4\sigma^2)\right\} \\ &= \sup_{t \geq 2}\left\{\sqrt{2\pi}\,\frac{t^3}{t^2\sqrt{2} - (\sqrt{2})^3}\,\exp(-t^2/4)\right\} = \sqrt{2\pi}\sup_{t \geq 2}\left\{t^3\exp(-t^2/4)\right\} \leq \sqrt{2\pi}\sqrt{8}e.\end{aligned}$$

To show (II) to (III), let $Z \sim N(0, \tau^2)$ from (II), we have

$$\begin{aligned}\mathbb{E}[X^{2k}] &= \int_0^\infty \mathbb{P}(X^{2k} \geq u)du = \int_0^\infty \mathbb{P}(|X| \geq u^{1/2k})du \\ &\leq c\int_0^\infty \mathbb{P}(|Z| \geq u^{1/2k})du = c\,\mathbb{E}[Z^{2k}] = c\frac{(2k)!}{2^k k!}\tau^{2k} \leq \frac{(2k)!}{2^k k!}(c\tau)^{2k}.\end{aligned}$$

To show (III) to (I), notice first that for any $k = 1, 2, \ldots$, by Cauchy-Schwarz inequality we have

$$\mathbb{E}[(\lambda X)^{2k+1}] \leq \sqrt{\mathbb{E}[\lambda^{2k}X^{2k}] \cdot \mathbb{E}[\lambda^{2k+2}X^{2k+2}]} \leq \frac{1}{2}\Big(\lambda^{2k}\mathbb{E}[X^{2k}] + \lambda^{2k+2}\mathbb{E}[X^{2k+2}]\Big).$$

[1] I think in Wainwright's book (Wainwright 2019) the constant $\sqrt{2\pi}$ is missing as a typo.

Thus, using θ from (III),

$$\begin{aligned}\mathbb{E}[e^{\lambda X}] &= 1 + \mathbb{E}[X] + \sum_{j=2}^{\infty} \frac{\mathbb{E}[(\lambda X)^j]}{j!} \\ &= 1 + \sum_{k=1}^{\infty} \frac{1}{2(2k+1)!}\Big(\lambda^{2k}\mathbb{E}[X^{2k}] + \lambda^{2k+2}\mathbb{E}[X^{2k+2}]\Big) + \sum_{k=1}^{\infty} \frac{1}{(2k)!}\lambda^{2k}\mathbb{E}[X^{2k}] \\ &\leq 1 + \sum_{k=1}^{\infty} \frac{2^k}{(2k)!}\lambda^{2k}\mathbb{E}[X^{2k}] \leq 1 + \sum_{k=1}^{\infty} \frac{1}{k!}(\theta^2\lambda^2)^k = \exp(\theta^2\lambda^2),\end{aligned}$$

completing the proof from (III) to (I) by identifying $\sigma = \sqrt{2}\theta$.

The remaining is to show (I) equivalent to (IV). For (I) $\Rightarrow$ (IV), notice

$$\begin{aligned}\frac{\sqrt{2\pi s}}{\sigma}\exp(sx^2/2\sigma^2) &= \int_{-\infty}^{\infty} \exp\Big(\lambda x - \frac{\lambda^2\sigma^2}{2s}\Big)d\lambda, \\ \int_{-\infty}^{\infty} \exp\Big(\frac{\lambda^2\sigma^2(s-1)}{2s}\Big)d\lambda &= \frac{1}{\sigma}\sqrt{\frac{2\pi s}{1-s}},\end{aligned}$$

then taking expectation on each of the above, with Fubini's theorem, (IV) is obtained. For (IV) $\Rightarrow$ (I), we refer to Chapter 2.4 in Wainwright (2019). $\square$

6.2 Hoeffding's inequality using Hoeffiding's lemma

The following Hoeffding's inequality is the best we can do for bounded random variables as the variance is usually much smaller than $\sum_{i=1}^n (b_i - a_i)^2/4$.

Theorem 6.1. *(Hoeffding's inequality using Hoeffding's lemma) Let $X_1, \ldots, X_n$ be independent random variables each with support $[a_i, b_i]$. Define $S_n = \sum_{i=1}^n (X_i - \mathbb{E}[X_i])$, then for any positive t we have*

$$\mathbb{P}(S_n \geq t) \leq \exp\Big\{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big\}. \tag{12}$$

Proof. (Sketch) Using Lemma 6.1, we have

$$\psi_{S_n}(\lambda) = \sum_{i=1}^n \psi_{X_i}(\lambda) \leq \frac{\lambda^2}{8}\sum_{i=1}^n (b_i - a_i)^2,$$

then compute the Cramér transform and use Theorem 2.1 we may obtain the result. $\square$

Lemma 6.1. *(Hoeffding's lemma) Let Y be some zero mean random variable with support $[a, b]$. Then for any real value λ,*

$$\psi_Y(\lambda) \leq \frac{(b-a)^2\lambda^2}{8}.$$

Proof. Notice that $\left|Y - \frac{a+b}{2}\right| \leq \frac{b-a}{2}$, we have

$$\mathrm{Var}(Y) = \mathrm{Var}(Y - (a+b)/2) \leq (b-a)^2/4.$$

Then the proof is complete once we integrate both sides of the following, given the fact that $\psi_Y(0) = \psi_Y'(0) = 0$,

$$\psi_Y(\lambda)'' = \mathrm{Var}(Y) \leq (b-a)^2/4.$$

$\square$

References

Wainwright, M. J. (2019), *Basic tail and concentration bounds*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, p. 21–57.