High-Dimensional Statistics: A Non-Asymptotic Viewpoint

Chapter 2: Basic tail and concentration bounds

Notes by Renxiong Liu

For high-dimensional statistics, we will often be confronted with theorems on "bounds". In this chapter, we explore some elementary techniques for obtaining both deviation and concentration inequalities. These techniques will serve as the starting points for large-deviation bounds and concentration of measure.

2.1 Classical bounds

The most elementary bound that we have seen is given by the *Markov inequality*, which often provides a relative loose result in terms of bounding.

Theorem 2.1 (Markov inequality)**.** *Let X be a non-negative random variable. For any $t > 0$, we have that*

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}X}{t} \tag{2.1}$$

Several natural corollary of *Markov inequality* (c.f. Theorem 2.1) is given as below.

Corollary 2.1. *Let X be a random variable such that $\mathbb{E}|X|^k < \infty$, then for all $t > 0$*

1. *(Chebyshevs inequality) Suppose $k \geq 2$. Let $\mu = \mathbb{E}X$, then*

$$\mathbb{P}(|X - \mu| > t) \leq \frac{var(X)}{t^2} \tag{2.2}$$

2. *Suppose $k \geq 1$, Let $\mu = \mathbb{E}X$, then for all $t > 0$*

$$\mathbb{P}(|X - \mu| > t) \leq \frac{\mathbb{E}|X - \mu|^k}{t^k}$$

The other classic bound is called *Chernoff bound*, which is a generic method to construct the bound.

Theorem 2.2 (Chernoff-bound)**.** *Let X be a random variable such that $\mathbb{E}|X| < \infty$. Then for any $t \in \mathbb{R}$, we have that*

$$\mathbb{P}(X - \mathbb{E}X > t) \leq \min_{\lambda > 0} \frac{1}{e^{\lambda t}} \mathbb{E}e^{\lambda(X - \mathbb{E}X)} \tag{2.3}$$

In next section, we will see that Chernoff bounding trick serves as the most basic technique to control the tail probability.

2.2 Sub-Gaussian variables and Hoeffding bound

In this section, we introduce sub-Gaussian random variables and present one widely discussed bound, named *Hoeffding bound*.

Definition 2.1 (sub-Gaussian variables)**.** A random variable X is sub-Gaussian with parameter σ if for any $\lambda \in \mathbb{R}$,

$$\mathbb{E}e^{\lambda(X-\mathbb{E}X)} \leq e^{\frac{1}{2}\sigma^2\lambda^2}$$

Note for any Gaussian random variable $X \in N(\mu, \sigma^2)$, X must be a sub-Gaussian with parameter σ. Hence the definition of sub-Gaussian variables extends the scope of discussed random variables from Gaussian to a larger class. Some examples are given below.

Example 2.1 (Rademacher variables)**.** A Rademacher random variable ϵ takes value $\{-1, 1\}$ with equal probability. In this example, we show that ϵ is sub-Gaussian variable with $\sigma = 1$.

Proof. We use series expansion to show this claim. We have that

$$\mathbb{E}e^{\lambda(\epsilon-\mathbb{E}\epsilon)} = \mathbb{E}e^{\lambda\epsilon} = \mathbb{E}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\epsilon^k = \mathbb{E}\sum_{k=0}^{\infty}\frac{\lambda^{2k}}{(2k)!}\epsilon^{2k} = \sum_{k=0}^{\infty}\frac{\lambda^{2k}}{(2k)!} \leq \sum_{k=0}^{\infty}\frac{\lambda^{2k}}{2^k k!} = e^{\frac{1}{2}\lambda^2}$$

This completes the proof.

Example 2.2 (bounded variables)**.** To simplify the proof, let X be a random variable such that $X \in [a, b]$ and $\mathbb{E}X = 0$. We show that X is sub-Gausian with $\sigma = (b - a)$.

Proof. We apply the *symmetrization argument* to show this claim. Let X' be an independent copy of X, that is, X, X' are i.i.d. We have that

$$\mathbb{E}e^{\lambda(X-\mathbb{E}X)} = \mathbb{E}e^{\lambda(X-\mathbb{E}X')} = \mathbb{E}e^{\lambda\mathbb{E}[X-X'|X]} \leq \mathbb{E}e^{\lambda(X-X')}$$

Here we apply *Jensen's inequality*. To proceed, let ϵ be a Rademacher variable and independent with X, X'. Note

$$\mathbb{E}e^{\lambda\epsilon(X-X')} = \mathbb{E}\{\mathbb{E}[e^{\lambda\epsilon(X-X')} \mid \epsilon]\} = \frac{1}{2}\mathbb{E}e^{\lambda\epsilon(X-X')} + \frac{1}{2}\mathbb{E}e^{\lambda\epsilon(X'-X)} = \mathbb{E}e^{\lambda(X-X')}$$

We have that

$$\mathbb{E}e^{\lambda(X-X')} = \mathbb{E}\{\mathbb{E}[e^{\lambda\epsilon(X-X')} \mid X, X']\} \leq \mathbb{E}\{e^{\frac{1}{2}\lambda^2(X-X')^2}\} \leq e^{\frac{1}{2}(b-a)^2\lambda^2}$$

This completes the proof.

Next, we present the useful Hoeffding bound and give some remarks on how to use that.

Theorem 2.3 (Hoeffding bound)**.** *Let $\{X_i\}_{i=1}^n$ be independent sub-Gaussian random variables with parameters $\{\sigma_i\}_{i=1}^n$. Then for any $t > 0$, we have that*

$$\mathbb{P}(\sum_{k=1}^{n}(X_k - \mathbb{E}X_k) > t) \leq \exp(-\frac{t^2}{2\sum_{k=1}^n \sigma_k^2}) \tag{2.4}$$

Proof. By Chernoff-bound trick, for $t > 0$, we have that

$$\mathbb{P}(\sum_{k=1}^{n}(X_k - \mathbb{E}X_k) > t) \leq \frac{1}{\exp(\lambda t)} \prod_{k=1}^{n} \mathbb{E}e^{\lambda(X_k - \mathbb{E}X_k)} \leq \exp(\frac{1}{2}\lambda^2 \sum_{k=1}^{n} \sigma_k^2 - \lambda t)$$

Note

$$\min_{\lambda > 0} \exp(\frac{1}{2}\lambda^2 \sum_{k=1}^{n} \sigma_k^2 - \lambda t) = \exp(-\frac{t^2}{2\sum_{k=1}^{n} \sigma_k^2})$$

This finishes the proof.

Remark. We leave out Theorem 2.6 in the book, which states some equivalent form of defining sub-Gaussian.

2.3 Sub-exponential variables and Bernstein bounds

In this section, we extend the scope of discussed random variables from sub-Gaussian to sub-exponential and present the Bernstain bound, which is widely used in literatures.

Definition 2.2 (sub-exponential variables)**.** A random variable X is sub-exponential with parameter (σ, α) if for any $|\lambda| \leq \frac{1}{\alpha}$,

$$\mathbb{E}e^{\lambda(X - \mathbb{E}X)} \leq e^{\frac{1}{2}\sigma^2\lambda^2}$$

It's easy to see that if X is a sub-Gaussian variable with parameter σ, then it's a sub-exponential variable with parameters $(\sigma, 0)$. In next example, we will show that a sub-exponential variable is not necessary a sub-Gaussian variable, which means the family of sub-exponential variables is larger than the family of sub-Gaussian variables.

Example 2.3 (sub-exponential is not sub-Gaussian)**.** Let $Z \sim N(0, 1)$ be a standard normal random variable. Denote $X = Z^2$, in this example, we show

(1). X is not sub-Gaussian

(2). X is sub-exponential with parameter $(\sigma, \alpha) = (2, 4)$.

Proof. To see this result, note $\mathbb{E}X = 1$, we have that

$$\mathbb{E}\exp(\lambda(X - 1)) = \int_{\mathbb{R}} e^{\lambda(z^2 - 1)} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \frac{e^{-\lambda}}{\sqrt{1 - 2\lambda}} \leq \exp(\frac{1}{2} \cdot 4\lambda^2), \text{ for } |\lambda| < \frac{1}{4}$$

which validates (2). To see why X is not sub-Gaussian, note

$$\mathbb{E}\exp(\lambda(X - 1)) = \frac{e^{-\lambda}}{\sqrt{1 - 2\lambda}}$$

when $\lambda \to 1/2^+$, we will have $\mathbb{E}\exp(\lambda(X - 1)) \to \infty$, which is a contradiction.

Theorem 2.4 (Sub-exponential tail bound)**.** *Let X be sub-exponential with parameter (σ, α). Then for any $t > 0$, we have that*

$$\mathbb{P}(X - \mathbb{E}X > t) \leq \begin{cases} \exp(-\frac{t^2}{2\sigma^2}), & \text{if } 0 \leq t \leq \frac{\sigma^2}{\alpha} \\ \exp(-\frac{t}{2\alpha}), & \text{if } t > \frac{\sigma^2}{\alpha} \end{cases} \tag{2.5}$$

Proof. By definition, for $\lambda \in (0, \frac{1}{\alpha})$, we have that

$$\mathbb{E}\exp(\lambda(X - \mathbb{E}X)) \leq \exp(\frac{1}{2}\sigma^2\lambda^2)$$

which implies that for $\lambda \in (0, \frac{1}{\alpha})$,

$$\mathbb{P}(X - \mathbb{E}X > t) \leq \frac{\mathbb{E}\exp(\lambda(X - \mathbb{E}X))}{\exp(\lambda t)} \leq \exp(\frac{1}{2}\sigma^2\lambda^2 - \lambda t)$$

Note for $G(\lambda) = \frac{1}{2}\sigma^2\lambda^2 - \lambda t$, $\lambda \in (0, \frac{1}{\alpha})$, we have that

$$\min_{\lambda \in (0, \frac{1}{\alpha})} G(\lambda) = \begin{cases} -\frac{t^2}{2\sigma^2}, & \text{if } t < \frac{\sigma^2}{\alpha} \\ -\frac{2t\alpha - \sigma^2}{2\alpha^2}, & \text{if } t > \frac{\sigma^2}{\alpha} \end{cases}$$

Moreover, when $t > \frac{\sigma^2}{\alpha}$, we have that $-\frac{2t\alpha - \sigma^2}{2\alpha^2} < -\frac{t}{2\alpha}$, this means

$$\mathbb{P}(X - \mathbb{E}X > t) \leq \begin{cases} \exp(-\frac{t^2}{2\sigma^2}), & \text{if } t < \frac{\sigma^2}{\alpha} \\ \exp(-\frac{t}{2\alpha}), & \text{if } t > \frac{\sigma^2}{\alpha} \end{cases}$$

This completes the proof.

Next, we introduce an important sufficient condition for a random variale to be sub-exponential, which is named *Bernsteins condition* in literatures.

Condition 2.1 (Bernsteins condition)**.** Let X be a random varible with arbitrary order of moments, i.e. for any $k \geq 1$, $\mathbb{E}|X|^k < \infty$. Denote $\mu = \mathbb{E}X$, $\sigma^2 = \text{var}(X)$. We say X satisfy the Bernsteins condition, if for any $k \geq 2$,

$$|\mathbb{E}(X - \mu)^k| \leq \frac{1}{2}\sigma^2 b^{k-2} k! \tag{2.6}$$

Remark. Motivation for *Bernsteins condition*: roughly speaking, we have that

$$\begin{aligned} \mathbb{E}\exp(\lambda(X - \mathbb{E}X)) &= \mathbb{E}\sum_{k=0}^{\infty} \frac{1}{k!}(\lambda(X - \mathbb{E}X))^k \leq 1 + \sum_{k=2}^{\infty} \frac{1}{k!}\lambda^k |\mathbb{E}(X - \mu)^k| \leq 1 + \lambda^2\sigma^2 \sum_{k=0}^{\infty} |\lambda b|^k \\ &= 1 + \frac{\lambda^2\sigma^2}{1 - |\lambda b|} \leq 1 + \frac{1}{2} \cdot 4\sigma^2\lambda^2, \quad \text{for } |\lambda| \leq \frac{1}{2b} \\ &\leq \exp(\frac{1}{2} \cdot 4\sigma^2\lambda^2), \quad \text{for } |\lambda| \leq \frac{1}{2b} \end{aligned}$$

Hence X is sub-exponential with parameter $(2\sigma, 2b)$.

Before proceeding, we summarize the relationship between the set of random variables of (1) sub-Gaussian, (2) sub-exponential and (3) variables satisfying *Bernsteins condition*:

$$(1) \Rightarrow (2) \text{ and } (3) \Rightarrow (2)$$

The converses are not true.

Next, we give a tighter bound based *Bernsteins condition*.

Theorem 2.5 (Bernstein-type bound)**.** *For any random variable satisfying condition 2.1, we have that*

$$\mathbb{E}\exp(\lambda(X-\mathbb{E}X)) \le \exp(\frac{1}{2}\cdot 4\sigma^2\lambda^2), \quad \textit{for } |\lambda| \le \frac{1}{2b} \tag{2.7}$$

Moreover, we have the concentration inequality:

$$\mathbb{P}(|X-\mu| > t) \le 2\exp(-\frac{t^2}{2(\sigma^2+bt)}), \quad \textit{for } t > 0 \tag{2.8}$$

The proof is trivial, so we leave out the details.

Remark. We leave out Theorem 2.13 in the book, which states some equivalent form of defining sub-exponential.

2.4 martingale-based methods

In the last few sections, the discussions are restricted to the sum of independent random variables. In literatures, the bounds are usually not on such linear combination and even relax the assumption of independece. In this section, we introduce the bounding techniques based on the martingale.

We first introduce some basic definitions that are related with martingale.

Definition 2.3 (Martingale)**.** We call $\{X_n\}_{n=1}^{\infty}$ a martingale with respect to filtration $\{\mathcal{F}_n\}_{n=1}^{\infty}$, if

- X_n is adapted to $\mathcal{F}_n$ for $n \ge 1$. That is, $X_n \in \mathcal{F}_n$ for $n \ge 1$.
- X_n is integrable with respect to measure $\mathbb{P}$. That is, $\mathbb{E}|X_n| < \infty$ for $n \ge 1$.
- $\mathbb{E}(X_{n+1} \mid \mathcal{F}_n) = X_n$, for $n \ge 1$.

Definition 2.4 (martingale difference sequence)**.** We call $\{D_n\}_{n=1}^{\infty}$ a martingale difference sequence with respect to filtration $\{\mathcal{F}_n\}_{n=1}^{\infty}$, if

- D_n is adapted to $\mathcal{F}_n$ for $n \ge 1$. That is, $D_n \in \mathcal{F}_n$ for $n \ge 1$.
- D_n is integrable with respect to measure $\mathbb{P}$. That is, $\mathbb{E}|D_n| < \infty$ for $n \ge 1$.
- $\mathbb{E}(D_{n+1} \mid \mathcal{F}_n) = 0$, for $n \ge 1$.

Remark. It's obvious to note that $X_n = \sum_{k=0}^{n} D_k$ is a martingale if and only if D_k is a martingale difference sequence.

Next, we illustrate the motivation for investigating the martingale difference sequence. Let $X = (X_1, \ldots, X_k, \ldots, X_n)$ be a set of independent random variables. (we will see the reason for assuming independence later.) Our goal is to find probability bound for $f(X) - \mathbb{E}f(X)$. The trick for martingal-based method is as follows. Denote $Y_k = \mathbb{E}(f(X) \mid X_1, \ldots, X_k)$ and let $Y_0 = \mathbb{E}(f(X))$.

It's easy to check $\{Y_k\}_{k=0}^n$ is a martingale with respect to $\{\mathcal{F}_k\}_{k=0}^n$, where $\mathcal{F}_k = \sigma(X_1, \ldots, X_k)$. Hence if we can bound the martingale difference: $D_k = Y_k - Y_{k-1}$, then we will achieve the goal.

Our first theorem on martingale-based method inherits the idea of sub-exponential.

Theorem 2.6. *Let $\{D_n\}_{n=1}^\infty$ be a martingale difference sequence with respect to filtration $\{\mathcal{F}_n\}_{n=1}^\infty$. Suppose that $\mathbb{E}(\exp(\lambda D_{n+1}) \mid \mathcal{F}_n) \le \exp(\frac{1}{2}\sigma_n^2\lambda^2)$, when $|\lambda| < \frac{1}{\alpha_n}$. Then $\sum_{k=1}^n D_k$ is sub-exponential with parameter $(\sqrt{\sum_{k=1}^n \sigma_n^2}, \max_{k=1,\ldots,n} \alpha_k)$.*

The trick of proof is by the property of conditional expectation, we leave the details out here. By using the concentration inequality of sub-exponential (c.f. Theorem 2.4), we will naturally get a concentration inequality for $\sum_{k=1}^n D_k$:

$$\mathbb{P}(|\sum_{k=1}^n D_k| > t) \le \begin{cases} 2\exp(-\frac{t^2}{2\sum_{k=1}^n \sigma_n^2}), & \text{if } 0 \le t \le \frac{\sum_{k=1}^n \sigma_n^2}{\max_{k=1,\ldots,n} \alpha_k} \\ 2\exp(-\frac{t}{2\max_{k=1,\ldots,n} \alpha_k}), & \text{if } t > \frac{\sum_{k=1}^n \sigma_n^2}{\max_{k=1,\ldots,n} \alpha_k} \end{cases} \tag{2.9}$$

One problem of using Theorem 2.6 is, it's hard to check the conditions. A stronger, but easy to check sufficient condition is given in the following theorem.

Theorem 2.7 (Azuma-Hoeffding)**.** *Let $\{D_n\}_{n=1}^\infty$ be a martingale difference sequence with respect to filtration $\{\mathcal{F}_n\}_{n=1}^\infty$. If for $k \ge 1$, there exists $a_k < b_k$, such that $D_k \in [a_k, b_k]$ almost surely, then we have that*

$$\mathbb{P}(|\sum_{k=1}^n D_k| > t) \le 2\exp(-\frac{2t^2}{\sum_{k=1}^n (b_k - a_k)^2}) \tag{2.10}$$

Proof. We apply Theorem 2.6 and its corollary (2.9) to show this claim. Similar to the proof for claim that bounded random variable is sub-Gaussian, we have can show $\{D_n\}$ satisfies $\mathbb{E}(\exp(\lambda D_{n+1}) \mid \mathcal{F}_n) \le \exp(\frac{1}{2} \cdot \frac{1}{4}(b_n - a_n)^2\lambda^2)$ for $\lambda \in \mathbb{R}$. Hence by Theorem 2.6, $\sum_{k=1}^n D_k$ is sub-exponential with parameter $(\frac{1}{2}\sqrt{\sum_{k=1}^n (b_n - a_n)^2}, 0)$. Lastly, by applying (2.9), we prove (2.10). This completes the proof.

Our next theorem comes back to what we discuss in the beginning of this section: we want to bound $f(X) - \mathbb{E}f(X)$. To proceed, we introduce the so-called *bounded-difference property*, which assume the bounded difference in function value for any changes in single coordinate. Given $x, x' \in \mathbb{R}^n$ and index $k \in \{1, 2, \ldots, n\}$, we define $x^{\backslash k}$ by

$$x_j^{\backslash k} = \begin{cases} x_j, & \text{if } j \ne k \\ x'_k, & \text{if } j = k \end{cases}$$

With notations above, the bounded-difference property is stated as follows.

Condition 2.2 (bounded-difference property)**.** We say $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the bounded difference property with parameters $(L_1, \ldots, L_n)$, if for any $k = 1, 2 \ldots, n$,

$$|f(x) - f(x^{\backslash k})| \le L_k$$

The next theorem closes the discussion of this section.

Theorem 2.8 (Bounded differences inequality)**.** *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ satisfies bounded-difference property with parameters $(L_1, \ldots, L_n)$. (c.f. Condition 2.2) Let $X = (X_1, \ldots, X_n)$ be a sequences of independent random variables. Then for any $t > 0$, it holds that*

$$\mathbb{P}(|f(X) - \mathbb{E}f(X)| > t) \leq 2\exp(-\frac{2t^2}{\sum_{k=1}^n L_k^2}) \tag{2.11}$$

Proof. We apply Azuma-Hoeffding Lemma (c.f. Theorem 2.7) to show this claim. Recall the trick we introduce in the beginning of this section, we denote $Y_k = \mathbb{E}(f(X) \mid X_1, \ldots, X_k)$ and let $Y_0 = \mathbb{E}(f(X))$. Define $D_k = Y_k - Y_{k-1}$, (2.11) indeed investigates the bound for $\mathbb{P}(|\sum_{k=1}^n D_k| > t)$.

Define

$$\begin{aligned} A_k &= \inf_x \mathbb{E}(f(X) \mid X_1, \ldots, X_{k-1}, x) - \mathbb{E}(f(X) \mid X_1, \ldots, X_{k-1}) \\ B_k &= \sup_x \mathbb{E}(f(X) \mid X_1, \ldots, X_{k-1}, x) - \mathbb{E}(f(X) \mid X_1, \ldots, X_{k-1}) \end{aligned}$$

Note

$$D_k = \mathbb{E}(f(X) \mid X_1, \ldots, X_k) - \mathbb{E}(f(X) \mid X_1, \ldots, X_{k-1})$$

We have that $D_k \in [A_k, B_k]$ almost surely. Next, we show that $B_k - A_k \leq L_k$ almost surely. To see this result, note that $X_1, X_2 \ldots, X_n$ are independent, we have that for any x,

$$\mathbb{E}(f(X) \mid X_1, \ldots, X_{k-1}, x) = \mathbb{E}_{(k+1):n} f(X_1, \ldots, X_{k-1}, x, X_{(k+1):n})$$

This means, for any x, y, we have that

$$\begin{aligned} & \mathbb{E}(f(X) \mid X_1, \ldots, X_{k-1}, x) - \mathbb{E}(f(X) \mid X_1, \ldots, X_{k-1}, y) \\ = \; & \mathbb{E}_{(k+1):n} \left\{ f(X_1, \ldots, X_{k-1}, x, X_{(k+1):n}) - f(X_1, \ldots, X_{k-1}, y, X_{(k+1):n}) \right\} \\ \leq \; & L_k \end{aligned}$$

By swapping the position of $\mathbb{E}(f(X) \mid X_1, \ldots, X_{k-1}, x)$ and $\mathbb{E}(f(X) \mid X_1, \ldots, X_{k-1}, y)$, we have that for any x, y

$$|\mathbb{E}(f(X) \mid X_1, \ldots, X_{k-1}, x) - \mathbb{E}(f(X) \mid X_1, \ldots, X_{k-1}, y)| \leq L_k$$

which implies $B_k - A_k \leq L_k$ almost surely. Hence by using (2.10), we obtain (2.11). This completes the proof.

Remark. Interesting examples, including example 2.23, 2.24 and 2.25, are left out.

2.5 Lipschitz functions of Gaussian variables

In last section, we eventually arrive at bounded differences inequality, which is inpractical in applications. In this section, we relax a little bit on the bounded difference assumption: we assume $f : \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz with respect to Euclidean norm. That is, for $x, y \in \mathbb{R}^n$, we have that

$$|f(x) - f(y)| \leq L\|x - y\|_2$$

Meanwhile, we add an extra requirement on the distribution of X: we assume $X = (X_1, \ldots, X_n)$ is the vector of i.i.d standard normal. With the two assumptions, we have the following theorem:

Theorem 2.9. *Let $X = (X_1, \ldots, X_n)$ be the vector of i.i.d standard normal, and $f : \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz with respect to Euclidean norm. Then $f(X) - \mathbb{E}f(X)$ is sub-Gaussian with parameter at most L. Moreover, we have that*

$$\mathbb{P}(|f(X) - \mathbb{E}f(X)| > t) \leq 2\exp(-\frac{t^2}{2L^2}) \tag{2.12}$$

We leave out the proof and examples 2.30-2.32.