Concentration inequalities in Statistical learning

Zhipeng Zhu

December 20, 2020

1 Background

Inequalities in statistics provide a means of bounding measures and quantities. They are usually used to specify bounds on quantities when these bounds are particularly difficult or intractable to compute. Inequalities play an important role in the algorithm of statistical learning and machine learning. They are involved in underpinning methods or approaches used in actual cases.

Here are several famous concentration inequalities involved in statistical learning.

- Markov's Inequality
- Chebyshev's Inequality
- Bounded Differences Inequality

In this project, I will focus on the Bounded Differences Inequality.

1.1 Chebyshev's Inequality

Chebyshev's Inequality is a basic tool to develop the bounded inequalities for different quantities. Usually it is stated for random variables, while sometimes it can be generalized for measure spaces. Here is the probabilistic statement.

Theorem 1.1. *For some $a \in X$ where $X \subseteq R$, let f be a non-negative function such that $\{f(x) \geq b, forallx \geq a\}$, where $b \in Y$ where $Y \subseteq R$. Then the following inequality holds,*

$$P(x \geq a) \leq \frac{\mathbb{E}f(x)}{b} \tag{1}$$

1.2 Chernoff's bound

The Chernoff's bound is a version of Markov's inequality, as well as the Chebyshev's one.

Suppose the function f is monotonically increasing. Thus, for every $x \geq a$, $f(x) \geq f(a)$. Substitute $b = f(a)$ in Chebyshev's Inequality, and we have:

$$P(x \geq a) \leq \frac{\mathbb{E}f(x)}{f(a)} \tag{2}$$

Markov's Inequality says that

$$P(x \geq a) \leq \frac{\mathbb{E}x}{a} \tag{3}$$

It holds for $a > 0$ and non-negative x,

$$P(|x - \mathbb{E}(x)| \geq a) \leq \frac{\mathbb{E}\left\{|x - \mathbb{E}(x)|^2\right\}}{a^2} = \frac{\mathrm{Var}\{x\}}{a^2} \tag{4}$$

The Chernoff's bound is then,

$$P(x \geq a) \leq \frac{\mathbb{E}e^{sx}}{e^{sa}} \tag{5}$$

1.3 Pre-Gaussian Random Variables

The Pre-Gaussian random variables develop from the Sub-Gaussian random variable. The Sub-Gaussian random variables are defined as

$$\log M_{X-\mu}(\lambda) \leq \frac{\lambda^2\sigma^2}{2} \tag{6}$$

The Pre-Gaussian has one more parameter than the Sub-Gaussian one.

Definition 1.1. *[Bar20] X is pre-Gaussian with parameters (σ^2, b) if, for all $|\lambda| < 1/b$,*

$$\log M_{X-\mu}(\lambda) \leq \frac{\lambda^2\sigma^2}{2} \tag{7}$$

Example:

- Sub-Gaussian X with parameter σ^2 is pre-Gaussian with parameters (σ^2, b) for all $b > 0$.

1.4 Martingale Difference Sequences

The martingales have the following definition.

Definition 1.2. *[Bar20] A sequence Y_n of random variables adapted to a filtration $\mathcal{F}_n$ is a martingale if, for all n,*

$$\begin{aligned} \mathbb{E}\,|Y_n| &< \infty \\ \mathbb{E}\,[Y_{n+1} \mid \mathcal{F}_n] &= Y_n \end{aligned} \tag{8}$$

Note:

- $\mathcal{F}_n$ is a filtration means these σ-fields are nested: $\mathcal{F}_n \subset \mathcal{F}_{n+1}$.
- Y_n is adapted to $\mathcal{F}_n$ means that each Y_n is measurable with respect to $\mathcal{F}_n$.

Then the martingale difference sequencs can be generated based on the definition of martingales.

Definition 1.3. *[Bar20] A sequence D_n of random variables adapted to a filtration $\mathcal{F}_n$ is a martingale Difference sequence if, for all n,*

$$\begin{aligned} \mathbb{E}\,|D_n| &< \infty \\ \mathbb{E}\,[D_{n+1} \mid \mathcal{F}_n] &= 0 \end{aligned} \tag{9}$$

2 Body

2.1 Bounds for Pre-Gaussian Random Variables

First of all, we want to check the bounds on Pre-Gaussian random variables. Based on the definition of Pre-Gaussian, we have the concentration version of definition that:

Theorem 2.1.

$$\textit{For any } t > 0, \max\{P(X \geq t), P(X \leq -t)\} \leq \exp(\frac{-t^2}{2\sigma^2}) \tag{10}$$

Here, we have an expansion of it.

Lemma 2.2. *[Bar20] For X pre-Gaussian with parameters (σ^2, b),*

$$P(X \geq \mu + t) \leq \begin{cases} \exp\left(-\frac{t^2}{2\sigma^2}\right) & \textit{if } 0 \leq t \leq \sigma^2/b \\ \exp\left(-\frac{t}{2b}\right) & \textit{if } t > \sigma^2/b \end{cases} \tag{11}$$

Because $t^2/\sigma^2 > t^2/(\sigma^2 + bt)$ and $t^2/bt > t^2/(\sigma^2 + bt)$, we have

$$P(X \geq \mu + t) \leq \exp\left(-\frac{t^2}{2(\sigma^2 + bt)}\right) \tag{12}$$

Proof. Assume $\mu = 0$. For $0 \leq \lambda < 1/b$,

$$\begin{aligned} P(X \geq t) &\leq \exp(-\lambda t)\mathbb{E}\exp(\lambda X) \\ &\leq \exp\left(-\lambda t + \frac{\lambda^2\sigma^2}{2}\right) \end{aligned} \tag{13}$$

Without the constraint $[0, 1/b)$ on λ, the minimum occurs at $\lambda* = t/\sigma^2$. Thus if

$$t/\sigma^2 < 1/b \Longleftrightarrow t < \sigma^2/b \tag{14}$$

we have

$$P(X \geq t) \leq \exp\left(-\lambda^* t + \lambda^{*2}\sigma^2/2\right) = \exp\left(-t^2/\left(2\sigma^2\right)\right) \tag{15}$$

The function $f : t \to -\lambda t + \frac{\lambda^2\sigma^2}{2}$ is monotonically decreasing in $[0, \lambda*]$, and obviously also in $[0, 1/b] \subset [0, \lambda*]$. If t is larger, the minimum will occur at $\lambda = 1/b$. Therefore, substituting the λ gives

$$P(X \geq t) \leq \exp\left(-t/b + \sigma^2/\left(2b^2\right)\right) \leq \exp(-t/(2b)) \tag{16}$$

where the second inequality follows from $t \geq \sigma^2/b$. □

For independent X_i, pre-Gaussian with parameters (σ_i^2, b_i), the sum $X = X_1 + \cdots + X_n$ is pre-Gaussian with parameters $\sum_i \sigma_i^2, max_i b_i$.

Actually, for $\mathbb{E}X_i = 0$, we have

$$\begin{aligned} M_X(\lambda) &= \prod_i \mathbb{E}\exp\left(\lambda X_i\right) \\ &\leq \prod_i \exp\left(\lambda^2\sigma_i^2/2\right) = \exp\left(\lambda^2 \sum_i \sigma_i^2/2\right) \end{aligned} \tag{17}$$

When $|\lambda| < 1/b_i$, for all i, the inequality holds.

Corollary 2.2.1. *[Bar20] For independent X_i, pre-Gaussian with parameters (σ_i^2, b_i), with mean μ_i,*

$$P\left(\tfrac{1}{n}\sum_{i=1}^n (X_i - \mu_i) \geq t\right) \leq \begin{cases} \exp\left(-nt^2/\left(2\sigma^2\right)\right) & \textit{for } 0 \leq t \leq \sigma^2/b \\ \exp(-nt/(2b)) & \textit{for } t > \sigma^2/b \end{cases} \tag{18}$$

where $\sigma^2 = \sum_i \sigma_i^2$ and $b = \max_i b_i$.

Consider a random variable X with mean μ, variance σ^2, and bound $|X-\mu| \leq b$. Then X is pre-Gaussian with parameters $(2\sigma^2, 2b)$. Hence,

$$P(X \geq \mu + t) \leq \exp\left(-\frac{t^2}{4\left(\sigma^2 + bt\right)}\right) \tag{19}$$

When we improve the constants from $4\left(\sigma^2 + bt\right)$ to $2(\sigma^2 + bt/3)$, we get the Bernstein's Inequality.

Theorem 2.3. *(Bernstein's Inequality)[Kut02] Let $\xi_1, \ldots, \xi_m$ be independent random variable, with $|\xi_k - \mathbb{E}\xi_k| \leq b$ for all k. Let $X = \sum_{k=1}^m \xi_k$, and let $\sigma^2 = Var(X)$. Let $\mu = \mathbb{E}X$. Then we have*

$$P(X \geq \mu + t) \leq \exp\left(-\frac{t^2}{2(\sigma^2 + bt/3)}\right) \tag{20}$$

2.2 Concentration Bounds for Martingales

Recall the definition of martingales from Definition 1.2. The martingale is a sequence of random variables adapted to a filtration if, for all integers n, we have

$$\begin{gathered} \mathbb{E}\left|Y_n\right| < \infty \\ \mathbb{E}\left[Y_{n+1} \mid \mathcal{F}_n\right] = Y_n \end{gathered} \tag{21}$$

For martingale difference sequence D_n s.t. $D_n = Y_n - Y_{n-1}$, we can construct bounds by the properties of Pre-Gaussian random variables which shown in the first section.

Theorem 2.4. *[Bar20] Consider a martingale Difference sequence D_n which is adapted to a filtration $\mathcal{F}_n$. Assume it satisfies*

$$\text{for } |\lambda| \leq 1/b_n \text{ a.s., } \mathbb{E}\left[\exp\left(\lambda D_n\right) \mid \mathcal{F}_{n-1}\right] \leq \exp\left(\lambda^2 \sigma_n^2/2\right) \tag{22}$$

Then $\sum_{i=1}^n D_i$ is pre-Gaussian, with $(\sigma^2, b) = (\sum_{i=1}^n \sigma_i^2, max_i b_i)$.

$$P\left(\left|\sum_i D_i\right| \geq t\right) \leq \begin{cases} 2\exp\left(-t^2/\left(2\sigma^2\right)\right) & \text{if } 0 \leq t \leq \sigma^2/b \\ 2\exp(-t/(2b)) & \text{if } t > \sigma^2/b \end{cases} \tag{23}$$

Proof. Given $|\lambda| < 1/b_n$,

$$\begin{aligned} \mathbb{E}\exp\left(\lambda \sum_i D_i\right) &= \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n-1} D_i\right) \mathbb{E}\left[\exp\left(\lambda D_n\right) \mid \mathcal{F}_{n-1}\right]\right] \\ &\leq \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n-1} D_i\right)\right] \exp\left(\lambda^2 \sigma_n^2/2\right) \end{aligned} \tag{24}$$

Iterating shows that $\sum_i D_i$ is pre-Gaussian. □

Lemma 2.5. *[Bar20] Consider a martingale difference sequence D_i that a.s. falls in an interval of length B_i. Then*

$$P\left(\left|\sum_i D_i\right| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_i B_i^2}\right) \tag{25}$$

Corollary 2.5.1. *(McDiarmid's Inequality) [Sri] Suppose $f: \mathcal{X}^n \to \mathbb{R}$ satisfies the following bounded differences inequality:*

for all $x_1, \ldots, x_n, x_i' \in \mathcal{X}$,

$$\left|f\left(x_1, \ldots, x_n\right) - f\left(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n\right)\right| \leq B_i \tag{26}$$

Then

$$P(|f(X) - \mathbb{E}f(X)| \geq t) \leq 2\exp\left(-\frac{2t^2}{\sum_i B_i^2}\right) \tag{27}$$

Proof. Use the Chernoff's bound (5), we have

$$P(X - \mathbb{E}[X] \geq t) \leq e^{-st} e^{\mathbb{E}[X - \mathbb{E}[Z]]} \tag{28}$$

Now let,

$$V_i = \mathbb{E}[X \mid x_1, \ldots, x_i] - \mathbb{E}[X \mid x_1, \ldots, x_{i-1}], \forall i = 1, .., n \tag{29}$$

Then $V = \sum_{i=1}^{n} V_i = X - \mathbb{E}[X]$.
Therefore,

$$P(Z - \mathbb{E}[X] \geq t) \leq e^{-st} \mathbb{E}\left[e^{\sum_{i=1}^{n} sV_i}\right] = e^{-st} \prod_{i=1}^{n} \mathbb{E}\left[e^{sV_i}\right] \tag{30}$$

Let V_i be bounded by the interval $[L_i, U_i]$. Because $|X - X_i'| \leq B_i$ (26), it follows that $|V_i| \leq B_i$. Thus, we have $|U_i - L_i| \leq B_i$.
Because

$$\mathbb{E}\left[e^{sV_i}\right] \leq e^{\frac{s^2 (U_i - L_i)^2}{8}} \leq e^{\frac{s^2 B_i^2}{8}} \tag{31}$$

We then have

$$P(X - \mathbb{E}[X] \geq t) \leq e^{-ts} \prod_{i=1}^{n} e^{\frac{s^2 B_i^2}{8}} = e^{s^2 \sum_{i=1}^{n} \frac{B_i^2}{8} - st} \tag{32}$$

Then we can minimize the bound respect to s and have

$$2s \sum_{i=1}^{n} \frac{B_i^2}{8} - t = 0 \Rightarrow s = \frac{4t}{\sum_{i=1}^{n} B_i^2} \tag{33}$$

Then the bound is

$$P(X - \mathbb{E}[X] \geq t) \leq e^{\left(\frac{4t}{\sum_{i=1}^{nt} B_i^2}\right)^2 \sum_{i=1}^{n} \frac{B_i^2}{8} - \left(\frac{4t^2}{\sum_{i=1}^{n} B_i^2}\right)} \Rightarrow P(X - \mathbb{E}[X] \geq t) \leq e^{-\frac{2t^2}{\sum_{i=1}^{n} B_i^2}} \tag{34}$$

It follows the Bounded Differences inequality. □

2.3 Applications of Bounded Differences Inequality

The bounded differences inequality has broad applications on calculating the tail-bounds for functions with non-independent random variables. Here are some examples that apply the bounded differences inequality.

Example 1: Rademacher Averages [Bar20]

For a set $A \subset \mathbb{R}^n$, Consider

$$Z = \sup_{a \in A} \langle \epsilon, a \rangle \tag{35}$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ is a sequence of i.i.d uniform $\{\pm 1\}$ random variables. Define the Rademacher complexity of A as $R(A) = \mathbb{E}Z/n$. (a measure of the size of A). The bounded difference approach implies that Z is concentrated around $R(A)$:

Theorem 2.6. *Z is sub-Gaussian with parameter $4 \sum_i \sup_{a \in A} a_i^2$.*

Proof. Rewrite $Z = f(\epsilon) = f(\epsilon_1, \ldots, \epsilon_n)$. Note that a change of ϵ_i will results in a change in Z of no more than $B_i = \sup_{a \in A} 2 |a_i|$. Then the result follows. □

Example 2: Exmpirical Process [Bar20]

For a class F of functions $f : \mathcal{X} \to [0, 1]$, suppose that $X_1, \ldots, X_n$ are i.i.d on $\mathcal{X}$. Consider that

$$Z = \sup_{f \in F} \left| \mathbb{E}f(X) - \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right| =: \| \underbrace{P - P_n}_{\text{emp proc}} \|_F \tag{36}$$

This is called a uniform law of large numbers, if Z converges to 0. It shows that Z is concentrated about $\mathbb{E}Z$:

Theorem 2.7. *Z is sub-Gaussian with parameter $1/n$.*

Proof. Rewrite $Z = g(X_1, \ldots, X_n)$. Note that a change of X_i will results in a change in Z of no more than $B_n = 1/n$. Then the result follows. □

Example 3: U Statistics

Let's define a function f on $\{Y_k\}_{k=1}^{\infty}$ s.t. $f(Y_1, \ldots, Y_n) = \frac{1}{\binom{n}{2}} \sum_{i<j} g\left(Y_i, Y_j\right)$, where $g : \mathbb{R}^2 \to \mathbb{R}$ is a symbolic function and $g(a, b) \leq b, \forall a, b$.

Then we can show that such function f satisfies the bounded differences inequality.

Proof.

$$\begin{aligned} f\left(Y_1, \ldots, Y_k, \ldots, Y_n\right) - f\left(Y_1, \ldots, Y_k', \ldots, Y_n\right) &= \frac{1}{\binom{n}{2}} \sum_{j \neq k} g\left(Y_j, Y_k\right) - g\left(Y_j, Y_k'\right) \\ &\leq \frac{2(2b)}{n(n-1)} \leq \frac{4b}{n} \end{aligned} \tag{37}$$

□

3 Open questions and research directions

3.1 Open questions

In 2000, Talagrand [Tal95] proposed more inequalities in dertermining bounds. Generally he extends the Bounded Differences Inequality to further applicable fields. For example, he extends it towards the form of integration. Besides, he also discusses the optimization (or sharpening) on the bounds based on different objects. Can we develop further specific inequalities for more narrowed objects? In other words, for some specific application and problems, can we simplify the Bounded Differences Inequality for more precise using? Also, can we use the original Bounded Differences Inequality to proof other concept?

For example the "Hamming Distance": [Maj17]

Let $x, x' \in \Omega := \Omega_1 \times \ldots \Omega_n$. We define

$$d_H\left(x, x'\right) := \left|\left\{i : x_i \neq x_i'\right\}\right| \tag{38}$$

and define the set $A_k := \{x : x \in \Omega, d_H(x, A) \leq k\}$

Can we prove that

$$P[X \in A] \cdot P\left[d_H(X, A) \geqslant t\right] \leqslant \exp\left(-t^2/2n\right) \tag{39}$$

3.2 Doob construction

A Doob martingale is a construction of a stochastic process to approximates a given random variable. It has the properties of the martingale with respect to the given filtration.

The Doob construction can be used to help prove the Bounded Differences Inequality by following brief steps. [Bar20]

$$\begin{aligned} Y_i &= \mathbb{E}\left[f(X) \mid X_1^i\right] \\ D_i &= Y_i - Y_{i-1} \\ f(X) - \mathbb{E} f(X) &= \sum_{i=1}^n D_i \end{aligned} \tag{40}$$

Then

$$\begin{aligned} D_i = Y_i - Y_{i-1} &= \mathbb{E}\left[f(X) \mid X_1^i\right] - \mathbb{E}\left[f(X) \mid X_1^{i-1}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[f(X) \mid X_1^i\right] - f(X) \mid X_1^{i-1}\right] \end{aligned} \tag{41}$$

Thus D_i is a random variable that falls in an interval of length no more than B_i.

Similarly, can we use the Doob construction in proof of extended bounded differences inequalities? For example, the inequalities proposed by Talagrand in 2020.

A Exercises

A.1 Exercise 1: Bernstein's Inequality

Prove the Bernstein's Inequality

If $\mathbb{P}(|X_i| \leq c) = 1$ and $\mathbb{E}(X_i) = \mu$, then for any $\epsilon > 0$,

$$\mathbb{P}\left(\left|\bar{X}_n - \mu\right| > \epsilon\right) \leq 2\exp\left\{-\frac{n\epsilon^2}{2\sigma^2 + 2c\epsilon/3}\right\} \tag{42}$$

Solution:

Lemma A.1. *[JW08] Suppose thath $|X| \leq c$ and $\mathbb{E}(X) = 0$, for any $t > 0$,*

$$\mathbb{E}(e^{tX}) \leq \exp\{t^2\sigma^2(\frac{e^{tc} - 1 - tc}{(tc)^2})\} \tag{43}$$

where $\sigma^2 = Var(X)$

Proof. Let $F = \sum_{r=2}^{\infty} \frac{t^{r-2}\mathbb{E}(X^r)}{r!\sigma^2}$. Then,

$$\mathbb{E}\left(e^{tX}\right) = \mathbb{E}\left(1 + tx + \sum_{r=2}^{\infty} \frac{t^r X^r}{r!}\right) = 1 + t^2\sigma^2 F \leq e^{t^2\sigma^2 F} \tag{44}$$

For $r \geq 2$, $\mathbb{E}\left(X^r\right) = \mathbb{E}\left(X^{r-2}X^2\right) \leq c^{r-2}\sigma^2$.

Therefore, we have

$$F \leq \sum_{r=2}^{\infty} \frac{t^{r-2}c^{r-2}\sigma^2}{r!\sigma^2} = \frac{1}{(tc)^2}\sum_{i=2}^{\infty}\frac{(tc)^r}{r!} = \frac{e^{tc} - 1 - tc}{(tc)^2} \tag{45}$$

Hence,

$$\mathbb{E}\left(e^{tX}\right) \leq \exp\left\{t^2\sigma^2\frac{e^{tc} - 1 - tc}{(tc)^2}\right\} \tag{46}$$

□

From Lemma A.1, we can assume that $\mu = 0$ for simplicity. Then

$$\mathbb{E}\left(e^{tX_i}\right) \leq \exp\left\{t^2\sigma_i^2\frac{e^{tc} - 1 - tc}{(tc)^2}\right\} \tag{47}$$

where $\sigma^2 = \mathbb{E}(X_i^2)$.

Then we have

$$\begin{aligned}\mathbb{P}\left(\bar{X}_n > \epsilon\right) &= \mathbb{P}\left(\sum_{i=1}^{n} X_i > n\epsilon\right) = \mathbb{P}\left(e^{t\sum_{i=1}^{n} X_i} > e^{tn\epsilon}\right)\\ &\leq e^{-tn\epsilon}\mathbb{E}\left(e^{t\sum_{i=1}^{n} X_i}\right) = e^{-tn\epsilon}\prod_{i=1}^{n}\mathbb{E}\left(e^{tX_i}\right)\\ &\leq e^{-tn\epsilon}\exp\left\{nt^2\sigma^2\frac{e^{tc} - 1 - tc}{(tc)^2}\right\}\end{aligned} \tag{48}$$

Take $t = (1/c)\log(1 + \epsilon c/\sigma^2)$ we will have"

$$\mathbb{P}\left(\bar{X}_n > \epsilon\right) \leq \exp\left\{-\frac{n\sigma^2}{c^2}h\left(\frac{c\epsilon}{\sigma^2}\right)\right\} \tag{49}$$

where $h(u) = (1 + u)\log(1 + u) - u$.

By noting that $h(u) \geq u^2/(2 + 2u/3)$ for $u \geq 0$, we just show the Bernstein's Inequality.

A.2 Exercise 2

Prove that for any $z > 0$, if $m \geq 3(z_3)\ln(z+3)$, then $\frac{m}{\ln m} > z$. [Kut02]

Solution:

Firstly, we note that

$$\frac{d}{dm}\frac{m}{\ln m} = \frac{\ln m - 1}{\ln^2 m} \tag{50}$$

Thus, the term $\frac{m}{\ln m}$ is increasing when $m > e$.

Then, given that $\ln(z+3) \geq \ln\ln(z+3)$, and $z > 0 \Rightarrow \ln(z+3) > \ln 3$. Hence,

$$\begin{aligned} \frac{m}{\ln m} &\geq \frac{3(z+3)\ln(z+3)}{\ln 3 + \ln(z+3) + \ln\ln(z+3)} \\ &> \frac{3(z+3)\ln(z+3)}{3\ln(z+3)} = z + 3 > z \end{aligned} \tag{51}$$

References

[Bar20] P. Bartlett. *Lecture 4*. CS281B/Stat241B. Statistical Learning Theory. 2020.

[JW08] H. L. John Lafferty and L. Wasserman. *Concentration of Measure*. 2008. URL: http://www.stat.cmu.edu/~larry/=sml/Concentration.pdf.

[Kut02] S. Kutin. "Extensions to McDiarmid's inequality when differences are bounded with high probability". In: (May 2002).

[Maj17] H. K. Maji. *Lecture 07: Independent Bounded Differences Inequality*. CS 59000: Mathematical Toolkit in Computer Science (Spring 2017). 2017.

[Sri] K. Sridharan. *A Gentle Introduction to Concentration Inequalities*.

[Tal95] M. Talagrand. "CONCENTRATION OF MEASURE AND ISOPERIMETRIC INEQUALITIES IN PRODUCT SPACES". In: *Mathématiques de l'Institut des Hautes Scientifiques* 81 (1995), pp. 73–205.