Concentration Inequalities

Jiuzhou Miao

School of Statistics and Mathematics, Zhejiang Gongshang University

March 5, 2025

Materials and resources

- Textbook: *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Martin J. Wainwright. Cambridge. 2019.
- Contents: 2, 4, 5, 6, 7, 8, 12, 13, 15
- Notes and slides: https://jiuzhoumiao.github.io/

Overview

- Tools and techniques:
  1. Concentration inequalities (Chapter 2).
  2. Uniform laws of large number (Chapter 4).
  3. Metric entropy (Chapter 5).
  4. Reproducing kernel Hilbert spaces (Chapter 12).
  5. Minimax lower bounds (Chapter 15).
- Models and estimators:
  1. Random matrices and covariance estimation (Chapter 6).
  2. High dimensional sparse linear models (Chapter 7).
  3. High dimensional principal component analysis (Chapter 8).
  4. Nonparametric least squares (Chapter 13).

Three regimes

- Let n be the sample size and d be the dimension of parameters.

Three regimes

- Let n be the sample size and d be the dimension of parameters.
- Classical asymptotics: n tends to infinity, but d is fixed.

Three regimes

- Let n be the sample size and d be the dimension of parameters.
- Classical asymptotics: n tends to infinity, but d is fixed.
- High-dimensional asymptotics: The pair (n, d) tends to infinity simultaneously while for some scaling function Ψ, the sequence $\Psi(n, d)$ remains fixed, or converges to some value $\alpha \in [0, \infty]$, e.g., $\Psi(n, d) = d/n$.

Three regimes

- Let n be the sample size and d be the dimension of parameters.
- Classical asymptotics: n tends to infinity, but d is fixed.
- High-dimensional asymptotics: The pair (n, d) tends to infinity simultaneously while for some scaling function Ψ, the sequence $\Psi(n, d)$ remains fixed, or converges to some value $\alpha \in [0, \infty]$, e.g., $\Psi(n, d) = d/n$.
- Non-asymptotics: The pair (n, d), as well as other problem parameters, are viewed as fixed, and high-probability statements are made as a function of them.

From Markov to Chernoff

- Markov's inequality: For a non-negative random variable X with finite mean, one has that for all $t > 0$,

$$\mathbb{P}(X \geq t) \leq t^{-1}\mathbb{E}(X).$$

From Markov to Chernoff

- Markov's inequality: For a non-negative random variable X with finite mean, one has that for all $t > 0$,

$$\mathbb{P}(X \geq t) \leq t^{-1}\mathbb{E}(X).$$

- Chebyshev's inequality: For a random variable X with finite variance, one has that for all $t > 0$,

$$\mathbb{P}\{|X - \mu| \geq t\} \leq t^{-2}\mathrm{Var}(X),$$

where $\mu = \mathbb{E}(X)$.

From Markov to Chernoff

- Markov's inequality: For a non-negative random variable X with finite mean, one has that for all $t > 0$,

$$\mathbb{P}(X \geq t) \leq t^{-1}\mathbb{E}(X).$$

- Chebyshev's inequality: For a random variable X with finite variance, one has that for all $t > 0$,

$$\mathbb{P}\{|X - \mu| \geq t\} \leq t^{-2}\mathrm{Var}(X),$$

where $\mu = \mathbb{E}(X)$.

- If X has a central moment of order k, then Markov's inequality yields a more sharp upper bound $t^{-k}\mathbb{E}\{|X - \mu|^k\}$.

From Markov to Chernoff

- Let $\varphi(\lambda) = \mathbb{E}\{e^{\lambda(X-\mu)}\}$ be the moment generating function of $X - \mu$. Assume that there is a constant $b > 0$ such that $\varphi(\lambda)$ exists for all $|\lambda| \le b$.

From Markov to Chernoff

- Let $\varphi(\lambda) = \mathbb{E}\{e^{\lambda(X-\mu)}\}$ be the moment generating function of $X - \mu$. Assume that there is a constant $b > 0$ such that $\varphi(\lambda)$ exists for all $|\lambda| \leq b$.
- For any $\lambda \in [0, b]$, Markov's inequality can be applied to the random variable $e^{\lambda(X-\mu)}$. i.e., for all $t > 0$,

$$\mathbb{P}(X - \mu \geq t) = \mathbb{P}\{e^{\lambda(X-\mu)} \geq e^{\lambda t}\} \leq e^{-\lambda t}\varphi(\lambda).$$

From Markov to Chernoff

- Let $\varphi(\lambda) = \mathbb{E}\{e^{\lambda(X-\mu)}\}$ be the moment generating function of $X - \mu$. Assume that there is a constant $b > 0$ such that $\varphi(\lambda)$ exists for all $|\lambda| \leq b$.
- For any $\lambda \in [0, b]$, Markov's inequality can be applied to the random variable $e^{\lambda(X-\mu)}$. i.e., for all $t > 0$,

$$\mathbb{P}(X - \mu \geq t) = \mathbb{P}\{e^{\lambda(X-\mu)} \geq e^{\lambda t}\} \leq e^{-\lambda t}\varphi(\lambda).$$

- Chernoff's inequality: For all $t > 0$,

$$\log \mathbb{P}(X - \mu \geq t) \leq \inf_{\lambda \in [0,b]} \{-\lambda t + \log \varphi(\lambda)\}.$$

From Markov to Chernoff

Proposition 1 (Gaussian tail bounds)

Let $X \sim \mathcal{N}(\mu, \sigma^2)$, then one has that for all $t > 0$,

(1) *One-sided inequality:*

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

(2) *Two-sided inequality:*

$$\mathbb{P}\big(|X - \mu| \geq t\big) \leq 2e^{-\frac{t^2}{2\sigma^2}}.$$

Sub-Gaussian random variables

Definition 2 (Sub-Gaussian random variable)

A random variable X with mean $\mu = \mathbb{E}(X)$ is sub-Gaussian if there is a positive number σ such that for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left\{e^{\lambda(X-\mu)}\right\} \leq e^{\sigma^2\lambda^2/2}.$$

- $-X$ is sub-Gaussian if and only if X is sub-Gaussian.

Sub-Gaussian random variables

Definition 2 (Sub-Gaussian random variable)

A random variable X with mean $\mu = \mathbb{E}(X)$ is sub-Gaussian if there is a positive number σ such that for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left\{e^{\lambda(X-\mu)}\right\} \leq e^{\sigma^2\lambda^2/2}.$$

- $-X$ is sub-Gaussian if and only if X is sub-Gaussian.
- If $X \sim \mathcal{N}(\mu, \sigma^2)$, then X is sub-Gaussian.

Sub-Gaussian random variables

Definition 2 (Sub-Gaussian random variable)

A random variable X with mean $\mu = \mathbb{E}(X)$ is sub-Gaussian if there is a positive number σ such that for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}\{e^{\lambda(X-\mu)}\} \le e^{\sigma^2\lambda^2/2}.$$

- $-X$ is sub-Gaussian if and only if X is sub-Gaussian.
- If $X \sim \mathcal{N}(\mu, \sigma^2)$, then X is sub-Gaussian.
- If X_1 and X_2 are independent sub-Gaussian variables with parameters σ_1 and σ_2, then $X_1 + X_2$ is sub-Gaussian with parameter $\sqrt{\sigma_1^2 + \sigma_2^2}$.

Sub-Gaussian tail bounds

Proposition 3 (Sub-Gaussian tail bounds)

Let X be a sub-Gaussian random variable with mean $\mu = \mathbb{E}(X)$ and sub-Gaussian parameter σ, then one has that for all $t > 0$,

$$\mathbb{P}\big(|X - \mu| \geq t\big) \leq 2e^{-\frac{t^2}{2\sigma^2}}.$$

Hoeffding's inequality

Theorem 4 (Hoeffding's inequality)

Let $\{X_i\}_{i=1}^n$ *are independent sub-Gaussian random variables with mean* $\mu_i = \mathbb{E}(X_i)$ *and sub-Gaussian parameter* σ_i*, then one has that for all* $t > 0$*,*

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

Rademacher random variables

- Let ε be a random variable with

$$\mathbb{P}(\varepsilon = 1) = \mathbb{P}(\varepsilon = -1) = 1/2.$$

  ε is called as Rademacher random variables.

Rademacher random variables

- Let ε be a random variable with

$$\mathbb{P}(\varepsilon = 1) = \mathbb{P}(\varepsilon = -1) = 1/2.$$

  ε is called as Rademacher random variables.

- ε is sub-Gaussian with parameter $\sigma = 1$.

Rademacher random variables

- Let ε be a random variable with

$$\mathbb{P}(\varepsilon = 1) = \mathbb{P}(\varepsilon = -1) = 1/2.$$

  ε is called as Rademacher random variables.
- ε is sub-Gaussian with parameter $\sigma = 1$.
- By taking expectations and using Taylor's expansion

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

Bounded random variables

Proposition 5 (Bounded random variables are sub-Gaussian)

Let X be a zero-mean and supported on some interval $[a, b]$, then one has that X is sub-Gaussian with parameter $\sigma = (b - a)/2$.

- I will provide the proofs for both cases that $\sigma = b - a$ and $\sigma = (b - a)/2$.

Sub-Exponential random variables

Definition 6 (Sub-Exponential random variable)

A random variable X with mean $\mu = \mathbb{E}(X)$ is sub-Exponential if there are non-negative parameters (ν, α) such that for all $|\lambda| < \alpha^{-1}$,

$$\mathbb{E}\{e^{\lambda(X-\mu)}\} \leq e^{\nu^2\lambda^2/2}.$$

- Sub-Gaussian variable (with parameter σ) is sub-Exponential $(\nu, \alpha) = (\sigma, 0)$, $1/0 = \infty$.

Sub-Exponential random variables

Definition 6 (Sub-Exponential random variable)

A random variable X with mean $\mu = \mathbb{E}(X)$ is sub-Exponential if there are non-negative parameters (ν, α) such that for all $|\lambda| < \alpha^{-1}$,

$$\mathbb{E}\{e^{\lambda(X-\mu)}\} \leq e^{\nu^2\lambda^2/2}.$$

- Sub-Gaussian variable (with parameter σ) is sub-Exponential $(\nu, \alpha) = (\sigma, 0)$, $1/0 = \infty$.
- However, the converse statement is not true.

Sub-Exponential but not sub-Gaussian

- Let $Z \sim \mathcal{N}(0,1)$ and $X = Z^2$. X is a Chi-square random variable with 1 degree of freedom.

Sub-Exponential but not sub-Gaussian

- Let $Z \sim \mathcal{N}(0,1)$ and $X = Z^2$. X is a Chi-square random variable with 1 degree of freedom.
- For $\lambda < 1/2$, one has that

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

but for $\lambda \geq 1/2$, the moment generating function is infinite, which reveals that X is not sub-Gaussian.

Sub-Exponential but not sub-Gaussian

- Let $Z \sim \mathcal{N}(0,1)$ and $X = Z^2$. X is a Chi-square random variable with 1 degree of freedom.
- For $\lambda < 1/2$, one has that

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

but for $\lambda \geq 1/2$, the moment generating function is infinite, which reveals that X is not sub-Gaussian.

- Following some calculus, one has that for $|\lambda| < 1/4$,

$$\frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \leq e^{2\lambda^2} = e^{4\lambda^2/2},$$

which shows that X is sub-Exponential with parameters $(\nu, \alpha) = (2, 4)$.

Sub-Exponential tail bounds

Proposition 7 (Sub-Exponential tail bounds)

Let $\{X_i\}_{i=1}^n$ be independent sub-Exponential random variables with mean $\mu_i = \mathbb{E}(X_i)$ and sub-Exponential parameters (ν_i, α_i), then one has that $\sum_{i=1}^n (X_i - \mu_i)$ is sub-Exponential with parameters (ν_*, α_*), where $\nu_* = \left(\sum_{i=1}^n \nu_i^2\right)^{1/2}$, $\alpha_* = \max_{1\le i\le n} \alpha_i$, and

$$\mathbb{P}\left\{\left|\sum_{i=1}^n (X_i - \mu_i)\right| \geq t\right\} \leq \begin{cases} 2e^{-\frac{t^2}{2\nu_*^2}} & , 0 < t \leq \frac{\nu_*^2}{\alpha} \\ 2e^{-\frac{t}{2\alpha_*}} & , t > \frac{\nu_*^2}{\alpha_*} \end{cases}.$$

Chi-square random variables

- Let $\{Z_i\}_{i=1}^n$ be independent $\mathcal{N}(0,1)$ random variables. We have shown that Z_i^2 is sub-Exponential with parameters $(\nu, \alpha) = (2, 4)$ for each i.

Chi-square random variables

- Let $\{Z_i\}_{i=1}^n$ be independent $\mathcal{N}(0,1)$ random variables. We have shown that Z_i^2 is sub-Exponential with parameters $(\nu, \alpha) = (2, 4)$ for each i.
- By using Proposition 7, one has that $\sum_{i=1}^n (Z_i^2 - 1)$ is sub-Exponential with parameters $(\nu_*, \alpha_*) = (2\sqrt{n}, 4)$ and

$$\mathbb{P}\left\{\left|n^{-1}\sum_{i=1}^n Z_i^2 - 1\right| \geq t\right\} = \mathbb{P}\left\{\left|\sum_{i=1}^n (Z_i^2 - 1)\right| \geq nt\right\}$$
$$\leq \begin{cases} 2e^{-\frac{nt^2}{8}} & , 0 < t \leq 1 \\ 2e^{-\frac{nt}{8}} & , t > 1 \end{cases}.$$

Johnson–Lindenstrauss embedding

- Let $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N\}$, $N \geq 2$ be given distinct vectors, with each vector lying in $\mathbb{R}^d$. If d is large, then it might be expensive to store and manipulate the dataset.

Johnson–Lindenstrauss embedding

- Let $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N\}$, $N \geq 2$ be given distinct vectors, with each vector lying in $\mathbb{R}^d$. If d is large, then it might be expensive to store and manipulate the dataset.
- The idea of dimensionality reduction is to construct a mapping $F : \mathbb{R}^d \to \mathbb{R}^m$, with the projected dimension m substan tially smaller than d, that preserves some "essential" features of the dataset.

Johnson–Lindenstrauss embedding

- Let $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N\}$, $N \geq 2$ be given distinct vectors, with each vector lying in $\mathbb{R}^d$. If d is large, then it might be expensive to store and manipulate the dataset.
- The idea of dimensionality reduction is to construct a mapping $F : \mathbb{R}^d \to \mathbb{R}^m$, with the projected dimension m substan tially smaller than d, that preserves some "essential" features of the dataset.
- We consider that F can preserve pairwise distances, or equivalently norms and inner products. i.e., for some $\delta \in (0, 1)$, F satisfies that for all pairs $\boldsymbol{u}_i \neq \boldsymbol{u}_j$,

$$(1-\delta)\|\boldsymbol{u}_i - \boldsymbol{u}_j\|_2^2 \leq \left\|F(\boldsymbol{u}_i) - F(\boldsymbol{u}_j)\right\|_2^2 \leq (1+\delta)\|\boldsymbol{u}_i - \boldsymbol{u}_j\|_2^2.$$

Johnson–Lindenstrauss embedding

Theorem 8 (Johnson-Lindenstrauss embedding)

For N vectors $\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_N\}$, $N \geq 2$ and some $\delta \in (0,1)$, there is a mapping $F : \mathbb{R}^d \to \mathbb{R}^m$ which satisfies that

$$(1-\delta)\|\boldsymbol{u}_i - \boldsymbol{u}_j\|_2^2 \leq \left\|F(\boldsymbol{u}_i) - F(\boldsymbol{u}_j)\right\|_2^2 \leq (1+\delta)\|\boldsymbol{u}_i - \boldsymbol{u}_j\|_2^2,$$

for all pairs $\boldsymbol{u}_i \neq \boldsymbol{u}_j$ with probability at least $1 - N^2 e^{-m\delta^2/8}$.

Bernstein's condition

Definition 9 (Bernstein's condition)

Given a random variable X with mean $\mu = \mathbb{E}(X)$ and variance $\sigma^2 = \mathbb{E}(X^2) - \mu^2$, we say that Bernstein's condition with parameter b holds if for $k = 2, 3, \ldots,$

$$\left|\mathbb{E}\{(X-\mu)^k\}\right| \leq 2^{-1} k! \sigma^2 b^{k-2}.$$

- One sufficient condition for Bernstein's condition to hold is that X be bounded.

Bernstein's condition

Definition 9 (Bernstein's condition)

Given a random variable X with mean $\mu = \mathbb{E}(X)$ and variance $\sigma^2 = \mathbb{E}(X^2) - \mu^2$, we say that Bernstein's condition with parameter b holds if for $k = 2, 3, \ldots,$

$$\left|\mathbb{E}\{(X-\mu)^k\}\right| \le 2^{-1} k! \sigma^2 b^{k-2}.$$

- One sufficient condition for Bernstein's condition to hold is that X be bounded.
- In particular, if $|X - \mu| \le b$, then it is straightforward to verify that Bernstein's condition holds.

Bernstein's inequality

Theorem 10 (Bernstein's inequality)

Let $\{X_i\}_{i=1}^n$ *are independent random variables with mean* $\mu_i = \mathbb{E}(X_i)$ *and variance* $\sigma_i^2 = \mathbb{E}(X_i^2) - \mu_i^2$*. If* X_i *satisfies Bernstein's condition with parameter* b *for* $1 \le i \le n$*, one has that:*

(1) *For all* $|\lambda| < 1/b$*,*

$$\mathbb{E}\{e^{\lambda \sum_{i=1}^n (X_i - \mu_i)}\} \le \exp\left\{\frac{2^{-1}\lambda^2 \sum_{i=1}^n \sigma_i^2}{1 - |\lambda| b}\right\}.$$

(2) *For all* $t > 0$*,*

$$\mathbb{P}\left\{\left|\sum_{i=1}^n (X_i - \mu_i)\right| \ge t\right\} \le 2\exp\left\{-\frac{t^2}{2(bt + \sum_{i=1}^n \sigma_i^2)}\right\}.$$

Motivations

- Let $\{X_k\}_{k=1}^n$ be i.i.d. random variables and $f : \mathbb{R}^n \to \mathbb{R}$. We want to derive the bounds of

$$\mathbb{P}\left[\left|f(X_1, \ldots, X_n) - \mathbb{E}\{f(X_1, \ldots, X_n)\}\right| \geq t\right]$$

for all $t > 0$.

Motivations

- Let $\{X_k\}_{k=1}^n$ be i.i.d. random variables and $f : \mathbb{R}^n \to \mathbb{R}$. We want to derive the bounds of

$$\mathbb{P}\left[\left|f(X_1,\ldots,X_n) - \mathbb{E}\{f(X_1,\ldots,X_n)\}\right| \geq t\right]$$

for all $t > 0$.

- Denote $Y_0 = \mathbb{E}\{f(X_1,\ldots,X_n)\}$, $Y_n = f(X_1,\ldots,X_n)$ and

$$Y_k = \mathbb{E}\{f(X_1,\ldots,X_n)\big|X_1,\ldots,X_k\}$$

for $k = 1,\ldots,n-1$.

Motivations

- Then one can rewrite $f(X_1, \dots, X_n) - \mathbb{E}\{f(X_1, \dots, X_n)\}$ by

$$Y_n - Y_0 = \sum_{k=1}^{n} (Y_k - Y_{k-1}) = \sum_{k=1}^{n} D_k.$$

Motivations

- Then one can rewrite $f(X_1, \ldots, X_n) - \mathbb{E}\{f(X_1, \ldots, X_n)\}$ by

$$Y_n - Y_0 = \sum_{k=1}^{n} (Y_k - Y_{k-1}) = \sum_{k=1}^{n} D_k.$$

- Here, $\{Y_k\}_{k=1}^n$ is a particular example of a martingale sequence, whereas $\{D_k\}_{k=1}^n$ is an example of a martingale difference sequence.

Martingale

Definition 11 (Filtration)

Let $\{\mathcal{A}_k\}_{k=1}^{\infty}$ be a sequence of σ-fields. We say that $\{\mathcal{A}_k\}_{k=1}^{\infty}$ is a filtration if $\mathcal{A}_k \subseteq \mathcal{A}_{k+1}$ for all $k \geq 1$. For a sequence of random variables $\{Y_k\}_{k=1}^{\infty}$, we say that $\{Y_k\}_{k=1}^{\infty}$ is adapted to the filtration $\{\mathcal{A}_k\}_{k=1}^{\infty}$ if Y_k is measurable with respect to $\mathcal{A}_k$ for all $k \geq 1$.

Definition 12 (Martingale)

Let $\{Y_k\}_{k=1}^{\infty}$ be a sequence of random variables adapted to the filtration $\{\mathcal{A}_k\}_{k=1}^{\infty}$. We say that $(Y_k, \mathcal{A}_k)_{k=1}^{\infty}$ is a martingale if for all $k \geq 1$, $\mathbb{E}\big(|Y_k|\big) < \infty$ and almost surely

$$\mathbb{E}(Y_{k+1}|\mathcal{A}_k) = Y_k.$$

Two special cases

- If $\mathcal{A}_k = \sigma(Y_1, \ldots, Y_k)$ in Definition 12, we say simply that $\{Y_k\}_{k=1}^{\infty}$ is a martingale.

Two special cases

- If $\mathcal{A}_k = \sigma(Y_1, \ldots, Y_k)$ in Definition 12, we say simply that $\{Y_k\}_{k=1}^{\infty}$ is a martingale.
- Let $\{X_k\}_{k=1}^{\infty}$ be another sequence of random variables. If $\mathcal{A}_k = \sigma(X_1, \ldots, X_k)$ in Definition 12, we say that $\{Y_k\}_{k=1}^{\infty}$ is a martingale with respect to $\{X_k\}_{k=1}^{\infty}$.

Partial sums as martingales

- Let $\{X_k\}_{k=1}^{\infty}$ be a sequence of i.i.d. random variables with mean zero. Denote $S_k = \sum_{i=1}^{k} X_i$ and $\mathcal{A}_k = \sigma(X_1, \ldots, X_k)$.

Partial sums as martingales

- Let $\{X_k\}_{k=1}^{\infty}$ be a sequence of i.i.d. random variables with mean zero. Denote $S_k = \sum_{i=1}^{k} X_i$ and $\mathcal{A}_k = \sigma(X_1, \ldots, X_k)$.
- One has that for all $k \geq 1$,

$$\mathbb{E}(|S_k|) \leq \sum_{i=1}^{k} \mathbb{E}(|X_k|) < \infty,$$

and almost surely

$$\begin{aligned}\mathbb{E}(S_{k+1}|\mathcal{A}_k) &= \mathbb{E}(S_k + X_{k+1}|X_1, \ldots, X_k)\\ &= S_k + \mathbb{E}(X_{k+1}|X_1, \ldots, X_k)\\ &= S_k.\end{aligned}$$

Doob construction

- Let $\{X_k\}_{k=1}^{n}$ be i.i.d. random variables and $f : \mathbb{R}^n \to \mathbb{R}$.

Doob construction

- Let $\{X_k\}_{k=1}^n$ be i.i.d. random variables and $f : \mathbb{R}^n \to \mathbb{R}$.
- For short hand, denote $X_1^k = (X_1, \dots, X_k)$ for $1 \le k \le n$. Denote $Y_0 = \mathbb{E}\{f(X_1^n)\}$, $Y_n = f(X_1^n)$ and $Y_k = \mathbb{E}\{f(X_1^n) \big| X_1^k\}$ for $k = 1, \dots, n-1$.

Doob construction

- Let $\{X_k\}_{k=1}^n$ be i.i.d. random variables and $f : \mathbb{R}^n \to \mathbb{R}$.
- For short hand, denote $X_1^k = (X_1, \ldots, X_k)$ for $1 \leq k \leq n$. Denote $Y_0 = \mathbb{E}\{f(X_1^n)\}$, $Y_n = f(X_1^n)$ and $Y_k = \mathbb{E}\{f(X_1^n) \big| X_1^k\}$ for $k = 1, \ldots, n-1$.
- Suppose that $\mathbb{E}\Big\{\big|f(X_1^n)\big|\Big\} < \infty$, then one has that for all $k \geq 1$,

$$\mathbb{E}(|Y_k|) \leq \mathbb{E}\left[\mathbb{E}\Big\{\big|f(X_1^n)\big| \Big| X_1^k\Big\}\right] = \mathbb{E}\Big\{\big|f(X_1^n)\big|\Big\} < \infty,$$

and almost surely

$$\begin{aligned} \mathbb{E}(Y_{k+1}|X_1^k) &= \mathbb{E}\left[\mathbb{E}\{f(X_1^n) | X_1^{k+1}\} \Big| X_1^k\right] \\ &= \mathbb{E}\{f(X_1^n) | X_1^k\} \\ &= Y_k. \end{aligned}$$

Likelihood ratio

- Let f and g be two mutually absolutely continuous density functions and $\{X_k\}_{k=1}^{\infty}$ be a sequence of random variables drawn i.i.d. according to density function f.

Likelihood ratio

- Let f and g be two mutually absolutely continuous density functions and $\{X_k\}_{k=1}^{\infty}$ be a sequence of random variables drawn i.i.d. according to density function f.
- For each $k \geq 1$, denote $Y_k = \prod_{i=1}^{k} \{g(X_i)/f(X_i)\}$.

Likelihood ratio

- Let f and g be two mutually absolutely continuous density functions and $\{X_k\}_{k=1}^{\infty}$ be a sequence of random variables drawn i.i.d. according to density function f.
- For each $k \geq 1$, denote $Y_k = \prod_{i=1}^{k} \{g(X_i)/f(X_i)\}$.
- Note that for each $i \geq 1$, one has that

$$\mathbb{E}\{g(X_i)/f(X_i)\} = \int_{-\infty}^{\infty} \{g(x)/f(x)\} f(x) \mathrm{d}x = 1,$$

one has that for each $k \geq 1$, $\mathbb{E}\big(|Y_k|\big) < \infty$ and almost surely

$$\mathbb{E}(Y_{k+1} | X_1^k) = Y_k \mathbb{E}\{g(X_{k+1})/f(X_{k+1})\} = Y_k.$$

Martingale difference sequence

Definition 13 (Martingale difference sequence)

Let $\{D_k\}_{k=1}^{\infty}$ be a sequence of random variables adapted to the filtration $\{\mathcal{A}_k\}_{k=1}^{\infty}$. We say that $(D_k, \mathcal{A}_k)_{k=1}^{\infty}$ is a martingale difference sequence if for all $k \geq 1$, $\mathbb{E}(|D_k|) < \infty$ and almost surely

$$\mathbb{E}(D_{k+1}|\mathcal{A}_k) = 0.$$

- Let $(Y_k, \mathcal{A}_k)_{k=0}^{\infty}$ be a martingale. One can construct a martingale difference sequence $(D_k, \mathcal{A}_k)_{k=1}^{\infty}$ by setting $D_k = Y_k - Y_{k-1}$ for $k \geq 1$.

Concentration inequalities for martingale difference sequences

Theorem 14 (Sub-exponential bounds for martingale difference sequences)

Let $(D_k, \mathcal{A}_k)_{k=1}^{\infty}$ be a martingale difference sequence, and suppose that $\mathbb{E}(e^{\lambda D_k}|\mathcal{A}_{k-1}) \leq e^{\nu_k^2\lambda^2/2}$ almost surely for $|\lambda| < \alpha_k^{-1}$. Then one has that:

(1) *$\sum_{k=1}^{n} D_k$ is sub-exponential with parameters (ν_*, α_*), which is the same as the notation given in Proposition 7.*

(2) *The sum satisfies the concentration inequality*

$$\mathbb{P}\left\{\left|\sum_{k=1}^{n} D_k\right| \geq t\right\} \leq \begin{cases} 2e^{-\frac{t^2}{2\nu_*^2}} & , 0 < t \leq \frac{\nu_*^2}{\alpha} \\ 2e^{-\frac{t}{2\alpha_*}} & , t > \frac{\nu_*^2}{\alpha_*} \end{cases}.$$

Concentration inequalities for martingale difference sequences

Theorem 15 (Azuma–Hoeffding)

Let $(D_k, \mathcal{A}_k)_{k=1}^{\infty}$ *be a martingale difference sequence for which there are constants* $(a_k, b_k)_{k=1}^{n}$ *such that* $D_k \in [a_k, b_k]$ *almost surely for all* $k = 1, \dots, n$. *Then one has that for all* $t > 0$,

$$\mathbb{P}\left\{\left|\sum_{k=1}^{n} D_k\right| \geq t\right\} \leq 2\exp\left\{-\frac{2t^2}{\sum_{i=1}^{n}(b_k - a_k)^2}\right\}.$$

Bounded differences inequality

Definition 16 (Bounded differences property)

Let $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^n$ that their i-th element are x_i and x_i' respectively. Define $\boldsymbol{x}^{\sim k} \in \mathbb{R}^n$ via

$$x_i^{\sim k} = \begin{cases} x_i, & i \neq k \\ x_k', & i = k \end{cases}.$$

Let $f : \mathbb{R}^n \to \mathbb{R}$. We say that f satisfies the bounded difference property with parameters $(L_1, \ldots, L_n)$ if for each $k = 1, \ldots, n$,

$$\left| f(\boldsymbol{x}) - f(\boldsymbol{x}^{\sim k}) \right| \leq L_k$$

for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^n$.

Bounded differences inequality

Theorem 17 (Bounded differences inequality)

Let $\mathbf{X} = (X_1, \ldots, X_n)^\top$ *be a random vector with independent components and* $f : \mathbb{R}^n \to \mathbb{R}$ *which satisfies the bounded difference property with parameters* $(L_1, \ldots, L_n)$*. Then one has that for all* $t > 0$*,*

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

Classical Hoeffding from bounded differences

- Let $\{X_i\}_{i=1}^n$ be independent bounded random variables on $[a,b]$. Consider $f(\boldsymbol{x}) = \sum_{i=1}^n (x_i - \mu_i)$, where $\mu_i = \mathbb{E}(X_i)$.

Classical Hoeffding from bounded differences

- Let $\{X_i\}_{i=1}^n$ be independent bounded random variables on $[a,b]$. Consider $f(\boldsymbol{x}) = \sum_{i=1}^n (x_i - \mu_i)$, where $\mu_i = \mathbb{E}(X_i)$.
- We have that

$$\left|f(\boldsymbol{x}) - f(\boldsymbol{x}^{\sim k})\right| = |x_k - x_k'| \leq b - a.$$

Classical Hoeffding from bounded differences

- Let $\{X_i\}_{i=1}^n$ be independent bounded random variables on $[a, b]$. Consider $f(\boldsymbol{x}) = \sum_{i=1}^n (x_i - \mu_i)$, where $\mu_i = \mathbb{E}(X_i)$.
- We have that

$$\left| f(\boldsymbol{x}) - f(\boldsymbol{x}^{\sim k}) \right| = |x_k - x_k'| \leq b - a.$$

- Then one has that for all $t > 0$,

$$\mathbb{P}\left\{ \left| \sum_{i=1}^n (X_i - \mu_i) \right| \geq t \right\} = \mathbb{P}\left\{ \left| f(\mathbf{X}) - \mathbb{E}\{f(\mathbf{X})\} \right| \geq t \right\}$$
$$\leq 2 \exp\left\{ -\frac{2t^2}{n(b-a)^2} \right\}.$$

U-statistics

- Let $\{X_i\}_{i=1}^n$ be i.i.d. random variables and $g : \mathbb{R}^2 \to \mathbb{R}$ be a symmetric function of its arguments.

U-statistics

- Let $\{X_i\}_{i=1}^n$ be i.i.d. random variables and $g : \mathbb{R}^2 \to \mathbb{R}$ be a symmetric function of its arguments.
- The quantity

$$U_n = \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} g(X_i, X_j)$$

is called as a pairwise U-statistic, which is an unbiased estimator of $\mathbb{E}\{g(X_1, X_2)\}$.

U-statistics

- Let $\{X_i\}_{i=1}^n$ be i.i.d. random variables and $g : \mathbb{R}^2 \to \mathbb{R}$ be a symmetric function of its arguments.
- The quantity

$$U_n = \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} g(X_i, X_j)$$

is called as a pairwise U-statistic, which is an unbiased estimator of $\mathbb{E}\{g(X_1, X_2)\}$.

- Assume that $\|g\|_\infty \le b$ for some $b > 0$ and let

$$f(\boldsymbol{x}) = \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} g(x_i, x_j).$$

U-statistics

- One has that

$$
\begin{aligned}
\left|f(\boldsymbol{x}) - f(\boldsymbol{x}^{\sim k})\right| &\le \binom{n}{2}^{-1} \sum_{i \neq j} \left|g(x_i, x_j) - g(x_i, x_j')\right| \\
&\le \binom{n}{2}^{-1} (n-1) \times 2b \\
&= 4bn^{-1}.
\end{aligned}
$$

U-statistics

- One has that

$$\begin{aligned}\left|f(\boldsymbol{x})-f(\boldsymbol{x}^{\sim k})\right| &\le \binom{n}{2}^{-1}\sum_{i\neq j}\left|g(x_i,x_j)-g(x_i,x_j')\right| \\ &\le \binom{n}{2}^{-1}(n-1)\times 2b \\ &= 4bn^{-1}.\end{aligned}$$

- Then one has that for all $t>0$,

$$\mathbb{P}\left\{\left|U_n-\mathbb{E}(U_n)\right|\ge t\right\}\le 2e^{-nt^2/(8b^2)}.$$

Rademacher complexity

- Let $\{\varepsilon_k\}_{k=1}^n$ be independent Rademacher random variables and $\mathcal{A} \subseteq \mathbb{R}^n$ be a collection of vectors.

Rademacher complexity

- Let $\{\varepsilon_k\}_{k=1}^n$ be independent Rademacher random variables and $\mathcal{A} \subseteq \mathbb{R}^n$ be a collection of vectors.
- For $\boldsymbol{a} = (a_1, \ldots, a_n)^\mathsf{T} \in \mathcal{A}$, define

$$Z(\mathcal{A}) = \sup_{\boldsymbol{a}\in\mathcal{A}} \sum_{k=1}^{n} a_k\varepsilon_k = \sup_{\boldsymbol{a}\in\mathcal{A}} \langle \boldsymbol{a}, \boldsymbol{\varepsilon} \rangle.$$

The quantity $\mathcal{R}(\mathcal{A}) = \mathbb{E}\{Z(\mathcal{A})\}$ is called as the Rademacher complexity of the collection $\mathcal{A}$.

Rademacher complexity

- Let $\{\varepsilon_k\}_{k=1}^n$ be independent Rademacher random variables and $\mathcal{A} \subseteq \mathbb{R}^n$ be a collection of vectors.
- For $\boldsymbol{a} = (a_1, \ldots, a_n)^\mathsf{T} \in \mathcal{A}$, define

$$Z(\mathcal{A}) = \sup_{\boldsymbol{a}\in\mathcal{A}} \sum_{k=1}^n a_k\varepsilon_k = \sup_{\boldsymbol{a}\in\mathcal{A}} \langle \boldsymbol{a}, \boldsymbol{\varepsilon}\rangle.$$

  The quantity $\mathcal{R}(\mathcal{A}) = \mathbb{E}\{Z(\mathcal{A})\}$ is called as the Rademacher complexity of the collection $\mathcal{A}$.
- Let $f(\boldsymbol{\varepsilon}) = \sup_{\boldsymbol{a}\in\mathcal{A}}\langle \boldsymbol{a}, \boldsymbol{\varepsilon}\rangle$. Since $f(\boldsymbol{\varepsilon}^{\sim k}) \geq \langle \boldsymbol{a}, \boldsymbol{\varepsilon}^{\sim k}\rangle$ for any $\boldsymbol{a} \in \mathcal{A}$, one has that

$$\langle \boldsymbol{a}, \boldsymbol{\varepsilon}\rangle - f(\boldsymbol{\varepsilon}^{\sim k}) \leq \langle \boldsymbol{a}, \boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^{\sim k}\rangle \leq a_k(\varepsilon_k - \varepsilon_k') \leq 2|a_k|.$$

Rademacher complexity

- Taking the supremum over $\mathcal{A}$ on both sides, one has that

$$f(\boldsymbol{\varepsilon}) - f(\boldsymbol{\varepsilon}^{\sim k}) \leq 2 \sup_{a \in \mathcal{A}} |a_k|$$

Rademacher complexity

- Taking the supremum over $\mathcal{A}$ on both sides, one has that

$$f(\varepsilon) - f(\varepsilon^{\sim k}) \leq 2 \sup_{a \in \mathcal{A}} |a_k|$$

- By symmetry, one can conclude that f satisfies the bounded difference inequality in coordinate k with parameter $2 \sup_{a \in \mathcal{A}} |a_k|$.

Rademacher complexity

- Taking the supremum over $\mathcal{A}$ on both sides, one has that

$$f(\varepsilon) - f(\varepsilon^{\sim k}) \leq 2 \sup_{a \in \mathcal{A}} |a_k|$$

- By symmetry, one can conclude that f satisfies the bounded difference inequality in coordinate k with parameter $2 \sup_{a \in \mathcal{A}} |a_k|$.
- Hence, $Z(\mathcal{A})$ is sub-Gaussian with parameter $2\sqrt{\sum_{k=1}^{n} \sup_{a \in \mathcal{A}} a_k^2}$.

Lipschitz functions

Definition 18 (Lipschitz functions)

Let $f : \mathbb{R}^n \to \mathbb{R}$ and $L > 0$. We say that f is L-Lipschitz with respect to the Euclidean norm $\|\cdot\|_2$ if

$$\big|f(\boldsymbol{x}) - f(\boldsymbol{y})\big| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|_2$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$.

Lipschitz functions of Gaussian variables

Theorem 19 (Concentration properties of Lipschitz functions of Gaussian variables)

Let $\mathbf{X} = (X_1, \dots, X_n)^\mathsf{T}$ be a vector of i.i.d. standard Gaussian variables and $f : \mathbb{R}_n \to \mathbb{R}$ be L-Lipschitz with respect to the Euclidean norm. Then $f(\mathbf{X}) - \mathbb{E}\{f(\mathbf{X})\}$ is sub-Gaussian with parameter L and for all $t > 0$,

$$\mathbb{P}\left\{\left|f(\mathbf{X}) - \mathbb{E}\{f(\mathbf{X})\}\right| \geq t\right\} \leq 2\exp\left\{-\frac{t^2}{2L^2}\right\}.$$

- I will prove a weaker version that the sub-Gaussian parameter is $\pi L/2$ when f is both Lipschitz and differentiable.

Gaussian complexity

- Let $\{\varepsilon_k\}_{k=1}^n$ be independent $\mathcal{N}(0,1)$ random variables and $\mathcal{A} \subseteq \mathbb{R}^n$ be a collection of vectors.

Gaussian complexity

- Let $\{\varepsilon_k\}_{k=1}^n$ be independent $\mathcal{N}(0,1)$ random variables and $\mathcal{A} \subseteq \mathbb{R}^n$ be a collection of vectors.
- For $\boldsymbol{a} = (a_1, \ldots, a_n)^\mathsf{T} \in \mathcal{A}$, define

$$Z(\mathcal{A}) = \sup_{\boldsymbol{a}\in\mathcal{A}} \sum_{k=1}^n a_k \varepsilon_k = \sup_{\boldsymbol{a}\in\mathcal{A}} \langle \boldsymbol{a}, \boldsymbol{\varepsilon} \rangle.$$

The quantity $\mathcal{R}(\mathcal{A}) = \mathbb{E}\{Z(\mathcal{A})\}$ is called as the Gaussian complexity of the collection $\mathcal{A}$.

Gaussian complexity

- Let $\{\varepsilon_k\}_{k=1}^n$ be independent $\mathcal{N}(0,1)$ random variables and $\mathcal{A} \subseteq \mathbb{R}^n$ be a collection of vectors.
- For $\boldsymbol{a} = (a_1, \ldots, a_n)^\mathsf{T} \in \mathcal{A}$, define

$$Z(\mathcal{A}) = \sup_{\boldsymbol{a} \in \mathcal{A}} \sum_{k=1}^n a_k \varepsilon_k = \sup_{\boldsymbol{a} \in \mathcal{A}} \langle \boldsymbol{a}, \boldsymbol{\varepsilon} \rangle.$$

  The quantity $\mathcal{R}(\mathcal{A}) = \mathbb{E}\{Z(\mathcal{A})\}$ is called as the Gaussian complexity of the collection $\mathcal{A}$.
- Let $f(\boldsymbol{\varepsilon}) = \sup_{\boldsymbol{a} \in \mathcal{A}} \langle \boldsymbol{a}, \boldsymbol{\varepsilon} \rangle$. One has that

$$\langle \boldsymbol{a}, \boldsymbol{\varepsilon} \rangle = \langle \boldsymbol{a}, \boldsymbol{\varepsilon}' \rangle + \langle \boldsymbol{a}, \boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}' \rangle \leq f(\boldsymbol{\varepsilon}') + \Big(\sup_{\boldsymbol{a} \in \mathcal{A}} \|\boldsymbol{a}\|_2\Big) \|\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}'\|_2.$$

Gaussian complexity

- Taking the supremum over $\mathcal{A}$ on both sides, one has that

$$f(\boldsymbol{\varepsilon}) - f(\boldsymbol{\varepsilon}') \le \Big(\sup_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a}\|_2\Big) \|\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}'\|_2.$$

Gaussian complexity

- Taking the supremum over $\mathcal{A}$ on both sides, one has that

$$f(\boldsymbol{\varepsilon}) - f(\boldsymbol{\varepsilon}') \leq \Big(\sup_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a}\|_2\Big) \|\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}'\|_2.$$

- By symmetry, one has that f is a $\sup_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a}\|_2$-Lipschitz function. Then $Z(\mathcal{A})$ is sub-Gaussian with parameter $\sup_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a}\|_2$.

Gaussian chaos variables

- Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be a symmetric matrix with $\mathbf{A} = (a_{ij})_{n \times n}$ and $\mathbf{X}, \mathbf{Y}$ be independent $\mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ random vectors with $\mathbf{X} = (X_1, \dots, X_n)^\mathsf{T}$ and $\mathbf{Y} = (Y_1, \dots, Y_n)^\mathsf{T}$. Define

$$Z_n = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} X_i Y_j = \mathbf{X}^\mathsf{T} \mathbf{A} \mathbf{Y},$$

called as a (decoupled) Gaussian chaos.

Gaussian chaos variables

- Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a symmetric matrix with $\mathbf{A} = (a_{ij})_{n\times n}$ and $\mathbf{X}, \mathbf{Y}$ be independent $\mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ random vectors with $\mathbf{X} = (X_1, \ldots, X_n)^\mathsf{T}$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)^\mathsf{T}$. Define

$$Z_n = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} X_i Y_j = \mathbf{X}^\mathsf{T}\mathbf{A}\mathbf{Y},$$

  called as a (decoupled) Gaussian chaos.
- One has that $\mathbb{E}(Z_n) = 0$, so it is natural to seek a tail bound on Z_n.

Gaussian chaos variables

- Condition on $\mathbf{X}$, one has that Z_n is a mean zero Gaussian random variable with variance $\mathbf{X}^\top \mathbf{A}^2 \mathbf{X} = \|\mathbf{A}\mathbf{X}\|_2^2$. Then one has that for all $\delta > 0$,

$$\mathbb{P}\big(|Z_n| \geq \delta \big| \mathbf{X}\big) \leq 2\exp\left\{-\frac{\delta^2}{2\|\mathbf{A}\mathbf{X}\|_2^2}\right\}.$$

Gaussian chaos variables

- Condition on $\mathbf{X}$, one has that Z_n is a mean zero Gaussian random variable with variance $\mathbf{X}^\top \mathbf{A}^2 \mathbf{X} = \|\mathbf{A}\mathbf{X}\|_2^2$. Then one has that for all $\delta > 0$,

$$\mathbb{P}\big(|Z_n| \geq \delta \big| \mathbf{X}\big) \leq 2 \exp\left\{-\frac{\delta^2}{2\|\mathbf{A}\mathbf{X}\|_2^2}\right\}.$$

- Let $f(\boldsymbol{x}) = \|\mathbf{A}\boldsymbol{x}\|_2$. One can easily show that f is a $\|\mathbf{A}\|_{\mathrm{op}}$-Lipschitz function, where

$$\|\mathbf{A}\|_{\mathrm{op}} = \max_{\|\boldsymbol{x}\|_2 = 1} \|\mathbf{A}\boldsymbol{x}\|_2.$$

Gaussian chaos variables

- By Jensen's inequality, one has that

$$\mathbb{E}(\|\mathbf{AX}\|_2) \leq \sqrt{\mathbb{E}(\|\mathbf{AX}\|_2^2)} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2} = \|\mathbf{A}\|_{\mathrm{F}}.$$

Gaussian chaos variables

- By Jensen's inequality, one has that

$$\mathbb{E}(\|\mathbf{AX}\|_2) \le \sqrt{\mathbb{E}(\|\mathbf{AX}\|_2^2)} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2} = \|\mathbf{A}\|_{\mathrm{F}}.$$

- Then for all $t > 0$, one has that

$$\begin{aligned}\mathbb{P}(\|\mathbf{AX}\|_2 \ge \|\mathbf{A}\|_{\mathrm{F}} + t) &\le \mathbb{P}\left\{\|\mathbf{AX}\|_2 \ge \mathbb{E}(\|\mathbf{AX}\|_2) + t\right\} \\ &\le 2\exp\left\{-\frac{t^2}{2\|\mathbf{A}\|_{\mathrm{op}}^2}\right\}.\end{aligned}$$

Gaussian chaos variables

- By taking $t^2 = \delta \|\mathbf{A}\|_{\mathrm{op}}$ and note that

$$\left(\|\mathbf{A}\|_{\mathrm{F}} + t\right)^2 \leq 2\|\mathbf{A}\|_{\mathrm{F}}^2 + 2t^2,$$

one has that

$$\mathbb{P}\left(\|\mathbf{AX}\|_2^2 \geq 2\|\mathbf{A}\|_{\mathrm{F}}^2 + 2\delta\|\mathbf{A}\|_{\mathrm{op}}\right) \leq 2\exp\left\{-\frac{\delta}{2\|\mathbf{A}\|_{\mathrm{op}}}\right\}.$$

Gaussian chaos variables

- By taking $t^2 = \delta\|\mathbf{A}\|_{\mathrm{op}}$ and note that

$$\big(\|\mathbf{A}\|_{\mathrm{F}} + t\big)^2 \le 2\|\mathbf{A}\|_{\mathrm{F}}^2 + 2t^2,$$

one has that

$$\mathbb{P}\big(\|\mathbf{A}\mathbf{X}\|_2^2 \ge 2\|\mathbf{A}\|_{\mathrm{F}}^2 + 2\delta\|\mathbf{A}\|_{\mathrm{op}}\big) \le 2\exp\left\{-\frac{\delta}{2\|\mathbf{A}\|_{\mathrm{op}}}\right\}.$$

- Finally, one has that

$$\begin{aligned}\mathbb{P}\big(|Z_n| \ge \delta\big) &= \mathbb{E}\Big\{\mathbb{P}\big(|Z_n| \ge \delta \big| \mathbf{X}\big)\Big\} \\ &\le 2\exp\left\{-\frac{\delta^2}{4\|\mathbf{A}\|_{\mathrm{F}}^2 + 4\delta\|\mathbf{A}\|_{\mathrm{op}}}\right\} + 2\exp\left\{-\frac{\delta}{2\|\mathbf{A}\|_{\mathrm{op}}}\right\} \\ &\le 4\exp\left\{-\frac{\delta^2}{4\|\mathbf{A}\|_{\mathrm{F}}^2 + 4\delta\|\mathbf{A}\|_{\mathrm{op}}}\right\}.\end{aligned}$$

Thank You