High-Dimensional Statistics
A Non-Asymptotic Viewpoint

Chengxin Gong, Peking University
https://wqgcx.github.io/

2022 年 10 月 15 日

目录

1 Introduction

1.1 Classical versus high-dimensional theory

- Law of large numbers: $\hat{\mu}_n := \frac{1}{n}\sum_{i=1}^n X_i$ converges in probability to $\mu = \mathbb{E}[X_1]$.
- Central limit theorem: $\sqrt{n}(\hat{\mu}_n - \mu)$ converges in distribution to a centered Gaussian with covariance matrix $\Sigma = \mathrm{cov}(X_1)$.
- In a classical theoretical framework, the ambient dimension d of the data space is typically viewed as fixed. Some essential facts: (1) The data sets arising in many parts of modern science and engineering have a "high-dimensional flavor", with d on the same order as, or possibly larger than, the sample size n; (2) For many of these applications, classical "larger n, fixed d" theory fails to provide useful predictions; (3) Classical methods can break down dramatically in high-dimensional regimes.

1.2 What can help us in high dimensions

- Sparsity in vectors: We know that each mean vector μ_j is relatively small, with only s of its d entries being non-zero.
- Structure in covariance matrices: If the covariance matrix Σ is assumed to be sparse but the positions were unknown, then a reasonable estimator would be the soft-thresholded version $\widetilde{\Sigma} := T_{\lambda_n}(\hat{\Sigma})$ of the sample covariance, where $\lambda_n = \sqrt{\frac{2\log d}{n}}$ and $T_\lambda(x) = I[|x| > \lambda](x - \lambda\mathrm{sgn}(x))$.
- Structured forms of regression: (1) sparse additive model: $f(x_1, \cdots, x_d) = \sum_{j\in S} g_j(x_j)$ where $S \subset \{1, 2, \cdots, d\}$ of cardinality $s = |S|$; (2) multiple-index model: $f(x_1, \cdots, x_d) = h(Ax)$ for some matrix $A \in \mathbb{R}^{s\times d}$; (3) projection pursuit regression: $f(x_1, \cdots, x_d) = \sum_{j=1}^M g_j(\langle a_j, x\rangle)$.

2 Basic tail and concentration bounds

2.1 Classical bounds

- Markov's inequality: Given a non-negative random variable X with finite mean, $\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}[X]}{t}$ for all $t > 0$.
- Chebyshev's inequality: For a random variable X with finite variance, $\mathbb{P}[|X - \mu| \geq t] \leq \frac{\mathrm{var}(X)}{t^2}$.
- $\mathbb{P}[|X - \mu| \geq t] \leq \frac{\mathbb{E}[|X-\mu|^k]}{t^k}$ for all $t > 0$. Chernoff bound: $\mathbb{P}[(X - \mu) \geq t] = \mathbb{P}[e^{\lambda(X-\mu)} \geq e^{\lambda t}] \leq \frac{\mathbb{E}[e^{\lambda(X-\mu)}]}{e^{\lambda t}} \Rightarrow \log\mathbb{P}[(X - \mu) \geq t] \leq \inf_{\lambda\in[0,b]}\{\log\mathbb{E}[e^{\lambda(X-\mu)}] - \lambda t\}$.
- Gaussian tail bounds: $X \sim \mathcal{N}(\mu, \sigma^2), \mathbb{E}[e^{\lambda X}] = e^{\mu\lambda + \frac{\sigma^2\lambda^2}{2}}$ for all $\lambda \in \mathbb{R}$. $\inf_{\lambda\geq 0}\{\log\mathbb{E}[e^{\lambda(X-\mu)}] - \lambda t\} = \inf_{\lambda\geq 0}\{\frac{\lambda^2\sigma^2}{2} - \lambda t\} = -\frac{t^2}{2\sigma^2} \Rightarrow \mathbb{P}[X \geq \mu + t] \leq e^{-\frac{t^2}{2\sigma^2}}$ for all $t \geq 0$.
- A random variable X with mean $\mu = \mathbb{E}[X]$ is sub-Gaussian if there is a positive number σ such that $\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{\sigma^2\lambda^2}{2}}$ for all $\lambda \in \mathbb{R}$.

- Any sub-Gaussian variable satisfies the concentration inequality: $\mathbb{P}[|X-\mu| \geq t] \leq 2e^{-\frac{t^2}{2\sigma^2}}$ for all $t \in \mathbb{R}$.

- Hoeffding bound: Suppose that the variables $X_i, i = 1, 2, \cdots, n$, are independent, and X_i has mean μ_i and sub-Gaussian parameter σ_i. Then for all $t \geq 0$, we have $P[\sum_{i=1}^n (X_i - \mu_i) \geq t] \leq \exp\{-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\}$. In particular, if $X_i \in [a, b]$ for all $i = 1, 2, \cdots, n$, then it is sub-Gaussian with parameter $\sigma = \frac{b-a}{2}$, so that $\mathbb{P}[\sum_{i=1}^n (X_i - \mu_i) \geq t] \leq e^{-\frac{2t^2}{n(b-a)^2}}$.

- Equivalent characterizations of sub-Gaussian variables (suppose $\mathbb{E}X = 0$): (1) $\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\lambda^2\sigma^2}{2}}$ for all $\lambda \in \mathbb{R}$ and some $\sigma \geq 0$; (2) $\mathbb{P}[|X| \geq s] \leq c\mathbb{P}[|Z| \geq s]$ for all $s \geq 0$ and some $c > 0$; (3) $\mathbb{E}[X^{2k}] \leq \frac{(2k)!}{2^k k!}\theta^{2k}$ for all $k = 1, 2, \cdots$ and some $\theta \geq 0$; (4) $\mathbb{E}[e^{\frac{\lambda X^2}{2\sigma^2}}] \leq \frac{1}{\sqrt{1-\lambda}}$ for all $\lambda \in [0, 1)$ and some $\sigma \geq 0$.

- A random variable X with mean $\mu = \mathbb{E}X$ is sub-exponential if there are non-negative parameters (ν, α) such that $\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{\nu^2\lambda^2}{2}}$ for all $|\lambda| < \frac{1}{\alpha}$.

- Sub-exponential tail bound: Suppose that X is sub-exponential with parameters (ν, α). Then $\mathbb{P}[X - \mu \geq t] \leq \begin{cases} e^{-\frac{t^2}{2\nu^2}} \text{ if } 0 \leq t \leq \frac{\nu^2}{\alpha} \\ e^{-\frac{t}{2\alpha}} \text{ for } t > \frac{\nu^2}{\alpha} \end{cases}$.

- Given a random variable X with mean μ and variance σ^2, Bernstein's condition with parameter b holds if $|\mathbb{E}[(X-\mu)^k]| \leq \frac{1}{2}k!\sigma^2 b^{k-2}$ for $k = 2, 3, 4, \cdots$. Bernstein's condition $\Rightarrow$ sub-exponential with parameters determined by σ^2 and b.

- Bernstein-type bound: For any random variable satisfying Bernstein condition, $\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{\lambda^2\sigma^2/2}{1-b|\lambda|}}$ for all $|\lambda| < \frac{1}{b}$. By setting $\lambda = \frac{t}{bt+\sigma^2} \in [0, \frac{1}{b})$, we derive the concentration inequality: $\mathbb{P}[|X-\mu| \geq t] \leq 2e^{-\frac{t^2}{2(\sigma^2+bt)}}$ for all $t \geq 0$.

- χ^2-variables: $X \sim \chi_n^2$ is sub-exponential with parameters $(\nu, \alpha) = (2\sqrt{n}, 4)$, thus $\mathbb{P}[|\frac{1}{n}X - 1| \geq t] \leq 2e^{-nt^2/8}$, for all $t \in (0, 1)$.

- Equivalent characterizations of sub-exponential variables (suppose $\mathbb{E}X = 0$): (1) There are non-negative numbers (ν, α) such that $\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\nu^2\lambda^2}{2}}$ for all $|\lambda| < \frac{1}{\alpha}$; (2) There is a positive number $c_0 > 0$ such that $\mathbb{E}[e^{\lambda X}] < \infty$ for all $|\lambda| \leq c_0$; (3) There are constants $c_1, c_2 > 0$ such that $\mathbb{P}[|X| \geq t] \leq c_1 e^{-c_2 t}$ for all $t > 0$; (4) The quantity $\gamma := \sup_{k \geq 2}[\frac{\mathbb{E}[X^k]}{k!}]^{1/k}$ is finite.

- One-sided Bernstein's inequality: If $X \leq b$ almost surely, then $\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \leq \exp\left(\frac{\frac{\lambda^2}{2}\mathbb{E}[X^2]}{1-\frac{b\lambda}{3}}\right)$ for all $\lambda \in [0, 3/b)$. Consequently, given n independent random variables such that $X_i \leq b$ almost surely, we have $\mathbb{P}[\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \geq n\delta] \leq \exp\left(-\frac{n\delta^2}{2(\frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i^2] + \frac{b\delta}{3})}\right)$.

2.2 Martingale-based methods

- Give a sequence $\{Y_k\}_{k=1}^\infty$ of random variables adapted to a filtration $\{\mathscr{F}_k\}_{k=1}^\infty$ ($\{\mathscr{F}_k\}_{k=1}^\infty$ is a sequence of σ-fields and $\mathscr{F}_k \subset \mathscr{F}_{k+1}$, Y_k is measurable w.r.t. the σ-field $\mathscr{F}_k$), the pair $\{(Y_k, \mathscr{F}_k)\}_{k=1}^\infty$ is a martingale if, for all $k \geq 1$, $\mathbb{E}[|Y_k|] < \infty$ and $\mathbb{E}[Y_{k+1}|\mathscr{F}_k] = Y_k$.

- Let $\{(D_k, \mathscr{F}_k)\}_{k=1}^{\infty}$ be a martingale difference sequence ($D_k := Y_k - Y_{k-1}$), and suppose that $E[e^{\lambda D_k}|\mathscr{F}_{k-1}] \leq e^{\lambda^2\nu_k^2/2}$ a.s. for any $|\lambda| < 1/\alpha_k$. Then (1) The sum $\sum_{k=1}^n D_k$ is sub-exponential with parameters $(\sqrt{\sum_{k=1}^n \nu_k^2}, \alpha*)$ where $\alpha* := \max_{k=1,\cdots,n} \alpha_k$; (2) Theu sum satisfies the concentration inequality $\mathbb{P}[|\sum_{k=1}^n D_k| \geq t] \leq \begin{cases} 2e^{-\frac{t^2}{2\sum_{k=1}^n \nu_k^2}} \text{ if } 0 \leq t \leq \frac{\sum_{k=1}^n \nu_k^2}{\alpha*} \\ 2e^{-\frac{t}{2\alpha*}} \text{ if } t > \frac{\sum_{k=1}^n \nu_k^2}{\alpha*} \end{cases}$.
- Azuma-Hoeffding: Let $\{(D_k, \mathscr{F}_k)\}_{k=1}^{\infty}$ be a martingale difference sequence for which there are constants $\{(a_k, b_k)\}_{k=1}^n$ such that $D_k \in [a_k, b_k]$ a.s. for all $k = 1, \cdots, n$. Then for all $t \geq 0$, $\mathbb{P}[|\sum_{k=1}^n D_k| \geq t] \leq 2e^{-\frac{2t^2}{\sum_{k=1}^n (b_k - a_k)^2}}$.
- Bounded differences inequality: Suppose that f satisfies the bounded difference property with parameters $(L_1, \cdots, L_n)$ ($|f(x) - f(x^{\backslash k})| \leq L_k$) and the random vector $X = (X_1, X_2, \cdots, X_n)$ has independent components. Then $\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2e^{-\frac{2t^2}{\sum_{k=1}^n L_k^2}}$ for all $t \geq 0$.

2.3 Lipschitz functions of Gaussian variables

- Let $(X_1, \cdots, X_n)$ be a vector of i.i.d. standard Gaussian variables, and let $f : \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz with respect to the Euclidean norm. Then the variable $f(X) - \mathbb{E}[f(X)]$ is sub-Gaussian with parameter at most L, and hence $\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2e^{-\frac{t^2}{2L^2}}$ for all $t \geq 0$.
- Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable. Then for any convex function $\phi : \mathbb{R} \to \mathbb{R}$, we have $\mathbb{E}[\phi(f(X) - \mathbb{E}[f(X)])] \leq \mathbb{E}[\phi(\frac{\pi}{2}\langle \nabla f(X), Y\rangle)]$ where $X, Y \sim \mathcal{N}(0, I_n)$ are standard multivariate Gaussian, and independent.
- Order Statistics: Given two random vectors $(X_1, \cdots, X_n)$ and $(Y_1, \cdots, Y_n)$. By reordering its entries in a non-decreasing manner, we have $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$. It can be shown that $|X_{(k)} - Y_{(k)}| \leq ||X - Y||_2$ for all $k = 1, \cdots, n$, so that each order Statistics is a 1-Lipschitz function. Thus $\mathbb{P}[|X_{(k)} - \mathbb{E}[X_{(k)}]| \geq \delta] \leq 2e^{-\frac{\delta^2}{2}}$.
- Gaussian complexity: Let $\{W_k\}_{k=1}^n$ be an i.i.d. sequence of $\mathcal{N}(0,1)$ variables. Given a collection of vectors $A \subset \mathbb{R}^n$, define the random variable $Z : \sup_{a\in A}[\sum_{k=1}^n a_k W_k] = \sup_{a\in A}\langle a, W\rangle$. $\mathbb{P}[|Z - \mathbb{E}[Z]| \geq \delta] \leq 2\exp\left(-\frac{\delta^2}{2D^2(A)}\right)$ where $D(A) = \sup_{a\in A} ||a||_2$.
- Gaussian chaos variables: Let $Q \in \mathbb{R}^{n\times n}$ be a symmetric matrix, and let $w, \widetilde{w}$ be independent zero-mean Gaussian random vectors with covariance matrix I_n. The random variable $Z : w^T Q \widetilde{w}$ is known as a decoupled Gaussian chaos. $P[|Z| \geq \delta] \leq 4\exp\left(-\frac{\delta^2}{4||Q||_F^2 + 4\delta||Q||_2}\right)$.

2.4 Exercises

- Exercise 2.1: (a) $P(X = 1) = 1, P(X \neq 1) = 0$; (b) $P(X = 1) = P(X = -1) = \frac{1}{2}, P(X \neq \pm 1) = 0$.
- Exercise 2.2: (b) Let $f(z) = \int_z^\infty \phi(z)dz - \phi(z)\left(\frac{1}{z} - \frac{1}{z^3}\right), g(z) = \phi(z)\left(\frac{1}{z} - \frac{1}{z^3} + \frac{3}{z^5}\right) - \int_z^\infty \phi(z)dz$. $f'(z) = -\frac{3}{z^4}\phi(z) < 0 \Rightarrow f(z) > f(\infty), g'(z) = -\frac{3}{5z^6}\phi(z) < 0 \Rightarrow g(z) > g(\infty) = 0$.

- Exercise 2.3: $\frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda\delta}} = \frac{\sum_{k=0}^{\infty}\frac{\lambda^k\mathbb{E}[X^k]}{k!}}{\sum_{k=0}^{\infty}\frac{\lambda^k\delta^k}{k!}} \geq \inf_k \frac{\frac{\lambda^k\mathbb{E}[X^k]}{k!}}{\frac{\lambda^k\delta^k}{k!}} = \inf_k \frac{\mathbb{E}[X^k]}{\delta^k}$.

- Exercise 2.4: (a) $\psi'(0) = \frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]}|_{\lambda=0} = \mathbb{E}[X] = \mu$; (b) $\psi''(\lambda) = \mathbb{E}_\lambda[(X - \mathbb{E}_\lambda[X])^2] = \mathbb{E}_\lambda[(X - \frac{b-a}{2} + \frac{b-a}{2} - \mathbb{E}_\lambda[X])^2] = \mathbb{E}_\lambda[(X - \frac{b-a}{2})^2] - (\frac{b-a}{2} - \mathbb{E}_\lambda[X])^2 \leq \frac{(b-a)^2}{4}$; (c) $\psi(\lambda) = \log\mathbb{E}[e^{\lambda X}] = \psi(0) + \psi'(0)\lambda + \frac{\psi''(\xi)}{2}\lambda^2 \leq \lambda\mu + \frac{(b-a)^2}{8}\lambda^2 \Rightarrow \log\mathbb{E}[e^{\lambda(X-\mu)}] \leq \frac{\sigma^2\lambda^2}{2}$ where $\sigma = \frac{b-a}{2}$.

- Exercise 2.5: (a) Let $f(\lambda) = e^{\frac{\lambda^2\sigma^2}{2}+\lambda\mu} - \mathbb{E}[e^{\lambda X}]$. Then $f(\lambda) \geq 0$ for all $\lambda \in \mathbb{R}$ and $f(0) = 0$, which means $f'(0) = 0 \Rightarrow \mathbb{E}[X] = \mu$; (b) $f(\lambda) \geq 0, f(0) = 0$ also means $f''(0) \geq 0 \Rightarrow \mathbb{E}[X^2] \leq \mu^2 + \sigma^2 \Rightarrow \mathrm{var}(X) \leq \sigma^2$; (c) $P(X = 2) = \frac{1}{3}, P(X = -1) = \frac{2}{3}$. $\mathrm{var}(X) = 2$, but $\sigma^2_{\min} \approx 2.164$.

- Exercise 2.6: Use one-sided Bernstein's inequality, $\mathbb{P}[Z_n \leq \mathbb{E}[Z_n] - \sigma^2\delta] \leq \exp\left(-\frac{n\sigma^4\delta^2}{2\mathbb{E}[X^4]}\right)$. $E[X^4] = \int_0^\infty 4x^3\mathbb{P}[|X| > x]dx \leq \int_0^\infty 8x^3e^{-\frac{x^2}{2\sigma^2}}dx = 16\sigma^4$, which yields $P[Z_n \leq \mathbb{E}[Z_n] - \sigma^2\delta] \leq \exp\left(-\frac{n\delta^2}{32}\right)$.

- Exercise 2.7: (a) $\log\mathbb{E}[e^{\lambda X_i}] = \log\mathbb{E}[\sum_{k=0}^\infty \frac{(\lambda X_i)^k}{k!}] \leq b^{k-2}\mathbb{E}[\sum_{k=2}^\infty \frac{\lambda^k X_i^2}{k!}] = \sigma^2\lambda^2\left\{\frac{e^{\lambda b}-1-\lambda b}{(\lambda b)^2}\right\}$; (b) $\mathbb{P}[\sum_{i=1}^n X_i \geq n\delta] = \mathbb{P}[e^{\lambda\sum_{i=1}^n X_i} \geq e^{\lambda n\delta}] \leq \exp\left\{n\sigma^2\lambda^2\left\{\frac{e^{\lambda b}-1-\lambda b}{(\lambda b)^2}\right\} - n\lambda\delta\right\}$. Let $\lambda = \frac{\log(1+\frac{b\delta}{\sigma^2})}{b}$, then RHS $= \exp\{-\frac{n\delta^2}{b^2}h(\frac{b\delta}{\sigma^2})\}$; (c) WLOG let $b = \sigma = 1$, one can show that $h(t) - \frac{t^2}{2(1+t/3)} \geq 0$.

- Exercise 2.8: (a) $P[Z \geq t] \leq 1 \wedge Ce^{-\frac{t^2}{2(\nu^2+bt)}} \leq 1 \wedge Ce^{-\frac{t^2}{4(\nu^2\vee bt)}} = (1 \wedge Ce^{-\frac{t^2}{4\nu^2}}) \vee (1 \wedge Ce^{-\frac{t^2}{4bt}}) \leq 1 \wedge Ce^{-\frac{t^2}{4\nu^2}} + 1 \wedge Ce^{-\frac{t^2}{4bt}}$. 从而 $\mathbb{E}[Z] = \int_0^\infty P(Z \geq t)dt \leq \int_0^\infty 1 \wedge Ce^{-\frac{t^2}{4\nu^2}}dt + \int_0^\infty 1 \wedge Ce^{-\frac{t^2}{4bt}}dt \leq 2\nu(\sqrt{\pi} + \sqrt{\log C}) + 4b(1 + \log C)$. (b) Note that $P[|\frac{1}{n}\sum_{i=1}^n X_i| \geq t] \leq 2e^{-\frac{t^2}{2[(\sigma/\sqrt{n})^2+(b/n)t]}}$. Let $C = 2, \nu = \frac{\sigma}{\sqrt{n}}$ and $b = \frac{b}{n}$.

3 Concentration of measure

3.1 Concentration by entropic techniques

- Given a convex function $\phi : \mathbb{R} \to \mathbb{R}$, define ϕ-entropy as $\mathbb{H}_\phi(X) = \mathbb{E}[\phi(X)] - \phi(\mathbb{E}[X])$. By Jensen's inequality and the convexity of ϕ, it is always non-negative and serves as a measure of variability.

- Throughout the remainder of this chapter, we focus on a slightly different choice of entropy functional: $\phi(u) := u\log u$ for $u > 0$ and $\phi(0) := 0$. For any non-negative random variable $Z := e^{\lambda X} \geq 0$, it defines the ϕ-entropy given by $\mathbb{H}[Z] = \mathbb{E}[Z\log Z] - E[Z]\log\mathbb{E}[Z] \Rightarrow \mathbb{H}[e^{\lambda X}] = \lambda\varphi_x'(\lambda) - \varphi_x(\lambda)\log\varphi_x(\lambda)$ where $\varphi_x(\lambda) = \mathbb{E}[e^{\lambda X}]$.

- Herbst argument: Suppose that the entropy $\mathbb{H}(e^{\lambda X})$ satisfies inequality $\mathbb{H}(e^{\lambda X}) \leq \frac{1}{2}\sigma^2\lambda^2\varphi_x(\lambda)$ for all $\lambda \in I$, where I can be either of the intervals $[0, \infty)$ or $\mathbb{R}$. Then X satisfies the bound $\log\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \leq \frac{1}{2}\lambda^2\sigma^2$ for all $\lambda \in I$.

- Bernstein entropy bound: Suppose that there are positive constants b and σ such that the entropy $\mathbb{H}(e^{\lambda X})$ satisfies the bound $\mathbb{H}(e^{\lambda X}) \leq \lambda^2\{b\phi_x'(\lambda) + \phi_x(\lambda)(\sigma^2 - b\mathbb{E}[X])\}$ for all $\lambda \in [0, 1/b)$. Then X satisfies the bound $\log\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \leq \sigma^2\lambda^2(1 - b\lambda)^{-1}$ for all $\lambda \in [0, 1/b)$.

- Let $\{X_i\}_{i=1}^n$ be independent random variables, each supported on the interval $[a,b]$ and let $f:\mathbb{R}^n \to \mathbb{R}$ be separately convex (for each index $k \in \{1,2,\cdots,n\}$, the univariate function $y_k \mapsto f(x_1,\cdots,x_{k-1},y_k,x_{k+1},\cdots,x_n)$ is convex for each fixed vector $(x_1,\cdots,x_{k-1},x_{k+1},\cdots,x_n) \in \mathbb{R}^{n-1}$), and L-Lipschitz with respect to the Euclidean norm. Then, for all $\delta > 0$, we have $\mathbb{P}[f(X) \geq \mathbb{E}[f(X)] + \delta] \leq \exp\left(-\frac{\delta^2}{4L^2(b-a)^2}\right)$.

- Entropy bound for univariate functions: Let $X, Y \sim \mathbb{P}$ be a pair of i.i.d. variables. Then for any function $g:\mathbb{R}\to\mathbb{R}$, we have $\mathbb{H}(e^{\lambda g(X)}) \leq \lambda^2 \mathbb{E}[(g(X)-g(Y))^2 e^{\lambda g(X)} I(g(x) \geq g(Y))]$ for all $\lambda > 0$. If in addition X is supported on $[a,b]$, and g is convex and Lipschitz, then $\mathbb{H}(e^{\lambda g(X)}) \leq \lambda^2 (b-a)^2 \mathbb{E}[(g'(X))^2 e^{\lambda g(X)}]$.

- Tensorization of entropy: Let $f:\mathbb{R}^n \to \mathbb{R}$, and let $\{X_k\}_{k=1}^n$ be independent random variables. Then $\mathbb{H}(e^{\lambda f(X_1,\cdots,X_n)}) \leq \mathbb{E}[\sum_{k=1}^n \mathbb{H}(e^{\lambda f_k(X_k)} | X^{\backslash k})]$.

3.2 A geometric perspective on concentration

- Given a set $A \subset \mathcal{X}$ and a point $x \in \mathcal{X}$, define the quantity $\rho(x,A) = \inf_{y\in A} \rho(x,y)$. Given parameter $\epsilon > 0$, the ϵ-enlargement of A is given by $A^\epsilon = \{x \in \mathcal{X} | \rho(x,A) < \epsilon\}$.

- The concentration function $\alpha : [0,\infty) \to \mathbb{R}_+$ associated with metric measure space $(\mathbb{P}, \mathcal{X}, \rho)$ is given by $\alpha_{\mathbb{P},(\mathcal{X},\rho)}(\epsilon) := \sup_{A\subset\mathcal{X}} \{1 - \mathbb{P}[A^\epsilon] \,|\, \mathbb{P}[A] \geq \frac{1}{2}\}$.

- Given a random variable $X \sim \mathbb{P}$ and concentration function $\alpha_{\mathbb{P}}$, any 1-Lipschitz function on $(\mathcal{X},\rho)$ satisfies $\mathbb{P}[|f(X) - m_f| \geq \epsilon] \leq 2\alpha_{\mathbb{P}}(\epsilon)$, where m_f is any median of f ($P[f(X) \geq m_f] \geq \frac{1}{2}$ and $\mathbb{P}[f(X) \leq m_f] \leq \frac{1}{2}$). Conversely, suppose that there is a function $\beta : \mathbb{R}_+ \to \mathbb{R}_+$ such that for any 1-Lipschitz function on $(\mathcal{X},\rho)$, $P[f(X) \geq \mathbb{E}[f(X)] + \epsilon] \leq \beta(\epsilon)$ for all $\epsilon \geq 0$. Then the concentration function satisfies the bound $\alpha_{\mathbb{P}}(\epsilon) \leq \beta(\epsilon/2)$.

- Let $\mathbb{P}$ be any strongly log-concave distribution with parameter $\gamma > 0$. Then for any function $f:\mathbb{R}^n \to \mathbb{R}$ that is L-Lipschitz with respect to Euclidean norm, we have $\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2\exp^{-\frac{\gamma t^2}{4L^2}}$.

3.3 Wasserstein distances and information inequalities

- Wasserstein metric: $W_\rho(\mathbb{Q},\mathbb{P}) = \sup_{||f||_{\text{Lip}}\leq 1} \left[\int f d\mathbb{Q} - \int f d\mathbb{P}\right]$.

- Kullback-Leibler divergence: $D(\mathbb{Q},\mathbb{P}) = \begin{cases} \mathbb{E}_{\mathbb{Q}}[\log \frac{d\mathbb{Q}}{d\mathbb{P}}] \text{ if } \mathbb{Q} \text{ is absolutely continuous w.r.t. } \mathbb{P} \\ +\infty \text{ otherwise} \end{cases}$.

- Total variation distance: $||\mathbb{Q} - \mathbb{P}||_{\text{TV}} = \sup_{A\subset\mathcal{X}} |\mathbb{Q}(A) - \mathbb{P}(A)|$.

- A distribution $\mathbb{M}$ on the product space $\mathcal{X} \bigotimes \mathcal{X}$ is a coupling of the pair $(\mathbb{Q},\mathbb{P})$ if its arginal distributions in the first and second coordinates coincide with $\mathbb{Q}$ and $\mathbb{P}$, respectively.

- Let $f:\mathcal{X}\to\mathbb{R}$ be any 1-Lipschitz function and $\mathbb{M}$ be any coupling. Then $\int \rho(x,x') d\mathbb{M}(x,x') \geq \int (f(x) - f(x')) d\mathbb{M}(x,x') = \int f(d\mathbb{P} - d\mathbb{Q})$. The Kantorovich-Rubinstein duality guarantees the

following important fact: if we minimize over all possible couplings, then this argument can be reversed, and we have the equivalence $\sup_{||f||_{\text{Lip}} \leq 1} \int f(d\mathbb{Q} - d\mathbb{P}) = \inf_{\mathbb{M}} \int_{\mathcal{X} \times \mathcal{X}} \rho(x, x') d\mathbb{M}(x, x') = \inf_{\mathbb{M}} \mathbb{E}_{\mathbb{M}}[\rho(X, X')]$.

- For a given metric ρ, the probability measure $\mathbb{P}$ is said to satisfy a ρ-transformation cost inequality with parameter $\gamma > 0$ if $W_p(\mathbb{Q}, \mathbb{P}) \leq \sqrt{2\gamma D(\mathbb{Q}||\mathbb{P})}$ for all probability measures $\mathbb{Q}$.

- Consider a metric measure space $(\mathbb{P}, \mathcal{X}, \rho)$, and suppose that $\mathbb{P}$ satisfies the ρ-transportation cost inequality. Then its concentration function satisfies the bound $\alpha_{\mathbb{P},(\mathcal{X},\rho)} \leq 2\exp(-\frac{t^2}{2\gamma})$. Moreover, for any $X \sim \mathbb{P}$ and any L-Lipschitz function $f : \mathcal{X} \to \mathbb{R}$, we have the concentration inequality $\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2\exp(-\frac{t^2}{2\gamma L^2})$.

- Suppose that, for each $k = 1, 2, \cdots, n$, the univariate distribution $\mathbb{P}_k$ satisfies a ρ_k-transportation cost inequality with parameter γ_k. Then the product distribution $\mathbb{P} = \bigotimes_{k=1}^n \mathbb{P}_k$ satisfies the transportation cost inequality $W_\rho(\mathbb{Q}, \mathbb{P}) \leq \sqrt{2(\sum_{k=1}^n \gamma_k) D(\mathbb{Q}||\mathbb{P})}$ for all distributions $\mathbb{Q}$ where the Wasserstein metric is defined using the distance $\rho(x, y) = \sum_{k=1}^n \rho_k(x_k, y_k)$.

- Let $\mathbb{P}$ be the distribution of a β-contractive Markov chain (there exists some $\beta \in [0, 1)$ such that $\max_{i=1,\cdots,n-1} \sup_{x_i, x_i'} ||\mathbb{K}_{i+1}(\cdot|x_i) - \mathbb{K}_{i+1}(\cdot|x_i')||_{\text{TV}} \leq \beta$ where $\mathbb{K}_{i+1}(x_{i+1}|x_i) = \mathbb{P}(X_{i+1} = x_{i+1}|X_i = x_i)$) over the discrete space $\mathcal{X}^n$. Then for any other distribution $\mathbb{Q}$ over $\mathcal{X}^n$, we have $W_\rho(\mathbb{Q}, \mathbb{P}) \leq \frac{1}{1-\beta}\sqrt{\frac{n}{2} D(\mathbb{Q}||\mathbb{P})}$, where the Wasserstein distance is defined with respect to the Hamming norm $\rho(x, y) = \sum_{i=1}^n I[x_i \neq y_i]$.

- Consider a vector of independent random variables $(X_1, \cdots, X_n)$, each taking values in $[0, 1]$, and let $f : \mathbb{R}^n \to \mathbb{R}$ be convex, and L-Lipschitz with respect to the Euclidean norm. Then for all $t \geq 0$, we have $\mathbb{P}[|f(X) - \mathbb{E}[f(x)]| \geq t] \leq 2e^{-\frac{t^2}{2L^2}}$.

3.4 Tail bounds for empirical processes

- Let $\mathscr{F}$ be a class of functions (each of the form $f : \mathcal{X} \to \mathbb{R}$), and let $(X_1, \cdots, X_n)$ be drawn from a product distribution $\mathbb{P} = \bigotimes_{i=1}^n \mathbb{P}_i$, where each $\mathbb{P}_i$ is supported on some set $\mathcal{X}_i \subset \mathcal{X}$. We then consider the random variable $Z = \sup_{f \in \mathscr{F}} \left\{\frac{1}{n} \sum_{i=1}^n f(X_i)\right\}$. The primary goal of this section is to derive a number of upper bounds on the tail event $\{Z \geq \mathbb{E}[Z] + \delta\}$.

- Functional Hoeffding theorem: For each $f \in \mathscr{F}$ and $i = 1, \cdots, n$, assume that there are real numbers $a_{i,f} \leq b_{i,f}$ such that $f(x) \in [a_{i,f}, b_{i,f}]$ for all $x \in \mathcal{X}_i$. Then for all $\delta \geq 0$, we have $\mathbb{P}[Z \geq \mathbb{E}[Z] + \delta] \leq \exp\left(-\frac{n\delta^2}{4L^2}\right)$, where $L^2 := \sup_{f \in \mathcal{F}} \{\frac{1}{n} \sum_{i=1}^n (b_{i,f} - a_{i,f})^2\}$.

- Talagrand concentration for empirical processes: Consider a countable class of function $\mathscr{F}$ uniformly bounded by b. Then for all $\delta > 0$, the random variable Z satisfies the upper tail bound $\mathbb{P}[Z \geq \mathbb{E}[Z] + \delta] \leq 2\exp\left(\frac{-n\delta^2}{8e\mathbb{E}[\Sigma^2] + 4b\delta}\right)$ where $\Sigma^2 = \sup_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^n f^2(X_i)$.

- Controlling the random variance: Let $Z = \sup_{f \in \mathscr{F}} \sum_{i=1}^n f(X_i)$, $\widetilde{Z} = Z - \mathbb{E}[Z]$ and $\Gamma(X) = \sup_{f \in \mathscr{F}} \sum_{i=1}^n f^2(X_i)$. For all $\lambda > 0$, we have $\mathbb{E}[\Gamma e^{\lambda \widetilde{Z}}] \leq (e - 1)\mathbb{E}[\Gamma]\mathbb{E}[e^{\lambda \widetilde{Z}}] + \mathbb{E}[\widetilde{Z} e^{\lambda \widetilde{Z}}]$.

4 Uniform laws of large numbers

4.1 Motivation

- Glivenko-Cantelli: For any distribution, the empirical CDF $\hat{F}_n$ is a strongly consistent estimator of the population CDF in the inform norm, meaning that $||\hat{F}_n - F||_\infty \overset{\text{a.s.}}{\to} 0$.