STA231C High-Dimensional Statistics Notes

Zhi Zhang

March 2020

Contents

Chapter 1

Measure of Concentration

1.1 Subgaussian Random Variable

Definition 1.1.1. We say a r.v. X with mean μ is subG with parameter $\sigma ge0$ if

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\frac{\lambda^2\sigma^2}{2}}$$

Johnson–Lindenstrauss lemma: random projection approximately preserve euclidean distance

1.2 Subexp random variable

Martingale methods tower property, strip step

$$f(x) - \mathbb{E}\left[f(x)\right] \sim \text{subexp}(||v||_2, ||\alpha||_\infty)$$

$$\mathbb{P}(|f(x) - \mathbb{E}[f(x)]| \geq t) \leq 2\exp\{-\min\left(\frac{t^2}{2||v||^2}, \frac{t}{2||\alpha_\infty||}\right)\}$$

Azuma-Hoeffding inequality, as long as the martingale difference bounded. We use it to prove bounded difference property.

The bounded difference inequality, McDiarmid's inequality, how to handle non-linear functions of independent random variables.

1.3 Gaussian random vectors

Gaussian concentration inequality

$$\mathbb{P}(|f(z) - \mathbb{E}[f(z)]| \geq t) \leq 2e^{\frac{-t^2}{2L^2}}$$

$$\mathbb{P}(|f(x) - \mathbb{E}[f(x)]| \geq t) \leq 2e^{\frac{-t^2}{2L^2||\sigma||_{op}}}$$

control of the largest singular value of the covariance matrix.

Fubini's theorem

Orthogonal linear map to the Gaussian r.v. preserve the Gaussian property

1.4 Quadratic forms in subG r.v.

$$\mathbb{P}(|X^\top AX - \mathbb{E}[X^\top AX]| \geq t) \leq small(t)$$

Theorem 1.4.1 (Hassan Wright Inequality)**.**

$$\mathbb{P}\left(|X^\top AX - \mathbb{E}[X^\top AX]| \geq t\right) \leq 2\exp\left(-c\min\left(\frac{t^2}{\tau^4\left\|A\right\|_p^2}, \frac{t}{\tau^2\left\|A\right\|_{op}}\right)\right)$$

$$\mathbb{E}\left[X^\top AX\right] = \operatorname{tr}(A)$$

1.5 Concentration Inequalities that are based only on moments

how can we develop concentration Inequalities for r.v that don't have mgf?

$$\mathbb{P}\left(|S - \mathbb{E}\left[S\right]| \geq t\right) \leq \frac{\mathbb{E}\left[|S - \mathbb{E}\left[S\right]|^r\right]}{r^t}$$

Rosenthal inequality allows higher order moments to be bounded in terms of variances.

Theorem 1.5.1 (Khinehines Theorem)**.**

$$\left\|\sum_{k=1}^n a_k\epsilon_k\right\|_r \leq c\sqrt{r}\left\|\sum_{k=1}^n a_k\epsilon_k\right\|_2$$

book "decoupling" by Víctor De la Peña and Evarist Gine Masdeu

Chapter 2

Uniform Laws of Large Numbers

2.1 basic

Theorem 2.1.1 (Glivenko–Cantelli Thm, fundamental theorem of statistics)**.** *let* $\hat{F}_n(t) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}\{x_k \le t\}$

$$\|h-g\|_\infty = \sup_{t\in R}|h(t)-g(t)|$$

$$\left\|\hat{F}_n - F\right\|_\infty \to 0$$

$$\left\|\hat{F}_n - F\right\| = \left\|\hat{P}_n - P\right\|_{\mathcal{F}}$$

2.2 Empirical Risk Minimization

$$R(\theta,\theta^*) = \int \log\left(\frac{p_{\theta^*}(x)}{p_\theta(x)}\right) p_{\theta^*}(x)dx = D_{KL}(p_{\theta^*}||p_\theta)$$

If $\hat{\theta}$ is obtained by ERM(eg. MLE), how can we show the excess risk $\epsilon(\theta,\hat{\theta^*})$ is small?

$$\epsilon(\hat{\theta}-\theta^*) \le 2\left\|\hat{P}_n - P\right\|_{\mathcal{F}}$$

$$\mathbb{E}\left[\left\|\hat{P}_n - P\right\|_{\mathcal{F}}\right] + \sigma$$

holds with $prob \ge 1 - \exp\left(-\frac{n\sigma^2}{2b^2}\right)$

Definition 2.2.1 (Rademacher Complexity)**.**

$$\mathrm{Rm}(\mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(x_i)\right]\right.$$

measure how large the class of the function

Lemma 2.2.1.

$$\mathbb{E}\left[\left\|\hat{P}_n - P\right\|_{\mathcal{F}}\right] \le 2\,\mathrm{Rm}\left(\mathcal{F}\right)$$

bounded difference property

$$|G(x_1,\cdots,x_n) - G(x_1,x_k',\cdots,x_n)| \le \frac{2b}{n}$$

Fact 2.2.1. moving sup inside the $\mathbb{E}$ makes the expectation bigger jessen inequality, moving out the expectation out of a convex function creates an upper bound

Fact 2.2.2. $\mathrm{Rm}\left(\mathcal{F}\right) \to 0$ is a sufficient condition for $\mathcal{F}$ to be a GC class

2.3 Glivenko–Cantelli Theorem

Lemma 2.3.1. $f(x) = \mathbb{1}\{x \leq t\}$, implies

$$\left\|\hat{F}_n - F\right\|_\infty \leq c\sqrt{\frac{\log(n+1)}{n}} + \delta$$

with prob $\geq 1 - \exp\left(\frac{-n\delta^2}{2}\right)$

Remark 2.3.1. Dvoretzky–Kiefer–Wolfowitz inequality, $P(\left\|\hat{F}_n - F\right\|_\infty \geq t) \leq 2e^{-2nt^2}$

Chapter 3

Metric Entropy

Fact 3.0.1. For any $T\mathbb{R}$, if $B(1)$ is the unit ball for a general norm $\|\|$, then $N(\delta, \rho, B(1)) \leq \left(\frac{1}{\delta} + 1\right)^d$ provided $\rho(\theta, \theta') = \|\theta - \theta'\|$, likewise, the metric entropy looks like $d\log(\frac{1}{\delta})$ as $\delta \to 0$

ideal, the class of function is as complex as the ball in the space.

Fact 3.0.2. Examples of metric entropy of sets of functions

$$\mathcal{F} = \{g : [0,1] \to [R] \big| |g(x) - gx' \leq L|x - x'||\}$$

then

$$\log N(\delta, \|\|_\infty, \mathcal{F}) \asymp \frac{L}{\delta}$$

where

$$\|g - g'\|_\infty = \sup_{x\in[0,1]} |g(x) - g'(x)|$$

Goal: bound things like $\mathbb{E}\left[\sup_{\theta\in T} X_\theta\right]$, T is the metric space with metric ρ

3.1 Dudley Entropy Integral

Goal: bound $\mathbb{E}\left[\sup_{\theta\in T} X_\theta\right]$ or $\mathbb{E}\left[\sup_{\theta,\theta'\in T}(X_\theta - X_{\theta'}\right]$

Definition 3.1.1. we say a process $\{X_\theta\}_{\theta\in T}$ is subG with a metric ρ, if $X_\theta - X_{\theta'} \sim subG(\rho(\theta, \theta'))$, equivalently, the

$$\mathbb{E}\left[e^{\lambda(X_\theta - X_{\theta'})}\right] \leq e^{\frac{\lambda^2\rho^2(\theta,\theta')}{2}}$$

Example 3.1.1. Let $z \sim N(0, I_d)$ as standard Gaussian vector, $T = unitl_2ballB_2(1)$, $\rho = \|\|_2$, and $X_\theta = \langle Z, \theta\rangle$, then $\{X_\theta\}$ is subG

$$X_\theta - X_{\theta'} = \langle Z, \theta - \theta'\rangle \sim N(0, \|\theta - \theta'\|_2^2)$$

$$\mathbb{E}\left[\sup_{\theta\in T} X_\theta\right] \leq \int_0^D \sqrt{\log N(\delta, \rho, T)d\delta}$$

where $D = dim(T)$

Remark 3.1.1.

$$\mathbb{E}\left[\sup_{\theta,\theta'} X_\theta - X_{\theta'}\right] \leq C$$

implies

$$\mathbb{E}\left[\sup_{\theta} X_\theta\right] \leq C$$

Fact 3.1.1.

$$\mathbb{E}\left[\sup_{\theta,\theta'} X_\theta - X_{\theta'}\right] = \left\|\sup_{\theta,\theta'} X_\theta - X_{\theta'}\right\|_{L_1}$$

weakly convergence for empirical process

Example 3.1.2. control of Gaussian weights

$$G(T) \leq \inf_{\delta\in[0,D]} |\sqrt{d}\delta + D\sqrt{\log N(\delta,\rho,T)}|$$

sum of squares of d Gaussian r.v. is d

Jensen's Inequality put inside "E" into a concave function (sqrt), or put outside E into a convex function (abs) gets an upper bound

Theorem 3.1.1 (Dudley entropy integral)**.** *Let $\{X_\theta\}_{\theta\in T}$ be a zero-mean subG process, with ρ, and $D = diam(T)$ then*

$$\mathbb{E}\left[\sup_{\theta\in T} X_\theta\right] \leq \int_0^D \sqrt{\log N(\delta,\rho,T)}d\delta$$

note: the larger of the δ, the smaller of points needed to pick from T to composite the δ cover.

Fact 3.1.2. If $Z_1, \cdots, Z_m$ are $subG(1)$ r.v.s, with all $\mathbb{E}[Z_i] = 0$ for $i = 1, \cdots, m$, then $\mathbb{E}[max_{1\leq i\leq m}]\, Z_i \lesssim \sqrt{\log(m)}$

Fact 3.1.3.

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum \epsilon_i f(x_i)|\right] &\leq \frac{1}{\sqrt{n}}\mathbb{E}\left[\sup_{f\in\mathcal{F}} X_f\right] \\
&\lesssim \frac{1}{\sqrt{n}}\int_0^D \sqrt{\log N(\delta,\rho,\mathcal{F})}d\delta \\
\text{Duadley gives } &\lesssim \frac{1}{\sqrt{n}}\int_0^D \sqrt{c(b,\mu)\log N(1+2\log(\frac{b}{\delta}))}d\delta \\
&\lesssim \frac{c'(b,\mu)}{\sqrt{n}}
\end{aligned}$$

3.2 Sudakov-Fernique-Sleplan's Inequality

Fact 3.2.1.

$$\phi(x) = \max_j x_j$$

the smoothed version of maximum

$$\phi_\beta(x) = \frac{1}{\beta}\log\left(\sum_{j=1}^d e^{\beta x_j}\right)$$

$$\phi(x) \leq \phi_\beta(x) \leq \phi(x) + \frac{\log(d)}{\beta}$$

Example 3.2.1 (Application: Gaussian Contraction Inequality)**.**

Fact 3.2.2. For independent random variables X and Y, the variance of their sum or difference is the sum of their variances: $\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$ $\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$

3.3 Gordon's inequality

Gordon's inequality allows us to control $\mathbb{E}[\sigma_{min}(A)]$, provided we ca ncompare with another Gaussian matrix B, think $Y_{u,v} = u^T Bv$.

Fact 3.3.1.

$$\sigma_{min}(A) = \inf_{u\in U}\sup_{v\in V} u^T Av$$

Chapter 4

Random Matrices and Covariance Estimation

4.1 Notation

$\Sigma \succeq 0$
For any sym matrix A

$$\lambda_{min}(A)I_d \preceq A \preceq \lambda_{max}I_d$$

$$\sigma_j(M) = \sqrt{\lambda_j(M^T M)}$$

Coward Fischer Formula

$$\lambda_{max}(A) = \sup_{\|u\|_2=1} u^T A u$$

$$\lambda_{inf}(A) = \inf_{\|u\|_2=1} u^T A u$$

$$\sigma_{max}(M) = \sup_{\|v\|_2=1} \|Mv\|_2 = \sup_{\|u\|_2=1} \sup_{\|v\|_2=1} u^T M v$$

$$\sigma_{min}(M) = \inf_{\|v\|_2=1} \|Mv\|_2 = \sup_{\|u\|_2=1} \inf_{\|v\|_2=1} u^T M v$$

$$\|x\|_2 = \sup_{\|u\|_2=1} u^T x$$

$$\|M\|_{op} = \sigma_{max}(M)$$

$$\|A\|_{op} = \max_j |\lambda_j(A)| = \sup_{\|u\|_2=1} |u^T A u|$$

Theorem 4.1.1 (Weyl's thm)**.**

$$|\lambda_j(A) - \lambda_j(B)| \leq \|A - B\|_{op}$$

$$|\sigma_j(M) - \sigma_j(M')| \leq \|M - M'\|_{op}$$

Fact 4.1.1. swap ping trick, suppose $M \in \mathbb{R}^{n\times d}$ $M' \in \mathbb{R}^{n\times d}$, then MM' has the same non-zero eigenvalues as $M'M$

4.2 covariance estimation

Population

$$\Sigma_{ij} = \operatorname{cov}(x_{1i}, x_{1j})$$

$$\mu = \mathbb{E}[x_1]$$

$$\Sigma = \mathbb{E}\left[(x-\mu)(x-\mu)^T\right]$$

Sample:

$$\hat{\Sigma} = 1/(n-1)\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$$

$$\bar{x} = 1/n \sum_{i=1} n x_i$$

Assume $\mu = 0$

$$\hat{\Sigma} = 1/(n)\sum_{i=1}^{n}(x_i)(x_i)^T$$

useful simplification b/c it allows to be written as a sum of ind rank-1 random matrices.

4.3 Extreme eigenvalues of Wishart Matrices

Theorem 4.3.1.

$$\mathbb{P}\left(\sqrt{\lambda_{max}(\hat{\Sigma})} \geq \sqrt{\lambda_{max}(\Sigma)(1+\delta)} + \sqrt{\operatorname{tr}(\Sigma)/n}\right) \leq e^{-n\delta^2/2}$$

Proof. part1: show $\sqrt{\lambda_{max}(\hat{\Sigma})}$ concentrates around $\mathbb{E}\left[\sqrt{\lambda_{max}(\hat{\Sigma})}\right]$ applying the G. concentration inequality to a certain lipschitz function

Claim 4.3.1.

$$f(z) = \Sigma_{max}(\frac{1}{\sqrt{n}}\Sigma^{1/2}Z^T)$$

is a lipschitz function

part2: upper bound ☐

Fact 4.3.1.

$$\hat{\Sigma} = 1/n\Sigma^{1/2}Z^T Z\Sigma^{1/2}$$

Corollary 4.3.1.

$$\left\|\hat{\Sigma} - \Sigma\right\|_{op} \leq (2\epsilon + \epsilon^2)\left\|\Sigma\right\|_{op}$$

holds w.p. at least $1 - 2e^{-n\delta^2/2}$

$$\epsilon = \sqrt{d/n} + \delta$$

Fact 4.3.2.

$$\mathbb{E}\left[\|g\|_2\right] = \mathbb{E}\left[\sqrt{\|g\|_2^2}\right] \leq \sqrt{\mathbb{E}\left[\|g\|_2^2\right]} = \sqrt{n}$$

Fact 4.3.3.

$$\|M\|_{op} \leq max l_1 norm of a row of M$$

Claim 4.3.2. If every row of $adj(\Sigma)$ has at most s non-zeros, then $\|adj(\Sigma)\|_{op} \leq s$

4.4 structure Σ

shareholding operator

$$Tr(M) = M_{i,j}, if|M_{ij} \geq \lambda| = 0, otherwise$$

the issue is that the $Tr(\hat{\Sigma})$ may not be P.S.D

Theorem 4.4.1. *Assume the previous setting for subG data still holds, for any $\delta > 0$, let $\lambda = \lambda_n = \frac{8\sqrt{\log(d)}}{\sqrt{n}} + \delta$, then the event*

$$\left\| T_\lambda(\hat{\Sigma}) - \Sigma \right\| \leq 2 \left\| adj(\Sigma) \right\|_{op} \lambda_n$$

holds wp at least $1 - 8\exp(-\frac{n}{16})(\delta \cup \delta^2)$

if Σ has at most s non-zeros per row, then $\left\| adj(\Sigma) \right\|_{op} \leq s$

$$\left\| T_\lambda(\hat{\Sigma}) - \Sigma \right\| \leq 2\left(8\sigma^2 \sqrt{\frac{s^2 \log(d)}{n}} + \delta\sigma^2 \right)$$

Bhatia book at matrix analysis tarce inequalities

4.5 random matrices

Definition 4.5.1. mgf:

$$\phi_Q(t) = \mathbb{E}\left[\exp(tQ)\right] \in \mathbb{R}^{d \times d}$$

Definition 4.5.2 (subG matrix)**.** V will play role of σ^2 in $subG(\sigma)$

$$\phi_Q(t) \preceq e^{\frac{t^2 V}{2}}, t \in \mathbb{R}$$

Definition 4.5.3 (operator monotonicity)**.** $f : \mathbb{R} \mapsto \mathbb{R}$ iff

$$A \preceq B \to f(A) \preceq f(B)$$

$$f(x) = \log(x)$$

is operator monotonicity but

$$f(x) = e^x$$

is not op mono. Even thought $f(x) = e^x$ is not op mono, we still get

$$\text{tr}(\exp(A)) \preceq \text{tr}(\exp(B))$$

if f is non-decreasing function

$$\text{tr}(f(A)) \preceq \text{tr}(f(B))$$

4.6 Matrix Chernoff

Theorem 4.6.1.

$$\mathbb{P}\left(\lambda_{max}(Q) \geq \delta\right) \leq \text{tr}(\phi_Q(t))e^{-t\delta}$$

$$\mathbb{P}\left(\lambda_{min}(Q) \geq \delta\right) \leq \text{tr}(\phi_Q(-t))e^{-t\delta}$$

$$\mathbb{P}\left(\left\| Q \right\|_{op} \geq \delta\right) \leq [\text{tr}(\phi_Q(t)) + \text{tr}(\phi_Q(-t))]e^{-t\delta}$$

Lemma 4.6.1. $S_n = Q_1 + \cdots Q_n$

$$\text{tr}(\phi_{S_n}(t)) \leq \text{tr}(\exp(\sum_{i=1}^{n} \log \phi_{Q_i}(t)))$$

Theorem 4.6.2 (Matrix Hoeffding). *Let $Q_1, \cdots Q_n$ be independent mean 0 sym random matrices in $\mathbb{R}^{d\times d}$ then for any $\delta > 0$, we have*

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n} Q_i\right\|_{op} \geq \delta\right) \leq 2de^{-\frac{n\delta^2}{2\|V\|_{op}}}$$

Chapter 5

Sparse Linear Models

5.1 Backgrounds

when $d > n$, the model is not identifiable, and the $X^T X$ will not be invertible.

To eliminate the identifiable issue, if we assume θ^* lines in a special set of parameters Θ_0, then it is possible that $\theta^* + v \notin \Theta_0$, for any $v \in null(X)$, $v \neq 0$, the set Θ_0 will corresponds to sparse or approximately sparse vectors in $\mathbb{R}^d$

$\|\theta^*\|_0 = card\{j \in \{1, \cdots d\} | \theta_j^* \neq 0\}$

5.2 Noisiness model, RNSP

$$y = X\theta^*$$

all-subsets regression, non-convex, we replace the zero norm with l_1 norm. l_1 norm is the one with smallest q that is convex.

$c(s)$ corresponds approximately sparse vectors with support s.

Remark 5.2.1. key question: If $\hat{\theta}$ is the solution to the l_1 minimization problem, when does the $\hat{\theta} = \theta^*$, the answer is this can be determined in terms of the "restricted null space property" of X

If

$$null(X) \cap c(s) = 0$$

then such X satisfies "RNBP" (restricted null space basis pursuit).

Theorem 5.2.1. *For any $\theta^* \in \mathbb{R}^d$ with support S the BP solution is unique and given by $\hat{\theta} = \theta^*$ iff the design matrix X satisfies the RNSP with respect to S.*