36-709: Advanced Statistical Theory I

Spring 2023 • Prof. Matey Neykov

Notes by Fan Pu Zeng

Contents

Lecture 1 — January 17, 2023 **4**
1 Gaussian Sequence Model 4
2 The MLE Can Be a Terrible Estimator 5

Lecture 2 — January 19, 2023 **8**
1 Concentration Inequalities 8
2 Sub-Gaussian Random Variables 10
3 Hoeffding's Bound 11

Lecture 3 — January 24, 2023 **13**
1 Sums of Sub-Gaussian Random Variables 13
2 Equivalent Characterizations of Sub-Gaussian Random Variables 13
3 Sub-Exponential Random Variables 14
4 Bernstein's Condition 16

Lecture 4 — January 26, 2023 **18**
1 Sums of Independent Sub-Exponential Random Variables 18
2 Equivalent Characterizations of Sub-Exponential Random Variables 18
3 Maximal Inequality 19
4 Lipschitz Functions of Gaussian Random Variables 20
5 Martingales 21

Lecture 5 — January 31, 2023 **23**
1 Concentration for Martingale Difference Sequences 23
2 Bounded Differences Inequality 24
3 Covering and Packing 25

Lecture 6 — February 2, 2023 **28**
1 Packing and Covering Numbers 28
2 Volumetric Argument 29
3 Example: Lipschitz Functions on the Unit Interval 30

Lecture 7 — February 7, 2023 **32**
1 Gaussian and Rademacher Complexity 32
2 Function Classes 34
3 Sub-Gaussian Processes 35

Lecture 8 — February 9, 2023 **37**
1 One-Step Discretization (continued) 37
2 The Canonical Gaussian Process 38
3 Chaining and the Dudley Integral 39

Lecture 9 — February 14, 2023 **42**
1 Dudley Entropy Integral 42
2 Gaussian Comparison Inequalities 44

Lecture 10 — February 16, 2023 **46**
1 Slepian's Inequality 46
2 Sudakov–Fernique Inequality 46
3 Gaussian Contraction Inequality 47
4 Sudakov Minoration 48
5 Principal Component Analysis (PCA) 49

Lecture 11 — February 21, 2023 **51**
1 Matrix Recap 51
2 Covariance Matrix Estimation 52

Lecture 12 — February 23, 2023 **55**
1 Spiked Covariance Model 55
2 The Davis–Kahan Theorem 55

Lecture 13 — February 28, 2023 **59**
1 Sparse PCA 59
2 Lasso 60

Lecture 14 — March 2, 2023 **64**
1 Lasso Error Bounds (Continued) 64
2 Restricted Eigenvalues for Random Designs 66
3 Prediction Error 67

Lecture 15 — March 14, 2023 **69**
1 Variable Selection 69

2 Inference after the Lasso 70

36-709: Advanced Statistical Theory I Spring 2023

Lecture 1 — January 17, 2023

Prof. Matey Neykov *Scribe: Fan Pu Zeng*

1 Gaussian Sequence Model

In the *Gaussian sequence model* we have one observation and n parameters:

$$\underbrace{Y}_{\text{observation}} = \underbrace{\mu}_{\text{mean}} + \underbrace{\xi}_{\text{noise}}, \qquad \xi \sim N(0, \sigma^2 I_n),$$

where the noise level σ is known. The goal is to estimate $\mu \in \mathbb{R}^n$.

The James–Stein estimator does much better than the MLE.

Suppose μ lies in a closed convex set K, and let $\Pi_K Y$ be the projection of Y onto K. Then

$$\|\Pi_K Y - \mu\| \leq \|Y - \mu\|.$$

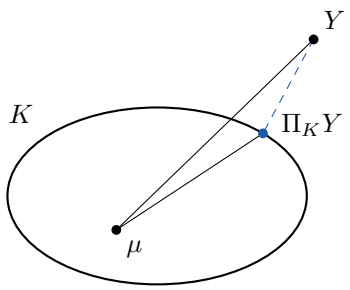


Figure 1: Projecting Y onto the convex set K brings it closer to any $\mu \in K$.

1.1 Comparing estimators

How do we compare estimators? One way is to look at

$$\sup_{\mu \in K} \mathbb{E} \|\hat{\nu}(Y) - \mu\|^2,$$

the "worst-case performance of $\hat{\nu}$ on K", where K is a closed convex set. This is a number, so it lets us compare two estimators $\hat{\nu}_1$ and $\hat{\nu}_2$.

The best possible one achieves

$$\inf_{\hat{\nu}} \sup_{\mu \in K} \mathbb{E} \|\hat{\nu}(Y) - \mu\|^2.$$

2 The MLE Can Be a Terrible Estimator

For some sets K, the MLE can be a terrible estimator. Consider the following example of a set K, an ellipsoid:

$$K = \Big\{x = (x_1, \dots, x_n) : \sum_{i=1}^{n-1} x_i^2 + \frac{x_n^2}{\sqrt{n}} \le 1\Big\}.$$

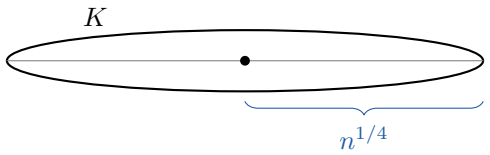


Figure 2: The ellipsoid K is long and thin: it reaches $n^{1/4}$ along the n-th coordinate and at most 1 along the others.

Let's compare two different estimators:

- the MLE;
- $\hat{\nu}(Y) = (\underbrace{0, 0, \dots, 0}_{n-1}, Y_n)$.

2.1 The estimator $\hat{\nu}$

We have

$$\mathbb{E}\,\|\hat{\nu} - \mu\|^2 = \sum_{i=1}^{n-1} \mu_i^2 + \mathbb{E}(Y_n - \mu_n)^2.$$

Assume $\sigma = 1$. Then $Y_n - \mu_n = \xi_n$, so $(Y_n - \mu_n)^2 = \xi_n^2 \sim \chi_1^2$, and $\mathbb{E}[\chi_1^2] = 1$. Hence

$$\mathbb{E}\,\|\hat{\nu} - \mu\|^2 = \underbrace{\sum_{i=1}^{n-1} \mu_i^2}_{\le 1} + 1 \le 2,$$

where $\sum_{i=1}^{n-1} \mu_i^2 \le 1$ since $\mu \in K$. So the error of our estimator is bounded by 2, a constant.

2.2 The MLE

The MLE is the projection

$$\Pi_K Y = \arg\min \|Y - \Pi_K Y\|^2 \quad \text{given that} \quad \sum_{i=1}^{n-1} (\Pi_K Y)_i^2 + \frac{(\Pi_K Y)_n^2}{\sqrt{n}} \le 1.$$

The Lagrangian is

$$\mathcal{L} = \|Y - \Pi_K Y\|^2 + \lambda\Big(\sum_{i=1}^{n-1} (\Pi_K Y)_i^2 + \frac{(\Pi_K Y)_n^2}{\sqrt{n}} - 1\Big).$$

Setting $\partial\mathcal{L}/\partial(\Pi_K Y)_i = 0$:

- For $i \in \{1, \dots, n-1\}$:

$$2(\Pi_K Y)_i - 2Y_i + 2\lambda(\Pi_K Y)_i = 0 \iff \boxed{(1+\lambda)(\Pi_K Y)_i = Y_i}.$$

- For $i = n$:

$$2(\Pi_K Y)_n - 2Y_n + \frac{2\lambda(\Pi_K Y)_n}{\sqrt{n}} = 0 \iff \boxed{\Big(1 + \frac{\lambda}{\sqrt{n}}\Big)(\Pi_K Y)_n = Y_n}.$$

We know the MLE satisfies the two equations above, but it is not in closed form.

Fix

$$\mu = (\underbrace{0, 0, \dots, 0}_{n-1}, n^{1/4}) \in K.$$

(We chose this point adversarially.) We want to lower bound $\sup_{\mu \in K} \mathbb{E}\,\|\Pi_K Y - \mu\|^2$.

For $i \le n-1$ we have $Y_i = \xi_i \sim N(0,1)$, so

$$\sum_{i=1}^{n-1} Y_i^2 \sim \chi^2_{n-1} \gtrsim n,$$

that is, at least n up to some constant.

Roughly speaking (a heuristic explanation): $\mathbb{E}\,\chi^2_{n-1} = n-1$, and such variables concentrate around their mean. Since $\operatorname{Var}\chi^2_{n-1} \sim n-1$,

$$\text{expectation} \pm \text{s.d.} = n - 1 \pm \sqrt{n-1} \gtrsim n.$$

(We discuss concentration inequalities next time.)

Squaring the first stationarity condition and summing,

$$(1+\lambda)^2 \ge (1+\lambda)^2 \sum_{i=1}^{n-1} (\Pi_K Y)_i^2 = \sum_{i=1}^{n-1} Y_i^2 \sim \chi^2_{n-1} \gtrsim n,$$

where the inequality holds because $\Pi_K Y \in K$, so $\sum_{i=1}^{n-1} (\Pi_K Y)_i^2 \le 1$. Therefore $\lambda \gtrsim \sqrt{n}$. In the second stationarity condition

$$\underbrace{\Big(1 + \frac{\lambda}{\sqrt{n}}\Big)}_{>1}(\Pi_K Y)_n = Y_n,$$

the factor is at least a constant larger than 1, so

$$(\Pi_K Y)_n = c\, Y_n = c(\mu_n + \xi_n) = c(n^{1/4} + \xi_n), \qquad c < 1.$$

Then

$$\begin{aligned}
\|\Pi_K Y - \mu\|^2 &= \sum_{i=1}^{n} \big((\Pi_K Y)_i - \mu_i\big)^2 \\
&\ge \big((\Pi_K Y)_n - \mu_n\big)^2 \\
&= \big(c(n^{1/4} + \xi_n) - n^{1/4}\big)^2 \\
&\gtrsim n^{1/2}
\end{aligned}$$

up to some constant, since $\xi_n \sim O(1)$. So the worst-case error of the MLE on this K grows like $\sqrt{n}$, while that of $\hat{\nu}$ is at most 2.

36-709: Advanced Statistical Theory I Spring 2023

Lecture 2 — January 19, 2023

Prof. Matey Neykov *Scribe: Fan Pu Zeng*

1 Concentration Inequalities

1.1 Markov's inequality

Markov's inequality is a tail bound; it is not actually a concentration inequality.

Theorem 1 (Markov's inequality)**.** *Let X be a non-negative random variable, and assume that $\mathbb{E}\,X$ exists. Then*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}\,X}{t}.$$

Proof. Writing $\mathbb{P}(X \geq x) = \mathbb{E}\,\mathbf{1}(X \geq x)$,

$$\begin{aligned}\mathbb{E}\,X = \int_0^\infty \mathbb{P}(X \geq x)\,dx &\geq \int_0^t \mathbb{P}(X \geq x)\,dx \\ &\geq \int_0^t \mathbb{P}(X \geq t)\,dx \qquad (\text{since } x \leq t) \\ &= t\,\mathbb{P}(X \geq t).\end{aligned}$$

■

We also have

$$\mathbb{E}\,X^k = \int_0^\infty kx^{k-1}\,\mathbb{P}(X \geq x)\,dx.$$

1.2 Chebyshev's inequality

Let X have finite variance, and let $\mu = \mathbb{E}\,X$. Applying Markov's inequality to $Y = (X-\mu)^2$ gives, for $t \geq 0$,

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

We can extend the same idea to higher k. Applying Markov's inequality to $Y = |X-\mu|^k$ gives, for $t \geq 0$,

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

assuming $\mathbb{E}|X-\mu|^k$ exists.

These are concentration inequalities: they are centered around the mean.

1.3 The Chernoff bound

The same idea, using Markov's inequality, can be applied to moment generating functions (MGFs). Suppose that

$$\varphi(\lambda) = \mathbb{E}\, e^{\lambda(X-\mu)}$$

exists for all $\lambda \le b$, for some $b > 0$. Apply Markov's inequality to $Y = e^{\lambda(X-\mu)}$: for $\lambda > 0$,

$$\begin{aligned}
\mathbb{P}(X - \mu \ge t) &= \mathbb{P}(\lambda(X-\mu) \ge \lambda t) \\
&= \mathbb{P}(\underbrace{e^{\lambda(X-\mu)}}_{Y} \ge e^{\lambda t}) \\
&\le \frac{\mathbb{E}\, e^{\lambda(X-\mu)}}{e^{\lambda t}} \qquad \text{(Markov's inequality).}
\end{aligned}$$

Now we can optimize for the best $\lambda > 0$ to minimize this:

$$\log \mathbb{P}(X - \mu \ge t) \le \inf_{\lambda \in (0,b]} \Big\{ \log \mathbb{E}\, e^{\lambda(X-\mu)} - \lambda t \Big\}.$$

This gives the *Chernoff bound*.

Actually, optimizing over k in the earlier bound

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

could yield even better bounds. But people still use the MGF method, as it is easier to derive. (See Exercise 2.3 in MJW for more.)

1.4 Applying the Chernoff bound to Gaussian random variables

Let $X \sim N(\mu, \sigma^2)$. Then

$$\mathbb{E}\, e^{\lambda X} = e^{\mu\lambda + \sigma^2\lambda^2/2}, \qquad \lambda \in \mathbb{R}.$$

Hence

$$\begin{aligned}
\inf_{\lambda \ge 0} \Big\{ \log \mathbb{E}\, e^{\lambda(X-\mu)} - \lambda t \Big\} &= \inf_{\lambda \ge 0} \Big\{ \frac{\sigma^2\lambda^2}{2} - \lambda t \Big\} \qquad \text{(a quadratic in } \lambda) \\
&= -\frac{t^2}{2\sigma^2},
\end{aligned}$$

attained at $\lambda = t/\sigma^2$. So we showed

$$\mathbb{P}(X - \mu \ge t) \le e^{-t^2/(2\sigma^2)} \qquad \forall\, t \ge 0. \tag{$\star$}$$

This bound is nearly optimal, up to some factors (Exercise 2.2 in MJW).

2 Sub-Gaussian Random Variables

Definition (Sub-Gaussian random variable)**.** A random variable X with mean $\mu = \mathbb{E}\,X$ is *sub-Gaussian* with parameter σ^2, written $\mathrm{SG}(\sigma^2)$, if there exists $\sigma > 0$ such that

$$\mathbb{E}\,e^{\lambda(X-\mu)} \le e^{\sigma^2\lambda^2/2} \qquad \forall\,\lambda \in \mathbb{R}.$$

That is, the MGF of the centered variable is smaller than the MGF of a Gaussian.

Why do we care about sub-Gaussian random variables? The bound $(\star)$ holds for any sub-Gaussian random variable: both have light tails. The parameter σ is called the *sub-Gaussian parameter* (or *variance proxy*).

Any Gaussian is sub-Gaussian, with σ equal to the standard deviation of the Gaussian. (So the class of sub-Gaussian random variables is non-empty.)

Fact 1. $-X$ is sub-Gaussian if and only if X is sub-Gaussian.

So, by applying $(\star)$ to $-X$,

$$\mathbb{P}(X \le \mu - t) \le e^{-t^2/(2\sigma^2)} \qquad \forall\,t \ge 0.$$

Thus any sub-Gaussian random variable satisfies

$$\mathbb{P}(|X-\mu| \ge t) \le 2e^{-t^2/(2\sigma^2)} \qquad \forall\,t \ge 0.$$

Example 1 (Rademacher random variable)**.** Let ε be -1 or $+1$, each with probability $1/2$. We claim that a Rademacher random variable is sub-Gaussian with $\sigma = 1$.

Here $\mu = 0$, and

$$\begin{aligned}
\mathbb{E}\,e^{\lambda\varepsilon} &= \underbrace{\tfrac{1}{2}(e^{\lambda} + e^{-\lambda})}_{\cosh(\lambda)} && \text{(hyperbolic cosine)}\\
&= \frac{1}{2}\Big\{\sum_{k=0}^{\infty}\frac{(-\lambda)^k}{k!} + \sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\Big\} && \text{(Taylor expansion)}\\
&= \sum_{k=0}^{\infty}\frac{\lambda^{2k}}{(2k)!} && \text{(odd degrees cancel out)}\\
&\le 1 + \sum_{k=1}^{\infty}\frac{\lambda^{2k}}{2^k k!}\\
&= e^{\lambda^2/2},
\end{aligned}$$

where the inequality uses $(2k)! \ge 2^k k!$. To see this, write

$$(2k)! = \underbrace{1\cdot 2\cdots k}_{k!}\cdot\underbrace{(k+1)\cdots 2k}_{\text{each factor}\ge 2}.$$

So

$$\boxed{\cosh(\lambda) \le e^{\lambda^2/2}},$$

a bound that appears a lot. So Rademacher random variables are sub-Gaussian with $\sigma = 1$.

Example 2 (Bounded random variables)**.** Let X be a mean-0 random variable supported on $[a, b]$, and let X' be an independent copy of X. Since $\mu = 0$, the MGF of the centered variable is just the MGF, and

$$\begin{aligned} \mathop{\mathbb{E}}_{X} e^{\lambda X} &= \mathop{\mathbb{E}}_{X} e^{\lambda(X - \mathbb{E}_{X'} X')} && (\text{since } \mathop{\mathbb{E}}_{X'} X' = 0) \\ &\le \mathop{\mathbb{E}}_{X,X'} e^{\lambda(X - X')}, \end{aligned}$$

since $\exp(-x)$ is convex, by Jensen's inequality.

Let ε be a Rademacher random variable independent of everything else. Then $X - X' \overset{d}{=} \varepsilon(X - X')$ (equal in distribution), so

$$\mathop{\mathbb{E}}_{X,X'} e^{\lambda(X-X')} = \mathop{\mathbb{E}}_{X,X'} \big[\mathop{\mathbb{E}}_{\varepsilon} e^{\lambda\varepsilon(X-X')}\big], \qquad \text{where} \quad \mathop{\mathbb{E}}_{\varepsilon} e^{\lambda\varepsilon(X-X')} = \tfrac{1}{2} e^{\lambda(X-X')} + \tfrac{1}{2} e^{\lambda(X'-X)}.$$

With $\tilde{\lambda} = \lambda(X - X')$, the inner expectation is the MGF of a Rademacher random variable, so $\mathbb{E}_{\varepsilon} e^{\tilde{\lambda}\varepsilon} \le e^{\tilde{\lambda}^2/2}$. Hence

$$\begin{aligned} \mathop{\mathbb{E}}_{X,X'} \big[\mathop{\mathbb{E}}_{\varepsilon} e^{\tilde{\lambda}\varepsilon}\big] &\le \mathop{\mathbb{E}}_{X,X'} e^{\lambda^2 (X-X')^2/2} \\ &\le e^{\lambda^2 (b-a)^2/2}, \end{aligned}$$

by the crude bound $|X - X'| \le b - a$.

Conclusion. Bounded random variables are sub-Gaussian with $\sigma = b - a$.

Remark 1. This is not the sharpest bound. Exercise 2.4 in MJW shows that one can take $\sigma = (b - a)/2$.

3 Hoeffding's Bound

Suppose $X_1 \sim \mathrm{SG}(\sigma_1^2)$ and $X_2 \sim \mathrm{SG}(\sigma_2^2)$ are independent. Then

$$X_1 + X_2 \sim \mathrm{SG}(\sigma_1^2 + \sigma_2^2).$$

This can be proven by computing their MGFs and using independence.

Proposition 1 (Hoeffding's bound)**.** Suppose that the X_i are independent, and that X_i has mean μ_i and $X_i \sim \mathrm{SG}(\sigma_i^2)$. Then for all $t \ge 0$,

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

This follows from $\sum_{i=1}^{n} X_i \sim \mathrm{SG}\big(\sum_{i=1}^{n} \sigma_i^2\big)$ and $(\star)$, $\mathbb{P}(X - \mu \ge t) \le e^{-t^2/(2\sigma^2)}$.

Hoeffding's bound is usually stated for bounded random variables, but it can also be stated for sub-Gaussian ones, like the above.

Bounded random variable version. Let $X_i \in [a,b]$ with $\mathbb{E}\, X_i = \mu_i$, so that $X_i \sim \mathrm{SG}(\sigma^2)$ with $\sigma = (b-a)/2$. Then

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

This is an immediate consequence of substituting $\sigma_i = (b-a)/2$.

Remark 2. Every bounded random variable on $[a,b]$ is sub-Gaussian with parameter $(b-a)/2$. (It won't be with $(b-a)/4$, for example.)

If $X \sim \mathrm{SG}(\sigma^2)$, then $X \sim \mathrm{SG}(\sigma'^2)$ for $\sigma' > \sigma$. But we want the smallest possible σ. It turns out that σ^2 cannot be smaller than the variance of the random variable.

Proposition 2.

(1) If $X \sim \mathrm{SG}(\sigma^2)$, then $\mathrm{Var}(X) \leq \sigma^2$, with $\mathrm{Var}(X) = \sigma^2$ for X Gaussian.

(2) Let $X \sim \mathrm{SG}(\sigma^2)$ and $Y \sim \mathrm{SG}(\tau^2)$. Then

(i) $\alpha X \sim \mathrm{SG}(\alpha^2\sigma^2)$;

(ii) $X + Y \sim \mathrm{SG}\big((\tau+\sigma)^2\big)$;

(iii) if X and Y are independent, then $X + Y \sim \mathrm{SG}(\sigma^2 + \tau^2)$.

Parts (i) and (iii) follow from the definition.

Proof of (1). By Taylor expansion,

$$\begin{aligned}
\mathbb{E}\, e^{\lambda(X - \mathbb{E}\, X)} &= 1 + \lambda \underbrace{\mathbb{E}(X - \mathbb{E}\, X)}_{0} + \frac{\lambda^2\, \mathbb{E}(X - \mathbb{E}\, X)^2}{2} + o(\lambda^2) \\
&= 1 + \frac{\lambda^2\, \mathrm{Var}(X)}{2} + o(\lambda^2) \\
&\leq e^{\lambda^2\sigma^2/2} = 1 + \frac{\lambda^2\sigma^2}{2} + o(\lambda^2).
\end{aligned}$$

Divide by λ^2 and let $\lambda \to 0$. So $\sigma^2 \geq \mathrm{Var}(X)$. ■

Lecture 3 — January 24, 2023

Prof. Matey Neykov *Scribe: Fan Pu Zeng*

1 Sums of Sub-Gaussian Random Variables

Previously we stated that if $X \sim \mathrm{SG}(\sigma^2)$ and $Y \sim \mathrm{SG}(\tau^2)$, then

$$X + Y \sim \mathrm{SG}\big((\sigma+\tau)^2\big),$$

with no independence assumption. Take X and Y to be centered. Then

$$\begin{aligned}
\mathbb{E}\, e^{\lambda(X+Y)} &= \mathbb{E}\big[e^{\lambda X} e^{\lambda Y}\big] \\
&\le \big(\mathbb{E}\, e^{\lambda p X}\big)^{1/p} \big(\mathbb{E}\, e^{\lambda q Y}\big)^{1/q} && \text{(Hölder, where } \tfrac{1}{p}+\tfrac{1}{q}=1\text{)} \\
&\le \big(e^{\lambda^2 p^2 \sigma^2/2}\big)^{1/p} \big(e^{\lambda^2 q^2 \tau^2/2}\big)^{1/q} \\
&= e^{\lambda^2(p\sigma^2 + q\tau^2)/2}.
\end{aligned}$$

Now choose

$$\arg\min_{p,q}\ p\sigma^2 + q\tau^2 \quad \text{subject to} \quad \frac{1}{p} + \frac{1}{q} = 1,$$

which gives $p = \tau/\sigma + 1$ and $q = \sigma/\tau + 1$, so that $p\sigma^2 + q\tau^2 = \sigma^2 + 2\sigma\tau + \tau^2$. Therefore

$$\mathbb{E}\, e^{\lambda(X+Y)} \le e^{\lambda^2(\sigma+\tau)^2/2}.$$

2 Equivalent Characterizations of Sub-Gaussian Random Variables

Theorem 2 (Equivalence of sub-Gaussian random variables)**.** *Given a zero-mean random variable X, the following are equivalent.*

(I) There exists $\sigma \ge 0$ such that

$$\mathbb{E}\, e^{\lambda X} \le e^{\lambda^2\sigma^2/2} \qquad \forall\, \lambda \in \mathbb{R}.$$

(II) There exist $c \ge 0$ and a Gaussian $Z \sim N(0, \tau^2)$ such that

$$\mathbb{P}(|X| \ge s) \le c\, \mathbb{P}(|Z| \ge s) \qquad \forall\, s \ge 0.$$

(Here c can depend on τ.)

(III) There exists a constant $\theta \ge 0$ such that

$$\mathbb{E}\, X^{2k} \le \frac{(2k)!}{2^k k!}\, \theta^{2k}, \qquad k = 1, 2, \ldots$$

That is, the even moments are bounded. The factor $\frac{(2k)!}{2^k k!}$ grows like k^k.

(IV) There exists $\sigma \geq 0$ such that

$$\mathbb{E}\, e^{\lambda X^2/(2\sigma^2)} \leq \frac{1}{\sqrt{1-\lambda}} \qquad \forall\, \lambda \in [0,1).$$

So sub-Gaussian random variables could be defined in any of these ways.

Remark 3. In (III), by Hölder's inequality with $p = \frac{2k}{2k-1}$,

$$\left(\mathbb{E}|X|^{2k-1}\right)^{\frac{1}{2k-1}} \leq \left(\mathbb{E}\, X^{2k}\right)^{\frac{1}{2k}},$$

so the odd moments are controlled as well.

3 Sub-Exponential Random Variables

Definition (Sub-exponential)**.** A random variable X with $\mu = \mathbb{E}\, X$ is *sub-exponential* if there exist (ν, α) with $\nu, \alpha \geq 0$ such that

$$\mathbb{E}\, e^{\lambda(X-\mu)} \leq e^{\nu^2\lambda^2/2} \qquad \forall\, |\lambda| \leq \alpha^{-1}.$$

The only difference from the sub-Gaussian definition is that λ is restricted to this range. So sub-Gaussian implies sub-exponential: sub-Gaussian random variables are sub-exponential with $\alpha = 0$ and $\nu = \sigma$. Sub-exponential random variables have bigger tails, and concentrate, but worse than sub-Gaussian ones.

Example 3. Let $Z \sim N(0,1)$ and $X = Z^2$, so that $\mathbb{E}\, X = 1$ (the mean of a χ^2 variable). Take $|\lambda| < 1/4$. Then

$$\begin{aligned}
\mathbb{E}\, e^{\lambda(X-1)} &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\lambda(z^2-1)} e^{-z^2/2}\, dz \\
&= \frac{e^{-\lambda}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \underbrace{e^{(\lambda-1/2)z^2}}_{\text{Gaussian kernel}}\, dz \\
&= \frac{e^{-\lambda}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2(1-2\lambda)^{-1}}}\, dz \\
&= \frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \\
&\leq e^{\lambda^2/(1-2\lambda)}.
\end{aligned}$$

The last inequality follows from

$$\boxed{-\log(1-u) - u \leq \frac{u^2}{2(1-u)} \qquad \forall\, u \in (0,1)}$$

(for example, by taking derivatives). Continuing,

$$\mathbb{E}\, e^{\lambda(X-1)} \leq e^{2\lambda^2} = e^{4\lambda^2/2} \qquad (\text{since } |\lambda| \leq 1/4).$$

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

Proposition 3. Suppose X is sub-exponential with (ν, α). Then

$$\mathbb{P}(X - \mu \geq t) \leq \begin{cases} e^{-t^2/(2\nu^2)}, & 0 \leq t \leq \nu^2/\alpha, \\ e^{-t/(2\alpha)}, & t > \nu^2/\alpha. \end{cases}$$

When t is large, the concentration is worse than sub-Gaussian. A similar inequality holds for $\mathbb{P}(X - \mu \leq -t)$, and for $\mathbb{P}(|X - \mu| \geq t)$, where you need to multiply by 2.

Proof. Without loss of generality, $\mu = 0$. By the Chernoff bound,

$$\mathbb{P}(X \geq t) \leq e^{-\lambda t}\, \mathbb{E}\, e^{\lambda X} \leq e^{g(\lambda,t)}, \qquad g(\lambda, t) = -\lambda t + \frac{\nu^2 \lambda^2}{2},$$

valid for $\lambda \in [0, \alpha^{-1}]$. We need, for every fixed t,

$$g^*(t) = \inf_{\lambda \in [0, \alpha^{-1}]} g(\lambda, t).$$

The unconstrained minimum of $g(\cdot, t)$ is at

$$\boxed{\lambda^* = t/\nu^2}.$$

Case $t \leq \nu^2/\alpha$. Then $\lambda^* \leq \alpha^{-1}$, so λ^* is also the constrained minimum, and

$$g^*(t) = g(\lambda^*, t) = -\frac{t^2}{2\nu^2}.$$

Case $t > \nu^2/\alpha$. Then

$$\lambda^* = \frac{t}{\nu^2} > \frac{\nu^2}{\alpha \nu^2} = \alpha^{-1},$$

so the minimum over the constrained set is at α^{-1} (Figure 3), and

$$\begin{aligned} g^*(t) = g(\alpha^{-1}, t) &= -\frac{t}{\alpha} + \frac{\nu^2}{2\alpha^2} \\ &\leq -\frac{t}{\alpha} + \frac{t}{2\alpha} \qquad (\text{since } t > \nu^2/\alpha) \\ &= -\frac{t}{2\alpha}. \end{aligned}$$

■

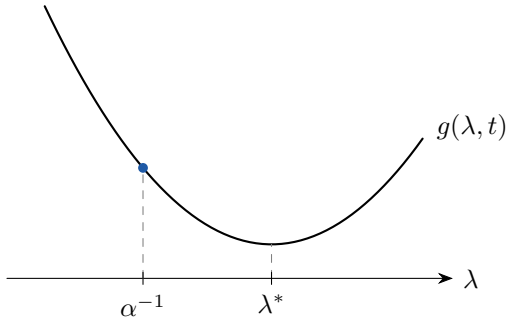


Figure 3: When $t > \nu^2/\alpha$, the unconstrained minimizer λ^* lies beyond α^{-1}, so the minimum over $[0, \alpha^{-1}]$ is at α^{-1}.

4 Bernstein's Condition

Bernstein's condition is a condition on the moments. Let X be a random variable with $\mu = \mathbb{E}\, X$ and $\sigma^2 = \mathbb{E}\, X^2 - \mu^2$. Suppose that

$$\mathbb{E}|X-\mu|^k \le \frac{1}{2} k!\, \sigma^2 b^{k-2} \qquad \forall\, k = 2, 3, \dots$$

for some constant $b > 0$.

We show that this implies X is sub-exponential. By the Taylor expansion of the MGF (the first-order term cancels out),

$$\begin{aligned}
\mathbb{E}\, e^{\lambda(X-\mu)} &= 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k\ge 3} \frac{\lambda^k\, \mathbb{E}(X-\mu)^k}{k!} \\
&\le 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^2\sigma^2}{2} \sum_{k\ge 3} (|\lambda| b)^{k-2} && \text{(Bernstein's condition)} \\
&= 1 + \frac{\lambda^2\sigma^2}{2} \underbrace{\sum_{k=0}^{\infty} (|\lambda| b)^k}_{1/(1-|\lambda| b)} \\
&= 1 + \frac{\lambda^2\sigma^2}{2(1-|\lambda| b)} \\
&\le e^{\lambda^2\sigma^2/(2(1-|\lambda| b))} && \text{(using } 1 + x \le e^x\text{)}.
\end{aligned}$$

Suppose $|\lambda| \le \frac{1}{2b}$. Then

$$\mathbb{E}\, e^{\lambda(X-\mu)} \le e^{\lambda^2\sigma^2} = e^{(\sqrt{2}\sigma)^2\lambda^2/2} \qquad \text{for } |\lambda| \le \frac{1}{2b},$$

so X is sub-exponential with $(\nu, \alpha) = (\sqrt{2}\sigma, 2b)$.

So Bernstein's condition is easier to check, as you don't need to evaluate the whole MGF, unlike the definition of sub-exponential.

Proposition 4. For every random variable satisfying Bernstein's condition, we have

- $\mathbb{E}\, e^{\lambda(X-\mu)} \le \exp\Big(\dfrac{\lambda^2\sigma^2}{2(1-|\lambda| b)}\Big)$ for all $|\lambda| < \frac{1}{b}$;
- $\mathbb{P}(|X-\mu| \ge t) \le 2\exp\Big(-\dfrac{t^2}{2(\sigma^2 + bt)}\Big)$ for $t \ge 0$.

Proof sketch. By the Chernoff bound,

$$\mathbb{P}(X - \mu \ge t) \le e^{-\lambda t}\, \mathbb{E}\, e^{\lambda(X-\mu)} \le \exp\Big(-\lambda t + \frac{\lambda^2\sigma^2}{2(1-\lambda b)}\Big).$$

Plug in $\lambda = \frac{t}{bt+\sigma^2} \in [0, \frac{1}{b})$. Then $1 - \lambda b = \frac{\sigma^2}{bt+\sigma^2}$, and the exponent becomes

$$-\frac{t^2}{bt+\sigma^2} + \frac{t^2}{2(bt+\sigma^2)} = -\frac{t^2}{2(\sigma^2+bt)}.$$

The bound on $\mathbb{P}(X - \mu \le -t)$ is similar. ∎

4.1 Bounded random variables

Note that a bounded random variable satisfies Bernstein's condition. Suppose $|X-\mu| \le 3b$ (possible since X is bounded). Then for $k = 2, 3, \ldots,$

$$\begin{aligned}\mathbb{E}|X-\mu|^k &= \mathbb{E}|X-\mu|^2|X-\mu|^{k-2} \\ &\le \mathbb{E}|X-\mu|^2(3b)^{k-2} \\ &\le \sigma^2 \frac{k!}{2} b^{k-2}.\end{aligned}$$

We just need to verify

$$\boxed{3^{k-2} \le \frac{k!}{2}},$$

which is true for $k = 3$, and definitely true for $k > 3$. So this satisfies Bernstein's condition, and

> For a bounded random variable,
>
> $$\mathbb{P}(|X-\mu| \ge t) \le 2\exp\Big(-\frac{t^2}{2(\sigma^2 + bt)}\Big).$$

Contrast this with Hoeffding's bound.

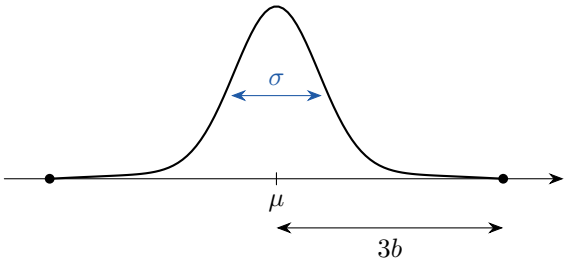


Figure 4: A bounded random variable whose spread σ is much smaller than its range, which is set by b.

Remark 4. If $\sigma \ll b$ and we look at small values of t, then the concentration that we get from Bernstein's inequality is superior to Hoeffding's.

Bennett's inequality is even better; see Exercise 2.7 in MJW.

Lecture 4 — January 26, 2023

Prof. Matey Neykov *Scribe: Fan Pu Zeng*

1 Sums of Independent Sub-Exponential Random Variables

Suppose $\{X_k\}_{k=1}^n$ are independent sub-exponential random variables with parameters (ν_k, α_k), and $\mathbb{E}\, X_k = \mu_k$. Consider $\sum_{k=1}^n (X_k - \mu_k)$. By independence,

$$\mathbb{E}\, e^{\lambda \sum_{k=1}^n (X_k - \mu_k)} = \prod_{k=1}^n \mathbb{E}\, e^{\lambda (X_k - \mu_k)} \leq \prod_{k=1}^n e^{\lambda^2 \nu_k^2/2} = e^{\lambda^2 \left(\sum_{k=1}^n \nu_k^2\right)/2},$$

where the k-th bound needs $|\lambda| \leq 1/\alpha_k$. All of them hold on

$$\bigcap_{k=1}^n \Big\{|\lambda| \leq \frac{1}{\alpha_k}\Big\} = \Big\{|\lambda| \leq \frac{1}{\max_{1\leq k\leq n} \alpha_k}\Big\}.$$

Therefore

$$\sum_{k=1}^n (X_k - \mu_k) \sim \text{sub-exp}\Big(\sqrt{\textstyle\sum_{k=1}^n \nu_k^2},\ \max_{1\leq k\leq n} \alpha_k\Big).$$

2 Equivalent Characterizations of Sub-Exponential Random Variables

Theorem 3. *The following are equivalent for a random variable X with $\mathbb{E}\, X = 0$.*

(I) There exist (ν, α) such that

$$\mathbb{E}\, e^{\lambda X} \leq e^{\lambda^2 \nu^2/2} \qquad \forall\, |\lambda| \leq \frac{1}{\alpha}.$$

(No centering is needed since $\mu = 0$.)

(II) There exists $c_0 > 0$ such that

$$\mathbb{E}\, e^{\lambda X} < \infty \qquad \forall\, |\lambda| < c_0.$$

(III) There exist $c_1, c_2 > 0$ such that

$$\mathbb{P}(|X| \geq t) \leq c_1 e^{-c_2 t}, \qquad t > 0.$$

(IV) The quantity

$$\gamma = \sup_{k\geq 2} \Big(\frac{\mathbb{E}|X|^k}{k!}\Big)^{1/k}$$

is finite. In other words, $\mathbb{E}|X|^k \leq k!\, \gamma^k$.

Some of these will be useful for the homework!

3 Maximal Inequality

Theorem 4 (Maximal inequality)**.** *Let* $X_1, \dots, X_n \sim \mathrm{SG}(\sigma^2)$ *with* $\mathbb{E}\, X_i = 0$*. Then*

$$\mathbb{E} \max_{1 \le i \le n} X_i \le \sigma\sqrt{2 \log n}, \qquad \mathbb{E} \max_{1 \le i \le n} |X_i| \le \sigma\sqrt{2 \log 2n}.$$

Furthermore, for $t > 0$*,*

$$\mathbb{P}\Big(\max_{1 \le i \le n} X_i \ge t\Big) \le n e^{-t^2/(2\sigma^2)}, \qquad \mathbb{P}\Big(\max_{1 \le i \le n} |X_i| \ge t\Big) \le 2n e^{-t^2/(2\sigma^2)}.$$

Remark 5. It is known that if $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$, then

$$\mathbb{E} \max_{1 \le i \le n} X_i \asymp \mathbb{E} \max_{1 \le i \le n} |X_i| \asymp \sigma\sqrt{\log n};$$

see Exercise 2.11 in MJW. Here "$\asymp$" means that there exist constants c, C such that

$$c\,\sigma\sqrt{\log n} \le \mathbb{E} \max_i X_i \le C\sigma\sqrt{\log n}.$$

Proof of Theorem 4. For every $s > 0$, rewriting,

$$\begin{aligned}
\mathbb{E} \max_i X_i &= \frac{1}{s}\, \mathbb{E} \log e^{s \max_i X_i} \\
&\le \frac{1}{s} \log \mathbb{E}\, e^{s \max_i X_i} && \text{(Jensen's inequality; log is concave)} \\
&= \frac{1}{s} \log \mathbb{E} \max_{1 \le i \le n} e^{sX_i} \\
&\le \frac{1}{s} \log \sum_{i=1}^n \mathbb{E}\, e^{sX_i} && \text{(the max is at most the sum)} \\
&\le \frac{1}{s} \log\big(n e^{s^2\sigma^2/2}\big) && \text{(since the } X_i \text{ are sub-Gaussian)} \\
&= \frac{\log n}{s} + \frac{s\sigma^2}{2}.
\end{aligned}$$

Now optimize over s. Choose $s = \sqrt{2 \log n}/\sigma$, which equates the two terms: one is decreasing in s and the other is increasing, and the optimum is where they meet. This gives

$$\mathbb{E} \max_i X_i \le \sigma\sqrt{2 \log n},$$

which shows the first part. For the tail bound,

$$\mathbb{P}\Big(\max_{1 \le i \le n} X_i \ge t\Big) = \mathbb{P}\Big(\bigcup_{i=1}^n \{X_i \ge t\}\Big) \le \sum_{i=1}^n \mathbb{P}(X_i \ge t) \le n e^{-t^2/(2\sigma^2)},$$

using that each X_i is sub-Gaussian.

To prove the remaining parts, note that

$$\max_{1 \le i \le n} |X_i| = \max_{1 \le i \le 2n} X_i, \qquad \text{where } X_{i+n} = -X_i \text{ for } i \in \{1, \dots, n\}.$$

Each $-X_i$ is also $\mathrm{SG}(\sigma^2)$. So replace n by $2n$, and you get the bounds $\sigma\sqrt{2 \log 2n}$ and $2n e^{-t^2/(2\sigma^2)}$. ∎

4 Lipschitz Functions of Gaussian Random Variables

Thus far we have considered concentration of random variables. Now let us study something more general: how functions of them concentrate.

Definition (Lipschitz function)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$. Then f is called *L-Lipschitz* (with respect to $\|\cdot\|_2$) if

$$|f(x) - f(y)| \le L \|x - y\|_2 \qquad \forall\, x, y \in \mathbb{R}^n.$$

Theorem 5. *Let $X = (X_1, \dots, X_n)$ be a vector of i.i.d. Gaussian random variables, and let $f : \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz with respect to $\|\cdot\|_2$. Then $f(X) - \mathbb{E}\, f(X)$ is sub-Gaussian with parameter at most L, and*

$$\mathbb{P}\big(|f(X) - \mathbb{E}\, f(X)| \ge t\big) \le 2e^{-t^2/(2L^2)}, \qquad t \ge 0.$$

The remarkable thing is that this bound is dimension-independent. This is very hard to prove in general, but it is true for i.i.d. Gaussians.

Example 4 (Gaussian complexity / Gaussian width)**.** Let $A \subset \mathbb{R}^n$, and let $\{w_k\}_{k=1}^n$ be i.i.d. $N(0,1)$. Define

$$Z = \sup_{a \in A} \sum_{k=1}^n a_k w_k = \sup_{a \in A} \langle a, w \rangle ,$$

the supremum of a dot product.

This is relevant to the least-squares estimator (LSE): the rate of convergence of the LSE can be expressed in terms of some function involving Gaussian widths.

We can view Z as a map

$$(w_1, \dots, w_n) \mapsto f(w_1, \dots, w_n), \qquad \text{where } f(w) = \sup_{a \in A} \langle a, w \rangle .$$

Let $w, w' \in \mathbb{R}^n$, and let a^* attain the supremum in $f(w)$. Since $-\sup_{a \in A} \langle a, w' \rangle \le -\langle a^*, w' \rangle$,

$$\begin{aligned} f(w) - f(w') &= \sup_{a \in A} \langle a, w \rangle - \sup_{a \in A} \langle a, w' \rangle \\ &\le \langle a^*, w \rangle - \langle a^*, w' \rangle \\ &= \langle a^*, w - w' \rangle \\ &\le \|a^*\|_2 \, \|w - w'\|_2 \qquad \text{(Cauchy–Schwarz)} \\ &\le \sup_{a \in A} \|a\|_2 \, \|w - w'\|_2 . \end{aligned}$$

Swapping w and w' gives the same bound, so

$$|f(w') - f(w)| \le \sup_{a \in A} \|a\|_2 \, \|w - w'\|_2 .$$

So f is a Lipschitz function with $L = \sup_{a \in A} \|a\|_2$, and by Theorem 5,

$$\mathbb{P}(|Z - \mathbb{E}\, Z| \ge t) \le 2 \exp\Big(-\frac{t^2}{2 \sup_{a \in A} \|a\|_2^2}\Big).$$

So a Lipschitz function of Gaussians (here Z) sharply concentrates around its mean.

5 Martingales

Let $\{\mathcal{F}_k\}_{k=1}^{\infty}$ be a collection of σ-fields such that $\mathcal{F}_k \subseteq \mathcal{F}_{k+1}$ for all $k \geq 1$. This is called a *filtration*.

Suppose $\{Y_k\}_{k=1}^{\infty}$ is a sequence where each Y_k is $\mathcal{F}_k$-measurable. Then we say $\{Y_k\}_{k=1}^{\infty}$ is *adapted* to $\{\mathcal{F}_k\}_{k=1}^{\infty}$.

Definition (Martingale)**.** Given a sequence $\{Y_k\}_{k=1}^{\infty}$ adapted to $\{\mathcal{F}_k\}_{k=1}^{\infty}$, we call $\{(Y_k, \mathcal{F}_k)\}_{k=1}^{\infty}$ a *martingale* if for all $k \geq 1$,

$$\mathbb{E}|Y_k| < \infty \qquad \text{and} \qquad \mathbb{E}[Y_{k+1} \mid \mathcal{F}_k] = Y_k.$$

Recall that $\mathbb{E}[Y_{k+1} \mid \mathcal{F}_k]$ is $\mathcal{F}_k$-measurable and satisfies

$$\int_A Y_{k+1}\, d\,\mathbb{P} = \int_A \mathbb{E}[Y_{k+1} \mid \mathcal{F}_k]\, d\,\mathbb{P} \qquad \forall\, A \in \mathcal{F}_k.$$

Also, $\mathbb{E}[Y \mid X] = \mathbb{E}[Y \mid \sigma(X)]$.

Example 5 (Doob's construction)**.** Let $X_1, \ldots, X_n$ be independent random variables, write $X = (X_1, \ldots, X_n)$, and let $f : \mathbb{R}^n \to \mathbb{R}$. Define

$$Y_k = \mathbb{E}[f(X) \mid X_1, \ldots, X_k], \qquad \mathcal{F}_k = \sigma(X_1, \ldots, X_k).$$

Then $(Y_k, \mathcal{F}_k)$ is a martingale if $\mathbb{E}|f(X)| < \infty$.

To verify $\mathbb{E}|Y_k| < \infty$:

$$\begin{aligned} \mathbb{E}\big|\mathbb{E}[f(X) \mid X_1, \ldots, X_k]\big| &\leq \mathbb{E}\ \mathbb{E}\big[|f(X)| \mid X_1, \ldots, X_k\big] && \text{(Jensen)} \\ &= \mathbb{E}|f(X)| && \text{(tower property)} \\ &< \infty. \end{aligned}$$

For the martingale property,

$$\begin{aligned} \mathbb{E}[Y_{k+1} \mid X_1, \ldots, X_k] &= \mathbb{E}\big[\mathbb{E}[f(X) \mid X_1, \ldots, X_k, X_{k+1}] \mid X_1, \ldots, X_k\big] \\ &= \mathbb{E}[f(X) \mid X_1, \ldots, X_k] \\ &= Y_k. \end{aligned}$$

So this is a martingale.

5.1 Martingale difference sequences

A *martingale difference sequence* (MDS) is a sequence $\{D_k\}_{k=1}^{\infty}$ adapted to a filtration $\{\mathcal{F}_k\}_{k=1}^{\infty}$ such that

$$\mathbb{E}|D_k| < \infty, \qquad \mathbb{E}[D_{k+1} \mid \mathcal{F}_k] = 0.$$

Let $\{Y_k\}_{k=1}^{\infty}$ be a martingale with respect to $\{\mathcal{F}_k\}_{k=1}^{\infty}$, and define

$$D_{k+1} = Y_{k+1} - Y_k.$$

Then

$$Y_n - Y_0 = \sum_{k=1}^{n} D_k.$$

For Doob's construction (Example 5), $Y_n = f(X)$ and $Y_0 = \mathbb{E}\, Y_1 = \mathbb{E}\, f(X)$, and

$$D_{k+1} = \mathbb{E}[f(X) \mid X_1, \ldots, X_{k+1}] - \mathbb{E}[f(X) \mid X_1, \ldots, X_k].$$

Theorem 6. *Let $\{(D_k, \mathcal{F}_k)\}_{k=1}^n$ be an MDS, and suppose that*

$$\mathbb{E}\big[e^{\lambda D_k} \mid \mathcal{F}_{k-1}\big] \le e^{\lambda^2 \nu_k^2/2} \qquad \forall\, |\lambda| \le 1/\alpha_k$$

almost surely. (The left-hand side is a random variable.) Let $\alpha^* = \max_{1 \le k \le n} \alpha_k$. Then:

(a) $\sum_{k=1}^n D_k$ *is* sub-exp$(\sqrt{\sum_k \nu_k^2},\, \alpha^*)$*;*

(b)

$$\mathbb{P}\Big(\Big|\sum_{k=1}^{n} D_k\Big| \ge t\Big) \le \begin{cases} 2e^{-t^2/(2\sum_k \nu_k^2)}, & 0 < t \le \dfrac{\sum_k \nu_k^2}{\alpha^*}, \\ 2e^{-t/(2\alpha^*)}, & t > \dfrac{\sum_k \nu_k^2}{\alpha^*}. \end{cases}$$

So we have a concentration bound on $\big|\sum_k D_k\big|$.

36-709: Advanced Statistical Theory I Spring 2023

Lecture 5 — January 31, 2023

Prof. Matey Neykov *Scribe: Fan Pu Zeng*

1 Concentration for Martingale Difference Sequences

We continue with the theorem from last time.

Theorem 7. *Let $\{(D_k, \mathcal{F}_k)\}_{k=1}^n$ be a martingale difference sequence (MDS), and suppose that almost surely*

$$\mathbb{E}[e^{\lambda D_k} \mid \mathcal{F}_{k-1}] \leq e^{\lambda^2 \nu_k^2/2} \qquad \forall\, |\lambda| \leq \frac{1}{\alpha_k}.$$

Let $\alpha^* = \max_{k\in[n]} \alpha_k$. Then

(a) $\sum_{k=1}^n D_k$ is sub-exp$(\sqrt{\sum_{k=1}^n \nu_k^2},\, \alpha^*)$;

(b) we also have a concentration bound on $\mathbb{P}(|\sum_{k=1}^n D_k| \geq t)$, as before.

Remark 6. Let $Y = \mu + \xi$, where $\mu \in K$ and $\xi \sim N(0, \sigma^2 I)$. If $\hat{\mu}$ is the MLE and $r_\mu = \mathbb{E}\,\|\hat{\mu} - \mu\|^2$, then $r_\mu \asymp \varepsilon_\mu^2$, where

$$\varepsilon_\mu = \arg\max_{\varepsilon}\ \sigma\, w\big(B(\mu, \varepsilon) \cap K\big) - \frac{\varepsilon^2}{2}.$$

Here $B(\mu, \varepsilon)$ is the ε-ball around μ, and

$$w\big(B(\mu, \varepsilon) \cap K\big) = \mathbb{E} \sup_{t \in B(\mu,\varepsilon)\cap K} \langle t, z\rangle\,, \qquad z \sim N(0, I),$$

is the Gaussian complexity of the set $B(\mu, \varepsilon) \cap K$.

Proof of Theorem 7. Let λ be such that $|\lambda| \leq 1/\alpha^*$. We want to show that the sum is sub-exponential. Since $\sum_{k=1}^{n-1} D_k$ is $\mathcal{F}_{n-1}$-measurable, by the iterative property of conditional expectation,

$$\begin{aligned}
\mathbb{E}\Big[e^{\lambda \sum_{k=1}^n D_k}\Big] &= \mathbb{E}\Big[e^{\lambda \sum_{k=1}^{n-1} D_k}\ \mathbb{E}[e^{\lambda D_n} \mid \mathcal{F}_{n-1}]\Big] \\
&\leq \mathbb{E}\Big[e^{\lambda \sum_{k=1}^{n-1} D_k}\, e^{\lambda^2 \nu_n^2/2}\Big] \\
&\leq \prod_{k=1}^n e^{\lambda^2 \nu_k^2/2} && \text{(by iterating)} \\
&= e^{\lambda^2 \sum_{k=1}^n \nu_k^2/2},
\end{aligned}$$

which shows that the sum is sub-exponential, with $\nu^2 = \sum_{k=1}^n \nu_k^2$. Part (b) then follows from the tail bound for sub-exponential random variables. ■

1.1 The Azuma–Hoeffding inequality

Corollary 1 (Azuma–Hoeffding)**.** *Let $\{(D_k, \mathcal{F}_k)\}$ be an MDS such that $D_k \in [a_k, b_k]$ almost surely for all $k \in [n] = \{1, \dots, n\}$. Then for all $t \geq 0$,*

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

This is similar to Hoeffding's inequality.

Proof. Bounded random variables are sub-Gaussian with parameter $(b_k - a_k)/2$, so

$$\mathbb{E}[e^{\lambda D_k} \mid \mathcal{F}_{k-1}] \leq e^{\lambda^2 (b_k - a_k)^2/8} \qquad \forall\, \lambda.$$

So Theorem 7 applies with $\nu_k = (b_k - a_k)/2$ and $\alpha_k = 0$: the sum $\sum_k D_k$ is sub-Gaussian with parameter $\nu^2 = \sum_k (b_k - a_k)^2/4$, and the sub-Gaussian tail bound gives $2\exp(-t^2/(2\nu^2)) = 2\exp(-2t^2/\sum_k (b_k - a_k)^2)$, which implies the stated bound. ∎

2 Bounded Differences Inequality

Recall from Doob's construction that

$$f(X) - \mathbb{E}\, f(X) = \sum_{k=1}^{n} \Big(\mathbb{E}[f(X) \mid X_1, \dots, X_k] - \mathbb{E}[f(X) \mid X_1, \dots, X_{k-1}]\Big).$$

For $x, x' \in \mathbb{R}^n$, define $x^{\setminus k} \in \mathbb{R}^n$ ("x slash k") by

$$x_j^{\setminus k} = \begin{cases} x_j, & j \neq k, \\ x'_k, & j = k. \end{cases}$$

Definition (Bounded difference property)**.** A function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the *bounded difference property* if there exist $(L_1, \dots, L_n)$ such that for all $k \in [n]$,

$$|f(x) - f(x^{\setminus k})| \leq L_k \qquad \forall\, x, x' \in \mathbb{R}^n. \tag{$\star$}$$

That is, if you perturb one coordinate, the change in f is bounded.

The idea is that f is "stable" in some sense.

Corollary 2 (Bounded differences inequality)**.** *Suppose f satisfies ($\star$) and $X = (X_1, \dots, X_n)$ has independent coordinates. Then*

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

Proof. Construct an MDS via Doob's construction:

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

Define

$$A_k = \inf_x \ \mathbb{E}[f(X) \mid X_1, \ldots, X_{k-1}, X_k = x] - \mathbb{E}[f(X) \mid X_1, \ldots, X_{k-1}],$$
$$B_k = \sup_x \ \mathbb{E}[f(X) \mid X_1, \ldots, X_{k-1}, X_k = x] - \mathbb{E}[f(X) \mid X_1, \ldots, X_{k-1}].$$

Clearly $A_k \le D_k \le B_k$, by definition. By the bounded difference property, we can prove that

$$B_k - A_k \le L_k.$$

Indeed, by independence, the conditional expectation $\mathbb{E}[f(X) \mid X_1, \ldots, X_{k-1}, X_k = x]$ is the average of $f(X_1, \ldots, X_{k-1}, x, X_{k+1}, \ldots, X_n)$ over $X_{k+1}, \ldots, X_n$, and changing x to another value changes f by at most L_k. Now we can just apply Azuma–Hoeffding (Corollary 1).

There is a small incorrectness here: A_k and B_k are random variables, but in Azuma–Hoeffding we require them to be fixed. But we only really need them fixed given $\mathcal{F}_{k-1}$, and A_k and B_k are measurable with respect to $\mathcal{F}_{k-1} = \sigma(X_1, \ldots, X_{k-1})$. ■

The bounded difference property could hold, for example, for Lipschitz functions on a compact set.

3 Covering and Packing

This material is in Chapter 5 of MJW.

Definition (Metric space)**.** A *metric space* is a pair (T, ρ), where $\rho : T \times T \to \mathbb{R}$ satisfies:

(a) it is non-negative: $\rho(\theta, \theta') \ge 0$ for all (θ, θ'), with equality if and only if $\theta = \theta'$;

(b) it is symmetric: $\rho(\theta, \tilde{\theta}) = \rho(\tilde{\theta}, \theta)$ for all $(\theta, \tilde{\theta})$;

(c) it satisfies the triangle inequality: $\rho(\theta, \tilde{\theta}) \le \rho(\theta, \theta') + \rho(\theta', \tilde{\theta})$.

Example 6.

- $\mathbb{R}^d$ with the Euclidean metric $\|\cdot\|_2$:

$$\rho(\theta, \tilde{\theta}) = \sqrt{\sum_{i=1}^{d} (\theta_i - \tilde{\theta}_i)^2}.$$

- $\{0, 1\}^d$ with the rescaled Hamming distance:

$$\rho_H(\theta, \tilde{\theta}) = \frac{1}{d} \sum_{i=1}^{d} \mathbf{1}(\theta_i \ne \tilde{\theta}_i).$$

- Function spaces: $L^2_\mu[0,1]$ with
$$\rho(f,g) = \|f-g\|_2 = \left[\int_0^1 (f(x)-g(x))^2 \, d\mu(x)\right]^{1/2}.$$
- $C[0,1]$ with
$$\rho(f,g) = \|f-g\|_\infty = \sup_{x\in[0,1]} |f(x)-g(x)|.$$

Given (T,ρ), how massive is the space?

Definition (Covering number)**.** A *δ-cover* of a set T with respect to ρ is a set $\{\theta_1,\dots,\theta_N\} \subset T$ such that for every $\theta \in T$ there exists $i \in [N]$ with $\rho(\theta,\theta_i) \le \delta$.

The *δ-covering number* is the smallest cardinality of a δ-cover, and is denoted by $N(\delta,T,\rho)$.

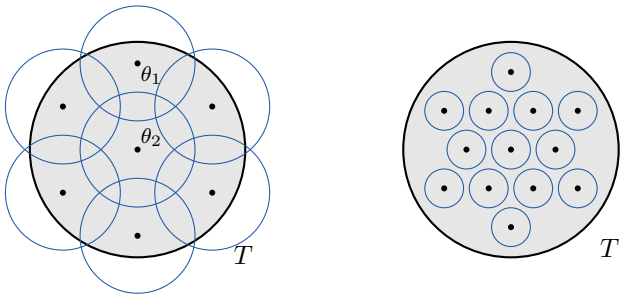
For some sets T, such a finite δ-cover may not exist. In this class, we will only look at *totally bounded* sets T, for which a finite δ-cover always exists for all $\delta > 0$. In $\mathbb{R}^n$, totally bounded is equivalent to bounded.

If $0 \le \delta \le \delta'$, then $N(\delta) \ge N(\delta')$.

In probability and statistics, $\log N(\delta,T)$ shows up more. It is called the *metric entropy*.

Definition (Packing number)**.** A *δ-packing* of T with respect to ρ is a set $\{\theta_1,\dots,\theta_M\} \subset T$ such that $\rho(\theta_i,\theta_j) > \delta$ for all $i \ne j \in [M]$.

The *δ-packing number* is the largest cardinality of a δ-packing, denoted $M(\delta,T,\rho)$.



Covering Packing

Figure 5: Left: a cover of T, whose balls around $\theta_1,\theta_2,\dots$ together contain T. Right: a packing of T, whose points are far enough apart that small balls around them do not overlap.

Why do we study this? Suppose $Y = \mu + \xi$, where $\xi \sim N(0,\sigma^2 I)$ and $\mu \in K$ for a convex and compact set K. Let
$$r^2 = \inf_{\hat\mu} \sup_{\mu\in K} \mathbb{E}\,\|\hat\mu - \mu\|_2^2 .$$
It turns out that $r \asymp \varepsilon$, where ε satisfies
$$\underbrace{\log M(\varepsilon, K, \|\cdot\|_2)}_{\text{packing entropy}} \asymp \frac{\varepsilon^2}{\sigma^2}.$$
This is roughly correct for many sets K.

Lemma 1. *For all $\delta > 0$, the packing and covering numbers satisfy*

$$M(2\delta, T, \rho) \leq N(\delta, T, \rho) \leq M(\delta, T, \rho).$$

36-709: Advanced Statistical Theory I Spring 2023

Lecture 6 — February 2, 2023

Prof. Matey Neykov *Scribe: Fan Pu Zeng*

Homework note. In Problem 2 of the homework, assume that X and Y have mean 0.

1 Packing and Covering Numbers

Recall the lemma from last time.

Lemma 2. *For all $\delta > 0$, we have*

$$M(2\delta, T, \rho) \overset{(a)}{\leq} N(\delta, T, \rho) \overset{(b)}{\leq} M(\delta, T, \rho),$$

where M is the packing number and N is the covering number.

Proof. (b) Suppose $E = \{\theta_1, \ldots, \theta_M\}$ is a maximal δ-packing set. Take $\theta \in T \setminus E$. Then there exists $i \in [M]$ such that $\rho(\theta, \theta_i) \leq \delta$. Indeed, if no such i exists, then $\{\theta_1, \ldots, \theta_M, \theta\}$ forms a δ-packing, which contradicts the maximality of E. So E is a δ-cover, and therefore $M \geq N(\delta, T, \rho)$, which shows (b).

(a) Suppose there exist a 2δ-packing set $\{\theta_1, \ldots, \theta_M\}$ and a δ-covering set $\{x_1, \ldots, x_N\}$ such that $M \geq N + 1$. We want to show that this is impossible.

Since there are more points in the 2δ-packing set than balls in the cover, by the pigeonhole principle there are two points $\theta_i, \theta_j \in B(x_k, \delta)$ for some k. Then

$$\begin{aligned} 2\delta < \rho(\theta_i, \theta_j) &\leq \rho(\theta_i, x_k) + \rho(x_k, \theta_j) \\ &\leq \delta + \delta = 2\delta, \end{aligned}$$

which is a contradiction. This shows (a). ∎

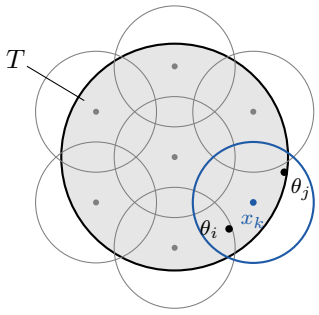


Figure 6: If a 2δ-packing had more points than a δ-cover has balls, two packing points θ_i, θ_j would lie in the same ball $B(x_k, \delta)$.

2 Volumetric Argument

Theorem 8 (Volumetric argument)**.** *Let $\Theta \subset \mathbb{R}^d$ and let $\|\cdot\|$ be an arbitrary norm. Then*

$$\Big(\frac{1}{\delta}\Big)^d \frac{\mathrm{Vol}(\Theta)}{\mathrm{Vol}(B)} \overset{(a)}{\leq} N(\delta, \Theta, \|\cdot\|) \leq M(\delta, \Theta, \|\cdot\|) \overset{(b)}{\leq} \frac{\mathrm{Vol}\big(\Theta + \frac{\delta}{2}B\big)}{\mathrm{Vol}\big(\frac{\delta}{2}B\big)} \overset{(c)}{\leq} \Big(\frac{3}{\delta}\Big)^d \frac{\mathrm{Vol}(\Theta)}{\mathrm{Vol}(B)},$$

where inequality (c) requires that Θ is convex and $\delta B \subset \Theta$. Here:

- *$B = \{x : \|x\| \leq 1\}$ is the unit ball of $\|\cdot\|$, so $\mathrm{Vol}(B)$ is the volume of the unit ball;*
- *"+" for sets means the Minkowski sum, $A + B = \{a + b : a \in A, b \in B\}$, and $cA = \{ca : a \in A\}$;*
- *volume means Lebesgue measure on $\mathbb{R}^d$.*

Proof. (a) Let $\{\theta_1, \ldots, \theta_{N(\delta)}\}$ be a minimal δ-cover. We know that $\Theta \subseteq \bigcup_{i=1}^{N(\delta)} B(\theta_i, \delta)$, so

$$\begin{aligned}
\mathrm{Vol}(\Theta) &\leq \mathrm{Vol}\Big(\bigcup_{i=1}^{N(\delta)} B(\theta_i, \delta)\Big) \\
&\leq \sum_{i=1}^{N(\delta)} \mathrm{Vol}\big(B(\theta_i, \delta)\big) \\
&= N(\delta)\, \delta^d\, \mathrm{Vol}(B).
\end{aligned}$$

Therefore

$$N(\delta) \geq \frac{\mathrm{Vol}(\Theta)}{\mathrm{Vol}(B)} \Big(\frac{1}{\delta}\Big)^d.$$

(b) Let $\{\theta_i\}_{i=1}^{M(\delta)}$ be a maximal δ-packing. For a δ-packing, the balls $B(\theta_i, \delta/2)$ are disjoint, and

$$\bigcup_{i=1}^{M(\delta)} B(\theta_i, \delta/2) \subseteq \Theta + \frac{\delta}{2}B.$$

So

$$\mathrm{Vol}\Big(\bigcup_{i=1}^{M(\delta)} B(\theta_i, \delta/2)\Big) \leq \mathrm{Vol}\Big(\Theta + \frac{\delta}{2}B\Big),$$

and by disjointness the left-hand side is

$$\sum_{i=1}^{M(\delta)} \mathrm{Vol}\big(B(\theta_i, \delta/2)\big) = M(\delta)\, \mathrm{Vol}\Big(\frac{\delta}{2}B\Big).$$

So

$$M(\delta) \leq \frac{\mathrm{Vol}\big(\Theta + \frac{\delta}{2}B\big)}{\mathrm{Vol}\big(\frac{\delta}{2}B\big)}.$$

(c) We use two facts.

(1) When $\delta B \subset \Theta$, we have $\Theta + \frac{\delta}{2}B \subset \Theta + \frac{1}{2}\Theta$.

(2) When Θ is convex, $\Theta + \frac{1}{2}\Theta = \frac{3}{2}\Theta$.

Together, (1) and (2) give $\Theta + \frac{\delta}{2}B \subseteq \frac{3}{2}\Theta$. Hence

$$\frac{\mathrm{Vol}(\Theta + \frac{\delta}{2}B)}{\mathrm{Vol}(\frac{\delta}{2}B)} \leq \frac{\mathrm{Vol}(\frac{3}{2}\Theta)}{\mathrm{Vol}(\frac{\delta}{2}B)} = \frac{(\frac{3}{2})^d \mathrm{Vol}(\Theta)}{(\frac{\delta}{2})^d \mathrm{Vol}(B)} = \Big(\frac{3}{\delta}\Big)^d \frac{\mathrm{Vol}(\Theta)}{\mathrm{Vol}(B)}.$$

Proof of (1). Let $z \in \Theta + \frac{\delta}{2}B$, so $z = x + y$ with $x \in \Theta$ and $y \in \frac{\delta}{2}B \subset \frac{1}{2}\Theta$. Then $z = x + y \in \Theta + \frac{1}{2}\Theta$.

Proof of (2). Take $\theta \in \frac{3}{2}\Theta$ and write $\theta = \frac{1}{3}\theta + \frac{2}{3}\theta$. Here $\frac{1}{3}\theta \in \frac{1}{2}\Theta$ and $\frac{2}{3}\theta \in \Theta$, so $\theta \in \Theta + \frac{1}{2}\Theta$. This shows $\frac{3}{2}\Theta \subseteq \Theta + \frac{1}{2}\Theta$.

Conversely, let $x \in \Theta + \frac{1}{2}\Theta$, and write $x = y + \frac{z}{2}$ where $y, z \in \Theta$. Then

$$\frac{2}{3}x = \frac{2}{3}y + \frac{1}{3}z \in \Theta$$

by convexity of Θ, so $x \in \frac{3}{2}\Theta$. ■

So if $\Theta = B$, then

$$\log M(\delta, B, \|\cdot\|) \asymp d \log \frac{1}{\delta}.$$

3 Example: Lipschitz Functions on the Unit Interval

This is Example 5.10 in MJW. Consider the class of Lipschitz functions

$$\mathcal{F}_L = \{g : [0,1] \to \mathbb{R} \mid g(0) = 0,\ |g(x) - g(x')| \leq L|x - x'|\ \ \forall\, x, x' \in [0,1]\}.$$

Proposition 5. For small δ,

$$\log N_\infty(\delta, \mathcal{F}_L) \asymp \frac{L}{\delta},$$

where N_∞ is the covering number in the sup norm $\|\cdot\|_\infty$.

For $\varepsilon > 0$, define $M = \lfloor 1/\varepsilon \rfloor$ and consider the points

$$x_i = (i-1)\varepsilon, \quad i = 1, \ldots, M, \qquad x_{M+1} = M\varepsilon \leq 1.$$

Definition. Define $\phi : \mathbb{R} \to \mathbb{R}_+$ by

$$\phi(u) = \begin{cases} 0, & u < 0, \\ u, & 0 \leq u \leq 1, \\ 1, & u > 1. \end{cases}$$

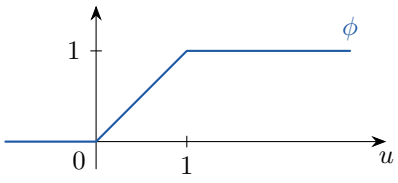


Figure 7: The ramp function ϕ.

We use ϕ to build exponentially many functions. For every $\beta \in \{-1, +1\}^M$, define $f_\beta : [0, 1] \to [-L, L]$ via

$$f_\beta(y) = \sum_{i=1}^{M} \beta_i L \phi\Big(\frac{y - x_i}{\varepsilon}\Big)\varepsilon.$$

One can check that f_β is L-Lipschitz, and $f_\beta(0) = 0$, so $f_\beta \in \mathcal{F}_L$.

For all $\beta \neq \beta'$,

$$\sup_{y \in [0,1]} |f_\beta(y) - f_{\beta'}(y)| \geq 2L\varepsilon,$$

which means the functions are $2L\varepsilon$-separated.

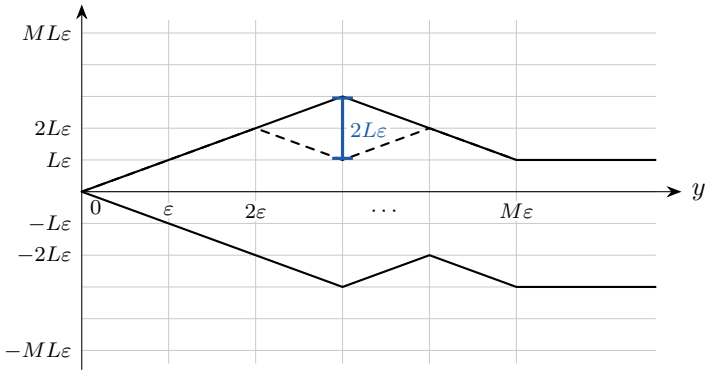


Figure 8: Three functions f_β with $M = 5$. Each is piecewise linear with slope $\pm L$ on each interval $[x_i, x_{i+1}]$, so it is L-Lipschitz, and it goes flat after $M\varepsilon$. Two functions that agree up to some point branch out there and differ by $2L\varepsilon$ one step later.

For example, say $y \in (x_2, x_3)$. Then

$$f_\beta(y) = \beta_1 \phi\Big(\frac{y - x_1}{\varepsilon}\Big)\varepsilon L + \beta_2 \phi\Big(\frac{y - x_2}{\varepsilon}\Big)\varepsilon L,$$

and the other terms vanish since $y < x_i$ for $i \geq 3$. Here $\frac{y - x_1}{\varepsilon} \geq 1$, so the first term equals $\beta_1 \varepsilon L$, and on (x_2, x_3) the function moves with slope $\beta_2 L$.

So we see that any two functions from this family must diverge by $2L\varepsilon$ at some place. Since we have $2^M = 2^{\lfloor 1/\varepsilon \rfloor}$ such functions, set $\varepsilon = \delta/L$, so that $2L\varepsilon = 2\delta$. Then

$$\log 2^M \gtrsim \lfloor 1/\varepsilon \rfloor = \lfloor L/\delta \rfloor \gtrsim L/\delta,$$

where the constant in the first step depends on the base of the logarithm. So there is a packing of $\mathcal{F}_L$ at scale about 2δ in the sup norm with 2^M elements, and by Lemma 2 this gives the lower bound $\log N_\infty(\delta, \mathcal{F}_L) \gtrsim L/\delta$.

Lecture 7 — February 7, 2023

Prof. Matey Neykov *Scribe: Fan Pu Zeng*

1 Gaussian and Rademacher Complexity

Given a process $\{X_\theta, \theta \in T\}$, how does this process depend on the structure of the set T?

Say $T \subset \mathbb{R}^d$, and define $\{G_\theta : \theta \in T\}$ by

$$G_\theta = \langle w, \theta \rangle = \sum_{i=1}^{d} w_i \theta_i, \qquad w_i \overset{\text{i.i.d.}}{\sim} N(0,1).$$

This is the *canonical Gaussian process* on the set T.

Definition (Gaussian complexity)**.** The *Gaussian complexity* of T is

$$\mathcal{G}(T) = \mathbb{E} \sup_{\theta \in T} \langle w, \theta \rangle .$$

So we have the set T and its Gaussian complexity $\mathcal{G}(T)$. If we understand one, can we understand something about the other?

Since $w/\|w\|_2$ is independent of $\|w\|_2$,

$$\mathcal{G}(T) = \underbrace{\mathbb{E} \sup_{\theta \in T} \left\langle \theta, \frac{w}{\|w\|_2} \right\rangle}_{\frac{w}{\|w\|_2} \sim U(S^{d-1})} \mathbb{E} \|w\|_2 ,$$

where $U(S^{d-1})$ is the uniform distribution on the unit sphere. Here

$$(1 - o(1))\sqrt{d} \le \mathbb{E} \|w\|_2 \le \sqrt{\mathbb{E} \|w\|_2^2} = \sqrt{d},$$

where the upper bound is Jensen's inequality.

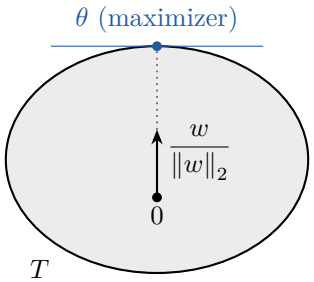


Figure 9: For a direction $w/\|w\|_2$, the supremum of $\langle \theta, w/\|w\|_2 \rangle$ over T is attained at the point θ where a line orthogonal to $w/\|w\|_2$ touches T.

If we substitute the w_i with $\varepsilon_i \overset{\text{i.i.d.}}{\sim} U(\{-1,1\})$, we get

$$R_\theta = \langle \theta, \varepsilon \rangle = \sum_{i=1}^{d} \theta_i \varepsilon_i,$$

which is called the *Rademacher process*.

Definition (Rademacher complexity)**.** The *Rademacher complexity* of T is

$$\mathcal{R}(T) = \mathbb{E} \sup_{\theta \in T} \langle \theta, \varepsilon \rangle .$$

Proposition 6. For any $T \subset \mathbb{R}^d$, we always have

$$\mathcal{R}(T) \le \sqrt{\frac{\pi}{2}}\, \mathcal{G}(T), \tag{1}$$

$$\mathcal{G}(T) \le 2\sqrt{\log d}\, \mathcal{R}(T), \tag{2}$$

where d is the dimension.

We usually talk about bounded T.

Inequality (1) is tight: take $T = [-1, 1]$, and check that $\mathcal{R}(T) = \sqrt{\pi/2}\, \mathcal{G}(T)$.

Proof of (1). We have $\mathbb{E}|w_i| = \sqrt{2/\pi}$, so multiplying and dividing by it,

$$\begin{aligned} \mathcal{R}(T) &= \sqrt{\frac{\pi}{2}}\ \mathop{\mathbb{E}}_{\varepsilon} \sup_{\theta \in T} \sum_{i=1}^{d} \varepsilon_i \theta_i\, \mathbb{E}|w_i| \\ &\le \sqrt{\frac{\pi}{2}}\ \mathop{\mathbb{E}}_{\varepsilon} \mathop{\mathbb{E}}_{w} \sup_{\theta \in T} \sum_{i=1}^{d} \varepsilon_i \theta_i |w_i|, \end{aligned}$$

by Jensen's inequality, since the supremum over θ is a convex function of the $|w_i|$. Since $\varepsilon_i |w_i| \overset{d}{=} w_i$ (equal in distribution), this is

$$\sqrt{\frac{\pi}{2}}\ \mathop{\mathbb{E}}_{w} \sup_{\theta \in T} \sum_{i=1}^{d} \theta_i w_i = \sqrt{\frac{\pi}{2}}\, \mathcal{G}(T). \qquad \blacksquare$$

Example 7 (Euclidean ball)**.** Let $T = B_2^d = \{x \in \mathbb{R}^d : \|x\|_2 \le 1\}$ be the unit ball in the Euclidean norm. Then

$$\mathcal{R}(B_2^d) = \mathbb{E} \sup_{\theta \in B_2^d} \langle \varepsilon, \theta \rangle = \mathbb{E}\, \|\varepsilon\|_2 = \sqrt{d},$$

using Cauchy–Schwarz, with equality achieved when we plug in $\theta = \varepsilon / \|\varepsilon\|_2$, which achieves the supremum. By the same argument,

$$\mathcal{G}(B_2^d) = \mathbb{E}\, \|w\|_2 \le \sqrt{\mathbb{E}\, \|w\|_2^2} = \sqrt{d}$$

by Jensen's inequality, since $\|w\|_2^2$ is a sum of d χ^2 random variables. It is a fact that $\mathbb{E}\, \|w\|_2 \ge (1 - o(1))\sqrt{d}$. So

$$\mathcal{R}(B_2^d) \asymp \mathcal{G}(B_2^d).$$

Example 8 (ℓ_1-ball)**.** Let $B_1^d = \{x \in \mathbb{R}^d : \|x\|_1 \le 1\}$. Since ℓ_1 and ℓ_∞ are dual norms,

$$\mathcal{R}(B_1^d) = \mathbb{E} \sup_{\theta \in B_1^d} \langle \varepsilon, \theta \rangle = \mathbb{E} \|\varepsilon\|_\infty = 1.$$

Indeed,

$$\sum_{i=1}^d \varepsilon_i \theta_i \le \max_{i \in [d]} |\varepsilon_i| \sum_{i=1}^d |\theta_i|, \qquad \text{and} \quad \sum_{i=1}^d |\theta_i| \le 1.$$

On the other hand,

$$\mathcal{G}(B_1^d) = \mathbb{E} \|w\|_\infty \asymp \sqrt{2 \log d};$$

from Exercise 2.11 in MJW,

$$\frac{\mathbb{E} \|w\|_\infty}{\sqrt{2 \log d}} = 1 \pm o(1).$$

So this also shows that $\mathcal{G}(T) \le 2\sqrt{\log d}\, \mathcal{R}(T)$ is tight.

2 Function Classes

Let $\mathcal{F}$ be a function class with domain $\mathcal{X}$, and let $x_1^n = \{x_1, \dots, x_n\}$ be a collection of "design points". We observe

$$y_i = f(x_i) + \varepsilon_i,$$

where $f \in \mathcal{F}$. We will want to evaluate the Gaussian complexity of $\mathcal{F}$. For $f : \mathcal{X} \to \mathbb{R}$, define

$$\mathcal{F}(x_1^n) = \big\{ \big(f(x_1), \dots, f(x_n)\big) : f \in \mathcal{F} \big\} \subseteq \mathbb{R}^n.$$

We will look at the Gaussian complexity of

$$\frac{\mathcal{F}(x_1^n)}{n}, \qquad \frac{\mathcal{F}(x_1^n)}{\sqrt{n}}.$$

Define

$$\|f - g\|_n = \sqrt{\frac{\sum_{i=1}^n \big(f(x_i) - g(x_i)\big)^2}{n}}.$$

When we measure how good an estimate $\hat{f}$ is, we want to evaluate

$$\frac{\sum_{i=1}^n \big(\hat{f}(x_i) - f(x_i)\big)^2}{n}.$$

Suppose that all functions $f \in \mathcal{F}$ satisfy $\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)| \le b$, which implies $\|f\|_n \le b$. (These are assumptions on $\mathcal{F}$, not on the design points.) Then, by scaling,

$$\begin{aligned}
\mathcal{G}\Big(\frac{\mathcal{F}(x_1^n)}{n}\Big) &= \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^n \frac{w_i}{\sqrt{n}} \frac{f(x_i)}{\sqrt{n}} \\
&\le \frac{\mathbb{E} \|w\|_2}{\sqrt{n}} \sup_{f \in \mathcal{F}} \|f\|_n \qquad \text{(Cauchy–Schwarz)} \\
&\le 1 \cdot b = b,
\end{aligned}$$

since

$$\|f\|_n = \sqrt{\frac{\sum_{i=1}^n f^2(x_i)}{n}} \le \sqrt{\frac{nb^2}{n}} = b.$$

This is a trivial bound, since we only used Cauchy–Schwarz.

Now we try to relate the Gaussian complexity to the covering and packing numbers of a set. Usually it is very difficult to find the Gaussian complexity, but for covering and packing we have more bounds.

3 Sub-Gaussian Processes

Definition (Sub-Gaussian process)**.** A collection of mean-zero random variables $\{X_\theta : \theta \in T\}$ is called a *sub-Gaussian process* with respect to a metric ρ_X on T if

$$\mathbb{E}\, e^{\lambda(X_\theta - X_{\tilde\theta})} \le e^{\frac{\lambda^2 \rho_X^2(\theta,\tilde\theta)}{2}} \qquad \forall\, \theta, \tilde\theta \in T,\ \lambda \in \mathbb{R}.$$

Here $\rho_X^2(\theta, \tilde\theta)$ acts as a proxy for the variance.

This implies

$$\mathbb{P}\big(|X_\theta - X_{\tilde\theta}| \ge t\big) \le 2e^{-\frac{t^2}{2\rho_X^2(\theta,\tilde\theta)}}.$$

Remark 7. The canonical Gaussian process and the Rademacher process are examples of sub-Gaussian processes with respect to $\rho_X(\theta, \tilde\theta) = \left\|\theta - \tilde\theta\right\|_2$.

The X_θ's are not necessarily i.i.d. For example, $X_\theta = \langle w, \theta\rangle$ and $X_{\theta'} = \langle w, \theta'\rangle$ are not i.i.d., since they use the same w.

Notation:

- $N_X(\delta, T)$ is the δ-covering number of T with respect to ρ_X;
- $N_2(\delta, T)$ is the δ-covering number of T with respect to $\|\cdot\|_2$ (for the Gaussian process).

We assume that the diameter

$$D = \sup_{\theta,\tilde\theta \in T} \rho_X(\theta, \tilde\theta) < \infty.$$

Proposition 7 (One-step discretization)**.** Let $\{X_\theta, \theta \in T\}$ be a zero-mean sub-Gaussian process with respect to ρ_X. Then for all $\delta \in [0, D]$ such that $N_X(\delta, T) \ge 10$, we have

$$\mathbb{E}\Big[\sup_{\theta,\tilde\theta\in T}(X_\theta - X_{\tilde\theta})\Big] \le 2\,\mathbb{E}\Big[\sup_{\substack{\gamma,\gamma'\in T\\ \rho_X(\gamma,\gamma')\le\delta}}(X_\gamma - X_{\gamma'})\Big] + 4\sqrt{D^2 \log N_X(\delta, T)}.$$

The left-hand side is the biggest possible difference.

MJW calls the first term on the right the "approximation error". It goes to 0 as $\delta \to 0$: as the covering radius δ gets smaller, $\rho_X(\gamma, \gamma')$ also becomes small, and similarly $X_\gamma - X_{\gamma'}$.

Note that if we fix $\theta_0 \in T$, then $\mathbb{E}\, X_{\theta_0} = 0$, so

$$\mathbb{E} \sup_{\theta \in T} X_\theta = \mathbb{E} \sup_{\theta \in T} (X_\theta - X_{\theta_0}) \leq \mathbb{E} \sup_{\theta, \tilde{\theta} \in T} (X_\theta - X_{\tilde{\theta}}).$$

Proof. Fix $\delta > 0$. Let $N = N_X(\delta, T)$, and take a minimal δ-cover $\{\theta_1, \ldots, \theta_N\}$ of T. By the definition of a δ-covering, for every $\theta \in T$ there exists θ_i such that $\rho_X(\theta, \theta_i) \leq \delta$. Then

$$\begin{aligned} X_\theta - X_{\theta_1} &= (X_\theta - X_{\theta_i}) + (X_{\theta_i} - X_{\theta_1}) \\ &\leq \underbrace{\sup_{\substack{\gamma, \gamma' \in T \\ \rho_X(\gamma, \gamma') \leq \delta}} (X_\gamma - X_{\gamma'}) + \max_{i \in [N]} |X_{\theta_i} - X_{\theta_1}|}_{L}. \end{aligned}$$

Making the same argument for $X_{\theta_1} - X_{\tilde{\theta}}$ gives $X_{\theta_1} - X_{\tilde{\theta}} \leq L$. So

$$X_\theta - X_{\tilde{\theta}} \leq 2L.$$

Taking the supremum over $\theta, \tilde{\theta}$ and then expectations,

$$\mathbb{E} \sup_{\theta, \tilde{\theta} \in T} (X_\theta - X_{\tilde{\theta}}) \leq 2\, \mathbb{E} \sup_{\substack{\gamma, \gamma' \in T \\ \rho_X(\gamma, \gamma') \leq \delta}} (X_\gamma - X_{\gamma'}) + 2\, \mathbb{E} \max_{i \in [N]} |X_{\theta_i} - X_{\theta_1}|.$$

Each $X_{\theta_i} - X_{\theta_1}$ is sub-Gaussian with parameter $\rho_X(\theta_i, \theta_1) \leq D$, so the bound on the expected maximum of sub-Gaussian random variables gives $\mathbb{E} \max_{i \in [N]} |X_{\theta_i} - X_{\theta_1}| \leq 2\sqrt{D^2 \log N}$, which proves the claim. ∎

36-709: Advanced Statistical Theory I Spring 2023

Lecture 8 — February 9, 2023

Prof. Matey Neykov *Scribe: Fan Pu Zeng*

1 One-Step Discretization (continued)

We continue the proof of the one-step discretization bound from last lecture. Let $\{\theta_1, \ldots, \theta_N\}$ be a minimal δ-cover of T with respect to ρ_X, and for $\theta \in T$ let

$$\theta_i = \arg\min_{j \in [N]} \rho_X(\theta, \theta_j).$$

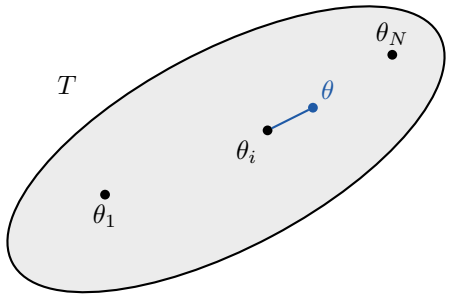


Figure 10: Each $\theta \in T$ is approximated by its nearest point θ_i of the δ-cover $\{\theta_1, \ldots, \theta_N\}$.

As shown last time, $X_\theta - X_{\theta_1}$ is at most

$$L = \sup_{\substack{\gamma, \gamma' \in T \\ \rho_X(\gamma, \gamma') \le \delta}} (X_\gamma - X_{\gamma'}) + \max_{i \in [N]} |X_{\theta_i} - X_{\theta_1}|.$$

Repeating this for a point $X_{\tilde{\theta}}$ gives $X_{\theta_1} - X_{\tilde{\theta}} \le L$, and so

$$\mathbb{E} \sup_{\theta, \tilde{\theta} \in T} (X_\theta - X_{\tilde{\theta}}) \le 2\,\mathbb{E} \sup_{\substack{\gamma, \gamma' \in T \\ \rho_X(\gamma, \gamma') \le \delta}} (X_\gamma - X_{\gamma'}) + 2\,\mathbb{E} \max_{i \in [N]} |\underbrace{X_{\theta_i} - X_{\theta_1}}_{\text{sub-Gaussian}}|.$$

Each $X_{\theta_i} - X_{\theta_1}$ is a sub-Gaussian random variable with parameter

$$\rho_X(\theta_i, \theta_1) \le D = \operatorname{diam}_X(T).$$

By the maximal property of sub-Gaussian random variables,

$$\mathbb{E} \max_{i \in [N]} |X_{\theta_i} - X_{\theta_1}| \le \sqrt{2D^2 \log 2N} \le 2\sqrt{D^2 \log N}$$

when N is large ($N \ge 2$). Substituting this back gives the bound $2\,\mathbb{E} \sup_{\rho_X(\gamma, \gamma') \le \delta} (X_\gamma - X_{\gamma'}) + 4\sqrt{D^2 \log N_X(\delta, T)}$ of the proposition.

The gist of the proof: when you have a sub-Gaussian process, if you can cover the set, you can approximate the supremum by that cover.

2 The Canonical Gaussian Process

For the canonical Gaussian process on T, the metric ρ_X is $\|\cdot\|_2$, and

$$\mathcal{G}(T) = \mathbb{E} \sup_{\theta \in T} \langle \xi, \theta \rangle, \qquad \xi \sim N(0, I).$$

Proposition 8. For the canonical Gaussian process on T,

$$\mathcal{G}(T) \le \min_{\delta \in [0, D]} \Big\{ \mathcal{G}(\tilde{T}(\delta)) + 2\sqrt{D^2 \log N_2(\delta, T)} \Big\},$$

where $D = \mathrm{diam}_2(T)$, $N_2(\delta, T)$ is the covering number of T in $\|\cdot\|_2$, and

$$\tilde{T}(\delta) = \{\gamma - \gamma' : \gamma, \gamma' \in T,\ \|\gamma - \gamma'\|_2 \le \delta\}.$$

This is the same as the previous result, except for a factor of 2: here we are only interested in the supremum over one parameter.

Usually we only care about the Gaussian complexity (or Gaussian width) up to constant factors.

We can bound the first term crudely. Since $\|\gamma - \gamma'\|_2 \le \delta$ for every element of $\tilde{T}(\delta)$,

$$\begin{aligned}
\mathcal{G}(\tilde{T}(\delta)) &= \mathbb{E} \sup_{\theta \in \tilde{T}(\delta)} \langle \xi, \theta \rangle \\
&\le \mathbb{E} \|\xi\|_2 \, \delta && \text{(Cauchy–Schwarz)} \\
&\le \sqrt{\mathbb{E} \|\xi\|_2^2} \, \delta && \text{(Jensen's inequality)} \\
&= \sqrt{d} \, \delta && \text{(expectation of a } \chi^2\text{)}.
\end{aligned}$$

So

$$\mathcal{G}(T) \le \min_{\delta \in [0, D]} \Big\{ \sqrt{d} \, \delta + 2\sqrt{D^2 \log N_2(\delta, T)} \Big\}.$$

Example 9 (Lipschitz functions)**.** Consider the class of Lipschitz functions

$$\mathcal{F}_L = \{g : [0,1] \to \mathbb{R} \mid g(0) = 0,\ |g(x) - g(x')| \le L|x - x'|\ \ \forall x, x' \in [0,1]\}.$$

For $\delta \in (0, \delta_0)$, we showed in the homework that

$$\log N_\infty(\delta, \mathcal{F}_L) \le \frac{cL}{\delta},$$

where N_∞ is the covering number in the sup norm.

Recall that

$$\mathcal{F}_L(x_1^n) = \{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}_L\}, \qquad x_1^n = (x_1, \ldots, x_n).$$

Observe that for $f, g \in \mathcal{F}_L$,

$$\sup_{x \in [0,1]} |f(x) - g(x)| \le \delta \implies \sqrt{\frac{\sum_{i=1}^n (f(x_i) - g(x_i))^2}{n}} \le \delta.$$

Therefore

$$N_2\Big(\delta, \frac{\mathcal{F}_L(x_1^n)}{\sqrt{n}}\Big) \le N_\infty(\delta, \mathcal{F}_L),$$

because a cover in the sup norm on the right-hand side is also a cover on the left-hand side, by the above.

For the diameter,

$$\begin{aligned} D &= \sup_{f,g\in\mathcal{F}_L} \sqrt{\frac{\sum_{i=1}^n \big(f(x_i) - g(x_i)\big)^2}{n}} \\ &\le \sup_{f,g\in\mathcal{F}_L} \sqrt{\frac{\sum_{i=1}^n 2f^2(x_i) + 2g^2(x_i)}{n}} \qquad\qquad (\text{by } (a-b)^2 \le 2(a^2+b^2)) \\ &\le 2L, \end{aligned}$$

because $\sup_{x\in[0,1]}|f(x)| \le L$ for all $f \in \mathcal{F}_L$. This follows from the Lipschitz property, since $f(0) = 0$ and

$$|f(x) - f(0)| \le Lx \le L.$$

So

$$\boxed{D \le 2L.}$$

Since the Gaussian width is linear,

$$\mathcal{G}\Big(\frac{\mathcal{F}_L(x_1^n)}{\sqrt{n}\sqrt{n}}\Big) = \frac{1}{\sqrt{n}}\,\mathcal{G}\Big(\frac{\mathcal{F}_L(x_1^n)}{\sqrt{n}}\Big).$$

Using the previous results (with $d = n$ and $D \le 2L$),

$$\mathcal{G}\Big(\frac{\mathcal{F}_L(x_1^n)}{n}\Big) \le \frac{1}{\sqrt{n}} \min_{\delta\in[0,D]} \Big\{\delta\sqrt{n} + 4L\sqrt{\frac{cL}{\delta}}\Big\}.$$

Taking $\delta \asymp n^{-1/3}$, this is $\delta + \frac{4L\sqrt{cL}}{\sqrt{n\delta}} \lesssim n^{-1/3}$. So

$$\boxed{\mathcal{G}\Big(\frac{\mathcal{F}_L(x_1^n)}{n}\Big) \lesssim n^{-1/3}.}$$

3 Chaining and the Dudley Integral

Definition (Dudley integral)**.** For a fixed set T, let

$$J(\delta, D) = \int_\delta^D \sqrt{\log N_X(u, T)}\, du,$$

where D is the diameter and $N_X(u, T)$ is the covering number.

Theorem 9 (Dudley entropy integral)**.** *Let $\{X_\theta : \theta \in T\}$ be a mean-zero sub-Gaussian process with respect to a pseudometric ρ_X (that is, $\rho_X(\theta, \theta') = 0$ is allowed even if $\theta \neq \theta'$). Then for all $\delta \in [0, D]$,*

$$\begin{aligned}\mathbb{E} \sup_{\theta \in T} X_\theta &\leq \mathbb{E} \sup_{\theta, \tilde{\theta} \in T} (X_\theta - X_{\tilde{\theta}}) \\ &\leq \mathbb{E} \sup_{\substack{\gamma, \gamma' \in T \\ \rho_X(\gamma, \gamma') \leq \delta}} (X_\gamma - X_{\gamma'}) + 64\sqrt{2}\, J\Big(\frac{\delta}{4}, D\Big).\end{aligned}$$

The first inequality is from last time.

Proof. As before,

$$\sup_{\theta, \tilde{\theta} \in T} (X_\theta - X_{\tilde{\theta}}) \leq 2 \sup_{\substack{\gamma, \gamma' \in T \\ \rho_X(\gamma, \gamma') \leq \delta}} (X_\gamma - X_{\gamma'}) + 2 \max_{i \in [N]} |X_{\theta_i} - X_{\theta_1}|,$$

where $U = \{\theta_1, \ldots, \theta_N\}$ is a minimal δ-cover of T and $N = N_X(\delta, T)$. Most of the work goes into the last term.

Let U_m be a minimal $\varepsilon_m = D2^{-m}$ cover of T with respect to ρ_X. These are covers at resolutions $D/2, D/4, \ldots$. Let $N_m = |U_m|$. Since $U \subset T$, U_m covers U too.

Since $|U| = N < \infty$, there exists $L \in \mathbb{N}$ such that

$$|U_{L-1}| < |U|, \quad |U_L| \geq |U| \qquad \text{if} \quad D2^{-(L-1)} \geq \delta. \tag{$*$}$$

Take L to be the maximal L such that $(*)$ holds. Once we reach L we stop constructing sets, and set $U_L = U$.

For all $m = 1, \ldots, L-1$, define $\pi_m : U \to U_m$ by

$$\pi_m(\theta) = \operatorname*{arg\,min}_{\beta \in U_m} \rho_X(\theta, \beta).$$

So $\pi_m(\theta)$ is the best approximation of $\theta \in U$ in U_m. Let

$$\gamma^L = \theta \in U_L, \qquad \gamma^{m-1} = \pi_{m-1}(\theta), \qquad m = L, L-1, \ldots, 2.$$

Goal: have something that telescopes, to bound $2 \max_{i \in [N]} |X_{\theta_i} - X_{\theta_1}|$.

Chaining: thinking of θ as a chain of points,

$$X_\theta - X_{\gamma^1} = \sum_{m=2}^{L} (X_{\gamma^m} - X_{\gamma^{m-1}}).$$

Each term satisfies

$$X_{\gamma^m} - X_{\gamma^{m-1}} = X_{\pi_m(\theta)} - X_{\pi_{m-1}(\theta)} \leq \max_{\beta \in U} |X_{\pi_m(\beta)} - X_{\pi_{m-1}(\beta)}|,$$

so for every $\theta \in U$,

$$|X_\theta - X_{\gamma^1}| \leq \sum_{m=2}^{L} \max_{\beta \in U} |X_{\pi_m(\beta)} - X_{\pi_{m-1}(\beta)}|.$$

Given $\tilde{\theta} \in U$, we can build a chain for $\tilde{\theta}$ in the same way. Then

$$\begin{aligned} |X_\theta - X_{\tilde{\theta}}| &\leq \big|(X_\theta - X_{\gamma^1}) + (X_{\gamma^1} - X_{\tilde{\gamma}^1}) + (X_{\tilde{\gamma}^1} - X_{\tilde{\theta}})\big| \\ &\leq |X_\theta - X_{\gamma^1}| + \overbrace{|X_{\gamma^1} - X_{\tilde{\gamma}^1}|}^{\text{keep}} + |X_{\tilde{\gamma}^1} - X_{\tilde{\theta}}|, \end{aligned}$$

and therefore

$$\max_{\theta,\tilde{\theta} \in U} |X_\theta - X_{\tilde{\theta}}| \leq \max_{\gamma^1, \tilde{\gamma}^1 \in U_1} |X_{\gamma^1} - X_{\tilde{\gamma}^1}| + 2 \sum_{m=2}^{L} \max_{\beta \in U} \big|X_{\pi_m(\beta)} - X_{\pi_{m-1}(\beta)}\big|.$$

Next: the proof is completed in Lecture 9.

Lecture 9 — February 14, 2023

Prof. Matey Neykov *Scribe: Fan Pu Zeng*

1 Dudley Entropy Integral

Theorem 10 (Dudley entropy integral)**.** *Let $\{X_\theta, \theta \in T\}$ be a mean-zero sub-Gaussian process with respect to ρ_X. Then for all $\delta \in [0, D]$,*

$$\mathbb{E} \sup_{\theta,\tilde{\theta}\in T} (X_\theta - X_{\tilde{\theta}}) \leq 2\,\mathbb{E} \sup_{\substack{\gamma,\gamma'\in T \\ \rho_X(\gamma,\gamma')\leq\delta}} (X_\gamma - X_{\gamma'}) + 64\sqrt{2}\int_{\delta/4}^{D/2} \sqrt{\log N_X(u,T)}\, du.$$

Proof. As before,

$$\sup_{\theta,\tilde{\theta}\in T} (X_\theta - X_{\tilde{\theta}}) \leq 2 \sup_{\substack{\gamma,\gamma'\in T \\ \rho_X(\gamma,\gamma')\leq\delta}} (X_\gamma - X_{\gamma'}) + 2\max_{i\in[N]} |X_{\theta_i} - X_{\theta_1}|,$$

where $U = \{\theta_1, \ldots, \theta_N\}$ is a minimal δ-cover of T.

Definition. Let U_m be a minimal $\varepsilon_m = D2^{-m}$ cover of T, where $D = \operatorname{diam}_X(T)$, and let $N_m = |U_m| = N_X(\varepsilon_m, T)$.

Since $U \subset T$, U_m also covers U. Since $|U| = N < \infty$, there exists $L \in \mathbb{N}$ such that

$$|U_{L-1}| < |U| \quad \text{but} \quad |U_L| \geq |U|, \qquad D2^{-(L-1)} \geq \delta. \tag{$*$}$$

L is the largest integer such that $(*)$ holds. Once we reach this L we stop constructing the U_m, and we set $U_L = U$.

For all $m = 1, \ldots, L-1$, define $\pi_m : U \to U_m$ by

$$\pi_m(\theta) = \arg\min_{\beta\in U_m} \rho_X(\theta, \beta)$$

(breaking ties arbitrarily). In words, $\pi_m(\theta)$ is the best approximation to $\theta \in U$ in U_m. Let

$$\gamma^L = \theta \in U = U_L, \qquad \gamma^{m-1} = \pi_{m-1}(\theta) \in U_{m-1}, \qquad m = L, L-1, \ldots, 2.$$

By *chaining*,

$$X_\theta - X_{\gamma^1} = \sum_{m=2}^{L} (X_{\gamma^m} - X_{\gamma^{m-1}}),$$

so

$$|X_\theta - X_{\gamma^1}| \leq \sum_{m=2}^{L} \max_{\beta\in U} |X_{\pi_m(\beta)} - X_{\pi_{m-1}(\beta)}|.$$

Given $\tilde{\theta} \in U$, build a chain $\{\tilde{\gamma}^1, \ldots, \tilde{\gamma}^L\}$, and we can find a similar bound for $|X_{\tilde{\theta}} - X_{\tilde{\gamma}^1}|$. Then

$$\begin{aligned}|X_\theta - X_{\tilde{\theta}}| &= \big|(X_\theta - X_{\gamma^1}) + (X_{\gamma^1} - X_{\tilde{\gamma}^1}) + (X_{\tilde{\gamma}^1} - X_{\tilde{\theta}})\big| \\ &\le |X_\theta - X_{\gamma^1}| + |X_{\tilde{\theta}} - X_{\tilde{\gamma}^1}| + |X_{\gamma^1} - X_{\tilde{\gamma}^1}|,\end{aligned}$$

and so

$$\max_{\theta,\tilde{\theta}\in U} |X_\theta - X_{\tilde{\theta}}| \le \max_{\gamma^1,\tilde{\gamma}^1 \in U_1} |X_{\gamma^1} - X_{\tilde{\gamma}^1}| + 2\sum_{m=2}^{L} \max_{\beta\in U} \big|X_{\pi_m(\beta)} - X_{\pi_{m-1}(\beta)}\big|.$$

The first term. Write $N(t) = N_X(t, T)$. We have $|U_1| \le N(D/2) = N_X(D/2, T)$ and $\rho_X(\gamma^1, \tilde{\gamma}^1) \le D$, so each $X_{\gamma^1} - X_{\tilde{\gamma}^1}$ is sub-Gaussian with parameter D. By the maximal inequality,

$$\mathbb{E} \max_{\gamma^1,\tilde{\gamma}^1\in U_1} |X_{\gamma^1} - X_{\tilde{\gamma}^1}| \le \sqrt{2D^2 \log\big(2N^2(D/2)\big)} \le 4\sqrt{2}\, D\sqrt{\log N(D/2)},$$

using $\log 2x \le 2\log x$ for $x \ge 2$; we have $N(D/2) \ge 2$.

Why is $N(D/2) \ge 2$? If $N(D/2) = 1$, then $|U_1| = 1$, so $\max_{\gamma^1,\tilde{\gamma}^1\in U_1} |X_{\gamma^1} - X_{\tilde{\gamma}^1}| = 0$, and then the bound is always true.

The chaining terms. We have

$$\max_{\beta\in U} \big|X_{\pi_m(\beta)} - X_{\pi_{m-1}(\beta)}\big| \le \max_{\substack{\gamma\in U_m,\, \tilde{\gamma}\in U_{m-1} \\ \rho_X(\gamma,\tilde{\gamma}) \le D2^{-(m-2)}}} |X_\gamma - X_{\tilde{\gamma}}|,$$

since

$$\begin{aligned}\rho_X\big(\pi_m(\beta), \pi_{m-1}(\beta)\big) &\le \rho_X\big(\pi_m(\beta), \beta\big) + \rho_X\big(\pi_{m-1}(\beta), \beta\big) \\ &\le D2^{-m} + D2^{-(m-1)} \le D2^{-(m-2)}.\end{aligned}$$

The number of pairs is

$$|U_{m-1}||U_m| \le |U_m|^2 = N^2(D2^{-m}).$$

So, by the maximal inequality again,

$$\begin{aligned}\mathbb{E} \max_{\beta\in U} \big|X_{\pi_m(\beta)} - X_{\pi_{m-1}(\beta)}\big| &\le \sqrt{2\big(D2^{-(m-2)}\big)^2 \log\big(2N^2(D2^{-m})\big)} \\ &\le 2\sqrt{2}\, D2^{-(m-2)}\sqrt{\log N(D2^{-m})},\end{aligned}$$

since $N(D/2) \ge 2 \implies N(D2^{-m}) \ge 2$.

Putting it together.

$$\begin{aligned}\mathbb{E} \max_{\theta,\tilde{\theta}\in U} |X_\theta - X_{\tilde{\theta}}| &\le 4\sqrt{2}\, D\sqrt{\log N(D/2)} + 2\sum_{m=2}^{L} 2\sqrt{2}\, D2^{-(m-2)}\sqrt{\log N(D2^{-m})} \\ &\le 8\sqrt{2}\sum_{m=1}^{L} D2^{-(m-1)}\sqrt{\log N(D2^{-m})}.\end{aligned}$$

Since $\log N(t)$ is decreasing in t,

$$D2^{-(m-1)}\sqrt{\log N(D2^{-m})} \le 4\int_{D2^{-(m+1)}}^{D2^{-m}} \sqrt{\log N(u)}\, du.$$

Why? Over the range of integration $u \le D2^{-m}$, so

$$4\int_{D2^{-(m+1)}}^{D2^{-m}} \sqrt{\log N(u)}\, du \ge 4\sqrt{\log N(D2^{-m})}\, \underbrace{(D2^{-m} - D2^{-(m+1)})}_{D2^{-(m+1)}} = D2^{-(m-1)}\sqrt{\log N(D2^{-m})}.$$

Summing over $m = 1, \dots, L$, the intervals $[D2^{-(m+1)}, D2^{-m}]$ fit together into $[D2^{-(L+1)}, D/2]$. Recall that L is such that

$$D2^{-(L-1)} \ge \delta \implies D2^{-(L+1)} \ge \frac{\delta}{4}.$$

Since $\max_{i\in[N]}|X_{\theta_i} - X_{\theta_1}| \le \max_{\theta,\tilde\theta\in U}|X_\theta - X_{\tilde\theta}|$, this gives

$$2\,\mathbb{E}\max_{\theta,\tilde\theta\in U}|X_\theta - X_{\tilde\theta}| \le 64\sqrt{2}\int_{\delta/4}^{D/2}\sqrt{\log N(u)}\, du,$$

which proves the theorem. ■

Idea: we can look at the set T at different resolutions.

Example 10 (B_2^n)**.** Let $T = B_2^n$, and use $\delta = 0$. Then

$$\mathcal{G}(T) \le C_1 \int_0^{\operatorname{diam}(T)/2} \sqrt{\log N_2(u, T)}\, du,$$

where C_1 is a constant. Since

$$N(B_2^n, \varepsilon) \le \Big(\frac{3}{\varepsilon}\Big)^n \qquad \text{for } \varepsilon \in (0, 1],$$

this gives

$$C_1 \int_0^{1/2} \sqrt{n\log(3/u)}\, du \asymp \sqrt{n}.$$

2 Gaussian Comparison Inequalities

Given two Gaussian processes $\{Y_\theta, \theta \in T\}$ and $\{X_\theta, \theta \in T\}$, and a real-valued function F, we are interested in comparing $F(X)$ with $F(Y)$.

Example 11. $F(X) = \sup_{\theta\in T} X_\theta$.

For what follows, we will assume that the processes take finitely many values (that is, T is finite).

Theorem 11. *Let $(X_1, \dots, X_N)$ and $(Y_1, \dots, Y_N)$ be a pair of centered Gaussian random vectors. Suppose there exist $A, B \subseteq [N] \times [N]$ with $A \cap B = \emptyset$ such that*

$$\begin{aligned}
&\mathbb{E}\, X_iX_j \le \mathbb{E}\, Y_iY_j && \forall (i,j) \in A,\\
&\mathbb{E}\, X_iX_j \ge \mathbb{E}\, Y_iY_j && \forall (i,j) \in B,\\
&\mathbb{E}\, X_iX_j = \mathbb{E}\, Y_iY_j && \forall (i,j) \notin A \cup B.
\end{aligned}$$

Let $F : \mathbb{R}^N \to \mathbb{R}$ be a twice differentiable function, and suppose that

$$\frac{\partial^2 F}{\partial u_i \partial u_j}(u) \geq 0 \qquad \forall (i,j) \in A,$$
$$\frac{\partial^2 F}{\partial u_i \partial u_j}(u) \leq 0 \qquad \forall (i,j) \in B.$$

Then

$$\mathbb{E}\, F(X) \leq \mathbb{E}\, F(Y).$$

Proof. Without loss of generality, $X \perp Y$. Even if there were a correlation structure, that is, a covariance matrix

$$\begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix} \xrightarrow{\text{linear transform}} \begin{bmatrix} \Sigma_{XX} & 0 \\ 0 & \Sigma_{YY} \end{bmatrix},$$

we can multiply by some matrix to get to this correlation structure.

Define an interpolating process

$$Z(t) = \sqrt{1-t}\, X + \sqrt{t}\, Y, \qquad \forall t \in [0,1].$$

The square roots make the interpolation variance preserving. Let $\phi : [0,1] \to \mathbb{R}$,

$$\phi(t) = \mathbb{E}\, F\big(Z(t)\big).$$

Aim: show $\phi'(t) \geq 0$, which implies

$$\underbrace{\phi(1)}_{\mathbb{E}\, F(Y)} \geq \underbrace{\phi(0)}_{\mathbb{E}\, F(X)}.$$

Next: the proof continues in the next lecture.

Lecture 10 — February 16, 2023

Prof. Matey Neykov *Scribe: Fan Pu Zeng*

1 Slepian's Inequality

Theorem 12 (Slepian's inequality)**.** *Let $X, Y \in \mathbb{R}^N$ be zero-mean Gaussian random vectors such that*

$$\begin{aligned} \mathbb{E}\, X_i X_j &\geq \mathbb{E}\, Y_i Y_j && \forall i \neq j, \\ \mathbb{E}\, X_i^2 &= \mathbb{E}\, Y_i^2 && \forall i \in [N]. \end{aligned}$$

Then

$$\mathbb{E} \max_{i \in [N]} X_i \leq \mathbb{E} \max_{i \in [N]} Y_i.$$

Proof. Define the *smooth max function*

$$F_\beta(x) = \beta^{-1} \log \sum_{j=1}^{N} e^{\beta x_j},$$

which is an approximate max function. One can prove that

$$\max_{j \in [N]} x_j \leq F_\beta(x) \leq \max_{j \in [N]} x_j + \frac{\log N}{\beta}.$$

This follows from $e^{\beta \max_j x_j} \leq \sum_{j=1}^N e^{\beta x_j} \leq N e^{\beta \max_j x_j}$ by taking logarithms and dividing by β. This is true for any numbers x_j, not just Gaussian random variables.

The mixed second derivatives are

$$\frac{\partial^2 F_\beta(x)}{\partial x_i \partial x_j} = -\frac{\beta e^{\beta x_i} e^{\beta x_j}}{\left(\sum_{k \in [N]} e^{\beta x_k}\right)^2} \leq 0 \qquad \forall i \neq j.$$

So we take $A = \emptyset$ and $B = \{(i,j) : i \neq j\}$ in the Gaussian comparison inequality from last time. Then

$$\mathbb{E} \max_{i \in [N]} X_i \leq \mathbb{E}\, F_\beta(X) \leq \mathbb{E}\, F_\beta(Y) \leq \mathbb{E} \max_{i \in [N]} Y_i + \frac{\log N}{\beta},$$

where the middle inequality is that comparison theorem. Taking $\beta \to \infty$ proves the claim. ■

2 Sudakov–Fernique Inequality

Theorem 13 (Sudakov–Fernique)**.** *Let $\{X_j, j = 1, \ldots, N\}$ and $\{Y_j, j = 1, \ldots, N\}$ be a pair of centered Gaussian processes (vectors). Suppose that*

$$\mathbb{E}(X_i - X_j)^2 \leq \mathbb{E}(Y_i - Y_j)^2 \qquad \forall (i,j) \in [N] \times [N].$$

Then

$$\mathbb{E} \max_{i \in [N]} X_i \leq \mathbb{E} \max_{i \in [N]} Y_i.$$

This has requirements closely related to those of Slepian's inequality: expanding the squares, the condition compares

$$\mathbb{E}\, X_i^2 - 2\,\mathbb{E}\, X_i X_j + \mathbb{E}\, X_j^2 \quad \text{and} \quad \mathbb{E}\, Y_i^2 - 2\,\mathbb{E}\, Y_i Y_j + \mathbb{E}\, Y_j^2.$$

If the variances are equal, it is the same as the previous theorem.

3 Gaussian Contraction Inequality

Let each ϕ_j be a 1-Lipschitz function,

$$|\phi_j(s) - \phi_j(t)| \leq |s - t| \qquad \forall s, t \in \mathbb{R},$$

with $\phi_j(0) = 0$. For $\theta \in \mathbb{R}^d$ define

$$\phi(\theta) = (\phi_1(\theta_1), \ldots, \phi_d(\theta_d)) \in \mathbb{R}^d.$$

Consider $T \subset \mathbb{R}^d$ and the image of T under ϕ,

$$\phi(T) = \{\phi(\theta) : \theta \in T\}.$$

Theorem 14 (Gaussian contraction inequality)**.** *With the setup above,*

$$\mathcal{G}(\phi(T)) \leq \mathcal{G}(T),$$

that is,

$$\mathbb{E}\Big[\sup_{\theta \in T} \sum_{j=1}^{d} w_j \phi_j(\theta_j)\Big] \leq \mathbb{E}\Big[\sup_{\theta \in T} \sum_{j=1}^{d} w_j \theta_j\Big].$$

This is a consequence of Sudakov–Fernique.

Proof sketch. Define

$$X_\theta = \sum_j w_j \phi_j(\theta_j), \qquad Y_\theta = \sum_j w_j \theta_j.$$

Use

$$\mathbb{E}(X_\theta - X_{\tilde{\theta}})^2 \leq \mathbb{E}(Y_\theta - Y_{\tilde{\theta}})^2$$

and apply Sudakov–Fernique. The inequality holds because $\mathbb{E}(X_\theta - X_{\tilde{\theta}})^2 = \sum_j \big(\phi_j(\theta_j) - \phi_j(\tilde{\theta}_j)\big)^2 \leq \sum_j (\theta_j - \tilde{\theta}_j)^2 = \mathbb{E}(Y_\theta - Y_{\tilde{\theta}})^2$, using that each ϕ_j is 1-Lipschitz.

We can also show that

$$\mathcal{R}(\phi(T)) \leq \mathcal{R}(T).$$

(MJW gives $\mathcal{R}(\phi(T)) \leq 2\mathcal{R}(T)$.) We cannot apply Sudakov–Fernique here, since we no longer have Gaussian random variables.

Example 12. Given a function class $\mathcal{F}$ and design points $x_1^n = (x_1, \dots, x_n)$, we want to evaluate the Gaussian complexity of

$$\mathcal{F}^2(x_1^n) = \{(f^2(x_1), \dots, f^2(x_n)) : f \in \mathcal{F}\} \subset \mathbb{R}^n.$$

Suppose $\|f\|_\infty \le b$ for all $f \in \mathcal{F}$. We claim

$$\mathcal{G}(\mathcal{F}^2(x_1^n)) \le 2b\,\mathcal{G}(\mathcal{F}(x_1^n)).$$

Proof. Define $\phi_b : \mathbb{R} \to \mathbb{R}$ by

$$\phi_b(t) = \begin{cases} t^2/2b & |t| \le b, \\ b/2 & |t| > b. \end{cases}$$

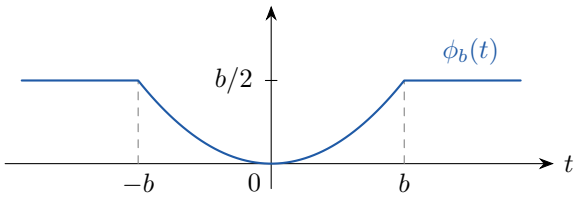


Figure 11: The function ϕ_b: equal to $t^2/2b$ on $[-b, b]$ and constant at $b/2$ outside.

Because $|f(x_i)| \le b$,

$$\phi_b(f(x_i)) = \frac{f^2(x_i)}{2b}.$$

So

$$\begin{aligned} \frac{1}{2b}\,\mathcal{G}(\mathcal{F}^2(x_1^n)) &= \mathbb{E}\Big[\sup_{f\in\mathcal{F}} \sum_{i=1}^n \frac{w_i f^2(x_i)}{2b}\Big] \\ &= \mathbb{E}\Big[\sup_{f\in\mathcal{F}} \sum_{i=1}^n w_i \phi_b(f(x_i))\Big] \\ &\le \mathbb{E}\Big[\sup_{f\in\mathcal{F}} \sum_{i=1}^n w_i f(x_i)\Big] = \mathcal{G}(\mathcal{F}(x_1^n)). \end{aligned}$$

For the inequality: $\phi_b(0) = 0$, and ϕ_b has derivative $2t/2b \le 1$ in absolute value, so it is 1-Lipschitz. Then we apply the Gaussian contraction inequality. ■

4 Sudakov Minoration

Theorem 15 (Sudakov minoration)**.** *Let $\{X_\theta, \theta \in T\}$ be a zero-mean Gaussian process defined on T, where $T \subset \mathbb{R}^d$ and $T \neq \emptyset$. Then*

$$\mathbb{E}\sup_{\theta\in T} X_\theta \ge \sup_\delta \frac{\delta}{2}\sqrt{\log M_X(\delta, T)},$$

where $M_X(\delta, T)$ is the δ-packing number of T with respect to

$$\rho_X(\theta, \tilde\theta) = \sqrt{\mathbb{E}(X_\theta - X_{\tilde\theta})^2}.$$

Unfortunately, this does not match the upper bound in general. For the canonical process $X_\theta = \langle w, \theta \rangle$, check that $\rho_X(\theta, \tilde{\theta}) = \left\| \theta - \tilde{\theta} \right\|_2$.

Proof. For every $\delta > 0$, let $\{\theta^1, \dots, \theta^M\}$ be a δ-packing of T, and let $Y_i = X_{\theta^i}$ for $i = 1, \dots, M$. Then

$$\mathbb{E}(Y_i - Y_j)^2 = \rho_X^2(\theta^i, \theta^j) > \delta^2 \qquad \forall i \neq j,$$

since the θ^i form a δ-packing. Construct $Z_i \overset{\text{i.i.d.}}{\sim} N(0, \delta^2/2)$. Then

$$\mathbb{E}(Z_i - Z_j)^2 = \frac{\delta^2}{2} + \frac{\delta^2}{2} = \delta^2 \qquad \forall i \neq j.$$

Therefore

$$\mathbb{E} \sup_{\theta \in T} X_\theta \geq \mathbb{E} \max_{i \in [M]} Y_i \overset{\text{S–F}}{\geq} \mathbb{E} \max_{i \in [M]} Z_i \geq \frac{\delta}{2} \sqrt{\log M},$$

where the second inequality is Sudakov–Fernique and the last is by Exercise 2.11 in MJW. ■

Example 13 (ℓ_1-ball)**.** Let

$$B_1^d = \{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq 1\}$$

(the cross polytope). Then

$$\mathcal{G}(B_1^d) = \mathbb{E} \sup_{\|\theta\|_1 \leq 1} \langle w, \theta \rangle = \mathbb{E} \|w\|_\infty \leq 2\sqrt{\log d} \qquad (d \geq 2).$$

The middle equality holds because ℓ_1 and ℓ_∞ are dual norms. Alternatively, see it as

$$\sum_{i=1}^d w_i \theta_i \leq \sum_{i=1}^d |w_i||\theta_i| \leq \max_{i \in [d]} |w_i|.$$

By Sudakov minoration, for all $\delta > 0$,

$$\frac{\delta}{2} \sqrt{\log M(\delta, B_1^d, \|\cdot\|_2)} \leq 2\sqrt{\log d}.$$

So this matches. Furthermore,

$$\log M(\delta, B_1^d, \|\cdot\|_2) \leq \frac{C \log d}{\delta^2} \qquad \text{for } \delta \gtrsim \frac{1}{\sqrt{d}}.$$

5 Principal Component Analysis (PCA)

- PCA is a standard technique for exploratory data analysis, visualization, and dimension reduction.
- Find the maximum variance component of a distribution $\iff$ find a low-dimensional space that captures the majority of the variance in the data.

Suppose $X \sim (0, \Sigma) \in \mathbb{R}^d$. We want to find $v \in S^{d-1} \subset \mathbb{R}^d$ such that $\langle v, X \rangle$ has maximum variance:

$$\begin{aligned} v^* &= \arg\max_{v \in S^{d-1}} \operatorname{Var}\big(\langle v, X \rangle\big) \\ &= \arg\max_{v \in S^{d-1}} \mathbb{E} \langle v, X \rangle^2 && \text{(the mean is 0)} \\ &= \arg\max_{v \in S^{d-1}} \langle v, \Sigma v \rangle . \end{aligned}$$

So v^* is the principal eigenvector of Σ, assuming it is unique.

Suppose we are interested in finding an orthonormal matrix $V \in \mathbb{R}^{d \times r}$,

$$V = \begin{bmatrix} | & & | \\ v_1 & \cdots & v_r \\ | & & | \end{bmatrix},$$

such that

$$\max \mathbb{E} \left\| V^T X \right\|_2^2 = \max \mathbb{E} \sum_{j=1}^{r} \langle v_j, X \rangle^2 .$$

This is achieved when the v_j are the top r eigenvectors of Σ. PCA gets the top r eigenvectors of $\hat{\Sigma}$, the sample covariance.

36-709: Advanced Statistical Theory I Spring 2023

Lecture 11 — February 21, 2023

Prof. Matey Neykov *Scribe: Fan Pu Zeng*

1 Matrix Recap

Singular value decomposition

Any $m \times n$ matrix A has an SVD

$$\underbrace{A}_{m\times n} = \underbrace{U}_{m\times r}\underbrace{D}_{r\times r}\underbrace{V^T}_{r\times n},$$

where $r \leq \min(m, n)$ is the rank of A, U and V are orthogonal matrices, and D is diagonal. Equivalently,

$$A = \sum_{j=1}^{r} \lambda_j u_j v_j^T, \qquad \lambda_j \geq 0,$$

where the λ_j are the *singular values* of A.

Note. Let λ be the eigenvalues of $A^T A$. The values $\sqrt{\lambda}$ for $\lambda > 0$ are the singular values of A.

Vector norms of a matrix

$$|A|_q = \Big(\sum_{i,j} |a_{ij}|^q\Big)^{1/q}.$$

These are not used often.

Spectral norm of a matrix

Suppose A has singular values $(\lambda_1, \ldots, \lambda_r, 0, \ldots, 0) = \lambda$. Then

$$\|A\|_q = \|\lambda\|_q .$$

- If $q = \infty$,
$$\|A\|_\infty = \|A\|_{\text{op}} = \max_i |\lambda_i| = \max_i \lambda_i = \|\lambda\|_\infty .$$
In Wainwright's book, $|||\cdot|||$ denotes the operator norm.
- If $q = 2$,
$$\|A\|_2 = \|A\|_F = \|\lambda\|_2 = \sqrt{\text{tr}(A^T A)},$$
the *Frobenius norm*.

Let A, B be $m \times n$ matrices with singular values

$$\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_{\min(m,n)}(A),$$
$$\lambda_1(B) \geq \lambda_2(B) \geq \cdots \geq \lambda_{\min(m,n)}(B).$$

Then *Weyl's inequality* says

$$\max_k |\lambda_k(A) - \lambda_k(B)| \leq \|A - B\|_{\text{op}}.$$

To see the connection with eigenvalues, consider

$$\begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix}, \qquad \begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix}.$$

These two matrices have eigenvalues equal to the singular values of A and B, respectively.

Inner product between matrices

For matrices whose dimensions make sense,

$$\langle A, B \rangle = \operatorname{tr}(A^T B) = \operatorname{tr}(B^T A) \leq \|A\|_p \|B\|_q, \qquad \text{where } \frac{1}{p} + \frac{1}{q} = 1,\ p, q \in [1, \infty].$$

This is *Hölder's inequality*.

2 Covariance Matrix Estimation

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} (0, \Sigma)$ with $X_i \in \mathbb{R}^d$, that is, $\mathbb{E}\, X = 0$ and $\mathbb{E}\, XX^T = \Sigma$. The *sample covariance* is

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n X_i X_i^T.$$

Theorem 16. *Let $Y \in \mathbb{R}^d$ be a random vector such that*

$$\mathbb{E}\, Y = 0, \qquad \mathbb{E}\, YY^T = I_d, \qquad v^T Y \sim \text{subG}(1) \quad \forall v \in S^{d-1}$$

(that is, for every unit vector v). Let $X_1, \ldots, X_n$ be n independent copies of the sub-Gaussian random vector $X = \Sigma^{1/2} Y$. Then

$$\mathbb{E}\, X = 0, \qquad \mathbb{E}\, XX^T = \Sigma, \qquad v^T X \sim \text{subG}\big(\|\Sigma\|_{\text{op}}^{1/2}\big) \quad \forall v \in S^{d-1}.$$

Furthermore, with probability at least $1 - \delta$,

$$\left\|\hat{\Sigma} - \Sigma\right\|_{\text{op}} \lesssim \|\Sigma\|_{\text{op}} \left(\sqrt{\frac{d + \log(1/\delta)}{n}} \vee \frac{d + \log(1/\delta)}{n} \right),$$

where $a \vee b = \max(a, b)$ and $\lesssim$ holds up to constants.

Remark 8. Y is a *sub-Gaussian vector* if $v^T Y \sim \text{subG}(1)$ for all $v \in S^{d-1}$.

Proof of the theorem. Clearly $\mathbb{E}\, X = 0$ and $\mathbb{E}\, XX^T = \Sigma$. We prove the other facts first. Writing $u = \Sigma^{1/2} v / \left\| v^T \Sigma^{1/2} \right\|_2$, which has norm 1,

$$\begin{aligned}\mathbb{E}\, e^{\lambda v^T X} &= \mathbb{E}\, e^{\lambda v^T \Sigma^{1/2} Y} = \mathbb{E}\, e^{\lambda \left\| v^T \Sigma^{1/2} \right\|_2 u^T Y} \\ &\le e^{\lambda^2 \left\| v^T \Sigma^{1/2} \right\|_2^2 / 2} \le e^{\lambda^2 \|\Sigma\|_{\mathrm{op}} / 2},\end{aligned}$$

where the first inequality holds since Y is 1-sub-Gaussian: $\mathbb{E}\, e^{\lambda u^T Y} \le e^{\lambda^2/2}$. This shows $v^T X \sim \mathrm{subG}(\|\Sigma\|_{\mathrm{op}}^{1/2})$.

Next,

$$\begin{aligned}\frac{\left\| \hat{\Sigma} - \Sigma \right\|_{\mathrm{op}}}{\|\Sigma\|_{\mathrm{op}}} &= \frac{\left\| \frac{1}{n} \sum_{i=1}^n X_i X_i^T - \Sigma \right\|_{\mathrm{op}}}{\|\Sigma\|_{\mathrm{op}}} \\ &= \frac{\left\| \Sigma^{1/2} \left(\frac{1}{n} \sum_{i=1}^n Y_i Y_i^T - I_d \right) \Sigma^{1/2} \right\|_{\mathrm{op}}}{\|\Sigma\|_{\mathrm{op}}} \\ &\le \frac{\left\| \Sigma^{1/2} \right\|_{\mathrm{op}} \left\| \frac{1}{n} \sum_{i=1}^n Y_i Y_i^T - I_d \right\|_{\mathrm{op}} \left\| \Sigma^{1/2} \right\|_{\mathrm{op}}}{\|\Sigma\|_{\mathrm{op}}} \\ &= \left\| \frac{1}{n} \sum_{i=1}^n Y_i Y_i^T - I_d \right\|_{\mathrm{op}},\end{aligned}$$

since $\left\| \Sigma^{1/2} \right\|_{\mathrm{op}}^2 = \|\Sigma\|_{\mathrm{op}}$. So we may assume without loss of generality that $\Sigma = I_d$.

Now we construct an ε-net. Let N be a 1/4-net of S^{d-1}, meaning it is a minimal 1/4-cover of S^{d-1}. We want to claim

$$\left\| \hat{\Sigma} - I_d \right\|_{\mathrm{op}} \le 2 \max_{x,y \in N} x^T (\hat{\Sigma} - I_d) y.$$

The idea is that we can use a union bound to take the max over a finite set. The net is bounded by

$$|N| \le \left(\frac{3}{1/4} \right)^d = 12^d.$$

(Previously we showed the bound $(3/\delta)^d$, where δ is the size of the smaller ball.)

Lemma 3. *Let N be a 1/4-net of S^{d-1}. Then*

$$\|A\|_{\mathrm{op}} \le 2 \max_{x,y \in N} x^T A y.$$

This trick is useful whenever you need to optimize over an uncountable set.

Proof. For $u, v \in S^{d-1}$,

$$u^T A v \le \max_{x \in N} x^T A v + \frac{1}{4} \max_{u \in S^{d-1}} u^T A v.$$

Why? There exists $x_0 \in N$ such that $\|u - x_0\|_2 \le 1/4$, and

$$u^T A v = \underbrace{x_0^T A v}_{\le \max_{x \in N} x^T A v} + \underbrace{\frac{(u - x_0)^T A v}{\|u - x_0\|_2} \|u - x_0\|_2}_{\le \frac{1}{4} \max_{u \in S^{d-1}} u^T A v}.$$

Continuing, by the same argument applied to v,

$$\begin{aligned} u^T A v &\le \max_{x,y\in N} x^T A y + \frac{1}{4}\max_{x\in N}\max_{v\in S^{d-1}} x^T A v + \frac{1}{4}\max_{u\in S^{d-1}} u^T A v \\ &\le \max_{x,y\in N} x^T A y + \frac{1}{2}\max_{u,v\in S^{d-1}} u^T A v, \end{aligned}$$

by relaxing the net constraint. Maximizing over all $u, v \in S^{d-1}$ on the left-hand side gives

$$\underbrace{\max_{u,v\in S^{d-1}} u^T A v}_{\|A\|_{\mathrm{op}}} \le 2\max_{x,y\in N} x^T A y,$$

after moving the $\frac{1}{2}\max_{u,v\in S^{d-1}} u^T A v$ term to the left-hand side and multiplying by 2. ■

With the lemma done, we continue the proof of the theorem. For all $t \ge 0$, by the lemma and the union bound,

$$\mathbb{P}(\|\hat{\Sigma} - I_d\|_{\mathrm{op}} > t) \le \sum_{x,y\in N} \mathbb{P}\Big(x^T(\hat{\Sigma} - I_d)y > \frac{t}{2}\Big). \tag{$\star$}$$

Here

$$x^T(\hat{\Sigma} - I_d)y = \frac{1}{n}\sum_{i=1}^n \Big\{(X_i^T x)(X_i^T y) - \underbrace{\mathbb{E}(X_i^T x)(X_i^T y)}_{x^T y}\Big\},$$

where $X_i^T x$ and $X_i^T y$ are each subG(1). From a homework problem (HW 1), we know that

$$(X_i^T x)(X_i^T y) \sim \mathrm{subE}(\kappa_1, \kappa_2),$$

where κ_1, κ_2 are absolute constants. By the concentration inequality for sub-exponential random variables,

$$\mathbb{P}\Big(x^T(\hat{\Sigma} - I_d)y > \frac{t}{2}\Big) \le \exp\Big(-n\Big(\frac{t^2}{4\cdot 2\kappa_1^2} \wedge \frac{t}{2\cdot 2\kappa_2}\Big)\Big).$$

So for any x, y we have the above result. Then from $(\star)$,

$$\sum_{x,y\in N} \mathbb{P}\Big(x^T(\hat{\Sigma} - I_d)y > \frac{t}{2}\Big) \le 144^d \exp\Big(-\frac{n}{2}\Big(\frac{t^2}{4\kappa_1^2} \wedge \frac{t}{2\kappa_2}\Big)\Big),$$

using $|N|^2 \le 144^d$. For this to be smaller than δ, choose

$$t = \left(\frac{2\kappa_2 \cdot 2d\log 144}{n} + \frac{2\kappa_2 \cdot 2\log(1/\delta)}{n}\right) \vee \left(\frac{(2\kappa_1)^2 \cdot 2d\log 144}{n} + \frac{(2\kappa_1)^2 \cdot 2\log(1/\delta)}{n}\right)^{1/2}.$$

With this choice, with probability at least $1 - \delta$,

$$\left\|\hat{\Sigma} - I_d\right\|_{\mathrm{op}} \le t \lesssim \sqrt{\frac{d + \log(1/\delta)}{n}} \vee \frac{d + \log(1/\delta)}{n},$$

which proves the theorem. ■

1 Spiked Covariance Model

In the *spiked covariance model*,

$$X_i \overset{\text{i.i.d.}}{\sim} N(0,\Sigma), \quad i \in \{1,\dots,n\}, \qquad \Sigma = \theta v v^T + I_d, \quad v \in S^{d-1}, \quad \theta \geq 0.$$

Here θvv^T is the *spike*, θ is the *signal strength*, and the principal eigenvector of Σ is v.

Roughly speaking, if $d \gtrsim n$, then $\min_{\varepsilon\in\{\pm1\}} \|\varepsilon\hat{v} - v\|_2$ can be very big. If $d \lesssim n$, PCA on $\hat{\Sigma}$ has error of order d/n.

Equivalently, we can write

$$X_i = \sqrt{\theta}\, v\, \xi_i + w_i, \qquad w_i \sim N(0, I_d), \quad \xi_i \sim N(0,1), \quad \xi_i \perp w_i.$$

Instead of using the loss function $\|\hat{v} - v\|_2$, we will use the loss

$$\min_{\varepsilon\in\{\pm1\}} \|\varepsilon\hat{v} - v\|_2, \qquad \text{where} \qquad \hat{v} = \arg\max_{w\in S^{d-1}} w^T\hat{\Sigma}w,$$

because when we get the principal eigenvector, we don't know what its sign is.

2 The Davis–Kahan Theorem

Let us introduce the principal angle.

Definition (Principal angle)**.** For $\|u\|_2 = \|v\|_2 = 1$, the *principal angle* between u and v is

$$\sphericalangle(u,v) = \arccos|u^Tv|.$$

Theorem 17 (Davis–Kahan $\sin\Theta$ theorem)**.** *Let A, B be two $d\times d$ PSD matrices. Let (λ_1,u_1), …, (λ_d,u_d) and (μ_1,v_1), …, (μ_d,v_d) be the eigenvalue–eigenvector pairs of A and B respectively, ordered so that $\lambda_1 \geq \cdots \geq \lambda_d$ and $\mu_1 \geq \cdots \geq \mu_d$. Then*

1. $$\sin\big(\sphericalangle(u_1,v_1)\big) \leq \frac{2\|A-B\|_{\text{op}}}{\max(\lambda_1-\lambda_2,\ \mu_1-\mu_2)};$$

2. $$\min_{\varepsilon\in\{\pm1\}} \|\varepsilon u_1 - v_1\|_2^2 \leq 2\sin^2\big(\sphericalangle(u_1,v_1)\big).$$

Proof. For all $x \in S^{d-1}$,

$$\begin{aligned} x^T A x = \sum_{j=1}^{d} (x^T u_j)^2 \lambda_j &\le (x^T u_1)^2 \lambda_1 + \lambda_2 \underbrace{\sum_{j=2}^{d} (x^T u_j)^2}_{=1-(x^T u_1)^2} \\ &= \cos^2 \sphericalangle(u_1, x)\, \lambda_1 + \lambda_2 \sin^2 \sphericalangle(u_1, x), \end{aligned}$$

since x is a unit vector and the u_j are orthonormal. Plugging in $x = v_1$, and using $u_1^T A u_1 = \lambda_1$,

$$\begin{aligned} u_1^T A u_1 - v_1^T A v_1 &\ge \lambda_1 - \lambda_1 \cos^2 \sphericalangle(u_1, v_1) - \lambda_2 \sin^2 \sphericalangle(u_1, v_1) \\ &= (\lambda_1 - \lambda_2) \sin^2 \sphericalangle(u_1, v_1). \end{aligned}$$

Now we also want to upper bound $u_1^T A u_1 - v_1^T A v_1$. Adding and subtracting $u_1^T B u_1$,

$$\begin{aligned} u_1^T A u_1 - v_1^T A v_1 &= u_1^T B u_1 - v_1^T A v_1 + u_1^T (A - B) u_1 \\ &\le v_1^T B v_1 - v_1^T A v_1 + u_1^T (A - B) u_1 \\ &= \left\langle A - B,\ u_1 u_1^T - v_1 v_1^T \right\rangle, \end{aligned}$$

where the inequality holds because v_1 is the principal eigenvector of B. For the last step, recall $\langle A, B \rangle = \operatorname{tr}(A^T B)$, and see that

$$\left\langle A - B,\ u_1 u_1^T \right\rangle = \operatorname{tr}((A - B) u_1 u_1^T) = u_1^T (A - B) u_1,$$

since we can move terms around inside the trace.

By Hölder's inequality for matrices, $\langle A, B \rangle \le \|A\|_p \|B\|_q$, so

$$\begin{aligned} \left\langle A - B,\ u_1 u_1^T - v_1 v_1^T \right\rangle &\le \|A - B\|_\infty \left\| u_1 u_1^T - v_1 v_1^T \right\|_1 \\ &= \|A - B\|_{\mathrm{op}} \left\| u_1 u_1^T - v_1 v_1^T \right\|_1, \end{aligned}$$

where $\|\cdot\|_1$ is the nuclear norm and $\|\cdot\|_\infty$ is the operator norm. The matrix $u_1 u_1^T - v_1 v_1^T$ has rank 2; call its two eigenvalues ν_1, ν_2. Then

$$|\nu_1| + |\nu_2| \le \sqrt{2} \sqrt{\nu_1^2 + \nu_2^2},$$

which can be checked by squaring. Hence

$$\left\langle A - B,\ u_1 u_1^T - v_1 v_1^T \right\rangle \le \sqrt{2}\, \|A - B\|_{\mathrm{op}} \left\| u_1 u_1^T - v_1 v_1^T \right\|_2. \tag{$\star$}$$

Since u_1, v_1 are unit vectors, summing the squared entries gives

$$\left\| u_1 u_1^T - v_1 v_1^T \right\|_2^2 = \left\| u_1 u_1^T - v_1 v_1^T \right\|_F^2 = 2 - 2(u_1^T v_1)^2 = 2 \sin^2\big(\sphericalangle(u_1, v_1)\big).$$

Substituting in, the right-hand side of ($\star$) is $2 \|A - B\|_{\mathrm{op}} \sin \sphericalangle(u_1, v_1)$. So

$$2 \|A - B\|_{\mathrm{op}} \sin \sphericalangle(u_1, v_1) \ge u_1^T A u_1 - v_1^T A v_1 \ge (\lambda_1 - \lambda_2) \sin^2 \sphericalangle(u_1, v_1),$$

and therefore

$$\sin \sphericalangle(u_1, v_1) \le \frac{2\|A - B\|_{\mathrm{op}}}{\lambda_1 - \lambda_2}.$$

Here $\sin \sphericalangle(u_1, v_1)$ is nonnegative, since we take $\sphericalangle(u_1, v_1) = \arccos|u_1^T v_1|$. By symmetry we get the second bound for free, that is,

$$\sin \sphericalangle(u_1, v_1) \le \frac{2\|A - B\|_{\mathrm{op}}}{\max(\lambda_1 - \lambda_2,\ \mu_1 - \mu_2)}.$$

Proof of the second part. Expanding,

$$\min_{\varepsilon \in \{\pm 1\}} \|\varepsilon u_1 - v_1\|_2^2 = \min_{\varepsilon \in \{\pm 1\}} \Big(\underbrace{\|\varepsilon u_1\|_2^2}_{1} - 2\langle \varepsilon u_1, v_1 \rangle + \underbrace{\|v_1\|_2^2}_{1} \Big) = 2 - 2|u_1^T v_1|.$$

Since $|u_1^T v_1| \le 1$ by Cauchy–Schwarz,

$$2 - 2|u_1^T v_1| \le 2 - 2(u_1^T v_1)^2 = 2\sin^2 \sphericalangle(u_1, v_1). \qquad \blacksquare$$

Now return to the spiked covariance model. Take A to be the population covariance and B the sample covariance:

$$A = \Sigma = \theta v v^T + I_d, \qquad B = \hat{\Sigma}.$$

The eigengap of Σ is the largest eigenvalue minus the second largest,

$$\mathrm{eigengap}(\Sigma) = (1 + \theta) - 1 = \theta.$$

The operator norm error $\left\|\hat{\Sigma} - \Sigma\right\|_{\mathrm{op}}$ is controlled, up to constants, by the bound from last lecture.

Corollary 3. *Let $Y \in \mathbb{R}^d$ be a random vector with $\mathbb{E}\, Y = 0$, $\mathbb{E}\, YY^T = I_d$, and $u^T Y \sim \mathrm{subG}(1)$ for all $u \in S^{d-1}$. Let the X_i be i.i.d. copies of $X = \Sigma^{1/2} Y$, where $\Sigma = \theta v v^T + I_d$. Then the largest eigenvector $\hat{v}$ of $\hat{\Sigma}$ satisfies*

$$\min_{\varepsilon \in \{\pm 1\}} \|\varepsilon \hat{v} - v\|_2 \lesssim \frac{1 + \theta}{\theta} \left(\sqrt{\frac{d + \log(1/\delta)}{n}} \vee \frac{d + \log(1/\delta)}{n} \right)$$

with probability at least $1 - \delta$.

Here $1 + \theta = \|\Sigma\|_{\mathrm{op}}$, the term in parentheses comes from the bound we had last class, and the θ in the denominator is the eigengap from the Davis–Kahan theorem.

Recall that PCA does poorly when $d \gtrsim n$. So we use some sparsity constraints. Assume v is k-sparse, with $\Sigma = \theta v v^T + I_d$, and let

$$\hat{v} = \operatorname*{arg\,max}_{\substack{u \in S^{d-1} \\ \|u\|_0 = k}} u^T \hat{\Sigma} u,$$

where $\|u\|_0$ is the number of non-zero entries of u. This estimator cannot be computed in practice.

Theorem 18. *Let $Y \in \mathbb{R}^d$ be a random vector with $\mathbb{E}\, Y = 0$, $\mathbb{E}\, YY^T = I_d$, and $u^T Y \sim \mathrm{subG}(1)$ for all $u \in S^{d-1}$. Let $X_1, \dots, X_n$ be i.i.d. copies of $X = \Sigma^{1/2} Y$, where $\Sigma = \theta v v^T + I_d$, $v \in S^{d-1}$ and $\|v\|_0 = k \le d/2$. Then $\hat{v}$ satisfies*

$$\min_{\varepsilon \in \{\pm 1\}} \|\varepsilon \hat{v} - v\|_2 \lesssim \frac{1+\theta}{\theta} \left(\sqrt{\frac{k \log(ed/k) + \log(1/\delta)}{n}} \vee \frac{k \log(ed/k) + \log(1/\delta)}{n} \right)$$

with probability at least $1 - \delta$.

The bound is better, but the downside is that we need to know an upper bound on k.

Proof sketch. We have

$$v^T \Sigma v - \hat{v}^T \Sigma \hat{v} \le \left\langle \hat{\Sigma} - \Sigma,\ \hat{v}\hat{v}^T - v v^T \right\rangle,$$

where v and $\hat{v}$ are both k-sparse. So

$$\left\langle \hat{\Sigma} - \Sigma,\ \hat{v}\hat{v}^T - v v^T \right\rangle = \langle \hat{\Sigma}(S) - \Sigma(S),\ \hat{v}(S)\hat{v}(S)^T - v(S)v(S)^T \rangle, \qquad |S| \le 2k,$$

where S is the union of the supports of v and $\hat{v}$, and $M(S)$, $u(S)$ denote restrictions to the coordinates in S.

Lecture 13 — February 28, 2023

Prof. Matey Neykov *Scribe: Fan Pu Zeng*

1 Sparse PCA

Recall the result from last lecture.

Theorem 19. *Let $Y \in \mathbb{R}^d$ be a mean-zero 1-sub-Gaussian random vector such that $\mathbb{E}YY^T = I_d$ (that is, Y is* isotropic*). Let $X_1, \dots, X_n$ be i.i.d. with $X_i \overset{d}{=} \Sigma^{1/2}Y$, where*

$$\Sigma = \theta vv^T + I_d \quad \text{with} \quad \|v\|_0 \le k \le d/2 \qquad \text{(sparse)}.$$

Then $\hat{v}$ satisfies

$$\min_{\varepsilon \in \{\pm 1\}} \|\varepsilon \hat{v} - v\|_2 \lesssim \frac{1+\theta}{\theta}\left(\sqrt{\frac{k\log(ed/k) + \log(1/\delta)}{n}} \vee \frac{k\log(ed/k) + \log(1/\delta)}{n}\right)$$

with probability at least $1-\delta$, where

$$\hat{v} = \underset{\substack{w \in S^{d-1} \\ \|w\|_0 = k}}{\arg\max}\, w^T \hat{\Sigma} w$$

is the constrained maximal eigenvector*.*

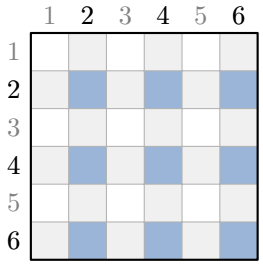
Proof sketch. We have

$$v^T\Sigma v - \hat{v}^T\Sigma\hat{v} \le \left\langle \hat{\Sigma} - \Sigma,\ \hat{v}\hat{v}^T - vv^T \right\rangle,$$

where $\hat{v}$ and v are both k-sparse. Let $S = \operatorname{supp}(\hat{v}) \cup \operatorname{supp}(v)$, so $|S| \le 2k$. So there exists $S \subset \{1, 2, \dots, d\}$ with $|S| \le 2k$ such that

$$\left\langle \hat{\Sigma} - \Sigma,\ \hat{v}\hat{v}^T - vv^T \right\rangle = \langle \hat{\Sigma}(S) - \Sigma(S),\ \hat{v}(S)\hat{v}(S)^T - v(S)v(S)^T \rangle.$$

Here $v(S)$ is v restricted to the coordinates in the set S, and $\hat{\Sigma}(S)$ is the submatrix of $\hat{\Sigma}$ with rows and columns in S.



$\hat{\Sigma}(S)$ for $S = \{2, 4, 6\}$

Figure 12: The submatrix $\hat{\Sigma}(S)$ keeps the entries of $\hat{\Sigma}$ whose row and column both lie in S.

So

$$v^T\Sigma v - \hat{v}^T\Sigma\hat{v} \le \left\|\hat{\Sigma}(S) - \Sigma(S)\right\|_{\mathrm{op}} \left\|\hat{v}(S)\hat{v}(S)^T - v(S)v(S)^T\right\|_1.$$

Following the same proof as for the Davis–Kahan $\sin\Theta$ theorem, we can get that

$$\sin(\sphericalangle(\hat{v}, v)) \le \underbrace{\frac{2}{(1+\theta)-1}}_{\text{eigengap of } \Sigma} \sup_{S:|S|\le 2k} \left\|\hat{\Sigma}(S) - \Sigma(S)\right\|_{\mathrm{op}}.$$

By a union bound,

$$\begin{aligned}\mathbb{P}\Big(\sup_{S:|S|\le 2k} \left\|\hat{\Sigma}(S) - \Sigma(S)\right\|_{\mathrm{op}} \ge t\left\|\Sigma\right\|_{\mathrm{op}}\Big) &\le \sum_{S:|S|\le 2k} \mathbb{P}\Big(\left\|\hat{\Sigma}(S) - \Sigma(S)\right\|_{\mathrm{op}} \ge t\left\|\Sigma(S)\right\|_{\mathrm{op}}\Big) \\ &\le \sum_{i=0}^{2k}\binom{d}{i} \cdot 144^{2k}\exp\Big(-\frac{n}{2}\Big(\frac{t^2}{(2\kappa_1)^2} \wedge \frac{t}{2\kappa_2}\Big)\Big),\end{aligned}$$

where the first inequality uses $\left\|\Sigma(S)\right\|_{\mathrm{op}} \le \left\|\Sigma\right\|_{\mathrm{op}}$, and the factor 144^{2k} comes from a covering of S^{2k-1}, as in the bound from last time.

Remark 9. Why is $\left\|\Sigma(S)\right\|_{\mathrm{op}} \le \left\|\Sigma\right\|_{\mathrm{op}}$? For $\left\|w\right\|_2 = 1$, consider $w^T\Sigma(S)w$. Let $v \in \mathbb{R}^d$ be the vector with $\mathrm{supp}(v) = S$ that equals w on the support S (and is 0 elsewhere). Then $\left\|v\right\|_2 = 1$ and $w^T\Sigma(S)w = v^T\Sigma v$. Taking the supremum over w gives $\left\|\Sigma(S)\right\|_{\mathrm{op}} \le \sup_{\|v\|_2=1} v^T\Sigma v = \left\|\Sigma\right\|_{\mathrm{op}}$.

How many choices of S are there? There are $\sum_{i=0}^{2k}\binom{d}{i}$. So the probability above is at most

$$\Big(\frac{ed}{2k}\Big)^{2k} \cdot 144^{2k}\exp\Big(-\frac{n}{2}\Big(\frac{t^2}{(2\kappa_1)^2} \wedge \frac{t}{2\kappa_2}\Big)\Big) \le \delta,$$

where the factor $144^{2k}\exp(\cdots)$ comes from the ε-net, each term of which is sub-exponential. We select t such that this inequality holds; the calculation is tedious and skipped.

2 Lasso

Suppose we observe $(y, X) \in \mathbb{R}^n \times \mathbb{R}^{n\times d}$, where

$$y = X\theta^* + w, \qquad w \in \mathbb{R}^n.$$

Consider solving

$$\hat{\theta} \in \arg\min_{\theta\in\mathbb{R}^d}\Big\{\frac{1}{2n}\left\|y - X\theta\right\|_2^2 + \lambda_n\left\|\theta\right\|_1\Big\}, \tag{1}$$

where $\lambda_n\left\|\theta\right\|_1$ is the ℓ_1 *regularizer*, which is convex.

Alternatively, one can solve

$$\min_{\theta\in\mathbb{R}^d}\Big\{\frac{1}{2n}\left\|y - X\theta\right\|_2^2\Big\} \quad \text{s.t.} \quad \left\|\theta\right\|_1 \le R \tag{2}$$

(the original formulation, by Tibshirani), or

$$\min_{\theta\in\mathbb{R}^d}\left\|\theta\right\|_1 \quad \text{s.t.} \quad \frac{1}{2n}\left\|y - X\theta\right\|_2^2 \le b^2. \tag{3}$$

These are all equivalent by Lagrangian duality!

2.1 The restricted eigenvalue condition

For some $\alpha \geq 1$ and $S \subset [d] = \{1, \ldots, d\}$, define the *cone*

$$C_\alpha(S) = \{\Delta \in \mathbb{R}^d \mid \|\Delta_{S^c}\|_1 \leq \alpha \|\Delta_S\|_1\},$$

where S^c is the complement of S.

Remark 10. What does "cone" mean? If $\Delta \in C_\alpha(S)$, then $k\Delta \in C_\alpha(S)$ for all $k \geq 0$.

Definition (Restricted eigenvalue condition)**.** X satisfies the *restricted eigenvalue condition* (RE) over S with parameters (κ, α) if

$$\frac{1}{n}\|X\Delta\|_2^2 \geq \kappa \|\Delta\|_2^2 \qquad \forall\, \Delta \in C_\alpha(S).$$

Equivalently,

$$\Delta^T \frac{X^T X}{n} \Delta \geq \kappa \|\Delta\|_2^2, \qquad \text{where} \quad \frac{X^T X}{n} = \frac{1}{n}\sum_{i=1}^n x_i x_i^T,$$

so this lower bounds the eigenvalue of $X^T X/n$, but only in the directions in the cone. We really need the restriction $\Delta \in C_\alpha(S)$ because it shows up in the proof.

2.2 Why restricted eigenvalues

Let's take the constrained lasso (2), where $R = \|\theta^*\|_1$ and θ^* is the true value of θ. Let

$$\mathcal{L}_n(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2.$$

We know $\mathcal{L}_n(\hat{\theta}) \leq \mathcal{L}_n(\theta^*)$. As $n \to \infty$, we expect that $\mathcal{L}_n(\hat{\theta}) \approx \mathcal{L}_n(\theta^*)$.

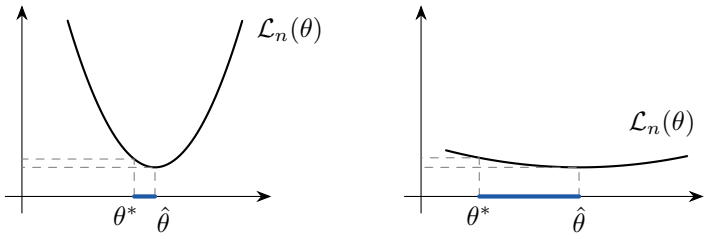


Figure 13: The same small gap $\mathcal{L}_n(\theta^*) - \mathcal{L}_n(\hat{\theta})$ keeps θ^* close to the minimizer $\hat{\theta}$ when the loss is sharply curved (left), but not when the landscape is flat (right).

So θ^* can be far away from the minimizer $\hat{\theta}$ when the landscape is flat. The shape of the function $\mathcal{L}_n(\theta)$ is related to its Hessian matrix

$$\nabla^2 \mathcal{L}_n(\theta) = \frac{X^T X}{n},$$

which is easy to compute for least squares. The curvature is like the smallest eigenvalue, so we don't want it to be too small. We want

$$\Delta^T \frac{X^T X}{n} \Delta \geq \kappa \|\Delta\|_2^2$$

for all Δ. But this is impossible when $d > n$. Hence we have the restricted eigenvalue condition.

2.3 Error bounds

Theorem 20. *Suppose*

(A1) θ^* *is sparse, and is supported on a set* $S \subset [d]$*;*

(A2) the design matrix satisfies RE with $S = \mathrm{supp}(\theta^*)$ *and parameters* $(\kappa, 3)$*, that is, with restricted eigenvalue* κ *and* $\alpha = 3$ *in* $\|\Delta_{S^c}\|_1 \le \alpha \|\Delta_S\|_1$.

Let $|S| = s$*. Then:*

(a) Any solution to (1) with $\lambda_n \ge 2 \left\| \frac{X^T w}{n} \right\|_\infty$ *satisfies*

$$\left\| \hat{\theta} - \theta^* \right\|_2 \le \frac{3}{\kappa} \sqrt{s}\, \lambda_n.$$

(b) Any solution of (2) with $R = \|\theta^*\|_1$ *satisfies*

$$\left\| \hat{\theta} - \theta^* \right\|_2 \le \frac{4}{\kappa} \sqrt{s} \left\| \frac{X^T w}{n} \right\|_\infty.$$

(c) Any solution of (3) with $b^2 \ge \frac{\|w\|_2^2}{2n}$ *satisfies*

$$\left\| \hat{\theta} - \theta^* \right\|_2 \le \frac{4}{\kappa} \sqrt{s} \left\| \frac{X^T w}{n} \right\|_\infty + \frac{2}{\kappa} \sqrt{b^2 - \frac{\|w\|_2^2}{2n}}.$$

Note that we don't observe w: it is the noise. Suppose, for example, that $w_i \sim N(0, \sigma^2)$ for $i = 1, \ldots, n$. The difficulty is that we don't know what w is, but all three conditions depend on w. Also, choosing $R = \|\theta^*\|_1$ in (b) is a big assumption.

In all three cases (a), (b), (c), we have

$$\left\| \hat{\theta} - \theta^* \right\|_1 \le 4\sqrt{s} \left\| \hat{\theta} - \theta^* \right\|_2.$$

2.4 Square-root lasso

The *square-root lasso* minimizes

$$\frac{1}{2n} \|y - X\theta\|_2 + \lambda_n \|\theta\|_1.$$

2.5 Toward the proof of part (a)

Call $\hat{\Delta} = \hat{\theta} - \theta^*$, and let

$$\mathcal{L}_n(\theta, \lambda_n) = \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda_n \|\theta\|_1.$$

Since $\hat{\theta}$ is the minimizer,

$$\mathcal{L}_n(\hat{\theta}, \lambda_n) \leq \mathcal{L}_n(\theta^*, \lambda_n) = \frac{1}{2n} \|w\|_2^2 + \lambda_n \|\theta^*\|_1 .$$

Since $y - X\hat{\theta} = w - X\hat{\Delta}$, the left-hand side is

$$\frac{1}{2n} \left\|w - X\hat{\Delta}\right\|_2^2 + \lambda_n \left\|\hat{\theta}\right\|_1 = \frac{1}{2n}\Big(\|w\|_2^2 - 2w^T X\hat{\Delta} + \left\|X\hat{\Delta}\right\|_2^2\Big) + \lambda_n \left\|\hat{\theta}\right\|_1 .$$

Rearranging,

$$\frac{1}{2n} \left\|X\hat{\Delta}\right\|_2^2 \leq \frac{w^T X\hat{\Delta}}{n} + \lambda_n(\|\theta^*\|_1 - \left\|\hat{\theta}\right\|_1).$$

We will get a lower bound on the left-hand side using the restricted eigenvalue condition.

36-709: Advanced Statistical Theory I Spring 2023

Lecture 14 — March 2, 2023

Prof. Matey Neykov *Scribe: Fan Pu Zeng*

1 Lasso Error Bounds (Continued)

We continue from last time, with the proof of part (a) of the lasso error bound: if θ^* is supported on S with $|S| = s$, X satisfies RE over S with parameters $(\kappa, 3)$, and $\lambda_n \geq 2\left\|\frac{X^T w}{n}\right\|_\infty$, then any solution of the lasso program

$$\min_{\theta\in\mathbb{R}^d}\Big\{\frac{1}{2n}\|y - X\theta\|_2^2 + \lambda_n \|\theta\|_1\Big\} \tag{1}$$

satisfies $\left\|\hat{\theta} - \theta^*\right\|_2 \leq \frac{3}{\kappa}\sqrt{s}\,\lambda_n$.

Proof of part (a). Let

$$\mathcal{L}(\theta, \lambda_n) = \frac{1}{2n}\|y - X\theta\|_2^2 + \lambda_n \|\theta\|_1, \qquad \hat{\Delta} = \hat{\theta} - \theta^*.$$

Then

$$\mathcal{L}(\hat{\theta}, \lambda_n) \leq \mathcal{L}(\theta^*, \lambda_n) = \frac{1}{2n}\|w\|_2^2 + \lambda_n \|\theta^*\|_1,$$

and

$$\mathcal{L}(\hat{\theta}, \lambda_n) = \frac{1}{2n}\Big(\|w\|_2^2 - 2w^T X\hat{\Delta} + \left\|X\hat{\Delta}\right\|_2^2\Big) + \lambda_n \left\|\hat{\theta}\right\|_1,$$

so

$$\frac{1}{2n}\left\|X\hat{\Delta}\right\|_2^2 \leq \frac{1}{n}w^T X\hat{\Delta} + \lambda_n(\|\theta^*\|_1 - \left\|\hat{\theta}\right\|_1).$$

For the last term,

$$\|\theta^*\|_1 - \left\|\hat{\theta}\right\|_1 = \|\theta_S^*\|_1 - \underbrace{\left\|\theta_S^* + \hat{\Delta}_S\right\|_1}_{\|\hat{\theta}_S\|_1} - \underbrace{\left\|\hat{\Delta}_{S^c}\right\|_1}_{\|\hat{\theta}_{S^c}\|_1},$$

because $\theta_{S^c}^* = 0$, so

$$\hat{\Delta}_{S^c} = \hat{\theta}_{S^c} - \theta_{S^c}^* = \hat{\theta}_{S^c},$$

and because the ℓ_1 norm is decomposable: $\left\|\hat{\theta}_S\right\|_1 + \left\|\hat{\theta}_{S^c}\right\|_1 = \left\|\hat{\theta}\right\|_1$.

Therefore

$$\begin{aligned} 0 \leq \frac{1}{n}\left\|X\hat{\Delta}\right\|_2^2 &\leq \frac{2w^T X\hat{\Delta}}{n} + 2\lambda_n(\|\theta_S^*\|_1 - \left\|\theta_S^* + \hat{\Delta}_S\right\|_1 - \left\|\hat{\Delta}_{S^c}\right\|_1) \\ &\leq 2\left\|\frac{X^T w}{n}\right\|_\infty \left\|\hat{\Delta}\right\|_1 + 2\lambda_n\Big(\|\theta_S^*\|_1 - (\|\theta_S^*\|_1 - \left\|\hat{\Delta}_S\right\|_1) - \left\|\hat{\Delta}_{S^c}\right\|_1\Big). \end{aligned}$$

The first term uses Hölder's inequality. The second uses the triangle inequality,

$$-\left\|\theta_S^* + \hat{\Delta}_S\right\|_1 \leq -\|\theta_S^*\|_1 - \left\|\hat{\Delta}_S\right\|_1.$$

Using the upper bound $2\left\|\frac{X^T w}{n}\right\|_\infty \leq \lambda_n$,

$$\begin{aligned}
\frac{1}{n}\left\|X\hat{\Delta}\right\|_2^2 &\leq \lambda_n \left\|\hat{\Delta}\right\|_1 + 2\lambda_n(\left\|\hat{\Delta}_S\right\|_1 - \left\|\hat{\Delta}_{S^c}\right\|_1) \\
&= \lambda_n(\left\|\hat{\Delta}_S\right\|_1 + \left\|\hat{\Delta}_{S^c}\right\|_1) + 2\lambda_n(\left\|\hat{\Delta}_S\right\|_1 - \left\|\hat{\Delta}_{S^c}\right\|_1) \\
&= 3\lambda_n \left\|\hat{\Delta}_S\right\|_1 - \lambda_n \left\|\hat{\Delta}_{S^c}\right\|_1.
\end{aligned}$$

Since the left-hand side is nonnegative,

$$\left\|\hat{\Delta}_{S^c}\right\|_1 \leq 3\left\|\hat{\Delta}_S\right\|_1,$$

that is, $\hat{\Delta} \in C_\alpha(S)$ for $\alpha = 3$. So by RE,

$$\begin{aligned}
\kappa\left\|\hat{\Delta}\right\|_2^2 &\leq \frac{1}{n}\left\|X\hat{\Delta}\right\|_2^2 \\
&\leq 3\lambda_n \left\|\hat{\Delta}_S\right\|_1 \\
&\leq 3\lambda_n\sqrt{s}\left\|\hat{\Delta}_S\right\|_2 \\
&\leq 3\lambda_n\sqrt{s}\left\|\hat{\Delta}\right\|_2,
\end{aligned}$$

and therefore

$$\left\|\hat{\Delta}\right\|_2 \leq \frac{3\lambda_n\sqrt{s}}{\kappa}.$$

■

We can also bound the ℓ_1 error:

$$\begin{aligned}
\left\|\hat{\Delta}\right\|_1 &= \left\|\hat{\Delta}_S\right\|_1 + \left\|\hat{\Delta}_{S^c}\right\|_1 \\
&\leq 4\left\|\hat{\Delta}_S\right\|_1 \\
&\leq 4\sqrt{s}\left\|\hat{\Delta}_S\right\|_2 \\
&\leq 4\sqrt{s}\left\|\hat{\Delta}\right\|_2.
\end{aligned}$$

Applying the previous result on $\left\|\hat{\Delta}\right\|_2$,

$$\left\|\hat{\Delta}\right\|_1 \leq \frac{12\lambda_n s}{\kappa}.$$

Example 14. Suppose X satisfies the RE property, the noise is $w_i \sim N(0, \sigma^2)$ i.i.d., and the columns X_j of X satisfy

$$\max_{j\in[d]} \frac{\|X_j\|_2}{\sqrt{n}} \leq C.$$

Then $\left\|\frac{X^T w}{n}\right\|_\infty$ is the maximum of d zero-mean Gaussian variables whose variances are at most $\frac{C^2\sigma^2}{n}$.

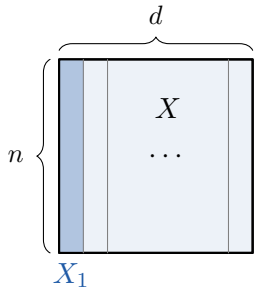


Figure 14: The design matrix X; each coordinate of $X^T w$ is the inner product of one column, such as X_1, with the noise w.

For the first column,

$$\frac{X_1^T w}{n} \sim N\Big(0, \frac{\|X_1\|_2^2}{n^2}\Big), \qquad \frac{\|X_1\|_2^2}{n^2} \le \frac{C^2\sigma^2}{n}.$$

So from a maximal inequality,

$$\mathbb{P}\Big(\Big\|\frac{X^T w}{n}\Big\|_\infty \ge C\sigma\Big(\sqrt{\frac{\log d}{n}} + \delta\Big)\Big) \le 2e^{-n\delta^2/2} \qquad \forall\, \delta > 0.$$

So if $\lambda_n = 2C\sigma(\sqrt{\frac{\log d}{n}} + \delta)$, then

$$\Big\|\hat\theta - \theta^*\Big\|_2 \le \frac{6C\sigma}{\kappa}\sqrt{s}\Big(\sqrt{\frac{\log d}{n}} + \delta\Big)$$

with probability at least $1 - 2e^{-n\delta^2/2}$.

This is nice, since $\Big\|\hat\theta - \theta^*\Big\|_2$ is small even in very high dimensions. The rate

$$\sqrt{\frac{s\log d}{n}}$$

is the "fast" lasso rate.

2 Restricted Eigenvalues for Random Designs

Theorem 21. *Let $X \in \mathbb{R}^{n\times d}$ have rows $X_i \overset{i.i.d.}{\sim} N(0,\Sigma)$. There exist universal constants $c_1 < 1 < c_2$ such that*

$$\frac{\|X\theta\|_2^2}{n} \ge c_1\Big\|\sqrt{\Sigma}\,\theta\Big\|_2^2 - c_2\,\rho^2(\Sigma)\frac{\log d}{n}\,\|\theta\|_1^2 \qquad \forall\,\theta\in\mathbb{R}^d$$

with probability at least $1 - \frac{e^{-n/32}}{1-e^{-n/32}}$, where $\rho(\Sigma) = \max_j \Sigma_{jj}$.

We want

$$\frac{\|X\theta\|_2^2}{n} \ge \kappa\,\|\theta\|_2^2 \qquad \forall\,\theta \in C_\alpha(S).$$

If $\theta \in C_\alpha(S)$, then $\|\theta_{S^c}\|_1 \le \alpha \|\theta_S\|_1$, so by the triangle inequality

$$\begin{aligned}\|\theta\|_1 &\le (1+\alpha)\|\theta_S\|_1 \\ &\le (1+\alpha)\sqrt{s}\ \|\theta_S\|_2 \\ &\le (1+\alpha)\sqrt{s}\ \|\theta\|_2 .\end{aligned}$$

Also

$$\left\|\sqrt{\Sigma}\,\theta\right\|_2^2 \ge \lambda_{\min}(\Sigma)\,\|\theta\|_2^2 .$$

Plugging both into the theorem,

$$\frac{\|X\theta\|_2^2}{n} \ge \left(c_1\lambda_{\min}(\Sigma) - c_2\,\rho^2(\Sigma)(1+\alpha)^2\frac{s\log d}{n}\right)\|\theta\|_2^2 \ge \frac{c_1\lambda_{\min}(\Sigma)}{2}\,\|\theta\|_2^2$$

for all $\theta \in C_\alpha(S)$, if

$$\frac{c_1\lambda_{\min}(\Sigma)}{2} > c_2\,\rho^2(\Sigma)(1+\alpha)^2\frac{s\log d}{n}.$$

So if s is sufficiently small, then the random design matrix X will satisfy RE w.h.p.

3 Prediction Error

Definition (Mean squared prediction error)**.** The *mean squared prediction error* (MSPE) is

$$\text{MSPE} := \frac{\left\|X(\hat\theta - \theta^*)\right\|_2^2}{n} = n^{-1}\sum_{i=1}^n \left\langle X_i, \hat\theta - \theta^*\right\rangle^2 .$$

Hold X fixed and generate new noise, $\tilde y = X\theta^* + \tilde w$. Then

$$\frac{1}{n}\,\mathbb{E}\left\|\tilde y - X\hat\theta\right\|_2^2 = \underbrace{\frac{1}{n}\left\|X(\hat\theta - \theta^*)\right\|_2^2}_{\text{MSPE}} + \sigma^2,$$

where σ^2 is a constant.

Theorem 22. *Consider the lasso, that is, program (1), with* $\lambda_n \ge 2\left\|\frac{X^T w}{n}\right\|_\infty$.

(a) (No assumptions.) Any optimal $\hat\theta$ *satisfies*

$$\frac{\left\|X(\hat\theta - \theta^*)\right\|_2^2}{n} \le 12\,\|\theta^*\|_1\,\lambda_n.$$

This is the slow rate.

(b) (Sparsity.) If θ^* *is supported on* S *with* $|S| = s$*, and* X *satisfies* $(\kappa, 3)$*-RE, then*

$$\frac{\left\|X(\hat\theta - \theta^*)\right\|_2^2}{n} \le \frac{9s\lambda_n^2}{\kappa}.$$

This is the fast rate *(here assuming* λ_n *is small).*

Proof. (a) As before,

$$\begin{aligned} 0 \le \frac{1}{2n}\left\|X\hat{\Delta}\right\|_2^2 &\le \frac{w^T X\hat{\Delta}}{n} + \lambda_n(\|\theta^*\|_1 - \left\|\hat{\theta}\right\|_1) \\ &\le \left\|\frac{X^T w}{n}\right\|_\infty \left\|\hat{\Delta}\right\|_1 + \lambda_n(\|\theta^*\|_1 - \left\|\hat{\theta}\right\|_1) \qquad (\star) \\ &\le \frac{\lambda_n}{2}(\left\|\hat{\theta}\right\|_1 + \|\theta^*\|_1) + \lambda_n(\|\theta^*\|_1 - \left\|\hat{\theta}\right\|_1), \end{aligned}$$

similar to the previous proof. Rearranging, $0 \le \frac{3\lambda_n}{2}\|\theta^*\|_1 - \frac{\lambda_n}{2}\left\|\hat{\theta}\right\|_1$, so

$$\left\|\hat{\theta}\right\|_1 \le 3\|\theta^*\|_1 \quad \Longrightarrow \quad \left\|\hat{\Delta}\right\|_1 \le \left\|\hat{\theta}\right\|_1 + \|\theta^*\|_1 \le 4\|\theta^*\|_1 .$$

Going back to $(\star)$,

$$\frac{\left\|X\hat{\Delta}\right\|_2^2}{2n} \le \frac{\lambda_n}{2}\left\|\hat{\Delta}\right\|_1 + \lambda_n \underbrace{(\|\theta^*\|_1 - \left\|\theta^* + \hat{\Delta}\right\|_1)}_{\le \|\theta^*\|_1 - (\|\theta^*\|_1 - \left\|\hat{\Delta}\right\|_1)} \le \frac{3\lambda_n}{2}\left\|\hat{\Delta}\right\|_1 \le \frac{3\lambda_n \cdot 4\|\theta^*\|_1}{2}.$$

The left-hand side is half the MSPE, so cancelling the 2 gives (a).

(b) From the proof of the previous theorem, we had

$$\frac{1}{n}\left\|X\hat{\Delta}\right\|_2^2 \le 3\lambda_n \left\|\hat{\Delta}_S\right\|_1 \le 3\lambda_n\sqrt{s}\,\left\|\hat{\Delta}_S\right\|_2 .$$

Under RE and sparsity, the previous theorem showed

$$\left\|\hat{\Delta}\right\|_2 \le \frac{3\lambda_n\sqrt{s}}{\kappa},$$

so

$$\underbrace{\frac{1}{n}\left\|X\hat{\Delta}\right\|_2^2}_{\text{MSPE}} \le \frac{9\lambda_n^2 s}{\kappa}. \qquad \blacksquare$$

Example 15. Let $w_i \sim N(0,\sigma^2)$, $\max_{j\in[d]}\frac{\|X_j\|_2}{\sqrt{n}} \le C$, and

$$\lambda_n = 2C\sigma\Big(\sqrt{\frac{\log d}{n}} + \delta\Big).$$

Then

$$\frac{\left\|X(\hat{\theta}-\theta^*)\right\|_2^2}{n} \le 24\|\theta^*\|_1 C\sigma\Big(\sqrt{\frac{\log d}{n}} + \delta\Big) \sim \frac{1}{\sqrt{n}}.$$

No RE or sparsity is required. If θ^* is sparse and we have RE, then

$$\left\|X(\hat{\theta}-\theta^*)\right\|_2^2 \lesssim \frac{C^2\sigma^2}{\kappa}\Big(\frac{2s\log d}{n} + s\delta^2\Big) \sim \frac{1}{n}.$$

Lecture 15 — March 14, 2023

Prof. Matey Neykov *Scribe: Fan Pu Zeng*

1 Variable Selection

Goal: figure out where θ^* (the regression coefficients) has non-zero coefficients. The model is

$$y = X\theta^* + w,$$

where w is noise and θ^* is s-sparse. We don't really need the coefficients that correspond to 0. This is a different problem from estimation or prediction.

Let θ^* be supported on $S = S(\theta^*)$, the *support* (the set of non-zero coefficients). Can we find $\hat{\theta}$ such that $S(\hat{\theta}) = S(\theta^*)$? If we know $\left\|\hat{\theta} - \theta^*\right\|_2$ is small, that doesn't necessarily mean that $S(\hat{\theta}) = S(\theta^*)$.

1.1 Assumptions

Let $S = S(\theta^*)$, and let X_S be the columns of X indexed by S.

(A3) **Lower eigenvalue assumption.**

$$\lambda_{\min}\Big(\frac{X_S^T X_S}{n}\Big) \geq c_{\min} > 0.$$

(A4) **Mutual incoherence.** There exists $\alpha \in [0, 1)$ such that

$$\max_{j \in S^c} \left\|(X_S^T X_S)^{-1} X_S^T X_j\right\|_1 \leq \alpha.$$

1.2 Intuition for the conditions

(A3) is about identifiability. It says that X_S is a full-rank matrix: if the smallest eigenvalue is 0, then X_S is not full rank, and we could find a different θ. If X_S is not a full-rank matrix, there will be many θ^*'s supported on S that produce the same data.

(A4) tries to impose that the columns of X are not too similar. Suppose we take X_j and we try to explain X_j with some linear combination of vectors in $\text{col}(X_S)$:

$$\hat{w} = \arg\min_w \|X_j - X_S w\|_2^2 = (X_S^T X_S)^{-1} X_S^T X_j$$

by regression. We would like $\hat{w}$ to be small. (The reason why it is required is that it shows up in the proof.)

1.3 Support recovery

Theorem 23 (MJW). *Suppose (A3) and (A4) hold. Then if λ_n is sufficiently big, the lasso has a unique solution $\hat{\theta}$, with support $\hat{S} \subseteq S$, and*

$$\left\|\hat{\theta}_S - \theta^*_S\right\|_\infty \leq \underbrace{\left\|\left(\frac{X_S^T X_S}{n}\right)^{-1}\frac{X_S^T w}{n}\right\|_\infty + \left|\!\left|\!\left|\frac{X_S^T X_S}{n}\right|\!\right|\!\right|_\infty \lambda_n}_{B(\lambda_n, X)},$$

where, for an $s \times s$ matrix A,

$$|||A|||_\infty = \max_{j\in[s]} \sum_{i=1}^{s} |A_{ij}|.$$

If $|\theta^*_i| > B(\lambda_n, X)$, then $|\hat{\theta}_i| > 0$ by the triangle inequality, since $|\hat{\theta}_i| \geq |\theta^*_i| - |\hat{\theta}_i - \theta^*_i| > B(\lambda_n, X) - B(\lambda_n, X) = 0$, so $i \in \hat{S}$. Hence, under the *minimal signal assumption*

$$\min_{i\in S} |\theta^*_i| > B(\lambda_n, X),$$

we get $\hat{S} = S$: if the smallest $|\theta^*_i|$ exceeds $B(\lambda_n, X)$, then every coordinate i of $\hat{\theta}_S$ corresponds to a non-zero coefficient, so $\hat{S} = S$ in those coordinates.

The two vectors look like

$$\theta^* = (0, \ldots, 0, \underbrace{\neq 0, \neq 0, \ldots}_{S}, 0, 0), \qquad \hat{\theta} = (0, \ldots, 0, \underbrace{\neq 0, \neq 0, \ldots}_{\hat{S}\subseteq S}, 0, 0).$$

So

$$\left\|\hat{\theta} - \theta^*\right\|_2 = \left\|\hat{\theta}_S - \theta^*_S\right\|_2 \leq \sqrt{s}\left\|\hat{\theta}_S - \theta^*_S\right\|_\infty.$$

Bounding the ℓ_∞ norm is a very strong statement: it also bounds the ℓ_1 and ℓ_2 norms.

2 Inference after the Lasso

How do we perform inference after the lasso? One way, taken from Javanmard and Montanari (JMLR 2014), is as follows.

1. Start with
$$\hat{\theta} = \arg\min_\theta \frac{1}{2n}\|y - X\theta\|_2^2 + \lambda_n \|\theta\|_1,$$
where λ_n is the tuning parameter from the lasso.
2. Set $\hat{\Sigma} = \frac{X^T X}{n}$.
3. Solve the following problem:
$$\min_m \; m^T \hat{\Sigma} m \quad \text{s.t.} \quad \left\|\hat{\Sigma} m - e_j\right\|_\infty \leq \mu_n,$$
where μ_n is a tuning parameter and $e_j = (0, 0, \ldots, 1, 0, \ldots, 0)^T$ has its 1 in coordinate j. Call the solution $\hat{m}_j$.

4. Set

$$\hat{\theta}_j^d = \hat{\theta}_j + \frac{1}{n}\hat{m}_j^T X^T (y - X\hat{\theta}).$$

Here d is for *debiased*. It turns out that one can show $\hat{\theta}_j^d$ has an asymptotically normal distribution. A nicety is that there is no sparsity assumption on Σ^{-1}.

We present the material in a stylized fashion; the full proof is more general (read the paper).

Theorem 24. *Suppose the rows of X are i.i.d. $N(0,\Sigma)$ and $\rho^2(\Sigma)\frac{s\log d}{n}$ is sufficiently small, where $\rho^2(\Sigma)$ is the maximum diagonal entry of Σ. (This implies that $\frac{X^TX}{n}$ satisfies RE w.h.p.) Let $\lambda_{\min}(\Sigma)$ and $\lambda_{\max}(\Sigma)$ be bounded. If $w_i \sim N(0,\sigma^2)$ and*

$$\frac{s\log d}{\sqrt{n}} = o(1),$$

where s is the sparsity of θ^* and d is the ambient dimension, then

$$\sqrt{n}(\hat{\theta}_j^d - \theta_j^*) \mid X \stackrel{d}{=} N\big(0, \sigma^2 \hat{m}_j^T \hat{\Sigma} \hat{m}_j\big) + o_p(1),$$

where $o_p(1)$ is little-o in probability. Here λ_n is selected as in MJW, with

$$\lambda_n \asymp \sigma\sqrt{\frac{\log d}{n}}, \qquad \mu_n \asymp \sqrt{\frac{\log d}{n}}.$$

If we just want $\left\|\hat{\theta} - \theta^*\right\|_2$ to be small, it is sufficient that $\frac{s\log d}{n} = o(1)$, since

$$\left\|\hat{\theta} - \theta^*\right\|_2 \lesssim \sqrt{\frac{s\log d}{n}} = o(1).$$

Recall that

$$\hat{m}_j = \arg\min_{m_j} \; m_j^T \hat{\Sigma} m_j \quad \text{s.t.} \quad \left\|\hat{\Sigma} m_j - e_j\right\|_\infty \le \mu_n,$$

which is a convex optimization problem. We could use any m_j that satisfies the constraint, but the minimizer is what gives the asymptotic performance.

Proof sketch. Since $y - X\hat{\theta} = X(\theta^* - \hat{\theta}) + w$, the debiased estimator satisfies

$$\begin{aligned}\hat{\theta}_j^d - \theta_j^* &= \underbrace{\hat{\theta}_j - \theta_j^*}_{e_j^T(\hat{\theta}-\theta^*)} + \frac{1}{n}\hat{m}_j^T X^T\big(X(\theta^* - \hat{\theta})\big) + \underbrace{\frac{\hat{m}_j^T X^T w}{n}}_{z} \\ &= \underbrace{(\hat{m}_j^T\hat{\Sigma} - e_j^T)(\theta^* - \hat{\theta})}_{\Delta_j} + z.\end{aligned}$$

Conditional on X,

$$\sqrt{n}\, z \mid X \stackrel{d}{=} N\big(0, \sigma^2 \hat{m}_j^T \hat{\Sigma} \hat{m}_j\big),$$

since $\sqrt{n}\, z = \hat{m}_j^T X^T w/\sqrt{n}$ is Gaussian with variance $\sigma^2 \hat{m}_j^T X^T X \hat{m}_j/n = \sigma^2 \hat{m}_j^T \hat{\Sigma} \hat{m}_j$. All we need to show is

$$\sqrt{n}\,\Delta_j = o_p(1).$$

By Hölder's inequality,

$$\sqrt{n}\,|\Delta_j| \le \sqrt{n}\,\underbrace{\left\|\hat{\Sigma}\hat{m}_j - e_j\right\|_\infty}_{\le \mu_n} \left\|\theta^* - \hat{\theta}\right\|_1 \le \sqrt{n}\,\mu_n \cdot \frac{12 s \lambda_n}{\kappa} = o_p(1),$$

using the lasso bound on $\left\|\theta^* - \hat{\theta}\right\|_1$ from MJW, and $\lambda_n \asymp \sigma\sqrt{\frac{\log d}{n}}$. We want μ_n to be small to satisfy this, but big enough for the condition $\left\|\hat{\Sigma}\hat{m}_j - e_j\right\|_\infty \le \mu_n$ to be satisfiable.

How to choose μ_n? Check when $m_j = \Sigma^{-1} e_j$ is in the constraint set. We have

$$\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^n X_i X_i^T m_j\Big] - e_j = 0,$$

since $\mathbb{E}\,\hat{\Sigma} = \Sigma$. Using sub-exponential concentration and a union bound, you can show

$$\left\|\frac{1}{n}\sum_{i=1}^n X_i X_i^T m_j - e_j\right\|_\infty \lesssim \sqrt{\frac{\log d}{n}}.$$

So with $\mu_n \asymp \sqrt{\frac{\log d}{n}}$, the bound above becomes

$$\sqrt{n}\,\mu_n \cdot \frac{12 s \lambda_n}{\kappa} \asymp \sqrt{n}\sqrt{\frac{\log d}{n}} \cdot s\sqrt{\frac{\log d}{n}} = \frac{s \log d}{\sqrt{n}} = o(1).$$