

Statistics 210B Lecture Notes

High-Dimensional Statistics

Professor: Song Mei
Scribe: Daniel Raban

Spring 2022

Contents

1	Introduction to High-Dimensional Statistics	6
1.1	Overview of the course	6
1.2	A motivating example: sparse estimation	6
1.3	Relationships with other statistical topics	9
2	Basic Concentration Inequalities	10
2.1	Concentration inequalities for sample averages	10
2.2	Markov's inequality	10
2.3	Chebyshev's inequality	11
2.4	Chernoff's inequality	13
2.5	Comparison of inequalities	15
2.6	Applying union bounds	16
3	Sub-Gaussian and Sub-Exponential Random Variables	18
3.1	Sub-Gaussian random variables	18
3.2	Hoeffding's inequality	19
3.3	Examples of sub-Gaussian random variables	20
3.4	Equivalent characterizations of sub-Gaussianity	21
3.5	Sub-exponential random variables	22
4	Bernstein's Inequality, the Johnson-Lindenstass Lemma, and More Concentration Inequalities	25
4.1	Bernstein condition for sub-exponentiality	25
4.2	Bernstein's inequality	26
4.3	An application: the Johnson-Lindenstrass Lemma	27
4.4	Equivalent characterizations of sub-exponentiality	28

4.5	Bennett's inequality	29
4.6	Maximal inequality	30
4.7	Truncation argument	30
5	Martingale Concentration Inequalities	32
5.1	Motivation and overview	32
5.2	Doob's martingale representation of $f(X_1, \dots, X_n)$	33
5.3	Martingale concentration	34
5.4	Applications	36
5.5	Freedman's inequality	38
5.6	Maximal Azuma-Hoeffding inequality	39
6	Gaussian Concentration	40
6.1	Freedman's inequality	40
6.2	Maximal Azuma-Hoeffding inequality	40
6.3	Gaussian concentration	41
6.4	Examples of Gaussian concentration	42
6.5	Gaussian complexity	43
6.6	Proof of the Gaussian concentration inequality (interpolation method)	44
6.7	Other methods for establishing concentration	45
7	Concentration Inequalities for Convex Functions	47
7.1	Overview	47
7.2	Concentration of separately convex, Lipschitz functions	48
7.3	Concentration of convex Lipschitz functions	48
7.4	Applications	49
7.4.1	Rachemacher complexity	49
7.4.2	Operator norm	50
7.5	Proof techniques: the Herbst argument and transportation	50
7.6	Concentration of Lipschitz functions of log-concave random variables	51
7.7	Proof technique: the isoperimetric inequality	51
8	Introduction to Empirical Process Theory	53
8.1	Convergence of CDFs and the Glivenko-Cantelli theorem	53
8.2	Uniform laws for more general function classes	54
8.3	Rademacher complexity	56
8.4	An upper bound of $\ \mathbb{P}_n - \mathbb{P}\ _{\mathcal{F}}$ via $\mathcal{R}_n(\mathcal{F})$	58
9	Bounds on Rademacher Complexity of Function Classes	60
9.1	Bounding $\mathbb{E}[\ \mathbb{P}_n - \mathbb{P}\ _{\mathcal{F}}]$ in terms of Rademacher complexity	60
9.2	Aside: the maximal inequality	62
9.3	Bounding Rademacher complexity using the maximal inequality	63

10 VC Dimension, Covering, and Packing	66
10.1 VC dimension	66
10.2 The metric entropy method	68
10.3 Covering and packing	68
11 Volume Bounds for Metric Entropy and the Chaining Method	71
11.1 Recap: one-step discretization bound	71
11.2 Volume bounds for metric entropy	71
11.3 The chaining method	74
12 The Metric Entropy Method for Function Spaces	77
12.1 Recap: controlling complexity via chaining	77
12.2 One step discretization and chaining bounds for Rademacher complexity of function classes	77
12.3 Useful metrics on $\mathcal{F} \subseteq L^1(\mathbb{P})$	78
12.4 The uniform entropy bound for empirical processes	80
12.5 Examples of bounding Rademacher complexity for different covering numbers	81
13 Examples of Rademacher Complexity Bounds for Function Classes	83
13.1 Recap: chaining bounds for Rademacher complexity of function classes	83
13.2 Examples of upper bounds for parametric and nonparametric function classes	84
13.3 Boolean function classes	86
13.4 Contraction inequalities	87
13.5 Further topics: Orlicz processes and bracketing numbers	89
14 Concentration of Sample Covariance of Gaussian Random Vectors	90
14.1 Eigenvalues of sample covariance of Gaussian random vectors	90
14.2 Proof of the Sudakov-Fernique inequality	93
14.3 More on Gaussian comparison inequalities	93
14.4 Concentration of sub-Gaussian sample covariance	94
15 Concentration of Sample Covariance of Sub-Gaussian and Bounded Random Vectors	96
15.1 Concentration of sample covariance of sub-Gaussian vectors	96
15.2 Concentration of sample covariance of bounded random vectors	98
15.3 Matrix Hoeffding/Bernstein inequality	98
15.3.1 Matrix Chernoff inequality	99
15.3.2 Sub-Gaussian and sub-exponential matrices	100
15.3.3 Tensorization of the matrix MGF	101

16 Introduction to Sparse Linear Regression	103
16.1 High-dimensional linear regression	103
16.2 Recovery in the noiseless setting	104
16.3 A sufficient condition for exact recovery	105
17 Sufficient Conditions for Exact Recovery in Sparse Linear Regression and Introduction to Noisy, Sparse Linear Regression	108
17.1 Recap: sparse linear regression via the restricted nullspace condition	108
17.2 Two sufficient conditions for the restricted nullspace property	108
17.2.1 The pairwise incoherence condition	109
17.2.2 The restricted isometry property	110
17.3 Estimation in the noisy setting	112
18 Efficient Error Estimation for Noisy, Sparse Linear Regression	114
18.1 Recap: introduction to noisy, sparse linear regression	114
18.2 The restricted eigenvalue condition	114
18.3 Bounds on ℓ_2 error	116
18.4 Proof of RE condition bounds	117
19 Restricted Eigenvalue Condition for Gaussian Random Matrices	121
19.1 Recap: Noisy, sparse linear estimation and the restricted eigenvalue condition	121
19.2 Restricted eigenvalue condition for Gaussian random matrices	121
19.3 LASSO oracle inequality	125
20 LASSO Prediction Error Bound and High-Dimensional Principal Component Analysis	127
20.1 Recap: overview of results for noisy, sparse linear regression	127
20.2 LASSO prediction error bound	128
20.3 Principal component analysis in high dimensions	129
20.4 General perturbation bound for eigenvectors	131
21 Principle Component Analysis for Spiked and Sparse Ensembles	134
21.1 Recap: estimation error bound for principle component analysis	134
21.2 Consequence for a spiked ensemble	134
21.3 Sparse principle component analysis	137
21.3.1 ℓ_1 -penalized estimation	137
21.3.2 The semidefinite programming relaxation estimator	138
21.3.3 The $s \ll n \ll s^2$ regime	138
21.4 Extra topics we will not cover	139

22	Examples of and Oracle Inequality for Non-Parametric Least Squares Regression	140
22.1	Recap: localized Gaussian complexity bound for non-parametric least squares	140
22.2	Applications of the localized Gaussian complexity bound	141
22.3	Oracle inequalities	143
22.4	Applications of the oracle inequality	145
23	L^2 Prediction Error Bounds for Nonparametric Function Regression	146
23.1	Recap: prediction error bounds for $\ \cdot\ _n$ compared to $\ \cdot\ _{L^2}$	146
23.2	Relation between $\ \cdot\ _n^2$ and $\ \cdot\ _{L^2}^2$	146
23.3	Naive bound	147
23.4	Using localization to get a faster rate	148
23.5	Uniform law for Lipschitz cost function	150
24	Introduction to Minimax Lower Bounds	152
24.1	Minimax risk and methods of obtaining lower bounds	152
24.2	Reduction to an M -ary testing problem	153
24.3	Some divergence measures	155
25	Methods for Proving Minimax Lower Bounds	158
25.1	Recap: Testing lemma and divergence measures for minimax lower bounds .	158
25.2	Le Cam's two points method	158
25.3	Mutual information	160
25.4	Fano's inequality	161
25.5	Yang-Barron's method	163

1 Introduction to High-Dimensional Statistics

1.1 Overview of the course

The first half of this course will cover theoretical tools used to establish theorems in high-dimensional statistics:

- Concentration inequalities
- Empirical process theory
- Gaussian process theory and random matrix theory

The second half of this course will cover statistical problems:

- Covariance estimation
- Sparse estimation problem
- Principal component analysis (PCA) in high dimension
- Non-parametric regression
- Minimax lower bounds

1.2 A motivating example: sparse estimation

Here is a motivating example:

Example 1.1 (High dimensional sparse estimation). Here is the assumption of our statistical model. We observe

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \in \mathbb{R}^n, \quad X = \begin{bmatrix} X_1^\top \\ \vdots \\ X_n^\top \end{bmatrix} \in \mathbb{R}^{n \times d}, \quad X_i \in \mathbb{R}^d.$$

We assume that the relationship $Y = X\theta^* + \omega$ holds, where $\theta^* \in \mathbb{R}^d$ with $\theta^* = \begin{bmatrix} \theta_1^* \\ \vdots \\ \theta_d^* \end{bmatrix}$ and $\omega \in \mathbb{R}^n$ is noise. In the high dimensional case, we have $n \ll d$, so standard linear regression will not be useful.

To deal with the problem in the high-dimensional case, we make the further assumption that θ^* is supported on $S \subseteq \{1, 2, \dots, d\}$, with $|S|$ denoted by s ; that is, θ_i^* can be nonzero only on the indices in S . This is called an **s -sparse** assumption. Our task is that given (Y, X) , we want to estimate θ^* .

We present results without proof, although we will develop these results later in the course.

(a) The naive estimator (assuming $\omega_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$) is

$$\hat{\theta}_{\text{LS}} := \arg \min_{\theta \in \mathbb{R}^d} \frac{1}{2n} \|Y - X\theta\|_2^2.$$

Classical theory tells us that

$$\begin{aligned} \mathbb{E}[\|\hat{\theta}_{\text{LS}} - \theta^*\|_2^2] &= \frac{\text{tr}(X^\top X)^{-1}}{n} \sigma^2 \\ &= \Theta\left(\frac{d}{n} \sigma^2\right) \end{aligned}$$

If $n \ll d$, then $\mathbb{E}[\|\hat{\theta}_{\text{LS}} - \theta^*\|_2^2] \gg 1$. This estimator, however does not use the assumption that $\theta^* \in \mathbb{R}^d$ is s -sparse.

(b) The LASSO estimator¹ is

$$\hat{\theta}_{\text{LASSO}} := \arg \min_{\theta \in \mathbb{R}^d} \frac{1}{2n} \|Y - X\theta\|_2^2 + \lambda_n \|\theta\|_1,$$

which has an L^1 penalty. Our goal is to show that

$$\|\hat{\theta}_{\text{LASSO}} - \theta^*\|_2 \lesssim c \sqrt{\frac{s \log d}{n}}.$$

We need the following condition:

Definition 1.1. The matrix X satisfies the **restricted eigenvalue (RE)**² condition over S with parameter (κ, α) if

$$\underbrace{\frac{1}{n} \|X\Delta\|_2^2}_{= \frac{1}{n} \langle \Delta, X^\top X \Delta \rangle} \geq \kappa \|\Delta\|_2^2 \quad \forall \Delta \in \mathbb{C}_\alpha(S) := \{\Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \leq \alpha \|\Delta_S\|_1\}.$$

This is a geometric assumption on $\mathcal{L}(\theta) = \frac{1}{2n} \|Y - X(\theta_* + \Delta)\|_2^2$. If X is (κ, α) -RE, then $\mathcal{L}(\Delta)$ is strongly convex in the cone $\mathbb{C}_\alpha(S)$.

Theorem 1.1. Suppose θ^* is supported on S , with $|S| = s$, and X satisfies the RE condition over S with parameter $(\kappa, 3)$. Further assume that $\lambda_n \geq 2 \| \frac{X^\top \omega}{n} \|_\infty$. Then

$$\|\hat{\theta}_{\text{LASSO}} - \theta^*\|_2 \leq \frac{3}{\kappa} \sqrt{s} \lambda_n.$$

¹LASSO comes from Tibshirani in 1994 and Chen, Donoho, and Saunders in 1994, as well.

²This condition was introduced by Bickel, Ritov, and Tsybakov in 2009.

What does this mean? The sparsity assumption is more natural; for example, if we are dealing with gene data in biology, we may assume that only a few genes will determine a trait. Let's now tackle a few questions about our assumptions:

1. When does RE hold?
2. How large is $2\|X^\top\omega\|_\infty/n$?
3. How can we compare the bound with the least squares estimator?

Make the assumption that $X_I \stackrel{\text{iid}}{\sim} N(0, \text{Id})$ (which can be generalized) and $\omega_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$. Here are the answers to our questions:

1.

Proposition 1.1. *Suppose $(X_i)_{i \in [n]} \stackrel{\text{iid}}{\sim} N(0, \text{Id})$. Fix $S \subseteq [d]$ with $|S| = s$. Then there exist universal constants $0 < c_1 < 1 < c_2$ such that when $n \geq c_2 s \log d$, we have*

$$\mathbb{P}\left(\frac{1}{2n}\|X\Delta\|_2^2 \geq c_1\|\Delta_2\|^2 \quad \forall \Delta \in \mathbb{C}_3(s)\right) \geq 1 - \frac{e^{-n/32}}{1 - e^{-n/32}}.$$

This tells us that the $(c_1, 3)$ -RE condition is satisfied with high probability (w.h.p.) as long as $n \geq s \log d$. To establish this proposition, we need to use empirical process theory and concentration inequalities.

2.

Lemma 1.1. *Suppose that $\max_{i \in [n]} \|x_i\|_2/\sqrt{n} \leq B_n$ and $\omega_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$. Then there is a universal constant c such that for all $t > 0$,*

$$\mathbb{P}\left(\frac{\|X^\top\omega\|_\infty}{n} \leq cB_n\sigma\left(\sqrt{\frac{2\log d}{n}} + t\right)\right) \geq 1 - 2e^{-nt^2/2}$$

Moreover, when $X_i \sim N(0, \text{Id})$, then for all $t \in (0, 1)$,

$$\mathbb{P}\left(\max_{i \in [n]} \frac{\|X_i\|_2^2}{n} \leq 1 + t\right) \geq 1 - ne^{-nt^2/8}.$$

This lemma tells us that

$$\frac{2\|X^\top\omega\|_\infty}{n} \leq \underbrace{\tilde{C}\sigma\sqrt{\frac{\log(d/\delta)}{n}}}_{\lambda_n}$$

with probability at least $1 - 2\delta$. To establish this lemma, we need concentration inequalities and empirical process theory.

3. Plug in $\lambda = \tilde{C}\sigma\sqrt{\frac{\log(d/\delta)}{n}}$ to get

$$\|\hat{\theta}_{\text{LASSO}} - \theta^*\|_2 \leq \frac{3}{\kappa}\sqrt{s}\lambda_n = \frac{3}{\kappa}\tilde{C}\sigma\sqrt{\frac{s\log(d/\delta)}{n}}$$

with probability at least $1 - 3\delta$. This means that as long as $n \gtrsim s\log(d/\delta)$,

$$\|\hat{\theta}_{\text{LASSO}} - \theta^*\|_2^2 \ll 1.$$

In comparison, $\mathbb{E}[\|\hat{\theta}_{\text{LS}} - \theta^*\|_2^2] = \Theta(\frac{d}{n}\sigma^2)$, which needs $n \geq d$ to be small.

1.3 Relationships with other statistical topics

Here are the relationships between this course and other courses:

- Stat 210A Theoretical Statistics: In statistical decision theory, we have a statistical model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ with a statistical procedure $\delta : D \rightarrow \Theta$ and a loss function $\ell : \Theta \times \Theta \rightarrow \mathbb{R}$. We can then calculate the risk function $R(\theta; \delta) = \mathbb{E}_\theta[\ell(\theta; \delta(Z))]$. We can compare risk functions for different procedures by looking at summarized statistics of the risk function:

- Bayes risk: We assume $\theta \sim \pi$, so $R_B(\pi; \delta) = \mathbb{E}_{\theta \sim \pi}[R(\theta; \delta)]$.
- Minimax: We can look at $R_M(\Theta; \delta) = \sup_{\theta \in \Theta} R(\theta; \delta)$.

In our example, LASSO is approximately minimax optimal

- CS 281A/Stat 241A Statistical learning theory: This focuses on a different (but related) collection of models (empirical risk minimization). We study them using a similar set of tools (concentration inequalities, empirical process theory).
- Stat 260 Mean field asymptotics in statistical learning: Here, we focus on the same collections of statistical models but study them in another regime ($n, d \rightarrow \infty$ with $n/d \rightarrow \text{constant}$ asymptotics). We use different collection of tools (statistical physics, AMP, Gaussian comparison). This needs stronger assumptions but gives more refined results.

Other useful courses are convex optimization and information theory. These courses are important in order to learn deep learning theory and reinforcement learning theory. In the next lecture, we will start learning about concentration inequalities.

2 Basic Concentration Inequalities

2.1 Concentration inequalities for sample averages

Suppose we have a random variable $X \sim \mathbb{P}_X$, sampled from the distribution \mathbb{P}_X . Let $\mu = \mathbb{E}_{X \sim \mathbb{P}_X}[X]$ be its expectation. In general, $|x - \mu|$ could be very large. However, in many scenarios (especially when X takes a special form), $|x - \mu|$ is very small with high probability.

Example 2.1. Let $X = \frac{1}{n} \sum_{i=1}^n Z_i$, where $Z_i \stackrel{\text{iid}}{\sim} \mathbb{P}_Z$ with $\mathbb{P}_Z \in \mathcal{P}([0, 1])$ (supported in $[0, 1]$). Then $\mathbb{E}[X] = \mathbb{E}[Z_i] =: \mu$. We will show in this lecture that

1. For all $t > 0$,

$$\mathbb{P}(|x - \mu| \geq t) = \mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n Z_i - \mu\right| \geq t\right) \leq \underbrace{2 \exp\left(-\frac{nt^2}{2}\right)}_{\xrightarrow{n \rightarrow \infty} 0}.$$

2. Equivalently, for any $0 < \delta < 1$,

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n Z_i - \mu\right| \geq \sqrt{\frac{2 \log(2/\delta)}{n}}\right) \leq \delta.$$

3. Equivalently,

$$\left|\frac{1}{n} \sum_{i=1}^n Z_i - \mu\right| < \sqrt{\frac{2 \log(2/\delta)}{n}}$$

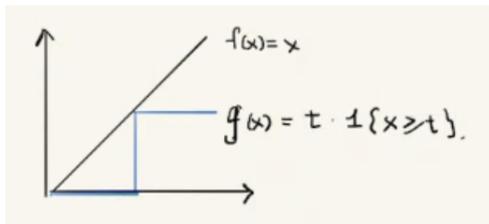
with probability at least $1 - \delta$, or **with high probability**.

2.2 Markov's inequality

Lemma 2.1 (Markov's inequality). *Let X be a nonnegative random variable. Then for all $t > 0$,*

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

Proof. Define $f(x) = x$ and $g(x) = t \mathbb{1}_{\{x \geq t\}}$. Then $f(x) \geq g(x)$.



Then

$$\mathbb{E}[X] \geq \mathbb{E}[t\mathbb{1}_{\{X \geq t\}}] = t\mathbb{P}(X \geq t). \quad \square$$

Markov's inequality is important because other concentration inequalities are consequences of Markov's inequality. For our example, we can apply Markov's inequality to $|X - \mu|$ with $X = \frac{1}{n} \sum_{i=1}^n Z_i$ to get

$$\begin{aligned} \mathbb{P}(|X - \mu| \geq t) &\leq \frac{\mathbb{E}[|X - \mu|]}{t} \\ &= \frac{\mathbb{E}[|\frac{1}{n} \sum_{i=1}^n Z_i - \mu|]}{t} \end{aligned}$$

Using Jensen's inequality, we can upper bound this by

$$= \frac{\mathbb{E}[|\frac{1}{n} \sum_{i=1}^n Z_i - \mu|^2]^{1/2}}{t}$$

Observe that $\mathbb{E}[(\frac{1}{n} \sum_{i=1}^n Z_i - \mu)^2] \leq n \mathbb{E}[(Z_i - \mu)^2]/n^2 \leq 1/n$. So we get

$$\begin{aligned} &\leq \frac{(1/n)^{1/2}}{t} \\ &= \frac{1}{\sqrt{nt}}. \end{aligned}$$

To rearrange this in terms of a tail probability δ , solve $\frac{1}{\sqrt{nt}} = \delta$:

$$\mathbb{P}\left(|X - \mu| \geq \frac{1}{\sqrt{n\delta}}\right) \leq \delta.$$

That is,

$$|X - \mu| < \frac{1}{\sqrt{n\delta}}$$

with probability at least $1 - \delta$. Here, we have gotten the correct $1/\sqrt{n}$ scaling, but the $1/\delta$ dependence is not optimal yet.

Remark 2.1. Letting $n \rightarrow \infty$ gives us a weak law of large numbers. However, if we sum these probabilities in n , we get a divergent sum, so we would need to be more careful if we wanted to use the Borel-Cantelli lemma to prove a strong law of large numbers.

2.3 Chebyshev's inequality

Lemma 2.2. *If $\text{Var}(X)$ exists, then for all $t > 0$,*

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{\text{Var}(X)}{t^2}.$$

Proof. Apply Markov's inequality:

$$\begin{aligned} \mathbb{P}(|X - \mathbb{E}[X]| \geq t) &\leq \mathbb{P}(|X - \mathbb{E}[X]|^2 \geq t^2) \\ &\leq \frac{\mathbb{E}[|X - \mathbb{E}[X]|^2]t^2}{t^2} \end{aligned} \quad \square$$

For our example, apply Chebyshev's inequality to $X = \frac{1}{n} \sum_{i=1}^n Z_i$ to get

$$\begin{aligned} \mathbb{P}\left(\left|\frac{1}{\sum_{i=1}^n Z_i} - \mu\right| \geq t\right) &\leq \frac{\text{Var}\left(\frac{1}{n} \sum_{i=1}^n Z_i\right)}{t^2} \\ &= \frac{\text{Var}(Z_i)}{nt^2} \\ &\leq \frac{1}{nt^2}. \end{aligned}$$

Solving $\delta = \frac{1}{nt^2}$, we get

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n Z_i - \mu\right| \geq \frac{1}{\sqrt{n}\sqrt{\delta}}\right) \leq \delta.$$

That is,

$$\left|\frac{1}{n} \sum_{i=1}^n Z_i - \mu\right| \geq \frac{1}{\sqrt{n}\sqrt{\delta}}$$

with probability at least $1 - \delta$. In comparison to our application of Markov's inequality, this gives a $1/\sqrt{\delta}$ dependence instead of a $1/\delta$ dependence, which is significant when δ is small.

In general, we have

Lemma 2.3. *For all $t > 0$,*

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

provided this k -th moment exists.

As an exercise, apply this to our example and carefully bound $\mathbb{E}[|\frac{1}{n} \sum_{i=1}^n Z_i - \mu|^k]$ to show that there is a constant $C_k < \infty$ such that

$$\left|\frac{1}{n} \sum_{i=1}^n Z_i - \mu\right| \leq \frac{C_k}{\sqrt{n}\delta^{1/k}}$$

with probability at least $1 - \delta$.

As another exercise, derive Cantelli's inequality using the same principle:

Lemma 2.4 (Cantelli's inequality).

$$\mathbb{P}(X - \mathbb{E}[X] \geq t) \leq \frac{\text{Var}(X)}{\text{Var}(X) + t^2}.$$

Proof. The events $\{X - \mu \geq t\} = \{f(x - \mu) \geq f(t)\}$ are the same, where $f(t) = (t + u)^2$ for some special choice of u . \square

2.4 Chernoff's inequality

Lemma 2.5 (Chernoff's inequality). *For all $t > 0$, we have*

$$\begin{aligned} \mathbb{P}(X \geq \mu + t) &\leq \inf_{\lambda} \frac{\mathbb{E}[e^{\lambda(X-\mu)}]^{-\lambda t}}{e} \\ &= e^{-h(t)}, \end{aligned}$$

where

$$h(t) = \sup_{\lambda} \lambda t - \log \mathbb{E}[e^{\lambda(X-\mu)}].$$

Proof. We will prove the inequality. We can upper bound the tail probability by rewriting this event:

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

This holds for all λ , so it holds for the inf over all λ . We get

$$\begin{aligned} \mathbb{P}(X - \mu \geq t) &= \inf_{\lambda} \mathbb{P}(e^{\lambda(X-\mu)} \geq e^{\lambda t}) \\ &\leq \inf_{\lambda} \frac{\mathbb{E}[e^{\lambda(X-\mu)}]}{e^{\lambda t}}, \end{aligned}$$

where we have used Markov's inequality. \square

Remark 2.2. To interpret the quantities in the bound, define the **moment generating function** of a random variable Z as

$$M_Z(\lambda) := \mathbb{E}[e^{\lambda Z}].$$

This is called the moment generating function because

$$\frac{d}{d\lambda} M_Z(\lambda)|_{\lambda=0} = \mathbb{E}_Z[Z e^{\lambda Z}]|_{\lambda=0} = \mathbb{E}[Z].$$

In general,

$$\frac{d^k}{d\lambda^k} M_Z(\lambda)|_{\lambda=0} = \mathbb{E}_Z[Z^k e^{\lambda Z}]|_{\lambda=0} = \mathbb{E}[Z^k],$$

the k -th moment.

Define the **cumulant generating function** of Z as

$$K_Z(\lambda) := \log \mathbb{E}[e^{\lambda Z}] = \log M_Z(\lambda).$$

This is called the cumulant generating function because it generates the **cumulants**

$$\kappa_k = \frac{d^k}{d\lambda^k} K_Z(\lambda) \Big|_{\lambda=0}.$$

For example, $\kappa_2 = \text{Var}(Z) \geq 0$. In fact, $K_Z''(\lambda) \geq 0$, so the cumulant generating function is always convex.

Define the Legendre transform f^* of $f : \mathbb{R} \rightarrow \mathbb{R}$ as

$$f^*(t) = \sup_{\lambda \in \mathbb{R}} \lambda t - f(\lambda).$$

Then $h(t)$ is the Legendre transform of $K_{X-\mu}(\lambda)$. The Legendre transform can be thought of as a dual³ in the sense that $f^{**}(\lambda) = (f^*)^*(\lambda) = f(\lambda)$ if f is convex.

For our example, apply Chernoff's inequality to $X = \frac{1}{n} \sum_{i=1}^n Z_i$. Here is a claim we will prove next lecture: If $Z \sim \mathbb{P}_Z \in \mathcal{P}([0, 1])$, then

$$\mathbb{E}[e^{\lambda(Z - \mathbb{E}[Z])}] \leq e^{\lambda^2/2}, \quad \forall \lambda \in \mathbb{R}.$$

Using this claim, we bound

$$\begin{aligned} \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n Z_i - \mu \geq t\right) &\leq \inf_{\lambda} \frac{\mathbb{E}[e^{\lambda(\frac{1}{n} \sum_{i=1}^n Z_i - \mu)}]}{e^{\lambda t}} \\ &= \inf_{\lambda} \frac{\mathbb{E}[\prod_{i=1}^n e^{\lambda \frac{1}{n} (Z_i - \mu)}]}{e^{\lambda t}} \end{aligned}$$

Using independence of the Z_i ,

$$\begin{aligned} &= \inf_{\lambda} \frac{\prod_{i=1}^n \mathbb{E}[e^{\lambda \frac{1}{n} (Z_i - \mu)}]}{e^{\lambda t}} \\ &= \inf_{\lambda} \frac{\mathbb{E}[e^{\lambda \frac{1}{n} (Z_i - \mu)}]^n}{e^{\lambda t}} \\ &\leq \inf_{\lambda} \frac{(e^{(\lambda/n)^2/2})^n}{e^{\lambda t}} \\ &= \inf_{\lambda} e^{\lambda^2/(2n) - \lambda t} \end{aligned}$$

³The Legendre transform is sometimes known as the **Fenchel dual**.

This exponent is quadratic in λ , so we can calculate that it is minimized at $\lambda_* = nt$.

$$\begin{aligned} &= e^{-(nt)^2/(2n) - nt \cdot t} \\ &= e^{-nt^2/2}. \end{aligned}$$

We will apply this line of reasoning again and again in this course.

Similarly, we have the lower bound

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n Z_i - \mu \leq -t\right) \leq e^{-nt^2/2}.$$

Combining these two tail inequalities, we get

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n Z_i - \mu\right| \geq t\right) \leq 2e^{-nt^2/2}.$$

This is the inequality we presented at the beginning of the lecture. If we solve $\delta = 2e^{-nt^2/2}$, we get

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n Z_i - \mu\right| \geq \sqrt{\frac{2 \log(2/\delta)}{n}}\right) \leq \delta.$$

That is,

$$\left|\frac{1}{n} \sum_{i=1}^n Z_i - \mu\right| < \sqrt{\frac{2 \log(2/\delta)}{n}}$$

with probability at least $1 - \delta$.

2.5 Comparison of inequalities

Here is a table comparing the different inequalities we have seen.

	Markov	Chebyshev	k -th moment	Chernoff
require	First moment	Second moment	k -th moment	Moment generating function
bound	$\frac{1}{\sqrt{n\delta}}$	$\frac{1}{\sqrt{n}\sqrt{\delta}}$	$\frac{1}{\sqrt{n}\delta^{1/k}}$	$\frac{\sqrt{2 \log(2/\delta)}}{\sqrt{n}}$

Using more moments, we get better bounds; using the MGF is like using all the moments of a random variable. These have the same dependence in n but different dependence in δ . What is the benefit of better dependence in δ ? This is useful for the union bound!

2.6 Applying union bounds

Lemma 2.6 (Union bound). *Suppose we have a collection of events $\{E_s\}_{s \in [d]}$. If $\mathbb{P}(E_s^c) \leq \frac{\delta}{d}$ for all s , then*

$$\mathbb{P}\left(\bigcup_{s \in [d]} E_s\right) \geq 1 - \delta.$$

So if we divide delta by the number of events d , we can use a good δ dependence to get a good union bound.

Remark 2.3. Here is a common mistake that happens in homework, exams, and even ICML and NeurIPS papers. Let $(Z_i^{(s)})_{i \in [n], s \in [d]} \stackrel{\text{iid}}{\sim} \mathbb{P}_Z \in \mathbb{P}([0, 1])$. Suppose someone proves that for all $s \in [d]$,

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n Z_i^{(s)} - \mu\right| \leq \sqrt{\frac{\log(1/\delta)}{n}}\right) \geq 1 - \delta.$$

The common mistake is to claim that

$$\mathbb{P}\left(\forall s \in [d], \left|\frac{1}{n} \sum_{i=1}^n Z_i^{(s)} - \mu\right| \leq \sqrt{\frac{\log(1/\delta)}{n}}\right) \geq 1 - \delta.$$

This is not true because it ignores the dependence on the dummy variable s . Instead, the correct thing to do is to say

$$\mathbb{P}\left(\forall s \in [d], \left|\frac{1}{n} \sum_{i=1}^n Z_i^{(s)} - \mu\right| \leq \sqrt{\frac{\log(d/\delta)}{n}}\right) \geq 1 - \delta.$$

This d is usually very large, such as exponential or doubly exponential in n .

So please avoid the following statement:

$$\forall s \in [d], \quad \left|\frac{1}{n} \sum_{i=1}^n Z_i^{(s)} - \mu\right| \leq \varepsilon n, \quad \text{with probability at least } 1 - \delta.$$

This is ambiguous if the probability applies to each individual s or all s at once. Instead, use this statement instead:

For individual bounds, write

(a) $\forall s \in [d], \mathbb{P}(\dots) \geq 1 - \delta.$

(b) $\forall s \in [d]$, with probability at least $1 - \delta$, the following event happens:

$$\left|\frac{1}{n} \sum_{i=1}^n Z_i^{(s)} - \mu\right| \leq \varepsilon n.$$

For union bounds use these:

(a) $\mathbb{P}(\forall s, \dots) \geq 1 - \delta.$

(b) With probability at least $1 - \delta$, the following event happens:

$$\forall s \in [d], \quad \left| \frac{1}{n} \sum_{i=1}^n Z_i^{(s)} - \mu \right| \leq \varepsilon n.$$

(c)

$$\sup_{s \in [d]} \left| \frac{1}{n} \sum_{i=1}^n Z_i^{(s)} - \mu \right| \leq \varepsilon n \quad \text{with probability at least } 1 - \delta.$$

Here are some exercises to do for using union bounds:

Suppose $(Z_i^{(s)})_{i \in [n], s \in [d]} \stackrel{\text{iid}}{\sim} \mathbb{P}_Z \in \mathcal{P}([0, 1]).$

- Markov's inequality implies that with probability $1 - \delta$, the following happens:

$$\forall s \in [d], \quad \left| \frac{1}{n} \sum_{i=1}^n Z_i^{(s)} - \mu \right| \leq \frac{d}{\sqrt{n}\delta}.$$

- Chebyshev's inequality implies that with probability $1 - \delta$, the following happens:

$$\forall s \in [d], \quad \left| \frac{1}{n} \sum_{i=1}^n Z_i^{(s)} - \mu \right| \leq \frac{\sqrt{d}}{\sqrt{n}\sqrt{\delta}}.$$

- Markov's inequality implies that with probability $1 - \delta$, the following happens:

$$\forall s \in [d], \quad \left| \frac{1}{n} \sum_{i=1}^n Z_i^{(s)} - \mu \right| \leq \frac{\sqrt{2 \log(2d/\delta)}}{\sqrt{n}}.$$

3 Sub-Gaussian and Sub-Exponential Random Variables

3.1 Sub-Gaussian random variables

Last time, we used Chernoff’s inequality to get an upper bound on the tail probability of $\frac{1}{n} \sum_{i=1}^n Z_i - \mu$, where Z_i are iid and supported in $[0, 1]$. We made a claim about the moment generating function of such random variables:

$$\mathbb{E}[e^{\lambda(Z - \mathbb{E}[Z])}] \leq e^{\lambda^2/2}.$$

We can abstract this into a definition:

Definition 3.1. A random variable with $\mu = \mathbb{E}[X]$ is σ -**sub-Gaussian**⁴ if there is a positive number σ such that

$$\mathbb{E}[e^{\lambda(X - \mu)}] \leq e^{\lambda^2 \sigma^2 / 2} \quad \forall \lambda \in \mathbb{R}.$$

Combining with Chernoff’s inequality, we have that if X is σ -sub-Gaussian, then

$$\begin{aligned} \mathbb{P}(X - \mu \geq t) &\leq \inf_{\lambda} \frac{\mathbb{E}[e^{\lambda(X - \mu)}]}{e^{\lambda t}} \\ &\leq \inf_{\lambda} e^{\lambda^2 \sigma^2 / 2 - \lambda t} \end{aligned}$$

This quadratic function in the exponent is minimized at $\lambda = t/\sigma^2$:

$$\begin{aligned} &= e^{(t/\sigma^2)^2 \cdot \sigma^2 / 2 - t^2 / \sigma^2} \\ &= e^{-t^2 / (2\sigma^2)}. \end{aligned}$$

Why is this called “sub-Gaussian”?

(a) If $G \sim N(\mu, \sigma^2)$, then

$$\mathbb{E}[e^{\lambda(G - \mu)}] = \int_{-\infty}^{\infty} e^{\lambda(x - \mu)} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) dx$$

We can combine the exponentials and complete the square in the exponent to solve this integral.

$$= e^{\lambda^2 \sigma^2 / 2}.$$

(b) If $G \sim N(0, 1)$, then

$$\lim_{t \rightarrow \infty} \frac{\mathbb{P}(G \geq t)}{\underbrace{\frac{1}{t} \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)}_{\phi(t)}} = 1.$$

⁴Some textbooks call this σ^2 -sub-Gaussian, and you should think of σ as a surrogate for variance.

In addition, if ϕ is the standard Gaussian probability density function, then

$$\frac{1}{t}\phi(t) \leq \mathbb{P}(G \geq t) \leq \left(\frac{1}{t} - \frac{1}{t^3} + \frac{3}{t^5}\right)\phi(t).$$

This is exercise 2.2 in Wainwright's textbook. To prove this, first show that $\phi(z) = -\frac{\phi'(z)}{z}$. Next, calculate $\int_t^\infty \phi(z) dz = \int_t^\infty -\frac{\phi'(z)}{z} dz$ by using integration by parts.

3.2 Hoeffding's inequality

Proposition 3.1 (Hoeffding's inequality). *Suppose $X_i, i = 1, \dots, n$ are independent, where X_i has mean μ and is σ_i -sub-Gaussian. Then*

1. $\sum_{i=1}^n X_i$ has mean $\sum_{i=1}^n \mu_i$ and is sub-Gaussian with parameter $\sqrt{\sum_{i=1}^n \sigma_i^2}$.

2.

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

Proof.

1.

$$\begin{aligned} \mathbb{E}[e^{\lambda \sum_{i=1}^n (X_i - \mu_i)}] &= \mathbb{E}\left[\prod_{i=1}^n e^{\lambda(X_i - \mu_i)}\right] \\ &= \prod_{i=1}^n \mathbb{E}[e^{\lambda(X_i - \mu_i)}] \\ &\leq \prod_{i=1}^n e^{\lambda^2 \sigma_i^2 / 2} \\ &= e^{\lambda^2 (\sum_{i=1}^n \sigma_i^2) / 2}. \end{aligned}$$

2. The second statement is by Chernoff's inequality, as above. □

Let $(X_i)_{i \in [n]} \stackrel{\text{oniid}}{\sim} X$ be σ -sub-Gaussian. Then

$$\begin{aligned} \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n X_i - \mu \geq t\right) &= \mathbb{P}\left(\sum_{i=1}^n (X_i - \mu) \geq nt\right) \\ &\leq \exp\left(-\frac{(nt)^2}{2n\sigma^2}\right) \\ &= \exp\left(-\frac{nt^2}{2\sigma^2}\right). \end{aligned}$$

- (a) How do we extract the order of $\frac{1}{n} \sum_{i=1}^n X_i - \mu$? Let $\delta = \exp(-\frac{nt^2}{2\sigma^2})$ and solve for t to get $t = \sigma \sqrt{\frac{2 \log(1/\delta)}{n}}$. Thus,

$$\frac{1}{n} \sum_{i=1}^n X_i \leq \mu + \sigma \sqrt{\frac{2 \log(1/\delta)}{n}} \quad \text{with probability at least } 1 - \delta.$$

To check for mistakes, look at the units: X_i , μ , and σ have the same units, while δ and n are unitless. Here, we can see that the units match up.

- (b) How many samples are needed to that $\frac{1}{n} \sum_{i=1}^n X_i - \mu \leq t$ with probability $1 - \delta$? Let $\delta = \exp(-\frac{nt^2}{2\sigma^2})$, and solve for n to get $n = \frac{2\sigma^2}{t^2} \log(1/\delta)$.

3.3 Examples of sub-Gaussian random variables

Example 3.1 (Rademacher random variables). Consider a **Rademacher random variable** $\varepsilon \sim \text{Unif}(\{\pm 1\})$. ε is 1-sub-Gaussian.

Proof.

$$\mathbb{E}[e^{\lambda\varepsilon}] = \frac{1}{2}e^\lambda + \frac{1}{2}e^{-\lambda}$$

We want to upper bound this by $e^{\lambda^2/2}$. One way is to use the Taylor expansion:

$$\begin{aligned} &= \frac{1}{2} \sum_{k=1}^{\infty} \frac{\lambda^k}{k!} + \frac{(-\lambda)^k}{k!} \\ &= \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{(2k!)}. \end{aligned}$$

If we take the Taylor expansion of $e^{\lambda^2/2}$, we get $1 + \sum_{k=1}^{\infty} \frac{\lambda^{2k}}{2^k k!}$. To compare the Taylor expansions, we only need to show that $(2k)! \geq 2^k k!$. \square

Example 3.2 (Bounded random variable). Let $X \in \mathcal{P}([a, b])$. We claim that X is $(b - a)$ -sub-Gaussian.⁵

Proof. Instead of a direct calculation, we use a series of tricks.

Trick 1: Let $X' \stackrel{d}{=} X$ with X, X' independent. Then

$$\mathbb{E}_X[e^{\lambda(X-\mu)}] = \mathbb{E}_X[e^{\lambda X - \mathbb{E}_X[X']}]$$

Trick 2: Use Jensen's inequality to get $e^{-\lambda \mathbb{E}[X']} \leq \mathbb{E}[e^{-\lambda X'}]$. This gives

$$\leq \mathbb{E}_{X, X'} \mathbb{E}[e^{\lambda(X-X')}]$$

⁵It is actually possible to show this with parameter $(b - a)/2$, but we will not show this fact in this lecture.

Trick 3: Introduce $\varepsilon \sim \text{Unif}(\{\pm 1\})$ with ε independent of (X, X') . Then $\varepsilon(X - X') \stackrel{d}{=} X - X'$.

$$= \mathbb{E}_{\varepsilon, X, X'} \mathbb{E}[e^{\lambda \varepsilon (X - X')}]$$

Using the tower property of conditional expectation,

$$= \mathbb{E}_{X, X'} [\mathbb{E}_{\varepsilon}[e^{\lambda \varepsilon (X - X')} \mid X, X']]$$

By the 1-sub-Gaussianity of ε ,

$$\leq \mathbb{E}_{X, X'} [e^{\lambda^2 (X - X')^2 / 2}]$$

Since $(X - X') \leq (b - a)^2$ by the boundedness of X, X' ,

$$\leq e^{\lambda^2 (b-a)^2 / 2}. \quad \square$$

Remark 3.1. These tricks will be useful in later lectures and in statistics research. This technique is known as **symmetrization**.

3.4 Equivalent characterizations of sub-Gaussianity

Here are some

Theorem 3.1 (HDP 2.6 or RV 2.5.1). *Let X be a random variable. Then the following are equivalent:*

(i) *The tails of X satisfy*

$$\mathbb{P}(|X| \geq t) \leq 2 \exp\left(-\frac{t^2}{\kappa_1^2}\right) \quad \forall t \geq 0.$$

(ii) *The moments of X satisfy*

$$\|X\|_{L^p} = (\mathbb{E}[|X|^p])^{1/p} \leq \kappa_2 \sqrt{p}, \quad \forall p \geq 1.$$

(iii) *The moment generating function of X^2 satisfies*

$$\mathbb{E}[\exp(\lambda^2 X^2)] \leq \exp(\kappa_3^2 \lambda^2) \quad \forall \lambda \text{ such that } |\lambda| \leq \frac{1}{\kappa_3}.$$

(iv) *The moment generating function of X^2 is bounded at some point:*

$$\mathbb{E}[\exp(X^2 / \kappa_4^2)] \leq 2.$$

Moreover, if $\mathbb{E}[X] = 0$, then properties (i)-(iv) are also equivalent to

5. *The moment generating function of X satisfies*

$$\mathbb{E}[\exp(\lambda X)] \leq \exp(\kappa_5^2 \lambda^2 / 2) \quad \forall \lambda \in \mathbb{R}.$$

Here, $\kappa_1, \dots, \kappa_5$ are universal constants.

Proof. Proof is an exercise. □

Remark 3.2. Some people define sub-Gaussian through property (i) instead of (v). It can also be defined in terms of **Orlicz norms**, which are covered in an exercise in Wainwright's book. We use the moment generating function definition because a *tensorization* property will be important to us later.

Proposition 3.2. *There is a universal constant κ such that if X is σ -sub-Gaussian and Z is a random variable bounded by 1, then ZX is $\kappa\sigma$ -sub-Gaussian.*

Remark 3.3. Z and X can be dependent!

Proof. We can use any of the characterizations (i), (ii), (ii), (iv) to prove this. (v) doesn't work as easily. □

3.5 Sub-exponential random variables

Let $G \sim N(0, 1)$. Then G^2 is not sub-Gaussian. This is because $\mathbb{E}[G^2] = 1$, and

$$\begin{aligned} \mathbb{E}[e^{\lambda(G^2-1)}] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\lambda(z^2-1)} e^{-z^2/2} dz \\ &= \begin{cases} \frac{e^{-\lambda}}{\sqrt{1-2\lambda}} & \lambda < 1/2 \\ \infty & \lambda \geq 1/2. \end{cases} \end{aligned}$$

We can still derive a good but weaker tail bound for this kind of random variable.

Definition 3.2. A random variable X is (ν, α) -**sub-exponential** if

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\lambda^2\nu^2/2} \quad \forall |\lambda| \leq \frac{1}{\alpha}.$$

We can see from this definition that sub-Gaussian variables are sub-exponential with any $\alpha > 0$.

Example 3.3. If $G \sim N(0, 1)$, then G^2 is $(2, 4)$ -sub-exponential.

Proof. We want to show that

$$\mathbb{E}[e^{\lambda(G^2-1)}] = \frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \leq e^{2\lambda^2} \quad \forall |\lambda| \leq \frac{1}{4}.$$

we can do this by comparing Taylor series. □

Combining this definition with Chernoff's inequality, we have that if X is (ν, α) -sub-exponential, then

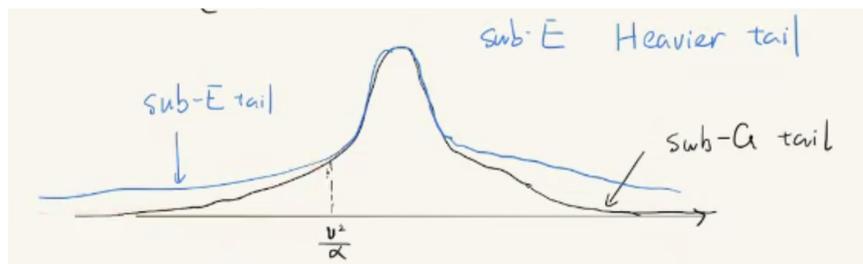
$$\begin{aligned} \mathbb{P}(X - \mu \geq t) &\leq \inf_{\lambda} \frac{\mathbb{E}[e^{\lambda(X-\mu)}]}{e^{\lambda t}} \\ &\leq \inf_{|\lambda| \leq 1/\alpha} \frac{e^{\nu^2 \lambda^2 / 2}}{e^{\lambda t}} \\ &= \exp\left(\inf_{\lambda \leq 1/\alpha} \nu^2 \lambda^2 / 2 - \lambda t\right) \end{aligned}$$

If this interval contains $\lambda = t/\nu^2$, then this is the minimum. Otherwise, the minimum will be on the boundary.

$$= \begin{cases} \exp(-\frac{t^2}{2\nu^2}) & \text{if } \frac{t}{\nu^2} \leq \frac{1}{\alpha} \\ \exp(\frac{\nu^2}{2\alpha^2} - \frac{t}{\alpha}) & \text{if } \frac{t}{\nu^2} > \frac{1}{\alpha} \end{cases}$$

The second expression is $\leq \exp(-\frac{t}{\alpha})$. So we can write this as

$$\leq \exp\left(-\min\left\{\frac{t^2}{2\nu^2}, \frac{t}{\alpha}\right\}\right).$$



Why is this called “sub-exponential?”

(a) If $Z \sim \text{Exp}(1/\alpha)$, then

$$\mathbb{P}(Z \geq t) = \exp\left(-\frac{t}{\alpha}\right).$$

(b) $\text{Exp}(1)$ is $(\sqrt{2}, 2)$ -sub-exponential: If $Z \sim \text{Exp}(1)$, then

$$Z \stackrel{d}{=} \frac{1}{2}(G_1^2 + G_2^2), \quad G_1, G_2 \stackrel{\text{iid}}{\sim} N(0, 1).$$

Then

$$\mathbb{E}[e^{\lambda(Z-1)}] = \mathbb{E}[e^{\frac{\lambda}{2}(G_1^2 + G_2^2 - 2)}]$$

for $|\lambda| \leq 1/2$,

$$= \mathbb{E}[e^{\frac{\lambda}{2}(G_1^2-1)}] \mathbb{E}[e^{\frac{\lambda}{2}(G_2^2-1)}]$$

$$\leq \frac{e^{-\lambda}}{1-\lambda}$$

$$\leq e^{\lambda^2}.$$

4 Bernstein's Inequality, the Johnson-Lindenstass Lemma, and More Concentration Inequalities

4.1 Bernstein condition for sub-exponentiality

A bounded random variable is sub-Gaussian and hence is sub-exponential, but we can get a tighter quantitative sub-exponential bound.

Proposition 4.1. *Suppose X has a mean μ and variance σ^2 . Suppose that $\mathbb{E}[(X - \mu)^k] \leq \frac{1}{2}k!\sigma^2b^{k-2}$ for all $k \geq 2$. Then X is $(\sqrt{2}\sigma, 2b)$ -sub-exponential.*

Note that the units in this inequality condition make sense. This condition is called the **Bernstein condition**.

Proof. We just need to show that the moment generating function is bounded: Do a Taylor expansion:

$$\begin{aligned} \mathbb{E}[e^{\lambda(X-\mu)}] &= 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty} \lambda^k \frac{\mathbb{E}[(X-\mu)^k]}{k!} \\ &\leq 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^2\sigma^2}{2} \sum_{k=3}^{\infty} (|\lambda|b)^{k-2} \end{aligned}$$

This is a geometric series, so we can simplify it.

$$\begin{aligned} &\leq 1 + \frac{\lambda\sigma^2/2}{1-b|\lambda|} \\ &\leq e^{(\lambda^2\sigma^2/2)/(1-b|\lambda|)} \end{aligned}$$

When $|\lambda| \leq \frac{1}{2b}$,

$$\leq e^{\lambda^2(\sqrt{2}\sigma)^2/2}. \quad \square$$

Now let X be a random variable with $\text{Var}(X) = \sigma^2$ and $0 \leq X \leq b$. Then

$$\begin{aligned} \mathbb{E}[|X - \mu|^k] &\leq \mathbb{E}[|X - \mu|^2 \cdot b^{k-2}] \\ &= \sigma^2 b^{k-2} \\ &\leq \frac{k!}{2} \sigma^2 b^{k-2}, \end{aligned}$$

so X is $(\sqrt{2}\sigma, 2b)$ -sub-exponential. Last time, we had that X is b -sub-Gaussian. So the sub-exponential tail bound here is stronger in the region where the sub-exponential and sub-Gaussian tail behaviors are similar.

4.2 Bernstein's inequality

Lemma 4.1 (Bernstein's inequality). *Let $\{X_i\}_{i \in [n]}$ be independent with $\mathbb{E}[X_i] = \mu_i$ and X_i (ν_i, α_i) -sub-exponential. Then $\sum_{i=1}^n (X_i - \mu_i)$ is sub exponential with parameters $\nu_* = \sqrt{\sum_{i=1}^n \nu_i^2}$ and $\alpha_* = \max_i \alpha_i$. Moreover,*

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

Proof.

$$\begin{aligned} \mathbb{E}[e^{\lambda \sum_{i=1}^n (X_i - \mu_i)}] &= \prod_{i=1}^n \mathbb{E}[e^{\lambda (X_i - \mu_i)}] \\ &\leq e^{\lambda^2 \sum_{i=1}^n \nu_i^2 / 2}. \end{aligned}$$

for all $\lambda \leq 1/\max_{i \in [n]} \alpha_i$. □

Let $(X_i)_{i \in [n]} \stackrel{\text{iid}}{\sim} X$ be (ν, b) -sub-exponential. Then

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n (X_i - \mu_i) \geq t\right) \leq e^{-n \min\{\frac{t^2}{2\nu^2}, \frac{t}{2b}\}}.$$

- (a) How do we extract the order of $\frac{1}{n} \sum_{i=1}^n X_i - \mu$? Set $\delta = \exp(-n \min\{\frac{t^2}{2\nu^2}, \frac{t}{2b}\})$, and solve for t to get

$$t = \max\left\{\nu \sqrt{\frac{2 \log(1/\delta)}{n}}, b \frac{2 \log(1/\delta)}{n}\right\}.$$

This tells us that

$$\frac{1}{n} \sum_{i=1}^n X_i - \mu \leq \max\left\{\nu \sqrt{\frac{2 \log(1/\delta)}{n}}, b \frac{2 \log(1/\delta)}{n}\right\} \quad \text{with probability at least } 1 - \delta.$$

For small δ , the first term is the dominant term while the second is a *burn-in term*.

- (b) How many samples do we need to have $\frac{1}{n} \sum_{i=1}^n X_i - \mu \leq t$ with probability $1 - \delta$? Set $\delta = \exp(-n \min\{\frac{t^2}{2\nu^2}, \frac{t}{2b}\})$ and solve for n to get

$$n = \max\left\{\frac{2\nu^2}{t^2} \log(1/\delta), \frac{2b}{t} \log(1/\delta)\right\}.$$

When t is small, the first term is dominant, while the second is of smaller order.

Example 4.1. Let X_i be iid with support in $[0, b]$ and $\text{Var}(X_i) \leq \nu^2$. We know that X_i is b -sub-Gaussian and (ν, b) -sub-exponential. In order for $|\frac{1}{n} \sum_{i=1}^n X_i - \mu| \leq \varepsilon$ with probability $1 - \delta$,

$$\text{sG}(1) \implies n \geq \frac{b^2}{\varepsilon^2} \log\left(\frac{1}{\delta}\right),$$

$$\text{sE}(\nu, 1) \implies n \geq \max\left\{\frac{\nu^2}{\varepsilon^2} \log\left(\frac{1}{\delta}\right), \frac{b}{\varepsilon} \log\left(\frac{1}{\delta}\right)\right\}.$$

When $\varepsilon \leq b$, $\frac{b}{\varepsilon} \log(\frac{1}{\delta}) \leq \frac{b^2}{\varepsilon^2} \log(\frac{1}{\delta})$. So the sub-exponential bound is a stronger bound.

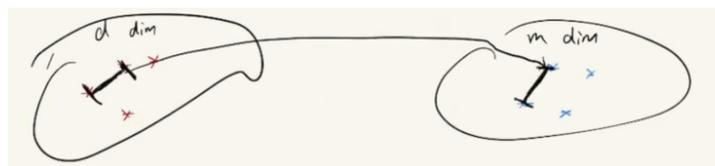
4.3 An application: the Johnson-Lindenstrass Lemma

Let $Y = \sum_{i=1}^n Z_i$ with $Z_i \sim N(0, 1)$. Then $Y \sim \chi^2(n)$. Last time, we showed that Z_i^2 is $\text{sE}(2, 4)$, so $Y \sim \text{sE}(2\sqrt{n}, 4)$. By Bernstein's inequality,

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n Z_i^2 - 1\right| \geq t\right) \leq 2e^{-nt^2/8} \quad \forall t \leq 1.$$

Here is a problem: Suppose we have $\{u_1, u_2, \dots, u_N\} \subseteq \mathbb{R}^d$ with a high dimension d . Can we find a $F : \mathbb{R}^d \rightarrow \mathbb{R}^m$ with some small m such that the distances are preserved? That is, we want

$$1 - \delta \leq \frac{\|F(u_i) - F(u_j)\|_2^2}{\|u_i - u_j\|_2^2} \leq 1 + \delta, \quad \forall i, j \in [N].$$



How small can we make m ? The Johnson-Lindenstrass says that we can achieve this by random projection.

Lemma 4.2 (Johnson-Lindenstrass). Let $X \in \mathbb{R}^{m \times d}$ have entries $X_{i,j} \stackrel{\text{iid}}{\sim} N(0, 1)$, and let $F : \mathbb{R}^d \rightarrow \mathbb{R}^m$ be defined as $R(u) = \frac{1}{\sqrt{m}} X \cdot u$. Then for any fixed $\{u_1, \dots, u_N\} \subseteq \mathbb{R}^d$, as long as $m \gtrsim \frac{1}{\varepsilon^2} \log(\frac{N}{\delta})$, then with probability $1 - \delta$, we have

$$1 - \varepsilon \leq \frac{\|F(u_i) - F(u_j)\|_2^2}{\|u_i - u_j\|_2^2} \leq 1 + \varepsilon, \quad \forall i, j \in [N].$$

Remark 4.1. The dimension that we can reduce to is of order $\log N$, where N is the number of points. So no matter the dimension d , we can always reduce the dimension to order $\log N$.

Proof. Denote $Y_{i,j} = \frac{\|F(u_i) - F(u_j)\|_2^2}{\|u_i - u_j\|_2^2}$. We claim that $Y_{i,j} \sim \chi^2(m)/m$. Then Bernstein's inequality will give

$$\mathbb{P}(|Y_{i,j} - 1| \geq t) \leq 2e^{-mt^2/\delta} \quad \forall t \leq 1.$$

Using a union bound on all $N(N-1) \leq N^2$ pairs $i \neq j$, we get

$$\mathbb{P}(\exists i, j \in [N] \text{ s.t. } |Y_{i,j} - 1| \geq t) \leq 2N^2 e^{-mt^2/8} \quad \forall t \leq 1.$$

Setting the right hand side equal to δ , we can solve for m to get

$$m \geq \frac{8}{t^2} \log\left(\frac{2N^2}{\delta}\right) = \frac{C}{t^2} \log\left(\frac{N}{\delta}\right).$$

Now let's verify the claim that $Y_{i,j} = \frac{\|F(u_i) - F(u_j)\|_2^2}{\|u_i - u_j\|_2^2} \sim \chi^2(m)/m$. Note that

$$\frac{1}{\sqrt{m}} X(u_i - u_j) \sim N\left(0, \frac{\|u_i - u_j\|_2^2}{m} I_m\right),$$

which implies that

$$\frac{\|X(u_i - u_j)\|_2^2}{m} \sim \|u_i - u_j\|_2^2 \chi^2(m)/m.$$

This proves the claim. □

Remark 4.2. If we use Markov's inequality instead of Bernstein's inequality, we get a worse bound.

4.4 Equivalent characterizations of sub-exponentiality

Theorem 4.1 (2.13 in HDS, 2.7.1 in HDP⁶). *The following statements are equivalent:*

(a)

$$\mathbb{P}(|X| \geq t) \leq 2 \exp(-t/\kappa_1), \quad \forall t \geq 0.$$

(b)

$$\|X\|_{L^p} = (\mathbb{E}[|X|^p])^{1/p} \leq \kappa_2 p, \quad \forall p \geq 1.$$

(c)

$$\mathbb{E}[\exp(\lambda|X|)] \leq \exp(\kappa_3 \lambda) \quad \forall \lambda \text{ s.t. } 0 \leq \lambda \leq \frac{1}{\kappa_3}.$$

(d)

$$\mathbb{E}[\exp(|X|/\kappa_4)] \leq 2.$$

⁶These two theorems actually say something slightly different.

Moreover, if $\mathbb{E}[X] = 0$, (a)-(d) are equivalent to

5.

$$\mathbb{E}[\exp(\lambda X)] \leq \exp(\lambda^2 \kappa_5^2 / 2) \quad \forall |\lambda| \leq \frac{1}{\kappa_5}.$$

Here, $\kappa_1, \dots, \kappa_5$ are universal constants.

We will not give the proof here, but you can check either textbook. Here is an example:

Example 4.2. Let $X_1 \sim \text{sG}(\sigma_1)$ and $X_2 \sim \text{sG}(\sigma_2)$ be not necessarily independent with $\mathbb{E}[X_1] = \mathbb{E}[X_2] = 0$. We claim that $X_1 X_2 \sim \text{sE}(K\sigma_1\sigma_2, K\sigma_1\sigma_2)$ for some universal K . For this, we can use property (b) above: First rescale X_1 and X_2 for simplicity. Using the Cauchy-Schwarz inequality,

$$\begin{aligned} \mathbb{E} \left[\left(\left| \frac{X_1}{\sigma_1} \right| \left| \frac{X_2}{\sigma_2} \right| \right)^p \right] &\leq \mathbb{E} \left[\left| \frac{X_1}{\sigma_1} \right|^{2p} \right]^{1/2} \mathbb{E} \left[\left| \frac{X_2}{\sigma_2} \right|^{2p} \right]^{1/2} \\ &= \left\| \frac{X_1}{\sigma_1} \right\|_{L^{2p}}^p \left\| \frac{X_2}{\sigma_2} \right\|_{L^{2p}}^p \end{aligned}$$

By the rescaling, $X_i/\sigma_i \sim \text{sG}(1)$ for $i = 1, 2$. The sub-Gaussian condition says that $\|G\|_{L^{2p}} \leq K(2p)^p$ for all p .

$$\begin{aligned} &\leq K^p(\sqrt{2p})^p \cdot K^p(\sqrt{2p})^p \\ &= K^{2p}(2p)^p. \end{aligned}$$

This tells us that $\left\| \frac{X_1 X_2}{\sigma_1 \sigma_2} \right\|_{L^p} \leq K^2 2p$ for all p .

4.5 Bennett's inequality

Here is a stronger bound for bounded random variables. Here, we don't require boundedness from below.

Lemma 4.3 (Bennett's inequality). *Let $(X_i)_{i \in [n]}$ be independent, where $X_i - \mathbb{E}[X_i] \leq b$ a.s., and $\nu_i^2 := \text{Var}(X_i)$ for all $i \in [n]$. Then*

$$\mathbb{P} \left(\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \geq t \right) \leq \exp \left(- \frac{\sum_{i=1}^n \nu_i^2}{b^2} h \left(\frac{bt}{\sum_{i=1}^n \nu_i^2} \right) \right),$$

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

Remark 4.3. This has a stronger assumption than Bernstein's inequality and provides a stronger bound than Bernstein's inequality for bounded random variables. However, it doesn't often improve much over Bernstein's inequality.

4.6 Maximal inequality

Lemma 4.4. *Let $(X_i)_{i \in [n]}$ be a sequence of random variables. For any convex, strictly increasing $\psi : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$, we have*

$$\begin{aligned}\mathbb{E} \left[\max_{i \in [n]} X_i \right] &\leq \psi^{-1} \left(\sum_{i=1}^n \mathbb{E}[\psi(X_i)] \right), \\ \mathbb{P} \left(\max_{i \in [n]} X_i \geq t \right) &\leq \sum_{i=1}^n \frac{\mathbb{E}[\psi(X_i)]}{\psi(t)}.\end{aligned}$$

Proof.

$$\mathbb{E} \left[\max_{i \in [n]} X_i \right] = \mathbb{E} \left[\psi^{-1} \left(\max_{i \in [n]} \psi(X_i) \right) \right]$$

Using Jensen's inequality,

$$\leq \psi^{-1} \left(\mathbb{E} \left[\max_{i \in [n]} \psi(X_i) \right] \right)$$

Upper bounding the maximum by the sum,

$$= \psi^{-1} \left(\sum_{i=1}^n \mathbb{E}[\psi(X_i)] \right). \quad \square$$

Example 4.3. For $X_i \sim \text{sG}(\sigma)$, take $\psi(u) = e^{\lambda u}$. Optimizing over λ , we get

$$\mathbb{E} \left[\max_{i \in [n]} X_i \right] \leq \sigma \sqrt{2 \log(n)}.$$

This gives an important intuition: n sub-Gaussian random variables have maximum of order $\sqrt{\log(n)}$.

4.7 Truncation argument

Here is a very useful technique in research for getting concentration inequalities for random variables which are not sub-Gaussian nor sub-exponential.

Example 4.4. Let $X_i = G_i^4$, where $(G_i)_{i \in [n]} \stackrel{\text{iid}}{\sim} N(0, 1)$. Then $\mathbb{E}[X_i] = \mathbb{E}[G_i^4] = 3$, but $\mathbb{E}[e^{\lambda X_i}]$ doesn't exist. However, we still want to upper bound $\frac{1}{n} \sum_{i=1}^n X_i - 3$.

Here is the technique:

Step 1: Find b_n such that

$$\mathbb{P} \left(\max_{i \in [n]} X_i \geq b_n \right) \leq \frac{\delta}{2}$$

and ε_n such that

$$\mathbb{E}[X_i \mathbb{1}_{\{X_i \geq b_n\}}] \leq \varepsilon_n.$$

Step 2: Apply Hoeffding/Bernstein and get

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n(X_i\mathbb{1}_{\{X_i\leq b_n\}} - \mathbb{E}[X_i\mathbb{1}_{\{X_i\leq b_n\}}]) \leq t_n\right) \geq 1 - \frac{\delta}{2}.$$

Step 3: Combining Steps 1 and 2 implies that

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n(X_i - \mathbb{E}[X_i]) \leq t_n + \varepsilon_n\right) \geq 1 - \delta.$$

As an exercise, figure out b_n, t_n, ε_n as a function of n and δ . The requirement is that $t_n + \varepsilon_n \sim \tilde{O}(\frac{1}{\sqrt{n}})$.

5 Martingale Concentration Inequalities

5.1 Motivation and overview

Our goal is to get a tail bound for $X_1 + \dots + X_n$, where the X_i are independent. Here is our solution so far:

- (a) Chernoff inequality bounded by MGF.
- (b) Bound MGF using sub-Gaussian and sub-exponential properties.
- (c) Many commonly used random variables are sub-Gaussian or sub-exponential.

What about more complicated structure?

1. Sometimes, we want to show concentration of $S_n = f(X_1, \dots, X_n) =: f(X_{1:n})$.
2. Sometimes, we want to show concentration of $S_n = \sum_{t=1}^T X_t$, where $\{X_t\}_{t \geq 1}$ is correlated. We can deal with this if it is a Martingale difference sequence.

This lecture, we will take the approach of a Martingale concentration inequality. We will use Markov's inequality on $e^{\lambda S_n}$ along with a conditional MGF bound and optimizing over λ . We will see

- (a) Doob's Martingale representation
- (b) Azuma-Hoeffding, Azuma- Bernstein, and bounded difference inequalities
- (c) Applications
- (d) Variants: Freedman's inequality and Doob's maximal inequality

Example 5.1. Suppose $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} P_X \in \mathcal{P}([a, b])$. We want to estimate $\theta = \mathbb{E}_{X, X' \stackrel{\text{iid}}{\sim} P_X} [g(X, X')]$, where we assume that $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ is symmetric (such as $g(x, x') = |x - x'|$ or $g(x, x') = \frac{1}{2}(x - x')^2$). In the latter case, $\theta = \text{Var}(X)$.

Hoeffding introduced ***U-statistics*** for estimating these parameters θ :

$$U(X_{1:n}) = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} g(X_i, X_j).$$

If we let

$$\hat{\mathbb{P}}_{X, X'} = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} \delta_{(X_i, X_j)}$$

be the empirical distribution, then $U(X_{1:n}) = \hat{E}_{(X, X')} [g(X, X')]$. The U statistic is an unbiased estimator of θ because

$$\mathbb{E}[U(X_{1:n})] = \mathbb{E}[g(X_i, X_j)] = \theta.$$

This has the smallest variance among all unbiased estimators.

Today, we will show the concentration bound

$$\mathbb{P}(|U - \theta| \geq t) \leq 2 \exp\left(-\frac{nt^2}{2\|g\|_\infty}\right).$$

This is significant because U is not a sum of independent random variables, so our previous technology does not work here.

5.2 Doob's martingale representation of $f(X_1, \dots, X_n)$

Now return to the setting where we are dealing with $f(X_1, \dots, X_n)$, where the X_i are independent. Define

$$Y_k = \mathbb{E}[f(X_{1:n}) \mid X_{1:k}] \quad k \geq 0.$$

We can think of conditioning on $X_{1:k}$ as conditioning on the σ -algebra $\mathcal{F}_k = \sigma(X_{1:k})$

Example 5.2. Here is the example to keep in mind: Let $f(X_{1:n}) = X_1 + \dots + X_n$ with independent X_i . Then

$$Y_k = X_1 + \dots + X_k + \mathbb{E}[X_{k+1}] + \dots + \mathbb{E}[X_n].$$

Further define the difference

$$D_k = Y_k - Y_{k-1}.$$

In the previous example, $D_k = X_k - \mathbb{E}[X_k]$. We can in general write

$$f(X) - \mathbb{E}[f(X)] = Y_n - Y_0 = \sum_{k=1}^n (Y_k - Y_{k-1}) = \sum_{k=1}^n D_k.$$

We call $\{Y_k\}$ a **martingale sequence** and $\{D_k\}$ a **martingale difference sequence**.

Let us recall what a martingale is.

Definition 5.1. A **filtration** is an increasing nested sequence of σ -algebras

$$\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \dots \subseteq \mathcal{F}_n \subseteq \dots$$

Often, we take $\mathcal{F}_k = \sigma(X_{1:k})$. If the filtration is not defined properly, the result you get may not be true.

Definition 5.2. If we have $\{Y_k\}_{k=1}^\infty$, where Y_k is \mathcal{F}_k -measurable, then we say that $\{Y_k\}$ is **$\{\mathcal{F}_k\}$ -adapted**.

Definition 5.3. $\{(Y_k, \mathcal{F}_k)\}_{k \geq 1}$ is a **martingale sequence** if

1. $\{Y_k\}$ is adapted to $\{\mathcal{F}_k\}$.

2. $\mathbb{E}[|Y_k|] < \infty$,
3. $\mathbb{E}[Y_k | \mathcal{F}_{k-1}] = Y_{k-1}$.

Martingales are often used to model gambling problems where your strategy can depend on the outcomes of the past. If you don't have a martingale, you can sometimes subtract the mean to get one.

Definition 5.4. $\{D_k\}_{k \geq 1}$ is a **martingale difference sequence** if $\{\sum_{k=1}^n D_k\}_{n \geq 1}$ is a martingale with respect to $\{\mathcal{F}_k\}_{k \geq 1}$.

Example 5.3. Let $\{X_i\}_{i \geq 1} \stackrel{\text{iid}}{\sim} P_X$, where $\mathbb{E}[|X|] < \infty$. Denote $\mu = \mathbb{E}_X[X]$ and $S_k = \sum_{s=1}^k X_s$. Then $\{(X_k - k\mu, \sigma(X_{1:k}))\}_{k \geq 1}$ is a martingale.

Proof. We only need to check the third property:

$$\begin{aligned} \mathbb{E}[S_k - k\mu | X_{1:k-1}] &= S_{k-1} - (k-1)\mu \\ &= Y_{k-1}. \end{aligned} \quad \square$$

Example 5.4 (Doob's martingale). Let $\{X_i\}_{i \geq 1}$ be independent⁷ and $\mathbb{E}[|f(X_1, \dots, X_n)|] < \infty$. Then $\{(Y_k = \mathbb{E}[f(X_{1:n}) | X_{1:k}], \sigma(X_{1:k}))\}_{k \geq 1}$ is a martingale sequence.

Proof. Again, we only check the third property:

$$\begin{aligned} \mathbb{E}[Y_{k+1} | \sigma(X_{1:k})] &= \mathbb{E}[\mathbb{E}[f(X_{1:n}) | X_{1:n+1}] | X_{1:k}] \\ &= \mathbb{E}[f(X_{1:n}) | X_{1:k}] \\ &= Y_k \end{aligned}$$

The second equality is by the tower property of conditional expectation. □

5.3 Martingale concentration

Most inequalities for an iid sum have a martingale version. Here is a martingale version of Bernstein's inequality.⁸

Theorem 5.1. Let $\{(D_k, \mathcal{F}_k)\}$ be a martingale difference sequence. If

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

then

1. $\sum_{k=1}^n D_k$ is $\text{sE}(\sqrt{\sum_{k=1}^n \nu_k^2}, \max_{k \leq n} \alpha_k)$.

⁷In class, we had this assumption, but I don't think it is actually needed.

⁸This inequality does not have a formal name, but you may call it an Azuma-Bernstein inequality.

2.

$$\mathbb{P}\left(\left|\sum_{k=1}^n D_k\right| \geq t\right) \leq 2 \exp\left(-\min\left\{\frac{t^2}{2\sum_{k=1}^n \nu_k^2}, \frac{t}{2\alpha_*}\right\}\right)$$

This condition is that a random variable given by the MGF is bounded. We will see later how to check this condition.

Proof. We can start with the Chernoff bound

$$\mathbb{P}\left(\sum_{k=1}^n D_k \geq t\right) \leq \inf_{\lambda} \frac{\mathbb{E}[e^{\lambda \sum_{k=1}^n D_k}]}{e^{\lambda t}}.$$

Then we can bound the moment generating function by using the tower property of conditional expectation

$$\mathbb{E}[e^{\lambda \sum_{k=1}^n D_k}] = \mathbb{E}[e^{\lambda \sum_{k=1}^{n-1} D_k} \mathbb{E}[e^{\lambda D_n} \mid \mathcal{F}_{n-1}]]$$

Using $\lambda \leq \frac{1}{\alpha_n}$,

$$\begin{aligned} &\leq \mathbb{E}[e^{\lambda \sum_{k=1}^{n-1} D_k} e^{\lambda^2 \nu_n^2 / 2}] \\ &= \mathbb{E}[e^{\lambda \sum_{k=1}^{n-1} D_k}] e^{\lambda^2 \nu_n^2 / 2} \end{aligned}$$

Iterating this argument, we get

$$\leq e^{\lambda^2 (\sum_{k=1}^n \nu_k^2) / 2}$$

for all $\lambda \leq \frac{1}{\max_{k \leq n} \alpha_k}$. □

Remark 5.1. In this theorem, the ν_k are deterministic. In the case where the ν_k are \mathcal{F}_{k-1} -measurable, we will get a related but different bound.

Here is a corollary which is sometimes easier to use than the previous theorem.

Corollary 5.1 (Azuma-Hoeffding inequality). *Let $\{(D_k, \mathcal{F}_k)\}$ be a martingale difference sequence. Suppose there exists $\{(a_k, b_k)\}_{k=1}^n$ such that $D_k \in (a_k, b_k)$ a.s., where b_k, a_k are \mathcal{F}_{k-1} -measurable and $|b_k - a_k| \leq L_k$. Then*

1. $\sum_{k=1}^n D_k$ is sG($\sqrt{\sum_{k=1}^n L_k^2 / 2}$).

2.

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

Proof. We have $\mathbb{E}[e^{\lambda D_k} \mid \mathcal{F}_{k-1}] \leq e^{\lambda^2 (b_k - a_k)^2 / 8}$. Use the same proof as before. □

Now specialize to Doob's martingale

$$D_k = \mathbb{E}[f(X_{1:n}) \mid X_{1:k}] - \mathbb{E}[f(X_{1:n}) \mid X_{1:k-1}].$$

Definition 5.5. $f(x_1, \dots, x_n)$ is a **bounded difference function** if for all $k \in [n]$, $x_{1:n}, x'_k$,

$$|f(x_{1:k-1}, x_k, x_{k+1:n}) - f(x_{1:k-1}, x'_k, x_{k+1:n})| \leq L_k.$$

This is a condition on how much the function changes if we change 1 coordinate. Here is a corollary of the Azuma-Hoeffding inequality

Corollary 5.2. *Suppose that $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is $L_{1:n}$ bounded and $X_{1:n}$ has independent components. Then for all $t \geq 0$,*

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

Proof. This is Azuma-Hoeffding with $\sum_{k=1}^n D_k = f(X_{1:n}) - \mathbb{E}[f(X_{1:n})]$. Here, there exist $A_k \leq D_k \leq B_k$, where $|B_k - A_k| \leq L_k$ because we can let

$$B_k = \sup_x \mathbb{E}[f(X_{1:n}) \mid X_{1:k-1}, X_k = x] - \mathbb{E}[f(X_{1:n}) \mid X_{1:k-1}],$$

$$A_k = \inf_x \mathbb{E}[f(X_{1:n}) \mid X_{1:k-1}, X_k = x] - \mathbb{E}[f(X_{1:n}) \mid X_{1:k-1}]. \quad \square$$

5.4 Applications

Example 5.5 (*U-statistics*). Here is how we can get a concentration inequality for U -statistics: Recall that

$$U(X_{1:n}) = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} |X_i - X_j|, \quad X_i \sim P_X \in \mathcal{P}([-b, b]).$$

Then

$$\begin{aligned} |U(X_{1:k-1}, X_k, X_{k+1:n}) - U(X_{1:k-1}, X'_k, X_{k+1:n})| &= \frac{1}{\binom{n}{2}} \left| \sum_{s \neq k} |X_s - X_k| - |X_s - X'_k| \right| \\ &\leq \frac{1}{\binom{n}{2}} \sum_{s \neq k} |X_k - X'_k| \\ &\leq \frac{2}{n(n-1)} \cdot (n-1) \cdot 2b \\ &\leq \frac{4b}{n}. \end{aligned}$$

So U is $(\frac{4b}{n}, \frac{4b}{n}, \dots, \frac{4b}{n})$ -bounded difference. This gives the tail bound

$$\mathbb{P}(|U(X_{1:n}) - \theta| \geq t) \leq 2 \exp\left(-\frac{2t^2}{n \frac{16}{n^2}}\right) = 2 \exp\left(-\frac{nt^2}{16}\right).$$

That is,

$$|U(X_{1:n}) - \theta| \lesssim b \sqrt{\frac{\log(2/\delta)}{n}} \quad \text{with probability } 1 - \delta.$$

Example 5.6 (Supremum of empirical process). Suppose we have samples $(Z_i)_{i \in [n]} \stackrel{\text{iid}}{\sim} P_Z$, where $Z_i = (X_i, Y_i)$. We can define the **loss function** $\ell : Z \times \Theta \rightarrow [0, 1]$ and the **empirical risk**

$$\widehat{R}_n(\theta) = \frac{1}{n} \sum_{k=1}^n \ell(Z_k; \theta).$$

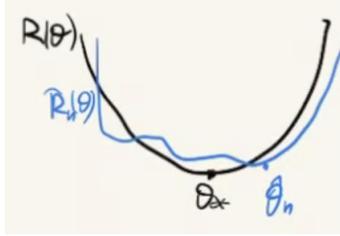
Correspondingly, we have the **population risk**

$$R(\theta) = \mathbb{E}[\widehat{R}_n | \theta] = \mathbb{E}[\ell(Z; \theta)]$$

In statistical learning theory, we are often concerned with the **excess risk**

$$\mathcal{E}[Z_{1:n}] := \sup_{\theta \in \Theta} R(\theta) - \widehat{R}_n(\theta).$$

We can use an **empirical risk minimizer** $\widehat{\theta}_n$, and we want to upper bound $R(\widehat{\theta}_n) \leq \widehat{R}_n(\widehat{\theta}_n) + \mathcal{E}(Z_{1:n})$.



We claim that $\mathcal{E}(Z_{1:n})$ is $(1/n, \dots, 1/n)$ -bounded difference. Then

$$|\mathcal{E}(Z_{1:n}) - \mathbb{E}[\mathcal{E}(Z_{1:n})]| \leq \sqrt{\frac{\log(2/\delta)}{2n}} \quad \text{with probability } 1 - \delta.$$

Proof. Fix $Z_{1:n}$, and let $\theta_* = \arg \max_{\theta \in \Theta} (R(\theta) - \widehat{R}_n(\theta))$. Then $\mathcal{E}(Z_{1:n}) = R(\theta_*) - \widehat{R}_n(\theta_*)$. We want to look at

$$|\mathcal{E}(Z_{1:n}) - \mathcal{E}(Z_{1:k-1}, Z'_k, Z_{k+1:n})| = \frac{1}{n} \sum_{i=1}^n (\ell(Z_i; \theta_*) - \mathbb{E}[\ell(Z_i; \theta_*)])$$

$$\begin{aligned}
& - \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i \neq k} (\ell(Z_i; \theta) - \mathbb{E}[\ell(Z_i; \theta)]) \\
& - \frac{1}{n} (\ell(Z'_k; \theta) - \mathbb{E}[\ell(Z'_k; \theta)]) \\
\leq & \frac{1}{n} \sum_{i=1}^n (\ell(Z_i; \theta_*) - \mathbb{E}[\ell(Z_i; \theta_*)]) \\
& - \frac{1}{n} \sum_{i \neq k} (\ell(Z_i; \theta_*) - \mathbb{E}[\ell(Z_i; \theta_*)]) \\
& - \frac{1}{n} (\ell(Z'_k; \theta_*) - \mathbb{E}[\ell(Z'_k; \theta_*)]) \\
= & \frac{1}{n} (\ell(Z_k; \theta_*) - \ell(Z'_k; \theta_*)) \\
\leq & \frac{1}{n}. \quad \square
\end{aligned}$$

Remark 5.2. This doesn't say anything about

$$\mathbb{E} \left[\sup_{\theta} \widehat{R}_n(\theta) - R(\theta) \right].$$

5.5 Freedman's inequality

Our “Azuma-Bernstein” inequality says that if $\mathbb{E}[e^{\lambda D_k} \mid \mathcal{F}_{k-1}] \leq e^{\lambda^2 \nu_k^2 / 2}$, then

$$\left| \frac{1}{n} \sum_{k=1}^n D_k \right| \leq \max \left\{ \sqrt{\frac{\frac{2}{n} \sum_{k=1}^n \nu_k^2}{n} \log \left(\frac{2}{\delta} \right)}, \frac{2\alpha_* \log \left(\frac{2}{\delta} \right)}{n} \right\} \quad \text{with probability } 1 - \delta.$$

However, sometimes ν_k^2 is not deterministic and instead is \mathcal{F}_{k-1} measurable.

Theorem 5.2 (Freedman's inequality). *Let $\{(D_k, \mathcal{F}_k)\}$ be a martingale difference sequence such that*

1. $\mathbb{E}[D_k \mid \mathcal{F}_{k-1}] = 0$.
2. $D_k \leq b$ a.s.

Then for all $\lambda \in (0, 1/b)$ and $\delta \in (0, 1)$,

$$\mathbb{P} \left(\sum_{t=1}^T X_t \leq \lambda \sum_{t=1}^T \mathbb{E}[D_k^2 \mid \mathcal{F}_{k-1}] + \frac{\log(1/\delta)}{\lambda} \right) \geq 1 - \delta.$$

This is useful in bandit and reinforcement learning research.⁹

⁹For example, see Theorem 1 in Beygelzimer, Langford, et. al. 2010.

5.6 Maximal Azuma-Hoeffding inequality

Recall Doob's maximal inequality for sub-martingales.

Lemma 5.1 (Doob's maximal inequality). *If $\{X_s\}_{s \geq 0}$ is a sub-martingale, i.e.*

$$X_s \leq \mathbb{E}[X_t \mid \mathcal{F}_s] \quad \forall s < t,$$

then for all $u > 0$,

$$\mathbb{P} \left(\sup_{0 \leq t \leq T} X_t \geq u \right) \leq \frac{\mathbb{E}[\max\{X_T, 0\}]}{u}.$$

This gives rise to a maximal version of the Azuma-Hoeffding inequality:

Theorem 5.3 (Maximal Azuma-Hoeffding inequality). *Let $\{(D_k, \mathcal{F}_k)\}$ be a martingale difference sequence, and suppose there exists $\{(a_k, b_k)\}_{k=1}^n$ such that $D_k \in (a_k, b_k)$ a.s. Then*

$$\mathbb{P} \left(\sup_{0 \leq k \leq n} \sum_{s=1}^k D_s \geq t \right) \leq \exp \left(- \frac{2t^2}{\sum_{k=1}^n (b_k - a_k)^2} \right).$$

6 Gaussian Concentration

6.1 Freedman’s inequality

Last time, we generalized the Hoeffding and Bernstein inequalities for independent random variables to Azuma-Hoeffding and “Azuma Bernstein inequalities for martingales.”

Our “Azuma-Bernstein” inequality says that if $\mathbb{E}[e^{\lambda D_k} \mid \mathcal{F}_{k-1}] \leq e^{\lambda^2 \nu_k^2 / 2}$, then

$$\left| \frac{1}{n} \sum_{k=1}^n D_k \right| \leq \max \left\{ \sqrt{\frac{\frac{2}{n} \sum_{k=1}^n \nu_k^2 \log \left(\frac{2}{\delta} \right)}{n}}, \frac{2\alpha_* \log \left(\frac{2}{\delta} \right)}{n} \right\} \quad \text{with probability } 1 - \delta.$$

However, sometimes ν_k^2 is not deterministic and $\nu_k^2 = \mathbb{E}[D_k^2 \mid \mathcal{F}_{k-1}]$ instead is \mathcal{F}_{k-1} measurable.

Theorem 6.1 (Freedman’s inequality). *Let $\{(D_k, \mathcal{F}_k)\}$ be a martingale difference sequence such that*

1. $\mathbb{E}[D_k \mid \mathcal{F}_{k-1}] = 0$.
2. $D_k \leq b$ a.s.

Then for all $\lambda \in (0, 1/b)$ and $\delta \in (0, 1)$,

$$\mathbb{P} \left(\sum_{t=1}^T X_t \leq \lambda \sum_{t=1}^T \mathbb{E}[D_k^2 \mid \mathcal{F}_{k-1}] + \frac{\log(1/\delta)}{\lambda} \right) \geq 1 - \delta.$$

This is useful in bandit and reinforcement learning research.¹⁰

6.2 Maximal Azuma-Hoeffding inequality

Recall Doob’s maximal inequality for sub-martingales.

Lemma 6.1 (Doob’s maximal inequality). *If $\{X_s\}_{s \geq 0}$ is a sub-martingale, i.e.*

$$X_s \leq \mathbb{E}[X_t \mid \mathcal{F}_s] \quad \forall s < t,$$

then for all $u > 0$,

$$\mathbb{P} \left(\sup_{0 \leq t \leq T} X_t \geq u \right) \leq \frac{\mathbb{E}[\max\{X_T, 0\}]}{u}.$$

This gives rise to a maximal version of the Azuma-Hoeffding inequality:

¹⁰For example, see Theorem 1 in Beygelzimer, Langford, et. al. 2010.

Theorem 6.2 (Maximal Azuma-Hoeffding inequality). *Let $\{(D_k, \mathcal{F}_k)\}$ be a martingale difference sequence, and suppose there exists $\{(a_k, b_k)\}_{k=1}^n$ such that $D_k \in (a_k, b_k)$ a.s. Then*

$$\mathbb{P}\left(\sup_{0 \leq k \leq n} \sum_{s=1}^k D_k \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{k=1}^n (b_k - a_k)^2}\right).$$

If we used the usual Azuma-Hoeffding inequality instead, we would need to use a union bound, which would give a factor of n in the bound. We can write this conclusion as

$$\sup_{0 \leq k \leq n} \sum_{s=1}^k D_k \leq \sqrt{\frac{C \log(1/\delta)}{n}}.$$

If we have the extra factor of n , we get an n/δ instead, which can sometimes be not a big deal for our bound since we are taking a log.

6.3 Gaussian concentration

Lemma 6.2. *Let $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} N(0, 1)$ and $f : \mathbb{R}^n \rightarrow \mathbb{R}$ such that f is L -Lipschitz in $\|\cdot\|_2$, i.e.*

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

Then

1. $f(X_{1:n}) - \mathbb{E}[f(X_{1:n})]$ is sG(L).
- 2.

$$\mathbb{P}(|f(X_{1:n}) - \mathbb{E}[f(X_{1:n})]| \geq t) \leq 2 \exp\left(-\frac{t^2}{2L^2}\right).$$

Remark 6.1. We need f to be Lipschitz as a whole function! It's not just sufficient for the function to be coordinate-wise Lipschitz.

Remark 6.2. If the X_i s are non-Gaussian, this doesn't always hold with only Lipschitz-ness.

There are many different proofs of this lemma, but none are very simple.

Proof 1: Gaussian interpolation method

Proof 2: Gaussian isoperimetric inequality

Proof 3: Gaussian log-Sobolev inequality + Herbst argument

Today, we will present a proof using the Gaussian interpolation method, which is useful in research. However, this is a technique where you need to develop some intuition to understand it.

6.4 Examples of Gaussian concentration

Example 6.1 (Order statistics). Let $(X_i)_{i \in [n]} \stackrel{\text{iid}}{\sim} N(0, 1)$. The order statistics are the random variables arranged in increasing order: $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$. Let $f_k(X_{1:n}) = X_{(k)}$. This is Lipschitz:

$$\begin{aligned} |f_k(X_{1:n}) - f_k(Y_{1:n})| &= |X_{(k)} - Y_{(k)}| \\ &\leq \sqrt{\sum_{k=1}^n |X_{(k)} - Y_{(k)}|^2} \end{aligned}$$

The **rearrangement inequality** says that if you sort the terms, the distance is greater than the distance of with unsorted terms.

$$\begin{aligned} &\leq \sqrt{\sum_{k=1}^n |X_k - Y_k|^2} \\ &= \|X - Y\|_2. \end{aligned}$$

This means that $L = 1$, so $X_{(k)} - \mathbb{E}[X_{(k)}]$ is sG(1). Therefore,

$$|X_{(k)} - \mathbb{E}[X_{(k)}]| \leq \sqrt{\log(2/\delta)} \quad \text{with probability } 1 - \delta.$$

If we apply this to $k = n$, we get

$$\left| \max_{i \in [n]} X_i - \underbrace{\mathbb{E} \left[\max_{i \in [n]} X_i \right]}_{\sqrt{2 \log n}} \right| = O_p(1).$$

Example 6.2 (Singular value of Gaussian random matrices). Let

$$X = \begin{bmatrix} X_{1,1} & \cdots & X_{1,d} \\ \vdots & & \vdots \\ X_{n,1} & \cdots & X_{n,d} \end{bmatrix} \in \mathbb{R}^{n \times d}, \quad X_{i,j} \stackrel{\text{iid}}{\sim} N(0, 1).$$

Let $f_k(X) = \sigma_k(X)$ be the k -th largest singular value of X . For example, $f_1(X) = \|X\|_{\text{op}}$. It can be shown that $\mathbb{E}[\|X\|_{\text{op}}] \approx \sqrt{n} + \sqrt{d}$. We can show that f_k is Lipschitz; what is the norm we want to be using for a matrix? Define the vectorized version of the matrix as $\text{vec}(X) := (X_{1,1}, X_{1,2}, \dots, X_{1,d}, X_{2,1}, \dots, X_{2,d}, \dots, X_{n,d})$. Then

$$\|\text{vec}(X) - \text{vec}(Y)\|_2 = \|X - Y\|_F = \sqrt{\sum_{i,j} (X_{i,j} - Y_{i,j})^2},$$

where $\|\cdot\|_F$ is the **Frobenius norm**. Now we have

$$|f_k(X) - f_k(Y)| \leq |\sigma_k(X) - \sigma_k(Y)|$$

Weyl's inequality, a deterministic linear algebra result, says that

$$\begin{aligned} &\leq \|X - Y\|_{\text{op}} \\ &\leq \|X - Y\|_F, \end{aligned}$$

so $L = 1$. Weyl's inequality can be proven by using the variational representation of singular values.

This calculation tells us that $f_k(X) - \mathbb{E}[f_k(X)]$ is sG(1), so

$$f_k(X) - \mathbb{E}[f_k(X)] \leq \sqrt{\log(2/\delta)} \quad \text{with probability } 1 - \delta.$$

Applying this to $k = 1$ gives

$$\left| \|X\|_{\text{op}} - \underbrace{\mathbb{E}[\|X\|_{\text{op}}]}_{\sqrt{n} + \sqrt{d}} \right| = O(1).$$

6.5 Gaussian complexity

Gaussian complexity is a very important notion in compressed sensing. Suppose we have a set $A \subseteq \mathbb{R}^n$. How do we measure its "size"? A reasonable size function S should at least satisfy $S(A) \leq S(B)$ if $A \subseteq B$. Here are some reasonable size functions:

1. Euclidean width: $D(A) = \max_{a \in A} \|a\|_2$.
2. Dimension: A line has dimension 1, and a plane has dimension 2.

Definition 6.1. Given a set A , let $W = (W_1, \dots, W_n)^\top \in \mathbb{R}^n$ with $W_i \stackrel{\text{iid}}{\sim} N(0, 1)$. The **Gaussian complexity** or "statistical dimension" of A is

$$\mathcal{G}(A) := \mathbb{E}_{W \sim N(0, I_n)} \left[\sup_{a \in A} \langle a, W \rangle \right].$$

Note that if we don't take the supremum in the expectation, the quantity would be 0. This quantity is always nonnegative.

Example 6.3. Let $B_p(r) = \{x \in \mathbb{R}^n : \|x\|_p \leq r\}$. Then

$$\mathcal{G}(B_p(r)) = \mathbb{E} \left[\sup_{\|x\|_p \leq r} \langle x, W \rangle \right]$$

If q is the conjugate exponent of p , so $\frac{1}{p} + \frac{1}{q} = 1$, this is the variational representation of the $\|\cdot\|_q$ norm:

$$r \mathbb{E}[\|W\|_q]$$

$$\approx rn^{1/q}.$$

Note that if $p_1 \leq p_2$, then $q_1 \geq q_2$, so $\mathcal{G}(B_{p_1}(r)) \leq \mathcal{G}(B_{p_2}(r))$.

We want to show that $f(W) := \sup_{a \in A} \langle a, W \rangle$ concentrates. Fix $w, w' \in \mathbb{R}^n$. Then

$$f(w) - f(w') = \sup_{a \in A} \langle a, w \rangle - \sup_{a \in A} \langle a, w' \rangle$$

Denote $a^* = \arg \max_a \langle a, w \rangle$

$$\begin{aligned} &= \langle a^*, w \rangle - \sup_{a \in A} \langle a, w' \rangle \\ &= \inf_{a \in A} \langle a^*, w \rangle - \langle a, w' \rangle \\ &\leq \langle a^* w - w' \rangle \\ &\leq \|a^*\| \|w - w'\|_2 \\ &\leq D(A) \|w - w'\|_2. \end{aligned}$$

The other side can be proven similarly, so f is $D(A)$ -Lipschitz. Concentration says that $f(W)$ is $\text{sG}(D(A))$.

Example 6.4. If we let $A = B_2(R)$, then

$$\mathbb{E}[f(W)] = \mathcal{G}(B_2(r)) = r\sqrt{n},$$

since $D(A) = r$.

6.6 Proof of the Gaussian concentration inequality (interpolation method)

Lemma 6.3. For all convex $\phi : \mathbb{R} \rightarrow \mathbb{R}$ and differentiable $f : \mathbb{R}^n \rightarrow \mathbb{R}$,

$$\mathbb{E}[\phi(f(X) - \mathbb{E}[f(Y)])] \leq \mathbb{E}[\phi(\frac{\pi}{2} \langle \nabla f(X), Y \rangle)],$$

where $X, Y \stackrel{\text{iid}}{\sim} N(0, I_n)$.

First, assume this lemma holds, and prove Gaussian concentration:

Proof. Take $\phi = \exp(\lambda \cdot)$. Then

$$\mathbb{E}[\exp(\lambda(f(X) - \mathbb{E}[f(Y)]))] \leq \mathbb{E}[\exp(\lambda \frac{\pi}{2} \langle \nabla f(X), Y \rangle)]$$

Observe that $\frac{\pi}{2} \langle \nabla f(X), Y \rangle$ is $N(0, \frac{\pi^2}{4} \|\nabla f(X)\|_2^2)$ given X .

$$\begin{aligned} &= \mathbb{E}_X [\exp(\frac{\lambda^2}{2} \frac{\pi^2}{4} \|\nabla f(X)\|_2^2)] \\ &\leq \exp\left(\frac{\lambda^2}{2} \frac{\pi^2}{4} L^2\right). \end{aligned}$$

This says that $f(X) - \mathbb{E}[f(X)]$ is $\text{sG}(\frac{\pi}{2}L)$. □

The above proof gives a worse constant, but the constant can be improved with different methods. Here is the proof of the lemma:

Proof. First, use conditioning and Jensen's inequality to say that.

$$\mathbb{E}[\phi(f(X) - \mathbb{E}[f(Y)])] \leq \mathbb{E}_{X,Y}[\phi(f(X) - f(Y))]$$

The idea is to use the integral representation of the Taylor expansion to interpolate between X and Y . Observe that if $Z(\theta) = X \cos \theta + Y \sin \theta$, then for every θ , $Z(\theta) \stackrel{d}{=} X \stackrel{d}{=} Y$ and $Z'(\theta) \stackrel{d}{=} X \stackrel{d}{=} Y$. Another important property is that $Z(\theta) \perp Z'(\theta)$; this is because $Z(\theta), Z'(\theta)$ are Gaussians with 0 covariance. Now

$$f(X) - f(Y) = \int_0^{\pi/2} \langle \nabla f(Z(\theta)), Z'(\theta) \rangle d\theta,$$

so we can write

$$\mathbb{E}[\phi(f(X) - f(Y))] = \mathbb{E} \left[\phi \left(\int_0^{\pi/2} \langle \nabla f(Z(\theta)), Z'(\theta) \rangle d\theta \right) \right]$$

Using Jensen's inequality, $\phi(\int \cdot d\theta) \leq \int \phi(\cdot) d\theta$ when $\int \cdot d\theta = 1$.

$$\begin{aligned} &\leq \frac{2}{\pi} \int_0^{\pi/2} \mathbb{E}[\phi(\frac{\pi}{2} \langle \nabla f(Z(\theta)), Z'(\theta) \rangle)] d\theta \\ &= \mathbb{E}[\phi(\frac{\pi}{2} \langle \nabla f(X), Y \rangle)]. \end{aligned}$$

□

This proof is very delicate, and the construction looks ad hoc, but it is actually very useful in a variety of situations.

6.7 Other methods for establishing concentration

1. Matrix concentration: If $(X_i)_{i \in [n]} \subseteq \mathbb{R}^{m \times d}$ with $X_i \stackrel{\text{iid}}{\sim} X$, can we find a bound for

$$\left\| \frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}[X_i] \right\|_{\text{op}} ?$$

The answer is yes; there is a matrix Bernstein inequality, Rudelson's inequality, and a matrix Freedman inequality. These involve the matrix MGF and Lieb's inequality. For more, see *An Introduction to Matrix Inequalities*, Tropp 2015, and *Introduction to Non-asymptotic analysis of random matrices*, Vershynin 2010.

2. Entropy method and the Herbst argument

Definition 6.2. The **Herbst** argument is that a sufficient condition for X to be $\text{sG}(\sigma)$ is to show that

$$\mathbb{H}(e^{\lambda X}) \leq \frac{\lambda^2 \sigma^2}{2} \mathbb{E}[e^{\lambda X}],$$

where \mathbb{H} is the entropy.

Why do we want to look at $\mathbb{H}(e^{\lambda X})$? This is because it has a good **tensorization property** when X_i are independent:

$$\mathbb{H}(e^{\lambda f(X_{1:n})}) \leq \mathbb{E} \left[\underbrace{\sum_{i=1}^n \mathbb{H}(e^{\lambda f_k(X_k)} \mid X^{\setminus k})}_{\substack{\text{easy to handle when} \\ f_k \text{ Lip.}, X_k \text{ bdd.}}} \right]$$

For this, see chapter 3.1 of Wainwright's textbook or chapter 3 of van Handel's textbook.

3. Isoperimetric inequality: This is a geometric property in \mathbb{R}^n with Lebesgue measure. If $A \subseteq \mathbb{R}^n$ has fixed volume and we want to minimize the perimeter, then the solution is when A is a ball. This generalizes to other measures:

$X \sim \mu =$	$N(0, I_n)$	$S^{n-1}(\sqrt{n})$	$\text{Unif}(\{\pm 1\}^n)$
	Half space	Spherical cap	Hamming ball

The isoperimetric inequality implies that $f(X)$ concentrates when f is Lipschitz. For this, see chapter 3.2 of Wainwright's book and also see Chapter 7 of the book by Lugosi, Massart, and Boucheron.

4. Transportation approach:

Lemma 6.4 (Bobkov-Gotze). *Given a measure $\mu \in \mathcal{P}(\mathbb{R}^n)$,*

$$X \sim \mu, \forall f \text{ 1-Lipschitz, } f(X) \text{ is } \text{sG}(\sigma) \iff W_1(\nu, \mu) \leq \sqrt{2\sigma^2 \text{KL}(\nu \parallel \mu)} \forall \nu \in \mathcal{P}(\mathbb{R}^n),$$

where W_1 is the transportation distance and KL is the relative entropy.

This property on the right also tensorizes in some way. For more on this, see chapter 3.3 in Wainwright's book or chapter 4 in van Handel's book.

7 Concentration Inequalities for Convex Functions

7.1 Overview

Let X_1, X_2, \dots, X_n be independent, and let $Z = f(X_{1:n})$, where $f : \mathbb{R}^n \rightarrow \mathbb{R}$. We have been asking the question: “when is there a high probability bound for $|Z - \mathbb{E}[Z]|$.”

Earlier, we had a solution in terms of the bounded differences inequality:

Theorem 7.1 (Bounded differences inequality). *Suppose that $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is $L_{1:n}$ bounded, i.e.*

$$|f(X_{1:n}) - f(x_{1:k-1}, x'_k, x_{k+1:n})| \leq L_k \quad \forall x_{1:n}, x_k,$$

and $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} N(0, 1)$. Then for all $t \geq 0$,

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

This martingale concentration method let us control U-statistics and the supremum of an empirical process.

Last lecture, we had the Gaussian concentration inequality:

Theorem 7.2 (Gaussian concentration). *Let $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} N(0, 1)$ and $f : \mathbb{R}^n \rightarrow \mathbb{R}$ such that f is L -Lipschitz in $\|\cdot\|_2$, i.e.*

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

Then

1. $f(X_{1:n}) - \mathbb{E}[f(X_{1:n})]$ is sG(L).

2.

$$\mathbb{P}(|f(X_{1:n}) - \mathbb{E}[f(X_{1:n})]| \geq t) \leq 2 \exp\left(-\frac{t^2}{2L^2}\right).$$

This let us discuss the singular values of a Gaussian random matrix and Gaussian complexity. To generalize this, the intuition is that we need 2 components:

1. We require the function f to not change much under perturbation of x .
2. We require the measure of X to behave sufficiently nicely.

7.2 Concentration of separately convex, Lipschitz functions

Theorem 7.3 (Concentration of separately convex, Lipschitz functions). *Suppose that*

1. *f is L -Lipschitz and coordinatewise convex:*

$$\partial_k^2 f(x_{1:n}) \geq 0 \quad \text{if } \partial_k^2 f \text{ exists}$$

2. *$(X_i)_{i \in [n]}$ independent with $X_i \in [a, b]$ a.s.*

Then

$$\mathbb{P}(f(X_{1:n}) - \mathbb{E}[f(X_{1:n})] \geq t) \leq \exp\left(-\frac{t^2}{2L^2(b-a)^2}\right).$$

This is a one-sided inequality; we don't have a lower tail bound here. To derive this result, we use the entropy method and the Herbst argument. This is covered in chapter 3.1 in Wainwright's textbook.

Remark 7.1. This has a stronger assumption than the bounded difference inequality, but it gives a stronger result.

7.3 Concentration of convex Lipschitz functions

Theorem 7.4 (Concentration of convex Lipschitz functions). *Suppose that*

1. *f is L -Lipschitz and convex:*

$$\nabla^2 f(x) \succ 0 \quad \text{if } \nabla^2 f \text{ exists}$$

2. *$(X_i)_{i \in [n]}$ independent with $X_i \in [a, b]$ a.s.*

Then $f(X_{1:n}) - \mathbb{E}[f]$ is $\text{sG}(L(b-a))$, so

$$\mathbb{P}(f(X_{1:n}) - \mathbb{E}[f(X_{1:n})] \geq t) \leq 2 \exp\left(-\frac{t^2}{2L^2(b-a)^2}\right).$$

Remark 7.2. Unlike the previous inequality, this one gives us an upper and lower tail bound. This has a stronger assumption than separate convexity, but it gives a stronger result.

To derive this result, Wainwright's book use a transportation approach. This is in chapter 3.6.

7.4 Applications

7.4.1 Rademacher complexity

If $A \subseteq \mathbb{R}^n$, how do we measure its size? We previously defined the **Gaussian complexity**

$$\mathcal{G}(A) : -\mathbb{E}_{W \stackrel{\text{iid}}{\sim} N(0,1)} \left[\sup_{a \in A} \langle W, a \rangle \right].$$

Definition 7.1. The **Rademacher complexity** is

$$\mathcal{R}(A) : -\mathbb{E}_{\varepsilon_i \stackrel{\text{iid}}{\sim} \text{Unif}(\{\pm 1\})} \left[\sup_{a \in A} \langle \varepsilon, a \rangle \right].$$

These notions are related, but they are useful in different situations.

Example 7.1. For all $1 < p < \infty$,

$$\begin{aligned} \mathcal{R}(B_p(r)) &= \mathbb{E}_\varepsilon \left[\sup_{\|a\|_p \leq r} \langle a, \varepsilon \rangle \right] = r \mathbb{E}_\varepsilon [\|\varepsilon\|_q] = rn^{1/q}, \\ \mathcal{G}(B_p(r)) &= rc_q n^{1/q}, \end{aligned}$$

where $\frac{1}{p} + \frac{1}{q} = 1$.

If $p = 1$, then

$$\begin{aligned} \mathcal{R}(B_1(r)) &= r \mathbb{E}_\varepsilon [\|\varepsilon\|_\infty] = r, \\ \mathcal{G}(B_1(r)) &= \mathbb{E}_W \left[\sup_{i \in [n]} |W_i| \right] \approx r \sqrt{2 \log n} + O(1). \end{aligned}$$

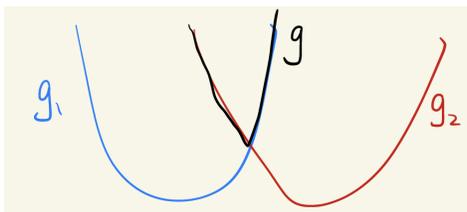
Here is an exercise from Wainwright's book.

Proposition 7.1. *There exist universal constants c, C such that for all $A \subseteq \mathbb{R}^n$,*

$$c\mathcal{R}(A) \leq \mathcal{G}(A) \leq C\sqrt{\log n}\mathcal{R}(A).$$

If we want to talk about concentration of Rademacher random variables, we can use the above concentration inequalities. Define $f(\varepsilon) = \sup_{a \in A} \langle \varepsilon, a \rangle$. Then $f(\varepsilon)$ is $D(A)$ -Lipschitz, where $D(A) = \sup_{a \in A} \|a\|_2$.

Lemma 7.1. *Let g_1, g_2 be convex functions. Then $g(x) = \max\{g_1(x), g_2(x)\}$ is convex.*



This implies that $f(\varepsilon)$ is convex. So we get that $f(\varepsilon) - \mathbb{E}[f(\varepsilon)]$ is $\text{sG}(2D(A))$. This tells us that

$$f(\varepsilon) \approx \mathcal{R}(A) + O(D(A)).$$

7.4.2 Operator norm

Let

$$X = \begin{bmatrix} X_{1,1} & \cdots & X_{1,d} \\ \vdots & & \vdots \\ X_{n,1} & \cdots & X_{n,d} \end{bmatrix} \in \mathbb{R}^{n \times d},$$

where $X_{i,j} \in [-1, 1]$ a.s. Then, if we let $f(x) = \|x\|_{\text{op}}$, then f is 1-Lipschitz and convex. So $f(X) = \mathbb{E}[f(X)]$ is $\text{sG}(2)$, which tells us that

$$\|X\|_{\text{op}} \simeq \mathbb{E}[\|X\|_{\text{op}}] + O(1).$$

7.5 Proof techniques: the Herbst argument and transportation

Here is how we can prove the above concentration inequalities.

1. Entropy method and the Herbst argument

Definition 7.2. The **Herbst** argument is that a sufficient condition for X to be $\text{sG}(\sigma)$ is to show that

$$\mathbb{H}(e^{\lambda X}) \leq \frac{\lambda^2 \sigma^2}{2} \mathbb{E}[e^{\lambda X}],$$

where \mathbb{H} is the entropy.

Why do we want to look at $\mathbb{H}(e^{\lambda X})$? This is because it has a good **tensorization property** when X_i are independent:

$$\mathbb{H}(e^{\lambda f(X_{1:n})}) \leq \mathbb{E} \left[\sum_{i=1}^n \underbrace{\mathbb{H}(e^{\lambda f_k(X_k)} \mid X^{\setminus k})}_{\substack{\text{easy to handle when} \\ f_k \text{ Lip.}, X_k \text{ bdd.}}} \right]$$

For this, see chapter 3.1 of Wainwright's textbook or chapter 3 of van Handel's textbook.

2. Transportation approach:

Lemma 7.2 (Bobkov-Gotze). *Given a measure $\mu \in \mathcal{P}(\mathbb{R}^n)$,*

$$X \sim \mu, \forall f \text{ 1-Lipschitz, } f(X) \text{ is } \text{sG}(\sigma) \iff W_1(\nu, \mu) \leq \sqrt{2\sigma^2 \text{KL}(\nu \parallel \mu)} \forall \nu \in \mathcal{P}(\mathbb{R}^n),$$

where W_1 is the transportation distance and KL is the relative entropy.

This property on the right also tensorizes in some way. For more on this, see chapter 3.3 in Wainwright's book or chapter 4 in van Handel's book.

7.6 Concentration of Lipschitz functions of log-concave random variables

Definition 7.3. A function $\psi : \mathbb{R}^n \rightarrow \mathbb{R}$ is r -strongly convex if $\nabla^2 \psi(x) \succeq rI_n$, if this exists.

Definition 7.4. If $\mu \in \mathcal{P}(\mathbb{R}^n)$, we say that μ is r -strongly log-concave if $\mu(x) = \exp(-\psi(x))$, where ψ is r -strongly convex.

Example 7.2. Let $p_\theta(x) = \frac{1}{Z(\theta)} \exp(\langle \theta, T(x) \rangle)$ be an exponential family. Suppose we have the prior

$$\pi(\theta) \sim N(0, I_n)$$

and the posterior

$$p(\theta | x) \propto p_\theta(x) \pi(\theta) = \frac{1}{\tilde{Z}(x)} \exp(\langle \theta, T(x) \rangle - \log Z(\theta) - \frac{1}{2} \|\theta\|_2^2)$$

So we may let

$$\psi(\theta) = -\langle \theta, T(x) \rangle + \log Z(\theta) + \frac{1}{2} \|\theta\|_2^2 + \log \tilde{Z}(x).$$

Note that

$$(\log Z(\theta))'' = \text{Cov}_\theta(T(X), T(X)) \geq 0.$$

Theorem 7.5 (Concentration of Lipschitz functions of log-concave random variables).

Suppose that

1. f is L -Lipschitz,
2. $X \sim \mu \in \mathcal{P}(\mathbb{R}^n)$, where μ is r -log-concave.

Then $f(X_{1:n}) - \mathbb{E}[f(X_{1:n})]$ is $\text{sG}(L/\sqrt{r})$.

7.7 Proof technique: the isoperimetric inequality

The isoperimetric inequality is a geometric property in \mathbb{R}^n with Lebesgue measure. If $A \subseteq \mathbb{R}^n$ has fixed volume and we want to minimize the perimeter, then the solution is when A is a ball. This generalizes to other measures:

$X \sim \mu =$		$N(0, I_n)$		$S^{n-1}(\sqrt{n})$		$\text{Unif}(\{\pm 1\}^n)$
		Half space		Spherical cap		Hamming ball

The isoperimetric inequality implies that $f(X)$ concentrates when f is Lipschitz. Suppose that $\mathbb{P}(A) = 1/2$, and take μ to be, for example, the Gaussian measure. Then define $A_\varepsilon = \{a : \exists b \in A \text{ s.t. } \|a - b\| \leq \varepsilon\}$. In this situation, perimeter is defined as

$$\lim_{\varepsilon \rightarrow 0} \frac{f(A_\varepsilon) - f(A)}{\varepsilon}.$$

Then, using the fact that $\mathbb{P}(\{x \in \mathbb{R}^n : x_1 \leq 0\}) = 1/2$, the isoperimetric inequality tells us that for all small enough ε ,

$$\mathbb{P}(A_\varepsilon) \geq \mathbb{P}(\{x \in \mathbb{R}^n : x_1 \leq \varepsilon\}) = 1 - \Phi(\varepsilon) \geq 1 - \exp\left(-\frac{t^2}{2}\right).$$

For more on this, see chapter 3.2 of Wainwright's book and also see Chapter 7 of the book by Lugosi, Massart, and Boucheron.

8 Introduction to Empirical Process Theory

8.1 Convergence of CDFs and the Glivenko-Cantelli theorem

Let $(X_i)_{i \in [n]} \stackrel{\text{iid}}{\sim} X$. X has CDF $F(t)$, i.e.

$$F(t) = \mathbb{P}(X \leq t).$$

We can also define the **empirical CDF**

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \leq t\}}.$$

This is the CDF of the empirical distribution of the X_i .

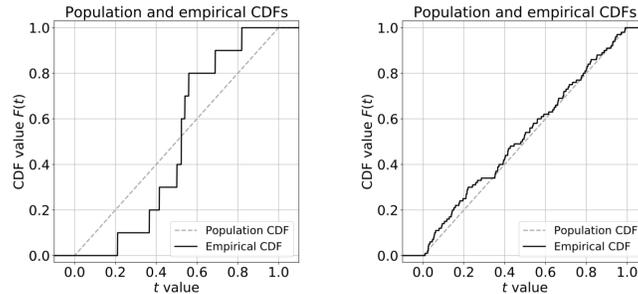
For any fixed t , the strong law of large numbers tells us that

$$\lim_{n \rightarrow \infty} \hat{F}_n(t) = F(t) \quad a.s.$$

If we are more ambitious, we may want convergence of functions. In this case, we look at the maximum difference,

$$\|F_n - F\|_\infty := \sup_{t \in [0,1]} |\hat{F}_n(t) - F(t)|.$$

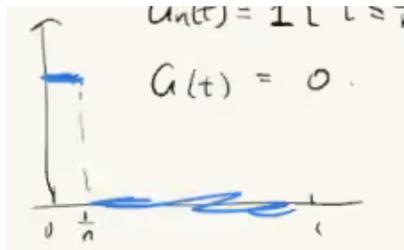
Here is a picture from Wainwright's book illustrating convergence of the empirical CDF to the uniform distribution on $[0, 1]$.



Why is convergence of the supremum norm stronger than pointwise convergence? In general,

$$\lim_{n \rightarrow \infty} G_n(t) = G(t) \forall t \not\Rightarrow \lim_{n \rightarrow \infty} \sup_t |G_n(t) - G(t)| = 0.$$

Example 8.1. Take $G_n(t) = \mathbb{1}_{\{t \leq 1/n\}}$.



Then for any $t > 0$, $G_n(t) \rightarrow 0$, but $\lim_{n \rightarrow \infty} \sup_t |G_n(t) - G(t)| = \infty$.

A classical result guarantees uniform convergence of the empirical CDF.

Theorem 8.1 (Glivenko-Cantelli, 1933). Let $X_i \stackrel{\text{iid}}{\sim} X$, where $F(t)$ is the CDF of X . Then

$$\lim_{n \rightarrow \infty} \|\hat{F}_n - F\|_\infty = 0 \quad \text{a.s.}$$

We will not prove this result. Instead, we will use empirical process theory, combined with concentration results to show something stronger:

$$\mathbb{P} \left(\|\hat{F}_n - F\|_\infty \geq 8\sqrt{\frac{\log(n+1)}{n}} + t \right) \leq \exp \left(-\frac{nt^2}{2} \right).$$

In other words,

$$\|\hat{F}_n - F\|_\infty \leq 8\sqrt{\frac{\log(n+1)}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}} \quad \text{with probability } 1 - \delta.$$

Why is this result stronger? If we let $n \rightarrow \infty$, we get convergence in probability. We can get a.s. convergence using the Borel-Cantelli lemma.

8.2 Uniform laws for more general function classes

Suppose $(X_i)_{i \in [n]} \stackrel{\text{iid}}{\sim} X \sim \mathbb{P}$, and suppose we have a **function class** $\mathcal{F} \subseteq \{f : \mathcal{X} \rightarrow \mathbb{R} : \mathbb{E}[|f(X)|] < \infty\}$.

Definition 8.1. The **empirical process** indexed by \mathcal{F} is

$$\left\{ \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)] \right) : f \in \mathcal{F} \right\}.$$

Define

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)] \right|.$$

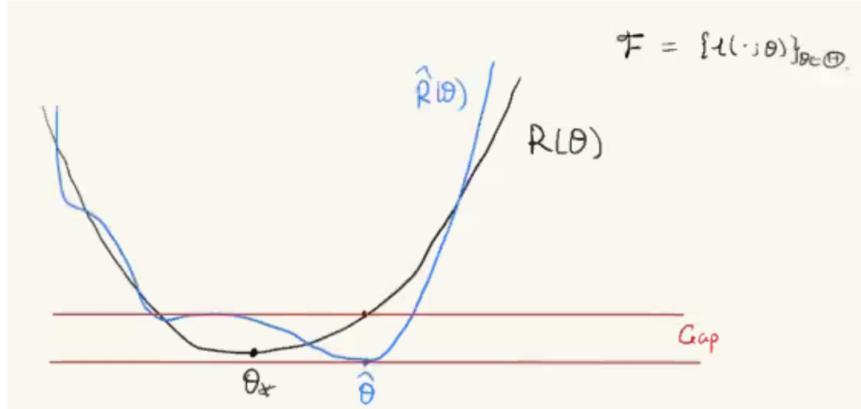
Here, $\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ is the **empirical measure**. This is the object we will study for the next portion of the course. If there is only 1 function f , we can deal with this using the law of large numbers and concentration inequalities. We will learn how to deal with this object using empirical process theory.

Why do we care about the maximum of empirical process in statistics and machine learning? Recall the following setup:

Data distribution	$(X_i)_{i \in [n]} \stackrel{\text{iid}}{\sim} \mathbb{P}$
Loss function	$L : \mathcal{X} \times \Theta \rightarrow \mathbb{R}$
Empirical risk	$\widehat{R}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(X_i; \theta)$
Population risk	$R(\theta) = \mathbb{E}_{X \sim \mathbb{P}}[\ell(X; \theta)]$
Empirical risk minimizer	$\widehat{\theta} = \arg \min_{\theta} \widehat{R}(\theta)$
Population risk minimizer	$\theta_* = \arg \min_{\theta} R(\theta)$
Excess risk	$E = R(\widehat{\theta}) - R(\theta_*)$

We train $\widehat{\theta}$ on the empirical risk, so we want the empirical risk to be close to the population risk. So to make sure training on our training data is accurate, we want to make the excess risk small. The excess risk has the following decomposition:

$$E = \underbrace{(R(\widehat{\theta}) - \widehat{R}_n(\widehat{\theta}))}_{\text{Gap}} + \underbrace{(\widehat{R}_n(\widehat{\theta}) - \widehat{R}_n(\theta_*))}_{\leq 0} + \underbrace{(\widehat{R}_n(\theta_*) - R(\theta_*))}_{\text{bound using Hoeffding}}$$



The Gap is

$$\text{Gap} = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\ell(X; \widehat{\theta}) - \ell(X_i; \widehat{\theta})].$$

We cannot use the strong law of large numbers to examine this because the $\ell(X_i; \widehat{\theta})$ are not independent random variables. We can fix this by replacing $\widehat{\theta}$ by the sup over θ :

$$\leq \sup_{\theta} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\ell(X; \theta) - \ell(X_i; \theta)] \right|.$$

Here, $f(X) = \ell(X; \theta)$, so we want to look at the function class $\mathcal{F} = \{\ell(\cdot; \theta) : \theta \in \Theta\}$.

Definition 8.2. We say that \mathcal{F} is a **Glivenko-Cantelli class** for \mathbb{P} if

$$\|P_n - \mathbb{P}\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)] \right| \xrightarrow{P} 0.$$

Example 8.2. The Glivenko-Cantelli theorem says that $\mathcal{F}_1 = \{\mathbb{1}_{\{x \leq t\}}\}_{t \in \mathbb{R}}$ is a Glivenko-Cantelli class for any $\mathbb{P} \in \mathcal{P}(\mathbb{R})$.

Example 8.3. Consider $\mathcal{F}_2 = \{\mathbb{1}_S : S \subseteq [0, 1] \text{ is a finite set}\}$, and assume that \mathbb{P} has density. This function class is *not* a Glivenko-Cantelli class. First note that $\mathcal{F}_1 \subseteq \mathcal{F}_2$, so if \mathcal{F}_2 is GC, then \mathcal{F}_1 is GC. So large function classes are less likely to be GC. To show that the function class is not GC, we can find a function in the function class which makes these two quantities different. Pick $S = \{X_i : i \in [n]\}$, so

$$\sup_{S \text{ finite}} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \in S\}} - \mathbb{E}[\mathbb{1}_{\{X_i \in S\}}] \right| \geq |1 - 0|.$$

This lower bound holds for every n , so this difference will never go to 0.

Our next goal is to study some methods for upper/lower bounding $\|P_n - \mathbb{P}\|_{\mathcal{F}}$. We will see

- Rademacher complexity and VC dimension (chapter 4 of Wainwright's book),
- Metric entropy method and chaining (chapter 5 of Wainwright's book).

8.3 Rademacher complexity

Recall that the Rademacher complexity of a set $A \subseteq \mathbb{R}^n$ is

$$\mathcal{R}(A) := \mathbb{E}_{\varepsilon \stackrel{\text{iid}}{\sim} \text{Unif}(\{\pm 1\})} \left\{ \sup_{a \in A} \langle a, \varepsilon \rangle \right\}$$

Definition 8.3. Given a function class \mathcal{F} and a fixed data set $(x_i)_{i \in [n]} \subseteq \mathcal{X}$, let

$$\mathcal{F}(x_{1:n}) := \{(f(x_1), \dots, f(x_n)) : f \in \mathcal{F}\} \subseteq \mathbb{R}^n.$$

The **Rademacher complexity** of the function class \mathcal{F} and the data set $(x_i)_{i \in [n]}$ is

$$\mathcal{R}(\mathcal{F}(x_{1:n})/n) := \mathbb{E}_{\varepsilon \stackrel{\text{iid}}{\sim} \text{Unif}(\{\pm 1\})} \left\{ \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right| \right\}.$$

If we write $\mathcal{A} = \pm \mathcal{F}(x_{1:n})/n$, then we can relate Rademacher complexity of sets and function classes by

$$\tilde{\mathcal{R}}(\mathcal{A}) = \mathcal{R}(\mathcal{F}(x_{1:n})/n),$$

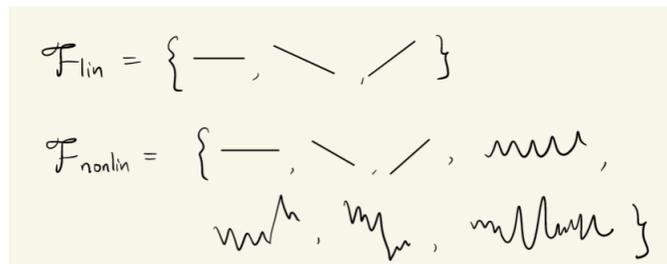
where $\tilde{\mathcal{R}}$ denotes the Rademacher complexity of a set.

Definition 8.4. Given a function class \mathcal{F} and a distribution $\mathbb{P} \in \mathcal{P}(\mathcal{X})$, let $(X_i)_{i \in [n]} \stackrel{\text{iid}}{\sim} \mathbb{P}$. The **Rademacher complexity** of the function class \mathcal{F} is

$$\mathcal{R}(\mathcal{F}) := \mathbb{E}_{X_i \stackrel{\text{iid}}{\sim} \mathbb{P}}[\mathcal{R}(\mathcal{F}(X_{1:n})/n)].$$

First, observe that if $\mathcal{F}_1 \subseteq \mathcal{F}_2$, then $\mathcal{R}_n(\mathcal{F}_1) \leq \mathcal{R}_n(\mathcal{F}_2)$, so this is a measure of the size of a function class.

Example 8.4. Consider comparing two function classes:



The notion of Rademacher complexity measures how well functions in the function class can align with Rademacher noise.



Here is the picture of what the comparison would look like:

$f_{x,lin} = \arg \max_{f \in \mathcal{F}_{lin}} \langle \varepsilon, f(x_{1:n}) \rangle / n,$	$f_{x,nonlin} = \arg \max_{f \in \mathcal{F}_{nonlin}} \langle \varepsilon, f(x_{1:n}) \rangle / n$
<p>align worse</p> <p>$\mathcal{R}_n(\mathcal{F}_{lin})$ smaller</p>	<p>align better</p> <p>$\mathcal{R}_n(\mathcal{F}_{nonlinear})$ larger</p>

Example 8.5. Let $\psi : \mathbb{R}^d \rightarrow \mathbb{R}^p$ be a fixed feature map, and consider the function class

$$\mathcal{F} = \{f(x) = \langle \psi(x), \theta \rangle : \|\theta\|_2 \leq B\}.$$

Then the Rademacher complexity of this function class is

$$\begin{aligned} \mathcal{R}_n(\mathcal{F}) &= \mathbb{E}_{X_i, \varepsilon_i} \left[\sup_{\|\theta\|_2 \leq B} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \langle \psi(X_i), \theta \rangle \right| \right] \\ &= \mathbb{E}_{X_i, \varepsilon_i} \left[\sup_{\|\theta\|_2 \leq B} \left| \varepsilon_i \left\langle \frac{1}{n} \sum_{i=1}^n \psi(X_i), \theta \right\rangle \right| \right] \\ &= \mathbb{E}_{X_i, \varepsilon_i} \left[\left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \psi(X_i) \right\|_2 \right] \cdot B \end{aligned}$$

Using Cauchy-Schwarz,

$$\begin{aligned} &\leq \mathbb{E}_{X_i, \varepsilon_i} \left[\left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \psi(X_i) \right\|_2^2 \right]^{1/2} \cdot B \\ &= \mathbb{E}_{X_i, \varepsilon_i} \left[\frac{1}{n^2} \sum_{i=1}^n \varepsilon_i^2 \|\psi(X_i)\|_2^2 \right]^{1/2} \cdot B \\ &= \frac{B}{\sqrt{n}} \mathbb{E}[\|\psi(X)\|_2^2]^{1/2}. \end{aligned}$$

Why introduce Rademacher complexity?

1. We will show that

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \approx \mathcal{R}_n(\mathcal{F}).$$

2. The Rademacher complexity is easier to upper bound. We will have tools to upper bound it, such as

- contraction inequality,
- VC dimension,
- fat-shattering dimension.

8.4 An upper bound of $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ via $\mathcal{R}_n(\mathcal{F})$

Proposition 8.1. For any function class \mathcal{F} and distribution \mathbb{P} ,

$$\mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}] \leq 2\mathcal{R}_n(\mathcal{F}).$$

Proof. Let $Y_i \stackrel{\text{iid}}{\sim} X_i$ be independent of X_i . Then

$$\begin{aligned} \mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}] &= \mathbb{E} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X_i)] \right| \right] \\ &= \mathbb{E}_{X_{1:n}} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}_{Y_{1:n}}[f(Y_i)] \right| \right] \\ &\leq \mathbb{E}_{X_{1:n}, Y_{1:n}} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n (f(X_i) - f(Y_i)) \right| \right] \end{aligned}$$

We can introduce a Rademacher random variable without changing the distribution.

$$\begin{aligned} &= \mathbb{E}_{X_{1:n}, Y_{1:n}, \varepsilon_{1:n}} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i (f(X_i) - f(Y_i)) \right| \right] \\ &\leq \mathbb{E}_{X_{1:n}, Y_{1:n}, \varepsilon_{1:n}} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right| - \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(Y_i) \right| \right] \\ &\leq 2\mathcal{R}_n(\mathcal{F}). \end{aligned} \quad \square$$

Next lecture, we will use a similar argument to show that if $\overline{\mathcal{F}} = \{f - \mathbb{E}[f] : f \in \mathcal{F}\}$, then

$$\mathcal{R}_n(\overline{\mathcal{F}}) \leq 2\mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}].$$

9 Bounds on Rademacher Complexity of Function Classes

9.1 Bounding $\mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}]$ in terms of Rademacher complexity

Last time, we were studying empirical processes defined by $X_i \stackrel{\text{iid}}{\sim} \mathbb{P} \in \mathcal{P}(\mathcal{X})$ and a function class $\mathcal{F} \subseteq \{f : \mathcal{X} \rightarrow \mathbb{R} : \mathbb{E}[|f(X)|] < \infty\}$. We want to bound the maximum of the empirical process,

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)] \right|.$$

We introduced the notion of Rademacher complexity for function classes: Given \mathcal{F} and $\{x_i\}_{i \in [n]}$, we let

$$\mathcal{R}(\mathcal{F}(x_{1:n})/n) = \mathbb{E}_{\varepsilon} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right| \right].$$

Then, given \mathcal{F} and \mathbb{P} ,

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{\varepsilon, X} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right| \right].$$

What is the relationship of Rademacher complexity and $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$? Define

$$\|\mathbb{S}_n\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right|.$$

Here is an upgraded version of what we showed last time.

Proposition 9.1. *For every convex, nondecreasing function $\Phi : \mathbb{R} \rightarrow \mathbb{R}$,*

$$\begin{aligned} \mathbb{E}_{X, \varepsilon}[\Phi(\tfrac{1}{2}\|\mathbb{S}_n\|_{\overline{\mathcal{F}}})] &\stackrel{\text{(a)}}{\leq} \mathbb{E}_X[\Phi(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}})] \\ &\stackrel{\text{(b)}}{\leq} \mathbb{E}_{X, \varepsilon}[\Phi(2\|\mathbb{S}_n\|_{\mathcal{F}})], \end{aligned}$$

where $\overline{\mathcal{F}} = \{f - \mathbb{E}[f] : f \in \mathcal{F}\}$.

Remark 9.1. Making $\Phi(t) = t$ retrieves the bound on $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ in terms of Rademacher complexity. We can also take the upper bound to also be $\overline{\mathcal{F}}$ because $\mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}] = \mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\overline{\mathcal{F}}}]$.

Proof. For (b),

$$\mathbb{E}_X[\Phi(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}})] = \mathbb{E}_X \left[\Phi \left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n (f(X_i) - \mathbb{E}[f(Y_i)]) \right| \right) \right]$$

Using Jensen's inequality,

$$\leq \mathbb{E}_{X,Y} \left[\Phi \left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n (f(X_i) - f(Y_i)) \right| \right) \right]$$

Since $f(X_i) - f(Y_i)$ has a symmetric distribution,

$$\begin{aligned} &= \mathbb{E}_{X,Y,\varepsilon} \left[\Phi \left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i (f(X_i) - f(Y_i)) \right| \right) \right] \\ &\leq \mathbb{E}_{X,Y,\varepsilon} \left[\Phi \left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right| + \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(Y_i) \right| \right) \right] \end{aligned}$$

Using Jensen's inequality again,

$$\begin{aligned} &\leq \frac{1}{2} \mathbb{E}_{X,\varepsilon} \left[\Phi \left(2 \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right| \right) \right] \\ &\quad + \frac{1}{2} \mathbb{E}_{Y,\varepsilon} \left[\Phi \left(2 \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(Y_i) \right| \right) \right] \\ &= \mathbb{E}_{X,\varepsilon} [\Phi(2\|S_n\|_{\mathcal{F}})]. \end{aligned}$$

For (a),

$$\mathbb{E}_{X,\varepsilon} [\Phi(\frac{1}{2}\|S_n\|_{\mathcal{F}})] = \mathbb{E}_{X,\varepsilon} \left[\Phi \left(\frac{1}{2} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i (f(X_i) - \mathbb{E}[f(Y_i)]) \right| \right) \right]$$

Using Jensen's inequality,

$$\leq \mathbb{E}_{X,Y,\varepsilon} \left[\Phi \left(\frac{1}{2} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i (f(X_i) - f(Y_i)) \right| \right) \right]$$

Since $f(X_i) - f(Y_i)$ has a symmetric distribution,

$$\begin{aligned} &= \mathbb{E}_{X,Y} \left[\Phi \left(\frac{1}{2} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n (f(X_i) - f(Y_i)) \right| \right) \right] \\ &= \mathbb{E}_{X,Y} \left[\Phi \left(\frac{1}{2} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n (f(X_i) - \mathbb{E}[f(X_i)]) - (f(Y_i) - \mathbb{E}[f(Y_i)]) \right| \right) \right] \\ &\leq \mathbb{E}_{X,Y} \left[\Phi \left(\frac{1}{2} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X_i)] \right| \right. \right. \\ &\quad \left. \left. + \frac{1}{2} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(Y_i) - \mathbb{E}[f(Y_i)] \right| \right) \right] \end{aligned}$$

Using Jensen's inequality again,

$$\begin{aligned}
&= \frac{1}{2} \mathbb{E}_X \left[\Phi \left(\frac{1}{2} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X_i)] \right| \right) \right] \\
&\quad + \frac{1}{2} \mathbb{E}_Y \left[\Phi \left(\frac{1}{2} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(Y_i) - \mathbb{E}[f(Y_i)] \right| \right) \right] \\
&= \mathbb{E}_X [\Phi(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}})]. \quad \square
\end{aligned}$$

Suppose that for all $f \in \mathcal{F}$, $\|f\|_{\infty} \leq b$. Then $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ is $(2b/n, \dots, 2b/n)$ -bounded difference. The bounded difference inequality then gives that $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ is $\text{sG}(b/\sqrt{n})$. In other words,

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} - \mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}] \leq b \sqrt{\frac{\log(2/\delta)}{n}} \quad \text{with probability } 1 - \delta.$$

This upper bound is typically smaller than $\mathcal{F}_n(\mathcal{F})$. This tells us that

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \begin{cases} \leq 2\mathcal{R}_n(\mathcal{F}) + b \sqrt{\frac{\log(2/\delta)}{n}} \\ \geq \frac{1}{2}\mathcal{R}_n(\overline{\mathcal{F}}) - b \sqrt{\frac{\log(2/\delta)}{n}}. \end{cases}$$

Note that

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = \|\mathbb{P}_n - \mathbb{P}\|_{\overline{\mathcal{F}}} \lesssim 2\mathcal{R}_n(\overline{\mathcal{F}}).$$

9.2 Aside: the maximal inequality

How do we upper bound the Rademacher complexity? Let's take a higher level picture and try to bound $\mathbb{E}[\sup_{\theta \in \Theta} X_{\theta}]$. In many cases, X_{θ} is sub-Gaussian for each fixed θ .

The simplest case is when Θ is finite. In this case, we have a **maximal inequality**: If for all $\theta \in \Theta$, $X_{\theta} \in \text{sG}(\sigma)$, then

$$\mathbb{E} \left[\max_{\theta \in \Theta} X_{\theta} \right] \leq \sigma \sqrt{2 \log |\Theta|}.$$

However, typically, this set Θ is infinite, so the maximal inequality cannot handle this case.

In the next lecture, we will discuss the metric entropy method, in which we approximate Θ by Θ_{ε} , where $|\Theta_{\varepsilon}| < \infty$ and

$$\sup_{\theta \in \Theta_{\varepsilon}} X_{\theta} \xrightarrow{\varepsilon \rightarrow 0} \sup_{\theta \in \Theta} X_{\theta}.$$

We will make this statement quantitative and precise. We will also introduce a different reduction, based on the concept of VC dimension.

9.3 Bounding Rademacher complexity using the maximal inequality

Use the special structure

$$\begin{aligned}\mathcal{R}_n(\mathcal{F}) &= \mathbb{E}_{X,\varepsilon} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{2} \sum_{i=1}^n \varepsilon_i f(X_i) \right| \right] \\ &= \mathbb{E}_X \left[\mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{2} \sum_{i=1}^n \varepsilon_i f(X_i) \right| \mid X_{1:n} \right] \right] \\ &= \mathbb{E}_X \left[\mathbb{E}_\varepsilon \left[\sup_{\nu \in \mathcal{F}(X_{1:n})} \left| \frac{1}{n} \langle \varepsilon, \nu \rangle \right| \mid X_{1:n} \right] \right]\end{aligned}$$

Bound the expectation by the supremum.

$$\leq \sup_{X_{1:n}} \mathbb{E}_\varepsilon \left[\sup_{\nu \in \mathcal{F}(X_{1:n})} \left| \frac{1}{n} \langle \varepsilon, \nu \rangle \right| \mid X_{1:n} \right]$$

If, for example, $\mathcal{F} \subseteq \{f : \mathcal{X} \rightarrow \{\pm 1\}\}$, then

$$\mathcal{F}(X_{1:n}) = \{(f(X_1), \dots, f(X_n)) : f \in \mathcal{F}\} \subseteq \{\pm 1\}^n.$$

Sometimes $|\mathcal{F}| = \infty$, but $|\mathcal{F}(X_{1:n})| < \infty$.

Example 9.1. Suppose $\mathcal{F} = \{\mathbb{1}_{\{X \leq t\}} : t \in \mathbb{R}\}$, so

$$\mathcal{F}(X_{1:n}) = \{(\mathbb{1}_{\{X_1 \leq t\}}, \mathbb{1}_{\{X_2 \leq t\}}, \dots, \mathbb{1}_{\{X_n \leq t\}}) : t \in \mathbb{R}\}.$$

Then if $X_1 < X_2 < \dots < X_n$,

$$\mathcal{F}(X_{1:n}) = \{(0, 0, \dots, 0), (1, 0, \dots, 0), (1, 1, 0, \dots, 0), \dots, (1, 1, \dots, 1)\},$$

so

$$\sup_{X_{1:n}} |\mathcal{F}(X_{1:n})| = n + 1.$$

Let's return to bounding

$$\mathbb{E}_\varepsilon \left[\sup_{\nu \in \mathcal{F}(X_{1:n})} \left| \frac{1}{n} \langle \varepsilon, \nu \rangle \right| \mid X_{1:n} \right].$$

We have that $\frac{1}{n} \langle \varepsilon, \nu \rangle = \frac{1}{n} \sum_{i=1}^n \varepsilon_i \nu_i$ is sG(σ_n), where

$$\sigma_n = \sup_{\nu \in \mathcal{F}(X_{1:n})} \frac{1}{n} \|\nu\|_2 = \sup_{f \in \mathcal{F}} \frac{1}{n} \sqrt{\sum_{i=1}^n f(X_i)^2}.$$

This tells us that the maximum of $|\mathcal{F}(X_{1:n})|$ is the number of mean 0 sG(σ_n) random variables. So the maximum inequality tells us that

$$\begin{aligned} \mathbb{E}_\varepsilon \left[\sup_{\nu \in \mathcal{F}(X_{1:n})} \left| \frac{1}{n} \langle \varepsilon, \nu \rangle \right| \mid X_{1:n} \right] &\leq \sigma_n \sqrt{2 \log(2|\mathcal{F}(X_{1:n})|)} \\ &\approx \underbrace{\sup_{f \in \mathcal{F}} \sqrt{\frac{\sum_{i=1}^n f(X_i)^2}{n}}}_{D_{\mathcal{F}}(X_{1:n})} \sqrt{\frac{2 \log(2|\mathcal{F}(X_{1:n})|)}{n}} \end{aligned}$$

Example 9.2. Let $\mathcal{F} = \{\mathbb{1}_{\{X \leq t\}} : t \in \mathbb{R}\}$ be the function class in the Glivenko-Cantelli theorem. Then

$$\begin{aligned} \sup_{X_{1:n}} |\mathcal{F}(X_{1:n})| &= n + 1, \\ \sup_{X_{1:n}} D_{\mathcal{F}}(X_{1:n}) &= \sup_{f \in \mathcal{F}} \sqrt{\frac{\sum_{i=1}^n 1^2}{n}} = 1. \end{aligned}$$

So we get

$$\mathcal{R}_n(\mathcal{F}) \leq \sqrt{\frac{2 \log(2(n+1))}{n}},$$

which bounds

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \lesssim 2 \sqrt{\frac{2 \log(2(n+1))}{n}} + \sqrt{\frac{\log(2/\delta)}{n}} \quad \text{with probability } 1 - \delta.$$

Remark 9.2. The above example gives a proof of the Glivenko-Cantelli theorem.

Remark 9.3. This $\log n$ factor is not sharp. Using other arguments, we will be able to show that the bound is actually of order $\sqrt{1/n}$. The issue here is that the maximal inequality is only sharp when the terms are independent. If X_i are sG(1), then

$$\sup_{i \in [n]} X_i = \begin{cases} O(\sqrt{\log n}) & \text{if the } X_i \text{ are independent} \\ X_1 = O(1) & \text{if } X_1 = X_2 = \dots = X_n. \end{cases}$$

Look at the bound

$$\Delta = \underbrace{D_{\mathcal{F}}(X_{1:n})}_{\text{typically } O(1)} \underbrace{\sqrt{\frac{2 \log(2|\mathcal{F}(X_{1:n})|)}{n}}}_{\text{want to vanish as } n \rightarrow \infty}.$$

Let's restrict our attention to $\mathcal{F} \subseteq \{f : \mathcal{X} \rightarrow \{\pm 1\}\}$. Here are two frequent behaviors of $|\mathcal{F}(X_{1:n})|$:

- (a) If $|\mathcal{F}(X_{1:n})| \lesssim O(n^\nu)$, then $\Delta = O(\sqrt{\frac{\nu \log n}{n}})$. This will go to 0 as $n \rightarrow \infty$, so this situation is good.
- (b) If $|\mathcal{F}(X_{1:n})| \lesssim O(\nu^n)$, then $\Delta = O(\sqrt{\frac{n \log \nu}{n}}) = O(\sqrt{\log \nu})$. This will not go to 0 as $n \rightarrow \infty$, so this situation is not good.

We want to be able to discriminate between these two cases. Since $\mathcal{F}(X_{1:n}) \subseteq \{\pm 1\}^n$, $|\mathcal{F}(X_{1:n})| \leq 2^n$. But when can we give a sharper upper bound?

Definition 9.1. \mathcal{F} has **polynomial discrimination** of order $\nu \geq 1$ if for all n and $X_{1:n}$,

$$|\mathcal{F}(X_{1:n})| \lesssim (n+1)^\nu.$$

Lemma 9.1. *Suppose \mathcal{F} has PD(ν). Then*

$$\mathcal{R}_n(\mathcal{F}) \leq 4 \left(\sup_{X_{1:n}} D_{\mathcal{F}}(X_{1:n}) \right) \sqrt{\frac{\nu \log(n+1)}{n}}.$$

Example 9.3. The function class $\{\mathbb{1}_{\{X \leq t\}} : t \in \mathbb{R}\}$ has PD(1), which implies the Glivenko-Cantelli theorem.

What kind of function classes have polynomial discrimination? Let $\psi : \mathcal{X} \rightarrow \mathbb{R}^d$.

Example 9.4. If $\mathcal{F} = \{\langle \psi(x), \theta \rangle + b : \theta \in \mathbb{R}^d, b \in \mathbb{R}\}$, then $|\mathcal{F}(X_{1:n})| = \infty$. So this does not have polynomial discrimination.

Example 9.5. If $\mathcal{F} = \{\mathbb{1}_{\{\langle \psi(x), \theta \rangle \geq b\}} : \theta \in \mathbb{R}^d, b \in \mathbb{R}\}$, then \mathcal{F} has PD($d+1$).

10 VC Dimension, Covering, and Packing

10.1 VC dimension

Last time we were discussing function classes with polynomial discrimination. Recall that a function class \mathcal{F} has $\text{PD}(\nu)$ if for all n and $X_{1:n}$, $|\mathcal{F}(X_{1:n})| \leq (n+1)^\nu$. If \mathcal{F} has $\text{PD}(\nu)$, then $\mathcal{R}_n(\mathcal{F}) \leq D\sqrt{\frac{\nu \log(n+1)}{n}}$. This gives the bound $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \lesssim D\sqrt{\frac{\nu \log(n+1)}{n}}$.

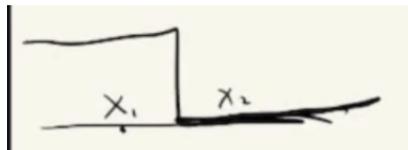
What function classes have polynomial discrimination? This question is answered by VC theory, named for Vapnik and Chervonenkis. If a function class has “VC dimension ν ,” then \mathcal{F} has $\text{PD}(\nu)$, which means that $\mathcal{R}_n(\mathcal{F}) \leq D\sqrt{\frac{\nu \log(n+1)}{n}}$.

Definition 10.1. Suppose $\mathcal{F} \subseteq \{F : \mathcal{X} \rightarrow \{0, 1\}\}$ is binary valued. We say that $x_{1:n}$ is **shattered** by \mathcal{F} if $|\mathcal{F}(x_{1:n})| = 2^n$. The **VC dimension**, $\nu(\mathcal{F})$, is the largest n such that there exists $x_{1:n}$ shattered by \mathcal{F} .

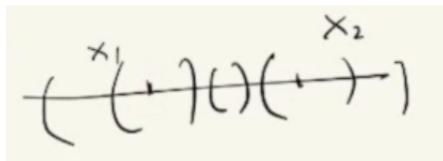
Note that $|\mathcal{F}(X_{1:n})| \leq 2^n$ always. So we want \mathcal{F} to be able to distinguish between points in a maximal sense.

Example 10.1. Let $\mathcal{F} = \{\mathbb{1}_{\{x \leq t\}} : t \in \mathbb{R}\}$. We claim that $\nu(\mathcal{F}) = 1$. Recall that $\mathcal{R}_n(\mathcal{F}) \leq 4\sqrt{\frac{\log(n+1)}{n}}$; this will also be implied by the VC-dimension. We have to show that there is some x_1 that is shattered by \mathcal{F} , and we have to show that no x_1, x_2 can be shattered by \mathcal{F} .

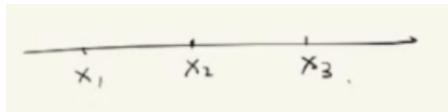
For $n = 1$, $\mathcal{F}(\{x_1\}) = \{0, 1\}$, so $\{x_1\}$ is shattered by \mathcal{F} . For $n = 2$, we want to show that $|\mathcal{F}(\{x_1, x_2\})| \leq 2^2 - 1$. If we assume, without loss of generality, that $x_2 > x_1$, this is because $\mathcal{F}(\{x_1, x_2\}) = \{(0, 0), (1, 1), (1, 0)\}$. Why does this not contain $(0, 1)$? This is because if one of these indicators gives 1 to x_2 , then it must give 1 to x_1 .



Example 10.2. Let $\mathcal{F} = \{\mathbb{1}_{\{s \leq x \leq t\}} : s < t \in \mathbb{R}\}$. We claim that $\nu(\mathcal{F}) = 2$. When $n = 2$, we want to find x_1, x_2 such that $|\mathcal{F}(\{x_1, x_2\})| = 2^2$. Here is how we can construct intervals to shatter a two point set:



Now suppose $x_1 < x_2 < x_3$. Then we cannot have $(1, 0, 1)$, since if an interval contains x_1, x_3 then it must contain x_2



Here is an example we will not prove.

Example 10.3. Let $\phi_1, \dots, \phi_p : \mathcal{X} \rightarrow \mathbb{R}$ be linear (which you can think of as feature maps), and consider $\mathcal{F} = \{\mathbb{1}_{\{\sum_{i=1}^p a_i \phi_i(x) \leq b\}} : a_i, b \in \mathbb{R}\}$. Then $\nu(\mathcal{F}) \leq p + 1$.

By definition, for all $n > \nu(\mathcal{F})$,

$$\sup_{x_{1:n}} |\mathcal{F}(x_{1:n})| \leq 2^n - 1.$$

Proposition 10.1 (Vapnik-Chervonenkis, Sauer-Shelah¹¹). For \mathcal{F} with VC dimension ν ,

$$\sup_{x_{1:n}} |\mathcal{F}(x_{1:n})| \leq \sum_{i=1}^{\nu} \binom{n}{i} \leq \min \left\{ (n+1)^\nu, \left(\frac{ne}{\nu} \right)^\nu \right\}.$$

By this proposition, we immediately have

$$\mathcal{R}_n(\mathcal{F}) \leq D \sqrt{\frac{\nu \log(n+1)}{n}}.$$

Here is an end-to-end result: If $\mathcal{F} = \{\mathbb{1}_{\{\sum_{i=1}^p a_i \phi_i(x) \leq b\}} : a_i, b \in \mathbb{R}\}$ and $(X_i)_{i \in [n]} \stackrel{\text{iid}}{\sim} \mathbb{P}$, then

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)] \right| \lesssim \sqrt{\frac{(p+1) \log n}{n}}.$$

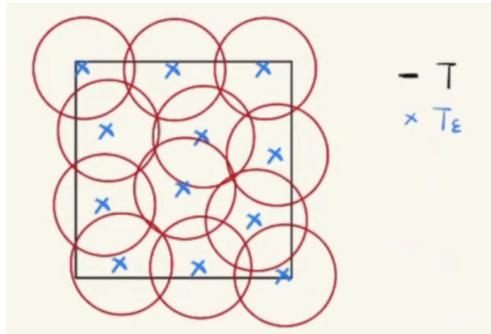
This $\log n$ factor can be eliminated later by the *chaining method*.

The proof of this proposition is a combinatorial argument; since the argument will not show up again, we will omit the proof, but you can look at the proof in the textbook.

¹¹This proposition was proven independently by Vapnik and Chervonenkis in 1971, by Sauer in 1972, and by Shelah by 1972.

10.2 The metric entropy method

Given a sub-Gaussian X_θ for all $\theta \in T$, we hope to upper bound $\mathbb{E}[\sup_{\theta \in T} X_\theta]$. How do we do this when $|T| = \infty$? The idea is to approximate T by a finite set T_ε as follows:



This gives

$$\mathbb{E} \left[\sup_{\theta \in T} X_\theta \right] \leq \mathbb{E} \left[\sup_{\tilde{\theta} \in T_\varepsilon} X_{\tilde{\theta}} \right] + \mathbb{E} \left[\sup_{\theta \in T, \tilde{\theta} \in T_\varepsilon} (X_\theta - X_{\tilde{\theta}}) \right].$$

We hope that

1. $|T_\varepsilon|$ is small.
2. $\mathbb{E}[\sup_{\theta \in T, \tilde{\theta} \in T_\varepsilon} (X_\theta - X_{\tilde{\theta}})]$ is small.

Given T and ρ , how can we find T_ε and bound $|T_\varepsilon|$?

10.3 Covering and packing

Definition 10.2. A **metric space** is a pair (T, ρ) , where $\rho : T \times T \rightarrow \mathbb{R}$ such that

1. $\rho(\theta, \theta') \geq 0$ for all $\theta, \theta' \in T$, with equality holding iff $\theta = \theta'$.
2. $\rho(\theta, \theta') = \rho(\theta', \theta)$.
3. $\rho(\theta, \theta') \leq \rho(\theta, \theta'') + \rho(\theta'', \theta')$.

Example 10.4. If $T = \mathbb{R}^d$, here are a few useful metrics:

$$\rho(\theta, \theta') = \|\theta - \theta'\|_2, \quad \rho(\theta, \theta') = \frac{1}{d} \sum_{i=1}^d \mathbb{1}_{\{\theta_i - \theta'_i\}}$$

The set T can be a function space, rather than a parameter space.

Example 10.5. Let $T = L^2(\mathcal{X}, \mu)$. Here are two metrics on T :

$$\rho(f, g) = \left(\int (f(x) - g(x))^2 d\mu(x) \right)^{1/2}, \quad \rho(f, g) = \|f - g\|_\infty.$$

Definition 10.3. $T_\varepsilon = \{\theta^1, \dots, \theta^N\}$ is an ε -**covering** of a set T if for all $\theta \in T$, there exists a $\theta^i \in T_\varepsilon$ such that $\rho(\theta, \theta^i) \leq \varepsilon$. The ε -**covering number** of T with respect to ρ is defined as

$$N(\varepsilon, T, \rho) := \inf\{N : |T_\varepsilon| = N, T_\varepsilon \text{ is an } \varepsilon\text{-covering of } T\}.$$

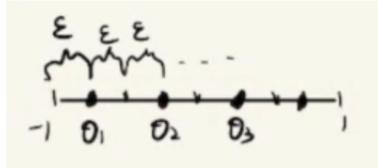
The maximal inequality gives

$$\mathbb{E} \left[\max_{\theta \in T_\varepsilon} X_\theta \right] \lesssim \sqrt{\log |T_\varepsilon|} \approx \sqrt{\log N(\varepsilon; T, \rho)}.$$

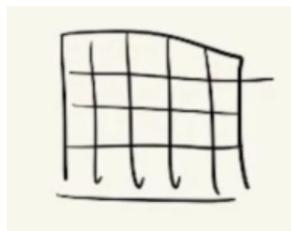
Definition 10.4. The function $\varepsilon \mapsto \log N(\varepsilon; T, \rho)$ for fixed (T, ρ) is called the **metric entropy of the set T** .

We will see examples that range from parametric families with $\log N(\varepsilon) \approx d \log(1 + 1/\varepsilon)$ to nonparametric families with $\log N(\varepsilon) \approx (1/\varepsilon)^\alpha$, where $\alpha \geq 0$.

Example 10.6. Let $T = [-1, 1]$ with $\rho(\theta, \theta') = |\theta - \theta'|$. Then $N(\varepsilon; T, \rho) \leq \frac{1}{\varepsilon} + 1$.



Example 10.7. If $T = [-1, 1]^d$ with $\rho(\theta, \theta') = \|\theta - \theta'\|_\infty$, then $N(\varepsilon; T, \rho) \leq (\frac{1}{\varepsilon} + 1)^d$.



Up to some constant, this bound is tight.

How about with other metrics? We may not be able to figure out a cover/packing. We can take a volume approach: We should expect

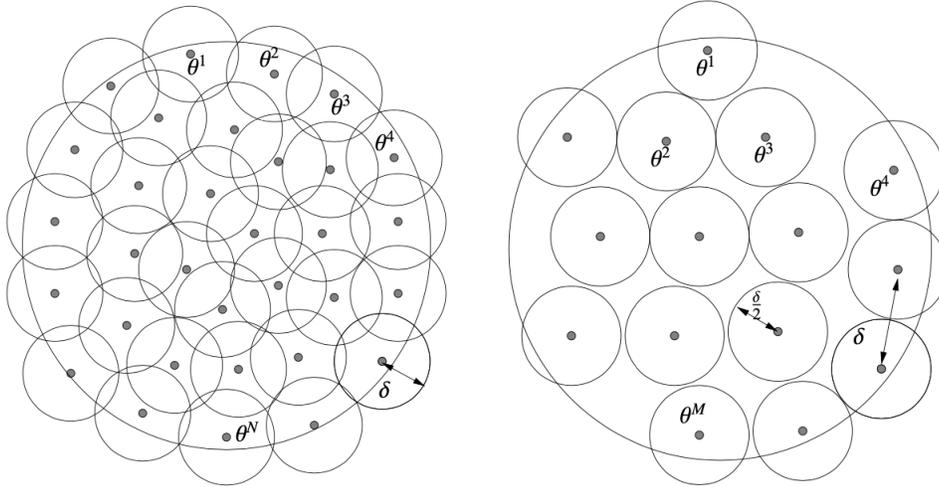
$$\log N(\varepsilon; T, \rho) \approx \log \left(\frac{\text{Vol}(T)}{\text{Vol}(B_\rho(\varepsilon))} \right).$$

To make this statement precise, we can introduce the idea of packing:

Definition 10.5. A set $\tilde{T}_\varepsilon = \{\theta^1, \dots, \theta^M\} \subseteq T$ is an ε -**packing** if for all $\theta^i, \theta^j \in \tilde{T}_\varepsilon$ with $i \neq j$, $\rho(\theta^i, \theta^j) > \varepsilon$. The ε -**packing number** is

$$M(\varepsilon; T, \rho) = \sup\{M : |\tilde{T}_\varepsilon| = M, \tilde{T}_\varepsilon \text{ is an } \varepsilon\text{-packing of } T\}.$$

This means that $B_\rho(\theta^i, \varepsilon/2) \cap B_\rho(\theta^j, \varepsilon/2) = \emptyset$. Here is a picture from Wainwright's textbook comparing packings and coverings:

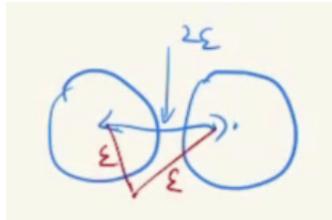


Lemma 10.1. For all $\varepsilon > 0$, we have

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

Proof. A maximal ε -packing gives an ε -covering. Suppose we have a maximal packing; then we cannot put another point into the packing, so the entire set T must be covered by the balls determined by the packing.

For a 2ε -packing with size M , all ε -coverings should have size at least M .



Otherwise, we would have a contradiction. □

11 Volume Bounds for Metric Entropy and the Chaining Method

11.1 Recap: one-step discretization bound

Last time, we began discussing the metric entropy method for obtaining bounds on empirical processes. We have a metric space (T, ρ) , and we want to control

$$\mathbb{E} \left[\sup_{\theta \in T} X_\theta \right] \quad \text{or} \quad \mathbb{E} \left[\sup_{\theta \in T} |X_\theta| \right],$$

where X_θ is usually mean 0 and sub-Gaussian. We introduced the metric entropy is $\log N(\varepsilon; T, \rho)$, where $N(\varepsilon; T, \rho) = \inf\{N : |T_\varepsilon| = N, T_\varepsilon \text{ is an } \varepsilon\text{-cover}\}$ is the ε -covering number.

Here is the one-step discretization bound that the maximal inequality gives us:

Lemma 11.1. *If $X_\theta \sim \text{sG}(\sigma)$ for all $\theta \in T$, then*

$$\begin{aligned} \mathbb{E} \left[\sup_{\theta \in T} |X_\theta| \right] &\lesssim \inf_{\varepsilon} \inf_{\varepsilon\text{-cover } T_\varepsilon} \mathbb{E} \left[\sup_{\theta \in T_\varepsilon} |X_\theta| \right] + \mathbb{E} \left[\sup_{\rho(\theta, \tilde{\theta}) \leq \varepsilon} |X_\theta - X_{\tilde{\theta}}| \right] \\ &\lesssim \inf_{\varepsilon} \sigma \sqrt{\log(N(\varepsilon; T, \rho))} + \mathbb{E} \left[\sup_{\rho(\theta, \tilde{\theta}) \leq \varepsilon} |X_\theta - X_{\tilde{\theta}}| \right] \end{aligned}$$

Today, we will mostly discuss the case where $T \subseteq \mathbb{R}^d$ is Euclidean space, and X_θ is some canonical random variable, such as $X_\theta = \langle \varepsilon, \theta \rangle$ or $X_\theta = \langle W, \theta \rangle$, which give the Radamacher and Gaussian complexities. We will give a volume-based method for bounding the covering number, give some examples, and then introduce the chaining method, which will give us a sharper bound.

In the next few lecture, we will extend this discussion to $T = \mathcal{F} \subseteq L^p(\mathbb{P})$ for $1 \leq p \leq \infty$, with $X_\theta = \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(Z_i)$ or $X_\theta = \frac{1}{n} \sum_{i=1}^n (f(Z_i) - \mathbb{E}[f(Z_i)])$. We will also relate this to and extend our VC theory.

11.2 Volume bounds for metric entropy

We want to understand the ε -covering number for $T \subseteq \mathbb{R}^d$. The intuition is that

$$\log N(\varepsilon; T, \rho) \asymp \log \frac{\text{Vol}(T)}{\text{Vol}(B_\rho(\varepsilon))},$$

so we can understand the covering number by understanding the volume. Here, we use the notation

$$B_\rho(\theta, \varepsilon) := \{\tilde{\theta} \in \mathbb{R}^d : \rho(\theta, \tilde{\theta}) \leq \varepsilon\}, \quad B_\rho(\varepsilon) := B_\rho(0, \varepsilon),$$

$$\text{Vol}(T) = \int \mathbb{1}_{\{x \in T\}} dx,$$

where dx is Lebesgue measure.

Last time, we introduced the notion of the ε -packing number

$$M(\varepsilon; T, \rho) = \sup\{M : |\tilde{T}_\varepsilon| = M, M \text{ is an } \varepsilon\text{-packing of } T\}.$$

This was related to the covering number by the following lemma.

Lemma 11.2. *For all $\varepsilon > 0$, we have*

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

Lemma 11.3.

$$\frac{\text{Vol}(T)}{\text{Vol}(B_\rho(\varepsilon))} \leq N(\varepsilon; T, \rho) \leq M(\varepsilon; T, \rho) \leq \frac{\text{Vol}(T + B_\rho(\varepsilon/2))}{\text{Vol}(B_\rho(\varepsilon/2))},$$

where $T + B_\rho(\varepsilon/2) = \{a + b : a \in T, b \in B_\rho(\varepsilon/2)\}$.

Proof. For the first inequality, let T_ε be an ε -covering, so $T \subseteq \bigcup_{\theta \in T_\varepsilon} B_\rho(\theta, \varepsilon)$. This tells us that

$$\begin{aligned} \text{Vol}(T) &\leq \text{Vol}\left(\bigcup_{\theta \in T_\varepsilon} B_\rho(\theta, \varepsilon)\right) \\ &\leq \sum_{\theta \in T_\varepsilon} \text{Vol}(B_\rho(\theta, \varepsilon)) \\ &\leq |T_\varepsilon| \text{Vol}(B_\rho(\varepsilon)). \end{aligned}$$

For the second inequality, let \tilde{T}_ε be a ε -packing, so the union of all the balls in the packing is contained in the set augmented by $\varepsilon/2$. That is, $\bigcup_{\theta \in \tilde{T}_\varepsilon} B(\theta, \varepsilon/2) \subseteq T + B_\rho(\varepsilon/2)$. This tells us that

$$\begin{aligned} \text{Vol}(T + B_\rho(\varepsilon/2)) &\geq \text{Vol}\left(\bigcup_{\theta \in \tilde{T}_\varepsilon} B(\theta, \varepsilon/2)\right) \\ &= |\tilde{T}_\varepsilon| \text{Vol}(B_\rho(\varepsilon)). \end{aligned}$$

Now take the sup over all packings. □

Example 11.1. Let $\rho = \|\cdot\|_p$ and $T = B_p(1) = \{x \in \mathbb{R}^d : \|x\|_p \leq 1\}$. Then

$$N(\varepsilon; T, \rho) \leq \frac{\text{Vol}(T + B(\varepsilon/2))}{\text{Vol}(B(\varepsilon/2))} = \frac{\text{Vol}(B_p(1 + \varepsilon/2))}{\text{Vol}(B_p(\varepsilon/2))}$$

Note that $\text{Vol}(B_\rho(r)) = c_{d,p} r^d$ for some constant $c_{d,p}$. We do not need to know the value of $c_{d,p}$ because we are looking at ratios of volumes. This gives

$$N(\varepsilon; T, \rho) \leq \frac{(1 + \varepsilon/2)^d}{(\varepsilon/2)^d} = \left(\frac{2}{\varepsilon} + 1\right)^d.$$

We also get the lower bound

$$N(\varepsilon; T, \rho) \geq \frac{\text{Vol}(B_p(1))}{\text{Vol}(B_p(\varepsilon))} = \frac{1^d}{\varepsilon^d} = \left(\frac{1}{\varepsilon}\right)^d.$$

So we get bounds on the metric entropy

$$d \log\left(\frac{1}{\varepsilon}\right) \leq \log N(\varepsilon; T, \rho) \leq d \log\left(\frac{2}{\varepsilon} + 1\right).$$

These bounds are of the same order. Note that the bounds do not depend on p because we are looking at the p -ball in the p -norm.

Example 11.2. Consider $W_i \stackrel{\text{iid}}{\sim} N(0, 1)$, so $\langle W, \theta \rangle \sim \text{sG}(\|\theta\|_2)$. Then we know that

$$\mathcal{G}(B_2(1)) = \mathbb{E} \left[\sup_{\theta \in B_2(1)} \langle W, \theta \rangle \right] = \mathbb{E}[\|W\|_2] \simeq \sqrt{d}.$$

Here is another way to get this computation:

$$\begin{aligned} \mathcal{G}(B_2(1)) &\leq C \left[\sup_{\theta \in B_2(1)} \underbrace{\|\theta\|_2}_{=1} \underbrace{\sqrt{\log N(\varepsilon; B_2(1), \|\cdot\|_2)}}_{\leq \sqrt{d \log(1+2/\varepsilon)}} + \mathbb{E}_W \left[\sup_{\|\theta - \theta'\|_2 \leq \varepsilon} |W_\theta - W_{\theta'}| \right] \right] \\ &\leq C \left[\sqrt{d \log(1+2/\varepsilon)} + \mathbb{E}_W \left[\sup_{\|\theta - \tilde{\theta}\|_2 \leq \varepsilon} \langle W, \theta - \tilde{\theta} \rangle \right] \right] \\ &= C \left[\sqrt{d \log(1+2/\varepsilon)} + \mathbb{E}_W \left[\sup_{\|r\|_2 \leq \varepsilon} \langle W, r \rangle \right] \right] \\ &= C \left[\sqrt{d \log(1+2/\varepsilon)} + \varepsilon \underbrace{\mathbb{E}_W \left[\sup_{\|\tilde{r}\|_2 \leq 1} \langle W, \tilde{r} \rangle \right]}_{\mathcal{G}(B_2(1))} \right]. \end{aligned}$$

This tells us that

$$\mathcal{G}(B_2(1)) \leq C \sqrt{d \log(1+2/\varepsilon)} + C\varepsilon \mathcal{G}(B_2(1)).$$

If we take $\varepsilon \leq \frac{1}{2C}$, then we get

$$\mathcal{G}(B_2(1)) \leq 2C \sqrt{d \log(1+4C)} \asymp \sqrt{d},$$

which is the same order as before.

11.3 The chaining method

We have been using the bound

$$\mathbb{E} \left[\sup_{\theta \in T} |X_\theta| \right] \lesssim \underbrace{\inf_{\varepsilon} \inf_{\varepsilon\text{-cover } T_\varepsilon} \mathbb{E} \left[\sup_{\theta \in T_\varepsilon} |X_\theta| \right]}_{\text{bdd by covering number}} + \underbrace{\mathbb{E} \left[\sup_{\rho(\theta, \tilde{\theta}) \leq \varepsilon} |X_\theta - X_{\tilde{\theta}}| \right]}_{\text{how to give tight control?}}$$

Controlling the right term can require ad-hoc arguments. The chaining method gives a way to bound this effectively.

Definition 11.1. $\{X_\theta\}_{\theta \in T}$ is a **sub-Gaussian process** with respect to ρ on T if

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

or, equivalently, $X_\theta - X_{\theta'}$ is $\text{sG}(\rho(\theta, \theta'))$.

Example 11.3. Let $T \subseteq \mathbb{R}^d$ with $\rho = \|\cdot\|_2$. Look at $X_\theta = \langle W, \theta \rangle$, where $W \sim N(0, I_d)$. To bound, the Gaussian complexity, we want to bound $\mathbb{E}[\sup_{\theta \in T} X_\theta]$. Then $X_\theta - X_{\theta'} = \langle W, \theta - \theta' \rangle \sim N(0, \|\theta - \theta'\|_2^2) \sim \text{sG}(\|\theta - \theta'\|_2)$.

Proposition 11.1. Let $\{X_\theta, \theta \in T\}$ be a mean 0 sub-Gaussian process with metric ρ . Then if $D = \sup_{\theta, \tilde{\theta} \in T}$,

$$\mathbb{E} \left[\sup_{\theta, \tilde{\theta}} (X_\theta - X_{\tilde{\theta}}) \right] \leq \inf_{\varepsilon \leq D} 2 \left[\sup_{\rho(r, r') \leq \varepsilon} (X_r - X_{r'}) \right] + 32 \underbrace{\int_{\varepsilon}^D \sqrt{\log N(u; T, \rho)} du}_{=: J(\varepsilon; D; T, \rho)}.$$

Here, $J(\varepsilon; D; T, \rho)$ is known as **Dudley's entropy integral**.

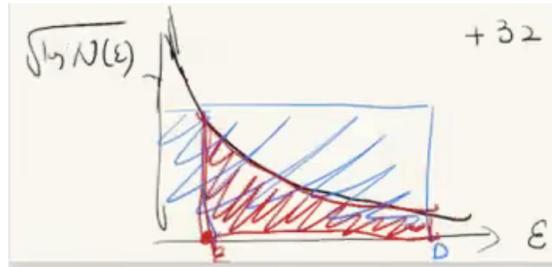
Remark 11.1. This gives an upper bound for $\mathbb{E}[\sup_{\theta \in T} X_\theta]$ because by the 0 mean condition and Jensen's inequality,

$$\begin{aligned} \mathbb{E} \left[\sup_{\theta \in T} X_\theta \right] &= \mathbb{E} \left[\sup_{\theta, \theta' \in T} (X_\theta - \mathbb{E}_{\theta'}[X_{\theta'}]) \right] \\ &\leq \mathbb{E} \left[\sup_{\theta, \tilde{\theta}} (X_\theta - X_{\tilde{\theta}}) \right]. \end{aligned}$$

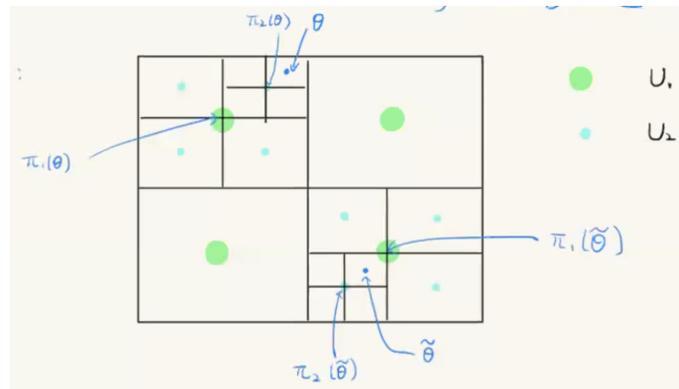
Remark 11.2. Compare this to the bound

$$\mathbb{E} \left[\sup_{\theta, \tilde{\theta}} (X_\theta - X_{\tilde{\theta}}) \right] \leq \inf_{\varepsilon \leq D} 2 \left[\sup_{\rho(r, r') \leq \varepsilon} (X_r - X_{r'}) \right] + 32D \sqrt{\log N(\varepsilon; T, \rho)}.$$

The integration gives a better bound because $\sqrt{\log N(\varepsilon)}$ is decreasing in ε .



Proof. Take a sequence of ε -coverings corresponding to $\varepsilon_m = D/2^m$ for $m = 0, 1, 2, 3, \dots, L$. Let U_m be the minimal ε_m -covering of T , so $|U_m| \leq N(\varepsilon_m; T_\rho)$. Then define the projection operation $\pi_m(\theta) = \arg \min_{\beta \in U_m} \rho(\theta, \beta)$.



This allows us to bound

$$|X_\theta - X_{\tilde{\theta}}| \leq |X_\theta - X_{\pi_2(\theta)}| + |X_{\pi_2(\theta)} - X_{\pi_1(\theta)}| + |X_{\pi_1(\theta)} - X_{\pi_1(\tilde{\theta})}| \\ + |X_{\pi_1(\tilde{\theta})} - X_{\pi_2(\tilde{\theta})}| + |X_{\pi_2(\tilde{\theta})} - X_{\tilde{\theta}}|.$$

Then we can take the expectation of $\sup_{\theta, \tilde{\theta}}$ on both sides. What is the purpose of having all these interpolation points? The first and the last terms have infinitely many choices, so these are the discretization terms, while the middle terms have only finitely many choices, so we can apply the maximal inequality.

$$\mathbb{E} \left[\sup_{\theta, \tilde{\theta} \in T} |X_\theta - X_{\tilde{\theta}}| \right] \leq \mathbb{E} \left[\sup_{\theta, \tilde{\theta} \in T} |X_{\pi_1(\theta)} - X_{\pi_1(\tilde{\theta})}| \right] + 2 \mathbb{E} \left[\sup_{\theta \in T} |X_{\pi_2(\theta)} - X_{\pi_1(\theta)}| \right] \\ + \dots + 2 \mathbb{E} \left[\sup_{\theta \in T} |X_{\pi_L(\theta)} - X_{\pi_{L-1}(\theta)}| \right] + 2 \mathbb{E} \left[\sup_{\theta \in T} |X_\theta - X_{\pi_L(\theta)}| \right]$$

These terms on the right correspond to $\varepsilon_0, \varepsilon_1, \dots, \varepsilon_{L-1}, \varepsilon_*$, respectively. This process will define a Riemann sum. For the remaining details, see the textbook. \square

Example 11.4. We want to bound the Gaussian complexity $\mathcal{G}(B_2(1)) = \mathbb{E}[\sup_{\theta \in B_2(1)} \langle W, \theta \rangle]$ using chaining. We get the bound

$$\begin{aligned}
 \mathcal{G}(B_2(1)) &\leq C \int_0^2 \sqrt{\underbrace{\log N(u; B_2(1), \|\cdot\|_2)}_{\leq d \log(2/u+1)}} du \\
 &\leq C \int_0^2 \sqrt{d \log(2/u+1)} du \\
 &= C\sqrt{d} \underbrace{\int_0^2 \sqrt{\log(2/u+1)} du}_{C'} \\
 &\asymp \sqrt{d}.
 \end{aligned}$$

12 The Metric Entropy Method for Function Spaces

12.1 Recap: controlling complexity via chaining

Last time, we were discussing the metric entropy method for obtaining bounds on empirical processes. We have a metric space (T, ρ) , and we want to control

$$\mathbb{E} \left[\sup_{\theta \in T} X_\theta \right] \quad \text{or} \quad \mathbb{E} \left[\sup_{\theta \in T} |X_\theta| \right],$$

where X_θ is usually mean 0 and sub-Gaussian. We introduced the metric entropy is $\log N(\varepsilon; T, \rho)$, where $N(\varepsilon; T, \rho) = \inf\{N : |T_\varepsilon| = N, T_\varepsilon \text{ is an } \varepsilon\text{-cover}\}$ is the ε -covering number.

We had the one step discretization bound given by the maximal inequality

$$\mathbb{E} \left[\sup_{\theta \in T} |X_\theta| \right] \lesssim \inf_{\varepsilon} \sigma \sqrt{\log(N(\varepsilon; T, \rho))} + \mathbb{E} \left[\sup_{\rho(\theta, \tilde{\theta}) \leq \varepsilon} |X_\theta - X_{\tilde{\theta}}| \right]$$

We introduced the condition of a process to be sG(ρ):

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

This condition allowed us to use the chaining bound

$$\mathbb{E} \left[\sup_{\theta \in T} |X_\theta| \right] \lesssim \inf_{\varepsilon} \sigma \int_{\varepsilon}^D \sqrt{\log N(u; T, \rho)} du + \mathbb{E} \left[\sup_{\rho(\theta, \tilde{\theta}) \leq \varepsilon} |X_\theta - X_{\tilde{\theta}}| \right].$$

Last time, we discussed examples where $T \subseteq \mathbb{R}^d$. We let $X_\theta = \langle \varepsilon, \theta \rangle$ or $X_\theta = \langle W, \theta \rangle$ to get bounds on the Rademacher/Gaussian complexity of Euclidean sets. Today, we will discuss examples where $T = \mathcal{F} \subseteq L^p$ for $1 \leq p \leq \infty$ is a function space. If we let

$$X_\theta = \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(Z_i) \quad \text{or} \quad X_\theta = \frac{1}{n} \sum_{i=1}^n (f(Z_i) - \mathbb{E}[f(Z_i)]),$$

then this gives us information about the Rademacher/Gaussian complexity of function spaces.

12.2 One step discretization and chaining bounds for Rademacher complexity of function classes

Recall that if $\{mcF \subseteq L^1(\mathbb{P})\}$ and $\varepsilon_i \stackrel{\text{iid}}{\sim} \text{Unif}(\{\pm 1\})$, then we defined the Rademacher complexity of function class as

$$\mathcal{R}_n(\mathcal{F}) := \mathbb{E}_{\varepsilon, X} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right| \right]$$

$$= \mathbb{E}_X[\mathcal{R}(\mathcal{F}(X_{1:n})/n)],$$

where we can think of this as the expectation of the empirical Rademacher complexity,

$$\mathcal{R}(\mathcal{F}(X_{1:n})/n) = \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right| \mid X_{1:n} \right],$$

where

$$\mathcal{F}(x_{1:n}) = (f(x_1), \dots, f(x_n)) : f \in \mathcal{F} \subseteq \mathbb{R}^n.$$

Recall that VC theory tells us that when the value of f is binary, $\mathcal{F}(x_{1:n})$ is a finite set. Then we can use the maximal inequality.

This lecture, we will control this using the metric entropy method. Rewrite

$$\mathcal{R}(\mathcal{F}(x_{1:n})/n) = \frac{1}{\sqrt{n}} \mathbb{E} \left[\sup_{f \in \mathcal{F}} |X_f| \right],$$

where

$$X_f := \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f(x_i).$$

Hoeffding's inequality tells us that $X_f \sim \text{sG}(\sqrt{\frac{1}{n} \sum_{i=1}^n f(x_i)^2})$.

To apply Dudley's entropy intergral bound on $\mathbb{E}[\sup_{\theta \in T} |X_\theta|]$, we need

1. A metric ρ on \mathcal{F} ,
2. X_f to be a sub-Gaussian process with respect to ρ ,
3. An upper bound for $N(u; \mathcal{F}, \rho)$,
4. (Optional) An upper bound for the discretization error.

12.3 Useful metrics on $\mathcal{F} \subseteq L^1(\mathbb{P})$

Here are four useful metrics

(a) $L^2(\mathbb{P})$ metric:

$$\|f - g\|_{L^2(\mathbb{P})}^2 = \int_{\mathcal{X}} (f(x) - g(x))^2 d\mathbb{P}(x).$$

(b) L^∞ metric: If $\text{supp } \mathbb{P} = \mathcal{X}$, then

$$\|f - g\|_{L^\infty} = \sup_{x \in \mathcal{X}} |f(x) - g(x)|.$$

(c) $L^2(\mathbb{P}_n)$ metric (given $x_{1:n}$):

$$\|f - g\|_{L^2(\mathbb{P}_n)}^2 = \int (f(x) - g(x))^2 d\mathbb{P}_n(x) = \frac{1}{n} \sum_{i=1}^n (f(x_i) - g(x_i))^2.$$

We can make this a random metric by using $X_{1:n}$.

This is equivalent to $\|\cdot\|_2$ on $\mathcal{F}(x_{1:n})/\sqrt{n} \subseteq \mathbb{R}^n$. Recall that

$$\mathcal{F}(x_{1:n}/\sqrt{n}) = \left\{ \frac{1}{\sqrt{n}}(f(x_1), \dots, f(x_n)) \in \mathbb{R}^n : f \in \mathcal{F} \right\}.$$

Then if $f(x_{1:n})/\sqrt{n}, g(x_{1:n})/\sqrt{n} \in \mathcal{F}(x_{1:n})/\sqrt{n}$,

$$\|f(x_{1:n})/\sqrt{n} - g(x_{1:n})/\sqrt{n}\|_2^2 = \frac{1}{n} \sum_{i=1}^n (f(x_i) - g(x_i))^2.$$

(d) Parametric metric: If $\mathcal{F} = \{f_\theta : \theta \in T \subseteq \mathbb{R}^d\}$, a metric ρ on T induces a metric ρ on \mathcal{F} by

$$\rho(f_\theta, f_{\tilde{\theta}}) := \rho(\theta, \tilde{\theta}).$$

Here are the relationships between these metrics:

- For any measure \mathbb{P} , $\|f - g\|_{\mathbb{P}} \leq \|f - g\|_{\infty}$. In particular, this says that $\|f - g\|_{\mathbb{P}_n} \leq \|f - g\|_{\infty}$ for all $x_{1:n}$.
- When $\mathcal{F} = \{f_\theta : \theta \in T \subseteq \mathbb{R}^d\}$, suppose that $|f_{\theta_1} - f_{\theta_2}(x)| \leq \Gamma(x)\rho(\theta_1, \theta_2)$. Then

$$\|f_{\theta_1} - f_{\theta_2}\|_{L^2(\mathbb{P})} \leq \|\Gamma\|_{L^2(\mathbb{P})}\rho(\theta_1, \theta_2),$$

$$\|f_{\theta_1} - f_{\theta_2}\|_{L^\infty} \leq \|\Gamma\|_{L^\infty}\rho(\theta_1, \theta_2).$$

Example 12.1. Let $\mathcal{F} = \{f_\theta(x) = 1 - e^{-\theta x}, x \in [0, 1] : \theta \in [0, 1]\}$. Then, using Taylor expansion and the intermediate value theorem,

$$|f_{\theta_1}(x) - f_{\theta_2}(x)| = \left| x e^{-\xi x} |\theta_1 - \theta_2| \right| \leq |x| \cdot |\theta_1 - \theta_2|.$$

This tells us that

$$\|f_{\theta_1} - f_{\theta_2}\|_{L^2(\mathbb{P})} \leq \|x\|_{L^2(\mathbb{P})} |\theta_1 - \theta_2|.$$

$$\|f_{\theta_1} - f_{\theta_2}\|_{L^\infty} \leq |\theta_1 - \theta_2|.$$

When x is not restricted to a bounded domain, we will not get a bound for the L^∞ norm

We care about inequalities between metrics because they introduce inequalities between covering numbers.

Lemma 12.1. *If ρ_1, ρ_2 are two metrics on T and $\rho_1(\theta_1, \theta_2) \leq \rho_2(\theta_1, \theta_2)$ for all $\theta_1, \theta_2 \in T$, then*

$$N(\varepsilon; T, \rho_1) \leq N(\varepsilon; T, \rho_2).$$

As a consequence,

$$N(\varepsilon; \mathcal{F}, L^2(\mathbb{P}_n)) \leq N(\varepsilon; \mathcal{F}, L^\infty), \quad N(\varepsilon; \mathcal{F}, L^2(\mathbb{P})) \leq N(\varepsilon; \mathcal{F}, L^\infty).$$

If $|f_{\theta_1}(x) - f_{\theta_2}(x)| \leq \Gamma(x)\rho(\theta_1, \theta_2)$, then

$$N(\varepsilon; \mathcal{F}, L^\infty) \leq N(\varepsilon; T, \|\Gamma\|_\infty \rho), \quad N(\varepsilon; \mathcal{F}, L^\infty) \leq N(\varepsilon; T, \|\Gamma\|_{L^2(\mathbb{P})} \rho).$$

Note that we can express this rescaling either in the metric or as a scaling factor in front of ε .

12.4 The uniform entropy bound for empirical processes

In what metrics might $X_f = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f(X_i)$ be a sub-Gaussian process?

$$\begin{aligned} \mathbb{E}[e^{\lambda(X_f - X_g)} \mid X_{1:n}] &= \mathbb{E}[e^{(\lambda/\sqrt{n}) \sum_{i=1}^n \varepsilon_i (f(X_i) - g(X_i))} \mid X_{1:n}] \\ &= \prod_{i=1}^n \mathbb{E}[e^{(\lambda/\sqrt{n}) \varepsilon_i (f(X_i) - g(X_i))} \mid X_i] \\ &\leq \prod_{i=1}^n e^{(\lambda^2/n)(f(X_i) - g(X_i))^2/2} \\ &= e^{(\lambda^2/2) \frac{1}{n} \sum_{i=1}^n (f(X_i) - g(X_i))^2/2} \end{aligned}$$

$$\begin{aligned} \text{Since } \frac{1}{n} \sum_{i=1}^n (f(X_i) - g(X_i))^2/2 - \|f - g\|_{\mathbb{P}_n} &\leq \|f - g\|_\infty, \\ &\leq e^{(\lambda^2/2) \|f - g\|_\infty}. \end{aligned}$$

This tells us that $(X_f)_{f \in \mathcal{F}}$ is a sub-Gaussian process with respect to the metric $\|\cdot\|_{L^2(\mathbb{P}_n)}$. The inequalities between metrics tell us that this is also then sub-Gaussian with respect to $\|\cdot\|_{L^\infty}$.

Now, if $D = \sup_{f, g \in \mathcal{F}} \|f - g\|_{L^2(\mathbb{P}_n)} =: \|\mathcal{F}\|_{\mathbb{P}_n}$ is the diameter,

$$\mathbb{E} \left[\sup_{f \in \mathcal{F}} |X_f| \right] \leq \int_0^D \sqrt{\log N(u; \mathcal{F}, L^2(\mathbb{P}_n))} du.$$

Then the empirical Rademacher complexity is bounded above by

$$\mathcal{R}(\mathcal{F}(X_{1:n})/n) = \frac{1}{\sqrt{n}} \mathbb{E} \left[\sup_{f \in \mathcal{F}} |X_f| \right]$$

Using the change of variables $u = \|\mathcal{F}\|_{\mathbb{P}_n} \tilde{u}$,

$$\begin{aligned} &\lesssim \frac{1}{\sqrt{n}} \int_0^{\|\mathcal{F}\|_{\mathbb{P}_n}} \sqrt{\log N(\|\mathcal{F}\|_{\mathbb{P}_n} \tilde{u}; \mathcal{F}, L^2(\mathbb{P}_n))} d\|\mathcal{F}\|_{\mathbb{P}_n} \tilde{u} \\ &= \frac{\|\mathcal{F}\|_{\mathbb{P}_n}}{\sqrt{n}} \int_0^1 \sqrt{\log N(\|\mathcal{F}\|_{\mathbb{P}_n} u; \mathcal{F}, L^2(\mathbb{P}_n))} du \\ &\leq \frac{\|\mathcal{F}\|_{\mathbb{P}_n}}{\sqrt{n}} \int_0^1 \sup_Q \sqrt{\log N(\|\mathcal{F}\|_Q u; \mathcal{F}, L^2(Q))} du. \end{aligned}$$

When we take the expectation of the empirical Rademacher complexity and use Cauchy-Schwarz, we get

$$\mathbb{E}[\mathcal{F}(\mathcal{F}(X_{1:n})/\sqrt{n})] \leq \frac{\|\mathcal{F}\|_{\mathbb{P}}}{\sqrt{n}} \int_0^1 \sup_X \sqrt{\log N(\|\mathcal{F}\|_Q u; \mathcal{F}, L^2(Q))} du.$$

We can summarize this in the following proposition:

Proposition 12.1 (Uniform entropy bound).

$$\mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}] \lesssim \mathcal{R}_n(\mathcal{F}) \lesssim \frac{\|\mathcal{F}\|_{\mathcal{P}}}{\sqrt{n}} \int_0^1 \sup_Q \sqrt{\log N(\|\mathcal{F}\|_Q u; \mathcal{F}, L^2(Q))} du.$$

This is not in Wainwright's textbook, but you can find it as Theorem 4.7 in *A Gentle Introduction to Empirical Process Theory and Applications* by Bodhisattva Sen.

12.5 Examples of bounding Rademacher complexity for different covering numbers

Example 12.2. Suppose we have $\log N(u) \asymp d \log(1 + 1/u)$. Then

$$\mathcal{R}_n(\mathcal{F}) \leq \frac{1}{\sqrt{n}} \int_0^1 \sqrt{d \log(1 + 1/u)} du \lesssim \sqrt{\frac{d}{n}}.$$

Example 12.3. If $\log N(u) \asymp 1/u$, then

$$\mathcal{R}_n(\mathcal{F}) \lesssim \frac{1}{\sqrt{n}} \int_0^1 \sqrt{\frac{1}{u}} du \lesssim \frac{1}{\sqrt{n}}.$$

Example 12.4. If $\log N(u) \asymp \frac{1}{u^d}$, where $d \geq 2$, then

$$\mathcal{R}_n(\mathcal{F}) \lesssim \frac{1}{\sqrt{n}} \int_0^1 \sqrt{\frac{1}{u^d}} du = \infty.$$

However, we can get a better bound in the last example by using the following proposition.

Proposition 12.2.

$$\sup_{\mathbb{P}} \mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}] \lesssim \mathcal{R}_n(\mathcal{F}) \lesssim \|\mathcal{F}\|_{\infty} \inf_{\varepsilon} \varepsilon + \frac{1}{\sqrt{n}} \int_{\varepsilon}^1 \sqrt{\log N(\|\mathcal{F}\|_{\infty} u; \mathcal{F}, L^{\infty})} du.$$

How can we upper bound $\mathbb{E}_{\varepsilon_i}[\sup_{\|f-g\|_{\mathbb{P}_n} \leq \varepsilon} |\sum_{i=1}^n \varepsilon_i (f(X_i) - g(X_i))|]$? We know that we can bound

$$\mathbb{E}_{\varepsilon_i} \left[\sup_{\|f-g\|_{L^{\infty}} \leq \varepsilon} \sum_{i=1}^n \varepsilon_i (f(X_i) - g(X_i)) \right] \leq \sqrt{n} \varepsilon.$$

If we use this bound, then when $\log N(u) \lesssim \frac{1}{u^d}$ with $d \geq 2$, we get

$$\mathcal{R}_n(\mathcal{F}) \lesssim \inf_{\varepsilon} \varepsilon + \frac{1}{\sqrt{n}} \int_{\varepsilon}^1 \frac{1}{u^{d/2}} du.$$

13 Examples of Rademacher Complexity Bounds for Function Classes

13.1 Recap: chaining bounds for Rademacher complexity of function classes

Last time, we were using the metric entropy method to bound the Rademacher complexity of a function class. We considered 4 metrics on \mathcal{F} :

$$\|\cdot\|_{L^2(\mathbb{P})}, \quad \|\cdot\|_{L^\infty}, \quad \|\cdot\|_{L^2(\mathbb{P}_n)}, \quad \rho \text{ on parameter space.}$$

Relationships of these metrics gave us relationships between the covering numbers:

$$\begin{aligned} N(\varepsilon; \mathcal{F}, \|\cdot\|_{L^2(\mathbb{P}_n)}) &\leq \sup_{\mathbb{P}} N(\varepsilon; \mathcal{F}, \|\cdot\|_{L^2(\mathbb{P})}) \\ &\leq N(\varepsilon; \mathcal{F}, \|\cdot\|_{L^\infty}) \end{aligned}$$

And if the function class \mathcal{F} is a Lipschitz parametrization,

$$\leq N(\varepsilon; T, \rho)$$

If we let $X_f = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f(x_i)$, then we can show that

$$\mathbb{E}[e^{\lambda(X_f - X_g)}] \leq e^{(\lambda^2/2)\|f-g\|_{\mathbb{P}_n}^2} \leq e^{(\lambda^2/2)\|f-g\|_\infty^2},$$

which tells us that $\{X_f\}_{f \in \mathcal{F}}$ is a sub-Gaussian process with respect to the $L^2(\mathbb{P}_n)$ or L^∞ metric.

We showed two results:

Proposition 13.1. *Let $\mathcal{R}_n(\mathcal{F}) := \mathbb{E}_{\varepsilon_i, X_i} [\sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i)|]$. Then*

1.

$$\mathcal{R}_n(\mathcal{F}) \lesssim \frac{D_{\mathbb{P}}}{\sqrt{n}} \int_0^1 \sup_Q \sqrt{\log N(D_Q u; \mathcal{F}, L^2(Q))} du,$$

2.

$$\mathcal{R}_n(\mathcal{F}) \lesssim \frac{D_\infty}{\sqrt{n}} \inf_\varepsilon \varepsilon + \frac{1}{\sqrt{n}} \int_\varepsilon^1 \sup_Q \sqrt{\log N(D_\infty u; \mathcal{F}, L^\infty)} du,$$

where $D_{\mathbb{P}} = \sup_{f \in \mathcal{F}} \|f\|_{L^2(\mathbb{P})}$ and $D_\infty = \sup_{f \in \mathcal{F}} \|f\|_\infty$.

13.2 Examples of upper bounds for parametric and nonparametric function classes

Here are some examples for upper bounds of Rademacher complexity for function classes.

Example 13.1. Let $\mathcal{F} = \{f_\theta(x) = 1 - e^{-\theta x}, x \in [0, 1] : \theta \in [0, 1]\}$ be a parametric function class. Then taking the derivative gives us

$$|f_{\theta_1}(x) - f_{\theta_2}(x)| \leq \sup_{\theta \in [\theta_1, \theta_2]} \underbrace{|xe^{-\theta x}|}_{\leq x} |\theta_1 - \theta_2| \leq |\theta_1 - \theta_2|$$

The covering number for the unit interval with $|\cdot|$ is bounded as

$$N(\varepsilon; [0, 1], |\cdot|) \leq \frac{1}{2\varepsilon} + 1,$$

so we get a covering number bound for the parametric function class

$$N(\varepsilon; \mathcal{F}, L^\infty) \leq N(\varepsilon; [0, 1], |\cdot|) \leq \frac{1}{2\varepsilon} + 1.$$

Using the chaining bound with $D_\infty = \sup_{f \in \mathcal{F}} \|f\|_\infty = \sup_{f \in \mathcal{F}} \sup_{x \in [0, 1]} |1 - e^{-\theta x}| \leq 1$,

$$\begin{aligned} \mathcal{R}_n(\mathcal{F}) &\leq \frac{D_\infty}{\sqrt{n}} \int_0^1 \sqrt{\log N(uD_\infty; \mathcal{F}, L^\infty)} du = \frac{1}{\sqrt{n}} \int_0^1 \sqrt{\log N(u; \mathcal{F}, L^\infty)} du \\ &= \frac{1}{\sqrt{n}} \int_0^1 \sqrt{\log(\frac{1}{2u} + 1)} du \\ &\lesssim \frac{C}{\sqrt{n}}. \end{aligned}$$

Example 13.2 (Lipschitz parameterization). Consider a function class $\mathcal{F} = \{f_\theta : \mathcal{X} \rightarrow \mathbb{R} : \theta \in B_2^d(1) \text{ with } \|f_0(x)\|_\infty = c_0 = 0. \text{ If } |f_{\theta_1}(x) - f_{\theta_2}(x)| \leq L\|\theta_1 - \theta_2\|_2, \text{ then we can use the bound}$

$$\log N(\varepsilon; \mathcal{F}, L^\infty) \leq \log N(\varepsilon; B_2^d(1), \|\cdot\|_2) \lesssim d \log \lesssim d \log \left(\frac{1}{\varepsilon} + 1 \right)$$

to get

$$\mathcal{R}_n(\mathcal{F}) \lesssim L \frac{D_\infty}{\sqrt{n}} \int_0^1 \sqrt{\log N(\varepsilon; \mathcal{F}, L^\infty)} d\varepsilon,$$

where $D_\infty = \sup_\theta \|f_\theta\|_\infty \leq c_0 + L = L$

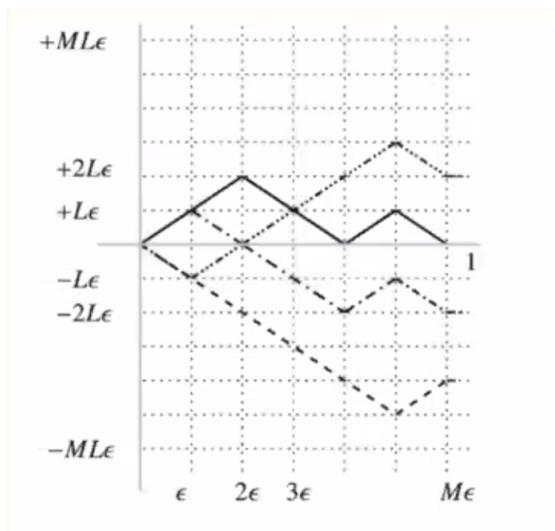
$$\begin{aligned} &\lesssim \frac{L}{\sqrt{n}} \int_0^1 \frac{1}{\sqrt{n}} \int_0^1 \sqrt{d \log(1/n)} du \\ &\lesssim L \sqrt{\frac{d}{n}}. \end{aligned}$$

If we have a nonparametric function class, it may have infinite Rademacher complexity. So in general, we will want some sort of smoothness condition to make the complexity finite.

Example 13.3 (Nonparametric class with smoothness/convexity). Consider the nonparametric function class $\mathcal{F}_L = \{g : [0, 1] \rightarrow \mathbb{R} \mid g(0) = 0, g \text{ is } L\text{-Lip}\}$. Then

$$\log N(\varepsilon; \mathcal{F}_L, L^\infty) \asymp \frac{L}{\varepsilon},$$

which we can see from the following figure in Wainwright's book that shows how to bound the packing number:



In particular, if $f \neq g$, then

$$M(L\varepsilon; \mathcal{F}, L^\infty) \geq 2^{1/\varepsilon}.$$

Taking log and rescaling ε , we get

$$\log M(L\varepsilon; \mathcal{F}, L^\infty) \geq 2^{1/\varepsilon} \geq \frac{2L}{\varepsilon} \log 2.$$

On the other hand, we can get an upper bound by seeing that these functions cover the function class.

Here, we have $\|\mathcal{F}\|_\infty = \sup_{f \in \mathcal{F}} |f| = L$, so the one-step discretization bound gives

$$\begin{aligned} \mathcal{R}_n(\mathcal{F}_L) &\lesssim \inf_{\varepsilon} \varepsilon + \frac{1}{\sqrt{n}} \sqrt{\log N(\varepsilon; \mathcal{F}, \|\cdot\|_\infty)} \\ &= \inf_{\varepsilon} \varepsilon + \frac{1}{\sqrt{n\varepsilon}} \end{aligned}$$

$$\asymp \frac{1}{n^{1/3}}$$

The chaining bound gives

$$\begin{aligned} \mathcal{R}_n(\mathcal{F}_L) &\lesssim \inf_{\varepsilon} \varepsilon + \frac{1}{\sqrt{n}} \int_{\varepsilon}^1 \sqrt{\frac{1}{u}} du \\ &\asymp \frac{1}{\sqrt{n}}. \end{aligned}$$

So in this case, the one-step discretization bound gives a sharper bound than the chaining method.

Example 13.4 (Nonparametric class, general d). Consider a nonparametric function class with general d :

$$\mathcal{F}_L^d = \{g : [0, 1]^d \rightarrow \mathbb{R} : g(0) = 0, g \text{ is } L\text{-Lip in } \|\cdot\|_{\infty}\}.$$

We can show that

$$\log N(\varepsilon; \mathcal{F}_L^d, L^{\infty}) \asymp \left(\frac{L}{\varepsilon}\right)^d.$$

The calculation of the resulting bounds on the Rademacher complexity is left for homework.

13.3 Boolean function classes

Consider a Boolean function class $\mathcal{F} \subseteq \{f : \mathcal{X} \rightarrow \{0, 1\}\}$, VC theory tells us that \mathcal{F} has $\text{PD}(\nu)$, where $\nu = \text{VC}(\mathcal{F})$. Using the maximal inequality, we have the bound

$$\mathcal{R}_n(\mathcal{F}) \lesssim \sqrt{\frac{\nu \log(n+1)}{n}}.$$

We have mentioned that the log factor in this bound makes the bound not tight.

Proposition 13.2. *For a boolean function class with $\nu = \text{VC}(\mathcal{F})$,*

$$\sup_{\mathbb{P}} \log(N(\varepsilon; \mathcal{F}, L^2(\mathbb{P}))) \lesssim \nu \log\left(\frac{e}{\varepsilon}\right)$$

for $\varepsilon < 1$.

For a sharp but difficult proof of this bound, see theorem 2.6.4 from [Van der Vaart and Wellner, 1996]. A weaker but easier version of this bound can be found in the notes [Sen, Theorem 7.9].

If we use the chaining argument, we get the bound

$$\mathcal{F}_n(\mathcal{F}) \lesssim \frac{1}{\sqrt{n}} \int_0^1 \sqrt{\nu \log(e/\varepsilon)} d\varepsilon \propto \sqrt{\frac{\nu}{n}}.$$

Example 13.5. Specialize to the function class $\mathcal{F} = \{\mathbb{1}_{x \leq t} : t \in \mathbb{R}\}$, which we first examined when looking at empirical processes. This has VC-dimension 1, so

$$\mathcal{R}_n(\mathcal{F}) \lesssim \sqrt{\frac{1}{n}}.$$

This tells us that

$$\mathbb{P} \left(\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \geq \frac{c}{\sqrt{n}} + \frac{\varepsilon}{\sqrt{n}} \right) \leq 2e^{-\varepsilon^2/2}.$$

Remark 13.1. This is not the tightest version of this bound. The tightest bound, given by Dvoretzky, Kiefer, Wolfowitz, and Massart, is

$$\mathbb{P} \left(\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \geq \frac{\varepsilon}{\sqrt{n}} \right) \leq 2e^{-\varepsilon^2/2}.$$

13.4 Contraction inequalities

Consider d functions $\phi_j : \mathbb{R} \rightarrow \mathbb{R}$ which are L -Lipschitz with $\phi_j(0) = 0$. We can think of $\phi_j(\theta)$ as a loss function $L(y; \theta)$.

Proposition 13.3 (Talagrand-Ledoux concentration). *Let $T \subseteq \mathbb{R}^d$, and let $\{\phi_j\}$ be centered Lipschitz. Then*

$$\mathbb{E} \left[\sup_{\theta \in T} \sum_{j=1}^d \varepsilon_j \phi_j(\theta_j) \right] \leq L \mathbb{E} \left[\sup_{\theta \in T} \sum_{j=1}^d \varepsilon_j \theta_j \right],$$

$$\mathbb{E} \left[\sup_{\theta \in T} \left| \sum_{j=1}^d \varepsilon_j \phi_j(\theta_j) \right| \right] \leq 2L \mathbb{E} \left[\sup_{\theta \in T} \left| \sum_{j=1}^d \varepsilon_j \theta_j \right| \right].$$

The interpretation is that the right hand side is $\mathcal{R}(T)$. The left hand side is $\mathcal{R}(\phi(T))$. This says that if we apply a contraction map to a space, the Rademacher complexity will not increase.

The textbook has a proof for when ε_i are iid Gaussian random variables. This is given by the Gaussian comparison inequality.

Example 13.6. Let $Z_i = (X_i, Y_i) \stackrel{\text{iid}}{\sim} \mathbb{P} \in \mathcal{P}(B_2(M) \times \{\pm 1\})$ for $i \in [n]$. For logistic regression, we want a logistic loss function:

$$M_\theta(Z) := \log(1 + \exp(-y\theta^\top x)).$$

Taking the expectation gives

$$M(\theta) = \mathbb{E}_Z[m_\theta(Z)].$$

We also let $\Theta = B_2(r)$. Compare the empirical and population risk:

$$\begin{aligned} E &:= \mathbb{E} \left[\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^n m_\theta(Z_i) - M(\theta) \right| \right] \\ &\leq 2 \mathbb{E} \left[\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i m_\theta(Z_i) \right| \right] \end{aligned}$$

We are looking at the function class $\mathcal{F} = \{m_\theta(z_i) : \theta \in \Theta\}$. If we want to replace $m_\theta(z_i)$ by $\theta^\top x_i$, then we can use the contraction inequality. This is because $\log(1 + e^x)$ is 1-Lipschitz (by $\frac{d}{dx} \log(1 + e^x) = \frac{e^x}{1+e^x} \leq 1$). So we can write $\phi_i(\tilde{\theta}_i) = \log(1 + \exp(-y_i \tilde{\theta}_i)) - \log 2$. This depends on Y_i , and $\tilde{\theta}_i = \theta^\top X_i$ depends on X_i , to use the contraction inequality, we first condition on Y and X :

$$\begin{aligned} &= 2 \mathbb{E}_{Y,X} \left[\mathbb{E}_\varepsilon \left[\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i m_\theta(Z_i) \right| \mid Y, X \right] \right] \\ &= 2 \mathbb{E}_{Y,X} \left[\mathbb{E}_\varepsilon \left[\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i (\phi(\tilde{\theta}_i) + \log 2) \right| \mid Y, X \right] \right] \end{aligned}$$

First, use the triangle inequality to get rid of the log 2:

$$\leq 2 \mathbb{E}_{Y,X} \left[\mathbb{E}_\varepsilon \left[\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i (\phi(\tilde{\theta}_i) + \log 2) \right| \mid Y, X \right] \right] + (\dots)$$

Now apply the contraction inequality with $\tilde{\Theta} = \{(\langle \theta, x_i \rangle, \dots, \langle \theta, x_n \rangle) : \theta \in \Theta\} \subseteq \mathbb{R}^n$.

$$\begin{aligned} &\leq 4 \mathbb{E}_{Y,X} \left[\mathbb{E}_\varepsilon \left[\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \tilde{\theta}_i \right| \mid Y, X \right] \right] + (\dots) \\ &= 4 \mathbb{E}_{Y,X} \left[\mathbb{E}_\varepsilon \left[\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \langle X_i, \theta \rangle \right| \mid Y, X \right] \right] + (\dots) \\ &= 4 \mathbb{E}_{\varepsilon, X} \left[\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \langle X_i, \theta \rangle \right| \right] + (\dots) \\ &= 4 \mathbb{E}_{\varepsilon, X} \left[\sup_{\theta \in \Theta} \left| \left\langle \frac{1}{n} \sum_{i=1}^n \varepsilon_i X_i, \theta \right\rangle \right| \right] + (\dots) \\ &= 4r \mathbb{E}_{\varepsilon, X} \left[\left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i X_i \right\|_2 \right] + (\dots) \end{aligned}$$

$$\begin{aligned}
&\leq 4r \mathbb{E}_{\varepsilon, X} \left[\left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i X_i \right\|_2^2 \right]^{1/2} + (\dots) \\
&= 4r \left(\frac{\mathbb{E}[\|X\|_2^2]}{n} \right)^{1/2} + (\dots) \\
&\leq 4 \frac{rM}{\sqrt{n}} + (\dots).
\end{aligned}$$

13.5 Further topics: Orlicz processes and bracketing numbers

There is a generalization of sub-Gaussian using the Orlicz norm.

Definition 13.1. Let $\psi_q(t) := \exp(t^q) - 1$ for $q \in [1, 2]$. The q -**Orlicz norm** is

$$\|X\|_{\psi_q} := \inf\{\lambda > 0 : \mathbb{E}[\psi_q(|X|/\lambda)] \leq 1\}.$$

We can prove concentration inequalities, the maximal inequality, the one-step discretization bound, and the chaining bounding in terms of Orlicz norms.

In empirical process theory, there is another notion of covering called the bracketing number. This is discussed in the notes by Sen and in Chapter 2 of Van der Waart and Wellner.

Definition 13.2. Given two functions $\ell(\cdot)$ and $u(\cdot)$, the **bracket**

$$[L, u] = \{f \in \mathcal{F} : \ell(x) \leq f(x) \leq u(x) \forall x \in \mathcal{X}\}.$$

An ε -**bracket** is a bracket $[L, u]$ with $\|\ell - u\| \leq \varepsilon$.

Definition 13.3. The **bracketing number** $N_{[]}(\varepsilon; \mathcal{F}, \|\cdot\|)$ is the minimum number of ε -brackets needed to cover \mathcal{F} , i.e.

$$N_{[]}(\varepsilon; \mathcal{F}, \|\cdot\|) = \min\{N : \{[\ell_i, u_i]_{i \in [N]}\} \text{ covers } \mathcal{F} \text{ and } \|\ell_i - u_i\| \leq \varepsilon\}.$$

Proposition 13.4. Let $\mathcal{R}_n(\mathcal{F}) := \mathbb{E}_{\varepsilon_i, X_i} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right| \right]$. Then

$$\mathcal{R}_n(\mathcal{F}) \lesssim \frac{D_{\mathbb{P}}}{\sqrt{n}} \int_0^1 \sqrt{\log N_{[]} (D_{\mathbb{P}} u; \mathcal{F}, L^2(\mathbb{P}))}.$$

Notice that here, unlike the our bound in terms of covering numbers, does not require us to take the sup over distributions Q . Regardless, usually, if you can prove a bound using the bracketing number, you can prove it using the covering number.

14 Concentration of Sample Covariance of Gaussian Random Vectors

14.1 Eigenvalues of sample covariance of Gaussian random vectors

Last time, we started to talk about the eigenvalues of sample covariance matrices of Gaussian random vectors. We had $(x_i)_{i=1}^n \stackrel{\text{iid}}{\sim} N(0, \Sigma)$, where $\Sigma \in S^{d \times d}$ is a positive definite $d \times d$ matrix. We have

$$X = \begin{bmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{bmatrix} \in \mathbb{R}^{n \times d}, \quad \widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n x_i x_i^\top = \frac{1}{n} X^\top X \in \mathbb{S}^{d \times d}.$$

We had the following theorem about the singular values of the random matrix.

Theorem 14.1.

1. $\mathbb{P}(\sigma_{\max}(X)/\sqrt{n} \geq \gamma_{\max}(\sqrt{\Sigma})(1 + \tau) + \sqrt{\text{tr}(\Sigma)/n}) \leq e^{-nt^2/2}.$
2. $\mathbb{P}(\sigma_{\min}(X)/\sqrt{n} \leq \gamma_{\min}(\sqrt{\Sigma})(1 + \tau) - \sqrt{\text{tr}(\Sigma)/n}) \leq e^{-nt^2/2}.$

The proof strategy was the following:

Proof. For simplicity, take $\Sigma = 1$. We had three main steps:

- (a) Concentration: $\mathbb{P}(|\sigma_k(X) - \mathbb{E}[\sigma_k(X)]| \geq t) \leq 2e^{-t^2/2}.$
- (b) $\mathbb{E}[\sigma_{\max}(X)] \leq \sqrt{n} + \sqrt{d}.$
- (c) $\mathbb{E}[\sigma_{\min}(X)] \geq \sqrt{n} - \sqrt{d}.$

Now we will give the details.

To prove (a), we need to show that the singular values are Lipschitz. By Weyl's inequality,

$$|\sigma_k(x_1) - \sigma_k(x_2)| \leq \|X_1 - X_2\|_{\text{op}} \leq \|X_1 - X_2\|_F.$$

This implies that $\sigma_k(X)$ is 1-Lipschitz in $\|\cdot\|_F$, the Frobenius norm. Therefore, we get Gaussian concentration, i.e. $\sigma_k(X) - \mathbb{E}[\sigma_k(X)]$ is sG(1).

To prove (b), we wanted an upper bound of $\sigma_{\max}(X)$, using the variational formulation

$$\sigma_{\max} = \sup_{(u,v) \in S^{n-1} \times S^{d-1}} \underbrace{\langle u, Xv \rangle}_{Z_{u,v}}.$$

We introduced the following inequality

Lemma 14.1 (Sudakov-Fernique inequality). *Let $\{Z_\theta\}_{\theta \in T}, \{Y_\theta\}_{\theta \in T}$ be two continuous Gaussian processes on a separable space T with $\mathbb{E}[Z_\theta] = \mathbb{E}[Y_\theta]$. If $\mathbb{E}[(Z_\theta - Z_{\theta'})^2] \leq \mathbb{E}[(Y_\theta - Y_{\theta'})^2]$ for all $\theta, \theta' \in T$, then*

$$\mathbb{E} \left[\max_{\theta \in T} Z_\theta \right] \leq \mathbb{E} \operatorname{squamax}_{\theta \in T} Y_\theta.$$

We will prove this later, but first, let's see how this helps us. Define $Z_{u,v} = \langle u, X_v \rangle$, where $X_{i,j} \stackrel{\text{iid}}{\sim} N(0, 1)$, and define

$$Y_{u,v} = \sum_{i=1}^n u_i g_i + \sum_{j=1}^d v_j h_j = \langle u, g \rangle + \langle v, h \rangle, \quad \text{iid } N(0, 1), h_j \stackrel{\text{iid}}{\sim} N(0, 1).$$

We check the second moment conditions:

$$\mathbb{E}[Z_{u,v} Z_{u',v'}] = \mathbb{E}[\langle X, uv^\top \rangle \langle X, u'(v')^\top \rangle]$$

In the summations, all but the diagonal terms will vanish.

$$\begin{aligned} &= \langle u, v^\top, u'(v')^\top \rangle \\ &= \langle u, u' \rangle \langle v, v' \rangle. \end{aligned}$$

This tells us that

$$\begin{aligned} \mathbb{E}[(Z_{u,v} - Z_{u',v'})^2] &= \underbrace{\mathbb{E}[Z_{u,v}^2]}_{=1} - 2\mathbb{E}[Z_{u,v} Z_{u',v'}] + \underbrace{\mathbb{E}[Z_{u',v'}^2]}_{=1} \\ &= 2 - 2\langle u, u' \rangle \langle v, v' \rangle. \end{aligned}$$

For Y , we have

$$\begin{aligned} \mathbb{E}[(Y_{u,v} - Y_{u',v'})^2] &= \underbrace{\mathbb{E}[Y_{u,v}^2]}_{=1} - \underbrace{2\mathbb{E}[Y_{u,v} Y_{u',v'}]}_{=2(\langle u, u' \rangle + \langle v, v' \rangle)} + \underbrace{\mathbb{E}[Y_{u',v'}^2]}_{=1} \\ &= 4 - 2(\langle u, u' \rangle + \langle v, v' \rangle). \end{aligned}$$

Then

$$\mathbb{E}[(Y_{u,v} - Y_{u',v'})^2] - \mathbb{E}[(Z_{u,v} - Z_{u',v'})^2] = 2(1 - \langle u, u' \rangle)(1 - \langle v, v' \rangle) \geq 0$$

Now, applying the Sudakov-Fernique inequality gives

$$\begin{aligned} \mathbb{E} \left[\max_{(u,v) \in S^{n-1} \times S^{d-1}} \langle u, Xv \rangle \right] &\leq \mathbb{E} \left[\max_{(u,v) \in S^{n-1} \times S^{d-1}} (\langle u, g \rangle + \langle v, h \rangle) \right] \\ &= \mathbb{E} \left[\max_{(u,v) \in S^{n-1} \times S^{d-1}} \langle u, g \rangle \right] + \mathbb{E} \left[\max_{(u,v) \in S^{n-1} \times S^{d-1}} \langle v, h \rangle \right] \end{aligned}$$

$$\begin{aligned}
&= \mathbb{E}[\|g\|_2] + \mathbb{E}[\|h\|_2] \\
&\leq \mathbb{E}[\|g\|_2^2]^{1/2} + \mathbb{E}[\|h\|_2^2]^{1/2} \\
&= \sqrt{n} + \sqrt{d}.
\end{aligned}$$

For (c), we want to show a lower bound for $\sigma_{\min}(X)$. We want to show that $\sigma_{\min} \geq \sqrt{n} - \sqrt{d}$ (with $n \geq d$). We use the variational representation

$$\sigma_{\min}(X) = \min_{v \in S^{d-1}} \max_{u \in S^{n-1}} \underbrace{\langle u, Xv \rangle}_{Z_{u,v}}$$

Here is another Gaussian process inequality which is a sort of generalization of Sudakov-Fernique.

Theorem 14.2 (Gordon's inequality). *Let $(Z_{s,t})_{s \in S, t \in T}, (Y_{s,t})_{s \in S, t \in T}$ be two Gaussian processes with $\mathbb{E}[Z_{s,t}] = \mathbb{E}[Y_{s,t}]$, and suppose that*

$$\begin{cases} \mathbb{E}[(Z_{s,t_1} - Z_{s,t_2})^2] \geq \mathbb{E}[(Y_{s,t_1} - Y_{s,t_2})^2] & \forall t_1, t_2 \in T, s \in S, \\ \mathbb{E}[(Z_{s_1,t_1} - Z_{s_2,t_2})^2] \leq \mathbb{E}[(Y_{s_1,t_1} - Y_{s_2,t_2})^2] & \forall s_1 \neq s_2 \in S, t_1, t_2 \in T. \end{cases}$$

Then

$$\mathbb{E} \left[\max_{s \in S} \min_{t \in T} Z_{s,t} \right] \leq \mathbb{E} \left[\max_{s \in S} \min_{t \in T} Y_{s,t} \right].$$

Take $Y_{u,v} = \langle g, u \rangle + \langle h, v \rangle$. Check that $Z_{u,v}$ and $Y_{u,v}$ satisfy the conditions in the theorem. Then

$$\begin{aligned}
-\mathbb{E}[\sigma_{\min}(X)] &= \mathbb{E} \left[\max_{v \in S^{d-1}} -\|Xv\|_2 \right] \\
&= \mathbb{E} \left[\max_{v \in S^{d-1}} \min_{u \in S^{n-1}} \langle u, -Xv \rangle \right] \\
&\leq \mathbb{E} \left[\max_{v \in S^{d-1}} \min_{u \in S^{n-1}} \langle g, u \rangle + \langle h, v \rangle \right]
\end{aligned}$$

where g, h are iid Gaussian random vectors.

$$\begin{aligned}
&= \mathbb{E} \left[\max_{v \in S^{d-1}} \langle h, v \rangle \right] + \mathbb{E} \left[\min_{u \in S^{n-1}} \langle g, u \rangle \right] \\
&= \underbrace{\mathbb{E}[\|h\|_2]}_{\approx \sqrt{d}} - \underbrace{\mathbb{E}[\|g\|_2]}_{\approx \sqrt{n}}.
\end{aligned}$$

So we get that

$$\mathbb{E}[\sigma_{\min}(X)] \geq \sqrt{n} - \sqrt{d}. \quad \square$$

14.2 Proof of the Sudakov-Fernique inequality

Now we will prove the Sudakov-Fernique inequality using the Gaussian interpolation method. Here is a simpler version of the inequality for when the index set is finite.

Lemma 14.2 (Sudakov-Fernique inequality). *Let $X, Y \in \mathbb{R}^n$ be two continuous Gaussian random vectors with $\mathbb{E}[X] = \mathbb{E}[Y]$. If $\mathbb{E}[(X_i - X_j)^2] \leq \mathbb{E}[(Y_i - Y_j)^2]$ for all i, j , then*

$$\mathbb{E} \left[\max_{i \in [n]} X_i \right] \leq \mathbb{E} \left[\max_{i \in [n]} Y_i \right].$$

Proof. Without loss of generality, we may take X, Y to be independent. Let $\mu = \mathbb{E}[X] = \mathbb{E}[Y]$, and define

$$\tilde{X} = X - \mu, \tilde{Y} = Y - \mu, \in \mathbb{R}^n \quad Z(\theta) = \cos \theta \tilde{X} + \sin \theta \tilde{Y}.$$

Fix $\beta > 0$, and define the soft max function $F_\beta : \mathbb{R}^n \rightarrow \mathbb{R}$ by $F_\beta(x) = \beta^{-1} \log(\sum_{i=1}^n e^{\beta x_i})$. The parameter β determines how soft this “soft max” function is; when $\beta \rightarrow \infty$, this will be the max function. For $\theta \in [0, \pi/2]$, let $\varphi(\theta) = \mathbb{E}[F_\beta(Z(\theta))]$. The idea is that $\varphi(0) \approx \mathbb{E}[\max_{i \in [n]} X_i]$ and $\varphi(\pi/2) \approx \mathbb{E}[\max_{i \in [n]} Y_i]$, and these will be exact as we let $\beta \rightarrow \infty$.

Using Fubini’s theorem and the chain rule, we can calculate the derivative

$$\varphi'(\theta) = \mathbb{E} \left[\sum_{i=1}^n \partial_{x_i} F_\beta(Z(\theta)) (-\sin \theta \tilde{X}_i + \cos \theta \tilde{Y}_i) \right]$$

Using integration by parts or Stein’s lemma,

$$\cos \theta \sin \theta \mathbb{E} \left[\sum_{i,j=1}^n \partial_{x_i, x_j}^2 F_\beta(Z(\theta)) \right] (\mathbb{E}[\tilde{Y}_i \tilde{Y}_j] - \mathbb{E}[\tilde{X}_i \tilde{Y}_j])$$

Define $p_i(x) = \partial_{x_i} F_\beta(x) = e^{\beta x_i} / \sum_{j=1}^n e^{\beta x_j}$, which is a probability distribution on \mathbb{R}^n . Using some algebra with p_i , we can show that $\varphi'(\theta) \geq 0$. This means that φ is increasing, so $\varphi(0) \leq \varphi(\pi/2)$. Then we let $\beta \rightarrow \infty$ to get the inequality. \square

The details of the algebra in the proof are contained in chapter 5 of Wainwright’s book.

14.3 More on Gaussian comparison inequalities

Here are some comments on these Gaussian comparison inequalities, which are very useful in many cases. There is a more general statement of Gordon’s inequality, which contains both an expectation version and a probabilistic version:

Theorem 14.3 (Gordon's inequality). *Let S, T be finite sets (or separable sets with continuous processes). Let $(X_{s,t})_{s \in S, t \in T}, (Y_{s,t})_{s \in S, t \in T}$ be two Gaussian processes with $\mathbb{E}[X_{s,t}] = \mathbb{E}[Y_{s,t}] = 0$, and suppose that*

$$\begin{cases} \mathbb{E}[(X_{s,t_1} - X_{s,t_2})^2] \geq \mathbb{E}[(Y_{s,t_1} - Y_{s,t_2})^2] & \forall t_1, t_2 \in T, s \in S, \\ \mathbb{E}[(X_{s_1,t_1} - X_{s_2,t_1})^2] \leq \mathbb{E}[(Y_{s_1,t_1} - Y_{s_2,t_1})^2] & \forall s_1 \neq s_2 \in S, t_1 \in T. \end{cases}$$

Then

1. For any deterministic function $Q(s, t)$,

$$\mathbb{E} \left[\max_{s \in S} \min_{t \in T} X_{s,t} + Q(s, t) \right] \leq \mathbb{E} \left[\max_{s \in S} \min_{t \in T} Y_{s,t} + Q(s, t) \right].$$

2. If we further have $\mathbb{E}[X_{s,t}^2] = \mathbb{E}[Y_{s,t}^2]$, then for all $\tau \in \mathbb{R}$ and functions $Q(s, t)$, we have

$$\mathbb{P} \left(\min_{s \in S} \max_{t \in T} (X_{s,t} + Q(s, t)) \geq \tau \right) \leq \mathbb{P} \left(\min_{s \in S} \max_{t \in T} (Y_{s,t} + Q(s, t)) \geq \tau \right).$$

For the probabilistic version of the inequality, it is better to assume the mean is zero, but we do not need this for the expectation version.

This inequality can be used to derive the Gaussian contraction inequality: $\mathcal{G}(\phi(T)) \leq \mathcal{G}(T)$ if ϕ is 1-Lipshitz. We can also use it to prove the following.

Theorem 14.4 (Sudakov minorization). *Let $\{X_\theta\}_{\theta \in T}$ be mean 0 Gaussian process on T . Then*

$$\mathbb{E} \left[\sup_{\theta \in T} X_\theta \right] \geq \sup_{\varepsilon > 0} \frac{\varepsilon}{2} \sqrt{\log M(\varepsilon; T, \rho_X)},$$

where $M(\varepsilon; T, \rho_X)$ is the packing number of T with metric $\rho_X(\theta, \theta') = \sqrt{\text{Var}(X_\theta - X_{\theta'})}$.

These applications are shown in chapter 5 of Wainwright's book.

14.4 Concentration of sub-Gaussian sample covariance

Now, we generalize our analysis to the case where x_i are sub-Gaussian random vectors, $\mathbb{E}[x_i x_i^\top] = \Sigma \in S^{d \times d}$ is a positive definite $d \times d$ matrix. Here, we still have

$$X = \begin{bmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{bmatrix} \in \mathbb{R}^{n \times d}, \quad \widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n x_i x_i^\top = \frac{1}{n} X^\top X \in S^{d \times d}.$$

In this context, similar concentration results will hold.

Definition 14.1. We say a mean 0 random variable $x \in \mathbb{R}^d$ is **sub-Gaussian**(σ) if

$$\mathbb{E}[e^{\lambda \langle v, x \rangle}] \leq e^{\lambda^2 \|v\|_2^2 \sigma^2 / 2} \quad \forall \lambda \in \mathbb{R}, v \in \mathbb{R}^d.$$

Remark 14.1. This is not the same as saying that each entry of the vector is sub-Gaussian. But if we suppose $x \in \mathbb{R}^d$ with x_i independent sG(σ), then x is sG(σ):

$$\begin{aligned} \mathbb{E} \left[e^{\lambda \sum_{i=1}^n v_i x_i} \right] &= \prod_{i=1}^n \mathbb{E}[e^{\lambda v_i x_i}] \\ &\leq \prod_{i=1}^n e^{\lambda^2 v_i^2 \sigma^2 / 2} \\ &= e^{\lambda^2 \|v\|_2^2 \sigma^2 / 2}. \end{aligned}$$

Theorem 14.5. Let $(x_i)_{i \in [n]}$ be independent mean zero sG(σ). Then with probability at least $1 - \delta$, we have

$$\|\widehat{\Sigma} - \Sigma\|_{\text{op}} \leq C\sigma^2 \left(\sqrt{\frac{d + \log(1/\delta)}{n}} + \frac{d + \log(1/\delta)}{n} \right).$$

The upper bound is of the same order as the Gaussian case. The only difference is that we lose a universal constant C .

15 Concentration of Sample Covariance of Sub-Gaussian and Bounded Random Vectors

15.1 Concentration of sample covariance of sub-Gaussian vectors

Last time, we were talking about concentration of sub-Gaussian sample covariance. If we have $X_i \stackrel{\text{iid}}{\sim} \mathbb{P} \in \mathcal{P}(\mathbb{R}^d)$ and covariance matrix $\mathbb{E}[X_i X_i^\top] = \Sigma \in S_+^{d \times d}$. Then we can estimate Σ by the sample covariance matrix $\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n X_i X_i^\top = \frac{1}{n} X^\top X \in S_+^{d \times d}$.

Definition 15.1. We say a mean 0 random variable $x \in \mathbb{R}^d$ is **sub-Gaussian**(σ) if

$$\mathbb{E}[e^{\lambda \langle v, x \rangle}] \leq e^{\lambda^2 \|v\|_2^2 \sigma^2 / 2} \quad \forall \lambda \in \mathbb{R}, v \in \mathbb{R}^d.$$

A sufficient condition for $X \in \mathbb{R}^d$ to be $\text{sG}(\sigma)$ is that X_i are independent with $X_i \sim \text{sG}(\sigma)$.

Theorem 15.1. Let $(X_i)_{i \in [n]}$ be independent, mean 0 $\text{sG}(\sigma)$. Then with probability at least $1 - \delta$, we have

$$\|\widehat{\Sigma} - \Sigma\|_{\text{op}} \leq C\sigma^2 \left(\sqrt{\frac{d + \log(1/\delta)}{n}} + \frac{d + \log(1/\delta)}{n} \right).$$

Proof. Here is the high level intuition of the proof:

We can represent

$$\begin{aligned} \|\widehat{\Sigma} - \Sigma\|_{\text{op}} &= \sup_{v \in S^{d-1}} |\langle v, (\widehat{\Sigma} - \Sigma)v \rangle| \\ &= \sup_{v \in S^{d-1}} \left| \frac{1}{n} \sum_{i=1}^n (\langle X_i, v \rangle^2 - \mathbb{E}[\langle X_i, v \rangle^2]) \right|. \end{aligned}$$

- (a) Fix v . Then $\frac{1}{n} \sum_{i=1}^n (\langle X_i, v \rangle^2 - \mathbb{E}[\langle X_i, v \rangle^2])$ has a sub-exponential tail bound.
- (b) If we let $|\Omega_\varepsilon = N|$ be the size of an ε -cover of the sphere, then we get the metric entropy bound (instead of a union bound over all the points on the sphere)

$$\sup_{v \in \Omega_\varepsilon} \left| \frac{1}{n} \sum_{i=1}^n (\langle X_i, v \rangle^2 - \mathbb{E}[\langle X_i, v \rangle^2]) \right| \lesssim \sqrt{\frac{\log(N_\varepsilon/\delta)}{n}} + \frac{\log(N_\varepsilon/\delta)}{n}.$$

- (c) Show that $N_\varepsilon \asymp d$.
- (d) Last, show that the discretization error is multiplicative.

Now for the actual proof:

Let $\Omega_\varepsilon = \{v^1, \dots, v^{N_\varepsilon}\}$ be an ε -covering of S^{d-1} in the $\|\cdot\|_2$ norm. Then $|\Omega_\varepsilon| \leq (1+2/\varepsilon)^d$. We claim that for every matrix $A \in \mathbb{R}^{d \times d}$,

$$\|A\|_{\text{op}} \leq \frac{1}{1-2\varepsilon-\varepsilon^2} \sup_{v \in \Omega_\varepsilon} |\langle v, Av \rangle|.$$

This claim holds because

$$\|A\|_{\text{op}} = \sup_{v \in S^{d-1}} |v, Av\rangle|.$$

Then for all $v \in S^{d-1}$, there is a $v^j \in \Omega_\varepsilon$ such that $\|v - v^j\|_2 \leq \varepsilon$. We can then compare

$$\langle v, Av \rangle = \langle v^j, Av^j \rangle + 2\langle v - v^j, Av^j \rangle + \langle v - v^j, A(v - v^j) \rangle.$$

Using this algebra, we get the bound

$$\sup_{v \in S^{d-1}} |\langle v, Av \rangle| \leq \sup_{w \in \Omega_\varepsilon} |\langle w, Aw \rangle| + (2\varepsilon + \varepsilon^2)\|A\|_{\text{op}}.$$

Rearranging this gives the claim:

$$\|A\|_{\text{op}} \leq \frac{1}{1-2\varepsilon-\varepsilon^2} \sup_{v \in \Omega_\varepsilon} |\langle v, Av \rangle|.$$

Take $\varepsilon = 1/8$, so we have a covering with $|\Omega_\varepsilon| \leq 17^d$. Then

$$\|\widehat{\Sigma} - \Sigma\|_{\text{op}} \leq 2 \sup_{v \in \Omega_{1/8}} |\langle v, (\widehat{\Sigma} - \Sigma)v \rangle|.$$

Now look at the tail bound of $|\langle v, (\widehat{\Sigma} - \Sigma)v \rangle|$ for fixed v . Then

$$|\langle v, (\widehat{\Sigma} - \Sigma)v \rangle| = \left| \frac{1}{n} \sum_{i=1}^n (\langle v, X_i \rangle^2 - \mathbb{E}[\langle v, X_i \rangle^2]) \right|.$$

By assumption, $\langle v, X_i \rangle / \sigma$ is $\text{sG}(1)$, so $(\langle v, X_i \rangle^2 - \mathbb{E}[\langle v, X_i \rangle^2]) / \sigma^2$ is $\text{sE}(1, 1)$. Therefore, $(\frac{1}{n} \sum_{i=1}^n (\langle v, X_i \rangle^2 - \mathbb{E}[\langle v, X_i \rangle^2]) / \sigma^2)$ is $\text{sE}(1/\sqrt{n}, 1/n)$.

Thus, we get the sub-exponential tail bound

$$\mathbb{P}(|\langle v, (\widehat{\Sigma} - \Sigma)v \rangle| \geq \sigma^2 t) \leq 2 \exp(-n \min(t^2, t)).$$

Using a union bound, we get

$$\mathbb{P}(|\langle v, (\widehat{\Sigma} - \Sigma)v \rangle| \geq \sigma^2 t) \leq 2 \exp(-n \min(t^2, t) + d \log 17).$$

Now pick $t = C \max\left\{\sqrt{\frac{d+\log(1/\delta)}{n}}, \frac{d+\log(1/\delta)}{n}\right\}$, so we get

$$\mathbb{P}\left(\sup_{v \in \Omega_{1/8}} |\langle v, (\widehat{\Sigma} - \Sigma)v \rangle| \leq C\sigma^2 \max\left\{\sqrt{\frac{d+\log(1/\delta)}{n}}, \frac{d+\log(1/\delta)}{n}\right\}\right) \geq 1 - \delta.$$

That is, with high probability,

$$\|\widehat{\Sigma} - \Sigma\|_{\text{op}} \leq C\sigma^2 \max\left\{\sqrt{\frac{d+\log(1/\delta)}{n}}, \frac{d+\log(1/\delta)}{n}\right\}. \quad \square$$

15.2 Concentration of sample covariance of bounded random vectors

Theorem 15.2. *Let $X_i \stackrel{\text{iid}}{\sim} X \in \mathbb{R}^d$, and let the covariance matrix $\mathbb{E}[XX^\top] = \Sigma$. Suppose that $\|X\|_2^2 \leq b$ almost surely. Then with probability $1 - \delta$,*

$$\|\widehat{\Sigma} - \Sigma\|_{\text{op}} \lesssim \sqrt{\frac{b\|\Sigma\|_2 \log(d/\delta)}{n}} + \frac{b}{n} \log(d/\delta).$$

Example 15.1. Let $X \sim \text{Unif}(S^{d-1}(\sqrt{d}))$, and let $\Sigma = \mathbb{E}[XX^\top] = \text{Id}$. Then we have $b = d$, so the theorem gives

$$\|\widehat{\Sigma} - \Sigma\|_2 \lesssim \sqrt{\frac{d \log d}{n}} + \frac{d}{n} \log d.$$

The proof of this theorem follows from a matrix Bernstein inequality, which we will now prove.

15.3 Matrix Hoeffding/Bernstein inequality

In general, let $X_1, X_2, \dots, X_n \in \mathbb{R}$ be independent $\text{sG}(\sigma)$ random variables. Then the scalar Hoeffding inequality says

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}[X_i])\right| \geq t\right) \leq 2 \exp\left(-\frac{nt^2}{2\sigma^2}\right).$$

The matrix Hoeffding inequality says

Theorem 15.3. *Let $Q_1, Q_2, \dots, Q_n \in S^{d \times d}$ be independent $\text{sG}(V)$, where $V \in S_+^{d \times d}$. Then*

$$\mathbb{P}\left(\left\|\frac{1}{n} \sum_{i=1}^n (Q_i - \mathbb{E}[Q_i])\right\|_{\text{op}} \geq t\right) \leq 2d \exp\left(-\frac{nt^2}{2\|V\|_{\text{op}}^2}\right).$$

We get an extra factor of d in the bound. Notice that when $d = 1$, notice that this reduces to the scalar Hoeffding inequality. Let's review the proof of the scalar Hoeffding inequality:

Use the scalar Chernoff inequality to get

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n(X_i - \mathbb{E}[X_i]) \geq t\right) \leq \inf_{\lambda \geq 0} \frac{\mathbb{E}[e^{\lambda \sum_{i=1}^n (X_i - \mathbb{E}[X_i])}]}{e^{\lambda t n}}$$

Using the scalar tensorization of the MGF,

$$= \inf_{\lambda \geq 0} \frac{\prod_{i=1}^n \mathbb{E}[e^{\lambda(X_i - \mathbb{E}[X_i])}]}{e^{\lambda t n}}$$

Now we use a scalar MGF bound from the sub-Gaussian definition.

$$\begin{aligned} &\leq \inf_{\lambda \geq 0} \prod_{i=1}^n e^{\lambda^2 \sigma^2 / 2} e^{-\lambda t} \\ &= e^{-\frac{nt^2}{2\sigma^2}}. \end{aligned}$$

The proof of the scalar Bernstein inequality is similar.

15.3.1 Matrix Chernoff inequality

Here is a Matrix Chernoff inequality:

Lemma 15.1. *Let $Q \in S^{d \times d}$ be a symmetric matrix. Then*

$$\mathbb{P}(\lambda_{\max}(Q) \geq t) \leq \inf_{\lambda \geq 0} \frac{\mathbb{E}[\text{tr}(e^{\lambda Q})]}{e^{\lambda t}}.$$

Let $Q \in S^{d \times d}$ be a symmetric matrix with eigendecomposition $Q = U\Lambda U^\top$. If we let $f : \mathbb{R} \rightarrow \mathbb{R}$, we define $f(Q) := U \text{diag}(f(\lambda_1), \dots, f(\lambda_d))U^\top \in S^{d \times d}$, so $e^Q = U \text{diag}(e^{\lambda_1}, \dots, e^{\lambda_d})U^\top$. If f is an analytic function with Taylor expansion $f(x) = \sum_{i=1}^{\infty} \frac{f^{(i)}(0)}{i!} x^i$, then

$$f(Q) = \sum_{i=1}^{\infty} \frac{f^{(i)}(0)}{i!} Q^i.$$

In particular,

$$e^Q = \sum_{i=0}^{\infty} \frac{1}{i!} Q^i.$$

Proof. For $\lambda \geq 0$,

$$\begin{aligned} \mathbb{P}(\lambda_{\max}(Q) \geq t) &= \mathbb{P}(e^{\lambda \lambda_{\max}(Q)} \geq e^{\lambda t}) \\ &= \mathbb{P}(\lambda_{\max}(e^{\lambda Q}) \geq e^{\lambda t}) \end{aligned}$$

Use Markov's inequality.

$$\leq \frac{\mathbb{E}[\lambda_{\max}(e^{\lambda Q})]}{e^{\lambda t}}$$

The largest eigenvalue of a positive definite matrix is upper bounded by its trace.

$$\leq \frac{\mathbb{E}[\text{tr}(e^{\lambda Q})]}{e^{\lambda t}}. \quad \square$$

15.3.2 Sub-Gaussian and sub-exponential matrices

Definition 15.2. A matrix $Q \in S^{d \times d}$ with $\mathbb{E}[Q] = 0$ is **sub-Gaussian**(V) for $V \in S_+^{d \times d}$ if

$$\Phi_Q(\lambda) = \mathbb{E}[e^{\lambda Q}] \preceq e^{\lambda^2 V/2} \quad \forall \lambda \in \mathbb{R}.$$

This is not equivalent to the definition we have given for vectors.

Example 15.2. Let $Q = \varepsilon B$, where $B \in S^{d \times d}$ and $\varepsilon \sim \text{Unif}(\{\pm 1\})$. Then

$$\begin{aligned} \mathbb{E}[e^{\lambda Q}] &= \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \mathbb{E}[Q^k] \\ &= \sum_{i=0}^{\infty} \frac{\lambda^{2i}}{(2i)!} \mathbb{E}[Q^{2i}] \\ &= \sum_{i=1}^{\infty} \frac{\lambda^{2i}}{(2i)!} B^{2i} \\ &\prec \sum_{i=1}^{\infty} \frac{1}{i!} \left(\frac{\lambda^2 B^2}{2} \right)^i \\ &= e^{\lambda^2 B^2/2}. \end{aligned}$$

Similarly, we can define sub-exponential matrices.

Definition 15.3. A matrix $Q \in S^{d \times d}$ with $\mathbb{E}[Q] = 0$ is **sub-exponential**(V, α) for $V \in S_+^{d \times d}$ and $\alpha \in \mathbb{R}_{\geq 0}$ if

$$\Phi_Q(\lambda) = \mathbb{E}[e^{\lambda Q}] \preceq e^{\lambda^2 V/2} \quad \forall |\lambda| \leq \frac{1}{\alpha}.$$

Here is a sufficient condition: Define $\text{Var}(Q) = \mathbb{E}[Q^2] - (\mathbb{E}[Q])^2 \in S_+^{d \times d}$. If $\mathbb{E}[Q] = 0$ and $\|Q\|_{\text{op}} \leq b$ a.s., then $Q \sim \text{sE}(\text{Var}(Q), b)$. This is proved in Wainwright's textbook.

Example 15.3. Let $\|X_i\|_2 \leq \sqrt{b}$, so $\mathbb{E}[X_i X_i^\top] = \Sigma$. Then if we let $Q = X_i X_i^\top - \Sigma$, then $\|Q\|_{\text{op}} \leq b$. This gives

$$\text{Var}(Q) = \mathbb{E}[(X_i X_i^\top - \Sigma)^2] \preceq b\Sigma,$$

so $Q \sim \text{sE}(b\Sigma, b)$.

15.3.3 Tensorization of the matrix MGF

Now we know how to give an upper bound of the matrix MGF. The last step is to see the tensorization of the matrix MGF. The scalar MGF tensorizes as

$$\mathbb{E}[e^{\lambda \sum_{i=1}^n X_i}] = \prod_{i=1}^n \mathbb{E}[e^{\lambda X_i}].$$

This is not true for matrices:

$$\mathbb{E}[e^{\lambda \sum_{i=1}^n Q_i}] \neq \prod_{i=1}^n \mathbb{E}[e^{\lambda Q_i}],$$

since $e^{A+B} \neq e^A e^B$. However, this lemma solves the problem:

Lemma 15.2. *Let Q_1, \dots, Q_n be independent. Then*

$$\mathrm{tr}(\mathbb{E}[e^{\lambda \sum_{i=1}^n Q_i}]) \leq \mathrm{tr}(e^{\sum_{i=1}^n \log \mathbb{E}[e^{\lambda Q_i}]}).$$

To prove this, we use the following general matrix inequality:

Lemma 15.3 (Lieb's inequality, 1973). *Let $H \in S^{d \times d}$. Then the function $f : S_+^{d \times d} \rightarrow \mathbb{R}$ sending $A \mapsto \mathrm{tr}(e^{H + \log A})$ is concave.*

This inequality was originally proven for the use of quantum information theory. Using Lieb's inequality, the lemma is just the repeated application of this concavity and Jensen's inequality. Now we can prove the matrix Hoeffding inequality:

Proof. Let Q_i be independent $\mathrm{sG}(V_i)$ random matrices with $\mathbb{E}[Q_i] = 0$. Use the matrix Chernoff inequality to get

$$\mathbb{P}\left(\lambda_{\max}\left(\frac{1}{n} \sum_{i=1}^n Q_i\right) \geq t\right) \leq \inf_{\lambda \geq 0} \mathbb{E}[\mathrm{tr}(e^{\lambda \sum_{i=1}^n Q_i})] e^{-\lambda t}$$

Using the Matrix tensorization of the MGF,

$$\leq \inf_{\lambda \geq 0} \mathrm{tr}(e^{\sum_{i=1}^n \log \mathbb{E}[e^{\lambda Q_i}]} e^{-\lambda t}$$

Now apply the matrix sub-Gaussian upper bound and the inequality $\log A \preceq \log B$ if $A \prec B$ (which is not true in general for every monotone function) to get

$$\begin{aligned} & \inf_{\lambda \geq 0} \mathrm{tr}(e^{\sum_{i=1}^n (\lambda^2/2) V_i}) e^{-\lambda t} \\ & \leq d \inf_{\lambda \geq 0} e^{(\lambda^2/2)n \|V\|_{\mathrm{op}}} e^{-\lambda t} \\ & = d e^{-\frac{nt^2}{2\|V\|_{\mathrm{op}}}}. \end{aligned} \quad \square$$

This gives the matrix Hoeffding and matrix Bernstein inequalities:

Theorem 15.4 (Matrix Hoeffding inequality). *Let $Q_i \stackrel{\text{ind}}{\sim} \text{sG}(V_i)$ with $\mathbb{E}[Q_i] = 0$. Then*

$$\mathbb{P} \left(\left\| \frac{1}{n} \sum_{i=1}^n Q_i \right\|_{\text{op}} \geq t \right) \leq 2d \exp \left(-\frac{nt^2}{2\sigma^2} \right),$$

where $\sigma^2 = \left\| \frac{1}{n} \sum_{i=1}^n V_i \right\|_{\text{op}}$.

Theorem 15.5 (Matrix Bernstein inequality). *Let $Q_i \stackrel{\text{ind}}{\sim} \text{sE}(V_i, \alpha_i)$ with $\mathbb{E}[Q_i] = 0$. Then*

$$\mathbb{P} \left(\left\| \frac{1}{n} \sum_{i=1}^n Q_i \right\|_{\text{op}} \geq t \right) \leq 2d \exp \left(-n \min \left\{ \frac{t^2}{2\sigma^2}, \frac{t}{2\alpha_*} \right\} \right),$$

where $\sigma^2 = \left\| \frac{1}{n} \sum_{i=1}^n V_i \right\|_{\text{op}}$ and $\alpha_* = \max_{i \in [n]} \alpha_i$.

Remark 15.1. These are symmetric versions of these inequalities. We can prove non-symmetric versions by taking $A \in \mathbb{R}^{n \times d}$ and considering

$$Q = \begin{bmatrix} 0 & A \\ A^\top & 0 \end{bmatrix} \in \mathbb{R}^{(n+d) \times (n+d)}.$$

The singular values of A are related to the eigenvalues of Q .

Going back to the sample covariance, we have $\|X_i\|_2^2 \leq b$ and $\mathbb{E}[X_i X_i^\top] = \Sigma$. Then $\widehat{\Sigma} - \Sigma \sim \text{sE}(b\Sigma, b)$, which gives us the matrix Bernstein bound

$$\mathbb{P}(\|\widehat{\Sigma} - \Sigma\|_{\text{op}} \geq t) \leq 2d \exp \left(-n \min \left\{ \frac{t^2}{2b\|\Sigma\|_{\text{op}}}, \frac{t}{2b} \right\} \right).$$

So with high probability,

$$\|\widehat{\Sigma} - \Sigma\|_{\text{op}} \lesssim \sqrt{\frac{b\|\Sigma\|_{\text{op}} \log(d/\delta)}{n}} + \frac{b}{n} \log(d/\delta).$$

16 Introduction to Sparse Linear Regression

16.1 High-dimensional linear regression

Consider the following high-dimensional linear model, with $y = X\theta^* + w \in \mathbb{R}^n$, where

$X \in \mathbb{R}^{n \times d}$ is the **design matrix** and $y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ is the **response**. We write the design

matrix as

$$X = \begin{bmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{bmatrix}, \quad x_i \in \mathbb{R}^d, i = 1, \dots, n$$

and the parameter as

$$\theta^* = \begin{bmatrix} \theta_1^* \\ \vdots \\ \theta_n^* \end{bmatrix}.$$

We interpret

$$w = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix}$$

as noise. We can also write the problem in the scalar form

$$y_i = \langle x_i, \theta^* \rangle + w_i \quad i = 1, \dots, n.$$

Our task is that we observe (X, y) , and we want to estimate $\theta^* \in \mathbb{R}^d$.

The classical asymptotic regime is that the dimension d is fixed, and the sample size n is large. We will focus on the high dimensional regime, in which both d and n are large, and $d > n$. In high dimensions, least squares will not give a consistent estimate. We need some further assumptions on θ^* and X so that consistent estimation is possible in high dimensions.

We will assume a *sparsity assumption*.

Definition 16.1. For $\theta^* \in \mathbb{R}^d$, define the **support** as

$$S(\theta^*) = \{j \in [d] : \theta_j^* \neq 0\}.$$

We will assume that $|S(\theta^*)| \leq s$. If $S(\theta^*)$ is known, then $n \geq s$ is enough for consistent estimation. We can look at the least-squares problem

$$\min_{\theta_S} \|y - X_S \theta_S\|_2^2, \quad \text{where } \theta_S = (\theta_i)_{i \in S} \in \mathbb{R}^{|S|},$$

$$X_S = \begin{bmatrix} x_{1,S}^\top \\ \vdots \\ x_{n,S}^\top \end{bmatrix} \in \mathbb{R}^{n \times |S|}, \quad \text{where } x_{i,S} = (x_{i,j})_{j \in S} \in \mathbb{R}^{|S|}.$$

which will have a unique minimizer.

Because of this, we will focus on when $S(\theta^*)$. We will show that $n \geq s \log(d/s)$. The interesting regime for this problem is when $s \ll n \ll d$.

16.2 Recovery in the noiseless setting

In the noiseless setting, we have

$$y = X\theta^* \in \mathbb{R}^n, \quad \theta^* \in \mathbb{R}^d,$$

where θ^* is s -sparse. Our task is to recover θ^* given (X, y) . If $n < d$, there will be infinite solutions θ such that $y = X\theta$. The **null space** of X is

$$\text{Null}(X) := \{\Delta \in \mathbb{R}^d : X\Delta = 0\}.$$

For all $\Delta \in \text{Null}(X)$, $\theta = \theta^* + \Delta$ satisfies $y = X\theta$. The **feasible space** of $y = X\theta$ is the affine space $\theta^* + \text{Null}(X) = \{\theta^* + \Delta : \Delta \in \text{Null}(X)\}$. This gives infinitely many solutions.

To find θ^* , we can use **ℓ_0 -norm minimization**:

$$\min_{\theta: y=X\theta} \|\theta\|_0, \quad \|\theta\|_0 = \sum_{i=1}^d \mathbb{1}_{\{\theta_i \neq 0\}}.$$

However, this is computationally hard because this norm is not convex. To solve this problem, we need to search over $S \subseteq [d]$, where $|S|$ is from $1, 2, \dots, s$, and look at whether there is a solution of $y = X_S \theta_S$. The complexity of this problem is

$$\Theta \left(\sum_{k=1}^{s-1} \binom{d}{k} \right) \approx d^s,$$

which is exponential in the sparsity. We would prefer polynomial complexity.

Instead, it is more efficient to consider the convex relaxation of ℓ_1 -norm minimization:

$$\min_{y=X\theta} \|\theta\|_1 = \sum_{i=1}^d |\theta_i|.$$

This problem was called **basis pursuit** in the original 1994 paper by Chen, Donoho, and Saunders.¹² If we consider the convex dual problem, then we get the **LASSO** problem, as

¹²This paper was not published until 1998.

introduced by Tibshirani. This ℓ_1 -norm minimization problem can be reformulated as a linear program and solved efficiently.

Our question is as follows: What is the condition such that the solution

$$\hat{\theta} := \arg \min_{\theta} \{\|\theta\|_1 : y = X\theta\}$$

equals the original θ^* ?

16.3 A sufficient condition for exact recovery

Fix $\theta^* \in \mathbb{R}^d$ with $S(\theta) = s$. We want some condition of $X \in \mathbb{R}^{n \times d}$ such that

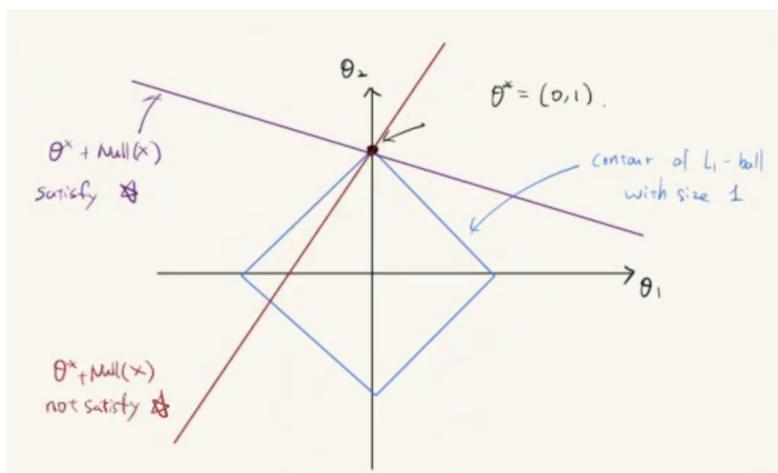
$$\arg \min_{\theta} \{\|\theta\|_1 : X\theta^* = X\theta\} = \theta^*.$$

Notice that $X\theta^* = X\theta$ means that $\theta \in \text{Null}(X) + \theta^*$, so this condition can be reformulated as

$$\forall \theta \in \theta^* + \text{Null}(X) \setminus \{\theta^*\}, \quad \|\theta\|_1 > \|\theta^*\|_1.$$

When will this property hold?

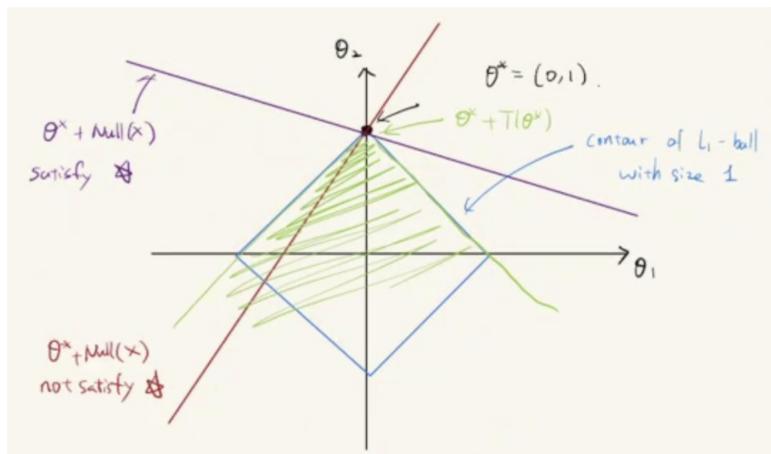
Example 16.1. To gain some intuition, consider the case with $d = 2$, $n = 1$, and with $\dim \text{Null}(X) = 1$. Then $\theta^* + \text{Null}(X)$ is an affine space passing through θ^* .



We can define the **tangent cone**

$$T(\theta^*) := \{\Delta \in \mathbb{R}^d : \|\theta^* + t\Delta\|_1 \leq \|\theta^*\|_1 \text{ for some } t > 0\}.$$

This enters the picture as follows.



We can see from the picture that we will not have exact recovery exactly when $\theta^* + \text{Null}(X)$ intersects the tangent cone at more than one point.

We will get exact recovery when

$$\theta^* + \text{Null}(X) \cap \theta^* + T(\theta^*) = \{\theta^*\},$$

which is equivalent to the condition

$$\text{Null}(X) \cap T(\theta^*) = \{0\}.$$

This is a necessary and sufficient condition for exact recovery of θ^* . This condition involves the interplay between properties of X and properties of θ^* .

Let's see how to reformulate this tangent cone. In our example, $d = 2$, $S = \{2\}$, and $\theta^* = (0, 1)^\top$. Then

$$\begin{aligned} T(\theta^*) &= \{(\Delta_1, \Delta_2) : \exists t > 0, \|(0, 1) + (t\Delta_1 + \Delta_2)\|_1 \leq \|(0, 1)\|_1\} \\ &= \{(\Delta_1, \Delta_2) : \exists t > 0, t|\Delta_1| + |1 + t\Delta_2| \leq 1\} \\ &= \{(\Delta_1, \Delta_2) : |\Delta_1| \leq |\Delta_2|, \Delta_2 \leq 0\}. \end{aligned}$$

In general, suppose $S(\theta^*) = S \subseteq [d]$. Then we can express the tangent cone as

$$T(\theta^*) = \{\Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1, \Delta_i, \theta_i^* \leq 0 \forall i \in S\}, \quad \Delta_S = (\Delta_i)_{i \in S}, \Delta_{S^c} = (\Delta_i)_{i \in S^c}.$$

Define the cone

$$\mathbb{C}(S) := \{\Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1\}.$$

Then $T(\theta^*) \subseteq \mathbb{C}(S)$ for any $S(\theta^*) = S$. A sufficient condition for exact recovery is that

$$\mathbb{C}(S) \cap \text{Null}(X) = \{0\}.$$

Definition 16.2. Let $X \in \mathbb{R}^{n \times d}$ with $S \subseteq [d]$. We say that X satisfies the **restricted nullspace property with respect to S** ($\text{RN}(S)$) if

$$\mathbb{C}(S) \cap \text{Null}(X) = \{0\}.$$

Theorem 16.1. *The following are equivalent:*

(a) For all $\theta^* \in \mathbb{R}^d$ with $S(\theta^*) = S$,

$$\arg \min_{\theta} \{\|\theta\|_1 : X\theta = X\theta^*\} = \theta^*.$$

(b) X satisfies the $\text{RN}(S)$, i.e.

$$\mathbb{C}(S) \cap \text{Null}(X) = \{0\}.$$

Earlier, we said that $\text{RN}(S)$ was only a *sufficient* condition for exact recovery. But this theorem says that it is necessary to have exact recovery for *any* θ^* with $S(\theta^*) = S$.

Proof. (b) \implies (a): Let $\hat{\theta} \in \arg \min_{\theta} \{\|\theta\|_1 : X\theta = X\theta^*\}$. Then $\|\hat{\theta}\|_1 \leq \|\theta^*\|_1$. Now suppose we define $\hat{\Delta} = \hat{\theta} - \theta^* \in \text{Null}(X)$; we want to show that $\hat{\Delta} \in \mathbb{C}(S)$. Then we have

$$\begin{aligned} \|\theta_S^*\|_1 &= \|\theta_1^*\| \\ &\geq \|\theta^* + \hat{\Delta}\|_1 \\ &= \|\theta_S^* + \hat{\Delta}_S\|_1 + \underbrace{\|\theta_{S^c}^* + \hat{\Delta}_{S^c}\|_1}_{=0} \end{aligned}$$

Using the triangle inequality,

$$\geq \|\theta_S^*\|_1 - \|\hat{\Delta}_S\|_1 + \|\hat{\Delta}_{S^c}\|_1.$$

Cancelling $\|\theta_S^*\|_1$ on both sides, we get $\|\hat{\Delta}_{S^c}\|_1 \leq \|\hat{\Delta}_S\|_1$. That is, $\hat{\Delta} \in \mathbb{C}(S) \cap \text{Null}(X)$. By our assumption, this means $\hat{\Delta} = 0$, so $\hat{\theta} = \theta^*$.

(a) \implies (b): Let $\tilde{\theta} \in \text{Null}(X) \setminus \{0\}$. We want to construct a θ^* so that to recover θ^* , we need $\text{RN}(S)$. We will not prove this direction because it is mostly more algebra. \square

What are examples of matrices satisfying $\text{RN}(S)$? For a random matrix $X \in \mathbb{R}^{n \times d}$ with $X_{i,j} \stackrel{\text{iid}}{\sim} N(0,1)$, $\text{RN}(S)$ is satisfied with high probability as long as $n \gtrsim s \log(d/s)$. This is one of the main components of **compressed sensing**. If you want to estimate a sparse signal, you can apply a random matrix and solve this ℓ_1 minimization problem.

17 Sufficient Conditions for Exact Recovery in Sparse Linear Regression and Introduction to Noisy, Sparse Linear Regression

17.1 Recap: sparse linear regression via the restricted nullspace condition

Our model is a the high dimensional sparse linear model, $y = X\theta^* \in \mathbb{R}^n$, where $X \in \mathbb{R}^{n \times d}$, $\theta^* \in \mathbb{R}^d$ and the support of θ^* has cardinality $|S(\theta^*)| \leq s$. Given (y, X) , we want to recover θ^* . When $d > n$, we want

$$\hat{\theta} := \arg \min_{y=X\theta} \|\theta\|_1.$$

When can we have exact recovery? Last time, we had the following condition.

Definition 17.1 (Restricted nullspace). Let $S \subseteq [d]$. $X \in \mathbb{R}^{n \times d}$ satisfies **RN**(S) if $\mathbb{C}(S) \cap \text{Null}(X) = \{0\}$, where

$$\mathbb{C}(S) := \{\Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1\}.$$

Theorem 17.1. *The following are equivalent:*

1. For all $\theta^* \in \mathbb{R}^d$ with $S(\theta^*) = S$,

$$\arg \min_{\theta} \{\|\theta\|_1 : X\theta^* = X\theta\} = \theta^*.$$

2. X satisfies **RN**(S), i.e. $\text{Null}(X) \cap \mathbb{C}(S) = \{0\}$.

However, it is hard to verify the restricted nullspace property for a matrix, since we need to check all subsets of $[d]$ of cardinality s . How can we find examples of matrices satisfying this property?

17.2 Two sufficient conditions for the restricted nullspace property

The intuition is that if $d < n$ (which is not the case we want to solve), X is full-rank, so we can take $X^\top X/n = I_d$. This implies that $\text{Null}(X) = \{0\}$ because

$$\|Xv\|_2^2/n = v^\top (X^\top X/n)v = v^\top I_d v = \|v\|_2^2.$$

Since $\mathbb{C}(S)$ is basically $\{\theta : S(\theta) = S\}$, we can restrict to S . So as long as we have $(X^\top X)_{s,s}/n = I_S$, if $v \in \{\theta : S(\theta) = s\} \cap \text{Null}(X)$, we can say

$$v^\top (X^\top X/n)v = v_s^\top (X^\top X/n)_{s,s}v_s = \|v_s\|_2^2.$$

This equals 0, so we get $v_S = 0$; i.e. $v = 0$.

This motivates the following definitions.

Definition 17.2. Let $\Gamma = X^\top X/n - I_d$. The **pairwise incoherence**¹³ is

$$\delta_{\text{PW}}(X) = \max_{i,j} |\Gamma_{i,j}| = \max_{i,j} |(X^\top X/n - I_d)_{i,j}|.$$

The **restricted isometry constant**¹⁴ is

$$\delta_s(X) = \max_{|S| \leq s} \|\Gamma_{S,S}\|_{\text{op}} = \max_{|S| \leq s} \|X_S^\top X_S/n - I_S\|_{\text{op}},$$

where $X_S \in \mathbb{R}^{n \times s}$ is the matrix where we only keep the columns in S .

Note that $\delta_d = \|\Gamma\|_{\text{op}}$.

17.2.1 The pairwise incoherence condition

Proposition 17.1 (Incoherence implies RN(S)). *If $\delta_{\text{PW}}(X) \leq \frac{1}{3s}$, then X satisfies RN(S) for any $|S| \leq s$.*

Proof. Assume that $\delta_{\text{PW}}(X) \leq \frac{1}{3s}$, and take any $\theta \in \text{Null}(X) \setminus \{0\}$; we want to show that $\theta \notin \mathbb{C}(S)$. Let $S \subseteq [d]$ with $|S| \leq s$. That is, our goal is to show that $\|\theta_{S^c}\|_1 > \|\theta_S\|_1$. The nullspace condition gives

$$0 = \|X\theta\|_2^2$$

We now want to decompose this into θ_S and θ_{S^c} so these two quantities appear. Writing $\theta_S \in \mathbb{R}^d$,

$$\begin{aligned} &= \|X(\theta_{S^c} + \theta_S)\|_2^2 \\ &= \theta_S^\top X_S^\top X_S \theta_S + 2\theta_{S^c}^\top X_{S^c}^\top X_S \theta_S + \underbrace{\|X_{S^c} \theta_{S^c}\|_2^2}_{\geq 0}. \end{aligned}$$

This implies that

$$\theta_S^\top X_S^\top X_S \theta_S \leq 2|\theta_{S^c}^\top X_{S^c}^\top X_S \theta_S|.$$

We can normalize by n to get

$$\theta_S^\top (X_S^\top X_S/n) \theta_S \leq 2|\theta_{S^c}^\top (X_{S^c}^\top X_S/n) \theta_S|.$$

The left hand side is

$$\theta_S^\top (X_S^\top X_S/n) \theta_S \geq \lambda_{\min}(X_S^\top X_S/n) \|\theta_S\|_2^2$$

Using the fact that $\|\theta\|_1^2 \leq \|\theta\|_0 \|\theta\|_2^2$, we get the lower bound

$$\geq \lambda_{\min}(X_S^\top X_S/n) \|\theta_S\|_1^2 / s.$$

¹³The pairwise incoherence was introduced in 2001 by Donoho and Huo.

¹⁴The restricted isometry constant was introduced by Candès and Tao in 2005

To upper bound the right hand side, we use the fact that $a^\top Ab \leq \|a\|_1 \|Ab\|_\infty \leq \|a\|_1 \|A\|_{\max} \|b\|_1$. Then

$$2|\theta_{S^c}^\top (X_{S^c}^\top X_S/n)\theta_S| \leq \|\theta_S\|_1 \|\theta_{S^c}\|_1 \|X_{S^c}^\top X_S/n\|_{\max}/$$

Putting these inequalities together gives

$$\frac{\|\theta_{S^c}\|_1}{\|\theta_S\|_1} \geq \frac{\lambda_{\min}(X_s^\top X_s/n)}{2s\|X_{S^c}^\top X_S/n\|_{\max}}.$$

So far, we have not used the pairwise incoherence. We claim that the pairwise incoherence condition $\delta_{\text{PW}}(X) < \frac{1}{3s}$ makes the right hand side > 1 . The key is to observe that $\|X_{S^c}^\top X_S/n\|_{\max} \leq \delta_{\text{PW}}(X)$ and that $\lambda_{\min}(X_s^\top X_s/n) \geq 2/3$ if the pairwise incoherence condition is satisfied. \square

17.2.2 The restricted isometry property

Here is another condition that implies the restricted nullspace property.

Proposition 17.2 (Restricted isometry property implies $\text{RN}(S)$). *If $\delta_{2s}(X) \leq 1/3$, then X satisfies $\text{RN}(S)$ for any $|S| \leq s$.*

This is proposition 7.11 in Wainwright's textbook, and we will not provide the proof here.

Remark 17.1. In general, we have the algebraic inequality

$$\delta_{\text{PW}}(X) \leq \delta_S(X) \leq s\delta_{\text{PW}}(X).$$

The pairwise incoherence is computable in polynomial time, while the weaker RIP condition needs time $\sum_{k=1}^s \binom{d}{k}$. Here is an exercise which shows that we can satisfy these conditions randomly.

Proposition 17.3. *Let $X \in \mathbb{R}^{n \times d}$ and $X_{i,j} \stackrel{\text{iid}}{\sim} N(0, 1)$. Then*

- (a) *If $n \gtrsim s^2 \log d$, then $\delta_{\text{PW}}(X) \leq \frac{1}{3s}$ with high probability.*
- (b) *If $n \gtrsim s \log(\frac{ed}{s})$, then $\delta_{2s}(X) \leq \frac{1}{3}$ with high probability.*

Here is the idea of the proof.

Proof.

(a) Write

$$\begin{aligned}\delta_{\text{PW}} &= \max_{i,j} |(X^\top X/n - I_d)_{i,j}| \\ &= \max_{i,j} \left| \frac{1}{n} \sum_{k=1}^n x_{i,k} x_{j,k} - \delta_{i,j} \right|\end{aligned}$$

Note that $\mathbb{E}[\frac{1}{n} \sum_{i=k}^n X_{i,k} X_{j,k}] = \delta_{i,j}$, so $I_{i,j} = \frac{1}{n} \sum_{i=k}^n X_{i,k} X_{j,k} - \delta_{i,j}$ will be $\text{sE}(\frac{1}{\sqrt{n}} \frac{1}{n})$ for fixed i, j . Then Bernstein's inequality gives

$$\mathbb{P}(|I_{i,j}| \geq t) \leq 2 \exp(-cn \min(t, t^2)).$$

Using a union bound, we get

$$\mathbb{P}\left(\max_{i,j} |I_{i,j}| \geq t\right) \leq 2d^2 \exp(-cn \min(t, t^2)).$$

Now, if we let $t = \frac{1}{3s}$, call the right hand side δ , and solve for n , we get the condition $n \gtrsim s^2 \log(d.\delta)$.

(b) The proof is similar, using the matrix version of concentration. □

Remark 17.2. Certain random matrix distributions will satisfy $\text{RN}(S)$ but not the RIP or coherence. For example, we will show later that if $X_i \stackrel{\text{iid}}{\sim} N(0, \Sigma)$, where $\Sigma = (1 - \mu)I_d + \mu \mathbf{1}\mathbf{1}^\top$, then X still satisfies $\text{RN}(S)$ with high probability. Here is a figure from

Wainwright's textbook:

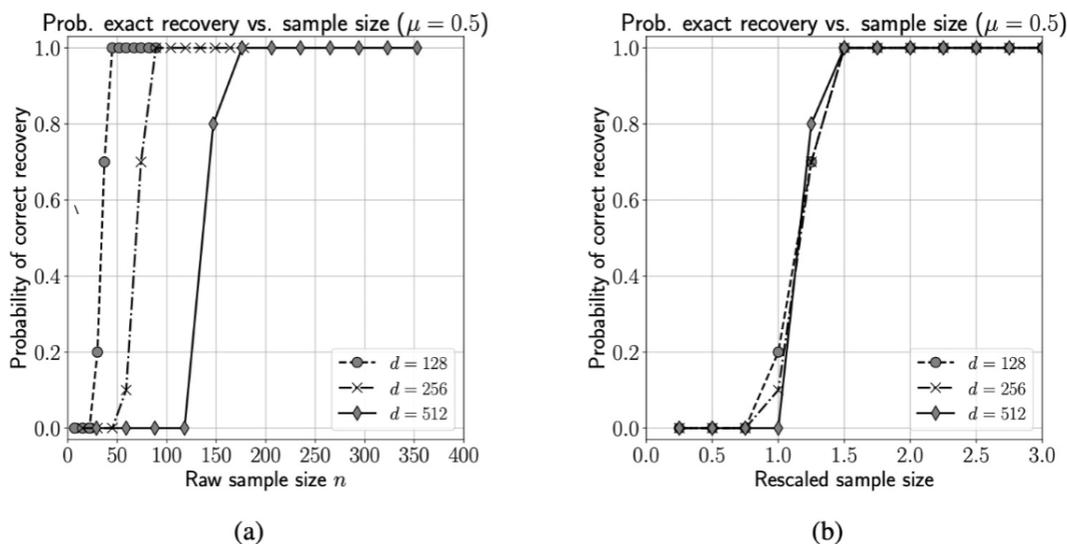


Figure 7.4 (a) Probability of basis pursuit success versus the raw sample size n for random design matrices drawn with i.i.d. rows $X_i \sim \mathcal{N}(0, \Sigma)$, where $\mu = 0.5$ in the model (7.17). Each curve corresponds to a different problem size $d \in \{128, 256, 512\}$ with sparsity $s = \lceil 0.1d \rceil$. (b) The same results replotted versus the rescaled sample size $n/(s \log(ed/s))$. The curves exhibit a phase transition at the same value of this rescaled sample size.

Here, there is a phase transition threshold which needs to be identified with an asymptotic analysis that we will not cover.

17.3 Estimation in the noisy setting

Now we will change our model to $y = X\theta^* + w \in \mathbb{R}^n$, where

$$w = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix} \in \mathbb{R}^n, \quad X \in \mathbb{R}^{n \times d}, \quad X = \begin{bmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{bmatrix}, \quad \theta^* = \begin{bmatrix} \theta_1^* \\ \vdots \\ \theta_n^* \end{bmatrix}.$$

We assume the sparsity condition $|S(\theta^*)| \leq s$. Given (y, X) , we want to estimate θ^* . This time, we want to minimize $\|\theta\|_1$ subject to the constraint that $\|y - X\theta\| \leq b^2$.

Here are three equivalent formulations of the **LASSO problem**, which we use for our estimation:

1. The λ **formulation**:

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

2. **1-norm constrained formulation**:

$$\arg \min_{\theta} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 \right\} \quad \text{s.t. } \|\theta\|_1 \leq R$$

3. The **error constrained formulation**:

$$\arg \min_{\theta} \{ \|\theta\|_1 \} \quad \text{s.t. } \frac{1}{2n} \|y - X\theta\|_2^2 \leq b^2.$$

These are equivalent in the sense that for all $\lambda_n > 0$, there is an $R < \infty$ such that the solution for the 1-norm constrained formulation with parameter R is a solution of the λ formulation. Similarly, we can go the other way. This equivalence requires a condition on X and is just convex duality.

How can we bound the estimation error? We will discuss this next time.

18 Efficient Error Estimation for Noisy, Sparse Linear Regression

18.1 Recap: introduction to noisy, sparse linear regression

We are investigating sparse linear regression, with the model $y = X\theta^* + w \in \mathbb{R}^n$, where

$$w = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix} \in \mathbb{R}^n, \quad X \in \mathbb{R}^{n \times d}, \quad X = \begin{bmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{bmatrix}, \quad \theta^* = \begin{bmatrix} \theta_1^* \\ \vdots \\ \theta_n^* \end{bmatrix}.$$

We assume the sparsity condition $|S(\theta^*)| \leq s$. Given (y, X) , our task is to estimate θ^* . We had three formulations of the LASSO problem:

1. The λ **formulation**:

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

2. **1-norm constrained formulation**:

$$\arg \min_{\theta} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 \right\} \quad \text{s.t. } \|\theta\|_1 \leq R$$

3. The **error constrained formulation**:

$$\arg \min_{\theta} \{\|\theta\|_1\} \quad \text{s.t. } \frac{1}{2n} \|y - X\theta\|_2^2 \leq b^2.$$

Given these three formulations, how can we give a tight upper bound of the estimation error $\|\hat{\theta} - \theta_*\|_2$? Last time, in the noiseless setting, we had the restricted nullspace condition $\text{Null}(X) \cap \mathbb{C}(S) = \{0\}$, which was sufficient for exact recovery of θ_* . In this noisy setting, we will have the **restricted eigenvalue condition**, which will be sufficient for efficient estimation.

18.2 The restricted eigenvalue condition

Recall the \mathbb{C} cone

$$\mathbb{C}(S) := \{\Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1\}.$$

We can modify this by adding a parameter:

$$\mathbb{C}_\alpha(S) := \{\Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \leq \alpha \|\Delta_S\|_1\}.$$

In this extended definition, $\mathbb{C}(S) = \mathbb{C}_1(S)$. If we let $\alpha \rightarrow 0$, we get

$$\mathbb{C}_0(S) = \Delta \in \mathbb{R}^d : S(\Delta) = S\}.$$

Later we will focus on the \mathbb{C}_α cone for $\alpha = 3$.

Definition 18.1. $X \in \mathbb{R}^{n \times d}$ satisfies the **restricted eigenvalue condition** over $S \subseteq [d]$ with parameter (κ, α) (denoted $\text{RE}(S, (\kappa, \alpha))$) if

$$\langle \Delta, (\frac{1}{n}X^\top X)\Delta \rangle = \frac{1}{n}\|X\Delta\|_2^2 \geq \kappa\|\Delta\|_2^2 \quad \forall \Delta \in \mathbb{C}_\alpha(S).$$

This is called the restricted eigenvalue condition because the condition

$$\langle \Delta, (\frac{1}{n}X^\top X)\Delta \rangle \geq \kappa\|\Delta\|_2^2 \quad \forall \Delta \in \mathbb{R}^d$$

is equivalent to $\lambda_{\min}(\frac{1}{n}X^\top X) \geq \kappa$.

Here is some intuition. We can think of the RE condition as a sort of strong convexity for the objective function. Suppose we define the objective function

$$L_n(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2,$$

which we want to minimize to get a minimizer $\hat{\theta}$. The Hessian is

$$\nabla^2 L_n(\theta) = \frac{1}{n}X^\top X \in \mathbb{R}^d.$$

When the sample size is large, we know that there is concentration:

$$\sup_{\theta \in \mathbb{R}^d} |L_n(\theta) - \mathbb{E}[L_n(\theta)]| \leq \text{small},$$

but we want to bound $\|\hat{\theta} - \theta^*\|_2$. If the Hessian is lower bounded by a large number, then the objective function will grow very fast around the minimizer. On the other hand, a

weak bound may mean that the objective function grows too slowly around the minimizer.

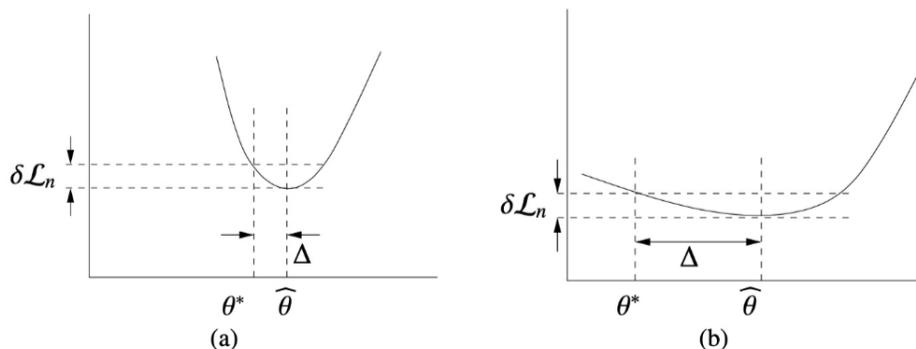


Figure 7.5 Illustration of the connection between curvature (strong convexity) of the cost function, and estimation error. (a) In a favorable setting, the cost function is sharply curved around its minimizer $\hat{\theta}$, so that a small change $\delta\mathcal{L}_n := \mathcal{L}_n(\theta^*) - \mathcal{L}_n(\hat{\theta})$ in the cost implies that the error vector $\Delta = \hat{\theta} - \theta^*$ is not too large. (b) In an unfavorable setting, the cost is very flat, so that a small cost difference $\delta\mathcal{L}_n$ need not imply small error.

18.3 Bounds on ℓ_2 error

Our setting is

$$Y = X\theta^* + w, \quad X \in \mathbb{R}^{n \times d}, \theta^* \in \mathbb{R}^d, w \in \mathbb{R}^n,$$

where $s \ll n \ll d$. We make two assumptions:

(A1): $S(\theta^*) = S \subseteq [d]$, where $|S| = s$.

(A2): X satisfies $\text{RE}(S, (\kappa, \alpha = 3))$.

(A2) is a bit of an abstract condition. Later, we will show that Gaussian random matrices satisfy (A2) when the sample size n is larger enough than the sparsity level s .

Theorem 18.1. *Under assumptions (A1) and (A2),*

(a) λ formulation: Take the Lagrangian parameter $\lambda_n \geq 2\| \frac{X^\top w}{n} \|_\infty$. Then

$$\|\hat{\theta} - \theta^*\|_2 \leq \frac{3}{\kappa} \sqrt{\lambda_n}$$

(b) 1-norm constraint formulation: Take $R = \|\theta^*\|_1$. Then

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

(c) *Error constraint formulation:* Let $b^2 \geq \frac{\|w\|_2^2}{2n}$. Then

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

In all these cases, we have the 1-norm bound

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

Remark 18.1. This theorem is fully deterministic. There is no probability happening, and this theorem is entirely due to algebra.

Remark 18.2. The bound $\frac{1}{\kappa}$ is independent of n .

Remark 18.3. People generally think that the λ formulation is best because the bound is not so sensitive to the choice of the hyperparameter λ_n . In the second formulation, it is also difficult to pick R because we do not know what $\|\hat{\theta}^*\|_1$ is.

In all cases, the error bound is $\sqrt{s} \left\| \frac{X^\top w}{n} \right\|_\infty$, and it is difficult to know what the typical size of this is. We make a further assumption: Assume X is deterministic with $\text{RE}(S, (\kappa, 3))$ with $\max_{j \in [d]} \frac{\|x_j\|_2}{\sqrt{n}} \leq C$, where $x_j \in \mathbb{R}^n$ is the j -th column of X . Let $w \sim \text{sG}(\sigma)$ with $\mathbb{E}[w] = 0$.

If these assumptions hold, then we claim that

$$\left\| \frac{X^\top w}{n} \right\|_\infty = \max_{i \in [d]} |\langle X_i, w \rangle / n| \lesssim \sigma \sqrt{\frac{\log d}{n}}.$$

Here, $\langle x_j, w \rangle / n \sim \text{sG}(\sigma \sqrt{1/n})$. This tells us that

$$\|\hat{\theta} - \theta^*\|_2 \lesssim \sqrt{s} \left\| \frac{X^\top w}{n} \right\|_\infty \lesssim \sqrt{\frac{s \log d}{n}}.$$

So we will have efficient estimation as long as $n \gg (\sigma^2 \vee 1) s \log d$.

18.4 Proof of RE condition bounds

The overall strategy is two steps:

1. Derive a basic inequality (the zero order optimality condition)
2. Algebraic manipulation.

Proof. (b): let's prove the 1-norm constraint formulation,

$$\hat{\theta} = \arg \min_{\theta} \frac{1}{2n} \|y - X\theta\|_2^2 \quad \text{s.t. } \|\theta\|_1 \leq \|\theta^*\|_1 = R.$$

By the optimality of $\hat{\theta}$, we know

$$\frac{1}{2n} \|y - X\hat{\theta}\|_2^2 \leq \frac{1}{2n} \|y - X\theta^*\|_2^2.$$

This is the zero order optimality condition. (The first order optimality condition for optimizing $f(x)$ subject to $g(x) \leq 0$ is $\nabla f(\hat{x}) = \lambda \nabla g(\hat{x})$, where λ is a scalar.) Here the right hand side is $\frac{1}{2n} \|w\|_2^2$, and the left hand side is $\frac{1}{2n} \|w + X(\theta^* - \hat{\theta})\|_2^2$. So we have

$$\begin{aligned} \|w\|_2^2 &\geq \|w + X(\theta^* - \hat{\theta})\|_2^2 \\ &= \|w\|_2^2 + 2\langle w, X(\theta^* - \hat{\theta}) \rangle + \|X(\theta^* - \hat{\theta})\|_2^2 \end{aligned}$$

Denote $\hat{\Delta} = \hat{\theta} - \theta^*$, which is what we want to bound. We can solve this to get

$$\|X\hat{\Delta}\|_2^2 \leq 2\langle w, X\hat{\Delta} \rangle.$$

Thus, our basic inequality is:

$$\frac{1}{n} \|X\hat{\Delta}\|_2^2 \leq \frac{2}{n} w^\top X\hat{\Delta}.$$

If $\hat{\Delta} \in \mathbb{C}_\alpha(S)$, the left hand side can be lower bounded by

$$\frac{1}{n} \|X\hat{\Delta}\|_2^2 \geq \kappa \|\hat{\Delta}\|_2^2,$$

using the restricted eigenvalue condition. To check why $\hat{\Delta} \in \mathbb{C}_\alpha(S)$, note that the condition $\|\hat{\theta}\|_1 \leq \|\theta^*\|_1$ tells us that $\hat{\Delta} \in \mathbb{C}(S) \subseteq \mathbb{C}_3(S)$.

The right hand side can be upper bounded by viewing the scalar $w^\top X\hat{\Delta}$ as the product of the vectors $w^\top X$ and $\hat{\Delta}$:

$$\frac{2}{n} w^\top X\hat{\Delta} \leq \frac{2}{n} \|X^\top w\|_\infty \cdot \|\hat{\Delta}\|_1.$$

Since $\hat{\Delta} \in \mathbb{C}(S)$, we can efficiently bound the 1-norm in terms of the 2-norm:

$$\|\hat{\Delta}\|_1 = \|\hat{\Delta}_{S^c}\|_1 + \|\hat{\Delta}_S\|_1 \leq 2\|\hat{\Delta}_S\|_1 \leq 2\sqrt{s}\|\hat{\Delta}_S\|_2 \leq 2\sqrt{s}\|\hat{\Delta}\|_2.$$

Using this in our inequality and dividing by κ on both sides gives

$$\|\hat{\Delta}\|_2 \leq \frac{4\sqrt{s}}{\kappa} \left\| \frac{X^\top w}{n} \right\|_\infty. \quad \square$$

Remark 18.4. If instead of bounding by $\|X^\top w\|_\infty \cdot \|\hat{\Delta}\|_1$, we try to bound by $\|X^\top w\|_2 \cdot \|\hat{\Delta}\|_2$, then we get $\|\hat{\Delta}\|_2 \leq \frac{2}{\kappa} \|X^\top w/n\|_2 \sim \sqrt{\frac{d}{n}}$. This is worse than the rate $\sqrt{\frac{\log d}{n}}$.

The proof of (c) follows the same lines:

Proof. The error-constraint formulation

$$\hat{\theta} = \arg \min_{\theta} \{\|\theta\|_1\} \quad \text{s.t.} \quad \frac{1}{2n} \|y - X\theta\|_2^2 \leq b^2.$$

gives (using $y - X\hat{\theta} = w - X\hat{\Delta}$).

$$\begin{cases} \|\hat{\theta}\|_1 \leq \|\theta^*\|_1, \\ \frac{1}{2n} \|w + X\hat{\Delta}\|_2^2 \leq \frac{1}{2n} \|w\|_{\text{@}}^2 + (b^2 - \frac{1}{2n} \|w\|_2^2) \end{cases}$$

The algebra proceeds the same as for (b), but we have to keep track of the additive term $\frac{2}{\sqrt{\kappa}} \sqrt{b^2 - \frac{\|w\|_2^2}{n}}$. \square

The proof of (a) has slightly different reasoning:

Proof. We first show that when $\lambda_n \geq 2\|X^\top w/n\|_\infty$, we have $\hat{\Delta} \in \mathbb{C}_3(S)$. By optimality, we have

$$\frac{1}{2n} \|w + X\hat{\Delta}\|_2^2 + \lambda_n \|\theta^* + \hat{\Delta}\|_1 \leq \frac{1}{2n} \|w\|_2^2 + \lambda_n \|\theta^*\|_1.$$

This gives us the Lagrangian basic inequality

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

We can upper bound the right hand side by

$$\begin{aligned} &\leq \left\| \frac{X^\top w}{n} \right\|_\infty (\|\hat{\Delta}\|_1 + \lambda_n (\|\theta_S^*\|_1 - \|\theta_S^* + \hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1)) \\ &\leq \left\| \frac{X^\top w}{n} \right\|_\infty (\|\hat{\Delta}\|_1 + \lambda_n (\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1)) \\ &\leq \frac{\lambda_n}{2} (3\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1). \end{aligned}$$

This upper bound must be nonnegative, so

$$\|\hat{\Delta}_{S^c}\|_1 \leq 3\|\hat{\Delta}_S\|_1,$$

which means that $\hat{\Delta} \in \mathbb{C}_3(S)$. Now, by the RE condition and this bound we have shown,

$$\frac{\kappa}{2} \|\hat{\Delta}\|_2^2 \leq \frac{\lambda_n}{2} (3\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1)$$

$$\leq \frac{\lambda_n}{2} 3\sqrt{s} \|\widehat{\Delta}\|_2.$$

Canceling a factor of $\|\widehat{\Delta}\|_2$ on both sides, we get $\|\widehat{\Delta}\|_2 \leq \frac{3\lambda_n}{\kappa} \sqrt{s}$. □

Next time, we will show that the RE condition is satisfied with high probability for Gaussian random matrices.

19 Restricted Eigenvalue Condition for Gaussian Random Matrices

19.1 Recap: Noisy, sparse linear estimation and the restricted eigenvalue condition

Let's continue our analysis of noisy, sparse linear regression. Our model is $y = X\theta^* + w \in \mathbb{R}^n$, where

$$w \in \mathbb{R}^n, \quad X = \begin{bmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{bmatrix} \in \mathbb{R}^{n \times d}, \quad \theta^* \in \mathbb{R}^d, \quad |S(\theta^*)| \leq s.$$

We looked at the λ formulation of the LASSO problem, where

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

We also looked at the 1-norm constrained and error-constrained formulations of the problem. We defined the \mathbb{C}_α cone

$$\mathbb{C}_\alpha(S) = \{\Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \leq \alpha \|\Delta_S\|_1\}.$$

Using this cone, we defined the restricted eigenvalue condition for efficient bounds on estimation.

Definition 19.1. $X \sim \text{RE}(S, (\kappa, \alpha))$ if

$$\frac{1}{n} \|X\Delta\|_2^2 \geq \kappa \|\Delta\|_2^2 \quad \forall \Delta \in \mathbb{C}_\alpha(S).$$

We proved the following result.

Theorem 19.1. *Assume that $\text{RE}(s, (\kappa, 3))$. With a proper choice of hyperparameter, we have*

$$\|\hat{\theta} - \theta^*\|_2 \lesssim \frac{1}{\kappa} \sqrt{s} \left\| \frac{X^\top w}{n} \right\|_\infty \lesssim \sigma \sqrt{\frac{s \log d}{n}}.$$

Now we would like to answer the question: when does RE hold?

19.2 Restricted eigenvalue condition for Gaussian random matrices

Theorem 19.2. *Let $X_i \stackrel{\text{iid}}{\sim} N(0, \Sigma)$, where $\Sigma \in S_+^{d \times d}$. There exist universal constants $c_1 < 1 < c_2$ such that*

$$\frac{\|X\Delta\|_2^2}{n} \geq c_1 \|\sqrt{\Sigma}\Delta\|_2^2 - c_2 \rho^2(\Sigma) \frac{\log d}{n} \|\Delta\|_1^2 \quad \forall \Delta \in \mathbb{R}^d$$

with probability at least $1 - \frac{e^{-n/32}}{1 - e^{-n/32}}$. Here, $\rho^2(\Sigma) = \max_{i \in [d]} \Sigma_{i,i}$.

We think of this as a generalized RE condition. Let's show that this implies $\text{RE}(S, (\kappa, 3))$ for every S with cardinality $\leq s$. For all $\Delta \in \mathbb{C}_3(S)$, we want to show that $\|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1$. Given the inequality $\|\Delta\|_1^2 \leq 4s\|\Delta\|_2^2$, we can lower bound the right hand side in the theorem:

$$\begin{aligned} c_1\|\sqrt{\Sigma}\Delta\|_2^2 - c_2\rho^2(\Sigma)\frac{\log d}{n}\|\Delta\|_1^2 &\geq c_1\lambda_{\min}(\Sigma)\|\Delta\|_2^2 - c_2\rho^2(\Sigma)\frac{\log d}{n}4s\|\Delta\|_2^2 \\ &= \underbrace{\left(c_1\lambda_{\min}(\Sigma) - 4c_2\rho^2(\Sigma)\frac{s\log d}{n}\right)}_{\geq \frac{1}{2}c\lambda_{\min}(\Sigma)}\|\Delta\|_2^2 \end{aligned}$$

If $n \geq s \log d \frac{8c_2}{c_1} \frac{\rho^2(\Sigma)}{\lambda_{\min}(\Sigma)}$, we have the inequality $4c_2\rho^2(\Sigma)\frac{s\log d}{n} \leq \frac{c_1}{2}\lambda_{\min}(\Sigma)$. We can use it to lower bound the bracketed part.

$$\geq \frac{1}{2}c\lambda_{\min}(\Sigma)\|\Delta\|_2^2.$$

Proof. Let's prove the theorem in the case where $\Sigma = I_d$, so $X_i \stackrel{\text{iid}}{\sim} N(0, I_d)$. Our goal is the inequality

$$\frac{\|X\Delta\|_2^2}{n} + c'_2\frac{\log d}{n}\|\Delta\|_1^2 \geq c'_1\|\Delta\|_2^2 \quad \forall \Delta \in \mathbb{R}^d.$$

Call $\|X\Delta\|_2^2$ the “ X norm of Δ .” We want to relate this to the 1-norm and 2 norm of Δ . A sufficient condition is to have

$$\frac{\|X\Delta\|_2}{\sqrt{n}} + c_2\sqrt{\frac{\log d}{n}}\|\Delta\|_1 \geq c_1\|\Delta\|_2 \quad \forall \Delta \in \mathbb{R}^d$$

because if $a, b > 0$, then $a + b \leq c \implies a^2 + b^2 \leq c^2$. This inequality is invariant to scaling Δ , so it is sufficient to show that

$$\frac{\|X\Delta\|_2}{\sqrt{n}} + c_2\sqrt{\frac{\log d}{n}}\|\Delta\|_1 \geq c_1 \quad \forall \|\Delta\|_2 = 1.$$

So we want to check that

$$\frac{\|X\Delta\|_2}{\sqrt{n}} \geq c_1 - c_2\sqrt{\frac{\log d}{n}}\|\Delta\|_1 \quad \forall \|\Delta\|_2 = 1.$$

It is sufficient to show this for all Δ with bounded 1-norm:

$$\frac{\|X\Delta\|_2}{\sqrt{n}} \geq c_1 - c_2\sqrt{\frac{\log d}{n}}r \quad \forall \|\Delta\|_2 = 1, \|\Delta\|_1 \leq r$$

for all $r > 0$. This means we can show that

$$\inf_{\|\Delta\|_2=1, \|\Delta\|_1 \leq r} \frac{\|X\Delta\|_2}{\sqrt{n}} \geq c_1 - c_2\sqrt{\frac{\log d}{n}}r \quad \forall r > 0.$$

The intuition is that we want to apply the Gaussian comparison inequality, for which we need a $\|X\Delta\|_2$ on the left hand side and no Δ dependence on the right hand side. We have 3 steps:

Step 1: Expectation bound for fixed $r > 0$ (Gaussian comparison inequality)

$$\mathbb{E} \left[\inf_{\|\Delta\|_2=1, \|\Delta\|_1 \leq r} \frac{\|X\Delta\|_2}{\sqrt{n}} \right] \geq c_1 - c_2 \sqrt{\frac{\log d}{n}} r$$

Step 2: Concentration for fixed $r > 0$ (Gaussian concentration)

$$G_r = \left\{ \inf_{\|\Delta\|_2=1, \|\Delta\|_1 \leq r} \frac{\|X\Delta\|_2}{\sqrt{n}} \geq c_1 - c_2 \sqrt{\frac{\log d}{n}} r \right\}$$

occurs with high probability.

Step 3: Union bound over $r > 0$ (Peeling argument): Let $G = \bigcap_{r>0} G_r$, so that $G^c = \bigcup_{r>0} G_r^c$. Then we can calculate

$$\mathbb{P}(G^c) \leq \sum_{r>0} \mathbb{P}(G_r^c).$$

We need to discretize the sum to get a bound that works.

We provide the rest of the proof in lemmas. □

Lemma 19.1 (Gaussian comparison). *There exist constants c_1, c_2 such that*

$$\mathbb{E} \left[\inf_{\|\Delta\|_2=1, \|\Delta\|_1 \leq r} \frac{\|X\Delta\|_2}{\sqrt{n}} \right] \geq c_1 - c_2 \sqrt{\frac{\log d}{n}} r$$

Proof. By the variational representation of the norm,

$$\mathbb{E} \left[\inf_{\|\Delta\|_2=1, \|\Delta\|_1 \leq r} \frac{\|X\Delta\|_2}{\sqrt{n}} \right] = \mathbb{E} \left[\inf_{\Delta \in S^{d-1}(1) \cap B_1(r)} \sup_{u \in S^{n-1}} \frac{\langle u, X\Delta \rangle}{n} \right].$$

By Gordon's inequality,

$$\mathbb{E} \left[\inf_{\Delta \in S} \sup_{u \in T} \langle u, X\Delta \rangle \right] \geq \mathbb{E} \left[\inf_{\Delta \in S} \sup_{u \in T} \langle h, \Delta \rangle + \langle g, u \rangle \right],$$

for any S, T , where $X_{i,j}, g_i, h_i \stackrel{\text{iid}}{\sim} N(0, 1)$. So we get

$$\mathbb{E} \left[\inf_{\|\Delta\|_2=1, \|\Delta\|_1 \leq r} \frac{\|X\Delta\|_2}{\sqrt{n}} \right] \geq \mathbb{E} \left[\inf_{\Delta \in S^{d-1}(1) \cap B_1(r)} \sup_{\|u\|_2=1} \frac{\langle h, \Delta \rangle}{\sqrt{n}} + \frac{\langle g, u \rangle}{\sqrt{n}} \right]$$

$$\begin{aligned}
&= \mathbb{E} \left[\inf_{\Delta} \frac{\langle h, \Delta \rangle}{\sqrt{n}} + \sup_{\|u\|_2=1} \frac{\langle g, u \rangle}{\sqrt{n}} \right] \\
&= \mathbb{E} \left[\inf_{\|\Delta\|_2=1, \|\Delta\|_1 \leq r} \frac{\langle h, \Delta \rangle}{\sqrt{n}} \right] + \mathbb{E} \left[\sup_{\|u\|_2=1} \frac{\langle g, u \rangle}{\sqrt{n}} \right]
\end{aligned}$$

Since $\mathbb{E}[\|g_2\|^2/n] = 1$, the expectation of the square root will be close to 1. We have the lower bound $\mathbb{E}[\|g\|_2/\sqrt{n}] \geq 1/4$. The first term on the other hand, can be expressed as $-\mathbb{E} \left[\sup_{\|\Delta\|_2=1, \|\Delta\|_1 \leq r} \frac{\langle -h, \Delta \rangle}{\sqrt{n}} \right] \geq -\mathbb{E} \left[\sup_{\|\Delta\|_1 \leq r} \frac{\langle -h, \Delta \rangle}{\sqrt{n}} \right] = -\mathbb{E} \left[\frac{\| -h \|_\infty}{\sqrt{n}} \right] r \geq -2\sqrt{\frac{\log d}{n}} r$. So we get

$$\geq \frac{1}{4} - 2\sqrt{\frac{\log d}{n}} r. \quad \square$$

Lemma 19.2 (Concentration). *Let $X_{i,j} \stackrel{\text{iid}}{\sim} N(0, 1)$. The the event*

$$G_r = \left\{ \inf_{\|\Delta\|_2=1, \|\Delta\|_1 \leq r} \frac{\|X\Delta\|_2}{\sqrt{n}} \geq c_1 - c_2\sqrt{\frac{\log d}{n}} r \right\}$$

occurs with high probability.

Proof. Define the function

$$f(X) = \inf_{\|\Delta\|_2=1, \Delta \in S} \frac{\|X\Delta\|_2}{\sqrt{2}}.$$

We want to show that f is Lipschitz for the Frobenius norm, so we can use the Gaussian concentration lemma. Define $\Delta^* = \arg \min \|X_2\Delta\|_2/\sqrt{n}$. Then

$$\begin{aligned}
f(X_1) - f(X_2) &\leq \frac{\|X_1\Delta^*\|_1}{\sqrt{n}} - \frac{\|X_2\Delta^*\|_2}{\sqrt{n}} \\
&\leq \frac{\|(X_1 - X_2)\Delta^*\|_1}{\sqrt{n}} \\
&\leq \frac{\|X_1 - X_2\|_{\text{op}} \|\Delta^*\|_1}{\sqrt{n}} \\
&\leq \frac{\|X_1 - X_2\|_F}{\sqrt{n}}
\end{aligned}$$

This means that f is $\frac{1}{\sqrt{n}}$ -Lipschitz in $\|X\|_F$, so $f(X)$ is $\text{sG}(1/\sqrt{n})$. Then

$$\mathbb{P}(f(X) \leq E[f(X)] - t) \leq e^{-nt^2/2},$$

so

$$G_r := \left\{ \inf_{\|\Delta\|_2=1, \|\Delta\|_1 \leq r} \frac{\|X\Delta\|_2}{\sqrt{n}} \geq c_1 - c_2\sqrt{\frac{\log d}{n}} r \right\}$$

occurs with high probability. □

Lemma 19.3 (Peeling argument). *Let the bad event be*

$$G^c = \left\{ \exists \Delta, \|\Delta\|_2 = 1 \text{ s.t. } \frac{\|X\Delta\|_2}{\sqrt{n}} \leq c_1 - c_2 \sqrt{\frac{\log d}{n}} \|\Delta\|_1 \right\}.$$

then $G^c \subseteq \bigcup_{m=m_{\min}}^{m_{\max}} G_{2^{m+1}}^c$, so $\mathbb{P}(G^c) \leq \sum_{m=m_{\min}}^{m_{\max}} \mathbb{P}(G_{2^{m+1}}^c)$.

Proof. Note that $\|\Delta\|_2 \leq \|\Delta\|_1 \leq \sqrt{d}\|\Delta\|_2$, so we get $1 \leq \|\Delta\|_1 \leq \sqrt{d}$. We discretize the interval in the log scale:

$$[1, \sqrt{d}] = \bigcup_{m=0}^{m_{\max}} [2^m, 2^{m+1}), \quad m_{\max} = \log_2(\sqrt{d}) \approx \log d.$$

The we can write

$$\begin{aligned} G^c &\subseteq \bigcup_{m=m_{\min}}^{m_{\max}} \left\{ \exists \Delta, \|\Delta\|_2 = 1, 2^m \leq \|\Delta\|_1 \leq 2^{m+1} \text{ s.t. } \frac{\|X\Delta\|_2}{\sqrt{n}} \leq c_1 - c_2 \sqrt{\frac{\log d}{n}} 2^m \right\} \\ &\subseteq \underbrace{\left\{ \inf_{\|\Delta\|_2=1, \|\Delta\|_1 \leq 2^{m+1}} \frac{\|X\Delta\|_2}{\sqrt{n}} \leq c_1 - \frac{c_2}{2} \sqrt{\frac{\log d}{n}} \right\}}_{G_{2^{m+1}}^c}. \end{aligned}$$

So we have shown that $G^c \subseteq \bigcup_{m=m_{\min}}^{m_{\max}} G_{2^{m+1}}^c$. □

19.3 LASSO oracle inequality

We have shown that we can efficiently bound the approximation error of θ^* if θ^* is sparse. But what if θ^* is not exactly sparse but is instead approximately sparse? That is, what if $\theta_{S^c}^* \neq 0$ but $\|\theta_{S^c}^*\|_1$ is small?

Definition 19.2. We say that an estimator $\hat{\theta}$ satisfies an **oracle inequality** with respect to the risk R , set Θ , and model $\{\mathbb{P}_\theta : \theta \in \Theta^*\}$ ($\Theta \subseteq \Theta^*$), if there exist constants c and $\varepsilon_n(\mathbb{P}_{\theta^*}, \Theta)$ such that for any $\theta^* \in \Theta^*$, then

$$R(\hat{\theta}; \theta^*) \leq c \underbrace{\inf_{\theta \in \Theta} R(\theta; \theta^*)}_{\text{approx. error/oracle risk}} + \underbrace{\varepsilon_n(\mathbb{P}_{\theta^*}, \Theta)}_{\text{statistical error}}.$$

We hope that c is not too large and that ε_n is small. If $\theta^* \in \Theta$, then

$$\inf_{\theta \in \Theta} R(\theta; \theta^*) = 0.$$

Let $\Theta = \{\Delta \in \mathbb{R}^d : \|\Delta\|_0 \leq s\}$ be the set of s -sparse vectors and let $R(\theta; \theta^*) = \|\theta - \theta^*\|_2$. Then if θ^* is s -sparse, $\inf_{\theta \in \Theta} R(\theta; \theta^*) = 0$. If θ^* is not s -sparse, then

$$\inf_{\theta \in \Theta} R(\theta; \theta^*) > 0.$$

We use our generalized RE condition:

$$\frac{\|X\Delta\|_2^2}{n} \geq c_1 \|\sqrt{\Sigma}\Delta\|_2^2 - c_2 \rho^2(\Sigma) \frac{\log d}{n} \|\Delta\|_1^2, \quad \forall \Delta \in \mathbb{R}^d.$$

Theorem 19.3 (LASSO oracle inequality). *Assume the generalized RE condition holds for $X \in \mathbb{R}^{n \times d}$. Let $\hat{\theta}$ be solution to the λ formulation of LASSO with $\lambda_n \geq 2\| \frac{X^\top w}{n} \|_\infty$. Then for any S with $|S| \leq \frac{c_1}{64c_2} \frac{\bar{\kappa}}{\rho^2(\Sigma)} \frac{n}{\log d}$ (where $\bar{\kappa} = \lambda_{\min}(\Sigma)$),*

$$\|\hat{\theta} - \theta^*\|_2^2 \leq \underbrace{\frac{144 \lambda_n^2}{c_1^2 \bar{\kappa}^2} |S|}_{\text{statistical error} \lesssim \sigma^2 \frac{s \log d}{n}} + \underbrace{\frac{16 \lambda_n}{c_1 \bar{\kappa}} \|\theta_{S^c}^*\|_1 + \frac{32c_2 \rho^2(\Sigma) \log d}{c_1 \bar{\kappa} n} \|\theta_{S^c}^*\|_1^2}_{\text{approx. error/oracle risk} \lesssim \varepsilon_n + \varepsilon_n^2},$$

where $\varepsilon_n = \sqrt{\frac{\log d}{n}} \|\theta_{S^c}^*\|_1$.

Proof. This is a deterministic inequality, so the proof is to derive a basic inequality and then use some algebra. The proof is in the textbook. \square

20 LASSO Prediction Error Bound and High-Dimensional Principal Component Analysis

20.1 Recap: overview of results for noisy, sparse linear regression

Let's finish up our analysis of noisy, sparse linear regression. Our model is $y = X\theta^* + w \in \mathbb{R}^n$, where

$$w \in \mathbb{R}^n, \quad X = \begin{bmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{bmatrix} \in \mathbb{R}^{n \times d}, \quad \theta^* \in \mathbb{R}^d, \quad |S(\theta^*)| \leq s.$$

We looked at the λ formulation of the LASSO problem, where

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

We also looked at the 1-norm constrained and error-constrained formulations of the problem. We defined the \mathbb{C}_α cone

$$\mathbb{C}_\alpha(S) = \{\Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \leq \alpha \|\Delta_S\|_1\}.$$

Using this cone, we defined the restricted eigenvalue condition for efficient bounds on estimation.

Definition 20.1. $X \sim \text{RE}(S, (\kappa, \alpha))$ if

$$\frac{1}{n} \|X\Delta\|_2^2 \geq \kappa \|\Delta\|_2^2 \quad \forall \Delta \in \mathbb{C}_\alpha(S).$$

We proved the following result, upper bounding the estimation error.

Theorem 20.1. *Assume that $\text{RE}(s, (\kappa, 3))$. With a proper choice of hyperparameter, we have*

$$\|\hat{\theta} - \theta^*\|_2 \lesssim \frac{1}{\kappa} \sqrt{s} \left\| \frac{X^\top w}{n} \right\|_\infty \lesssim \sigma \sqrt{\frac{s \log d}{n}}.$$

We also showed that Gaussian random matrices satisfy this condition with high probability.

Theorem 20.2. *Let $X \in \mathbb{R}^{n \times d}$ have iid $N(0, 1)$ entries. If $n \gtrsim s \log d$, then with high probability, $X \sim \text{RE}(S, (\kappa, 3))$ for all $|S| \leq s$.*

20.2 LASSO prediction error bound

Instead of bounding $\|\hat{\theta} - \theta^*\|_2$, we would like to bound the **prediction error** (with fixed design):

$$\frac{1}{n} \mathbb{E}_{\tilde{w}}[\|\tilde{y} - X\hat{\theta}\|_2^2] = \frac{1}{n} \|X(\hat{\theta} - \theta^*)\|_2^2 + \sigma^2,$$

where $\tilde{y} = X\theta^* + \tilde{w}$ and $\tilde{w} \sim N(0, \sigma^2 I_d)$. We can upper bound $\frac{1}{n} \|X(\hat{\theta} - \theta^*)\|_2^2 \leq \|\hat{\theta} - \theta^*\|_2^2 \|X^\top X/n\|_{\text{op}}$; however, this is not always a good bound because $\|X^\top X/n\|_{\text{op}}$, which has order d/n (which blows up for $n \ll d$). Instead, we want to bound the prediction error directly

Theorem 20.3 (Prediction error bound). *Let θ^* be s -sparse. Assume that the hyperparameter in the λ -formulation of the LASSO problem is $\lambda_n \geq 2\|\frac{X^\top w}{n}\|_\infty$. Then*

1. Any optimal solution $\hat{\theta}$ satisfies the bound

$$\frac{1}{n} \|X(\hat{\theta} - \theta^*)\|_2^2 \leq 12\|\theta^*\|_1 \lambda_n.$$

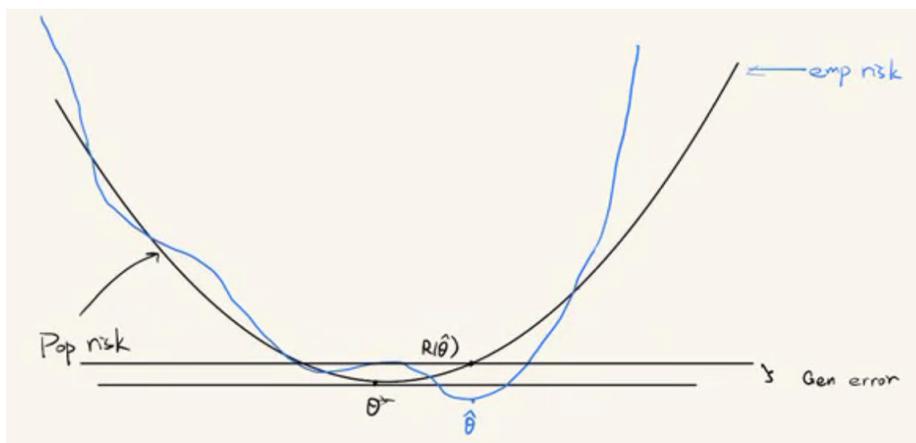
2. If X satisfies $\text{RE}(S, (\kappa, 3))$, then

$$\frac{1}{n} \|X(\hat{\theta} - \theta^*)\|_2^2 \leq \frac{9}{\kappa} s \lambda_n^2.$$

Proof. As before, the proof is a basic inequality, plus some algebra. □

Remark 20.1. The first bound is $\lesssim \|\theta^*\|_1 \sqrt{\frac{\log d}{n}}$, so we get decay $O(1/\sqrt{n})$. This is called the **slow rate bound**. The second bound is $\lesssim s(\sqrt{\frac{\log d}{n}})^2$, so we get decay $O(1/n)$. This is called the **fast rate bound**. Usually, without imposing any geometric assumptions, we get a slower rate bound than we get with such assumptions.

This phenomenon occurs in many settings such as in the empirical risk minimization problem.



The setting is that we have data $(z_i)_{i \in [n]} \stackrel{\text{iid}}{\sim} \mathbb{P}_z$ and a loss function $\ell : \Theta \times Z \rightarrow \mathbb{R}$. The **empirical risk** is

$$\widehat{R}_n(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(\theta; X_i),$$

and the **population risk** is

$$R(\theta) = \mathbb{E}[\ell(\theta; Z_i)].$$

If we take $\widehat{\theta} = \arg \min_{\theta} \widehat{R}_n(\theta)$, the minimizer of the empirical risk, then our **generalization error** is

$$R(\widehat{\theta}) - R(\theta^*).$$

Without geometric assumptions, we can show a **uniform convergence bound**

$$R(\widehat{\theta}) - R(\theta^*) \leq 2 \sup_{\theta \in \Theta} |\widehat{R}_n(\theta) - R(\theta)|.$$

Suppose $\Theta = B(0, 10\|\theta^*\|)$. The upper bound of such an empirical process usually scales linearly in $\|\theta^*\|$, which does not give a very sharp prediction error bound.

Here is what we get with a geometric assumption. Assume that $\kappa \|\widehat{\theta} - \theta^*\|_2^2 \leq (R(\widehat{\theta}) - R(\theta^*))$. Here, κ is a **strong convexity parameter**. With this assumption, we can show an upper bound that is like

$$R(\widehat{\theta}) - R(\theta^*) \leq 2 \sup_{\theta \in B(\theta^*, \|\widehat{\theta} - \theta^*\|_2)} |\widehat{R}_n(\theta) - R(\theta)| \lesssim \|\widehat{\theta} - \theta^*\|_2 \sqrt{\frac{d \log d}{n}}.$$

This is nice because it scales linearly in the estimation error, which is usually smaller than $\|\theta^*\|$. We can bound $\|\widehat{\theta} - \theta^*\|_2 \lesssim \sqrt{\frac{d \log d}{n}}$. Applying the geometric assumption gives the bound

$$R(\widehat{\theta}) - R(\theta^*) \leq \frac{d \log d}{n}.$$

20.3 Principal component analysis in high dimensions

Suppose we observe covariates $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} X \in \mathbb{R}^d$ with $\mathbb{E}[X] = 0$ and $\text{Cov}(X) = \Sigma \in S_+^{d \times d}$. Let the eigenvalues of Σ be $\lambda_1(\Sigma) \geq \lambda_2(\Sigma) \geq \dots \geq \lambda_d(\Sigma) \geq 0$. We can find an orthonormal basis of eigenvectors $v_1(\Sigma), \dots, v_d(\Sigma) \in \mathbb{R}^d$ such that $\Sigma v_i = \lambda_i v_i$ for all $i \in [d]$. If we let $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_d) \in \mathbb{R}^{d \times d}$ and $B = [v_1, \dots, v_d] \in \mathbb{R}^{d \times d}$, then we can write $\Sigma = V \Lambda V^\top$.

The statistical interpretation of v_1 is that

$$\begin{aligned} v_1 &\in \arg \max_{\|v\|_2=1} \text{Var}(\langle x, v \rangle) && X \in \mathbb{R}^d, \mathbb{E}[X] = 0. \\ &= \arg \max_{\|v\|_2=1} \langle v, \mathbb{E}[X X^\top] v \rangle \end{aligned}$$

$$= \arg \max_{\|v_2\|=1} \langle v, \Sigma v \rangle.$$

More generally, if we let $V_k = [v_1, \dots, v_k] \in \mathbb{R}^{d \times k}$, then

$$V_k \in \arg \max_{\substack{U \in \mathbb{R}^{d \times k} \\ \text{partial orth.}}} \underbrace{\mathbb{E}[\|U^\top X\|_2^2]}_{\sum_{i=1}^k \text{Var}(\langle X, u_i \rangle)}.$$

Here is our statistical question: Given samples $\{X_i\}_{i \in [n]} \stackrel{\text{iid}}{\sim} X \in \mathbb{R}^d$, how can we estimate the principal components? Straightforwardly, we can use the eigenvectors of the sample covariance. If we define the sample covariance matrix

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n X_i X_i^\top, \quad \mathbb{E}[\widehat{\Sigma}] = \Sigma,$$

then our estimator is

$$\widehat{\theta} = \arg \max_{\theta} \langle \theta, \widehat{\Sigma} \theta \rangle.$$

By comparison, the ground truth is

$$\theta^* = \arg \max_{\|\theta\|_2=1} \langle \theta, \Sigma \theta \rangle.$$

How close is $\widehat{\theta}$ to θ^* ? We want to translate the closeness of Σ and $\widehat{\Sigma}$ to closeness of θ and θ^* . To quantify this, recall Weyl's eigenvalue perturbation inequality:

Lemma 20.1 (Weyl's inequality). *For any matrices $\widehat{\Sigma}, \Sigma$,*

$$|\lambda(\widehat{\Sigma}) - \lambda_i(\Sigma)| \leq \|\widehat{\Sigma} - \Sigma\|_{\text{op}}.$$

The proof of this fact comes from the variational characterization of the eigenvalues.

For a perturbation inequality for the eigenvectors, we also need the first eigen-gap to be large.

Definition 20.2. Let $\lambda_1(\Sigma) \geq \lambda_2(\Sigma) \geq \dots \geq \lambda_d(\Sigma)$ be the eigenvalues of Σ . Then k -th **eigen-gap** is $\nu_k = \lambda_k - \lambda_{k+1}$.

We will write $\nu = \nu_1$ to refer to the first eigen-gap. You can think of having a large eigen-gap as similar to the restricted eigenvalue condition for LASSO. The parameter ν plays a similar role to κ in LASSO, where $\text{RE}(\mathcal{S}, (\kappa, 3))$ means that $\Delta^\top \frac{X^\top X}{n} \Delta \geq \kappa \|\Delta\|_2^2$.

Example 20.1. Here is an example of instability of a matrix with a small eigen-gap. Suppose we have a diagonal matrix

$$Q_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1.01 \end{bmatrix}.$$

The eigenvalues are $\lambda_1(Q_0) = 1.01$ and $\lambda_2(Q_0) = 1$, so the eigengap is $\nu(Q_0) = 0.01$. In this case, $\theta^*(Q_0) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. Now look at the perturbation

$$Q_\varepsilon = Q_0 + \varepsilon \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & \varepsilon \\ \varepsilon & 1.01 \end{bmatrix},$$

where ε is small. If $\varepsilon = 0.01$, then $\theta^*(Q_\varepsilon) \approx \begin{bmatrix} 0.53 \\ 0.85 \end{bmatrix}$, which is far from $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$.

20.4 General perturbation bound for eigenvectors

Theorem 20.4. *Let $\Sigma \in S_+^{d \times d}$, and let $\theta^* \in \mathbb{R}^d$ be an eigenvector for $\lambda_1(\Sigma)$. Let $\nu = \lambda_1(\Sigma) - \lambda_2(\Sigma) > 0$ be the first eigen-gap. Let the perturbation $P \in S^{d \times d}$ be such that $\|P\|_{\text{op}} < \nu/2$, and let $\widehat{\Sigma} = \Sigma + P$. If $\widehat{\theta} \in \mathbb{R}^d$ is an eigenvector for $\lambda_1(\widehat{\Sigma})$, then*

$$\|\widehat{\theta} - \theta^*\|_2 \leq \frac{2\|\widetilde{P}\|_2}{\nu - 2\|P\|_{\text{op}}}.$$

Here

$$\widetilde{P} = U^\top P U = \begin{bmatrix} \widetilde{P}_{1,1} & \widetilde{P}^\top \\ \widetilde{P} & \widetilde{P}_{2,2} \end{bmatrix} \in \mathbb{R}^{d \times d},$$

where U is the orthogonal matrix such that $\Sigma = U \Lambda U^\top$ and the blocks of \widetilde{P} have sizes

$$\begin{bmatrix} 1 \times 1 & d \times (d-1) \\ (d-1) \times 1 & (d-1) \times (d-1) \end{bmatrix}.$$

If $\|P\|_{\text{op}}$, then we get the bound

$$\|\widehat{\theta} - \theta^*\|_2 \leq \frac{4}{\nu} \|\widetilde{P}\|_2 \leq \frac{4}{\nu} \|P\|_{\text{op}}.$$

To prove this, first let $\widehat{\Delta} = \widehat{\theta} - \theta^*$, and define the quantity

$$\begin{aligned} \Psi(\widehat{\Delta}; P) &= \langle \widehat{\theta}, P \widehat{\theta} \rangle - \langle \theta^*, P \theta^* \rangle \\ &= \langle \widehat{\Delta}, P \widehat{\Delta} \rangle + 2\langle \widetilde{\Delta}, P \theta^* \rangle. \end{aligned}$$

Here is the basic inequality of PCA:

Lemma 20.2 (PCA basic inequality).

$$\nu \cdot (1 - \langle \widehat{\theta}, \theta^* \rangle^2) \leq |\psi(\widehat{\Delta}; P)|.$$

The left hand side measures the distance between $\widehat{\theta}$ and θ^* . We first prove this basic inequality:

Proof. The zero order optimality condition for $\widehat{\theta}$ says that $\widehat{\theta} = \arg \max_{\theta} \langle \theta, \widehat{\Sigma} \theta \rangle$. Then

$$\langle \widehat{\theta}, \widehat{\Sigma} \widehat{\theta} \rangle \geq \langle \theta^*, \widehat{\Sigma} \theta^* \rangle.$$

Recall that $\widehat{\Sigma} = \Sigma + P$. We can express this inequality as

$$\langle \widehat{\theta}, \Sigma \widehat{\theta} \rangle + \langle \widehat{\theta}, P \widehat{\theta} \rangle \geq \langle \theta^*, \Sigma \theta^* \rangle + \langle \theta^*, P \theta^* \rangle.$$

Putting the like terms on each side gives

$$\langle \theta^*, \Sigma \theta^* \rangle - \langle \widehat{\theta}, \Sigma \widehat{\theta} \rangle \leq \langle \widehat{\theta}, P \widehat{\theta} \rangle - \langle \theta^*, P \theta^* \rangle.$$

The right hand side is $\psi(\widehat{\Delta}; P)$.

To figure out the left hand side, write $\widehat{\theta} = \rho \theta^* + \sqrt{1 - \rho^2} z$, where $\|z\|_2 = 1$, $\langle z, \theta^* \rangle = 0$. Then $\rho = \langle \widehat{\theta}, \theta^* \rangle$. We can then expand

$$\begin{aligned} \langle \widehat{\theta}, \Sigma \widehat{\theta} \rangle &= \langle \rho \theta^* + \sqrt{1 - \rho^2} z, \Sigma(\rho \theta^* + \sqrt{1 - \rho^2} z) \rangle \\ &= \rho^2 \underbrace{\langle \theta^*, \Sigma \theta^* \rangle}_{=\lambda_1} + 2\rho \sqrt{1 - \rho^2} \underbrace{\langle \theta^*, \Sigma z \rangle}_{=0} + (1 - \rho^2) \underbrace{\langle z, \Sigma z \rangle}_{\leq 2}. \end{aligned}$$

The bound on the last term is because $\langle z, \Sigma z \rangle \leq \sup_{\|z\|_2=1, \langle z, \theta^* \rangle=0} \langle z, \Sigma z \rangle = \lambda_2$.

$$\leq \rho^2 \lambda_1 + (1 - \rho^2) \lambda_2.$$

So the left hand side is

$$\begin{aligned} \langle \theta^*, \Sigma \theta^* \rangle - \langle \widehat{\theta}, \Sigma \widehat{\theta} \rangle &\geq \lambda_1 - (\rho^2 \lambda_1 + (1 - \rho^2) \lambda_2) \\ &= (\lambda_1 - \lambda_2)(1 - \rho^2) \\ &= \nu(1 - \rho^2). \end{aligned}$$

So we get

$$\nu(1 - \langle \widehat{\theta}, \theta^* \rangle^2) \leq \Psi(\widehat{\Delta}; P). \quad \square$$

Proof. Given the basic inequality, we now upper bound

$$\Psi(\widehat{\Delta}; P) = \langle \widehat{\theta}, P \widehat{\theta} \rangle - \langle \theta^*, P \theta^* \rangle.$$

Write $\Sigma = U \Lambda U^\top$ and $P = U \widetilde{P} U^\top$. We know that $U^\top \theta^* = e_1$, the first standard basis vector, so

$$U^\top \widehat{\theta} = U^\top (\rho \theta^* + \sqrt{1 - \rho^2} z) = \rho e_1 + \sqrt{1 - \rho^2} \underbrace{U^\top z}_{=: \widetilde{z}},$$

where $\|\tilde{z}\|_2 = 1$. Then

$$\begin{aligned}
\Psi(\hat{\Delta}; P) &= \langle U^\top \hat{\theta}, \tilde{P}U^\top \hat{\theta} \rangle - \langle U^\top \theta^*, \tilde{P}U^\top \theta^* \rangle \\
&= \langle \rho e_1 + \sqrt{1 - \rho^2} \tilde{z}, \tilde{P}(\rho e_1 + \sqrt{1 - \rho^2} \tilde{z}) \rangle - \langle e_1, \tilde{P}e_1 \rangle \\
&= \rho^2 \langle e_1, \tilde{P}e_1 \rangle + 2\rho\sqrt{1 - \rho^2} \langle \tilde{z}, \tilde{P}e_1 \rangle + (1 - \rho^2) \langle \tilde{z}, \tilde{P}\tilde{z} \rangle - \langle e_1, \tilde{P}e_1 \rangle \\
&= (1 - \rho^2) \underbrace{\langle e_1, \tilde{P}e_1 \rangle}_{\leq \|P\|_{\text{op}}} + (1 - \rho^2) \langle \tilde{z}, \tilde{P}\tilde{z} \rangle + 2\rho\sqrt{1 - \rho^2} \underbrace{\langle \tilde{z}, \tilde{P}e_1 \rangle}_{\leq \|P\|_2}.
\end{aligned}$$

So, using the basic inequality, we get

$$\nu(1 - \rho^2) \leq 2(1 - \rho^2)\|P\|_{\text{op}} + 2\rho\sqrt{1 - \rho^2}\|\tilde{P}\|_2.$$

We can solve this to get

$$\sqrt{1 - \rho^2} \leq \frac{2\rho\|\tilde{P}\|_2}{\nu - 2\|P\|_{\text{op}}}$$

So

$$\begin{aligned}
\|\hat{\theta} - \theta^*\|_2 &= \sqrt{2(1 - \rho)} \\
&\leq \frac{\sqrt{2}\rho}{\sqrt{1 + \rho}} \frac{2\|\tilde{P}\|_2}{\nu - 2\|P\|_{\text{op}}} \\
&\leq \frac{2\|\tilde{P}\|_2}{\nu - 2\|P\|_{\text{op}}}.
\end{aligned}$$

□

21 Principle Component Analysis for Spiked and Sparse Ensembles

21.1 Recap: estimation error bound for principle component analysis

In high-dimensional principal component analysis, we observe $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} X \in \mathbb{R}^d$, where $\mathbb{E}[X] = 0$ and $\text{Cov}(X) = \Sigma \in \mathbb{R}^{n \times d}$. We have the empirical covariance matrix

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n x_i x_i^\top.$$

The ground truth is

$$\theta^* = \arg \max_{\|\theta\|_2=1} \langle \theta, \Sigma \theta \rangle,$$

while our estimator is

$$\widehat{\theta} = \arg \max_{\|\theta\|_2=1} \langle \theta, \widehat{\Sigma} \theta \rangle.$$

We want to upper bound the estimation error $\|\widehat{\theta} - \theta^*\|_2$.

Last time, he had the following theorem:

Theorem 21.1. *Let $\Sigma \in S_+^{d \times d}$, and let $\theta^* \in \mathbb{R}^d$ be an eigenvector for $\lambda_1(\Sigma)$. Let $\nu = \lambda_1(\Sigma) - \lambda_2(\Sigma) > 0$ be the first eigen-gap. Let the perturbation $P \in S^{d \times d}$ be such that $\|P\|_{\text{op}} < \nu/2$, and let $\widehat{\Sigma} = \Sigma + P$. If $\widehat{\theta} \in \mathbb{R}^d$ is an eigenvector for $\lambda_1(\widehat{\Sigma})$, then*

$$\|\widehat{\theta} - \theta^*\|_2 \leq \frac{2\|\widetilde{P}\|_2}{\nu - 2\|P\|_{\text{op}}}.$$

Here

$$\widetilde{P} = U^\top P U = \begin{bmatrix} \widetilde{P}_{1,1} & \widetilde{P}^\top \\ \widetilde{P} & \widetilde{P}_{2,2} \end{bmatrix} \in \mathbb{R}^{d \times d},$$

where U is the orthogonal matrix such that $\Sigma = U \Lambda U^\top$ and the blocks of \widetilde{P} have sizes

$$\begin{bmatrix} 1 \times 1 & d \times (d-1) \\ (d-1) \times 1 & (d-1) \times (d-1) \end{bmatrix}.$$

21.2 Consequence for a spiked ensemble

In the spiked covariance model, introduced by Jonstone in 2001, we estimate $\theta^* \in \mathbb{R}^d$ with $\|\theta^*\|_2 = 1$. We observe $x_i = \sqrt{\nu} \xi_i \theta^* + w_i$, where

$$\xi_i \in \mathbb{R}, \quad \mathbb{E}[\xi_i] = 0, \quad \mathbb{E}[\xi_i^2] = 1,$$

$$w_i \in \mathbb{R}^d \mathbb{E}[w_i] = 0, \quad \mathbb{E}[w_i w_i^\top] = I_d.$$

The w_i and ξ_i are independent. If we calculate the covariance structure of x_i , we have

$$\begin{aligned} \mathbb{E}[x_i x_i^\top] &= \mathbb{E}(\sqrt{\nu} \xi_i \theta^* + w_i)(\sqrt{\nu} \xi_i \theta^* + w_i^\top) \\ &= \nu \theta^* (\theta^*)^\top + I_d. \end{aligned}$$

This is Σ . The largest eigenvalue is $\lambda_{\max}(\Sigma) = \nu + 1$. The second largest eigenvalue is $\lambda_2(\Sigma)$. So $\nu = \lambda_{\max}(\Sigma) - \lambda_2(\Sigma)$ is the eigengap, and the leading eigenvector of Σ is θ^* . We estimate θ by

$$\hat{\theta} = \arg \max_{\|\theta\|_2=1} \langle \theta, \Sigma \theta \rangle.$$

Our theorem gives us the following bound on $\|\hat{\theta} - \theta^*\|_2$.

Corollary 21.1. *Assume $\xi \sim \text{sG}(1)$ and $w_i \sim \text{sG}(1)$. If $n > d$ and $\sqrt{\frac{\nu+1}{\nu^2}} \sqrt{\frac{d}{n}} \leq \frac{1}{128}$, then*

$$\|\hat{\theta} - \theta^*\|_2 \lesssim \sqrt{\frac{\nu+1}{\nu^2}} \sqrt{\frac{d}{n}}$$

with high probability.

If you want this to be $\leq \varepsilon$, you need $n \gtrsim d \frac{\nu+1}{\nu^2}$. For large ν , $\|\hat{\theta} - \theta^*\|_2 \sim \frac{1}{\sqrt{\nu}}$.

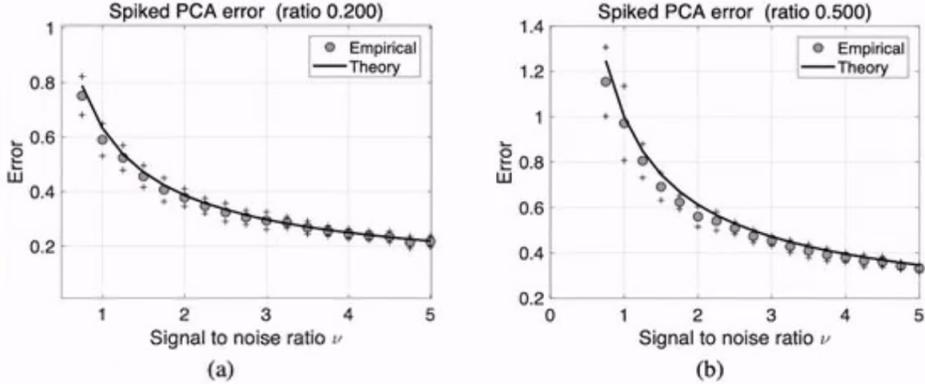


Figure 8.4 Plots of the error $\|\hat{\theta} - \theta^*\|_2$ versus the signal-to-noise ratio, as measured by the eigengap ν . Both plots are based on a sample size $n = 500$. Dots show the average of 100 trials, along with the standard errors (crosses). The full curve shows the theoretical bound $\sqrt{\frac{\nu+1}{\nu^2}} \sqrt{\frac{d}{n}}$. (a) Dimension $d = 100$. (b) Dimension $d = 250$.

Proof. Recall that the theorem says that $\|\hat{\theta} - \theta\|_2 \leq \frac{2\|\tilde{P}\|_2}{\nu-2\|P\|_{\text{op}}}$. We need to upper bound $\|\tilde{P}\|_2$ and $\|P\|_{\text{op}}$.

$$P = \hat{\Sigma} - \Sigma$$

$$\begin{aligned}
&= \frac{1}{n} \sum_{i=1}^n (\sqrt{\nu} \xi \theta^* + w_i)(\sqrt{\nu} \xi_i \theta^* + w_i)^\top - (\nu \theta^* (\theta^*)^\top + I_d) \\
&= \left(\frac{1}{n} \sum_{i=1}^n \xi_i^2 - 1 \right) \nu \theta^* (\theta^*)^\top + \left(\frac{1}{n} \sum_{i=1}^n w_i w_i^\top - I_d \right) + \left(\frac{1}{n} \sum_{i=1}^n \xi_i w_i^\top \right) (\theta^*)^\top + \text{transpose}.
\end{aligned}$$

So we get

$$\|P\|_{\text{op}} \leq \underbrace{\left| \frac{1}{n} \sum_{i=1}^n \xi_i^2 - 1 \right|}_a \nu + \underbrace{\left\| \frac{1}{n} \sum_{i=1}^n w_i w_i^\top - I_d \right\|_{\text{op}}}_c + 2\sqrt{\nu} \underbrace{\left\| \frac{1}{n} \sum_{i=1}^n \xi_i w_i \right\|_2}_b.$$

We can also bound

$$\|\tilde{P}\|_2 \leq \sqrt{\nu} \underbrace{\left\| \frac{1}{n} \sum_{i=1}^n \xi_i w_i \right\|_2}_b + \underbrace{\left\| \frac{1}{n} \sum_{i=1}^n w_i w_i^\top - I_d \right\|_{\text{op}}}_c,$$

so we just need to bound a, b, c .

By sub-exponential concentration, $a \lesssim \sqrt{\frac{1}{n}}$. The term c is a random matrix with mean 0, and using a metric entropy argument with matrix concentration gives $c \lesssim \sqrt{\frac{d}{n}}$.

Similarly, we can show that $b \lesssim \sqrt{\frac{d}{n}}$. Given these upper bounds, we get

$$\|P\|_{\text{op}} \lesssim \nu \sqrt{\frac{1}{n}} + (\sqrt{\nu} + 1) \sqrt{\frac{d}{n}},$$

$$\|\tilde{P}\|_2 \lesssim (\sqrt{\nu} + 1) \sqrt{\frac{d}{n}}.$$

So if $\sqrt{\frac{d}{n}} \lesssim \frac{\nu}{\sqrt{\nu+1}}$, then $\nu - 2\|P\|_{\text{op}} \geq \frac{\nu}{2}$. This gives the bound

$$\|\hat{\theta} - \theta^*\|_2 \lesssim \frac{2\|\tilde{P}\|_2}{\nu/2} \lesssim \sqrt{\frac{\nu+1}{\nu^2}} \sqrt{\frac{d}{n}}.$$

Here, we give an example of how to use the metric entropy bound for the term b .

$$\begin{aligned}
\left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i w_i \right\|_2 &= \sup_{\|\nu\|_2=1} \left\langle \nu, \frac{1}{n} \sum_{i=1}^n \varepsilon_i w_i \right\rangle \\
&= \sup_{\|\nu\|_2=1} \frac{1}{n} \sum_{i=1}^n \underbrace{\varepsilon_i}_{\text{sG}(1)} \underbrace{\langle w_i, \nu \rangle}_{\text{sG}(1)}. \\
&\qquad\qquad\qquad \underbrace{\hspace{10em}}_{\text{sE}(1,1)}
\end{aligned}$$

This tells us that

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \langle w_i, \nu \rangle\right| \geq t\right) \leq 2 \exp(-n \min(t, t^2)) \quad \forall \nu \in S^{d-1}.$$

Now let $\Omega_{1/4}$ be a $1/4$ -cover of S^{d-1} , so $|\Omega_{1/4}| \leq C^d$ for a constant C . Show that this implies

$$\sup_{\nu \in S^{d-1}} |\langle \nu, a \rangle| \leq 2 \sup_{\nu \in \Omega_{1/4}} |\langle \nu, a \rangle|.$$

So we can use a union bound with

$$\begin{aligned} \mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i w_i\right\|_2 \geq t\right) &\leq \mathbb{P}\left(2 \sup_{\nu \in \Omega_{1/4}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i \langle w_i, \nu \rangle \geq t\right) \\ &\leq C^d \exp(-n \min\{t, t^2\}). \end{aligned} \quad \square$$

21.3 Sparse principle component analysis

This is an active research direction. It has been well-studied, but there are some important properties that are not well-understood. We assume that $\theta = \arg \max_{\|\theta\|_2=1} \langle \theta, \Sigma \theta \rangle$ is s -sparse, where $s \ll n \ll d$.

In the sparse spiked covariance model, $\theta^* \in \mathbb{R}^d$, $\|\theta^*\|_2 = 1$, and $|S(\theta^*)| \lesssim s$. We observe

$$x_i = \sqrt{\nu} \xi_i \theta^* + w_i, \quad i \in [n],$$

where $\xi_i \sim \text{sG}(1)$ and $w_i \sim \text{sG}(1)$. We have two theoretical questions:

- What should the sample size be to get a consistent estimator? We will see that as long as $n \gg s$, there is a consistent estimator.
- What is the sample size for a computationally efficient (polynomial time) consistent estimator? The best known computationally efficient estimator has $n \gg s^2$.
- What happens for $s \ll n \ll s^2$? This is an active research direction. It is conjectured that there exists a computational and statistical gap.

21.3.1 ℓ_1 -penalized estimation

To answer part (a), we solve the estimation problem with an added ℓ_1 penalty.

- The 1 -norm constrained formulation is

$$\hat{\theta} = \arg \max_{\substack{\|\theta\|_2=1 \\ \|\theta\|_1 \leq R}} \langle \theta, \hat{\Sigma} \theta \rangle.$$

- The λ -penalized formulation is

$$\hat{\theta} = \arg \max_{\|\theta\|_2=1} \langle \theta, \widehat{\Sigma} \theta \rangle - \lambda_n \|\theta\|_1.$$

In this formulation, we need $\|\theta\|_1 \leq (\frac{n}{\log d})^{1/4}$ for theoretical analysis.

Theorem 21.2. *Assume $n \gtrsim s \log d \cdot \min\{1, \frac{\nu^2}{\nu+1}\}$. Take $\lambda_n \asymp \sqrt{\nu+1} \sqrt{\frac{\log d}{n}}$. Then*

$$\|\hat{\theta} - \theta^*\|_2 \lesssim \sqrt{\frac{\nu+1}{\nu^2}} \sqrt{\frac{s \log d}{n}}.$$

So the required sample size is $\gtrsim s \log d$.

Proof. Here are the steps:

1. Use a basic inequality from the zero order optimality condition to derive a deterministic upper bound of $\|\hat{\theta} - \theta^*\|_2$ by assuming a deterministic assumption on X . This is like imposing the RE condition for LASSO.
2. Prove a concentration inequality and plug in the bound. □

21.3.2 The semidefinite programing relaxation estimator

The 1-norm constrained formulation

$$\max_{\substack{\|\theta\|_2=1 \\ \|\theta\|_1 \leq R}} \langle \theta, \widehat{\Sigma} \theta \rangle$$

is equivalent, by a change of variable $\Theta = \theta\theta^\top \in \mathbb{R}^{d \times d}$ to

$$\max_{\substack{\text{tr}(\Theta)=1 \\ \sum_{j,k} |\Theta_{j,k}| \leq R^2 \\ \text{rank}(\Theta)=1}} \langle \widehat{\Sigma}, \Theta \rangle.$$

The only nonconvex constraint is the rank constraint. If we drop the rank constraint, then the optimization problem becomes convex.

Theorem 21.3 (Amini, Wainwright, 2008). *If $n \gg s^2 \log d$, then the semidefinite programing solution has rank 1 and is consistent.*

21.3.3 The $s \ll n \ll s^2$ regime

What do we know in this regime?

Theorem 21.4 (Berthet, Rigollet, 2013). *For $s \ll n \ll s^2$, sparse PCA is computationally harder or equivalent to the **planted clique problem** in the hard regime.*

It is conjectured that no polynomial time algorithm can solve this problem.

21.4 Extra topics we will not cover

This completes our discussion of the material in chapter 7 and 8 of Wainwright's book. We will not cover chapters 9, 10, or 11, which generalize the material in chapters 7 and 8. Some topics these chapters discuss are

- Logistic LASSO
- Phase retrieval (used in imaging science)
- Matrix sensing
- Matrix completion (used in recommendation systems)

Example 21.1. As an example, we will explain matrix completion. We want to estimate $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$, where $\Theta^* = UV^\top$, $U \in \mathbb{R}^{d_1 \times r}$, $V \in \mathbb{R}^{d_2 \times r}$, and $r \ll \min\{d_1, d_2\}$. We can, for example, think of $\Theta_{i,j}$ as the score of user i given to movie j . Then U_i is user i 's feature, and V_j is movie j 's feature.

We observe $\{M_{i,j} = \Theta_{i,j}^* + \varepsilon_{i,j}\}_{(i,j) \in \Omega}$, and we want to estimate $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$. How many samples is required?

The MLE estimator is

$$\min_{\text{rank}(\Theta) \leq r} \|M_{i,j} - \Theta_{i,j}\|_2^2.$$

This rank constraint is not convex, so we can relax it to a constraint $\|\Theta\|_* \leq r$ on the nuclear norm.

22 Examples of and Oracle Inequality for Non-Parametric Least Squares Regression

22.1 Recap: localized Gaussian complexity bound for non-parametric least squares

We are studying non-parametric regression. Our model is that we observe $x_i \in \mathcal{X}$ and $y_i \in \mathbb{R}$, where

$$y_i = f^*(z_i) + \sigma \cdot w_i, \quad i \in [n]$$

and $f^* \in \mathcal{F} \subseteq \{f : \mathcal{X} \rightarrow \mathbb{R}\}$ is in a designated function class. The noise is $w_i \stackrel{\text{iid}}{\sim} N(0, 1)$.

We consider the non-parametric least squares problem, which has the constrained form

$$\hat{f} \in \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$$

Our goal is to bound the prediction error

$$\|\hat{f} - f^*\|_{L^2(\mathbb{P}_n)} = \frac{1}{n} \sum_{i=1}^n (\hat{f}(x_i) - f^*(x_i))^2.$$

Last time, we proved the following localized Gaussian complexity bound.

Theorem 22.1. *Suppose that $\mathcal{F}^* = \mathcal{F} - \{f^*\}$ is star shaped. Then*

$$\mathbb{E}_{w_i} [\|\hat{f}_n - f^*\|_n^2] \lesssim \delta_n^2,$$

where δ_n^2 solves $\mathcal{G}_n(\delta; \mathcal{F}^*) = \delta^2 / (2\sigma)$, which is

$$\mathcal{G}_n(\delta; \mathcal{F}^*) := \mathbb{E} \left[\sup_{\substack{g \in \mathcal{F}^* \\ \|g\|_n \leq \delta}} \left| \frac{1}{n} \sum_{i=1}^n w_i g(x_i) \right| \right].$$

The chaining method gives us a bound

$$\mathcal{G}_n(\delta; \mathcal{F}^*) \lesssim \frac{\delta^2}{4\sigma} + \frac{16}{\sqrt{n}} \int_{\frac{\delta^2}{4\sigma}}^{\delta} \sqrt{\log N_n(t; B_n(\delta; \mathcal{F}^*))} dt.$$

Let's look at some concrete examples for this localized Gaussian complexity bound.

22.2 Applications of the localized Gaussian complexity bound

Example 22.1. Let $\mathcal{F}_{1:n} = \{f_\theta(\cdot) = \langle \cdot, \theta \rangle : \theta \in \mathbb{R}^d\}$, and let

$$y_i = \langle x_i, \theta^* \rangle + \sigma \cdot w_i, \quad i \in [n],$$

where $\theta^* \in \mathbb{R}^d$. Our estimator is

$$\hat{\theta} = \min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n (y_i - \langle x_i, \theta \rangle)^2,$$

so

$$f_{\hat{\theta}} = \arg \min_{f_\theta \in \mathcal{F}_{1:n}} \frac{1}{n} \sum_{i=1}^n (y_i - f_\theta(x_i))^2.$$

We will show that

$$\begin{aligned} \|f_{\hat{\theta}} - f_{\theta^*}\|_n^2 &= \frac{1}{n} \sum_{i=1}^n \langle x_i, \hat{\theta} - \theta^* \rangle^2 \\ &= \frac{\|X(\theta^* - \hat{\theta})\|_2^2}{n} \\ &\lesssim \sigma^2 \cdot \frac{\text{rank}(X)}{n} \\ &\lesssim \sigma^2 \frac{d}{n}. \end{aligned}$$

We have the upper bound proportional to $\frac{1}{\sqrt{n}} \int_{\frac{\delta^2}{4\delta}}^{\delta} \sqrt{\log N_n(t; B_n(\delta; \mathcal{F}^*))} dt$, so we just need to calculate this covering number. This ball is

$$B_n(\delta; \mathcal{F}_{1:n}) = \left\| f_\theta(x) = \langle x, \theta \rangle : \sqrt{\frac{1}{n} \sum_{i=1}^n \langle x_i, \theta \rangle^2} \leq \delta \right\|,$$

which is isomorphic to the δ -ball in the range of X (where $\dim \text{range}(X) = \text{rank}(X)$). Using a volume argument, the covering number is

$$N_n(t; B_n(\delta; \mathcal{F}_{1:n})) \leq r \cdot \log \left(1 + \frac{2\delta}{t} \right), \quad r = \text{rank}(X).$$

So the metric entropy integral is upper bounded by

$$\frac{\sqrt{r}}{\sqrt{n}} \int_{\frac{\delta^2}{4\delta}}^{\delta} \sqrt{\log \left(1 + \frac{2\delta}{t} \right)} dt \leq c \cdot \delta \sqrt{rn}.$$

We have $c\delta\sqrt{\frac{r}{n}} = \frac{\delta^2}{4\sigma}$, so solving gives $\delta_n = c\sigma\sqrt{\frac{r}{n}}$. So $\delta_n^2 = c\sigma\sqrt{\frac{r}{n}}$, and we get

$$\mathbb{E}_w[\|f_{\hat{\theta}} - f_{\theta^*}\|_n^2] \lesssim \sigma\sqrt{\frac{r}{n}}.$$

Example 22.2 (Lipschitz function class). Let $\mathcal{F}_{\text{Lip}}(L) = \{f : [0, 1] \rightarrow \mathbb{R} : f(0) = 0, f \text{ is } L\text{-Lipschitz}\}$. Then

$$\mathcal{F}^* \subseteq \mathcal{F}_{\text{Lip}}(L) - \mathcal{F}_{\text{Lip}}(L) = \mathcal{F}_{\text{Lip}}(2L).$$

We have upper bounded the metric entropy of this function class as

$$\log N(\varepsilon; \mathcal{F}(2L), \|\cdot\|_\infty) \lesssim \frac{L}{\varepsilon},$$

where $\|f\|_\infty = \sup_{x \in \mathcal{X}}$, so $\|f\|_n = (\frac{1}{n} \sum_{i=1}^n f(x_i)^2)^{1/2} \leq \|f\|_\infty$. This tells us that

$$\log N(\varepsilon; \mathcal{F}(2L), \|\cdot\|_n) \leq \log N(\varepsilon; \mathcal{F}(2L), \|\cdot\|_\infty) \lesssim \frac{L}{\varepsilon}.$$

So the metric entropy integral is

$$\begin{aligned} \frac{1}{\sqrt{n}} \int_{\frac{\delta^2}{4\sigma}}^{\delta} \sqrt{\log N_n(t; \mathcal{F}(2L), \|\cdot\|_\infty)} dt &\leq \frac{1}{\sqrt{n}} \int_{\frac{\delta^2}{4\sigma}}^{\delta} \sqrt{\frac{L}{t}} dt \\ &= \sqrt{\frac{L}{n}} \left(2\sqrt{t} \Big|_{\frac{\delta^2}{4\sigma}}^{\delta} \right) \\ &= c \sqrt{\frac{L}{n}} (\sqrt{\delta} - \sqrt{\delta^2/(4\sigma)}) \\ &\leq c \sqrt{\frac{L}{n}} \sqrt{\delta}. \end{aligned}$$

Solving $\sqrt{\frac{L\delta}{n}} = \delta^2$ gives $\delta^2 \lesssim (\frac{L\sigma^2}{n})^{2/3}$.

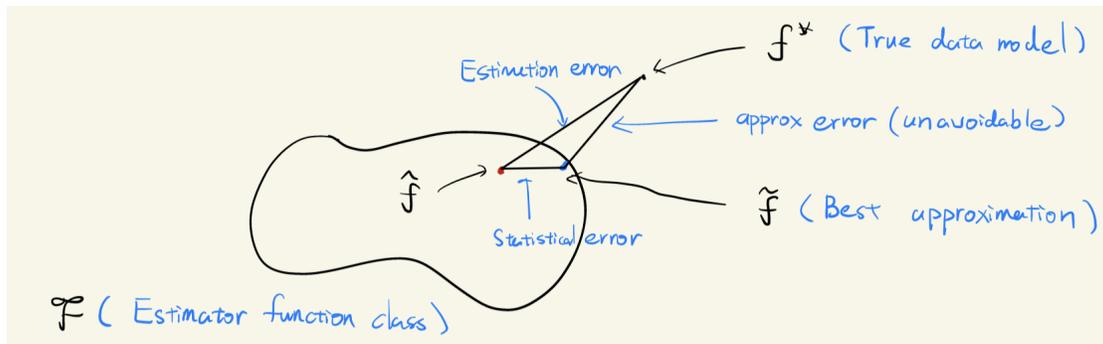
Example 22.3. What if $\log N \asymp \frac{1}{\varepsilon^d}$ for $d \geq 3$ (Lipschitz in d dimensions)? Then

$$\begin{aligned} \frac{1}{\sqrt{n}} \int_{\varepsilon}^{\delta} \frac{1}{t^{d/2}} dt &= \frac{1}{\sqrt{n}} \frac{2}{d-2} \frac{-1}{t^{d/2-1}} \Big|_{\varepsilon}^{\delta} \\ &\leq \frac{1}{\sqrt{n}} \frac{2}{d-2} \frac{1}{\varepsilon^{d/2-1}}. \end{aligned}$$

Take $\varepsilon = \frac{\delta^2}{4\sigma}$ and compare $\frac{1}{\sqrt{n}} \frac{2}{d-2} \frac{1}{\varepsilon^{d/2-1}} = \varepsilon$ to get $\varepsilon \lesssim \frac{1}{n^{4/d}}$. This gives $\delta^2 \lesssim \frac{1}{n^{4/d}}$.

22.3 Oracle inequalities

In practice, we may encounter the situation $f^* \notin \mathcal{F}$, like if we fit a linear model to something which is not exactly linear.



Suppose $\tilde{f} \in \mathcal{F}$ is closest to f^* . We hope that \hat{f} is close to \tilde{f} when we have a lot of samples. That is, we hope that

$$\|\hat{f} - f^*\| \lesssim \inf_{f \in \mathcal{F}} \|f - f^*\| + \varepsilon_n,$$

where $\varepsilon_n \rightarrow 0$ as $n \rightarrow \infty$. We would also like ε_n to decay as fast as possible. This kind of bound gives us a justification that our nonparametric regression gives us a best approximation to the function f^* .

Define $\partial\mathcal{F} = \mathcal{F} - \mathcal{F} = \{f - g : f, g \in \mathcal{F}\}$. Assume that $\partial\mathcal{F}$ is star-shaped; we can always take the star hull to make this true, so this is not a stringent assumption.

Theorem 22.2. Let $\delta_n = \inf\{\delta > 0 : \mathcal{G}_n(\delta; \partial\mathcal{F}) \leq \frac{\delta^2}{2\sigma}\}$. Then there exist constants c_0, c_1, c_2 such that the event

$$\{\|\hat{f} - f^*\|_n^2 \leq \inf_{\gamma \in (0,1)} \left[\frac{1 + \text{gamma}}{1 - \gamma} \|f - f^*\|_n^2 + \frac{c_0}{\gamma(1 - \gamma)} \delta_n t \right] \quad \forall f \in \mathcal{F}$$

occurs with probability at least $1 - c_1 e^{-c_2 \frac{nt\delta_n}{\sigma^2}}$.

This says that

$$\|\hat{f} - f^*\|_n^2 \lesssim \inf_{f \in \mathcal{F}} \|f - f^*\|_n^2 + \delta_n^2,$$

so we can integrate this probability bound to get an expectation bound:

$$\mathbb{E}[\|\hat{f} - f^*\|_n^2] \lesssim \inf_{f \in \mathcal{F}} \|f - f^*\|_n^2 + \delta_n^2 + \frac{\sigma^2}{n}.$$

Note that if $f^* \in \mathcal{F}$, then the first term is 0, so this recovers the prediction error bound in the previous theorem.

Proof. We start from a basic inequality:

$$\frac{1}{2n} \sum_{i=1}^n (y_i - \widehat{f}(x_i))^2 \leq \frac{1}{2n} \sum_{i=1}^n (y_i - f^*(x_i))^2.$$

This tells us that

$$\frac{1}{2} \|\widehat{f} - f^*\|_n^2 \leq \frac{1}{2} \|\widetilde{f} - f^*\|_n^2 + \underbrace{\left| \frac{1}{n} \sum_{i=1}^n w_i (\widehat{f}(x_i) - \widetilde{f}(x_i)) \right|}_{(*)}.$$

We want to upper bound the right term; this is basically the same thing we did for the previous prediction error bound, but with \widetilde{f} instead of f^* . Recall that by definition, $\mathcal{F}_n(\delta; \partial\mathcal{F}) = \mathbb{E}[\sup_{\substack{g \in \partial\mathcal{F} \\ \|g\|_n \leq \delta}} |\frac{1}{n} \sum_{i=1}^n w_i g(x_i)|]$ and $\mathcal{G}_n(\delta; \partial\mathcal{F}) \asymp \delta_n^2$.

The simple case is when $\|\widehat{f} - \widetilde{f}\|_n \leq \delta$. In this case,

$$(*) \lesssim \mathcal{G}_n(\delta_n; \partial\mathcal{F}) \asymp \delta_n^2.$$

The harder case is when $\|\widehat{f} - \widetilde{f}\|_n \geq \delta_n$. In this case, our goal is to show that $(*) \lesssim \delta_n \|\widehat{f} - \widetilde{f}\|_n$.

$$(*) = \left| \frac{1}{n} \sum_{i=1}^n w_i (\widehat{f}(x_i) - \widetilde{f}(x_i)) \frac{\delta_n}{\|\widehat{f} - \widetilde{f}\|_n} \right| \frac{\|\widehat{f} - \widetilde{f}\|_n}{\delta_n}$$

$=: g(x_i)$

Since $\partial\mathcal{F}$ is star-shaped, we have $g \in \partial\mathcal{F}$. Also observe that $\|g\|_n \leq \delta_n$.

$$\lesssim \sup_{\substack{g \in \partial\mathcal{F} \\ \|g\|_n \leq \delta_n}} \left| \frac{1}{n} \sum_{i=1}^n w_i g(x_i) \right| \frac{\|\widehat{f} - \widetilde{f}\|_n}{\delta_n}$$

If we have an argument to show that this quantity concentrates around its mean, we get

$$\begin{aligned} &\lesssim \mathcal{G}_n(\delta_n; \partial\mathcal{F}) \frac{\|\widehat{f} - \widetilde{f}\|_n}{\delta_n} \\ &= \delta_n \|\widehat{f} - \widetilde{f}\|_n. \end{aligned}$$

Using this line of argument, we can show that

$$\|\widehat{f} - f^*\|_n \leq \|\widetilde{f} - f^*\|_n + 2 \max\{\delta_n^2, \delta_n \|\widehat{f} - \widetilde{f}\|_n\}$$

The way to deal with the last term is to use the inequality

$$\begin{aligned} \delta_n \|\widehat{f} - \widetilde{f}\|_n &\leq \delta_n (\|\widehat{f} - f^*\|_n + \|\widetilde{f} - f^*\|_n) \leq \frac{1}{\varepsilon} \delta_n^2 + \varepsilon (\|\widehat{f} - f^*\|_n + \|\widetilde{f} - f^*\|_n)^2 \\ &\leq \frac{1}{\varepsilon} \delta_n^2 + 2\varepsilon \|\widehat{f} - f^*\|_n^2 + 2\varepsilon \|\widetilde{f} - f^*\|_n^2. \end{aligned}$$

Here, we are using the Fenchel-Young inequality, $ab = (a/\sqrt{\varepsilon})(b\sqrt{\varepsilon}) \leq (\frac{a}{\sqrt{\varepsilon}})^2 + (\sqrt{\varepsilon}b)^2$. \square

22.4 Applications of the oracle inequality

Example 22.4. Suppose $\{\phi_m\}_{m=1}^\infty$ is an orthogonal basis of $L^2(\mathbb{P})$, and let $\mathcal{F}_{\text{ortho}}(1, T) := \{f = \sum_{m=1}^T \beta_m \phi_m : \sum_{m=1}^T \beta_m^2 \leq 1\}$. If $f^* = \sum_{m=1}^\infty \theta_m^* \phi_m$, then $f^* \notin \mathcal{F}_{\text{ortho}}$. Using this oracle inequality, we can get

$$\|\widehat{f} - f^*\|_n^2 \lesssim \sum_{m>T} (\theta_m^*)^2 + \sigma^2 \frac{T}{n}.$$

The intuition is that if we have n samples, we can choose $T = \varepsilon n$ so that the right term is small. Then the error is roughly the contribution of the first term.

Example 22.5. Let $y_i = \langle x_i, \theta_* \rangle + \varepsilon_i$, and let $f_{\theta^*} = \langle \cdot, \theta_* \rangle$. Then consider the function class $\mathcal{F}_{\text{sparse}}(s) = \{f_\theta = \langle \cdot, \theta \rangle : \theta \in \mathbb{R}^d, \|\theta\|_0 \leq s\}$. Our estimator is then

$$\widehat{\theta} = \arg \min_{\|\theta\|_0 \leq s} \|y - X\theta\|_2^2.$$

This is the ℓ_0 -variant of LASSO, which is not efficiently computable. Even if the model is not s -sparse, we get

$$\frac{\|X(\widetilde{\theta} - \theta^*)\|_2^2}{n} \leq \inf_{\|\theta\|_0 \leq s} \frac{\|X(\theta - \theta^*)\|_2^2}{n} + \frac{\delta_n^2}{n}.$$

Here, we know that

$$\delta_n^2 \lesssim \sigma^2 \frac{s \log(ed/s)}{n}.$$

In section 13.4.1 of Wainwright's book, there is a discussion of oracle inequalities for regularized estimators.

23 L^2 Prediction Error Bounds for Nonparametric Function Regression

23.1 Recap: prediction error bounds for $\|\cdot\|_n$ compared to $\|\cdot\|_{L^2}$.

We have been studying non-parametric function regression, where we observe $x_i, y_i \in \mathbb{R}$ with $y_i = f^*(x_i) + w_i$ for $i \in [n]$. We assume $f^* \in \mathcal{F} \subseteq \{f : \mathcal{X} \rightarrow \mathbb{R}\}$ for some specific function class \mathcal{F} and take the noise to be $w_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$.

For the non-parametric least squares problem, we have the constrained form

$$\hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2.$$

Our goal is to bound the prediction error,

$$\|\hat{f} - f^*\|_{L^2(\mathbb{P}_n)}^2 = \frac{1}{n} \sum_{i=1}^n (\hat{f}(x_i) - f^*(x_i))^2.$$

We proved a prediction error bound that relies on a critical equation.

Theorem 23.1. *Let $\mathcal{F}^* = \mathcal{F} - \{f^*\}$ be star-shaped. Then $\mathbb{E}_w[\|\hat{f}_n - f^*\|_n^2] \lesssim \delta_n^2$, where δ_n solves the critical equation $\mathcal{G}_n(\delta; \mathcal{F}^*) = \delta^2 \cdot (2\sigma)$.*

What if we want to look at the behavior of \hat{f} on a new dataset $\tilde{x} \sim \mathbb{P}$ instead of x_i in the original dataset? If we have $y_i = f^*(x_i) + \tilde{w}_i$, where $\tilde{w}_i \sim N(0, \sigma^2)$ and $(\tilde{x}_i, \tilde{y}_i) \stackrel{\text{iid}}{\sim} (x_i, y_i)$, we can see that

$$\mathbb{E}_{\tilde{x}_i, \tilde{y}_i} [(\hat{f}(\tilde{x}_i) - \tilde{y}_i)^2] = \sigma^2 + \|\hat{f} - f^*\|_{L^2}^2.$$

So in many cases, we want to control the L^2 distance between \hat{f} and f^* .

23.2 Relation between $\|\cdot\|_n^2$ and $\|\cdot\|_{L^2}^2$

Let $f \in \mathcal{F}$. Then if the function f does not depend on our training data set,

$$\begin{aligned} \mathbb{E}_X[\|f\|_n^2] &= \mathbb{E}_x \left[\frac{1}{n} \sum_{i=1}^n f(x_i)^2 \right] \\ &= \mathbb{E}[f(x)^2] \\ &= \|f\|_{L^2}^2. \end{aligned}$$

Now suppose that $\hat{f}(x) = h(x; \{x_i, y_i\}_{i \in [n]})$ depends on our training data set. Then

$$\mathbb{E}_{x_i}[\|\hat{f} - f^*\|_n^2] = \mathbb{E}_x \left[\frac{1}{n} \sum_{i=1}^n (\hat{f}(x_i; \{x_i, y_i\}_{i \in [n]}) - f^*(\tilde{x}))^2 \right] \neq \mathbb{E}_x[(\hat{f}(\tilde{x}; \{x_i, y_i\}_{i \in [n]}) - f^*(\tilde{x}))^2]$$

We hope to show a result like

$$\|\widehat{f} - f^*\|_{L^2}^2 \lesssim \underbrace{\|\widehat{f} - f^*\|_n^2}_{\delta_n^2} + \varepsilon_n^2,$$

where $\varepsilon_n^2 \rightarrow 0$ as $n \rightarrow \infty$.

Today, we will show two bounds:

1. Naive bound: If we do not care about how fast $\varepsilon_n \rightarrow 0$, we can get a bound by using a global uniform bound, a global Rademacher complexity bound, and using bounded difference concentration.
2. Tighter bound: We will use
 - (a) the local uniform bound
 - (b) local Rademacher complexity
 - (c) a tighter concentration inequality, known as the Talagrand concentration inequality.

23.3 Naive bound

Let $f = \widehat{f} - f^* \in \mathcal{F}^*$. Then

$$\begin{aligned} \left| \|\widehat{f} - f^*\|_{L^2(\mathbb{P}_n)}^2 - \|\widehat{f} - f^*\|_{L^2(\mathbb{P})}^2 \right| &\leq \sup_{g \in \mathcal{F}^*} \left| \|g\|_n^2 - \|g\|_{L^2}^2 \right| \\ &= \sup_{g \in \mathcal{F}^*} \left| \frac{1}{n} \sum_{i=1}^n g(x_i)^2 - \mathbb{E}[g(x)^2] \right| \\ &=: Z \end{aligned}$$

We first try to find a bound on the expectation of Z :

$$\begin{aligned} \mathbb{E}[Z] &= \mathbb{E} \left[\sup_{g \in \mathcal{F}^*} \left| \frac{1}{n} \sum_{i=1}^n g(x_i)^2 - \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n g(\tilde{x}_i)^2 \right] \right| \right] \\ &\leq \mathbb{E} \left[\sup_{g \in \mathcal{F}^*} \left| \frac{1}{n} \sum_{i=1}^n (g(x_i)^2 - g(\tilde{x}_i)^2) \right| \right] \end{aligned}$$

Since the distribution of this is symmetric about 0,

$$= \mathbb{E} \left[\sup_{g \in \mathcal{F}^*} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon^2 (g(x_i)^2 - g(\tilde{x}_i)^2) \right| \right]$$

$$\leq 2 \mathbb{E} \left[\sup_{g \in \mathcal{F}^*} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 g(x_i)^2 \right| \right]$$

If this just had g instead of g^2 , this quantity would be the Rademacher complexity. So we want to bound this by the Rademacher complexity. Write $\phi(t) = t^2$, so

$$\leq 2 \mathbb{E} \left[\sup_{g \in \mathcal{F}^*} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 \phi(g(x_i)) \right| \right]$$

The function ϕ is $2\|\mathcal{F}^*\|_\infty$ -Lipschitz, where we can assume that $\|\mathcal{F}^*\|_\infty = 1$.

$$\leq 4 \underbrace{\mathbb{E} \left[\sup_{g \in \mathcal{F}^*} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i g(x_i) \right| \right]}_{\overline{\mathcal{R}}_n(\mathcal{F}^*) \text{ Rademacher complexity}} .$$

We can use chaining to bound this.

Now let's bound the distance from the mean. Using the bounded difference inequality, $|Z - \mathbb{E}[Z]| \sim \text{sG}(\|g\|_\infty^2/n)$, so

$$\|\hat{f} - f^*\|_{L^2}^2 \lesssim \underbrace{\|\hat{f} - f^*\|_n^2}_{\delta_n^2} + \overline{\mathcal{R}}_n(\mathcal{F}^*) + O(1/\sqrt{n}).$$

If \mathcal{F}^* is parametric with d parameters, then $\delta_n^2 \asymp \frac{d}{n}$ and $\overline{\mathcal{R}}_n(\mathcal{F}^*) \asymp \sqrt{\frac{d}{n}}$.

23.4 Using localization to get a faster rate

We will present some heuristics, rather than something completely rigorous. The rigorous treatment is in Chapter 14 of Wainwright's textbook. Suppose we already know that $\|\hat{f} - f^*\|_{L^2(\mathbb{P})} \leq r$. We can think about r decaying to 0 as $n \rightarrow \infty$. It may seem strange to assume that the L^2 norm is bounded when this is what we want to prove, but the idea is that we will get a more refined bound. So we can iterate this bound to get a nice final result

Letting $g = \hat{f} - f^* \in \mathcal{F}^*$,

$$\begin{aligned} \left| \|\hat{f} - f^*\|_{L^2(\mathbb{P}_n)}^2 - \|\hat{f} - f^*\|_{L^2(\mathbb{P})}^2 \right| &\leq \sup_{\substack{g \in \mathcal{F}^* \\ \|g\|_{L^2} \leq r}} \left| \frac{1}{n} \sum_{i=1}^n g(x_i)^2 - \mathbb{E}[g(x_i)^2] \right| \\ &=: Z(r). \end{aligned}$$

Now we bound the expectation using the same line of argument as before:

$$\mathbb{E}[Z(r)] \leq 4 \underbrace{\mathbb{E} \left[\sup_{\substack{g \in \mathcal{F}^* \\ \|g\|_{L^2} \leq r}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i g(x_i) \right| \right]}_{\overline{\mathcal{R}}_n(r; \mathcal{F}^*) \text{ localized Rademacher complexity}} .$$

We now show that $\overline{\mathcal{R}}_n(r; \mathcal{F}^*) \lesssim \varepsilon_n \cdot r$, where $\varepsilon_n = \inf\{\varepsilon : \overline{\mathcal{R}}_n(\varepsilon; \mathcal{F}^*) \leq \frac{\varepsilon^2}{16b}\}$ and $b = \sup_{g \in \mathcal{F}^*} \|g\|_\infty = 1$. This is because for any $r \geq \varepsilon_n$, $\frac{\overline{\mathcal{R}}_n(r; \mathcal{F}^*)}{r}$ is non-increasing (as long as \mathcal{F}^* is star-shaped). This tells us that

$$\frac{\overline{\mathcal{R}}_n(r; \mathcal{F}^*)}{r} \leq \frac{\overline{\mathcal{R}}_n(\varepsilon_n; \mathcal{F}^*)}{\varepsilon_n} = \frac{\varepsilon_n}{16}.$$

This means that $\overline{\mathcal{R}}_n(r; \mathcal{F}^*) \lesssim \varepsilon_n r$.

Now let's see how this implies an upper bound for the prediction error of the L^2 norm. Suppose that $Z(r) \approx \mathbb{E}[Z(r)]$ for any $r \in \mathbb{R}$ (this should be made quantitative with the Talagrand concentration inequality or a tighter concentration inequality). Then

$$\begin{aligned} \left| \underbrace{\|\hat{f} - f^*\|_{L^2(\mathbb{P}_n)}^2}_{a^2} - \underbrace{\|\hat{f} - f^*\|_{L^2(\mathbb{P})}^2}_{b^2} \right| &\lesssim \overline{\mathcal{R}}_n(r; \mathcal{F}^*) \\ &\lesssim \varepsilon_n r \\ &= \varepsilon_n \underbrace{\|\hat{f} - f^*\|_{L^2}}_b \end{aligned}$$

This is heuristic because the quantity $\|\hat{f} - f^*\|_{L^2}$ is random and depends on the training data set. However, we can use the iterative argument to make sense of this argument. We now have

$$|a^2 - b^2| \leq \varepsilon_n b \leq \frac{b^2}{4} + 4\varepsilon_n^2,$$

which gives $b^2 \lesssim a^2 + \varepsilon_n^2$. So we get that

$$\|\hat{f} - f^*\|_{L^2}^2 \lesssim \underbrace{\|\hat{f} - f^*\|_n^2}_{\delta_n^2} + \varepsilon_n^2.$$

This tells us that the upper bound of the prediction error in terms of the L^2 norm is of the same order as the upper bound of the prediction error in terms of the $L^2(\mathbb{P}_n)$ norm. If T is parametric with d parameters, then $\delta_n^2 \asymp \varepsilon_n^2 \asymp \frac{d}{n}$.

Here, our proof is different in two ways from the treatment in the textbook.

1. The first way is that we have assumed that our concentration inequality does not destroy our bound. If we just use the bounded differences inequality, we get the naive bound

$$|Z(r) - \mathbb{E}[Z(r)]| \lesssim \sqrt{\frac{1}{n}} = \eta_n.$$

The issue with this is that $Z(r)$ is $O(1/n)$ and $\mathbb{E}[Z(r)]$ is $O(1/n)$. Instead, we need to use the Talagrand inequality.

2. The second difference is that we have assumed beforehand that $\|\widehat{f} - f^*\|_{L^2(\mathbb{P})} \leq r$. The textbook instead uses a peeling argument. We actually want to find a bound on $\sup_r |Z(r) - \mathbb{E}[Z(r)]|$. To use a union bound, we need to discretize r , and a clever way to do so is to use a log scale, rather than a uniform grid.

In the end, we get the following theorem, which we state informally. This is Corollary 14.15 in the textbook.

Theorem 23.2. *Let*

$$\widehat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{2n} \sum_{i=1}^n (y_i - f(x_i))^2.$$

Then

$$\|\widehat{f} - f^*\|_{L^2} \lesssim \varepsilon_n^2 + \delta_n^2,$$

where

$$\varepsilon_n = \inf \left\{ r : \overline{\mathcal{R}}_n(r; \mathcal{F}) \lesssim \frac{r^2}{b} \right\}, \quad \delta_n = \inf \left\{ \delta : \mathcal{G}_n(\delta; \mathcal{F}^*) \lesssim \frac{\delta}{b} \right\}.$$

Here, ε_n is deterministic, as $\overline{\mathcal{R}}_n$ is averaged over $(x_i)_{i \in [n]}$. On the other hand, δ_n is random, as \mathcal{G}_n depends on $(x_i)_{i \in [n]}$.

23.5 Uniform law for Lipschitz cost function

More generally, we may want to consider cost functions which are not the squared error. Suppose we have $(x_i, y_i)_{i \in [n]} \stackrel{\text{iid}}{\sim} \mathbb{P} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ with a function class $\mathcal{F} \subseteq \{f : \mathcal{X} \rightarrow \widehat{\mathcal{Y}}\}$. Let the loss be $\mathcal{L} : \widehat{\mathcal{Y}} \times \mathcal{Y} \rightarrow \mathbb{R}$. Then we have the empirical risk

$$\mathbb{P}_n \mathcal{L}(f(x), y) = \frac{1}{n} \sum_{i=1}^n \mathcal{L}(f(x_i), y_i),$$

with empirical risk minimizer

$$\widehat{f} = \arg \min_{f \in \mathcal{F}} \mathbb{P}_n \mathcal{L}(f(x), y)$$

and population risk minimizer

$$f^* = \arg \min_f \underbrace{\mathbb{P} \mathcal{L}(f(x), y)}_{\mathbb{E}_{x,y}[\mathcal{L}(f(x), y)]}.$$

Our goal is to bound $\|\widehat{f} - f^*\|_{L^2}^2$.

We assume the loss is L -Lipschitz:

$$|\mathcal{L}(z, y) - \mathcal{L}(z', y)| \leq L|z - z'|.$$

Another assumption, which is harder to check, is that L is r -**strongly convex**: If we let $L_f(x, y) := L(f(x), y)$, then we require

$$\mathbb{P} \left(\mathcal{L}_f - L_{f^*} - \frac{\partial \mathcal{L}}{\partial z} \Big|_{f^*} (f - f^*) \right) \geq \frac{r}{2} \|f - f^*\|_{L^2}^2.$$

Example 23.1 (Logistic regression). Let $\mathcal{Y} = \{\pm 1\}$, $\mathcal{L}(\hat{y}, y) = \log(1 + e^{-2y\hat{y}})$, and

$$\mathbb{P}(y | x) = \frac{1}{1 + e^{-2yf^*(x)}}.$$

Then $\mathcal{L}(\hat{y}, y)$ is 1-Lipschitz in \hat{y} . Under mild conditions, $\mathbb{P}\mathcal{L}_f$ is r -strongly convex.

Here is Theorem 14.20 in the textbook.

Theorem 23.3. *Assume that \mathcal{F} is 1-uniformly bounded and star-shaped, with population minimizer f^* . Let $\delta_n = \inf\{\delta > \frac{c}{\sqrt{n}} : \overline{\mathcal{R}}_n(\delta; \mathcal{F}^*) \leq \delta^2\}$.*

(a) *If \mathcal{L} is L -Lipschitz in $\hat{\mathcal{Y}}$, then with high probability,*

$$\sup_{f \in \mathcal{F}} \frac{|\mathbb{P}_n(\mathcal{L}_f - \mathcal{L}_{f^*}) - \mathbb{P}(\mathcal{L}_f - \mathcal{L}_{f^*})|}{\|f - f^*\|_{L^2} + \delta_n} \leq 10L \cdot \delta_n.$$

(b) *If $\mathbb{P}\mathcal{L}_f$ is also r -strongly convex, then with high probability, for all \hat{f} such that $\mathbb{P}_n(\mathcal{L}_{\hat{f}} - \mathcal{L}_{f^*}) \leq 0$, we have*

$$\|\hat{f} - f^*\|_2^2 \leq \left(\frac{20L}{r} + 1 \right)^2 \delta_n^2$$

and

$$\mathbb{P}(\mathcal{L}_{\hat{f}} - \mathcal{L}_{f^*}) \leq 10L \left(\frac{20L}{r} + 1 \right)^2 \delta_n^2.$$

Remark 23.1. Statement (b) is a direct consequence of statement (a), using the r -strong convexity condition. The proof of (a) also relies on a local Rademacher complexity bound. We can bound $\sup_{f \in \mathcal{F}} |\mathbb{P}_n(\mathcal{L}_f - \mathcal{L}_{f^*}) - \mathbb{P}(\mathcal{L}_f - \mathcal{L}_{f^*})|$ using the Rademacher complexity, and we can get a faster rate using local Rademacher complexity.

This concludes our discussion of nonparametric function estimation. Next time, we will move on to minimax lower bounds.

24 Introduction to Minimax Lower Bounds

24.1 Minimax risk and methods of obtaining lower bounds

In the last few lectures, we were talking about upper bounds for error of statistical estimators. Now we will prove some lower bounds, which tell us that for a certain number of samples, you cannot have vanishing estimation error.

In statistical decision theory, we have a class of distributions \mathcal{P} and a parameter/function of distributions $\theta : \mathcal{P} \rightarrow \Theta$. If this is a one to one mapping, we write $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$. Then we have **statistical estimators**, which are mappings $\hat{\theta} : \mathcal{X} \rightarrow \Theta$. Suppose there is a **semimetric**¹⁵ $\rho(\theta, \theta') : \Theta \times \Theta \rightarrow \mathbb{R}$, such as

$$\rho(\theta, \theta') = \|\theta - \theta'\|_2, \quad \rho(f, f') = \|f - f'\|_{L^2}.$$

If $\Phi : [0, \infty) \rightarrow [0, \infty)$ is increasing, the **risk** is

$$R(\hat{\theta}; \theta(P)) = \mathbb{E}_{X \sim P}[\Phi(\rho(\hat{\theta}(X); \theta(P)))].$$

In this framework, the **loss function** is $\ell = \Phi \circ \rho$.

Definition 24.1. The **minimax risk** with n samples is

$$\mathcal{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) = \inf_{\hat{\theta}: \mathcal{X} \rightarrow \Theta} \sup_{P \in \mathcal{P}} R(\hat{\theta}; \theta(P))$$

The inf and the sup mean that we are taking the best estimator for the worst model.

- (a) If $R(\hat{\theta})$ achieves \mathcal{M}_n , it is good enough.
- (b) If $R(\hat{\theta}) \gg \mathcal{M}_n$, we should either find a better estimator or a sharper lower bound.

Example 24.1. Let $\Theta = \mathbb{R}^d$ with $\mathbb{P}_\theta = N(\theta, \sigma^2 I_d)$, $\theta \in \mathbb{R}^d$, where σ^2 is known. Our sample is $(x_i)_{i \in [n]} \stackrel{\text{iid}}{\sim} \mathbb{P}_\theta$, so $x_{1:n} \sim \mathbb{P}_\theta^n$. Our metric is $\rho(\theta, \theta') = \|\theta - \theta'\|_2$, and we pick $\Phi(t) = t^2$. Consider the estimator $\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n x_i$. Then

$$R(\hat{\theta}_n; P_\theta) = \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n x_i - \theta \right\|_2^2 \right] = \sigma^2 \frac{d}{n},$$

which tells us that

$$\mathcal{M}_n \leq \sigma^2 \frac{d}{n}.$$

However, we can prove the same value as a lower bound. Our goal in this lecture and the next is to show that $\mathcal{M}_n \geq c\sigma^2 \frac{d}{n}$ for some constant c .

¹⁵For a semimetric, we may allow $\theta \neq \theta'$ to still have $\rho(\theta, \theta') = 0$.

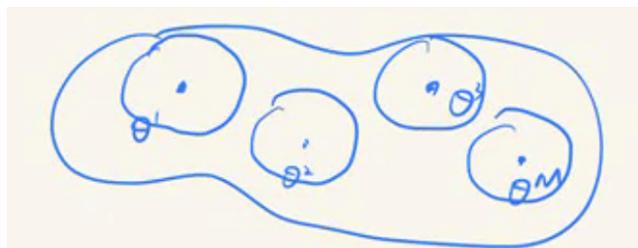
Remark 24.1. Here are some methods of showing lower bounds for estimation error, some of which we have already seen.

- (a) Bayesian decision theory: \mathcal{M}_n is the Bayes risk of the least favorable prior.
- (b) Cramer-Rao lower bound: For unbiased estimators, there is a lower bound given in terms of the Fisher information. If this does not depend on the Fisher information, then it is a minimax lower bound.
- (c) Bayes Cramer-Rao (Van-Tree's inequality): This gives a "local minimax" lower bound.
- (d) Reduction to a testing problem: We will study this now. We first need some tools from information theory.

24.2 Reduction to an M -ary testing problem

The idea is to find a testing problem easier than the estimation problem. A lower bound for the testing problem will imply a lower bound for estimation.

Step 1: Construct a 2δ -separated set of Θ in the ρ -metric.



So we require $\rho(\theta^i, \theta^j) \geq 2\delta$ for all $i \neq j$. This is the same as a packing, except we allow \geq instead of $>$. If our separated set is $\{\theta^1, \theta^2, \dots, \theta^M\}$, we get $\{\mathbb{P}_{\theta^1}, \mathbb{P}_{\theta^2}, \dots, \mathbb{P}_{\theta^M}\}$.

Step 2: Sample $(J, Z) \in [M] \times \mathcal{X}$. The joint distribution is

$$\begin{cases} J \sim \text{Unif}(\{1, 2, \dots, M\}) \\ Z \mid J = j \sim \mathbb{P}_{\theta^j}. \end{cases}$$

Step 3: Let \mathbb{Q} be the joint distribution of (J, Z) . Then the marginal distribution of Z is

$$\bar{\mathbb{Q}} = \frac{1}{M} \sum_{j=1}^M \mathbb{P}_{\theta^j}.$$

Our testing problem is that we want to find a $\psi : \mathcal{X} \rightarrow [M]$ such that $\mathbb{Q}(\psi(Z) \neq J)$ is small. If $M = 2$, this is standard binary hypothesis testing. The testing error is

$$\mathbb{Q}(\psi(Z) \neq J) = \frac{1}{2} \underbrace{[\mathbb{P}_{\theta^1}(\psi(Z) \neq 1)]}_{\text{Type I error}} + \underbrace{[\mathbb{P}_{\theta^2}(\psi(Z) \neq 2)]}_{\text{Type II error}}.$$

This is different from the traditional hypothesis testing setup in that instead of fixing the Type I error and minimizing the Type II error, we want to minimize the average of these errors.

Proposition 24.1 (From estimation to testing). *Let Ψ be increasing and $\{\theta^1, \dots, \theta^M\}$ be 2δ -separated for $\delta > 0$. Then*

$$\mathcal{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta) \inf_{\psi} \mathbb{Q}(\psi(Z) \neq J).$$

This works for all $\delta > 0$, so we can pick the δ which gives the best lower bound. In general, $\Phi(\delta)$ is increasing with δ , but the testing error $\inf_{\psi} \mathbb{Q}(\psi(Z) \neq J)$ is decreasing with δ . We can choose $\delta = \delta_n$ such that $\inf_{\psi} \mathbb{Q}(\psi(Z) \neq J) = \frac{1}{2}$; any constant would work here. Then the minimax lower bound will be

$$\mathcal{M}_n \geq \frac{1}{2} \Phi(\delta_n).$$

Proof. Fix P and $\hat{\theta}$. By Markov's inequality,

$$\begin{aligned} \mathbb{E}[\Phi(\rho(\hat{\theta}, \theta))] &\geq \Phi(\delta) \mathbb{P}(\Phi(\rho(\hat{\theta}, \theta)) \geq \Phi(\delta)) \\ &= \Phi(\delta) \mathbb{P}(\rho(\hat{\theta}, \theta) \geq \delta). \end{aligned}$$

We now want to relate this probability with the testing error. We have

$$\begin{aligned} \sup_{P \in \mathcal{P}} \mathbb{P}(\rho(\hat{\theta}, \theta) \geq \delta) &\geq \sup_{\theta \in \{\theta^1, \dots, \theta^M\}} \mathbb{P}_{\theta}(\rho(\hat{\theta}, \theta) \geq \delta) \\ &\geq \frac{1}{M} \sum_{j=1}^M \mathbb{P}_{\theta^j}(\rho(\hat{\theta}, \theta^j) \geq \delta) \\ &= \mathbb{Q}(\rho(\hat{\theta}, \theta^J) \geq \delta). \end{aligned}$$

Define a test ψ via $\hat{\theta}$: Let

$$\psi(z) = \arg \min_{L \in [M]} \rho(\hat{\theta}(Z), \theta^L).$$

This gives the θ^j which is the closest to our estimate $\hat{\theta}(Z)$. With this definition,

$$\{\psi(Z) \neq J\} \subseteq \{\rho(\hat{\theta}(Z), \theta^J) \geq \delta\}.$$

This means we can lower bound the above \mathbb{Q} probability:

$$\inf_{\hat{\theta}} \mathbb{Q}(\rho(\hat{\theta}(Z), \theta^J) \geq \delta) \geq \inf_{\psi} \mathbb{Q}(\psi(Z) \neq J). \quad \square$$

How do we choose $\{\theta^1, \dots, \theta^M\}$? Moreover, how do we lower bound $\inf_{\psi} \mathbb{Q}(\psi(Z) \neq J)$? Here are two general methods.

1. $M = 2$: Le Cam's method

- Two points method
- Convex hull method

2. $M \geq 3$:

- Assoaud's method
- Fano's method

Le Cam's method is the most classical one, so we will start with it. Fano's method is the most important and useful method for high-dimensional models.

24.3 Some divergence measures

Here are some basic tools for these methods. Let \mathbb{P}, \mathbb{Q} be two probability distributions on \mathcal{X} . How can we measure their distance?

Definition 24.2. The **total variation distance** is

$$\|\mathbb{P} - \mathbb{Q}\|_{\text{TV}} := \sup_{A \subseteq \mathcal{X}} |\mathbb{P}(A) - \mathbb{Q}(A)| = \frac{1}{2} \int_{\mathcal{X}} |p(x) - q(x)| d\mu(x),$$

where p, q are the densities of \mathbb{P}, \mathbb{Q} , if they exist.

Definition 24.3. The **Kullback-Leibler divergence** is

$$D(\mathbb{Q} \parallel \mathbb{P}) := \int_{\mathcal{X}} q(x) \log \frac{q(x)}{p(x)} d\lambda(x).$$

There is a more general definition of the K-L divergence that does not require \mathbb{Q}, \mathbb{P} to have densities with respect to Lebesgue measure. This is not a distance because $D(\mathbb{Q} \parallel \mathbb{P}) \neq D(\mathbb{P} \parallel \mathbb{Q})$, but it has distance-like properties, such as $D(\mathbb{P} \parallel \mathbb{Q}) \geq 0$ with $D(\mathbb{P} \parallel \mathbb{Q}) = 0$ iff $\mathbb{P} = \mathbb{Q}$.

Definition 24.4. The **Hellinger distance** is

$$\mathbb{H}^2(\mathbb{P} \parallel \mathbb{Q}) := \int_{\mathcal{X}} (\sqrt{p(x)} - \sqrt{q(x)})^2 d\nu(x).$$

Here are some relationships between these notions of distance:

Proposition 24.2 (Pinsker's inequality).

$$\|\mathbb{P} - \mathbb{Q}\|_{\text{TV}} \leq \sqrt{\frac{1}{2} D(\mathbb{P} \parallel \mathbb{Q})}.$$

Proposition 24.3 (Le Cam's inequality).

$$\|\mathbb{P} - \mathbb{Q}\|_{\text{TV}} \leq \sqrt{\mathbb{H}^2(\mathbb{P} \parallel \mathbb{Q})} \underbrace{\sqrt{1 - \frac{\mathbb{H}^2(\mathbb{P} \parallel \mathbb{Q})}{4}}}_{\leq 1}.$$

Proposition 24.4.

$$\mathbb{H}^2(\mathbb{P} \parallel \mathbb{Q}) \leq \frac{1}{2} D(\mathbb{P} \parallel \mathbb{Q}).$$

We will see that the TV distance is related to the testing error for a binary testing situation. On the other hand, the KL-divergence and Hellinger distance have good tensorization properties: If we let

$$\mathbb{P}^{1:n} = \mathbb{P}_1 \times \mathbb{P}_2 \times \cdots \times \mathbb{P}_n, \quad \mathbb{Q}^{1:n} = \mathbb{Q}_1 \times \mathbb{Q}_2 \times \cdots \times \mathbb{Q}_n,$$

then

$$\begin{aligned} D(\mathbb{P}^{1:n} \parallel \mathbb{Q}^{1:n}) &= \sum_{i=1}^n D(\mathbb{P}_i \parallel \mathbb{Q}_i), \\ \frac{1}{2} \mathbb{H}^2(\mathbb{P}^{1:n} \parallel \mathbb{Q}^{1:n}) &= 1 - \prod_{i=1}^n \left(1 - \frac{1}{2} \mathbb{H}^2(\mathbb{P}_i \parallel \mathbb{Q}_i)\right). \end{aligned}$$

Example 24.2 (Gaussian distribution). For a Gaussian distribution, we have the density

$$p_\theta = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right), \quad \theta \in \mathbb{R}.$$

The K-L divergence is

$$\begin{aligned} D(\mathbb{P}_\theta \parallel \mathbb{P}_{\theta'}) &= \int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right) \log \frac{\exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right)}{\exp\left(-\frac{(x-\theta')^2}{2\sigma^2}\right)} dx \\ &= \mathbb{E}_{X \sim \mathbb{P}_\theta} \left[-\frac{(X-\theta)^2}{2\sigma^2} + \frac{(X-\theta')^2}{2\sigma^2} \right] \\ &= \frac{(\theta')^2}{2\sigma^2} - \frac{\theta^2}{2\sigma^2} + \frac{1}{\sigma^2} \mathbb{E}_{X \sim \mathbb{P}_\theta}[(\theta - \theta')X] \\ &= \frac{(\theta')^2}{2\sigma^2} - \frac{\theta^2}{2\sigma^2} + \frac{1}{\sigma^2}(\theta - \theta')\theta \\ &= \frac{(\theta - \theta')^2}{2\sigma^2}. \end{aligned}$$

Using Pinsker's inequality and the tensorization property of the K-L divergence,

$$\|\mathbb{P}_\theta^n - \mathbb{P}_{\theta'}^n\|_{\text{TV}} \leq \sqrt{\frac{1}{2} D(\mathbb{P}_\theta^n \parallel \mathbb{P}_{\theta'}^n)}$$

$$\begin{aligned} &\leq \sqrt{\frac{n}{2} D(\mathbb{P}_\theta \parallel \mathbb{P}_{\theta'})} \\ &\leq \sqrt{\frac{n(\theta - \theta')^2}{4\sigma^2}}. \end{aligned}$$

We can also calculate the Hellinger distance

$$\mathbb{H}^2(\mathbb{P}_\theta \parallel \mathbb{P}_{\theta'}) = 1 - \exp\left(-\frac{(\theta - \theta')^2}{8\sigma^2}\right).$$

More generally, for $\theta \in \mathbb{R}^d$ and $\mathbb{P}_\theta = N(\theta, \sigma^2 I_d)$, we get

$$D(p_\theta \parallel p_{\theta'}) = \frac{\|\theta - \theta'\|_2^2}{2\sigma^2},$$

$$\mathbb{H}^2(\mathbb{P}_\theta \parallel \mathbb{P}_{\theta'}) = 1 - \exp\left(-\frac{\|\theta - \theta'\|_2^2}{8\sigma^2}\right).$$

25 Methods for Proving Minimax Lower Bounds

25.1 Recap: Testing lemma and divergence measures for minimax lower bounds

We have been studying minimax lower bounds. We have a semi-meric $\rho : \Theta \times \Theta \rightarrow \mathbb{R}_{\geq 0}$ and a 2δ -separated set $\{\theta^1, \dots, \theta^M\} \subseteq \Theta$. In our testing situation, we have the joint distribution

$$Q : \begin{cases} J \sim \text{Unif}(\{1, 2, \dots, M\}) \\ Z \mid J = j \sim \mathbb{P}_{\theta^j}. \end{cases}$$

We have an increasing function Φ , as well. We proved the following result:

Proposition 25.1 (From estimation to testing). *Let Ψ be increasing and $\{\theta^1, \dots, \theta^M\}$ be 2δ -separated for $\delta > 0$. Then*

$$\mathcal{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta) \inf_{\psi} \mathbb{Q}(\psi(Z) \neq J).$$

We also defined the total variation distance the K-L divergence, and the Hellinger distance

$$\|\mathbb{P} - \mathbb{Q}\|_{\text{TV}} = \frac{1}{2} \int_{\mathcal{X}} |p(x) - q(x)| dx,$$

$$D(\mathbb{P} \parallel \mathbb{Q}) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} dx,$$

$$\mathbb{H}^2(\mathbb{P} \parallel \mathbb{Q}) = \int_{\mathcal{X}} (\sqrt{p(x)} - \sqrt{q(x)})^2 dx.$$

These had the following relationships:

$$\|\mathbb{P} - \mathbb{Q}\|_{\text{TV}} \leq \sqrt{\frac{1}{2} D(\mathbb{P} \parallel \mathbb{Q})},$$

$$\|\mathbb{P} - \mathbb{Q}\|_{\text{TV}} \leq \sqrt{\mathbb{H}^2(\mathbb{P} \parallel \mathbb{Q})} \underbrace{\sqrt{1 - \frac{\mathbb{H}^2(\mathbb{P} \parallel \mathbb{Q})}{4}}}_{\leq 1}.$$

$$\mathbb{H}^2(\mathbb{P} \parallel \mathbb{Q}) \leq \frac{1}{2} D(\mathbb{P} \parallel \mathbb{Q}).$$

25.2 Le Cam's two points method

Take $M = 2$. Then $J \sim \text{Unif}(\{0, 1\})$, and $Z \mid J = j \sim \mathbb{P}_j$, and $\bar{\mathbb{Q}} = \frac{1}{2}\mathbb{P}_0 + \frac{1}{2}\mathbb{P}_1$. We claim that

$$\inf_{\psi} \mathbb{Q}(\psi(Z) \neq J) = \frac{1}{2}(1 - \|\mathbb{P}_0 - \mathbb{P}_1\|_{\text{TV}}).$$

Proof. For any ψ , we can find an A such that

$$\psi(x) = \begin{cases} 1 & x \in A \\ 0 & x \in A^c. \end{cases}$$

Then

$$\begin{aligned} \mathbb{Q}(\psi(Z) = J) &= \frac{1}{2}\mathbb{P}_1(A) + \frac{1}{2}\mathbb{P}_0(A^c) \\ &= \frac{1}{2}(\mathbb{P}_1(A) - \mathbb{P}_0(A)) + \frac{1}{2}. \end{aligned}$$

If we take the supremum over all ψ , we get

$$\begin{aligned} \sup_{\psi} \mathbb{Q}(\psi(Z) = J) &= \sup_A \frac{1}{2}(\mathbb{P}_1(A) - \mathbb{P}_0(A)) + \frac{1}{2} \\ &= \frac{1}{2}\|\mathbb{P}_1 - \mathbb{P}_0\|_{\text{TV}} + \frac{1}{2} \end{aligned}$$

The probability of the bad event is then

$$\inf_{\psi} \mathbb{Q}(\psi(Z) \neq J) = \frac{1}{2} - \frac{1}{2}\|\mathbb{P}_1 - \mathbb{P}_0\|_{\text{TV}}. \quad \square$$

This gives the following theorem.

Theorem 25.1 (Le Cam's two points lower bound). *For all $\delta > 0$ and $\mathbb{P}_0, \mathbb{P}_1 \in \mathcal{P}$ with $\rho(\theta(\mathbb{P}_0), \theta(\mathbb{P}_1)) \geq 2\delta$,*

$$\mathcal{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \frac{\Phi(\delta)}{2}(1 - \|\mathbb{P}_1 - \mathbb{P}_0\|_{\text{TV}}).$$

For the generalization to Le Cam's convex hull method, read chapter 15.2.2 in Wainwright's textbook.

Example 25.1 (Gaussian location family, $d = 1$). Our model is $\mathcal{P} = \{\mathbb{P}_{\theta} = N(\theta, \sigma^2) : \theta \in \mathbb{R}\}$, where σ is known. We have the semimetric $\rho(\theta', \theta) = |\theta' - \theta|$ and $\Phi(t) = t^2$. Our sample is $X_{1:n} \sim \mathbb{P}_{\theta}^n$. The true minimax risk is $\mathcal{M}_n = \frac{\sigma^2}{n}$. Here is a lower bound by Le Cam's method:

Consider $\mathbb{P}_{2\delta}$ and \mathbb{P}_0 , so $\rho(2\delta, 0) \geq 2\delta$. Then

$$\mathcal{M}_n(\theta(\mathcal{P}); |\theta - \theta'|^2) \geq \frac{\delta^2}{2}(1 - \|\mathbb{P}_{2\delta}^n - \mathbb{P}_0^n\|_{\text{TV}}),$$

where the n only appears in the bound as the fact that the measures are product measures. We want to lower bound $1 - \|\mathbb{P}_{2\delta}^n - \mathbb{P}_0^n\|_{\text{TV}}$ by $1/2$. We have by Pinsker's inequality and the tensorization property of K-L divergence

$$\|\mathbb{P}_{2\delta}^n - \mathbb{P}_0^n\|_{\text{TV}}^2 \leq \frac{1}{2}D(\mathbb{P}_{2\delta}^n \parallel \mathbb{P}_0^n)$$

$$\begin{aligned}
&= \frac{1}{2}nD(\mathbb{P}_{2\delta} \parallel \mathbb{P}_0) \\
&= \frac{1}{2}n \frac{(2\delta)^2}{2\sigma^2} \\
&= \frac{n\delta^2}{\sigma^2}.
\end{aligned}$$

Now choose $\frac{n\delta_n^2}{\sigma^2} = \frac{1}{2}$, so $\delta_n^2 = \frac{\sigma^2}{2n}$. Then $\|\mathbb{P}_{2\delta_n}^n - \mathbb{P}_0^n\|_{\text{TV}} \leq \frac{1}{2}$, and we get the minimax lower bound

$$\mathcal{M}_n \geq \frac{\delta_n^2}{2} \cdot \frac{1}{2} = \frac{\sigma^2}{16n}.$$

Up to constants, this is sharp.

Here is the problem with Le Cam's method. If we take $\theta \in \mathbb{R}^d$ with $\mathbb{P}_\theta = N(\theta, \sigma^2 I_d)$ for $d \geq 2$, then we will get the lower bound

$$\mathcal{M}_n \geq \frac{\sigma^2}{16n},$$

even though the actual minimax risk is $\mathcal{M}_n = \sigma^2 \frac{d}{n}$.

25.3 Mutual information

Here, we will develop some tools for Fano's method, which is a sharper method for lower bounding the minimax risk. Suppose we have two random variables $(X, Y) \sim \mathbb{P}_{X,Y}$. We want a measure of their dependence/independence (not the same as correlation). If X is independent of Y , we have

$$\mathbb{P}_{X,Y} = \mathbb{P}_X \times \mathbb{P}_Y = \int_{\mathcal{Y}} \mathbb{P}_{X,Y}(x, y) dy \times \int_{\mathcal{X}} P_{X,Y}(x, y) dx.$$

To get a measure of independence, we should look at the distance between these two objects:

$$D\left(\mathbb{P}_{X,Y}, \int_{\mathcal{Y}} \mathbb{P}_{X,Y}(x, y) dy \times \int_{\mathcal{X}} P_{X,Y}(x, y) dx\right).$$

Definition 25.1. The **mutual information** between X and Y is

$$I(X; Y) := D(\mathbb{P}_{X,Y} \parallel \mathbb{P}_X \times \mathbb{P}_Y).$$

Remark 25.1. The mutual information is always ≥ 0 . Although the K-L divergence is not symmetric, we have $I(X; Y) = I(Y; X)$.

If X and Y are independent, $I(X; Y) = 0$, and if $Y = f(X)$, the mutual information is maximized.

Recall that

$$Q : \begin{cases} J \sim \text{Unif}(\{1, 2, \dots, M\}) \\ Z | J = j \sim \mathbb{P}_{\theta^j}. \end{cases}$$

Then

$$\begin{aligned} I(J; Z) &= D(\mathbb{Q}_{2,J} \| \mathbb{Q}_2 \times \mathbb{Q}_J) \\ &= \frac{1}{M} \sum_{j=1}^M D(\mathbb{P}_{\theta^j} \| \text{ba}\mathbb{Q}), \end{aligned}$$

where

$$\bar{Q} = \frac{1}{M} \sum_{j=1}^M \mathbb{P}_{\theta^j}.$$

Suppose $\theta^j = \theta$ for all j . Then $I(J; Z) = 0$. Conversely, if the θ^j are far away from each other, then $I(J; Z)$ will be large.

Here are two upper bounds of $I(J; Z)$ we will now prove:

Proposition 25.2.

$$I(J; Z) \leq \frac{1}{M^2} \sum_{j,k=1}^M D(\mathbb{P}_{\theta^j} \| \mathbb{P}_{\theta^k}) \leq \max_{j,k} D(\mathbb{P}_{\theta^j} \| \mathbb{P}_{\theta^k}).$$

Lemma 25.1 (Yang-Barron's bound). *Let $N_{\text{KL}}(\varepsilon; \mathcal{P})$ be an ε -cover of \mathcal{P} in $\sqrt{D_{\text{KL}}}$. Then*

$$I(Z; J) \leq \inf_{\varepsilon > 0} \varepsilon^2 + \log N_{\text{KL}}(\varepsilon; \mathcal{P})$$

25.4 Fano's inequality

Let

$$Q : \begin{cases} J \sim \text{Unif}(\{1, 2, \dots, M\}) \\ Z | J = j \sim \mathbb{P}_{\theta^j}. \end{cases}$$

Lemma 25.2.

$$\inf_{\psi} \mathbb{Q}(\psi(Z) \neq J) \geq 1 - \frac{I(Z; J) + \log 2}{\log M}.$$

The proof is in Section 15.4 and requires some ideas such as the entropy. This does not require any restriction on the \mathbb{P}_{θ^j} . This lower bound gives us

Proposition 25.3. Let $\{\theta^1, \dots, \theta^M\}$ be 2δ -separated in the semimetric ρ . Then

$$\mathcal{M}_n(\theta(\mathcal{P}); \Phi \circ \rho) \geq \Phi(\delta) \left(1 - \frac{I(Z; J) + \log 2}{\log M} \right).$$

When using this lower bound, we will find δ_n such that

$$1 - \frac{I(Z; J) + \log 2}{\log M} \geq \frac{1}{2}.$$

Then we will get

$$\mathcal{M}_n \geq \frac{1}{2} \Phi(\delta_n).$$

So we need to upper bound $I(Z; J)$.

A simple upper bound is given by

$$\begin{aligned} I(J; Z) &= \frac{1}{M} \sum_{j=1}^M D(\mathbb{P}_{\theta^j} \parallel \frac{1}{M} \sum_{\ell=1}^M \mathbb{P}_{\theta^\ell}) \\ &\leq \frac{1}{M^2} \sum_{j,\ell=1}^M D(\mathbb{P}_{\theta^j} \parallel \mathbb{P}_{\theta^\ell}) \end{aligned}$$

Where we have used Jensen's inequality to show that the K-L divergence is convex in the second argument.

$$\leq \max_{j,\ell} D(\mathbb{P}_{\theta^j} \parallel \mathbb{P}_{\theta^\ell})$$

Example 25.2 (Gaussian location family, $d \geq 2$). Our model is $\mathcal{P} = \{\mathbb{P}_\theta = n(\theta, \sigma^2 I_d) : \theta \in \mathbb{R}^d\}$, where σ is known. Our semimetric is $\rho(\theta', \theta) = \|\theta' - \theta\|_2$ with $\Phi(t) = t^2$. The true minimax risk is

$$\mathcal{M}_n = \inf_{\hat{\theta}} \sup_{\theta} \mathbb{E}[\|\hat{\theta} - \theta\|_2^2] = \sigma^2 \frac{d}{n}.$$

The lower bound by Fano's method gives

$$\begin{aligned} \mathcal{M}_n &\geq \Phi(\delta) \left(1 - \frac{I(Z; J) + \log 2}{\log M} \right) \\ &\geq \Phi(\delta) \left(1 - \frac{\max_{j,k} D(\mathbb{P}_{\theta^j}^n \parallel \mathbb{P}_{\theta^k}^n) + \log 2}{\log M} \right) \end{aligned}$$

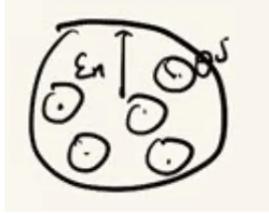
Our goal is to find the largest $\delta_n, M, \{\theta^1, \dots, \theta^M\}$ such that

(a) $\|\theta^j - \theta^k\|_2 \geq 2\delta_n$

(b)

$$\frac{\max_{j,k} D(\mathbb{P}_{\theta^j}^n \parallel \mathbb{P}_{\theta^k}^n) + \log 2}{\log M} \leq \frac{1}{2}.$$

Here is our construction: Let $\varepsilon_n = \sigma\sqrt{\frac{d}{n}}$ and $\delta_n = \frac{1}{100}\varepsilon_n = \frac{1}{100}\sigma\sqrt{\frac{d}{n}}$. Let $\{\theta^1, \dots, \theta_M\}$ be a maximal $2\delta_n$ packing of $B(0, \varepsilon_n) = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq \varepsilon_n\}$.



By a volume argument, we can get upper and lower bounds of M :

$$\log M \asymp d \log \left(c \frac{\varepsilon_n}{\delta_n} \right) \asymp c \cdot d.$$

To upper bound the K-L divergence on top, we have

$$\begin{aligned} \max_{j,k} D(\mathbb{P}_{\theta^j}^n \parallel \mathbb{P}_{\theta^k}^n) &= n \max_{j,k} D(\mathbb{P}_{\theta^j} \parallel \mathbb{P}_{\theta^k}) \\ &= n \max_{j,k} \frac{n \|\theta^j - \theta^k\|_2^2}{2\sigma^2} \\ &\leq \frac{n\varepsilon_n^2}{2\sigma^2} \\ &= c \cdot d \end{aligned}$$

Our quantities only depend on the ratio between ε_n and δ_n , so we can adjust the constant in front of δ_n to get the desired upper bound of $\frac{1}{2}$.

We then get

$$\mathcal{M}_n \geq \Phi(\delta_n) \frac{1}{2} = \frac{1}{2} \cdot \left(\frac{1}{100} \right)^2 \sigma^2 \frac{d}{n} = c\sigma^2 \frac{d}{n}.$$

25.5 Yang-Barron's method

The bound on $I(J; Z)$ by the max of the K-L divergences is generally only good when we have a parametric problem. For nonparametric problems, we want to use a better bound.

Lemma 25.3 (Yang-Barron's bound). *Let $N_{\text{KL}}(\varepsilon; \mathcal{P})$ be an ε -cover of \mathcal{P} in $\sqrt{D_{\text{KL}}}$. Then*

$$I(Z; J) \leq \inf_{\varepsilon > 0} \varepsilon^2 + \log N_{\text{KL}}(\varepsilon; \mathcal{P})$$

To apply this bound, we have two steps:

1. Choose $\varepsilon_n > 0$ such that

$$\varepsilon_n^2 \geq \log N_{\text{KL}}(\varepsilon_n; \mathcal{P}).$$

2. Choose the largest $\delta_n > 0$ such that

$$\log M(\delta_n; \rho, \Omega) \geq 4\varepsilon_n^2 + 2 \log 2.$$