Lecture notes on: Information-theoretic methods for high-dimensional statistics[*]

Yihong Wu[†]

January 14, 2020

[*] Work in progress and apologies for many mistakes. My sincere gratitude to the students at University of Illinois who took ECE598YW in Spring 2016 and contributed to scribing the initial version of the lecture notes.

[†] Department of Statistics and Data Science, Yale University, New Haven, USA, yihong.wu@yale.edu

Contents

Part I

Introductions to statistical decision theory

Outline:

- Introduction of the framework

  Parametric model vs. non-parametric model

- Best estimator

1.1 Basics of Statistical Decision Theory

- **Statistical Experiment:** A collection of probability distributions (over a common measurable space $(\mathcal{X}, \mathcal{F})$).

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\}$$

- **Data:**

$$X \sim P_\theta \qquad \text{for some } \theta \in \Theta$$

  X could be a random variable, vector, process, etc, depending on $\mathcal{X}$.

- **Objective:**

$$\begin{aligned} T : \ & \Theta \to \mathcal{Y} \\ & \theta \mapsto T(\theta) \end{aligned}$$

  The value $T(\theta)$ is what we want to estimate, which can be θ itself, or a relevant aspect of θ, e.g., a function of θ such as its norm $\|\theta\|$.

- **Estimator (Decision Rule):**

$$\hat{T} : \mathcal{X} \to \hat{\mathcal{Y}}$$

  Note the that $\hat{\mathcal{Y}}$ need not be the same as $\mathcal{Y}$.

  Remark 1.1. $\hat{T}$ can be a deterministic or randomized estimator:

  - deterministic estimator: $\hat{T} = \hat{T}(X)$.
  - randomized estimator: $\hat{T} = \hat{T}(X, U)$, where U denotes external randomness independent of X. In this case $\hat{T}$ should be viewed as a conditional probability distribution $P_{\hat{T}|X}$ (Markov transition kernel).

The problem in statistical experiment is the following game: By choosing the parameter θ, nature picks a distribution that generates the data X. The statistician observes the data and computes an estimation $\hat{T}$ of $T(\theta)$. The goal is for $\hat{T}$ to be close to T. To that end, we need to introduce a metric to quantify how good $\hat{T}$ is:

- **Loss Function:**

$$\begin{aligned} l :&\mathcal{Y} \times \hat{\mathcal{Y}} \to \mathbb{R} \\ &T \times \hat{T} \mapsto l(T, \hat{T}) \end{aligned}$$

  Since we are dealing with loss, all the negative (converse) results are lower bound and all the positive (achievable) results are upper bound.

  Note: Since X is a random variable, the estimator is also a random variable. Hence, $l(T, \hat{T})$ is a random variable. Therefore, to make sense of "minimizing the loss", we define the following:

- **Risk:**

$$R_\theta(\hat{T}) = \mathbb{E}_\theta[l(T, \hat{T})] = \int P_\theta(dx) P_{\hat{T}|X}(d\hat{t}|x) l(t(\theta), \hat{t}),$$

  which we refer to as the risk of $\hat{T}$ at θ. Note that the expected risk depends on the strategy as well as where the truth is. The subscript indicates the distribution with respect to which the expectation is taken.

The following diagram summarizes the process:

$$\begin{tikzcd}
\textit{param.} \arrow[r] & \theta \arrow[r, "P_\theta"] \arrow[d] & X \arrow[d] & \textit{data} \arrow[l] \\
& T \arrow[r, dotted, dash, "loss"'] & \hat{T} &
\end{tikzcd}$$

Example 1.1.
Gaussian Location Model (GLM): or Normal Mean Model, Additive Gaussian-Noise Channel

– Model:

$$\mathcal{P} = \{\mathcal{N}(\theta, I_p) : \theta \in \Theta\}$$

  where I_p is the p-dimensional identity matrix and $\Theta \subset \mathbb{R}^p$. Equivalently,

$$X = \theta + Z \qquad Z \sim \mathcal{N}(0, I_p), \theta \in \Theta \subset \mathbb{R}^p.$$

  - $p = 1$: scalar case
  - $p > 1$: vector case
    We also encompass matrix case: By arranging a p^2-dimensional vector into a $p \times p$ matrix. In this case $\Theta \subset \mathbb{R}^{p \times p}$.

– Objective: Examples of the objective include $T(\theta) = \theta$, $\|\theta\|_2$, $\theta_{max} = \max_{i \in [p]} \theta_i$, where $[p] = \{1, \cdots, p\}$.

– Loss function: Examples of the loss function include the following:

$$\begin{aligned} &l(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_2^2, \|\theta - \hat{\theta}\|_1, \cdots \\ &\text{In the matrix case} : l(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_F^2, \|\theta - \hat{\theta}\|_{op}, \cdots \end{aligned}$$

– Estimator: Examples of the estimator include the following:

$$\text{Maximum Likelihood Estimator: } \hat{\theta} = X$$
$$\text{James-Stein estimator: } \hat{\theta}_{JS} = \left(1 - \frac{p-2}{\|X\|_2^2}\right) X$$

The choice of the estimator mainly depends on the objective.

– Parameter space: Examples of the parameter space include the following:

a) $\Theta = \mathbb{R}^p$: unstructured.

b) $\Theta = \{\text{all } k\text{-sparse vectors}\} = \{\theta \in \mathbb{R}^p : \|\theta\|_0 \leq k\}$, where $\|\theta\|_0 \triangleq |\{i : \theta_i \neq 0\}|$ denotes the size of the support.

$\Theta = l_q$-norm balls, $0 \leq q \leq \infty$, where $\|\theta\|_q = (\sum |\theta_i|^q)^{\frac{1}{q}}$.

c) Matrix case: low-rank matrices: $\Theta = \{\theta : rank(\theta) \leq r\}$.

Note that by definition, more structure (smaller paramater space) always leads to smaller risk; but it need not simplify the computation issue.

– Testing: We have two scenarios and based on the observed data X, we want to determine which one is the true scenario.

∗ Simple Hypothesis:

$$\begin{aligned} H_0 : &\quad \theta = \theta_0 \\ H_1 : &\quad \theta = \theta_1 \end{aligned}$$

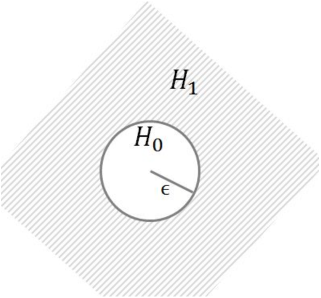
For instance θ_0 could be the all zero vector and θ_1 could be all one vector. Then this corresponds to sending a single bit repeatedly in Gaussian noise.

parameter space $= \Theta = \{\theta_0, \theta_1\} = \hat{\Theta} =$ decision space

$l(\theta, \hat{\theta}) = 1_{\{\theta \neq \hat{\theta}\}}$: This is Hamming loss (zero-one loss).

∗ Composite Hypothesis:
Example 1: One of the hypothesis is composite.

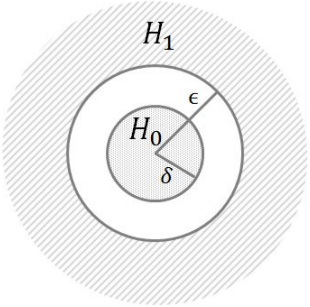
$$\begin{aligned} H_0 : &\quad \theta = 0 \\ H_1 : &\quad \|\theta\|_2 \geq \epsilon \end{aligned}$$



Here, H_0 and H_1 could be interpreted as pure noise case and the case where signal is present, respectively.

$\Theta = \{0\} \cup \{\theta : \|\theta\|_2 \geq \epsilon\}$

Example 2: Both hypothesis are composite.



$$\begin{aligned} H_0: &\quad \|\theta\|_2 \le \delta \\ H_1: &\quad \|\theta\|_2 \ge \epsilon \end{aligned}$$

Here, H_0 and H_1 could be interpreted as the case with weak signal and strong signal, respectively.

Remark 1.2 (Parametric model versus non-parametric model)**.** According to statistical conventions, parametric model refers to the case that the parameter of interest is finite-dimensional while non-parametric model refers to the case that the parameter is infinite-dimensional.

In this class, we are mostly interested in high-dimensional parametric model.

Parametric Model

Examples of parametric model:

- GLM or more generally exponential family. We start with distribution P on $\mathbb{R}^p$, and for $\theta \in \mathbb{R}^p$, consider the tilted distribution

  $$dP_\theta = \frac{e^{\langle \theta, X \rangle}}{\int P(dx) e^{\langle \theta, x \rangle}} dP$$

- Covariance matrix estimation:

  $X = (X_1, \cdots, X_n) \overset{iid}{\sim} \mathcal{N}(0, \Sigma)$. In this case, Σ is our parameter and $P_\theta = \mathcal{N}(0, \Sigma)^{\otimes n}$.

  If we want to estimate Σ, we can use the loss function $l(\Sigma, \hat{\Sigma}) = \|\Sigma - \hat{\Sigma}\|$.

  If we want to estimate a function of Σ, $T : \Sigma \to v$ (principle component) we can use loss function $l(v, \hat{v}) = d(\text{span}(v), \text{span}(\hat{v})\| = \|vv' - \hat{v}\hat{v}'\|$.

- Stochastic block model: We observe the graph G of n vertices bisected partitioned into two communities: C and C^c. So

  $$\begin{aligned} &X = G \\ &\Theta = \{C : C \subset [n], |C| = n/2\} \end{aligned}$$

  Given the partition, the graph is generated as follows: for nodes i and j, the probability that i is connected to j depends on whether they belong to the same partition or not, namely,

  $$P(i \sim j) = \begin{cases} p & \text{if } i, j \in C \text{ } or \text{ } C^c \\ q & o.w. \end{cases}$$

  Given the graph G, the goal is to estimate the communities with respect to the following loss function:

  $$l(C, \hat{C}) = 1_{\{C \neq \hat{C}\}} \text{ } or \text{ } |C \Delta \hat{C}|$$

- Large alphabet: Estimating a discrete distribution.

$$\begin{aligned}&\mathcal{P} = \{\text{all distributions on } [k]\}\\&X = (X_1, \cdots, X_n) \sim P \in \mathcal{P}\\&l(P, \hat{P}) = \|P - \hat{P}\| \ or \ D(P\|\hat{P})\end{aligned}$$

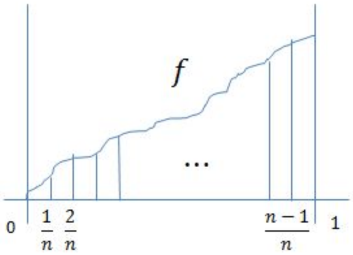
Non-parametric Model

Examples of non-parametric model:

- Density estimation: Here the parameter is a pdf, for example:

$$\begin{aligned}&f \in \mathcal{F} = \{\text{smooth, log concave, monotone}\}\\&X = (X_1, \cdots, X_n) \overset{iid}{\sim} f \ on \ \mathbb{R}^p\\&l(f, \hat{f}) = \|f - \hat{f}\|_2^2\end{aligned}$$

- Regression: We observe noisy samples at discrete points. The parameter is the unknown function f.



$$X = (X_i) = f(i/n) + Z_i$$

- White Gaussian noise model: we observe a wave form:

$$\begin{aligned}dX_t &= f(t)dt + dB_t\\X_t &= \int_0^t f(\tau)d\tau + B_t\end{aligned}$$

where B_t is a Brownian motion. Equivalently, if $f \in L^2$ where $\{\phi_i\}$ is an orthonormal basis, then

$$X_i = \langle X, \phi_i \rangle = \theta_i + Z_i \qquad i = 0, 1, \cdots$$

This is called Gaussian Sequence Model (which is GLM with $p = \infty$).

Remark 1.3. Testing:

simple vs. simple

$$H_0 : \theta = \theta_0 \ vs. \ H_1 : \theta = \theta_1 \qquad \Theta = \{\theta_0, \theta_1\}$$

simple vs. composite

$$H_0 : \theta = \theta_0 \ vs. \ H_1 : \theta \in \Theta_1 \qquad \Theta = \{\theta_0\} \cup \Theta_1$$

composite vs. composite

$$H_0 : \theta \in \Theta_0 \; vs. \; H_1 : \theta \in \Theta_1 \qquad \Theta = \Theta_0 \cup \Theta_1$$

$\hat{T}(X) \in \{0, 1\} \qquad l(\theta, \hat{T}) = 1_{\{\theta \notin \Theta_{\hat{T}}\}}.$

Remark 1.4. Confidence interval/region/bond: For example to estimate a function we output a region in which the function lies w.h.p.

$$\hat{T} = \text{some subset}$$
$$l(\theta, \hat{T}) = 1_{\{\theta \notin \hat{T}\}} + \text{size of } \hat{T}$$

Remark 1.5. We frequently deal with *independent sampling model.* In this case:

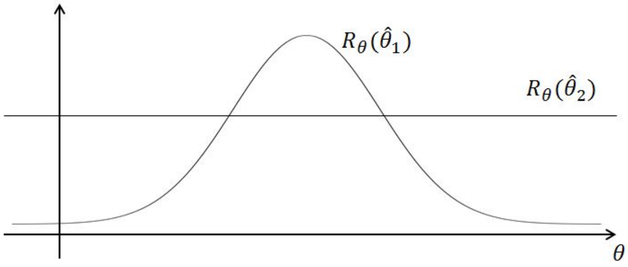
$$X = \underbrace{(X_1 \cdots, X_n)}_{\text{i.i.d. samples}}$$
$$\mathcal{P} = \{P_\theta^{\otimes n} : \theta \in \Theta\}$$

1.2 How to define the "best estimator"

One of the main objectives of this course is to investigate the fundamental limit, that is, to find the performance of the best estimator. We use the risk of an estimator to quantify its performance. As mentioned in the framework, for an estimator $\hat{\theta}$, we define the risk as follows:

$$R_\theta(\hat{\theta}) = \mathbb{E}_\theta[l(\theta, \hat{\theta})]$$

Note that, $R_\theta(\hat{\theta})$ could be viewed as a function of θ. As an example, the following figure depicts the risk curves for two different estimators.



To find the best estimator, we first need to define the figure of merit.

Naive Method: Find the estimator which is better that all other estimators at all points, i.e., find $\hat{\theta}$, such that

$$R_\theta(\hat{\theta}) \leq R_\theta(\hat{\theta}'), \qquad \forall \theta, \forall \hat{\theta}'.$$

It is easy to see that this method is typically too greedy to be realistic and an estimator that satisfies the requirement above does not exist. For example, consider $\theta_1 \neq \theta_2$ in Θ and $l(\theta, \hat{\theta})$ is some norm. Consider the estimator $\hat{\theta}_1 = \theta_1$ which throws away data and always spits out θ_1. Then $R_{\theta_1}(\hat{\theta}) \leq R_{\theta_1}(\hat{\theta}_1) = 0$ means $\hat{\theta} = \theta_1$, which means it cannot beat $\hat{\theta}_2 = \theta_2$ now. Therefore, we need other methods to compare estimators.

Method 1 Limit the class of competitors (of $\hat{\theta}$):

In some cases, by restricting the class of estimators, we can find a strategy which is uniformly the best. For example,

- Restricting to unbiased estimators: Frequently it is good to have be biased.
- Restricting to invariant estimators

Method 1 is difficult to generalize to high dimensional problems.

Method 2 Bayes approach: average-case analysis.

Method 3 Minimax approach: worst-case analysis

As mentioned before, finding a curve that dominates all other curves at all points is not always feasible. Hence, in Methods 2 and 3, we summarize a curve to a number so that we can compare them. In Method 2, we give weights to each point and take the average. The weights are called the prior. But the problem is which prior to choose. In Method 3, we consider the worst prior. For example, according to Method 3, in the figure above, $\hat{\theta}_2$ is a better strategy.

§ 2. Minimax risk and Bayes risk

Recall from last lecture:

Model: A set of probability distributions $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, where θ is the parameter (finite or infinite dimensional) that specifies the distribution.

The estimation problem: Nature chooses θ and generate data X from the distribution P_θ. Upon observing X, the statistician estimates a functional $T(\theta)$ of θ, by $\hat{T}$. In this lecture, for simplicity, we focus on estimating θ itself and thus $T(\theta) = \theta$. In the following parts, we consider deterministic estimator of θ denoted by $\hat{\theta}(X)$ as well as randomized estimator given by the transition kernel $P_{\theta|X}$. Equivalently, we can write $\hat{\theta} = \hat{\theta}(X, U)$, where U is a random variable that is independent from X. For all practical purposes (e.g., X takes value in a standard Borel space), we can choose U to be uniform on $[0, 1]$. (Why?)

Risk: $R_\theta(\hat{\theta}) = \mathbb{E}_\theta \ell(\theta, \hat{\theta})$, which quantifies the quality of the estimator $\hat{\theta}$ at θ.

Remark 2.1 (Convex loss $\implies$ deterministic estimator)**.** If $\hat{\theta} \mapsto \ell(\theta, \hat{\theta})$ is convex, then randomization does not help. The proof of this claim is just based on the Jensen's inequality: for any randomized estimator $\hat{\theta}$, we have

$$R_\theta(\hat{\theta}) = \mathbb{E}\ell(\theta, \hat{\theta}) \geq \mathbb{E}\ell(\theta, \mathbb{E}[\hat{\theta}|X]),$$

where $\mathbb{E}[\hat{\theta}|X]$ is a deterministic estimator.

2.1 Bayes risk

The Bayes approach is an average-case analysis by considering the average risk of an estimator over all $\theta \in \Theta$. Concretely, we set a probability distribution (prior) π on Θ. Then, the **average risk** (w.r.t π) is defined as

$$R_\pi(\hat{\theta}) = \mathbb{E}_{\theta \sim \pi} R_\theta(\hat{\theta}) = \mathbb{E}_{\theta, X} \ell(\theta, \hat{\theta}).$$

The **Bayes risk** for a prior π is the minimum that the average risk can achieve, i.e.

$$R^*_\pi = \inf_{\hat{\theta}} R_\pi(\hat{\theta}).$$

Example 2.1 (Quadratic loss and MMSE)**.** Let $\theta, \hat{\theta} \in \mathbb{R}$, $\theta \sim \pi$. Consider quadratic loss $\ell(\theta, \hat{\theta}) = \|\hat{\theta} - \theta\|_2^2$, then the Bayes risk is the minimum mean-square error (MMSE)

$$R^*_\pi = \mathbb{E}\|\theta - \mathbb{E}[\theta|X]\|_2^2,$$

where the Bayes estimator is the conditional mean $\hat{\theta}(X) = \mathbb{E}[\theta|X]$.

Example 2.2 (Gaussian Location Model)**.** $X = \theta + Z, Z \sim \mathcal{N}(0, 1), \theta \in \mathbb{R}$. Consider the Gaussian prior distribution: $\theta \sim \pi = \mathcal{N}(0, \sigma^2)$. Then $\mathbb{E}[\theta|X] = \frac{\sigma^2}{1+\sigma^2} X$ and

$$R^*_\pi = \frac{\sigma^2}{\sigma^2 + 1}. \tag{2.1}$$

Similarly, for multivariate GLM: $X = \theta + Z, Z \sim \mathcal{N}(0, \mathbf{I}_p)$, if $\theta \sim \pi = \mathcal{N}(0, \sigma^2 \mathbf{I}_p)$, then we have

$$R_\pi^* = \frac{\sigma^2}{\sigma^2 + 1} p. \tag{2.2}$$

If $R_\pi^* = \inf_{\hat{\theta}} R_\pi(\hat{\theta})$ is attained by $\hat{\theta}$, $\hat{\theta}$ is called *Bayes estimator*. Bayes estimator is always deterministic – this fact holds for any loss function. To see this, note that for any randomized estimator $\hat{\theta} = \hat{\theta}(X, U)$, its risk is lower bounded by

$$R_\pi(\hat{\theta}) = \mathbb{E}_{\theta,X,U} \ell(\theta, \hat{\theta}(X, U)) = \mathbb{E}_U R_\pi(\hat{\theta}(\cdot, U)) \geq \inf_u R_\pi(\hat{\theta}(\cdot, u))$$

where for any u, $\hat{\theta}(\cdot, u)$ is a deterministic estimator.

An alternative way to appreciate this is the following: Note that for any randomized estimator understood as a Markov kernel $P_{\hat{\theta}|X}$, the average risk $R_\pi(\hat{\theta})$ is an affine functional of $P_{\hat{\theta}|X}$. Maximizing a convex (e.g., affine) function over a convex constraint set is always achieved at the extremal points. In this case the extremal points of Markov kernels are simply delta measures, which corresponds to deterministic estimators.

The usual critisim to the Bayes approach is which prior to pick. A framework related to this but not discussed in this case is the empirical Bayes approach, where one "estimates" the prior from the data instead of choosing a prior a priori. Instead, we take a frequentist viewpoint by considering the worst-case situation:

2.2 Minimax risk

We have the risk of $\hat{\theta}$ at a given point $\theta : R_\theta(\hat{\theta})$. The **minimax risk** is defined as

$$R^* = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} R_\theta(\hat{\theta}). \tag{2.3}$$

If there exists $\hat{\theta}$ s.t. $\sup_{\theta \in \Theta} R_\theta(\hat{\theta}) = R^*$, then the estimator $\hat{\theta}$ is minimax (minimax optimal).

Finding the value of the minimax risk R^* entails

$$\text{Minimax upper bound:} \quad \exists \hat{\theta}, \forall \theta, R_\theta(\hat{\theta}) \leq r \Leftrightarrow R^* \leq r \tag{2.4}$$

$$\text{Minimax lower bound:} \quad \forall \hat{\theta}, \exists \theta, R_\theta(\hat{\theta}) \geq r \Leftrightarrow R^* \geq r \tag{2.5}$$

This task is frequently difficult especially in high dimensions. Instead of the exact minimax risk, it is often useful to find a constant-factor approximation, which we call **minimax rate**

$$R^* \asymp \psi, \tag{2.6}$$

that is, $c\psi \leq R^* \leq C\psi$ for some universal constants $c, C \geq 0$. Establishing ψ is a minimax rate still entails upper and lower bounds (2.4) and (2.5), albeit within multiplicative constant factors.

In practice, minimax lower bounds are rarely established via the obvious recipe (2.5). Throughout this course, all lower bound techniques essentially boil down to lower bounding the minimax risk by Bayes risk with a smarly chosen prior.

Theorem 2.1 (Minimax risk $\geq$ worst-case Bayes risk)**.**

$$R^* \geq R_B^* \triangleq \sup_\pi R_\pi^*.$$

Proof. Two (equivalent) ways to understand this fact:

1. "max ≥ mean": $\forall \hat{\theta}$, $R_\pi(\hat{\theta}) = \mathbb{E}_{\theta\sim\pi} R_\theta(\hat{\theta}) \leq \sup_{\theta\in\Theta} R_\theta(\hat{\theta})$;

2. "min max ≥ max min":
$$R^* = \inf_{\hat{\theta}} \sup_{\theta\in\Theta} R_\theta(\hat{\theta}) = \inf_{\hat{\theta}} \sup_{\pi\in\mathcal{M}(\Theta)} R_\pi(\hat{\theta}) \geq \sup_{\pi\in\mathcal{M}(\Theta)} \inf_{\hat{\theta}} R_\pi(\hat{\theta}) = \sup_\pi R^*_\pi,$$
where $\mathcal{M}(\Theta)$ is the set of all probability distributions on Θ. □

Example 2.3 (Minimax > worst-case Bayes)**.** Let $\theta, \hat{\theta} \in \mathbb{N} \triangleq \{1, 2, ...\}$ and $\ell(\theta, \hat{\theta}) = \mathbf{1}\{\hat{\theta} < \theta\}$, i.e., the statistician loses one dollar if the nature's choice exceeds the statistician's guess and loses nothing if otherwise. Consider the extreme case of blind guessing (i.e., no data is available, say, $X = 0$). Then $\forall \hat{\theta}$, we have $R_\theta(\hat{\theta}) = \mathbb{P}(\hat{\theta} < \theta)$. Furthermore, we have $R^* \geq \lim_{\theta\to\infty} \mathbb{P}(\hat{\theta} < \theta) = 1$, which is clearly achievable. On the other hand, for any prior π on $\mathbb{N}$, $R_\pi(\hat{\theta}) = \mathbb{P}(\hat{\theta} < \theta)$ and we let $\hat{\theta} \to \infty$. Therefore, we have $R^*_\pi = 0$. Therefore in this case
$$R^* = 1 > R^*_B = 0.$$

Example 2.4 (Gaussian Linear Model)**.** In the scalar case, this experiment is given by:
$$X \sim \mathcal{N}(\theta, 1), \quad \ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2, \quad \theta, \hat{\theta} \in \mathbb{R}$$

To get a minimax upper bound, we choose $\hat{\theta} = X$ and thus $R_\theta(\hat{\theta}) = 1$. Therefore, $R^* \leq 1$. To get a minimax lower bound, we set a prior distribution for θ, i.e., $\pi \sim \mathcal{N}(0, \sigma^2)$. Using (2.1), we have $R^* \geq R^*_\pi = \frac{\sigma^2}{\sigma^2+1}$ for all $\sigma > 0$ and thus $R^* \geq \sup_\pi R^*_\pi = 1$.

The p-dimensional case: $X = \theta + Z \in \mathbb{R}^p$, $Z \sim \mathcal{N}(0, \mathbf{I}_p)$, $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^2$. Similarly, using (2.1) as a lower bound and using $\hat{\theta} = X$ for the upper bound, we have $R^* = p$.

Exercise 2.1. Show that the minimax quadratic risk of the GLM model $X \sim \mathcal{N}(\theta, 1)$ with parameter space $\theta \geq 0$ is the same as the unconstrained case.

This might be a bit surprising because it is reasonable to think that the thresholded estimator $X_+ = \max(X, 0)$ can improve the performance. This is indeed correct in the sense that it achieves a better risk pointwise (at every $\theta \geq 0$); however, in the worst case the gain is asymptotically diminishing.

2.3 Minimax and Bayes risk: an optimization perspective

In the last lecture, we proved that minimax risk is always lower bounded from the worst-case Bayes risk, namely
$$R^* \geq R^*_B \triangleq \sup_{\pi\in\mathcal{M}(\Theta)} R^*_\pi. \tag{2.7}$$

This result can be interpreted from an optimization perspective. More precisely:

- The RHS of (2.7) is exactly the dual program of the LHS;
- The inequality (2.7) is simply *weak duality*;
- If *strong duality* holds, then (2.7) holds with equality, i.e., minimax theorem holds.

For simplicity, we consider the case where Θ is a finite set. Recall the minimax risk

$$R^* = \min_{\hat{\theta}} \max_{\theta \in \Theta} \mathbb{E}_\theta[\ell(\theta, \hat{\theta})],$$

which is in fact a convex optimization problem. Indeed, $P_{\hat{\theta}} \mapsto \mathbb{E}_\theta[\ell(\theta, \hat{\theta})]$ is affine and supremum of affine functions are convex. Let us write down its dual problem. First we can rewrite

$$\begin{aligned} R^* = \min_{\hat{\theta}, t} \quad & t \\ \text{s.t} \quad & \mathbb{E}_\theta[\ell(\theta, \hat{\theta})] \le t, \quad \forall \theta \in \Theta. \end{aligned}$$

Attach dual variable $\pi_\theta \ge 0$ to each inequality constraint. The Lagrangian of the above minimization problem is

$$\begin{aligned} L(\hat{\theta}, t, \pi) &= t + \sum_{\theta \in \Theta} \pi_\theta \left(\mathbb{E}_\theta[\ell(\theta, \hat{\theta})] - t \right) \\ &= \left(1 - \sum_{\theta \in \Theta} \pi_\theta \right) t + \sum_{\theta \in \Theta} \pi_\theta \mathbb{E}_\theta[\ell(\theta, \hat{\theta})]. \end{aligned}$$

Note that unless $\sum_{\theta \in \Theta} \pi_\theta = 1$, $\min_{t \in \mathbb{R}} L(\hat{\theta}, t, \pi)$ is $-\infty$. Thus the dual variable $\pi = \{\pi_\theta : \theta \in \Theta\}$ is a probability measure (prior) and the dual problem is

$$\begin{aligned} \max_\pi \min_{\hat{\theta}, t} L(\hat{\theta}, t, \pi) &= \max_{\substack{\pi : \pi_\theta \ge 0, \\ \sum_{\theta \in \Theta} \pi_\theta = 1}} \min_{\hat{\theta}} R_\pi(\hat{\theta}) \\ &= \max_{\pi \in \mathcal{M}(\Theta)} R_\pi^*. \end{aligned}$$

Hence, $R^* \ge R_\pi^*$, for all $\pi \in \mathcal{M}(\Theta)$. Note that, as usual, the dual program is a concave maximization problem.

Theorem 2.2 (Minimax theorem)**.**

$$R^* = R_B^*$$

in either of the following cases:

- *Θ is a finite set and the data X takes values in a finite set $\mathcal{X}$.*
- *Θ is a finite set and the loss function ℓ is bounded from below, i.e.,* $\inf_{\theta, \hat{\theta}} \ell(\theta, \hat{\theta}) > -\infty$

Proof. The first case directly follows from the fact that strong duality holds for finite-dimensional linear programming.

For the second case, we start by showing that if $R^* = \infty$, then $R_B^* = \infty$. To see this, consider the uniform prior π on Θ. Then for any estimator $\hat{\theta}$, there exists $\theta \in \Theta$ such that $R(\theta, \hat{\theta}) = \infty$. Then $R_\pi(\hat{\theta}) \ge \frac{1}{|\Theta|} R(\theta, \hat{\theta}) = \infty$.

Next we assume that $R^* < \infty$. Then $R^* \in \mathbb{R}$ since ℓ is bounded from below (say, by a) by assumption. Define

$$\begin{aligned} &S = \{R(\cdot, \hat{\theta}) \in \mathbb{R}^\Theta : \hat{\theta} \text{ is a randomized estimator}\} = \{\text{set of all possible risk vectors}\} \\ &T = \{t \in \mathbb{R}^\Theta : t_i < R^*\}. \end{aligned}$$

Note that both S and T are convex (why?) subsets of Euclidean space $\mathbb{R}^\Theta$ and $S \cap T = \emptyset$ by definition of R^*. By the separation hyperplane theorem, there exists a non-zero $\pi \in \mathbb{R}^\Theta$ and $c \in \mathbb{R}$, such that $\inf_{s \in S} \langle \pi, s \rangle \geq c \geq \sup_{t \in T} \langle \pi, t \rangle$. Obviously, π must be componentwise positive, for otherwise $\sup_{t \in T} \langle \pi, t \rangle = \infty$. Therefore by normalization we may assume that π is a probability vector, i.e., a prior on Θ. Then $R_B^* \geq R_\pi^* = \inf_{s \in S} \langle \pi, s \rangle \geq \sup_{t \in T} \langle \pi, t \rangle \geq R^*$, completing the proof. □

§ 3. Minimax risk of GLM and four extensions

In this lecture, we discuss several extension of the Gaussian local model (GLM) to illustrate the following concepts such as *tensor product of experiments* and *sample complexity*.

Recall the scalar GLM we discussed in the last lecture, where $X = \theta + Z$, where $\theta \in \mathbb{R}, Z \sim \mathcal{N}(0, \sigma^2)$ and the loss function is quadratic $\ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$. Then we have

$$R^* = \sigma^2. \tag{3.1}$$

This follows from

- Lower bound: If $\theta \sim \mathcal{N}(0, \sigma_0^2)$, we know $R_\pi = \frac{\sigma_0^2 \sigma^2}{\sigma_0^2 + \sigma^2}$. Letting $\sigma_0 \to \infty$ yields $R^* \geq \sigma^2$.
- Upper bound: Let the estimator be $\hat{\theta} = X$. Thus $R_\theta(\hat{\theta}) = \sigma^2$ for all θ. Hence $R^* \leq \sigma^2$.

3.1 Multivariate version and tensor product of experiments

We observe $X = \theta + Z$, where $\theta \in \mathbb{R}^p, Z \sim \mathcal{N}(0, \sigma^2 I_p)$ and the loss function $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_2^2$. Then

$$R^* = p\sigma^2. \tag{3.2}$$

This can be obtained using similar argument to the univariate case:

- Lower bound: $\theta \sim \mathcal{N}(0, \sigma_0^2 I_p)$ and $\sigma_0 \to \infty$.
- Upper bound: take $\hat{\theta} = X$.

The multivariate GLM can be viewed as a tensor product of the univariate GLM, and their minimax risks satisfy a general relationship. We discuss this notion below:

Minimax risk for tensor product of the experiment Given statistical experiments $\mathcal{P}_i = \{P_{\theta_i} \colon \theta_i \in \Theta_i\}$ and the corresponding loss function ℓ_i, for $i \in [p]$, consider their tensor product, which is the following statistical experiment:

$$\begin{aligned}
\mathcal{P} = \left\{ P_\theta = \prod_{i=1}^{p} P_{\theta_i} \colon \theta = \{\theta_1, \ldots, \theta_p\} \in \Theta \triangleq \prod_{i=1}^{p} \Theta_i \right\}&, \\
X = (X_1, \ldots, X_p) \text{ where } X_i \overset{\text{ind}}{\sim} P_{\theta_i}&, \\
\ell(\theta, \hat{\theta}) = \sum_{i=1}^{n} \ell_i(\theta_i, \hat{\theta}_i), \forall \theta, \hat{\theta} \in \Theta&.
\end{aligned}$$

Then the minimax risk of the tensor product experiment is related to the minimax risk $R^*(\mathcal{P}_i)$ and worst-case Bayes risks $R_B^*(\mathcal{P}_i) \triangleq \sup_\pi R_\pi(\mathcal{P}_i)$ of individual experiments as follows:[1]

[1] Here the minimax risk is defined allowing randomized procedures.

Theorem 3.1 (Minimax risk of tensor product)**.**

$$\sum_{i=1}^{p} R_B^*(\mathcal{P}_i) \leq R^*(\mathcal{P}) \leq \sum_{i=1}^{p} R^*(\mathcal{P}_i). \tag{3.3}$$

Consequently, if minimax theorem holds for each experiment, i.e., $R^*(\mathcal{P}_i) = R_B^*(\mathcal{P}_i)$, we have

$$R^*(\mathcal{P}) = \sum_{i=1}^{p} R^*(\mathcal{P}_i). \tag{3.4}$$

Proof. The right inequality simply follows by separately estimating θ_i based on X_i, namely, $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)$. For the left inequality, consider a product prior $\pi = \prod_{i=1}^{p} \pi_i$. Then X_i's are independent. For any $\hat{\theta}_i = \hat{\theta}_i(X_1, \ldots, X_p, U_i)$, where U_i is independent external randomness, we can rewrite $\hat{\theta}_i = \hat{\theta}_i(X_i, \tilde{U}_i)$, where $\tilde{U}_i = (X_{\setminus i}, U_i) \perp\!\!\!\perp X_i$ serves as randomization. Therefore the Bayes risk of $\hat{\theta}_i$ satisfies: $\mathbb{E}[\ell(\theta_i, \hat{\theta}_i)] \geq R_{\pi_i}^*$. Summing over i and taking suprema over priors π_i's yields the left inequality of (3.4). □

Remark 3.1 (Minimax risk of tensor product < sum of minimax risks)**.** The right inequality of (3.4) can be strict. This might appear surprising since X_i only carries information about θ_i and it is intuitive to estimate θ_i based solely on X_i. Nevertheless, the following is a counterexample:

Consider $X = \theta Z$, where $\theta \in \mathbb{N}$, $Z \sim \text{Bern}(\frac{1}{2})$. The estimator $\hat{\theta}$ takes values in $\mathbb{N}$ as well and the loss function is $\ell(\theta, \hat{\theta}) = \mathbf{1}\{\hat{\theta} < \theta\}$, i.e., whoever guesses the greater number wins. The minimax risk for this experiment is equal to $\mathbb{P}[Z = 0] = \frac{1}{2}$. To see this, note that if $Z = 0$, then all information about θ is erased. Therefore for any (randomized) estimator $P_{\hat{\theta}|X}$, the risk is lower bounded by $R_\theta(\hat{\theta}) = \mathbb{P}[\hat{\theta} < \theta] \geq \mathbb{P}[\hat{\theta} < \theta, Z = 0] = \frac{1}{2}\mathbb{P}[\hat{\theta} < \theta | X = 0]$. Therefore sending $\theta \to \infty$ yields $\sup_\theta R_\theta(\hat{\theta}) \geq \frac{1}{2}$. This is achievable by $\hat{\theta} = X$. Clearly, this is a case where minimax theorem does not hold, which is very similar to the trivial example given in the last lecture.

Next consider the tensor product of two copies of this experiment. We show that the minimax risk is strictly less than one. For $i = 1, 2$, let $X_i = \theta_i Z_i$, where $Z_1, Z_2 \overset{\text{i.i.d.}}{\sim} \text{Bern}(\frac{1}{2})$. Consider the following estimator $\hat{\theta}_1 = \hat{\theta}_2 = X_1 \vee X_2$. Then for any $\theta_1, \theta_2 \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}[\ell(\theta, \hat{\theta})] &= \mathbb{P}[\hat{\theta}_1 < \theta_1] + \mathbb{P}[\hat{\theta}_2 < \theta_2] = \mathbb{P}[Z_1 = 0, Z_2 < \theta_1/\theta_2] + \mathbb{P}[Z_2 = 0, Z_1 < \theta_2/\theta_1] \\ &= \frac{1}{2}(\mathbb{P}[Z_2 < \theta_1/\theta_2] + \mathbb{P}[Z_1 < \theta_2/\theta_1]) \leq \frac{3}{4}.\end{aligned}$$

Remark 3.2 (Non-uniqueness of minimax estimator)**.** In general, minimax risk achieving strategies need not be unique. For instance, consider Example 3.1 where $\hat{\theta} = X$ is the maximum likelihood estimator as well as the minimax. On the other hand, the risk of the James-Stein estimator

$$\hat{\theta}_{JS} = \left(1 - \frac{p-2}{\|X\|^2}\right) X$$

dominates that of MLE everywhere (see Fig. 3.1). Therefore $\hat{\theta}_{JS}$ also achieves $R^* = p$ for $p \geq 3$.

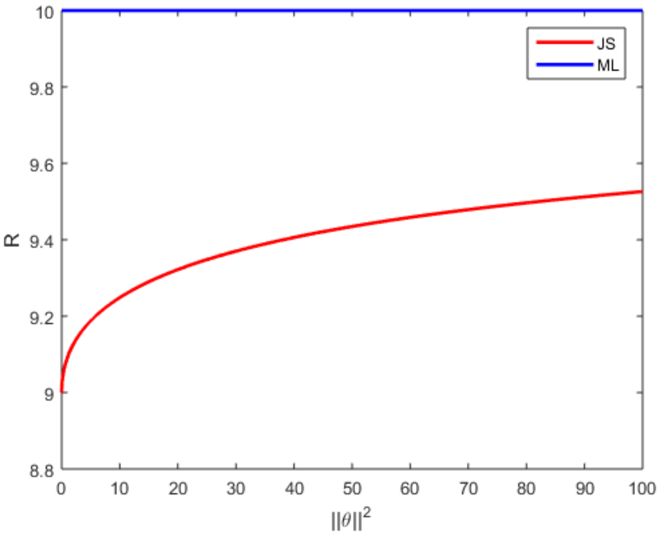


Figure 3.1: Risks of MLE and JS estimators for $p = 10$.

3.2 Multiple samples and sample complexity

We now consider a variant of GLM where we observe $X = (X_1, \ldots, X_n)$ where $X_i = \theta + Z_i, Z_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2 I_p), \theta \in \mathbb{R}^p$. In this case, we have

$$R^* = \frac{p\sigma^2}{n}. \tag{3.5}$$

To see this, note that for the case of i.i.d. Gaussian random variables, $\bar{X}$ is a sufficient statistic of X for θ, because the joint pdf $p_{X_1,\ldots,X_n|\theta}$ is of the form $h(X)\, g_\theta(\bar{X})$, and hence by Fisher's factorization criterion, $\theta \to \bar{X} \to (X_1, \ldots, X_n)$. Therefore the model reduces to $\bar{X} \sim \mathcal{N}(\theta, \frac{\sigma^2}{n} I_p)$, which is the single-sample multivariate case and the minimax risk follows from (3.2).

Sample complexity Given the experiment $\{P_\theta : \theta \in \Theta\}$, consider the experiment

$$\mathcal{P}_n = \left\{P_\theta^{\otimes n} : \theta \in \Theta\right\}.$$

Note this is not the tensor product of the given experiment because all samples are generated by a common parameter. It is easy to see that $n \mapsto R^*(\mathcal{P}_n)$ is decreasing since we can always discard samples. Typically, $R^*(\mathcal{P}_n) \to 0$ as $n \to \infty$. Thus it is natural to consider how fast $R^*(\mathcal{P}_n)$ decreases with n (convergence rate). Equivalently, one can ask what is the minimum number of samples to attain a prescribed error ϵ even in the worst case. This motivates the following definition.

Definition 3.1 (Sample complexity). Given an error margin $\epsilon > 0$, we define the *sample complexity* of the statistical model as

$$n^*(\epsilon) \triangleq \min\{n \in \mathbb{N} : R^*(\mathcal{P}_n) \leq \epsilon\}.$$

In machine learning and related fields, it is customary and useful to consider high-probability bound instead of average risk bound and it is useful to define the sample complexity to be the minimum number of samples required to achieve a prescribed loss with high confidence. In other words, given $\epsilon > 0$ and $0 < \delta < 1$, the sample complexity $n^*(\epsilon, \delta)$ is the smallest n such that there exists $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ satisfying

$$\mathbb{P}_\theta(\ell(\theta, \hat{\theta}) \le \epsilon) \ge 1 - \delta, \quad \forall \theta \in \Theta.$$

This is in fact just a special case of Definition 3.1 with the loss function ℓ replaced by $\mathbf{1}\{\ell(\theta, \hat{\theta}) \ge \epsilon\}$.

Remark 3.3. For the multi-sample GLM with unit variance, we know that $R^* = \frac{p}{n}$. Hence the sample complexity is given by $n^*(\epsilon) = \lceil \frac{p}{\epsilon} \rceil$. Here we notice that the sample complexity grows linearly with the dimension p. This is the common wisdom that "the sample size need to scale at least proportionally to the number of parameters", also known as "counting the degrees of freedom". Indeed in high dimensions we typically expect the sample complexity to grow with the ambient dimension. However, such claim of linear growth should be taken with a grain of salt because it highly depends on what loss function and what is target we are estimating. For example, consider the matrix case $\theta \in \mathbb{R}^{p \times p}$ and let ϵ be a small constant. Then

- For quadratic loss, namely, $\|\theta - \hat{\theta}\|_F^2$, then we have $R^* = \frac{p^2}{n}$ and hence $n^*(\epsilon) = \Theta(p^2)$.
- If the loss function is $\|\theta - \hat{\theta}\|_{op}^2$, then we have $R^* \asymp \frac{p}{n}$ and hence $n^*(\epsilon) = \Theta(p)$.
- If we only want to estimate the scalar functional $\|\theta\|_{\ell_\infty}$, then $n^*(\epsilon) = \Theta(\sqrt{\log p})$.

3.3 Nonparametric extension

The result we obtained on the minimax risk of GLM can be in fact generalized to the following nonparametric setting. Consider the class of distributions (which need have density) on the real line with bounded variance:

- Model: $\mathcal{P} = \{P \in \mathcal{M}(\mathbb{R}) : \mathrm{var}_P \le 1\}$, where var_P denotes the variance of the distribution P.
- Data: $X = (X_1, \dots, X_n) \overset{iid}{\sim} P$ for some $P \in \mathcal{P}$.
- Objective: We wish to estimate $\theta(P)$ where $\theta(P) =$ mean of the distribution P.
- Loss function: $\ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ for $\theta, \hat{\theta} \in \mathbb{R}$.

Then the minimax risk is

$$R^*(\mathcal{P}) = \frac{1}{n}.$$

Proof. Restricting the analysis to the subcollection of Gaussian distributions $\mathcal{P}_G = \{\mathcal{N}(\theta, 1) : \theta \in \mathbb{R}\}$, we know that $R^*(\mathcal{P}_G) = \frac{1}{n}$. Hence $R^*(\mathcal{P}) \ge \frac{1}{n}$. On the other hand, for the estimator $\hat{\theta} = \bar{X}$,

$$R_\theta(\hat{\theta}) = \mathbb{E}[(\theta(P) - \hat{\theta})^2] = \mathbb{E}[(\theta(P) - \bar{X})^2] = \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^n (\theta(P) - X_i)^2\right] \le \frac{1}{n}.$$

Hence $\sup_{P \in \mathcal{P}} R_\theta(\hat{\theta}) \le \frac{1}{n}$ and $R^*(\mathcal{P}) \le \frac{1}{n}$. Thus $R^*(\mathcal{P}) = \frac{1}{n}$. □

3.4 Non-quadratic loss

One can also consider non-quadratic loss functions such as $\|\theta - \hat{\theta}\|_1$ when $\theta \in \mathbb{R}^p$ or $\|\theta - \hat{\theta}\|_{op}$ when $\theta \in \mathbb{R}^{p \times p}$, etc., where R^* will no longer be given by (3.5). We will prove the following result later in the course (see Lecture 9).

Theorem 3.2. *For the Gaussian location model where $X = (X_1, \ldots, X_n) \overset{iid}{\sim} \mathcal{N}(\theta, I_p)$ and $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^2$ for some arbitrary norm $\|\cdot\|$, one has*

$$R^* = \frac{\mathbb{E}[\|Z\|^2]}{n}.$$

Thus (3.5) can be seen as a direct consequence of this theorem. In this case, the sample complexity $n^*(\epsilon)$ scales as $\frac{\mathbb{E}[\|Z\|^2]}{\epsilon}$, depending on the norm.

Part II

f-divergences, information inequalities, and large-sample asymptotics

4.1 f-divergences

Outline of the topics.

- *Notion of dissimilarity between distributions:* Common f-divergences such as KL-divergence, Hellinger distance, total variation distance, χ^2-distance, etc.
- *Notion of dependence between distributions*: Mutual information.
- *Data processing principle*
- *Fisher information & minimax bounds*

We now define the f-divergence between probability distributions over a measurable space $(\mathcal{X}, \mathcal{F})$, introduced by Csiszár [Csi67]. Roughly speaking, all f-divergences quantify the difference between a pair of distributions, each with different operational meaning.

Definition 4.1 (f-divergence)**.** Let P and Q be two probability distributions on $\mathcal{X}$. Then for any convex function $f : (0, \infty) \to \mathbb{R}$ such that it is strictly convex[1] at 1 and $f(1) = 0$, the f-divergence of Q from P with $P \ll Q$ is defined as

$$D_f\left(P \| Q\right) \triangleq \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right]. \tag{4.1}$$

Remark 4.1. When $\mathcal{X}$ is discrete, $D_f\left(P \| Q\right) = \sum_{x \in \mathcal{X}} Q(x) f\left(\frac{P(x)}{Q(x)}\right)$.

Remark 4.2. The notation $\frac{dP}{dQ}$ stands for the relative density (Radon-Nikodym derivative) of P with respect to Q, whenever $P \ll Q$, which is a function from $\mathcal{X}$ to $\mathbb{R}$. For conciseness, we sometimes abbreviate $\frac{dP}{dQ}$ by $\frac{P}{Q}$. The defining property of $\frac{dP}{dQ}$ is its utility in *change of measure*:

$$\mathbb{E}_P\left[f(X)\right] = \mathbb{E}_Q\left[f(X)\frac{dP}{dQ}(X)\right]$$

Remark 4.3 (General definition of f-divergence)**.** If P is not absolutely continuous with respect to Q, (4.1) is not well-defined. The general definition of f-divergence is as follows:

$$D_f(P \| Q) \triangleq \mathbb{E}_Q\left[f\left(\frac{dP/d\mu}{dQ/d\mu}\right)\right] \tag{4.2}$$

where μ is a dominating probability measure (e.g., $\mu = (P + Q)/2$) of P and Q, i.e., both $P \ll \mu$ and $Q \ll \mu$, with the understanding that

[1] By strict convexity at 1, we mean for all $s, t \in (0, \infty)$ and $\alpha \in (0, 1)$ such that $\alpha s + \bar{\alpha} t = 1$, we have $\alpha f(s) + (1 - \alpha) f(t) > f(1)$.

- $f(0) = f(0+)$,
- $0f(\frac{0}{0}) = 0$, and
- $0f(\frac{a}{0}) = \lim_{x \downarrow 0} xf(\frac{a}{x})$ for $a > 0$.

This definition is not merely for completeness, for example, we will show later that $d_{\text{TV}}(P, Q) = 1$ iff $P \perp Q$.

The following are the common f-divergences ("big four") that we would be frequently used in this course.

- **Kullback-Leibler (KL) divergence:** aka relative entropy, $f(x) = x \log x$,

  $$D(P\|Q) \triangleq \mathbb{E}_Q \left[\frac{P}{Q} \log \frac{P}{Q}\right] = \mathbb{E}_P \left[\log \frac{P}{Q}\right].$$

  It is worth noting that, in general $D(P\|Q) \neq D(Q\|P)$. When $f(x) = -\log x$, we obtain $D_f(P\|Q) = \mathbb{E}_Q \left[-\log \frac{P}{Q}\right] = D(Q\|P)$.

- **Total variation:** $f(x) = \frac{1}{2}|x - 1|$,

  $$d_{\text{TV}}(P, Q) \triangleq \frac{1}{2}\mathbb{E}_Q \left[\left|\frac{P}{Q} - 1\right|\right] = \frac{1}{2} \int |dP - dQ|.$$

  Moreover, $d_{\text{TV}}(\cdot, \cdot)$ is a metric on the space of probability distributions, and hence it is a symmetric function of P and Q.

- **χ^2-divergence**: $f(x) = (x - 1)^2$,

  $$\chi^2(P\|Q) \triangleq \mathbb{E}_Q \left[\left(\frac{P}{Q} - 1\right)^2\right] = \int \frac{(P - Q)^2}{Q} = \int \frac{P^2}{Q} - 1.$$

  Note that we can also choose $f(x) = x^2 - 1$. Indeed different f can lead to the same divergence.

- **Squared Hellinger distance:** $f(x) = (1 - \sqrt{x})^2$,

  $$H^2(P, Q) \triangleq \mathbb{E}_Q \left[\left(1 - \sqrt{\frac{P}{Q}}\right)^2\right] = \int \left(\sqrt{P} - \sqrt{Q}\right)^2.$$

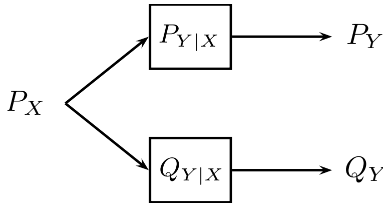
  Note that $H^2(P, Q) = H^2(Q, P)$.

Theorem 4.1 (Properties of f-divergences)**.**

- ***Non-negativity:*** *$D_f(P\|Q) \geq 0$ with equality if and only if $P = Q$.*
- ***Joint convexity:*** *$(P, Q) \mapsto D_f(P\|Q)$ is a jointly convex function. Consequently, $P \mapsto D_f(P\|Q)$ and $Q \mapsto D_f(P\|Q)$ are also convex functions.*
- ***Conditioning increases f-divergence:*** *Define the conditional f-divergence:*

  $$D_f\left(P_{Y|X}\|Q_{Y|X}|P_X\right) \triangleq \mathbb{E}_{X\sim P_X}\left[D_f\left(P_{Y|X}\|Q_{Y|X}\right)\right],$$

  Let $P_X \xrightarrow{P_{Y|X}} P_Y$ and $P_X \xrightarrow{Q_{Y|X}} Q_Y$, i.e.,



Then

$$D_f\left(P_Y \| Q_Y\right) \leq D_f\left(P_{Y|X} \| Q_{Y|X} | P_X\right).$$

Note: For the last property, one can view P_Y and Q_Y as the output distributions after passing P_X through the channel transition matrices $P_{Y|X}$ and $Q_{Y|X}$ respectively. The above relation tells us that the average f-divergence between the corresponding channel transition rows is at least the f-divergence between the output distributions.

Proof.

- $D_f(P\|Q) = \mathbb{E}_Q\left[f\left(\frac{P}{Q}\right)\right] \geq f\left(\mathbb{E}_Q\left[\frac{P}{Q}\right]\right) = f(1) = 0$, where the inequality follows from the Jensen's inequality. By strict convexity at 1, equality holds if and only if $P = Q$.
- For any convex function f on $\mathbb{R}_+$, it follows that $(p, q) \mapsto qf\left(\frac{p}{q}\right)$ is convex on $\mathbb{R}_+^2$ (the perspective of f). Since $D_f(P\|Q) = \mathbb{E}_Q\left[f\left(\frac{P}{Q}\right)\right]$, $D_f(P\|Q)$ is jointly convex.
- This follows directly from the joint-convexity of $D_f\left(P\|Q\right)$ and the Jensen's inequality.

□

Recall the definition of f-divergences from last time. If a function $f : \mathbb{R}^+ \to \mathbb{R}$ satisfies the following properties:

- f is a convex function.
- $f(1) = 0$.
- f is strictly convex at $x = 1$, i.e. for all x, y, α such that $\alpha x + \overline{\alpha} y = 1$, the inequality $f(1) < \alpha f(x) + \overline{\alpha} f(y)$ is strict.

Then the functional that maps pairs of distributions to $\mathbb{R}_+$ defined by

$$D_f(P\|Q) \triangleq \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right]$$

is an f-divergence.

4.2 Data processing inequality

Theorem 4.2. *Consider a channel that produces Y given X based on the law $P_{Y|X}$ (shown below). If P_Y is the distribution of Y when X is generated by P_X and Q_Y is the distribution of Y when X is generated by Q_X, then for any f-divergence $D_f(\cdot\|\cdot)$,*

$$D_f(P_Y\|Q_Y) \leq D_f(P_X\|Q_X).$$

$$\begin{tikzcd}
P_X \arrow[dr] & & P_Y \\
& \boxed{P_{Y|X}} \arrow[ur] \arrow[dr] & \\
Q_X \arrow[ur] & & Q_Y
\end{tikzcd}$$

One interpretation of this result is that processing the observation x makes it more difficult to determine whether it came from P_X or Q_X.

Proof.

$$D_f(P_X\|Q_X) = \mathbb{E}_{Q_X}\left[f\left(\frac{P_X}{Q_X}\right)\right] \overset{\text{(a)}}{=} \mathbb{E}_{Q_{XY}}\left[f\left(\frac{P_{XY}}{Q_{XY}}\right)\right] = \mathbb{E}_{Q_Y}\left[\mathbb{E}_{Q_{X|Y}} f\left(\frac{P_{XY}}{Q_{XY}}\right)\right]$$

$$\text{Jensen's inequality} \rightarrow \geq \mathbb{E}_{Q_Y}\left[f\left(\mathbb{E}_{Q_{X|Y}}\frac{P_{XY}}{Q_{XY}}\right)\right]$$

$$= \mathbb{E}_{Q_Y}\left[f\left(\mathbb{E}_{P_{X|Y}}\frac{P_Y}{Q_Y}\right)\right] \overset{\text{(b)}}{=} \mathbb{E}_{Q_Y}\left[f\left(\frac{P_Y}{Q_Y}\right)\right] = D_f(P_Y\|Q_Y).$$

Note that (a) means $D_f(P_X\|Q_X) = D_f(P_{XY}\|Q_{XY})$; (b) can be alternatively understood by noting that $\mathbb{E}_Q[\frac{P_{XY}}{Q_{XY}}|Y]$ is precisely the relative density $\frac{P_Y}{Q_Y}$, by checking the definition of change of measure, i.e., $\mathbb{E}_P[g(Y)] = \mathbb{E}_Q[g(Y)\frac{P_{XY}}{Q_{XY}}] = \mathbb{E}_Q[g(Y)\mathbb{E}[\frac{P_{XY}}{Q_{XY}}|Y]]$ for any g. □

Remark 4.4. $P_{Y|X}$ can be a deterministic map so that $Y = f(X)$. More specifically, if $f(X) = \mathbf{1}_E(X)$ for any event E, then Y is Bernoulli with parameter $P(E)$ or $Q(E)$ and the data processing inequality gives

$$D_f(P_X\|Q_X) \geq D_f(\text{Bern}(P(E))\|\text{Bern}(Q(E))). \tag{4.3}$$

This is how we prove the converse direction of large deviation.

Example 4.1. If $X = (X_1, X_2)$ and $f(X) = X_1$, then we have $D_f(P_{X_1X_2}\|Q_{X_1X_2}) \geq D_f(P_{X_1}\|Q_{X_1})$. As seen from the proof of Theorem 4.2, this is in fact equivalent to data processing inequality.

Remark 4.5. If $D_f(P\|Q)$ is an f-divergence, then $D_{\tilde{f}}(P\|Q)$ with $\tilde{f}(x) := xf(\frac{1}{x})$ is also an f-divergence and $D_f(P\|Q) = D_{\tilde{f}}(Q\|P)$. Example: $D_f(P\|Q) = D(P\|Q)$ then $D_{\tilde{f}}(P\|Q) = D(Q\|P)$.

Proof. First we verify that $\tilde{f}$ has all three properties required for $D_{\tilde{f}}(\cdot\|\cdot)$ to be an f-divergence.

- For $x, y \in \mathbb{R}^+$ and $\alpha \in [0,1]$ define $c = \alpha x + \overline{\alpha} y$ so that $\frac{\alpha x}{c} + \frac{\overline{\alpha} y}{c} = 1$. Observe that

  $$\tilde{f}(\alpha x + \overline{\alpha} y) = cf\left(\frac{1}{c}\right) = cf\left(\frac{\alpha x}{c}\frac{1}{x} + \frac{\overline{\alpha} y}{c}\frac{1}{y}\right) \leq c\frac{\alpha x}{c} f\left(\frac{1}{x}\right) + c\frac{\overline{\alpha} y}{c} f\left(\frac{1}{y}\right) = \alpha\tilde{f}(x) + \overline{\alpha}\tilde{f}(y).$$

  Thus $\tilde{f} : \mathbb{R}^+ \to \mathbb{R}$ is a convex function.

- $\tilde{f}(1) = f(1) = 0$.

- For $x, y \in \mathbb{R}^+$, $\alpha \in [0,1]$, if $\alpha x + \overline{\alpha} y = 1$, then by strict convexity of f at 1,

  $$0 = \tilde{f}(1) = f(1) = f\left(\alpha x\frac{1}{x} + \overline{\alpha} y\frac{1}{y}\right) < \alpha x f\left(\frac{1}{x}\right) + \overline{\alpha} y f\left(\frac{1}{y}\right) = \alpha\tilde{f}(x) + \overline{\alpha}\tilde{f}(y).$$

  So $\tilde{f}$ is strictly convex at 1 and thus $D_{\tilde{f}}(\cdot\|\cdot)$ is a valid f-divergence.

Finally,

$$D_f(P\|Q) = \mathbb{E}_Q\left[f\left(\frac{P}{Q}\right)\right] = \mathbb{E}_P\left[\frac{Q}{P}f\left(\frac{P}{Q}\right)\right] = \mathbb{E}_P\left[\tilde{f}\left(\frac{Q}{P}\right)\right] = D_{\tilde{f}}(Q\|P). \qquad \square$$

4.3 Total variation and hypothesis testing

Recall that the choice of $f(x) = \frac{1}{2}|x-1|$ gives rise to the total variation distance,

$$D_f(P\|Q) = \frac{1}{2}\mathbb{E}_Q\left|\frac{P}{Q}-1\right| = \frac{1}{2}\int |P-Q|,$$

where $\int |P-Q|$ is a short-hand understood in the usual sense, namely, $\int |\frac{dP}{d\mu} - \frac{dQ}{d\mu}|d\mu$ where μ is a dominating measure, e.g., $\mu = P+Q$, and the value of the integral does not depends on μ.

We will denote total variation by $d_{\mathrm{TV}}(P,Q)$ or $\mathrm{TV}(P,Q)$.

Theorem 4.3. *The following definitions for total variation are equivalent:*

1.
$$d_{\mathrm{TV}}(P,Q) = \sup_E P(E) - Q(E), \tag{4.4}$$
where the supremum is over all measurable set E.

2. *Given an observation X, the minimal sum of Type-I and Type-II error probabilities for testing $X \sim P$ versus $X \sim Q$ is given by $1 - d_{\mathrm{TV}}(P,Q)$,*
$$\min_\phi \{P(\phi = 1) + Q(\phi = 0)\} = 1 - d_{\mathrm{TV}}(P,Q), \tag{4.5}$$
where the minimum is over all decision rule $\phi \in \{0,1\}$ as a (deterministic or random) function of the observation X.[2] *Furthermore,*[3]
$$d_{\mathrm{TV}}(P,Q) = 1 - \int P \wedge Q. \tag{4.6}$$

3. *Provided the diagonal $\{(x,x) : x \in \mathcal{X}\}$ is measurable,*
$$d_{\mathrm{TV}}(P,Q) = \inf_{\substack{P_{XY}:\\ P_X=P, P_Y=Q}} \mathbb{P}[X \neq Y]. \tag{4.7}$$

4. *Let $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}, \|f\|_\infty \le 1\}$. Then*
$$d_{\mathrm{TV}}(P,Q) = \frac{1}{2}\sup_{f\in\mathcal{F}} \mathbb{E}_P f(x) - \mathbb{E}_Q f(x). \tag{4.8}$$

Remark 4.6 (Variational representation)**.** The equation (4.4) and (4.8) provide sup-representation of total variation, which will be extended to general f-divergences (later). Note that (4.7) is an inf-representation of total variation in terms of couplings, meaning total variation is the Wasserstein distance with respect to Hamming distance. The benefit of variational representations is that choosing a particular coupling in (4.7) gives an upper bound on $d_{\mathrm{TV}}(P,Q)$, and choosing a particular f in (4.8) yields a lower bound.

[2] The extension of (4.5) from from simple to composite hypothesis testing is in (18.1)

[3] Throughput the course $a \wedge b = \min\{a,b\}$ and $a \vee b = \max\{a,b\}$. Here again $\int P \wedge Q$ is a short-hand understood per the usual sense, namely, $\int (\frac{dP}{d\mu} \wedge \frac{dQ}{d\mu})d\mu$ where μ is any dominating measure.

Remark 4.7 (Operational meaning)**.** In the binary hypothesis test for $H_0 : X \sim P$ or $H_1 : X \sim Q$, Theorem 4.3 shows that $1 - d_{\text{TV}}(P, Q)$ is the sum of false alarm and missed detection probabilities. This can be seen either from (4.4) where E is the decision region for deciding P or from (4.6) since the optimal test (for average probability of error) is the likelihood ratio test $\frac{dP}{dQ} > 1$. In particular,

- $d_{\text{TV}}(P, Q) = 1 \Leftrightarrow P \perp Q$, the probability of error is zero since essentially P and Q have disjoint supports.
- $d_{\text{TV}}(P, Q) = 0 \Leftrightarrow P = Q$ and the minimal sum of error probabilities is one, meaning the best thing to do is to flip a coin.

4.4 Motivating example: Hypothesis testing with multiple samples

Observation: "Not all f-divergences are born equal"

1. Different f-divergences have different operational significance. For example, as we saw in Section 4.3, testing two hypothesis boils down to total variation, which determines the fundamental limit (minimum average probability of error). Later in the course we will encounter another f-divergence: $L(P\|Q) = \int \frac{(P-Q)^2}{P+Q}$, which is useful for estimation under quadratic loss.
2. Some f-divergence is easier to evaluate than others. For example, for product distributions, Hellinger distance and χ^2-divergence **tensorize** in the sense that they are easily expressible in terms of those of the one-dimensional marginals; however, computing the total variation between product measures is frequently difficult. Another example is that computing the χ^2-divergence from a mixture of distributions to a simple distribution is convenient.

Therefore the punchline is that it is often fruitful to bound one f-divergence by another and this sometimes leads to tight characterizations. In this section we consider a specific useful example to drive this point home. Then in the next lecture we develop inequalities between f-divergences systematically.

Consider a binary hypothesis test where data $X = (X_1, \ldots, X_n)$ are i.i.d drawn from either P or Q and the goal is to test

$$H_0 : X \sim P^{\otimes n} \quad \text{vs} \quad H_1 : X \sim Q^{\otimes n}.$$

As mentioned before, $1 - d_{\text{TV}}(P^{\otimes n}, Q^{\otimes n})$ gives minimal Type-I+II probabilities of error, achieved by the maximum likelihood test. By the data processing inequality, $d_{\text{TV}}(P^{\otimes m}, Q^{\otimes m}) \leq d_{\text{TV}}(P^{\otimes n}, Q^{\otimes n})$ for $m < n$. From this we see that $d_{\text{TV}}(P^{\otimes n}, Q^{\otimes n})$ is an increasing sequence in n (and bounded by 1 by definition) and hence converges. One would hope that as $n \to \infty$, $d_{\text{TV}}(P^{\otimes n}, Q^{\otimes n})$ converges to 1 and consequently, the probability of error in the hypothesis test converges to zero. It turns out that if the distributions P, Q are independent of n, then large deviation theory gives

$$d_{\text{TV}}(P^{\otimes n}, Q^{\otimes n}) = 1 - \exp(-nC(P, Q) + o(n)), \tag{4.9}$$

where the constant $C(P, Q) = -\log \inf_{0 \leq \alpha \leq 1} \int P^\alpha Q^{1-\alpha}$ is the **Chernoff Information** of P, Q. It is clear from this that $d_{\text{TV}}(P^{\otimes n}, Q^{\otimes n}) \to 1$ as $n \to \infty$, and, in fact, exponentially fast.

However, as frequently encountered in high-dimensional statistical problems, if the distributions $P = P_n$ and $Q = Q_n$ depend on n, then the large-deviation approach that leads to (4.9) is no longer valid. In such a situation, total variation is still relevant for hypothesis testing, but its behavior as $n \to \infty$ is not obvious nor easy to compute. In this case, understanding how a more computationally tractable f-divergence is related to total variation may give insight on hypothesis testing without needing to directly compute the total variation. It turns out Hellinger distance is precisely suited for this task – see Theorem 4.4 below.

Recall that the squared Hellinger distance, $H^2(P,Q) = \mathbb{E}_Q\left[\left(1-\sqrt{\frac{P}{Q}}\right)^2\right]$ is an f-divergence with $f(x) = (1-\sqrt{x})^2$, which provides a sandwich bound for total variation

$$0 \le \frac{1}{2}H^2(P,Q) \le d_{\text{TV}}(P,Q) \le H(P,Q)\sqrt{1-\frac{H^2(P,Q)}{4}} \le 1. \tag{4.10}$$

This can be proved by elementary manipulations and a systematic proof will be explained in the next lecture. Direct consequences of this bound are:

- $H^2(P,Q) = 2$, if and only if $d_{\text{TV}}(P,Q) = 1$.
- $H^2(P,Q) = 0$ if and only if $d_{\text{TV}}(P,Q) = 0$.
- Hellinger consistency ⇔ TV consistency, namely $H^2(P_n,Q_n) \to 0 \Leftrightarrow d_{\text{TV}}(P_n,Q_n) \to 0$.

Theorem 4.4. *For any sequence of distributions P_n and Q_n, as $n \to \infty$,*[4]

$$\begin{aligned} d_{\text{TV}}(P_n^{\otimes n}, Q_n^{\otimes n}) \to 0 &\Leftrightarrow H^2(P_n,Q_n) = o\left(\frac{1}{n}\right) \\ d_{\text{TV}}(P_n^{\otimes n}, Q_n^{\otimes n}) \to 1 &\Leftrightarrow H^2(P_n,Q_n) = \omega\left(\frac{1}{n}\right) \end{aligned}$$

Proof. Because the observations $X = (X_1, X_2, ...X_n)$ are i.i.d, the joint distribution factors

$$\begin{aligned} H^2(P_n^{\otimes n}, Q_n^{\otimes n}) &= 2 - 2\mathbb{E}_{Q_n^{\otimes n}}\left[\sqrt{\prod_{i=1}^n \frac{P_n}{Q_n}(X_i)}\right] \\ \text{By independence} \to &= 2 - 2\prod_{i=1}^n \mathbb{E}_{Q_n}\left[\sqrt{\frac{P_n}{Q_n}(X_i)}\right] = 2 - 2\left(\mathbb{E}_{Q_n}\left[\sqrt{\frac{P_n}{Q_n}}\right]\right)^n \\ &= 2 - 2\left(1 - \frac{1}{2}H^2(P_n,Q_n)\right)^n. \end{aligned}$$

Then $d_{\text{TV}}(P_n^{\otimes n}, Q_n^{\otimes n}) \to 0$ if and only if $H^2(P_n^{\otimes n}, Q_n^{\otimes n}) \to 0$, which happens precisely when $H^2(P_n,Q_n) = o(\frac{1}{n})$.

Similarly, $d_{\text{TV}}(P_n^{\otimes n}, Q_n^{\otimes n}) \to 1$ if and only if $H^2(P_n^{\otimes n}, Q_n^{\otimes n}) \to 2$ which happens precisely when $H^2(P_n,Q_n) = \omega(\frac{1}{n})$. □

Remark 4.8. The proof of Theorem 4.4 relies on two ingredients:

1. Sandwich bound (4.10).
2. Tensorization properties of Hellinger:

$$H^2\left(\prod_{i=1}^n P_i, \prod_{i=1}^n Q_i\right) = 2 - 2\prod_{i=1}^n\left(1 - \frac{H^2(P_i,Q_i)}{2}\right) \tag{4.11}$$

[4] For positive sequences $\{a_n\}, \{b_n\}$, we say $a_n = \omega(b_n)$ if $b_n = o(a_n)$.

Note that there are other f-divergences that are also tensorizable, e.g., χ^2-divergences:

$$\chi^2\left(\prod_{i=1}^n P_i, \prod_{i=1}^n Q_i\right) = \prod_{i=1}^n \left(1+\chi^2(P_i, Q_i)\right) - 1; \tag{4.12}$$

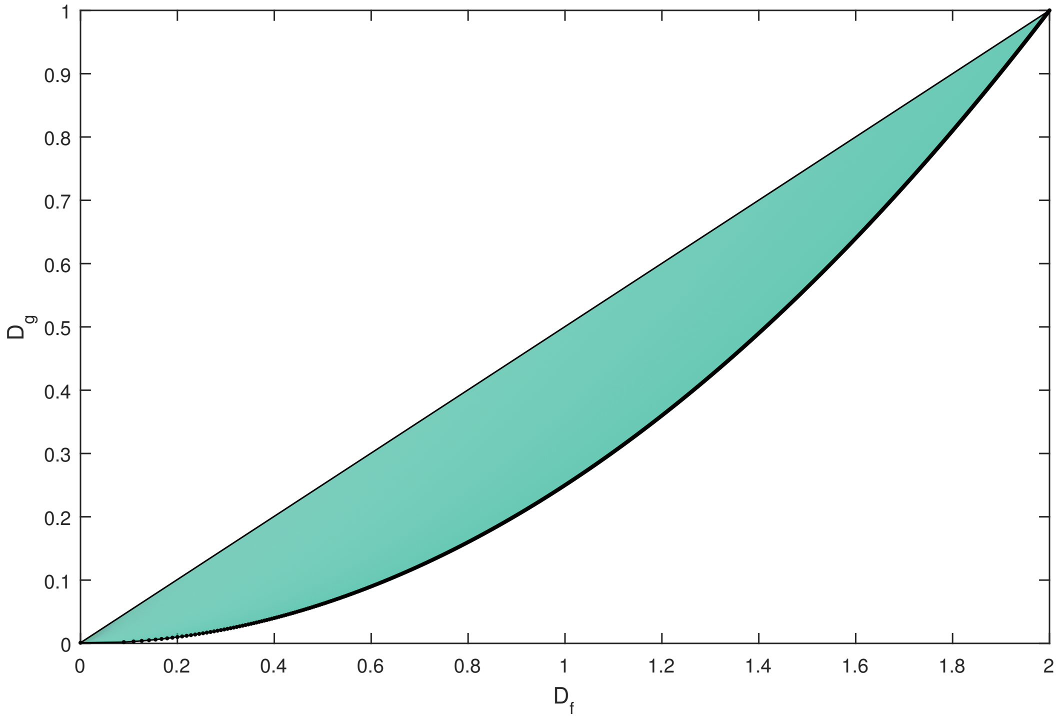
however, no sandwich inequality like (4.10) exists for d_{TV} and χ^2 and hence there is no χ^2-version of Theorem 4.4. Asserting the non-existence of such inequalities requires understanding the relationship between these two f-divergences (see next lecture).

4.5 Inequalities between f-divergences

We will discuss two methods for finding inequalities between f-divergences.

- ad hoc approach: case-by-case proof using results like Jensen's inequality, max $\leq$ mean $\leq$ min, Cauchy-Schwarz, etc.
- systematic approach: **joint range** of f-divergences.

Definition 4.2. The *joint range* between two f-divergences $D_f(\cdot\|\cdot)$ and $D_g(\cdot\|\cdot)$ is the range of the mapping $(P,Q) \mapsto (D_f(P\|Q), D_g(P\|Q))$, i.e., the set $\mathcal{R} \subset \mathbb{R}_+ \times \mathbb{R}_+$ where $(x,y) \in \mathcal{R}$ if there exist distributions P, Q on some common measurable space such that $x = D_f(P\|Q)$ and $y = D_g(P\|Q)$.



The green region in the above figure shows what a joint range between $D_f(\cdot\|\cdot)$ and $D_g(\cdot\|\cdot)$ might look like. By definition of $\mathcal{R}$, the lower boundary gives the sharpest lower bound of D_g in terms of D_f, namely:

$$D_f(P\|Q) \geq V(D_g(P\|Q)), \quad \text{where } V(t) \triangleq \inf\{D_f(P\|Q) : D_g(P\|Q) = t\};$$

similarly, the upper boundary gives the best upper bound. As will be discussed in the next lecture, the sandwich bound (4.10) correspond to precisely the lower and upper boundaries of the joint range of H^2 and d_{TV}, therefore not improvable. It is important to note, however, that $\mathcal{R}$ may be an unbounded region and some of the boundaries may not exist, meaning it is impossible to bound one by the other, such as χ^2 versus d_{TV}.

To gain some intuition, we start with the ad hoc approach by proving *Pinsker's inequality*, which bounds total variation from above by the KL divergence.

Theorem 4.5 (Pinsker's inequality)**.**

$$D(P\|Q) \geq 2\log e d_{\text{TV}}^2(P,Q). \tag{4.13}$$

Proof. First we show that, by the data processing inequality, it suffices to prove the result for Bernoulli distributions. For any event E, let $Y = \mathbf{1}\{X \in E\}$ which is Bernoulli with parameter $P(E)$ or $Q(E)$. By data processing inequality, $D(P\|Q) \geq d(P(E)\|Q(E))$. If Pinsker's inequality is true for all Bernoulli random variables, we have

$$\sqrt{\frac{1}{2}D(P\|Q)} \geq d_{\text{TV}}(\text{Bern}(P(E)), \text{Bern}(Q(E)) = |P(E) - Q(E)|$$

Taking the supremum over E gives $\sqrt{\frac{1}{2}D(P\|Q)} \geq \sup_E |P(E) - Q(E)| = d_{\text{TV}}(P,Q)$, in view of Theorem 4.3.

The binary case follows easily from Taylor's theorem (with integral remainder form):

$$d(p\|q) = \int_q^p \frac{p-t}{t(1-t)}dt \geq 4\int_q^p (p-t)dt = 2(p-q)^2$$

and $d_{\text{TV}}(\text{Bern}(p), \text{Bern}(q)) = |p-q|$. □

Remark 4.9. Pinsker's inequality is known to be sharp in the sense that the constant "2" in (4.13) is not improvable, i.e., there exist $\{P_n, Q_n\}$ such that $\frac{\text{LHS}}{\text{RHS}} \to 2$ as $n \to \infty$. (Why?) Nevertheless, this does not mean that (4.13) itself is not improvable because it might be possible to subtract some higher-order term from the RHS. This is indeed the case and there are many refinements of Pinsker's inequality. But what is the best inequality? Settling this question rests on characterizing the joint range and the lower boundary. This is the topic of next lecture.

In the last lecture we proved the Pinkser's inequality that $D(P\|Q) \geq 2d_{\rm TV}^2(P,Q)$ in an ad hoc manner. The downside of ad hoc approaches is that it is hard to tell whether those inequalities can be improved or not. However, the key step when we proved the Pinkser's inequality, reduction to the case for Bernoulli random variables, is inspiring: is it possible to reduce inequalities between any two f-divergences to the binary case?

5.1 Inequalities via joint range

A systematic method is to prove those inequalities via their joint range. For example, to prove a lower bound of $D(P\|Q)$ by a function of $d_{\rm TV}(P,Q)$ that $D(P\|Q) \geq F(d_{\rm TV}(P,Q))$ for some $F : [0,1] \mapsto [0,\infty]$, the best choice, by definition, is the following:

$$F(\epsilon) \triangleq \inf_{(P,Q):d_{\rm TV}(P,Q)=\epsilon} D(P\|Q).$$

The problem boils to the characterization of the region $\{(d_{\rm TV}(P,Q), D(P\|Q)) : P,Q\} \subseteq \mathbb{R}^2$, their joint range, whose lower boundary is the function F.

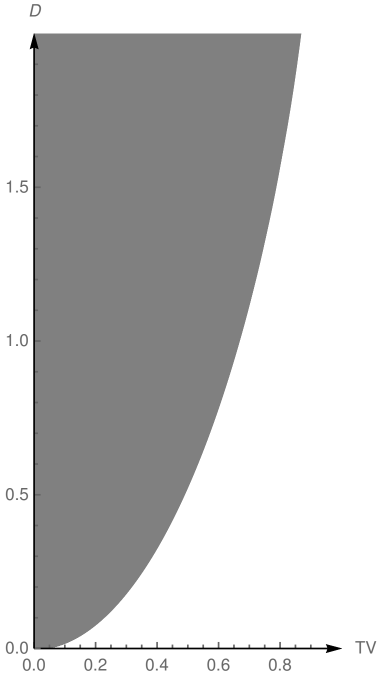


Figure 5.1: Joint range of $d_{\rm TV}$ and D.

Definition 5.1 (Joint range)**.** Consider two f-divergences $D_f(P\|Q)$ and $D_g(P\|Q)$. Their joint range is a subset of $\mathbb{R}^2$ defined by

$$\begin{aligned}
\mathcal{R} &\triangleq \{(D_f(P\|Q), D_g(P\|Q)) : P, Q \text{ are probability measures on some measurable space}\}\,,\\
\mathcal{R}_k &\triangleq \{(D_f(P\|Q), D_g(P\|Q)) : P, Q \text{ are probability measures on } [k]\}\,.
\end{aligned}$$

The region $\mathcal{R}$ seems difficult to characterize since we need to consider P, Q over all measurable spaces; on the other hand, the region $\mathcal{R}_k$ for small k is easy to obtain. The main theorem we will prove is the following [HV11]:

Theorem 5.1 (Harremoës-Vajda '11)**.**

$$\mathcal{R} = \mathrm{co}(\mathcal{R}_2).$$

It is easy to obtain a parametric formula of $\mathcal{R}_2$. By Theorem 5.1, the region $\mathcal{R}$ is no more than the convex hull of $\mathcal{R}_2$.

Theorem 5.1 implies that $\mathcal{R}$ is a convex set; however, as a warmup, it is instructive to prove convexity of $\mathcal{R}$ directly, which simply follows from the arbitrariness of the alphabet size of distributions. Given any two points $(D_f(P_0\|Q_0), D_g(P_0\|Q_0))$ and $(D_f(P_1\|Q_1), D_g(P_1\|Q_1))$ in the joint range, it is easy to construct another pair of distributions (P, Q) by alphabet extension that produces any convex combination of those two points.

Theorem 5.2. *$\mathcal{R}$ is convex.*

Proof. Given any two pairs of distributions (P_0, Q_0) and (P_1, Q_1) on some space $\mathcal{X}$ and given any $\alpha \in [0, 1]$, we define another pair of distributions (P, Q) on $\mathcal{X} \times \{0, 1\}$ by

$$\begin{aligned}
P &= \bar{\alpha}(P_0 \times \delta_0) + \alpha(P_1 \times \delta_1),\\
Q &= \bar{\alpha}(Q_0 \times \delta_0) + \alpha(Q_1 \times \delta_1).
\end{aligned}$$

In other words, we construct a random variable $Z = (X, B)$ with $B \sim \mathrm{Bern}(\alpha)$, where $P_{X|B=i} = P_i$ and $Q_{X|B=i} = Q_i$. Then

$$\begin{aligned}
D_f(P\|Q) &= \mathbb{E}_Q\left[f\left(\frac{P}{Q}\right)\right] = \mathbb{E}_B\left[\mathbb{E}_{Q_{Z|B}}\left[f\left(\frac{P}{Q}\right)\right]\right] = \bar{\alpha}D_f(P_0\|Q_0) + \alpha D_f(P_1\|Q_1),\\
D_g(P\|Q) &= \mathbb{E}_Q\left[g\left(\frac{P}{Q}\right)\right] = \mathbb{E}_B\left[\mathbb{E}_{Q_{Z|B}}\left[g\left(\frac{P}{Q}\right)\right]\right] = \bar{\alpha}D_g(P_0\|Q_0) + \alpha D_g(P_1\|Q_1).
\end{aligned}$$

Therefore, $\bar{\alpha}(D_f(P_0\|Q_0), D_g(P_0\|Q_0)) + \alpha(D_f(P_1\|Q_1), D_g(P_1\|Q_1)) \in \mathcal{R}$ and thus $\mathcal{R}$ is convex. □

Theorem 5.1 is proved by the following two lemmas:

Lemma 5.1 (non-constructive/existential)**.** $\mathcal{R} = \mathcal{R}_4$.

Lemma 5.2 (constructive/algorithmic)**.**

$$\mathcal{R}_{k+1} = \mathrm{co}(\mathcal{R}_2 \cup \mathcal{R}_k) \qquad \textit{for any } k \geq 2$$

and hence

$$\mathcal{R}_k = \mathrm{co}(\mathcal{R}_2), \qquad \textit{for any } k \geq 3.$$

5.1.1 Representation of f-divergences

To prove Lemma 5.1 and Lemma 5.2, we first express f-divergences by means of expectation over the likelihood ratio.

Lemma 5.3. *Given two f-divergences $D_f(\cdot\|\cdot)$ and $D_g(\cdot\|\cdot)$, their joint range is*

$$\mathcal{R} = \left\{ \begin{pmatrix} \mathbb{E}[f(X)] + \tilde{f}(0)(1-\mathbb{E}[X]) \\ \mathbb{E}[g(X)] + \tilde{g}(0)(1-\mathbb{E}[X]) \end{pmatrix} : X \geq 0, \mathbb{E}[X] \leq 1 \right\},$$

$$\mathcal{R}_k = \left\{ \begin{pmatrix} \mathbb{E}[f(X)] + \tilde{f}(0)(1-\mathbb{E}[X]) \\ \mathbb{E}[g(X)] + \tilde{g}(0)(1-\mathbb{E}[X]) \end{pmatrix} : \begin{array}{c} X \geq 0, \mathbb{E}[X] \leq 1, X \textit{ takes at most } k-1 \textit{ values,} \\ \textit{or } X \geq 0, \mathbb{E}[X] = 1, X \textit{ takes at most } k \textit{ values} \end{array} \right\},$$

where $\tilde{f}(0) \triangleq \lim_{x\to 0} xf(1/x)$ and $\tilde{g}(0) \triangleq \lim_{x\to 0} xg(1/x)$.

In the statement of Lemma 5.3, we remark that $\tilde{f}(0)$ and $\tilde{g}(0)$ are both well-defined (possibly $+\infty$) by the convexity of $x \mapsto xf(1/x)$ and $x \mapsto xg(1/x)$ (from the last lecture).

Before proving above lemma, we look at the following two examples to understand the correspondence between a point in the joint range and a random variable. The first example is the simple case that $P \ll Q$, when the likelihood ratio of P and Q (or Radon-Nikodym derivative defined on the union of the spaces of P and Q) is well-define.

Example 5.1. Consider the following two distributions P, Q on [3]:

	1	2	3
P	0.34	0.34	0.32
Q	0.85	0.1	0.05

Then $D_f(P\|Q) = 0.85f(0.4) + 0.1f(3.4) + 0.05f(6.4)$, which is $\mathbb{E}[f(X)]$ where X is the likelihood ratio of P and Q taking 3 values with the following pmf:

x	0.4	3.4	6.4
$\mu(x)$	0.85	0.1	0.05

On the other direction, given the above pmf of a non-negative, unit-mean random variable $X \sim \mu$ that takes 3 values, we can construct a pair of distribution by $Q(x) = \mu(x)$ and $P(x) = x\mu(x)$.

In general cases P is not necessarily absolutely continuous w.r.t. Q, and the likelihood ratio X may not always exist. However, it is still well-defined on the event $\{Q > 0\}$.

Example 5.2. Consider the following two distributions P, Q on [2]:

	1	2
P	0.4	0.6
Q	0	1

Then $D_f(P\|Q) = f(0.6) + 0f(\frac{0.4}{0})$, where $0f(\frac{p}{0})$ is understood as

$$0f\left(\frac{p}{0}\right) = \lim_{x\to 0} xf\left(\frac{p}{x}\right) = p \lim_{x\to 0} \frac{x}{p} f\left(\frac{p}{x}\right) = p\tilde{f}(0).$$

Therefore $D_f(P\|Q) = f(0.6) + 0.4\tilde{f}(0) = \mathbb{E}[f(X)] + \tilde{f}(0)(1-\mathbb{E}[X])$ where X is defined on $\{Q > 0\}$:

x	0.6
$\mu(x)$	1

On the other direction, given above pmf of a non-negative random variable $X \sim \mu$ with $\mathbb{E}[X] \le 1$ that takes 1 value, we let $Q(x) = \mu(x)$, let $P(x) = x\mu(x)$ on $\{Q > 0\}$ and let P have an extra point mass $1 - \mathbb{E}[X]$.

Proof of Lemma 5.3. We first prove it for $\mathcal{R}$. Given any pair of distributions (P, Q) that produces a point of $\mathcal{R}$, let p, q denote the densities of P, Q under some dominating measure μ, respectively. Let

$$X = \frac{p}{q} \text{ on } \{q > 0\}, \quad \mu_X = Q, \tag{5.1}$$

then $X \ge 0$ and $\mathbb{E}[X] = P[q > 0] \le 1$. Then

$$\begin{aligned} D_f(P\|Q) = \int_{\{q>0\}} f\left(\frac{p}{q}\right) \mathrm{d}Q + \int_{\{q=0\}} \frac{q}{p} f\left(\frac{p}{q}\right) \mathrm{d}P &= \int_{\{q>0\}} f\left(\frac{p}{q}\right) \mathrm{d}Q + \tilde{f}(0)P[q = 0] \\ &= \mathbb{E}[f(X)] + \tilde{f}(0)(1 - \mathbb{E}[X]), \end{aligned}$$

Analogously,

$$D_g(P\|Q) = \mathbb{E}[g(X)] + \tilde{g}(0)(1 - \mathbb{E}[X]),$$

On the other direction, given any random variable $X \ge 0$ and $\mathbb{E}[X] \le 1$ where $X \sim \mu$, let

$$\mathrm{d}Q = \mathrm{d}\mu, \quad \mathrm{d}P = X\mathrm{d}\mu + (1 - \mathbb{E}[X])\delta_*, \tag{5.2}$$

where $*$ is an arbitrary symbol outside the support of X. Then

$$\begin{pmatrix} D_f(P\|Q) \\ D_g(P\|Q) \end{pmatrix} = \begin{pmatrix} \mathbb{E}[f(X)] + \tilde{f}(0)(1 - \mathbb{E}[X]) \\ \mathbb{E}[g(X)] + \tilde{g}(0)(1 - \mathbb{E}[X]) \end{pmatrix}.$$

Now we consider $\mathcal{R}_k$. Given two probability measures P, Q on $[k]$, the likelihood ratio defined in (5.1) takes at most k values. If $P \ll Q$ then $\mathbb{E}[X] = 1$; if $P \not\ll Q$ then X takes at most $k - 1$ values.

On the other direction, if $\mathbb{E}[X] = 1$ then the construction of P, Q in (5.2) are on the same support of X; if $\mathbb{E}[X] < 1$ then the support of P is increased by one. □

5.1.2 Proof of Theorem 5.1

Aside: Fenchel-Eggleston-Carathéodory's theorem: Let $S \subseteq \mathbb{R}^d$ and $x \in \mathrm{co}(S)$. Then there exists a set of $d+1$ points $S' = \{x_1, x_2, \ldots, x_{d+1}\} \in S$ such that $x \in \mathrm{co}(S')$. If S is connected, then d points are enough.

Proof of Lemma 5.1. It suffices to prove that

$$\mathcal{R} \subseteq \mathcal{R}_4.$$

Let $S \triangleq \{(x, f(x), g(x)) : x \ge 0\}$ which is a connected set. For any pair of distributions (P, Q) that produces a point of $\mathcal{R}$, we construct a random variable X as in (5.1), then $(\mathbb{E}[X], \mathbb{E}[f(X)], \mathbb{E}[g(X)]) \in \mathrm{co}(S)$. By Fenchel-Eggleston-Carathéodory's theorem,[1] there exists $(x_i, f(x_i), g(x_i))$ and the corresponding weight α_i for $i = 1, 2, 3$ such that

$$(\mathbb{E}[X], \mathbb{E}[f(X)], \mathbb{E}[g(X)]) = \sum_{i=1}^{3} \alpha_i (x_i, f(x_i), g(x_i)).$$

[1] To prove Theorem 5.1, it suffices to invoke the basic Carathéodory's theorem, which proves a weaker version of Lemma 5.1 that $\mathcal{R} = \mathcal{R}_5$.

We construct another random variable X' that takes value x_i with probability α_i. Then X takes 3 values and

$$(\mathbb{E}[X], \mathbb{E}[f(X)], \mathbb{E}[g(X)]) = (\mathbb{E}[X'], \mathbb{E}[f(X')], \mathbb{E}[g(X')]). \tag{5.3}$$

By Lemma 5.3 and (5.3),

$$\begin{pmatrix} D_f(P\|Q) \\ D_g(P\|Q) \end{pmatrix} = \begin{pmatrix} \mathbb{E}[f(X)] + \tilde{f}(0)(1-\mathbb{E}[X]) \\ \mathbb{E}[g(X)] + \tilde{g}(0)(1-\mathbb{E}[X]) \end{pmatrix} = \begin{pmatrix} \mathbb{E}[f(X')] + \tilde{f}(0)(1-\mathbb{E}[X']) \\ \mathbb{E}[g(X')] + \tilde{g}(0)(1-\mathbb{E}[X']) \end{pmatrix} \in \mathcal{R}_4.$$

□

We observe from Lemma 5.3 that $D_f(P\|Q)$ only depends on the distribution of X for some $X \geq 0$ and $\mathbb{E}[X] \leq 1$. To find a pair of distributions that produce a point in $\mathcal{R}_k$ it suffices to find a random variable $X \geq 0$ taking k values with $\mathbb{E}[X] = 1$, or taking $k-1$ values with $\mathbb{E}[X] \leq 1$. In Example 5.1 where (P, Q) produces a point in $\mathcal{R}_3$, we want to show that it also belongs to $\text{co}(\mathcal{R}_2)$. The decomposition of a point in $\mathcal{R}_3$ is equivalent to the decomposition of the likelihood ratio X that

$$\mu_X = \alpha\mu_1 + \bar{\alpha}\mu_2.$$

A solution of such decomposition is that $\mu_X = 0.5\mu_1 + 0.5\mu_2$ where μ_1, μ_2 has the following pmf:

x	0.4	3.4
$\mu_1(x)$	0.8	0.2

x	0.4	6.4
$\mu_2(x)$	0.9	0.1

Then by (5.2) we obtain two pairs of distributions

P_1	0.32	0.68
Q_1	0.8	0.2

P_2	0.36	0.64
Q_2	0.9	0.1

We obtain that

$$\begin{pmatrix} D_f(P\|Q) \\ D_g(P\|Q) \end{pmatrix} = 0.5 \begin{pmatrix} D_f(P_1\|Q_1) \\ D_g(P_1\|Q_1) \end{pmatrix} + 0.5 \begin{pmatrix} D_f(P_2\|Q_2) \\ D_g(P_2\|Q_2) \end{pmatrix}.$$

Proof of Lemma 5.2. It suffices to prove the first statement, namely, $\mathcal{R}_{k+1} = \text{co}(\mathcal{R}_k)$ for any $k \geq 2$. Since $\mathcal{R}_k \subseteq \mathcal{R}_{k+1}$ by definition, it remains to show $\mathcal{R}_{k+1} \subseteq \text{co}(\mathcal{R}_k)$.

Given any pair of distributions (P, Q) that produces a point of $(D_f(P\|Q), D_g(P\|Q)) \in \mathcal{R}_{k+1}$, we construct a random variable X as in (5.1) that takes at most $k+1$ values. Let μ denote the distribution of X. We consider two cases that $\mathbb{E}_\mu[X] < 1$ and $\mathbb{E}_\mu[X] = 1$ separately.

- $\mathbb{E}_\mu[X] < 1$. Then X takes at most k values since otherwise $P \ll Q$. Denote the smallest value of X by x and then $x < 1$. Suppose $\mu(x) = q$ and then μ can be represented as

  $$\mu = q\delta_x + \bar{q}\mu',$$

  where μ' is supported on at most $k-1$ values of X other than x. Let $\mu_2 = \delta_x$. We need to construct another probability measure μ_1 such that

  $$\mu = \alpha\mu_1 + \bar{\alpha}\mu_2,$$

  - $\mathbb{E}_{\mu'}[X] \leq 1$. Let $\mu_1 = \mu'$ and let $\alpha = \bar{q}$.
  - $\mathbb{E}_{\mu'}[X] > 1$. Let $\mu_1 = p\delta_x + \bar{p}\mu'$ where $p = \frac{\mathbb{E}_{\mu'}[X]-1}{\mathbb{E}_{\mu'}[X]-x}$ such that $\mathbb{E}_{\mu_1}[X] = 1$. Let $\alpha = \frac{\mathbb{E}_\mu[X]-x}{1-x}$.

- $\mathbb{E}_\mu[X] = 1$.[2] Denote the smallest value of X by x and the largest value by y, respectively, and then $x \le 1, y \ge 1$. Suppose $\mu(x) = r$ and $\mu(y) = s$ and then μ can be represented as

$$\mu = r\delta_x + (1 - r - s)\mu' + s\delta_y,$$

where μ' is supported on at most $k-1$ values of X other than x, y. Let $\mu_2 = \beta\delta_x + \bar{\beta}\delta_y$ where $\beta = \frac{y-1}{y-x}$ such that $\mathbb{E}_{\mu_2}[X] = 1$. We need to construct another probability measure μ_1 such that

$$\mu = \alpha\mu_1 + \bar{\alpha}\mu_2,$$

  - $\mathbb{E}_{\mu'}[X] \le 1$. Let $\mu_1 = p\delta_y + \bar{p}\mu'$ where $p = \frac{1-\mathbb{E}_{\mu'}[X]}{y-\mathbb{E}_{\mu'}[X]}$ such that $\mathbb{E}_{\mu_1}[X] = 1$. Let $\bar{\alpha} = r/\beta$.
  - $\mathbb{E}_{\mu'}[X] > 1$. Let $\mu_1 = p\delta_x + \bar{p}\mu'$ where $p = \frac{\mathbb{E}_{\mu'}[X]-1}{\mathbb{E}_{\mu'}[X]-x}$ such that $\mathbb{E}_{\mu_1}[X] = 1$. Let $\bar{\alpha} = s/\bar{\beta}$.

Applying the construction in (5.2) with μ_1 and μ_2, we obtain two pairs of distributions (P_1, Q_1) supported on k values and (P_2, Q_2) supported on two values, respectively. Then

$$\begin{aligned}
\begin{pmatrix} D_f(P\|Q) \\ D_g(P\|Q) \end{pmatrix} &= \begin{pmatrix} \mathbb{E}_\mu[f(X)] + \tilde{f}(0)(1 - \mathbb{E}_\mu[X]) \\ \mathbb{E}_\mu[g(X)] + \tilde{g}(0)(1 - \mathbb{E}_\mu[X]) \end{pmatrix} \\
&= \alpha \begin{pmatrix} \mathbb{E}_{\mu_1}[f(X)] + \tilde{f}(0)(1 - \mathbb{E}_{\mu_1}[X]) \\ \mathbb{E}_{\mu_1}[g(X)] + \tilde{g}(0)(1 - \mathbb{E}_{\mu_1}[X]) \end{pmatrix} + \bar{\alpha} \begin{pmatrix} \mathbb{E}_{\mu_2}[f(X)] + \tilde{f}(0)(1 - \mathbb{E}_{\mu_2}[X]) \\ \mathbb{E}_{\mu_2}[g(X)] + \tilde{g}(0)(1 - \mathbb{E}_{\mu_2}[X]) \end{pmatrix} \\
&= \alpha \begin{pmatrix} D_f(P_1\|Q_1) \\ D_g(P_1\|Q_1) \end{pmatrix} + \bar{\alpha} \begin{pmatrix} D_f(P_2\|Q_2) \\ D_g(P_2\|Q_2) \end{pmatrix}.
\end{aligned}$$

□

Remark 5.1. Theorem 5.1 can be viewed as a consequence of Krein-Milman's theorem. Consider the space of $\{P_X : X \ge 0, \mathbb{E}[X] \le 1\}$, which has only two types of extreme points:

1. $X = x$ for $0 \le x \le 1$;
2. X takes two values x_1, x_2 with probability α_1, α_2, respectively, and $\mathbb{E}[X] = 1$.

For the first case, let $P = \text{Bern}(x)$ and $Q = \delta_1$; for the second case, let $P = \text{Bern}(\alpha_2 x_2)$ and $Q = \text{Bern}(\alpha_2)$.

5.2 Examples

5.2.1 Hellinger distance versus total variation

The upper and lower bound we mentioned in the last lecture is the following [Tsy09, Sec. 2.4]:

$$\frac{1}{2}H^2 \le d_{\text{TV}} \le H\sqrt{1 - H^2/4}. \tag{5.4}$$

Equivalently, in terms of Hellinger affinity $\alpha \triangleq \int \sqrt{PQ}$, we have the so-called Bhattacharya bound:

$$1 - \alpha \le d_{\text{TV}} \le \sqrt{1 - \alpha^2}.$$

[2] The author is grateful to Pengkun Yang for correcting the error in the original proof.

The joint range $\mathcal{R}_2$ of (d_{TV}, H^2) has a parametric formula

$$\left\{(2(1-\sqrt{pq}-\sqrt{\bar{p}\bar{q}}), |p-q|) : 0 \le p \le 1, 0 \le q \le 1\right\}$$

and is the gray region in Fig. 5.2. The joint range $\mathcal{R}$ is the convex hull of $\mathcal{R}_2$ (grey region, non-convex) and exactly described by (5.4); so (5.4) is not improvable. Indeed, with t ranges from 0 to 1,

- the upper boundary is achieved by $P = \text{Bern}(\frac{1+t}{2}), Q = \text{Bern}(\frac{1-t}{2})$,
- the lower boundary is achieved by $P = (1-t, t, 0), Q = (1-t, 0, t)$.

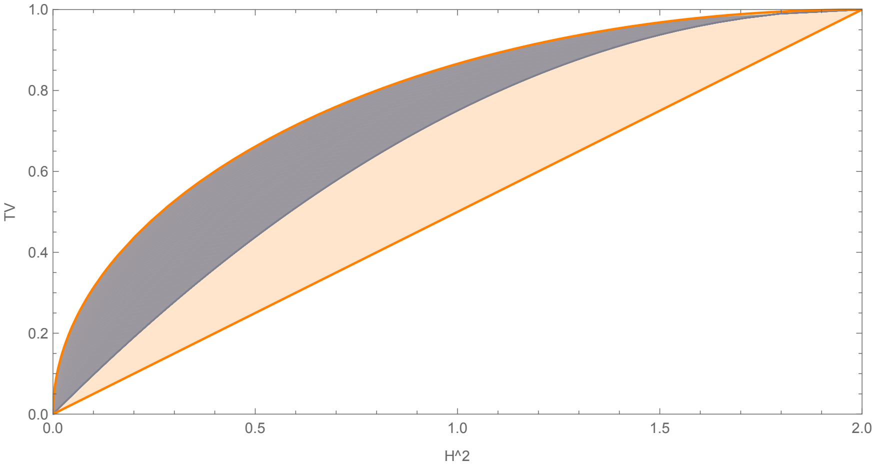


Figure 5.2: Joint range of d_{TV} and H^2.

5.2.2 KL divergence versus total variation

Pinsker's inequality states that

$$D(P\|Q) \ge 2d_{\text{TV}}^2(P,Q). \tag{5.5}$$

There are various kinds of improvements of Pinsker's inequality. Now we know that the best lower bound is the lower boundary of Fig. 5.1, which is exactly the boundary of $\mathcal{R}_2$. Therefore a paremetric formula of the lower boundary is easy to write down, but there is no known close-form expression. Here is a corollary (weaker bound) due to Vadja [Vaj70]:

$$D(P\|Q) \ge \log\frac{1+d_{\text{TV}}(P,Q)}{1-d_{\text{TV}}(P,Q)} - \frac{2d_{\text{TV}}(P,Q)}{1+d_{\text{TV}}(P,Q)}.$$

Consequences:

- The original Pinsker's inequality shows that $D \to 0 \Rightarrow d_{\text{TV}} \to 0$.
- $d_{\text{TV}} \to 1 \Rightarrow D \to \infty$. Thus $D = O(1) \Rightarrow d_{\text{TV}}$ is bounded away from one. This is not obtainable from Pinsker's inequality.

Also from Fig. 5.1 we know that it is impossible to have an upper bound of $D(P\|Q)$ using a function of $d_{\text{TV}}(P,Q)$ due to the lack of upper boundary.

For more examples see [Tsy09, Sec. 2.4].

§ 6. Variational representation, HCR and CR lower bounds.

Last lecture we discussed systematic methods to find the best inequalities between different f-divergence via their joint range. We showed that examining the binary cases is sufficient to derive optimal inequalities. In this lecture we will further discuss lower bounds for statistical estimation using f-divergences.

Outline:

- Variational representation of f-divergences.
  - Convexity.
  - Lower semi-continuity.
- (Specializing to χ^2) Lower bounds for statistical estimation.
  - Hammersley-Chapman-Robbins (HCR) lower bound.
  - Cramér-Rao (CR) lower bound.
  - Bayesian Hammersley-Chapman-Robbins (HCR) lower bound.
  - Bayesian Cramér-Rao (CR) lower bound.

6.1 Variational representation of f-divergences

We begin with an example regarding the total variation metric.

Example 6.1 (Total variation)**.** Let $(\mathcal{X}, \mathcal{F})$ a measure space and P, Q two probability distributions. In previous lectures we saw how by choosing $f(x) = \frac{1}{2}|x-1|$ the f-divergence becomes the total variation metric. In particular, we saw that:

$$d_{\mathrm{TV}}(P, Q) = D_f(P\|Q) = \frac{1}{2}\int |P - Q| = \sup_{E\in\mathcal{F}} |P(E) - Q(E)| = \frac{1}{2}\sup_{\|f\|_\infty \le 1} |\mathbb{E}_P f(x) - \mathbb{E}_Q f(x)|.$$

It should be noted that the requirement of f to be convex in the definition of f-divergence is essential. In Euclidean spaces any convex function can be represented as the pointwise supremum of a family of affine functions and vice versa, every supremum of a family of affine functions produces a convex function. Take $f(x) = \frac{1}{2}|x-1|$ as an example. We see that it can be written as a pointwise supremum of $f_1(x) = \frac{1}{2}(x-1)$ and $f_2(x) = \frac{1}{2}(1-x)$. This remark can be used not only as a geometric interpretation of convex functions but as a definition of convexity. For f-divergences which are convex functions of probability measures, its variational representation amounts to writing it as a pointwise supremum of affine functions.

6.1.1 Convex conjugate

Let $f : (0, +\infty) \to \mathbb{R}$ be a convex function. The convex conjugate f^* of f is defined by:

$$f^*(y) = \sup_{x \in \mathbb{R}} [xy - f(x)]. \tag{6.1}$$

Two important properties of the convex conjugates are

1. f^* is also convex (which holds regardless of f being convex or not);
2. Biconjugation: $(f^*)^* = f$.

In particular, the definition of f^* yields the following (Young-Fenchel inequality)

$$f(x) \geq xy - f^*(y), \tag{6.2}$$

where the last inequality holds for any y.

Using the notion of convex conjugate, we obtain a variational representation of f-divergence in terms of the convex conjugate of f:[1]

$$D_f(P\|Q) = \mathbb{E}_Q\left[f\left(\frac{P}{Q}\right)\right] = \sup_{g:\mathcal{X}\to\mathbb{R}} \mathbb{E}_P[g(X)] - \mathbb{E}_Q[f^*(g(X))], \tag{6.3}$$

where g is such that both expectations are finite (of course). This representation is insightful for many reasons. For example, we get the following properties for free:

1. Convexity: First of all, note that $D_f(P\|Q)$ is expressed as a supremum of affine functions (since the expectation is a linear operation). As a result, we get that $(P, Q) \mapsto D_f(P\|Q)$ is convex, which was proved in previous lectures using different method.
2. Weak lower semicontinuity: We begin with an example. Assume $\{X_i\}$ are i.i.d. Rademachers (± 1). Then, by the central limit theorem we have that

   $$\frac{\sum_{i=1}^n X_i}{\sqrt{n}} \xrightarrow{\text{D}} \mathcal{N}(0,1);$$

   however,

   $$D_f\left(\frac{P_{X_1+X_2+\ldots+X_n}}{\sqrt{n}} \middle\| \mathcal{N}(0,1)\right) \not\to 0,$$

   since the former distribution is discrete and the latter is continuous. Therefore the best we can hope for f-divergence is semicontinuity. Indeed, if $\mathcal{X}$ is a nice space (e.g., Euclidean space), in (6.3) we can restrict the function g to continuous bounded functions, in which case $D_f(P\|Q)$ is expressed as a supremum of weakly continuous functionals (note that $f^* \circ g$ is also continuous and bounded since f^* is continuous) and is hence weakly lower semi-continuous, i.e., for any sequence of distributions P_n and Q_n such that $P_n \xrightarrow{w} P$ and $Q_n \xrightarrow{w} Q$, we have

   $$\liminf_{n\to\infty} D_f(P_n\|Q_n) \geq D_f(P\|Q).$$

[1] Equivalently, one can take the supremum over all kernels $P_{Z|X}$ where Z is $\mathbb{R}$-valued.

Example 6.2 (Total variation)**.** By using $f(x) = \frac{1}{2}|x-1|$ in the formula of f-divergence we get the total variation metric given by

$$d_{TV}(P,Q) = \frac{1}{2}\int |P-Q|.$$

By using the definition of convex conjugate it is easy to see that

$$f^*(y) = \sup_x \left\{ xy - \frac{1}{2}|x-1| \right\} = \begin{cases} +\infty & \text{if } |y| > \frac{1}{2} \\ y & \text{if } |y| \leq \frac{1}{2} \end{cases}$$

Thus (6.3) gives

$$d_{TV}(P,Q) = \sup_{g:\mathcal{X}\to\mathbb{R}} \mathbb{E}_P[g(X)] - \mathbb{E}_Q[f^*(g(X))] = \sup_{g:|g|\leq\frac{1}{2}} \mathbb{E}_P[g(X)] - \mathbb{E}_Q[g(X)], \tag{6.4}$$

where in the last equality we restricted the supremum to functions bounded by 1/2, since any other function would make the term inside the supremum equal to $-\infty$.

Example 6.3 (KL-divergence)**.** By using $f(x) = x\log x$ in the formula of f-divergence we get the KL-divergence

$$D(P\|Q) = \mathbb{E}_P\left[\log\frac{P}{Q}\right].$$

By using differentiation to find the supremum it is easy to see that $f^*(y) = e^{y-1}$. Plugging in the formula of f-divergence we get

$$D(P\|Q) = 1 + \sup_{g:\mathcal{X}\to\mathbb{R}} \mathbb{E}_P[g(X)] - \mathbb{E}_Q[e^{g(X)}]. \tag{6.5}$$

In comparison, the famous Donsker-Varadhan representation is

$$D(P\|Q) = \sup_g \mathbb{E}_P[g(X)] - \log\mathbb{E}_Q[e^{g(X)}], \tag{6.6}$$

which is stronger than (6.5) in the sense that for each g, the RHS of (6.6) is at least that of (6.5), since $\log(1+t) \leq t$.

Example 6.4 (χ^2-divergence)**.** By using $f(x) = (x-1)^2$ in the formula of f-divergence we get the χ^2-divergence

$$\chi^2(P\|Q) = \mathbb{E}_Q\left[\left(\frac{P}{Q}-1\right)^2\right] = \mathrm{var}_Q\left(\frac{P}{Q}\right).$$

By using differentiation to find the supremum it is easy to see that $f^*(y) = y + \frac{y^2}{4}$. Hence

$$\chi^2(P\|Q) = \sup_{g:\mathcal{X}\to\mathbb{R}} \mathbb{E}_P[g(X)] - \mathbb{E}_Q\left[g(X) + \frac{g^2(X)}{4}\right].$$

Finally by a change of variable $h(x) = \frac{1}{2}g(x) + 1$ we get

$$\chi^2(P\|Q) = \sup_{h:\mathcal{X}\to\mathbb{R}} 2\mathbb{E}_P[h(X)] - \mathbb{E}_Q[h^2(X)] - 1. \tag{6.7}$$

It is not hard to see that we also have a more symmetric version which is directly related to bias and variance tradeoff:

$$\chi^2(P\|Q) = \sup_{h:\mathcal{X}\to\mathbb{R}} \frac{(\mathbb{E}_P[h(X)] - \mathbb{E}_Q[h(X)])^2}{\mathrm{var}_Q(h(X))}. \tag{6.8}$$

The representation (6.7) will be used repeatedly for the derivation of the Hammersley-Chapman-Robbins (HCR) lower bound as well as their Bayesian version in the next section.

6.2 Hammersley-Chapman-Robbins (HCR) lower bound

In this section, we apply the variational representation for the χ^2-divergence to probability distributions P and Q on $\mathbb{R}$.[2] By limiting the choice of function h to affine functions, the equality (6.7) becomes an inequality. In particular, let $h(x) = ax + b$ and optimize over $a, b \in \mathbb{R}$, we have

$$\chi^2(P\|Q) \geq \sup_{a,b\in\mathbb{R}} \left\{2(a\mathbb{E}_P(X) + b) - \mathbb{E}_Q[(aX + b)^2] - 1\right\} = \frac{(\mathbb{E}_P[X] - \mathbb{E}_Q[X])^2}{\operatorname{var}_Q(X)}. \tag{6.9}$$

Note: The inequality (6.9) can be interpreted as follows: On the left hand side of the inequality we have the χ^2-divergence, a measure of the dissimilarity between two distributions. Looking at the right hand side we see that if the two distributions are centered at very distant locations, then the right hand side will be large. Due to (6.9), this will lead to a bigger χ^2-divergence something that was in fact expected.

The reason that the variance with respect to the Q distribution appears in the denominator is to quantify how different the two means are *relatively*. Indeed, the standard deviation must appear as a normalizing factor because the LHS is a numerical number. Also, the bound only involves the variance under Q not P, which is consistent with the asymmetry of χ^2-divergence.

Using (6.7) we now derive the HCR lower bound on the variance of an estimator (possibly randomized). To this end, assume that data $X \sim P_\theta$, where $\theta \in \Theta \subset \mathbb{R}$. We use quadratic cost to quantify the difference between the real and the predicted parameter, i.e., $\ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$. Then the risk of estimator $\hat{\theta}$ when the real parameter is θ is given by $R_\theta(\hat{\theta}) = \mathbb{E}_\theta[(\theta - \hat{\theta})^2]$. Now, fix $\theta \in \Theta$. For any other $\theta' \in \Theta$ we will use (6.9) with $Q = P_\theta$ and $P = P'_\theta$. As a result we have that

$$\chi^2(P_{\theta'}\|P_\theta) \geq \chi^2(P_{\hat{\theta}}\|Q_{\hat{\theta}}) \geq \frac{(\mathbb{E}_\theta[\hat{\theta}] - \mathbb{E}_{\theta'}[\hat{\theta}])^2}{\operatorname{var}_\theta(\hat{\theta})}$$

Where the first inequality arises by using the data processing inequality and the second inequality by (6.9). Finally, by swapping the denominator with the left hand side and taking the supremum over all $\theta' \neq \theta$, and since $\operatorname{var}_\theta(\hat{\theta})$ is not a function of θ', we derive the final result.

Theorem 6.1 (Hammersley-Chapman-Robbins (HCR) lower bound)**.** *For the quadratic loss, any estimator $\hat{\theta}$ satisfies*

$$R_\theta(\hat{\theta}) \geq \operatorname{var}_\theta(\hat{\theta}) \geq \sup_{\theta'\neq\theta} \frac{(\mathbb{E}_\theta[\hat{\theta}] - \mathbb{E}_{\theta'}[\hat{\theta}])^2}{\chi^2(P_{\theta'}\|P_\theta)}, \quad \forall\theta \in \Theta. \tag{6.10}$$

6.3 Cramér-Rao (CR) lower bound

We now derive the Cramér-Rao lower bound as a consequence of the HCR lower bound. To this end, we restrict the problem to unbiased estimators, where an estimator $\hat{\theta}$ is said to be unbiased if $\mathbb{E}_\theta[\hat{\theta}] = \theta$ for all $\theta \in \Theta$. Then by applying the HCR lower bound we have that

$$\operatorname{var}_\theta(\hat{\theta}) \geq \sup_{\theta'\neq\theta} \frac{(\theta - \theta')^2}{\chi^2(P_{\theta'}\|P_\theta)} \geq \lim_{\theta'\to\theta} \frac{(\theta - \theta')^2}{\chi^2(P_{\theta'}\|P_\theta)}. \tag{6.11}$$

Here, we bypass the supremum by sending θ' to θ. However, when $\theta' \to \theta$ both the numerator and denominator will go to zero. Doing this, we hope that the denominator will go to zero quadratically

[2] This can always be assumed by allowing the likelihood ratio function $\frac{dP}{dQ}$ which is a sufficient statistic.

as the numerator does. Remember that

$$\chi^2(P_{\theta'}\|P_\theta) = \int \frac{(P_\theta - P_{\theta'})^2}{P_\theta}.$$

Then by using the Taylor expansion for P_θ around θ' we get that

$$P_\theta - P_{\theta'} = (\theta - \theta')\frac{dP_\theta}{d\theta} + o[(\theta - \theta')^2],$$

for θ near θ'. Combining the above while ignoring the little-o terms we get that

$$\chi^2(P_{\theta'}\|P_\theta) = (\theta - \theta')^2 \int \frac{(\frac{dP_\theta}{d\theta})^2}{P_\theta}.$$

Plugging back in (6.11) we get the CR lower bound.

Theorem 6.2 (Cramér-Rao (CR) lower bound)**.** *For any unbiased estimator $\hat{\theta}$ and any $\theta \in \Theta$*

$$\mathrm{var}_\theta(\hat{\theta}) \geq \frac{1}{I(\theta)},$$

where $I(\theta)$ is the Fisher information given by

$$I(\theta) = \int \frac{(\frac{dP_\theta}{d\theta})^2}{P_\theta}.$$

An intuitive interpretation of $I(\theta)$ is that it is a measure of the information the data contains for the estimation of the parameter when its true value is θ.

Example 6.5 (GLM)**.** Let $\theta \in \mathbb{R}$ and $X \sim P_\theta = \mathcal{N}(\theta, 1)$. Define the standard normal distribution by $\Phi(x)$. Note that $P_\theta(x) = \Phi(x - \theta)$. Next we calculate the Fisher information. By shifting x to θ, note that

$$I(\theta) = \int \frac{(\frac{dP_\theta(x)}{d\theta})^2}{P_\theta(x)} dx = \int \frac{(\frac{d}{d\theta}\Phi(x-\theta))^2}{\Phi(x-\theta)} dx = I(0).$$

Thus, $I(\theta) = I(0)$ for all $\theta \in \Theta$. In general, in any case where we have the model $X = \theta + Noise$, where the noise is standard normal (location model) we have that the fisher information is the same everywhere.

Remark
Another useful way of seeing the Fisher information is the following:

$$I(\theta) = \int \frac{(\frac{\partial P_\theta(x)}{\partial\theta})^2}{P_\theta(x)} \partial x = \mathbb{E}_\theta\left[\left(\frac{\frac{\partial P_\theta(X)}{\partial\theta}}{P_\theta(X)}\right)^2\right] = \mathbb{E}_\theta\left[\left(\frac{\partial \log P_\theta(X)}{\partial\theta}\right)^2\right] = \mathrm{var}_\theta\left[\frac{\partial \log P_\theta(X)}{\partial\theta}\right],$$

where the last equality holds after noticing that

$$\mathbb{E}_\theta\left[\frac{\partial \log P_\theta(X)}{\partial\theta}\right] = 0.$$

6.4 Biased estimators

Many times restricting ourselves to unbiased estimators proves to be very limiting. As a result, biased estimators need to be considered. Then it is useful to see how the HCR bound can be applied in this case. Define the bias of an estimator $\hat{\theta}$ by $b(\theta) = \mathbb{E}_\theta[\hat{\theta}] - \theta$. Assuming the risk function is quadratic it is easy to see that for a biased estimator by directly using HCR then

$$R_\theta(\hat{\theta}) = b^2(\theta) + \mathrm{var}_\theta(\hat{\theta}) \geq b^2(\theta) + \sup_{\theta' \neq \theta} \frac{(b(\theta') + \theta' - b(\theta) - \theta)^2}{\chi^2(P_{\theta'} \| P_\theta)}.$$

By using the same taylor expansion trick and assuming that $b(\theta)$ is differentiable we finally get that for an estimator $\hat{\theta}$ and any $\theta \in \Theta$

$$R_\theta(\hat{\theta}) \geq b^2(\theta) + \frac{(1 + b'(\theta))^2}{I(\theta)}.$$

Using this inequality we can find a lower bound on the worst case mini-max risk. In particular, we have that

$$R^* = \inf_{\hat{\theta}} \sup_\theta R_\theta(\hat{\theta}) \geq \inf_b \left[\sup_\theta \left(b^2(\theta) + \frac{(1 + b'(\theta))^2}{I(\theta)} \right) \right],$$

where in the last inequality we also used the fact that the choice of the estimator affects our quantity only through the bias.

6.5 Bayesian CR lower bound

Previously in this lecture we used the HCR bound to derive the CR lower bound. In order to derive the Bayesian version of the CR lower bound a similar approach can be used: first prove the Bayesian HCR and then derive the Bayesian CR lower bound as a result.

Theorem 6.3 (Bayesian CR lower bound)**.** *Assume that the loss function is quadratic, i.e., $\ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$. Also, for any estimator $\hat{\theta}$ (possibly randomized), and for any prior $\pi \in M(\Theta)$ define the Bayes risk of $\hat{\theta}$ by $R_\pi(\hat{\theta}) = \int R_\theta(\hat{\theta})\pi(d\theta) = \int \mathbb{E}_\theta[(\hat{\theta} - \theta)^2]\pi(d\theta)$. Then we have that*

$$R_\pi^* = \inf_{\hat{\theta}} R_\pi(\hat{\theta}) \geq \frac{1}{\mathbb{E}_{\theta \sim \pi}[I(\theta)] + I(\pi)},$$

where $I(\pi)$ the Fisher information of π, i.e.,

$$I(\pi) = \int \frac{(\pi'(\theta))^2}{\pi(\theta)} d\theta.$$

Recall the Chi-squared divergence and Hammersley-Chapman-Robbins (HCR) bound from last class. Suppose that P, Q are two probability distribution defined on some space $\mathcal{X}$ and X is an $\mathcal{X}$-valued random variable. The Chi-squared divergence has the following variational representation.

$$\chi^2(P\|Q) = \sup_{g:\mathcal{X}\to\mathbb{R}} 2\mathbb{E}_P[g(X)] - \mathbb{E}_Q[g^2(X)] - 1.$$

Furthermore, if $\mathcal{X} = \mathbb{R}$, choosing affine function g yields

$$\chi^2(P\|Q) \geq \frac{(\mathbb{E}_P[X] - \mathbb{E}_Q[X])^2}{\mathrm{var}_Q[X]} \tag{7.1}$$

which gives the HCR bound.

7.1 HCR Lower Bound

We are now continuing on the HCR lower bound from the last class. We here illustrate an example of HCR lower bound on estimation.

Example 7.1 (Estimation)**.** Let $\theta \in \mathbb{R}$ be an unknown parameter and let $X \in \mathbb{R}$ be a random variable (data) whose distribution depends on θ. Suppose $\hat{\theta}$ is an unbiased estimate of θ based on X. The relationships can be shown as

$$\theta \to X \to \hat{\theta}.$$

The estimation loss $\ell(\theta, \hat{\theta})$ is defined as $\ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$. Let $P = P_{\theta'}$, $Q = P_\theta$, and then the risk is lower bounded by

$$R_\theta(\hat{\theta}) \geq \mathrm{var}_\theta(\hat{\theta}) \geq \frac{(\mathbb{E}_\theta\hat{\theta} - \mathbb{E}_{\theta'}\hat{\theta})^2}{\chi^2(P_{\theta'}\|P_\theta)}.$$

Suppose $\hat{\theta}$ is an unbiased estimate of θ, then

$$R_\theta(\hat{\theta}) \geq \sup_{\theta\neq\theta} \frac{(\theta - \theta')^2}{\chi^2(P_{\theta'}\|P_\theta)} \geq \lim_{\theta'\to\theta} \frac{(\theta' - \theta)^2}{\chi^2(P_{\theta'}\|P_\theta)}.$$

7.2 Fisher information

The Fisher information is a way of measuring the amount of information that an observable random variable X carries about an unknown, deterministic parameter θ upon which the distribution of the observation X depends. Assume the probability density function of random variable X conditional on the value of θ is p_θ. The Fisher information is defined as

Definition 7.1 (Fisher information)**.** The Fisher information of the parameteric family of densitities $\{p_\theta : \theta \in \Theta\}$ (with respect to μ) at θ is

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial \log p_\theta}{\partial \theta}\right)^2\right] = \int \left(\frac{\partial p_\theta}{\partial \theta}\right)^2 \frac{1}{p_\theta} d\mu. \tag{7.2}$$

Theorem 7.1 (Fisher information)**.** *Assume that p_θ is twice differentiable with respect to θ and satisfies the regularity condition:*

$$\int \frac{\partial^2 p_\theta}{\partial \theta^2} d\mu = \frac{\partial^2}{\partial \theta^2} \int p_\theta d\mu = 0.$$

The Fisher information can be written as

$$I(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2 \log p_\theta}{\partial \theta^2}\right]$$

Proof. Since

$$\frac{\partial^2 \log p_\theta}{\partial \theta^2} = \frac{\frac{\partial^2 p_\theta}{\partial \theta^2}}{p_\theta} - \left(\frac{\frac{\partial p_\theta}{\partial \theta}}{p_\theta}\right)^2 = \frac{\frac{\partial^2 p_\theta}{\partial \theta^2}}{p_\theta} - \left(\frac{\partial \log p_\theta}{\partial \theta}\right)^2$$

and

$$\mathbb{E}\left[\frac{\partial^2 p_\theta}{\partial \theta^2}\frac{1}{p_\theta}\right] = 0$$

by assumption, we have

$$I(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial \theta}\log p_\theta\right)^2\right] = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial \theta^2}\log p_\theta\right]. \qquad \square$$

Theorem 7.2 (Fisher information: mutiple sample)**.** *Suppose random sample $X_1, \dots, X_n$ independently and identically drawn from a distribution p_θ. The Fisher information $I_n(\theta)$ provided by random samples $X_1, \dots, X_n$ is*

$$I_n(\theta) = nI(\theta),$$

where $I(\theta)$ is Fisher information provided by a single sample X_1.

Proof. We first denote the joint pdf of $X_1, \dots, X_n$ as

$$p_\theta(x_1, \dots, x_n) = \prod_{i=1}^n p_\theta(x_i).$$

Then the Fisher information $I_n(\theta)$ provided by $X_1, \dots, X_n$ is

$$I_n(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial p_\theta(X_1, \dots, X_n)}{\partial \theta}\right)^2\right] = \int \dots \int \left(\frac{\partial p_\theta(x_1, \dots, x_n)}{\partial \theta}\right)^2 p_\theta(x_1, \dots, x_n) dx_1 dx_2 \dots dx_n,$$

which is an n-dimensional integral. Thus, by Theorem 7.1, the Fisher information provided by $X_1, \dots, X_n$ can be calculated as

$$I_n(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2 \log p_\theta(X_1, \dots, X_n)}{\partial \theta^2}\right] = -\mathbb{E}_\theta\left[\sum_{i=1}^n \frac{\partial^2 \log p_\theta(X_i)}{\partial \theta^2}\right] = -\sum_{i=1}^n \mathbb{E}_\theta\left[\frac{\partial^2 \log p_\theta(X_i)}{\partial \theta^2}\right] = nI(\theta).$$

$\square$

7.3 Variations of HCR/CR lower bound

This section contains the following three versions of HCP/CR lower bound:

- Multiple Samples Version
- Multivariate Version
- Functional Version

7.3.1 Multiple-sample version

Suppose θ is some unknown, deterministic parameter and $X_1, \ldots, X_n$ are n random variables iid drawn from the distribution P_θ. The estimate $\hat{\theta}$ comes from $X_1, \ldots, X_n$. The relationships is shown as follows:

$$\theta \to X_1, \ldots, X_n \to \hat{\theta}.$$

Then the risk is lower bound by

$$R_\theta(\hat{\theta}) \geq \mathrm{var}_\theta(\hat{\theta}) \geq \frac{(\mathbb{E}_\theta \hat{\theta} - \mathbb{E}_{\theta'} \hat{\theta})^2}{\chi^2(P_{\theta'}^{\otimes n} \| P_\theta^{\otimes n})}.$$

For the HCR lower bound,

$$R_\theta(\hat{\theta}) \geq \sup_{\theta \neq \theta'} \frac{(\theta - \theta')^2}{(1 + \chi^2(P_\theta \| P_{\theta'}))^n - 1} \overset{\theta' \to \theta}{\geq} \frac{1}{nI(\theta)}.$$

7.3.2 Multivariate Version

We next show the multi-dimensional version of

$$\chi^2(P\|Q) \geq \frac{(\mathbb{E}_P X - \mathbb{E}_Q X)^2}{\mathrm{var}_Q X}.$$

Suppose P, Q are two distributions defined on $\mathbb{R}^p$, then

$$\chi^2(P\|Q) = \sup_{g:\mathbb{R}^p \to \mathbb{R}} [2\mathbb{E}_P g(X) - \mathbb{E}_Q g^2(X) - 1].$$

Furthter, if $g(X) = \langle a, X \rangle + 1$, then

$$\chi^2(P\|Q) \geq 2\mathbb{E}_P \langle a, X \rangle + 1 - \mathbb{E}_Q(\langle a, X \rangle + 1)^2.$$

If we further assume $\mathbb{E}_Q X = 0$, then we have

$$\chi^2(P\|Q) \geq 2 \langle a, \mathbb{E}_P X \rangle - a^T \mathbb{E}_Q[XX^T]a.$$

Therefore, we finally have

$$\chi^2(P\|Q) \geq (\mathbb{E}_P X - \mathbb{E}_Q X)^T \mathrm{cov}_Q^{-1}(X)(\mathbb{E}_P X - \mathbb{E}_Q X)$$

Let the loss function $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_2^2$ and $\hat{\theta}$ be the unbiased estimate of θ, *i.e.,* $\mathbb{E}_\theta \hat{\theta} = \theta$. Then

$$(\theta' - \theta)^T \mathrm{cov}_\theta^{-1}(\hat{\theta})(\theta' - \theta) \leq \chi^2(P_{\theta'} \| P_\theta) \overset{\theta' \to \theta}{=} (\theta' - \theta)^T \mathbf{I}(\theta)(\theta' - \theta) + \|\theta' - \theta\|_2^2,$$

where the equality follows from the Taylor expansion and Fisher information matrix is given as

$$\mathbf{I}(\theta) = \int \frac{\nabla P_\theta (\nabla P_\theta)^T}{P_\theta}.$$

If we take $\theta' = \theta + \epsilon u$ for an arbitrary unit vector u and $\epsilon \to 0$, we have

$$u^T \mathrm{cov}_\theta^{-1}(\hat{\theta}) u \leq u^T \mathbf{I}(\theta) u,$$

which is equivalent to

$$\mathrm{cov}_\theta(\hat{\theta}) \succeq \mathbf{I}^{-1}(\theta),$$

and further indicates

$$R_\theta(\hat{\theta}) = \mathrm{tr}(\mathrm{cov}_\theta(\hat{\theta})) \geq \mathrm{tr}(\mathbf{I}^{-1}(\theta)). \tag{7.3}$$

Then we have

$$\mathbb{E}\|\theta - \hat{\theta}\|_2^2 = \sum_{i=1}^{p} \mathbb{E}(\hat{\theta}_i - \theta_i)^2 \geq \sum_{i=1}^{p} \frac{1}{I_i}, \tag{7.4}$$

where $I_i \triangleq \mathbf{I}_{ii}(\theta)$, since

$$\sum_{i=1}^{p} \frac{1}{I_i(\theta)} \leq \mathrm{tr}(I^{-1}(\theta)).$$

Note that if we apply the one-dimensional CRLB for each coordinate we would get (7.4) which is weaker than (7.3).

Finally, similar to Theorem 7.1, assuming the corresponding regularity of the Hessian, the Fisher information matrix can be written as

$$\mathbf{I}(\theta) = \mathbb{E}_\theta[(\nabla \log P_\theta)(\nabla \log P_\theta)^T] = \mathrm{cov}_\theta(\nabla \log P_\theta) = -\left(\mathbb{E}_\theta\left[\frac{\partial^2 \log P_\theta}{\partial\theta_i \partial\theta_j}\right]\right).$$

7.3.3 Functional Version

Assume that θ is an unknown parameter, that random variable X comes from the distribution P_θ and that $\hat{T}(X)$ is an estimation for $T(\theta)$, where $T : \Theta \to \mathbb{R}$. The relationship is shown as follows:

$$\theta \to X \to \hat{T}.$$

If we further assume $\hat{T}(\theta)$ is an unbiased estimation for $T(\theta)$, then

$$\mathrm{var}_\theta(\hat{T}) \geq \frac{\|\nabla T\|_2^2}{I(\theta)}$$

7.4 Bayesian Cramér-Rao Lower Bound via data processing inequality

The class will introduce two methods of proving Bayesian Cramér-Rao lower bound.

- Method 1: $\chi^2 \to$ Bayesian HCR $\to$ Bayesian CR (next).
- Method 2: Classical Method (see Theorem 8.1).

The notation used in this section is shown as follows:

- $\Theta = \mathbb{R}$
- $\ell(\theta, \hat\theta) = (\theta - \hat\theta)^2$.
- π is a "nice" prior on $\mathbb{R}$

The relationship can be described as follows:

$$\pi \to \theta \to X \to \hat\theta.$$

Theorem 7.3 (Bayesian Cramér-Rao Lower Bound)**.** *Assuming suitable regularity conditions, then*

$$R^* \ge R^*_\pi = \inf_{\hat\theta} \mathbb{E}_\pi(\theta, \hat\theta)^2 \ge \frac{1}{\mathbb{E}_{\theta\sim\pi} I(\theta) + I(\pi)},$$

where R^*_π is the Bayes risk and $I(\pi) = \int \frac{\pi'^2}{\pi}$ is the Fisher information of the prior.

Proof. Consider the following comparison of experiments:

$$Q : \pi \longrightarrow \theta \xrightarrow{P_\theta = Q_{X|\theta}} X \longrightarrow \hat\theta,$$

$$P : \tilde\pi \longrightarrow \theta \xrightarrow{\tilde P_\theta = P_{X|\theta}} X \longrightarrow \hat\theta.$$

Then

$$\begin{aligned}
\chi^2(P_{\theta X} \| Q_{\theta X}) &\ge \chi^2(P_{\theta\hat\theta} \| Q_{\theta\hat\theta}) && \text{data processing inequality} \\
&\ge \chi^2(P_{\theta-\hat\theta} \| Q_{\theta-\hat\theta}) && \text{data processing inequality} \\
&\ge \frac{(\mathbb{E}_P(\theta-\hat\theta) - \mathbb{E}_Q(\theta-\hat\theta))^2}{\operatorname{var}_\pi(\hat\theta - \theta)}. && \text{by (7.1)}
\end{aligned}$$

Let T_δ denote the pushforward of shifting by δ, that is, $T_\delta(P_A) = P_{A+\delta}$. Let us choose

$$Q_\theta = \pi, Q_{X|\theta} = P_\theta, P_\theta = T_\delta \pi, P_{X|\theta} = P_{\theta-\delta},$$

then $P_X = Q_X$ which further indicates $P_{\hat\theta} = Q_{\hat\theta}$ and the mean of $\hat\theta$ under distribution of P equals to the mean under the distribution under Q. Hence $\mathbb{E}_P(\theta - \hat\theta) - \mathbb{E}_Q(\theta - \hat\theta) = \delta$! For the Bayesian HCR lower bound,

$$R^*_\pi \ge \sup_{\delta \ne 0} \frac{\delta^2}{\chi^2(P_{X\theta} \| Q_{X\theta})} \ge \lim_{\delta\to 0} \frac{\delta^2}{\chi^2(P_{X\theta} \| Q_{X\theta})} = \frac{1}{I(\pi) + \mathbb{E}_{\theta\sim\pi}[I(\theta)]}. \tag{7.5}$$

The last step is justified as follows:

$$\begin{aligned}
\chi^2(P_{X\theta} \| Q_{X\theta}) &= \int \frac{(P_{X\theta} - Q_{X\theta})^2}{Q_{X\theta}} = \int \frac{[P_\theta(P_{X|\theta} - Q_{X|\theta}) + (P_\theta - Q_\theta)Q_{X|\theta}]^2}{Q_{X\theta}} \\
&= \int \frac{P_\theta^2}{Q_\theta} \int \frac{(P_{X|\theta} - Q_{X|\theta})^2}{Q_{X|\theta}} + \int \frac{(P_\theta - Q_\theta)^2}{Q_\theta^2} + 2 \int \frac{P_\theta(P_\theta - Q_\theta)}{Q_\theta} \int (P_{X|\theta} - Q_{X|\theta}) \\
&= \chi^2(P_\theta \| Q_\theta) + \mathbb{E}\left[\chi^2(P_{X|\theta} \| Q_{X|\theta}) \cdot \left(\frac{P_\theta}{Q_\theta}\right)^2\right]
\end{aligned}$$

Then applying

- $\chi^2(P_\theta \| Q_\theta) = \chi^2(T_{\delta\pi} \| \pi) = \delta^2[I(\pi) + o(1)]$ by Taylor expansion,
- $\chi^2(P_{X|\theta} \| Q_{X|\theta}) = [I(\theta) + o(1)]\delta^2$ by Taylor expansion,

we obtain (7.5). □

7.5 Information Bound

In this section, we introduce the local version of the minimax lower bound. The local minimax risks is defined in a quadratic form: $\inf_{\hat{\theta}} \sup_{|\theta-\theta_0|\le\epsilon} \mathbb{E}(\hat{\theta}-\theta)^2$. Further, we have

$$\begin{aligned}\inf_{\hat{\theta}} \sup_{|\theta-\theta_0|\le\epsilon} \mathbb{E}(\hat{\theta}-\theta)^2 &\ge \frac{1}{I(\theta)+n\mathbb{E}_{\theta\sim\pi}[I(\theta)]} \\ &= \frac{1+o(1)}{n\mathbb{E}_{\theta\sim\pi}[I(\theta)]}\end{aligned}$$

If $\theta \mapsto I(\theta)$ is continuous, then

$$\mathbb{E}_{\theta\sim\pi}[I(\theta)] = I(\theta_0) + o(1) = \frac{1+o(1)}{nI(\theta)}.$$

Assume the random variable Z coming from the distribution π, $Z \sim \pi$. Let $I(Z) \triangleq I(\pi)$. For constant $\alpha, \beta \neq 0$, then $I(Z+\alpha) = I(Z)$ and $I(\beta Z) = \frac{I(Z)}{\beta^2}$. If the π has the distribution of form $\cos^2\frac{\pi x}{2}$, then $\min_{\pi:[-1,1]} I(\pi) = \pi^2$. If the distribution π has the form of $\cos^2\frac{\pi(x-\theta_0)}{2\epsilon}$, then $I(\theta) = \frac{\pi^2}{\epsilon}$. Then we have

$$\inf_{\hat{\theta}} \sup_{|\theta_0-\theta|\le\epsilon} \mathbb{E}(\hat{\theta}-\theta)^2 \ge R^*_\pi \ge \frac{1}{n\mathbb{E}_{\theta\sim\pi}[I(\theta)]+I(\pi)}.$$

Now if we pick $\epsilon = n^{-1/4}$, we have

$$R^* \ge \inf_{\hat{\theta}} \sup_{|\theta-\theta_0|\le n^{-1/4}} \mathbb{E}_\theta(\theta-\hat{\theta})^2 \ge \frac{1}{nI(\theta)+o(\sqrt{n})} \overset{Optimize}{\Longrightarrow} R^* \ge \frac{1+o(1)}{n\inf_{\theta_0\in\Theta} I(\theta_0)}.$$

§ 8. Bayesian Cramér-Rao (continued), MLE

8.1 Example: Gaussian Location Model (GLM)

Let $X_i = \theta + Z_i$, where $Z_i \sim \mathcal{N}(0,1)$, and $\theta \sim \pi = \mathcal{N}(0,s)$. Given i.i.d. observations $X = (X_1, X_2, \cdots, X_n)$, we have

$$\begin{aligned}
\chi^2(P_{\theta X}||Q_{\theta X}) &= \chi^2(P_{\theta\bar{X}}||Q_{\theta\bar{X}}) \\
&= \chi^2(P_\theta||Q_\theta) + \mathbb{E}_Q\left[\left(\frac{P_\theta}{Q_\theta}\right)^2 \chi^2(P_{\bar{X}|\theta}||Q_{\bar{X}|\theta})\right] \\
&= (e^{\delta^2/s} - 1) + e^{\delta^2/s}(e^{n\delta^2} - 1) \\
&= e^{\delta^2(n+\frac{1}{s})} - 1.
\end{aligned}$$

The first line follows from the fact that $\bar{X}$ is a sufficient statistic ($\theta \to \bar{X} \to X$), and the information processing inequality. The second line follows from Lecture 7 (last equation, Page 5). The third line follows from

$$\chi^2\left(\mathcal{N}(\theta,\sigma^2)||\mathcal{N}(\theta+\delta,\sigma^2)\right) = e^{\delta^2/\sigma^2} - 1.$$

Therefore, by Bayesian HCR and Bayesian Cramér-Rao Lower Bound:

$$R_\pi^* \geq \sup_{\delta\neq 0} \frac{\delta^2}{e^{\delta^2(n+\frac{1}{s})} - 1} = \lim_{\delta\to 0} \frac{\delta^2}{e^{\delta^2(n+\frac{1}{s})} - 1} = \frac{1}{n+\frac{1}{s}} = \frac{s}{sn+1}.$$

In this case, the lower bound is exact! (It has been verified that $R_\pi^* = \frac{s}{sn+1}$.) The minimax lower bound is $R^* \geq \sup_s R_\pi^* = \frac{1}{n}$.

8.2 Classical Proof of Bayesian Cramér-Rao Lower Bound

Theorem 8.1 (Same statement as Theorem 7.3)**.** *If $X \sim P_\theta$, $\theta \sim \pi$, we have*

$$\mathbb{E}[(\hat{\theta}(X) - \theta)^2] \geq \frac{1}{I(\pi) + \mathbb{E}_{\theta\sim\pi}[I(\theta)]}.$$

Alternative Proof. Note that

$$\int \hat{\theta}(x)\frac{\partial}{\partial\theta}(P_\theta(x)\pi(\theta))\ \mathrm{d}\theta = 0, \tag{8.1}$$

$$\int \theta\frac{\partial}{\partial\theta}(P_\theta(x)\pi(\theta))\ \mathrm{d}\theta = -\int P_\theta(x)\pi(\theta)\ \mathrm{d}\theta, \tag{8.2}$$

where the first equation follows from the regularity condition, and the second equation follows from integration by part.

Therefore,

$$\begin{aligned}\mathbb{E}\left[(\hat{\theta}(X)-\theta)\frac{\partial \log(P_\theta(X)\pi(\theta))}{\partial\theta}\right] &= \int \mu(\mathrm{d}x)\int(\hat{\theta}(x)-\theta)\frac{\partial(P_\theta(x)\pi(\theta))}{\partial\theta}\frac{P_\theta(x)\pi(\theta)}{P_\theta(x)\pi(\theta)}\mathrm{d}\theta\\ &= \int \mu(\mathrm{d}x)\int P_\theta(x)\pi(\theta)\mathrm{d}\theta\\ &= 1,\end{aligned}$$

where the second line follows from (8.1) and (8.2).

By Cauchy-Schwarz inequality,

$$1 = \mathbb{E}\left[(\hat{\theta}(X)-\theta)\frac{\partial \log(P_\theta(X)\pi(\theta))}{\partial\theta}\right] \le \mathbb{E}\left[(\hat{\theta}(X)-\theta)^2\right]\mathbb{E}\left[\left(\frac{\partial \log(P_\theta(X)\pi(\theta))}{\partial\theta}\right)^2\right].$$

Hence

$$\mathbb{E}\left[(\hat{\theta}(X)-\theta)^2\right] \ge \frac{1}{\mathbb{E}\left[\left(\frac{\partial \log P_\theta(X)}{\partial\theta}+\frac{\partial\log\pi(\theta)}{\partial\theta}\right)^2\right]} = \frac{1}{\mathbb{E}[I(\theta)]+I(\pi)}.$$

□

8.3 An Alternative Information Inequality

If we choose a uniform prior in Theorem 8.1, the resulting lower bound is zero since the Fisher information of uniform distribution is infinity. Nevertheless, it is possible to obtain an alternative information inequality involving $\mathbb{E}_{\theta\sim\text{uniform}}[I(\theta)]$; however, it should be pointed out that the lower bound applies to the minimax risk (not Bayes risk with respect to uniform prior) since the proof in act involves two prior: uniform on the interval and uniform over the two endpoints.

Theorem 8.2. *Assume the usual regularity condition:*

$$\int \frac{\partial p_\theta}{\partial x}dx = 0.$$

Then

$$R^* = \inf_{\hat{\theta}}\sup_{\theta\in[\theta_0-\epsilon,\theta_0+\epsilon]}\mathbb{E}_\theta[(\theta-\hat{\theta})^2] \ge \frac{1}{(\epsilon^{-1}+\sqrt{n\bar{I}})^2}$$

where $\overline{I}$ denotes the average Fisher information:

$$\overline{I} = \frac{1}{2\epsilon}\int_{\theta_0-\epsilon}^{\theta_0+\epsilon} I(\theta)\ \mathrm{d}\theta.$$

Proof. See Problem 2 in Homework 1. □

Remark 8.1. Theorem 8.2 is a strict improvement of the inequality of Chernoff-Rubin-Stein:[1]

$$\inf_{\hat{\theta}}\sup_{\theta\in[\theta_0-\epsilon,\theta_0+\epsilon]}\mathbb{E}_\theta[(\theta-\hat{\theta})^2] \ge \max_{0<\delta<1}\min\left\{\frac{\delta^2}{4},\frac{1-\epsilon}{n\bar{I}}\right\} = \frac{1}{(\epsilon^{-1}+\sqrt{n\bar{I}+1})^2}.$$

Both this and Theorem 8.2 suffice to prove the optimal minimax lower bound.

[1] This is given in [Che56, Lemma 1] without proof, which Chernoff credited to Rubin and Stein.

8.4 Maximum Likelihood Estimator (MLE) and asymptotic efficiency

We *sketch* the analysis of MLE in the classical large-sample asymptotics. Let $X = (X_1, X_2, \cdots, X_n) \overset{\text{i.i.d.}}{\sim} P_{\theta_0}$, define maximum likelihood estimator:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta\in\Theta} L_\theta(X),$$

where

$$L_\theta(X) = \log P_\theta^{\otimes n}(X) = \sum_{i=1}^n \log P_\theta(X_i).$$

Intuition:

$$\mathbb{E}_{\theta_0}\left[L_\theta(X) - L_{\theta_0}(X)\right] = \mathbb{E}_{\theta_0}\left[\sum_{i=1}^n \log\frac{P_\theta(X_i)}{P_{\theta_0}(X_i)}\right] = -nD(P_{\theta_0}||P_\theta) \le 0.$$

So as long as $\theta_0 \neq \theta$, $L_\theta(X) - L_{\theta_0}(X)$ is a random walk with negative drift. From here the consistency of MLE follows upon assuming appropriate regularity conditions.

Assuming more conditions one can obtain asymptotic normality and $\sqrt{n}$-consistency of MLE. Next, we derive a local quadratic approximation of the log-likelihood function. By Taylor expansion,

$$L_\theta(X) = L_{\theta_0}(X) + \sum_{i=1}^n \left.\frac{\partial \log P_\theta(X_i)}{\partial\theta}\right|_{\theta=\theta_0}(\theta-\theta_0) + \frac{1}{2}\sum_{i=1}^n \left.\frac{\partial^2 \log P_\theta(X_i)}{\partial\theta^2}\right|_{\theta=\theta_0}(\theta-\theta_0)^2 + o((\theta-\theta_0)^2). \tag{8.3}$$

Recall that

$$\mathbb{E}\left[\frac{\partial \log P_\theta(X_i)}{\partial\theta}\right] = 0, \quad \mathbb{E}\left[\left(\frac{\partial \log P_\theta(X_i)}{\partial\theta}\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2 \log P_\theta(X_i)}{\partial\theta^2}\right] = I(\theta).$$

By the Central Limit Theorem,

$$\frac{1}{\sqrt{nI(\theta_0)}}\sum_{i=1}^n \frac{\partial \log P_\theta(X_i)}{\partial\theta} \xrightarrow{\text{d.}} \mathcal{N}(0,1).$$

By the Weak Law of Large Numbers,

$$\sum_{i=1}^n \frac{\partial^2 \log P_\theta(X_i)}{\partial\theta^2} = -nI(\theta_0) + o_P(n).$$

Substituting these quantities into (8.3), we obtain a local quadratic approximation of the log-likelihood function:

$$L_\theta(X) \approx L_{\theta_0}(X) + \sqrt{nI(\theta_0)}\cdot Z\cdot(\theta-\theta_0) - \frac{1}{2}nI(\theta_0)(\theta-\theta_0)^2,$$

where $Z \sim \mathcal{N}(0,1)$. Maximizing the right-hand side, we obtain:

$$\hat{\theta}_{\text{MLE}} \approx \theta_0 + \frac{Z}{\sqrt{nI(\theta_0)}}.$$

Therefore, MLE achieves the locally minimax lower bound $R^* \ge \frac{1+o(1)}{nI(\theta_0)}$ (see Section 7.5 in Lecture 7).

Remark 8.2. The general asymptotic theory of MLE and achieving information bound is due to Hájek and LeCam.

8.5 Bayesian Lower Bounds for Functional Estimation

Next, we derive the Bayesian Cramér-Rao lower bound for functional estimation $\widehat{T}(X)$.

Theorem 8.3. *Let $T : \mathbb{R}^p \to \mathbb{R}$, and*

$$\begin{tikzcd} \theta \arrow[r] \arrow[d] & X \arrow[d] \\ T(\theta) & \widehat{T}(X) \end{tikzcd}$$

Then we have

$$R_\pi^* \geq (\nabla T)' I^{-1} \nabla T.$$

Proof. By similar arguments in previous lectures,

$$\chi^2(P_{\theta X} || Q_{\theta X}) \geq \chi^2(P_{T-\widehat{T}} || Q_{T-\widehat{T}}) \geq \frac{\left(\mathbb{E}_P[T - \widehat{T}] - \mathbb{E}_Q[T - \widehat{T}]\right)^2}{\mathrm{Var}_Q[T - \widehat{T}]}. \tag{8.4}$$

Let $Q(\theta) = \pi(\theta)$, and $P(\theta) = \pi(\theta - \epsilon u)$, where $u \in \mathbb{R}^p$. In order to make the marginal distribution of $P_X = Q_X$, let $P_\theta(x) = Q_{\theta - \epsilon u}(x)$. Hence the numerator and the denominator in (8.4) satisfy:

$$\begin{aligned} \left(\mathbb{E}_P[T - \widehat{T}] - \mathbb{E}_Q[T - \widehat{T}]\right)^2 &= (\mathbb{E}_P[T] - \mathbb{E}_Q[T])^2 \\ &= \left(\int \pi(\theta) T(\theta + \epsilon u) \, \mathrm{d}\theta - \int \pi(\theta) T(\theta) \, \mathrm{d}\theta\right)^2 \\ &= \left(\int \pi(\theta) \langle \nabla T, \epsilon u \rangle + o(\epsilon)\right)^2 \\ &= \epsilon^2 \langle \mathbb{E}_\pi \nabla T, u \rangle^2 + o(\epsilon^2), \end{aligned} \tag{8.5}$$

$$\mathrm{Var}_Q[T - \widehat{T}] \leq \mathbb{E}_Q[(T - \widehat{T})^2] = R_\pi. \tag{8.6}$$

The left-hand side of (8.4) satisfies

$$\begin{aligned} \chi^2(P_{\theta X} || Q_{\theta X}) &= \chi^2(P_\theta || Q_\theta) + \mathbb{E}_Q\left[\chi^2(P_{X|\theta} || Q_{X|\theta}) \left(\frac{P_\theta}{Q_\theta}\right)^2\right] \\ &= \int \frac{(\pi(\theta - \epsilon u) - \pi(\theta))^2}{\pi(\theta)} \mathrm{d}\theta + \mathbb{E}_\pi\left[\int \frac{(Q_{\theta - \epsilon u}(x) - Q_\theta(x))^2}{Q_\theta(x)} \mathrm{d}x \left(\frac{\pi(\theta - \epsilon u)}{\pi(\theta)}\right)^2\right] \\ &= \int \frac{\epsilon^2 u' (\nabla \pi)(\nabla \pi)' u}{\pi(\theta)} \mathrm{d}\theta + \mathbb{E}_\pi\left[\int \frac{\epsilon^2 u' (\nabla_\theta Q)(\nabla_\theta Q)' u}{Q_\theta(x)} \mathrm{d}x\right] + o(\epsilon^2) \\ &= \epsilon^2 u' \left(I(\pi) + \mathbb{E}_\pi[I(\theta)]\right) u + o(\epsilon^2). \end{aligned} \tag{8.7}$$

Substituting (8.5), (8.6), and (8.7) into (8.4), we have

$$R_\pi^* \geq \frac{\langle \mathbb{E}_\pi \nabla T, u \rangle^2}{u' \left(I(\pi) + \mathbb{E}_\pi[I(\theta)]\right) u}$$

Locally, $\mathbb{E}_\pi \nabla T(\theta) \approx \nabla T(\theta_0)$, and $I(\pi) + \mathbb{E}_\pi[I(\theta)] \approx I(\theta_0)$. Hence

$$\boxed{R_\pi^* \geq \sup_u \frac{\langle \nabla T(\theta_0), u \rangle^2}{u' I(\theta_0) u} = (\nabla T(\theta_0))' I^{-1}(\theta_0) \nabla T(\theta_0).}$$

The maximum is attained when $u = I^{-1}(\theta_0)\nabla T(\theta_0)$.[2] □

Remark 8.3. The maximum likelihood estimator satisfies $T(\hat{\theta}_{\text{MLE}}) = T(\theta_0 + \frac{1}{\sqrt{n}}Z)$, where $Z \sim \mathcal{N}(0, I^{-1}(\theta_0))$. Hence

$$T(\hat{\theta}_{\text{MLE}}) \sim N\left(T(\theta_0), \frac{1}{n}(\nabla T(\theta_0))' I^{-1}(\theta_0)(\nabla T(\theta_0))\right).$$

The maximum likelihood estimator again asymptotically achieves the locally minimax lower bound.

8.6 Example: Classical asymptotics of entropy estimation

Corollary 8.1. *Let $X_1, \cdots, X_n \overset{i.i.d.}{\sim} p \in \mathcal{M}_k$, where $\mathcal{M}_k$ denotes the set of probability distributions over $[k] = \{1, \ldots, k\}$. Then the minimax quadratic risk of entropy estimation satisfies*

$$R^* = \inf_{\widehat{H}} \sup_{P \in \mathcal{M}_k} \mathbb{E}[(\widehat{H} - H)^2] = \frac{1}{n}\left(\max_{p \in \mathcal{M}_k} V(p) + o(1)\right), \quad n \to \infty$$

where

$$\begin{aligned} H(p) &= \sum_{i=1}^{k} p_i \log \frac{1}{p_i} = \mathbb{E}\left[\log \frac{1}{p(X)}\right], \\ V(p) &= \operatorname{Var}\left(\log \frac{1}{p(X)}\right) \end{aligned}$$

Note: $\max_{p \in \mathcal{M}_k} V(p) \leq \log^2 k$ for all $k \geq 3$ (see [PPV10, Eq. (464)]).

Proof. We have $H : \Theta \to \mathbb{R}^+$, where $\theta = (p_1, p_2, \cdots, p_{k-1})$.[3] Therefore,

$$\frac{\partial H}{\partial p_i} = \log \frac{p_k}{p_i}, \quad i = 1, 2, \cdots, k-1.$$

Next, we compute the Fisher Information matrix:

$$I(\theta)_{ij} = -\mathbb{E}\left[\frac{\partial^2 \log p(X)}{\partial p_i \partial p_j}\right] = \begin{cases} \frac{1}{p_i} + \frac{1}{p_k} & \text{if } i = j \\ \frac{1}{p_k} & \text{if } i \neq j \end{cases}.$$

Therefore,

$$I(\theta) = \begin{bmatrix} \frac{1}{p_1} & & \\ & \ddots & \\ & & \frac{1}{p_{k-1}} \end{bmatrix} + \frac{1}{p_k}\mathbf{1}\mathbf{1}'.$$

By Matrix Inversion Lemma,[4] we have

$$I^{-1}(\theta) = \begin{bmatrix} p_1 & & \\ & \ddots & \\ & & p_{k-1} \end{bmatrix} + \begin{bmatrix} p_1 \\ \vdots \\ p_{k-1} \end{bmatrix} \begin{bmatrix} p_1 & \cdots & p_{k-1} \end{bmatrix}.$$

[2] This can be shown, for example, by letting $\tilde{u} = I^{-\frac{1}{2}}(\theta_0)u$.

[3] $p_k = 1 - p_1 - \cdots - p_{k-1}$.

[4] $(A + UCV)^{-1} = A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$.

Therefore,

$$\begin{aligned}
\nabla H' I^{-1}(\theta) \nabla H &= \sum_{i=1}^{k-1} p_i \log^2 \frac{p_k}{p_i} - \left(\sum_{i=1}^{k-1} p_i \log \frac{p_k}{p_i} \right)^2 \\
&= \sum_{i=1}^{k} p_i \log^2 \frac{1}{p_i} + \log^2 \frac{1}{p_k} - 2 \sum_{i=1}^{k} p_i \log \frac{1}{p_i} \log \frac{1}{p_k} - \left(\left(\sum_{i=1}^{k} p_i \log \frac{1}{p_i} \right) - \log \frac{1}{p_k} \right)^2 \\
&= \sum_{i=1}^{k} p_i \log^2 \frac{1}{p_i} - \left(\sum_{i=1}^{k} p_i \log \frac{1}{p_i} \right)^2 \\
&= \mathbb{E}\left[\log^2 \frac{1}{p(X)} \right] - \left(\mathbb{E}\left[\log \frac{1}{p(X)} \right] \right)^2 = \mathrm{Var}\left[\log \frac{1}{p(X)} \right] = V(p).
\end{aligned}$$

Given n samples, the Fisher Information matrix is $nI(\theta)$. By Theorem 8.3,

$$R^* \geq \frac{1 + o(1)}{n} \nabla H' I^{-1}(\theta) \nabla H = \frac{1 + o(1)}{n} V(p). \qquad \square$$

Part III

Unstructured estimation problems in high dimensions

§ 9. Exact minimax risk for Gaussian location model, LeCam's method

In this lecture we consider estimation problems with no prior assumption on the structure of the parameter space. Examples of structures include sparsity, smoothness and low-rankness.

Let $X = (X_1, \dots, X_n) \overset{i.i.d}{\sim} P_\theta$ be n samples drawn from distribution P_θ parametrized by $\theta \in \Theta$, where Θ is $\mathbb{R}^p$. Given a loss function $\ell : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^+$, the minimax risk is

$$R_n^*(\Theta) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \ell(\theta, \hat{\theta}).$$

Two obvious observations:

- More structures lead to smaller risk. Formally, if $\Theta' \subset \Theta$, then $R_n^*(\Theta') \leq R_n^*(\Theta)$.[1] Without assuming any prior structure, $\Theta = \mathbb{R}^p$, and we denote $R_n^*(\mathbb{R}^p) = R_{n,p}^*$.

- More samples lead to smaller risk. Formally, $n \mapsto R_n^*(\Theta)$ is decreasing and typically vanishing as $n \to \infty$. In the classical large-sample asymptotic regime as studied in Lecture 8, the speed is usually "parametric“, e.g., $\frac{1}{n}$ under the quadratic risk. In comparison, the focus in this course is understanding the dependency on dimension and other structural parameters without assuming large sample size. This is captured by the minimax rate. For example, we say $R_{n,p}^* \asymp \Psi_{n,p}$, when $c \leq \frac{R_{n,p}^*}{\Psi_{n,p}} \leq c'$, $\forall\, n, p$ for some universal constants c and c'.

9.1 Log-concavity, Anderson's lemma and exact minimax risk in GLM

Definition 9.1 (Gaussian location model (GLM))**.** Let $X_1, \dots, X_n$ be iid drawn from $\mathcal{N}(\theta, I_p)$ with $\theta \in \mathbb{R}^p$. The goal is to estimate the mean θ. Let $\hat{\theta}$ denote the estimator and $R_{n,p}^*$ denote the minimax risk under loss function $\ell(\theta, \hat{\theta})$.

Theorem 9.1. *Under GLM with quadratic loss function $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_2^2 = \sum_{i=1}^p (\theta_i - \hat{\theta}_i)^2$, then*

$$R_{n,p}^* = \frac{p}{n}, \quad \forall n, p \in \mathbb{N}.$$

Proof. We upper bound and lower bound $R_{n,p}^*$ by $\frac{p}{n}$ in order to show equality. Let us have an estimator $\bar{X} = \frac{\sum X_i}{n} \sim \mathcal{N}(\theta, \frac{1}{n} I_p)$. Hence the risk $R_{n,p}^*$ is upper bounded by the risk obtained when using estimator $\hat{\theta} = \bar{X}$. We can compute the risk for using $\hat{\theta} = \bar{X}$ as $\frac{p}{n}$. So,

$$R_{n,p}^* \leq \frac{p}{n} \tag{9.1}$$

[1] Note that this does not mean that achieving $R_n^*(\Theta')$ is computationally easier than $R_n^*(\Theta)$!

We lower bound the minimax risk $R^*_{n,p}$ by Bayes risk with prior $\pi \sim \mathcal{N}(0, sI_p)$. We can compute $R^*_\pi = \frac{sp}{sn+1}$. So,

$$\begin{aligned} R^*_{n,p} &\geq R^*_\pi \\ &\overset{\lim s\to\infty}{=} \frac{p}{n} \end{aligned} \tag{9.2}$$

Combining the upper bound and lower bound in (9.1) and (9.2), we complete the proof. □

The limitation of the above proof technique is that it only works for quadratic loss function. We next discuss a more general theorem which works over a larger range of loss functions.

Definition 9.2 (Bowl-shaped)**.** A function $\rho : \mathbb{R}^d \to \mathbb{R}_+$ is called bowl-shaped when all its sublevel sets $K_c = \{x : \rho(x) < c\}$ for all $c \in \mathbb{R}$ are convex and symmetric (i.e. $K_c = -K_c$).

Theorem 9.2. *Consider GLM with loss functions $\ell(\theta, \hat{\theta}) = \rho(\theta - \hat{\theta})$, where $\rho : \mathbb{R}^p \to \mathbb{R}_+$ is bowl-shaped and lower-semicontinous. Then*

$$R^*_{n,p} = \mathbb{E}\rho\left(\frac{Z}{\sqrt{n}}\right),$$

where $Z \sim \mathcal{N}(0, I_p)$.

Corollary 9.1. *Let $\rho = \|.\|^q, q \geq 1$, then under GLM,*

$$R^*_{n,p} = \frac{1}{n^{q/2}}\mathbb{E}\|Z\|^q.$$

Example 9.1. Applications of Corollary 9.1:

- If $\rho = \|.\|_2^2$, then $R^*_{n,p} = \frac{1}{n}\mathbb{E}\|Z\|^2 = \frac{p}{n}$.
- If $\rho = \|.\|_\infty$, then $\mathbb{E}\|Z\|_\infty \asymp \sqrt{\log p}$ and $R^*_{n,p} \asymp \sqrt{\frac{\log p}{n}}$.
- If $\theta \in \mathbb{R}^{p\times p}$ is a matrix, and $\rho = \|.\|_{op}^2 = \sigma_{\max}(\cdot)$, then $\mathbb{E}\|Z\|_{op} \asymp \sqrt{p}$ and $R^*_{n,p} \asymp \frac{p}{n}$
- If $\theta \in \mathbb{R}^{p\times p}$ is a matrix, and $\rho = \|.\|_F^2$, $R^*_{n,p} = \frac{p^2}{n}$.

Proof of Theorem 9.2. (Upper bound) Consider the estimator $\hat{\theta} = \bar{X} = \frac{1}{n}\sum_i X_i \sim \mathcal{N}(\theta, \frac{1}{n}I_p)$. Then $\theta - \hat{=}\sqrt{\frac{1}{n}}Z$ where $Z \sim \mathcal{N}(0, I_p)$. Thus

$$R^*_{n,p} \geq \mathbb{E}[\ell(\theta, \bar{X})] = \mathbb{E}[\rho(\theta - \bar{X})] = \mathbb{E}[\rho(\frac{1}{\sqrt{n}}Z)]. \tag{9.3}$$

(Lower bound) We lower bound the minimax risk $R^*_{n,p}$ by Bayes risk R^*_π with prior $\pi = \mathcal{N}(0, sI_p)$:

$$
\begin{aligned}
R^*_{n,p} &\geq R^*_\pi \\
&= \inf_{\hat{\theta}} \mathbb{E}_\pi[\rho(\theta - \hat{\theta})] \\
&= \inf_{\hat{\theta}} \mathbb{E}[\mathbb{E}[\rho(\theta - \hat{\theta})|X]] \\
&= \mathbb{E}[\inf_{\hat{\theta}} \mathbb{E}[\rho(\theta - \hat{\theta})|X]] \\
&\overset{(a)}{=} \mathbb{E}[\mathbb{E}[\rho(\theta - \mathbb{E}[\theta|X])|X]] \\
&\overset{(b)}{=} \mathbb{E}[\rho(\sqrt{\frac{s}{1+sn}}Z)] \\
&\overset{s\to\infty}{=} \lim_{s\to\infty} \mathbb{E}[\rho(\sqrt{\frac{s}{1+sn}}Z)] \\
&\overset{(c)}{\leq} \mathbb{E}[\lim_{s\to\infty} \rho(\sqrt{\frac{s}{1+sn}}Z)] \\
&\overset{(d)}{=} \mathbb{E}[\rho(\lim_{s\to\infty} \sqrt{\frac{s}{1+sn}}Z)] \\
&= \mathbb{E}[\rho(\frac{1}{\sqrt{n}}Z)] \qquad (9.4)
\end{aligned}
$$

where (a) follows from Anderson's Lemma 9.1, (b) uses $Z \sim \mathcal{N}(0, I_p)$ or $\sqrt{\frac{s}{1+sn}}Z = (\theta - \mathbb{E}[\theta|X]) \sim \mathcal{N}(0, \frac{s}{1+sn}I_p)$ since $\theta|X \sim \mathcal{N}(\frac{sn}{1+sn}, \frac{s}{1+sn}I_p)$, (c) follows from Fatou's Lemma, and (d) follows since $\rho(\cdot)$ is a lower-semicontinuous function.

Combining the upper bound and lower bounds in (9.3) and (9.4), we can say that $R^*_{n,p} = \mathbb{E}[\rho(\frac{1}{\sqrt{n}}Z)]$. □

Lemma 9.1 (Anderson)**.** *Let $X \sim \mathcal{N}(0, \Sigma)$, and $\rho : \mathbb{R}^p \to \mathbb{R}_+$ is a bowl-shaped loss function, then*

$$\min_{y\in\mathbb{R}^p} \mathbb{E}[\rho(y + X)] = \mathbb{E}[\rho(X)].$$

In order to prove Lemma 9.1, it suffices to consider ρ being indicator functions. This is done in the next lemma, which we prove later for simpler exposition.

Lemma 9.2. *Let $K \in \mathbb{R}^p$ be a symmetric convex set and $X \sim \mathcal{N}(0, \Sigma)$ for some covariance matrix Σ. Then $\forall y \in \mathbb{R}, \mathbb{P}(X + y \in K) \leq \mathbb{P}(X \in K)$.*

Proof of Lemma 9.1. Denote the sub-level set set $K_c = \{x \in \mathbb{R}^p : \rho(x) < c\}$. Since ρ is bowl-shaped, K_c is convex and symmetric, which satisfies the conditions of Lemma 9.2. So,

$$
\begin{aligned}
\mathbb{E}[\rho(y + x)] &= \int_0^\infty \mathbb{P}(\rho(y + x) \geq c)dc, \\
&= \int_0^\infty (1 - \mathbb{P}(y + x \in K_c))dc, \\
&\geq \int_0^\infty (1 - \mathbb{P}(x \in K_c))dc, \\
&= \int_0^\infty \mathbb{P}(\rho(x) \geq c)dc, \\
&= \mathbb{E}[\rho(x)].
\end{aligned}
$$

Hence, $\min_{y\in\mathbb{R}^p}\mathbb{E}[\rho(y+x)]=\mathbb{E}[\rho(x)]$. □

Before going into the proof of Lemma 9.2, we need the following definition.

Definition 9.3. A measure μ on $\mathbb{R}^p$ is said to be *log-concave* if

$$\mu(\lambda A+(1-\lambda)B)\geq\mu(A)^\lambda\mu(B)^{1-\lambda}$$

for all measurable $A,B\subset\mathbb{R}^p$ and any $\lambda\in[0,1]$.

The following result characterizes log-concavity of measures in terms of that of its density. See [Rin76] for a proof.

Theorem 9.3 (Prékopa). *A measure μ is log-concave if and only if μ has a density f with respect to the Lebesgue measure, such that f is a log-concave function.*

Example 9.2. Examples of log-concave measures:

- Lebesgue measure: Let $\mu=\mathrm{vol}$ be the Lebesgue measure on $\mathbb{R}^p$, which satisfies Theorem 9.3 ($f\equiv 1$). Then
$$\mathrm{vol}(\lambda A+(1-\lambda)B)\geq\mathrm{vol}(A)^\lambda\mathrm{vol}(B)^{1-\lambda},\tag{9.5}$$
which implies[2] the Brunn-Minkowski inequality:
$$\mathrm{vol}(A+B)^{\frac{1}{p}}\geq\mathrm{vol}(A)^{\frac{1}{p}}+\mathrm{vol}(B)^{\frac{1}{p}}.\tag{9.6}$$
- Gaussian distribution: Let $\mu=\mathcal{N}(0,\Sigma)$, with a log-concave density f since $\log f(x)=-\frac{p}{2}\log(2\pi)-\frac{1}{2}\log\det(\Sigma)-\frac{1}{2}x'\Sigma^{-1}x$ is concave.

Proof of Lemma 9.2. By Theorem 9.3, the distribution of X is log-concave. Then

$$\mathbb{P}[X\in K]\overset{(a)}{=}\mathbb{P}\Big[X\in\frac{1}{2}(K+y)+\frac{1}{2}(K-y)\Big]\tag{9.7}$$
$$\overset{(b)}{\geq}\sqrt{\mathbb{P}[X\in K-y]\mathbb{P}[X\in K+y]}\tag{9.8}$$
$$\overset{(c)}{=}\mathbb{P}[X+y\in K],\tag{9.9}$$

where (a) follows from $\frac{1}{2}(K+y)+\frac{1}{2}(K-y)=\frac{1}{2}K+\frac{1}{2}K=K$ since K is convex; (b) follows from the definition of log-concavity in Definition 9.3 with $\lambda=\frac{1}{2}$, $A=K-y=\{x-y:x\in K\}$ and $B=K+y$; (c) follows from $\mathbb{P}[X\in K+y]=\mathbb{P}[X\in -K-y]=\mathbb{P}[X+y\in K]$ since X has a symmetric distribution and K is symmetric ($K=-K$). □

9.2 LeCam's two-point argument

In this section we study a general method to obtain a lower bound on the minimax risk $R^*_{n,p}(\Theta)$.

[2] Applying (9.5) to $A'=\mathrm{vol}(A)^{-1/p}A$, $B'=\mathrm{vol}(B)^{-1/p}B$ (both of which have unit volume), and $\lambda=\frac{\mathrm{vol}(A)^{1/p}}{\mathrm{vol}(A)^{1/p}+\mathrm{vol}(B)^{1/p}}$.

Theorem 9.4 (LeCam's Method/two-point argument)**.** *Suppose the loss function* $\ell : \Theta \times \Theta \to \mathbb{R}_+$ *satisfies the following* α*-triangle inequality*

$$\ell(\theta_0, \theta_1) \leq \alpha(\ell(\theta_0, \theta) + \ell(\theta_1, \theta)), \quad \forall \ \theta_0, \theta_1, \theta,$$

for some $\alpha > 0$*, then*

$$R^*(\Theta) \triangleq \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \ell(\theta, \hat{\theta}) \geq \sup_{\theta_0, \theta_1 \in \Theta} \frac{\ell(\theta_0, \theta_1)}{4\alpha}(1 - d_{\mathrm{TV}}(P_{\theta_0}, P_{\theta_1})) \tag{9.10}$$

Proof. In general, testing is "easier" than estimation in the statistical sense that one can often convert an estimator to a test. Hence, in LeCam's method, we convert the estimation problem to a hypothesis testing problem by discretizing the set Θ and obtain a lower bound on the worst-case risk R^*.

For simplicity, let us break Θ into two points $\Theta' = \{\theta_1, \theta_2\} \subset \Theta$. Consider the problem, when the distribution P_θ is either P_{θ_1} or P_{θ_2}. Let us consider the risk in this problem using test ψ, where,

$$\psi = \begin{cases} \theta_0 & \ell(\theta_0, \hat{\theta}) \leq \ell(\theta_1, \hat{\theta}) \\ \theta_1 & \ell(\theta_1, \hat{\theta}) < \ell(\theta_0, \hat{\theta}) \end{cases}$$

for any estimate $\hat{\theta}$ for problem $\theta \in \Theta$.

Let us denote the minimax risk obtained in this problem as $R^*(\Theta')$. Since, we are considering a simpler problem of $\theta = \theta_1$ or $\theta = \theta_2$ rather than $\theta \in \Theta$, the risk $R^*(\Theta')$ forms a lower bound to the risk R^*. So,

$$R^* \geq R^*(\Theta') \overset{(b)}{=} R^*_{\theta_0} \vee R^*_{\theta_1}. \tag{9.11}$$

where (b) follows from the definition of minimax risk.

Now, let $\epsilon = \ell(\theta_0, \theta_1)$. The probability of false alarm is defined $P_{\theta_0}(\psi = \theta_1)$ and probability of miss is defined as $P_{\theta_1}(\psi = \theta_0)$. Now,

$$\begin{aligned} P_{\theta_0}(\psi = \theta_1) &= P_{\theta_0}(\ell(\hat{\theta}, \theta_1) \leq \ell(\hat{\theta}, \theta_0)) \\ &\overset{(a)}{\leq} P_{\theta_0}(\ell(\hat{\theta}, \theta_0) \geq \frac{\epsilon}{2\alpha}) \\ &\overset{(b)}{\leq} \frac{2\alpha}{\epsilon} \mathbb{E}_{\theta_0}[\ell(\hat{\theta}, \theta_0)] \end{aligned} \tag{9.12}$$

where (b) follows from Markov's inequality, (a) follows because the α-triangle inequality and separation assumption: $\epsilon \leq \ell(\theta_0, \theta_1) \leq \alpha(\ell(\theta_0, \theta) + \ell(\theta_1, \theta))$.

Similarly, we can establish that the probability of miss detection:

$$P_{\theta_1}[\psi = \theta_0] \leq \frac{2\alpha \mathbb{E}_{\theta_1}[\ell(\hat{\theta}, \theta_1)]}{\epsilon}. \tag{9.13}$$

Now, since $1 - d_{\mathrm{TV}}$ is the minimal total probability of error, we have

$$\begin{aligned} 1 - d_{\mathrm{TV}}(P_{\theta_1}, P_{\theta_0}) &\leq P_{\theta_1}[\psi = \theta_0] + P_{\theta_0}(\psi = \theta_1) \\ &\overset{(a)}{\leq} \frac{2\alpha}{\epsilon}(\mathbb{E}_{\theta_1}[\ell(\hat{\theta}, \theta_1)] + \mathbb{E}_{\theta_0}[\ell(\hat{\theta}, \theta_0)]) \\ &= \frac{2\alpha}{\epsilon}(R_{\theta_0}(\hat{\theta}) + R_{\theta_1}(\hat{\theta})) \\ &\leq \frac{4\alpha}{\epsilon}(R_{\theta_0}(\hat{\theta}) \vee R_{\theta_1}(\hat{\theta})) \\ &\leq \frac{4\alpha}{\epsilon}(R_{\theta_0}(\hat{\theta}) \vee R_{\theta_1}(\hat{\theta})), \end{aligned} \tag{9.14}$$

where (a) follows from (9.13) and (9.12).

Combining (9.14) with (9.11), we can say that

$$R^* \geq \frac{\epsilon}{4\alpha}(1 - d_{\mathrm{TV}}(P_{\theta_1}, P_{\theta_0}))$$

Optimizing over the pair θ_0, θ_1 gives (9.10). □

Example for Theorem 10.1: Suppose $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^q,\ \ q \geq 1$. Then we can easily show that $l(\cdot)$ satisfies $2^{q-1}-$triangle inequality. So, by Theorem 10.1, when $q = 2$,$R^*_{n,p} \geq \sup_{\theta_0,\theta_1} \frac{1}{8}\|\theta_0 - \theta_1\|^2(1 - d_{\mathrm{TV}}(P_{\theta_0}, P_{\theta_1}))$.

Recap:

Theorem 10.1 (Le Cam's Method)**.** *If* $\ell(\theta_0, \theta_1) \leq \alpha\{\ell(\theta_0, \hat{\theta}) + \ell(\theta_1, \hat{\theta_1})\}, \forall \hat{\theta}$ *then*

$$R^* = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta[\ell(\theta, \hat{\theta})] \geq \frac{\ell(\theta_0, \theta_1)}{4\alpha}(1 - d_{\mathrm{TV}}(P_{\theta_0}, P_{\theta_1})) \tag{10.1}$$

Note:

- For n samples, the total variation increases and hence we get a smaller lower bound.
- For different loss functions we have:

$$l = \|.\| \Rightarrow \alpha = 1$$
$$l = \|.\|^q \Rightarrow \alpha = 2^{q-1}$$

- For quadratic loss $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_2^2$, using Theorem 10.1, we have:

$$R^* \geq \frac{\|\theta_0 - \theta_1\|_2^2}{8}(1 - d_{\mathrm{TV}}(P_{\theta_0}, P_{\theta_1})) \tag{10.2}$$

  Can we improve the constant 8 in the above inequality? The answer is YES as we shall show in the next section that it can be replaced by 4 which is the best possible.

10.1 Le Cam's method for quadratic loss

Let Θ be an inner product space. Consider the quadratic loss $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_2^2 = \langle \theta - \hat{\theta}, \theta - \hat{\theta} \rangle$.

Theorem 10.2 (Reduction of factor in (10.2) from 8 to 4)**.**

$$R^* = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta[\|\theta - \hat{\theta}\|_2^2] \geq \frac{\|\theta_0 - \theta_1\|_2^2}{4}(1 - d_{\mathrm{TV}}(P_{\theta_0}, P_{\theta_1})). \tag{10.3}$$

Remark 10.1. The constant 4 is clearly the best possible. In the extreme case of $P_{\theta_0} = P_{\theta_1}$ (non-identifiable), the best estimate is $\arg\min \ell(\theta_0, \hat{\theta}) \vee \ell(\theta_0, \hat{\theta})$, which in the quadratic case is $\frac{1}{2}(\theta_0 + \theta_1)$.

Proof. We use minimiax risk $\geq$ Bayes risk:

$$R^* \geq R^*_\pi$$

Using $\pi = \bar{\lambda}\delta_{\theta_0} + \lambda\delta_{\theta_1}$ as the prior, where $\lambda \in [0,1], \bar{\lambda} = 1 - \lambda$, we have:

$$R_\pi^* = \inf_{\hat{\theta}} \bar{\lambda} E_{\theta_0} \|\theta_0 - \hat{\theta}\|_2^2 + \bar{\lambda} E_{\theta_1} \|\theta_1 - \hat{\theta}\|_2^2 \tag{10.4}$$

$$\Rightarrow R_\pi^* = \int_X \mu(dx)\{\inf_{\hat{\theta}} \bar{\lambda} P_{\theta_0}(x)\|\theta_0 - \hat{\theta}(x)\|_2^2 + \lambda P_{\theta_1}\|\theta_1 - \hat{\theta}(x)\|_2^2\} \tag{10.5}$$

We first consider the following general problem:

$$\begin{aligned}
&\Rightarrow \inf_{\hat{\theta}}\{\bar{\alpha}\|\theta_0 - \hat{\theta}\|_2^2 + \alpha\|\theta_1 - \hat{\theta}\|_2^2\} \\
&\Rightarrow \inf_{\hat{\theta}}\{\|\hat{\theta}\|_2^2 - 2\hat{\theta}(\bar{\alpha}\theta_0 + \alpha\theta_1) + \|\bar{\alpha}\theta_0 + \alpha\theta_1\|_2^2 - \|\bar{\alpha}\theta_0 + \alpha\theta_1\|_2^2 + \bar{\alpha}\|\theta_0\|_2^2 + \alpha\|\theta_1\|_2^2\} \\
&\Rightarrow \inf_{\hat{\theta}}\{\alpha\bar{\alpha}\|\theta_0 - \theta_1\|_2^2 + \|\hat{\theta} - (\bar{\alpha}\theta_0 + \alpha\theta_1)\|_2^2\} = \alpha\bar{\alpha}\|\theta_0 - \theta_1\|_2^2
\end{aligned}$$

So we basically have the conditional mean as the estimate for the above problem which is intuitively correct. We now normalize (10.4) and use the above result to get:

$$\begin{aligned}
R_\pi^* &= \lambda\hat{\lambda}\|\theta_0 - \theta_1\|_2^2 \int_X \mu(dx)\frac{P_{\theta_0}P_{\theta_1}}{\bar{\lambda}P_{\theta_0} + \lambda P_{\theta_1}} \\
&= \lambda\hat{\lambda}\|\theta_0 - \theta_1\|_2^2 E_{\theta_0}\{\frac{P_{\theta_1}}{\bar{\lambda}P_{\theta_0} + \lambda P_{\theta_1}}\}
\end{aligned}$$

Now, we observe that $\bar{\lambda}P_{\theta_0} + \lambda P_{\theta_1} \leq P_{\theta_0} \vee P_{\theta_1}$. Using this fact, we have:

$$\begin{aligned}
R_\pi^* &\geq \lambda\hat{\lambda}\|\theta_0 - \theta_1\|_2^2(\int_X \mu(dx)(P_{\theta_0} \vee P_{\theta_1})) \\
&= \frac{1}{4}\|\theta_0 - \theta_1\|_2^2(1 - d_{\mathrm{TV}}(P_{\theta_0}, P_{\theta_1}))
\end{aligned}$$

where we used $\lambda = \bar{\lambda} = \frac{1}{2}$. □

10.2 Two-point method

For two-point method, we strip off the uncertainty by choosing only 2 possible values of the parameters. So we have:

$$\begin{aligned}
R_\pi^* &\geq R^*(\{\theta_0, \theta_1\}) \\
&= \sup_\pi R_\pi^*
\end{aligned}$$

where the last equality follows from minimax theorem (which holds here since we consider a finite set of parameters). Now, for the optimal Bayes Risk we have:

$$\begin{aligned}
R_\pi^* &= \inf_{\hat{\theta}:\mathcal{X}\to\Theta} \bar{\lambda}E_{\theta_0}\ell(\theta_0, \hat{\theta}) + \lambda E_{\theta_1}\ell(\theta_1, \hat{\theta}) \\
&= E_{\theta_0} \inf_{\hat{\theta}:\mathcal{X}\to\Theta}\{\bar{\lambda}\ell(\theta_0, \hat{\theta}) + \lambda\frac{P_{\theta_1}}{P_{\theta_0}}\ell(\theta_1, \hat{\theta})\}
\end{aligned}$$

Note: We could change the order of expectation and infimum in the above equation as the infimum is over $\hat{\theta}$ which depends only on data.

We now define $\bar{\lambda}\ell(\theta_0, \hat{\theta}) + \lambda\frac{P_{\theta_1}}{P_{\theta_0}}\ell(\theta_1, \hat{\theta}) \; = F(\frac{P_{\theta_1}}{P_{\theta_0}})$. Therefore, we have:

$$\Rightarrow R_\pi^* = E_{\theta_0}\{F(\frac{P_{\theta_1}}{P_{\theta_0}})\}$$

Example 10.1 (Quadratic Loss Function)**.** If $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_2^2$, then $R_\pi^* =$ Expected value of an f-divergence between P_{θ_0} and P_{θ_1}.

We can choose an f-divergence which suits our needs.

So for two-point method, we have:

$$\begin{aligned} R^*(\Theta) &\geq R^*(\{\theta_0, \theta_1\}) \\ &\geq \text{Function of (separation between } \theta_0 \text{ and } \theta_1\text{, separation between } P_{\theta_0} \text{ and } P_{\theta_1}) \end{aligned}$$

Remark 10.2. Since the separation between P_{θ_0} and P_{θ_1} is quantified using f-divergences, we can lower bound the minimax risk in terms of f-divergences other than total variation as well as follows:

- Using Le Cam's method, we can find a bound using total variation and then replace total variation with other f-divergences like χ^2 or hellinger distance.

- We can also use some other f-divergence directly instead of using total variation.

10.3 How good is Le Cam's bound?

In this section, we try to understand how tight Le Cam's bound is. To gain insight, we first consider the following example:

Example 10.2 (p-dimensional, n-sample Gaussian Location Model)**.** For p-dimensional, n-sample GLM, we use $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ as the estimate. So we have $\bar{X} \sim N(\theta, \frac{1}{n}I_p)$. We also know from Theorem 9.1 that for $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_2^2$, we have $R^* = \frac{p}{n}$. Let us compare this result with the lower bound obtained using Le Cam's method:

$$\begin{aligned} R^* &\geq \sup_{\theta_0, \theta_1 \in R^p} \frac{1}{4}\|\theta_0 - \theta_1\|_2^2 (1 - d_{\text{TV}}(N(\theta_0, \frac{1}{n}I_p), N(\theta_1, \frac{1}{n}I_p))) \\ &= \sup_{\theta \in R^p} \frac{1}{4}\|\theta\|_2^2 (1 - d_{\text{TV}}(N(0, \frac{1}{n}I_p), N(\theta, \frac{1}{n}I_p))) \\ &= \frac{1}{4n} \sup_{\theta \in R^p} \|\theta\|_2^2 (1 - d_{\text{TV}}(N(0, I_p), N(\theta, I_p))) \\ &= \frac{1}{4n} \sup_{s > 0} s^2 (1 - d_{\text{TV}}(N(0,1), N(s,1))), \end{aligned}$$

where the second step follows from the fact that we can replace θ_0 by 0 and θ_1 by θ with out any loss of generality, and the last step follows from the following: to reduce the total variation to one dimension, we simply rotate the vector θ to reduce the problem to that of one-dimensional total variation calculation

$$\begin{aligned} d_{\text{TV}}(N(0, I_p), N(\theta, I_p)) &= d_{\text{TV}}(N(0, I_p), N(\|\theta\| e_1, I_p) \\ &= d_{\text{TV}}(N(0,1), N(\|\theta\|, 1)). \end{aligned}$$

Clearly, the upper bound above doesn't scale with the dimension p.

How to scale R^* with p? We observe that we have considered a similar model as previous lectures and hence using tensorization of 1-dimensional n-sample GLM, we can conclude R^* should linearly grow in p. Explanation: Since $\ell(\theta, \hat{\theta}) = \sum_{i=1}^p \ell(\theta_i, \hat{\theta}_i)$, and each dimension of vector θ is estimated using corresponding dimension of the vector $\bar{X}$. Hence, as each dimension has a constant lower bound, the vector should have a lower bound scaling linearly with p as its lower bound is the sum of respective one-dimensional lower bounds. Therefore, we have $pR^*_{\pi_{1-d}} \leq R^*_p \leq pR^*_{1-d}$.

To improve upon the lower bound obtained using Le Cam's method, we consider more than two points to obtain the minimax bound. In next section, we shall discuss Assouad's Lemma which consider a hypercube instead of a line.

10.4 Assouad's Lemma

Lemma 10.1 (Assouad's Lemma)**.** *If each coordinate consists of binary testing, i.e. $\theta \in \{0,1\}^p \subset \Theta = R^p$ and $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_1$, then:*

$$R^* \geq \frac{p}{4}(1 - \max_{d(\theta,\theta')=1} d_{\text{TV}}(P_\theta, P_{\theta'}))$$

Proof. Since minimiax risk is greater than Bayes risk, we have $R^* \geq R^*_\pi$. Also we consider a uniform prior over $\{0,1\}^p$. We also define $\tilde{\theta}_i$ as follows:

$$\tilde{\theta}_i = \begin{cases} 0, & \hat{\theta}_i < \frac{1}{2} \\ 1, & \text{otherwise} \end{cases}$$

Therefore, $\forall \hat{\theta} : X \to R^p$, we have:

$$\begin{aligned} E\|\theta - \hat{\theta}\|_1 &= \sum_{i=1}^p E|\theta_i - \hat{\theta}_i| \\ &\geq \frac{1}{2}\sum_{i=1}^p E|\theta_i - \tilde{\theta}_i| \\ &= \frac{1}{2}\sum_{i=1}^p P(\theta_i \neq \tilde{\theta}_i) \\ &\geq \frac{1}{2}\sum_{i=1}^p \inf_{\hat{\theta}_i = \hat{\theta}_i(X)} P(\theta_i \neq \hat{\theta}_i) \end{aligned}$$

Since, $\theta_i \in \{0,1\}$, we have:

$$E\|\theta - \hat{\theta}\|_1 \geq \frac{1}{4}\sum_{i=1}^p (1 - d_{\text{TV}}(P_{X|\theta_i=0}, P_{X|\theta_i=1})) \tag{10.6}$$

We now try to upper bound the total variation expression in the above inequality. From Bayes rule, we get:

$$d_{\text{TV}}(P_{X|\theta_i=0}, P_{X|\theta_i=1}) = d_{\text{TV}}(\frac{1}{2^{p-1}}\sum_{\theta:\theta_i=1} P_\theta, \frac{1}{2^{p-1}}\sum_{\theta:\theta_i=0} P_\theta)$$

Using convexity of total variation, we have:

$$\begin{aligned} d_{\text{TV}}(P_{X|\theta_i=0}, P_{X|\theta_i=1}) &\leq \frac{1}{2^{p-1}}\sum_{\theta_{\setminus i}\in\{0,1\}^{p-1}} d_{\text{TV}}(P_{\{\theta_{\setminus i},1\}}, P_{\{\theta_{\setminus i},0\}}) \\ &\leq \max_{d(\theta,\theta')=1} d_{\text{TV}}(P_\theta, P_{\theta'}) \end{aligned}$$

Using the above result in (10.5) and using the fact that $\ell(\theta, \hat{\theta} = \sum_{i=1}^{p} \ell(\theta_i, \hat{\theta}_i)$, we get:

$$R^* \geq \frac{l(0,1)p}{4}(1 - \max_{d(\theta,\theta')=1} d_{\text{TV}}(P_\theta, P_{\theta'}))$$

For l_1 loss function, $l(0,1) = 1$, hence we obtain the result. □

Example 10.3 (p-dimensional, n-sample Gaussian Location Model(GLM))**.** We consider $\ell(\theta, \hat{\theta}) = \sum_{i=1}^{p}(\theta_i - \hat{\theta}_i)^2, \theta \in \{0, \epsilon\}^p$. Using Assoud's Lemma, we get:

$$\begin{aligned} R^* &\geq \frac{\epsilon^2 p}{4}\{1 - \max_{\theta,\theta' \in \{0,\epsilon\}^p, d(\theta,\theta')=1} d_{\text{TV}}(N(\theta, \frac{1}{n}I_p), N(\epsilon, \frac{1}{n}I_p))\} \\ &= \frac{\epsilon^2 p}{4}\{1 - d_{\text{TV}}(N(0, \frac{1}{n}I_p), N(\epsilon, \frac{1}{n}I_p))\} \end{aligned}$$

Using $\epsilon = \frac{1}{\sqrt{n}}$ and scaling by $\frac{1}{n}$, we get:

$$R^* \geq \frac{kp}{n}$$

where $k = 1 - d_{\text{TV}}(N(0,1), N(1,1))$ is a constant (~ 0.7).

In the next lecture we will talk more about Assouad's Lemma which considers a hypercube of parameters. We will also introduce Fano's Lemma which uses a pyramid of parameters instead of a hypercube.

§ 11. Mutual Information Method

Quick review: Assouad's lemma

In Assouad's lemma discussed in the last lecture, we made the following assumptions:

- Parameter space is a hypercube embedded in $\mathbb{R}^p$, i.e. $\Theta = \{\theta_0, \theta_1\}^p$.
- Loss function $l(\theta, \hat{\theta})$ is *separable*, i.e. $l(\theta, \hat{\theta}) = \sum_{i=1}^p l(\theta_i, \hat{\theta}_i)$, (e.g. Hamming, ℓ_2 squared.)
- and satisfies α-triangle inequality, i.e. $l(\theta_0, \theta_1) \le \alpha \left[l(\theta_0, \hat{\theta}) + l(\theta_1, \hat{\theta}) \right]$.

Letting $\pi \sim \text{Unif}(\Theta)$, we could proceed as:

$$\begin{aligned}
\mathcal{R}^* \ge \mathcal{R}^*_\pi = \inf_{\hat{\theta}} \sum_{i=1}^p \mathbb{E}_{\theta\sim\pi}[l(\theta_i, \hat{\theta}_i)] &= \sum_{i=1}^p \inf_{\hat{\theta}_i} \mathbb{E}_{\theta\sim\pi}[l(\theta_i, \hat{\theta}_i)] \\
&\overset{\text{Le Cam}}{\ge} \sum_{i=1}^p \frac{l(\theta_0, \theta_1)}{4\alpha} \left[1 - d_{\text{TV}}(P_{X|\theta_i=\theta_0}, P_{X|\theta_i=\theta_1})\right] \\
&\overset{\text{convexity}}{\ge} \frac{p \cdot l(\theta_0, \theta_1)}{4\alpha} \left[1 - \max_{d_H(\theta,\theta')=1} d_{\text{TV}}(P_\theta, P_{\theta'})\right].
\end{aligned}$$

where the last line could be thought of as a "deteriorated" version.

Example 11.1 (Gaussian Location Model)**.** As usual, let $Z \sim \mathcal{N}(0, I_p)$, $\Theta = \left\{0, \frac{1}{\sqrt{n}}\right\}^p$, and $l(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_2^2$ (which satisfies 2-triangle inequality). Then, Assouad's lemma gives us:

$$\mathcal{R}^* \ge \frac{p}{8n}\left[1 - \max_{d_H(\theta,\theta')=1} d_{\text{TV}}(P_\theta, P_{\theta'})\right] = \frac{p}{8n}\left[1 - d_{\text{TV}}\left(\mathcal{N}\left(0, \frac{1}{n}\right), \mathcal{N}\left(\frac{1}{\sqrt{n}}, \frac{1}{n}\right)\right)\right] \simeq \frac{0.3}{8}\frac{p}{n}$$

which is not very good compared to $\frac{p}{n}$.

Along with the above example, the fact that the loss function is not always separable (e.g. ℓ_∞) necessitates a search a more versatile method. In this lecture, we discuss the "mutual information method" where the most important measure of information would be, of course, the mutual information $I(X;Y)$.

11.1 Mutual Information $I(X;Y)$

Recall that KL-divergence was defined using the function $f(x) = x \log x$:

$$D(P\|Q) \triangleq \mathbb{E}_Q\left[\frac{P}{Q}\log\frac{P}{Q}\right] = \mathbb{E}_P\left[\log\frac{P}{Q}\right].$$

Now, the mutual information can be defined using KL-divergence.

Definition 11.1 (Mutual Information)**.** Given a joint probability distribution P_{XY}, the mutual information between X and Y is defined as

$$I(X;Y) \triangleq D(P_{XY}\|P_X P_Y),$$

the distance between the original distribution and the hypothetical distribution assuming that X and Y are independent.

Mutual information has the following useful properties:

Proposition 11.1 (Properties of Mutual Information)**.** *Followings are true:*

1. $I(X;Y) = D(P_{Y|X}\|P_Y|P_X) = \mathbb{E}_{x\sim P_X}[D(P_{Y|X=x}\|P_Y)]$
2. *(Symmetry)* $I(X;Y) = I(Y;X)$.
3. *(Measure of dependency)* $I(X;Y) \geq 0$ *with equality iff* $X \perp\!\!\!\perp Y$.
4. *(I vs H: Y discrete)* $I(X;Y) = H(Y) - H(Y|X)$, *where* $H(Y)$ *denotes the Shannon entropy* $H(Y) \triangleq \sum_y P_Y(y)\log\frac{1}{P_Y(y)}$.
5. *(I vs h: Y continuous)* $I(X;Y) = h(Y) - h(Y|X)$, *where* $h(Y)$ *denotes the differential entropy* $h(Y) \triangleq \int f_Y(y)\log\frac{1}{f_Y(y)}dy$.

Example 11.2 (Additive noise: binary)**.** Let $Y = X \oplus Z$, where $X \sim \text{Bern}(\delta)$, $Z \sim \text{Bern}(\epsilon)$, $X \perp\!\!\!\perp Z$, and $\oplus$ denotes the XOR operation (binary addition). Then,

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) = H(Y) - H(X\oplus Z|X) = H(Y) - H(Z) \\ &= H(\text{Bern}(\delta \star \epsilon)) - H(\text{Bern}(\epsilon)) = h(\delta \star \epsilon) - h(\epsilon). \end{aligned}$$

where $h(t) = t\log\frac{1}{t} + (1-t)\log\frac{1}{1-t}$ is the binary entropy function (not differential entropy!) and the convolution operation $\star$ on $[0,1]$ for Bernoulli random variables is $\delta \star \epsilon = \delta\bar{\epsilon} + \bar{\delta}\epsilon$.

Example 11.3 (Additive noise: Gaussian)**.** Let $Y = X + Z$ where $X \sim \mathcal{N}(0,S)$, $Z \sim \mathcal{N}(0,1)$, $X \perp\!\!\!\perp Z$. Then,

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y|X) = h(Y) - h(Z) \\ &= h(\mathcal{N}(0,1+S)) - h(\mathcal{N}(0,1)) = \frac{1}{2}\log(1+S). \end{aligned}$$

Alternatively, we could do

$$I(X;Y) = D(P_{Y|X}\|P_Y|P_X) = \mathbb{E}_{x\sim\mathcal{N}(0,S)}[D(\mathcal{N}(x,1)\|\mathcal{N}(0,1+S))].$$

to arrive the same conclusion.

Like f-divergence, the mutual information has a very useful property when applied on Markov chains: the data processing inequality. In fact, the data processing inequality of mutual information is a direct consequence of that of KL-divergence.

Theorem 11.1 (Data processing inequality for M.I.)**.** *Let* $X \to Y \to Z$ *forms a Markov chain. Then,*

$$I(X;Z) \leq I(X;Y).$$

Proof. For the same kernel $P_{Z|Y}$, we have $P_{Y|X=x} \xrightarrow{P_{Z|Y}} P_{Z|X=x}$ for each x and similarly $P_Y \xrightarrow{P_{Z|Y}} P_Z$. Hence applying the data processing inequality for KL divergence yield

$$I(X;Z) = D(P_{Z|X} \| P_Z | P_X) \le D(P_{Y|X} \| P_Y | P_X) = I(X;Y).$$

□

Remark 11.1. For the longer Markov chain $W \to X \to Y \to Z$, we have $I(W;Z) \le I(X;Y)$.

Remark 11.2. For other f-divergences, we can define $I_f(X;Y) \triangleq D_f(P_{Y|X} \| P_Y | P_X)$ which naturally satisfies the data processing inequality on Markov chain.

For a detailed explanation on the materials presented in this section, please refer to [PW15, Ch.2.1-2.2] or [CT06].

11.2 Mutual information method: minimax lower bound

Here's the main idea of the mutual information method: As usual, we are trying to estimate the parameter θ distributed by some prior π, using the estimator $\hat{\theta}$ using the experiment X as its input. In other words, we have a Markov chain $\theta \to X \to \hat{\theta}$.

Then we can upper-bound the mutual information between θ and $\hat{\theta}$ as follows:

$$I(\theta, \hat{\theta}) \le I(\theta; X) \le \sup_{\pi \in \mathcal{M}(\Theta)} I(\theta; X),$$

where the first inequality is due to the data processing inequality of mutual information. The second inequality could be used to drop the assumption that we know the prior π, and is useful when the data X does not provide enough information about θ.

For the lower bound, we have the following:

$$I(\theta, \hat{\theta}) \ge \inf_{P_{\hat{\theta}|\theta} : \mathbb{E}l(\theta, \hat{\theta}) \le \mathcal{R}^*_\pi} I(\theta; \hat{\theta}),$$

for any 'good' $\hat{\theta}$ that satisfies $\mathbb{E}l(\theta, \hat{\theta}) \le \mathcal{R}^*_\pi$. This could be interpreted as a minimum amount of information required for an estimation task.

Also notice the followings:

- This line of inequalities is akin to the converse proof of joint-source channel coding in information theory, with the capacity-like upper bound and rate-distortion-like lower bound.
- Only the lower bound is related to the loss function.
- Sometimes we need a smart choice of the prior.

11.3 Extremization of the mutual information

A good news is that we have the convexity and concavity of the mutual information at hand, which could help us find the infimum and supremum of the mutual information. In specific, we have the following property cf. [PW15, p.28]:

Proposition 11.2 (Convexity and Concavity of mutual information)**.** *Consider the notation* $I(P_X, P_{Y|X}) = I(X;Y)$. *Then*

- *For fixed $P_{Y|X}$, $P_X \mapsto I(P_X, P_{Y|X})$ is concave.*
- *For fixed P_X, $P_{Y|X} \mapsto I(P_X, P_{Y|X})$ is convex.*

The upper bound, or the maximization part, is the following task: given $P_{Y|X}$, we want to find $\max_{P_X \in \mathcal{P}} I(X;Y)$ where $\mathcal{P}$ is a convex set.

Example 11.4 (GLM, upper bound)**.** Again let $Y = X + Z$, where $Z \sim \mathcal{N}(0, I_p)$ and $X \perp\!\!\!\perp Z$. However, in this case we do not know the prior distribution of X. Rather, we consider a convex set of priors $\mathcal{P} = \{P_X : \mathbb{E}\|X\|_2^2 \leq p \cdot s\}$, the signals with constrained average per-dimension power. Then, by the well-known formula for Gaussian channel capacity cf. [PW15, p.33]

$$\max_{P_X \in \mathcal{P}} I(X; X + Z) = \frac{p}{2} \log(1 + s).$$

The lower bound, or the minimization part, is: given P_X, we want to find $\min_{P_{Y|X} \in \mathcal{P}} I(X;Y)$.

Example 11.5 (GLM, lower bound)**.** We are only assuming that $X \sim \mathcal{N}(0, s \cdot I_p)$. In the case of the squared distortion, it is known that [PW15, p.33]

$$\min_{P_{Y|X} : \|Y - X\|^2 \leq p \cdot \epsilon} I(X;Y) = \begin{cases} \frac{p}{2} \log\left(\frac{s}{\epsilon}\right) & \epsilon < s \\ 0 & \text{otherwise.} \end{cases}$$

For non-Gaussian cases, it is in general difficult to find the bounds exactly, and in the following lectures we would discuss the further bounding on both bounds. But before that, we provide several more examples.

Example 11.6 (Bernoulli, lower bound)**.** Let $X \sim \text{Bern}(\delta)^{\otimes p}$. Then for $\epsilon < \delta < \frac{1}{2}$,

$$\min_{P_{Y|X} : \mathbb{E} d_H(X,Y) \leq p \cdot \epsilon} I(X;Y) = p\left[h(\delta) - h(\epsilon)\right].$$

Example 11.7 (p-dim, n-sample GLM, quadratic loss, combining bounds)**.** Let $\theta \sim \mathcal{N}(0, S \cdot I_p)$, and $\theta \to X \to \hat{\theta}$ hold. Following the usual assumptions, we have $P_{X|\theta} \sim \mathcal{N}(\theta, \frac{1}{n} I_p)$. Then, from the upper bound we know

$$I(\theta, \hat{\theta}) \leq I(\theta; X) = \frac{p}{2} \log(1 + S \cdot n).$$

From the lower bound, we have:

$$I(\theta, \hat{\theta}) \geq \min_{P_{\hat{\theta}|\theta} : \|\hat{\theta} - \theta\|_2^2 \leq \mathcal{R}_\pi^*} I(\theta; \hat{\theta}) = \frac{p}{2} \log \frac{S}{\mathcal{R}_\pi^* / p}.$$

Combining the preceeding two displays we get

$$\mathcal{R}_\pi^* \geq \frac{S \cdot p}{1 + S \cdot n}$$

which becomes $\mathcal{R}^* \geq \frac{p}{n}$ as $S \to \infty$ and, surprisingly, recovers the exact minimax risk in Theorem 9.1 without loss of any constant factor.,

Note: Statistical estimation task could be represented as a Markov chain $\theta \to X \to \hat{\theta}$ where $P_{X|\theta}$ is given by the model and $P_{\hat{\theta}|X}$ is the estimator we design. In comparison, in data transmission we have the Markov chain $\theta \to X \to Y \to \hat{\theta}$ where $P_{Y|X}$ is fixed by the channel and we design the "encoder" $P_{X|\theta}$ and decoder $P_{\hat{\theta}|Y}$.

11.4 Coming next

Starting from the next lecture, we discuss various methods to further upper and lower bound $I(\theta; \hat{\theta})$. In specific:

- Fano's method is again about reducing the estimation into testing, thereby forming the Markov chain $\theta \to X \to \hat{\theta} \to \hat{\theta}_{\text{test}}$, and investigating the value $\min I(\theta; \hat{\theta}_{\text{test}})$.

- Mutual information would be view as an information radius, and we would use the fact that radius is upper bounded by diameter, which would be more easily characterized.

§ 12. Mutual Information Method: Continued

12.1 Recap: Mutual Information Method

We have several equivalent definitions of mutual information from last class, capturing a measure of how far X and Y are from independence, or how much information about Y is provided by X:

$$\begin{aligned} I(X;Y) &= D(P_{XY} \| P_X P_Y) \\ &= D(P_{Y|X} \| P_Y | P_X) = \mathbb{E}_{x \sim P_X}[D(P_{Y|X=x} \| P_Y)] \\ &= \inf_{Q : X \perp\!\!\!\perp Y \text{under } Q} D(P_{XY} \| Q_{XY}) \end{aligned}$$

Given the normal model $\theta \to X \to \hat{\theta}$, where θ generates the data X which generates an estimate $\hat{\theta}$, we can use the mutual information method to bound $I(\theta; \hat{\theta})$. In particular, as we saw last time, the following chain of inequalities always holds:

$$\min_{P_{\hat{\theta}|\theta} : \mathbb{E}\ell(\theta, \hat{\theta}) \leq R_\pi^*} I(\theta; \hat{\theta}) \leq I(\theta; \hat{\theta}) \leq I(\theta; X) \leq \max_{P_\theta \in \mathcal{M}(\theta)} I(\theta; X)$$

We like to think of the left-most lower bound as the "cost" of an estimation task, which depends only on the prior and the loss function, but not on how the data is collected. We think of $\max_{P_\theta \in \mathcal{M}(\theta)} I(\theta; X)$ as the "capacity" of the model, which depends only on the model itself. Last lecture, we were able to compute the cost and capacity exactly for the Gaussian Location Model. In general, we may not be able to exactly compute the cost and capacity, so we will focus on methods for bounding them in this lecture.

12.2 Tensorization of Mutual Information

First, we would like to develop tools for bounding the mutual information of not just random variables, but random vectors as well. The chain rule for mutual information gives us an intuitive way to express the mutual information of a random vector as a sum of the mutual information of one-dimensional random variables:

Theorem 12.1 (Mutual Information Chain Rule)**.** *Let the random vector $X = (X_1, \ldots, X_k)$ be jointly distributed with Y. Then:*

$$\begin{aligned} I(X;Y) &= I(X_1, X_2, \ldots, X_k; Y) \\ &= I(X_1;Y) + I(X_2;Y|X_1) + \ldots + I(X_k; Y|X^{k-1}) \end{aligned}$$

The proof of the chain rule follows from telescoping logs. For more information, see section 2.5 of [CT06]. In general, we cannot remove the conditioning and bound $I(X;Y)$ from above or below by $\sum_i I(X_i;Y)$. However, in some situations it is possible.

Example 12.1 (Tensorization in extremization problem)**.** Suppose $X = (X_1, \ldots, X_k)$ and $Y = (Y_1, \ldots, Y_k)$ are random vectors, and each coordinate of Y depends only on the corresponding coordinate of X:

$$\begin{gathered} X_1 \to Y_1 \\ X_2 \to Y_2 \\ \vdots \\ X_k \to Y_k \end{gathered}$$

Then the conditional distribution of Y given X factors:

$$P_{Y|X} = \prod_{i=1}^{k} P_{Y_i|X_i}$$

So long as the channels are decoupled like this, we have:

$$I(X;Y) \leq \sum_{i=1}^{k} I(X_i, Y_i)$$

with equality if the X_i are independent from each other. Therefore, in particular:

$$\max_{P_X} I(X;Y) = \sum_{i=1}^{k} \max_{P_{X_i}} I(X_i, Y_i)$$

We can also consider a minimization problem for $I(X;Y)$. For example, if the coordinates of X are independent, i.e.:

$$P_X = \prod_{i=1}^{k} P_{X_i}$$

then we get can a lower bound on the mutual information:

$$I(X;Y) \geq \sum_{i=1}^{k} I(X_i, Y_i)$$

Equality holds when the coordinates of Y depend only on the corresponding coordinates of X, so $\min_{P_{Y|X}} I(X;Y)$ is achieved at the product of minimizers:

$$\min_{P_{Y|X}} I(X;Y) = \sum_{i=1}^{k} \min_{P_{Y_i|X_i}} I(X_i;Y_i)$$

In GLM, we could get nice bounds through the product structure. Otherwise, if there is no product structure, we would need to use the chain rule, which can be more difficult.

12.3 Capacity as Information Radius

To start, let us consider another way of thinking about mutual information.

Theorem 12.2 (Another Representation of Mutual Information)**.**

$$I(X;Y) = \min_Q D(P_{Y|X}\|Q|P_X)$$

Proof. For any Q we have:

$$\begin{aligned} I(X;Y) &= D(P_{Y|X}\|P_Y|P_X) \\ &= \mathbb{E}\log \frac{P_{Y|X}}{Q}\frac{Q}{P_Y} \\ &= D(P_{Y|X}\|Q|P_X) - D(P_Y\|Q) \end{aligned}$$

We get the desired result by noting that $D(P_Y\|Q) \geq 0$ and optimizing over Q. In particular, we can bound the mutual information using a convenient choice of Q, as we will see in the next example:

Example 12.2 (GLM)**.** Suppose $X \sim P_\theta = \mathcal{N}(\theta, 1)$. Then, choosing the best possible Gaussian Q and applying the above bound, we have:

$$\begin{aligned} I(\theta, X) &\leq \mathbb{E}_\theta D(P_\theta\|Q) \\ &= \inf_{\mu\in\mathbb{R}, s\geq 0} D(\mathcal{N}(\theta,1)\|\mathcal{N}(\mu, S)) \\ &= \frac{1}{2}\log(1+\mathrm{Var}(X)) \end{aligned}$$

where the solution to the minimization problem comes from the well-known formula for Gaussian channel capacity [PW15, p. 28].

Geometric Interpretation

The above representation of mutual information has a nice geometric picture, as follows: Let $\mathcal{X}$ be some space, let $\ell : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ be a loss function, and let A be a subset of $\mathcal{X}$.

Definition 12.1 (Radius of a Set)**.** The radius of A is the smallest ball that covers A. Note that we do not require the center y of the ball to be contained in A:

$$\mathrm{rad}(A) \triangleq \inf_{y\in\mathcal{X}} \sup_{x\in A} \ell(x,y)$$

Definition 12.2 (Diameter of a Set)**.** The diameter of A is the largest loss between two points in A:

$$\mathrm{diam}(A) \triangleq \sup_{x,y\in A} \ell(x,y)$$

Remark 12.1. Note that $\mathrm{rad}(A) \leq \mathrm{diam}(A)$. If ℓ satisfies the triangle inequality, then we further have $\mathrm{rad}(A) \geq \frac{1}{2}\mathrm{diam}(A)$.

Nothing above required ℓ to be a valid metric. In fact, we will be examining the following case where ℓ is not symmetric and does not satisfy the triangle inequality:

- $A = \{P_\theta : \theta \in \Theta\} \triangleq \mathcal{P}$
- $\ell(P, Q) = D(P\|Q)$
- $\mathrm{rad}(\mathcal{P}) = \inf_Q \sup_{P\in\mathcal{P}} D(P\|Q)$
- $\mathrm{diam}(\mathcal{P}) = \sup_{P,Q\in\mathcal{P}} D(P\|Q)$

By bounding the radius of $\mathcal{P}$, we can now upper bound the capacity of $\mathcal{P}$.

Theorem 12.3 (Capacity Bounded by Radius). *Suppose we have the model $\theta \to X$, where $\mathcal{P} = \{P_\theta\}$ is defined as above. Let $C(\mathcal{P})$ be the capacity of $\mathcal{P}$. Then:*

$$C(P) \le \mathrm{rad}(\mathcal{P}) \le \mathrm{diam}(\mathcal{P})$$

Proof. Using Theorem 12.2, we have:

$$\begin{aligned}
C(\mathcal{P}) &= \sup_{P_\theta \in \mathcal{M}(\theta)} I(\theta; X) \\
&= \sup_{P_\theta \in \mathcal{M}(\theta)} \inf_Q D(P_{X|\theta}\|Q|P_\theta) \\
&\le \inf_Q \sup_{P_\theta \in \mathcal{M}(\theta)} D(P_{X|\theta}\|Q|P_\theta) \\
&= \inf_Q \sup_{\theta\in\Theta} D(P_\theta\|Q) \\
&= \mathrm{rad}(\mathcal{P}) \\
&\le \mathrm{diam}(\mathcal{P}) = \sup_{\theta,\theta'\in\Theta} D(P_\theta\|P_{\theta'})
\end{aligned}$$

Note: In fact, if $\mathcal{P}$ is convex, then we have equality in the third step, which would give us $C(\mathcal{P}) = \mathrm{rad}(\mathcal{P})$. This is a result of Kemperman (cf. [PW15, Theorem 4.5]).

Example 12.3 (GLM, bounded mean). Let $\mathcal{P} = \{P_\theta\} = \{\mathcal{N}(\theta, n^{-1}) : |\theta| \le \delta\}$. We can bound the radius of P_θ, taking $Q \sim \mathcal{N}(0, n^{-1})$:

$$\begin{aligned}
\mathrm{rad}(\mathcal{P}) &= \inf_Q \sup_{|\theta|\le\delta} D\left(\mathcal{N}(\theta, n^{-1})\|Q\right) \\
&\le \sup_{|\theta|\le\delta} D\left(\mathcal{N}(\theta, n^{-1})\|\mathcal{N}(0, n^{-1})\right) \\
&= \sup_{|\theta|\le\delta} \frac{n}{2}\theta^2 \\
&= \frac{n\delta^2}{2}
\end{aligned}$$

We have used the fact that the KL divergence between two normal distributions with mean u and v and identical variance σ^2 is $\frac{1}{2\sigma^2}|u-v|^2$. We can also compute the diameter quite easily:

$$\begin{aligned}
\mathrm{diam}(\mathcal{P}) &= \sup_{\theta,\theta'\in[\pm\delta]} D(\mathcal{N}(\theta, n^{-1})\|\mathcal{N}(\theta', n^{-1})) \\
&= \frac{n}{2} \sup_{\theta,\theta'\in[\pm\delta]} |\theta-\theta'|^2 \\
&= 2n\delta^2
\end{aligned}$$

Note that in this case, using the diameter instead of the radius only loses a factor of 4.

Now, we can proceed to the more general bounded GLM:

Theorem 12.4 (Bounded GLM)**.** *Let $X \sim P_\theta = \mathcal{N}(\theta, \frac{1}{n}I_p)$. Let $\ell(\theta, \theta') = \|\theta - \theta'\|_2^2$ (quadratic loss), and let $\Theta = B_2(0, \rho) \subset \mathbb{R}^p$. Then:*

$$R^* \asymp \frac{p}{n} \wedge \rho^2$$

Remark 12.2. The interpretation is that if ρ^2 is small and either we do not have enough samples or dimension is very high so that $\frac{p}{n}$ is smaller than ρ^2, then we should discard all the your data and declare zero as the estimate, because data do not provide better resolution than the prior information.

Proof. (Upper bound) We are already done here. Using $\overline{X}$ as an estimator, we have from previous lectures that (up to constant factors for these bounds):

$$R^* \leq \frac{p}{n}$$

Using 0 as an estimator, we just showed:

$$R^* \leq \rho^2$$

Therefore $R^* \leq \frac{p}{n} \wedge \rho^2$.

(Lower bound) First, to make things simpler, we will consider the case where $p = 1$. Before, when obtaining a lower bound on minimax risk, we used a Gaussian prior. However, we cannot use such a prior in this case because the Gaussian distribution is not supported on a ball of radius ρ. Instead, we will choose a uniform prior $\pi \sim \text{Uniform}(-r, r)$, with $r < \rho$. As before, we have:

$$\min_{P_{\hat{\theta}|\theta}:\mathbb{E}\ell(\theta,\hat{\theta})\leq D} I(\theta; \hat{\theta}) \leq I(\theta; \hat{\theta}) \leq I(\theta; X) \leq \text{rad}(\{\mathcal{N}(\theta, \frac{1}{n}) : |\theta| \leq r\})$$

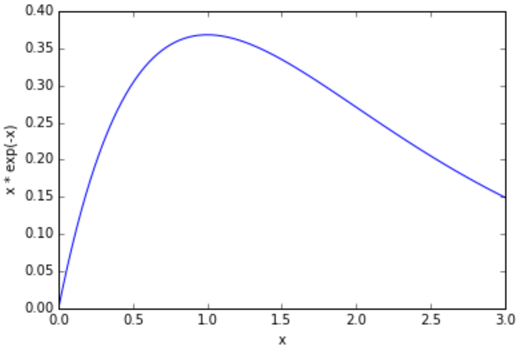
We already have that the radius above is bounded by $\frac{nr^2}{2}$. However, the cost $\mathfrak{C} = \min_{P_{\hat{\theta}|\theta}:\mathbb{E}\ell(\theta,\hat{\theta})\leq D} I(\theta; \hat{\theta})$ is much harder to calculate. We will therefore use a trick called the Shannon lower bound to bound $\mathfrak{C}$. The Shannon lower bound says that the cost given a non-Gaussian prior is not too far away from the cost given a Gaussian prior, provided that the prior is fairly Gaussian-like:

$$\mathfrak{C} \geq \mathfrak{C}\mid_{\theta\sim\text{Gaussian}} - D(\text{unif}(-r, r)\|\mathcal{N}(0, r^2/3))$$

Note: The quantity $\frac{r^2}{3}$ above is the variance of the uniform distribution.

We have (from last lecture) that the cost given a Gaussian prior is $\frac{1}{2}\log\frac{r^2/3}{D}$. Furthermore, we have that $D(\text{unif}(-r, r)\|\mathcal{N}(0, \frac{r^2}{3})) = D(\text{unif}(-1, 1)\|\mathcal{N}(0, \frac{1}{3})) = c_1$ is a constant that does not depend on r. Therefore, for some other constant c:

$$\begin{aligned}\mathfrak{C} &\geq \frac{1}{2}\log\frac{r^2/3}{D} - c_1\\ &= \frac{1}{2}\log\frac{r^2 c}{D}\end{aligned}$$



To complete the lower bound, remember that $\frac{1}{2}\log\frac{r^2c_2}{D} \le \mathfrak{C} \le \frac{nr^2}{2}$, so:

$$\begin{aligned} R^* \geq R^*_\pi &\geq cr^2\exp(-nr^2), \forall r \in [0,\rho] \\ &\geq \sup_{r\in[0,\rho]} cr^2\exp(-nr^2) \\ &\asymp \frac{1}{n} \wedge \rho^2 \end{aligned}$$

To justify the last step, do a change of variables $x = nr^2$, so the expression becomes $\frac{1}{n}\sup_{0\le x\le n\rho^2} x\exp(-x)$. If we examine the function $x\exp(-x)$, we see that it achieves a global maximum of $\frac{1}{e}$ at $x = 1$. Therefore, if $x < 1$ we should choose $x\exp(-x)$, and if $x \geq 1$ we should choose $\frac{1}{e}$. This gives us:

$$\frac{1}{n}\sup_{0\le x\le n\rho^2} x\exp(-x) = \frac{1}{n}(n\rho^2 e^{-n\rho^2} \wedge \frac{1}{e})$$

□

Recap: In order to get the upper bound on the minimax risk, we used the radius, which can be thought of as the maximum distance between a central estimate and any other point in the space of distributions. The lower bound on the minimax risk came from the Shannon lower bound, which is based on how different the selected prior distribution is from a Gaussian distribution.

To extend the lower bound to an arbitrary dimension p, start with a uniform prior over a ball of radius r, calculate its variance, and use the Shannon lower bound again.

In the last class, we learned minimax risk bounding technique by data processing inequality of mutual information such that for $\theta - X - \hat{\theta}$,

$$\inf_{P_{\hat{\theta}|\theta}:\mathbb{E}[\ell(\theta,\hat{\theta})]\leq R_\pi^*} I(\theta;\hat{\theta}) \leq I(\theta;\hat{\theta}) \leq I(\theta;X) \leq \text{capacity} = \sup_{P_\theta} I(\theta;X). \tag{13.1}$$

Because the exact characterization of the LHS is intractable in most cases, we need an appropriate technique that further lower bounds the LHS, which is called the Shannon lower bound. Another technique to get a minimax lower bound, called Fano's method, will be discussed as well.

13.1 Shannon lower bound

13.1.1 Shannon lower bound

Suppose that the loss function is rth power of an arbitrary norm over $\mathbb{R}^p$, i.e., $\ell(\theta,\hat{\theta}) = \|\theta-\hat{\theta}\|^r$, and let $R_\pi^* = D$. Then, the LHS can be written as

$$\begin{aligned}
\inf_{P_{\hat{\theta}|\theta}:\mathbb{E}[\ell(\theta,\hat{\theta})]\leq D} I(\theta;\hat{\theta}) &= \inf_{P_{\hat{\theta}|\theta}:\mathbb{E}[\|\theta-\hat{\theta}\|^r\leq D} I(\theta;\hat{\theta}) \\
&= \inf_{P_{\hat{\theta}|\theta}:\mathbb{E}[\|\theta-\hat{\theta}\|^r\leq D} h(\theta) - h(\theta|\hat{\theta}) \\
&= \inf_{P_{\hat{\theta}|\theta}:\mathbb{E}[\|\theta-\hat{\theta}\|^r\leq D} h(\theta) - h(\theta-\hat{\theta}|\hat{\theta}) \\
&\geq \inf_{P_{\hat{\theta}|\theta}:\mathbb{E}[\|\theta-\hat{\theta}\|^r\leq D} h(\theta) - h(\theta-\hat{\theta}) \\
&= h(\theta) - \sup_{\mathbb{E}\|W\|^r\leq D} h(W) \triangleq \text{SLB}.
\end{aligned}$$

where $W \triangleq \theta - \hat{\theta}$ and the very last quantity is called the Shannon lower bound. To evaluate the supremum term, any convex optimization technique such as Lagrange multiplier can be applied.

A special case of the lower bound for Euclidean norm is given by

$$\text{SLB} = h(\theta) - \sup_{\mathbb{E}\|W\|_2^2\leq D} h(W) = h(\theta) - h\left(\mathcal{N}\left(0, \frac{D}{p}I_p\right)\right) = h(\theta) - \frac{p}{2}\log\left(2\pi e\frac{D}{p}\right),$$

where we used the fact that Gaussian maximizes differential entropy when the second moment is bounded.

Theorem 13.1 (Shannon's Lower Bound). *Let $\|\cdot\|$ be an arbitrary norm on $\mathbb{R}^p$ and $r > 0$. Then*

$$\inf_{P_{\hat{\theta}|\theta}:\mathbb{E}\|\theta-\hat{\theta}\|^r\leq D} I(\theta;\hat{\theta}) \geq h(\theta) - \log\left\{V_p \cdot \left(\frac{Dre}{p}\right)^{\frac{p}{r}} \cdot \Gamma\left(1+\frac{p}{r}\right)\right\},$$

where V_p is the volume of the unit radius ball, i.e.,

$$V_p \triangleq vol(B_{\|\cdot\|}) = vol(\{x \in \mathbb{R}^p : \|x\| \le 1\}).$$

The proof will be given in homework.

Note: The Shannon lower bound is asymptotically tight as $D \to 0$.

Example 13.1 (GLM)**.** Consider the p-dimensional n-sample GLM, i.e., $(X_1, \cdots, X_n) \overset{iid}{\sim} \mathcal{N}(\theta, I_p)$ or equivalently $\bar{X} \sim \mathcal{N}\left(\theta, \frac{1}{n} I_p\right)$. Then the minimax risk with respect to $\|\cdot\|^r$ is

$$R^* \gtrsim \frac{1}{(cn)^{r/2}} V_p^{-r/p}.$$

Proof. Take a prior $\theta \sim \pi = \mathcal{N}(0, sI_p)$. Then the inequality chain (13.1) is rewritten as

$$\begin{aligned}\frac{p}{2}\log(1+ns) &\ge I(\theta, X) \ge I(\theta; \hat{\theta}) \ge \inf_{P_{\hat{\theta}|\theta}: \mathbb{E}[\ell(\theta,\hat{\theta})] \le R_\pi^*} I(\theta;\hat{\theta}) \\ &\ge \text{SLB} = \frac{p}{2}\log(2\pi e s) - \log\left\{V_p \cdot \left(\frac{R_\pi^* r e}{p}\right)^{\frac{p}{r}} \cdot \Gamma\left(1+\frac{p}{r}\right)\right\}.\end{aligned}$$

Then, rearranging terms, taking limit $s \to \infty$, and using the Stirling's formula we get

$$R_\pi^* \gtrsim \frac{1}{(cn)^{r/2}} V_p^{-r/p} \Rightarrow R^* \gtrsim \frac{1}{(cn)^{r/2}} V_p^{-r/p}. \tag{13.2}$$

□

Note that for $r = 2$,

$$R^* \gtrsim \frac{1}{n} V_p^{-2/p},$$

while the exact bound (see Sec. 3.2) is $R^* = \frac{\mathbb{E}\|Z\|^2}{n} = \frac{p}{n}$. In the next example, we will see volumes for ℓ_q norm.

Example 13.2 (ℓ_q-norm)**.** Consider ℓ_q-norm, i.e., for $1 \le q \le \infty$

$$\|x\|_q = \left(\sum_{i=1}^p |x_i|^q\right)^{1/q}.$$

See the volume for several q's.

- ($q = 2$) (Cont'd from the previous) Note that $R^* = \frac{p}{n}$ for the quadratic loss $\|\cdot\|_2^2$. The n-dimensional volume of a unit Euclidean ball B_2 is given by

  $$V_p(B_2)^{1/p} = \frac{\pi^{1/2}}{\left(\Gamma\left(1+\frac{p}{2}\right)\right)^{1/p}} \asymp \frac{1}{\sqrt{p}},$$

  which follows from the Stirling's approximation,

  $$\left(\Gamma\left(1+\frac{p}{2}\right)\right)^{1/p} \asymp \left(\left(\frac{p}{2e}\right)^{p/2}\left(\frac{p}{2}\right)^{1/2}\right)^{1/p} \asymp \left(\frac{p}{2e}\right)^{1/2}\left(\frac{p}{2}\right)^{1/2p} \asymp \sqrt{p}.$$

  Plugging in (13.2) with $r = 2$,

  $$R^* \gtrsim \frac{1}{n} V_p^{-1/2} = \frac{p}{n}.$$

  Hence in this case the SLB is tight.

- ($1 \le q < \infty$) Consider ℓ_q norm, where $1 \le q < \infty$, the volume of a unit ℓ_q ball is given by

$$V_p(B_q) = \frac{\left[2\Gamma\left(1+\frac{1}{q}\right)\right]^p}{\Gamma\left(1+\frac{p}{q}\right)}.$$

  So using (13.2) and the Stirling's formula, the minimax bound for a loss function $\|\cdot\|_q^2$ is given by

$$R^* \gtrsim \frac{p^{2/q}}{n}.$$

  Another way to get the same bound is that

$$R^* \gtrsim \frac{1}{n}\mathbb{E}\|Z\|_q^2 \asymp \frac{p^{2/q}}{n}.$$

  Here the property that if $Z \sim \mathcal{N}(0, I_p)$, $\|Z\|_q^q = \Theta_P(p)$ is used.

- ($q = \infty$) Recall a unit hypercube in $\mathbb{R}^p$. Then, $V_p(B_\infty) = 2^p$, hence, $R^* \gtrsim \frac{1}{n}$ by the SLB. On the other hand, we know the exact risk,

$$R^* = \frac{1}{n}\mathbb{E}\|Z\|_\infty^2 \asymp \frac{\log p}{n}.$$

  So in this case the SLB is not tight. Here, the equality follows from the fact that if $Z \sim \mathcal{N}(0, I_p)$, $\|Z\|_\infty = \Theta_P(\sqrt{\log p})$.

Note: In the case that we have restriction on θ such that $\theta \in \Theta \subset \mathbb{R}^p$, where Θ is a convex set with non-empty interior, the only thing to be changed is the SLB part. Upper bound by capacity remains unchanged. As an example of uniform prior over some $\Theta \subset \mathbb{R}^p$,

$$\text{capacity} \ge \text{SLB} = h(\theta) - \log[\cdots R_\pi^* \cdots] = \log vol(\Theta) - \log[\cdots R_\pi^* \cdots].$$

We get the bound of minimax risk connecting this SLB with capacity formula.

Also note that the exact characterization of $R^*(\Theta)$ is open even for a convex set Θ.

13.1.2 Gaussian width of a convex body K

Suppose $Z \sim N(0, I_p)$ and a set K is convex and symmetric. Define the *Gaussian width* of K

$$w(K) \triangleq \mathbb{E}\left[\sup_{x\in K}\langle x, Z\rangle\right].$$

Lemma 13.1 (Urysohn)**.**

$$\text{vol}(K)^{1/p} \lesssim \frac{w(K)}{p}.$$

Urysohn's lemma helps us characterize the bound of minimax risk. In our case, $K = B_{\|\cdot\|}$, then

$$w(K) = \mathbb{E}\left[\sup_{x\in K}\langle x, Z\rangle\right] = \mathbb{E}\left[\sup_{\|x\|\le 1}\langle x, Z\rangle\right] = \mathbb{E}\|Z\|_*,$$

which is in fact the expected *dual norm* of Z. From the lemma, we have $V_p^{1/p} \lesssim \frac{\mathbb{E}\|Z\|_*}{p}$. Therefore,

$$R^* \gtrsim \frac{1}{n}V_p^{-2/p} \gtrsim \frac{1}{n}\left(\frac{p}{\mathbb{E}\|Z\|_*}\right)^2.$$

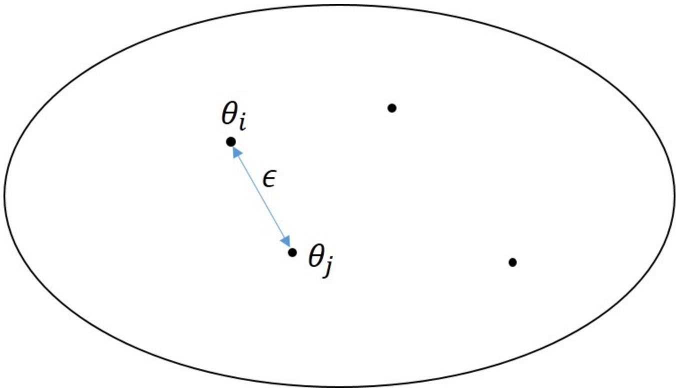


Figure 13.1: Discretization

13.2 Fano's method

Recall the inequality chain,

$$\inf_{P_{\hat{\theta}|\theta}:\mathbb{E}\|\hat{\theta}-\theta\|\le R^*_\pi} I(\theta;\hat{\theta}) \le I(\theta;\hat{\theta}) \le I(\theta;X) \le \text{capacity}.$$

In this section, we discuss Fano's method that reduces the LHS to multiple hypothesis testing problem, which is easier to compute.

The steps are followings:

1. (Discretize) Instead of Θ, consider a discrete subset $\tilde{\Theta} = \{\theta_1, \cdots, \theta_n\} \subset \Theta$. Points are picked to satisfy $\|\theta_i - \theta_j\| \ge \epsilon$ for all $i \ne j$. Figure 13.1 visualizes this discretization.

2. (Reduce to multiple hypothesis testing) Assume uniform prior such that $\theta \sim \pi = \text{unif}(\{\theta_1, \cdots, \theta_n\})$ and let f be a quantizer that maps $\theta \in \Theta$ to $\theta_i \in \tilde{\Theta}$, the closest point to θ. Note that $f(\theta) = \theta$ because θ is drawn over $\tilde{\Theta}$. So by data processing inequality for $\theta - X - \hat{\theta} - f(\hat{\theta})$,

$$I(\theta;\hat{\theta}) \ge I(\theta; f(\hat{\theta})).$$

Note that $I(\theta; f(\hat{\theta}))$ is a function of joint probability mass over discrete space $\tilde{\Theta} \times \tilde{\Theta}$.

Let's see the error events $\{\theta \ne f(\hat{\theta})\}$. Let say the true source is $\theta = \theta_k$. If error happens, it implies our estimate $\hat{\theta}$ closer to $\theta_j = f(\hat{\theta})$ than θ_k for some j. In other words, if error happens,

$$\|\hat{\theta} - f(\hat{\theta})\| \le \|\hat{\theta} - \theta_k\|.$$

So due to triangular inequality, the error event implies

$$\begin{aligned} \epsilon \le \|f(\hat{\theta}) - \theta_k\| = \|f(\hat{\theta}) - \hat{\theta} + \hat{\theta} - \theta_k\| &\le \|f(\hat{\theta}) - \hat{\theta}\| + \|\hat{\theta} - \theta_k\| \\ &\le 2\|\hat{\theta} - \theta_k\|, \\ \Rightarrow \quad \frac{\epsilon}{2} &\le \|\hat{\theta} - \theta_k\|. \end{aligned}$$

Hence,

$$P_e \triangleq \Pr(\theta_k \ne f(\hat{\theta})) \le \Pr\left(\|\hat{\theta} - \theta_k\| \ge \frac{\epsilon}{2}\right) \le \frac{\mathbb{E}\|\hat{\theta} - \theta_k\|}{\epsilon/2} \le \frac{R^*_\pi}{\epsilon/2} = \frac{2R^*_\pi}{\epsilon}$$

$$\Rightarrow \inf_{P_{\hat{\theta}|\theta}:\mathbb{E}\|\hat{\theta}-\theta\|\le R^*_\pi} I(\theta;\hat{\theta}) \ge \inf_{P_e \le \frac{2R^*_\pi}{\epsilon}} I(\theta;\hat{\theta}) \ge \inf_{P_e \le \frac{2R^*_\pi}{\epsilon}} I(\theta; f(\hat{\theta})).$$

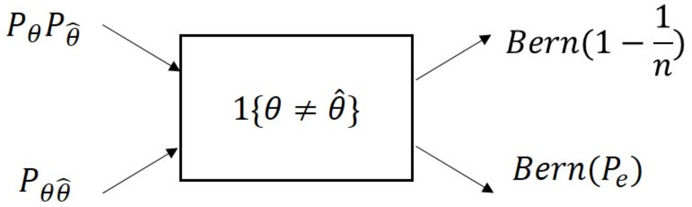


Figure 13.2: Data processing kernel for Fano's inequality

So, we reduce the LHS to a multiple hypothesis test problem where $\theta, f(\hat{\theta})$ are both discrete.

3. (Apply Fano's inequality) Recall the data processing inequality for KL divergence by Figure 13.2. Here our processor is $1\{\theta \neq \hat{\theta}\}$, and we can further lower bound as

$$\begin{aligned}
I(\theta; f(\hat{\theta})) &= D\left(P_{\theta, f(\hat{\theta})} \| P_\theta P_{f(\hat{\theta})}\right) \\
&\geq D\left(\text{Bern}(P_e) \| \text{Bern}\left(1 - \frac{1}{n}\right)\right) \\
&= P_e \log \frac{P_e}{1 - \frac{1}{n}} + (1 - P_e) \log \frac{1 - P_e}{\frac{1}{n}} \\
&= -h(P_e) + \log n - P_e \log(n-1) \\
&\geq -\log 2 + \log n - P_e \log n, \\
\Rightarrow P_e &\geq 1 - \frac{I(\theta; f(\hat{\theta})) + \log 2}{\log n},
\end{aligned}$$

where $h(\cdot)$ is a binary entropy function. So finally we reach the bound

$$\begin{aligned}
\frac{2R^*}{\epsilon} \geq \frac{2R^*_\pi}{\epsilon} &\geq P_e \geq 1 - \frac{I(\theta; f(\hat{\theta})) + \log 2}{\log n} \\
\Rightarrow \quad R^* &\geq \frac{\epsilon}{2}\left(1 - \frac{I(\theta; f(\hat{\theta})) + \log 2}{\log n}\right).
\end{aligned}$$

Note: The situation of the Fano's inequality in class is that

1. θ uniformly takes M values.
2. Markov chain $\theta - X - \hat{\theta}$ holds.

Then, the Fano's inequality says that

$$\begin{aligned}
I(\theta; X) &\geq -\log 2 + \log M - P_e \log(M-1) \\
&\geq -\log 2 + (1 - P_e) \log M, \\
\Rightarrow P_e &\geq 1 - \frac{I(\theta; X) + \log 2}{\log M}.
\end{aligned}$$

The Fano's inequality intuitively means that when the mutual information is fixed, P_e cannot be less than a certain value. On the other hand, when P_e is fixed, the mutual information must be greater than a certain value.

Note: We can also use the Fano inequality as following:

$$I(\theta;X) \geq \min_{P_e \leq \frac{2R_\pi^*}{\epsilon}} I(\theta;X),$$

and similarly as above,

$$\begin{aligned} \frac{2R^*}{\epsilon} \geq \frac{2R_\pi^*}{\epsilon} &\geq P_e \geq 1 - \frac{I(\theta;X) + \log 2}{\log n} \\ \Rightarrow \quad R^* &\geq \frac{\epsilon}{2}\left(1 - \frac{I(\theta;X) + \log 2}{\log M}\right). \end{aligned}$$

Note: If the loss function is $\|\cdot\|^2$,

$$\min_{P_e \leq \frac{2R_\pi^*}{\epsilon}} I(\theta;X) \quad \Rightarrow \quad R^* \geq \left(\frac{\epsilon}{2}\right)^2 \left(1 - \frac{I(\theta;X) + \log 2}{\log M}\right).$$

§ 14. Packing, covering, and consequences on minimax risk

Last lecture, we lower bounded $\min_{\|\theta-\hat{\theta}\|} I(\theta;\hat{\theta})$ using Shannon lower bound, and we saw that for the p dimensional n sample GLM,

$$R^*(\mathbb{R}^p) \gtrsim \frac{1}{n\text{vol}^{\frac{2}{p}}(B_{\|\cdot\|})}$$

with respect to a loss $\ell(\theta,\hat{\theta}) = \|\theta-\hat{\theta}\|^2$ and an arbitrary norm $\|\cdot\|$.

To understand why some sort of volume shows up, we further extend the lower bound obtained using Fano's method. We first introduce the concept of **packing**, **covering**, relate them to the notion of volume, and then plug them into the lower bound obtained using the Fano's inequality. When applied to GLM, this alternative method gives the same dependence on the dimension and the sample size for ℓ_q norms with $q < \infty$.

14.1 Covering and Packing

Let $(V, \|\cdot\|)$ be a normed space and $\Theta \subset V$.

Definition 14.1 (ϵ-covering)**.** We say $\{V_1, ..., V_N\}$ is an ϵ-covering of Θ if $\Theta \subset \cup_{i=1}^N B(V_i, \epsilon)$, or equivalently, $\forall \theta \in \Theta$, $\exists i$ such that $\|\theta - V_i\| \leq \epsilon$.

Definition 14.2 (ϵ-packing)**.** We say $\{\theta_1, ..., \theta_M\} \subset \Theta$ is an ϵ-packing of Θ if $\min_{i\neq j} \|\theta_i - \theta_j\| > \epsilon$,[1] or equivalently, the balls $\{B(\theta_i, \epsilon/2) : j \in [M]\}$ are disjoint.

Upon defining ϵ-covering and ϵ-packing, one naturally asks what is the minimal number of ϵ-balls one needs in order to cover Θ, and what is the maximal number of $\epsilon/2$-balls one can pack in Θ. Those numbers are defined as covering and packing numbers, which, similar to volume and width, measures the "massiveness" of a set.

Definition 14.3 (Covering number)**.** $N(\Theta, \|\cdot\|, \epsilon) \triangleq \min\{n : \exists \epsilon\text{-covering over } \Theta \text{ of size } n\}$.

Definition 14.4 (Packing number)**.** $M(\Theta, \|\cdot\|, \epsilon) \triangleq \max\{m : \exists \epsilon\text{-packing of } \Theta \text{ of size } m\}$.

Remark 14.1. Some basic remarks.

- $M(\Theta, \|\cdot\|, \epsilon)$ and $N(\Theta, \|\cdot\|, \epsilon)$ are often abbreviated as $M(\epsilon)$, $N(\epsilon)$.
- For ϵ-covering, the balls need not be disjoint.
- $N(\Theta, \|\cdot\|, \epsilon)$ is a decreasing function of ϵ when the norm and Θ are fixed. That is, if $\epsilon_0 < \epsilon_1$, and $\{V_1, ..., V_N\}$ is an ϵ-covering of Θ, then $\Theta \subset \cup_{i=1}^N B(V_i, \epsilon_0) \subset \cup_{i=1}^N B(V_i, \epsilon_1)$.

[1] Notice we imposed strict inequality for convenience.

- Metric entropy: $\log M(\epsilon)$ and $\log N(\epsilon)$.
- $N(\epsilon) < \infty \ \forall \epsilon > 0 \Leftrightarrow \Theta$ is totally bounded (In topology, a metric space is said to be totally bounded if for every $\epsilon > 0$ there is a finite covering of the space by ϵ-balls). For example, a metric space is compact iff it is complete and totally bounded. Hence a compact metric space is totally bounded.

Remark 14.2. Note that in Definition 14.1 we do not require the ϵ-covering to be a subset of Θ. It turns out imposing this restriction does not change the behavior of the covering number that much. Similar to Definition 14.3, denote by $N'(\Theta, \|\cdot\|, \epsilon)$ the smallest ϵ-covering of Θ that is included in Θ. Then we have

$$N(\Theta, \|\cdot\|, \epsilon) \leq N'(\Theta, \|\cdot\|, \epsilon) \leq N(\Theta, \|\cdot\|, \epsilon/2) \tag{14.1}$$

The left inequality is obvious. To see the right inequality,[2] let $\{\theta_1, \ldots, \theta_N\}$ be an $\frac{\epsilon}{2}$-covering of Θ. For each i, let $\theta_i' = \arg\min_{u\in\Theta} \|\theta_i - u\|$. Then $\{\theta_1', \ldots, \theta_N'\} \subset \Theta$ constitutes an ϵ-covering. Indeed, for any $\theta \in \Theta$, we have $\|\theta - \theta_i\| \leq \epsilon/2$ for some θ_i. Then $\|\theta - \theta_i'\| \leq \|\theta - \theta_i\| + \|\theta_i - \theta_i'\| \leq 2\|\theta - \theta_i\| \leq \epsilon$.

The relation between the packing number and the covering number is described in the following theorem.

Theorem 14.1 (Kolomogrov-Tikhomirov)**.**

$$M(\Theta, \|\cdot\|, 2\epsilon) \leq N(\Theta, \|\cdot\|, \epsilon) \leq M(\Theta, \|\cdot\|, \epsilon). \tag{14.2}$$

Proof. First prove the right inequality. Suppose $E = \{\theta_1, ..., \theta_M\}$ is a maximal packing. Then $\forall \theta \in \Theta \backslash E$, $\exists i$ such that $\|\theta - \theta_i\| \leq \epsilon$ (if this does not hold for θ then we can construct a bigger packing with $\theta_{M+1} = \theta$). Hence E is automatically an ϵ-covering (which is also a subset of Θ). Since $N(\Theta, \|\cdot\|, \epsilon)$ is the minimal size of all possible coverings, we have $M(\Theta, \|\cdot\|, \epsilon) \geq N(\Theta, \|\cdot\|, \epsilon)$.

We next prove the left inequality by contradiction. Suppose there exists a 2ϵ-packing $\{\theta_1, ..., \theta_M\}$ and an ϵ-covering $\{x_1, ..., x_N\}$ such that $M \geq N + 1$. Then by the pigeonhole principle, there exist distinct θ_i and θ_j belonging to the same ϵ-ball $B(x_k, \epsilon)$. This means that the distance between θ_i and θ_j cannot be more than the diameter of the ball, i.e., $\|\theta_i - \theta_j\| \leq 2\epsilon$, which leads to a contradiction since $\|\theta_i - \theta_j\| > 2\epsilon$ for a 2ϵ-packing. Hence the size of any 2ϵ-packing is less or equal to the size of any ϵ-covering. Hence $M(\Theta, \|\cdot\|, 2\epsilon)$, the maximal size of a 2ϵ-packing is at most $N(\Theta, \|\cdot\|, \epsilon)$, the minimal size of an ϵ-covering. □

When V is the d-dimensional Euclidean space, we can extend the previous theorem by further lower/upper bounding the covering/packing numbers. The result is given as follows.

Theorem 14.2. *Let $\Theta \subset V = \mathbb{R}^d$ and let $\|\cdot\|$ be an arbitrary norm. Then*

$$\left(\frac{1}{\epsilon}\right)^d \frac{\text{vol}(\Theta)}{\text{vol}(B)} \overset{(a)}{\leq} N(\Theta, \|\cdot\|, \epsilon) \leq M(\Theta, \|\cdot\|, \epsilon) \overset{(b)}{\leq} \frac{\text{vol}(\Theta + \frac{\epsilon}{2}B)}{\text{vol}(\frac{\epsilon}{2}B)} \underset{\substack{\Theta \text{ convex} \\ \epsilon B \subset \Theta}}{\overset{(c)}{\leq}} \frac{\text{vol}(\frac{3}{2}\Theta)}{\text{vol}(\frac{\epsilon}{2}B)} = \left(\frac{3}{\epsilon}\right)^d \frac{\text{vol}(\Theta)}{\text{vol}(B)}.$$

where B denotes the unit norm ball and $+$ denotes the Minkowski sum, i.e., $A + B = \{a + b : a \in A, b \in B\}$.

[2] Another way to see this is from Theorem 14.1: note that (b) in (14.2) yields a ϵ-covering that is included in Θ. Together with (a), we get $N'(\epsilon) \leq M(\epsilon) \leq N(\epsilon/2)$.

Proof. First prove (a). For a covering of minimal size, $\Theta \subset \cup_{i=1}^{n} B(X_i, \epsilon)$. Hence

$$\mathrm{vol}(\Theta) \leq \mathrm{vol}(\cup_{i=1}^{N(\epsilon)} B(X_i, \epsilon)) \leq \sum_{i=1}^{N(\epsilon)} \mathrm{vol}(B(X_i, \epsilon)).$$

Since $\mathrm{vol}(B(X_i, \epsilon)) = \epsilon^d \mathrm{vol}(B)$, we have $\mathrm{vol}(\Theta) \leq N(\epsilon)\epsilon^d \mathrm{vol}(B)$. Hence (a) is proved.

Next we prove (b). For an ϵ-packing, the balls $B(\theta_i, \epsilon/2)$ are disjoint, and $\cup_{i=1}^{M(\epsilon)} B(\theta_i, \epsilon/2) \subset \Theta + \frac{\epsilon}{2}B$. Taking the volume on both sides, we have

$$\mathrm{vol}(\Theta + \frac{\epsilon}{2}B) \geq \mathrm{vol}(\cup_{i=1}^{M(\epsilon)} B(\theta_i, \epsilon/2)) = M(\epsilon)\mathrm{vol}(\frac{\epsilon}{2}B).$$

This proves (b).

To prove (c), we prove two statements. (1) When $\epsilon B \subset \Theta$, $\Theta + \frac{\epsilon}{2}B \subset \Theta + \frac{1}{2}\Theta$, and (2) when Θ is convex, $\Theta + \frac{1}{2}\Theta = \frac{3}{2}\Theta$.

To prove (1), notice for any $z \in \Theta + \frac{\epsilon}{2}B$, we have $z = x + y$ where $x \in \frac{\epsilon}{2}B$ and $y \in \Theta$. Since $x \in \frac{\epsilon}{2}B \Rightarrow x \in \Theta$, we immediately have $z \in \Theta + \frac{1}{2}\Theta$.

To prove (2), first notice that $\forall \theta \in \frac{3}{2}\Theta$, $\theta = \frac{1}{3}\theta + \frac{2}{3}\theta$. Since $\frac{1}{3}\theta \in \frac{1}{2}\Theta$, and $\frac{2}{3}\theta \in \Theta$, $\frac{3}{2}\Theta \subseteq \Theta + \frac{1}{2}\Theta$. On the other hand, for any $x \in \Theta + \frac{1}{2}\Theta$, we have $x = y + \frac{1}{2}z$ with $y, z \in \Theta$. When Θ is convex, $\frac{2}{3}x = \frac{2}{3}y + \frac{1}{3}z \in \Theta$. Hence $x \in \frac{3}{2}\Theta$, implying $\Theta + \frac{1}{2}\Theta \subseteq \frac{3}{2}\Theta$.

With (1) and (2), (c) follows immediately. □

Remark 14.3. Why is Theorem 14.1 cool?

- (a) is a converse, saying that the minimal covering size cannot be too small. When combined with $N(\epsilon) \leq M(\epsilon)$, this turns into an existential statement: It is possible to construct a packing of size at least $\mathrm{vol}(\Theta)/\mathrm{vol}(B(\epsilon))$. From the proof we see that this corresponds to a *greedy* construction. Furthermore, for Hamming space and Hamming distance, this is exactly the *Gilbert-Varshanov* bound.
- (b) is a converse, saying that the maximal packing size cannot be too large. When combined with $N(\epsilon) \leq M(\epsilon)$, this turns into an existence statement: there exists a small covering.

Example 14.1 (Metric entropy of norm balls)**.** Let $\|\cdot\|$ be an arbitrary norm on $\mathbb{R}^d$ and let $B = B_{\|\cdot\|} = \{x \in \mathbb{R}^d : \|x\| \leq 1\}$ be the corresponding unit norm ball. Consider the covering number of $B_{\|\cdot\|}$ with respect to the same norm, namely, $N(\epsilon) = N(B_{\|\cdot\|}, \|\cdot\|, \epsilon)$. When $\epsilon \geq 1$, $N(\epsilon) = 1$. When $\epsilon < 1$, applying Theorem 14.2 we have

$$\left(\frac{1}{\epsilon}\right)^d = \frac{\mathrm{vol}(\Theta)}{\mathrm{vol}(B_2)} \leq N(\epsilon) \leq M(\epsilon) \leq \frac{\mathrm{vol}((1+\frac{\epsilon}{2})B)}{\mathrm{vol}(\frac{\epsilon}{2}B)} = \left(1 + \frac{2}{\epsilon}\right)^d \leq \left(\frac{3}{\epsilon}\right)^d.$$

Hence $d \log \frac{1}{\epsilon} \leq \log N(\epsilon) \leq d \log \frac{3}{\epsilon}$. This relationship holds for all norms as long as the covering is done with respect to the same norm.

If we fix the dimension d and let $\epsilon \to 0$, then because all norms on Euclidean space are equivalent (with constant factors of each other where the constants depend on dimension), whenever Θ has non-empty interior, $\log N(\Theta, \|\cdot\|, \epsilon) = (d + o(1)) \log \frac{1}{\epsilon}$. In particular, for a different norm $\|\cdot\|'$, $N(B_{\|\cdot\|}, \|\cdot\|', \epsilon)$ is still approximately $d \log \frac{1}{\epsilon}$ when ϵ is sufficiently small. But how small is small enough depends on the dimension d and there are some interesting high-dimensional phenomena when the covering number is not determined by volumetric methods. See Lecture 15 for details.

14.2 Applying metric entropy & Fano's inequality to minimax risk

We now apply metric entropy and Fano's inequality to lower bound the minimax risk. The key idea is to reduce estimation over Θ to testing between a packing $E = \{\theta_1, ..., \theta_M\}$ within $T \subset \Theta$. Then $R^*(\Theta) \geq R^*(T) \geq R^*_\pi$ where π is equi-probable over E.

Let $E = \{\theta_1, ..., \theta_M\}$ be an ϵ-packing on $T \subset \Theta$. Let $\tilde{\theta}$ be the quantized version of $\hat{\theta}$ restricted to E. Then we have the Markov chain $\theta \to X \to \hat{\theta} \to \tilde{\theta}$. Consider the quadratic loss function $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^2$. Recall that

$$\mathrm{rad}_{\mathrm{KL}}(T) = \inf_Q \sup_{\theta \in T} D(P_\theta \| Q),$$

and

$$\mathrm{diam}_{\mathrm{KL}}(T) = \sup_{\theta, \theta' \in T} D(P_{\theta'} \| P_\theta).$$

We immediately have

$$\begin{aligned}
\mathbb{E}\left[\|\theta - \hat{\theta}\|^2\right] &\overset{Markov}{\geq} \left(\frac{\epsilon}{2}\right)^2 \mathbb{P}\left[\|\theta - \hat{\theta}\| \geq \frac{\epsilon}{2}\right] \geq \left(\frac{\epsilon}{2}\right)^2 \mathbb{P}\left[\theta \neq \tilde{\theta}\right] \\
&\overset{Fano}{\geq} \left(\frac{\epsilon}{2}\right)\left(1 - \frac{I(\theta; X) + \log 2}{\log M(\epsilon)}\right) \\
&\geq \frac{\epsilon^2}{4}\left(1 - \frac{\mathrm{rad}_{\mathrm{KL}}(T) + \log 2}{\log M(\epsilon)}\right) \\
&\geq \sup_{T \subset \Theta, \epsilon > 0} \frac{\epsilon^2}{4}\left(1 - \frac{\mathrm{diam}_{\mathrm{KL}}(T) + \log 2}{\log \frac{\mathrm{vol}(T)}{\mathrm{vol}(\epsilon B)}}\right),
\end{aligned} \tag{14.3}$$

where in the last step, the inequality holds true for all choices of T and ϵ, and the supremum is placed to obtain a better bound.

For GLM, we can use the above method (Fano+packing) to obtain the same result (up to constant factor) by Shannon Lower Bound.

Example 14.2 (p-dimensional n-sample GLM)**.** Let $\Theta = \mathbb{R}^p$, and $T = B_2(s)$. Then $\mathrm{diam}_{\mathrm{KL}}(T) = \sup_{\theta, \theta' \in T} D(P_\theta \| P_{\theta'}) = \sup_{\theta, \theta' \in T} D(N(\theta, I_p)^{\otimes n} \| N(\theta', I_p)^{\otimes n}) = \sup_{\theta, \theta'} \frac{n}{2}\|\theta - \theta'\|^2 = \frac{n}{2}\mathrm{diam}^2(T) = \frac{n}{2}s^2$. By (14.3), we have

$$R^* \geq \frac{\epsilon^2}{4}\left(1 - \frac{\mathrm{diam}_{\mathrm{KL}}(T) + \log 2}{\log \frac{\mathrm{vol}(T)}{\mathrm{vol}(\epsilon B)}}\right) = \frac{\epsilon^2}{4}\left(1 - \frac{\frac{n}{2}s^2 + \log 2}{\log \frac{s^p \mathrm{vol}(B_2)}{\epsilon^p \mathrm{vol}(B_{\|\cdot\|})}}\right).$$

We now choose ϵ and s. Denote $\mathrm{vol}(B_{\|\cdot\|}) = V$, and recall that $\mathrm{vol}^{1/p}(B_2) \asymp \frac{1}{\sqrt{p}}$, i.e., $c_1 \frac{1}{\sqrt{p}} < \mathrm{vol}^{1/p}(B_2) < c_2 \frac{1}{\sqrt{p}}$. If we choose

$$s = c_3 \sqrt{\frac{p}{n}}, \ \epsilon = c_4 \frac{1}{\sqrt{n} V^{1/p}},$$

then

$$R^* \geq \frac{c_4^2}{4nV^{2/p}}\left(1 - \frac{\frac{c_3^2 p}{2} + \log 2}{p \log \frac{c_1 c_3}{c_4}}\right) \geq \frac{c_4^2}{4nV^{2/p}}\left(1 - \frac{\frac{c_3^2}{2} + \log 2}{\log \frac{c_1 c_3}{c_4}}\right).$$

As long as we choose c_1, c_2, c_3, c_4 such that $(\frac{c_3^2}{2} + \log 2)/\log \frac{c_1 c_3}{c_4} < c < 1$, we have

$$R^* \geq \frac{c_4^2(1-c)}{4nV^{2/p}} \gtrsim \frac{1}{nV^{2/p}}. \tag{14.4}$$

Remark 14.4. When the specified norm is $\|\cdot\|_\infty$, the norm ball becomes a cube, and the volume is (for fixed values of p)

$$V = 2^p.$$

Hence $R^* \gtrsim \frac{1}{n}$; however, we know $R^* \asymp \frac{\log p}{n}$ and we lose the dependence on the dimension p.

So what should be blamed? It turns out our mutual information method and, in fact, its further relaxation via packing plus Fano's inequality are both tight in this case. What is loose is the volume ratio bound on packing number in Theorem 14.2. In the next lecture, we will prove

$$\log N(B, \|\cdot\|, \epsilon) \asymp \begin{cases} p \log \frac{1}{\epsilon\sqrt{p}}, & \epsilon \lesssim \frac{1}{\sqrt{p}} \\ \frac{1}{\epsilon^2} \log(p\epsilon^2), & \epsilon \gtrsim \frac{1}{\sqrt{p}} \end{cases}.$$

This will lead to the tight result $R^* \asymp \frac{\log p}{n}$.

In this lecture we study the upper and lower bounds on $M(B_1, \|\cdot\|_2, \epsilon)$.

From the last lecture Theorem 14.2, we know that for any $\Theta \subset \mathbb{R}^d$ and any $\epsilon > 0$,

$$\frac{\mathrm{vol}(\Theta)}{\mathrm{vol}(\epsilon B)} \le M(\Theta, \|\cdot\|, 2\epsilon) \le N(\Theta, \|\cdot\|, \epsilon) \le M(\Theta, \|\cdot\|, \epsilon) \le \frac{\mathrm{vol}(\Theta + \frac{\epsilon}{2}B)}{\mathrm{vol}(\frac{\epsilon}{2}B)}.$$

where B is the ball of radius 1 measured by $\|\cdot\|$. Therefore,

$$M(B_1, \|\cdot\|_2, \epsilon) \le \frac{\mathrm{vol}(B_1 + \frac{\epsilon}{2}B_2)}{\mathrm{vol}(\frac{\epsilon}{2}B_2)} \le \frac{\mathrm{vol}((1+\frac{\epsilon\sqrt{d}}{2})B_1)}{\mathrm{vol}(\frac{\epsilon}{2}B_2)} = \left(\frac{1+\frac{\epsilon\sqrt{d}}{2}}{\frac{\epsilon}{2}}\right)^d \left(\frac{c_1}{\sqrt{d}}\right)^d \le \left(1 + \frac{c_2}{\epsilon\sqrt{d}}\right)^d,$$

where we have used the fact that $B_2 \subset \sqrt{d}B_1$ by Cauchy-Schwarz inequality, $\mathrm{vol}(B_1) = 2^d/d!$ and hence $\mathrm{vol}(B_1)^{1/d} \asymp \frac{1}{d}$, and $\mathrm{vol}(B_2)^{1/d} \asymp \frac{1}{\sqrt{d}}$. On the other hand,

$$M(B_1, \|\cdot\|_2, \epsilon) \ge \frac{\mathrm{vol}(B_1)}{\mathrm{vol}(\epsilon B_2)} = \left(\frac{1}{\epsilon}\right)^d \frac{\mathrm{vol}(B_1)}{\mathrm{vol}(B_2)} = \left(\frac{c}{\epsilon\sqrt{d}}\right)^d.$$

From last lecture we know that volume bound is tight when ϵ is sufficiently small (i.e., $\epsilon \to 0$). However, in high dimension, *how small is sufficiently small depends on the dimensionality.* The volume bound derived above is useful only when $\epsilon \lesssim \frac{1}{\sqrt{d}}$. It turns out when $\epsilon \gg \frac{1}{\sqrt{d}}$ (which could still be small), we need different methods, and the full picture is the following

Lemma 15.1. *There exist absolute constants c_1, c_2, such that*

$$d\log\left(1 + \frac{c_2}{\epsilon^2 d}\right) \wedge \frac{1}{\epsilon^2}\log\left(1 + c_2\epsilon^2 d\right) \lesssim \log M(B_1, \|\cdot\|_2, \epsilon) \lesssim d\log\left(1 + \frac{c_1}{\epsilon^2 d}\right) \wedge \frac{1}{\epsilon^2}\log\left(1 + c_1\epsilon^2 d\right).$$

That is to say,

$$\log M(B_1, \|\cdot\|_2, \epsilon) \asymp \begin{cases} \frac{1}{\epsilon^2}\log\left(\epsilon^2 d\right) & \epsilon \gtrsim \frac{1}{\sqrt{d}} \\ d & \epsilon \asymp \frac{1}{\sqrt{d}} \\ d\log\frac{1}{\epsilon^2 d} & \epsilon \lesssim \frac{1}{\sqrt{d}} \end{cases}. \tag{15.1}$$

15.1 Upper bound via Sudakov minoration

Recall that the Gaussian width of $\Theta \subset \mathbb{R}^d$ is defined as[1]

$$w(\Theta) = \mathbb{E}\sup_{\theta\in\Theta}\langle\theta, Z\rangle, \qquad \text{where } Z \sim N(0, \mathbf{I}_d).$$

[1] To avoid measurability difficulty, $w(\Theta)$ should be understood as $\sup_{T\subset\Theta, |T|<\infty} \mathbb{E}\max_{\theta\in T}\langle\theta, Z\rangle$.

Theorem 15.1 (Sudakov minoration)**.** *For any $\Theta \subset \mathbb{R}^d$ and any $\epsilon > 0$,*

$$w(\Theta) \gtrsim \epsilon\sqrt{\log M(\Theta, \|\cdot\|_2, \epsilon)}.$$

The proof of Theorem 15.1 relies on Slepian's Gaussian comparison lemma:

Lemma 15.2 (Slepian's lemma)**.** *Let $X = (X_1, \dots, X_n)$ and $Y = (Y_1, \dots, Y_n)$ be Gaussian random vectors. If $\mathbb{E}(Y_i - Y_j)^2 \leq \mathbb{E}(X_i - X_j)^2$ for all i, j, then $\mathbb{E}\max Y_i \leq \mathbb{E}\max X_i$.*

For a self-contained proof see [Cha05].[2] See also [Pis99, Lemma 5.7, p. 70] for a simpler proof of a weaker version $\mathbb{E}\max X_i \leq 2\mathbb{E}\max Y_i$, which suffices for our purposes though.

Proof of Theorem 15.1 assuming Slepian. Let $\{\theta_1, \dots, \theta_M\}$ be an he optimal ϵ-packing of Θ. Let $X_i = \langle\theta_i, Z\rangle$ for $i \in [M]$, where $Z \sim \mathcal{N}(0, I_d)$. Let $Y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \epsilon^2/2)$. Then for any pair i, j, X_i and X_j are jointly Gaussian, and

$$\mathbb{E}(X_i - X_j)^2 = (\theta_i - \theta_j)'\mathbb{E}[ZZ'](\theta_i - \theta_j) = \|\theta_i - \theta_j\|_2^2 \geq \epsilon^2 = \mathbb{E}(Y_i - Y_j)^2.$$

It follows from Lemma 15.2 that

$$\mathbb{E}\max_{1\leq i\leq M} X_i \geq \mathbb{E}\max_{1\leq i\leq M} Y_i \asymp \epsilon\sqrt{\log M}.$$

This completes the proof because $\mathbb{E}\sup_{\theta\in\Theta}\langle\theta, Z\rangle \geq \mathbb{E}\max_{1\leq i\leq M} X_i$. □

We can apply this theorem to $\Theta = B_1$. In this case, by the definition of the dual norm,

$$w(B_1) = \mathbb{E}\sup_{x\in\mathbb{R}^d:\,\|x\|_1\leq 1}\langle x, Z\rangle = \mathbb{E}\|Z\|_\infty \asymp \sqrt{\log d}.$$

The theorem then implies that

$$\log M(B_1, \|\cdot\|_2, \epsilon) \lesssim \frac{\log d}{\epsilon^2}. \tag{15.2}$$

This bound is almost optimal: When $\epsilon \gg 1/\sqrt{d}$, this upper bound is (in fact optimal and) much better than what we get from the volume argument, which is

$$\log M(B_1, \|\cdot\|_2, \epsilon) \lesssim d\log\left(1 + \frac{c}{\epsilon\sqrt{d}}\right).$$

However, (15.2) is not always sharp. For example, when $\epsilon \asymp 1/\sqrt{d}$, it gives $d\log d$ and we know (even from volume bound) that the correct behavior is d. This suggests we need a more refined bound that interpolates between volume and Sudakov.

[2] If you took ECE 534 last fall, you should revisit Problem 4 of `http://maxim.ece.illinois.edu/teaching/fall15a/homework/hw4.pdf` which follows [Cha05].

15.2 Upper bound via Maurey's empirical method

We can construct a covering of B_1 using the probabilistic method. Let $\{e_i : i = 1, \ldots, d\}$ be the standard basis of $\mathbb{R}^d$. For an arbitrary $x \in B_1$, define a d-dimensional random vector Z as

$$Z = \begin{cases} \operatorname{sgn}(x_i)e_i & \text{w.p. } |x_i| \\ 0 & \text{w.p. } 1 - \|x\|_1 \end{cases}$$

Then Z has the property that $\mathbb{E}Z_i = x_i$ for $i = 1, \ldots, d$, hence $\mathbb{E}Z = x$, and $\mathrm{Var}[Z_i] = \mathbb{E}(Z_i - x_i)^2$ for $i = 1, \ldots, d$. Let $Z_{(1)}, \ldots, Z_{(k)}$ be i.i.d. copies of Z, and let $\bar{Z} = \frac{1}{k}\sum_{j=1}^k Z_{(j)}$. Then

$$\mathbb{E}\|\bar{Z} - x\|_2^2 = \sum_{i=1}^d \mathbb{E}(\bar{Z}_i - x_i)^2 = \sum_{i=1}^d \mathrm{Var}[\bar{Z}_i] = \frac{1}{k}\sum_{i=1}^d \mathrm{Var}[Z_i] = \frac{1}{k}\mathbb{E}\|Z - x\|_2^2 \le \frac{1}{k}\mathbb{E}\|Z - x\|_1^2 \le \frac{1}{k},$$

where we have used the facts that $\mathrm{Var}[\bar{Z}_i] = \frac{1}{k}\mathrm{Var}[Z_i]$ and $\|Z - x\|_2 \le \|Z - x\|_1 \le 1$. If we choose $k = 1/\epsilon^2$, then $\mathbb{E}\|\bar{Z} - x\|_2 \le \sqrt{\mathbb{E}\|\bar{Z} - x\|_2^2} \le \epsilon$. So there is a realization $\bar{z}$ of $\bar{Z}$ such that

$$\|\bar{z} - x\|_2 \le \epsilon.$$

Now we examine how many distinct values $\bar{Z}$ can take regardless of x, which gives the size of the packing. Note that

$$\bar{Z} = \frac{1}{k}\sum_{j=1}^k Z_{(j)} = \frac{1}{k}(K_1, \ldots, K_d),$$

where

$$\sum_{i=1}^d K_i \le k, \qquad \text{with } K_i \in \mathbb{Z}, \text{ and } 0 \le |K_i| \le k \text{ for } i = 1, \ldots, d. \tag{15.3}$$

For any $(K_1, \ldots, K_d)$ satisfying inequality (15.3), we get a solution for the following inequality

$$\sum_{i=1}^d K_i^+ + K_i^- \le k, \qquad \text{with } K_i^+, K_i^- \in \mathbb{Z}, \text{ and } 0 \le K_i^+, K_i^- \le k \text{ for } i = 1, \ldots, d, \tag{15.4}$$

by setting $K_i^+ = K_i$ and $K_i^- = 0$ if $K_i \ge 0$, and setting $K_i^+ = 0$ and $K_i^- = -K_i$ if $K_i < 0$. Therefore, the number of values $\bar{Z}$ can take is upper bounded by the number of solutions of inequality (15.4). Note that there are $\binom{k+2d-1}{2d-1}$ solutions for

$$\sum_{i=1}^d K_i^+ + K_i^- = k, \qquad \text{with } K_i^+, K_i^- \in \mathbb{Z}, \text{ and } 0 \le K_i^+, K_i^- \le k \text{ for } i = 1, \ldots, d,$$

because the solutions are all possible types of the sequences of length k with alphabet size $2d$. It follows that the number of solutions of inequality (15.4) is

$$\binom{0 + 2d - 1}{2d - 1} + \binom{1 + 2d - 1}{2d - 1} + \ldots + \binom{k + 2d - 1}{2d - 1} = \binom{k + 2d}{2d} = \binom{k + 2d}{k},$$

which is an upper bound on the number of $\bar{Z}$'s regardless of x. We thus have shown the existence of an ϵ-covering of B_1 in $\|\cdot\|_2$ with cardinality upper bounded by

$$\binom{\frac{1}{\epsilon^2} + 2d}{2d} = \binom{\frac{1}{\epsilon^2} + 2d}{\frac{1}{\epsilon^2}}.$$

Therefore,

$$\log N(B_1, \|\cdot\|_2, \epsilon) \le 2d \log\left(1 + \frac{1}{2\epsilon^2 d}\right) \wedge \frac{1}{\epsilon^2} \log\left(1 + 2d\epsilon^2\right).$$

We can see that the first upper bound recovers the result from the volume argument, while the second upper bound is even stronger than the result obtained from Sudakov's minoration.

15.3 Lower bound via packing Hamming spheres

Let $S_k = \{x \in \{0,1\}^d : w_{\mathrm{H}}(x) = k\}$ be the Hamming sphere of radius k. For the ρ-packing of S_k in Hamming distance $\|\cdot\|_{\mathrm{H}}$, using Theorem 14.2 we have

$$\log M(S_k, \|\cdot\|_{\mathrm{H}}, \rho) \ge \frac{|S_k|}{|B_{\mathrm{H}}(\rho)|} = \frac{\binom{d}{k}}{\sum_{i=0}^{\rho} \binom{d}{i}}.$$

This leads to the following lemma.

Lemma 15.3 (Gilbert-Varshamov)**.** *There exist constants c_1 and c_2 such that for all $d \in \mathbb{N}$ and any $k \in [d]$,*

$$\log M(S_k, \|\cdot\|_{\mathrm{H}}, c_1 k) \ge c_2 k \log \frac{ed}{k}.$$

Now we construct a packing of B_1 based on a packing of S_k. Let $\{x_1, \ldots, x_M\}$ be a $c_1 k$-packing of S_k. Let $\theta_i = x_i/k$. Then $\theta_i \in B_1$ for $i = 1, \ldots, M$, and

$$\|\theta_i - \theta_j\|_2^2 = \frac{1}{k^2}\|x_i - x_j\|_{\mathrm{H}} \ge \frac{c_1}{k}.$$

Therefore, $\{\theta_1, \ldots, \theta_M\}$ is a $\sqrt{c_1/k}$-packing of B_1 in $\|\cdot\|_2$. Choosing $k = 1/\epsilon^2$, it follows from Lemma 15.3 that

$$\log M(B_1, \|\cdot\|_2, \sqrt{c_1}\epsilon) \ge \frac{c_2}{\epsilon^2} \log\left(ed\epsilon^2\right)$$

for some constants c_1 and c_2.

To summarize, combining the upper and lower bounds, we have

$$\log N(B_1, \|\cdot\|_2, \epsilon) \asymp \begin{cases} \frac{1}{\epsilon^2}\log\left(\epsilon^2 d\right) & \epsilon \gtrsim \frac{1}{\sqrt{d}} \\ d & \epsilon \asymp \frac{1}{\sqrt{d}} \\ d \log \frac{1}{\epsilon^2 d} & \epsilon \lesssim \frac{1}{\sqrt{d}} \end{cases}. \qquad (15.5)$$

15.4 Duality of metric entropy

First we define a more general notoin of covering number. For $K, T \subset \mathbb{R}^d$, define the covering number of K using translates of T as

$$N(K, T) = \min\{N : \exists x_1, \ldots, x_N \in \mathbb{R}^d \text{ such that } K \subset \cup_{i=1}^N T + x_i\}.$$

An amazing theorem of Artstein-Milman-Szarek [AMS04] establishes the following duality result for metric entropy: There exist constants α and β such that for any symmetric convex body K,

$$\frac{1}{\beta} \log N\left(B_2, \frac{\epsilon}{\alpha} K^\circ\right) \le \log N(K, \epsilon B_2) \le \log N(B_2, \alpha\epsilon K^\circ),$$

where B_2 is the usual unit ℓ_2-ball,

$$K^\circ = \left\{ y : \sup_{x \in K} \langle x, y \rangle \le 1 \right\}$$

is the polar body of K. For example, $B_p^\circ = B_q$ whenever $\frac{1}{p} + \frac{1}{q} = 1$. Therefore by duality, (15.5) also applies to $\log N(B_2, \|\cdot\|_\infty, \epsilon)$, which is what is needed for application to minimax risk.

15.5 Example: Sharp rate for ℓ_∞ loss

Finally, we use the results in this lecture to derive the minimax lower bound for the p-dimension, n-sample Gaussian location model with respect to the distortion function $\|\theta - \hat{\theta}\|_\infty^2$.

We can construct an ϵ-packing of $B_2(\delta)$ in $\|\cdot\|_\infty$. From the Fano's method,

$$\begin{aligned}
R^* &\gtrsim \epsilon^2 \left(1 - \frac{\mathrm{diam}_{\mathrm{KL}}(\{N(\theta, \frac{1}{n}\mathbf{I}_p), \theta \in B_2(\delta)\}) + \log 2}{\log M(B_2(\delta), \|\cdot\|_\infty, \epsilon)} \right) \\
&= \epsilon^2 \left(1 - \frac{n\delta^2 + \log 2}{\log M(B_2(\delta), \|\cdot\|_\infty, \epsilon)} \right) \\
&\gtrsim \epsilon^2 \left(1 - \frac{n\delta^2 + \log 2}{\log M(B_1, \|\cdot\|_2, \epsilon/\delta)} \right) \\
&\gtrsim \epsilon^2 \left(1 - \frac{n\delta^2 + \log 2}{\frac{\delta^2}{\epsilon^2} \log\left(1 + \frac{p\epsilon^2}{\delta^2}\right)} \right)
\end{aligned}$$

where we have used that fact that $\mathrm{diam}_{\mathrm{KL}}(\{N(\theta, \frac{1}{n}\mathbf{I}_p), \theta \in B_2(\delta)\}) = n\mathrm{diam}^2_{\|\cdot\|_2}(B_2(\delta))$, the duality theorem, and the lower bound on $\log M(B_1, \|\cdot\|_2, \epsilon/\delta)$. Choosing $\epsilon = c_1\sqrt{\frac{\log p}{n}}$ and $\delta = c_2\epsilon$ with appropriate c_1 and c_2 such that the parenthesis in the lower bound is a positive constant, we obtain

$$R^* \gtrsim \frac{\log p}{n}.$$

An alternative proof of this result is by choosing the packing set as $\tau\{e_1, \dots, e_p\}$ for some $\tau > 0$ to be determined later. This set is a τ-packing of $\mathbb{R}^d$ in $\|\cdot\|_\infty$, because $\|\tau(e_i - e_j)\|_\infty = \tau$ for all pairs $\{i, j\}$. We also have $\|\tau(e_i - e_j)\|_2^2 = 2\tau^2$. Then by Fano's method,

$$R^* \ge \tau^2 \left(1 - \frac{2n\tau^2 + \log 2}{\log p} \right).$$

Choosing $\tau = c\sqrt{\frac{\log p}{n}}$ with some appropriate constant c such that the parenthesis in the above bound is a positive constant, we obtain

$$R^* \gtrsim \frac{\log p}{n}.$$

§ 16. Yang-Barron's construction for density estimation

So far we have been mostly focusing on parametric problems in finite (possibly high) dimensions. In this lecture, we shift our attention to nonparametric problems in infinite-dimensional. We consider the problem of density estimation, which can be formulated as follows. The main goal of this lecture is to give an exposition of the scheme of Yang and Barron [YB99], who constructed an estimator by averaging a sequence of predictive density estimates and give KL risk guarantee in terms the KL covering number.

Given $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p \in \mathcal{P}$, we obtain an estimate $\hat{p} = \hat{p}(\cdot|X_1, \ldots, X_n)$. The loss function is the KL divergence $D(p\|\hat{p})$. The average risk is thus

$$\mathbb{E}_p D(p\|\hat{p}) = \int D\left(p\|\hat{p}(\cdot|x^n)\right) p^{\otimes n}(dx^n).$$

Our task is to upper bound the minimax risk

$$\inf_{\hat{p}} \sup_{p \in \mathcal{P}} \mathbb{E}_p D(p\|\hat{p}), \tag{16.1}$$

We note that the term "density" is a little misleading, and hence quoted, because the elements of $\mathcal{P}$ need not have a density. Crucially, the estimator in (16.1), such as the one due to Yang and Barron, need not be a member of $\mathcal{P}$; in other words, we allow *improper* estimates.

16.1 Bounding Capacity with Covering Number

This section introduces a bound on capacity using covering number, which is useful in terms of both its conclusion and its proof. Before it is formally states, here is a recap on some important concepts.

- KL divergence:
$$D(P\|Q) = \mathbb{E}_P\left[\log \frac{P}{Q}\right].$$

- Mutual information:
$$I(X;Y) = D\left(P_{XY}\|P_X P_Y\right) = \inf_Q D\left(P_{Y|X}\|Q|P_X\right), \tag{16.2}$$
where the infimum is achieved at $Q = P_Y = \mathbb{E}_X[p_{Y|X}]$.

- Capacity: Denote $\mathcal{P} = \left\{p_{Y|X=x} : x \in \mathcal{X}\right\}$, then the capacity is defined as
$$C = \sup_{P_X} I(X;Y) \leq \text{radius} = \inf_Q \sup_{x \in \mathcal{X}} D\left(P_{Y|X=x}\|Q\right), \tag{16.3}$$
with "=" if $\mathcal{P}$ is convex.

- Covering number for sets of distributions

$$\begin{aligned} N(\varepsilon) &= \min \# \text{ of ``balls'' that covers } \mathcal{P} \\ &= \min \left\{ N : \exists Q_1, \ldots, Q_N \text{ s.t. } \forall x \in \mathcal{X}, \exists i \in [N], D(P_{Y|X=x} \| Q_i) \le \varepsilon \right\}. \end{aligned}$$

Now we are ready to state the lemma.

Lemma 16.1.

$$C \le \inf_{\varepsilon > 0} \{\varepsilon + \log N(\varepsilon)\} \tag{16.4}$$

There are two ways of proving this lemma.

Proof. #1

Fix ε, let $N = N(\varepsilon)$, $\exists Q_1, \ldots, Q_N$ that form an ε-cover. $\forall x \in \mathcal{X}$, let $i(x) = \text{argmin}_{i \in [N]} D\left(P_{Y|X=x} \| Q_i\right)$, and thus $D\left(P_{Y|X=x} \| Q_{i(x)}\right) \le \varepsilon$.

Fix any P_X,

$$\begin{aligned} I(X;Y) &= I(X, i(X); Y) \\ &\overset{(1)}{=} I(i(X); Y) + I(X; Y | i(X)) \\ &\le H(i(X)) + I(X; Y | i(X)) \\ &\overset{(2)}{\le} \log N + \varepsilon, \end{aligned}$$

where (1) is derived from the chain rule of mutual information. (2) is derived from that $H(i(X))$ is the entropy of a distribution with N outcomes, whose maximum is achieved when all the outcomes are equiprobable; and that

$$\begin{aligned} I(X; Y | i(X)) &= \inf_Q D\left(P_{Y|X} \| Q | i(X)\right) \\ &\le D\left(P_{Y|X} \| Q_{i(X)} | i(X)\right) \le \varepsilon. \end{aligned}$$

□

Proof. #2

$$\begin{aligned} I(X;Y) &= \inf_Q D\left(P_{Y|X} \| Q | P_X\right) \\ &\le D\left(P_{Y|X} \| \frac{1}{N} \sum_{i=1}^{N} Q_i | P_X\right) \\ &= \mathbb{E}_X \left[D\left(P_{Y|X} \| \frac{1}{N} \sum_{i=1}^{N} Q_i\right) \right] \\ &= \mathbb{E}_X \left[\mathbb{E}_{P_{Y|X}} \log \frac{P_{Y|X}}{\frac{1}{N} \sum_{i=1}^{N} Q_i} \right] \\ &\le \mathbb{E}_X \left[\mathbb{E}_{P_{Y|X}} \log \frac{P_{Y|X}}{\frac{1}{N} Q_{i(X)}} \right] \\ &= \log N + \mathbb{E}_{P_X} \left[D\left(P_{Y|X} \| Q_{i(X)}\right) \right] \\ &\le \log N + \varepsilon. \end{aligned}$$

□

Remark 16.1. "=" in equation (16.4) holds if $\mathcal{P}$ is convex, and thus C =radius (from equation (16.3)). It is easy to verify it with a special case ε=radius, where $N(\varepsilon) = 1$, and both sides of equation (16.4) equal to radius.

Remark 16.2. For n samples $X^n = (X_1, \ldots, X_n) \overset{i.i.d}{\sim} p_X$, note that

$$D\left(P^{\otimes n} \| Q^{\otimes n}\right) = nD(P\|Q).$$

Denote $N_n(\varepsilon)$ is the covering number for $\mathcal{P}^{\otimes n}$, and $N(\varepsilon)$ for $\mathcal{P}$. The product distributions of a ε/n-cover for $\mathcal{P}$ form a ε-cover for $\mathcal{P}^{\otimes n}$. Therefore

$$N_n(\varepsilon) \leq N\left(\frac{\varepsilon}{n}\right).$$

In Gaussian case, for instance, KL-divergence is represented by Euclidean norm, and thus

$$N(\epsilon) \sim \left(\frac{1}{\varepsilon}\right)^d.$$

According to equation (16.4),

$$\begin{aligned} C_n &\lesssim \inf_{\varepsilon > 0} \left\{\varepsilon + d \log \frac{n}{\varepsilon}\right\} \\ &= d \inf_{\varepsilon > 0} \left\{\frac{\varepsilon}{d} + \log \frac{n/d}{\varepsilon/d}\right\} \\ &\overset{\varepsilon' = \varepsilon/d}{=} d \inf_{\varepsilon' > 0} \left\{\varepsilon' + \log \frac{n/d}{\varepsilon'}\right\} \\ &= d \log\left(1 + \frac{n}{d}\right). \end{aligned} \tag{16.5}$$

16.2 An Upper Bound on the Bayes Risk

This section introduces an upper bound on the Bayes risk, which inspires the upper bound on the minimax risk, as will be shown in the next section.

Consider the standard Bayes setting where

$$X^n = (X_1, \ldots, X_n) \overset{i.i.d}{\sim} p_\theta, \text{ and } \theta \sim \pi,$$

and the estimate, $\hat{p}(\cdot|X^n)$, is a function of X^n. The Bayes risk is given by

$$\mathbb{E}_{\theta, X^n}\left[D(p_\theta \| \hat{p}(\cdot|X^n))\right] = \int \pi(d\theta) p_\theta^{\otimes n}(dx^n) D(p_\theta \| \hat{p}(\cdot|x^n)).$$

Lemma 16.2. *The Bayes risk is*

$$\inf_{\hat{p}} \mathbb{E}_{\theta, X^n}\left[D(p_\theta \| \hat{p}(\cdot|X^n))\right] = I(\theta; X_{n+1}|X^n),$$

where X_{n+1} is identically distributed to and independent of $X_1, \ldots, X_n$. The infimum is achieved when

$$\hat{p}(\cdot|X^n) = p_{X_{n+1}|X^n},$$

which is the Bayes estimator.

Proof. First note that p_θ and $\hat{p}(\cdot|X^n)$ are distributions for a new data, which can be denoted as X_{n+1}. Taking the infimum over $\hat{p} = \hat{p}(\cdot|\cdot)$ of the Bayes risk,

$$\begin{aligned}
\inf_{\hat{p}} \mathbb{E}_{\theta, X^n}\left[D(p_\theta \| \hat{p}(\cdot|X^n))\right] &= \inf_{\hat{p}} \int \pi(d\theta) p_\theta^{\otimes n}(dx^n) D(p_\theta \| \hat{p}(\cdot|X^n = x^n)) \\
&= \int p_{X^n}(dx^n) \inf_{\hat{p}} \mathbb{E}_{\theta|X^n = x^n}\left[D(p_\theta \| \hat{p})\right] \\
&= \int p_{X^n}(x^n) \inf_{\hat{p}} D\left(p_{X_{n+1}|\theta} \| \hat{p} | p_{\theta|X^n = x^n}\right) \\
&\overset{(1)}{=} \int p_{X^n}(x^n) D\left(p_{X_{n+1}|\theta} \| p_{X_{n+1}|X^n} | p_{\theta|X^n = x^n}\right) \\
&= I(\theta; X_{n+1}|X^n).
\end{aligned}$$

(1) is derived from equation (16.2). Specifically, fix $X^n = x^n$, the infimum of $D\left(p_{X_{n+1}|\theta} \| \hat{p} | p_{\theta|X^n = x^n}\right)$ is achieved when

$$\hat{p} = \mathbb{E}_{\theta|X^n}\left[p_{X_{n+1}|\theta}\right] = p_{X_{n+1}|X^n}.$$

□

With lemma 16.2, we can derive an upper bound in terms of capacity, fix any prior $\pi(\theta)$,

$$\begin{aligned}
C_{n+1} \equiv \sup_{\pi(\theta)} I(\theta; X^{n+1}) &\geq I(\theta; X^{n+1}) \\
&\overset{(1)}{=} I(\theta; X_1) + I(\theta; X_2|X_1) + \cdots + I(\theta; X_{n+1}|X^n) \\
&\overset{(2)}{\geq} (n+1) I(\theta; X_{n+1}|X^n)
\end{aligned}$$

(1) is due to the chain rule of mutual information; (2) is due to the fact that the mutual information with one extra observation diminishes as the number of existing observations increases, namely

$$I(\theta; X_{n+1}|X^n) \leq I(\theta; X_n|X^{n-1}).$$

Therefore, from equation (16.5), we have a bound for optimal Bayes risk, which holds for any prior $\pi(\theta)$:

$$I(\theta; X_{n+1}|X^n) \leq \frac{C_{n+1}}{n+1} \lesssim \frac{d}{n} \log\left(1 + \frac{n}{d}\right). \tag{16.6}$$

16.3 An Upper Bound for Minimax Risk

This section introduces a theorem which states that the bound in equation (16.6) also holds for minimax risk, and its proof is inspired by the Bayes case.

Theorem 16.1 (Yang-Barron)**.**

$$\inf_{\hat{p}} \sup_{\theta \in \Theta} \mathbb{E}_\theta D(p_\theta \| \hat{p}) \leq \inf_{\varepsilon > 0} \frac{1}{n} \log N(\varepsilon) + \varepsilon \sim \frac{d}{n} \log\left(1 + \frac{n}{d}\right).$$

Proof. Choose the following estimate

$$\hat{p}(\cdot|X^n) = \frac{1}{n} \sum_{i=1}^{n} p_{X_i|X^{i-1}}(\cdot|X^{i-1}),$$

where

$$p_{X_i|X^{i-1}} = \frac{\int \pi(d\theta) \prod_{j=1}^{i} p_\theta(X_j)}{\int \pi(d\theta) \prod_{j=1}^{i-1} p_\theta(X_j)}.$$

Hence the estimator is a function of $\pi(\theta)$. Note that $\pi(\theta)$ here is used only to define an estimator; it has nothing to do with Bayes setting. The rest of the proof bounds the worst case risk of $\hat{p}$ induced by an appropriate $\pi(\theta)$.

Fix θ, the risk for $\hat{p}$ can be upper bounded by

$$\begin{aligned}\mathbb{E}_{p_\theta} D(p_\theta \| \hat{p}) &= \mathbb{E}_{p_\theta} D\left(p_\theta \| \frac{1}{n}\sum_{i=1}^{n} p_{X_i|X^{i-1}}\right) \\ &\overset{(1)}{\leq} \frac{1}{n}\sum_{i+1}^{n} D\left(p_\theta \| p_{X_i|X^{i-1}}\right) \\ &\overset{(2)}{=} \frac{1}{n} D\left(p_\theta^{\otimes n} \| p_{X^n}\right).\end{aligned} \tag{16.7}$$

where (1) is due to the convexity of KL divergence; (2) is the chain rule of KL divergence:

$$\begin{aligned}D\left(P_{X^N} \| Q_{X^N}\right) &= \mathbb{E}\left[\log \frac{P_{X^N}}{Q_{X^N}}\right] \\ &= \mathbb{E}\left[\frac{\prod_{i=1}^{n} P_{X^i|X^{i-1}}}{\prod_{i=1}^{n} Q_{X^i|X^{i-1}}}\right] \\ &= \sum_{i=1}^{n} D\left(P_{X^i|X^{i-1}} \| Q_{X^i|X^{i-1}}\right).\end{aligned}$$

Fix ε, denote $N = N(\varepsilon)$ as the covering number. Let $G = \{\theta_1, \ldots, \theta_N\}$ be a set whose corresponding p_θs form an ε-covering of $\mathcal{P}$. Choose the $\hat{p}$ induced by $\pi(\theta) \sim \text{uniform}(G)$. Then

$$\begin{aligned}D\left(p_\theta^{\otimes n} \| p_{X^n}\right) &= D\left(p_\theta^{\otimes n} \| \frac{1}{N}\sum_{i=1}^{N} p_{\theta_i}^{\otimes n}\right) \\ &= \mathbb{E}\log\left[\frac{p_\theta^{\otimes n}}{\frac{1}{N}\sum_{i=1}^{N} p_{\theta_i}^{\otimes n}}\right] \\ &\leq \mathbb{E}\log\left[\frac{p_\theta^{\otimes n}}{\frac{1}{N} p_{\theta_{i(X)}}^{\otimes n}}\right] \\ &\leq \log N + n\varepsilon.\end{aligned} \tag{16.8}$$

Combining equations (16.7) and (16.8), we can bound the minimax risk:

$$\inf_{p} \sup_{\theta\in\Theta} \mathbb{E}_{p_\theta} D\left(p_\theta \| \hat{p}\right) \leq \sup_{\theta\in\Theta} \mathbb{E}_{p_\theta} D\left(p_\theta \| \hat{p}\right) \leq \frac{1}{n}\left(\log N + n\varepsilon\right) = \frac{1}{n}\log N + \varepsilon.$$

Since it holds for $\forall\varepsilon$, taking the infimum of both sides, and noticing that $N(\varepsilon) \sim (1/\varepsilon)^d$ concludes the proof. □

In Gaussian case, the minimax risk is the canonical d/n (KL-divergence reduces to Euclidean norm), so Theorem 16.1 is loose with an additional log factor $\log(1 + n/d)$. In the next lecture, we will obtain a tighter bound, which is polynomial with respect to $1/n$, given some additional Lipschitz continuity constraint.

§ 17. Application to Smooth Density Estimation

In last lecture, we studied the minimax risk of a parameterized density estimation and its upper bound. We are given n i.i.d. samples X_1, ..., X_n generated from P_θ, where $P_\theta \in \mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is the density to be estimated. Let the loss function between a true distribution P_θ and an estimated distribution $\hat{P}$ be their KL-divergence, i.e.,

$$\ell(P_\theta, \hat{P}) = D(P_\theta \| \hat{P}).$$

One can bound the minimax risk R^* of this estimation problem by,

$$R_n^* = \inf_{\hat{P}} \sup_{\theta \in \Theta} \mathbb{E}_\theta D(P_\theta \| \hat{P}) \leq \frac{C_n}{n}, \tag{17.1}$$

where C_n is the capacity over θ and X^n, i.e.,

$$C_n = \sup_{\pi \in M(\Theta)} I(\theta; X^n) = \inf_{\epsilon > 0} \{n\epsilon + \log N_{KL}(\epsilon)\},$$

where $N_{KL}(\epsilon)$ is the covering number of $\mathcal{P}$.

Further, we can use the chain rule in mutual information to learn the properties of C_n. For any prior π over Θ, one has,

$$R_\pi^* = I(\theta; X_{n+1} | X^*) = I(\theta; X^{n+1}) - I(\theta; X^n).$$

Taking the supremum over π on both sides, one has,

$$\begin{aligned} R_n^* = \sup R_\pi^* &= \sup_\pi \left(I(\theta; X^{n+1}) - I(\theta; X^n) \right) \\ &\geq \sup_\pi I(\theta; X^{n+1}) - \sup_\pi I(\theta; X^n) = C_{n+1} - C_n. \end{aligned}$$

Therefore we can have a lower bound over R_n^* as well.

Remark 17.1. There are some properties of $\{C_n\}$:

- $\{C_n\}$ is subadditive and increasing, i.e.,

  $$C_{n+m} \leq C_n + C_m, \qquad \forall m,\ n \in \mathbb{Z}_+.$$

  and therefore $\frac{C_n}{n}$ has a limit for $n \to \infty$. By Fekete's lemma,

  $$\lim_{n \to \infty} \frac{C_n}{n} = \inf_{n \geq 1} \frac{C_n}{n}.$$

- If we let $\Delta_n = C_{n+1} - C_n$, one can rewrite C_n by,

$$C_n = \sum_{k=1}^{n-1} \Delta_k,$$

  and therefore,

$$\Delta_n \leq \frac{\sum_{k=1}^{n-1} \Delta_k}{n}.$$

In today's lecture, we use the bound in (17.1) to study the minimax risk of a nonparameterized density estimation. As a leading example, consider the problem of estimating a smooth probability density function. To be precise, we are interested in estimating a pdf $f \in \mathcal{P}_\beta$ with smoothness parameter $\beta > 0$, where f belongs to $\mathcal{P}_\beta$ iff,

- f is a pdf on $[0, 1]$ and is upper bounded by a constant, say, 2.
- $f^{(m)}$ α-Hölder continuous, i.e.,

$$|f^{(m)}(x) - f^{(m)}(y)| \leq |x - y|^\alpha, \qquad \forall\, x, y \in (0, 1),$$

  where $\alpha \in (0, 1]$, $m \in \mathbb{Z}$ and $\beta = \alpha + m$.

For example, if $\beta = 1$, then $\mathcal{P}_1$ is simply the set of pdfs which are Lipshitz and bounded by 2.

Theorem 17.1. *Given n i.i.d. samples X_1, ..., X_n randomly generated from a pdf $f \in \mathcal{P}_\beta$, the minimax risk of an estimation $\hat{f}$ of f under the quadratic loss function $\ell(f, \hat{f}) = \|f - \hat{f}\|_2^2 = \int_0^1 (f(x) - \hat{f}(x))^2 dx$ satisfies*

$$R^*(\mathcal{P}_\beta) = \inf_{\hat{f}} \sup_{f \in \mathcal{P}_\beta} \mathbb{E}\|f - \hat{f}\|_2^2 \asymp n^{-\frac{2\beta}{1+\beta}}. \tag{17.2}$$

Before we goes into the proof for Theorem 17.1, we makes some remarks.

Remark 17.2. The larger β is, the smoother the pdfs are, the faster R^* decays with n.

Remark 17.3. If f is defined over $[0, 1]^d$, the bound turns into,

$$R_n^*(\mathcal{P}_\beta) = \inf_{\hat{f}} \sup_{f \in \mathcal{P}_\beta} \mathbb{E}\|f - \hat{f}\|_2^2 \asymp n^{-\frac{2\beta}{d+\beta}}$$

Now we prove Theorem 17.1.

Proof. First, we claim we can use the minimax risk over a set of lower bounded pdfs to bound $R_n^*(\mathcal{P}_\beta)$. The idea is in Lemma 17.1

Lemma 17.1. *Let $\mathcal{F}$ be the set of pdfs that are lower bounded, i.e., $\mathcal{F} = \left\{f : f \geq \frac{1}{2}\right\}$. Let $\mathcal{P}$ be an arbitrary set of pdfs on $[0, 1]$ and let $\tilde{\mathcal{P}} = \mathcal{P} \cap \mathcal{F}$. Then*

$$R_n^*(\mathcal{P}) \leq R_n^*(\tilde{\mathcal{P}}) \leq 16 R_n^*(\mathcal{P}).$$

Proof of Lemma 17.1. Since $\tilde{\mathcal{P}}_\beta \subset \mathcal{P}_\beta$, the lower bound is obvious,

$$R_n^*(\tilde{\mathcal{P}}_\beta) \leq R_n^*(\mathcal{P}_\beta). \tag{17.3}$$

We will construct an estimator to show,

$$R_n^*(\mathcal{P}_\beta) \leq 16 R_n^*(\tilde{\mathcal{P}}_\beta). \tag{17.4}$$

Let X_1, ..., X_n be the n i.i.d. samples from $f \in \mathcal{P}_\beta$ we have, and let U_1, ..., U_n be n i.i.d. samples uniformly generated from $[0,1]$. We define n i.i.d. random variables Z_1, ..., Z_n as,

$$Z_i = \begin{cases} U_i & \text{w.p. } \frac{1}{2}, \\ X_i & \text{otherwise.} \end{cases}$$

Thus, it is equivalent to think Z_1, ..., Z_n are i.i.d. samples from $g = \frac{1}{2}(1+f) \in \tilde{\mathcal{P}}_\beta$. Let $\hat{g}$ be an estimator of g from Z^n. Let $\tilde{g}$ be its projection in $\mathcal{F}$, i.e.,

$$\tilde{g} = \arg\min_{h \in \mathcal{F}} \|h - \hat{g}\|.$$

Note $g \in \mathcal{F}$, and we can bound the distance between $\tilde{g}$ and g by,

$$\|\tilde{g} - g\| \leq \|\hat{g} - g\| + \|\tilde{g} - \hat{g}\| \leq 2\|\hat{g} - g\|.$$

Let $\hat{f} = 2\tilde{g} - 1$, which is a valid pdf since $\tilde{g}$ is lower bounded by $\frac{1}{2}$. As a result, for every pdf $f \in \mathcal{P}_\beta$, there is a corresponding $g = \frac{1}{2}(1+f) \in \tilde{\mathcal{P}}_\beta$ which has a good estimator $\hat{g}$, and one can construct a good estimator $\hat{f}$ from $\hat{g}$ in the sense that,

$$\|\hat{f} - f\| = 2\|\tilde{g} - g\| \leq 4\|\hat{g} - g\|.$$

Therefore,

$$\begin{aligned} R_n^*(\mathcal{P}_\beta) &= \inf_{\hat{f}} \sup_{f \in \mathcal{P}_\beta} \mathbb{E}\|\hat{f} - f\|_2^2 \\ &\leq 16 \inf_{\hat{g}} \sup_{f \in \mathcal{P}_\beta} \mathbb{E} \left\| \hat{g} - \frac{1}{2}(1+f) \right\|_2^2 \\ &\leq 16 \inf_{\hat{g}} \sup_{g \in \tilde{\mathcal{P}}_\beta} \mathbb{E}\, \|\hat{g} - g\|_2^2 = R_n^*(\tilde{\mathcal{P}}_\beta), \end{aligned}$$

where the first inequality is due to the construction of $\hat{f}$, and the second inequality is due to $\left\{\frac{1}{2}(1+f) : f \in \mathcal{P}_\beta\right\} \subset \tilde{\mathcal{P}}_\beta$. Therefore from (17.3) and (17.4) Lemma 17.1 follows. □

It is then equivalent to prove,

$$R_n^*(\tilde{\mathcal{P}}_\beta) = \inf_{\hat{f}} \sup_{f \in \mathcal{P}_\beta} \mathbb{E}\|f - \hat{f}\|_2^2 \asymp n^{-\frac{2\beta}{d+\beta}}$$

Upper bound First we use the capacity to upper bound the minimax risk. On one hand, It is known that for any bounded pdf f and g,

$$\|f-g\|_1^2 \gtrsim \|f-g\|_2^2,$$

and the total variation between f and g is bounded by its KL-divergence,

$$D(f\|g) \geq 2d_{\text{TV}}^2(f,g) = \frac{1}{2}\|f-g\|_1^2.$$

Therefore, we have for any bounded pdf f and g,

$$\|f-g\|_2^2 \lesssim D(f\|g)$$

As a result,

$$R_n^*(\tilde{\mathcal{P}}_\beta) = \inf_{\hat{g}} \sup_{g\in\tilde{\mathcal{P}}_\beta} \mathbb{E}\|g-\hat{g}\|_2^2 \leq \inf_{\hat{g}} \sup_{g\in\tilde{\mathcal{P}}_\beta} \mathbb{E}D(g\|\hat{g}) = R_{n,KL}^*(\tilde{\mathcal{P}}_\beta). \tag{17.5}$$

On the other hand, one can bound the minimax risk under KL-divergence by (17.1), where the capacity between g and X^n can be computed via,

$$\begin{aligned} C_n &\leq \inf_{\epsilon>0}\{\log N_{KL}(\epsilon) + n\epsilon\} \\ &\asymp \inf_{\epsilon>0}\{\log N_2(\sqrt{\epsilon}) + n\epsilon\} \\ &\asymp \inf_{\epsilon>0}\{\epsilon^{-\frac{1}{2\beta}} + n\epsilon\} = n^{\frac{1}{1+2\beta}}. \end{aligned}$$

The second equality is due to the connection between the KL-divergence and the L_2 distance. The third equality comes from Kolomogrov-Tikhomirov's Theorem. Therefore with (17.5), the upper bound is proved by showing,

$$R_n^*(\tilde{\mathcal{P}}_\beta) \lesssim \frac{C_n}{n} \lesssim n^{-\frac{2\beta}{1+2\beta}}. \tag{17.6}$$

Lower bound Next we lower bound $R_n^*(\mathcal{P}_\beta)$ by Fano's inequality. Due to the relation between covering and packing numbers, we know,

$$\log M(\tilde{\mathcal{P}}_\beta, \|\cdot\|_2, \epsilon) \asymp \log N(\tilde{\mathcal{P}}_\beta, \|\cdot\|_2, \epsilon) \asymp \epsilon^{-1/\beta},$$

where the second equality is due to Kolomogrov-Tikhomirov's Theorem. Let $\epsilon = n^{-\frac{\beta}{1+2\beta}}$. Fano's inequality tells us,

$$\begin{aligned} R_n^*(\tilde{\mathcal{P}}_\beta) &\gtrsim \epsilon^2\left(1 - \frac{I(g;X^n)+\log 2}{\log M(\tilde{\mathcal{P}}_\beta, \|\cdot\|_2, \epsilon)}\right) \\ &\gtrsim \epsilon^2\left(1 - \frac{C_n}{\log M(\tilde{\mathcal{P}}_\beta, \|\cdot\|_2, \epsilon)}\right) \\ &\gtrsim \epsilon^2\left(1 - \frac{n^{\frac{1}{1+2\beta}}}{\epsilon^{-\frac{1}{\beta}}}\right) \\ &\asymp \epsilon^2 = n^{-\frac{2\beta}{1+2\beta}}. \end{aligned} \tag{17.7}$$

The proof is done via (17.6) and (17.7). □

We make some remarks on the proof.

Remark 17.4. We have learned two ways to construct a density estimator:

- The mean of predictive density estimators;
- The maximum likelihood estimator.

None of those is computationally efficient. In practice, kernel density estimator (KDE) is proposed: let $X_1, ..., X_n$ be the n samples, one can estimate the density by its histogram,

$$\hat{\pi} = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}.$$

This estimator, however, is not a pdf. To address this issue, one put a kernel instead of a spike over each sample points, i.e.,

$$\hat{f}_\Omega = \Omega \otimes \hat{\pi},$$

where Ω is the kernel function and is chosen to satisfy the smooth constraint of the pdfs.

Remark 17.5. The result in Theorem 17.1 can be generalized to L_p for $p < \infty$, following the same analysis.

§ 18. Density estimation via pairwise comparison à la Le Cam-Birgé

When we prove the lower bound in Theorem 17.1, we use the reasoning that if an ϵ-covering of a set Θ cannot be tested, then $\theta \in \Theta$ cannot be estimated more than precision ϵ, thereby establishing a minimax *lower bound* in terms of the KL metric entropy. Conversely, we can ask the following question:

> Is it possible to construct an estimator based on tests, and produce a minimax *upper bound* in terms of the metric entropy?

For Hellinger loss, the answer is yes, although the metric entropy involved is with respect to the Hellinger distance not KL divergence. The basic construction is due to Le Cam and further developed by Birgé. The main idea is as follows: Fix an ϵ-covering $\{P_1, \ldots, P_N\}$ of the set of distributions $\mathcal{P}$. Given n samples drawn from $P \in \mathcal{P}$, let us test which ball P belongs to; this allows us to estimate P up to Hellinger loss ϵ. This can be realized by a pairwise comparison (tournament) of testing the (composite) hypothesis $P \in B(P_i, \epsilon)$ versus $P \in B(P_j, \epsilon)$. This program can be further refined to involve on the local entropy of the model.

18.1 Composite hypothesis testing and Hellinger distance

Consider the following general problem. Let $\mathcal{P}$ and $\mathcal{Q}$ be two (*not* necessarily convex) classes of distributions. Given iid samples $X_1, \ldots, X_n$ drawn from some distribution P, we want to test, according some decision rule $\phi = \phi(X_1, \ldots, X_n) \in \{0, 1\}$, whether $P \in \mathcal{P}$ ($\phi = 0$) or $P \in \mathcal{Q}$ ($\phi = 1$). By the minimax theorem, the optimal error is given by the total variation between the worst-case mixtures:

$$\min_{\phi} \left\{ \sup_{P \in \mathcal{P}} P(\phi = 1) + \sup_{Q \in \mathcal{Q}} Q(\phi = 0) \right\} = 1 - d_{\rm TV}(\mathrm{co}(\mathcal{P}^{\otimes n}), \mathrm{co}(\mathcal{Q}^{\otimes n})), \tag{18.1}$$

wherein the notations are explained as follows:

- $\mathcal{P}^{\otimes n} \triangleq \{P^{\otimes n} : P \in \mathcal{P}\}$ consists of all n-fold products of distributions in $\mathcal{P}$;
- $\mathrm{co}(\cdot)$ denotes the convex hull, that is, the set of all mixtures. For example, for a parametric family, $\mathrm{co}(\{P_\theta : \theta \in \Theta\}) = \{P_\pi : \pi \in \Delta(\Theta)\}$, where $P_\pi = \int P_\theta \pi(d\theta)$ is the mixture under the mixing distribution π, and $\Delta(\Theta)$ denotes the collection of all probability distributions (priors) on Θ.
- The distance d (e.g. total variation $d_{\rm TV}$) between for two sets A and B is defined as usual by that of the closest pairs: $d(A, B) \triangleq \inf_{a \in A, b \in B} d(a, b)$.

The optimal test that achieves (18.1) is the likelihood ratio given by the worst-case mixtures, that is, the closest[1] pair of mixture (P_n^*, Q_n^*) such that $d_{\rm TV}(P_n^*, Q_n^*) = d_{\rm TV}(\mathrm{co}(\mathcal{P}^{\otimes n}), \mathrm{co}(\mathcal{Q}^{\otimes n}))$.

The exact result (18.1) is unwieldy as the RHS involves finding the least favorable priors over the n-fold product space. The following result, due to Le Cam, is a "single-letter" upper bound in terms of the Hellinger separation. It is the consequence of the more general tensorization property of Rényi divergence (of which Hellinger is a special case). Here we give a straightforward argument.

Theorem 18.1.

$$\min_{\phi} \left\{ \sup_{P\in\mathcal{P}} P(\phi = 1) + \sup_{Q\in\mathcal{Q}} Q(\phi = 0) \right\} \le \exp\left\{ -\frac{n}{2} \inf_{P\in\mathcal{P}, Q\in\mathcal{Q}} H^2(P,Q) \right\}, \tag{18.2}$$

For this, we need the following key lemma:

Lemma 18.1. *For any sets of distributions $\mathcal{P}_i$ and $\mathcal{Q}_i$,*

$$1 - \frac{1}{2} H^2 \left(\mathrm{co}\left(\bigotimes_{i=1}^n \mathcal{P}_i \right), \mathrm{co}\left(\bigotimes_{i=1}^n \mathcal{Q}_i \right) \right) \le \prod_{i=1}^n \left(1 - \frac{1}{2} H^2(\mathrm{co}(\mathcal{P}_i), \mathrm{co}(\mathcal{Q}_i)) \right).$$

Proof. Denote $\mathrm{Aff} \triangleq 1 - \frac{1}{2}H^2$. By induction it is sufficient to consider $n = 2$ and we aim to show

$$\mathrm{Aff}(\mathrm{co}(\mathcal{P}_1 \otimes \mathcal{P}_2), \mathrm{co}(\mathcal{Q}_1 \otimes \mathcal{Q}_2)) \le \mathrm{Aff}(\mathrm{co}(\mathcal{P}_1), \mathrm{co}(\mathcal{Q}_1)) \cdot \mathrm{Aff}(\mathrm{co}(\mathcal{P}_2), \mathrm{co}(\mathcal{Q}_2)). \tag{18.3}$$

Any element of $\mathrm{co}(\mathcal{P}_1 \otimes \mathcal{P}_2)$ corresponds to a joint distribution $P_{X_1X_2}$ where

- X_1 and X_2 are conditionally independent given some latent variable Z, i.e., $P_{X_1X_2|Z} = P_{X_1|Z} \otimes P_{X_2|Z}$;
- For any z, $P_{X_1|Z=z} \in \mathcal{P}_1$ and $P_{X_2|Z=z} \in \mathcal{P}_2$.

Similarly, we have $Q_{X_1X_2} \in \mathrm{co}(\mathcal{Q}_1 \otimes \mathcal{Q}_2)$ with $Q_{X_1X_2|Z} = Q_{X_1|Z} \otimes Q_{X_2|Z}$, $Q_{X_1|Z=z} \in \mathcal{Q}_1$, $Q_{X_2|Z=z} \in \mathcal{Q}_2$ for any z.

Let $\alpha_i = \mathrm{Aff}(\mathrm{co}(\mathcal{P}_i), \mathrm{co}(\mathcal{Q}_i))$ for $i = 1, 2$. Then

$$\begin{aligned}
\mathrm{Aff}(P_{X_1X_2}, Q_{X_1X_2}) &= \int \sqrt{P_{X_1X_2}(dx_1dx_2) Q_{X_1X_2}(dx_1dx_2)} \\
&= \int \sqrt{P_{X_1}(dx_1) Q_{X_1}(dx_1)} \underbrace{\left\{ \int \sqrt{P_{X_2|X_1=x_1}(dx_2) Q_{X_2|X_1=x_1}(dx_2)} \right\}}_{\mathrm{Aff}(P_{X_2|X_1=x_1}, Q_{X_2|X_1=x_1})} \\
&\overset{\text{(a)}}{\le} \alpha_2 \cdot \mathrm{Aff}(P_{X_1}, Q_{X_1}) \overset{\text{(b)}}{\le} \alpha_1 \alpha_2
\end{aligned}$$

where (a) follows from the crucial observation that $P_{X_2|X_1=x_1} = \int P_{X_2|Z=z} P_{Z|X_1=x_1}(dz) \in \mathrm{co}(\mathcal{P}_2)$ and $Q_{X_2|X_1=x_1} \in \mathrm{co}(\mathcal{Q}_2)$ likewise; (b) follows from the fact $P_{X_1} \in \mathrm{co}(\mathcal{P}_1)$ and $Q_{X_1} \in \mathrm{co}(\mathcal{Q}_1)$. This proves the desired (18.3). □

Now we finish the proof of Theorem 18.1:

[1] In case the closest pair does not exist, we can replace it by an infimizing sequence.

Proof. From (18.1) we get

$$
\begin{aligned}
1 - d_{\mathrm{TV}}(\mathrm{co}(\mathcal{P}^{\otimes n}), \mathrm{co}(\mathcal{Q}^{\otimes n})) &\overset{(a)}{\leq} 1 - \frac{1}{2}H^2(\mathrm{co}(\mathcal{P}^{\otimes n}), \mathrm{co}(\mathcal{Q}^{\otimes n})) \\
&\overset{(b)}{\leq} \left(1 - \frac{1}{2}H^2(\mathrm{co}(\mathcal{P}), \mathrm{co}(\mathcal{Q}))\right)^n \leq \exp\left(-\frac{n}{2}H^2(\mathrm{co}(\mathcal{P}), \mathrm{co}(\mathcal{Q}))\right)
\end{aligned}
$$

where (a) follows from (4.10); (b) follows from Lemma 18.1. □

In the sequel we will apply Theorem 18.1 to two disjoint Hellinger balls (both are convex).

18.2 Hellinger guarantee on Le Cam-Birgé's pairwise comparison estimator

The idea of constructing estimator based on pairwise tests is due to Le Cam ([LC86], see also [vdV02, Section 10]) and Birgé [Bir83]. We are given n i.i.d. samples $X_1, X_2, \cdots, X_n$ generated from P, where $P \in \mathcal{P}$ is the distribution to be estimated. Here let us emphasize that $\mathcal{P}$ need **not** be a convex set. Let the loss function between the true distribution P and the estimated distribution $\hat{P}$ be their squared Hellinger distance, i.e.

$$\ell(P, \hat{P}) = H^2(P, \hat{P}).$$

Then, we have the following result:

Theorem 18.2 (Le Cam-Birgé)**.** *Denote by $N_H(\mathcal{P}, \epsilon)$ the ϵ-covering number of the set $\mathcal{P}$ under the Hellinger distance (cf. Definition 14.3). Let ϵ_n be such that*

$$n\epsilon_n^2 \geq \log N_H(\mathcal{P}, \epsilon_n) \vee 1.$$

Then there exists an estimator $\hat{P} = \hat{P}(X_1, \ldots, X_n)$ taking values in $\mathcal{P}$ such that for any $t \geq 1$,

$$\sup_{P \in \mathcal{P}} P[H(P, \hat{P}) > 4t\epsilon_n] \lesssim e^{-t^2} \tag{18.4}$$

and, consequently,

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P[H^2(P, \hat{P})] \lesssim \epsilon_n^2 \tag{18.5}$$

Remark 18.1. Let's compare this result with lower and upper bounds in terms of the KL covering number:

1. Lower bound: Based on Fano's inequality (cf. Section 14.2) we have a minimax lower bound ϵ_n^2 for the KL divergence loss in terms of the KL covering number, where ϵ_n^2 is given by

   $$\log N_{KL}(\mathcal{P}, \epsilon_n^2) \asymp n\epsilon_n^2. \tag{18.6}$$

   Note that (**??**) and (18.6) look tantalizingly similar, except that the covering number is with respect to Hellinger and KL, and the loss function is for Hellinger and KL respectively, so they are not directly comparable.

2. Upper bound: Recall Theorem 16.1 due to Yang-Barron which provides a general upper bound for density estimation under the KL divergence loss (stronger than Hellinger) in terms of the KL covering number. It should be remarked that the estimator of Yang-Barron in general does *not* belong to the original space $\mathcal{P}$ (unless $\mathcal{P}$ is convex, which fails for most parameter models); in the language of learning theory, such an estimator is called *improper*. In contrast, the Le Cam-Birgé estimator is proper, just like MLE; in the case of a parametric family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, this means the construction yields a parameter estimator $\hat{\theta} \in \Theta$ for the loss function $\ell(\theta, \hat{\theta}) \triangleq H(P_\theta, P_{\hat{\theta}})$.

Proof of Theorem 18.2. It suffices to prove the high-probability bound (18.4). Abbreviate $\epsilon = \epsilon_n$ and $N = N_H(\mathcal{P}, \epsilon_n)$. Let $P_1, \cdots, P_N$ be a maximal ϵ-packing of $\mathcal{P}$ under the Hellinger distance, which also serves as an ϵ-covering (cf. Theorem 14.1). Thus, $\forall i \neq j$,

$$H(P_i, P_j) \geq \epsilon,$$

and for $\forall P \in \mathcal{P}$, $\exists i \in [N]$, s.t.

$$H(P, P_i) \leq \epsilon,$$

Denote $B(P, \epsilon) = \{Q : H(P, Q) \leq \epsilon\}$ denote the ϵ-Hellinger ball centered at P. Crucially, *Hellinger ball is convex*[2] thanks to the convexity of squared Hellinger distance as an f-divergence (cf. Theorem 4.1). Indeed, for any $P', P'' \in B(P, \epsilon)$ and $\alpha \in [0, 1]$,

$$H^2(\bar{\alpha}P' + \alpha P'', P) \leq \bar{\alpha}H^2(P', P) + \alpha H^2(P'', P) \leq \epsilon^2.$$

Next, consider the following **pairwise comparison problem**, where we test two Hellinger balls (composite hypothesis) against each other:

$$\begin{cases} H_i : P \in B(P_i, \epsilon) \\ H_j : P \in B(P_j, \epsilon) \end{cases}$$

for all $i \neq j$, s.t. $H(P_i, P_j) \geq \delta = 4\epsilon$.

Since both $B(P_i, \epsilon)$ and $B(P_j, \epsilon)$ are convex, applying Theorem 18.1 yields a test $\psi_{ij} = \psi_{ij}(X_1, \ldots, X_n)$, with $\psi_{ij} = 0$ corresponding to declaring $P \in B(P_i, \epsilon)$, and $\psi_{ij} = 1$ corresponding to declaring $P \in B(P_j, \epsilon)$, such that $\psi_{ij} = 1 - \psi_{ji}$ and the following large deviation bound holds: for all i, j, s.t. $H(P_i, P_j) \geq \delta$,

$$\sup_{P \in B(P_i, \epsilon)} P(\psi_{ij} = 1) \leq e^{-\frac{n}{8}H(P_i, P_j)^2}, \tag{18.7}$$

where we used the triangle inequality of Hellinger distance: for any $P \in B(P_i, \epsilon)$ and any $Q \in B(P_j, \epsilon)$,

$$H(P, Q) \geq H(P_i, P_j) - 2\epsilon \geq H(P_i, P_j)/2 \geq 2\epsilon.$$

For $i \in [N]$, define the random variable

$$T_i \triangleq \begin{cases} \max_{j \in [N]} H^2(P_i, P_j) & \text{s.t. } \psi_{ij} = 1, \quad H(P_i, P_j) > \delta; \\ 0, & \text{no such } j \text{ exists.} \end{cases}$$

[2] Note that this is not entirely obvious because $P \mapsto H(P, Q)$ is not convex (for example, consider $p \mapsto H(\text{Bern}(p), \text{Bern}(0.1))$.

Basically, T_i records the maximum distance from P_i to those P_j outside the δ-neighborhood of P_i that is confusable with P_i given the present sample. Our density estimator is defined as

$$\hat{P} = P_{i^*}, \quad \text{where} \quad i^* \in \arg\min_{i \in [N]} T_i. \tag{18.8}$$

Now for the proof of correctness, assume that $P \in B(P_1, \epsilon)$. The intuition is that, we should expect, typically, that $T_1 = 0$, and furthermore, $T_j \geq \delta^2$ for all j such that $H(P_1, P_j) \geq \delta$. Note that by the definition of T_i and the symmetry of the Hellinger distance, for any pair i, j such that $H(P_i, P_j) \geq \delta$, we have

$$\max\{T_i, T_j\} \geq H(P_i, P_j).$$

Consequently,

$$\begin{aligned} H(\hat{P}, P_1)\mathbf{1}\{H(\hat{P}, P_1) \geq \delta\} &= H(P_{i_*}, P_1)\mathbf{1}\left\{H(P_{i_*}, P_1) \geq \delta\right\} \\ &\leq \max\{T_{i_*}, T_1\}\mathbf{1}\left\{\max\{T_{i_*}, T_1\} \geq \delta\right\} = T_1\mathbf{1}\left\{T_1 \geq \delta\right\}, \end{aligned}$$

where the last equality follows from the definition of i_* as a global minimizer in (18.8). Thus, for any $t \geq 1$,

$$\begin{aligned} P[H(\hat{P}, P_1) \geq t\delta] &\leq P[T_1 \geq t\delta] \\ &\leq N(\epsilon)e^{-2n\epsilon^2 t^2} \qquad (18.9) \\ &\lesssim e^{-t^2}, \qquad (18.10) \end{aligned}$$

where (18.9) follows from (18.7) and (18.10) uses the assumption that $n\epsilon^2 \geq 1$ and $N \leq e^{n\epsilon^2}$. □

18.3 Refinement using local entropy

Just like Theorem 16.1, while they are often tight for nonparametric problems where the metric entropy grows superlogarithmically, for finite-dimensional models the direct application of Theorem 18.2 results in a slack by a log factor. For example, for a d-dimensional parametric family, e.g., the Gaussian location model or its finite mixtures, the metric entropy usually behaves as $\log N_H(\epsilon) \asymp d \log \frac{1}{\epsilon}$. Thus when $n \gtrsim d$, Theorem 18.2 entails choosing $\epsilon_n^2 \asymp \frac{d}{n} \log \frac{n}{d}$, which falls short of the parametric rate $\mathbb{E}[H^2(\hat{P}, P)] \lesssim \frac{d}{n}$ which are typically achievable.

As usual, such a log factor can be removed using the local entropy argument. To this end, define the local Hellinger entropy:

$$N_{\text{loc}}(\mathcal{P}, \epsilon) \triangleq \sup_{P \in \mathcal{P}} \sup_{\eta \geq \epsilon} N_H(B(P, \eta) \cap \mathcal{P}, \eta/2). \tag{18.11}$$

Theorem 18.3 (Le Cam-Birgé: local entropy version)**.** *Let ϵ_n be such that*

$$n\epsilon_n^2 \geq \log N_{\text{loc}}(\mathcal{P}, \epsilon_n) \vee 1.$$

Then there exists an estimator $\hat{P} = \hat{P}(X_1, \ldots, X_n)$ taking values in $\mathcal{P}$ such that for any $t \geq 2$,

$$\sup_{P \in \mathcal{P}} P[H(P, \hat{P}) > 4t\epsilon_n] \leq e^{-t^2} \tag{18.12}$$

and hence

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P[H^2(P, \hat{P})] \lesssim \epsilon_n^2 \tag{18.13}$$

Remark 18.2 (Doubling dimension)**.** Suppose that for some $d > 0$, $\log N_{\text{loc}}(\mathcal{P}, \epsilon) \le d \log \frac{1}{\epsilon}$ holds for all sufficiently large small ϵ; this is the case for finite-dimensional models where the Hellinger distance is comparable with the vector norm by the usual volume argument (Theorem 14.2). Then we say the *doubling dimension* (also known as the *Le Cam dimension* [vdV02]) of $\mathcal{P}$ is at most d; this terminology comes from the fact that the local entropy concerns covering Hellinger balls using balls of half the radius. Then Theorem 18.3 shows that it is possible to achieve the "parametric rate" $O(\frac{d}{n})$. In this sense, the doubling dimension serves as the effective dimension of the model $\mathcal{P}$.

Lemma 18.2. *For any $P \in \mathcal{P}$ and $\eta \ge \epsilon$ and $k \ge \mathbb{Z}_+$,*

$$N_H(B(P, 2^k\eta) \cap \mathcal{P}, \eta/2) \le N_{\text{loc}}(\mathcal{P}, \epsilon)^k \tag{18.14}$$

Proof. We proceed by induction on k. The base case of $k = 0$ follows from the definition (18.11). For $k \ge 1$, assume that (18.14) holds for $k - 1$ for all $P \in \mathcal{P}$. To prove it for k, we construct a cover of $B(P, 2^k\eta) \cap \mathcal{P}$ as follows: first cover it with $2^{k-1}\eta$-balls, then cover each ball with $\eta/2$-balls. By the induction hypothesis, the total number of balls is at most

$$N_H(B(P, 2^k\eta) \cap \mathcal{P}, 2^{k-1}\eta) \cdot \sup_{P' \in \mathcal{P}} N_H(B(P', 2^{k-1}\eta) \cap \mathcal{P}, \eta/2) \le N_{\text{loc}}(\epsilon) \cdot N_{\text{loc}}(\epsilon)^{k-1}$$

completing the proof. □

We now prove Theorem 18.3:

Proof. We analyze the same estimator (18.8) following the proof of Theorem 18.2, except that the estimate (18.9) is improved as follows: Define the Hellinger shell $A_k \triangleq \{P : 2^k\delta \le H(P_1, P) < 2^{k+1}\delta\}$ and $G_k \triangleq \{P_1, \ldots, P_N\} \cap A_k$. Recall that $\delta = 4\epsilon$. Given $t \ge 2$, let $\ell = \lfloor \log_2 t \rfloor$ so that $2^\ell \le t < 2^{\ell+1}$. Then

$$\begin{aligned}
P[T_1 \ge t\delta] &\le \sum_{k \ge \ell} P[2^k\delta \le T_1 < 2^{k+1}\delta] \\
&\overset{\text{(a)}}{\le} \sum_{k \ge \ell} |G_k| e^{-\frac{n}{8}(2^k\delta)^2} \\
&\overset{\text{(b)}}{\le} \sum_{k \ge \ell} N_{\text{loc}}(\epsilon)^{k+3} e^{-2n\epsilon^2 4^k} \\
&\overset{\text{(c)}}{\lesssim} e^{-4^\ell} \le e^{-t^2}
\end{aligned}$$

where (a) follows from from (18.7); (c) follows from the assumption that $\log N_{\text{loc}} \le n\epsilon^2$ and $k \ge \ell \ge \log_2 t \ge 1$; (b) follows from the following reasoning: since $\{P_1, \ldots, P_N\}$ is an ϵ-packing, we have

$$|G_k| \le M(A_k, \epsilon) \le N(A_k, \epsilon/2) \le N(B(P_1, 2^{k+1}\delta) \cap \mathcal{P}, \epsilon/2) \le N_{\text{loc}}(\epsilon)^{k+3}$$

where the first and the last inequalities follow from Theorem 14.1 and Lemma 18.2 respectively. □

Part IV

Structured high-dimensional estimation problems

§ 19. Denoising sparse vectors: lower bound

This lecture focuses on the problem of denoising for a sparse vector. Let's begin with our favorite example, the Gaussian Location model.

Example 19.1 (GLM)**.** Consider the p-dimensional n-sample GLM. We have

$$Y_i = \theta + Z_i,$$

where $\theta \in \Theta \subseteq \mathbb{R}^p$, $i \in [n]$. We have n i.i.d copies of Y, and the noise $Z \sim \mathcal{N}(0, I_p)$. We consider the quadratic minimax loss for this estimation problem,

$$R_n^*(\Theta) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2.$$

Here, we add some structure for the parameter space Θ to study the so called **denoising by sparsity** problem, let

$$\Theta = \{\text{all } k\text{-sparse vectors}\} = B_0(k) = \{\theta \in \mathbb{R}^p, \|\theta\|_0 \leq k\}, \quad k \in [p],$$

where $\|\theta\|_0 = |\{i : \theta_i \neq 0\}|$ is the number of nonzero entries of θ, indicating the sparsity of θ. We want to analysis the asymptotic behavior of $R_n^*(B_0(k))$.

Remark 19.1. The set $B_0(k)$ can be written as a union of linear subspace of $\mathbb{R}^p$.

$$B_0(k) = \bigcup_{S \subseteq [p], |S| \leq k} \{\theta, \theta_{S^C} = 0\}.$$

Remark 19.2. To study the behavior of $R_n^*(B_0(k))$, it is sufficient to consider one sample and the risk $R_1^*(B_0(k))$. Indeed, we have

$$R_n^*(B_0(k)) = \frac{1}{n} R_1^*(B_0(k)).$$

Proof. Since $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ is the sufficient statistics of this problem, and $\bar{Y} \sim \mathcal{N}(\theta, \frac{1}{n} I_p)$. Given n i.i.d. samples, it is sufficient to solve the following one-dimensional problem,

$$\bar{Y} = \theta + \frac{1}{\sqrt{n}} Z \quad \Leftrightarrow \quad \sqrt{n}\bar{Y} = \sqrt{n}\theta + Z,$$

where $Z \sim \mathcal{N}(0, I_p)$. Since $\sqrt{n} B_0(k) = B_0(k)$, estimating $\sqrt{n}\theta$ has the same minimax risk for estimating θ given one sample. Thus,

$$\begin{aligned} R_1^*(B_0(k)) &= \inf_{\hat{\theta}} \sup_{\sqrt{n}\theta \in \sqrt{n} B_0(k)} \mathbb{E}_\theta \|\sqrt{n}\hat{\theta} - \sqrt{n}\theta\|_2^2 \\ &= n \inf_{\hat{\theta}} \sup_{\theta \in B_0(k)} \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2 \\ &= n R_n^*(B_0(k)). \end{aligned}$$

Thus, in the ensuing discussion, we only consider the case of $n = 1$. □

Remark 19.3 (The Oracle Lower bound)**.** Note that for $n = 1$, given the information of the position of nonzero entries, we have the following lower bound which has been proved before

$$R_1^*(B_0(k)) \geq k.$$

Theorem 19.1. *The minimax risk for sparse denoising is*

$$R_1^*(B_0(k)) \asymp k + \log \binom{p}{k} \asymp k \log \frac{ep}{k}.$$

Furthermore, if $k = o(p)$ and $p \to \infty$, we have the following results (proved in homework)

$$R_1^*(B_0(k)) = \big(2 + o(1)\big) k \log \frac{p}{k}.$$

19.1 Lower bound for denoising by sparsity

We will prove the lower bound in Theorem 19.1 by the mutual information method. Consider the following subset of Hamming space:

$$B = \{b \in \{0,1\}^p : w_H(b) = k\},$$

where $w_H(b)$ is the Hamming weights of b. Suppose that b is drawn uniformly from the set B, and $\theta = \tau b$, where

$$\tau = \sqrt{\frac{1}{100} \log \frac{p}{k}}.$$

Thus, we have the following Markov chain which represents our problem model,

$$b \to \theta \to Y \to \hat{\theta} \to \hat{b}.$$

Denote the set $G = \tau B$, so $\theta \in G$. The mutual information is upper bounded by the radius of set G,

$$\begin{aligned} I(\theta; \hat{\theta}) \leq I(\theta; Y) &\leq \mathrm{rad}_{kL}(\mathcal{N}(\theta, I_p), \theta \in G) \\ &\leq \sup_{\theta \in G} D(P_\theta \| P_0) \\ &= \sup_{\theta \in G} \frac{1}{2} \|\theta\|_2^2 = \frac{k\tau^2}{2}. \end{aligned}$$

To give a lower bound for $I(\theta; \hat{\theta})$, consider

$$\hat{b} = \arg\min_{b \in B} \|\hat{\theta} - \tau b\|_2^2.$$

Since $\hat{b}$ is the minimizer of $\|\hat{\theta} - \tau b\|_2^2$, we have,

$$\|\tau \hat{b} - \theta\|_2 \leq \|\tau \hat{b} - \hat{\theta}\|_2 + \|\theta - \hat{\theta}\|_2 \leq 2\|\theta - \hat{\theta}\|_2.$$

Thus,

$$\tau^2 d_H(b, \hat{b}) = \|\tau \hat{b} - \theta\|_2^2 \leq 4\|\theta - \hat{\theta}\|_2^2,$$

where d_H denotes the Hamming distance between b and $\hat{b}$. Suppose that $\mathbb{E}\|\hat{\theta} - \theta\|_2^2 = \epsilon\tau^2 k$. Then we have $\mathbb{E}d_H(b, \hat{b}) \le 4\epsilon k$. Our goal is to show that ϵ is at least a small constant by the mutual information method:

$$\begin{aligned}
I(\hat{b}; b) &\ge \min_{\mathbb{E}d_H(b,\hat{b}) \le 4\epsilon k} I(\hat{b}; b) \\
&= H(b) - \max_{\mathbb{E}d_H(b,\hat{b}) \le 4\epsilon k} H(b|\hat{b}) \\
&= \log \binom{p}{k} - \max_{\mathbb{E}d_H(b,\hat{b}) \le 4\epsilon k} H(b \oplus \hat{b}|\hat{b}) \\
&\ge \log \binom{p}{k} - \max_{\mathbb{E}w_H(b \oplus \hat{b}) \le 4\epsilon k} H(b \oplus \hat{b}).
\end{aligned}$$

Note the following fact[1],

$$\max_{\mathbb{E}w_H(W) = m, W \in \{0,1\}^p} H(W) = ph\left(\frac{m}{p}\right).$$

Combine this with the previous bound, we get

$$I(\hat{b}; b) \ge \log \binom{p}{k} - ph(\frac{4\epsilon k}{p}).$$

On the other hand, we have

$$I(\hat{b}; b) \le I(\theta; Y) \le \frac{k}{2}\tau^2 = \frac{k}{200} \log \frac{p}{k}.$$

Note that $h(\alpha) \asymp -\alpha \log \alpha$ for $\alpha < \frac{1}{4}$. WLOG, since $k \le \frac{p}{16}$, we have $\epsilon \ge c_0$ for some universal constant c_0. Therefore

$$R^* \ge \epsilon\tau^2 k \gtrsim k \log \frac{p}{k}.$$

Combining with the result in the oracle lower bound, we have the desired.

$$R^* \gtrsim k + k \log \frac{p}{k}.$$

Note: For constant sparsity, i.e., $k = O(1)$, we cannot let each coordinate of θ to be i.i.d. Bernoulli random variable, i.e. $\theta_i \sim \text{Bern}(\frac{k}{p})$. Since in this case, $\|\theta\|_0 \sim \text{Binomial}(p, \frac{k}{p})$, which, for large p and fixed k, is close to Poisson distribution Poi(k). Then with a constant probability, θ is not k sparse.

Remark 19.4. For the case $k = o(p)$, the sharp asymptotics is

$$R^*_{k,p} \ge (2 + o_p(1))k \log \frac{p}{k}.$$

To prove this result, we need to first show that for the case $k = 1$,

$$R^*_{1,p} \ge (2 + o_p(1)) \log p.$$

Next, show that for any k, the minimax risk is lower bounded by the Bayesian risk with the block prior. The block prior is that we divide the p-coordinate into k blocks, and pick one coordinate from each p/k-coordinate uniformly. With this prior, one can show

$$R^*_{k,p} \ge kR^*_{1,p/k} = (2 + o_p(1))k \log \frac{p}{k}.$$

[1] It can be easily verified that the maximum is achieved with the distribution Bern($\frac{m}{p}$)$^{\otimes p}$, write this distribution as $q(w) = (\frac{m}{p})^{w_H(w)}(1 - \frac{m}{p})^{p - w_H(w)}$. For any $p(w)$ satisfies $\mathbb{E}(w_H(W)) = m$, we have $H(W) = -D(p\|q) + \mathbb{E}_p[\log \frac{1}{q(w)}] \le \mathbb{E}_p[\log \frac{1}{q(w)}] = m \log \frac{p}{m} + (p - m) \log \frac{p}{p-m} = ph(\frac{m}{p})$.

Let $\theta \in \Theta = B_0(k) = \{\theta \in \mathbb{R}^p : \|\theta\|_0 \leq k\}$, be a sparse vector. We observe $Y = \theta + Z$, where $Z \sim \mathcal{N}(0, I_p)$. Recall that the last lecture obtains the upper bound on the minimax risk for the problem using the mutual information method as

$$R_n^*(\Theta) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta[\|\hat{\theta} - \theta\|_2^2] \gtrsim k \log\left(\frac{p}{k}\right).$$

This lecture focuses on obtaining the upper bound to the minimax error by analyzing the risk corresponding to (1) the maximum likelihood estimator (2) thresholding estimators.

20.1 Upper bound for denoising by sparsity

In this subsection, we will prove the upper bound for 19.1. We will use the following results on the maxima of Gaussian, proved in our homework.

$$Y = \theta + Z, \quad Z \sim \mathcal{N}(0, I_p),$$

then,

$$\|Z\|_\infty \leq \sqrt{2 \log p} + o_p(1).$$

Given this result, it is natural to consider the following minimization problem,

- ℓ_0-minimization

$$\hat{\theta} = \arg\min \|\theta\|_0, \qquad \text{s.t.} \quad \|y - \theta\|_\infty \leq \tau = \sqrt{2 \log p}.$$

- ℓ_1-minimization

$$\hat{\theta} = \arg\min \|\theta\|_1, \qquad \text{s.t.} \quad \|y - \theta\|_\infty \leq \tau = \sqrt{2 \log p}.$$

However, we can only show the estimator given by these two constraint minimization problem satisfy (see Section 20.3)

$$\sup_{\|\theta\|_0 \leq k} \mathbb{E}_\theta \|\theta - \hat{\theta}\|_2^2 \lesssim k \log p,$$

which does not match the desired result (inside the log). Thus, we will look at the Maximum Likelihood estimator. For Gaussian distribution,

$$P_\theta(y) \propto \exp\left(-\frac{\|y - \theta\|_2^2}{2}\right).$$

Thus, the MLE is equivalent to the minimum distance rule,

$$\hat{\theta}_{\text{MLE}} = \arg\min_{\|\theta\|_0 \leq k} \|y - \theta\|_2^2.$$

We can show for this constrained least square problem,

$$\sup_{\|\theta\|_0 \leq k} \mathbb{E}_\theta \|\theta - \hat{\theta}_{\text{MLE}}\|_2^2 \lesssim k \log \frac{ep}{k}.$$

Proof. Let $h = \hat{\theta}_{\text{MLE}} - \theta$. Thus,

$$\|Z - h\|_2^2 = \|\hat{\theta}_{\text{MLE}} - y\|_2^2 \leq \|\theta - y\|_2^2 = \|z\|_2^2.$$

It is equivalent to

$$\begin{aligned}\|h\|_2^2 &\leq 2\langle h, z\rangle \\ &\leq 2 \sup_{\|u\|_0 \leq 2k} \langle u, z\rangle \\ &= 2\|h\|_2 \sup_{\|u\|_0 \leq 2k, u \in S^{p-1}} \langle u, z\rangle,\end{aligned}$$

where S^{p-1} is the unit sphere in $\mathbb{R}^p$. Let $A = S^{p-1} \cap B_0(2k)$, then $\mathbb{E}\sup_{u\in A}\langle u, z\rangle \triangleq w(A)$ is the Gaussian width defined before. We have shown

$$\mathbb{E}\|h\|_2 \leq 2w(A).$$

Remark 20.1. Estimators, $\hat{\theta}$ are typically efficiently computable for the denoising problem defined above. Further, adaptive estimators that function in the absence of knowledge of k can be defined.

20.2 Maximum Likelihood estimator and risk upper bound

20.2.1 MLE and Basic Inequality

The maximum likelihood estimator for the denoising problem under additive Gaussian noise is given by

$$\hat{\theta}_{\text{MLE}}(y) \in \arg\min_{\tilde{\theta} \in B_0(k)} \|y - \tilde{\theta}\|_2^2. \tag{20.1}$$

We now show that $\forall \theta \in B_0(k)$,

$$\mathbb{E}\|\hat{\theta}_{\text{MLE}} - \theta\|_2^2 \lesssim k \log\left(\frac{p}{k}\right)$$

holds both, under expectation and with high probability. For ease of notation, we shall henceforth refer to the ML estimator as $\hat{\theta}$.

We observe that the ground truth θ is a feasible solution of (20.1). Since the estimator minimizes the ℓ_2 distance, we have

$$\|Z - h\|_2^2 = \|y - \hat{\theta}\|_2^2 \leq \|y - \theta\|_2^2 = \|Z\|_2^2,$$

where $h = \hat{\theta} - \theta$. Thus $\|h\|_0 \leq 2k$. Hence we have

$$\begin{aligned}\|h\|_2^2 \leq 2\langle h, Z\rangle &= 2\|h\|_2 \left\langle Z, \frac{h}{\|h\|_2}\right\rangle \\ &\leq 2\|h\|_2 \sup_{u \in S^{p-1} \cap B_0(2k)} \langle Z, u\rangle \\ \Leftrightarrow \|h\|_2 &\leq 2 \sup_{u \in S^{p-1} \cap B_0(2k)} \langle Z, u\rangle\,.\end{aligned} \tag{20.2}$$

20.2.2 Risk upper bound through Gaussian width

Let $G = S^{p-1} \cap B_0(2k)$. Thus, from (20.2), we have

$$\mathbb{E}\left[\|h\|_2\right] \leq 2\mathbb{E}\left[\sup_{u \in G} \langle u, Z \rangle\right] = 2w(G),$$

where $w(\cdot)$ is the Gaussian width. We know that Sudakov minoration lower bounds the Gaussian width as

$$w(G) \gtrsim \epsilon \sqrt{\log\left(N(G, \|\cdot\|_2, \epsilon)\right)} \asymp \sqrt{k \log\left(\frac{ep}{k}\right)},$$

as long as $\epsilon \asymp 1$. The above result follows from the Gilbert-Varshamov lower bound via packing Hamming spheres.

However we are interesting in an upper bound for the Gaussian width here. One way to obtain this is using Dudley's entropy integral method [Dud67],

$$\begin{aligned} w(G) &\lesssim \int_0^{\text{rad}(G)} \sqrt{\log(N(G, \|\cdot\|_2, \epsilon))} d\epsilon \\ &\lesssim \int_0^1 \sqrt{\log\left(\frac{1}{\epsilon}\right)^k \binom{p}{2k}} d\epsilon \qquad (20.3) \\ &\asymp \sqrt{k \log \frac{pe}{k}}, \end{aligned}$$

where (20.3) follows from the fact that the vectors projected onto the set of support vectors lie on S^{2k-1} and the fact that there are $\binom{p}{2k}$ possible support vector combinations.

20.2.3 Risk upper bound through covering argument

We now provide an alternate proof to show that the upper bound is held with high probability (consequently in expectation). Let J represent a set of indices. Let us partition G as

$$G = \cup_{|J|=2k} G_J = \cup_{|J|=2k} \left\{x \in \mathbb{R}^p : \text{supp}(x) = J, x_J \in S^{2k-1}\right\}.$$

Hence, we have

$$\sup_{u \in G} \langle u, Z \rangle = \max_{|J|=2k} \sup_{u \in G_J} \langle u, Z \rangle = \max_{|J|=2k} \|Z_J\|_2.$$

Fix an index set J such that $|J| = 2k$. Let $\mathcal{U} = \{u_1, \ldots, u_N\}$ be an ϵ-net of G_J. Thus, the set of vectors form a cover of a $2k$ dimensional sphere. Thus,

$$N = N(S^{2k-1}, \|\cdot\|_2, \epsilon) \leq \left(\frac{3}{\epsilon}\right)^{2k}.$$

Now, $\forall u \in G_J$, $\exists i \in [N]$ such that $\|u - u_i\|_2 \leq \epsilon$. Thus, $\exists r \in \sqrt{2}G_J$ such that $u = u_i + r$. Thus we have

$$\sup_{u \in G_J} \langle u, Z \rangle \leq \max_{i \in [N]} \langle u_i, Z \rangle + \sup_{r \in \sqrt{2}G_J} \langle r, Z \rangle.$$

Now, we know that

$$\sup_{r \in \sqrt{2}G_J} \langle r, Z \rangle \leq \sqrt{2}\epsilon \sup_{u \in G_J} \langle u, Z \rangle.$$

Using this, we have

$$\sup_{u\in G_J} \langle u, Z\rangle \lesssim \max_{i\in[N]} \langle u_i, Z\rangle ,$$

when ϵ is an appropriately chosen constant. Here $\langle u_i, Z\rangle \sim \mathcal{N}(0,1)$ as $\|u_i\|_2 = 1$.

Since $\binom{p}{2k}$ choices of index sets are possible, we bound the tail probability using union bound as

$$\begin{aligned}
\mathbb{P}\left[\sup_{u\in G} \langle u, Z\rangle > t\right] &\leq \sum_{|J|=2k} \mathbb{P}\left[\sup_{u\in G_J} \langle u, Z\rangle > t\right] \\
&\leq \sum_{|J|=2k} \sum_{i\in[N]} \mathbb{P}\left[\langle u_i, Z\rangle > t\right] \\
&\leq \binom{p}{2k} \exp(ck) Q(t) \leq \exp\left(2k \log \frac{p}{2k}\right) \exp(ck) \exp\left(-\frac{t^2}{2}\right),
\end{aligned}$$

where the last step follows from bounding the size of the ϵ-net and the Q-function. Thus, for $t \asymp \sqrt{k \log \frac{ep}{k}}$, (scaled by an appropriately large constant), the tail probability is arbitrarily low. Thus, with high probability,

$$\sup_{u\in G} \langle u, Z\rangle \lesssim \sqrt{k \log \frac{pe}{k}}.$$

20.2.4 Risk upper bound using tail bound for χ^2 distribution

As observed earlier,

$$\sup_{u\in G} \langle u, Z\rangle = \max_{|J|=2k} \|Z_J\|_2.$$

Since $Z \sim \mathcal{N}(0, I_p)$, $\|Z_J\|_2^2 \sim \chi^2_{2k}$ for a given J. We first study a few properties of the χ^2 random variable.

Let $L \sim \chi^2_m$. Then, $\mathbb{E}\left[L\right] = m$, $\text{Var}(L) \asymp m$ *i.e.*$\sigma_L \asymp \sqrt{m}$.

Theorem 20.1 ([?]). *If $L \sim \chi^2_m$, then*

$$\begin{aligned}
\mathbb{P}\left[L - m > S\sqrt{m} + S^2\right] &\leq \exp\left(\frac{-S^2}{2}\right) \\
\mathbb{P}\left[L - m < -S\sqrt{m}\right] &\leq \exp\left(\frac{-S^2}{2}\right).
\end{aligned}$$

Now, applying the above concentration inequality for $m = 2k$, $S = \sqrt{ck \log \frac{p}{k}}$, we have

$$\mathbb{P}\left[\|Z_J\|_2^2 > 2k + k\sqrt{c \log \frac{p}{k}} + ck \log \frac{p}{k}\right] \leq \mathbb{P}\left[\|Z_J\|_2^2 > k \log \frac{pe}{k}\right] \qquad \leq \exp\left(\frac{-ck \log \frac{p}{k}}{2}\right).$$

Thus, with high probability,

$$\sup_{u\in G} \langle u, Z\rangle \lesssim \sqrt{k \log \frac{pe}{k}}.$$

20.3 Thresholding schemes and Risk upper bounds

20.3.1 Hard and Soft thresholding

For the denoising problem defined above, the *hard thresholding* estimate corresponding to the threshold τ is given by

$$\hat{\theta}_{\mathrm{HT}}(y)_i = \begin{cases} y_i, & \text{if } |y_i| > \tau \\ 0, & \text{if } |y_i| \leq \tau \end{cases}$$

Similarly, the *soft thresholding* estimate is given by

$$\hat{\theta}_{\mathrm{ST}}(y)_i = \begin{cases} y_i - \tau, & \text{if } y_i > \tau \\ y_i + \tau, & \text{if } y_i < -\tau \\ 0, & \text{if } |y_i| \leq \tau \end{cases}$$

The HT estimate is not continuous and the corresponding risk function does not vary monotonically. On the other hand, the soft thresholding avoids both these issues.

The hard and soft thresholding estimators can alternatively be written in the form of penalized objective functions. Consider the problem defined as follows:

$$\theta'(y) = \operatorname*{arg\,min}_{\tilde{\theta} \in \mathbb{R}^p} \|y - \tilde{\theta}\|_2^2 + \lambda \|\tilde{\theta}\|_0.$$

Then, for appropriately chosen penalty factor λ, $\theta'(y) = \hat{\theta}_{\mathrm{HT}}(y)$. Similarly, for the problem

$$\theta'(y) = \operatorname*{arg\,min}_{\tilde{\theta} \in \mathbb{R}^p} \|y - \tilde{\theta}\|_2^2 + \lambda \|\tilde{\theta}\|_1,$$

for appropriately chosen λ, $\theta'(y) = \hat{\theta}_{ST}(y)$.

Note: Under such thresholding schemes, we may not necessarily obtain a k-sparse vector as we desire. However, we shall ignore this fact as we are interested in only the risk upper bounds.

20.3.2 ℓ_∞-constrained procedure

Consider the following ℓ_∞-constrained formulation of the problem

$$\hat{\theta}(y) \in \operatorname*{arg\,min}_{\tilde{\theta} \in \mathbb{R}^p : \|y - \tilde{\theta}\|_\infty \leq \tau} \|\tilde{\theta}\|_0.$$

We observe that the hard thresholding estimate is a feasible solution to the above problem. (The set that minimizes the above objective function is in reality a continuum of points.) The constraint of interest is that $\|y - \tilde{\theta}\|_\infty \leq \tau$. Thus, setting $\tilde{\theta}_i = 0$ when $|y_i| \leq \tau$ and $\tilde{\theta}_i = y_i$ when $|y_i| > \tau$ satisfies the constraint. Further, this estimate also minimizes the ℓ_0 norm and thus $\hat{\theta}(y)$ is a feasible solution.

Theorem 20.2. *For all $\theta \in B_0(k)$, $\hat{\theta}$ a feasible solution to the above problem, for $\tau = \sqrt{2 \log p}$, with high probability,*

$$\|\hat{\theta} - \theta\|_2^2 \leq 16 k \log p.$$

Proof. We shall decompose the proof into three steps.

Step 1: Set τ to ensure feasibility of ground truth.

Since $Y = \theta + Z$,

$$\|y - \theta\|_\infty = \|Z\|_\infty \lesssim \sqrt{2 \log p} \quad \text{whp.}$$

Thus we observe that the ground truth is feasible high probability.

Step 2: Analyze structure of error.

The error is given by $h = \hat{\theta} - \theta$. Since θ is a feasible solution,

$$\|\hat{\theta}\|_0 \leq \|\theta\|_0 \leq k.$$

Thus, $\|h\|_0 \leq 2k$.

Step 3: Bound ℓ_2 norm.

$$\begin{aligned}
\|h\|_2^2 &\leq \|h\|_\infty^2 \|h\|_0 \\
&\leq 2k\|\hat{\theta} - \theta\|_\infty^2 \\
&\leq 2k(\|\hat{\theta} - y\|_\infty + \|y - \theta\|_\infty)^2 \qquad (20.4) \\
&\leq 8k\tau^2 = 16k \log p,
\end{aligned}$$

where (20.4) follows from the triangle inequality. We note that all the above statements hold with high probability following the statement of feasibility. □

Remark 20.2. How to convert the high-probability bound in Theorem 20.2 to one in expectation, i.e., $\mathbb{E}\|\hat{\theta} - \theta\|_2^2 \lesssim k \log p$? Note that we need to account for the rare-event that that $\|Z\|_\infty$ is large and we hope to prove this *without* assuming that θ is bounded.

Similarly, consider the problem

$$\hat{\theta}(y) \in \operatorname*{arg\,min}_{\tilde{\theta} \in \mathbb{R}^p : \|y - \tilde{\theta}\|_\infty \leq \tau} \|\tilde{\theta}\|_1.$$

We observe here that for any $\tilde{\theta}$ satisfying the constraint, $\|\tilde{\theta}\|_1 \geq \sum_{i=1}^p (|y| - \tau)\mathbf{1}\left\{|y| > \tau\right\}$. The soft thresholding estimate satisfies the above bound and the constraint and is thus a feasible solution to the problem.

Theorem 20.3. *For all $\theta \in B_0(k)$, $\hat{\theta}$ a feasible solution to the above problem, for $\tau = \sqrt{2 \log p}$, with high probability,*

$$\|\hat{\theta} - \theta\|_2^2 \leq 32k \log p.$$

Proof. We proceed in similar fashion to the earlier proof.

Step 1: Set τ to ensure feasibility of ground truth.

Since $Y = \theta + Z$,

$$\|y - \theta\|_\infty = \|Z\|_\infty \lesssim \sqrt{2 \log p} \quad \text{whp.}$$

Thus we observe that the ground truth is feasible with high probability.

Step 2: Analyze structure of error.

The error is given by $h = \hat{\theta} - \theta$. Thus $\|h\|_\infty \leq 2\tau$. Let $J = \operatorname{supp}(\theta)$. Define the cone

$$C_J = \left\{x \in \mathbb{R}^p : \|x_{J^c}\|_1 \leq \|x_J\|_1\right\}. \qquad (20.5)$$

We now have

$$\|h_J\|_1 - \|h_{J^c}\|_1 = \sum_{i \in J} |\hat{\theta}_i - \theta_i| - \sum_{i \in J^c} |\hat{\theta}_i| \geq \|\theta\|_1 - \|\hat{\theta}\|_1 \geq 0,$$

which follows from the triangle inequality and the feasibility of θ. Thus $h \in C_J$.

Step 3: Bound ℓ_2 norm.

$$\begin{aligned}
\|h\|_2^2 &\leq \|h\|_1\|h\|_\infty && (20.6)\\
&\leq 4\tau\|h_J\|_1 \\
&\leq 4\tau\sqrt{k}\|h_J\|_2 && (20.7)\\
&\leq 4\tau\sqrt{k}\|h\|_2 \\
\Leftrightarrow \|h\|_2^2 &\leq 32k\log p,
\end{aligned}$$

where (20.6) and (20.7) follow from Holder's inequality and Cauchy-Schwarz inequality respectively. □

Remark 20.3 (Approximate Sparsity)**.** Let J be a set of indices of size k. Let $h \in C_J = \{x \in \mathbb{R}^p : \|x_{J^c}\|_1 \leq \|x_J\|_1\}$. Consider the set of k largest elements in h_{J^c} indexed by the set K. Then,

$$\|h_{(J\cup K)^c}\|_2^2 \geq \frac{1}{2}\|h\|_2^2.$$

Proof. For every element, we have

$$|h_{J^c}^{(i)}| \leq \frac{1}{i}\|h_{J^c}\|_1.$$

Thus,

$$\begin{aligned}
\|h_{J^c}\|_2^2 \leq \sum_{i=k+1}^{p-k} |h_{J^c}^{(i)}|^2 &\leq \sum_{i=k+1}^{p-k} \frac{1}{i^2}\|h_{J^c}\|_1^2 \\
&\leq \frac{1}{k}k\|h_{J^c}\|_2^2 \leq \|h_{J\cup K}\|_2^2,
\end{aligned}$$

which follows from Cauchy-Schwarz inequality and the fact that $h \in C_J$. □

Remark 20.4.

1. When the vector is sufficiently sparse, specifically $k = o(p)$,

   $$R^* \leq (2 + o(1))k\log\frac{p}{k}.$$

   Further, the bound can be achieved in the adaptive case too.

2. If $k = \Theta(p)$, i.e, $\frac{k}{p} \to \alpha \in (0, 1]$ as $p \to \infty$, then,

   $$R^* = p\left(\beta(\alpha) + o(1)\right),$$

   where $\beta(\alpha)$ is a constant dependent on α.

§ 21. Minimax rates for sparse linear regression

In the last lecture we analyzed the k-sparse Gaussian location model in high dimension (the denoising problem) and proved minimax rate for estimating the location parameter. In this lecture we extend the earlier ideas to sparse linear regression in high dimension. We prove a minimax lower bound and then obtain upper bounds on the risk of a few procedures.

21.1 Problem setup: Sparse linear regression

The sparse linear regression model is

$$Y_{n\times 1} = X_{n\times p}\theta_{p\times 1} + Z, \quad Z \sim \mathcal{N}(0, I_n), \tag{21.1}$$

where $X \in \mathbb{R}^{n\times p}$ is the design matrix and $\theta \in \mathbb{R}^p$ is an unknown k-sparse parameter vector. In this lecture we are concerned with the case when $n << p$ but $n \geq k$, i.e., we have more predictors in the design matrix than we have samples.

Interpretation: Y is a noisy linear combination of the columns of the design matrix X. The goal here is to estimate θ, given Y and X. Note that the system has more unknowns than the number of equations and hence is indeterminate even without the noise. Estimation is made possible due to the k-sparsity structure.

Note: We consider here random design matrices only. More precisely we have

$$X_{ij} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 1/n),$$

so that the columns have roughly unit norm.

The next theorem proves a lower bound on the minimax risk for estimating θ in the k-sparse regression model.

Theorem 21.1. *The minimax risk for estimating θ in the model defined by Equation (21.1) is lower bounded by*

$$R^* = R^*(p, k, n) = \inf_{\hat{\theta}} \sup_{\|\theta\|_0 \leq k} \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2 \gtrsim k \log \frac{ep}{k}, \quad \forall n.$$

Proof. Note that (X_i, Y_i)'s are i.i.d sampled from a distribution P_θ. We show that the KL-diameter for two different θ is exactly same as that in p-dimensional gaussian location model, i.e.,

$$D(P_{\theta_0} \| P_{\theta_1}) = \frac{1}{2}\|\theta_0 - \theta_1\|^2.$$

We derive the result as follows,

$$\begin{aligned}
D(P_{X_i,Y_i|\theta_0}\|P_{X_i,Y_i|\theta_1}) &= \mathbb{E}_{X_i}[D(P_{Y_i|X_i,\theta_0}\|P_{Y_i|X_i,\theta_1})] \\
&= \mathbb{E}_{X_i}[D(\mathcal{N}(\langle X_i,\theta_0\rangle,1)\|\mathcal{N}(\langle X_i,\theta_1\rangle,1))] \\
&= \mathbb{E}_{X_i}[\frac{1}{2}\langle X_i,\theta_0-\theta_1\rangle^2] \\
&= \frac{1}{2}\mathbb{E}_{X_i}[(\theta_0-\theta_1)'X_i'X_i(\theta_0-\theta_1)] \\
&= \frac{1}{2}(\theta_0-\theta_1)'\mathbb{E}[X_i'X_i](\theta_0-\theta_1) \\
&= \frac{1}{2n}\|\theta_0-\theta_1\|^2.
\end{aligned}$$

Hence,

$$D(P_{\theta_0}\|P_{\theta_1}) = D(P^{\otimes n}_{X_i,Y_i|\theta_0}\|P^{\otimes n}_{X_i,Y_i|\theta_1}) = \frac{1}{2}\|\theta_0-\theta_1\|^2.$$

The result then follows from the analysis of denoising (regression with identity design) from previous lecture. □

From the previous discussion we have, $R^* \gtrsim k\log\frac{ep}{k}$ for any n, which is identical to the minimax rate of the denoising problem for any n. This is reasonable, because even with full observation $n \gtrsim p$, which roughly corresponds to the denoising problem we cannot beat this rate. Surprisingly, as long as $n \gtrsim k\log\frac{ep}{k}$, the denoising rate is attainable and we have $R^* \asymp k\log\frac{ep}{k}$, achieved by, e.g., the maximum likelihood estimator. This is proved in Section 21.2.

Note that MLE is computationally expensive. A computable alternative procedure is the Dantzig selector [CT07]. As analyzed in Section 21.3 It is guaranteed to achieve the rate $R^* \lesssim k\log p$ as long as $n \geq k\log\frac{ep}{k}$. This falls slightly sort of the optimal rate. However unlike MLE, the procedure is completely adaptive and can be cast as a linear programming problem. More recently a procedure called SLOPE [SC15] has been proposed which achieves the optimal rate. In particular its risk coincides with the minimax rate with sharp constant, namely, $R^* \leq (2+o(1))k\log\frac{ep}{k}$ as $p\to\infty$ and $k=o(p)$, provided that $n \gtrsim k\log\frac{ep}{k}$.

21.2 Analysis of MLE

The MLE in this case is defined to be any solution (may or may not be unique) to the constrained least-square, which is a combinatorial optimization problem:

$$\hat{\theta}_{\text{MLE}} \in \arg\min_{\|\theta\|_0\leq k} \|Y-X\theta\|_2^2. \tag{21.2}$$

Unfortunately the optimization problem is NP-hard in the worst case and we only know how to solve it through exhaustive search.

Theorem 21.2. *Whenever $n \geq Ck\log\frac{ep}{k}$ for some sufficiently large constant C, $\forall\theta\in B_0(k)$.*

$$\|\hat{\theta}_{\text{MLE}}-\theta\|_2^2 \lesssim k\log\frac{ep}{k}, \tag{21.3}$$

$$\|X(\hat{\theta}_{\text{MLE}}-\theta)\|_2^2 \lesssim k\log\frac{ep}{k}, \tag{21.4}$$

hold with high probability.

Proof. Since $\hat{\theta}$ is a minimizer, we have

$$\|Y - X\hat{\theta}\|_2^2 \leq \|Y - X\theta\|_2^2 = \|Z\|_2^2.$$

On the left hand side we have,

$$\|Y - X\hat{\theta}\|_2^2 = \|Y - X\theta + X\theta - X\hat{\theta}\|_2^2 = \|Z - Xh\|_2^2,$$

where $h = \hat{\theta} - \theta$. Hence we have,

$$\|Z - Xh\|_2^2 \leq \|Z\|_2^2,$$

which leads to the basic inequality

$$\begin{aligned} \|Xh\|_2^2 &\leq 2\langle Z, Xh\rangle \\ &= 2Z'Xh \\ &\leq 2\|h\|_2 \sup_{u \in S^{p-1} \cap B_0(2k)} Z'Xu. \end{aligned}$$

[TODO: alternatively, we can do

$$\|Xh\|_2^2 \leq 2\langle Z, Xh\rangle = 2\|Xh\|_2 \left\langle Z, \frac{Xh}{\|Xh\|_2} \right\rangle \leq 2\|Xh\|_2 \sup_{v \in S^{p-1} \cap XB_0(2k)} Z'v.$$

and hope to use union bound and metric entropy to show the supremum is also $\sqrt{\log\binom{p}{k}}$. Then we get $\|Xh\|_2^2 \lesssim k \log \frac{ep}{k}$ without assumption on X.]

Note that the left hand side is not the estimation error, instead it is the prediction error $\|Xh\|_2^2 = \|X\hat{\theta} - X\theta\|_2^2$. Hence to conclude both (21.3) and (21.4) from the basic inequality, it suffices to show

(a) $\|h\|_2 \lesssim \|Xh\|_2$, (Restricted isometry property)

(b) $\sup_{u \in S^{p-1} \cap B_0(2k)} Z'Xu \lesssim \sqrt{k \log \frac{ep}{k}}$, with high probability.

We first prove (b).

$$\sup_{u \in S^{p-1} \cap B_0(2k)} Z'xu = \|Z\| w(G) \lesssim \sqrt{k \log \frac{ep}{k}},$$

where $w(G)$ is the Gaussian width of the set $S^{p-1} \cap B_0(2k)$ and from last lecture we know, $w(G) \lesssim \sqrt{k \log \frac{ep}{k}}$.

For (a) we will show that

$$\inf_{\|h\|_0 \leq k} \frac{\|Xh\|_2}{\|h\|_2} \gtrsim c \quad \text{if } n \gtrsim k \log \frac{ep}{k},$$

where c is a constant. First note that,

$$\inf_{u \neq 0} \frac{\|Au\|_2}{\|u\|_2} = \sigma_{\min}(A).$$

For any feasible h, $Xh = X_J h_J$, where $J = \text{supp}(h)$ is the support of h and X_J is the $n \times k$ matrix whose columns are the columns of X that corresponds to the rows in the support J. Then we have

$$\inf_{\|h\|_0 \leq k} \frac{\|Xh\|_2}{\|h\|_2} = \min_{|J| \leq k} \sigma_{\min}(X_J).$$

For a fixed J, $\sigma_{\min}(X_J)$ concentrates to $1-\sqrt{\frac{k}{n}}$. Hence an union bound gives,

$$\mathbb{P}\left[\min_{|J|\leq k}\sigma_{\min}(X_J)<t\right]\leq\binom{p}{k}\mathbb{P}[\sigma_{\min}(X_{[k]})<t].$$

Using the tail bound

$$\mathbb{P}\left[\sigma_{\min}(X_{[k]})<1-\sqrt{\frac{k}{n}}-\frac{t}{\sqrt{n}}\right]\leq\exp(-t^2/2),$$

and choosing $t=4k\log\frac{ep}{k}$ and $n=100k\log\frac{ep}{k}$, we have $1-\sqrt{\frac{k}{n}}-\frac{t}{\sqrt{n}}\geq 0.5$ and consequently, $\mathbb{P}[\min_{|J|\leq k}\sigma_{\min}(X_J)<0.5]\to 0$. This completes the proof. □

21.3 Dantzig selector

The Dantzig selector can written as the following optimization problem,

$$\min\|\theta\|_1,\quad \text{s.t } \|X'(Y-X\theta)\|_\infty\leq\tau. \tag{21.5}$$

This optimization problem can be efficiently solved as a linear programming. Another computable procedure for k-sparse regression is the LASSO, which can be written as the following optimization problem

$$\min\|Y-X\theta\|_2^2+\lambda\|\theta\|_1,\quad \theta\in\mathbb{R}^n. \tag{21.6}$$

Note that one of the reasons X' is added to the constraint in the Dantzig selector in (21.7) is to make the solution rotation-invariant. Precisely, if $U\in O(n)$ be an $n\times n$ orthogonal rotation matrix, then $UY=UX\theta+UZ$. Note that in this case, $\hat{\theta}(X,Y)=\hat{\theta}(UX,UY)$.

Theorem 21.3. *Let the Dantzig selector $\hat{\theta}_{\text{DS}}$ denote a minimizer of (21.7). As long as $n\geq Ck\log\frac{ep}{k}$ for some sufficiently large constant C, for any $\theta\in B_0(k)$,*

$$\|\hat{\theta}_{\text{DS}}-\theta\|_2^2\lesssim k\log p, \tag{21.7}$$

$$\|X(\hat{\theta}_{\text{DS}}-\theta)\|_2^2\lesssim k\log p, \tag{21.8}$$

hold with high probability.

Proof. Paralleling that of Theorem 20.3 for the denoising problem, the proof is divided in three steps.

Step 1: Set τ to guarantee the ground truth is feasible. We choose

$$\|X'Z\|_\infty\leq\tau=\sqrt{2\log p},$$

so that θ is feasible.

Step 2: Structure of the error $h=\hat{\theta}-\theta$. This step is identical to the proof of Theorem 20.3 because the objective function is the same (ℓ_1 norm). Let J be the support of θ. Recall the cone (20.5), that is,

$$C_J\triangleq\{h:\|h_{J^c}\|_1\leq\|h_J\|_1\}. \tag{21.9}$$

Since $\|\hat{\theta}\|_1\leq\|\theta\|_1$, we have $h\in C_J$. Furthermore, we have

$$\|X'Xh\|_\infty\leq 2\tau$$

Indeed, due to feasility of θ and θ_{DS},

$$\begin{aligned}\|X'Xh\|_\infty &= \|X'X(\hat\theta-\theta)\|_\infty\\ &= \|X'(Y-X\theta)-X'(Y-X\hat\theta)\|_\infty\\ &\le \|X'(Y-X\theta)\|_\infty + \|X'(Y-X\hat\theta)\|_\infty\\ &\le 2\tau.\end{aligned}$$

Step 3: Bound The risk. We have,

$$\begin{aligned}\|Xh\|_2^2 &= \langle Xh, Xh\rangle\\ &= h'X'Xh\\ &= \big\langle X'Xh, h\big\rangle\\ &\le \|X'Xh\|_\infty\|h\|_1 \quad \text{(Hölder)}\\ &\le 2\tau 2\|h_J\|_1\\ &\le 4\sqrt{k}\tau\|h_J\|_2 \quad \text{(Cauchy-Schwartz)}\\ &\le 4\sqrt{k}\tau\|h\|_2.\end{aligned}$$

Now we need to show one last thing to complete the proof, which play the same role of RIP in the analysis of MLE in Theorem 21.2:

$$\|h\|_2^2 \lesssim \|Xh\|_2^2 \quad \text{w.h.p } \forall h\in C_J, \tag{21.10}$$

which shows that X preserves the distance on the cone C_J (instead of all sparse vectors). To prove (21.10) we use the special feature of the cone C_J defined in (21.9), that for any $h\in C_J$, half of the energy of h is on $2k$ coordinates, i.e. h is almost $2k$-sparse.

For notational simplcity, suppose the entries h are ordered in decreasing magnitudes: Let $J=\mathrm{supp}(\theta)$ and h_J corresponds to the first k coordinates. We divide the remaining h after first block into blocks of size k and name the blocks $K_1, K_2, \ldots$ and the vectors $h_1, h_2, \ldots$, such that $h_{K_i}=h_i$.

Let $a = h_{J\cup K_1} = h_J + h_1$. By construction, it has more than 1/2 of the energy, i.e.,

$$\|a\|_2^2 \ge \frac{1}{2}\|h\|_2^2$$

and define,

$$b \triangleq h_{(J\cup K_1)^c} = \sum_{i\ge 2} h_i.$$

Then

$$\begin{aligned}\|Xh\|_2^2 &= \|Xa+Xb\|_2^2\\ &\ge \|Xa\|_2^2 + 2\langle Xa, Xb\rangle \end{aligned}\tag{21.11}$$

Since X satisfies the restricted isometry property, for $n \ge ck\log\frac{ep}{k}$, there exists $c_1(c)$ with $c_1\to 1$ if $c\to\infty$, such that

$$\|Xa\|_2^2 \ge c_1\|a\|_2^2 \ge \frac{c_1}{2}\|h\|_2^2.$$

Now we need to just show that the cross term $\langle Xa, Xb\rangle$ is small in magnitude. For this we use the following restricted decorrelation lemma,

Lemma 21.1. *Let $n \geq ck \log \frac{ep}{k}$. Then, with high probability, for all $u, v \in B_0(2k)$ we have,*

$$|\langle Xa, Xb\rangle| \leq c_2\|u\|\|v\|,$$

where $c_2 = c_2(c)$ and $c_2 \to 0$ if $c \to \infty$.

Then we have,

$$\begin{aligned}
\langle Xa, Xb\rangle &= \sum_{j\geq 2}\langle Xa, Xh_j\rangle \\
&\leq \sum_{j\geq 2}\langle Xa, hX_j\rangle \\
&\leq \sum_{j\geq 2}\|a\|_2\|h_j\|_2 \quad \text{(Lemma 21.1)} \\
&\leq c_2\|h\|_2\sum_{j\geq 2}\sqrt{k}\|h_j\|_\infty \\
&\leq c_2\|h\|_2\sum_{j\geq 2}\sqrt{k}\frac{\|h_{j-1}\|_1}{k} \quad \text{(By ordering)} \\
&\leq \frac{c_2}{\sqrt{k}}\|h\|_2(\sum_{j\geq 2}\|h_{j-1}\|_1) \\
&\leq \frac{c_2}{\sqrt{k}}\|h\|_2\|h_{J^c}\|_1 \quad \text{(Property of cone)} \\
&\leq \frac{c_2}{\sqrt{k}}\|h\|_2\|h_J\|_1 \\
&\leq \frac{c_2}{\sqrt{k}}\|h\|_2^2.
\end{aligned}$$

Reverting back to (21.11) we have,

$$\|Xh\|_2^2 \geq (c_1/2 - c_2)\|h\|_2^2 \gtrsim \|h\|_2^2.$$

This completes the proof. □

Remark 21.1 (Adaptivity issues)**.** Note that the Dantzig selector procedure is adaptive to k, but not to σ. To see this consider the following high dimensional k-sparse regression problem,

$$Y = X\theta + Z, \quad Z \sim \mathcal{N}(0, \sigma^2 I_n).$$

If σ is known then we can set $\tau = \sigma\sqrt{2\log p}$, but typically σ is not known.

A similar problem arises with LASSO as well. In (21.6), if σ is known then the optimal $\lambda = 2\sigma\sqrt{\log p}$, but if σ is unknown then λ is a tuning parameter. As a remedy for this another procedure called square root LASSO was proposed which can be written as the following optimization problem,

$$\min \|Y - X\theta\|_2 + \lambda\|\theta\|_1, \quad \theta \in \mathbb{R}^n.$$

The optimal $\lambda = \sqrt{\log p}$ even when σ is unknown. However the downside is that this optimization problem is not easy to solve.

Part V

Functional estimation and composite hypothesis testing

§ 22. Functional estimation & testing

In this chapter, we will be interested in analyzing the sample complexity and minimax rates for the functional estimation problem introduced earlier. We will also consider the hypothesis testing paradigm so that we can utilize important tools such as LeCam's method for proving bounds and analyzing the minimax rate.

Formally, the setting is as follows: Assume that θ is an unknown parameter in the parameter space Θ and θ generates the data X according to the distribution P_θ. For a fixed real valued functional T on θ, i.e, $T : \Theta \to \mathbb{R}$, we wish to estimate T based on the observations through the estimator $\hat{T}(X)$.

In the estimation paradigm, the setting can be pictorially represented as follows:

$$\begin{tikzcd}
\theta \arrow[r, "P_\theta"] \arrow[d] & X \arrow[d] \\
T = T(\theta) & \hat{T}(X)
\end{tikzcd}$$

In the hypothesis testing paradigm, we are interested in determining the class of parameters that gave rise to the observations. Formally, given $t_0, t_1 \in \mathbb{R}$, the problem is formulated as:

$$\begin{aligned}
H_0 &: T \leq t_0, \\
H_1 &: T \geq t_1.
\end{aligned}$$

Equivalently, we can also think of the above hypothesis testing as a composite hypothesis testing of θ as follows:

$$\begin{aligned}
H_0 &: \theta \in \Theta_0 = \{\theta : T(\theta) \leq t_0\}, \\
H_1 &: \theta \in \Theta_1 = \{\theta : T(\theta) \geq t_1\}.
\end{aligned}$$

Example 22.1. Consider the Gaussian location model $X \sim \mathcal{N}\left(\theta, I_p\right), \theta \in \mathbb{R}^p$. Let $T : \mathbb{R}^p \to \mathbb{R}$ be given by $T(\theta) = \|\theta\|$. A possible test would be determining if $\|\theta\|$ is too small or too large given some thresholds. Specifically,

$$\begin{aligned}
H_0 &: \|\theta\| \leq 1, \\
H_1 &: \|\theta\| \geq 3.
\end{aligned}$$

22.1 Lower bounds on minimax risk for functional estimation

Since T takes only real values where as Θ can be arbitrary high dimensional space, such as Euclidean space $\mathbb{R}^p$, T can be thought of as a low dimensional representation of the parameter space Θ. Thus

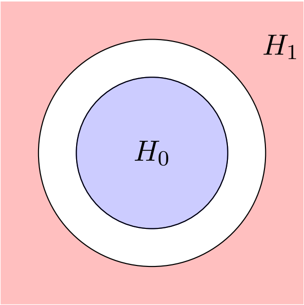


Figure 22.1: $\|\theta\| \leq 1$ vs. $\|\theta\| \geq 3$.

it suggests if we can employ techniques such as LeCam's two-point argument to prove lower bounds on the minimax risk estimation of $T(\theta)$.

To this end, recall the LeCam's two-point method discussed in Lecture 9. The key idea in the two-point argument is the fact that if we can estimate a parameter, we can also test it efficiently. We reduced the task of estimation to that of the binary hypothesis testing, i.e, for fixed $\theta_0, \theta_1 \in \Theta$,

$$\begin{aligned} H_0 &: \theta = \theta_0, \\ H_1 &: \theta = \theta_1 \end{aligned}$$

to derive lower bounds on the minimax risk R^* (under quadratic loss). In particular, we showed that

$$R^* = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[(\theta - \hat{\theta})^2 \right] \gtrsim \sup_{\theta_0 \neq \theta_1} \|\theta_0 - \theta_1\|^2 \left(1 - d_{\mathrm{TV}}\left(P_{\theta_0}, P_{\theta_1}\right)\right).$$

In a similar vein, we can consider the following binary hypothesis testing for T (for some $t_0, t_1 \in \mathbb{R}$)

$$\begin{aligned} H_0 &: T \leq t_0, \\ H_1 &: T \geq t_1 \end{aligned} \tag{22.1}$$

to obtain a lower bound on the minimax risk for estimation of T. Specifically, if $\pi_0 \in \mathcal{M}(\Theta_0), \pi_1 \in \mathcal{M}(\Theta_1)$ are any two priors on Θ_0 and Θ_1 respectively, we obtain

$$R^* = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[(T(\theta) - \hat{T}(X))^2 \right] \gtrsim (t_0 - t_1)^2 \left(1 - d_{\mathrm{TV}}\left(P_{\pi_0}, P_{\pi_1}\right)\right). \tag{22.2}$$

Thus our task reduces to finding two priors π_0, π_1 so that the lower bound in (22.2) would be maximized, or roughly speaking, we want to pick two priors that would ensure maximum confusion in testing of the two hypothesis.

Now we study a closely related concept of sample complexity for the analysis of the same.

22.2 Estimation of $\|\theta\|$ in GLM

Our aim is to prove that for the p-dimensional GLM where the data $X \sim \mathcal{N}(\theta, \frac{1}{n} I_p), \theta \in \mathbb{R}^p$, the minimax risk R^* for the estimation of $T(\theta) = \|\theta\|$ obeys $R^* \asymp \frac{\sqrt{p}}{n}$.

First we give a preview of this result and other estimation tasks in terms of a closely related concept: sample complexity. The proofs of these results is similar to that of those concerning average and minimax risk. Recall from Lecture 3, where we defined the sample complexity to be the minimum number of samples required to achieve a prescribed estimation error, either in expectation or in probability with high confidence.

Estimation tasks	Sample complexity
$T(\theta) = \theta$	$n^* \asymp p$
$T(\theta) = \theta_1$	$n^* \asymp 1$
$T(\theta) = \theta_{\max}$	$n^* \asymp \log p$
$T(\theta) = \|\theta\|_2$	$n^* \asymp \sqrt{p}$

One important observation is the fact that to estimate $\|\theta\|_2$, one can employ a plug-in estimator where we first estimate θ and then compute $\|\theta\|_2$. However, this naive procedure requires as many samples as that are required to estimate θ. Instead, we can perform much better by using only $\sqrt{p}$ samples to estimate $\|\theta\|_2$.

Instead of the setting in (22.1), where both the hypotheses are composite, we consider a simplified testing scenario where only one of the hypotheses is composite and hence more tractable.

$$\begin{aligned} H_0 &: \theta = 0, \\ H_1 &: \|\theta\|_2 \geq \rho \end{aligned}$$

Pictorially,

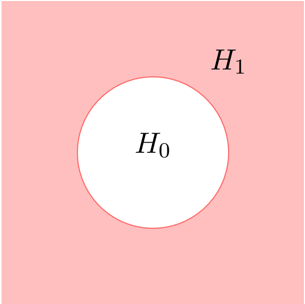


Figure 22.2: $\theta = 0$ vs. $\|\theta\| \geq \rho$.

We can further simplify this to the observation of one sample case making use of the fact that, to incur a minimum probability of error (say 0.01), $\max \rho$ for n-sample GLM $= \frac{\max \rho \text{ for 1 sample GLM}}{\sqrt{n}}$. Thus our model reduces to

$$X \sim \mathcal{N}(\theta, I_p), \theta \in \mathbb{R}^p.$$

22.2.1 Draw backs of two-point argument

A naive application of LeCam's two-point argument for the estimation of $\|\theta\|$ through the binary hypothesis testing of $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$ would yield

$$\begin{aligned} R^* &\gtrsim (\|\theta_0\| - \|\theta_1\|)^2 \left(1 - d_{\text{TV}}\left(\mathcal{N}(\theta_0, I_p), \mathcal{N}(\theta_1, I_p)\right)\right) \\ &= 2\left(\|\theta_0\| - \|\theta_1\|\right)^2 Q\left(\frac{\|\theta_0 - \theta_1\|}{2}\right). \end{aligned}$$

Since $\|\theta_0\| - \|\theta_1\| \leq \|\theta_0 - \theta_1\|$ by triangle inequality and $\sup \|\theta_0 - \theta_1\|_2 Q\left(\frac{\|\theta_0-\theta_1\|_2}{2}\right) \asymp 1$, we see that this approach does not yield any useful lower bound. Thus we lose the much needed dependence of the minimax risk R^* on the dimension of our data, p.

Forgoing the two-point approach wherein we assume uniform distribution on two fixed parameters θ_0, θ_1, we want to choose a prior π supported on $\{\theta : \|\theta\|_2 \geq t\}$ such that the total variation distance $d_{\mathrm{TV}}(P_0, P_\pi)$ is bounded away from 1. In other words, we want to choose a prior π so that P_π closely resembles P_0 in the sense that the probability of error for testing is bounded away from 0. Note that P_π denotes the distribution on the data X which is given by $P_\pi = \pi * \mathcal{N}(0, I_p)$ whereas $P_0 = \mathcal{N}(0, I_p)$.

Recall from Lecture 5, the following chain of inequalities for KL divergence, χ^2 distance and total variation obtained using the concept of joint range. We have

$$\chi^2(P\|Q) \geq \log\left(\chi^2(P\|Q) + 1\right) \geq D(P\|Q) \geq d_{\mathrm{TV}}(P,Q) \log \frac{1 + d_{\mathrm{TV}}(P,Q)}{1 - d_{\mathrm{TV}}(P,Q)}$$

for any two distributions P, Q. This relation suggests that a sufficient condition to ensure $d_{\mathrm{TV}}(P_0, P_\pi)$ to be bounded away from 1, or equivalently $1 - d_{\mathrm{TV}}(P_0, P_\pi) \gtrsim 0$, is to make $\chi^2(P_0, P_\pi) \lesssim 1$. In this regard, we need the following lemma which gives an alternative characterization of χ^2-distance.

Lemma 22.1 (Ingster-Suslina method). *Let Θ be a parameter space and for each $\theta \in \Theta$, let P_θ be a family of probability distributions on a measure space $\mathcal{X}$ and let Q also be a distribution on $\mathcal{X}$. Then*

$$\chi^2(P_\pi\|Q) = \mathbb{E}[G(\theta, \tilde{\theta})] - 1,$$

where $\theta, \tilde{\theta} \overset{i.i.d.}{\sim} \pi$, $G(\theta, \tilde{\theta}) = \int \frac{P_\theta P_{\tilde{\theta}}}{Q}$ and $P_\pi = \int P_\theta \pi(d\theta)$ is the mixture.

Proof. For any two distributions P and Q, we have

$$\chi^2(P\|Q) = \int \frac{(P-Q)^2}{Q} = \mathrm{Var}_Q\left(\frac{P}{Q}\right) = \mathbb{E}_Q\left(\frac{P}{Q}\right)^2 - 1 = \int \frac{P^2}{Q} - 1.$$

Thus, $\chi^2(P_\pi\|Q) = \int \frac{P_\pi^2}{Q} - 1$ and

$$\begin{aligned}\int \frac{P_\pi^2}{Q} = \int \frac{P_\pi P_\pi}{Q} &= \int \frac{\int P_\theta(x)\pi(d\theta) \int P_{\tilde{\theta}}(x)\pi(d\tilde{\theta})}{Q(x)} \mu(dx) \\ &\overset{\text{Fubini}}{=} \int\int \pi(d\theta)\pi(d\tilde{\theta}) \int \frac{P_\theta P_{\tilde{\theta}}}{Q}\mu(dx) \\ &= \mathbb{E}[G(\theta, \tilde{\theta})].\end{aligned}$$

□

In the case of GLM, $\mathbb{E}[G(\theta, \tilde{\theta})]$ can be computed explicitly and hence we obtain the following corollary.

Corollary 22.1. *For $P_\theta = \mathcal{N}(\theta, I_p)$ and $Q = \mathcal{N}(0, I_p)$, then*

$$\chi^2(\mathbb{E}_{\theta\sim\pi}[\mathcal{N}(\theta, I_p)] \| \mathcal{N}(\theta, I_p)) = \mathbb{E}[\exp\langle\theta, \tilde{\theta}\rangle] - 1.$$

Proof. Since $P_\theta = \mathcal{N}(\theta, I_p)$, we have

$$\begin{aligned}
G(\theta,\tilde\theta) &= \int \frac{\frac{1}{(2\pi)^{p/2}}\exp\left(\frac{-\|x-\theta\|^2}{2}\right)\frac{1}{(2\pi)^{p/2}}\exp\left(\frac{-\|x-\tilde\theta\|^2}{2}\right)}{\frac{1}{(2\pi)^{p/2}}\exp\left(\frac{-\|x\|^2}{2}\right)} \\
&= \int \frac{1}{(2\pi)^{p/2}}\exp\left(-\frac{1}{2}\left(\|x-\theta\|^2+\|x-\tilde\theta\|^2-\|x\|^2\right)\right) \\
&= \int \frac{1}{(2\pi)^{p/2}}\exp\left(-\frac{1}{2}\left(\|x\|^2-2\langle x,\theta+\tilde\theta\rangle+\|\theta+\tilde\theta\|^2-2\langle\theta,\tilde\theta\rangle\right)\right) \\
&= \exp(\langle\theta,\tilde\theta\rangle)\int \frac{1}{(2\pi)^{p/2}}\exp\left(-\frac{\|x-\theta-\tilde\theta\|^2}{2}\right) \\
&= \exp(\langle\theta,\tilde\theta\rangle).
\end{aligned}$$

□

Example 22.2 (Gaussian mixtures)**.** Consider the symmetric two-component Gaussian distribution $\frac{1}{2}N(-\theta,1)+\frac{1}{2}N(\theta,1)$, where $\theta \geq 0$. Next we show that

$$D\left(\frac{1}{2}N(-\theta,1)+\frac{1}{2}N(\theta,1)\middle\| N(0,1)\right) = O(\theta^4), \quad \theta\to 0 \tag{22.3}$$

This should be contrasted with the behavior usual Gaussian mean model $D(N(\theta,1)\|N(0,1)) = \Theta(\theta^2)$. We see that by symmetrization, the distribution now is significantly closer to the standard normal (indeed, the mean is zero now). In fact, θ now is much more difficult to estimate in the sense that, with n independent samples, the minimax quadratic risk scales as $\Theta(n^{-1/2})$, much slower than the usual parametric rate $\Theta(n^{-1})$:

- The lower bound simply follows from Le Cam's two-point argument in conjunction with the observation (22.3).
- To show the upper bound, with n independent samples $X_1,\ldots,X_n$, first estimate θ^2 by $\widehat{\theta^2} = 1/n\sum_i(X_i^2-1)$. Then instead of directly taking the square root, compute $\widehat\theta = \sqrt{\widehat{\theta^2}+1/n}$ to improve stability.

To show (22.3), we bound KL by χ^2 from above and invoke Corollary 22.1:

$$\chi^2\left(\frac{1}{2}N(-\theta,1)+\frac{1}{2}N(\theta,1)\middle\| N(0,1)\right) = \mathbb{E}\exp(X\tilde X)-1 = \mathbb{E}\exp(\theta^2 B)-1 = \cosh(\theta^2)-1 = \Theta(\theta^4),$$

where $X,\tilde X \overset{\text{i.i.d.}}{\sim} \pm\theta$ equally likely and B is Rademacher.

§ 23. Functional Estimation: LeCam's Method, Risk Upper Bound

In the previous lecture we considered the problem of functional estimation and the idea of using LeCam's method by averaging over multiple points to obtain better estimates of the lower bound for minimax risk. In this lecture, we first use LeCam's method to obtain the lower bound and later describe an estimator obtain the matching upper bound for the minimax risk for the estimation of the ℓ_2-norm of GLM.

Consider the p-dimensional GLM. Let $\theta \in \Theta = \mathbb{R}^p$ and $X \sim \mathcal{N}(\theta, \frac{1}{n} I_p)$. Let $T(\theta) = \|\theta\|_2$. Then,

$$R^*(\Theta) = \inf_{\hat{T}} \sup_{\theta \in \mathbb{R}^p} \mathbb{E}_\theta[(\hat{T} - T)^2] \asymp \frac{\sqrt{p}}{n}.$$

Owing to the scaling property, it suffices to prove the result for the 1-sample GLM.

23.1 LeCam's Method Lower Bound

In order to employ LeCam's method, consider the binary detection problem defined by

$$\begin{cases} H_0 : \theta = 0 \\ H_1 : \|\theta\|_2 \geq \rho \end{cases}.$$

Let $\pi(\cdot)$ be a distribution on $\{\theta : \|\theta\|_2 \geq \rho\}$, $P_0 = \mathcal{N}(0, I_p)$ and

$$P_\pi = \int \mathcal{N}(\theta, I_p) \pi(d\theta).$$

Then by LeCam's method we saw in the previous lecture that

$$R^* \geq \rho^2 \left(1 - d_{\mathrm{TV}}(P_0, P_\pi)\right) \gtrsim \rho^2,$$

when $1 - d_{\mathrm{TV}}(P_0, P_\pi) \gtrsim 0$. From the bounds on the total variational distance, we know that the above condition is satisfied when $\chi^2(P_\pi, P_0) \lesssim 1$, i.e., the χ^2 distance is bounded.

From the Ingster-Suslina method, we know that

$$\chi^2(P_\pi, P_0) = \mathbb{E}\left[G(\theta, \tilde{\theta})\right] - 1,$$

where $\theta, \tilde{\theta} \overset{i.i.d}{\sim} \pi$ and

$$G(\theta, \tilde{\theta}) = \int \frac{P_\theta(dx) P_{\tilde{\theta}}(dx)}{P_0(dx)}.$$

For the GLM,

$$G(\theta, \tilde{\theta}) = \exp\left(\left\langle \theta, \tilde{\theta} \right\rangle\right).$$

Remark 23.1. As an aside, we note that

$$\mathbb{E}\left[\exp\left(\left\langle \theta, \tilde{\theta} \right\rangle\right)\right] \geq \exp\left(\mathbb{E}\left[\left\langle \theta, \tilde{\theta} \right\rangle\right]\right) = \exp\left(\left\langle \mathbb{E}\theta, \mathbb{E}\tilde{\theta} \right\rangle\right) = \exp\left(\|\mathbb{E}\theta\|_2^2\right) > 1.$$

We now consider three priors and bound the χ^2 distance in each case.

23.1.1 Uniform distribution on sphere

Let $\theta, \tilde{\theta} \overset{i.i.d}{\sim} \text{Unif}(\rho S^{p-1})$. Let $\theta = \rho u, \tilde{\theta} = \rho \tilde{u}$ where $\|u\|_2^2 = \|\tilde{u}\|_2^2 = 1$. Hence,

$$\mathbb{E}\left[\exp\left(\left\langle \theta, \tilde{\theta} \right\rangle\right)\right] = \mathbb{E}\left[\exp\left(\rho^2 \left\langle u, \tilde{u} \right\rangle\right)\right].$$

We now exploit the fact that the inner product of directions in high dimensions is small. Let $u = \frac{Z}{\|Z\|_2}, \tilde{u} = \frac{\tilde{Z}}{\|\tilde{Z}\|_2}$, where $Z, \tilde{Z} \overset{i.i.d}{\sim} \mathcal{N}(0, I_p)$ and let $\rho^2 = c\sqrt{p}$. Then,

$$\mathbb{E}\left[\exp\left(\left\langle \theta, \tilde{\theta} \right\rangle\right)\right] = \mathbb{E}\left[\exp\left(c\sqrt{p}\frac{\left\langle Z, \tilde{Z} \right\rangle}{\|Z\|_2 \|\tilde{Z}\|_2}\right)\right] = \mathbb{E}\left[\exp\left(cY\right)\right],$$

where

$$Y = \frac{\sqrt{p}\left\langle Z, \tilde{Z} \right\rangle}{\|Z\|_2 \|\tilde{Z}\|_2}.$$

Now, by the Central Limit Theorem and the fact that $\|Z\|_2 = O_P(\sqrt{p})$,

$$\frac{\left\langle Z, \tilde{Z} \right\rangle}{\sqrt{p}} \xrightarrow{D} \mathcal{N}(0,1), \quad \frac{\|Z\|_2}{\sqrt{p}} \xrightarrow{P} 1, \quad \frac{\|\tilde{Z}\|_2}{\sqrt{p}} \xrightarrow{P} 1.$$

Thus, from Slutsky's theorem, $Y \xrightarrow{D} \mathcal{N}(0,1)$. But we are interested in the convergence of the MGF of Y. Here is a useful result:

Lemma 23.1 ([Koz47, Theorem 1])**.** *Let $X_n \xrightarrow{\text{D}} X$. Define*

$$M(x) \triangleq \sup_n \mathbb{P}\left[|X_n| \geq x\right]. \tag{23.1}$$

If $M(x)$ satisfies

$$\lim_{x \to \infty} M(x) e^{|t|x} = 0 \tag{23.2}$$

for all $|t| \leq \alpha$ for some $\alpha > 0$. Then for all $|t| \leq \alpha$, the moment generating functions $\varphi_n(t) = \mathbb{E}\left[\exp(tX_n)\right]$ and $\varphi(t) = \mathbb{E}\left[\exp(tX)\right]$ are finite. Moreover, $\varphi_n(t) \to \varphi(t)$ as $n \to \infty$.

Here, we have $Y \in [-\sqrt{p}, \sqrt{p}]$. Thus, by Hoeffding's inequality the tail of Y is exponentially bounded. Thus, the MGF of Y converges to the MGF of $\mathcal{N}(0,1)$ which is given by

$$\mathbb{E}\left[\exp\left(sX\right)\right] = \exp\left(\frac{1}{2}s^2\right), \quad \text{when } X \sim \mathcal{N}(0,1).$$

Thus $1 - d_{\text{TV}}(P_0, P_\pi) \gtrsim 0$ and $R^* \gtrsim \sqrt{p}$.

23.1.2 Uniform distribution on hypercube

Let $\rho = cp^{\frac{1}{4}}$ and $\theta, \tilde{\theta} \overset{\text{i.i.d}}{\sim} \text{Unif}\left(cp^{-\frac{1}{4}}\{\pm 1\}^p\right)$.

$$\mathbb{E}\left[\exp\left(\left\langle \theta, \tilde{\theta} \right\rangle\right)\right] = \mathbb{E}\left[\exp\left(\frac{c^2}{\sqrt{p}}\left\langle W, \tilde{W} \right\rangle\right)\right] = \mathbb{E}\left[\exp\left(\frac{c^2}{\sqrt{p}} G_p\right)\right],$$

where $G_p = \left\langle W, \tilde{W} \right\rangle = \sum_{i=1}^{p} W_i \tilde{W}_i$. Now,

$$\mathbb{E}\left[\exp\left(W_i \tilde{W}_i\right)\right] = \frac{1}{2}\left(\exp\left(\frac{c^2}{\sqrt{p}}\right) + \exp\left(-\frac{c^2}{\sqrt{p}}\right)\right).$$

Using Taylor's expansion, we have

$$\frac{\exp(x) + \exp(-x)}{2} = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \cdots \leq \sum_n \frac{(x^2)^n}{n!} = \exp(x^2).$$

Thus we have

$$\mathbb{E}\left[\exp\left(\left\langle \theta, \tilde{\theta} \right\rangle\right)\right] = \left(\frac{1}{2}\left(\exp\left(\frac{c^2}{\sqrt{p}}\right) + \exp\left(-\frac{c^2}{\sqrt{p}}\right)\right)\right)^p \leq \exp(c^4).$$

This consequently implies that for a sufficiently small constant, the χ^2 distance is small as well. Thus, $1 - d_{\mathrm{TV}}(P_0, P_\pi) \gtrsim 0$ and $R^* \gtrsim \sqrt{p}$.

23.1.3 Uniform prior on sparse vectors

Let us consider the binary hypothesis test given by

$$\begin{cases} H_0 : \theta = 0 \\ H_1 : \|\theta\|_2 \geq \rho, \ \theta \in \mathbb{R}_+^p \end{cases}.$$

Now, the priors considered earlier can't be used. In this context, we shall use sparse vectors and a uniform prior to bound the χ^2 distance.

Consider the set of k-sparse vectors and let $\theta, \tilde{\theta} \overset{\text{i.i.d}}{\sim} \mathrm{Unif}\{\theta \in \{0, \epsilon\}^p : |\mathrm{supp}(\theta)| = k\}$. Let $I = \mathrm{supp}(\theta), \tilde{I} = \mathrm{supp}(\tilde{\theta})$. Let $\rho = cp^{\frac{1}{4}}$. Then

$$\|\theta\|_2 = \epsilon\sqrt{k} = \rho = cp^{\frac{1}{4}}.$$

Let $k = \sqrt{p}$. Then $\epsilon = c$. Then

$$\mathbb{E}\left[\exp\left(\left\langle \theta, \tilde{\theta} \right\rangle\right)\right] = \mathbb{E}\left[\exp\left(c^2 \left\langle 1_I, 1_{\tilde{I}} \right\rangle\right)\right] = \mathbb{E}\left[\exp\left(c^2 |\mathrm{supp}(I \cap \tilde{I}|\right)\right].$$

Owing to the symmetry of the problem, it suffices to fix I to be $\{1, \ldots, k\}$ and consider the expectation with respect to the uniform distribution on $\tilde{I}$. Thus $B = |\mathrm{supp}(I \cap \tilde{I})|$ is distributed as Hypergeometric$(p, \sqrt{p}, \sqrt{p})$.

Theorem 23.1 ([Hoe63, Theorem 4]). *Let the population $C = \{c_1, \ldots, c_N\}$. Let $X_1, \ldots, X_n$ denote a random sample without replacement from C and $Y_1, \ldots, Y_n$ denote a random sample with replacement. Let $f(\cdot)$ be a continuous and convex function. Then,*

$$\mathbb{E}\left[f\left(\sum_{i=1}^{n} X_i\right)\right] \leq \mathbb{E}\left[f\left(\sum_{i=1}^{n} Y_i\right)\right].$$

As a corollary of the above theorem, we have

Corollary 23.1. *Let $B \sim$ Hypergeometric$(p, \sqrt{p}, \sqrt{p})$ and $\tilde{B} \sim$ Binom$(\sqrt{p}, \frac{1}{\sqrt{p}})$. Then,*

$$\mathbb{E}\left[\exp\left(sB\right)\right] \leq \mathbb{E}[\exp(s\tilde{B})] = \left(1 - \frac{1}{\sqrt{p}} + \frac{1}{\sqrt{p}}\exp(s)\right)^{\sqrt{p}}.$$

Thus, we have,

$$\mathbb{E}[\exp(\langle\theta, \tilde{\theta}\rangle)] \leq \left(1 + \frac{1}{\sqrt{p}}\left(\exp(c^2) - 1\right)\right)^{\sqrt{p}} \leq \exp\left(\exp\left(c^2\right) - 1\right).$$

Hence for a sufficiently small c, we see that the TV distance is bounded away from 1 and thus $R^* \gtrsim \sqrt{p}$.

23.2 Risk Upper Bound

Having obtained the risk lower bound using LeCam's method, we now seek an estimator that achieves the matching upper bound on the risk. That is, given $X \sim \mathcal{N}(0, I_p)$, we seek to obtain an estimator $\hat{T} = \hat{T}(X)$ of $T = \|\theta\|_2$, such that

$$\sup_{\theta \in \mathbb{R}^p} \mathbb{E}_\theta\left[\left(\hat{T} - T\right)^2\right] \lesssim \sqrt{p}.$$

We shall first consider the plug-in estimator $\hat{T} = \|X\|_2$. Here we note from the triangle inequality that

$$|\hat{T} - T| = |\|X\|_2 - \|\theta\|_2| \leq \|Z\|_2 = O_P(\sqrt{p}).$$

Consequently, $\mathbb{E}_\theta[(\hat{T} - T)^2] \lesssim p$. However, we can verify that this bound is tight - consider the case where $\theta = 0$. This increased risk can be attributed to the presence of a bias in the estimator. That is, we have

$$\mathbb{E}_\theta\left[\|X\|_2^2\right] = \mathbb{E}_\theta\left[\|Z + \theta\|_2^2\right] = \mathbb{E}_\theta\left[\|Z\|_2^2\right] + \|\theta\|_2^2 + \mathbb{E}_\theta\left[2\left\langle Z, \theta\right\rangle\right] = p + \|\theta\|_2^2.$$

In order to negate this bias, define the estimator $\hat{T} = \sqrt{\left(\|X\|_2^2 - p\right)_+}$, where $(x)_+ = \max(x, 0)$. We shall split the analysis of risk of the estimator into two cases.

Case 1: $\|\theta\|_2 \leq p^{\frac{1}{4}}$

Here we have

$$R_\theta = \mathbb{E}_\theta\left[\left(\hat{T} - \|\theta\|_2\right)^2\right] \leq 2\mathbb{E}_\theta\left[\hat{T}^2\right] + 2\|\theta\|_2^2 \leq 2\mathbb{E}\left[|S|\right] + O(\sqrt{p}),$$

where $S = \|X\|_2^2 - p$. We now note that

$$\mathbb{E}_\theta\left[\left|\|X\|_2^2 - p\right|\right] \leq \|\theta\|_2^2 + 2\mathbb{E}_\theta\left[\left|\left\langle\theta, Z\right\rangle\right|\right] + \mathbb{E}_\theta\left[\left|\|Z\|_2^2 - p\right|\right] = O_P(\sqrt{p}),$$

owing to the Central Limit Theorem and the fact that $\|\theta\|_2^2 \leq \sqrt{p}$. Using this, we have $R_\theta \lesssim \sqrt{p}, \forall \|\theta\|_2 \leq p^{\frac{1}{4}}$.

Case 2: $\|\theta\|_2 \geq p^{\frac{1}{4}}$

In this case, let us rewrite the estimation error as follows

$$\hat{T} - T = \sqrt{S_+} - \|\theta\|_2 = \frac{S_+ - \|\theta\|_2^2}{\sqrt{S_+} + \|\theta\|_2}.$$

Thus, we have

$$
\begin{aligned}
|\hat{T} - T| &\leq \frac{|\left(\|X\|_2^2 - p\right)_+ - \|\theta\|_2^2|}{\|\theta\|_2} \leq \frac{|\|X\|_2^2 - p - \|\theta\|_2^2|}{\|\theta\|_2} \\
&= \frac{|\|Z\|_2^2 + \|\theta\|_2^2 + 2\langle\theta, Z\rangle - p - \|\theta\|_2^2|}{\|\theta\|_2} \\
&\leq \frac{|\|Z\|_2^2 - p|}{\|\theta\|_2} + \frac{|2\langle\theta, Z\rangle|}{\|\theta\|_2},
\end{aligned}
$$

where the last step follows from the triangle inequality. Further, we have

$$|\|Z\|_2^2 - p| = O_P(\sqrt{p})$$

and

$$\frac{|2\langle\theta, Z\rangle|}{\|\theta\|_2} = |2\left\langle\frac{\theta}{\|\theta\|_2}, Z\right\rangle| = O_P(1),$$

as $\left\langle\frac{\theta}{\|\theta\|_2}, Z\right\rangle \sim \mathcal{N}(0,1)$. Thus, using the fact that $\|\theta\|_2 \geq p^{\frac{1}{4}}$, we have

$$|\hat{T} - T| \lesssim p^{\frac{1}{4}} \Leftrightarrow R_\theta \lesssim \sqrt{p}.$$

Thus, summarizing the two cases, we observe that

$$\sup_{\theta\in\mathbb{R}^p} \mathbb{E}_\theta[(\hat{T} - T)^2] \lesssim \sqrt{p}$$

and thus $R^* \asymp \sqrt{p}$.

Example 23.1 (Covariance model and independence testing)**.** Let $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \Sigma)$ where Σ is a $p \times p$-dimension Covariance matrix which is to be estimated. Under this model,

- estimating Σ with $l(\hat{\Sigma}, \Sigma) = \|\hat{\Sigma} - \Sigma\|_{\text{op}}$ needs $\Theta(\sqrt{p})$ samples;
- estimating $T(\Sigma) = \|\Sigma\|_{\text{op}}$ with $l(\hat{T}, T) = (\hat{T} - T)^2$ also needs $\Theta(\sqrt{p})$ samples.

Example 23.2 (Looseness of χ^2-method and sharp constant by truncated χ^2)**.** Let $X \sim \mathcal{N}(\theta, I_p)$ and $T = T(\theta) = \theta_{\max} \triangleq \max_{i\in[p]} \theta_i$. Let $l(\hat{T}, T) = \left(\hat{T} - T\right)^2$. Then,

$$R^* = \inf_{\hat{T}} \sup_{\theta\in\mathbb{R}^p} \mathbb{E}_\theta\left[\left(\hat{T} - T\right)^2\right] = \frac{1}{2}(1 + o(1)) \log p, \text{ as } p \to \infty.$$

The results of the above examples are proved in the next lecture.

Outline:

- GLM: estimating $\theta_{\max}$. More careful application of χ^2-method yields the sharp constant.
- Covariance matrix (independence testing): estimating a scalar functional can require as many samples needed as estimating the whole parameter.
- Uniformity testing: Is lottery fair?

24.1 GLM: estimating $\theta_{\max}$

The model of the observations are the same as before: $X = \theta + Z$ where $Z \sim N(0, I_p)$. We want to estimate the magnitude of θ, i.e., $T(\theta) = \theta_{\max}$. We will show the minimax risk with sharp constant in high dimensions:

$$\inf_{\hat{T}} \sup_{\theta \in \mathbb{R}^p} \mathbb{E}_\theta (\hat{T} - \theta_{\max})^2 = \left(\frac{1}{2} + o(1)\right) \log p, \quad p \to \infty.$$

Upper bound: Let's first analyze the maximum likelihood estimator, namely, $X_{\max}$. Consider $\theta = \alpha e_1$ Then $X_{\max} = \max\{\alpha + Z_1, Z_2, \ldots, Z_p\} \approx \max\left\{\alpha + Z_1, \sqrt{2\log p}\right\}$. The picture is the blue curve in Fig. 24.1. A better idea in this case is to decrease $X_{\max}$ by $\sqrt{2\log p}/2$, which will reduce the worst case error.

Let $\hat{T} = X_{\max} - \frac{\sqrt{2\log p}}{2}$. WLOG, consider $\theta_{\max} = \theta_1$. Then

$$\hat{T} - \theta_{\max} = \max_i \left\{X_{\max} - \frac{\sqrt{2\log p}}{2} - \theta_{\max}\right\} \leq \max_i Z_i - \sqrt{\frac{\log p}{2}} \overset{\text{w.h.p.}}{\leq} \sqrt{\frac{\log p}{2}}(1 + o(1)),$$

$$\hat{T} - \theta_{\max} \geq X_1 - \frac{\sqrt{2\log p}}{2} - \theta_{\max} = Z_1 - \sqrt{\frac{\log p}{2}} \geq O_P\left(-\sqrt{\frac{\log p}{2}}(1 + o(1))\right).$$

Lower bound: Consider two hypotheses:

$$H_0 : \theta = 0, \qquad H_1 : \theta_{\max} \geq \tau.$$

Put a prior on H_1: $\theta \sim \text{Uniform}\{\tau e_1, \tau e_2, \ldots, \tau e_p\}$. Then under H_0 the sample $X \sim P_0 = N(0, I_p)$ and under H_1 the sample $X \sim P_\pi = \frac{1}{p}\sum_{i=1}^p N(\tau e_i, I_p)$. The goal is to show that $d_{\text{TV}}(P_0, P_\pi) \to 0$ when $\tau = \sqrt{(2-\epsilon)\log p}$ for any $\epsilon > 0$.

In this problem, directly applying χ^2-method yields the minimax rate but not the sharp constant: Let $\theta = \tau e_I$ and $\tilde{\theta} = \tau e_{\tilde{I}}$, where $I, \tilde{I} \overset{\text{i.i.d.}}{\sim} \text{Uniform}[p]$.

$$\chi^2(P_\pi \| P_0) = \mathbb{E}\exp\langle\theta, \tilde{\theta}\rangle - 1 = \mathbb{E}\exp\left(\tau^2 \mathbf{1}_{\{I \neq \tilde{I}\}}\right) - 1 = \frac{\exp(\tau^2) - 1}{p}.$$

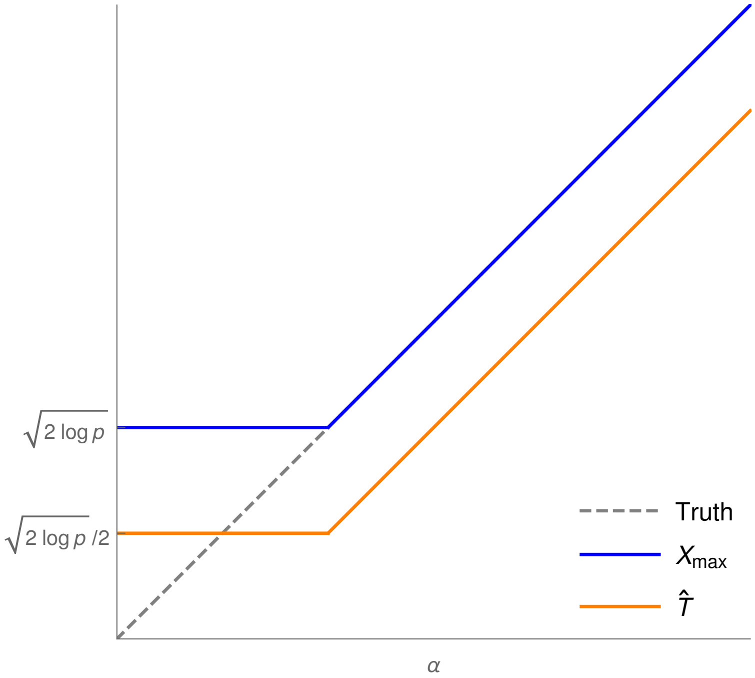


Figure 24.1: Maximum likelihood estimator and improvement via de-biasing.

Therefore $\chi^2(P_\pi \| P_0) \to 0 \Leftrightarrow \frac{\tau}{\sqrt{\log p}} < 1$ and we conclude that $R^* \geq \frac{1+o(1)}{4} \log p$.

We can apply χ^2-method more carefully by conditioning on some high probability event. The main idea is that low probability event has vanishing contribution on the total variation distance but may contribute a lot to the χ^2 distance. Let $\tau = \sqrt{(2-\epsilon)\log p}$ and let

$$E = \left\{ \max_i X_i \leq \sqrt{2\log p} \right\}.$$

Since $\max_i Z_i \leq \sqrt{2\log p}$ with high probability, and $Z_i = O_P(1)$ for any fixed i, E is an high probability event under *both* P_0 *and* P_π. Denote by P_0^E and P_π^E the probability measure conditioned on E, that is, $P_0^E(\cdot) = \frac{P_0(\cdot \cap E)}{P_0(E)}$. Note that

$$d_{\rm TV}(P_0, P_0^E) = 1 - P_0(E), \quad d_{\rm TV}(P_\pi, P_\pi^E) = 1 - P_\pi(E). \tag{24.1}$$

By triangle inequality, it suffices to show that $d_{\rm TV}(P_0^E, P_\pi^E) \to 0$. By the formula for conditional probability, the likelihood ratio is

$$\frac{P_\pi^E}{P_0^E} = \frac{P_0(E)}{P_\pi(E)} \frac{P_\pi}{P_0} \mathbf{1}_E.$$

Applying χ^2-method on P_0^E and P_π^E, we obtain that

$$\begin{aligned}\int \frac{P_\pi^2}{P_0}\mathbf{1}_E &= \mathbb{E}_{\theta,\tilde{\theta}\sim\pi}\left[\int \frac{P_\theta P_{\tilde{\theta}}}{P_0}\mathbf{1}_E\right] = \mathbb{E}_{\theta\sim\pi}\mathbb{E}_{X\sim P_\pi}\left[\exp\left(-\frac{\|\theta\|_2^2}{2} + \langle\theta, X\rangle\right)\mathbf{1}_E\right]\\ &= \mathbb{E}_{X\sim P_\pi}\mathbb{E}_I\left[\exp\left(-\tau^2/2 + \tau\langle X, e_I\rangle\right)\mathbf{1}_E\right]\\ &= \left(1-\frac{1}{p}\right)\mathbb{E}\left[\exp\left(-\tau^2/2+\tau N(0,1)\right)\mathbf{1}_E\right] + \frac{1}{p}\mathbb{E}\left[\exp\left(-\tau^2/2+\tau X_1\right)\mathbf{1}_E\right]\\ &\le \left(1-\frac{1}{p}\right) + \frac{1}{p}\exp\left(\left(-\frac{2-\epsilon}{2} + \sqrt{2(2-\epsilon)}\right)\log p\right).\end{aligned}$$

Note that $-(2-\epsilon)/2+\sqrt{2(2-\epsilon)} < 1$ as long as $\epsilon > 0$. Therefore $\int \frac{P_\pi^2}{P_0}\mathbf{1}_E = 1+o(1)$ and consequently

$$\begin{aligned}\chi^2(P_\pi^E\|P_0^E) = o(1) &\implies d_{\mathrm{TV}}(P_\pi^E, P_0^E) = o(1)\\ &\overset{(24.1)}{\implies} d_{\mathrm{TV}}(P_\pi, P_0) = o(1)\\ &\overset{\text{LeCam}}{\implies} R^* \ge \frac{1+o(1)}{2}\log p,\end{aligned}$$

where we applied LeCam's method for quadratic risk in Theorem 10.2.

24.2 Covariance matrix model

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(0, \Sigma)$, where Σ is the covariance matrix with size $p \times p$. A sufficient statistic for Σ is the sample covariance matrix:

$$S = \frac{1}{n}\sum_{i=1}^n X_i X_i'.$$

Let $\Theta = \left\{\Sigma : \|\Sigma\|_{op} \le \lambda\right\}$. The minimax risk for estimating Σ under the operator norm is

$$R_1^* \triangleq \inf_{\hat{\Sigma}}\sup_{\Sigma\in\Theta}\mathbb{E}\|\hat{\Sigma}-\Sigma\|_{op}^2 \asymp \lambda^2\left(1 \wedge \frac{p}{n}\right).$$

Even if we only want to estimate the operator norm, a scalar functional of Σ, the difficulty in terms of the minimax rate is the same as estimating Σ itself:

$$R_2^* \triangleq \inf_{\widehat{\|\Sigma\|_{op}}}\sup_{\Sigma\in\Theta}\mathbb{E}\left(\widehat{\|\Sigma\|_{op}} - \|\Sigma\|_{op}\right)^2 \asymp \lambda^2\left(1\wedge\frac{p}{n}\right).$$

Note that $\|\hat{\Sigma}\|_{op}$ is a viable estimator for $\|\Sigma\|_{op}$. By the triangle inequality of the operator norm,

$$R_2^* \lesssim R_1^*.$$

It suffices to show an upper bound for estimating Σ and the same lower bound for estimating $\|\Sigma\|_{op}$.

Upper bound for estimating Σ: Note a trivial upper bound that $R_1^* \leq \lambda^2$. It remains to show that $R_1^* \lesssim \lambda^2 p/n$ when $n \gtrsim p$. Consider the sufficient statistic S. We want to show that for any $\|\Sigma\|_{op} \leq \lambda$,

$$\|S - \Sigma\|_{op} \overset{\text{w.h.p}}{\leq} \lambda\sqrt{\frac{p}{n}},$$

when $n \gtrsim p$. Let $X_i = \Sigma^{1/2} Z_i$ then $Z_i \overset{\text{i.i.d.}}{\sim} N(0, I_p)$ and $S = \Sigma^{1/2}(\frac{1}{n}\sum_{i=1}^n Z_i Z_i')\Sigma'^{1/2}$. Let $\tilde{S} \triangleq \frac{1}{n}\sum_{i=1}^n Z_i Z_i'$ then

$$\|S - \Sigma\|_{op} = \|\Sigma^{1/2}(\tilde{S} - I_p)\Sigma'^{1/2}\|_{op} \leq \|\Sigma^{1/2}\|_{op}\|\tilde{S} - I_p\|_{op}\|\Sigma'^{1/2}\|_{op} = \lambda\|\tilde{S} - I_p\|_{op}.$$

We use the result that, with high probability,

$$\|\tilde{S} - I_p\|_{op}^2 \lesssim \sqrt{\frac{p}{n}} + \frac{p}{n}.$$

The intuition for the above result is that

$$\|\tilde{S} - I_p\|_{op}^2 \leq \sup_{\|v\|=1} \|\tilde{S}v\|^2 + 1 - 2\inf_{\|v\|=1}\|\tilde{S}v\| \approx (1 + \sqrt{p/n})^2 + 1 - 2(1 - \sqrt{p/n}) = 4\sqrt{\frac{p}{n}} + \frac{p}{n}.$$

When $n \gtrsim p$ we have $\|\tilde{S} - I_p\|_{op} \overset{\text{w.h.p.}}{\lesssim} \sqrt{p/n}$.

Lower bound for estimating $\|\Sigma\|_{op}$: Let $a, b > 0$ be two parameters to be specified in the end. Consider two hypotheses:

$$H_0 : \Sigma = \Sigma_0 = aI, \qquad H_1 : \Sigma = \Sigma_v = aI + bvv',$$

where under the alternative Σ is a rank-one perturbation from the identity matrix. Then the operator norms under H_0 and H_1 are separated by b. Put a prior on H_1 that $v \sim \text{Uniform}\left\{\frac{\pm 1}{\sqrt{p}}\right\}^p$.

Applying the χ^2-method, we obtain that

$$\begin{aligned}\chi^2 + 1 &= \mathbb{E}_{v,\tilde{v}} \int \frac{N(0,\Sigma_v)^{\otimes n} N(0,\Sigma_{\tilde{v}})^{\otimes n}}{N(0,\Sigma_0)^{\otimes n}} = \mathbb{E}_{v,\tilde{v}} \left(\int \frac{N(0,\Sigma_v)N(0,\Sigma_{\tilde{v}})}{N(0,\Sigma_0)}\right)^n \\ &= \mathbb{E}_{v,\tilde{v}} \left(\sqrt{\frac{|\Sigma_0|}{|\Sigma_v||\Sigma_{\tilde{v}}||\Sigma_v^{-1} + \Sigma_{\tilde{v}}^{-1} - \Sigma_0^{-1}|}}\right)^n = \mathbb{E}_{v,\tilde{v}} \left(\det\left(I_p - \frac{b^2}{a^2} vv'\tilde{v}\tilde{v}'\right)\right)^{-n/2} \\ &= \mathbb{E}_{v,\tilde{v}} \left(\det\left(I_p - \frac{b^2}{a^2}\langle v', \tilde{v}\rangle v\tilde{v}'\right)\right)^{-n/2}.\end{aligned}$$

Applying matrix determinant lemma that $\det(A + uv') = (1 + v'A^{-1}u)\det(A)$ yields that

$$\chi^2 + 1 = \mathbb{E}_{v,\tilde{v}} \left(1 - \frac{b^2}{a^2}\langle v', \tilde{v}\rangle^2\right)^{-n/2} \leq \mathbb{E}_{v,\tilde{v}} \exp\left(\frac{nb^2}{2a^2}\langle v', \tilde{v}\rangle^2\right).$$

Note that the distribution of $\langle v', \tilde{v}\rangle$ is the same as $\frac{1}{p}\sum_{i=1}^p R_i$ where R_i is an i.i.d. Rademacher random variable taking values ± 1 with probability $1/2$. Then $\langle v', \tilde{v}\rangle$ is concentrated on $[-\frac{1}{\sqrt{p}}, \frac{1}{\sqrt{p}}]$

(this can be made rigorous through Hungarian coupling). The problem boils down to the following simple optimization:

$$\begin{aligned} &\max \ b \\ &\text{s.t. } 0 \le a \le a+b \le \lambda, \\ &\frac{nb^2}{a^2 p} \le c, \end{aligned}$$

for some constant c. The optimal solution is

$$b = \frac{\lambda}{1+\sqrt{n/cp}} \asymp \lambda\left(1 \wedge \sqrt{\frac{p}{n}}\right).$$

24.3 Uniformity testing: Is the lottery fair?

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P$ where P is a distribution on $[k]$. Consider two hypotheses:

$$H_0 : P = \text{Uniform}[k], \qquad H_1 : d_{\text{TV}}(P, \text{Uniform}[k]) \ge \epsilon.$$

A test is a function $\psi : [k]^n \to \{0,1\}$ and we want the probability of error to be

$$P_0^{\otimes n}(\psi = 1) + \sup_{P \in H_1} P^{\otimes n}(\psi = 0) \le 1\%.$$

The sample complexity $n^*(k,\epsilon)$ is defined by the minimum sample size n such that a satisfactory test exists.

Theorem 24.1 ([Pan08])**.**

$$n^*(k,\epsilon) \asymp \frac{\sqrt{k}}{\epsilon^2}.$$

Remark 24.1. In comparison, estimating P by $\hat{P}$ such that $\mathbb{E}d_{\text{TV}}(P,\hat{P}) \le \epsilon$ requires $\asymp k/\epsilon^2$ samples, achieved by the empirical distribution. In other words, in order for the empirical distribution $\hat{P}$ to be ϵ-close to the true distribution P, we need a lot more samples.

Remark 24.2 (Sufficient statistics for symmetric functionals)**.** To estimate a distribution or any functional thereof, a sufficient statistic is the histogram $(N_1, \ldots, N_k)$ where N_i records the number of appearances of symbol i. Since the total variation distance is permutation invariant (symmetric), a further sufficient statistic is the histogram of histogram, commonly known as profiles or fingerprints, $(\varphi_1, \ldots, \varphi_n)$, where φ_i counts the number of symbols that appear exactly i times, i.e.,

$$\varphi_i = \sum_j \mathbf{1}\{N_j = i\}.$$

Note that here the sufficiency should be understood operationally, in that for estimating symmetric functionals, it does not deteriorate the worst-case risk if one restrict to estimators as functions of the profiles alone. This sufficiency does not come with a factorization theorem in the usual sense of Fisher and Neyman, since this summary is not sufficient for other estimating tasks, e.g., estimating P_1.

More generally, we have the following result, which is in the same spirit of the theory of invariant estimators:

Lemma 24.1. *Let $\theta \in \Theta \subset \mathbb{R}^d$ where Θ is closed under permutation. Let the observation $X \sim P_\theta$ is also $\mathbb{R}^d$ valued. Let $F(\theta)$ be a real-valued functional to be estimated. Assume that: for any permutation $\sigma \in S_d$, for any $\theta \in \Theta$*

1. *Permutation invariance of the model: $\sigma(X) \sim P_{\sigma(\theta)}$.*
2. *Permutation invariance of the functional: $F(\sigma(\theta)) = F(\theta)$.*

Then the minimax estimator is permutation invariant.

Proof. Given any proposed estimator $\hat{F}$, define $\tilde{F}$ by averaging the input over all permutations:

$$\tilde{F}(x) \triangleq \mathbb{E}_{\boldsymbol{\sigma}}[\hat{F}(\boldsymbol{\sigma}(x))] = \frac{1}{d!}\sum_{\sigma \in S_d} \hat{F}(\sigma(x)),$$

where $\boldsymbol{\sigma}$ is an independent random permutation. We show that the worst-case risk of $\tilde{F}$ is no worse than that of $\hat{F}$. Fix $\theta_0 \in \Theta$. The idea is to consider the prior which is the random permutation of θ_0. Denote the risk of an estimator $\hat{F}$ at θ by $R(\hat{F}, \theta) = \mathbb{E}_\theta[(F(\theta) - \hat{F}(X))^2]$. Then the risk of $\tilde{F}$ at θ_0 is controlled by

$$\begin{aligned} R(\tilde{F}, \theta_0) &= \mathbb{E}_{\theta_0}[(\mathbb{E}_{\boldsymbol{\sigma}}[\hat{F}(\boldsymbol{\sigma}(X))] - F(\theta_0))^2] \\ &\overset{\text{(a)}}{=} \mathbb{E}_{\theta_0}[(\mathbb{E}_{\boldsymbol{\sigma}}[\hat{F}(\boldsymbol{\sigma}(X))] - F(\boldsymbol{\sigma}(\theta_0)))^2] \\ &\overset{\text{(b)}}{\le} \mathbb{E}_{\theta_0}\mathbb{E}_{\boldsymbol{\sigma}}[(\hat{F}(\boldsymbol{\sigma}(X)) - F(\boldsymbol{\sigma}(\theta_0)))^2] \\ &\overset{\text{(c)}}{=} \mathbb{E}_{\boldsymbol{\sigma}}[R(\hat{F}, \boldsymbol{\sigma}(\theta_0))] \\ &\le \sup_{\theta \in \Theta} R(\hat{F}, \theta) \end{aligned}$$

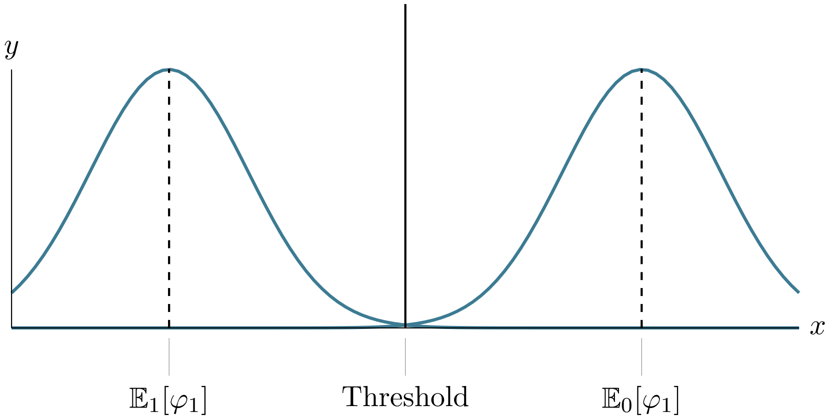
where (a) and (c) follows from the permutation invariance of the functional and the model, respectively; (b) is Jensen's inequality. Taking supremum over θ_0 of the LHS shows that the modified $\tilde{F}$ is at least as good as $\hat{F}$ in the worst case. □

Upper bound: Our test statistic is φ_1. This is related to "birthday paradox": consider k days and n people,

$$\mathbb{P}[\text{no identical birthday}] = \frac{k}{k}\frac{k-1}{k}\cdots\frac{k-n+1}{k} = \exp\left(\sum_{i=1}^{n-1}\log(1-i/k)\right) \approx \exp(-n^2/2k).$$

When $n \lesssim \sqrt{k}$ then $\varphi_1 \approx n$. The intuition is that the coincidence is least likely under uniform distribution: φ_1 is large (close to n) under H_0 and φ_1 is small under H_1.

By definition $\varphi_1 = \sum_{i=1}^{k} \mathbf{1}_{N_i=1}$. We can compute that $\mathbb{E}_0[\varphi_1] - \mathbb{E}_1[\varphi_1] \gtrsim \frac{n^2\epsilon^2}{k}$ and $\mathrm{Var}_0[\varphi_1] \lesssim \frac{n^2}{k}$. If $n \gtrsim \frac{\sqrt{k}}{\epsilon^2}$ then $\sqrt{\mathrm{Var}_0[\varphi_1]} \lesssim \mathbb{E}_0[\varphi_1] - \mathbb{E}_1[\varphi_1]$. Under H_1 we can also compute that $\sqrt{\mathrm{Var}_0[\varphi_1]} \lesssim \mathbb{E}_0[\varphi_1] - \mathbb{E}_1[\varphi_1]$. The picture is shown as below and the detailed computation is referred to [Pan08].



Lower bound: Consider two hypotheses:

$$H_0 : P = \text{Uniform}[k], \qquad H_1 : P = P_I = (p_1, \dots, p_k),$$

where $I \subseteq [k]$ is of size $k/2$ and

$$p_i = \begin{cases} \frac{1+\epsilon}{k}, & i \in I, \\ \frac{1-\epsilon}{k}, & i \notin I. \end{cases}$$

Put a uniform prior on H_1 where I is chosen uniformly at random from all subsets of size $k/2$. The goal is to show that

$$d_{\text{TV}}\left(\frac{1}{\binom{k}{k/2}} \sum_{|I|=k/2} P_I^{\otimes n}, \text{Uniform}[k]^{\otimes n} \right) < c$$

for some constant $c < 1$. A sufficient condition is that

$$\chi^2 \left(\frac{1}{\binom{k}{k/2}} \sum_{|I|=k/2} P_I^{\otimes n} \middle\| \text{Uniform}[k]^{\otimes n} \right) < \infty.$$

Applying the Ingster-Suslina method (Lemma 22.1):

$$\begin{aligned} \chi^2 + 1 = \mathbb{E}_{I,\tilde{I}} \int \frac{P_I^{\otimes n} P_{\tilde{I}}^{\otimes n}}{P_0^{\otimes n}} &= \mathbb{E}_{I,\tilde{I}} \left(\sum \frac{P_I P_{\tilde{I}}}{P_0} \right)^n = \mathbb{E}_{I,\tilde{I}} \left(\frac{4\epsilon^2 |I \cap \tilde{I}|}{k} + 1 - \epsilon^2 \right)^n \\ &\le \mathbb{E}_{I,\tilde{I}} \exp\left(n\epsilon^2 \left(\frac{4|I \cap \tilde{I}|}{k} - 1 \right) \right), \end{aligned}$$

where $I \cap \tilde{I} \sim \text{Hypergeometric}(k, k/2, k/2)$. Applying the convex stochastic dominance of the binomial distribution over the hypergeometric distribution (Corollary 23.1), we obtain that

$$\begin{aligned} \chi^2 + 1 \le \mathbb{E}_{I,\tilde{I}} \exp\left(n\epsilon^2 \left(\frac{4\text{Binom}(k, 1/2)}{k} - 1 \right) \right) &= \left(\frac{\exp(2n\epsilon^2/k) + \exp(-2n\epsilon^2/k)}{2} \right)^{k/2} \\ &\le \exp\left(\frac{1}{2} \left(\frac{2n\epsilon^2}{k} \right)^2 \frac{k}{2} \right) < \infty, \end{aligned}$$

when $n \lesssim \frac{\sqrt{k}}{\epsilon^2}$, where we used the inequality that $\frac{e^x + e^{-x}}{2} \le e^{x^2/2}$ (by Taylor expansion).

§ 25. Approximation-theoretic methods for functional estimation

In this lecture, we discuss the powerful idea of "approximation theory" and its use in functional estimation, by considering our favorite example of p-dimensional Gaussian Location Model (GLM):

- $X \sim \mathcal{N}(\theta, I_p), \quad \theta \in [-1,1]^p \triangleq \Theta.$
- $T(\theta) = \|\theta\|_1, \quad \ell(T,\hat{T}) = (T-\hat{T})^2.$

Notice that we already investigated the case where the functional was $T(\theta) = \|\theta\|_2$, where we knew that $R^* \asymp \sqrt{p}$. However, in this case where the statistic $T(\cdot) = \|\cdot\|_1$ has a kink at 0, we should adopt elegant techniques of the approximation theory. In fact, what we are going to prove (spoiler alert!) is the following:

$$R^* \triangleq \inf_{\hat{T}} \sup_{\theta\in\Theta} \mathbb{E}[(T-\hat{T})^2] \asymp \left(p \cdot \frac{\log\log p}{\log p}\right)^2.$$

25.1 Upper bound

25.1.1 Unbiased estimator does not exist.

Benchmark: Suppose we use a naïve estimator of $\hat{T}(X) = \|X\|_1$. This estimator is definitely not the best; when $\theta = 0$, then $T(\theta) = 0$ but $\hat{T} = \|Z\|_1 \asymp p$ and therefore suffering the quadratic risk of $\Theta(p^2)$!

In the above benchmark, we observed that it is 'bias' that is the main culprit. So let's see if we could get an 'unbiased' estimator of $T(\theta)$.

Unfortunately, that is not possible even in the simplest 1-dimensional scenario: Having $X \sim \mathcal{N}(\theta,1)$ we want to find an estimator $\hat{T}(X)$ such that $\mathbb{E}_\theta[\hat{T}(X)] = |\theta|$. If we expand the (lhs),

$$\int_{-\infty}^{\infty} \hat{T}(x) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2}\right) dx = |\theta|.$$

It is easy to see that the (lhs) is differentiable in θ everywhere regardless of what $\hat{T}$ is (thanks to the smoothness of the pdf of $\mathcal{N}(\theta,1)$). In fact, it is an analytic function[1] of θ. On the other hand, (rhs) is not differentiable in θ at $\theta = 0$, and is not an analytic function.

Hence, it is impossible to have an unbiased estimator of $T(\theta)$. However, we certainly can bound the magnitude of the bias, by 'approximating' $T(\theta)$ with a set of analytic functions. Namely, we are going to use the friendly **polynomials** to approximate $T(\theta) = |\theta|$.

[1] infinitely differentiable with the everywhere convergent Taylor series!

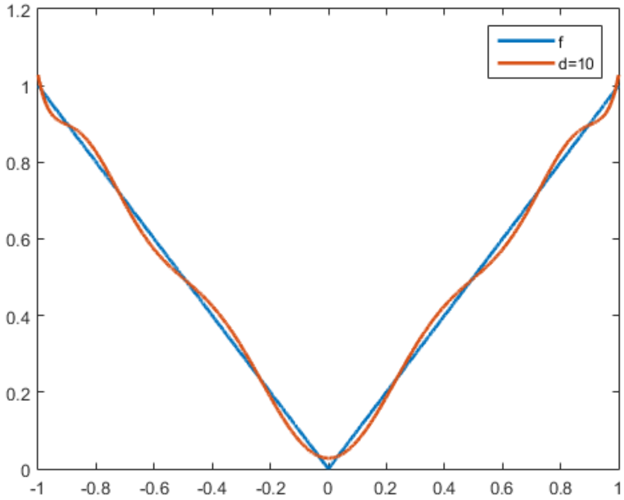


Figure 25.1: Example of approximating $f(x) = |x|$ with a degree-10 polynomial.

25.1.2 Bounding bias: Polynomial approximation

Let us denote the set of *degree-d polynomials* by $\mathcal{P}_d \triangleq \{q : \deg(q) \le d\}$, where $q(x) = \sum_{i=0}^d a_i x^i$. To contend with bias, we are interested in the task of **uniform approximation**: Given $f : [0,1] \to \mathbb{R}$, we want to find a polynomial q which achieves

$$\begin{aligned} E_d(f) &\triangleq \inf_{\deg(q)\le d} \|q - f\|_\infty \\ &= \inf_{\deg(q)\le d} \sup_{x\in[0,1]} |q(x) - f(x)|. \end{aligned}$$

This is an infinite-dimensional linear programming, and at the same time, $(d+1)$-dimensional convex optimization.

Here are some useful facts.

- *Stone-Weierstrass theorem*: Polynomials are *dense* in $C[0,1]$. In other words, for all continuous $f : [0,1] \to \mathbb{R}$, we know that $E_d(f) \xrightarrow{d\to\infty} 0$. This follows from the explicit construction below.
- *Bernstein's construction*: Uniform convergence is possible for any continuous $f : [0,1] \to \mathbb{R}$, i.e. there exists a sequence of polynomials indexed by the degree $\{b_d\}_{d=0}^\infty$ which uniformly converges to f as $d \to \infty$. Here's one way to do it, due to Sergei Bernstein. For a function f, let's define *Bernstein polynomial* of f of degree d as

  $$B_d(x) \triangleq \sum_{k=0}^d f\left(\frac{k}{d}\right)\binom{d}{k} x^k (1-x)^{d-k} = \sum_{k=0}^d f\left(\frac{k}{d}\right) \mathbb{P}\left[\mathrm{Binom}(d,x) = k\right] = \mathbb{E}_{Y\sim\mathrm{Binom}(d,x)}[f(Y/d)].$$

  Since $Y/d \to x$ in distribution and f is a continous (hence bounded) function on $[0,1]$, then $B_d(x) \to f(x)$ for any $x \in [0,1]$ (pointwise convergence). To upgrade to uniform convergence, simply invoke the uniform continuity of f and Chebyshev inequality.[2]

[2] Indeed, $|\mathbb{E}[f(Y/d) - f(x)]| \le \mathbb{E}[|f(Y/d) - f(x)|\mathbf{1}\{|Y/d - x| \le \epsilon\}] + \|f\|_\infty \mathbb{P}\left[|Y/d - x| > \epsilon\right] \le \sup_{|x'-x|\le\epsilon} |f(x') - f(x)| + \|f\|_\infty/d$, where ϵ is arbitrary.

[add lower bound with $f(x) = |x-1|$ and show that for 1-Lip, $1/\sqrt{d}$ is tight.]

- *Jackson-type inequality*: For a smoother f, we can do better. Whenever $\mathrm{Lip}(f) \leq 1$, Jackson's inequality implies that there exists a $q \in \mathcal{P}_d$ such that $\max_{x\in[-1,1]} |q(x) - |x|| \leq \frac{C}{d}$ for some constant C. While the details would not be stated here, the main idea is to use Fourier series.

Equipped with these tools, we now have the following program in mind: We have the statistic $T(\theta) = \sum_{i=1}^{p} |\theta_i|$, which can be approximated by some $\tilde{T}(\theta) = \sum_{i=1}^{p} q_d(\theta_i)$ where q_d is a degree-d polynomial function. While $T(\theta)$ cannot be estimated unbiasedly, we can unbiasedly estimate $\tilde{T}(\theta)$ with some estimator $\hat{T}(X)$. Then, the bias of this estimator in estimating true $T(\theta)$ would be

$$\mathbb{E}[\hat{T}(X) - T(\theta)] = \tilde{T}(\theta) - T(\theta) \leq \frac{Cp}{d}.$$

Now, the whole story is about estimating degree-d polynomial q_d unbiasedly. In other words, we want to unbiasedly estimate **moments** of the empirical distribution of $\frac{1}{p}\sum_i \theta_i$.

25.1.3 Estimating moments: Orthogonal polynomials

In this subsection, we introduce a systematic way of estimating the moments of θ when we can observe $X \sim \mathcal{N}(\theta, 1)$. We are going to use a basis for the space of polynomials, called **orthogonal polynomials**.

First, we need to define an inner product of two functions. We define it as a 'weighted' integral of the product of two functions. Here, we choose the weight $\phi(x)$ to be a pdf of $\mathcal{N}(0,1)$, which would make this definition of inner product useful under the additive Gaussian noise assumption. Our inner product is defined as follows: when $f, q : \mathbb{R} \to \mathbb{R}$,

$$\langle f, q\rangle \triangleq \int_{-\infty}^{+\infty} f(x)q(x)\phi(x)dx = \mathbb{E}[f(Z)q(Z)], \quad Z \sim \mathcal{N}(0,1).$$

Now, we apply the Gram-Schmidt process to the family of monomials $\{1, x, x^2, x^3, \ldots\}$ to form an orthogonal basis $\{H_0, H_1, H_2, \ldots\}$, which we call **Hermite polynomials**. In other words, we define Hermite polynomials as follows:

$$H_k \triangleq x^k - \sum_{i=0}^{k-1} \frac{\langle H_i, x^k\rangle}{\langle H_i, H_i\rangle} H_i,$$

for example, $H_0 = 1, H_1 = x, H_2 = x^2 - 1, \ldots$. Note that we only did orthogonalization not normalization.

Here are some properties of Hermite polynomials:

- Orthogonality: $\langle H_k, H_i\rangle = k!\mathbf{1}\{k = i\}$.
- Basis for L_2: $\{H_0, H_1, H_2, \ldots\}$ forms an orthogonal basis for $L_2(\mathbb{R}, \varphi)$.
- Basis for $\mathcal{P}_d$: $\{H_0, H_1, H_2, \ldots, H_d\}$ forms an orthogonal basis for $\mathcal{P}_d$.
- Monic: the leading coefficient of each Hermite polynomial is 1.

Lemma 25.1 (Estimating monomials). *If $X \sim \mathcal{N}(\theta, 1)$, then $\mathbb{E}[H_k(X)] = \theta^k$, i.e. $H_k(X)$ is an unbiased estimator of θ^k.*

Proof. Instead of using their explicit formula, we use the defining properties of Hermite polynomials. Let $Z \sim \mathcal{N}(0,1)$. Note that $H_k(X) = H_k(\theta + Z)$ is a deg-k polynomial in Z, which can be expressed as a linear combination of the basis $\{H_0, \ldots, H_k\}$ as $H_k(\theta+Z) = \sum_{i=0}^{k} a_i H_i(Z)$. Note that $a_i = a_i(\theta)$ is a polynomial in θ and $\deg(a_i) \le k - i$. Since H_k is monic, we have $a_0 = \theta^k$. By orthogonality and $H_0 = 1$, we have

$$\mathbb{E}[H_k(X)] = \sum_{i=0}^{k} a_i \mathbb{E}[H_i(Z)] = a_0 = \theta^k.$$

□

25.1.4 Recipe of an estimator: Selecting right d

The story so far is as follows: We fix some $d \in \mathbb{N}$, and find the best approximation of $|x|$ on $[-1,1]$ of degree d, $q(x) = \sum_{i=0}^{d} a_i x^i$ by the Jackson-type inequality. Then, we have $\mathbb{E}\left[\sum_{i=0}^{d} a_i H_i(x_j)\right] = q(\theta_j)$ for each coordinate $j \in [p]$. Now, our estimator for $\|\theta\|_1$ is

$$\hat{T} \triangleq \sum_{j=1}^{p} \left[\sum_{i=1}^{d} a_i H_i(x_j) \right].$$

Then, we know that

- (*systematic error*) $|\text{bias}(\hat{T})| \le \frac{Cp}{d}$ (proved in the previous sections) and
- (*stochastic error*) $\text{Var}(\hat{T}) = \sum_{j=1}^{p} \text{Var}\left(\sum_{i=1}^{d} a_i H_i(x_j)\right) \lesssim p \cdot d!$ where $d!$ is from the norm of the Hermite polynomials. We also used the fact that a_i are not too big, usually $\le C \cdot d$ for some constant C.

We see that as d grows, the bound on bias decreases while the bound on variance increases. We might hope to find a 'sweet spot,' where $\text{bias}^2 \asymp \text{Var} \Leftrightarrow \frac{p^2}{d^2} \approx p \cdot d! \Leftrightarrow d \approx \frac{\log p}{\log\log p}$ (using Stirling's expansion) so that the squared error is minimized. Plugging in, we get want we wanted:

$$\mathbb{E}\left(\hat{T} - \|\theta\|_1\right)^2 \lesssim \left(p \cdot \frac{\log\log p}{\log p}\right)^2.$$

The method we used might seem a little bit ad hoc, but since we can show that this is the best (as will be followed in the next section), we are happy with it. For details, the readers may refer to the beautiful paper of Cai and Low [CL11], and the basic idea goes back to Nemirovski [?] and Lepski et al. [LNS99].

25.2 Lower bound

25.2.1 Composite vs. Composite: 1-dimensional reduction

We now show that the rate of $\left(p \cdot \frac{\log\log p}{\log p}\right)^2$ is actually unbeatable. The problem is estimating one number $T(\theta) = \|\theta\|_1$, and this is a collapsed version of the high-dimensional object. The strategy we used for ℓ_2 norm was to convert it to a test and to run Le Cam's method: $H_0 : \theta = 0$ vs. $H_1 : \|\theta\|_1 \ge \rho$. Unfortunately, this doesn't work; the best we can get this way is a rate of $p^{3/2}$.

What is to blame, is that we previously simplified the test into 'simple vs. composite' from the original 'composite vs. composite', i.e. $H_0 : \|\theta\|_1 \leq a, H_1 : \|\theta\|_1 \geq b$ (it may suffice to consider $\approx a$ vs. $\approx b$). The reason for the simplification was that the second moment calculation was too complicated in the latter case. However, with polynomial ideas, the calculation becomes a doable task and we can get $\left(p \cdot \frac{\log \log p}{\log p}\right)^2$.

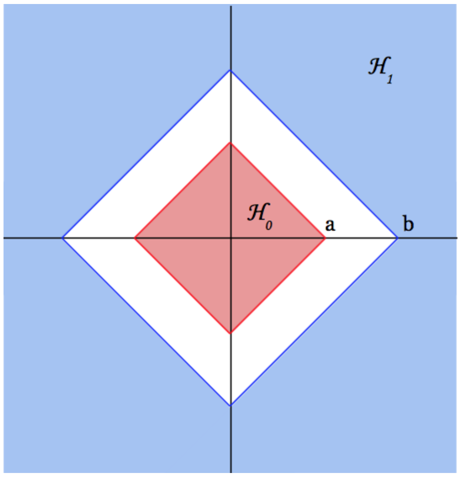


Figure 25.2: $H_0 : \|\theta\|_1 \leq a, H_1 : \|\theta\|_1 \geq b$

Let's see how far we could proceed with the classic method. Let $\Theta_0 \triangleq \{\theta : \|\theta\|_1 \approx a\}, \Theta_1 \triangleq \{\theta : \|\theta\|_1 \approx b\}$. Then, for any prior $P \in \mathcal{M}(\Theta_0), Q \in \mathcal{M}(\Theta_1)$, we have

$$R^* \gtrsim (b-a)^2(1 - d_{\text{TV}}(P * \mathcal{N}(0, I_p), Q * \mathcal{N}(0, I_p)))$$

and we may hope something like $d_{\text{TV}}(P * \mathcal{N}(0, I_p), Q * \mathcal{N}(0, I_p)) \leq 0.1$, or simply bounded away from 1. However, each distribution is p-dimensional Gaussian convolved with another p-dimensional distribution, and d_{TV} is not easy to calculate.

Then, can we reduce it to 1-dimension? In that case, the hypotheses should look like $H_0 : \theta = (\theta_1, \ldots, \theta_p) \overset{\text{i.i.d.}}{\sim} P, H_1 : \theta = (\theta_1, \ldots, \theta_p) \overset{\text{i.i.d.}}{\sim} Q$, and the data generating takes the form of $x = (x_1, \ldots, x_p) \overset{\text{i.i.d.}}{\sim} (P \text{ or } Q) * \mathcal{N}(0, 1)$, and we might use the tensorization $d_{\text{TV}}(P_X^{\otimes p}, Q_x^{\otimes p}) \leq p \cdot d_{\text{TV}}(P_X, Q_X) \leq 0.1$ to proceed further.

One problem is that if we take i.i.d. θ from P or Q, θ might not lie exactly on the spheres Θ_0, Θ_1. But this is actually a small problem and can be fixed. If we look at $T(\theta)$:

$$\text{under } H_0 : T(\theta) = \sum_{i=1}^{p} |\theta_i| \overset{\text{CLT}}{=} p \cdot \mathbb{E}_P|\theta| + o_p(\sqrt{p}),$$
$$\text{under } H_1 : T(\theta) = \sum_{i=1}^{p} |\theta_i| \overset{\text{CLT}}{=} p \cdot \mathbb{E}_Q|\theta| + o_p(\sqrt{p}).$$

And if we can choose P, Q s.t. $\mathbb{E}_P|\theta|$ and $\mathbb{E}_Q|\theta|$ to be separated by $\asymp p\frac{\log \log p}{\log p} >> \sqrt{p}$, then we are happy even though this is not exactly a sphere vs. sphere.

25.2.2 Choosing P, Q: Sufficient condition of moments

Now the question is how to choose P, Q. We can formally formulate a problem as follows:

$$\begin{aligned}
&\text{maximize : } \mathbb{E}_Q|\theta| - \mathbb{E}_P|\theta| \\
&\text{subject to : } d_{\mathrm{TV}}(P * \mathcal{N}(0,1), Q * \mathcal{N}(0,1)) \le \frac{0.1}{p} \\
&\text{variables : } P, Q \in \mathcal{M}([-1,1]).
\end{aligned}$$

This is a convex problem, but it doesn't mean that it is easily solvable; it is infinite-dimensional. Thus we are going to consider a sufficient condition replacing $d_{\mathrm{TV}}(P * \mathcal{N}(0,1), Q * \mathcal{N}(0,1)) \le \frac{0.1}{p}$.

Before explicitly stating the sufficient condition, consider the following problem called **moment problem**: Suppose two distributions have exactly same moments. Are these two distributions the same? Quick answer is no (when supported on $\mathbb{R}$), and yes (when supported on $[0,1]$). Rephrasing the latter one, if two distributions have equal first $1,000$ moments, then their relative distance should be very small, even after convolving with $\mathcal{N}(0,1)$ (think about the Taylor's expansion).

So the idea is to have the condition:

$$\mathbb{E}_P\theta^i = \mathbb{E}_Q\theta^i \quad i = 1, \ldots, d.$$

for a sufficiently large d to ensure $d_{\mathrm{TV}}(P * \mathcal{N}(0,1), Q * \mathcal{N}(0,1)) \le \frac{0.1}{p}$. It turns out that whenever the above condition is satisfied, we have $d_{\mathrm{TV}}(P * \mathcal{N}(0,1), Q * \mathcal{N}(0,1)) \approx \frac{1}{(c \cdot d!)}$ for some constant c, and thus it is sufficient to have $d \asymp \frac{\log p}{\log\log p}$. One way to show this is to use Taylor expansion and expand with Hermite polynomials.

Now, the revised problem is:

$$\begin{aligned}
&\text{maximize : } \mathbb{E}_Q|\theta| - \mathbb{E}_P|\theta| \\
&\text{subject to : } \mathbb{E}_P\theta^i = \mathbb{E}_Q\theta^i \quad i = 1, \ldots, d, \\
&\text{variables : } P, Q \in \mathcal{M}([-1,1]).
\end{aligned}$$

Note that the revised problem is again infinite-dimensional, but no more a convex programming but a linear programming.

25.2.3 Dual: again approximation of $|\theta|$ by degree-d polynomial

To solve the revised problem, we are going to show that it's dual problem is the approximation of $|\theta|$ by a degree-d polynomial q_d. More specifically, the dual problem is

$$2 \cdot E_d(|\theta|) = 2 \cdot \inf_{q_d \in \mathcal{P}_d} \sup_{\theta \in [-1,1]} |q_d(\theta) - |\theta||.$$

In fact, we are going to show more general result: for any f, $2 \cdot E_d(f) \ge \max_{P,Q} \left[\mathbb{E}_Q f(\theta) - \mathbb{E}_P f(\theta)\right]$ (we are showing only the useful direction). Note that the other direction, saying that the duality gap is zero, is not true for free since this is an infinite-dimensional problem.

The proof is essentially by Lagrangian. For each of d constraints of the primal problem, attach λ_i, $i = 1, \ldots, d$. Then, the Lagrangian is:

$$\begin{aligned}
L(P, Q, \lambda_1, \ldots, \lambda_d) &= \mathbb{E}_Q f(\theta) - \mathbb{E}_P f(\theta) - \sum_{i=1}^d \lambda_i (\mathbb{E}_Q\theta^i - \mathbb{E}_P\theta^i) \\
&= \mathbb{E}_Q\left[f(\theta) - \sum_{i=1}^d \lambda_i\theta^i\right] - \mathbb{E}_P\left[f(\theta) - \sum_{i=1}^d \lambda_i\theta^i\right].
\end{aligned}$$

Given the constants $\{\lambda_i\}$, the problem of choosing P, Q is merely a choosing of best θ, i.e.

$$\max_{P,Q} L(P, Q, \lambda_1, \ldots, \lambda_d) = \max_{\theta}[f(\theta) - q_d(\theta)] - \min_{\theta}[f(\theta) - q_d(\theta)]$$

where q_d is a polynomial using λ_i as coefficients (we played a little trick of adding and subtracting λ_0). Note that this value is greater than the solution of the primal solution. Now, minimizing over $\{\lambda_i\}$,

$$\min_{q_d \in \mathcal{P}_d} \left[\max_{\theta}[f(\theta) - q_d(\theta)] - \min_{\theta}[f(\theta) - q_d(\theta)] \right] = 2 \cdot \min_{q_d \in \mathcal{P}_d} \|f - q_d\|_\infty.$$

Using the results from the previous section on $E_d(|\theta|)$, we are done!

Part VI

Advanced topics

§ 26. Adaptation and Aggregation

26.1 Overview

We consider an estimation problem, where parameter set can be partitioned into smaller sets, each representing a different model. If the model that corresponds to the true parameter is unknown, can we design an adaptive estimator almost as good as the oracle estimator? See [AJ00, Yan00], etc. The material of this lecture is from the beautiful paper of Leung and Barron [LB06].

Let $R^*(\Theta) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}\ell(\theta, \hat{\theta})$, and $\Theta = \bigcup_m \Theta_m$, where m is the index for models. Examples include:

- *Sparse linear regression.* For sparse linear regression, $Y = X\theta + Z$, models can be indexed by: 1) support $m = \text{supp}(\theta)$, or 2) sparsity level $m = \|\theta\|_0$.
- *Density estimation.* For density estimation, the model index could be the smoothness parameter $m = \alpha$, i.e., $f \in \text{Hölder}(\alpha)$.

There are several flavors for this problem. Given oracle estimators $\{\hat{\theta}_m\}$, our goal is to find an adaptive estimator $\hat{\theta}$ that satisfies one of the following three oracle inequalities ($\forall \theta$):

(MS) $\mathbb{E}\ell(\theta, \hat{\theta}) \le \min_m \mathbb{E}\ell(\theta, \hat{\theta}_m) + \text{penalty}$.

(C) Convex combination: $\mathbb{E}\ell(\theta, \hat{\theta}) \le \min_{\hat{\theta} \in \text{co}(\{\theta_m\})} \mathbb{E}\ell(\theta, \hat{\theta}) + \text{penalty}$.

(L) Linear combination: $\mathbb{E}\ell(\theta, \hat{\theta}) \le \min_{\hat{\theta} \in \text{span}(\{\theta_m\})} \mathbb{E}\ell(\theta, \hat{\theta}) + \text{penalty}$.

This lecture focus on the model selection criterion.

The key for designing such an adaptive estimator is a unbiased risk estimator. Suppose $\hat{\gamma}_m = \hat{\gamma}_m(\text{data})$ is an unbiased risk estimator, i.e.,

$$\mathbb{E}_\theta[\hat{\gamma}_m] = R_\theta(\hat{\theta}_m) = \mathbb{E}_\theta[\ell(\theta, \hat{\theta}_m)].$$

Using $\hat{\gamma}$ as a proxy of the true risk, one can adopt one the following strategies for the adaptive estimator:

1. Pick the best one. Suppose $m^* = \arg\min_m \hat{\gamma}_m$, then output $\hat{\theta}_{m^*}$. An variation of this strategy is to split the samples: use the first part of the data to compute $\hat{\gamma}_m$ (the unbiased risk estimators), and the second part of the data to compute the "best" oracle estimator $\hat{\theta}_{m^*}$.
2. Mixture of all oracle estimators. Let $\hat{\theta} = \sum_m w_m \hat{\theta}_m$, where the weights w_m is larger if the risk proxy $\hat{\gamma}_m$ is smaller.

26.2 Stein's Unbiased Risk Estimator (SURE)

We discuss the Stein's unbiased risk estimator in this section.

Theorem 26.1 (Stein's Identity)**.** *Suppose $Z \sim N(0,1)$. For any continuously differentiable g,*

$$\mathbb{E}[g'(Z)] = \mathbb{E}[Zg(Z)],$$

provided both expectations are finite.

Proof. Integration by parts. □

Corollary 26.1. *Suppose $Y \sim N(\mu, 1)$. Under the same assumption of Theorem 26.1,*

$$\mathbb{E}[g'(Y)] = \mathbb{E}[(Y-\mu)g(Y)].$$

Corollary 26.2. *For one-dimensional GLM $Y \sim N(\mu,1)$, an unbiased risk estimator for $\hat{\mu}$ is*

$$\hat{\gamma} = (\hat{\mu} - Y)^2 + 2\hat{\mu}' - 1.$$

For p-dimensional GLM $Y \sim N(\mu, I_p)$, an unbiased risk estimator for $\hat{\mu}$ is

$$\hat{\gamma} = \|\hat{\mu} - Y\|^2 + 2\left\langle \nabla\hat{\mu}, \mathbf{1} \right\rangle - p.$$

Proof.

$$\mathbb{E}[\|\hat{\mu}-\mu\|^2] = \mathbb{E}[\|\hat{\mu}-Y\|^2] + \mathbb{E}[\|Y-\mu\|^2] + \mathbb{E}[2\left\langle Y-\mu, \hat{\mu}-Y \right\rangle].$$

Clearly, $\|\hat{\mu}-Y\|^2$ is an unbiased estimator of the first term, the second term equals p. By Corollary 26.1, the third term has an unbiased estimator: $2\sum_{i=1}^{p}(\partial_i\hat{\mu} - 1) = 2(\langle \nabla\hat{\mu}, \mathbf{1}\rangle - p)$. □

Example: SURE can be applied to the following GLM problem. The data follows $Y \sim N(\mu, I_n)$, where the mean μ belongs to a union of subspaces, $\mu \in \bigcup_m E_m$. An special case is sparse linear regression: $Y = X\theta + Z$. In this case, $m = \operatorname{supp}(\theta)$ and $E_m = \operatorname{span}(X_m)$. The oracle mean estimators is

$$\hat{\mu}_m = \operatorname{proj}_{E_m} Y = X_m(X_m'X_m)^{-1}X_m'Y.$$

We can derive the corresponding risk estimators $\hat{\gamma}_m$ using SURE. In the next section, we discuss how $\hat{\mu}_m$ and $\hat{\gamma}_m$ can be aggregated to form an adaptive estimator $\hat{\mu}$.

26.3 Main Result

In this section, we discuss one result using the second strategy in Section 26.1:

$$\hat{\mu} = \sum_m w_m\hat{\mu}_m,$$

where the weights w_m are computed from data, and satisfy $w_m \geq 0$ and $\sum_m w_m = 1$. In particular, we use the following exponential weighting:

$$w_m = \frac{1}{Z(\beta)}\pi_m \exp(-\frac{\beta}{2}\hat{\gamma}_m),$$

where the normalization constant is $Z(\beta) = \sum_m \pi_m \exp(-\frac{\beta}{2}\hat{\gamma}_m)$. The prior distribution satisfies $\pi_m \geq 0$ and $\sum_m \pi_m = 1$. The exponential weighting can be interpreted as follows:

- When $\beta = \infty$, $\hat{\mu} = \hat{\mu}_{m^*}$, where $m^* = \arg\min_m \hat{\gamma}_m$. This case reduces to the first strategy (without weighting). It fully relies on the data, and ignores the prior distribution.
- When $\beta = 0$, the aggregation uses the prior distribution only.
- When $0 < \beta < \infty$, e.g., $\beta = \frac{1}{2}, 1$, this corresponds to a Bayesian estimator.

Lemma 26.1 (Key Lemma)**.** *If $\hat{\mu}_m$ is linear in Y, then for every $\beta \le \frac{1}{2}$, we have*

$$\hat{\gamma} \le \sum_m w_m \hat{\gamma}_m.$$

Here, $\hat{\gamma}$ is the SURE for the weighted estimator $\hat{\mu}$, and $\hat{\gamma}_m$ is the SURE for the estimator $\hat{\mu}_m$, which is derived using model m.

Proof Sketch.

$$\hat{\gamma} = \sum_m w_m \hat{\gamma}_m - (1 - 2\beta) \cdot \sum_m w_m \|\mu - \hat{\mu}_m\|^2.$$

Clearly, $1 - 2\beta \ge 0$, and $\sum_m w_m \|\mu - \hat{\mu}_m\|^2 \ge 0$. □

Corollary 26.3 (Main Result)**.** *If $\hat{\theta}_m$ is linear in Y, and $\beta \le \frac{1}{2}$, then*

$$R_\theta(\hat{\theta}) \le \min_m \left\{ R_\theta(\hat{\theta}_m) + \frac{2}{\beta} \log \frac{1}{\pi_m} \right\}.$$

For the special case where $\pi_m = \frac{1}{M}$ and $\beta = \frac{1}{2}$, we have

$$R_\theta(\hat{\theta}) \le \min_m R_\theta(\hat{\theta}_m) + 4 \log M.$$

Proof. Recall that $w_m = \frac{1}{Z(\beta)} \pi_m \exp(-\frac{\beta}{2}\hat{\gamma}_m)$. By Lemma 26.1,

$$\begin{aligned}
\hat{\gamma} &\le \sum_m w_m \hat{\gamma}_m \\
&= \sum_m w_m \frac{2}{\beta} \log \frac{\pi_m}{w_m Z(\beta)} \\
&= \frac{2}{\beta} \sum_m w_m \log \frac{\pi_m}{w_m} + \frac{2}{\beta} \log \frac{1}{Z(\beta)} \\
&= -\frac{2}{\beta} D(w||\pi) + \frac{2}{\beta} \log \frac{1}{Z(\beta)} \\
&\le \frac{2}{\beta} \log \frac{1}{\max_m \pi_m \exp(-\frac{\beta}{2}\hat{\gamma}_m)},
\end{aligned}$$

where the last inequality is due to the fact that $D(w||\pi) \ge 0$ and $Z(\beta) = \sum_m \pi_m \exp(-\frac{\beta}{2}\hat{\gamma}_m) \ge \max_m \pi_m \exp(-\frac{\beta}{2}\hat{\gamma}_m)$. The corollary follows from taking expectation over both sides:

$$\begin{aligned}
R_\theta(\hat{\theta}) &= \mathbb{E}_\theta \hat{\gamma} \\
&\le \mathbb{E}_\theta \left\{ \min_m \left\{ \frac{2}{\beta} \log \frac{1}{\pi_m} + \hat{\gamma}_m \right\} \right\} \\
&\le \min_m \left\{ \mathbb{E}_\theta \hat{\gamma}_m + \frac{2}{\beta} \log \frac{1}{\pi_m} \right\} \\
&= \min_m \left\{ R_\theta(\hat{\theta}_m) + \frac{2}{\beta} \log \frac{1}{\pi_m} \right\}.
\end{aligned}$$

□

Example: In sparse linear regression, let $\Theta = \mathbb{R}^p$, $m = \text{supp}(\theta) \subset [p]$, and $X = I_p$. Clearly, uniform prior is not good enough. If we choose prior $\pi_m = \frac{1}{\binom{p}{|m|}(p+1)}$, then

$$R_\theta(\hat{\theta}) \le \|\theta\|_0 + 4\log\binom{p}{\|\theta\|_0} + 4\log(1+p).$$

An even better prior is $\pi_m \propto \frac{1}{\binom{p}{|m|}e^{|m|+1}}$.

26.4 Bayesian origin of exponential weighting

This section presents an explanation for the exponential weighting ($\beta = 1$).

Suppose the prior distribution on m is π_m, and the prior distribution of μ given m is $\lambda_m(\mu)$. Hence, the prior distribution of μ is $\sum_m \pi_m \lambda_m(\mu)$. The Bayesian estimators are

$$\hat{\mu} = \mathbb{E}[\mu|Y], \quad \hat{\mu}_m = \mathbb{E}[\mu|Y, m].$$

Hence

$$\hat{\mu} = \mathbb{E}[\mu|Y] = \mathbb{E}[\mathbb{E}[\mu|Y, m]|Y] = \sum_m \hat{\mu}_m P(m|Y).$$

By the Bayes rule, the correct weights are

$$w_m = P(m|Y) = \frac{\pi_m P(Y|m)}{\sum_m \pi_m P(Y|m)}. \tag{26.1}$$

Next, we show that $P(Y|m)$ has an exponential form, establishing a motivation for exponential weighting. Suppose $\lambda_m(\mu)$ is a Gaussian prior (e.g., $N(0, sP_m)$). Then we have

$$P(y|m) = \int \phi(y-\mu)\lambda_m(\mathrm{d}\mu),$$

where $\phi(\cdot)$ is the Gaussian PDF. The convolution of two Gaussian PDF is still a Gaussian PDF of the form

$$P(y|m) = c\exp(-\frac{1}{2}\|y-\mu_m\|^2). \tag{26.2}$$

If $\hat{\mu}_m$ is linear in Y, then the SURE satisfies

$$\hat{\gamma}_m = \|\hat{\mu}_m - Y\|^2 + 2\langle\nabla\hat{\mu}_m, \mathbf{1}\rangle - p = \|\hat{\mu}_m - Y\|^2 + \text{const}. \tag{26.3}$$

Therefore, comparing (26.1), (26.2) and (26.3), we conclude that the weighting deduced from Bayesian estimators is

$$w_m \propto \pi_m P(Y|m) \propto \pi_m \exp(-\frac{1}{2}\hat{\gamma}_m),$$

which corresponds to the exponential weighting with $\beta = 1$.

Bibliography

[AJ00] Arkadii Nemirovski Anatoli Juditsky. Functional aggregation for nonparametric regression. *The Annals of Statistics*, 28(3):681–712, 2000.

[AMS04] Shiri Artstein, Vitali Milman, and Stanisław J Szarek. Duality of metric entropy. *Annals of mathematics*, pages 1313–1328, 2004.

[Bir83] L. Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 65(2):181–237, 1983.

[Cha05] Sourav Chatterjee. An error bound in the Sudakov-Fernique inequality. *arXiv preprint arXiv:0510424*, 2005.

[Che56] Herman Chernoff. Large-sample theory: Parametric case. *The Annals of Mathematical Statistics*, 27(1):1–22, 1956.

[CL11] T. T. Cai and M. G. Low. Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *The Annals of Statistics*, 39(2):1012–1041, 2011.

[Csi67] I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Sci. Math. Hungar.*, 2:229–318, 1967.

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of information theory, 2nd Ed.* Wiley-Interscience, New York, NY, USA, 2006.

[CT07] Emmanuel Candés and Terence Tao. The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351, 2007.

[Dud67] Richard M. Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290 – 330, 1967.

[Hoe63] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[HV11] P. Harremoës and I. Vajda. On pairs of f-divergences and their joint range. *IEEE Trans. Inf. Theory*, 57(6):3230–3235, Jun. 2011.

[Koz47] W. Kozakiewicz. On the convergence of sequences of moment generating functions. *Annals of Mathematical Statistics*, pages 61–69, 1947.

[LB06] Gilbert Leung and Andrew R Barron. Information theory and mixing least-squares regressions. *IEEE Trans. Inf. Theory*, 52(8):3396–3410, 2006.

[LC86] Lucien Le Cam. *Asymptotic methods in statistical decision theory.* Springer-Verlag, New York, NY, 1986.

[LNS99] Oleg Lepski, Arkady Nemirovski, and Vladimir Spokoiny. On estimation of the L_r norm of a regression function. *Probability Theory and Related Fields*, 113(2):221–253, 1999.

[Pan08] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Trans. Inf. Theory*, 54(10):4750–4755, 2008.

[Pis99] G. Pisier. *The volume of convex bodies and Banach space geometry.* Cambridge University Press, 1999.

[PPV10] Y. Polyanskiy, H. V. Poor, and S. Verdú. Channel coding rate in the finite blocklength regime. *IEEE Trans. Inf. Theory*, 56(5):2307–2359, May 2010.

[PW15] Y. Polyanskiy and Y. Wu. Lecture notes on information theory. Feb 2015. http://www.ifp.illinois.edu/~yihongwu/teaching/itlectures.pdf.

[Rin76] Yosef Rinott. On convexity of measures. *The Annals of Probability*, 4(6):1020–1026, 1976.

[SC15] Weijie Su and Emmanuel Candés. SLOPE is adaptive to unknown sparsity and asymptotically minimax. *arXiv preprint arXiv:1503.08393*, 2015.

[Tsy09] A. B. Tsybakov. *Introduction to Nonparametric Estimation.* Springer Verlag, New York, NY, 2009.

[Vaj70] Igor Vajda. Note on discrimination information and variation (corresp.). *IEEE Transactions on Information Theory*, 16(6):771–773, 1970.

[vdV02] Aad van der Vaart. The statistical work of Lucien Le Cam. *The Annals of Statistics*, pages 631–682, 2002.

[Yan00] Yuhong Yang. Combining different procedures for adaptive regression. *Journal of Multivariate Analysis*, 74(1):135–161, Jul 2000.

[YB99] Y. Yang and A. R. Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999.