High Dimensional Probability and Statistics **2nd Semester, 2021-2022**

Lecture 9: Minimax Lower Bounds

Instructor: Ke Wei *Scribe: Ke Wei (Updated: 2022/05/15)*

Motivation: Consider a set of probability distributions define on $\mathcal{X}$ and indexed by Θ, denoted $\mathcal{P} = \{\mathbb{P}_\theta,\ \theta \in \Theta\}$. For example, θ can denote certain parameter of a distribution or the corresponding probability density function. Given a set of i.i.d data $(X_1, \cdots, X_n)$ sampled from $\mathbb{P}_\theta$ where θ is not known a prior, a fundamental statistical problem is to estimate θ from $\mathcal{D}$.

Let $\widehat{\theta} : (X_1, \cdots, X_n) \to \Theta$ be an estimation procedure. The concentration inequalities and other probability tools presented earlier can help establish an upper bound of the estimation error in terms of[1]

$$\Phi\left(\rho\left(\widehat{\theta}, \theta\right)\right),$$

where $\rho(\cdot, \cdot)$ is a (semi)metric defined on Θ and $\Phi : [0, \infty) \to [0, \infty)$ is an increasing function. As an example, for a univariate mean estimation problem, $\rho(\theta, \theta') = |\theta - \theta'|$ and $\Phi(t) = t^2$ yields the squared error. On the other hand, it is worth investigating whether the estimation error of $\widehat{\theta}$ is optimal. To this end, we study the lower bound of the estimation error based on the *minimax risk*, defined by

$$\mathfrak{M}_n(\Theta) = \inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[\Phi\left(\rho\left(\widehat{\theta}, \theta\right)\right)\right], \tag{9.1}$$

where the subscript θ means that $X_1, \cdots, X_n$ are sampled from $\mathbb{P}_\theta$. That is, for a fixed estimation procedure we consider the worst case error by taking the supremum over all the distributions, and then study the smallest worst case error achievable by *any procedure*.

There are two methods for obtaining the minimax lower bound: Bayesian analysis and reduction to hypothesis testing. We will focus on the latter one since it is more versatile and can be applied to most situations. *To gain some intuition of the hypothesis testing method,consider the minimax risk of estimating a scalar parameter in terms of the risk function $|\theta - \theta'|$. Suppose there are two point θ_1 and θ_2 such that $|\theta_1 - \theta_2| \geq \delta$. If whatever method we use to test which point the observed data comes from the probability of testing error is a constant, then the estimation error for any procedure should be greater than a multiple of δ since with constant probability we are likely to mistaken one from the other.* Of course we can also consider the problem of testing multiple points. Thus, overall the problem is about how to choose the testing points such that they are as far away as possible while the probability of testing error for any testing method remains a constant.

In this lecture we discuss two standard techniques for establishing the lower bounds of the minimax risks based on testing, including the Le Cam and Fano methods. Roughly speaking, Le Cam method is based on binary testing and Fano methods are based on multiway hypothesis testing. There is another method which is not covered in this lecture, known as Assouad method, for lower bounding the minimax risk. Assouad method is based on the multiple binary hypothesis testing when the risk function is separable, see for example Chapter 8 and 9 of [2].

Agenda:

[1] We use $\widehat{\theta}$ to denote $\widehat{\theta}(X_1, \cdots, X_n)$ for simplicity.

- Reduction to hypothesis testing
- Some divergence measures
- Le Cam method
- Fano methods

9.1 Reduction to Hypothesis Testing

Let $\{\theta_1, \cdots, \theta_m\}$ be a 2δ-packing of the space Θ under the (semi)metric ρ, i.e., $\rho(\theta_i, \theta_j) \geq 2\delta$ for all $i \geq j$. Define $\mathbb{P}_j^n = \mathbb{P}_{\theta_j} \times \cdots \times \mathbb{P}_{\theta_j}$. First, by the Markov inequality we have

$$\mathbb{E}_{\theta_j}\left[\Phi\left(\rho(\widehat{\theta}, \theta_j)\right)\right] \geq \Phi(\delta) \cdot \mathbb{P}_j^n\left[\Phi\left(\rho(\widehat{\theta}, \theta_j)\right) \geq \Phi(\delta)\right] \geq \Phi(\delta) \cdot \mathbb{P}_j^n\left[\rho(\widehat{\theta}, \theta_j) \geq \delta\right],$$

where we note that $\widehat{\theta} = \widehat{\theta}(X_1, \cdots, X_n)$, and $\mathbb{P}_j^n$ indicates that $(X_1, \cdots, X_n)$ are sampled from $\mathbb{P}_{\theta_j}$. In addition, the second inequality is due to the fact that Φ is increasing. It follows that

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta\left[\Phi\left(\rho\left(\widehat{\theta}, \theta\right)\right)\right] \geq \max_{\theta_j} \mathbb{E}_{\theta_j}\left[\Phi\left(\rho(\widehat{\theta}, \theta_j)\right)\right] \geq \Phi(\delta)\left(\frac{1}{m}\sum_{j=1}^m \mathbb{P}_j^n\left[\rho(\widehat{\theta}, \theta_j) \geq \delta\right]\right).$$

Next we will show $\frac{1}{m}\sum_{j=1}^m \mathbb{P}_j^n\left[\rho(\widehat{\theta}, \theta) \geq \delta\right]$ can be lower bounded by hypothesis testing error.

In the hypothesis testing, a test function is a map from a set of i.i.d data sampled from one of $\{\mathbb{P}_{\theta_j},\ j = 1, \cdots, m\}$ to $\{1, \cdots, m\}$, which is used to infer from which probability distribution the data comes from. Given an estimation procedure $\widehat{\theta}$, we can define a test function naturally as follows:

$$\widehat{\Psi}(X_1, \cdots, X_n) = \arg\min_{\ell \in [m]} \rho(\widehat{\theta}(X_1, \cdots, X_n), \theta_\ell),$$

where the tier is broken arbitrarily. Since $\{\theta_1, \cdots, \theta_m\}$ is a 2δ-packing of Θ, it is clear that (see Figure 9.1)

$$\rho(\widehat{\theta}, \theta_j) < \delta \quad \Rightarrow \quad \widehat{\Psi} = j.$$

Thus, when $(X_1, \cdots, X_n)$ are sampled from $\mathbb{P}_{\theta_j}$, we have[2]

$$\mathbb{P}_j^n\left[\widehat{\Psi} \neq j\right] \leq \mathbb{P}_j^n\left[\rho(\widehat{\theta}, \theta_j) \geq \delta\right].$$

Consequently,

$$\frac{1}{m}\sum_{j=1}^m \mathbb{P}_j^n\left[\rho(\widehat{\theta}, \theta_j) \geq \delta\right] \geq \frac{1}{m}\sum_{j=1}^m \mathbb{P}_j^n\left[\widehat{\Psi} \neq j\right].$$

[2]We also use $\widehat{\Psi}$ to denote $\widehat{\Psi}(X_1, \cdots, X_n)$ for simplicity.

Moreover, we have

$$\sup_{\theta\in\Theta}\mathbb{E}_\theta\left[\Phi\left(\rho\left(\widehat{\theta},\theta_j\right)\right)\right]\geq\Phi(\delta)\left(\frac{1}{m}\sum_{j=1}^m\mathbb{P}_j^n\left[\widehat{\Psi}\neq j\right]\right).$$

Taking the infimum over all estimation procedures $\widehat{\theta}$ on the lefthand side and the infimum over all test functions yields the following proposition.

Proposition 9.1 *Under the setup of the above test problem, the minimax risk* (9.1) *is lower bounded as*

$$\mathfrak{M}_n(\Theta)\geq\Phi(\delta)\inf_{\Psi}\left(\frac{1}{m}\sum_{j=1}^m\mathbb{P}_j^n\left[\Psi(X_1,\cdots,X_n)\neq j\right]\right),\tag{9.2}$$

where the infimum ranges over all test functions. Note that δ is parameter that is free to choose and it denotes the minimum distance between θ_i and θ_j for all $i\neq j$.

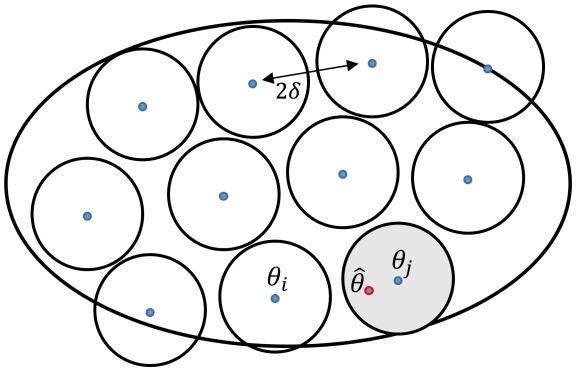

Figure 9.1: An illustration of 2δ-packing.

Consider a joint distribution (J,Z^J), where J is uniform distributed in $\{1,\cdots,m\}$ and given $J=j$, $Z^j=(X_1,\cdots,X_n)$ obeys the distribution of $\mathbb{P}_j^n$. It is clear that the joint distribution obeys

$$\mathbb{Q}\left[Z^J\in\cdot,J=j\right]=\frac{1}{m}\mathbb{P}_j^n\left[Z^j\in\cdot\right],$$

and the marginal distributions are given by

$$\mathbb{Q}_J\left[J=j\right]=\frac{1}{m}\quad\text{and}\quad\mathbb{Q}_Z\left[Z^J\in\cdot\right]=\frac{1}{m}\sum_{j=1}^m\mathbb{P}_j^n\left[Z^j\in\cdot\right].$$

Moreover, for any test function Ψ, we have

$$\mathbb{Q}\left[\Psi(Z^J)\neq J\right]=\sum_{j=1}^m\mathbb{Q}\left[\Psi(Z^J)\neq J,J=j\right]$$

$$= \sum_{j=1}^{m} \mathbb{Q}\left[\Psi(Z^J) \neq j, J = j\right]$$
$$= \frac{1}{m} \sum_{j=1}^{m} \mathbb{P}_j^n(\Psi(Z^j) \neq j).$$

Therefore, we can rewrite (9.2) as

$$\mathfrak{M}_n(\Theta) \geq \Phi(\delta) \inf_{\Psi} \mathbb{Q}\left[\Psi(Z^J) \neq J\right], \tag{9.3}$$

which will be used in the sequel for conciseness.

Remark 9.2 *In words, reduction to hypothesis testing lower bounds the best achievable estimation error by a multiple of the failing probability of test. It is not hard to imagine that the smallest mis-test probability fundamentally relies on how close $\mathbb{P}_j^n$ are, which enables us to provides a bound independent of the test function. Moreover, the lower bound in* (9.2) *or* (9.3) *is a function of the separation δ, which trades off between $\Phi(\delta)$ (increases as δ increases) and the probability of test error $\inf_\Psi \mathbb{Q}\left[\Psi(Z^J) \neq J\right]$ (relying on δ implicitly, decreases as δ increases).* In order to obtain a desirably large lower bound, one usually chooses the largest[3] δ such that $\inf_\Psi \mathbb{Q}\left[\Psi(Z^J) \neq J\right]$ is greater than a constant[4] (for example 1/2) and then uses the corresponding $\Phi(\delta)$ to provide lower bound. *As we have explained in the motivation part, the intuition is that if the parameters are far away (i.e, by choosing the largest possible δ) but it is still difficult to distinguish the related distributions from the observations (i.e., probability of testing error is constant), then the estimation error must be lower bounded by related function of the parameter distance since we can mistaken one for the other. Next, we will present two concrete methods: the Le Cam and Fano methods.*

9.2 Some Divergence Measures

We first take a detour and present some inequalities for divergence measures and their consequences for product distributions. Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability distributions defined on $\mathcal{X}$. Assume they have densities $p(x)$ and $q(x)$ respectively with respect to some underlying base measure μ. The three related divergences are

- KL divergence: $D(\mathbb{Q}\|\mathbb{P}) = \int_{\mathcal{X}} q(x) \log \frac{q(x)}{p(x)} \mu(dx)$,
- TV distance: $\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} = \sup_{A \subset \mathcal{X}} |\mathbb{P}(A) - \mathbb{Q}(A)| = \frac{1}{2} \int_{\mathcal{X}} |p(x) - q(x)| \mu(dx)$,
- Hellinger distance: $H^2(\mathbb{P}\|\mathbb{Q}) = \int_{\mathcal{X}} \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \mu(dx)$.

Recall that KL divergence and TV distance have also been mentioned in Lecture 3. The three divergence measures are related as follows.

Lemma 9.3 *For two distributions $\mathbb{P}$ and $\mathbb{Q}$, we have*

1. $\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} \leq \sqrt{\frac{1}{2} D(\mathbb{Q}\|\mathbb{P})}$,

[3] As can be seen δ may rely on other parameters, such as the number of samples.

[4] That is, choose the largest possible δ that the testing problem is still sufficiently challenging.

2. $\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} \leq H(\mathbb{P}\|\mathbb{Q})\sqrt{1 - \frac{H^2(\mathbb{P}\|\mathbb{Q})}{4}}$.

Proof: The proof for the first inequality can be found in Lecture 3. The second inequality can be proved by the Cauchy-Schwarz inequality (**check this!**). ■

Recall that $\mathbb{P}^n$ (respectively, $\mathbb{Q}^n$) is the product distribution on the product space $\mathcal{X}^n$ (i.e., the distribution of n i.i.d random variables). It is desirable to express the distance between $\mathbb{P}^n$ and $\mathbb{Q}^n$ in terms of $\mathbb{P}$ and $\mathbb{Q}$. For TV distance, it is difficult to express $\|\mathbb{P}^n - \mathbb{Q}^n\|_{\mathrm{TV}}$ in terms of $\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}}$. For KL divergence and the Hellinger distance we have the following lemma.

Lemma 9.4 *For two distributions $\mathbb{P}$ and $\mathbb{Q}$ and the corresponding , we have*

1. $D(\mathbb{Q}^n\|\mathbb{P}^n) = nD(\mathbb{Q}\|\mathbb{P})$,
2. $H^2(\mathbb{P}^n\|\mathbb{Q}^n) \leq nH^2(\mathbb{P}\|\mathbb{Q})$.

Proof: The first inequality can be proved directly using the fact that the density functions for $\mathbb{P}^n$ and $\mathbb{Q}^n$ are $p(x_1)\cdots p(x_n)$ and $q(x_1)\cdots q(x_n)$ respectively. Additionally, it can be shown that (**check this!**)

$$\frac{1}{2}H^2(\mathbb{P}^n\|\mathbb{Q}^n) = 1 - \left(1 - \frac{1}{2}H^2(\mathbb{P}\|\mathbb{Q})\right)^n.$$

Then the second inequality follows immediately since $(1-x)^n \geq 1 - nx$ for $x \in [0,1]$. ■

9.3 Le Cam Method

Le Cam method provides lower bounds on the minimax using the simple binary hypothesis testing. This section explores this connection based on the total variation distance.

Lemma 9.5 *In the case of binary hypothesis testing, we have*

$$\inf_{\Psi} \mathbb{Q}\left[\Psi(Z^J) \neq J\right] = \frac{1}{2}\left(1 - \|\mathbb{P}_1^n - \mathbb{P}_2^n\|_{\mathrm{TV}}\right),$$

where $\mathbb{P}_1^n$ and $\mathbb{P}_2^n$ are product distributions corresponding to θ_1 and θ_2, respectively.

Proof: For any test function Ψ defined on $\mathcal{X}^n$, let

$$A = \{(x_1, \cdots, x_n) \in \mathcal{X}^n : \ \Psi(x_1, \cdots, x_n) = 1\}.$$

and A^c be the complementary on which $\Psi = 2$. Then we have

$$\begin{aligned}
\sup_{\Psi} \mathbb{Q}\left[\Psi(Z^J) = J\right] &= \sup_{A} \frac{1}{2}\left(\mathbb{P}_1^n[A] + \mathbb{P}_2^n[A^c]\right) \\
&= \frac{1}{2} + \frac{1}{2}\sup_{A}\left(\mathbb{P}_1^n[A] - \mathbb{P}_2^n[A]\right) \\
&= \frac{1}{2} + \frac{1}{2}\|\mathbb{P}_1^n - \mathbb{P}_2^n\|_{\mathrm{TV}}.
\end{aligned}$$

Noting that $\sup_{\Psi} \mathbb{Q}\left[\Psi(Z^J) = J\right] = 1 - \inf_{\Psi} \mathbb{Q}\left[\Psi(Z^J) \neq J\right]$, the claim follows immediately. ■
Combining the above lemma and Proposition 9.1 together yields the following minimax risk bound.

Proposition 9.6 *We have*

$$\mathfrak{M}_n(\Theta) \geq \frac{\Phi(\delta)}{2}\left(1 - \|\mathbb{P}_1^n - \mathbb{P}_2^n\|_{\mathrm{TV}}\right)$$

for ***any pair*** *of distributions θ_1 and θ_2 satisfying $\rho(\theta_1, \theta_2) \geq 2\delta$.*

Note that as δ decreases $\|\mathbb{P}_1^n - \mathbb{P}_2^n\|_{\mathrm{TV}}$ decreases, and the binary hypothesis testing problem becomes more challenging. In practice, we roughly attempt to choose the largest possible δ such that $\|\mathbb{P}_1^n - \mathbb{P}_2^n\|_{\mathrm{TV}}$ is a small constant so that we can still mistaken the θ_1 and θ_2 (yielding the lower bound of the estimation error depending on δ) .

Example 9.7 *Let $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \mathbb{R}\}$ be a family of normal distributions $\mathcal{N}(\theta, \sigma^2)$ with fixed variance σ^2. We study the minimax risk of estimating θ from i.i.d samples $\{X_k\}_{k=1}^n$ drawn from $\mathbb{P}_\theta$. We consider two parameters $\theta_1 = 0$ and $\theta_2 = \theta$ satisfying $\theta = 2\delta$. In order to apply the Le Cam method, we need to bound $\|\mathbb{P}_\theta^n - \mathbb{P}_0^n\|_{\mathrm{TV}}$. Given two probability distributions $\mathbb{P}$ and $\mathbb{Q}$ defined over $\mathcal{X}$, respectively with their probability densities $p(x)$ and $q(x)$ under some base measure μ, it can be easily shown that (****check this!****)*

$$\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}}^2 \leq \frac{1}{4}\left(\int_{\mathcal{X}} \frac{p^2(x)}{q(x)}\mu(dx) - 1\right).$$

Using this result for $\mathbb{P}_0^n$ and $\mathbb{P}_\theta^n$ on $\mathcal{X} = \mathbb{R}^n$ yields that

$$\|\mathbb{P}_\theta^n - \mathbb{P}_0^n\|_{\mathrm{TV}}^2 \leq \frac{1}{4}\left(\exp\left(n\theta^2/\sigma^2\right) - 1\right) = \frac{1}{4}\left(\exp\left(4n\delta^2/\sigma^2\right) - 1\right).$$

Taking $\delta = \frac{1}{2}\frac{\sigma}{\sqrt{n}}$ yields that

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta\left[|\widehat{\theta} - \theta|^2\right] \geq \frac{\delta^2}{2}\left(1 - \sqrt{e-1}/2\right) \geq \frac{\delta^2}{6} = \frac{1}{24}\frac{\sigma^2}{n}.$$

*The scale σ^2/n is sharp, and the sample mean $\widehat{\theta} = \frac{1}{n}\sum_{k=1}^n X_k$ satisfies this bound (****check this!****).*

9.4 Fano Methods

The Fano methods provide lower bounds based on the multiway hypothesis testing and the Fano inequality in information theory.

9.4.1 Information Theory Basics

Information theory is essentially about studying the information or randomness stored in probability distributions. Here we provide some basic materials in information theory that is needed for lower bounding the minimax risk. More details about information can be found in the book *Elements of Information Theory*. The fundamental notion in information theory is Shannon entropy.

Definition 9.8 (Shannon entropy) *Let $X \sim \mathbb{Q}$ where $\mathbb{Q}$ is a probability distribution on $\mathcal{X}$ with density $q(x)$ with respect to some base measure μ. The Shannon entropy of[5] X is*

$$H(X) = -\int_{\mathcal{X}} q(x) \log q(x) \mu(dx). \tag{9.4}$$

Lets gain some intuition of Shannon entropy by looking at discrete random variables. When X is a discrete random variable, we can take $\mathcal{X}$ as a finite set and take μ as a counting measure on $\mathcal{X}$. In this case, the definition (9.4) reduces to the discrete entropy[6]

$$H(X) = -\sum_{x \in \mathcal{X}} q(x) \log q(x). \tag{9.5}$$

It measures on average how many bits are needed to store the probability distribution of $\mathcal{X}$. More precisely, to represent the probability for $X = x$ (i.e., $q(x)$), we need $\log 1/(q(x))$ bits since it corresponds to $1/q(x)$ possibilities. Thus, on average we need $H(x)$ bits to store the distribution of X. In the simplest uniform distribution case $q(x) = 1/|\mathcal{X}|$, $\log |\mathcal{X}|$ bits are needed to denote all $|\mathcal{X}|$ possibilities.

Lemma 9.9 *For **discrete entropy**, we have $0 \leq H(X) \leq \log |\mathcal{X}|$.*

It is worth noting that for differential entropy (i.e., entropy of continuous random variables), $H(X) \geq 0$ is not always true since $q(x)$ can be greater than 1 (for example consider a uniform distribution over a small interval). The upper bound $\log |\mathcal{X}|$ is achieved by the uniform distribution on $\mathcal{X}$, i.e., $\mathbb{Q}(X = x) = \frac{1}{|\mathcal{X}|}$.

Proof: The lower bound $H(X) \geq 0$ follows from $q(x) \leq 1$ and the upper bound follows from Jensen inequality. ■

We can also define the conditional entropy, which is the amount of information left in a random variable after observing another.

Definition 9.10 (Conditional entropy) *Given a pair of random variables(X, Y) on $(\mathcal{X}, \mathcal{Y})$ with joint distribution $\mathbb{Q}_{X,Y}$, the conditional entropy of $X|Y$ is defined as*

$$H(X|Y) = \mathbb{E}_Y \left[-\int_{\mathcal{X}} q(x|Y) \log q(x|Y) \mu(dx) \right].$$

In addition, given two random variables, we can define the mutual information between them.

Definition 9.11 (Mutual information) *Given a pair of random variables(X, Y) on $(\mathcal{X}, \mathcal{Y})$ with joint distribution $\mathbb{Q}_{X,Y}$, let $\mathbb{Q}_X$ and $\mathbb{Q}_Y$ denote the respect marginal distributions. The mutual information of X and Y is defined as*

$$I(X, Y) = D(\mathbb{Q}_{X,Y} \| \mathbb{Q}_X \mathbb{Q}_Y).$$

[5] Shannon entropy is actually a function of probability distributions since there are many different random variables obeying the same distribution. Here we just follow the standard practice in information theory and treat it as a function of random variables.

[6] Note that, for continuous random variables, the Shannon entropy is often referred to as the differential entropy.

We first note that $I(X,Y) \geq 0$, and $I(X,Y) = 0$ if and only if X and Y are independent. Thus, it can be thought as a way to measure the amount of dependence between X and Y. When X and Y are independent, $I(X;Y) = 0$.

We have the following properties about entropy, conditional entropy and mutual information.

Lemma 9.12 *We have*

1. $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$,
2. $H(X,Y|Z) = H(X|Z) + H(Y|X,Z) = H(Y|Z) + H(X|Y,Z)$,
3. $I(X,Y) = H(X) + H(Y) - H(X,Y)$
4. $I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$,
5. $H(X|Y) \leq H(X)$, $H(Y|X) \leq H(Y)$,
6. $H(Y|X) = 0$ *if* $Y = f(X)$, *i.e., when* Y *is a function of* X.

Proof: Whenever it is possible, we will assume the existence of the (conditional) density functions in the proofs for conciseness.

The first two identities are known as the chain rule for entropy. We only prove the first equality in 1 and 2 since the other two can be proved similarly. Noting that $q(y|x) = \frac{q(x,y)}{q(x)}$, we have

$$\begin{aligned} H(Y|X) &= -\int_{\mathcal{X}} \left(\int_{\mathcal{Y}} q(y|x) \log q(y|x) \mu(dy) \right) q(x) \mu(dx) \\ &= -\int_{\mathcal{X}} \left(\int_{\mathcal{Y}} \frac{q(x,y)}{q(x)} \log \frac{q(x,y)}{q(x)} \mu(dy) \right) q(x) \mu(dx) \\ &= H(X,Y) - H(X). \end{aligned}$$

Similarly, noting that $q(y|x,z) = \frac{q(x,y,z)}{q(x,z)} = \frac{q(x,y|z)q(z)}{q(x,z)} = \frac{q(x,y|z)}{q(x|z)}$ and $q(x,z) = q(x|z)q(z)$, we have

$$\begin{aligned} H(Y|X,Z) &= -\int_{\mathcal{X}} \int_{\mathcal{Z}} \left(\int_{\mathcal{Y}} \frac{q(x,y|z)}{q(x|z)} \log \frac{q(x,y|z)}{q(x|z)} \mu(dy) \right) q(x|z) q(z) \mu(dx) \mu(dz) \\ &= -\int_{\mathcal{X}} \int_{\mathcal{Z}} \left(\int_{\mathcal{Y}} q(x,y|z) \log q(x,y|z) \mu(dy) \right) q(z) \mu(dx) \mu(dz) \\ &\quad + \int_{\mathcal{X}} \int_{\mathcal{Z}} \left(\int_{\mathcal{Y}} q(x,y|z) \log q(x|z) \mu(dy) \right) q(z) \mu(dx) \mu(dz) \\ &= -\int_{\mathcal{X}} \int_{\mathcal{Z}} \left(\int_{\mathcal{Y}} q(x,y|z) \log q(x,y|z) \mu(dy) \right) q(z) \mu(dx) \mu(dz) \\ &\quad + \int_{\mathcal{Z}} \left(\int_{\mathcal{X}} \left(\int_{\mathcal{Y}} q(x,y|z) \mu(dy) \right) \log q(x|z) \mu(dx) \right) q(z) \mu(dz) \\ &= H(X,Y|Z) - H(X|Z). \end{aligned}$$

Expanding the expression for $I(X,Y)$,

$$I(X,Y) = \int_{\mathcal{Y}} \int_{\mathcal{X}} q(x,y) \log \frac{q(x,y)}{q(x)q(y)} \mu(dx) \mu(dy),$$

yields 3 straightforwardly.

Combining 1 and 2 together yields 4, and 5 follows from 4 directly. Note that 5 means the conditional entropy is always less than or equal to the entropy. That is, considering the entropy under certain condition only decreases the uncertainty of a random variable. Moreover, if X and Y are independent, then $H(X|Y) = H(X)$, so in this situation observing Y will not reduce the uncertainty in X.

When $Y = f(X)$, $Y|(X = x)$ is deterministic or it is a discrete random variable only taking one value at $f(x)$. Evidently, we have $H(Y|X) = 0$. This means there is no uncertainty in Y once X is observed and hence $H(Y|X) = 0$. ■

Now we are ready to present and prove the Fano inequality in information theory. Let X be random variable on a finite set $\mathcal{X}$. Assume we observe a different random variable Y, and want to estimate $\mathbb{Q}[\Psi(Y) \neq X]$, where $\mathbb{Q}$ is the joint distribution of X and Y, and $\Psi(\cdot)$ is a test function.

Lemma 9.13 (Fano inequality) *We have*

$$\mathbb{Q}[\Psi(Y) \neq X] \geq \frac{H(X|Y) - \log 2}{\log |\mathcal{X}|}. \tag{9.6}$$

Proof: Let E be the random variable such that $E = 1$ if $\Psi(Y) \neq X$ and $E = 0$ otherwise. The proof follows by expanding $H(X, E|Y)$ in two different ways given in 2 of Lemma 9.12.

Letting $h = -p \log p - (1 - p) \log(1 - p)$, we have

$$\begin{aligned} H(X, E|Y) &= H(E|Y) + H(X|E, Y) \\ &= \underbrace{H(E|Y)}_{\leq H(E)} + \underbrace{\mathbb{Q}[E = 1] H(X|E = 1, Y)}_{\leq \mathbb{Q}[E=1] \log(|\mathcal{X}|-1)} + \underbrace{\mathbb{Q}[E = 0] H(X|E = 0, Y)}_{=0} \\ &\leq h(\mathbb{Q}[\Psi(Y) \neq X]) + \mathbb{Q}[\Psi(Y) \neq X] \log(|\mathcal{X}| - 1), \end{aligned}$$

where we have used the fact that conditioned on $E = 1, Y = y$, X can only take $|\mathcal{X}| - 1$ possible values and conditioned on $E = 0, Y = y$, $X = \Psi(y)$ is deterministic. On the other hand,

$$H(X, E|Y) = H(X|Y) + H(E|X, Y) = H(X|Y),$$

where $H(E|X, Y) = 0$ due to 6 of Lemma 9.12. Combining the above two inequalities together and further noting $h(p) \leq \log 2$ for all $p \in [0, 1]$ concludes the proof. ■

9.4.2 Fano Lower Bound on Minimax Risk

Recall that the minimax risk can be lower bounded by $\Phi(\delta) \inf_\Psi \mathbb{Q}\left[\Psi(Z^J) \neq J\right]$, where the random variable Z^J is generated by first sampling J uniformly from $[m] = \{1, \cdots, m\}$ and then generating Z^J according to $\mathbb{P}_j^n$ (here $\mathbb{P}_j^n, j = 1, \cdots, m$ are the product distributions which corresponds to the 2δ-separated set $\{\theta_j\}_{j=1}^m$), see Section 9.1 for details. Intuitively, $\mathbb{Q}\left[\Psi(Z^J) \neq J\right]$ should relate to the dependence between Z^J and J. For example, if Z^J is independent of J, it would be impossible to tell J from Z^J. Since $I(Z^J, J)$ provides one way to characterize the dependence between Z^J and J in terms of the KL divergence. It is reasonable to bound $\mathbb{Q}\left[\Psi(Z^J) \neq J\right]$ by $I(Z^J, J)$ and then provide a minimax lower bound based on it. Indeed, we have the following theorem.

Theorem 9.14 *Under the setting for the construction of J and Z^J in Section 9.1, we have*

$$\mathfrak{M}_n(\Theta) \geq \Phi(\delta)\left(1 - \frac{I(Z^J, J) + \log 2}{\log m}\right). \tag{9.7}$$

Proof: It suffices to show that

$$\mathbb{Q}\left[\Psi(Z^J) \neq J\right] \geq 1 - \frac{I(Z^J, J) + \log 2}{\log m}. \tag{9.8}$$

To this end, letting $X = J$ and $Y = Z^J$ in (9.6) and further noting $H(J|Z^J) = H(J) - I(Z^J, J) = \log m - I(Z^J, J)$ shows (9.8). ■

In order to apply Theorem 9.14, we need to further upper bound $I(Z^J, J)$. The local Fano method and global Fano method establish the lower minimax risk bound by upper bounding $I(Z^J, J)$ in different ways.

9.4.3 Local Fano Method

The mutual information can be written in terms of the component distributions $\{\mathbb{P}_j^n\}_{j=1}^m$ and the mixture distribution $\mathbb{Q}_Z = \frac{1}{m}\sum_{j=1}^m \mathbb{P}_j^n$ as follows

$$I(Z^J, J) = \frac{1}{m}\sum_{j=1}^m D(\mathbb{P}_j^n \| \mathbb{Q}_Z). \tag{9.9}$$

Letting $p_j^n(x_1, \cdots, x_n)$ be the density of $\mathbb{P}_j^n$ under some base measure $\mu(dx_1 \cdots dx_n)$ and noting that $\frac{1}{m}$ is the density of $\mathbb{Q}_J$ under the counting measure $\mu(dj)$, the density of the joint distribution $\mathbb{Q}$ under the base measure $\mu(dx_1 \cdots dx_n)\ \mu(dj)$ is given by $\frac{1}{m}p_j^n(x_1, \cdots, x_n)$, and the density of $\mathbb{Q}_Z$ is given by $\frac{1}{m}\sum_{j=1}^m p_j^n(x_1, \cdots, x_n)$. Thus a simple calculation yields,

$$\begin{aligned}
I(Z^J, J) &= \int_{\mathcal{X}^n \times [m]} \frac{1}{m} p_j^n(x_1, \cdots, x_n) \log \frac{\frac{1}{m}p_j^n(x_1, \cdots, x_n)}{\left(\frac{1}{m}\sum_{i=1}^m p_i^n(x_1, \cdots, x_n)\right)\frac{1}{m}} \mu(dx_1 \cdots dx_n)\mu(dj) \\
&= \frac{1}{m}\sum_{j=1}^m \int_{\mathcal{X}^n} p_j^n(x_1, \cdots, x_n) \log \frac{p_j^n(x_1, \cdots, x_n)}{\left(\frac{1}{m}\sum_{i=1}^m p_i^n(x_1, \cdots, x_n)\right)} \mu(dx_1 \cdots dx_n) \\
&= \frac{1}{m}\sum_{j=1}^m D(\mathbb{P}_j^n \| \mathbb{Q}_Z),
\end{aligned}$$

which proves (9.9). In addition, we have

$$\begin{aligned}
D(\mathbb{P}_j^n \| \mathbb{Q}_Z) &= D(\mathbb{P}_j^n \| \frac{1}{m}\sum_{i=1}^m \mathbb{P}_i^n) \\
&= \int_{\mathcal{X}^n} p_j^n(x_1, \cdots, x_n) \log \frac{p_j^n(x_1, \cdots, x_n)}{\frac{1}{m}\sum_{i=1}^m p_i^n(x_1, \cdots, x_n)} \mu(dx_1 \cdots dx_n) \\
&= -\int_{\mathcal{X}^n} p_j^n(x_1, \cdots, x_n) \log \frac{\frac{1}{m}\sum_{i=1}^m p_i^n(x_1, \cdots, x_n)}{p_j^n(x_1, \cdots, x_n)} \mu(dx_1 \cdots dx_n)
\end{aligned}$$

$$\leq \frac{1}{m}\sum_{i=1}^{m}\int_{\mathcal{X}^n} p_j^n(x_1,\cdots,x_n)\log\frac{p_j^n(x_1,\cdots,x_n)}{p_i^n(x_1,\cdots,x_n)}\mu(dx_1\cdots dx_n)$$
$$= \frac{1}{m}\sum_{i=1}^{m} D(\mathbb{P}_j^n\|\mathbb{P}_i^n),$$

where the fourth line follows from the Jensen inequality. Inserting this inequality into (9.9) yields

$$I(Z^J, J) \leq \frac{1}{m^2}\sum_{j,i=1}^{m} D(\mathbb{P}_j^n\|\mathbb{P}_i^n). \tag{9.10}$$

Therefore, we have the following proposition.

Proposition 9.15 *Under the setting for the construction of J and Z^J in Section 9.1, we have*

$$\mathfrak{M}_n(\Theta) \geq \Phi(\delta)\left(1 - \frac{\frac{1}{m^2}\sum_{j,i=1}^{m} D(\mathbb{P}_j^n\|\mathbb{P}_i^n) + \log 2}{\log m}\right). \tag{9.11}$$

To apply the bound in (9.11), we need to construct a family of distributions $\{\mathbb{P}_j\}_{j=1}^m$ corresponding to $\{\theta_j\}_{j=1}^m$ such that

- $\rho(\theta_j, \theta_\ell) \geq 2\delta$, and m can be as large as possible,
- $D(\mathbb{P}_j^n\|\mathbb{P}_i^n)$ is sufficiently small.

Due to the second constraint, we cannot construct a packing of the entire space Θ; otherwise, $\max_{i,j} D(\mathbb{P}_j^n\|\mathbb{P}_i^n)$ would be large. Instead, the *local Fano method* construct a packing of local subset by first construct a packing set of a fixed radius and then shrinking the packing sets by δ, which leaves the packing number unchanged but gives us the room to choose a δ that is sufficiently small such that $D(\mathbb{P}_j^n\|\mathbb{P}_i^n)$ can be sufficiently small such that $1 - \frac{\frac{1}{m^2}\sum_{j,i=1}^{m} D(\mathbb{P}_j^n\|\mathbb{P}_i^n)+\log 2}{\log m}$ is larger than a small constant. Let illustrate this with two examples.

Example 9.16 *We consider the mean estimation of multivariate normal distributions (in contrast to Example 9.7) $\mathcal{N}(\theta, \sigma^2 I_d)$, where $\theta \in \mathbb{R}^d$. It is not hard to show that the mean squared error of the sample mean estimator is of the order $\frac{d\sigma^2}{n}$ (**check this!**). In this example we will show that the minimax risk of the means squared error is $\gtrsim \frac{d\sigma^2}{n}$*

To this end, let $\{x_1,\cdots,x_m\}$ be a 1/2 packing of the unit ℓ_2-ball with $\log m \geq d\log 2$. Define $\theta_j = 4\delta x_j$. Then it is trivial that $\|\theta_i - \theta_j\|_2 \geq 2\delta$ and $\|\theta_i - \theta_j\|_2 \leq 8\delta$. In addition, we have

$$D(\mathbb{P}_j^n\|\mathbb{P}_i^n) = nD(\mathbb{P}_j\|\mathbb{P}_i) = nD(\mathcal{N}(\theta_j,\sigma^2 I_d)\|\mathcal{N}(\theta_i,\sigma^2 I_d)) = \frac{n}{2\sigma^2}\|\theta_j - \theta_i\|_2^2 \leq \frac{32n\delta^2}{\sigma^2}.$$

It follows that

$$\frac{\frac{1}{m^2}\sum_{j,i=1}^{m} D(\mathbb{P}_j^n\|\mathbb{P}_i^n) + \log 2}{\log m} \leq \frac{\frac{32n\delta^2}{\sigma^2} + \log 2}{d\log 2} \lesssim \frac{1}{2},$$

if we choose $\delta^2 \asymp \frac{d}{n}\sigma^2$. Thus, we conclude that

$$\inf_{\widehat{\theta}}\sup_{\theta\in\mathbb{R}^d}\mathbb{E}\left[\|\widehat{\theta} - \theta\|_2^2\right] \gtrsim \frac{d\sigma^2}{n}.$$

Example 9.17 *Consider the model $Y = A\theta^* + w$, where $A \in \mathbb{R}^{n\times d}$ is fixed and $rank(A) = d$, and $w \sim \mathcal{N}(0, \sigma^2 I_n)$. We want to lower bound the minimax risk when estimating θ^* from Y under the (semi)metric*

$$\rho(\theta, \theta') = \frac{\|A(\theta - \theta')\|_2}{\sqrt{n}}.$$

Define the set $S = \{x \in \text{range}(A) : \|x\|_2 = 1\}$. We can construct a $1/2$-packing of S with the packing number m satisfying $\log m \geq d \log 2$. Let $\{x_1, \cdots, x_m\}$ denote the packing set, the goal is to construct a set $\{\theta_1, \cdots, \theta_m\}$ such that $\rho(\theta_i, \theta_j) \geq 2\delta$. To this end, define θ_j to be the vector such that $A\theta_j = 4\delta\sqrt{n}x_j$. Then, it is easy to verify that

$$\rho(\theta_i, \theta_j) = \frac{\|A(\theta_i - \theta_j)\|_2}{\sqrt{n}} = 4\delta\|x_i - x_j\|_2,$$

and consequently, $2\delta \leq \rho(\theta_i, \theta_j) \leq 8\delta$.

Note that the observations $Y = (Y_1, \cdots, Y_n) \sim \mathcal{N}(A\theta, \sigma^2 I_n)$. By the divergence property of multivariable Gaussian distribution, we have

$$D(\mathbb{P}_j^n \| \mathbb{P}_i^n) = \frac{1}{2\sigma^2}\|A(\theta_j - \theta_i)\|_2^2 \leq \frac{32n\delta^2}{\sigma^2}.$$

It follows that

$$\frac{\frac{1}{m^2}\sum_{j,i=1}^m D(\mathbb{P}_j^n \| \mathbb{P}_i^n) + \log 2}{\log m} \leq \frac{\frac{32n\delta^2}{\sigma^2} + \log 2}{d \log 2} \lesssim \frac{1}{2},$$

if we choose $\delta^2 \asymp \frac{d}{n}\sigma^2$. Thus, we conclude that

$$\inf_{\widehat{\theta}} \sup_{\theta\in\mathbb{R}^d} \mathbb{E}\left[\frac{\|A(\widehat{\theta} - \theta)\|_2^2}{n}\right] \gtrsim \frac{d\sigma^2}{n}.$$

*This bound is sharp in order which can be achieved for example by the least-squares estimator (**check this!**).*

9.4.4 Global Fano Method

Recall from (9.9) that

$$I(Z^J, J) = \frac{1}{m}\sum_{j=1}^m D(\mathbb{P}_j^n \| \mathbb{Q}_Z), \quad \mathbb{Q}_Z = \frac{1}{m}\sum_{j=1}^m \mathbb{P}_j^n.$$

Thus, if we can construct a packing of $\mathcal{P}^n$ in terms of the KL divergence, it is likely to bound $I(Z^J, J)$ using the packing of the all the distributions. This leads to the global Fano method, also known as Yang-Barron method.

Lemma 9.18 *Let N_{KL} be the ε-covering number of $\mathcal{P}^n$ under the square root KL divergence. Then we have*

$$I(Z^J, J) \leq \inf_{\varepsilon>0}\left\{\varepsilon^2 + \log N_{\text{KL}}\right\}. \tag{9.12}$$

Proof: We first claim that

$$\frac{1}{m}\sum_{j=1}^{m} D(\mathbb{P}_j^n \| \mathbb{Q}_Z) \le \frac{1}{m}\sum_{j=1}^{m} D(\mathbb{P}_j^n \| \mathbb{Q}), \quad \text{for any } \mathbb{Q}.$$

That is, the average distribution minimizes the KL divergence. Indeed, we have

$$\begin{aligned}
\frac{1}{m}\sum_{j=1}^{m} D(\mathbb{P}_j^n \| \mathbb{Q}_Z) &= \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}_{\mathbb{P}_j^n}\left[\log \frac{d\mathbb{P}_j^n}{d\mathbb{Q}_Z}\right] = \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}_{\mathbb{P}_j^n}\left[\log\left(\frac{d\mathbb{P}_j^n}{d\mathbb{Q}}\frac{d\mathbb{Q}}{d\mathbb{Q}_Z}\right)\right] \\
&= \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}_{\mathbb{P}_j^n}\left[\log \frac{d\mathbb{P}_j^n}{d\mathbb{Q}}\right] + \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}_{\mathbb{P}_j^n}\left[\log \frac{d\mathbb{Q}}{d\mathbb{Q}_Z}\right] \\
&= \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}_{\mathbb{P}_j^n}\left[\log \frac{d\mathbb{P}_j^n}{d\mathbb{Q}}\right] - \mathbb{E}_{\mathbb{Q}_Z}\left[\log \frac{d\mathbb{Q}_Z}{d\mathbb{Q}}\right] \\
&\le \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}_{\mathbb{P}_j^n}\left[\log \frac{d\mathbb{P}_j^n}{d\mathbb{Q}}\right]. \\
&= \frac{1}{m}\sum_{j=1}^{m} D(\mathbb{P}_j^n \| \mathbb{Q}).
\end{aligned}$$

Consequently,

$$I(Z^J, J) \le \frac{1}{m}\sum_{j=1}^{m} D(\mathbb{P}_j^n | \mathbb{Q}) \le \max_{j=1,\cdots,m} D(\mathbb{P}_j^n | \mathbb{Q})$$

for any $\mathbb{Q}$. Thus, it suffices to obtain a bound by a particular $\mathbb{Q}$.

To this end, let $\{\mathbb{Q}_1, \cdots, \mathbb{Q}_N\}$ be a ε-net of $\mathcal{P}^n$ under the square-root KL distance and define $\mathbb{Q} = \frac{1}{N}\sum_{k=1}^{N} \mathbb{Q}_k$. By construction, there exists a $\mathbb{Q}_{k_j}$ such that $D(\mathbb{P}_j^n \| \mathbb{Q}_{k_j}) \le \varepsilon^2$. Then,

$$\begin{aligned}
D(\mathbb{P}_j^n \| \mathbb{Q}) &= \mathbb{E}_{\mathbb{P}_j^n}\left[\log \frac{d\mathbb{P}_j^n}{d\mathbb{Q}}\right] \\
&= \mathbb{E}_{\mathbb{P}_j^n}\left[\log \frac{d\mathbb{P}_j^n}{\frac{1}{N}\sum_{k=1}^{N} d\mathbb{Q}_k}\right] \\
&\le \mathbb{E}_{\mathbb{P}_j^n}\left[\log \frac{d\mathbb{P}_j^n}{\frac{1}{N} d\mathbb{Q}_{k_j}}\right] \\
&\le \varepsilon^2 + \log N.
\end{aligned}$$

Since this bound holds for any $\mathbb{P}_j^n$ and any $\varepsilon > 0$, the claim follows. ∎

Combing Lemma 9.18 with Theorem 9.14 yields the following proposition.

Proposition 9.19 *Under the setting for the construction of J and Z^J in Section 9.1, we have*

$$\mathfrak{M}_n(\Theta) \ge \Phi(\delta)\left(1 - \frac{\left(\varepsilon^2 + \log N_{\mathrm{KL}}\right) + \log 2}{\log m}\right). \tag{9.13}$$

Recall that m in (9.13) is the number of θ_j such that $\rho(\theta_i, \theta_j) \geq 2\delta$, so it relies on δ and when δ is prescribed we may choose $\{\theta_j\}_{j=1}^m$ to be global packing of Θ so that m is maximized. Note that there are two parameters ε and δ to be determined in (9.13). A typical way to choose them is

- choose ε such that $\varepsilon^2 \geq \log N_{\mathrm{KL}}$,
- choose largest possible δ such that $\log m \geq 4\varepsilon^2 + 2\log 2$,

so that $1 - \frac{\left(\varepsilon^2 + \log N_{\mathrm{KL}}\right) + \log 2}{\log m} \geq \frac{1}{2}$.

Example 9.20 *Consider the family of density functions*

$$\mathcal{F} = \{f : [0,1] \to [c_0, c_1] : \ \|f''\|_\infty \leq c_2 \text{ and } \int_0^1 f(x)dx = 1\},$$

where $0 < c_0 < 1 < c_1, c_2 > 1$ are constants. We study the minimax risk of estimating a density function from i.i.d data $X_1, \cdots, X_n \sim \mathbb{P}_f$ under the Hellinger distance

$$\rho(f, g) = H(f\|g) := H(\mathbb{P}_f \| \mathbb{P}_g) = \sqrt{\int_0^1 \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 dx}.$$

Note that

$$\begin{aligned} D(\mathbb{P}_f \| \mathbb{P}_g) &= \int_0^1 f(x) \log \frac{f(x)}{g(x)} dx \\ &\leq \int_0^1 f(x) \left(\frac{f(x)}{g(x)} - 1\right) dx \\ &= \int_0^1 \frac{(f(x) - g(x))^2}{g(x)} dx \\ &\leq \frac{1}{c_0} \int_0^1 (f(x) - g(x))^2 dx, \end{aligned}$$

and

$$\begin{aligned} \rho(f, g)^2 &= \int_0^1 \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 dx \\ &\leq \frac{1}{4c_0^2} \int_0^1 \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 \left(\sqrt{f(x)} + \sqrt{g(x)}\right)^2 dx \\ &= \frac{1}{4c_0^2} \int_0^1 (f(x) - g(x))^2 \, dx. \end{aligned}$$

Therefore, both the squared KL divergence and $\rho(\cdot, \cdot)$ can be bounded by the L_2 distance. Consequently, in order to apply the global Fano method, we only need to understand the metric entropy in the L_2-norm. Since $f \in \mathcal{F}$ is second order smooth with $\|f''\|_\infty \leq c_2$, it can be shown that (See Example 5.11 of [1]),

$$\log N(\mathcal{F}, \|\cdot\|_2, \alpha) \asymp \left(\frac{1}{\alpha}\right)^{1/2}.$$

Since $D(\mathbb{P}_f^n \| \mathbb{P}_g^n) = nD(\mathbb{P}_f \| \mathbb{P}_g)$, $\sqrt{D(\mathbb{P}_f^n \| \mathbb{P}_g^n)} \le \varepsilon$ *if* $\sqrt{D(\mathbb{P}_f \| \mathbb{P}_g)} \le \varepsilon/\sqrt{n}$. *It follows that,*

$$\log N_{\mathrm{KL}} \asymp \left(\frac{\sqrt{n}}{\varepsilon}\right)^{1/2}$$

Thus, in order for $\varepsilon^2 \ge \log N_{\mathrm{KL}}$, *we may choose* $\varepsilon^2 \asymp (n)^{\frac{1}{5}}$. *Moreover, since* $\log m \asymp \left(\frac{1}{\delta}\right)^{1/2}$, *for the above choice of* ε, *we may choose* $\delta \asymp n^{-\frac{2}{5}}$ *such that* $\log m \ge 4\varepsilon^2 + 2\log 2$. *Finally, it can be concluded that*

$$\inf_{\widehat{f}} \sup_{f} H^2(\widehat{f} \| f) \gtrsim n^{-\frac{4}{5}}.$$

Reading Materials

[1] Martin Wainwright, *High Dimensional Statistics – A non-asymptotic viewpoint*, Chapters 15.

[2] John Duchi, *Lecture notes for Statistics 311/Electrical Engineering 377: Information Theory and Statistics*, Chapter 7, 10.

[3] Francis Bach, *Learning Theory from First Principles*, Chapter 11.1.