Chapter 4: *Distances between measures*

Zetai Cen

Last updated: July 9, 2024

Contents

1 Introduction **2**

2 Divergence Measures **3**

2.1 Total variation, Hellinger, and Kullback-Leibler 4

2.2 Useful inequalities 5

3 Minimax Lower Bounds **7**

3.1 A general bound: Bayes risk 8

3.2 Lower bound method I: Le Cam's 9

3.3 Lower bound method II: Fano's 10

3.4 Lower bound method III: Assouad's 10

4 Examples on Minimax **11**

5 Appendix A **12**

5.1 Proof of lemmas 12

1 Introduction

Keywords: total variation; Hellinger; Kullback-Leibler; Fano's lemma; Assouad's inequality; minimax rates of convergence; maximum likelihood in one dimension via Hellinger distance.

1. total variation, Kullback-Leibler, Hellinger;

2. Le Cam, Fano, Assouad;

We will be partly using Chapter 15 (Minimax lower bounds) in Wainwright (2019). As hinted by the term "minimax", we are interested in lower bounding some minimax rates for which divergence measures (between distributions) can be useful. It motivates us to introduce three important measures, as shown in bullet-point 1 above.

In the second part, we will talk about three techniques to derive minimax lower bound in general, corresponding to bullet-point 2 above. We also use the following materials in preparing this note:

- the book chapter by Yu (1997);

- the monograph by Guntuboyina (2011);

- the Lecture-8 note for STAT583 in University of Washington (https://sites.stat.washington.edu/people/fanghan/teaching/STAT583/minimax.pdf);

- the Chapter-5 note for the MIT open course, 18.S997 (Spring 2015) (https://ocw.mit.edu/courses/18-s997-high-dimensional-statistics-spring-2015/);

- the Lecture-2 note for ECE598 (Spring 2016) at Yale (http://www.stat.yale.edu/~yw562/teaching/598/lec02.pdf);

- the monograph by Canonne (2023);

- the Lecture-7 note for ECE 5630 at Cornell (http://people.ece.cornell.edu/zivg/ECE_5630_Lectures7.pdf).

2 Divergence Measures

To begin from a general level, let us first define the following.

Definition 2.1. ***(f-divergence)*** *Let any two probability measures $\mathbb{P}$ and $\mathbb{Q}$ be on some measurable space and some base measure ν satisfying $\mathbb{P} \ll \nu$ and $\mathbb{Q} \ll \nu$, with $p = d\mathbb{P}/d\nu$ and $q = d\mathbb{Q}/d\nu$. Let the convex function $f : [0, \infty) \to (-\infty, \infty]$ be such that $f(x) < \infty$ for all $x > 0$, $f(1) = 0$ and $f(0) = \lim_{t\to\infty} f(t)$. The f-divergence given the generator f is defined as*

$$\mathbb{D}_f(\mathbb{P}\|\mathbb{Q}) := \int f\Big(\frac{d\mathbb{P}}{d\mathbb{Q}}\Big) d\mathbb{Q} = \int f\Big(\frac{p}{q}\Big)\, q\, d\nu.$$

Such an f-divergence can induce many divergence measures useful in bounding minimax rates. Intuitively, $\mathbb{D}_f(\mathbb{P}\|\mathbb{Q})$ gauges the discrepancy between $\mathbb{P}$ and $\mathbb{Q}$, which is the heuristic of requiring $f(1) = 0$. In fact, we have the following properties for any f-divergence.

Proposition 2.1. *Given any f-divergence such that f is described as in Definition 2.1, we have the following.*

1. *(Linearity) Given any sequence of $a_i \geq 0$ and f_i satisfying Definition 2.1, we have*

$$\mathbb{D}_{\sum_i a_i f_i}(\mathbb{P}\|\mathbb{Q}) = \sum_i a_i \mathbb{D}_{f_i}(\mathbb{P}\|\mathbb{Q}).$$

2. *(Non-negativity) Any f-divergence is non-negative.*

3. *(Reversal divergence) Define the convex inversion of f as $g(x) := xf(1/x)$ on $[0, \infty)$, then g satisfies the condition defined as in Definition 2.1. Thus the "reverse" of $\mathbb{D}_f(\mathbb{P}\|\mathbb{Q})$ is obtained as $\mathbb{D}_g(\mathbb{P}\|\mathbb{Q})$, in the sense that, if $\mathbb{P}$ and $\mathbb{Q}$ are absolutely continuous with each other, we have*

$$\mathbb{D}_g(\mathbb{Q}\|\mathbb{P}) = \mathbb{D}_f(\mathbb{P}\|\mathbb{Q}).$$

Proof of Proposition 2.1. #1 is direct from definition. For #2, we have

$$\mathbb{D}_f(\mathbb{P}\|\mathbb{Q}) = \mathbb{E}^{\mathbb{Q}}\Big[f\Big(\frac{d\mathbb{P}}{d\mathbb{Q}}\Big)\Big] \geq f\Big\{\mathbb{E}^{\mathbb{Q}}\Big[\frac{d\mathbb{P}}{d\mathbb{Q}}\Big]\Big\} = f(1) = 0.$$

For #3, $g(1) = f(1) = 0$. For convexity, it suffices to check for any $x, y \in [0, \infty)$

$$g''(x) = \Big\{f\Big(\frac{1}{x}\Big) - xf'\Big(\frac{1}{x}\Big)\frac{1}{x^2}\Big\}' = -f'\Big(\frac{1}{x}\Big)\frac{1}{x^2} + \frac{1}{x^3}f''\Big(\frac{1}{x}\Big) + \frac{1}{x^2}f'\Big(\frac{1}{x}\Big) = \frac{1}{x^3}f''\Big(\frac{1}{x}\Big) \geq 0,$$

where the last inequality used the convexity of f. It remains to show the divergence generated by g is the reversal of that by f. We have

$$\mathbb{D}_g(\mathbb{Q}\|\mathbb{P}) = \int f\Big(\frac{d\mathbb{P}}{d\mathbb{Q}}\Big)\frac{d\mathbb{Q}}{d\mathbb{P}} d\mathbb{P} = \mathbb{D}_f(\mathbb{P}\|\mathbb{Q}).$$

□

An immediate result from Proposition 2.1 is that if the constructed $g = f$, we conclude by #3 that the f-divergence is symmetric. On the other hand, if $g \neq f$, we may construct a symmetric f-divergence to use $(f+g)/2$ by #1.

As we discuss in Section 3, divergence measures are useful to lower bound the minimax rate. Noticing that we will generalise f-divergence to the following measures, it is natural to ask: can we derive minimax lower bounds using f-divergence in general? The answer is affirmative by, e.g. Guntuboyina (2011), but we do not pursue such discussion here.

2.1 Total variation, Hellinger, and Kullback-Leibler

Definition 2.2. ***(Total variation distance)*** *In the setting of Definition 2.1, we call the f-divergence the total variation distance if $f(x) = |x-1|/2$ (of course, restricted on $[0,\infty)$). In other words, let the measurable space considered here be $(\mathcal{X}, \mathcal{A})$, we have*

$$\mathbf{TV}(\mathbb{P}, \mathbb{Q}) := \frac{1}{2}\int \Big|\frac{p-q}{q}\Big|\, q\, d\nu = \frac{1}{2}\int |p-q|\, d\nu = \sup_{\mathcal{T}\in\mathcal{A}} \Big|\int_{\mathcal{T}} (p-q)\, d\nu\Big| = \sup_{\mathcal{T}\in\mathcal{A}} \Big|\mathbb{P}(\mathcal{T}) - \mathbb{Q}(\mathcal{T})\Big|.$$

The name already tells the symmetry property of the total variation distance, which is either trivial from the last expression in Definition 2.2 or easily confirmed by the convex inversion of f.

Definition 2.3. ***(Hellinger distance)*** [1] *In the setting of Definition 2.1, the squared Hellinger distance is obtained by taking $f(x) = (\sqrt{x}-1)^2$. In other words, let the measurable space considered here be $(\mathcal{X}, \mathcal{A})$, we have*

$$\mathbf{H}(\mathbb{P}, \mathbb{Q}) := \Big(\int (\sqrt{p} - \sqrt{q})^2\, d\nu\Big)^{1/2} = \Big(\int (\sqrt{d\mathbb{P}} - \sqrt{d\mathbb{Q}})^2\Big)^{1/2}.$$

It is also direct from the expression that the Hellinger distance is symmetric.

Definition 2.4. ***(Kullback-Leibler divergence)*** *In the setting of Definition 2.1, we call the f-divergence the Kullback-Leibler divergence (K-L divergence) if $f(x) = x \log x$. In other words, let the measurable space considered here be $(\mathcal{X}, \mathcal{A})$, we have*

$$\mathbf{KL}(\mathbb{P}\|\mathbb{Q}) := \int \log\Big(\frac{p}{q}\Big) p\, d\nu = \int \log\Big(\frac{d\mathbb{P}}{d\mathbb{Q}}\Big)\, d\mathbb{P}.$$

Unlike the total variation distance or the Hellinger distance, the K-L divergence is asymmetric on its inputs and hence termed "divergence". Additionally, we also give the definition for another type of f-divergence which is also asymmetric.

Definition 2.5. ***(χ^2 divergence)*** *In the setting of Definition 2.1, we call the f-divergence the χ^2 divergence if $f(x) = (x-1)^2$. In other words, let the measurable space considered here be $(\mathcal{X}, \mathcal{A})$, we have*

$$\chi^2(\mathbb{P}\|\mathbb{Q}) := \int \frac{1}{q}(p-q)^2\, d\nu = \int \Big(\frac{d\mathbb{P}}{d\mathbb{Q}} - 1\Big)^2\, d\mathbb{Q}.$$

[1] Somehow I found different definitions for Hellinger distance, either the one shown here or with 1/2 in front of the integral.

2.2 Useful inequalities

Proposition 2.2. *(Pinsker's inequality) For all distributions $\mathbb{P}$ and $\mathbb{Q}$,*

$$\mathbf{TV}(\mathbb{P},\mathbb{Q}) \leq \sqrt{\frac{\mathbf{KL}(\mathbb{P}\|\mathbb{Q})}{2}}.$$

Proof of Proposition 2.2. Consider first the binary distribution such that $\mathbb{P},\mathbb{Q}$ are Bernoulli distribution with parameters p,q. Since $\mathbf{TV}(\mathbb{P},\mathbb{Q}) = |p-q|$, we are left to show

$$2(p-q)^2 \leq p\log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right).$$

For $p,q \in \{0,1\}$, the result is trivial, so we assume $p,q \in (0,1)$. We introduce a function $f:(0,1)\to\mathbb{R}$ as $f(x) = p\log(x) + (1-p)\log(1-x)$, then the RHS above is simply $f(p)-f(q)$. Thus,

$$f(p)-f(q) = \int_q^p f'(x)dx = \int_q^p \frac{p-x}{x(1-x)}dx \geq \int_q^p 4(p-x)dx = 2(p-q)^2,$$

where the inequality used $x(1-x)\leq 1/4$ for $x\in(0,1)$.

For general $\mathbb{P},\mathbb{Q}$ on an arbitrary sigma field $\mathcal{A}$, let X,Y be the random variables following $\mathbb{P},\mathbb{Q}$ and fix any $\mathcal{T}\in\mathcal{A}$. Construct two random variables $\mathbb{1}_{\mathcal{T}}(X)$ and $\mathbb{1}_{\mathcal{T}}(Y)$ which can be seen as Bernoulli distributed with parameters $\mathbb{P}(\mathcal{T})$ and $\mathbb{Q}(\mathcal{T})$, respectively. Denote the distributions of $\mathbb{1}_{\mathcal{T}}(X)$ and $\mathbb{1}_{\mathcal{T}}(Y)$ as $\mathbb{P}_*$ and $\mathbb{Q}_*$, then

$$2\left(\mathbb{P}(\mathcal{T})-\mathbb{Q}(\mathcal{T})\right)^2 = 2\,\mathbf{TV}^2(\mathbb{P}_*,\mathbb{Q}_*) \leq \mathbf{KL}(\mathbb{P}_*\|\mathbb{Q}_*) \leq \mathbf{KL}(\mathbb{P}\|\mathbb{Q}),$$

where the last inequality used the data processing inequality which we did not include it here (for now). Since the above holds for all $\mathcal{T}$, take the supremum over $\mathcal{A}$ we have

$$2\,\mathbf{TV}^2(\mathbb{P},\mathbb{Q}) = \sum_{\mathcal{T}\in\mathcal{A}} 2\left(\mathbb{P}(\mathcal{T})-\mathbb{Q}(\mathcal{T})\right)^2 \leq \mathbf{KL}(\mathbb{P}\|\mathbb{Q}).$$

□

Proposition 2.3. *(Le Cam's inequality) For all distributions $\mathbb{P}$ and $\mathbb{Q}$,*

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

Proof of Proposition 2.3.

$$\begin{aligned}\mathbf{TV}^2(\mathbb{P},\mathbb{Q}) &= \frac{1}{4}\Big(\int |p-q|\,d\nu\Big)^2 = \frac{1}{4}\Big(\int |\sqrt{p}-\sqrt{q}|\cdot|\sqrt{p}+\sqrt{q}|\,d\nu\Big)^2\\ &\leq \frac{1}{4}\Big\{\int (\sqrt{p}-\sqrt{q})^2\,d\nu\Big\}\Big\{\int \left[2p+2q-(\sqrt{p}-\sqrt{q})^2\right]d\nu\Big\}\\ &= \frac{1}{4}\,\mathbf{H}^2(\mathbb{P},\mathbb{Q})\Big\{4-\mathbf{H}^2(\mathbb{P},\mathbb{Q})\Big\}.\end{aligned}$$

□

Notice first the RHS of the Le Cam's inequality is obviously upper bounded by $\mathbf{H}(\mathbb{P}, \mathbb{Q})$. Moreover, it actually holds that

$$\mathbf{H}(\mathbb{P}, \mathbb{Q}) \leq \sqrt{\mathbf{KL}(\mathbb{P} \| \mathbb{Q})}, \tag{2.1}$$

so Le Cam's inequality with the above argument also gives $\mathbf{TV}(\mathbb{P}, \mathbb{Q}) \leq \sqrt{\mathbf{KL}(\mathbb{P} \| \mathbb{Q})}$ which is weaker than Pinsker's inequality. To show (2.1), we have

$$\begin{aligned} \mathbf{KL}(\mathbb{P} \| \mathbb{Q}) &= 2 \int -\log \Big(\frac{\sqrt{q}}{\sqrt{p}}\Big) p \, d\nu \geq 2 \int p \Big(1 - \frac{\sqrt{q}}{\sqrt{p}}\Big) \, d\nu \\ &= \int (p + q - 2\sqrt{pq}) \, d\nu = \int (\sqrt{p} - \sqrt{q})^2 \, d\nu = \mathbf{H}^2(\mathbb{P}, \mathbb{Q}), \end{aligned}$$

where the first inequality used $x = e^{\log x} \geq 1 + \log x$.

3 Minimax Lower Bounds

We start by introducing the minimax risk for statistical decision problems.

Context setup. Let $\mathcal{P}$ be a set of distributions, $X = \{X_1, \dots, X_n\}$ a set of observations from some unknown $\mathbb{P} \in \mathcal{P}$, and $\theta(\mathbb{P})$ an unknown parameter of $\mathbb{P}$.

Decision rule. Given $\widehat{\theta}(X)$ as an estimator of $\theta(\mathbb{P})$ and a specified loss function $L(\widehat{\theta}(X), \theta(\mathbb{P}))$, we define the **risk** for $\widehat{\theta}$ as $\mathbb{E}^{\mathbb{P}}\left[L(\widehat{\theta}(X), \theta(\mathbb{P}))\right]$.

Minimax rate. The **minimax risk** is essentially "minimax" risk, formally defined as

$$\mathcal{R}_{\text{minimax}} = \inf_{\widehat{\theta}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}^{\mathbb{P}}\left[L(\widehat{\theta}(X), \theta(\mathbb{P}))\right].$$

To make the above concept more concrete, two examples are given. For convenience, we use in short $\theta \equiv \theta(\mathbb{P})$ hereafter.

Example 3.1. *Let $\mathcal{P} = \{N(\theta, 1) \mid \theta \in \mathbb{R}\}$ and consider estimating θ with loss function $L(\widehat{\theta}, \theta) = (\widehat{\theta} - \theta)^2$. Note that θ is a scalar here, and the minimax risk is*

$$\mathcal{R}_{\text{minimax}} = \inf_{\widehat{\theta}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}^{\mathbb{P}}\left[(\widehat{\theta} - \theta)^2\right].$$

Example 3.2. *Let $\mathcal{P}$ be the set of all bivariate distributions and we observe $(X_1, Y_1), \dots, (X_n, Y_n)$ from $\mathbb{P} \in \mathcal{P}$.*

We are interested in the regression function $\theta = \mathbb{E}^{\mathbb{P}}[Y \mid X = x] =: m(x)$, and the loss function $L(\widehat{m}, m) = \int (\widehat{m}(x) - m(x))^2 dx$. θ is a function in this case, and the minimax risk is

$$\mathcal{R}_{\text{minimax}} = \inf_{\widehat{\theta}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}^{\mathbb{P}}\left[\int (\widehat{m}(x) - m(x))^2 dx\right].$$

The exact rate of minimax risk can be difficult to obtain directly, so a useful alternative is to find a good enough lower bound and hope that a proposed estimator achieves such bound, i.e. is said to be minimax optimal. Thus, studying minimax lower bound is indeed of practical interest.

If we are able to further upper bound the minimax risk, then we obtain the exact rate if the two bounds match. However, observe trivially for any $\widehat{\theta}(X)$ that

$$\mathcal{R}_{\text{minimax}} \le \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}^{\mathbb{P}}\left[L(\widehat{\theta}(X), \theta)\right],$$

so the maximum risk of any estimator gives an upper bound, while finding the lower bound is often harder.

3.1 A general bound: Bayes risk

While the minimax risk represents the lowest maximum risk, the Bayes risk deals with the lowest average risk with respect to some prior π of θ. Formally, let the **average risk for** π be defined as

$$\mathcal{R}_\pi = \mathbb{E}_{\theta\sim\pi}\mathbb{E}^{\mathbb{P}}\big[L(\widehat{\theta}(X),\theta)\big],$$

then the **Bayes risk** is simply the lowest of the average risk, i.e.

$$\mathcal{R}_{\text{Bayes},\pi} = \inf_{\widehat{\theta}} \mathcal{R}_\pi = \inf_{\widehat{\theta}} \mathbb{E}_{\theta\sim\pi}\mathbb{E}^{\mathbb{P}}\big[L(\widehat{\theta}(X),\theta)\big].$$

Notice first by Tonelli's Theorem we also have $\mathcal{R}_\pi = \mathbb{E}_X\mathbb{E}_{\theta|X}\big[L(\widehat{\theta}(X),\theta)\big]$, so minimising the Bayes risk is equivalent to minimising the posterior loss. Compared to the minimax risk, the Bayes risk is often easy to compute. We give two examples below.

Example 3.3. *Consider an arbitrary $\mathcal{P}$ and $\theta,\widehat{\theta}\in\mathbb{R}$, $\theta\sim\pi$. Given quadratic loss $L(\widehat{\theta},\theta)=(\widehat{\theta}-\theta)^2$, it is easy to see the posterior mean, $\widehat{\theta}=\mathbb{E}[\theta\mid X]$, minimises the posterior loss $\mathbb{E}_{\theta|X}\big[(\widehat{\theta}-\theta)^2\big]$, so the posterior mean is a Bayes estimator (i.e. an estimator reaching the Bayes risk) which is our primary interest. The Bayes risk is then simply the minimum mean squared error (over the joint).*

Example 3.4. *The same setting as Example 3.1. Consider a Gaussian prior $\theta\sim\pi=N(0,\sigma^2)$, then the Bayes estimator according to Example 3.3 is $\mathbb{E}_{\theta|X}[\theta]$. Assume only observe one sample X for simplicity.*

Given Gaussian prior and Gaussian likelihood, the posterior is such that

$$\begin{aligned} p(\theta\mid X) &\propto \exp\Big\{-\frac{\theta^2}{2\sigma^2}\Big\}\exp\Big\{-\frac{(X-\theta)^2}{2}\Big\} \propto \exp\Big\{-\frac{\theta^2(1+\sigma^2)-2\sigma^2X\theta}{2\sigma^2}\Big\} \\ &\propto \exp\Big\{-\frac{\theta^2-2\sigma^2X\theta/(1+\sigma^2)}{2\sigma^2/(1+\sigma^2)}\Big\} \propto \exp\Big\{-\frac{[\theta-\sigma^2X/(1+\sigma^2)]^2}{2\sigma^2/(1+\sigma^2)}\Big\}, \end{aligned}$$

so $\theta\mid X\sim N(\sigma^2X/(1+\sigma^2),\sigma^2/(1+\sigma^2))$. The Bayes estimator is hence $\sigma^2X/(1+\sigma^2)$, and the Bayes risk is obtained by plugging in $\mathcal{R}_\pi$ the Bayes estimator such that

$$\mathcal{R}_{\text{Bayes},\pi} = \mathbb{E}_X\mathbb{E}_{\theta|X}\Big\{\Big(\sigma^2X/(1+\sigma^2)-\theta\Big)^2\Big\} = \mathbb{E}_X\text{Var}_{\theta|X}(\theta) = \frac{\sigma^2}{1+\sigma^2}.$$

It is not hard to see $\mathcal{R}_{\text{Bayes},\pi}=\sigma^2/(1+n\sigma^2)$ if we observe $X=(X_1,\ldots,X_n)$ instead.

The reason we introduce Bayes estimator is its importance in bounding the minimax from below. For any prior π, we have

$$\mathcal{R}_{\text{Bayes},\pi} \le \mathbb{E}_{\theta\sim\pi}\mathbb{E}^{\mathbb{P}}\big[L(\widehat{\theta}_{\text{minimax}},\theta)\big] \le \mathbb{E}_{\theta\sim\pi}\sup_{\mathbb{P}\in\mathcal{P}}\mathbb{E}^{\mathbb{P}}\big[L(\widehat{\theta}_{\text{minimax}},\theta)\big] = \sup_{\mathbb{P}\in\mathcal{P}}\mathbb{E}^{\mathbb{P}}\big[L(\widehat{\theta}_{\text{minimax}},\theta)\big],$$

where $\widehat{\theta}_{\text{minimax}}$ is any minimax estimator and the last expression is essentially $\mathcal{R}_{\text{minimax}}$. Since such lower bound holds for all priors, intuitively the “worst” prior gives the tightest minimax lower bound. To present this idea more formally, we have the following proposition.

Proposition 3.1. *Let $\widehat{\theta}$ be a Bayes estimator for some prior π.*

1. *If for all $\mathbb{P}$ we have the risk upper bounded such that*

$$\mathbb{E}^{\mathbb{P}}\big[L(\widehat{\theta},\theta)\big] \le \mathcal{R}_{\text{Bayes},\pi},$$

   then $\widehat{\theta}$ is minimax optimal, and we call π a ***least favourable prior****.*

2. *If the risk $\mathbb{E}^{\mathbb{P}}\big[L(\widehat{\theta},\theta)\big]$ is constant, then the conclusion in #1 also holds.*

Proof of Proposition 3.1. The result #1 is direct from the fact that

$$\mathcal{R}_{\text{Bayes},\pi} \le \mathcal{R}_{\text{minimax}} \le \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}^{\mathbb{P}}\big[L(\widehat{\theta},\theta)\big] \le \mathcal{R}_{\text{Bayes},\pi},$$

where we take supremum of $\mathbb{E}^{\mathbb{P}}\big[L(\widehat{\theta},\theta)\big]$ over all $\mathbb{P}$ in the statement of Proposition 3.1 for the last inequality.

For #2, let $\mathbb{E}^{\mathbb{P}}\big[L(\widehat{\theta},\theta)\big] = c$ for some arbitrary constant, then $\mathcal{R}_{\text{Bayes},\pi} = \int c d\pi = c$. The proof is completed by applying the result in #1. □

3.2 Lower bound method I: Le Cam’s

Proposition 3.2. *Let $\mathcal{P}$ be a set of distributions. For any pair $\mathbb{P}_0, \mathbb{P}_1 \in \mathcal{P}$,*

$$\mathcal{R}_{\text{minimax}} \ge \frac{\Delta}{4}\int \big[p_0^n(x)\wedge p_1^n(x)\big]dx \ge \frac{\Delta}{8}\exp\big\{-n\,\mathbf{KL}(\mathbb{P}_0\|\mathbb{P}_1)\big\},$$

where $\Delta = L(\theta(\mathbb{P}_0),\theta(\mathbb{P}_1))$ for some specified loss function L.

The second inequality in Proposition 3.2 is useful as it might be complicated to compute $\int\big[p^n(x)\wedge q^n(x)\big]dx$. An alternative can also be

$$\int \big[p^n(x)\wedge q^n(x)\big]dx \ge \frac{1}{2}\Big(1-\frac{1}{2}\int |p_0(x)-p_1(x)|dx\Big)^{2n}.$$

Before getting into the proof of Proposition 3.2, we introduce some useful lemmas.

Lemma 3.1. *Given two distributions $\mathbb{P}_0, \mathbb{P}_1$, the sum of errors $\mathbb{P}_0(\psi=1)+\mathbb{P}_1(\psi=0)$ is minimised over all tests ψ by the Neyman-Pearson test*

$$\psi_*(x) = \begin{cases} 1, & \text{if } p_1(x) \ge p_0(x) \\ 0, & \text{if } p_1(x) < p_0(x). \end{cases}$$

Lemma 3.2. *For the Neyman-Pearson test* ψ_**,*

$$\mathbb{P}_0(\psi_* \neq 0) + \mathbb{P}_1(\psi_* \neq 1) = \int \big[p_0(x) \wedge p_1(x)\big] dx.$$

Lemma 3.3. *For any* $\mathbb{P}, \mathbb{Q}$*, we have*

$$\int p(x) \wedge q(x) dx \geq \frac{1}{2} \exp\Big(- \mathbf{KL}(\mathbb{P} \| \mathbb{Q})\Big).$$

Proof of Proposition 3.2. Denote for simplicity $\theta_0 \equiv \theta(\mathbb{P}_0)$ and $\theta_1 \equiv \theta(\mathbb{P}_1)$. Consider first $n = 1$, so we only observe X. The key step for the following is to read an estimator $\widehat{\theta}$ into a test statistic ψ such that

$$\psi(X) = \begin{cases} 1, & \text{if } L(\widehat{\theta}, \theta_1) \leq L(\widehat{\theta}, \theta_0) \\ 0, & \text{if } L(\widehat{\theta}, \theta_1) > L(\widehat{\theta}, \theta_0). \end{cases}$$

If $\mathbb{P} = \mathbb{P}_0$ and $\psi = 1$, then $\Delta = L(\theta_0, \theta_1) \leq L(\widehat{\theta}, \theta_0) + L(\widehat{\theta}, \theta_1) \leq 2L(\widehat{\theta}, \theta_0)$. Thus,

$$\mathbb{E}^{\mathbb{P}_0}\big[L(\widehat{\theta}, \theta_0)\big] \geq \mathbb{E}^{\mathbb{P}_0}\big[L(\widehat{\theta}, \theta_0)\mathbb{1}\{\psi = 1\}\big] \geq \frac{\Delta}{2}\mathbb{P}_0(\psi = 1).$$

Similarly, $\mathbb{E}^{\mathbb{P}_1}\big[L(\widehat{\theta}, \theta_1)\big] \geq \frac{\Delta}{2}\mathbb{P}_1(\psi = 0)$. Combine those two results, we have

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}^{\mathbb{P}}\big[L(\widehat{\theta}, \theta)\big] \geq \max_{\mathbb{P} \in \{\mathbb{P}_0, \mathbb{P}_1\}} \mathbb{E}^{\mathbb{P}}\big[L(\widehat{\theta}, \theta)\big] \geq \frac{\Delta}{2}\Big\{\mathbb{P}_0(\psi = 1) \vee \mathbb{P}_1(\psi = 0)\Big\}.$$

Taking infimum over all $\widehat{\theta}$ on both sides,

$$\begin{aligned} \inf_{\widehat{\theta}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}^{\mathbb{P}}\big[L(\widehat{\theta}, \theta)\big] &\geq \inf_{\psi} \frac{\Delta}{2}\Big\{\mathbb{P}_0(\psi = 1) \vee \mathbb{P}_1(\psi = 0)\Big\} \geq \inf_{\psi} \frac{\Delta}{4}\Big\{\mathbb{P}_0(\psi = 1) + \mathbb{P}_1(\psi = 0)\Big\} \\ &= \frac{\Delta}{4}\Big\{\mathbb{P}_0(\psi_* = 1) + \mathbb{P}_1(\psi_* = 0)\Big\} = \frac{\Delta}{4}\int \big[p_0(x) \wedge p_1(x)\big] dx, \end{aligned}$$

where the second inequality used the fact that the maximum is larger than the average, the second last equality used Lemma 3.1 with the Neyman-Pearson test ψ_*, and the last used Lemma 3.2.

This completes the proof for the first inequality in Proposition 3.2 for $n = 1$. For relaxing the restriction of $n = 1$, replace p_0 and p_1 with the joint version $p_0^n(x) = \prod_{i=1}^n p_0(x_i)$ and $p_1^n(x) = \prod_{i=1}^n p_1(x_i)$, respectively, and we are done.

For the second inequality in Proposition 3.2, it is direct from Lemma 3.3 and the fact that $\mathbf{KL}(\mathbb{P}_0^n \| \mathbb{P}_1^n) = n\,\mathbf{KL}(\mathbb{P}_0 \| \mathbb{P}_1)$. □

3.3 Lower bound method II: Fano's

3.4 Lower bound method III: Assouad's

4 Examples on Minimax

5 Appendix A

5.1 Proof of lemmas

Proof of Lemma 3.1. TBD

Proof of Lemma 3.2.

$$\begin{aligned}&\mathbb{P}_0(\psi_* \neq 0) + \mathbb{P}_1(\psi_* \neq 1) = \mathbb{P}_0(p_1(x) \geq p_0(x)) + \mathbb{P}_1(p_1(x) < p_0(x)) \\ &= \int \mathbb{1}(p_1(x) \geq p_0(x))\, p_0(x)\, dx + \int \mathbb{1}(p_1(x) < p_0(x))\, p_1(x)\, dx = \int \big[p_0(x) \wedge p_1(x)\big] dx.\end{aligned}$$

□

Proof of Lemma 3.3. Notice first $\int (p(x) \vee q(x))dx + \int (p(x) \wedge q(x))dx = 2$, then

$$\begin{aligned}&2\int (p(x) \wedge q(x))dx \geq \Big(2 - \int (p(x) \wedge q(x))dx\Big) \int (p(x) \wedge q(x))dx \\ &= \Big(\int (p(x) \vee q(x))dx\Big)\Big(\int (p(x) \wedge q(x))dx\Big) \geq \Big(\int \sqrt{(p(x) \vee q(x))(p(x) \wedge q(x))}dx\Big)^2 \\ &= \Big(\int \sqrt{p(x)q(x)}dx\Big)^2 = \exp\left(2\log \int p(x)\sqrt{\frac{q(x)}{p(x)}}dx\right) \\ &\geq \exp\left(\int p(x) \log\left\{\frac{q(x)}{p(x)}\right\}dx\right) = \exp\Big(-\mathbf{KL}(\mathbb{P}\|\mathbb{Q})\Big).\end{aligned}$$

□

References

Canonne, C. L. (2023), ‘A short note on an inequality between kl and tv’. arXiv:2202.07198v2.

Guntuboyina, A. (2011), ‘Lower bounds for the minimax risk using f-divergences, and applications’. arXiv:1002.0042v2.

Wainwright, M. J. (2019), *Basic tail and concentration bounds*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, p. 21–57.

Yu, B. (1997), *Assouad, Fano, and Le Cam*, Springer New York, New York, NY, pp. 423–435.