A NOTE ON MINIMAX OPTIMALITY

Wei Xiong *

June 17, 2022

1 Introduction

We are interested in the minimax optimality of various estimation procedures. In particular, we are interested in obtaining the mathching lower bound on estimation rates. This note is for Wainwright [2019].

1.1 Problem Setup

Given a class of distribution $\mathcal{P}$, let θ denote a mapping from a distribution $P \in \mathcal{P}$ to a parameter $\theta(P)$ taking value in some space Ω. We aim to estimate $\theta(P)$ based on a collection of samples $\{X_i\}$ i.i.d. drawn from P. Beyond the parameter estimation for the parametric distribution family, we may consider P supported on $[0,1]$ with differentiable density function f and estimate:

$$\theta(P) = \int_0^1 (f'(t))^2 dt \in \mathbb{R}.$$

Minimax risk. An estimator $\hat{\theta}$ can be viewed as a measurable function from $\mathcal{X}^n$ to the parameter space Ω. To assess the quality of any estimator, we consider a semi-metric $\rho(\hat{\theta}, \theta^*)$ where $\theta^* = \theta(P)$. Here θ^* is fixed, whereas $\hat{\theta}$ is a random variable so $\rho(\hat{\theta}, \theta^*)$ is also random. For any *fixed* θ^*, $\hat{\theta} \equiv \theta^*$ has zero risk but is meaningless as it can behave poorly for other choices of the parameter. To circumvent this and related difficulties, one may use the Bayesian approach to view θ^* as a random variable with some prior distribution. Another approach is to consider the *worst-case* risk $\sup_{P\in\mathcal{P}} \mathbb{E}_P \rho(\hat{\theta}, \theta(P))$. An optimal estimator in this sense defines a quantity known as *minimax risk*, namely,

$$R(\theta(\mathcal{P}); \rho) := \inf_{\hat{\theta}} \sup_{P\in\mathcal{P}} \mathbb{E}_P \rho(\hat{\theta}, \theta(P)), \tag{1.1}$$

where the infimum is taken over all possible estimators. If the estimator is based on n i.i.d. samples from P, we use R_n to denote the associated minimax risk. We can also extend the definition with a icreasing function ϕ:

$$R(\theta(\mathcal{P}); \phi(\rho)) := \inf_{\hat{\theta}} \sup_{P\in\mathcal{P}} \mathbb{E}_P \phi(\rho(\hat{\theta}, \theta(P))). \tag{1.2}$$

*The Hong Kong University of Science and Technology; email: wxiongae@connect.ust.hk.

A common choice is $\phi(t) = t^2$ which leads to the mean-squared error associated with ρ.

1.2 Preliminaries

We first introduce the notion of divergence measure. Let P and Q be two distributions on $\mathcal{X}$ with density p and q w.r.t. some base measure v. The *total-variation (TV) distance* is defined as

$$\|P - Q\|_{TV} := \sup_{A \in \mathcal{X}} |P(A) - Q(A)| = \frac{1}{2} \int_{\mathcal{X}} |p(x) - q(x)| v(dx), \tag{1.3}$$

where the equivalence of two definitions is explored in Wainwright [2019], Exercise 3.13. We also define the *Kullback–Leibler divergence* as

$$D(Q||P) = \int_{\mathcal{X}} q(x) \log \frac{q(x)}{p(x)} v(dx). \tag{1.4}$$

Unlike the total variation distance, KL-divergence is not actually a distance because it fails to be symmetric. A third distance is the *squared Hellinger distance*, given by

$$H^2(P||Q) := \int \left(\sqrt{p(x)} - \sqrt{q(x)} \right)^2 v(dx). \tag{1.5}$$

We have the following useful lemma.

Lemma 1. *For all distributions P and Q, we have*

$$\begin{aligned} \|P - Q\|_{TV} &\leq \sqrt{\frac{1}{2} D(Q||P)}, \\ \|P - Q\|_{TV} &\leq H(P||Q) \sqrt{1 - \frac{H^2(P||Q)}{4}}, \end{aligned} \tag{1.6}$$

where the first inequality is referred to as the Pinsker inequality and the second one is the Le Cam's inequality.

Packing. A δ-packing of a metric space (T, ρ) is a set $\{\theta^1, \cdots, \theta^M\} \subset T$ s.t. $\rho(\theta^i, \theta^j) \geq \delta$ for all $i \neq j$. The δ-packing number $M(\delta; T, \rho)$ is the cardinality of the largest δ-packing. (The inequality is not strict for convenience in the subsequent calculation.)

2 Le Cam's Method

The Le Cam's method relates the minimax risk to a Hypothesis testing whose risk can be further lower bounded by the divergence introduced in Section 1.2. This can be used to establish the statistical lower bound for the problems of interests.

Given a 2δ-packing of Ω: $\{\theta^j \in \Omega : j = 1, \cdots, M\}$, we select a representative distribution P_j for each θ^j to obtain $\{P_{\theta^j} : j = 1, \cdots, M\}$. We consider the following mixture distribution generating process.

- Sample a random integer J uniformly from $[M]$;
- Given $J = j$, sample $Z \sim P_{\theta^j} := P_j$.

We then denote Q as the joint distribution of (Z, J). Then, the marginal distribution of Z is $\bar{Q} := \frac{1}{M}\sum_{j=1}^{M} P_j$. We consider a *testing function* $\psi : \mathcal{Z} \to [M]$ for the M-ary hypothesis testing problem. So the associated probability of error is given by

$$Q(\psi(Z) \neq J).$$

We have

Lemma 2 (From estimation to testing.)**.** *For any increasing function ϕ and choice of 2δ-separated set, the minimax risk is lower bounded as*

$$R(\theta(\mathcal{P}); \phi(\rho)) \geq \phi(\delta) \inf_{\psi} Q[\psi(Z) \neq J]. \tag{2.1}$$

Note that $\phi(\delta)$ becomes smaller as $\delta \to 0$. On the other hand, as $\delta \to 0$, the testing becomes more difficult and so that we should expect $Q[\psi(Z) \neq J]$ grows as δ decreases. We usually pick a δ^* sufficiently small to ensure that $R(\theta(\mathcal{P}); \phi(\rho)) \geq c\phi(\delta^*)$ for some constant $c > 0$.

Proof. For any $P \in \mathcal{P}$ with $\theta = \theta(P)$, we have

$$\mathbb{E}_P \phi(\rho(\hat{\theta}, \theta)) \geq \phi(\delta) P[\phi(\rho(\hat{\theta}, \theta)) \geq \phi(\delta)] \geq \phi(\delta) P[\rho(\hat{\theta}, \theta) > \delta],$$

where the first inequality follows from Markov's inequality and the second one is because ϕ is increasing. Since the superimum is always larger than the average (of a subset), we have

$$\sup_{P \in \mathcal{P}} P[\rho(\hat{\theta}, \theta) > \delta] \geq \frac{1}{M}\sum_{j=1}^{M} P_j[\rho(\hat{\theta}, \theta^j) > \delta] = Q[\rho(\hat{\theta}, \theta^J) > \delta],$$

where the last inequality uses the joint distribution of (Z, J). It reduces to bound $Q[\rho(\hat{\theta}, \theta^J) > \delta]$ then. We define the test associated with an estimator $\hat{\theta}$ by

$$\psi(Z) := \operatorname*{argmin}_{\ell \in [M]} \rho(\theta^\ell, \hat{\theta}),$$

where the ties are broken arbitrarily. We claims that $\{\rho(\theta^j, \hat{\theta}) < \delta\}$ implies that the test is correct. This is because for any other $k \in [M]$,

$$\rho\left(\theta^k, \hat{\theta}\right) \geq \underbrace{\rho\left(\theta^k, \theta^j\right)}_{\geq 2\delta} - \underbrace{\rho\left(\theta^j, \hat{\theta}\right)}_{< \delta} > 2\delta - \delta = \delta > \rho(\theta^j, \hat{\theta}),$$

where we use $\{\phi^k; k \in [M]\}$ is a 2δ-packing. By the decision rule of the test, we must have $\psi(Z) = j$. This implies that

$$P_j[\rho(\hat{\theta}, \theta^j) > \delta] \geq P_j(\psi(Z) \neq j),$$

and

$$Q\left[\rho\left(\hat{\theta},\theta^J\right)\geq\delta\right]=\frac{1}{M}\sum_{j=1}^{M}P_j\left[\rho\left(\hat{\theta},\theta^j\right)\geq\delta\right]\geq Q[\psi(Z)\neq J].$$

Combined with

$$\mathbb{E}_P\phi(\rho(\hat{\theta},\theta))\geq\phi(\delta)P[\phi(\rho(\hat{\theta},\theta))\geq\phi(\delta)]\geq\phi(\delta)P[\rho(\hat{\theta},\theta)>\delta],$$

and

$$\sup_{P\in\mathcal{P}}P[\rho(\hat{\theta},\theta)>\delta]\geq\frac{1}{M}\sum_{j=1}^{M}P_j[\rho(\hat{\theta},\theta^j)>\delta]=Q[\rho(\hat{\theta},\theta^J)>\delta],$$

we have

$$\sup_{P\in\mathcal{P}}\mathbb{E}_P\phi(\rho(\hat{\theta},\theta))\geq\phi(\delta)Q[\psi(Z)\neq J].$$

Finally, we take a infimum on the RHS for ψ to obtain the desired lemma. □

2.1 Binary Testing

In the binary testing case, the mixture distribution is

$$\bar{Q}=\frac{1}{2}P_0+\frac{1}{2}P_1.$$

For a fixed $\psi:\mathcal{Z}\to\{0,1\}$, the associated probability of error is

$$Q[\psi(Z)\neq J]=\frac{1}{2}P_0(\psi(Z)\neq 0)+\frac{1}{2}P_1(\psi(Z)\neq 1). \tag{2.2}$$

Lemma 3 (Error probability to TV distance)**.** *We have*

$$\inf_{\psi}Q[\psi(Z)\neq J]=\frac{1}{2}[1-\|P_1-P_0\|_{TV}]. \tag{2.3}$$

This implies that

$$R(\theta(\mathcal{P});\phi(\rho))\geq\frac{\phi(\delta)}{2}[1-\|P_1-P_0\|_{TV}]. \tag{2.4}$$

Proof. We define $A:=\{x\in\mathcal{X}:\psi(x)=1\}$. Then,

$$\sup_{\psi}Q[\psi(Z)=J]=\sup_{A\subset\mathcal{X}}[\frac{1}{2}P_1(A)+\frac{1}{2}P_0(A^c)]=\sup_{A\subset\mathcal{X}}\frac{1}{2}[P_1(A)-P_0(A)]+\frac{1}{2}=\frac{1}{2}\|P_0-P_1\|_{TV}+\frac{1}{2}.$$

The result follows from

$$\sup_{\psi}Q[\psi(Z)=J]=1-\inf_{\psi}Q[\psi(Z)\neq J].$$

□

2.2 Parametric Examples

We consider $P_\theta = \{N(\theta, \sigma^2) : \theta \in \mathbb{R}\}$ with fixed σ^2 and consider $\rho(\theta, \theta') := |\theta - \theta'|$ and $\rho(\theta, \theta') := |\theta - \theta'|^2$. We are given a collection $Z = (Y_1, \cdots, Y_n)$ of i.i.d. samples drawn from a $N(\theta, \sigma^2)$ distribution and we use P_θ^n to denote this product distribution. We set $\theta = 2\delta$ to ensure that 0 and θ are 2δ-separate. Here δ is fixed and is to be specified later. We have (proved at the end of this example)

$$\|P_\theta^n - P_0^n\|_{TV}^2 \le \frac{1}{4}[e^{4n\delta^2/\sigma^2} - 1]$$

and thus

$$1 - \|P_\theta^n - P_0^n\|_{TV} \ge 1 - \frac{1}{2}\sqrt{e^{4n\delta^2/\sigma^2} - 1}.$$

Setting $\delta = \frac{1}{2}\frac{\sigma}{\sqrt{n}}$, Eqn. (2.4) yields

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta[\hat{\theta} - \theta\ |] \ge \frac{\delta}{2}\left\{1 - \frac{1}{2}\sqrt{e-1}\right\} \ge \frac{\delta}{6} = \frac{1}{12}\frac{\sigma}{\sqrt{n}}$$

and

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

It remains to bound the total-variation distance. First, we have

$$\|P, Q\|_{TV}^2 \le \frac{1}{2}\int p \log \frac{p}{q}\, \mathrm{d}\nu \le \frac{1}{2}\log \int \frac{p^2}{q}\, \mathrm{d}\nu \le \frac{1}{2}\left(\int \frac{p^2}{q}\, \mathrm{d}\nu - 1\right),$$

where the first inequality uses Lemma 1; the second inequality uses Jesen's inequality, and the last one is because of $\ln z \le z - 1$. It remains to bound the integral for $N(\theta, \sigma^2)$ and $N(0, \sigma^2)$:

$$\left[\int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left[(x - 2\theta)^2 - 2\theta^2\right]\right\} dx\right]^n = \exp\left\{\left(\frac{\sqrt{n}\theta}{\sigma}\right)^2\right\},$$

where we use the integral of the density is 1 in the equality. Note that the constant here is inferior. On the other hand, the scalings $\sigma/\sqrt{n}$ and σ^2/n are sharp because the sample mean estimator $\bar{X}_n$ satisfies

$$\sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta\left[\left|\bar{X}_n - \theta\right|\right] = \sqrt{\frac{2}{\pi}}\frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta\left[\left(\bar{X}_n - \theta\right)^2\right] = \frac{\sigma^2}{n}.$$

References

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.