Lecture-23: Reduction to hypothesis testing

1 Introduction

We study three commonly used techniques for proving minimax lower bounds, (a) Le Cam's method, (b) Assouad's lemma, and (c) Fano's method. Compared to the results for large-sample asymptotics in smooth parametric models, the approach here is more generic, less tied to mean-squared error, and applicable in nonasymptotic settings such as nonparametric or high-dimensional problems.

The common rationale of all three methods is reducing statistical estimation to hypothesis testing. Specifically, we lower bound the minimax risk $R^*(\Theta)$ for the parameter space Θ in the following steps.
Step 1. We notice that $R^*(\Theta) \geqslant R^*(\Theta')$ for any subcollection $\Theta' \subset \Theta$.
Step 2. From mutual information method, we have for any choice of prior π,

$$R^*_\pi \geqslant \phi^{-1}(I(\theta;X)) \geqslant \phi^{-1}(C).$$

Step 3. Choose a suitable prior $\pi \in \mathcal{M}(\Theta')$, to obtain

$$R^*(\Theta) \geqslant R^*(\Theta') \geqslant R^*_\pi \geqslant \phi^{-1}(C).$$

Remark 1. Le Cam, Assouad, and Fano's methods amount to choosing Θ' to be a two-point set, a hypercube, or a packing, respectively. In particular, Le Cam's method reduces the estimation problem to binary hypothesis testing. This method is perhaps the easiest to evaluate; however, the disadvantage is that it is frequently loose in estimating high-dimensional parameters. To capture the correct dependency on the dimension, both Assouad's and Fano's method rely on reduction to testing multiple hypotheses.

Remark 2. All three methods in fact follow from the common principle of the mutual information method (MIM), corresponding to different choice of priors. The limitation of these methods, compared to the MIM, is that, due to the looseness in constant factors, they are ineffective for certain problems such as estimation better than chance discussed.

1.1 Le Cam's two-point method

Definition 1.1. Let $\alpha > 0$ and a parameter space Θ with any three parameters $\theta_0, \theta_1, \theta \in \Theta$. We call $L : \Theta \times \Theta \to \mathbb{R}_+$ an α-metric on Θ, if it satisfies
(a) symmetry, i.e. $L(\theta_0,\theta_1) = L(\theta_2,\theta_1)$,
(b) positivity, i.e. $L(\theta_0,\theta_1) \geqslant 0$ with equality iff $\theta_0 = \theta_1$, and
(c) α-triangle inequality, i.e. $L(\theta_0,\theta_1) \leqslant \alpha(L(\theta_0,\theta) + L(\theta_1,\theta))$.

Theorem 1.2. *Consider a simple statistical decision theory setting with $\Theta = \hat{\Theta}$, and loss function $L : \Theta \times \Theta \to \mathbb{R}_+$ that is an α-metric on parameter space Θ. Then, the minimax risk $R^*(\Theta) \triangleq \inf_{\hat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}_\theta L(\theta,\hat{\theta})$ satisfies*

$$R^*(\Theta) \geqslant \sup_{\theta_0,\theta_1\in\Theta} \frac{L(\theta_0,\theta_1)}{2\alpha}(1 - \mathrm{TV}(P_{\theta_0}, P_{\theta_1})). \tag{1}$$

Proof. Fix parameters $\theta_0, \theta_1 \in \Theta$, a loss function L as defined in theorem hypothesis, and estimator $\hat{\theta} : \mathcal{X} \to \Theta$. We define $p(X) \triangleq \frac{L(\theta_1,\hat{\theta}(X))}{L(\theta_0,\hat{\theta}(X))+L(\theta_1,\hat{\theta}(X))}$, and the following randomized test $\tilde{\theta} : \mathcal{X} \times [0,1] \to \{\theta_0,\theta_1\}$ for any independent uniform random variable $U : \Omega \to [0,1]$, such that

$$\tilde{\theta}(X,U) \triangleq \theta_0 \mathbb{1}_{\{U\leqslant p\}} + \theta_1 \mathbb{1}_{\{U>p\}}.$$

We observe that the probability of errors are

$$E_{X\sim P_{\theta_0}} \mathbb{1}_{\{\tilde{\theta}=\theta_1\}} = \mathbb{E}_{X\sim P_{\theta_0}}(1 - p(X)), \qquad \mathbb{E}_{X\sim P_{\theta_1}} \mathbb{1}_{\{\tilde{\theta}=\theta_0\}} = \mathbb{E}_{X\sim P_{\theta_1}} p(X).$$

Using α-triangle inequality for loss function L, we observe that $\bar{p}(X) \leqslant \alpha \frac{L(\theta_0,\hat{\theta}(X))}{L(\theta_0,\theta_1)}$ and $p(X) \leqslant \alpha \frac{L(\theta_1,\hat{\theta}(X))}{L(\theta_0,\theta_1)}$. Therefore, it follows for $i \in \{0,1\}$

$$\mathbb{E}_{X\sim P_{\theta_i}}[L(\theta_i,\tilde{\theta}(X,U))] = L(\theta_0,\theta_1)\mathbb{E}_{X\sim P_{\theta_0}} \mathbb{1}_{\left\{\tilde{\theta}=\theta_1\right\}} \leqslant \alpha \mathbb{E}_{X\sim P_{\theta_i}}[L(\theta_i,\hat{\theta}(X))].$$

We assume $\theta \sim \pi$ taking the prior $\pi \triangleq \frac{1}{2}(\delta_{\theta_0} + \delta_{\theta_1})$. We can write the expectation of loss to obtain the following lower bound on the Bayes risk,

$$R_\pi(\Theta) \geqslant \mathbb{E}_{\theta\sim\pi}\mathbb{E}_{X\sim P_\theta}[L(\theta,\hat{\theta}(X))] = \sum_{i=0}^{1} \pi_{\theta_i}\mathbb{E}_{X\sim P_{\theta_i}} L(\theta_i,\hat{\theta}(X)) \geqslant \frac{1}{\alpha}\sum_{i=0}^{1} \pi_{\theta_i}\mathbb{E}_{X\sim P_{\theta_i}} L(\theta_i,\tilde{\theta}(X,U)).$$

Using the symmetry of loss function L and from them minimum average probability of error in binary hypothesis testing in Theorem A.2, we obtain

$$R_\pi(\Theta) \geqslant \frac{L(\theta_0,\theta_1)}{\alpha}\sum_{i=0}^{1} \pi_{\theta_i} P_{X\sim P_{\theta_i}} \mathbb{1}_{\left\{\theta_i\neq\tilde{\theta}(X,U)\right\}} = \frac{L(\theta_0,\theta_1)}{\alpha} P\left\{\tilde{\theta} \neq \theta\right\} \geqslant \frac{L(\theta_0,\theta_1)}{2\alpha}(1 - \mathrm{TV}(P_{\theta_0},P_{\theta_1})).$$

□

Example 1.3 (Binary hypothesis testing). Consider a binary hypothesis testing problem with $\Theta \triangleq \{\theta_0,\theta_1\}$ and the Hamming loss $L(\theta,\hat{\theta}) = \mathbb{1}_{\{\theta\neq\hat{\theta}\}}$, where $\theta,\hat{\theta} \in \Theta$ and $\alpha = 1$. Let P_e be the probability of error as defined in (2). Then the left side of (1) is the minimax probability of error, and the right side of (1) is the optimal average probability of error since $P_e(\hat{\theta}) \geqslant 1 - \mathrm{TV}(P,Q)$ for any estimator $\hat{\theta} : \mathcal{X} \times [0,1] \to \Theta$ from (3).

Remark 3. Binary hypothesis testing is an example where the bound (1) is tight up to constants. In fact, these two quantities can coincide, for example for Gaussian location model.

Remark 4. Another special case of interest is the quadratic loss $L(\theta,\hat{\theta}) = \left\|\theta - \hat{\theta}\right\|^2$, where $\theta,\hat{\theta} \in \mathbb{R}^d$, which satisfies the α-triangle inequality with $\alpha = 2$. In this case, the leading constant $\frac{1}{4}$ in (1) makes sense, because in the extreme case of $\mathrm{TV} = 0$ where P_{θ_0} and P_{θ_1} cannot be distinguished, the best estimate is simply $\frac{\theta_0+\theta_1}{2}$.

Remark 5. The inequality (1) can also be deduced based on properties of f-divergences and their joint range. We consider the prior $\pi = \frac{1}{2}(\delta_{\theta_0} + \delta_{\theta_1})$. Then the Bayes estimator is the posterior mean and given by $\frac{\theta_0 dP_{\theta_0} + \theta_1 dP_{\theta_1}}{dP_0 + dP_1}$ and the Bayes risk is given by

$$R^*_\pi = \frac{1}{2}\left\|\theta_0 - \theta_1\right\|^2 \int_{\mathcal{X}} \frac{dP_{\theta_0} dP_{\theta_1}}{dP_{\theta_0} + dP_{\theta_1}} = \frac{1}{4}\left\|\theta_0 - \theta_1\right\|^2 (1 - \mathrm{LC}(P_{\theta_0},P_{\theta_1})) \geqslant \frac{1}{4}\left\|\theta_0 - \theta_1\right\|^2 (1 - \mathrm{TV}(P_{\theta_0},P_{\theta_1})),$$

where $\mathrm{LC}(P_{\theta_0},P_{\theta_1}) = \int_{\mathcal{X}} \frac{(dP_{\theta_0} - dP_{\theta_1})^2}{dP_{\theta_0} + dP_{\theta_1}}$ is the Le Cam divergence.

Remark 6. Let $P,Q \ll \mu$ for measures $P,Q,\mu \in \mathcal{M}(\mathcal{X})$, so that we can define relative densities $p \triangleq \frac{dP}{d\mu}$ and $q \triangleq \frac{dQ}{d\mu}$. Then, we observe that $|p(x) - q(x)| \leqslant \frac{(p(x)-q(x))^2}{p(x)+q(x)}$ since $\frac{|p(x)-q(x)|}{p(x)+q(x)} \leqslant 1$. It follows that $\mathrm{LC} \leqslant \mathrm{TV}$.

A Total variation distance

Definition A.1 (Binary hypothesis testing). The *binary hypothesis testing* problem is formulated as follows. One is given an observation $X : \Omega \to \mathcal{X}$ with two possible hypotheses. The null-hypothesis H_0 implies that $X \sim P$, and the alternative hypothesis H_1 implies that $X \sim Q$. The goal is to decide, on the basis of observation X alone, which of the two hypotheses holds. In other words, we want to find a possibly randomized decision function $\phi : \mathcal{X} \times [0,1] \to \{0,1\}$ such that the probability of error is minimized, where probability of error is defined as the sum of two types of probabilities of error

$$P_e(\phi) \triangleq P\left\{\phi(X,U) = 1\right\} + Q\left\{\phi(X,U) = 0\right\}. \tag{2}$$

Theorem A.2 (Minimum probability of error). *For total variation distance* $\mathrm{TV} : \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \mathbb{R}_+$ *and family of real-valued functions* $\mathcal{G} \triangleq \left\{f \in \mathbb{R}^{\mathcal{X}} : \|f\|_\infty \leqslant 1\right\}$*, the following representations are true for* $P, Q \in \mathcal{M}(\mathcal{X})$*.*

(a) ***sup-representation.*** $\mathrm{TV}(P,Q) = \sup_{X^{-1}(E)\in\mathcal{F}} P\{X \in E\} - Q\{X \in E\} = \frac{1}{2}\sup_{f\in\mathcal{G}} \mathbb{E}_P f(X) - \mathbb{E}_Q f(X)$*. In particular, the minimal total error probability in* (2) *is given by*

$$\min\left\{P\{\phi(X,U)=1\} + Q\{\phi(X,U)=0\} : \phi \in \{0,1\}^{\mathcal{X}\times[0,1]}\right\} = 1 - \mathrm{TV}(P,Q). \tag{3}$$

(b) ***inf-representation.*** *If the diagonal* $\{X = Y\} \in \mathcal{F}$ *is measurable, then*

$$\mathrm{TV}(P,Q) = \min\left\{P_{X,Y}\{X \neq Y\} : P_{X,Y} \in \mathcal{M}(\mathcal{X}\times\mathcal{X}), P_X = P, P_Y = Q\right\}, \tag{4}$$

where minimization is over joint distributions $P_{X,Y}$ *with the property* $P_X = P$ *and* $P_Y = Q$*, which are called couplings of* P *and* Q*.*

Proof. Let $P, Q \ll \mu$ for some dominating measure $\mu \in \mathcal{M}(\mathcal{X})$ and denote the conditional densities $p \triangleq \frac{dP}{d\mu}, q \triangleq \frac{dQ}{d\mu}$. By definition, $\mathrm{TV}(P,Q) = \frac{1}{2}\int_{\mathcal{X}} |p(x) - q(x)|\, d\mu(x)$.

(a) For any $f \in \mathcal{G}$, we have $\int_{\mathcal{X}} f(x)(p(x)-q(x))d\mu \leqslant \int_{\mathcal{X}} |p(x)-q(x)|\, d\mu = 2\mathrm{TV}(P,Q)$. This implies that $\mathrm{TV}(P,Q) \geqslant \frac{1}{2}(\mathbb{E}_P f(X) - \mathbb{E}_Q f(X))$ for all $f \in \mathcal{G}$. For any $E \in \sigma(X)$, we can define $f \triangleq 2\mathbb{1}_E - 1 \in \mathcal{G}$ to obtain $\frac{1}{2}(\mathbb{E}_P f(X) - \mathbb{E}_Q f(X)) = P\{X \in E\} - Q\{X \in E\}$. It follows that

$$\mathrm{TV}(P,Q) \geqslant \sup_{f\in\mathcal{G}} \frac{1}{2}(\mathbb{E}_P f(X) - \mathbb{E}_Q f(X)) \geqslant \sup_{X^{-1}(E)\in\mathcal{F}} (P\{X \in E\} - Q\{X \in E\}).$$

For the converse, we take $E \triangleq \{x \in \mathcal{X} : p(x) > q(x)\}$ and notice that

$$0 = \int_{\mathcal{X}} (p(x)-q(x))d\mu = \int_E (p(x)-q(x))d\mu - \int_{E^c} (q(x)-p(x))d\mu.$$

It follows that $\int_E (p(x)-q(x))d\mu(x) = \int_{E^c}(q(x)-p(x))d\mu(x)$ and hence this choice of E attains the supremum, i.e.

$$\mathrm{TV}(P,Q) = \frac{1}{2}\int_{\mathcal{X}} |p(x)-q(x)|\, d\mu(x) = \int_E (p(x)-q(x))d\mu(x) = P\{X \in E\} - Q\{X \in E\}.$$

For any $E \in \sigma(X)$, we define a detector $\phi(X,U) \triangleq \mathbb{1}_{\{X\notin E\}}$ to obtain $P_e = P\{X \notin E\} + Q\{X \in E\} = 1 - (P\{X \in E\} - Q\{X \in E\}) \geqslant 1 - \mathrm{TV}(P,Q)$, where the equality is achieved for $E = \{x \in \mathcal{X} : p(x) > q(x)\}$.

(b) For the inf-representation, we notice that given a coupling $P_{X,Y}$ such that marginals $P_X = P$ and $P_Y = Q$, we have $\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)] = \mathbb{E}[f(X) - f(Y)] \leqslant 2P_{X,Y}\{X = Y\}$ for any $f : \mathcal{X} \to \mathbb{R}$ such that $\|f\|_\infty \leqslant 1$. Since $\mathrm{TV}(P,Q) = \frac{1}{2}\sup_{f\in\mathbb{R}^{\mathcal{X}}:\|f\|_\infty\leqslant 1} \mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)] \leqslant P_{X,Y}\{X \neq Y\}$, it follows that the inf-representation is always an upper bound. To show that this bound is tight, we construct the maximal coupling. We define probability $\pi \triangleq \int_{\mathcal{X}} (p(x) \wedge q(x))d\mu(x)$, and the following densities

$$r(x) \triangleq \frac{1}{\pi} p(x) \wedge q(x), \quad p_1(x) \triangleq \frac{1}{1-\pi}(p(x) - p(x)\wedge q(x)), \quad q_1(x) \triangleq \frac{1}{1-\pi}(q(x) - p(x)\wedge q(x)).$$

We assume that $U : \Omega \to [0,1]$ is an independent uniform random variable, and $V, W, Z : \Omega \to \mathbb{R}$ are independent random variables with densities p_1, q_1, r respectively, We define the coupling as

$$X \triangleq Z\mathbb{1}_{\{U\leqslant\pi\}} + V\mathbb{1}_{\{U>\pi\}}, \qquad Y \triangleq Z\mathbb{1}_{\{U\leqslant\pi\}} + W\mathbb{1}_{\{U>\pi\}}.$$

That is, $X = Y = Z$ with probability π, where Z is random and sampled from a distribution with density r, and with probability $1-\pi$, the random variables X, Y are independently from densities p_1, q_1 respectively. We observe that $P_X, P_Y \ll \mu$ and the relative densities are given as

$$\frac{dP_X}{d\mu} = \pi r + (1-\pi)p_1 = p, \qquad \frac{dP_Y}{d\mu} = \pi r + (1-\pi)q_1 = q.$$

That is, the joint distribution $P_{X,Y}$ is indeed a coupling of P and Q. Further, since $\mathrm{TV}(P,Q) = 1 - \int_{\mathcal{X}}(p \wedge q)d\mu$, we get

$$P_{X,Y}\{X \neq Y\} = 1 - \pi = \mathrm{TV}(P,Q).$$

□