Lecture-17: Minimax bounds

1 Introduction

When solving a statistical learning problem, there are often many procedures to choose from. This leads to the following question: how can we tell if one statistical learning procedure is better than another? One answer is provided by *minimax theory* which is a set of techniques for finding the minimum, worst case behavior of a procedure. We provide upper and lower bounds on error achievable by any algorithm regardless of its complexity and storage. Estimation problems under consideration are:

1. hypothesis testing,
2. point estimation, and
3. regression for parameter estimation

2 Definitions and notations

We denote a set of distributions by $\mathcal{P}$ over input space $\mathcal{X}$, such that $X : \Omega \to \mathcal{X}^m$ is an *i.i.d.* random vector with some distribution $P \in \mathcal{P}$. Let $\theta : \mathcal{P} \to \Omega$ be some function of probability distribution P, where Ω can be finite or infinite dimensional and the map may not be bijective. For example $\theta(P)$ can be the mean, variance, or density of P. We want to estimate $\theta(P)$ from an *i.i.d.* sample X with an unknown distribution $P \in \mathcal{P}$. There are policies, where we consider dependent samples. We measure the quality of estimator using the metric $\rho : \Omega \times \Omega \to \mathbb{R}_+$.

Example 2.1. Let $\hat{\theta}$ be an estimate of θ from sample S, then $\rho \triangleq \mathbb{E}\left\|\theta - \hat{\theta}\right\|$ is a potential cost function. If $\hat{\theta}_S = \theta_0$ then if actual distribution P from which sample comes from is P_{θ_0} and the estimate is perfect. If P_θ is away from P_{θ_0}, then the error is $\rho(\theta,\theta_0)$.

Remark 1. The Bayesian approach is to find the parameter $\hat{\theta}$ that minimizes the Bayesian risk $\mathbb{E}_\pi[\rho(\theta,\hat{\theta})]$ given a prior distribution π.

Definition 2.2 (Minimax approach). Find the parameter θ that minimizes the worst case performance of the estimator. That is, minimax estimator is the one that minimizes the worst case risk

$$R_m(\hat{\theta}) \triangleq \sup_{P \in \mathcal{P}} \mathbb{E}_P[\rho(\hat{\theta},\theta(P))].$$

Remark 2. We also want to get lower bounds on the worst case risk. Minimax estimators may not provide good performance in real life.

Definition 2.3 (Minimax risk). The minimax risk is defined as $R_m \triangleq \inf_{\hat{\theta}} R_m(\hat{\theta})$.

Example 2.4. Let $N(\theta,1)$ denote a Gaussian distribution with mean θ and variance 1, and the family of distributions $\mathcal{P} = \{N(\theta,1) : \theta \in \mathbb{R}\}$. Consider estimating mean θ with metric $\rho(a,b) = (a-b)^2$. The minimax risk is $R_m = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P(\hat{\theta} - \theta)^2$.

Example 2.5. Consider an *i.i.d.* labeled sample $S \in (\mathcal{X} \times \mathcal{Y})^m$ from distribution $P \in \mathcal{P}$, and regression function $m(x) \triangleq \mathbb{E}_P[Y \mid \{X = x\}]$. Consider estimating regression function with metric $\rho(m_1,m_2) = \int (m_1(x) - m_2(x))^2 dx$. The minimax risk is $R_m = \inf_{\hat{m}} \sup_{P \in \mathcal{P}} \mathbb{E}_P(\hat{m}(x) - m(x))^2 dx$.

Definition 2.6. For real valued parameter θ, if the minimax estimator is unique, then it is **admissible**.

Remark 3. If minimax estimator has additional good properties, such as being a Bayes estimate, then it is often a good estimator.

Definition 2.7 (Admissible estimator). An estimator $\hat{\theta}$ of parameter θ is **admissible**, if no other estimator is better than $\hat{\theta}$ uniformly over $\mathcal{P}$. That is, there exists no other θ' such that for all $P \in \mathcal{P}$

$$\mathbb{E}_P[\rho(\hat{\theta},\theta(P))] \geqslant \mathbb{E}_P[\rho(\theta',\theta(P))].$$

3 Bounding the minimax risk

The way we find risk R_m is to find an upper bound and a lower bound. Upper bound can be obtained by considering some estimator and getting bounds on its error. Let $\hat{\theta}$ be any estimator, then we observe that

$$R_m = \inf_{\hat{\theta}} R_m(\hat{\theta}) \leqslant R_m(\hat{\theta}) = U_m.$$

So the maximum risk of any estimator provides an upper bound U_m. Finding a lower bound L_m is harder. We will consider two methods: the Le Cam method and the Fano method. If the lower and upper bound are close, then we have succeeded. For example, if $L_m = c_m^{-\alpha}$ and $U_m = Cm^{-\alpha}$ for some positive constants c, C and α, then we have established that the minimax rate of convergence is $m^{-\alpha}$.

Remark 4. Lower bounds can be used to show if a particular estimator is *optimal*. If we don't have an estimator which matches the lower bound, then we know we have work to do. Either, the lower bound may not be tight enough and we may look for better bound. Or, we may look for better estimator which can meet the lower bound.

4 Lower bound on minimax risk

For a map $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$, we can write the minimax bound for $\theta(P)$ with metric $\Phi \circ \rho$

$$R_m(\Phi \circ \rho) \triangleq \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[\Phi \circ \rho(\theta,\hat{\theta})].$$

Remark 5 (LeCam's bound via testing). We have a problem of estimation which can be lower bounded by addressing a testing problem. We can define the set of possible parameters $\theta(\mathcal{P}) \triangleq= \{\theta(P) : P \in \mathcal{P}\}$, and take M points from $\theta(\mathcal{P})$ as $\left\{\theta^{(1)},\dots,\theta^{(M)}\right\}$ such that $\rho(\theta^{(i)},\theta^{(j)}) \geqslant 2\delta$ for all $i \neq j \in [M]$ and fixed $\delta > 0$.

Consider the following setting. Let $J : \Omega \to [M]$ be a uniform random variable, and the conditional distribution of labeled sample Z given $J = j$ is denoted by $P_{\theta^{(j)}}$ for all $j \in [M]$. We denote the joint distribution of pair (Z, J) by Q. That is,

$$Q(z,j) = P\{Z = z, J = j\} = \frac{1}{M} P_{\theta^{(j)}}(z).$$

Marginal distribution of Z is written as

$$\bar{Q}(z) \triangleq P\{Z = z\} = \frac{1}{M} \sum_{j \in [M]} P_{\theta^{(j)}}(z).$$

We consider the following hypothesis testing problem. Given a sample of Z, find the label $j \in [M]$ from which Z is generated. Let $\psi : \mathcal{Z}^m \to [M]$ be the classifier for this problem.

Proposition 4.1. *If Φ is nondecreasing, then for any classifier $\psi : \mathcal{Z}^m \to [M]$, we can write*

$$R_m(\Phi \circ \rho) \geqslant \Phi(\delta) \inf_{\psi} Q\{\psi(z) \neq J\}.$$

Proof. Consider a fixed $P \in \mathcal{P}$ and $\theta = \theta(P)$. Using the nonnegativity and monotone nondecreasing property of Φ, we can lower bound the mean

$$\mathbb{E}_P \Phi(\rho(\hat{\theta},\theta)) \geqslant \mathbb{E}_P[\Phi(\rho(\hat{\theta},\theta)) \mathbb{1}_{\{\rho(\hat{\theta},\theta) \geqslant \delta\}}] \geqslant \Phi(\delta) P\{\rho(\hat{\theta},\theta) \geqslant \delta\}.$$

We next focus on lower bounding the probability term. Since reducing the set decreases the supremum, we get

$$\sup_{P\in\mathcal{P}} P\left\{\rho(\hat{\theta},\theta)\geqslant\delta\right\}\geqslant\sup_{j\in[M]} P_{\theta^{(j)}}\left\{\rho(\hat{\theta},\theta^{(j)})\geqslant\delta\right\}\geqslant\frac{1}{M}\sum_{j=1}^{M}P_{\theta^{(j)}}\left\{\rho(\hat{\theta},\theta^{(j)})\geqslant\delta\right\}.$$

For any given estimator $\hat{\theta}$ of θ form the sample $z\in\mathcal{Z}^m$, we define M-ary classifier ψ for the testing problem as

$$\psi(z)\triangleq\arg\min_{\ell\in[M]}\rho(\theta^{(\ell)},\hat{\theta}).$$

The probability of error for classifier ψ is $P_{\theta^{(j)}}\left\{\psi(z)\neq j\right\}$ when z has distribution $P_{\theta^{(j)}}$. We observe that

$$P_{\theta^{(j)}}\left\{\psi(z)\neq j\right\}\leqslant P_{\theta^{(j)}}\left\{\rho(\hat{\theta},\theta^{(j)})\geqslant\delta\right\}.$$

From the definition of distribution Q, we can write

$$Q\left\{\rho(\hat{\theta},\theta^{(J)})\geqslant\delta\right\}=\sum_{j=1}^{M}P\left\{J=j\right\}P_{\theta^{(j)}}\left\{\rho(\hat{\theta},\theta^{(j)})\geqslant\delta\right\}=\frac{1}{M}\sum_{j=1}^{M}P_{\theta^{(j)}}\left\{\rho(\hat{\theta},\theta^{(j)})\geqslant\delta\right\}.$$

Combining the two results, we get $Q\left\{\rho(\hat{\theta},\theta^{(J)})\geqslant\delta\right\}\geqslant Q\left\{\psi(z)\neq J\right\}$, and the result follow. □

4.1 Distance measure on family of distributions

Definition 4.2. Total variation distance between two distribution $P,Q\in\mathcal{P}$ can be defined as

$$\|P-Q\|_{\mathrm{TV}}\triangleq\sup_{A}|P(A)-Q(A)|.$$

Remark 6. If P,Q have densities p,q respectively, then

$$\|P-Q\|_{\mathrm{TV}}=\frac{1}{2}\int|p(x)-q(x)|dx.$$

Definition 4.3. The Kullback-Leibler distance between two distributions $P,Q\in\mathcal{P}$ is defined as

$$\mathrm{KL}(P\|Q)\triangleq\int\log\Big(\frac{dP}{dQ}\Big)dP.$$

Remark 7. If P,Q have densities p,q respectively, then

$$\mathrm{KL}(P\|Q)=\int p(x)\log\frac{p(x)}{q(x)}dx.$$

Definition 4.4. The squared Hellinger distance between two distributions $P,Q\in\mathcal{P}$ is defined as

$$H^2(P\|Q)\triangleq\int(\sqrt{p(x)}-\sqrt{q(x)})^2dx.$$

Theorem 4.5 (Pinsker). $\|P-Q\|_{\mathrm{TV}}\leqslant\sqrt{\frac{1}{2}D(Q\|P)}$.

Lemma 4.6 (Le Cam). $\|P-Q\|_{\mathrm{TV}}\leqslant H(P\|Q)\sqrt{1-\frac{1}{4}H^2(P\|Q)}$.

Remark 8. Taking $M=2$ in the lower bound in Proposition 4.1, we get

$$Q\left\{\psi(Z)\neq J\right\}=\frac{1}{2}P_0\left\{\psi(z)\neq 0\right\}+\frac{1}{2}P_1\left\{\psi(z)\neq 1\right\}.$$

Classifier $\psi:\mathcal{Z}^m\to\{0,1\}$ can be identified as a partition $\{A^c,A\}$ corresponding to the decision regions where $A\triangleq\{z\in\mathcal{Z}^m:\psi(z)=1\}$. We can write

$$\sup_{\psi}Q\left\{\psi(z)=J\right\}=\sup_{A}\Big[\frac{1}{2}P_0(A^c)+\frac{1}{2}P_1(A)\Big]=\frac{1}{2}+\frac{1}{2}\sup_{A}(P_1(A)-P_0(A))=\frac{1}{2}+\frac{1}{2}\|P_1-P_0\|_{\mathrm{TV}}.$$

It follows that $\inf_{\psi}Q\left\{\psi(z)\neq J\right\}=1-\sup_{\psi}Q\left\{\psi(z)=J\right\}=\frac{1}{2}(1-\|P_1-P_0\|_{\mathrm{TV}})$. Together with Pinsker's inequality, we have for $M=2$,

$$R_m(\Phi\circ\rho)\geqslant\frac{\Phi(\delta)}{2}(1-\|P_1-P_0\|_{\mathrm{TV}})\geqslant\frac{\Phi(\delta)}{2}(1-\sqrt{\frac{1}{2}D(P_0\|P_1)}).$$

Example 4.7. Consider the estimation problem for Gaussian random variables considered in Example 2.4 with fixed and known variance σ^2 and unknown mean θ. We apply Le Cam's method for $M = 2$ and fixed $\delta > 0$, taking parameters in $\{0,2\delta\}$. Then,

$$\|P_0^m - P_1^m\|_{\text{TV}} \leqslant \frac{1}{4}(e^{\frac{4m\delta^2}{\sigma^2}} - 1)$$

Fixing $\delta = \frac{\sigma}{2\sqrt{m}}$ and recalling that $\Phi(\delta) = \delta^2$, we get

$$\inf_{\hat{\theta}} \sup_{\theta\in\mathbb{R}} \mathbb{E}_\theta(\theta - \hat{\theta})^2 \geqslant \frac{\delta^2}{2}(1 - \frac{1}{2}\sqrt{e-1}) \geqslant \frac{\sigma^2}{24m}.$$

Using the empirical mean estimator $\hat{\theta} = \frac{1}{m}\sum_{i=1}^m X_i$, we observe that $\sup_{\theta\in\mathbb{R}} \mathbb{E}_\theta(\hat{\theta} - \theta)^2 = \frac{\sigma^2}{m}$. The lower bound is *tight* in terms of the order, though the constant $\frac{1}{24}$ is not tight.

Example 4.8. Consider the estimation problem for Gaussian random variables with unknown mean θ and variance σ^2. In this case, for $\rho(\theta,\hat{\theta}) = (\theta - \hat{\theta})^2$, we observe that the loss function

$$\mathbb{E}_\theta \rho(\theta,\hat{\theta}) = \text{Var}(\hat{\theta}) + (\text{bias})^2.$$

For the MLE of mean θ for this problem, we can show that $\text{Var}_\theta(\hat{\theta}) \approx o\big(\frac{1}{m}\big)$, and $(\text{bias})^2 \approx o\big(\frac{1}{m^2}\big)$. That is, the bias term is negligible when compared to $\text{Var}_\theta(\hat{\theta})$ as the sample size increases, and hence $R_m \approx \text{Var}_\theta(\hat{\theta})$. Recall that, $\text{Var}_\theta(\hat{\theta}) \approx \frac{1}{mI(\theta)}$ for MLE, where $I(\theta)$ is the Fisher information matrix.

Remark 9. We can also show that for any estimator θ' and MLE estimator $\hat{\theta}$, we have $R(\theta,\theta') \geqslant R(\theta,\hat{\theta})$. That is MLE is approximately minimax estimator. In general, under some regularity conditions, the parametric estimation problem, the mean square error (MSE) decays as $\frac{1}{m}$. These conditions are satisfied by Gaussian distribution. However, there are examples which do not satisfy these regularity conditions and we get a faster rate of decay of MSE.

Example 4.9. Let U_θ be a uniform distribution over an interval $[\theta,\theta+1]$ parametrized by θ, and consider the set of distributions $\{U_\theta = U[\theta,\theta+1] : \theta \in \mathbb{R}\}$. For this example, those regularity conditions are not satisfied. We can show that

$$\inf_{\hat{\theta}} \sup_{\theta\in\mathbb{R}} \mathbb{E}_\theta(\theta - \hat{\theta})^2 \geqslant \frac{1 - \frac{1}{\sqrt{2}}}{128m^2}.$$

This rate is achieved by $\hat{\theta}_m \triangleq \min\{X_1,\dots,X_m\}$.

4.2 Tightening the lower bound using family of distributions

For $M = 2$, the lower bound was obtained by taking two points $\{\theta_0,\theta_1\} \in \theta(\mathcal{P})$ such that $\rho(\theta_0,\theta_1) \geqslant 2\delta$. Instead of taking two points, we take two classes of $\theta(\mathcal{P})$ which are 2δ separated. That is, let $\Theta_0,\Theta_1 \subseteq \theta(\mathcal{P})$ where

$$\inf_{(\theta_0,\theta_1)\in\Theta_0\times\Theta_1} \rho(\theta_0,\theta_1) \geqslant 2\delta.$$

We denote the set of distributions $\mathcal{P}_0 \triangleq \{P \in \mathcal{P} : \theta(P) \in \Theta_0\}$ and $\mathcal{P}_1 \triangleq \{P \in \mathcal{P} : \theta(P) \in \Theta_1\}$, and convex hull of a set of distributions $\mathcal{P}$ by $\text{conv}(\mathcal{P})$. Recall that $\text{conv}(\mathcal{P})$ is the smallest convex set containing $\mathcal{P}$.

Lemma 4.10. *For $\mathcal{P}_0,\mathcal{P}_1 \subseteq \mathcal{P}$ such that $\rho(\theta_0,\theta_1) \geqslant 2\delta$ for all $\theta_0 \in \theta(\mathcal{P}_0)$ and $\theta_1 \in \theta(\mathcal{P}_1)$. Then,*

$$\inf_{\hat{\theta}} \sup_{P\in\mathcal{P}} \mathbb{E}_P \rho(\hat{\theta},\theta(P)) \geqslant \frac{\delta}{2} \sup_{P_0\in\text{conv}(\mathcal{P}_0),P_1\in\text{conv}(\mathcal{P}_1)} (1 - \|P_0 - P_1\|_{\text{TV}}).$$

Remark 10. This gives tighter lower bound and improves over the constant $\frac{1}{24}$ in the lower bound $\frac{\sigma^2}{24m}$ for mean estimation of Gaussian random variables with unknown mean and known variance σ^2.

Example 4.11. Consider a sample $S \in (\mathcal{X} \times \mathcal{Y})^m$ where the input space $\mathcal{X} = [0,1]$, random sequence $X : \Omega \to [0,1]^m$ is *i.i.d.* uniform and $Y_i \triangleq c(X_i) + \epsilon_i$ for an unknown concept $c : \mathcal{X} \times \mathcal{Y}$ and *i.i.d.* zero mean unit variance Gaussian random sequence $\epsilon : \Omega \to \mathbb{R}^{\mathbb{N}}$ independent of X. We are interested in estimating $c(0)$, where the concept is from the family of maps $\mathcal{H} \triangleq \left\{h \in \mathcal{Y}^{\mathcal{X}} : |h(x) - h(x')| \leqslant L\,|x - x'|, x, x' \in \mathcal{X}\right\}$ is L-Lipschitz. For given hypothesis h, density of labeled example (x,y) is $p_h = p(x)p(y \mid x)$ where $p(y \mid x)$ is a Gaussian random variable with mean $h(x)$ and unit variance. We observe that $\mathcal{P} = \{P_h : h \in \mathcal{H}\}$.

We fix $\delta > 0$ and lower bound the minimax risk by taking two points $P_0, P_1 \in \mathcal{P}$ such that $D(P_0\|P_1) \geqslant 2\delta = \frac{\log 2}{m}$. Recall that family of distributions $\mathcal{P}$ is generated by hypothesis set $\mathcal{H}$, and we choose two hypotheses $\{h_0, h_1\} \in \mathcal{H}$ such that $h_0 = 0$ and $h_1(x) = L(\epsilon - x)\mathbb{1}_{[0,\epsilon]}(x)$ for all $x \in \mathcal{X}$ and a fixed $\epsilon \geqslant 0$. It is easy to check that $h_0, h_1 \in \mathcal{H}$ since they are both L-Lipschitz. We can show that $D(P_0\|P_1) = \frac{L^2\epsilon^3}{6}$ and thus taking $\epsilon \triangleq \left(\frac{6\log 2}{L^2 m}\right)^{\frac{1}{3}}$, we obtain $D(P_0\|P_1) = \frac{\log 2}{m}$. The lower bound on minimax risk is given by

$$R_m(\Phi \circ \rho) \geqslant \frac{\Phi(\delta)}{2}(1 - \sqrt{\frac{1}{2}D(P_0\|P_1)}) \geqslant \frac{\Phi(\delta)}{2}(1 - \sqrt{\frac{1}{2}\frac{1}{m}\log 2}) \geqslant \frac{\Phi(\delta)}{2}.$$