CS861: Theoretical Foundations of Machine Learning Lecture 8 - 09/22/2023
University of Wisconsin–Madison, Fall 2023

Lecture 08: Introduction to Minimax Optimality

Lecturer: Kirthevasan Kandasamy *Scribed by: Xindi Lin, Zhihao Zhao*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the instructor.*

In this lecture, we introduce Minimax Optimality and some related concepts. We will first introduce minimax optimality, then discuss some topics beyond point estimation.

1 Minimax Optimality

Recall that we cannot find an estimator that is uniformly better than other estimators in most cases, i.e.there is no $\widehat{\theta}$ better than any other $\widehat{\theta}'$ in the sense that

$$R(P, \widehat{\theta}) \leq R(P, \widehat{\theta}'), \qquad \forall P \in \mathcal{P},$$

Therefore, we wish to find an estimator which minimizes the maximum risk,

$$\sup_{P \in \mathcal{P}} R(P, \widehat{\theta}).$$

Definition 1. *The minimax risk R^* of a point estimation problem is defined as follows,*

$$R^* = \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} R(P, \hat{\theta}) = \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_{S \sim P} \left[\ell(\theta(P), \widehat{\theta}(S)) \right]$$

Note that R^* depends on $\mathcal{P}$, loss function ℓ, and the size of S. An estimator $\widehat{\theta}^*$ which achieves the minimax risk, i.e. $\sup_{P \in \mathcal{P}} R(P, \widehat{\theta}^*) = R^*$ is said to be ***minimax-optimal****.*

How do you compute the minimax risk? Classically, this was done via a concept called the "least favorable prior", which involved finding a Bayes' estimator with constant (frequentist) risk.

In this class, we will follow the following recipe for computing the minimax risk, which we will also apply to general estimation problems.

1. Design a "good estimator" $\hat{\theta}$, and upper bound its risk by U_n, then

   $$R^* \leq \sup_{P \in \mathcal{P}} R(P, \widehat{\theta}) \leq U_n.$$

2. Design a "good prior" Λ on $\mathcal{P}$ and lower bound the Bayes' risk, say by L_n.

   Recall that for any estimator $\widehat{\theta}$,

   $$\sup_{P \in \mathcal{P}} R(P, \widehat{\theta}) \geq \mathbb{E}_{P \sim \Lambda} \left[R(P, \widehat{\theta}) \right] \geq \mathbb{E}_{P \sim \Lambda} \left[R(P, \widehat{\theta}_\Lambda) \right] \geq L_n.$$

   where $\widehat{\theta}_\Lambda$ is the Bayes' estimator. the first inequality holds since "sup $\geq$ average" and the second inequality holds since the Bayes' estimator minimizes the Bayes' risk. By taking the infimum over all estimators, we have $R^* \geq L_n$.

3. Sometimes, we may need to restrict to a sub-class $\mathcal{P}^{'} \subset \mathcal{P}$, and construct our prior Λ on $\mathcal{P}'$. We have,

$$\sup_{P \in \mathcal{P}} R(P, \widehat{\theta}) \geq \sup_{P \in \mathcal{P}'} R(P, \widehat{\theta}).$$

4. If $U_n = L_n$, then U_n is the **minimax risk** and $\widehat{\theta}$ is **minimax-optimal**.
5. It is not always possible to achieve exact equality. However, if $U_n > L_n$, but $U_n \in O(L_n)$, then U_n is the **minimax rate** and $\widehat{\theta}$ is **rate-optimal**.

Example 1. Let $S = \{X_1, .., X_n\}$ drawn i.i.d. from $\mathcal{N}(\mu, \sigma^2)$, $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2); \mu \in \mathbb{R}\}$, we will show that $\widehat{\theta}_{SM}(S) = \frac{1}{n} \sum_{i=1}^n X_i$ is minimax-optimal.
Proof First, we find the upper bound

$$\sup_{P \in \mathcal{P}} R(P, \widehat{\theta}) \overset{def}{=} \sup_{\mu \in \mathbb{R}} \mathbb{E}_{S \sim \mathcal{N}(\mu, \sigma^2)} \left[(\mu - \widehat{\theta}(S))^2 \right] = \sup_{\mu \in \mathbb{R}} \frac{\sigma^2}{n} = \frac{\sigma^2}{n} \quad \Longrightarrow \quad R^* \leq \frac{\sigma^2}{n}.$$

Then we find the lower bound via Bayes' risk. Consider the prior $\Lambda = \mathcal{N}(0, \tau^2)$, from our example in the last lecture,

$$R^* \geq L_n = \left(\frac{1}{\tau^2} + \frac{1}{\sigma^2/n} \right)^{-1} \quad \text{holds for every } \tau^2,$$
$$\Longrightarrow R^* \geq \sup_{\tau^2 > 0} \left(\frac{1}{\tau^2} + \frac{1}{\sigma^2/n} \right)^{-1} = \frac{\sigma^2}{n}.$$

Combining two bounds together, we can conclude that $\widehat{\theta}_{SM}$ is minimax-optimal and $\frac{\sigma^2}{n}$ is the minimax risk. □

Example 2. Let $S = \{X_1, .., X_n\}$ drawn i.i.d. from *Bernoulli*(θ), where $\theta = \mathbb{E}_{X \sim P}[X]$. Let $\mathcal{P} = \{$*Bernoulli*$(\theta); \theta \in [0,1]\}$. Let us consider $\hat{\theta}(S) = \frac{1}{n} \sum_{i=1}^n X_i$.

First, the upper bound is found as follows,

$$R(P, \hat{\theta}) = \mathbb{E}_{S \sim P} \left[(\theta - \widehat{\theta}(S))^2 \right] = \frac{\text{Variance}}{n} = \frac{\theta(1-\theta)}{n},$$

$$\sup_{P \in \mathcal{P}} R(P, \hat{\theta}) = \sup_{\theta \in [0,1]} \frac{\theta(1-\theta)}{n} = \frac{1}{4n} \quad \Longrightarrow \quad R^* \leq \frac{1}{4n}.$$

To find the lower bound, we use $\Lambda = \text{Beta}(a, b)$ as the prior, then we have the following Bayes' risk,

$$\frac{1}{(n+a+b)^2} \left[\left(n - (a+b)^2 \right) \mathbb{E}_\theta \left[\theta^2 \right] + (n - 2a(a+b)) \mathbb{E}_\theta \left[\theta \right] + a^2 \right].$$

By choosing $a = b = \frac{\sqrt{n}}{2}$, we get

$$L_n = \frac{a^2}{(n+a+b)^2} = \frac{n/4}{(n+a+b)^2} = \frac{1}{4(\sqrt{n}+1)^2} = \frac{1}{4n + 8\sqrt{n} + 4}$$

We have $U_n > L_n$, but $U_n \in O(L_n) \Longrightarrow \widehat{\theta}_{SM}$ is rate-optimal and $\frac{1}{n}$ is the minimax-rate.
As a side note, it can be shown that

$$\hat{\theta}^* = \frac{\sqrt{n}}{1+\sqrt{n}} \left(\frac{1}{n} \sum_{i=1}^n X_i \right) + \frac{1}{2} \left(\frac{1}{1+\sqrt{n}} \right)$$

is minimax-optimal and $\frac{1}{4(\sqrt{n}+1)^2}$ is the minimax risk.

Example 3. $S = \{X_1, .., X_n\}$ drawn i.i.d. from $P \in \mathcal{P}$. $\mathcal{P}$ ={ all distribution with variance bounded by B^2}. We will show that $\widehat{\theta}_{SM}(S) = \frac{1}{n}\sum_{i=1}^n X_i$ is minimax-optimal.
Proof For the upper bound,

$$\sup_{P\in\mathcal{P}} R(P, \widehat{\theta}) = \sup_{P\in\mathcal{P}} \mathbb{E}_{S\sim P}\left[(\widehat{\theta}(S) - \theta)^2\right] = \sup_{P\in\mathcal{P}} \frac{\text{variance}}{n} = \frac{B^2}{n} \quad \implies \quad R^* \leq \frac{B^2}{n}.$$

The lower bound can be found by choosing a sub-class $\mathcal{P}' = \{\mathcal{N}(\mu, B^2); \mu \in \mathbb{R}\}$,

$$R^* = \inf_{\widehat{\theta}} \sup_{P\in\mathcal{P}} R(P, \widehat{\theta}) \geq \inf_{\widehat{\theta}} \sup_{P\in\mathcal{P}'} R(P, \widehat{\theta}) = \frac{B^2}{n}.$$

Combining two bounds together, we know that $\frac{B^2}{n}$ is the minimax risk and $\hat{\theta}_{SM}$ is minimax-optimal. □

We will now study minimaxi optimality for general estimation problems. The following lessons from point estimation will be useful going forward.

1) Often, the easiest way to lower bound the maximum risk is to lower bound it via the average risk, by using the fact that the maximum is larger than the average.

2) We should be careful in how we choose a subset $\mathcal{P}$ and prior Λ, so that the lower bound is tight.

3) We still need to design a good estimator to establish the upper bound.

2 Beyond Point Estimation

We will now extend the ideas beyond point estimation. Our estimation problem will have the following components:

1. A family of distributions $\mathcal{P}$

2. A dataset S of n i.i.d points drawn from $P \in \mathcal{P}$

3. A function(parameter) $\theta : \mathcal{P} \to \Theta$. We wish to estimate $\theta(P)$ from S.

4. An estimator $\hat{\theta} = \hat{\theta}(S) \in \Theta$

5. A loss function ℓ, $\ell = \Phi \circ \rho$, satisfies the following conditions:

   - $\rho : \Theta \times \Theta \to \mathbb{R}_+$ satisfies the following properties for all $\theta_1, \theta_2, \theta_3 \in \Theta$,
     (i) $\rho(\theta_1, \theta_1) = 0$,
     (ii) $\rho(\theta_1, \theta_2) = \rho(\theta_2, \theta_1)$,
     (iii) $\rho(\theta_1, \theta_2) \leq \rho(\theta_1, \theta_3) + \rho(\theta_3, \theta_2)$.
   - $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ is non-decreasing.

   When we estimate $\theta(P)$ with $\widehat{\theta}$, the loss is $\ell(\theta(P), \widehat{\theta}) = \Phi(\rho(\theta(P), \widehat{\theta}))$.

6. The risk of an estimato r $\widehat{\theta}$ is,

$$R(P, \hat{\theta}) = \mathbb{E}_{S\sim P}\left[\Phi \circ \rho(\theta(P), \widehat{\theta}(S))\right] = \mathbb{E}_S\left[\Phi \circ \rho(\theta, \hat{\theta})\right].$$

Note that, as before we have overloaded notation so that θ denotes the parameter $\theta \in \Theta$ and the function $\theta : \mathcal{P} \to \Theta$. Similarly, $\widehat{\theta}$ denotes the estimate $\widehat{\theta} \in \Theta$ and the estimator, which maps the data to Θ.

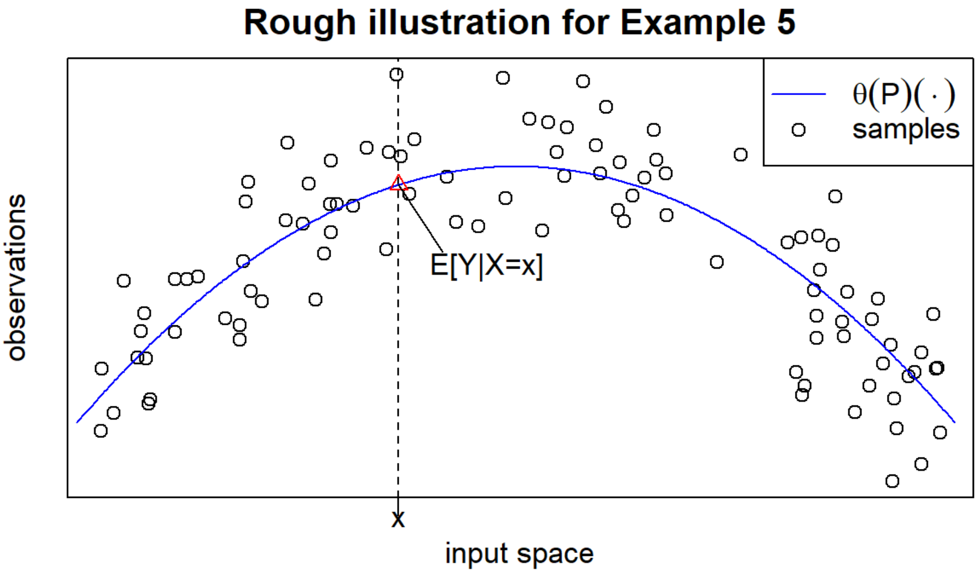

Figure 1: Figure to explain $\theta(P)(\cdot)$. Black dots are the observed data, blue curve is the regression function $\theta(P)(\cdot)$, For every x, the point of intersection, which is marked in red, is the regression result $\mathbb{E}[Y|X=x]$.

Definition 2. *We can now define the minimax risk R^* as follows,*

$$R^* = \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_S \left[\Phi \circ \rho \left(\theta(P), \widehat{\theta}(S) \right) \right].$$

Example 4. Normal mean estimation Let $S = \{X_1, .., X_n\}$ drawn i.i.d. from $\mathcal{N}(\mu, \sigma^2)$, where σ^2 is known. Here, $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2); \mu \in \mathbb{R}\}$,. We wish to estimate $\theta(P) = \mathbb{E}_{X \sim P}[X]$, therefore $\Theta = \mathbb{R}$,. If we use the squared loss $\ell(\theta_1, \theta_2) = (\theta_1 - \theta_2)^2$,, then $\rho = |\theta_1 - \theta_2|$ and $\Phi(t) = t^2$.

Example 5. (Regression) Here, $\mathcal{P}$ is the set of all distributions with support on $\mathcal{X} \times \mathbb{R}$, where is the input space. = The parameter space $\Theta = \{h : \mathcal{X} \to \mathbb{R}\}$, is the class of functions mapping $\mathcal{X}$ to . We wish to estimate the regression function, which is given by

$$\theta(P)(\cdot) = \underbrace{\mathbb{E}[Y|X = \cdot]}_{\text{regression function: see Fig 1}} = \int y dP(y|X = x)$$

We have illustrated this in Figure 1. If we use the L_2 loss, $\ell(\theta_1, \theta_2) = \int_{\mathcal{X}} (\theta_1(x) - \theta_2(x))^2 \, dx$, then, we have

$$\rho(\theta_1, \theta_2) = \sqrt{\int_{\mathcal{X}} (\theta_1(x) - \theta_2(x))^2 \, dx} \triangleq \|\theta_1 - \theta_2\|_2, \quad \text{and} \quad \Phi(t) = t^2.$$