CS861: Theoretical Foundations of Machine Learning Lecture 7 - 09/20/2023
University of Wisconsin–Madison, Fall 2023

Lecture 07: Lower Bounds for Point Estimation

Lecturer: Kirthevasan Kandasamy *Scribed by: Joseph Salzer, Tony Chang Wang*

In this lecture, we describe the general framework for point estimation and introduce the concept of average (Bayesian) risk optimality.

1 Statistical Lower Bounds

We should always ask ourselves, is our learning algorithm / estimator optimal? For example, our previous lectures have shown that the ERM process has its error decrease at a rate of around $\widetilde{O}(1/\sqrt{n})$. It is natural to ask if this rate of convergence can be improved any further. Such questions can be answered by studying statistical lower bounds as they allow us to characterize the difficulty of a learning problem.

We will start with point estimation, explained below, and then extend the ideas for regression, density estimation, and classification in subsequent lectures.

2 Point Estimation

We are interested in estimating a single parameter of a distribution (e.g mean of the distribution) using data drawn from that distribution.

A point estimation problem consists of **the following components**:

1. A family of distributions: $\mathcal{P}$.
2. A data set S: i.i.d for some distribution $P \in \mathcal{P}$.
3. A parameter of interest: $\theta = \theta(P) \in \mathbb{R}$, where P is the distribution of our data S.
4. An estimator for the parameter based on the drawn data set S: $\hat{\theta} = \hat{\theta}(S) \in \mathbb{R}$.
5. A loss function $\ell : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$, which measures how well $\hat{\theta}$ estimates the parameter θ.
6. Risk: $R(P, \hat{\theta}) = \mathbb{E}_{S\sim P}\left[\ell(\theta(P), \hat{\theta}(S)\right]$, where we take the expectation on the loss function over data S drawn from the distribution P.

$$R(P, \hat{\theta}) = \mathbb{E}_{S\sim P}\left[\ell(\theta(P), \hat{\theta}(S)\right] = \int_{S\sim P} \ell(\theta(P), \hat{\theta}(s))dP(s)$$

Example 1 (Normal Mean Estimation)**.**

1. Distribution family $\mathcal{P} = \{N(\mu, \sigma^2) : \mu \in \mathbb{R}\}$, where σ^2 is known.
2. Data $S = \{x_1, x_2, \ldots, x_n\}$ i.i.d from P, $P \in \mathcal{P}$.

3. Parameter $\theta = \theta(P) = \mathbb{E}_{X\sim P}[X]$, $P \in \mathcal{P}$
4. Loss function: $\ell(\theta_1, \theta_2) = (\theta_1 - \theta_2)^2$, for any two parameters θ_1, θ_2
5. Risk: $R_{P,\hat{\theta}} = \mathbb{E}_{\underbrace{S \sim P}_{\text{data}}}\left[\ell(P, \hat{\theta}(S))\right]$, $P \in \mathcal{P}$, $\hat{\theta}$ a parameter estimator
6. In HW0, we showed two estimators $\hat{\theta}_1$:

$$\hat{\theta}_1(S) = \frac{1}{n}\sum_{i=1}^{n} x_i$$
$$R(\theta, \hat{\theta}_1) = \mathbb{E}_{S\sim P}\left[(\theta - \theta_1\hat{(}S))^2\right] = \frac{\sigma^2}{n}$$

and $\hat{\theta}_2$

$$\hat{\theta}_2(S) = \frac{\alpha}{n}\sum_{i=1}^{n} x_i, \alpha \in [0, 1)$$
$$R(\theta, \hat{\theta}_2) = \mathbb{E}_{S\sim P}\left[(\theta - \hat{\theta}_2(S))^2\right] = \theta^2(1-\alpha)^2 + \frac{\alpha^2\sigma^2}{n}$$

(In both cases, when we take the expectation, θ is fixed, and $\hat{\theta}$ is the random variable since $\hat{\theta} = \hat{\theta}(S)$, and the expectation is w.r.t data S)

Notice that $R(\hat{\theta}_1, \theta) > R(\hat{\theta}_2, \theta)$ for some θ, while $R(\hat{\theta}_2, \theta) > R(\hat{\theta}_1, \theta)$ for some other θ, which brings difficulty to find a optimal estimator $\hat{\theta}$ that estimate well for an arbitrary parameter θ.

Extending this example, we see that the estimator $\theta = \mu$ for some $\mu \in$ will achieve 0 risk when $\theta = \mu$ but will perform poorly elsewhere. This illustrates that we cannot find a uniformly good estimator $\widehat{\theta}^{\star}$ which minimizes $R(\hat{}, \theta)$ for all θ.

Hence, it is customary to resort to other versions of optimality. The following to notions are common in the literature:

1. **Minimax Optimality.** Find an estimator $\hat{\theta}$ that minimizes the maximum risk over the class of distribution $\mathcal{P}$:

$$\min_{\hat{\theta}} \sup_{P\in\mathcal{P}} R(\theta(P), \hat{\theta})$$

2. **Average Risk Optimality.** Find an estimator $\hat{\theta}$ which minimizes the average risk over some distribution on $\mathcal{P}$. We will study this next.

3 Average Risk Optimality

Before we define what **average risk optimality** is, we have to introduce a probability measure Λ over the family of distributions $\mathcal{P}$. This measure Λ ought to be seen as an assignment of various weights to the parameter of interest $\theta := \theta(P) \in \mathbb{R}$, before the data is observed. In Bayesian terminology, Λ is our **prior** distribution and θ is a random variable.

Definition 1 (Average Risk)**.** *The average risk of a parameter is given by:*

$$\overline{R}_{\Lambda}(\theta) := \int R(P, \theta) d\Lambda(P) = \mathop{\mathbb{E}}_{P\sim\Lambda}[R(P, \theta)]$$

Now, we are equipped to define the **Bayes estimator** and **Bayes risk**:

Definition 2 (Bayes Estimator)**.** *An estimator $\hat{\theta}_\Lambda$ which minimizes the average risk is called the Bayes estimator (provided it exists).*

Definition 3 (Bayes Risk)**.** *The average risk of the Bayes estimator is called the Bayes risk. Namely,*

$$\overline{R}_\Lambda(\hat{\theta}_\Lambda) = \min_\theta \overline{R}_\Lambda(\theta)$$

It may be difficult to find the Bayes estimator from the definition of average risk alone. Instead we can apply the tower property for conditional expectation and rewrite the average risk as

$$\begin{aligned}\overline{R}_\Lambda(\theta) &= \mathop{\mathbb{E}}_{P\sim\Lambda}\left[\mathop{\mathbb{E}}_{S\sim P}[\ell(\theta(P),\hat{\theta}(S))|P]\right] \\ &= \mathop{\mathbb{E}}_{S}\left[\mathop{\mathbb{E}}_{P}[\ell(\theta(P),\hat{\theta}(S))|S]\right]\end{aligned}$$

Thus, in order to find the Bayes estimator, it suffices to find the value of $\hat{\theta}(S)$ that minimizes the conditional risk $\mathbb{E}_P[\ell(\theta(P),\hat{\theta}(S))|S]$, with the expectation taken over the posterior distribution. As an aside, we have dropped the distributions that S and P are coming from in the second equality above. This is due to an abuse of notation. The inner expectation is with respect to the posterior distribution of the parameter given the data. The outer expectation is with respect to the marginal distribution of S:

$$\mathbb{P}(S\in A) = \int P(S\in A)d\Lambda(P)$$

Lemma 1. *Under the squared error loss, $\ell(\theta_1,\theta_2) = (\theta_1-\theta_2)^2$, the Bayes estimator is the posterior mean. Namely,*

$$\hat{\theta}_\Lambda(S) = \mathbb{E}[\theta(P)|S]$$

Proof Let $\hat{\theta}(S) \coloneqq \mathbb{E}[\theta(P)|S]$ be the posterior mean. Consider any other estimator $\hat{\theta}'$ with conditional average risk:

$$\begin{aligned}\mathbb{E}[(\hat{\theta}'-\theta)^2|S] &= \mathbb{E}[(\hat{\theta}'-\hat{\theta}+\hat{\theta}-\theta)^2|S] &(1)\\ &= \underbrace{\mathbb{E}[(\hat{\theta}'-\hat{\theta})^2|S]}_{\geq 0} + \mathbb{E}[(\hat{\theta}-\theta)^2|S] + \underbrace{2\,\mathbb{E}[(\hat{\theta}'-\hat{\theta})(\hat{\theta}-\theta)|S]}_{\underbrace{2(\hat{\theta}'-\hat{\theta})\,\mathbb{E}((\hat{\theta}-\theta)|S)}_{=0 \text{ since } \mathbb{E}(\theta|S)=\hat{\theta}}} &(2)\\ &\geq \mathbb{E}[(\hat{\theta}-\theta)^2|S] &(3)\end{aligned}$$

Clearly $\hat{\theta}$ minimizes the conditional average risk (thereby minimizing the average risk) and is thus the Bayes estimator. □

We will now provide two examples on how to explicitly determine both the Bayes estimator and Bayes risk.

Example 2. Let our data be given by $S = \{X_i\}_1^n$. Now suppose $X_i|\theta \overset{\text{iid}}{\sim} N(\theta,\sigma^2)$ and $\theta\sim\Lambda \overset{\text{d}}{=} N(\mu,\tau^2)$ with σ^2, μ, and τ^2 known. Due to normal-normal conjugacy, we have $\theta|S\sim N(\tilde{\mu},\tilde{\tau}^2)$ where

$$\tilde{\mu} \coloneqq \frac{\sigma^2/n}{\tau^2+\sigma^2/n}\mu + \frac{\tau^2}{\tau^2+\sigma^2/n}\left(\frac{1}{n}\sum_{i=1}^n X_i\right)$$

$$\tilde{\tau}^2 \coloneqq \left(\frac{1}{\tau^2} + \frac{1}{\sigma^2/n}\right)^{-1}$$

Since $\tilde{\mu}$ is the posterior mean, it is the Bayes estimator. Then the Bayes risk is given by

$$\begin{aligned}\overline{R}_\Lambda(\tilde{\mu}) &= \mathop{\mathbb{E}}_{S}\underbrace{\left[\mathop{\mathbb{E}}_{\theta}[(\theta-\tilde{\mu})^2|S]\right]}_{\text{posterior variance}} \\ &= \mathop{\mathbb{E}}_{S}(\tilde{\tau}^2) \\ &= \tilde{\tau}^2\end{aligned}$$

As an aside, recall that the outer expectation is with respect to the marginal (unconditional) distribution of S. In particular, $X_i \sim N(\mu, \sigma^2+\tau^2)$. But since the posterior variance does not depend on the data (unlike in our next example), the Bayes risk is just the posterior variance.

Example 3. Again, let our data be given by $S = \{X_i\}_1^n$. Now suppose $X_i|\theta \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$ and $\theta \sim \Lambda \overset{\text{d}}{=} \text{Beta}(a,b)$ with a, b known. We know, via Bernoulli-Beta conjugacy, that the posterior distirbution is given by

$$\theta|S \sim \text{Beta}(n\bar{X}+a, b+n-n\bar{X}) \text{ where } \bar{X} \coloneqq \frac{1}{n}\sum_{i=1}^n X_i$$

Thus, the Bayes estimator can be found by taking the mean of the posterior Beta distribution. Namely,

$$\hat{\theta}_\Lambda = \mathbb{E}(\theta|S) = \frac{n\bar{X}+a}{n+a+b}$$

To find the Bayes risk, we could try to follow the same process as in the example above. However, unlike in that example, our posterior variance depends on our data through $n\bar{X}$; indeed our posterior variance is given by $\frac{(n\bar{X}+a)(b+n-n\bar{X})}{(a+b+n)^2(a+b+n+1)}$. As such, our outer expectation would be with respect to the marginal distribution of $n\bar{X}$ (unconditional on θ) which is difficult to calculate. Instead we can find the Bayes risk as

$$\begin{aligned}\overline{R}_\Lambda(\hat{\theta}_\Lambda) &= \mathop{\mathbb{E}}_{\theta\sim\text{Beta}(a,b)}\left[\mathop{\mathbb{E}}_{n\bar{X}|\theta\sim\text{Binomial}(n,\theta)}\left(\frac{n\bar{X}+a}{n+a+b}-\theta\right)^2\right] \\ &= \frac{1}{(n+a+b)^2}\mathop{\mathbb{E}}_{\theta\sim\text{Beta}(a,b)}\left[\theta^2((a+b)^2-n)+\theta(n-2a(a+b))+a^2\right]\end{aligned}$$

which can be further reduced using the first and second moments of the Beta distribution.