CS861: Theoretical Foundations of Machine Learning Lecture 7 - 09/18/2024

University of Wisconsin–Madison, Fall 2024

Lecture 07: Average Risk and Minimax Optimality

Lecturer: Kirthevasan Kandasamy *Scribed by: Jialuo Li, Jiaqi Tang*

In this lecture, we first study average (Bayesian) risk optimality and then turn to minimax optimality. We provide recipes and examples for finding these two risks.

1 Average (Bayesian) Risk Optimality

For Average risk, we will introduce Λ over probability space of θ and define,

$$\widehat{R_n}(\theta) = E_{P\sim\Lambda}[R(P,\widehat{\theta})|P]$$

$$= E_{P\sim\Lambda}[\ell(\theta(P),\widehat{\theta}(S))|P]$$

In the Baysian Paradym, Λ is socalled the prior. And we view $\theta(P)$ as a random variable P which is sampled from Λ. An estimator $\widehat{\theta}_n$ which minimizes $\widehat{R}_n$, is so called Bayes optimal estimator, if it exists. The minimum value is called the Bayes risk.

$$\bar{R}_\Lambda(\widehat{\theta}) = \mathbb{E}_S[E_P(\ell(\theta),\widehat{\theta}(S))|S]$$

If you find $\widehat{\theta}$ minimizes $\bar{R}_\Lambda(\widehat{\theta})$ for all S, then $\widehat{\theta}$ is the Bayes estimator. E_S is a conditional expectaction defined over $P(S\in A) = \int P_p(S\in A)d\Lambda(P)$ large P is the probability measure, and P is the random variable defined as above.

Proof

Let $\widehat{\theta}(S) = E(\theta(P)|S)$ then, consider any other estimator $\widehat{\theta}'$

$$\begin{aligned}
E_p([\widehat{\theta}'(S)-\theta(P)]^2|S) &= E_p[(\widehat{\theta}'-\theta)^2|S] \\
&= E_p[(\widehat{\theta}'-\theta)^2 + (\widehat{\theta}-\theta)^2 + 2(\widehat{\theta}'-\widehat{\theta})(\widehat{\theta}-\theta)|S] \\
&= \underbrace{E_p((\widehat{\theta}'-\widehat{\theta})^2|S)}_{\geq 0} + E_P[(\widehat{\theta}-\theta)^2|S] + \underbrace{\underbrace{2\,\mathbb{E}[(\widehat{\theta}'-\widehat{\theta})(\widehat{\theta}-\theta)|S]}_{2(\widehat{\theta}'-\widehat{\theta})\,\mathbb{E}((\widehat{\theta}-\theta)|S)}}_{=0 \text{ since } \mathbb{E}(\theta|S)=\widehat{\theta}} \\
&\geq E_p[(\widehat{\theta}-\theta)^2|S] = \bar{R}_\Lambda(\widehat{\theta})
\end{aligned}$$

□

Example 1. Normal mean estimation: Now suppose $X_i|\theta \overset{\text{iid}}{\sim} N(\theta,\sigma^2)$ and goal parameter $\theta\sim\Lambda \overset{\text{d}}{=} N(\mu,\tau^2)$ with σ^2, μ, and τ^2 known. Due to normal-normal conjugate, we have that $\mu|S\sim N(\tilde{\nu},\tilde{\tau}^2)$, where $\tilde{\nu} = \frac{\frac{\sigma^2}{n}}{\tau^2+\frac{\sigma^2}{n}}\nu + \frac{\frac{\sigma^2}{n}}{\tau^2+\frac{\sigma^2}{n}}(\frac{1}{n}\sum_{i=1}^n X_i)$, $\tilde{\tau} = (\frac{1}{\tau^2}+\frac{n}{\sigma^2})^{-1}$.Therefore, the Bayes estimator for normal mean is $\widehat{\mu}(S) = E_p[\mu|S] = \tilde{\nu}$ which is the posterior mean

Then, calculate Bayes' risk

$$\begin{aligned}
\bar{R_\Lambda}(\widehat{\mu}_n) &= E_P[E_S\left(\widehat{\mu}-\mu\right)^2|S] \\
&= E_S[\underbrace{E_P\left(\widehat{\mu}-\mu\right)^2|S]}_{\text{Posterior Variance}} \\
&= E_s[\tilde{\tau}^2] \\
&= \tilde{\tau}^2
\end{aligned}$$

In most case, Bayes risk is equal to the expectation of posterior variance. But in this case, posterior distribution is already known in the assumption. That is $X_i \sim N(\mu, \tau^2+\sigma^2)$. It does not depend on data. Thus the Bayes risk equals to the posterior variance.

Example 2. Again, let our data be given by $S = \{X_i\}$. $i = [1, 2..., n]$ 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. By Bernoulli-Beta conjugacy, the posterior distirbution is given by

$$\theta|S \sim \text{Beta}\left(\sum_{i=1}^n X_i + a, b + n - \sum_{i=1}^n X_i\right)$$

By Lemma 1, The optimal Bayes risk is just the posterior variance.

$$\begin{aligned}
(\widehat{R}_n|\widehat{\theta}_n) &= \underset{\theta\sim\text{Beta}(a,b)}{\mathbb{E}}\left[\underset{\sum_{i=1}^n X_i|\theta\sim\text{Binomial}(n,\theta)}{\mathbb{E}}\left(\frac{\sum_{i=1}^n X_i + a}{n+a+b} - \theta\right)^2\right] \\
&= \frac{1}{(n+a+b)^2}\underset{\theta\sim\text{Beta}(a,b)}{\mathbb{E}}\left[\theta^2((a+b)^2 - n) + \theta(n - 2a(a+b)) + a^2\right]
\end{aligned}$$

2 Minimax Optimality

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,*

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

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 a ***minimax-optimal*** *estimator.*

How do you compute the minimax risk? Classically, this was done via a concept called the "least favorable prior", which involves finding a Bayes' estimator with constant frequentist risk. In this class, we will instead use the following recipe:

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

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

2. Design a prior Λ with $\text{supp}(\Lambda) \subseteq \mathcal{P}$ and lower bound the Bayes' risk by L_n. This is a lower bound for R^*, because for any estimator $\widehat{\theta}$,

$$\sup_{P\in\mathcal{P}} R(P, \widehat{\theta}) \underbrace{\geq}_{\max\leq\text{average}} \mathbb{E}_{P\sim\Lambda}\left[R(P, \widehat{\theta})\right] \underbrace{\geq}_{\text{Bayes' estimator}} \mathbb{E}_{P\sim\Lambda}\left[R(P, \widehat{\theta}_\Lambda)\right] \geq L_n.$$

where $\widehat{\theta}_\Lambda$ is the Bayes' estimator. By taking the infimum over all estimators, we have $R^* = \inf_{\widehat{\theta}} \sup_{P\in\mathcal{P}} R(P, \widehat{\theta}) \geq L_n$.

3. If $U_n = L_n$, then L_n is the **minimax risk** and $\widehat{\theta}$ is **minimax-optimal**.
4. If $U_n \in O(L_n)$, then L_n is the **minimax rate** and $\widehat{\theta}$ is **rate-optimal** (sometimes simply minimax optimal).

Example 3. (Normal Mean Estimation) 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 $\widehat{\mu}(S) = \frac{1}{n}\sum_{i=1}^n X_i$ is minimax-optimal estimator of μ.
First, we find the upper bound

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

Then we find the lower bound via Bayes' risk. Consider the Bayes' risk under the prior $\Lambda = \mathcal{N}(0, \tau^2)$. From the previous example,

$$L_n = \widetilde{\tau}^2 = \left(\frac{1}{\tau^2} + \frac{1}{\sigma^2/n}\right)^{-1}$$

Therefore, by our recipe, $R^* \geq \left(\frac{1}{\tau^2} + \frac{1}{\sigma^2/n}\right)^{-1}$. Since it holds true for all τ, we get $R^* \geq \frac{\sigma^2}{n}$ by taking the supremum over τ.

Combining two bounds together, we can conclude that $\widehat{\mu}(S)$ is minimax-optimal and $\frac{\sigma^2}{n}$ is the minimax risk.

Example 4. $S = \{X_1, .., X_n\}$ drawn i.i.d. from $P \in \mathcal{P}$. $\mathcal{P} =\{$ all distribution with variance at most $\sigma^2\}$. We will show that $\widehat{\mu}(S) = \frac{1}{n}\sum_{i=1}^n X_i$ is minimax-optimal estimator of $\mathbb{E}[X]$.
For the upper bound,

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

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

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

Combining two bounds together, we know that $\frac{\sigma^2}{n}$ is the minimax risk and $\widehat{\mu}$ is minimax-optimal.

Example 5. (Bernoulli Mean Estimation) Let $\mathcal{P} = \{\text{Bernoulli}(\mu); \mu \in [0,1]\}$. Let $S = \{X_1, .., X_n\}$ drawn i.i.d. from $p \in \mathcal{P}$. Let us consider the sample mean $\widehat{\mu}(S) = \frac{1}{n}\sum_{i=1}^n X_i$.

First, the upper bound is found as follows,

$$U_n = \sup_{P \in \mathcal{P}} R(P, \widehat{\mu}) = \sup_{\mu \in [0,1]} \mathbb{E}_{S \sim \text{Bern}(\mu)}[(\mu - \widehat{\mu}(S))^2] = \sup_{\mu \in [0,1]} \frac{\mu(1-\mu)}{n} = \frac{1}{4n}$$

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

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

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

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

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

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

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