ECE598: Information-theoretic methods in high-dimensional statistics Spring 2016

Lecture 2: Minimax risk and Bayes risk

Lecturer: Yihong Wu *Scribe: Pan Li, Jan 26, 2016 [Ed. Feb 10]*

Recall from last lecture:

Model: A set of probability distributions $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, where θ is the parameter (finite or infinite dimensional) that specifies the distribution.

The estimation problem: Nature chooses θ and generate data X from the distribution P_θ. Upon observing X, the statistician estimates a functional $T(\theta)$ of θ, by $\hat{T}$. In this lecture, for simplicity, we focus on estimating θ itself and thus $T(\theta) = \theta$. In the following parts, we consider deterministic estimator of θ denoted by $\hat{\theta}(X)$ as well as randomized estimator given by the transition kernel $P_{\theta|X}$. Equivalently, we can write $\hat{\theta} = \hat{\theta}(X, U)$, where U is a random variable that is independent from X. For all practical purposes (e.g., X takes value in a standard Borel space), we can choose U to be uniform on $[0, 1]$. (Why?)

Risk: $R_\theta(\hat{\theta}) = \mathbb{E}_\theta \ell(\theta, \hat{\theta})$, which quantifies the quality of the estimator $\hat{\theta}$ at θ.

Remark 2.1 (Convex loss $\implies$ deterministic estimator)**.** If $\hat{\theta} \mapsto \ell(\theta, \hat{\theta})$ is convex, then randomization does not help. The proof of this claim is just based on the Jensen's inequality: for any randomized estimator $\hat{\theta}$, we have

$$R_\theta(\hat{\theta}) = \mathbb{E}\ell(\theta, \hat{\theta}) \geq \mathbb{E}\ell(\theta, \mathbb{E}[\hat{\theta}|X]),$$

where $\mathbb{E}[\hat{\theta}|X]$ is a deterministic estimator.

2.1 Bayes risk

The Bayes approach is an average-case analysis by considering the average risk of an estimator over all $\theta \in \Theta$. Concretely, we set a probability distribution (prior) π on Θ. Then, the **average risk** (w.r.t π) is defined as

$$R_\pi(\hat{\theta}) = \mathbb{E}_{\theta\sim\pi} R_\theta(\hat{\theta}) = \mathbb{E}_{\theta,X} \ell(\theta, \hat{\theta}).$$

The **Bayes risk** for a prior π is the minimum that the average risk can achieve, i.e.

$$R^*_\pi = \inf_{\hat{\theta}} R_\pi(\hat{\theta}).$$

Example 2.1 (Quadratic loss and MMSE)**.** Let $\theta, \hat{\theta} \in \mathbb{R}$, $\theta \sim \pi$. Consider quadratic loss $\ell(\theta, \hat{\theta}) = \|\hat{\theta} - \theta\|_2^2$, then the Bayes risk is the minimum mean-square error (MMSE)

$$R^*_\pi = \mathbb{E}\|\theta - \mathbb{E}[\theta|X]\|_2^2,$$

where the Bayes estimator is the conditional mean $\hat{\theta}(X) = \mathbb{E}[\theta|X]$.

Example 2.2 (Gaussian Location Model)**.** $X = \theta + Z, Z \sim \mathcal{N}(0,1), \theta \in \mathbb{R}$. Consider the Gaussian prior distribution: $\theta \sim \pi = \mathcal{N}(0, \sigma^2)$. Then $\mathbb{E}[\theta|X] = \frac{\sigma^2}{1+\sigma^2} X$ and

$$R_\pi^* = \frac{\sigma^2}{\sigma^2 + 1}. \tag{2.1}$$

Similarly, for multivariate GLM: $X = \theta + Z, Z \sim \mathcal{N}(0, \mathbf{I}_p)$, if $\theta \sim \pi = \mathcal{N}(0, \sigma^2 \mathbf{I}_p)$, then we have

$$R_\pi^* = \frac{\sigma^2}{\sigma^2 + 1} p. \tag{2.2}$$

If $R_\pi^* = \inf_{\hat{\theta}} R_\pi(\hat{\theta})$ is attained by $\hat{\theta}$, $\hat{\theta}$ is called *Bayes estimator*. Bayes estimator is always deterministic – this fact holds for any loss function. To see this, note that for any randomized estimator $\hat{\theta} = \hat{\theta}(X, U)$, its risk is lower bounded by

$$R_\pi(\hat{\theta}) = \mathbb{E}_{\theta,X,U} \ell(\theta, \hat{\theta}(X,U)) = \mathbb{E}_U R_\pi(\hat{\theta}(\cdot, U)) \geq \inf_u R_\pi(\hat{\theta}(\cdot, u))$$

where for any u, $\hat{\theta}(\cdot, u)$ is a deterministic estimator.

An alternative way to appreciate this is the following: Note that for any randomized estimator understood as a Markov kernel $P_{\hat{\theta}|X}$, the average risk $R_\pi(\hat{\theta})$ is an affine functional of $P_{\hat{\theta}|X}$. Maximizing a convex (e.g., affine) function over a convex constraint set is always achieved at the extremal points. In this case the extremal points of Markov kernels are simply delta measures, which corresponds to deterministic estimators.

The usual critisim to the Bayes approach is which prior to pick. A framework related to this but not discussed in this case is the empirical Bayes approach, where one "estimates" the prior from the data instead of choosing a prior a priori. Instead, we take a frequentist viewpoint by considering the worst-case situation:

2.2 Minimax risk

We have the risk of $\hat{\theta}$ at a given point θ : $R_\theta(\hat{\theta})$. The **minimax risk** is defined as

$$R^* = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} R_\theta(\hat{\theta}). \tag{2.3}$$

If there exists $\hat{\theta}$ s.t. $\sup_{\theta \in \Theta} R_\theta(\hat{\theta}) = R^*$, then the estimator $\hat{\theta}$ is minimax (minimax optimal).

Finding the value of the minimax risk R^* entails

$$\text{Minimax upper bound:} \qquad \exists \hat{\theta}, \forall \theta, R_\theta(\hat{\theta}) \leq r \Leftrightarrow R^* \leq r \tag{2.4}$$

$$\text{Minimax lower bound:} \qquad \forall \hat{\theta}, \exists \theta, R_\theta(\hat{\theta}) \geq r \Leftrightarrow R^* \geq r \tag{2.5}$$

This task is frequently difficult especially in high dimensions. Instead of the exact minimax risk, it is often useful to find a constant-factor approximation, which we call **minimax rate**

$$R^* \asymp \psi, \tag{2.6}$$

that is, $c\psi \le R^* \le C\psi$ for some universal constants $c, C \ge 0$. Establishing ψ is a minimax rate still entails upper and lower bounds (2.4) and (2.5), albeit within multiplicative constant factors.

In practice, minimax lower bounds are rarely established via the obvious recipe (2.5). Throughout this course, all lower bound techniques essentially boil down to lower bounding the minimax risk by Bayes risk with a sagaciously chosen prior.

Theorem 2.1 (Minimax risk $\ge$ worst-case Bayes risk)**.**

$$R^* \ge R_B^* \triangleq \sup_\pi R_\pi^*.$$

Proof. Two (equivalent) ways to understand this fact:

1. "max $\ge$ mean": $\forall \hat{\theta}$, $R_\pi(\hat{\theta}) = \mathbb{E}_{\theta\sim\pi} R_\theta(\hat{\theta}) \le \sup_{\theta\in\Theta} R_\theta(\hat{\theta})$;
2. "min max $\ge$ max min":

$$R^* = \inf_{\hat{\theta}} \sup_{\theta\in\Theta} R_\theta(\hat{\theta}) = \inf_{\hat{\theta}} \sup_{\pi\in\mathcal{M}(\Theta)} R_\pi(\hat{\theta}) \ge \sup_{\pi\in\mathcal{M}(\Theta)} \inf_{\hat{\theta}} R_\pi(\hat{\theta}) = \sup_\pi R_\pi^*,$$

where $\mathcal{M}(\Theta)$ is the set of all probability distributions on Θ. □

Example 2.3 (Minimax $>$ worst-case Bayes)**.** Let $\theta, \hat{\theta} \in \mathbb{N} \triangleq \{1, 2, ...\}$ and $\ell(\theta, \hat{\theta}) = \mathbf{1}\{\hat{\theta} < \theta\}$, i.e., the statistician loses one dollar if the nature's choice exceeds the statistician's guess and loses nothing if otherwise. Consider the extreme case of blind guessing (i.e., no data is available, say, $X = 0$). Then $\forall \hat{\theta}$, we have $R_\theta(\hat{\theta}) = \mathbb{P}(\hat{\theta} < \theta)$. Furthermore, we have $R^* \ge \lim_{\theta\to\infty} \mathbb{P}(\hat{\theta} < \theta) = 1$, which is clearly achievable. On the other hand, for any prior π on $\mathbb{N}$, $R_\pi(\hat{\theta}) = \mathbb{P}(\hat{\theta} < \theta)$ and we let $\hat{\theta} \to \infty$. Therefore, we have $R_\pi^* = 0$. Therefore in this case

$$R^* = 1 > R_B^* = 0.$$

Example 2.4 (Gaussian Linear Model)**.** This experiment is given by:

$$X \sim \mathcal{N}(\theta, 1), \quad \ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2, \quad \theta, \hat{\theta} \in \mathbb{R}$$

To get a minimax upper bound, we choose $\hat{\theta} = X$ and thus $R_\theta(\hat{\theta}) = 1$. Therefore, $R^* \le 1$. To get a minimax lower bound, we set a prior distribution for θ, i.e., $\pi \sim \mathcal{N}(0, \sigma^2)$. Using (2.1), we have $R^* \ge R_\pi^* = \frac{\sigma^2}{\sigma^2+1}$ for all $\sigma > 0$ and thus $R^* \ge \sup_\pi R_\pi^* = 1$.

p-dimensional case: $X = \theta + Z \in \mathbb{R}^p, Z \sim \mathcal{N}(0, \mathbf{I}_p), \ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^2$. Similarly, using (2.1) as a lower bound and using $\hat{\theta} = X$ for the upper bound, we have $R^* = p$.