36-705: Intermediate Statistics **Fall 2017**

Lecture 17: October 9

Lecturer: Siva Balakrishnan

In the last lecture we discussed Bayes estimators and minimax estimators which are optimal from different standpoints. Minimax estimators have lowest maximum risk, which Bayes estimators have lowest average (with respect to a distribution $\pi(\theta)$) risk.

We then discussed that Bayes estimators are often easy to compute: for instance for the squared-loss the posterior mean is the Bayes estimator. On the other hand, the minimax estimator is often difficult to compute directly.

We will study two ways in which to use Bayes estimators to find minimax estimators. One involves tightly bounding the minimax risk and the other involves identifying what is called a least favorable prior.

It is worth keeping in mind the trade-off: Bayes estimators although easy to compute are somewhat subjective (in that they depend strongly on the prior π). Minimax estimators although more challenging to compute are not subjective, but do have the drawback that they are protecting against the worst-case which might lead to pessimistic conclusions, i.e. the minimax risk might be much higher than the Bayes risk for a "nice" prior.

In 36-702, you will learn about ways to achieve a relaxed goal of computing estimators that achieve the minimax rate, i.e. estimators for which the risk goes to zero at the same rate as the minimax estimator. Formally,

$$\sup_{\theta\in\Theta} R(\theta,\widehat{\theta}) \asymp \inf_{\widetilde{\theta}} \sup_{\theta\in\Theta} R(\theta,\widetilde{\theta}) \quad n\to\infty \tag{17.1}$$

where $a_n \asymp b_n$ means that both a_n/b_n and b_n/a_n are both bounded as $n\to\infty$.

17.1 Minimax Estimators through Bayes Estimators

Our goal is to compute a minimax estimator $\widehat{\theta}$ that satisfies:

$$\sup_{\theta\in\Theta} R(\theta,\widehat{\theta}) \le \inf_{\widetilde{\theta}} \sup_{\theta\in\Theta} R(\theta,\widetilde{\theta}).$$

We will let θ_{minimax} denote a minimax estimator.

17.1.1 Bounding the Minimax Risk

One strategy to find the minimax estimator is by finding (upper and lower) bounds on the minimax risk that match. Then the estimator that achieves the upper bound is a minimax estimator.

Upper bounding the minimax risk is straightforward. Given an estimator $\widehat{\theta}_{\text{up}}$ we can compute its maximum risk and use it to upper bound the minimax risk, i.e.

$$\inf_{\widetilde{\theta}} \sup_{\theta \in \Theta} R(\theta, \widetilde{\theta}) \leq R(\theta, \widehat{\theta}_{\text{up}}).$$

We saw at the end of the last lecture that the Bayes risk of the Bayes estimator for any prior π lower bounds the minimax risk, i.e. fix a prior π and suppose that $\widehat{\theta}_{\text{low}}$ is the Bayes estimator with respect to π, then we have that:

$$B_\pi(\widehat{\theta}_{\text{low}}) \leq B_\pi(\theta_{\text{minimax}}) \leq \sup_{\theta} R(\theta, \theta_{\text{minimax}}) = \inf_{\widetilde{\theta}} \sup_{\theta \in \Theta} R(\theta, \widetilde{\theta}).$$

Let us see an example of this in action.

Example: We will prove a classical result that if we observe independent draws from a d-dimensional Gaussian, $X_1, \ldots, X_n \sim N(\theta, I_d)$, then the average:

$$\widehat{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

is a minimax estimator of θ with respect to the squared loss. I will do the entire calculation for the d-dimensional case – if you find this confusing try to first work out the case when $d = 1$.

First, let us compute the upper bound. We note that,

$$\widehat{\theta} \sim N(\theta, I_d/n),$$

so that its risk:

$$R(\theta, \widehat{\theta}) = \mathbb{E}[\sum_{i=1}^{d} (\widehat{\theta}_i - \theta_i)^2] = \mathbb{E}[\sum_{i=1}^{d} Z_i^2],$$

where $Z_i \sim N(0, 1/n)$. This yields that,

$$\inf_{\widetilde{\theta}} \sup_{\theta \in \Theta} R(\theta, \widetilde{\theta}) \leq R(\theta, \widehat{\theta}) = \frac{d}{n}.$$

Let us now try to lower bound the minimax risk using the Bayes risk. Let us take the prior to be zero-mean Gaussian, i.e. we take $\pi = N(0, c^2 I_d)$. You can convince yourself that the likelihood $p(X_1, \ldots, X_n | \theta) \propto p(\widehat{\theta} | \theta)$ (you can do this directly or appeal to sufficiency). This in turn gives us that the posterior,

$$p(\theta | X_1, \ldots, X_n) \propto p(\widehat{\theta} | \theta) \pi(\theta) \propto p(\theta | \widehat{\theta}),$$

is the same as the posterior in the following setting:

$$\begin{aligned}\theta &\sim N(0, c^2 I_d)\\ \widehat{\theta} &\sim N(\theta, I_d/n),\end{aligned}$$

so that in order to compute the posterior mean we note that,

$$\begin{pmatrix}\theta\\ \widehat{\theta}\end{pmatrix} \sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{bmatrix} c^2 I_d & c^2 I_d\\ c^2 I_d & (c^2+1/n) I_d\end{bmatrix}\right]$$

We can now compute the posterior (using standard conditional Gaussian formulae), and obtain its mean:

$$\mathbb{E}[\theta|\widehat{\theta}] = \frac{c^2}{c^2+1/n}\widehat{\theta}.$$

Now, the Bayes risk of this estimator provides us a lower bound on the minimax risk. To compute the Bayes risk we note that,

$$R(\theta, \widehat{\theta}) = \mathbb{E}_{X_1,\ldots,X_n} \| \frac{c^2}{c^2+1/n}\widehat{\theta} - \theta\|_2^2.$$

Above we noted that $\widehat{\theta} = \theta + Z$, where $Z \sim N(0, I_d/n)$, so we have

$$R(\theta, \widehat{\theta}) = \mathbb{E}_Z \| \frac{c^2}{c^2+1/n} Z - \frac{\theta}{n(c^2+1/n)}\|_2^2.$$

Let us denote $\beta := c^2 + 1/n$. Then we obtain that,

$$R(\theta, \widehat{\theta}) = \frac{\|\theta\|_2^2}{n^2\beta^2} + \frac{c^4}{\beta^2}\mathbb{E}\|Z\|_2^2 = \frac{\|\theta\|_2^2}{n^2\beta^2} + \frac{c^4}{\beta^2}\frac{d}{n}.$$

The Bayes risk further averages this over $\theta \sim N(0, c^2 I_d)$ to obtain that,

$$B_\pi\left(\frac{c^2}{c^2+1/n}\widehat{\theta}\right) = \frac{c^2 d}{n^2\beta^2} + \frac{c^4}{\beta^2}\frac{d}{n} = \frac{c^2 d}{n\beta} = \frac{d}{n(1+1/c^2)}.$$

Since c was arbitrary we can take the limit as $c \to \infty$ to obtain that the minimax risk is upper and lower bounded by d/n and conclude that the average $\widehat{\theta}$ is minimax.

17.1.2 Least Favorable Prior

The other way to obtain Bayes estimators by constructing what are called least favorable priors.

Theorem 17.1 *Let $\widehat{\theta}$ be the Bayes estimator for some prior π. If*

$$R(\theta, \widehat{\theta}) \leq B_\pi(\widehat{\theta}) \text{ for all } \theta \tag{17.2}$$

then $\widehat{\theta}$ is minimax and π is called a **least favorable prior***.*

Proof: Suppose that $\widehat{\theta}$ is not minimax. Then there is another estimator $\widehat{\theta}_0$ such that $\sup_\theta R(\theta, \widehat{\theta}_0) < \sup_\theta R(\theta, \widehat{\theta})$. Since the average of a function is always less than or equal to its maximum, we have that $B_\pi(\widehat{\theta}_0) \leq \sup_\theta R(\theta, \widehat{\theta}_0)$. Hence,

$$B_\pi(\widehat{\theta}_0) \leq \sup_\theta R(\theta, \widehat{\theta}_0) < \sup_\theta R(\theta, \widehat{\theta}) \leq B_\pi(\widehat{\theta}) \tag{17.3}$$

which is a contradiction.

Theorem 17.2 *Suppose that $\widehat{\theta}$ is the Bayes estimator with respect to some prior π. If the risk is constant then $\widehat{\theta}$ is minimax.*

Proof: The Bayes risk is $B_\pi(\widehat{\theta}) = \int R(\theta, \widehat{\theta})\pi(\theta)d\theta = c$ and hence $R(\theta, \widehat{\theta}) \leq B_\pi(\widehat{\theta})$ for all θ. Now apply the previous theorem.

Example 17.3 *Consider the Bernoulli model with squared error loss. We showed previously that the estimator*

$$\widehat{p} = \frac{\sum_{i=1}^n X_i + \sqrt{n/4}}{n + \sqrt{n}}$$

has a constant risk function. This estimator is the posterior mean, and hence the Bayes estimator, for the prior $\mathrm{Beta}(\alpha, \beta)$ *with $\alpha = \beta = \sqrt{n/4}$. Hence, by the previous theorem, this estimator is minimax.*

17.2 Asymptotic theory

We are going to spend the rest of this lecture (and likely much of the next lecture) discussing asymptotic theory for the MLE. We suppose that we obtain a sample $X_1, \ldots, X_n \sim p(X; \theta)$ and are interested in estimating θ.

Analogous to the asymptotic theory we developed for the average of i.i.d. random variables we will be interested in two questions:

1. **Consistency:** Does the MLE converge in probability to θ, i.e. does $\widehat{\theta}_{\text{MLE}} \xrightarrow{p} \theta$? This is analogous to the LLN.
2. **Asymptotic distribution:** What can we say about the distribution of $\sqrt{n}(\widehat{\theta}_{\text{MLE}} - \theta)$? This is analogous to the CLT.

We will begin with the question of consistency.

17.3 Consistency of the MLE

The main take-home from this section is that under somewhat mild conditions the MLE is a consistent estimator. We will try to develop the necessary conditions and build some intuition about the MLE and about what consistency entails.

17.3.1 MLE as Empirical Risk Minimization

We have discussed previously the idea of empirical risk minimization, where we construct an estimator by minimizing an empirical estimate of the risk. We looked at the particular case of classification with the 0/1 loss. The MLE can be viewed as a special case of ERM with a different loss function.

Suppose we define the risk function:

$$R_n(\widehat{\theta}, \theta) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{p(X_i; \theta)}{p(X_i; \widehat{\theta})},$$

then we can observe that minimizing this risk function is identical to maximizing the likelihood. Notice that we introduced an extra $p(X_i; \theta)$ term but this does not affect anything. Of course, if this is the empirical risk it is natural to wonder what the associated population risk is. This is given as:

$$R(\widehat{\theta}, \theta) = \mathbb{E}_\theta \log \frac{p(X; \theta)}{p(X; \widehat{\theta})},$$

which is known as the Kullback-Leibler divergence, i.e. the population risk is the KL divergence $\mathrm{KL}(p(X;\theta) \| p(X; \widehat{\theta}))$.

Notice that, the empirical risk is a sum of i.i.d terms so by the LLN we have that for any fixed $\widetilde{\theta}$

$$R_n(\widetilde{\theta}, \theta) \xrightarrow{p} R(\widetilde{\theta}, \theta).$$

To analyze empirical risk minimization we needed a *uniform* LLN and we will need exactly this to show consistency.

An important property of the KL divergence is that it is zero iff $p(X; \theta) = p(X; \widehat{\theta})$ almost everywhere (i.e. they are equal except on sets of measure 0).

The main thing to remember is the connection between MLE and KL divergence.

17.3.2 Conditions for consistency

Condition 1: Identifiability: A basic requirement for constructing any consistent estimator is that the model be identifiable, i.e. if $\theta_1 \neq \theta_2$ then it must be the case that $p(X;\theta_1) \neq p(X;\theta_2)$.

We will in general require something slightly stronger than this:

Condition 2: Strong identifiability: We assume that for every $\epsilon > 0$

$$\inf_{\widetilde{\theta}:|\widetilde{\theta}-\theta|\geq\epsilon} \mathrm{KL}(p(X;\theta)\|p(X;\widetilde{\theta})) > 0.$$

This condition is essentially the same as Condition 1, except that it does not allow the difference between the two distributions to be vanishingly small. The two conditions are equivalent if θ is restricted to lie in a compact set.

Condition 3: Uniform LLN: Assume that,

$$\sup_{\widetilde{\theta}} |R_n(\widetilde{\theta},\theta) - R(\widetilde{\theta},\theta)| \xrightarrow{p} 0.$$

This condition is a uniform LLN. As we have seen before it holds for instance if the Rademacher complexity of the class of functions of the form: $f_{\widetilde{\theta}}(X) = \log p(X;\widetilde{\theta})/p(X;\theta)$ is not too large. In 36-702 you will explore this idea further.

Theorem 17.4 *Suppose that Conditions 2 and 3 above hold, then the MLE is consistent.*

Proof: Fix an $\epsilon > 0$. Using the strong identifiability condition we see that for every $\epsilon > 0$, we have that there is an $\eta > 0$ such that,

$$\mathrm{KL}(p(X;\theta)\|p(X;\widetilde{\theta})) \geq \eta,$$

if $|\widetilde{\theta} - \theta| \geq \epsilon$. We will show that for the MLE $\widehat{\theta}$, we have that $\mathrm{KL}(p(X;\theta)\|p(X;\widehat{\theta})) \leq \eta$, as $n \to \infty$ in probability. This in turn implies that $|\widehat{\theta} - \theta| \leq \epsilon$ which implies that $\widehat{\theta} \xrightarrow{p} \theta$.

If remains to show that $\mathrm{KL}(p(X;\theta)\|p(X;\widehat{\theta})) \leq \eta$, as $n \to \infty$. Notice that,

$$\mathrm{KL}(p(X;\theta)\|p(X;\widehat{\theta})) = R(\widehat{\theta},\theta) = R(\widehat{\theta},\theta) - R_n(\widehat{\theta},\theta) + R_n(\widehat{\theta},\theta) \overset{\text{(i)}}{\leq} R(\widehat{\theta},\theta) - R_n(\widehat{\theta},\theta) \xrightarrow{p} 0,$$

where the final convergence simply uses Condition 3. The inequality (i) follows since,

$$R_n(\widehat{\theta},\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{p(X_i;\theta)}{p(X_i;\widehat{\theta})} \leq 0,$$

since $\widehat{\theta}$ is the MLE. ■

17.4 Inconsistency of the MLE

The MLE can fail to be consistent. When the model is not identifiable it is clear that we cannot have consistent estimators.

The other possible failure is the failure of the uniform law. This typically happens when the parameter space is too large. Here is a simple example:

Example: Suppose that we measure some outcome (say their blood sugar) for n individuals using a machine. We do it twice for every individual so that we can assess the variability of the machine, i.e. suppose we observe:

$$\begin{aligned} &Y_{11}, Y_{12} \sim N(\mu_1, \sigma^2) \\ &\vdots \\ &Y_{n1}, Y_{n2} \sim N(\mu_n, \sigma^2), \end{aligned}$$

and want to estimate σ^2. Even though we only want to estimate σ^2 the model has a growing number of parameters $\mu_1, \ldots, \mu_n, \sigma^2$ and the MLE for σ^2 will depend on estimating μ_i. Formally, we can see that the MLE for the means is:

$$\widehat{\mu}_i = \frac{Y_{i1} + Y_{i2}}{2}.$$

The log-likelihood for σ^2 can be written as:

$$\mathcal{LL}(\sigma^2, \mu) = -n \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left[(Y_{i1} - \mu_i)^2 + (Y_{i2} - \mu_i)^2 \right],$$

which is maximized when we take:

$$\widehat{\sigma}^2 = \frac{1}{2n} \sum_{i=1}^{n} \left[(Y_{i1} - \widehat{\mu}_i)^2 + (Y_{i2} - \widehat{\mu}_i)^2 \right] = \frac{1}{4n} \sum_{i=1}^{n} (Y_{i1} - Y_{i2})^2.$$

Notice that,

$$\mathbb{E}[\widehat{\sigma}^2] = \frac{\sigma^2}{2},$$

so by the LLN the MLE is inconsistent. One could easily fix this in this problem (by multiplying the MLE by 2) but more generally this could be tricky. We note that in this type of problem where the number of parameters is not fixed (and grows with the sample size) it is not even clear how to define convergence of the log-likelihood since its limit changes with the sample size.