LECTURE 6 NOTES

1. Minimax estimators. Instead of considering the average risk, another option is to consider the worst-case risk: $\sup_{\theta\in\Theta} R_\delta(\theta)$.

EXAMPLE 1.1. *Let* $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \text{Ber}(p)$*. The MLE of* p *is* $\hat{p}_{\text{ML}} = \bar{\mathbf{x}}$*, and the Bayes estimator under a* $\text{beta}(a,b)$ *prior is*

$$\hat{p}_B = \frac{n\bar{\mathbf{x}} + a}{a+b+n} \tag{1.1}$$

The worst-case MSE of the MLE is

$$\sup_{p\in[0,1]} \frac{p(1-p)}{n} = \frac{1}{4n}.$$

Recall the MSE of the Bayes estimator is

$$\text{MSE}_{\hat{p}_B}(p) = \frac{np(1-p)}{(a+b+n)^2} + \left(\frac{np+a}{a+b+n} - p\right)^2.$$

By choosing $a = b = \sqrt{\frac{n}{4}}$*, the MSE is constant in* p*:*

$$\text{MSE}_{\hat{p}_B}(p) = \frac{n}{4(n+\sqrt{n})^2}.$$

Figure 1 plots the risk of the MLE and the risk of the Bayes estimator. We see that the worst-case MSE of the Bayes estimator is smaller than that of the MLE. However, especially if n *is large, the MLE has smaller risk than the Bayes estimator except in a small neighborhood of* $p = \frac{1}{2}$*.*

DEFINITION 1.2. *The* minimax risk *of a parametric model is*

$$\inf_\delta \sup_{\theta\in\Theta} \text{Risk}_\delta(\theta).$$

The minimax risk is a property of a parametric model: it is the worst-case risk of the estimator with the smallest worst-case risk. It is interpreted as a measure of the hardness of estimating θ in the parametric model. An estimator whose worst-case risk is the minimax risk is a *minimax estimator*. That is,

$$\sup_{\theta\in\Theta} R_\delta(\theta) = \inf_\delta \sup_{\theta\in\Theta} \text{Risk}_\delta(\theta).$$

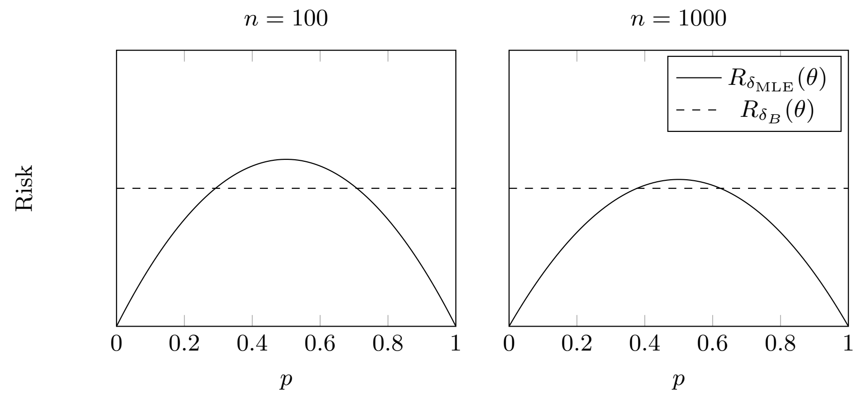

Fig 1: Risk of the MLE and Bayes estimator in Example 1.1

Finding minimax estimators is hard. Sometimes, we settle for an asymptotically minimax estimator:

$$\sup_{\theta\in\Theta} R_\delta(\theta) \sim \inf_\delta \sup_{\theta\in\Theta} R_\delta(\theta) \text{ as } n \to \infty,$$

where $a_n \sim b_n$ means $\frac{a_n}{b_n} \to 1$. Sometimes, even finding an asymptotically minimax estimator is hard, and we settle for a minimax rate optimal estimator

$$\sup_{\theta\in\Theta} R_{\delta_n}(\theta) \asymp \inf_{\delta_n} \sup_{\theta\in\Theta} R_{\delta_n}(\theta) \text{ as } n \to \infty,$$

where $a_n \asymp b_n$ means $\frac{a_n}{b_n}$ and $\frac{b_n}{a_n}$ remain bounded as n grows.

DEFINITION 1.3. *A prior π is a* least favorable prior *if*

$$\inf_\delta \mathbf{E}_\pi\big[R_\delta(\boldsymbol{\theta})\big] \ge \inf_\delta \mathbf{E}_{\pi'}\big[R_\delta(\boldsymbol{\theta})\big]$$

for any other prior π'.

THEOREM 1.4. *Let δ_B be the Bayes estimator for some prior π. If*

$$\sup_{\theta\in\Theta} \mathrm{Risk}_{\delta_B}(\theta) = \mathbf{E}_\pi\big[R_{\delta_B}(\boldsymbol{\theta})\big],$$

then δ_B is a minimax estimator, and π is a least favorable prior.

PROOF. We first show that δ_B is not minimax by contradiction. Suppose δ_B is not minimax: there is another estimator $\delta_{\wedge\vee}$ whose worse-case risk is smaller than the worse-case risk of δ_B. Since the Bayes risk is smaller than the minimax risk,

$$\mathbf{E}_\pi\big[R_{\delta_{\wedge\vee}}(\boldsymbol{\theta})\big] \le \sup_{\theta\in\Theta} \mathrm{Risk}_{\delta_{\wedge\vee}}(\theta) < \sup_{\theta\in\Theta} \mathrm{Risk}_{\delta_B}(\theta) = \mathbf{E}_\pi\big[R_{\delta_B}(\boldsymbol{\theta})\big],$$

which violates the assumption that δ_B is Bayes.

We turn our attention to showing that π is the least favorable prior. For any other prior π', we have

$$\begin{aligned}
&\inf_\delta \mathbf{E}_{\pi'}\big[R_\delta(\boldsymbol{\theta})\big] \\
&\quad\le \inf_\delta \sup_{\theta\in\Theta} \mathrm{Risk}_\delta(\theta) && \text{(Bayes risk is at most minimax risk)} \\
&\quad= \sup_{\theta\in\Theta} \mathrm{Risk}_{\delta_B}(\theta) && (\delta_B \text{ is minimax}) \\
&\quad= \mathbf{E}_\pi\big[R_{\delta_B}(\boldsymbol{\theta})\big] && \text{(by assumption)} \\
&\quad= \inf_\delta \mathbf{E}_\pi\big[R_\delta(\boldsymbol{\theta})\big], && (\delta_B \text{ is Bayes})
\end{aligned}$$

which implies π is least favorable. □

Corollary 1.5. *Let δ_B be the Bayes estimator for some prior π. If its (frequentist) risk does not depend on θ, then δ_B is minimax.*

Proof. If the risk function is a constant, then the Bayes and minimax risks are both the same constant. By Theorem 1.4, δ_B is minimax. □

Example 1.6 (Example 1.1 continued). *In Example 1.1, we showed that the MSE of the Bayes estimator does not depend on p. By Theorem 1.4, the Bayes estimator is minimax. We remark that the MLE is not minimax.*

Example 1.7. *Let $\mathbf{x} \sim \mathcal{N}(\mu, I_d)$. We show that, unsurprisingly, a minimax estimator of μ is $\hat\mu = \mathbf{x}$. Recall the Bayes estimator of μ under the prior $\mathcal{N}(0, b^2 I_d)$ is*

$$\hat\mu_B = \frac{b^2\mathbf{x}}{b^2+1} = \big(1 - \frac{1}{b^2+1}\big)\mathbf{x}.$$

It is possible to show that the Bayes risk is $\big(\frac{b^2}{b^2+1}\big)d$. Indeed,

$$\begin{aligned}
\mathbf{E}\big[\|\hat\mu_B - \boldsymbol{\mu}\|_2^2 \mid \boldsymbol{\mu}\big] &= \mathbf{E}\Big[\big\|\tfrac{b^2\mathbf{x}}{b^2+1} - \tfrac{b^2\boldsymbol{\mu}}{b^2+1}\big\|_2^2 \mid \boldsymbol{\mu}\Big] + \mathbf{E}\Big[\big\|\tfrac{b^2\boldsymbol{\mu}}{b^2+1} - \boldsymbol{\mu}\big\|_2^2 \mid \boldsymbol{\mu}\Big] \\
&= \big(\tfrac{b^2}{b^2+1}\big)^2 d + \big(\tfrac{b^2}{b^2+1} - 1\big)^2 \|\boldsymbol{\mu}\|_2 \\
&= \big(\tfrac{b^2}{b^2+1}\big)^2 d + \big(-\tfrac{1}{b^2+1}\big)^2 \|\boldsymbol{\mu}\|_2 .
\end{aligned}$$

Integrating over the prior,

$$\mathbf{E}_{\mathcal{N}(0,b^2)}\big[\|\hat\mu_B - \boldsymbol{\mu}\|_2^2\big] = \big(\tfrac{b^2}{b^2+1}\big)^2 d + \big(\tfrac{1}{b^2+1}\big)^2 b^2 d = \tfrac{b^4+b^2}{(b^2+1)^2} d$$

Recall the minimax risk is at least the Bayes risk:

$$\begin{aligned}\inf_\delta \sup_{\theta\in\Theta} R_\delta(\theta) &\geq \inf_\delta \mathbf{E}_{\mathcal{N}(0,b^2 I_d)}\big[R_\delta(\theta)\big] \\ &= \mathbf{E}_{\mathcal{N}(0,b^2 I_d)}\big[R_{\hat{\mu}_B}(\boldsymbol{\mu})\big] \\ &= \tfrac{b^4+b^2}{(b^2+1)^2} d.\end{aligned} \tag{1.2}$$

Since we have (1.2) for any b^2,

$$\inf_\delta \sup_{\theta\in\Theta} R_\delta(\theta) \geq \sup_{b^2>0} \tfrac{b^4+b^2}{(b^2+1)^2} d = d. \tag{1.3}$$

It is easy to check that the risk of the MLE $\hat{\mu}_{\text{MLE}} = \mathbf{x}$ is d for any μ:

$$\text{MSE}_{\hat{\mu}_{\text{MLE}}}(\mu) = \mathbf{E}_\mu\big[\|\mathbf{x}-\mu\|_2^2\big] = \mathbf{E}_\mu\big[(\mathbf{x}-\mu)^T(\mathbf{x}-\mu)\big],$$

which, by the trace trick, is

$$\begin{aligned}&= \mathbf{E}_\mu\big[\text{tr}\big((\mathbf{x}-\mu)(\mathbf{x}-\mu)^T\big)\big] \\ &= \text{tr}\Big(\mathbf{E}_\mu\big[\big((\mathbf{x}-\mu)(\mathbf{x}-\mu)^T\big]\Big) \\ &= \text{tr}(I_d) = d.\end{aligned}$$

Since the worst-case risk of the MLE attains (1.3), it is minimax.

To wrap up, we mention that

- minimaxity does not imply admissibility: a minimax estimator has the best worst-case performance, but its performance at other parameters may be suboptimal. However, any estimator that dominates a minimax estimator is also minimax. Thus unique minimax estimators are admissible.
- admissibility also does not imply minimaxity: an estimator is admissible if it has (strictly) smaller risk than any other estimator at a single $\theta \in \Theta$, but its risk at any other θ may be arbitrarily bad. However, an admissible estimator with constant risk is minimax.

2. Shrinkage estimators. We begin with a famous example.

EXAMPLE 2.1. *Consider estimating the mean of a (multivariate) Gaussian distribution from an observation $\mathbf{x} \sim \mathcal{N}(\mu, I_d)$. A minimax estimator is the MLE: $\hat{\mu}_{\text{ML}} = \mathbf{x}$. However, when $d \geq 3$, Stein showed that the MLE is inadmissible! Later, James and Stein showed that the estimator*

$$\hat{\mu}_{\text{JS}} := \left(1 - \frac{d-2}{\|\mathbf{x}\|_2^2}\right)\mathbf{x} \tag{2.1}$$

has smaller MSE than the MLE for any $\mu \in \mathbf{R}^d$. *The estimator is biased: it shrinks the MLE towards the origin.*

Stein's original argument of why it is possible to improve upon the MLE is simple. Intuitively, a good estimator of μ *should have roughly the same norm as* μ. *However, the MSE of the MLE is*

$$\begin{aligned}\mathbf{E}\Big[\|\mathbf{x}\|_2^2\Big] &= \mathbf{E}\Big[\|\mu + (\mathbf{x} - \mu)\|_2^2\Big] \\ &= \|\mu\|_2^2 + \mathbf{E}\Big[\|\mathbf{x} - \mu\|_2^2\Big] \\ &= \|\mu\|_2^2 + d.\end{aligned}$$

The preceding calculation suggests $\|\mathbf{x}\|_2^2$ *is likely larger than* $\|\mu\|_2^2$, *especially if* d *is large, which, in turn suggests shrinking* $\mathbf{x}$.

Yuekai Sun
Berkeley, California
November 23, 2015