STA732
Statistical Inference

Lecture 12: Minimax optimality

Yuansi Chen

Spring 2023

Duke University

https://www2.stat.duke.edu/courses/Spring23/sta732.01/

Recap from Lecture 11

1. Hierarchical Bayes is built from putting hyperprior on the prior, good for pooling multiple similar datasets
2. Empirical Bayes simply replaces the prior parameter in Bayes estimator with its empirical estimate
3. James-Stein estimator makes sample mean inadmissible
4. We should care about shrinkage, especially when dimension is large

Goal of Lecture 12

1. Calculate the risk of James-Stein estimator
2. Minimax risk, minimax estimator
3. Least favorable priors

5.1, 5.2 of Lehmann and Casella

Calculate the risk of James-Stein estimator

JS estimator better than sample mean

In the non-Bayesian Gaussian sequence model, n data points, $X \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, \sigma^2 \mathbb{I}_d), \theta \in \mathbb{R}^d$ (fixed), $\sigma^2 > 0$ (known), for $d \geq 3$, the sample mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

is inadmissible for estimating θ under squared error loss

The JS estimator

$$\delta_{\mathsf{JS}}(X) = \left(1 - \frac{(d-2)\sigma^2/n}{\left\|\bar{X}\right\|_2^2}\right) \bar{X}$$

has strictly lower risk uniformly

Stein's lemma

Useful tool for computing risk in Gaussian estimation problems.

Stein's Lemma, univariate, Lem 11.1 in Keener

Suppose $X \sim \mathcal{N}(\theta, \sigma^2)$, $h : \mathbb{R} \to \mathbb{R}$, differentiable, $\mathbb{E}\,|h'(X)| < \infty$, then

$$\mathbb{E}[(X-\theta)h(X)] = \sigma^2 \mathbb{E}[h'(X)]$$

proof idea: write down the intergrals for $\theta = 0, \sigma^2 = 1$ first

Multivariate Stein's Lemma, Thm 11.3 in Keener

Suppose $X \sim \mathcal{N}(\theta, \sigma^2 \mathbb{I}_d), \theta \in \mathbb{R}^d$, $h : \mathbb{R}^d \to \mathbb{R}^d$, differentiable, $\mathbb{E} \|Dh(X)\|_F < \infty$.

$$\begin{aligned} \mathbb{E}\left[(X-\theta)^\top h(X)\right] &= \sigma^2 \sum_{i=1}^d \mathbb{E} \frac{\partial h_i}{\partial x_i}(X) \\ &= \sigma^2 \mathbb{E}\,\mathrm{Tr}(Dh(X)) \end{aligned}$$

Stein's unbiased risk estimator (SURE)

We can use Stein's lemma to get unbiased estimator of the risk under squared error loss for any $\delta(X)$, apply with $h(X) = X - \delta(X)$.

$$\begin{aligned} R(\theta, \delta) &= \mathbb{E}_\theta \left[\|X - \theta - h(X)\|_2^2 \right] \\ &= \mathbb{E}_\theta \|X - \theta\|_2^2 + \mathbb{E}_\theta \|h(X)\|_2^2 - 2\mathbb{E}_\theta \left[(X - \theta)^\top h(X) \right] \\ &= d + \mathbb{E}_\theta \|h(X)\|_2^2 - 2\mathbb{E}_\theta \operatorname{Tr}(Dh(X)) \end{aligned}$$

We get an unbiased estimator for the risk

$$\hat{R}(X) = d + \|h(X)\|_2^2 - 2 \operatorname{Tr}(Dh(X))$$

Calculate the risk of James-Stein

proof idea: apply SURE

Why the detour to SURE?

The risk of James-Stein can be calculate directly if we are good enough with integrals with Gaussian. Why are we taking an indirect route?

- The proof via SURE is elegant and avoid the calculation of integrals
- SURE might be useful in other context where one needs a risk estimator

Minimax risk, minimax estimator

Where we are

We are at the second approach of arguing for "the best" estimator in point estimation: global measure of optimality

- We finished average risk optimality: Bayes optimal
- We begin minimax risk optimality

Minimax estimator

In minimax estimation, our global measure of risk is the worst-case risk.

Def. minimax estimator

Given $X \sim P_\theta$, where $\theta \in \Omega$, and a loss function $L(\theta, d)$, we look for an estimator δ that minimizes the worse-case risk

$$\sup_{\theta \in \Omega} R(\theta, \delta).$$

Any minimizer δ is called a minimax estimator.

Minimax risk

The minimum achievable sup-risk is called the minimax risk of the estimation problem

$$r^* = \inf_{\delta} \sup_{\theta \in \Omega} R(\theta, \delta)$$

minimax risk is just a single number!

Game theory interpretation of the minimax risk

A game between data analyst and Nature

- Data analyst chooses estimator δ
- Given the estimator δ, Nature chooses parameter θ to maximize the risk

  Nature chooses θ, not X!

Comparison to Bayes

- In minimax estimation, Nature plays adversarially
- Compared to Bayes estimation, Nature plays a specific mixed strategy: θ is drawn from a fixed prior distribution

How to upper and lower bound the minimax risk?

Upper bound

Choose any estimator δ_0, compute its worst-case risk, we have

$$r^* \leq \sup_{\theta \in \Omega} R(\theta, \delta_0)$$

Lower bound

Key observation: average-case risk $\leq$ worst-case risk

Take any Bayes estimator δ_Λ with prior Λ with Bayes risk r_Λ, we have

$$r_\Lambda = \inf_\delta \int R(\theta, \delta) d\Lambda(\theta) \leq \inf_\delta \max_{\theta \in \Omega} R(\theta, \delta) \leq r^*$$

One vague strategy to compute the minimax risk

The previous slide gives a vague strategy to compute the minimax risk

1. Find δ_0 with a small worst-case risk
2. Find prior Λ with large Bayes risk
3. Repeat step 1 and 2 until the upper and lower bounds match

Least favorable prior

Least favorable prior

Recall the Bayes risk under prior Λ is

$$r_\Lambda = \inf_\delta \int R(\theta, \delta) d\Lambda(\theta).$$

Def. least favorable prior

A prior Λ is a least favorable prior if $r_\Lambda \geq r_{\Lambda'}$ for any other prior Λ'.

The Bayes risk for the least favorable prior should give the tightest lower bound for the minimax risk

When an estimator has equal Bayes risk and worst-case risk

Thm. 5.1.4 in Lehmann and Casella

Suppose δ_Λ is Bayes for Λ satisfying

$$r_\Lambda = \sup_{\theta \in \Omega} R(\theta, \delta_\Lambda).$$

That is, the Bayes risk of δ_Λ is also the worst-case risk. Then

1. δ_Λ is minimax
2. Λ is a least favorable prior
3. If δ_Λ is the unique Bayes estimator for λ (a.s. for all P_θ), then it is the unique minimax estimator

This theorem shows that to find a minimax estimator it is sufficient to find a Bayes estimator with Bayes risk equal to its worst-case risk

Proof idea:

$$\sup_{\theta \in \Omega} R(\theta, \delta) \geq \int R(\theta, \delta) d\Lambda(\theta) \geq \int R(\theta, \delta_\Lambda) d\Lambda(\theta) = \sup_{\theta \in \Omega} R(\theta, \delta_\Lambda)$$

$$r_{\Lambda'} = \inf_{\delta} \int R(\theta, \delta) d\Lambda'(\theta) \leq \int R(\theta, \delta_\Lambda) d\Lambda'(\theta) \leq \sup_{\theta \in \Omega} R(\theta, \delta_\Lambda) = r_\Lambda$$

Sufficient conditions for checking minimax via the above theorem

Cor. 5.1.5 in Lehmann and Casella

If a Bayes estimator δ_Λ has constant risk as a function of θ (that is, $R(\theta, \delta_\Lambda) = R(\theta', \delta_\Lambda), \forall \theta, \theta'$), then δ_Λ is minimax

Cor. 5.1.6 in Lehmann and Casella

Given a Bayes estimator δ_Λ, define

$$w_\Lambda = \left\{ \theta : R(\theta, \delta_\Lambda) = \sup_{\theta'} R(\theta', \delta_\Lambda) \right\}.$$

If $\Lambda(w_\Lambda) = 1$, then δ_Λ is minimax.

Draw a picture

Example: Binomial

Suppose $X \sim$ Binomial(n, θ) for some $\theta \in (0, 1)$. We use squared error loss.

- Is $\frac{X}{n}$ minimax?
- If not, find a minimax estimator

  Hint: think about the Bayes estimators under Beta prior

Least favorable prior sequence

Motivation

From last section, if we

- find a least favorable prior
- show that its Bayes risk equals to the worst-case risk

then we find the minimax estimator which is the corresponding Bayes estimator.

It turns out that minimax estimators may not be Bayes!

Sometimes the least favorable prior is not a proper prior

Motivating example: minimax for normal mean estimation

Suppose $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, \sigma^2)$ with σ^2 known. We use squared error loss.

- Compute the risk of $\bar{X}$
- Is $\bar{X}$ Bayes with some prior? $\bar{X}$ is not Bayes but a limit of Bayes estimators
- Is $\bar{X}$ minimax?

Least favorable prior sequence

Def. least favorable prior sequence

Let $\{\Lambda_m\}$ be a sequence of priors with minimal average risks $\left\{r_{\Lambda_m}\right\}$ where $r_{\Lambda_m} = \inf_\delta \int R(\theta, \delta) d\Lambda_m(\theta)$. Then, $\{\Lambda_m\}$ is a least favorable sequence of priors if there is a real number r such that $r_{\Lambda_m} \to r < \infty$ and $r \geq r_{\Lambda'}$ for any prior Λ'

Remark: less restrictive than the definition of a least-favorable prior. Useful when the space of achievable risk is not compact

Thm. 5.1.12 in Lehmann and Casella

Suppose there is a real number r such that $\{\Lambda_m\}$ is a sequence of priors with $r_{\Lambda_m} \to r < \infty$. Let δ be any estimator such that $\sup_{\theta \in \Omega} R(\theta, \delta) = r$. Then

1. δ is minimax
2. $\{\Lambda_m\}$ is a least-favorable prior sequence

Proof of Thm. 5.1.12:

$$\sup_\theta R(\theta, \delta') \geq \int R(\theta, \delta') d\Lambda_m(\theta) \geq r_{\Lambda_m}$$

$$r_{\Lambda'} = \int R(\theta, \delta_{\Lambda'}) d\Lambda'(\theta) \leq \int R(\theta, \delta) d\Lambda'(\theta) \leq \sup_\theta R(\theta, \delta) = r$$

Back to our example of normal mean estimation

Use Thm. 5.1.12 to show that $\bar{X}$ is minimax

Summary

- Minimax risk $r^* = \min_\delta \max_{\theta\in\Omega} R(\theta, \delta)$
- To get an upper bound of the minimax risk, consider any estimator δ_0

$$r^* \le \sup_{\theta\in\Omega} R(\theta, \delta_0)$$

- To get a lower bound, consider any prior Λ

$$r^* \ge \int R(\theta, \delta_\Lambda) d\Lambda(\theta)$$

- Find least favorable prior/ least favorable prior sequence, can help us prove one estimator is minimax

Other comments on minimax

- Minimax estimators depend on the loss, depend on Ω
- Minimax estimators may be hard to find
- However, minimax lower bounds are often used in Stat theory to characterize hardness
- Critique on minimax optimality: A problem might be easy throughout most of parameters space but very hard in some bizzare corner you rarely encounter in practice.

What is next?

- Identification of minimax estimators via submodels
- Is minimax estimator admissible?

Thank you