6.S951 Modern Mathematical Statistics Fall 2024

Lecture 3 — September 12, 2024

Prof. Stephen Bates

Scribe: Andrew Yao

1 Outline

Agenda:

1. Minimax continued
2. Admissibility
3. Gaussian linear model
    - Bayes and minimax (and sufficiency!)

Last time:

1. Statistical decision theory framework
2. Sufficiency
3. Bayes-optimal estimators
4. Minimax optimality
    - Hardness lower bound via Bayes
    - Bare-hands upper bound (come up with estimator that hits lower bound)

So far the goal has been to develop formal language to discuss statistical problems.

2 Minimax continued

Minimax risk is always bigger than Bayes risk.

Corollary 1 (Bayes with constant risk over Θ is minimax)**.** *Let A^* be Bayes optimal with respect to Q. If $R(A^*;\theta)$ is constant in θ then A^* is minimax optimal.*

Proof. $R_M(A^*) = \sup_{\theta\in\Theta} R(A^*;\theta) = \int_{\theta\in\Theta} R(A^*;\theta)dQ(\theta) = R_B(A^*;Q)$ and $R_M(A) \geq R_B(A;Q) \geq R_B(A^*;Q)$ for all estimators A. $\square$

Example (binomial minimax). Suppose $X \sim \text{Binom}(n, \theta)$ and $L(a, \theta) = (a-\theta)^2$. Suppose $A^{\text{mean}}(x) = \frac{x}{n}$. Then $R(A^{\text{mean}}; \theta) = \frac{\theta(1-\theta)}{n}$ because A^{mean} is unbiased. This estimator is not minimax so we want to improve upon it.

Consider an estimator $A'(x) = a\frac{x}{n} + b$ for $a, b \in \mathbb{R}$. Then,

$$R(A'; \theta) = \text{variance} + \text{bias}^2 = \text{Var}(A') + (\mathbb{E}[A'(x)] - \theta)^2 = \frac{a^2}{n}\theta(1-\theta) + ((a-1)\theta + b)^2.$$

Choose $a = \frac{\sqrt{n}}{\sqrt{n}+1}$ and $b = \frac{1}{2(\sqrt{n}+1)}$. Observe that $R(A'; \theta) = \frac{1}{4(\sqrt{n}+1)^2}$ for all θ. Furthermore,

$$A'(x) = \frac{\sqrt{n}}{\sqrt{n}+1} \cdot \frac{x}{n} + \frac{1}{\sqrt{n}+1} \cdot \frac{1}{2}$$

is a convex combination of $\frac{x}{n}$ and $\frac{1}{2}$.

Observe that A' is Bayes when $Q \sim \text{Beta}\left(\frac{\sqrt{n}}{2}, \frac{\sqrt{n}}{2}\right)$. From Corollary 1, A' is minimax optimal. Figure 1 depicts the risk values for A^{mean} and A' for a large value of n.

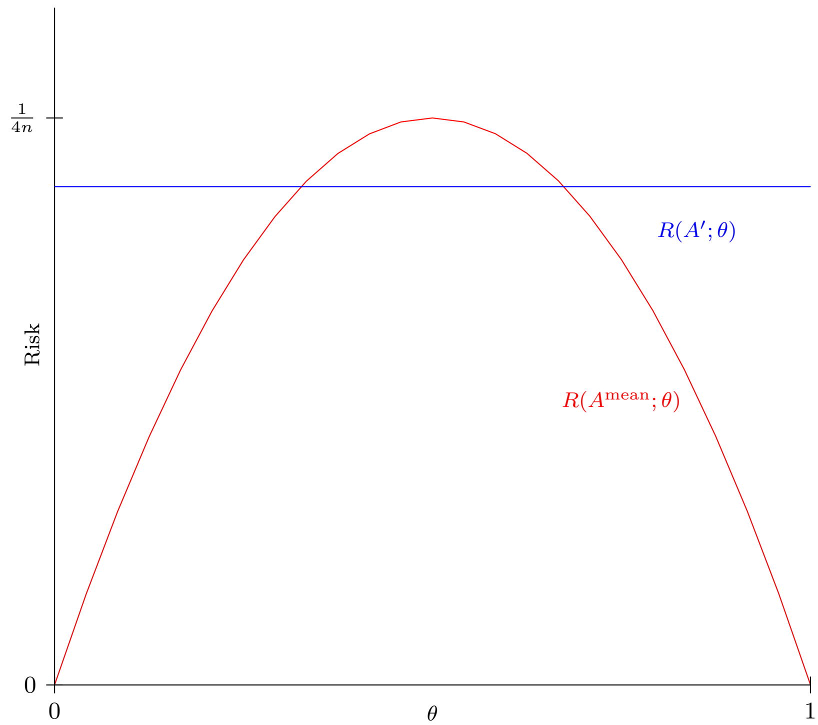

Figure 1: The red curve is $R(A^{\text{mean}}; \theta)$ and the blue line is $R(A'; \theta)$. The value of n depicted is $n = 225$. As n increases the blue line will approach the vertex of the red curve.

3 Admissibility

Definition 2. *An estimator A is **inadmissible** if there exists A' such that $R(A'; \theta) \leq R(A; \theta)$ for all $\theta \in \Theta$ and there exists $\theta \in \Theta$ such that $R(A'; \theta) < R(A; \theta)$. Furthermore, A is **admissibile** if*

A is not inadmissible.

Theorem 3. *A unique Bayes estimator A^* with respect to prior Q is admissible.*

Proof. Suppose A^* is not admissible for the sake of contradiction. Then there exists $A \neq A^*$ such that $R(A;\theta) \leq R(A^*;\theta)$ for all $\theta \in \Theta$. Then

$$R_B(A;Q) = \int R(A;\theta)dQ(\theta) = \int R(A^*;\theta)dQ(\theta) = R_B(A^*;Q).$$

Since $A \neq A^*$, this contradicts the uniqueness of A^*. □

Example (Gaussian mean). $\hat{\theta}^{\text{median}}$ is not admissible since $\hat{\theta}^{\text{median}}$ is dominated by $\hat{\theta}^{\text{mean}}$. However, $\hat{\theta}^{\text{reg}}$ is admissible by Theorem 3 because the estimator is Bayes. Furthermore $\hat{\theta}^{\text{mean}}$ is admissible although the proof is not obvious.

4 Gaussian linear model

4.1 Model

Suppose $Y_i = Z_i^T\theta + \varepsilon_i$, $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$, and $z_i \in \mathbb{R}^d$ is fixed for $1 \leq i \leq n$ and $\theta \in \Theta = \mathbb{R}^d$. In matrix form, $\vec{Y} = Z\theta + \vec{\varepsilon}$ where $Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n$, $\vec{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \in \mathbb{R}^n$, and $Z = \begin{pmatrix} -z_1^T- \\ \vdots \\ -z_n^T- \end{pmatrix} \in \mathbb{R}^{n\times d}$. Equivalently, $\vec{Y} \sim \mathcal{N}(Z\theta, \sigma^2 I_n)$.

Example (Least-squares estimator (OLS)). $\hat{\theta}^{\text{LS}} = \text{argmin}_{\theta\in\Theta}\|Y - Z\theta\|_2^2 = (Z^TZ)^{-1}Z^TY$. Observe $Z = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ implies that $\hat{\theta}^{\text{LS}} = \frac{1}{n}\sum_{i=1}^n Y_i$ so the least-squares estimator generalizes the mean estimator.

Proposition 4. *$\hat{\theta}^{LS}$ is the maximum likelihood estimator in the Gaussian linear model.*

Proof. First observe that $P_\theta(y) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n(y_i - z_i^T\theta)^2\right) = \exp\left(-\frac{1}{2\sigma^2}\|Y - Z\theta\|_2^2\right)$ and then use the previous example. □

Proposition 5. *$\hat{\theta}^{LS} \sim \mathcal{N}(\theta, (Z^TZ)^{-1}\sigma^2)$.*

Proof. $\hat{\theta}^{\text{LS}}$ is multivariate Gaussian because it is a linear function of the multivariate Gaussian distribution Y. We have that

$$\begin{aligned}
\mathbb{E}[\hat{\theta}^{\text{LS}}] &= \mathbb{E}[(Z^TZ)^{-1}Z^TY] = (Z^TZ)^{-1}Z^T\mathbb{E}[Y] = (Z^TZ)^{-1}Z^TZ\theta = \theta, \\
\text{Var}[\hat{\theta}^{\text{LS}}] &= \text{Var}[(Z^TZ)^{-1}Z^T(Z\theta+\varepsilon)] = (Z^TZ)^{-1}Z^TZ(Z^TZ)^{-1}\sigma^2 = (Z^TZ)^{-1}\sigma^2,
\end{aligned}$$

which finishes the proof. □

4.2 Anderson's lemma

Definition 6. *A set $S \subset \mathbb{R}^d$ is* ***symmetric*** *if $-S = S$. A function $\ell : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ is* ***bowl-shaped*** *if its level sets $\{\theta : \ell(\theta) \leq c\}$ are convex and symmetric for all $c \in \mathbb{R}_{\geq 0}$.*

Example. Examples of bowl-shaped functions are $\ell(x) = \|x\|_2^2$ and $\ell(x) = \|x\|_1$. Furthermore there exist non-convex functions f such that $\ell(x) = f(\|x\|_2)$ is bowl-shaped.

Theorem 7 (Anderson's lemma)**.** *Suppose $\ell : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ is bowl-shaped and $\varepsilon \sim \mathcal{N}(0, \Sigma) \in \mathbb{R}^d$. Then $R_1(x) = \mathbb{E}[\ell(x + \vec{\varepsilon})]$ is minimized at $x = 0$.*

Proof. The convex case is straightforward. For the general case please refer to the textbook. □

4.3 Bayes in Gaussian linear model

Suppose $Q \sim \mathcal{N}(0, \tau^2 I_d)$ is the prior and $L(a, \theta) = \ell(a - \theta)$ for bowl-shaped ℓ is the loss function. The posterior is

$$\mathbb{P}(\theta|Y) \propto \exp\left(-\frac{1}{2\sigma^2}\|Y - Z\theta\|_2^2 - \frac{1}{2\tau^2}\|\theta\|_2^2\right).$$

1. Note that $\log(\mathbb{P}(\theta|Y)) = C - \frac{1}{2\sigma^2}\|Y - Z\theta\|_2^2 - \frac{1}{2\tau^2}\|\theta\|_2^2$, where C is a constant, is a quadratic in θ so the posterior is Gaussian.

2. Since the posterior is Gaussian, its mean is its mode. Denote this quantity by $\hat{\theta}_{\text{posterior}}$. We have that

$$\hat{\theta}_{\text{posterior}} = \text{argmin}_{\theta \in \mathbb{R}^d}\left(\frac{1}{2\sigma^2}\|Y - Z\theta\|_2^2 + \frac{1}{2\tau^2}\|\theta\|_2^2\right) = \text{argmin}_{\theta \in \mathbb{R}^d}\left(\|Y - Z\theta\|_2^2 + \frac{\sigma^2}{\tau^2}\|\theta\|_2^2\right),$$

which is a ridge-regression problem. Denotes its solution by $\hat{\theta}^{\text{ridge}}$.

3. To compute the posterior variance, we only need to inspect the degree two terms of $\log(\mathbb{P}(\theta|Y))$. Particularly,

$$\log(\mathbb{P}(\theta|Y)) = -\frac{1}{2\sigma^2}\theta^T z^T z\theta - \frac{1}{2\tau^2}\theta^T\theta + c_1^T\theta + c_2$$

for some constants $c_1 \in \mathbb{R}^d$ and $c_2 \in \mathbb{R}$. Then, the posterior variance, which we denote by $\Sigma_\tau \in \mathbb{R}^{d \times d}$, is $\Sigma_\tau = \left(\frac{Z^T Z}{\sigma^2} + \frac{I_d}{\tau^2}\right)^{-1}$.

5 Conclusions

1. $\theta|Y$ has distribution $\mathcal{N}(\hat{\theta}^{\text{ridge}}, \Sigma_\tau)$.

2. The Bayes optimal estimator satisfies $\hat{\theta}^{\text{Bayes}} = \hat{\theta}^{\text{ridge}}$ by Anderson's lemma.

3. The Bayes risk equals $R_B(\hat{\theta}^{\text{Bayes}}) = \mathbb{E}[\ell(\Sigma_\tau^{\frac{1}{2}} W)]$, where $W \sim \mathcal{N}(0, I_d)$.