6.S951 Modern Mathematical Statistics Fall 2024

Lecture 2 — September 10, 2024

Prof. Stephen Bates

Scribe: Andi Qu

1 Outline

Last time:

1. Statistical decision theory framework
2. Sufficiency
   (a) How to find it: Fisher-Neyman factorization theorem
   (b) Rao-Blackwell theorem

Agenda:

1. Sufficiency parting thoughts
2. Bayes-optimal estimators
3. From Bayes to minimax

2 Sufficiency Parting Thoughts

Recall that a statistic $T : \mathcal{X} \mapsto \mathbb{R}^k$ is **sufficient** in $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ if the conditional distribution of $X \mid T = t$ is the same for all $\theta \in \Theta$.

Note that sufficient statistics are *not* unique.

2.1 Example (1D Gaussian Mean)

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 1)$. From Lecture 1, the sample mean $T_1(X) = \overline{X}$ is a sufficient statistic, but so is the whole dataset $T_2(X) = (X_1, \ldots, X_n)$.

In this example, $T_1(X)$ is one-dimensional while $T_2(X)$ is n-dimensional. The point of sufficient statistics is **data reduction**, so in a sense, $T_1(X)$ is a "better" statistic than $T_2(X)$.

2.2 Example (Order Statistics)

Let $\mathcal{X} \in \mathbb{R}^n$ and $\mathcal{P} = \{P_\theta^n : P_\theta \text{ is any distribution on } \mathbb{R}\}$. Assuming i.i.d. samples, the sorted vector of the samples $T(X) = (X_{(1)}, \ldots, X_{(n)})$ is a sufficient statistic.

By Rao-Blackwell, this result implies that order does not matter when dealing with i.i.d. samples.

3 Bayes-Optimal Estimators

In Lecture 1, we looked at two estimators for a 1D Gaussian mean:

1. The sample mean: $\hat{\theta}^{\text{mean}} = \frac{1}{n}\sum_{i=1}^N X_i$.
2. The *regularized* sample mean: $\hat{\theta}^{\text{reg}} = \hat{\theta}^{\text{mean}}/2$.

If we plot the risk of these two estimators, we see that neither one dominates the other everywhere. Given this ambiguity, how can we decide which estimator is "better"? One way is by comparing their Bayes risks.

Definition 1. *Let Q be a distribution on Θ. The **Bayes risk** of a statistical procedure A is:*

$$\begin{aligned} R_B(A; Q) &= \int_\Theta \int_\mathcal{X} L(A(x), \theta)\, dP_\theta(x)\, dQ(\theta) \\ &= \int_\Theta R(A; \theta)\, dQ(\theta) \end{aligned}$$

Definition 2. *A statistical procedure $A^\star$ is **Bayes-optimal** if:*

$$R_B(A^\star; Q) = \inf_A R_B(A; Q)$$

If $A^\star$ is Bayes-optimal, we say "$A^\star$ is Bayes" for short.

3.1 Finding Bayes Estimators

Theorem 3. *$A^\star$ is Bayes if:*

$$A^\star(x) \in \operatorname*{argmin}_a \mathbb{E}[L(a, \theta) \mid X = x]$$

Furthermore, if $\mathcal{A}$ is convex, $L(a; \theta)$ is strictly convex in a, and $\mathbb{E}[L(a, \theta) \mid X = x] < \infty$ for some a and x, then $A^\star$ is the unique Bayes estimator.

Proof. First, we prove optimality. We can flip the order of integration for Bayes risk to get:

$$\begin{aligned} R_B(A; Q) &= \int_\Theta \int_\mathcal{X} L(A(x), \theta)\, dP_\theta(x)\, dQ(\theta) \\ &= \int_\mathcal{X} \int_\Theta L(A(x), \theta)\, dQ(\theta)\, dP_\theta(x) \\ &= \mathbb{E}[\mathbb{E}[L(A(x), \theta) \mid X = x]] \end{aligned}$$

Let A be any other procedure or estimator. By definition:

$$\mathbb{E}[L(A^\star(x), \theta) \mid X = x] \leq \mathbb{E}[L(A(x), \theta) \mid X = x]$$

Taking the expectation of both sides, we get $R_B(A^\star; Q) \leq R_B(A; Q)$. Therefore, $A^\star$ is Bayes.

Next, we prove uniqueness. Let $F_x(a) = \mathbb{E}[L(a, \theta) \mid X = x]$. By our assumptions, $F_x(a)$ is strictly convex in a.

$\implies$ $A^\star$ is uniquely defined.

Now suppose we have another Bayes estimator A, so $\mathbb{E}[F_x(A(x)) - F_x(A^\star(x))] = 0$.

By definition of $A^\star$, we have $F_x(A(x)) - F_x(A^\star(x)) \geq 0$. The expectation of a non-negative random variable is zero if and only if that random variable is zero with probability 1.

$\implies$ $F_x(A(x)) = F_x(A^\star(x))$

But since $A^\star$ uniquely minimizes F_x, we must have $A = A^\star$. □

How can we interpret a Bayes estimator? By "being a Bayesian":

1. We start with a prior belief in the form of a distribution Q over Θ.
2. We see evidence/data x and form a posterior distribution on $\theta \mid X = x$.
3. We act optimally according to the posterior.

3.2 Example (Gaussian Mean Bayes Optimality)

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, \sigma^2)$ for some known σ. (More formally, $\mathcal{P} = \{P_\theta^n : P_\theta \text{ is } \mathcal{N}(\theta, \sigma^2)\}$ and $\mathcal{X} = \mathbb{R}^n$.) Also let $L(a, \theta) = (a - \theta)^2$.

Consider the Bayes risk w.r.t. prior $Q : \mathcal{N}(\mu_0, \tau^2)$.

Fact 4. *The conditional distribution $\theta \mid X$ is $\mathcal{N}\left((1 - B)\overline{X} + B\mu_0, (1 - B)\frac{\sigma^2}{n}\right)$, where $B = \frac{\sigma^2/n}{\sigma^2/n + \tau^2}$ represents how concentrated the prior was.*

Fact 5. *If $L(a, \theta) = (a - \theta)^2$, then $A^\star(x) = \mathbb{E}[\theta \mid X = x]$ is Bayes.*

$\implies$ $A^\star(x) = (1 - B)\overline{X} + B\mu_0$ is Bayes.

Depending on μ_0 and τ^2, it is possible for $\hat{\theta}^{\text{reg}}$ to be Bayes! Suppose $\mu_0 = 0$ and $\tau^2 = \frac{\sigma^2}{n}$.

Under this prior, $A^\star(x) = \frac{\overline{X}}{2} = \hat{\theta}^{\text{reg}}$.

3.3 Example (Beta-Binomial Conjugacy)

Let $X \sim \text{binom}(n, \theta)$, $L(a, \theta) = (a - \theta)^2$, and Q be Beta(a, b).

$\implies$ The posterior distribution is Beta$(a + X, b + n - X)$.

$\implies$ $A^\star(x) = \mathbb{E}[\theta \mid X = x] = \frac{X + a}{a + b + n}$ is Bayes.

4 From Bayes to Minimax

Recall that minimax risk is defined as the worst-case risk:

$$R_M(A) = \sup_{\theta \in \Theta} R(A; \theta)$$

Observation 6. *Since R_B averages R, we must have $R_M(A) \geq R_B(A; Q)$ for any Q.*

From this observation, we get the following inequality:

$$\inf_A R_M(A) \geq \inf_A R_B(A; Q) \geq R_B(A^\star; Q)$$

This inequality gives us a general strategy for finding minimax estimators:

1. Find some Q that maximizes $R_B(A^\star; Q)$.
2. Find some A such that $R_M(A) = R_B(A^\star; Q)$.
3. Conclude that A is minimax (and celebrate).

4.1 Example (1D Gaussian Mean)

Claim 7. *$\hat{\theta}^{mean} = \frac{1}{n}\sum_{i=1}^n X_i$ is minimax for $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\theta, \sigma^2)$.*

Proof. Following the above strategy:

1. First, we want to maximize $R_B(A^\star; Q)$.

$$\begin{aligned} R_B(A^\star; Q) &= \int_\Theta \int_{\mathcal{X}} (A^\star(x) - \theta)^2 \, dP_\theta(x) \, dQ(\theta) \\ &= \mathbb{E}[\mathbb{E}[(A^\star(x) - \theta)^2 \mid X = x]] \\ &= \mathbb{E}[\mathrm{Var}(\theta \mid X = x)] \\ &= (1 - B)\frac{\sigma^2}{n} \end{aligned}$$

   $\implies R_M(A) \geq (1 - B)\frac{\sigma^2}{n}$

   We want $B \to 0$ to maximize the RHS. We achieve this by setting $\tau^2 = \infty$ in the prior distribution Q.

2. Next, we note that $R_M(A^{\text{mean}}) = \frac{\sigma^2}{n}$ exactly. (The proof is left as an exercise.)

3. Therefore, $\hat{\theta}^{\text{mean}}$ is minimax.

□

4.2 A Surprising Example (Binomial Minimax)

Let $X \sim \text{binom}(n, \theta)$.

Claim 8. *$\hat{\theta}^{mean} = \frac{X}{n}$ is **not** minimax. The estimator $\hat{\theta} = \frac{X+\sqrt{n/4}}{n+\sqrt{n}}$ yields a lower minimax risk.*

Proof. Left as an exercise. (Or read Wasserman chapter 12.2.) □