6.S951 Modern Mathematical Statistics Fall 2024

Lecture 1 — September 5, 2024

Prof. Stephen Bates

Scribe: Hyewon Jeong, Jenny Huang

1 Outline

Topics that will be covered in this semester include:

- **Statistical Optimality:** procedures (positive results) and hardness results (lower bounds, fundamental limits of what can be learned) – 1st and 3rd quarters of the course.
- **Statistical Inference:** procedures that return the best guess with some form of correctness (certainty) assessment (e.g. a confidence interval) – 2nd and 4th quarters of the course.

In this lecture, we will cover Statistical Decision Theory (Section 2) and Sufficiency (Section 3), which covers Statistically lossless data reduction.

Agenda for today's class:

1. Statistical Decision Theory
2. Sufficiency (lossless data reduction)

Last time: N/A.

2 Statistical Decision Theory

What estimator (e.g., maximum likelihood estimator, the method of moments estimator, the posterior mean) can we choose for learning from noisy data?

Definition 1 (Statistical Model)**.** *A statistical model is a family of probability distributions,* $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$*, where* P_θ *is a probability distribution over sample space,* $\mathcal{X}$*, parameterized by* $\theta \in \Theta$*, where* Θ *is called the parameter space.*

Remark 1. Θ may be parametric (finite-dimensional, $\theta \in \mathbb{R}^d$) or non-parametric (infinite-dimensional).

Intuition 1. Given some observed data $\mathcal{X} \sim P_\theta$, the goal is to estimate θ or to take some action based on θ.

Let $\mathcal{A}$ be the space of actions, the *decision space*, and let $A : \mathcal{X} \mapsto \mathcal{A}$ be a mapping from the sample space to the space of actions. If $\mathcal{A} = \Theta$, the problem is called *estimation*, where $A(X) = \hat{\theta}(X)$.

Definition 2 (Statistic)**.** *A statistic, $T(X)$, is some function of the data, $T : \mathcal{X} \mapsto \mathbb{R}^k$.*

The quality of an action is measured with a *loss function.*

Definition 3 (Loss Function)**.** *The loss function is a function mapping the action space to the space of non-negative real numbers, $\mathcal{L} : A(X) \times \Theta \mapsto \mathbb{R}_{\geq 0}$ (e.g. mean squared error $\mathcal{L}(a, \theta) = ||a - \theta||^2$).*

The loss is *random*; it depends on the particular data set.

Definition 4 (Risk)**.** *The Risk of the estimator is the expected loss, $R : A(X) \times \Theta \mapsto \mathbb{R}_{\geq 0}$.*

$$R(A, \theta) = \mathbb{E}_{X \sim P_\theta}[\mathcal{L}(A(X), \theta)]$$

The risk is a property of a statistical procedure, A. It is a function over the parameter space, Θ.

The goal is to find procedures with low risk.

Example (Gaussian Mean).

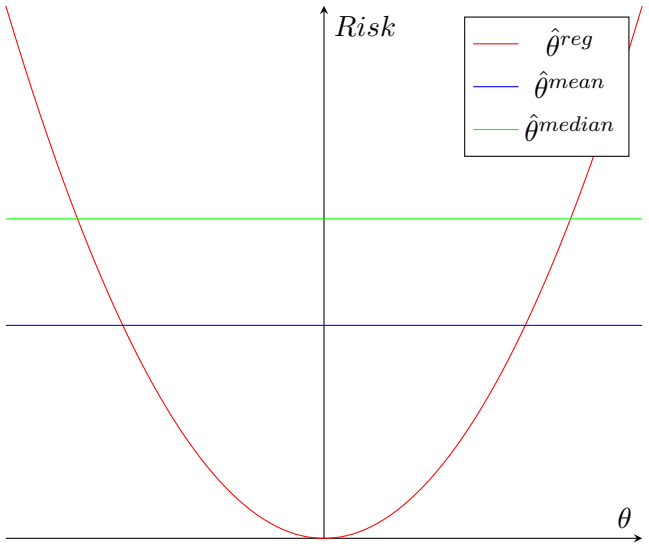
Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\theta, 1)$ be i.i.d. samples following $\mathcal{N}(\theta, 1)$, with

$$\mathcal{X} = \mathbb{R}^n, \quad \mathcal{P} = \{P_\theta^n : \mathcal{N}(\theta, 1)\}$$

Consider the following estimators:

- $\hat{\theta}_{\text{mean}}(X) = \frac{1}{n} \sum_{i=1}^n X_i$ (Mean)
- $\hat{\theta}_{\text{median}}(X) = \text{Median}(X_1, \ldots, X_n)$ (Median)
- $\hat{\theta}_{\text{reg}}(X) = (1 - \gamma)\hat{\theta}_{\text{mean}}(X)$ (Regularization estimator)

Here one can ask: *which procedure $\mathcal{A}$ are good procedures?* Key complexity is that the risk depends on θ. Let's take an example from Gaussian distribution:



Observe. The mean has lower risk than the median for all $\theta \in \Theta$. The regularization estimator has a lower risk for θ near zero but a higher risk as the magnitude of θ becomes large.

2.1 Optimality

We need definition of "doing good" or "optimal", either by 1) averaging over θ space (*Bayes Risk)* or 2) by measuring worse case behavior over all plausible scenario (*Minimax Risk).*

1. Bayes risk: Average over $\theta \in \Theta$ (procedures). $R_B(A) = \mathbb{E}_{\theta \sim Q}[R(\mathcal{A}, \theta)]$.
2. Minimax risk: The minimum worst case risk. $R_M(A) = max_{\theta \in \Theta} R(A, \theta)$.

Admissibility: A statistical procedure, A, is admissible if it is not dominated ($R(A, \theta) \leq R(A', \theta)$ for all $\theta \in \Theta$) by any $A' \in \mathcal{A}$ (ex: the median is *not* admissible).

Example (Non-parametric Mean)

Let $\mathcal{P} = \{P_\theta^n : P_\theta$ is any distribution on $\mathbb{R}$ (with finite mean and variance)$\}$ and $\{X_1, ..., X_n\}$ be n i.i.d. samples from the distribution. Consider the mean squared error loss function $\mathcal{L}(A; \theta) = (A(X) - \mu(P_\theta))^2$ where $\mu(P_\theta) = \mathbf{E}_{X \sim P_\theta} X_1$ then $A^{\text{mean}}(X) = \frac{1}{n}\Sigma_{i=1}^n X_i$. The risk, R, is derived as:

$$R(A^{\text{mean}}, \theta) = \mathbf{E}[\frac{1}{n}\Sigma_{i=1}^n X_i - \mu(P_\theta)]^2 = Var(\frac{1}{n}\Sigma X_i) = \frac{1}{n}Var(X_i)$$

Example (Ising Model) Consider a set of lattice slices where within each lattice site, there's a discrete variable assigned to denote spin of each atom i: $\mathcal{X} = \{-1, 1\}^n$. This is with parameter space $\Theta = \mathbf{R}^{\binom{n}{k}}$ where $\theta \in \Theta$ follows $\theta = \{\theta_{ij}, 1 \leq i \leq j \leq n\}$. The probability distribution is: $P_\theta = \frac{1}{C(\theta)} exp(\Sigma_{1 \leq i < j \leq n} \theta_{ij} x_i x_j)$, where $C(\theta)$ is a constant dependent on θ. If the spin of i and j is in the same direction, then the component $x_i x_j$ will have a positive value.

3 Sufficiency

Definition 5 (Sufficiency)**.** *A statistic $T : \mathcal{X} \mapsto \mathbb{R}^k$ is sufficient if the conditional distribution of $X|T(x) = t$ does not depend on θ.*

Intuition 2. A sufficient statistic is a compression of the data that has all of the information about the parameter. If we know $T(x)$ but not θ, we can simulate a new dataset that is "as good as" the original dataset in the statistical sense.

In particular, if we draw $x' \sim x|T(x) = t$, then $x' \stackrel{d}{=} x$.

Intuition 3. $T(x)$ is a sufficient statistic if x is independent of θ given $T(x)$, i.e. $x \perp\!\!\!\perp \theta | T(x)$. [1]

3.1 Fisher–Neyman factorization theorem

How can we tell when a statistic is sufficient? Given a probability density, can we tell whether a statistic is sufficient?

[1] If θ were a random variable, then this statement would be formal, but θ need not be random for the intuition of this statement to hold.

Theorem 6. *Fisher-Neyman Factorization: Let $X \in \mathbb{R}^d$ denote the data and let P_θ denote a density (with respect to a Lebesgue measure) for all $\theta \in \Theta$. Then a statistic $T : \mathcal{X} \mapsto \mathbb{R}^k$ is sufficient if and only if*

$$P_\theta(X) = g(T(X), \theta) f(X)$$

for some $g : \mathbf{R}^k \times \Theta \mapsto \mathbb{R}_{\geq 0}$ and $f : \mathcal{X} \mapsto \mathbb{R}_{\geq 0}$. [2]

Remark. θ only touches the data through $T(X)$.

Proof. Left as an exercise. □

Example (Gaussian Mean).

Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\theta, 1)$ be i.i.d. samples, with

$$\begin{aligned} P_\theta(\vec{X}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(X_i-\theta)^2}{2}} \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{i=1}^n X_i^2 + \theta \sum_{i=1}^n X_i - \frac{n\theta^2}{2}} \\ &= \underbrace{e^{\theta \sum_{i=1}^n X_i - \frac{n\theta^2}{2}}}_{g(T(X),\theta)} \underbrace{\frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{i=1}^n X_i^2}}_{f(X)}. \end{aligned}$$

Here, $T(X) = \sum_{i=1}^n X_i$.

3.2 Rao-Blackwell Theorem

We can always take an estimator and make it into a statistically better one by *only* using a sufficient statistic.

Theorem 7. *Rao-Blackwell Theorem: Let action space, $\mathcal{A}$, be convex. Let $\mathcal{L}(a, \theta)$ be convex in a for all $\theta \in \Theta$. Let $A : \mathcal{X} \mapsto \mathcal{A}$ be a statistical procedure, and let T be a sufficient statistic. Consider $A'(X) := \mathbb{E}[A(x)|T(x)]$. Then $R(A', \theta) \leq R(A, \theta)$ for all $\theta \in \Theta$.*

Proof. Left as an exercise (Hint: apply Jensen's Inequality). □

Given a sufficient statistic, the Rao-Blackwell theorem gives a recipe for creating a lower-risk statistical procedure.

Remark. In the case where $\mathcal{L}(a, \theta) = (a - \theta)^2$, $A'(X)$ has the same bias but lower variance than $A(X)$.

Example (Gaussian Mean).

[2] This theorem also holds for p.m.f.'s (discrete distributions).

Let $A(X) = median(X)$. Recall that the mean, $T(X) = \frac{1}{n}\sum_{i=1}^{n} X_i$, is sufficient. Given the mean, we can simulate data based on the mean, take its median, and that would be just as good as the original median.

$$A(X)|T(X) = t \sim f(t, \cdot)$$

for some distribution, $f(t, \cdot)$, centered at t.

Remark. Taking the median of the original data is equivalent to taking the mean and performing a symmetric sampling operation around the mean.