STAT 581: Advanced Theory of Statistical Inference I **Autumn 2026**

Lecture 1: Elementary decision theory

Instructor: Yen-Chi Chen

⊙ We thank previous instructors: Jon Wellner, Alex Luedtke, Fang Han, and Andrea Rotnitzky.

⊙ Much of this note is modified from Jon Wellner's lecture note https://sites.stat.washington.edu/jaw/COURSES/580s/581/lectnotes.18.html

1.1 Elementary Decision Theory

In the elementary decision theory, we consider the following setup:

- **Data:** $X \in \mathcal{X} \subset \mathbb{R}^d$ is a random variable (or a random vector) from a distribution $p_\theta(x)$ indexed by a parameter.
- **Parameter/Nature:** $\theta \in \Theta$, the parameter that determines the distribution generating our data.
- **Action:** $a \in \mathcal{A}$, where $\mathcal{A}$ is the action space. The action space is a finite or continuous space.

Decision rule. We consider a *stochastic decision* in the form of the conditional probability distribution $D(a|x)$. We call $D(a|x)$ a *decision rule* and denote $\mathcal{D}$ to be the collection of possible decision rules. When $\mathcal{A}$ is finite, we denote $\rho(a|x)$ to be the conditional probability of taking action a when we observe $X = x$. When $\mathcal{A}$ is continuous, we denote $\rho(a|x)$ to be the conditional probability density of taking action a when we observe $X = x$.

The decision rule includes *non-random decision rules* that we assign a probability of 1 to a specific value $\delta(x) \in \mathcal{A}$ when $X = x$. For such decision rule, the underlying probability measure is a point mass at $\delta(x)$, so we will just denote such rule as $\delta(x)$.

Statistical decision theory tries to find the best decision rule inside $\mathcal{D}$. To evaluate the performance of a rule D, we consider the *loss function*

$$L : \Theta \times \mathcal{A} \to \mathbb{R}.$$

In the above decision model, we have two layers of randomness, the sampling randomness in X from p_θ, and the randomness in the decision rule $D(a|x)$. So the loss will be a random quantity, which is not ideal for evaluating the performance of a rule.

To resolve this problem, we consider the *risk function*, which is the expected loss:

$$R(\theta, D) = \int_{\mathcal{X}} \int_{\mathcal{A}} L(\theta, a) D(da|x) p_\theta(x) dx = \begin{cases} \int \sum_{a\in\mathcal{A}} L(\theta, a)\rho(a|x)p_\theta(x)dx, & \text{if } \mathcal{A} \text{ is finite} \\ \int L(\theta, a)\rho(a|x)da\, p_\theta(x)dx, & \text{if } \mathcal{A} \text{ is continuous.} \end{cases}$$

With the risk function, we can then compare decision rules and discuss the *optimal decision rule*.

Example 1.1 (Point estimate) *The conventional point estimation can be viewed as a decision problem. Suppose we observe $X_1, \cdots, X_n \sim N(\mu, 1)$ and our goal is to estimate the unknown parameter $\mu \in \mathbb{R}$. In* *this case, the action space $\mathcal{A} = \mathbb{R}$ is our estimate of μ. A common loss function for this problem is a squared loss:*

$$L(\mu, a) = (\mu - a)^2.$$

The sample mean is a non-random decision rule that $\delta(x) = \bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Namely, $D(a|x)$ is a point mass at $a = \frac{1}{n}\sum_{i=1}^n x_i$.

For a given decision rule D, its risk depends on the underlying parameter θ, which is somewhat not ideal since the parameter θ is generally not observed. To remove the effect of θ in the risk function, we consider two frameworks:

- **'Average' case scenario: Bayesian framework.** In the Bayesian framework, we introduce a prior distribution π on the parameter space. And our goal is to find the *Bayes rule*: the decision rule that minimizes the *Bayes risk*:

  $$R_\pi(D) = \int_\Theta R(\theta, D)\pi(\theta)d\theta.$$

  Namely, the Bayes risk is the average risk over the prior distribution on Θ.

  The Bayes rule can be viewed as the optimal rule under *average* behavior of the risk function. Here the average is over the parameter space via the prior distribution.

- **'Worst' case scenario: Minimax framework.** The minimax framework does not place a prior on θ. Instead, it considers the *worst case* scenario of θ. The *minimax risk* is

  $$R^*(D) = \sup_{\theta \in \Theta} R(\theta, D)$$

  and our goal is to find the *minimax rule:* the decision rule that minimizes this worst case risk value.

Admissibility. A decision rule D is called *inadmissible* if there exists another decision rule D' such that $R(\theta, D') \leq R(\theta, D)$ for some θ and there is at least one $\theta' \in \Theta$ such that $R(\theta', D') < R(\theta', D)$. If a decision rule is not inadmissible, it is admissible.

Clearly, a decision rule D we pick should be admissible, otherwise why not just choose the other rule D' that making D inadmissible? However, later we shall see that some conventional estimator, such as the sample mean, may not be admissible!

Remark 1.2 *There are two minimaxity in Statistics. The* **minimax rule** *in the decision theory that we will cover here in STAT 581. This minimax rule is the classical minimaxity dated back to Wald and the goal is to find the* exact *minimal rule/estimator under a parametric model. The other minimaxity is the* **minimax lower bound (minimax rate)** *for nonparametric or high-dimensional models. It is a modern research topic about the limit of experiments and* convergence rate *of estimators in the worst case pioneered by Le Cam, Fano, and Assouad, and will be covered in STAT 582.*

Example 1.3 (Bus waiting problem) *Now we examine a simple example on bus waiting problem. Suppose I live near Seattle Children's hospital. One morning I wake up very late, and the class will start in* 7 *minutes. Fortunately, there is a bus stop just outside my home, so I have two possible ways to school: taking the bus or walking to school. Walking to school takes a constant* 9 *minutes so I will always be late for* 2 *minutes. Taking the buss may be faster but there is a huge uncertainty here. To simplify the problem, we assume that there are only four possible scenarios: no traffic (bus takes* 3 *mins), light traffic (bus takes* 6 *mins), regular traffic (bus takes* 9 *mins), and heavy traffic (but takes* 12 *mins). When I go outside, I can see the traffic and then make the decision on how to commute to school. My goal is to minimize the amount of* late time*.*

In this case, the action space has two possible elements, so we denote it as a binary number:

$$A = \begin{cases} 0, & \textit{if we take the bus,} \\ 1, & \textit{if we walk.} \end{cases}$$

The parameter space is the traffic scenario, which we denote it as

$$\theta = \begin{cases} 3, & \textit{if no traffic,} \\ 6, & \textit{if light traffic,} \\ 9, & \textit{if regular traffic,} \\ 12, & \textit{if heavy traffic.} \end{cases}$$

And the loss function is the following table:

$L(\theta, a)$	$\theta = 3$	$\theta = 6$	$\theta = 9$	$\theta = 12$
$A = 0$	0	0	2	5
$A = 1$	2	2	2	2

Suppose my perception of the traffic outside of my home is a binary random variable X such that

$$X = \begin{cases} 0, & \textit{if no so many traffic,} \\ 1, & \textit{if I see many traffic.} \end{cases}$$

And for simplicity, we use a simple Bernoulli to model the distribution of X given parameter θ:

$$X \sim \mathsf{Ber}(\theta/15).$$

Under this setup, now we investigate how to find a good decision rule.

For simplicity, we consider a non-random decision rule $D_\delta(x)$ that place a point mass at $\delta(x) \in \{0, 1\}$. For many problems, analyzing non-random decision rule is useful since a stochastic decision rule can be viewed as a randomized *non-random decision rule.*

One thing you may immediately notice is: the loss function is in fact independent of the data–it is purely a characteristic of the action and the parameter. However, it will be a random quantity if we have choose a decision rule $\delta(x)$. For a given θ and a decision rule $D_\delta(x)$, the risk is

$$R(\theta, D_\delta) = \sum_{x=0,1} L(\theta, \delta(x))P_\theta(x) = L(\theta, \delta(1))\left(\frac{\theta}{15}\right) + L(\theta, \delta(0))\left(1 - \frac{\theta}{15}\right).$$

A straightforward calculation shows that

$$\begin{aligned} R(3, D_\delta) &= L(3, \delta(1)) \times 0.2 + L(3, \delta(0)) \times 0.8 \\ R(6, D_\delta) &= L(6, \delta(1)) \times 0.4 + L(6, \delta(0)) \times 0.6 \\ R(9, D_\delta) &= L(9, \delta(1)) \times 0.6 + L(9, \delta(0)) \times 0.4 \\ R(12, D_\delta) &= L(12, \delta(1)) \times 0.8 + L(12, \delta(0)) \times 0.2. \end{aligned}$$

Now we run into a problem: the risk value depends on θ! Clearly, when $\theta = 3, 6, 9$, the optimal strategy is to choose $\delta(x) = 0$ regardless of $x = 0$ or $x = 1$. On the other hand, when $\delta = 12$, the optimal strategy is to choose $\delta(x) = 1$, again regardless of $x = 0$ or $x = 1$.

The problem is: we do NOT know θ, *so we need a way to remove its effect and design a good rule* $\delta(x)$.

The Bayesian framework and the minimax framework offer solution to this problem. In the Bayesian framework, we impose a prior on θ, *so we can* average *these risks to obtain a good decision rule (known as the Bayes rule). In the minimax framework, we consider the worst case* θ *and try to find a rule to minimize this worst case scenario.*

You can show that the minimax rule is $\delta_{MM}(x) = 1$ *(we always walk to school) while the Bayes rule depends on the prior distribution. Suppose we consider a uniform prior, i.e.,* $P(\theta = 3) = P(\theta = 6) = P(\theta = 9) = P(\theta = 12) = \frac{1}{4}$, *our Bayes rule is*

$$\delta_\pi(x) = \begin{cases} 0, & \textit{if } X = 0, \\ 1, & \textit{if } X = 1. \end{cases}$$

Namely, if we do not see traffic, we take the bus. If we see traffics, we walk to school.

You can easily show that both the minimax and the above Bayes rule are admissible. An inadmissible rule is:

$$\delta_{\mathsf{bad}}(x) = \begin{cases} 1, & \textit{if } X = 0, \\ 0, & \textit{if } X = 1. \end{cases}$$

This rule $\delta_{\mathsf{bad}}(x)$ *is dominated by the Bayes rule, i.e.,* $R(\theta, \delta_\pi(x)) \leq R(\theta, \delta_{\mathsf{bad}}(x))$ *for all* θ *and when* $\theta = 3, 6, 12$, *we have* $R(\theta, \delta_\pi(x)) < R(\theta, \delta_{\mathsf{bad}}(x))$.

1.1.1 (Neyman-Pearson) Hypothesis testing

The hypothesis testing can be viewed as a decision problem. However, it is a *constrained* decision problem that is not in the conventional Bayesian or minimax framework.

In the hypothesis testing problem, we partition the parameter space into two subsets: $\Theta = \Theta_0 \cup \Theta_1$ and $\Theta_0 \cap \Theta_1 = \emptyset$, where Θ_0 is the scenario where the null hypothesis is correct. Our action $A \in \{0, 1\}$ is the decision to reject the null hypothesis or not. $A = 1$ when we reject the null. The data is the random variable X from a model p_θ, so our decision rule of the test can be viewed as a binary random variable $D(X) \in \{0, 1\}$ whose distribution depends on the value of X. Note that $D(X)$ can be non-random.

To associate the type-1 and type-2 errors to the loss function, we consider the following losses:

$$L(\theta, a) = I(a = 1)I(\theta \in \Theta_0) + \ell_1 I(a = 0)I(\theta \in \Theta_1),$$

where $\ell_1 > 0$ is a constant. Namely, if we falsely reject null while it is true, the loss is 1. If we fail to reject the null while it is false, the loss is ℓ_1.

For a given decision rule, D, the type-1 error is

$$\sup_{\theta \in \Theta_0} P(\text{reject } H_0; \theta) = \sup_{\theta \in \Theta_0} P(D(X) = 1; \theta) = \sup_{\theta \in \Theta_0} \mathbb{E}\left[L(\theta, D)\right].$$

The type-2 error is

$$P(\text{fail to reject } H_0; \theta) = P(D(X) = 0; \theta) = \mathbb{E}\left[L(\theta, D)\right]$$

when $\theta \in \Theta_1$.

The *Neymann-Pearson* test framework tries to find the optimal decision rule such that

$$\min_{D \in \mathcal{D}} \mathbb{E}\left[L(\theta, D)\right]$$

for $\theta \in \Theta_1$ subject to the constraint that

$$\sup_{\theta \in \Theta_0} \mathbb{E}\left[L(\theta, D)\right] \leq \alpha,$$

i.e., the type-1 error is controlled at α.

1.1.2 Point estimation: non-random decision rule

In point estimation problem, i.e., our goal is to estimate θ, the action space reduces to the parameter space (our action is our estimator). In this case, it is very common that the loss function is convex in $a \in \mathcal{A}$ and $\mathcal{A}$ is a convex set. For instance, the square loss $L(\theta, a) = (\theta - a)^2$ or absolute loss $L(\theta, a) = |\theta - a|$ are both convex loss function.

Theorem 1.4 (Point estimation with convex loss: non-random decision rule) *In the point estimation problem with a convex loss, then for any random decision rule $D(a|x)$, its average is a non-random decision rule $\delta(x) = \int aD(da|x)$ such that*

$$R(\theta, \delta) \leq R(\theta, D)$$

for every θ. Therefore, it suffices to consider non-random decision rule for point estimation.

Proof: Let $A(x)$ be the outcome of the random decision rule $D(\cdot|x)$. Clearly, $\delta(x) = \mathbb{E}(A(x))$. You can easily show that due to the Jensen's inequality, the risk is

$$R(\theta, D) = \int \mathbb{E}[L(\theta, A(x))]p(x)dx \geq \int L(\theta, \mathbb{E}(A(x))) \equiv \int L(\theta, \delta(x))p(x)dx = R(\theta, \delta).$$

So the average action always has a lower risk for any θ, which completes the proof. ■

In short, Theorem 1.4 states:

> There is no benefits of using a random decision rule for point estimation under convex loss.

This lays the foundation for why in point estimation, we are mostly considering non-random decision rule as our estimator. This echoes the idea of Rao-Blackwell's theorem that once we know the sufficient statistics, including other information (statistics) will not improve the estimator and could even worsen the accuracy!

Therefore, in most of this note, we will focus on non-random decision rule. When considering a non-random decision rule, we will just use $\delta(x)$ in place of D_δ for abbreviation.

1.2 Bayesian inference

When using the Bayesian framework, we consider the Bayes risk given a prior distribution and the model generates our data $p_\theta(x)$ becomes a conditional distribution $p(x|\theta)$. For simplicity, we assume that the prior distribution $\Pi(\theta)$ has a density function $\pi(\theta)$ and write the expectation

$$\mathbb{E}(\phi(\theta)) = \int \phi(\theta)\Pi(d\theta) = \int \phi(\theta)\pi(\theta)d\theta.$$

Now consider a non-random decision rule $D_\delta(x)$ that place a point mass at $\delta(x)$, the Bayes risk can be written as

$$\begin{aligned} R_\pi(D_\delta) &= \int_\Theta R(\theta, D_\delta)\pi(\theta)d\theta \\ &= \int_\Theta \int_\mathcal{X} L(\theta, \delta(x))p(x|\theta)dx\pi(\theta)d\theta \\ &= \int_\mathcal{X} \left[\int_\Theta L(\theta, \delta(x))\pi(\theta|x)d\theta\right] p(x)dx, \end{aligned}$$

where $p(x) = \int p(x|\theta)\pi(\theta)d\theta$ is the marginal probability of X. The above form is useful because the quantity inside the bracket,

$$Q(a|x) = \int_\Theta L(\theta, a)\pi(\theta|x)d\theta$$

can often be optimized separately for each x, making it easy to obtain the Bayes rule.

A key quantity in the above analysis is the conditional distribution $\pi(\theta|x)$, which is known as the *posterior distribution* and is a critical concept in the Bayesian inference. So here we will briefly introduce the Bayesian inference.

Let X be the random variable representing our data and $p_\theta(x)$ be the model that generates X. In Bayesian inference, the parameter θ is viewed as a random variable from a prior distribution π. So the statistical model $p_\theta(x) = p(x|\theta)$ can be viewed as a conditional distribution.

The prior distribution π can be viewed as our belief of how likely θ to be at different values and the posterior distribution

$$\pi(\theta|x) \equiv p(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{p(x)}$$

represents our belief of θ after seeing the data $X = x$.

Since the posterior distribution is a conditional distribution of θ given x, the value of x is fixed for this distribution, so we have

$$\pi(\theta|x) \propto p(x|\theta)\pi(\theta) \propto f(\theta, x),$$

where $f(\theta, x)$ is called the *kernel* of the posterior distribution. Clearly, the kernel is not unique but once we know the kernel function, we can quickly obtain the posterior distribution via

$$\pi(\theta|x) = \frac{f(\theta, x)}{c_f(x)}, \qquad c_f(x) = \int f(\theta, x)d\theta.$$

A useful trick: we can just utilize the kernel function to find out what family the posterior distribution belongs to.

The prior $\pi(\theta)$ is called a conjugate prior to a statistical model $p(x|\theta)$ if the resulting posterior distribution and prior distribution belong to the same parametric family. Using a conjugate prior reduces the computation drastically since we only need to adjust the parameter values.

Example 1.5 (Conjugate prior 1: Beta-Binomial) *Consider* $X|\theta \sim \mathsf{Bin}(n, \theta)$ *and* $\theta \sim \mathsf{Beta}(\alpha, \beta)$*, where* N, α, β *are given. Then the posterior distribution is*

$$\pi(\theta|x) \propto p(x|\theta)\pi(\theta) \propto \binom{n}{x}\theta^x(1-\theta)^{n-x} \cdot \frac{1}{B(\alpha, \beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}.$$

So a kernel function to this posterior distribution is $f(\theta, x) = \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}$ *and it is clear that this belongs to the form of a PDF of a Beta distribution with parameter* $(\alpha + x, \beta + n - x)$*. Thus, the posterior* *distribution of θ given $X = x$ is Beta distribution with parameter $(\alpha + x, \beta + n - x)$, which belongs to the same family as the prior distribution.*

Example 1.6 (Conjugate prior 2: Poisson-Gamma) *Consider $X|\theta \sim \mathsf{Poisson}(\theta)$ and $\theta \sim \mathsf{Gamma}(\alpha, \lambda)$. Then the posterior distribution is*

$$\pi(\theta|x) \propto p(x|\theta)\pi(\theta) \propto \frac{1}{x!}\theta^x e^{-\theta} \cdot \frac{\lambda^\alpha \theta^{\alpha-1}}{\Gamma(\alpha)} e^{-\lambda\theta} \propto \theta^{x+\alpha-1}e^{-\theta(\lambda+1)}.$$

A kernel function is $f(\theta, x) = \theta^{x+\alpha-1}e^{-\theta(\lambda+1)}$, which clearly shows that it is a Gamma distribution with parameter $(\alpha + x, \lambda + 1)$.

Example 1.7 (Conjugate prior 3: Multinomial-Dirichlet) *Suppose $\theta \in \mathbb{R}^K$ satisfying $\theta_j \geq 0$ and $\sum_{j=1}^K \theta_j = 1$ (this is called a K-simplex). Given θ, the random vector $X \in \{0, 1, \cdots, N\}^K$ is from a multinomial distribution $\mathsf{Mult}(N, \theta)$ if $\sum_{j=1}^K X_j = N$ and its joint PMF is*

$$p(x_1, \cdots, x_K|\theta) = \frac{N!}{x_1! \cdots x_K!}\prod_{j=1}^K \theta_j^{x_j}.$$

You can easily see that the Binomial distribution is a special case of the multinomial distribution when $K = 2$.

To place a prior on feasible values of θ (K-simplex), we consider the Dirichlet distribution: $\theta \sim \mathsf{Dir}(\alpha)$, where $\alpha \in \mathbb{R}^K_{>0}$ and the PDF of θ is

$$\pi(\theta) = \frac{1}{D(\alpha)}\prod_{j=1}^K \theta_j^{\alpha_j - 1},$$

when $\theta_j \geq 0$ and $\sum_{j=1}^K \theta_j = 1$ and the density is 0 outside this feasible range. The Dirichlet distribution can be viewed as a generalized Beta distribution–when $K = 2$, the Dirichlet distribution reduces to Beta distribution.

In this case, the posterior distribution of θ given X is

$$\pi(\theta|x) \propto p(x|\theta)\pi(\theta) \propto \theta_j^{X_j + \alpha_j - 1},$$

so it is Dirichlet distribution with parameter $\alpha + X$.

Example 1.8 (Conjugate prior 4: Normal-normal) *Suppose we have random variable $X|\theta \sim N(\theta, \sigma^2)$ and we place a prior distribution on θ as $\theta \sim N(\mu, \tau^2)$. We assume that σ^2, μ, τ^2 are known. In this case, the posterior distribution of θ given X is*

$$\pi(\theta|x) \propto p(x|\theta)\pi(\theta) \propto e^{-\frac{1}{2\sigma^2}(x-\theta)^2} \cdot e^{-\frac{1}{2\tau^2}(\theta-\mu)^2},$$

which is a normal distribution

$$\theta|X \sim N\left(\frac{1/\tau^2}{1/\tau^2 + 1/\sigma^2}\mu + \frac{1/\sigma^2}{1/\tau^2 + 1/\sigma^2}X, \frac{1}{1/\tau^2 + 1/\sigma^2}\right). \tag{1.1}$$

1.3 Bayes rule

Recall that in the beginning of the previous section, we have the following decomposition of the Bayes risk for a non-random decision rule $\delta(x)$:

$$R_\pi(\delta) = \int_{\mathcal{X}} \left[\int_\Theta L(\theta, \delta(x))\pi(\theta|x)d\theta\right] p(x)dx.$$

Here are two insightful results that we can utilize.

Theorem 1.9 (Prevalence of non-random Bayes rule) *If the loss function $L(\theta, a)$ is convex for all θ and a, $\mathcal{D}$ is not constrained, $\mathcal{A}$ is a convex set, and there exists a Bayes rule $D_\pi \in \mathcal{D}$, then there exists a non-random Bayes rule $D_{\delta,\pi}$ such that it is Bayes rule and places a point mass at some location $\delta(x) \in \mathcal{A}$.*

Theorem 1.9 can be proved easily with Theorem 1.4.

Theorem 1.10 (Sufficiency for conditional risk minimization) *Consider the Bayesian framework where $X|\theta \sim p_\theta$ and $\theta \sim \pi$ and $L(\theta, a) \geq 0$ for all $\theta \in \Theta, a \in \mathcal{A}$. If there exist a non-random decision rule $D_\delta(x)$ placing a point mass at $\delta(x)$ that*

$$\int_\Theta L(\theta, \delta(x))\pi(\theta|x)d\theta = \min_{a\in\mathcal{A}} \int_\Theta L(\theta, a)\pi(\theta|x)d\theta$$

for almost every x, then $D_\delta(x)$ is a Bayes rule.

Theorem 1.10 shows that a simple way to find a Bayes rule is to consider the conditional quantity $Q(a|x) = \int_\Theta L(\theta, a)\pi(\theta|x)d\theta$ and choose $\delta(x)$ that minimizes the risk. Specifically, for each x, consider

$$\delta_\pi(x) = \mathsf{argmin}_{a\in\mathcal{A}} \int_\Theta L(\theta, a)\pi(\theta|x)d\theta. \tag{1.2}$$

Then $\delta_\pi(x)$ is the Bayes rule for the prior distribution π.

Example 1.11 (Mean square loss) *Consider the point estimation problem where we want to estimate $\Psi(\theta) \in \mathbb{R}$, a transformed quantity of the underlying parameter. Our action is our estimate. Consider the squared loss $L(\theta, a) = (\Psi(\theta) - a)^2$.*

What will our Bayes rule be in this case?

By equation (1.2), we only need to consider the conditional quantity

$$Q(a|x) = \int_\Theta (\Psi(\theta) - a)^2\pi(\theta|x)d\theta.$$

Clearly, this is a quadratic function of a, so the minimizer of it satisfies the first-order condition:

$$\frac{\partial}{\partial a}Q(a|x) = 0 = \int_\Theta \Psi(\theta)\pi(\theta|x)d\theta - a.$$

So we conclude the Bayes rule is

$$\delta_\pi(x) = \int_\Theta \Psi(\theta)\pi(\theta|x)d\theta,$$

which is the posterior mean.

Example 1.12 (Beta-Binomial revisited) *We now revisit the Beta-Binomial problem again. Consider $X|\theta \sim \mathsf{Bin}(n, \theta)$ and $\theta \sim \mathsf{Beta}(\alpha, \beta)$, where N, α, β are given. Then the posterior distribution is a Beta distribution with parameter $(\alpha + x, \beta + n - x)$.*

Suppose we are interested in estimating θ.

Under the square loss, the Bayes rule will be the posterior mean of $\mathsf{Beta}(\alpha + x, \beta + n - x)$, i.e.,

$$\delta_\pi(x) = \frac{\alpha + x}{\alpha + \beta + n} = \left(\frac{\alpha}{\alpha + \beta}\right)\frac{\alpha + \beta}{\alpha + \beta + n} + \left(\frac{x}{n}\right)\frac{n}{\alpha + \beta + n} = \textit{prior mean} \cdot W_n + \hat{\theta}_{MLE} \cdot (1 - W_n),$$

is a convex combination between the prior mean of θ and the maximum likelihood estimator $\hat{\theta}_{MLE} = x/n$ and the proportion of the prior $W_n = \frac{\alpha+\beta}{\alpha+\beta+n}$ is shrinking to 0 as the sample size $n \to \infty$.

1.4 Minimax rule

For Bayes rule, Theorems 1.9 and 1.10 are powerful and useful for finding a practical rule. The minimax rule, unfortunately, does not have such nice direct theoretical results. However, if we can change the problem of searching a minimax rule into searching a Bayes rule, we can utilize the above two theorems.

Recall that in the minimax framework, we consider a minimax risk of a decision rule D

$$R^*(D) = \sup_{\theta\in\Theta} R(\theta, D).$$

The minimax rule is D_{MM} such that $R^*(D)$ is minimized, i.e., for any other rule D', we have

$$R^*(D_{MM}) \leq R^*(D').$$

When we consider only the non-random rules, then the minimax rule is $\delta_{MM}(x)$ with

$$\sup_{\theta\in\Theta} R(\theta,\delta) \equiv R^*(\delta_{MM}) \leq R^*(\delta') \equiv \sup_{\theta\in\Theta} R(\theta,\delta') \tag{1.3}$$

for any other non-random decision rules δ'.

Before we proceed, we first introduce the concept of *least favorable prior.* A prior π_0 is called the least favorable prior if its Bayes risk

$$R_{\pi_0}(\delta_{\pi_0}) = \sup_{\pi} R_\pi(\delta_\pi).$$

Namely, this prior is like the worst prior that maximizes the Bayes risk.

Theorem 1.13 (Bayes rule ⇒ minimax rule) *Suppose π_0 is a prior distribution and δ_{π_0} is its Bayes rule. If the Bayes risk*

$$R_{\pi_0}(\delta_{\pi_0}) = \sup_{\theta\in\Theta} R(\theta, \delta_{\pi_0}),$$

then δ_{π_0} is a minimax rule and π_0 is the least favorable prior.

Proof: Minimaxity. For any other decision rule δ, we immediately have

$$\begin{aligned}
\sup_{\theta\in\Theta} R(\theta,\delta) &\geq \int R(\theta,\delta)\pi_0(\theta)d\theta \\
&\geq \int R(\theta,\delta_{\pi_0})\pi_0(\theta)d\theta \\
&\equiv R_{\pi_0}(\delta_{\pi_0}) \\
&= \sup_{\theta\in\Theta} R(\theta,\delta_{\pi_0}).
\end{aligned}$$

Thus, δ_{π_0} is minimax.

Least favorable prior. Consider another prior distribution ρ. Then its Bayes risk

$$\begin{aligned}
R_\rho(\delta_\rho) &\equiv \int R(\theta,\delta_\rho)\rho(\theta)d\theta \\
&\leq \int R(\theta,\delta_{\pi_0})\rho(\theta)d\theta \\
&\leq \sup_{\theta\in\Theta} R(\theta,\delta_{\pi_0}) \\
&= R_{\pi_0}(\delta_{\pi_0}).
\end{aligned}$$

■

Theorem 1.13 is a critical result that offers a simple way to find a minimax rule via a Bayes rule. The follow Corollary offers an even simpler way to find a minimax rule.

Corollary 1.14 (Constant risk Bayes rule ⇒ minimax rule) *If a Bayes rule δ_π has a constant risk, i.e., $R(\theta, \delta_\pi)$ is independent of θ, then it is minimax.*

Example 1.15 (Minimax rule in Beta-Binomial) *We now revisit the Beta-Binomial problem again. Consider $X|\theta \sim \mathsf{Bin}(n, \theta)$ and $\theta \sim \mathsf{Beta}(\alpha, \beta)$, where N, α, β are given. Then the posterior distribution is a Beta distribution with parameter $(\alpha + x, \beta + n - x)$. Suppose again we are interested in estimating θ.*

Under the square loss, the Bayes rule will be the posterior mean of $\mathsf{Beta}(\alpha + x, \beta + n - x)$, i.e.,

$$\delta_\pi(x) = \frac{\alpha + x}{\alpha + \beta + n} = \left(\frac{\alpha}{\alpha + \beta}\right) \frac{\alpha + \beta}{\alpha + \beta + n} + \left(\frac{x}{n}\right) \frac{n}{\alpha + \beta + n}.$$

Its risk under the square loss is

$$\begin{aligned}
R(\theta, \delta_\pi) &= \mathbb{E}((\theta - \delta_\pi(X))^2|\theta) \\
&= (\theta - \mathbb{E}(\delta_\pi(X)|\theta))^2 + \mathsf{Var}(\delta_\pi(X)|\theta) \\
&= \left(\theta - \frac{\alpha + n\theta}{\alpha + \beta + n}\right)^2 + \frac{n\theta(1-\theta)}{n^2} \frac{n^2}{(\alpha + \beta + n)^2} \\
&= \frac{1}{(n + \alpha + \beta)^2} \left[[(\alpha + \beta + n)\theta - \alpha - n\theta]^2 + n\theta(1-\theta)\right] \\
&= \frac{1}{(n + \alpha + \beta)^2} \left[\alpha^2 + (n - 2\alpha(\alpha + \beta)) \cdot \theta + ((\alpha + \beta)^2 - n) \cdot \theta^2\right].
\end{aligned}$$

To find the minimax rule, we can change our prior's parameter so that the above risk is a constant. This requires

$$n - 2\alpha(\alpha + \beta) = 0, \qquad (\alpha + \beta)^2 - n = 0,$$

which leads to the solution $\alpha = \beta = \sqrt{n}/2$!

Namely, if we choose the prior to be $\mathsf{Beta}(\sqrt{n}/2, \sqrt{n}/2)$, its risk function will be

$$R(\theta, \delta_\pi) = \frac{\alpha^2}{(n + \alpha + \beta)^2},$$

a constant. By Corollary 1.14, this Bayes rule is the minimax rule.

1.4.1 Finding minimax rule via a sequence of Bayes rules

Theorem 1.13 shows that we can find a minimax rule via a Bayes rule. This result can be further strengthen to allow a sequence of Bayes rules.

Theorem 1.16 (Verifying a minimax rule via a sequence of Bayes rules) *Let $\pi_1, \pi_2, \cdots$, be a sequence of prior distribution with Bayes rules $\delta_{\pi_1}, \delta_{\pi_2}, \cdots$. For the k-th prior distribution, let its Bayes risks be $r_k = R_{\pi_k}(\delta_{\pi_k})$.*

Assume that the sequence of the Bayes risk converges to a limit $\lim_{k\to\infty} r_k = r_\infty$ and there is a decision rule δ_0 satisfying

$$r_\infty = \sup_{\theta\in\Theta} R(\theta, \delta_0).$$

Then δ_0 is a minimax rule.

Theorem 1.16 is often used to verify if a given estimator/rule is minimax, so it is more like a confirmatory approach. In comparison, Theorem 1.13 shows a direct way to create a minimax estimator/rule from Bayes rule, so Theorem 1.13 is more like a constructive approach.

We can also generalize the definition of being least favorable to the sequence of priors and Theorem 1.16 also shows that this sequence of prior is least favorable.

Proof: The proof is almost the same as Theorem 1.13.

For any other decision rule δ, we immediately have

$$\sup_{\theta\in\Theta} R(\theta,\delta) \geq \int R(\theta,\delta)\pi_k(\theta)d\theta \geq \int R(\theta,\delta_{\pi_k})\pi_k(\theta)d\theta \equiv r_k.$$

Taking limit of $k\to\infty$ in both sides leads to

$$\sup_{\theta\in\Theta} R(\theta,\delta) \geq \lim_{k\to\infty} r_k \equiv r_\infty = \sup_{\theta\in\Theta} R(\theta,\delta_0).$$

Thus, δ_0 is minimax.

■

Theorem 1.16 suggests another way to find minimax rule when we have a candidate rule (estimator). For a candidate rule/estimator, if we can show that its risk is the limit of a sequence of prior distribution's risk, then this rule/estimator is minimax.

Example 1.17 (Sample mean is minimax under Gaussian: known variance) *An application of Theorem 1.16 is to show that the sample mean is minimax in a univariate Gaussian model. Suppose our data is IID $X_1,\cdots,X_n \sim N(\theta,\sigma^2)$, where σ^2 is known and θ, the population mean, is our parameter of interest.*

Consider the sample mean $\bar{X}_n$ as an estimator of the population mean θ. It is obvious that the sample mean has a risk $R(\theta,\bar{X}_n) = \frac{\sigma^2}{n}$.

Will the sample mean $\bar{X}_n$ be a minimax rule/estimator under the square loss $L(\theta,a) = (\theta-a)^2$? The answer is YES!

We will utilize the normal-normal conjugate prior. Suppose we have a prior $\theta \sim N(\mu,\tau^2)$, where μ,τ^2 are given. We denote this prior distribution as ϕ_{μ,τ^2}.

Using the fact that the sample mean $\bar{X}_n|\theta \sim N(\theta,\sigma^2/n)$ and equation (1.1), the posterior distribution of θ given the sample mean $\bar{X}_n$ is a normal distribution

$$\theta|\bar{X}_n \sim N\left(\frac{1/\tau^2}{1/\tau^2+n/\sigma^2}\mu + \frac{n/\sigma^2}{1/\tau^2+n/\sigma^2}\bar{X}_n, \frac{1}{1/\tau^2+n/\sigma^2}\right).$$

Thus, the Bayes rule, which is the posterior mean of θ given $\bar{X}_n$ is

$$\delta_{\phi_{\mu,\tau^2}}(\bar{X}_n) = \frac{1/\tau^2}{1/\tau^2+n/\sigma^2}\mu + \frac{n/\sigma^2}{1/\tau^2+n/\sigma^2}\bar{X}_n$$

and it has a Bayes risk

$$\begin{aligned}R_{\phi_{\mu,\tau^2}}(\delta_{\phi_{\mu,\tau^2}}) &= \mathbb{E}\left[(\theta - \delta_{\phi_{\mu,\tau^2}}(\bar{X}_n))^2\right] \\ &= \mathbb{E}\left[\mathbb{E}[(\theta - \mathbb{E}(\theta|\bar{X}_n))^2|\bar{X}_n]\right] \\ &= \mathbb{E}\left[\mathsf{Var}(\theta|\bar{X}_n)\right] \\ &= \frac{1}{1/\tau^2 + n/\sigma^2}.\end{aligned}$$

Thus, consider a sequence of prior distributions

$$N(\mu,1), N(\mu,4), N(\mu,9), N(\mu,16), \cdots, N(\mu,k^2), \cdots$$

and their Bayes risks will be approaching

$$\lim_{k\to\infty}\frac{1}{1/k^2 + n/\sigma^2} = \frac{\sigma^2}{n} = R(\theta, \bar{X}_n),$$

which is the risk of the sample mean. By Theorem 1.16, the sample mean is a minimax rule for square loss.

The above example shows that the sample mean is minimax when σ^2 is known. Now we generalize this result to cases where σ^2 is known. Before we proceed, we first introduce a useful lemma and generalize the risk to a nonparametric model $\mathcal{P}$. The set $\mathcal{P}$ is a collection of distributions. The loss function can be generalized to be $L : \mathcal{P} \times \mathcal{A} \to \mathbb{R}$, i.e., our loss is $L(P, a) \in \mathbb{R}$ for a distribution $P \in \mathcal{P}$ and an action $a \in \mathcal{A}$. For models indexed by a parameter θ, the above notation just reduces back to our original definition. The risk function of a decision rule δ is then $R(P, \delta) = \mathbb{E}_{X\sim P}(L(P, \delta(X))) = \int L(P, \delta(x))P(dx)$.

Lemma 1.18 *Consider two statistical models (collections of distributions) $\mathcal{P}_1 \subset \mathcal{P}_2$. If δ is a minimax rule under model $\mathcal{P}_1$ and we have*

$$\sup_{P\in\mathcal{P}_1} R(P,\delta) = \sup_{P\in\mathcal{P}_2} R(P,\delta),$$

then δ is also minimax under $\mathcal{P}_2$.

Proof: We prove by contradiction. δ is not minimax on $\mathcal{P}_2$.

This implies that there exist δ^* such that

$$\sup_{P\in\mathcal{P}_2} R(P,\delta^*) < \sup_{P\in\mathcal{P}_2} R(P,\delta).$$

Using the fact that $\mathcal{P}_1 \subset \mathcal{P}_2$, we have

$$\sup_{P\in\mathcal{P}_1} R(P,\delta^*) \le \sup_{P\in\mathcal{P}_2} R(P,\delta^*).$$

Thus, we conclude that

$$\sup_{P\in\mathcal{P}_1} R(P,\delta^*) \le \sup_{P\in\mathcal{P}_2} R(P,\delta^*) < \sup_{P\in\mathcal{P}_2} R(P,\delta) = \sup_{P\in\mathcal{P}_1} R(P,\delta),$$

which contradicts to the fact that δ is minimax on $\mathcal{P}_1$. ∎

Example 1.19 (Sample mean is minimax under Gaussian: unknown variance) *Now we come back to the Gaussian sample mean problem. We already know that the sample mean $\bar{X}_n$ is minimax for the model*

$$\mathcal{P}_1 = \{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma = \sigma_0\}$$

for a known σ_0. Moreover, the sample mean achieves the risk

$$\sup_{P \in \mathcal{P}_1} R(P, \delta) = \frac{\sigma_0^2}{n}.$$

Suppose we do not know σ^2 but we have an upper bound about it. This leads to the model

$$\mathcal{P}_2 = \{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma \leq \sigma_0\}.$$

Clearly, $\mathcal{P}_1 \subset \mathcal{P}_2$ and you can easily show that the maximal risk under $\mathcal{P}_2$ is

$$\sup_{P \in \mathcal{P}_2} R(P, \delta) = \frac{\sigma_0^2}{n} = \sup_{P \in \mathcal{P}_1} R(P, \delta).$$

Thus, by Lemma 1.18, $\bar{X}_n$ is also minimax under $\mathcal{P}_2$.

1.5 Admissibility

A rule δ is *inadmissible* if there exists another rule δ' such that $R(\theta, \delta') \leq R(\theta, \delta)$ for all θ and $R(\theta, \delta') < R(\theta, \delta)$ for some θ. When a rule is not inadmissible, it is admissible. Generally, checking the admissibility of a rule is very hard but here are some useful theories for this task.

Theorem 1.20 *Any unique Bayes rule is admissible. Any unique minimax rule is admissible.*

One thing to note here is: the uniqueness is for almost surely every x.

Theorem 1.21 (Uniqueness of a Bayes rule) *Consider a prior π and a Bayes rule of it. Define $Q_\pi(A) = \int P(X \in A|\theta)\pi(\theta)d\theta$ to be the marginal probability measure under this prior. Assume the followings:*

- *The loss function is square loss or convex in a.*
- *The Bayes risk is finite.*
- *$Q_\pi(A) = 0 \Rightarrow P_\theta(A) = 0$ for any measurable set A and $\theta \in \Theta$ (i.e., Q_π is a dominating measure for any P_θ).*

then the Bayes rule of π is unique, and hence, admissible.

Example 1.22 (Normal-normal example) *Consider again the n IID normal-normal example under square loss. The Bayes rule is the posterior mean, which is*

$$\delta_\pi(\bar{X}_n) = \frac{1/\tau^2}{1/\tau^2 + n/\sigma^2}\mu + \frac{n/\sigma^2}{1/\tau^2 + n/\sigma^2}\bar{X}_n.$$

Clearly, this risk is finite.

Moreover, you can easily see that Q is the dominating measure of any P_θ since they are both Gaussian. So the Bayes rule, the posterior mean, is unique Bayes and hence admissible.

1.5.1 Admissibility of sample mean ($d < 3$)

The previous result shows that the posterior mean under the normal-normal model is admissible. This almost tells us the admissibility of the sample mean. But we are still missing a piece: the result holds only for a fixed τ^2. Unlike minimaxity, admissibility cannot be achieved via a sequence of prior distributions (Section 1.4.1). So the question remains: *is the sample mean admissible?*

Theorem 1.23 *If $X_1, \cdots, X_n \sim N(\theta, \sigma^2)$ and σ^2 is known and $\theta \in \Theta = \mathbb{R}$, then $\bar{X}_n$ is an admissible estimator of θ under square loss.*

Proof: We will prove this by contradiction. The highlight idea is: while we cannot apply the minimax theory from Theorem 1.16 to admissibility, the fact that sample mean can be viewed as a limiting case of a sequence of Bayes rules is useful.

Without loss of generality, we just assume $\sigma^2 = 1$. Clearly, the risk of $\bar{X}_n$ is

$$R(\theta, \bar{X}_n) = \frac{1}{n}.$$

Assume that $\bar{X}_n$ is not admissible, so there exists another rule δ^* such that

$$R(\theta, \delta^*) \leq \frac{1}{n} - \epsilon$$

for some $\epsilon > 0$ and any $\theta \in [\theta_0, \theta_1]$, where $\theta_0 < \theta_1$ are some fixed quantity. The existence of this interval is due to the continuity of the risk function under the normal model.

Consider a prior π on θ as $\theta \in N(0, \tau^2)$. We already know that the Bayes rule under this prior δ_π is the posterior mean and it has a Bayes risk

$$r_\tau \equiv \int_\Theta R(\theta, \delta_\pi)\pi(\theta)d\theta = \frac{1}{1/\tau^2 + n}.$$

Now we consider the risk of δ^* under such prior:

$$r^* \equiv \int_\Theta \pi(\theta)d\theta$$

The fact that δ_π is the Bayes rule implies that

$$r_\tau \leq r^*. \tag{1.4}$$

Our strategy is very simple: we will show that $r^* < r_\tau$ when τ is sufficiently large, which contradicts to the fact that δ_π is the Bayes rule, i.e., equation (1.4) will be violated.

Since both r_τ, r^* are related to $\frac{1}{n}$, the risk of the sample mean, we consider the following ratio:

$$\begin{aligned}
\frac{\frac{1}{n} - r^*}{\frac{1}{n} - r_\tau} &= \frac{\frac{1}{\sqrt{2\pi\tau^2}} \int (\frac{1}{n} - R(\theta, \delta^*)) e^{-\frac{1}{2\tau^2}\theta^2} d\theta}{\frac{1}{n} - \frac{1}{1/\tau^2 + n}} \\
&= \frac{n(1 + n\tau^2)}{\sqrt{2\pi\tau^2}} \int \left(\frac{1}{n} - R(\theta, \delta^*)\right) e^{-\frac{1}{2\tau^2}\theta^2} d\theta \\
&\geq \frac{n(1 + n\tau^2)}{\sqrt{2\pi\tau^2}} \int \epsilon e^{-\frac{1}{2\tau^2}\theta^2} d\theta \\
&\geq \frac{n(1 + n\tau^2)}{\sqrt{2\pi\tau^2}} \epsilon \int_{\theta_0}^{\theta_1} e^{-\frac{1}{2\tau^2}\theta^2} d\theta.
\end{aligned}$$

When τ is sufficiently large, $e^{-\frac{1}{2\tau^2}\theta^2} \geq \frac{1}{2}$ for any $\theta \in [\theta_0, \theta_1]$.

Therefore, the above inequality becomes

$$\frac{\frac{1}{n} - r^*}{\frac{1}{n} - r_\tau} \geq \frac{n(1+n\tau^2)}{\sqrt{2\pi\tau^2}}\epsilon \int_{\theta_0}^{\theta_1} \frac{1}{2} d\theta = \frac{n(1+n\tau^2)}{2\sqrt{2\pi\tau^2}}\epsilon(\theta_1 - \theta_0) > 1$$

when τ is sufficiently large (notice that there is τ^2 in the numerator and $\sqrt{\tau^2} = \tau$ in the denominator). As a result, we conclude that

$$\frac{\frac{1}{n} - r^*}{\frac{1}{n} - r_\tau} > 1 \Leftrightarrow \frac{1}{n} - r^* > \frac{1}{n} - r_\tau \Leftrightarrow r^* < r_\tau,$$

which contradicts to equation (1.4) that δ_π is the Bayes estimator under $N(0, \tau^2)$. Therefore, such δ^* does not exist, so the sample mean $\bar{X}_n$ is admissible.

■

Remark. You can also show that the sample mean is admissible when $d = 2$. See page 170 of

> Ferguson, T. S. (2014). Mathematical statistics: A decision theoretic approach (Vol. 1). Academic press.

1.5.2 Inadmissibility of sample mean ($d \geq 3$)

We will now discuss the perhaps most surprising results in Statistics:

> *sample mean is NOT admissible when $d \geq 3$ (under square loss).*

For a multivariate $\theta \in \mathbb{R}^d$ and a multivariate estimator/action $\hat{\theta} \in \mathbb{R}^d$, we define the square loss to be

$$L(\theta, \hat{\theta}) = \sum_{j=1}^{d} (\theta_j - \hat{\theta}_j)^2 = \|\theta - \hat{\theta}\|^2.$$

Theorem 1.24 (Stein's theorem) *The sample mean $\bar{X}_n$ is inadmissible under square loss when $d \geq 3$.*

To prove Theorem 1.24, it suffices to provide an estimator that is better than the sample mean in the sense that the risk is not worse than sample mean and there exists some parameter θ that this estimator will be better than the sample mean.

To simplify the problem, we consider a simple Gaussian model at $n = 1$ and $\sigma^2 = 1$:

$$X \sim N(\theta, \mathbf{I}_d),$$

where $\mathbf{I}_d$ is the $d \times d$ identity matrix and $X, \theta \in \mathbb{R}^d$. Our goal is to estimate θ. The sample mean in this scenario corresponds to X.

Here we introduce the fast *James-Stein shrinkage estimator (JS estimator)*:

$$\hat{\theta}_{JS} = \begin{cases} \left(1 - \frac{d-2}{\|X\|^2}\right) X, & \text{if } X \neq 0 \\ 0, & \text{if } X = 0. \end{cases} \tag{1.5}$$

It turns out that JS estimator has a smaller risk than the sample mean $\bar{X}_n$ for estimating the population mean!

Remark 1.25 (JS estimator and Stein's paradox) *If our goal is to minimize mean square error, then JS estimator is better than sample mean for estimating the population mean when $d \geq 3$, i.e., when we consider three or more variables. This is a very strange result because each variable may be irrelevant to each other and JS estimator pull their information together and shrink the sample mean toward 0. To see how strange this is, suppose $d = 3$ and X_1 is the number of people coming to UW today, X_2 is the average temperature in Boston three days ago, X_3 is the number of tourists in Italy yesterday. These variables are irrelevant to each other, so when estimating the average of them, we shall not use information across them. So sample mean is suppose to be a good estimator. However, JS estimator pulls information across them (noticing the $\|X\|^2$ in the denominator) and achieves a better performance! This is known as the Stein's paradox/phenomenon.*

Now we show that $\hat{\theta}_{JS}$ has a smaller risk than X, i.e.,

$$R(\theta, \hat{\theta}_{JS}) < R(\theta, X) \equiv \mathbb{E}\left[\|\theta - X\|^2\right].$$

The risk of $\hat{\theta}_{JS}$ can be decomposed as

$$\begin{aligned}
R(\theta, \hat{\theta}_{JS}) &\equiv \mathbb{E}\left[\left\|\left(1 - \frac{d-2}{\|X\|^2}\right) X - \theta\right\|^2\right] \\
&= \mathbb{E}\left[\left\|-\frac{d-2}{\|X\|^2} X + (X-\theta)\right\|^2\right] \\
&= \mathbb{E}\left[\left\|-\frac{d-2}{\|X\|^2} X + (X-\theta)\right\|^2\right] \\
&= \mathbb{E}\left[\frac{(d-2)^2}{\|X\|^2}\right] - 2(d-2)\mathbb{E}\left[\frac{X^T(X-\theta)}{\|X\|^2}\right] + \mathbb{E}\left[\|X-\theta\|^2\right] \\
&= \mathbb{E}\left[\frac{(d-2)^2}{\|X\|^2}\right] - 2(d-2)\mathbb{E}\left[\frac{X^T(X-\theta)}{\|X\|^2}\right] + R(\theta, X).
\end{aligned} \tag{1.6}$$

Thus, all we need is to show that the first two terms are negative, i.e., we need

$$\mathbb{E}\left[\frac{(d-2)^2}{\|X\|^2}\right] < 2(d-2)\mathbb{E}\left[\frac{X^T(X-\theta)}{\|X\|^2}\right]. \tag{1.7}$$

To show equation (1.7), we use the following famous Stein's lemma:

Lemma 1.26 (Stein's Lemma) *Let $Y \sim N(\mu, \sigma^2 \mathbf{I}_d)$ be a multivariate normal random variable. Let $g : \mathbb{R}^d \to \mathbb{R}^d$ be a vector-value multivariate function such that each component*

$$\mathbb{E}\left|\frac{\partial}{\partial y_j} g_j(y)|_{y=Y}\right| < \infty.$$

we have

$$\mathbb{E}[(Y-\mu)^T g(Y)] = \sigma^2 \mathbb{E}[\nabla \cdot g(Y)],$$

where $\nabla \cdot g(y) = \sum_{j=1}^d \frac{\partial}{\partial y_j} g_j(y)$.

Note that this Lemma is a straight forward result using integration by parts under Gaussian model, so we omit its proof.

Now consider $g(y) = \frac{y}{\|y\|^2}$, so $g_j(y) = \frac{y_j}{\|y\|^2}$. Clearly,

$$\nabla \cdot g(y) = \sum_{j=1}^{d} \frac{\partial}{\partial y_j} g_j(y) = \sum_{j=1}^{d} \frac{1}{\|y\|^2} - \frac{2y_j^2}{\|y\|^4} = \frac{d}{\|y\|^2} - \frac{2\|y\|^2}{\|y\|^4} = \frac{(d-2)}{\|y\|^2}.$$

Lemma 1.26 implies that

$$\mathbb{E}\left[\frac{X^T(X-\theta)}{\|X\|^2}\right] = \mathbb{E}\left[(X-\theta)^T g(X)\right] = \mathbb{E}\left[\frac{(d-2)}{\|X\|^2}\right].$$

As a result, we conclude that the righ-hand-side of equation (1.7)

$$2(d-2)\mathbb{E}\left[\frac{X^T(X-\theta)}{\|X\|^2}\right] = 2(d-2)\mathbb{E}\left[\frac{(d-2)}{\|X\|^2}\right] = 2(d-2)^2\mathbb{E}\left[\frac{1}{\|X\|^2}\right] > \mathbb{E}\left[\frac{(d-2)^2}{\|X\|^2}\right]$$

so equation (1.7) holds.

Thus, $R(\theta, \hat{\theta}_{JS}) < R(\theta, X)$ for all θ in the Gaussian model, so the sample mean is NOT admissible.

Note that while JS estimator $\hat{\theta}_{JS}$ dominates the sample mean X, the JS estimator is inadmissible as well. See pages 356-357, and Section 5.7, pages 376-389 in

Lehmann, E. L., & Casella, G. (1998). Theory of point estimation. New York, NY: Springer New York.

Remark 1.27 *Although we have shown that the sample mean is inadmissible under the the square loss when* $d \geq 3$*, it is admissible for each variable under univariate square loss (Theorem 1.23). So admissibility depends on the loss function we are using–an estimator can be admissible under one loss function but not the other.*