6 Statistical Decision Theory

6.1 Basics

Let $\theta \in \Theta$ be an unknown quantity of interest. We will take a decision (or an action) d based on the observed data $x \in \mathcal{X}$, such as $d = \delta(x)$.

- The set $\mathcal{D}$ of all possible decisions is called a decision space.
- The function $\delta(x)$ is called a decision rule.

Example 1. Classification: Consider the problem of predicting $y_i \in \{0, 1\}$.

- The decision space is $\mathcal{D} = \{0, 1\}$ for 0-1 classification.
- The decision space is $\mathcal{D} = [0, 1]$ for probabilistic classification.

Example 2. Estimation: Let $\theta \in \Theta \subseteq \mathbb{R}^p$ be the parameter vector. We are interested in θ. The decision space is $\mathcal{D} = \Theta \subseteq \mathbb{R}^p$.

Definition 1 (Loss function)**.** A loss function $L(\theta, d)$ is any non-negative function $L : \Theta \times \mathcal{D} \to [0, \infty)$.

For example:

$$\begin{aligned} L_2 \text{ loss :} \quad & L(\theta - d) = (\theta - d)^2 \\ L_1 \text{ loss :} \quad & L(\theta - d) = |\theta - d|. \end{aligned}$$

Once we apply the loss function to $\delta(x)$, we should treat $L(\theta, \delta(x))$ as a realization from the random variable $L(\theta, \delta(X))$.

Definition 2 (Risk and Posterior Risk)**.** The (frequentist) risk is

$$R(\theta, \delta) = \mathrm{E}[L(\theta, \delta(X)) \mid \theta].$$

The posterior risk is $\mathrm{E}[L(\theta, \delta) \mid X = x]$.

Example 3. Let $X = \begin{bmatrix} X_1 & \cdots & X_n \end{bmatrix}^T$ be a vector of iid random variables from Bernoulli(p). We are interested in p.

- The sample space is $\mathcal{X} = [0, 1]$. The parameter space is $\Theta = [0, 1]$. The decision space is $\mathcal{D} = [0, 1]$.
- If we choose the loss function $L(\theta - d) = (\theta - d)^2$ and decision rule $\delta(X) = \bar{X}$, the frequentist risk is

  $$R(\theta, \delta) = \mathrm{E}[L(p, \delta(X)) \mid p] = \mathrm{E}\left[\left(p - \bar{X}\right)^2 \mid p\right] = \frac{p(1-p)}{n},$$

  where $\theta = p$ is treated as a fixed quantity when evaluating expectation.
- If the prior of p is $p \sim \mathrm{Beta}(a_0, b_0)$, then the posterior is

  $$p \mid x \sim \mathrm{Beta}\left(a_0 + \sum_{i=1}^n x_i, b_0 + n - \sum_{i=1}^n x_i\right).$$

  The posterior risk is

  $$\mathrm{E}[L(p, \delta) \mid X = x] = \mathrm{E}\left[\left(p - \bar{X}\right)^2 \mid X = x\right].$$

Definition 3 (Integrated Risk)**.** The integrated risk is the expectation of the risk with respect to the prior $\Lambda(\theta)$, given by

$$\mathrm{E}\left[L(\theta,\delta)\right] = \int R(\theta,\delta)\,d\Lambda(\theta) = \int \mathrm{E}\left[L(\theta,\delta(X)) \mid \theta\right] d\Lambda(\theta).$$

The decision that minimizes the integrated risk is called the Bayes decision rule (or Bayes estimator). The minimal integrated risk

$$\inf_{\delta} \mathrm{E}\left[L(\theta,\delta)\right]$$

is called the Bayes risk.

Theorem 1 (Find Bayes decision rule via posterior risk)**.** *Suppose that*

1. *there exists a decision rule with finite risk,*
2. *for almost all x, there exists a $\delta(x)$ minimizing the posterior risk $E[L(\theta,\delta) \mid X = x]$.*

Then, $\delta(x)$ is a Bayes decision rule.

Proof. Let a be any decision rule with finite risk (existence by Assumption 1). Then, $\mathrm{E}\left[L(\theta,a(X)) \mid X = x\right]$ is finite almost everywhere. Then, by Assumption 2,

$$\begin{aligned} \mathrm{E}\left[L(\theta,a(X)) \mid X = x\right] &\geq \mathrm{E}\left[L(\theta,\delta(X)) \mid X = x\right] \\ &\Downarrow \text{ Law of total expectation} \\ \mathrm{E}\left[L(\theta,a(X))\right] &\geq \mathrm{E}\left[L(\theta,\delta(X))\right], \end{aligned}$$

which means that $\delta(X)$ is Bayes. □

Theorem 2. *Suppose that there exists a decision rule with finite risk.*

1. *Consider the weighted L_2 loss*

$$L_W(\theta,d) = (\theta - d)^T W (\theta - d).$$

   Then, the Bayes decision rule is the posterior mean $E[\theta \mid X = x]$, where W does not depend on θ.

2. *Consider the absolute error loss*

$$L(\theta,d) = |\theta - d|.$$

   Then, the Bayes decision rule is the posterior median.

Example 4. Consider the L_2 loss. Find the Bayes estimator.

1. Let X_1, ..., X_n be an iid sample from Bernoulli (θ). Suppose that $\theta \sim \mathrm{Beta}(a,b)$. Then, the posterior is proportional to

$$\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \propto \theta^{a+\sum_{i=1}^n x_i - 1}(1-\theta)^{b+n-\sum_{i=1}^n x_i - 1},$$

   a Beta distribution Beta $(a + \sum_{i=1}^n x_i, b + n - \sum_{i=1}^n x_i)$. The posterior mean is

$$\delta = \frac{\alpha}{\alpha+\beta} = \frac{a + \sum_{i=1}^n x_i}{a+b+n}.$$

   If a decision rule has a finite risk, then it is the Bayes rule by the theorem. Consider the decision rule $\delta(x) = \bar{X}$. It has finite risk since

$$\mathrm{E}\left[L(\theta,\delta) \mid \theta\right] = \mathrm{E}\left[\left(\bar{X}-\theta\right)^2 \mid \theta\right] = \frac{\theta(1-\theta)}{n},$$

   which is finite for any θ.

2. Let X_1, ..., X_n be an iid sample from $N(\theta, 1)$. Suppose that $\theta \sim N(\mu_0, \sigma_0^2)$. The posterior is

$$\theta \mid x \quad \sim \quad N\left(\frac{\sigma_0^2 \sum_{i=1}^n x_i + \mu_0}{n\sigma_0^2 + 1}, \frac{\sigma_0^2}{n\sigma_0^2 + 1}\right).$$

The Bayes rule under the L_2 loss is $\frac{\sigma_0^2 \sum_{i=1}^n x_i + \mu_0}{n\sigma_0^2+1}$. We only need to find an estimator with finite risk. Consider just $\bar{X}$ such that $\mathrm{E}\left[\left(\bar{X} - \theta\right)^2 \mid \theta\right] = \theta/n$.

6.2 Point Estimation

We want our estimator to have a small frequentist risk.

Theorem 3 (Rao-Blackwell Theorem)**.** *Let T be a sufficient statistic for $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Let δ be an estimator of $g(\theta)$. Define $\eta(T) = E[\delta(X) \mid T]$. If $R(g(\theta), \delta) < \infty$, and $L(\theta, \cdot)$ is convex for all θ, then $R(g(\theta), \eta(T)) \leq R(g(\theta), \delta)$.*

The Rao-Blackwell Theorem in the Estimation section is a special case of the above Rao-Blackwell theorem, where we only consider unbiased estimators, the loss is the L_2 loss, and the frequentist risk is the variance.

Theorem 4 (Lehmann-Scheffé Theorem)**.** *Let T be a complete and sufficient statistic for a parameter θ. Let $\delta(X)$ be any unbiased estimator of $g(\theta)$. Then $\eta(T) = E[\delta(X) \mid T]$ is the unique unbiased of $g(\theta)$ that minimizes the frequentist risk $R(g(\theta), d)$, if $L(\theta, \cdot)$ is convex for all θ.*

Example 5. Consider X_1, ..., X_n from Bernoulli(θ). Note that

$$p(X \mid \theta) \quad = \quad \prod_{i=1}^n \theta^{X_i} (1-\theta)^{1-X_i} = \exp\left\{\sum_i X_i \log\left(\frac{\theta}{1-\theta}\right) + n\log(1-\theta)\right\}$$

Hence, $T = \sum_i X_i$ is sufficient and complete. Note that $\mathrm{E}\left[n^{-1}T \mid \theta\right] = \theta$. Hence, $\bar{X}$ is the unique unbiased of θ that minimizes any convex loss function.

In practice, we usually cannot compute the closed form expression of $\mathrm{E}[L(\theta, \delta(X)) \mid \theta]$. In supervised learning, we want to learn a function $h : x \to y$ from the data $\{(x_i, y_i), i = 1, ..., n\}$. The corresponding frequentist risk is $\mathrm{E}[L(Y, h(X)) \mid \theta]$. Hence we often minimize the empirical risk to estimate θ:

$$\hat{\theta} \quad = \quad \arg\inf_\theta \frac{1}{n} \sum_{i=1}^n L(y_i, h(x_i)).$$

Example 6. Some examples are

$$\begin{aligned}
&\arg\inf_\theta \frac{1}{n}\sum_{i=1}^n \left[\log q(y_i) - \log p(y_i \mid \theta(x_i))\right] = \arg\inf_\theta \frac{1}{n}\sum_{i=1}^n \log\left(\frac{q(y_i)}{p(y_i \mid \theta(x_i))}\right) \\
&\arg\inf_\theta \frac{1}{n}\sum_{i=1}^n (y_i - \theta_1 - \theta_2 x_i)^2 \\
&\arg\inf_\theta \frac{1}{n}\sum_{i=1}^n (y_i - \theta_1 - \theta_2 x_i)\left[\tau - 1(y_i - \theta_1 - \theta_2 x_i < 0)\right], \quad \text{known } \tau, \\
&\arg\inf_\theta \frac{1}{n}\sum_{i=1}^n \max\{0,\ 1 - y_i(\theta_1 + \theta_2 x_i)\}.
\end{aligned}$$

6.3 Admissible Estimator and Minimax

Definition 4 (Admissible Estimator)**.** A decision rule δ_0 is called inadmissible if there exits a decision rule δ_1 such that

$$\begin{aligned} R(\theta, \delta_0) &\geq R(\theta, \delta_1), \text{ for all } \theta \in \Theta, \\ R(\theta, \delta_0) &> R(\theta, \delta_1), \text{ for some } \theta \in \Theta. \end{aligned}$$

We say that δ_1 dominates δ_0. Otherwise, the decision rule δ_0 is called admissible.

Example 7. Let X_1, ..., X_n be independent random variables where $X_i \sim N(\theta_i, 1)$. The parameter is $\theta = \begin{bmatrix} \theta_1 & \cdots & \theta_n \end{bmatrix}^T \in \mathbb{R}^n$.

- An unbiased estimator of θ is $\delta_0(X) = X = \begin{bmatrix} X_1 & \cdots & X_n \end{bmatrix}^T$.
- The James-Stein estimator is

$$\delta_1(x) = \left(1 - \frac{n-2}{x^T x}\right) x.$$

- If we consider the L_2 loss, then the difference in the risk satisfies

$$\mathrm{E}[L(\theta, \delta_0(X)) \mid \theta] - \mathrm{E}[L(\theta, \delta_1(X)) \mid \theta] \geq \frac{(n-2)^2}{n-2+\theta^T\theta} > 0,$$

for all θ.

Definition 5 (Minimax)**.** A decision rule is minimax if it minimizes the maximum risk as

$$\inf_{d \in \mathcal{D}} \left[\sup_{\theta \in \Theta} R(\theta, d)\right] = \inf_{d \in \mathcal{D}} \left[\sup_{\theta \in \Theta} \mathrm{E}[L(\theta, d(X)) \mid \theta]\right].$$

Example 8. Suppose $X \mid \theta$ follows a 5-category multinomial distribution and $\theta \in \Theta = \{1, 2, 3\}$ indicates which distribution it is. The candidate distributions are

	x				
θ	1	2	3	4	5
1	0	0.05	0.05	0.8	0.1
2	0.05	0.05	0.8	0.1	0
3	0.9	0.05	0.05	0	0

Suppose that our decision space $\mathcal{D} = \Theta$. Consider

Our decision rule					
	Observed x				
δ	1	2	3	4	5
δ_1	$d = 3$	3	2	2	1
δ_2	3	2	2	1	1
δ_3	1	1	1	1	1

Loss function			
	Decision d		
θ	1	2	3
1	$L(\theta, d) = 0$	0.8	1
2	0.3	0	0.8
3	0.3	0.1	0

The frequentist risk is $R(\theta, \delta) = E[L(\theta, d) \mid \theta]$ as

$$R(\theta, \delta) = E[L(\theta, d) \mid \theta] = \sum_{x=1}^{5} L(\theta, \delta(x)) P(X = x \mid \theta)$$

For example,

$$\begin{aligned} R(\theta_1, \delta_1) &= 1 \cdot 0 + 1 \cdot 0.05 + 0.8 \cdot 0.05 + 0.8 \cdot 0.8 + 0 \cdot 0.1 = 0.73 \\ R(\theta_1, \delta_2) &= 1 \cdot 0 + 0.8 \cdot 0.05 + 0.8 \cdot 0.05 + 0 \cdot 0.8 + 0 \cdot 0.1 = 0.08 \end{aligned}$$

Hence, the risk matrix is

	δ		
θ	1	2	3
1	0.73	0.08	0
2	0.08	0.07	0.3
3	0.005	0.01	0.3

The maximum risk $\sup_{\theta\in\Theta} R(\theta, d)$ is attained as

	δ		
	1	2	3
$\sup_{\theta\in\Theta} R(\theta, d)$	0.73 $(\theta = 1)$	0.08 $(\theta = 1)$	0.3 $(\theta = 2, 3)$

The minmiax decision rule is δ_2.

Theorem 5 (Relation between minimax rule and admissible rule). *1. If there exists a unique minimax decision rule, then it is also admissible.*

2. If δ is admissible and has constant risk, then δ is minimax.

3. Suppose that $\mathcal{D}$ is convex and, for all $\theta \in \Theta$, the loss function $L(\theta, \cdot)$ is strictly convex. If δ_0 is admissible and has constant risk, then δ_0 is unique minimax.

Proof. We only prove Part I and Part II of the theorem. □

1. Minimax $\Rightarrow$ admissible: Let δ^* be the minimax decision rule. Suppose that it is not admissible. Then, there exists another decision rule δ such that

$$\begin{aligned} R(\theta, \delta^*) &\geq R(\theta, \delta), \text{ for all } \theta \in \Theta, \\ R(\theta_0, \delta^*) &> R(\theta_0, \delta), \text{ for a } \theta_0 \in \Theta. \end{aligned}$$

This implies that

$$\sup_{\theta\in\Theta} R(\theta, \delta^*) \geq \sup_{\theta\in\Theta} R(\theta, \delta).$$

Since δ^* is minimax, we should have

$$\inf_{d\in\mathcal{D}} \left[\sup_{\theta\in\Theta} R(\theta, d)\right] = \sup_{\theta\in\Theta} R(\theta, \delta^*) \overset{\text{above result}}{\geq} \sup_{\theta\in\Theta} R(\theta, \delta).$$

The only possible way for this to happen is that δ is also minimax, since the LHS is the minimal and should be $\leq$RHS. This contradicts the assumption that δ^* is the unique minimax decision rule.

2. Admissible $\Rightarrow$ Minimax: From 2), we already know that δ_0 satisfying the assumptions must be minimax. We only need to show that it is unique. Suppose that δ_0 is not unique minimax, that is, we can find a $\delta_1 \neq \delta_0$ such that δ_1 is also minimax as

$$\sup_{\theta\in\Theta} R(\theta, \delta_1) = \sup_{\theta\in\Theta} R(\theta, \delta_0) \underset{R(\theta,\delta_0) \text{ is constant}}{=} R(\theta_0, \delta_0) \text{ for any } \theta_0 \in \Theta.$$

Thus,

$$R(\theta_0, \delta_1) \leq \sup_{\theta\in\Theta} R(\theta, \delta_1) = R(\theta_0, \delta_0) \text{ for any } \theta_0 \in \Theta.$$

First consider the case where the equality holds: $R(\theta_0, \delta_1) = R(\theta_0, \delta_0)$. We define a new decision rule

$$\delta_2 = \frac{\delta_1 + \delta_0}{2}.$$

Such $\delta_2 \in \mathcal{D}$ if we assume $\mathcal{D}$ is convex. Thus,

$$\begin{aligned}
0 \leq R(\theta_0, \delta_2) &= \mathrm{E}\left[L(\theta_0, \delta_2(x)) \mid \theta_0\right] \\
L \text{ is strictly convex} \quad &< \mathrm{E}\left[\frac{1}{2}L(\theta_0, \delta_0(x)) + \frac{1}{2}L(\theta_0, \delta_1(x)) \mid \theta_0\right] \\
&= \frac{1}{2}R(\theta_0, \delta_0) + \frac{1}{2}R(\theta_0, \delta_1) \\
&= R(\theta_0, \delta_0)
\end{aligned}$$

This means that $R(\theta_0, \delta_2) < R(\theta_0, \delta_0)$, for any $\theta_0 \in \Theta$, which contradicts the assumption δ_0 is admissible. Hence, we must have

$$R(\theta_0, \delta_1) \quad < \quad \sup_{\theta \in \Theta} R(\theta, \delta_1) = R(\theta_0, \delta_0) \text{ for any } \theta_0 \in \Theta.$$

But this also contradicts the assumption δ_0 is admissible. Thus, we cannot find such δ_1.

6.4 Why Bayesian Statistics?

Theorem 6. *The Bayes decision rule is admissible if either set of the following conditions hold.*

1. $\lambda(\theta) > 0$ *for all* $\theta \in \Theta$, $R(\theta, \delta)$ *is continuous in* θ *for all* δ, *and*

$$\inf_{\delta \in \mathcal{D}} \int R(\theta, \delta)\, d\Lambda(\theta) \quad < \quad \infty.$$

2. *The Bayes decision rule is unique.*

3. $\mathcal{D}$ *is convex, the loss function* $L(\theta, \cdot)$ *is strictly convex for all* $\theta \in \Theta$, *and*

$$\inf_{\delta \in \mathcal{D}} \int R(\theta, \delta)\, d\Lambda(\theta) \quad < \quad \infty.$$

Proof. We only prove the first set of conditions. Condition set 1: Suppose that the Bayes rule δ_B is not admissible. Then there exists a δ_1 such that

$$\begin{aligned}
R(\theta, \delta_B) &\geq R(\theta, \delta_1), \text{ for all } \theta \in \Theta, \\
R(\theta_1, \delta_B) &> R(\theta_1, \delta_1), \text{ for some } \theta_1 \in \Theta.
\end{aligned}$$

Because $R(\theta, \delta)$ is continuous in θ for all δ, then there exists a neighborhood C of θ_1 such that

$$\begin{aligned}
R(\theta_1, \delta_B) &> R(\theta_1, \delta_1), \text{ for all } \theta \in C \subset \Theta, \\
\text{and} \quad \int_{\theta \in C} R(\theta, \delta_B)\, d\Lambda(\theta) &> \int_{\theta \in C} R(\theta, \delta_1)\, d\Lambda(\theta).
\end{aligned}$$

For $\theta \in C^c$, we should have

$$\int_{\theta \in C^c} R(\theta, \delta_B)\, d\Lambda(\theta) \quad \geq \quad \int_{\theta \in C^c} R(\theta, \delta_1)\, d\Lambda(\theta).$$

Hence,

$$\begin{aligned}
\int R(\theta, \delta_1)\, d\Lambda(\theta) &= \int_{\theta \in C} R(\theta, \delta_1)\, d\Lambda(\theta) + \int_{\theta \in C^c} R(\theta, \delta_1)\, d\Lambda(\theta) \\
&< \int_{\theta \in C} R(\theta, \delta_B)\, d\Lambda(\theta) + \int_{\theta \in C^c} R(\theta, \delta_B)\, d\Lambda(\theta) \\
&= \int R(\theta, \delta_B)\, d\Lambda(\theta) < \infty,
\end{aligned}$$

where the last inequality holds since $\inf_{\delta \in \mathcal{D}} \int R(\theta, \delta)\, d\Lambda(\theta) < \infty$. This contradicts the fact that δ_B is Bayes. $\square$

Theorem 7 (Blyth Theorem)**.** *Let Θ be an open set. Suppose that the set of decision rules with continuous $R(\theta, d)$ in θ forms a class $\mathcal{C}$ such that for any $d' \notin \mathcal{C}$ we can find a $d \in \mathcal{C}$ such that d dominates d'. Let δ be an estimator such that $R(\theta, \delta)$ is continuous of θ. Let $\{\Lambda_n\}$ be a sequence of priors such that*

1. *$\int R(\theta, \delta)\, d\Lambda_n(\theta) < \infty$ for all n,*
2. *for every nonemptry open set $\Theta_0 \in \Theta$, there exist constants $B > 0$ and N such that*
$$\int_{\Theta_0} d\Lambda_n(\theta) \quad \geq \quad B, \text{ for all } n \geq N,$$
3. *$\int R(\theta, \delta)\, d\Lambda_n(\theta) - \int R(\theta, \delta_n)\, d\Lambda_n(\theta) \to 0$ as $n \to \infty$, where δ_n is the Bayes rule under the prior Λ_n.*

Then, δ is admissible.

We have shown that the Bayes decision rule is admissible under some assumption. The Blyth theorem says that the admissible decision can be obtained such that

$$\lim_{n \to \infty} \int R(\theta, \delta)\, d\Lambda_n(\theta) - \int R(\theta, \delta_n)\, d\Lambda_n(\theta) \quad = \quad 0.$$

We can in fact claim that every admissible estimator is either a Bayes estimator or a limit of Bayes estimators as

$$\lim_{n \to \infty} \delta_n(x) \quad = \quad \delta_B(x), \text{ almost everywhere,}$$

under quite mild assumptions (e.g., $f(x \mid \theta) > 0$ for any $(x, \theta) \in \mathcal{X} \times \Theta$, $L(\theta, d)$ is continuous and strictly convex in d for every θ, among others). See Lehmann Theory of Point estimation Theorem 5.7.15 or Bayesian Choice Theorem 8.3.9.

Definition 6. A prior distribution Λ is least favorable if

$$\int R(\theta, \delta_B(\Lambda))\, d\Lambda(\theta) \quad \geq \quad \int R(\theta, \delta_B(\Lambda'))\, d\Lambda'(\theta)$$

for all prior distributions Λ'.

Theorem 8. *Let δ_B be the Bayes decision rule with respect to the prior $\pi(\theta)$. Suppose that*

$$\int R(\theta, \delta_B)\, d\Lambda(\theta) \quad = \quad \sup_\theta R(\theta, \delta_B).$$

Then, δ_B is minimax and $\pi(\theta)$ is least favorable. Further, if δ_B is the unique Bayes decision rule with respect to the prior $\pi(\theta)$, then it is the unique minimax estimator.

Proof. We only prove the minimax part. The assumption $\int R(\theta, \delta_B)\, d\Lambda(\theta) = \sup_\theta R(\theta, \delta_B)$ means that the minimum integrated risk equals to the maximum of the frequentist risk. Let δ be any other decision rule. Then

$$\begin{aligned} \sup_\theta R(\theta, \delta) = \int \left[\sup_\theta R(\theta, \delta)\right] d\Lambda(\theta) \quad &\geq \quad \int R(\theta, \delta)\, d\Lambda(\theta) \\ \text{definition of Bayes rule} \quad &\geq \quad \int R(\theta, \delta_B)\, d\Lambda(\theta) \\ \text{assumption} \quad &= \quad \sup_\theta R(\theta, \delta_B). \end{aligned} \tag{1}$$

Hence, δ_B is minimax since any other δ leads to $\sup_\theta R(\theta, \delta) \geq \sup_\theta R(\theta, \delta_B)$. $\square$

Corollary 1. *Let δ_B be the Bayes decision rule with respect to the proper prior $\Lambda(\theta)$. If δ_B has constant (frequentist) risk, then it is minimax.*

Proof. Since δ_B has constant frequentist risk (e.g., $R(\theta, \delta_B) = c$), then we trivially have

$$\begin{aligned} \int R(\theta, \delta_B)\, d\Lambda(\theta) &= c \int d\Lambda(\theta) = c, \text{ we need } \Lambda \text{ to be a proper prior.} \\ \sup_\theta R(\theta, \delta_B) &= c. \end{aligned}$$

Hence, the condition of the theorem (Bayes is minimax) is satisfied. The theorem means that δ_B is minimax. □

Example 9 (Minimax Estimator of Binomial Proportion)**.** Let $X_1, ..., X_n$ be an iid sample from Bernoulli (θ). Suppose that $\theta \sim$ Beta (a, b). Then, the posterior is Beta $(a + \sum_{i=1}^n x_i, b + n - \sum_{i=1}^n x_i)$. The Bayes estimator is the posterior mean as

$$\delta_B = \frac{a + \sum_{i=1}^n x_i}{a + b + n}$$

Its risk is

$$R(\theta, \delta_B) = E\left[\left(\frac{a + \sum_{i=1}^n x_i}{a + b + n} - \theta\right)^2 \mid \theta\right] = \frac{\left[(a+b)^2 - n\right]\theta^2 + [n - 2a(a+b)]\, p + a^2}{(a+b+n)^2}.$$

The numerator is a polynomial in θ. It is a constant if $(a+b)^2 = n$ and $n = 2a(a+b)$. In such a case,

$$R(\theta, \delta_B) = \frac{a^2}{(a+b+n)^2} \text{ is a constant.}$$

Hence, the Bayes decision rule is minimax. The solutions of a and b are $a = \sqrt{n}/2$ and $b = \sqrt{n}/2$.

Theorem 9. *Let $\{\Lambda_m\}$ be a sequence of proper prior distributions, and δ_m be the Bayes decision rule corresponding to the prior Λ_m. If δ is an estimator such that*

$$\sup_\theta R(\theta, \delta) = \lim_{m\to\infty} \int R(\theta, \delta_m)\, d\Lambda_m(\theta).$$

Then δ is minimax.

Proof. Suppose that d is any other decision rule. Then,

$$\begin{aligned} \sup_\theta R(\theta, d) &= \int \sup_\theta R(\theta, d)\, d\Lambda_m(\theta) \text{ we need proper priors here} \\ &\geq \int R(\theta, d)\, d\Lambda_m(\theta) \\ &\Downarrow \\ \sup_\theta R(\theta, d) &\geq \lim_{m\to\infty} \int R(\theta, d)\, d\Lambda_m(\theta). \end{aligned}$$

By the assumption of the theorem, we have

$$\begin{aligned} \sup_\theta R(\theta, \delta) &= \lim_{m\to\infty} \int R(\theta, \delta_m)\, d\Lambda_m(\theta) \\ \text{definition of Bayes rule} &\leq \lim_{m\to\infty} \int R(\theta, d)\, d\Lambda_m(\theta) \end{aligned}$$

Hence, we have

$$\sup_{\theta} R(\theta, \delta) \leq \lim_{m\to\infty} \int R(\theta, d)\, d\Lambda_m(\theta) \leq \sup_{\theta} R(\theta, d),$$

which means that δ is minimax. $\square$

Example 10 (Minimax for Normal Mean)**.** Let $X_1, ..., X_n$ be iid observations from $N\left(\theta, \sigma^2\right)$, where σ^2 is known. Consider the L_2 loss $L(\theta, d) = (\theta - d)^2$. The posterior is $\theta \mid x \sim N\left(\frac{\tau^2 \sum_{i=1}^n x_i + \sigma^2 \mu_0}{n\tau_m^2 + \sigma^2}, \quad \frac{\sigma^2 \tau_m^2}{n\tau_m^2 + \sigma^2}\right)$. Let $\delta(x) = \frac{\tau_m^2 \sum_{i=1}^n x_i + \sigma^2 \mu_0}{n\tau_m^2 + \sigma^2}$. Then

$$\mathrm{E}(\theta - \delta)^2 = \mathrm{E}_X\left\{\mathrm{E}_{\theta|X}\left[(\theta - \delta)^2 \mid X\right]\right\} = \mathrm{E}_X\left\{\frac{\sigma^2\tau_m^2}{n\tau_m^2 + \sigma^2}\right\} = \frac{\sigma^2\tau_m^2}{n\tau_m^2 + \sigma^2}.$$

If $\tau_m^2 \to \infty$ as $m \to \infty$, then $\mathrm{E}\left(\theta - \hat{\theta}\right)^2 \to \sigma^2/n$. By the theorem, $\bar{X}$ is minimax, since

$$R\left(\theta, \bar{X}\right) = \int \left(\theta - \bar{X}\right)^2 N\left(\theta, \sigma^2/n\right) d\bar{x} = \frac{\sigma^2}{n}.$$

Let $m(x; \Lambda)$ be the marginal likelihood of x under the prior $\Lambda(\theta)$. We define the frequentist risk between $p(x \mid \theta)$ and $m(x; \Lambda)$ as

$$R_n(\theta, \Lambda) = \mathrm{KL}\left(p(x \mid \theta), m(x; \Lambda)\right) = \int p(x \mid \theta) \log\left[\frac{p(x \mid \theta)}{m(x; \Lambda)}\right] d\mu(x).$$

The integrated risk is then

$$R_n(\Lambda) = \int R_n(\theta, \Lambda)\, d\Lambda(\theta) = \mathrm{E}\left[\mathrm{KL}\left(\pi(\theta \mid x), \pi(\theta)\right)\right],$$

which is the same as the mutual information of X and θ, and the expected Kullback-Leiber divergence.

Remark 1. Suppose that some regularity conditions are satisfied, including Θ is a compact set, the Fisher information equals to the negative expected Hessian, among others.

- It has been proved that, among all positive and continuous priors,

$$\sup_{\pi} R_n(\Lambda) - \inf_{p(x)} \sup_{\theta\in\Theta} \mathrm{KL}\left(p(x \mid \theta), p(x)\right) \to 0.$$

- It has also been proved that the Jeffreys prior $\lambda^*(\theta)$ is the unique continuous and positive prior such that

$$\sup_{\pi} R_n(\Lambda) - R_n(\lambda^*) \to 0.$$

Hence, asymptotically, Jeffreys prior maximizes the mutual information, is the least favorable prior, and the integrated risk equals to the minimax risk.