Chapter 5
Bayes Methods and Elementary Decision Theory

1. **Elementary Decision Theory**
2. **Structure of the risk body: the finite case**
3. **The finite case: relations between Bayes minimax, admissibility**
4. **Posterior distributions**
5. **Finding Bayes rules**
6. **Finding Minimax rules**
7. **Admissibility and Inadmissibility**
8. **Asymptotic theory of Bayes estimators**

Chapter 5

Bayes Methods and Elementary Decision Theory

1 Elementary Decision Theory

Notation 1.1. Let

- $\Theta \equiv$ the states of nature.
- $\mathbf{A} \equiv$ the action space.
- $\mathbf{X} \equiv$ the sample space of a random variable X with distribution P_θ.
- $P : \mathbf{X} \times \boldsymbol{\Theta} \mapsto [0,1]$; $P_\theta(X = x) =$ probability of observing $X = x$ when θ is true.
- $L : \Theta \times \mathbf{A} \mapsto \mathbb{R}^+$; a loss function.
- $d : \mathbf{A} \times \mathbf{X} \to [0,1]$, $d(a,x) = d(a|x) =$ probability of action a when $X = x$ (a decision function).
- $\mathbf{D} \equiv$ {all decision functions}.

The *risk* for the rule $d \in \mathbf{D}$ when $\theta \in \Theta$ is true is

$$\begin{aligned} R(\theta, d) &= E_\theta L(\theta, d(\cdot|X)) \\ &= \int_{\mathbf{X}} \int_{\mathbf{A}} L(\theta, a) d(da|X = x) P_\theta(dx) \\ &= \sum_{i=1}^{k} L(\theta, a_i) \left\{ \sum_{j=1}^{m} d(a_i, x_j) P_\theta(X = x_j) \right\} \qquad \text{in the discrete case.} \end{aligned}$$

We will call $R : \Theta \times \mathbf{D} \to \mathbb{R}^+$ the *risk function*.

A decision rule d is *inadmissible* if there is a rule d' such that $R(\theta, d') \le R(\theta, d)$ for all θ and $R(\theta, d') < R(\theta, d)$ for some θ. A decision rule d is *admissible* if it is not inadmissible.

Example 1.1 Hypothesis testing. $\mathbf{A} = \{0, 1\}$, $\Theta = \Theta_0 \cup \Theta_1$,

$$L(\theta, 0) = l_0 1_{\Theta_1}(\theta), \qquad L(\theta, 1) = l_1 1_{\Theta_0}(\theta).$$

Since **A** contains just two points, $d(1|x) = 1 - d(0|x) \equiv \phi(x)$; and then

$$\begin{aligned} P_\theta(\text{accept } H_0) &= E_\theta d(0|X) = E_\theta(1-\phi(X)), \\ P_\theta(\text{reject } H_0) &= E_\theta d(1|X) = E_\theta \phi(X), \end{aligned}$$

and

$$R(\theta, d) = l_1 E_\theta \phi(X) 1_{\Theta_0}(\theta) + l_0 E_\theta (1-\phi(X)) 1_{\Theta_1}(\theta).$$

The classical Neyman - Pearson hypothesis testing philosophy bounds $l_1\alpha \equiv \sup_{\theta\in\Theta_0} R(\theta,d)$ and tries to minimize

$$l_0(1-\beta_d(\theta)) \equiv R(\theta, d)$$

for each $\theta \in \Theta_1$.

Example 1.2 Estimation. $\mathbf{A} = \Theta$; typically $\Theta = \mathbb{R}^s$ for some s (but sometimes it is useful to take Θ to be a more general metric space (Θ, d)). A typical loss function (often used for mathematical simplicity more than anything else) is

$$L(\theta, a) = K|\theta - a|^2 \qquad \text{for some } \; K.$$

Then the risk of a rule d is:

$$\begin{aligned} R(\theta, d) &= KE_\theta|\theta - d(\cdot|X)|^2 \\ &= KE_\theta|\theta - d(X)|^2 \qquad \text{for a non -randomized rule } \; d \\ &= K\{Var_\theta(d(X)) + [\text{bias}_\theta(d(X))]^2\} \qquad \text{when } \; s = 1. \end{aligned}$$

Definition 1.1 A decision rule d_M is *minimax* if

$$\inf_{d\in\mathbf{D}} \sup_{\theta\in\Theta} R(\theta, d) = \sup_{\theta\in\Theta} R(\theta, d_M).$$

Definition 1.2 A probability distribution Λ over Θ is called a *prior distribution*.

Definition 1.3 For any given prior Λ and $d \in \mathbf{D}$, the *Bayes risk of d with respect to Λ* is

$$\begin{aligned} \mathcal{R}(\Lambda, d) &= \int_\Theta R(\theta, d) d\Lambda(\theta) \\ &= \sum_{i=1}^{l} R(\theta_i, d)\lambda_i \qquad \text{if } \; \Theta = \{\theta_1, \ldots, \theta_l\}. \end{aligned}$$

Definition 1.4 A *Bayes decision rule with respect to* Λ, d_Λ, is any rule satisfying

$$\mathcal{R}(\Lambda, d_\Lambda) = \inf_{d\in\mathbf{D}} \mathcal{R}(\Lambda, d) \equiv \text{Bayes risk}.$$

Example 1.3 $\Theta = \{1, 2\}$.

Urn 1: 10 red balls, 20 blue balls, 70 green balls.

Urn 2: 40 red balls, 40 blue balls, 20 green balls.

One ball is drawn from one of the two urns. Problem: decide which urn the ball came from if the losses $L(\theta, a)$ are given by:

θ/a	1	2
1	0	10
2	6	0

Let $d = (d_R, d_B, d_G)$ with d_x = probability of choosing urn 1 if color $X = x$ is observed. Then

$$R(1, d) \quad = \quad 10P_1(\text{action } 2) = 10\{.1(1 - d_R) + .2(1 - d_B) + .7(1 - d_G)\}$$

and, similarly,

$$R(2, d) = 6\{.4d_R + .4d_B + .2d_G\}.$$

if the prior distribution on the urns is given by $\lambda_1 = \lambda$ and $\lambda_2 = 1 - \lambda$, then the Bayes risk is

$$\mathcal{R}(\Lambda, d) = 10\lambda + (2.4 - 3.4\lambda)d_R + (2.4 - 4.4\lambda)d_B + (1.2 - 8.2\lambda)d_G.$$

This is minimized by choosing $d_x = 1$ if its coefficient is negative, 0 if its coefficient is positive. For example, if $\lambda = 1/2$, then the Bayes risk with respect to λ equals

$$5 + .7d_R + .2d_B - 2.9d_G,$$

which is minimized by $d_R = d_B = 0$, $d_G = 1$; i.e. the Bayes rule d_Λ with respect to $\lambda = 1/2$ is $d_\Lambda = (0, 0, 1)$. Note that the Bayes rule is in fact a non-randomized rule. This gives us the Bayes risk for $\lambda = 1/2$ as $\mathcal{R}(1/2, d_B) = 2.1$.

I claim that the minimax rule is $d_M = (0, 9/22, 1)$, which is a randomization of the two non-random rules $d_2 = (0, 0, 1)$, and $d_7 = (0, 1, 1)$. This is easily confirmed by computing $R(1, d_M) = 240/110 = R(2, d_M)$. Here is a table giving all of the nonrandomized rules and their risks:

X	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8
R	0	0	0	1	1	1	0	1
B	0	0	1	0	1	0	1	1
G	0	1	0	0	0	1	1	1
$R(1, d)$	10	3	8	9	7	2	1	0
$R(2, d)$	0	1.2	2.4	2.4	4.8	3.6	3.6	6

2 Structure of the Risk Body: the finite case

Suppose that

$$\Theta = \{\theta_1, \ldots, \theta_l\}, \qquad \mathcal{X} = \{x_1, \ldots, x_m\}, \qquad \mathbf{A} = \{a_1, \ldots, a_k\}.$$

Definition 2.1 A set $A \subset \mathbb{R}^l$ is *convex* if, for all $\lambda \in [0,1]$, $x, y \in A$, $\lambda x + (1-\lambda) y \in A$.

Lemma 2.1 Let $\lambda = (\lambda_1, \ldots, \lambda_t)$ be a probability distribution (so $\lambda_i \geq 0$ for $i = 1, \ldots, t$, and $\sum_{i=1}^t \lambda_i = 1$). Then for any $d_1, \ldots, d_t \in \mathbf{D}$, $\sum_{i=1}^t \lambda_i d_i \in \mathbf{D}$.

Proof. Set $d(a_l, x_j) = \sum_{i=1}^t \lambda_i d_i(a_l, x_j)$. Then $d(a_l, x_j) \geq 0$ and

$$\sum_{l=1}^k d(a_l, x_j) = \sum_{i=1}^t \lambda_i \sum_{l=1}^k d_i(a_l, x_j) = \sum_{i=1}^t \lambda_i = 1.$$

□

We call

$$\mathcal{R} = \{(R(\theta_1, d), \ldots, R(\theta_l, d)) \in \mathbb{R}^{+l} : \ d \in \mathbf{D}\}$$

the *risk body*.

Theorem 2.1 The risk body is convex.

Proof. Let $d_1, d_2 \in \mathbf{D}$ denote the rules corresponding to $R_i = (R(\theta_1, d_i), \ldots, R(\theta_l, d_i))$, $i = 1, 2$, and set $d = \lambda d_1 + (1-\lambda) d_2 \in \mathbf{D}$ by the lemma. Then

$$\begin{aligned} \lambda R(\theta_l, d_1) + (1-\lambda) R(\theta_l, d_2) &= \sum_{i=1}^k \sum_{j=1}^m L(\theta_l, a_i) \{\lambda d_1(a_i, x_j) + (1-\lambda) d_2(a_i, x_j)\} p_{\theta_l}(x_j) \\ &= R(\theta_l, d), \end{aligned}$$

so $d \in \mathbf{D}$ has risk point $\lambda R_1 + (1-\lambda) R_2 \in \mathcal{R}$. □

Theorem 2.2 Every $d \in \mathbf{D}$ may be expressed as a convex combination of the the nonrandomized rules.

Proof. Any $d \in \mathbf{D}$ may be represented as a $k \times m$ matrix of ≥ 0 real numbers whose columns add to 1. The nonrandomized rules are those whose entries are 0's and 1's. Set $d(a_i, x_j) = d_{ij}$, $d = (d_{ij})$, $1 \leq i \leq k$, $1 \leq j \leq m$.

There are k^m nonrandom decision rules: call them $\delta^{(r)} = (\delta_{ij}^{(r)})$, for $1 \leq r \leq k^m$. Given $d \in \mathcal{D}$, we want to find $\lambda_r \geq 0$ with $\sum_{r=1}^{k^m} \lambda_r = 1$ so that

$$\sum_{r=1}^{k^m} \lambda_r \delta^{(r)} = d.$$

Claim:

$$\lambda_r \equiv \prod_{j=1}^{m} \prod_{i=1}^{k} d_{ij}^{\delta_{ij}^{(r)}} = \prod_{j=1}^{m} \{\sum_{i=1}^{k} d_{ij} \delta_{ij}^{(r)}\}$$

work. (This is easily proved by induction on m.) □

Remark 2.1 The space of decision rules defined above, $\mathbf{D}$, with $d \in \mathbf{D}$ being a probability distribution over actions given that $X = x$, corresponds to the *behavioral decision rules* as discussed by Ferguson (1967), page 25. This differs from Ferguson's collection D^*, the *randomized decision functions*, which are probability distributions over the non-randomized rules.

3 The finite case: relations between Bayes, minimax, and admissibility

This section continues our examination of the special, but illuminating, case of a finite set Θ. In this case we can prove a number of results about Bayes and minimax rules and connections between them which carry over to more general problems with some appropriate reformulation.

Theorem 3.1 A minimax rule always exists.

Proof. If $0 \notin \mathcal{R}$, then for some $c > 0$ the cube with all sides of length c in the positive quadrant does not intersect $\mathcal{R}$. Let c increase until the square intersects $\mathcal{R}$; call this number c_0. Any decision rule with a risk point which intersects the c_0 square is minimax. (Note that the equation for the boundary of a cube is $\max_i R(\theta_i, d) = \text{constant}$.) $\Box$

Theorem 3.2 If $\lambda = (\lambda_1, \ldots, \lambda_l)$ is a prior, then Bayes decision rules have risk points on the hyperplane $\{x \in \mathbb{R}^l : \sum_i \lambda_i x_i = c^*\}$ where

$$c^* = \inf\{c \ge 0 : \text{ the plane determined by } \sum_i \lambda_i x_i = c \text{ intersects } \mathcal{R}\}.$$

Proof. The Bayes risk is $\mathcal{R}(\Lambda, d) = \sum_{i=1}^l \lambda_i R(\theta_i, d)$ is the Bayes risk and we want to minimize it. $\Box$

Lemma 3.1 An admissible rule is a Bayes rule for some prior λ.

Proof. See the picture! $\Box$

Note that *not every* Bayes rule is admissible.

Theorem 3.3 Suppose that d_0 is Bayes with respect to $\lambda = (\lambda_1, \ldots, \lambda_l)$ and $\lambda_i > 0$, $i = 1, \ldots, l$. Then d_0 is admissible.

Proof. Suppose that d_0 is *not* admissible; then there is a rule d better than d_0: i.e.

$$R(\theta_i, d) \le R(\theta_i, d_0) \qquad \text{for all } \ i$$

with $<$ for some i. Since $\lambda_i > 0$ for $i = 1, \ldots, l$,

$$\mathcal{R}(\Lambda, d) = \sum_{i=1}^l \lambda_i R(\theta_i, d) < \sum_{i=1}^l \lambda_i R(\theta_i, d_0) = \mathcal{R}(\Lambda, d_0),$$

contradicting the fact that d_0 is Bayes with respect to λ. $\Box$

Theorem 3.4 If $d_B \in \mathbf{D}$ is Bayes for λ and it has *constant risk*, then d_B is minimax.

Proof. Let r_0 be the constant risk. Assume d_B is *not* minimax and let d_M be the minimax rule (which exists). Then

(a) $$R(\theta_i, d_M) \le \max_{1\le j\le l} R(\theta_j, d_M) < \max_{1\le j\le l} R(\theta_j, d_B) = r_0$$

and

(b) $$\min_{d\in\mathbf{D}} \mathcal{R}(\lambda, d) = \mathcal{R}(\lambda, d_B) = \sum \lambda_i R(\theta_i, d_B) = r_0.$$

But (a) yields

(c) $$\mathcal{R}(\lambda, d_M) = \sum_i \lambda_i R(\theta_i, d_M) < \sum_i \lambda_i r_0 = r_0$$

which contradicts (b). $\square$

Example 3.1 We return to Example 1.3, and consider it from the perspective of Theorem 3.4. When the prior $\underline{\lambda} = (\lambda, 1-\lambda)$ with $\lambda = 6/11$, we find that the Bayes risk is given by

$$\begin{aligned}\mathcal{R}(\Lambda, d) &= 10\lambda + (2.4 - 3.4\lambda)d_R + (2.4 - 4.4\lambda)d_B + (1.2 - 8.2\lambda)d_G \\ &= 10 \cdot \frac{6}{11} + \frac{6}{11}d_R + 0 \cdot d_B - \frac{36}{11}d_G,\end{aligned}$$

so all the rules $d_\Lambda = (0, d_B, 1)$ are Bayes with respect to $\lambda = 6/11$. Can we find one with constant risk? The risks of these Bayes rules d_Λ are given by

$$R(1, d_\Lambda) = 1 + 2(1 - d_B),$$
$$R(2, d_\Lambda) = 0 + 2.4d_B + 1.2,$$

and these are equal if d_B satisfies $3 - 2d_B = 1.2 + 2.4d_B$, and hence $d_B = 9/22$. For this particular Bayes rule, $d_\Lambda = (0, 9/22, 1)$, and the risks are given by $R(1, d_\Lambda) = 24/11 = R(2, d_\Lambda)$. Thus the hypotheses of Theorem 3.4 are satisfied, and we conclude that $d_\Lambda = (0, 9/22, 1)$ is a minimax rule.

4 Posterior Distributions

Now we will examine Bayes rules more generally.

Definition 4.1 If $\boldsymbol{\theta} \sim \Lambda$, a prior distribution over Θ, then the conditional distribution of $\boldsymbol{\theta}$ given X, $P(\boldsymbol{\theta} \in A|X)$, $A \in \mathcal{B}(\Theta)$, is called the *posterior* distribution of θ. Write $\Lambda(\theta|x) = P(\boldsymbol{\theta} \le \theta|X = x)$ for the conditional distribution function if Θ is Euclidean.

If Λ has density λ with respect to ν and P_θ has density p_θ with respect to μ, then

$$\lambda(\theta|x) = \frac{p(x|\theta)\lambda(\theta)}{\int_\Theta p(x|\theta)\lambda(\theta)d\nu(\theta)} = \frac{p(x|\theta)\lambda(\theta)}{p(x)}$$

is the *posterior density* of $\boldsymbol{\theta}$ given $X = x$.

Definition 4.2 Suppose X has density $p(x|\theta)$, and $\lambda = \lambda(\cdot)$ is a prior density, $\lambda \in \mathcal{P}_\Lambda$. If the posterior density $\lambda(\cdot|x)$ has the same form (i.e. $\lambda(\cdot|x) \in \mathcal{P}_\Lambda$ for a.e. x), then λ is said to be a *conjugate prior* for $p(\cdot|\theta)$.

Example 4.1
A. (Poisson - Gamma). Suppose that $(X|\boldsymbol{\theta} = \theta) \sim \text{Poisson}(\theta)$, $\boldsymbol{\theta} \sim \Gamma(\alpha, \beta)$. Then

$$\begin{aligned}
\lambda(\theta|x) &= \frac{p(x|\theta)\lambda(\theta)}{p(x)} \propto p(x|\theta)\lambda(\theta) \\
&= e^{-\theta}\frac{\theta^x}{x!}\frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta} \\
&= \frac{\beta^\alpha}{x!\Gamma(\alpha)}\theta^{\alpha+x-1}e^{-(\beta+1)\theta},
\end{aligned}$$

so $(\boldsymbol{\theta}|X = x) \sim \Gamma(\alpha + x, \beta + 1)$, and gamma is a conjugate prior for Poisson.
B. (Normal mean - normal). (Example 1.3, Lehmann TPE, page 243). Suppose that $(X|\boldsymbol{\theta} = \theta) \sim N(\theta, \sigma^2)$, $\boldsymbol{\theta} \sim N(\mu, \tau^2)$. Then

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

Note that if X is replaced by $X_1, \ldots, X_n$ i.i.d. $N(\theta, \sigma^2)$, then by sufficiency,

$$(\boldsymbol{\theta}|\underline{X} = \underline{x}) \sim N\left(\frac{n/\sigma^2}{1/\tau^2 + n/\sigma^2}\overline{x} + \frac{1/\tau^2}{1/\tau^2 + n/\sigma^2}\mu, , \frac{1}{1/\tau^2 + n/\sigma^2}\right).$$

Example 4.2 If $(X|\boldsymbol{\theta} = \theta) \sim \text{Binomial}(n, \theta)$, $\boldsymbol{\theta} \sim B(\alpha, \beta)$, then

$$\begin{aligned}
\lambda(\theta|x) &= \frac{p(x|\theta)\lambda(\theta)}{p(x)} \propto p(x|\theta)\lambda(\theta) \\
&= \binom{n}{x}\theta^x(1-\theta)^{n-x}\theta^{\alpha-1}(1-\theta)^{\beta-1}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \\
&= \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\binom{n}{x},
\end{aligned}$$

so $(\boldsymbol{\theta}|X = x) \sim \text{Beta}(\alpha + x, \beta + n - x)$.

Example 4.3 (Multinomial - Dirichlet). Suppose that $(X|\boldsymbol{\theta} = \theta) \sim \text{Multinomial}_k(n, \underline{\theta})$ and $\boldsymbol{\theta} \sim$ Dirichlet(α); i.e.

$$\lambda(\theta) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\prod_{i=1}^k \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}$$

for $\theta_i \geq 0$, $\sum_{i=1}^k \theta_i = 1$, $\alpha_i > 0$. Then $(\underline{\boldsymbol{\theta}}|\underline{X} = \underline{x}) \sim \text{Dirichlet}(\underline{\alpha} + \underline{x})$,

$$E(\underline{\boldsymbol{\theta}}) = \frac{\underline{\alpha}}{\sum_{i=1}^k \alpha_i},$$

and

$$d_\Lambda(\underline{X}) \quad = \quad \frac{\underline{\alpha} + \underline{X}}{\sum \alpha_i + n} = \frac{\sum \alpha_i}{\sum \alpha_i + n} \frac{\underline{\alpha}}{\sum \alpha_i} + \frac{n}{\sum \alpha_i + n} \frac{\underline{X}}{n}.$$

5 Finding Bayes Rules

For convex loss functions we can restrict attention to non-randomized rules; see corollary 1.7.9, TPE, page 48. The following theorem gives a useful recipe for finding Bayes rules.

Theorem 5.1 Suppose that $\boldsymbol{\theta} \sim \Lambda$, $(X|\boldsymbol{\theta} = \theta) \sim P_\theta$, and $L(\theta, a) \geq 0$ for all $\theta \in \Theta$, $a \in \mathbf{A}$. If
(i) There exists a rule d_0 with finite risk, and
(ii) For a.e. x there exists $d_\Lambda(x)$ minimizing

$$E\{L(\boldsymbol{\theta}, d(\cdot|x))|X = x\} = \int_\Theta L(\theta, d(\cdot|x))\lambda(\theta|x)d\nu(\theta),$$

then $d_\Lambda(\cdot|x)$ is a Bayes rule.

Proof. For any rule d with finite risk

(a) $$E\{L(\boldsymbol{\theta}, d(X))|X\} \geq E\{L(\boldsymbol{\theta}, d_\Lambda(X))|X\} \qquad \text{a.s.}$$

by (ii). Hence

(b) $$\begin{aligned} \mathcal{R}(\Lambda, d) &= EE\{L(\boldsymbol{\theta}, d(X))|X\} = EL(\boldsymbol{\theta}, d(X)) \\ &\geq EE\{L(\boldsymbol{\theta}, d_\Lambda(X))|X\} = EL(\boldsymbol{\theta}, d_\Lambda(X)) \\ &= \mathcal{R}(\Lambda, d_\Lambda). \end{aligned}$$

□

Corollary 1 (Estimation with weighted squared - error loss). If $\Theta = \mathbf{A} = \mathbb{R}$ and $L(\theta, a) = K(\theta)|\theta - a|^2$, then

$$d_\Lambda(X) = \frac{E\{K(\boldsymbol{\theta})\boldsymbol{\theta}|X\}}{E\{K(\boldsymbol{\theta})|X\}} = \frac{\int_\Theta \theta K(\theta)d\Lambda(\theta|X)}{\int_\Theta K(\theta)d\Lambda(\theta|X)}.$$

When $K(\theta) = 1$, then

$$d_\Lambda(X) = \int \theta d\Lambda(\theta|X) = E\{\boldsymbol{\theta}|X\} \equiv \text{the posterior mean.}$$

Proof. For an arbitrary (nonrandomized) rule $d \in \mathcal{D}$,

$$\begin{aligned} \int K(\theta)|\theta - d(x)|^2 d\Lambda(\theta|x) &= \int K(\theta)|\theta - d_\Lambda(x) + d_\Lambda(x) - d(x)|^2 d\Lambda(\theta|x) \\ &= \int K(\theta)|\theta - d_\Lambda(x)|^2 d\Lambda(\theta|x) \\ &\quad + 2(d_\Lambda(x) - d(x)) \int K(\theta)\{\theta - d_\Lambda(x)\}d\Lambda(\theta|x) \\ &\quad + (d_\Lambda(x) - d(x))^2 \int K(\theta)d\Lambda(\theta|x) \\ &\geq \int K(\theta)|\theta - d_\Lambda(x)|^2 d\Lambda(\theta|x) \end{aligned}$$

with equality if $d(x) = d_\Lambda(x)$. □

Corollary 2 (Estimation with L_1-loss). If $\Theta = \mathbf{A} = \mathbb{R}$ and $L(\theta, a) = |\theta - a|$, then

$$d_\Lambda(x) = \text{ any median of } \Lambda(\theta|x).$$

Corollary 3 (Testing with $0-1$ loss). If $\mathbf{A} = \{0, 1\}$, $\Theta = \Theta_0 + \Theta_1$ (in the sense of disjoint union of sets), and $L(\theta, a_i) = l_i 1_{\Theta_i^c}(\theta)$, $i = 0, 1$, then any rule of the form

$$d_\Lambda(x) = \begin{cases} 1 & \text{if } P(\boldsymbol{\theta} \in \Theta_1|X = x) > (l_1/l_0)P(\boldsymbol{\theta} \in \Theta_0|X = x) \\ \gamma(x) & \text{if } P(\boldsymbol{\theta} \in \Theta_1|X = x) = (l_1/l_0)P(\boldsymbol{\theta} \in \Theta_0|X = x) \\ 0 & \text{if } P(\boldsymbol{\theta} \in \Theta_1|X = x) < (l_1/l_0)P(\boldsymbol{\theta} \in \Theta_0|X = x) \end{cases}$$

is Bayes with respect to Λ. Note that this reduces to a test of the Neyman - Pearson form when $\Theta_i = \{\theta_i\}$, $i = 0, 1$.

Proof. Let $\phi(x) = d(1|x)$. Then

$$\begin{aligned} E\{L(\boldsymbol{\theta}, \phi(x))|X = x\} &= \int_\Theta L(\theta, \phi(x))d\Lambda(\theta|x) \\ &= \int_\Theta \{l_1\phi(x)1_{\Theta_0}(\theta) + l_0(1 - \phi(x))1_{\Theta_1}(\theta)\}d\Lambda(\theta|x) \\ &= l_1\phi(x)P(\boldsymbol{\theta} \in \Theta_0|X = x) + l_0(1 - \phi(x))P(\boldsymbol{\theta} \in \Theta_1|X = x) \\ &= l_0P(\boldsymbol{\theta} \in \Theta_1|X = x) + \phi(x)\{l_1P(\boldsymbol{\theta} \in \Theta_0|X = x) - l_0P(\boldsymbol{\theta} \in \Theta_1|X = x)\} \end{aligned}$$

which is minimized by any rule of the form d_Λ. $\square$

Corollary 4 (Testing with linear loss). If $\Theta = \mathbb{R}$, $\mathbf{A} = \{0, 1\}$, $\Theta_0 = (-\infty, \theta_0]$, $\Theta_1 = (\theta_0, \infty)$, and

$$L(\theta, 0) = (\theta - \theta_0)1_{\Theta_1}(\theta); \qquad L(\theta, 1) = (\theta_0 - \theta)1_{\Theta_0}(\theta);$$

then

$$d_\Lambda(x) = \begin{cases} 1 & \text{if } E(\boldsymbol{\theta}|X = x) > \theta_0, \\ \gamma(x) & \text{if } E(\boldsymbol{\theta}|X = x) = \theta_0, \\ 0 & \text{if } E(\boldsymbol{\theta}|X = x) < \theta_0 \end{cases}$$

is Bayes with respect to Λ.

Proof. Again, by theorem 5.1 it suffices to minimize

$$\begin{aligned} &E\{L(\boldsymbol{\theta}, \phi(x))|X = x\} \\ &\quad = \int_\Theta \{\phi(x)(\theta_0 - \theta)1_{(-\infty,\theta_0]}(\theta) + (1 - \phi(x))(\theta - \theta_0)1_{(\theta_0,\infty)}(\theta)\}d\Lambda(\theta|x) \\ &\quad = \int_\Theta (\theta_0 - \theta)1_{(-\infty,\theta_0]}(\theta)d\Lambda(\theta|x) + (1 - \phi(x))\{E(\boldsymbol{\theta}|X = x) - \theta_0\} \end{aligned}$$

which is minimized for each fixed x by any rule of the form d_Λ. $\square$

Example 5.1 Suppose $X \sim$ Binomial(n, θ) with $\boldsymbol{\theta} \sim Beta(\alpha, \beta)$. Thus $(\boldsymbol{\theta}|X = x) \sim$ Beta$(\alpha + x, \beta + n - x)$, so that

$$E(\boldsymbol{\theta}) = \frac{\alpha}{\alpha + \beta}$$

and hence for $L(\theta, a) = (\theta - a)^2$ the Bayes rule is

$$\begin{aligned} d_\Lambda(X) &= E\{\boldsymbol{\theta}|X\} = \frac{\alpha + X}{\alpha + \beta + n} \\ &= \frac{\alpha + \beta}{\alpha + \beta + n} \cdot \frac{\alpha}{\alpha + \beta} + \frac{n}{\alpha + \beta + n} \frac{X}{n}. \end{aligned}$$

If the loss is $L(\theta, a) = (\theta - a)^2/\{\theta(1 - \theta)\}$ instead of squared - error loss, then, with $B(\alpha, \beta) \equiv \Gamma(\alpha + \beta)/(\Gamma(\alpha)\Gamma(\beta))$,

$$\begin{aligned} E\{\boldsymbol{\theta} K(\boldsymbol{\theta})|X\} &= \frac{B(\alpha + X, \beta + n - X - 1)}{B(\alpha + X, \beta + n - X)}, \\ E\{K(\boldsymbol{\theta})|X\} &= \frac{B(\alpha + X - 1, \beta + n - X - 1)}{B(\alpha + X, \beta + n - X)}, \end{aligned}$$

and hence the Bayes rule with respect to Λ for this loss function is

$$\begin{aligned} d_\Lambda(X) &= \frac{B(\alpha + X, \beta + n - X - 1)}{B(\alpha + X - 1, \beta + n - X - 1)} \\ &= \lambda_n(\alpha, \beta) \frac{\alpha - 1}{\alpha + \beta - 2} + (1 - \lambda_n(\alpha, \beta)) \frac{X}{n} \end{aligned}$$

where $\lambda_n(\alpha, \beta) = (\alpha + \beta - 2)/(\alpha + \beta + n - 2)$. Note that when $\alpha = \beta = 1$, the Bayes estimator for this loss function becomes the familiar maximum likelihood estimator $\hat{p} = X/n$.

Example 5.2 Suppose that $(\underline{X}|\underline{\boldsymbol{\theta}}) \sim$ Multinomial$_k(n, \underline{\boldsymbol{\theta}})$, and $\boldsymbol{\theta} \sim$ Dirichlet$(\underline{\alpha})$. Then

$$E(\underline{\boldsymbol{\theta}}) = \frac{\underline{\alpha}}{\sum \alpha_i},$$

and for squared error loss the Bayes rule is

$$d_\Lambda(\underline{X}) = E(\boldsymbol{\theta}|\underline{X}) = \frac{\underline{\alpha} + \underline{X}}{\sum \alpha_i + n}.$$

Example 5.3 Normal with normal prior. If $(X|\boldsymbol{\theta} = \theta) \sim N(\theta, \sigma^2)$, $\boldsymbol{\theta} \sim N(\mu, \sigma^2)$, then

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

Consequently

$$E(\boldsymbol{\theta}) = \mu,$$

and, for squared error loss the Bayes rule is

$$d_\Lambda(X) = E(\boldsymbol{\theta}|X) = \frac{1/\tau^2}{1/\tau^2 + 1/\sigma^2}\mu + \frac{1/\sigma^2}{1/\tau^2 + 1/\sigma^2}X.$$

This remains true if $L(\theta, a) = \rho(\theta - a)$ where ρ is convex and even (see e.g. Lehmann, TPE, page 244).

The following example of the general set-up is sometimes referred to as a "classification problem". Suppose that $\Theta = \{\theta_1, \ldots, \theta_k\}$ and $\mathbf{A} = \{a_1, \ldots, a_k\} = \Theta$. Suppose that $X \sim P_i \equiv P_{\theta_i}$, $i = 1, \ldots, k$ when θ_i is the true state of nature where P_i has density p_i with respect to a $\sigma-$finite dominating measure μ A simple loss function is given by

$$L(\theta_i, a_j) = 1\{i \neq j\}, \qquad i = 1, \ldots, k, \quad j = 1, \ldots, k.$$

Given a ("multiple") decision rule $d = (d(1|X), \ldots, d(k|X))$, the risk function is

$$\begin{aligned} R(\theta_i, d) &= \sum_{j=1}^{k} L(\theta_i, a_j) E_{\theta_i}\{d(j|X)\} = \sum_{j \neq i} E_{\theta_i}\{d(j|X)\} \\ &= 1 - E_{\theta_i}\{d(i|X)\}. \end{aligned}$$

Suppose that $\lambda = (\lambda_1, \ldots, \lambda_k)$ is a prior distribution on $\Theta = \{\theta_1, \ldots, \theta_k\}$. then the Bayes risk is

$$\begin{aligned} \mathcal{R}(\lambda, d) &= 1 - \sum_{i=1}^{k} \lambda_i E_{\theta_i}\{d(i|X)\} \\ &= \text{probability of missclassification using } d \\ &\qquad \text{when the distribution from which } X \text{ is drawn} \\ &\qquad \text{is chosen according to } \lambda. \end{aligned}$$

Here is a theorem characterizing the class of Bayes rules in this setting.

Theorem 5.2 Any rule d for which

$$d(i|X) = 0 \qquad \text{whenever} \quad \lambda_i p_i(X) < \max_j \lambda_j p_j(X)$$

for $i = 1, \ldots, k$ is Bayes with respect to λ. Equivalently, since $\sum_{i=1}^k d(i|X) = 1$ for all X,

$$d(i|X) = 1 \qquad \text{if} \quad \lambda_i p_i(X) > \lambda_j p_j(X), \quad \text{for } j \neq i.$$

Proof. Let d' be any other rule. We want to show that

$$\mathcal{R}(\lambda, d') - \mathcal{R}(\lambda, d) \geq 0$$

where d is any rule of the form given. But

$$\begin{aligned} \mathcal{R}(\lambda, d') - \mathcal{R}(\lambda, d) &= \sum_{i=1}^{k} \lambda_i \int d(i|x) p_i(x) d\mu(x) - \sum_{j=1}^{k} \lambda_j \int d'(j|x) p_j(x) d\mu(x) \\ &= \sum_{i=1}^{k} \sum_{j=1}^{k} \int d(i|x) d'(j|x) \{\lambda_i p_i(x) - \lambda_j p_j(x)\} d\mu(x) \\ &\geq 0 \end{aligned}$$

since whenever $\lambda_i p_i(x) < \lambda_j p_j(x)$ it follows that $d(i|x) = 0$. $\square$

Finally, here is a cautionary example.

Example 5.4 (Ritov-Wasserman). Suppose that $(X_1, R_1, Y_1), \ldots, (X_n, R_n, Y_n)$ are i.i.d. with a distribution described as follows. Let $\theta = (\theta_1, \ldots, \theta_B) \in [0,1]^B \equiv \Theta$ where B is large, e.g. 10^{10}. Let $\xi = (\xi_1, \ldots, \xi_B)$ be a vector of *known* numbers with $0 < \delta \le \xi_j \le 1 - \delta < 1$ for $j = 1, \ldots, B$. Furthermore, suppose that:

(i) $X_i \sim \text{Uniform}\{1, \ldots, B\}$.

(ii) $R_i \sim \text{Bernoulli}(\xi_{X_i})$.

(iii) If $R_i = 1$, $Y_i \sim \text{Bernoulli}(\theta_{X_i})$; if $R_i = 0$, Y_i is missing (i.e. not observed).

Our goal is to estimate

$$\psi = \psi(\theta) = P_\theta(Y_1 = 1) = \sum_{j=1}^{B} P(Y_1 = 1 | X_1 = j) P(X_1 = j) = \frac{1}{B} \sum_{j=1}^{B} \theta_j .$$

Now the likelihood contribution of (X_i, R_i, Y_i) is

$$f(X_i) f(R_i|X_i) f(Y_i|X_i, R_i) = \frac{1}{B} \xi_{X_i}^{R_i} (1 - \xi_{X_i})^{1-R_i} \theta_{X_i}^{Y_i R_i} (1 - \theta_{X_i})^{(1-Y_i)R_i},$$

and hence the likelihood for θ is

$$\begin{aligned} L_n(\theta) &= \prod_{i=1}^{n} \frac{1}{B} \xi_{X_i}^{R_i} (1 - \xi_{X_i})^{1-R_i} \theta_{X_i}^{Y_i R_i} (1 - \theta_{X_i})^{(1-Y_i)R_i} \\ &\propto \prod_{i=1}^{n} \theta_{X_i}^{Y_i R_i} (1 - \theta_{X_i})^{(1-Y_i)R_i} . \end{aligned}$$

Thus

$$\begin{aligned} l_n(\theta) &= \sum_{i=1}^{n} \{ Y_i R_i \log \theta_{X_i} + (1 - Y_i) R_i \log(1 - \theta_{X_i}) \} \\ &= \sum_{j=1}^{B} n_j \log \theta_j + \sum_{j=1}^{B} m_j \log(1 - \theta_j) \end{aligned}$$

where

$$n_j = \#\{i : \; Y_i = 1, \; R_i = 1, X_i = j\}, \qquad m_j = \#\{i : \; Y_i = 0, \; R_i = 1, X_i = j\}.$$

Note that $n_j = m_j = 0$ for most j since $B >> n$. Thus the MLE for most θ_j is not defined. Furthermore, for most θ_j the posterior distribution is the prior distribution (especially if the prior is a product distribution on $\Theta = [0,1]^B$!). Thus both MLE and Bayes estimation fail.

Here is a purely frequentist solution: the Horovitz- Thompson estimator of ψ is

$$\hat{\psi}_n = \frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{\xi_{X_i}} Y_i .$$

Note that

$$\begin{aligned}
E(\hat{\psi}_n) &= E\{\frac{R_1}{\xi_{X_1}}Y_1\} = EE\{\frac{R_1}{\xi_{X_1}}Y_1|X_1,R_1\} = E\left\{\frac{R_1}{\xi_{X_1}}E\{Y_1|X_1,R_1\}\right\} \\
&= E\left\{\frac{R_1}{\xi_{X_1}}\theta_{X_1}\right\} = EE\left\{\frac{R_1}{\xi_{X_1}}\theta_{X_1}|X_1\right\} = E\left\{\frac{\theta_{X_1}}{\xi_{X_1}}E\{R_1|X_1\}\right\} \\
&= E\left\{\frac{\theta_{X_1}}{\xi_{X_1}}\xi_{X_1}\right\} = E\{\theta_{X_1}\} \\
&= B^{-1}\sum_{j=1}^{B}\theta_j = \psi(\theta)\,.
\end{aligned}$$

Thus $\hat{\psi}_n$ is an unbiased estimator of $\psi(\theta)$. Moreover,

$$(1)\qquad Var(\hat{\psi}_n) = \frac{1}{n}\left\{\frac{1}{B}\sum_{j=1}^{B}\frac{\theta_j}{\xi_j} - \psi(\theta)^2\right\}.$$

Exercise 5.1 Show that (1) holds and hence that

$$Var(\hat{\psi}_n) \leq \frac{1}{n\delta}$$

under the assumption that $\xi_j \geq \delta > 0$ for all $1 \leq j \leq B$. [Hint: use the formula $Var(Y) = EVar(Y|X) + Var[E(Y|X)]$ twice.]

6 Finding Minimax Rules

Definition 6.1 A prior Λ_0 for which $\mathcal{R}(\Lambda, d_\Lambda)$ is maximized is called a *least favorable prior*:

$$\mathcal{R}(\Lambda_0, d_{\Lambda_0}) = \sup_{\Lambda} \mathcal{R}(\Lambda, d_\Lambda).$$

Theorem 6.1 Suppose that Λ is a prior distribution on Θ such that

$$(1) \qquad \mathcal{R}(\Lambda, d_\Lambda) = \int_\Theta R(\theta, d_\Lambda) d\Lambda(\theta) = \sup_\theta R(\theta, d_\Lambda).$$

Then:
(i) d_Λ is minimax.
(ii) If d_Λ is unique Bayes with respect to Λ, d_Λ is unique minimax.
(iii) Λ is least favorable.

Proof. (i) Let d be another rule. Then

$$\begin{aligned} \sup_\theta R(\theta, d) &\geq \int_\Theta R(\theta, d) d\Lambda(\theta) \\ \text{(a)} \qquad &\geq \int_\Theta R(\theta, d_\Lambda) d\Lambda(\theta) \qquad \text{since } d_\Lambda \text{ is Bayes wrt } \Lambda \\ &= \sup_\theta R(\theta, d_\Lambda) \qquad \text{by (1).} \end{aligned}$$

Hence d_Λ is minimax.
(ii). If d_Λ is unique Bayes, then $>$ holds in (a), so d_Λ is unique minimax.
(iii). Let Λ^* be some other prior distribution. Then

$$\begin{aligned} r_{\Lambda^*} &\equiv \int_\Theta R(\theta, d_{\Lambda^*}) d\Lambda^*(\theta) \leq \int_\Theta R(\theta, d_\Lambda) d\Lambda^*(\theta) \\ &\qquad \text{since } d_{\Lambda^*} \text{ is Bayes wrt } \Lambda^* \\ &\leq \sup_\theta R(\theta, d_\Lambda) \\ &= \mathcal{R}(\Lambda, d_\Lambda) \equiv r_\Lambda \qquad \text{by (1).} \end{aligned}$$

□

Corollary 1 If d_Λ is Bayes with respect to Λ and has constant risk, $R(\theta, d_\Lambda) = \text{constant}$, then d_Λ is minimax.

Proof. If d_Λ has constant risk, then (1) holds. □

Corollary 2 Let

$$\Theta_\Lambda \equiv \{\theta \in \Theta : \ R(\theta, d_\Lambda) = \sup_{\theta'} R(\theta', d_\Lambda)\}$$

be the set of θ's where the risk of d_Λ assumes its maximum. Then d_Λ is minimax if $\Lambda(\Theta_\Lambda) = 1$. Equivalently, d_Λ is minimax if there is a set Θ_Λ with $\Lambda(\Theta_\Lambda) = 1$ and $R(\theta, d_\Lambda) = \sup_{\theta'} R(\theta', d_\Lambda)$ for all $\theta \in \Theta_\Lambda$.

Example 6.1 Suppose $X \sim$ Binomial(n, θ) with $\boldsymbol{\theta} \sim$ Beta(α, β). Thus $(\boldsymbol{\theta}|X = x) \sim$ Beta$(\alpha + x, \beta + n - x)$, and as we computed in section 5,

$$d_\Lambda(X) = \frac{\alpha + X}{\alpha + \beta + n} = \frac{\alpha + \beta}{\alpha + \beta + n}\frac{\alpha}{\alpha + \beta} + \frac{n}{\alpha + \beta + n}\frac{X}{n}.$$

Consequently,

$$\begin{aligned} R(\theta, d_\Lambda) &= \left(\frac{n}{\alpha + \beta + n}\right)^2 \frac{\theta(1-\theta)}{n} + \left(\frac{\alpha + n\theta}{\alpha + \beta + n} - \theta\right)^2 \\ &= \frac{1}{(\alpha + \beta + n)^2}\left\{\alpha^2 + (n - 2\alpha(\alpha + \beta))\theta + ((\alpha + \beta)^2 - n)\theta^2\right\} \\ &= \frac{1}{(\alpha + \beta + n)^2}\alpha^2 \end{aligned}$$

if $2\alpha(\alpha + \beta) = n$ and $(\alpha + \beta)^2 = n$. But solving these two equations yields $\alpha = \beta = \sqrt{n}/2$. Thus for these choices of α and β, the risk of the Bayes rule is constant in θ, and hence

$$d_M(X) = \frac{1}{1 + \sqrt{n}}\frac{1}{2} + \frac{\sqrt{n}}{1 + \sqrt{n}}\frac{X}{n}$$

is Bayes with respect to $\Lambda =$ Beta$(\sqrt{n}/2, \sqrt{n}/2)$ and has risk

$$R(\theta, d_M) = \frac{1}{4(1 + \sqrt{n})^2} \le \frac{\theta(1-\theta)}{n} = R(\theta, X/n)$$

if

$$|\theta - 1/2| \le \frac{1}{2}\frac{\sqrt{1 + 2\sqrt{n}}}{1 + \sqrt{n}} \sim \frac{1}{\sqrt{2}n^{1/4}}.$$

Hence Beta$(\sqrt{n}/2, \sqrt{n}/2)$ is least favorable and d_M is minimax. Note that d_M is a consistent estimator of θ:

$$d_M(X) \to_p \theta$$

as $n \to \infty$, but its bias is of the order $n^{-1/2}$ (the bias equals $(1/2 - \theta)/(1 + \sqrt{n})$). Furthermore, d_M is a (locally) regular estimator of θ, but the limit distribution is *not* centered at zero if $\theta \ne 1/2$: under P_{θ_n} with $\theta_n = \theta + tn^{-1/2}$

$$\sqrt{n}(d_M(X) - \theta_n) \to_d N(1/2 - \theta, \theta(1-\theta)).$$

If a least-favorable prior does not exist, then we can still consider improper priors or limits of proper priors:

Let $\{\Lambda_k\}$ be a sequence of prior distributions, let d_k denote the Bayes estimator corresponding to Λ_k, and set

$$r_k \equiv \int_\Theta R(\theta, d_k) d\Lambda_k(\theta).$$

Suppose that

$$(2) \qquad r_k \to r < \infty \qquad \text{as} \quad k \to \infty.$$

Definition 6.2 The sequence of prior distributions $\{\Lambda_k\}$ with Bayes risks $\{r_k\}$ is said to be *least favorable* if $r_\Lambda \equiv \mathcal{R}(\Lambda, d_\Lambda) \le r$ for any prior distribution Λ.

Theorem 6.2 Suppose that $\{\Lambda_k\}$ is a sequence of prior distributions with Bayes risks satisfying

$$r_k \to r, \tag{3}$$

and d is an estimator for which

$$\sup_\theta R(\theta, d) = r. \tag{4}$$

Then:
A. d is mininax, and
B. $\{\Lambda_k\}$ is least - favorable.

Proof. A. Suppose d^* is any other estimator. Then

$$\sup_\theta R(\theta, d^*) \ge \int R(\theta, d^*) d\Lambda_k(\theta) \ge r_k$$

for all $k \ge 1$. Hence

$$\sup_\theta R(\theta, d^*) \ge r = \sup_\theta R(\theta, d) \qquad \text{by } (4),$$

so d is minimax.
B. If Λ is any prior distribution, then

$$r_\Lambda = \int R(\theta, d_\Lambda) d\Lambda(\theta) \quad \le \quad \int R(\theta, d) d\Lambda(\theta) \le \sup_\theta R(\theta, d) = r$$

by (4). $\Box$

Example 6.2 Let $X_1, \ldots, X_n$ be i.i.d. $N(\theta, \sigma^2)$ given $\boldsymbol{\theta} = \theta$, and suppose that $\boldsymbol{\theta} \sim N(\mu, \tau^2)$ and that σ^2 is known. Then it follows from (5.3) that

$$d_\Lambda(\underline{X}) = E\{\boldsymbol{\theta}|\underline{X}\} = \frac{1/\tau^2}{1/\tau^2 + n/\sigma^2}\mu + \frac{n/\sigma^2}{1/\tau^2 + n/\sigma^2}\overline{X}_n$$

and

$$\begin{aligned}
\mathcal{R}(\Lambda, d_\Lambda) &\equiv r_\Lambda = E\{\boldsymbol{\theta} - d_\Lambda(\underline{X})\}^2 \\
&= EE\{[\boldsymbol{\theta} - d_\Lambda(\underline{X})]^2|\underline{X}\} \\
&= EVar\{\boldsymbol{\theta}|\underline{X}\} \\
&= \frac{1}{1/\tau^2 + n/\sigma^2} \to \frac{\sigma^2}{n} \qquad \text{as } \tau^2 \to \infty \\
&= R(\theta, \overline{X})
\end{aligned}$$

for all $\theta \in \Theta = \mathbb{R}$. Hence by theorem 6.2, $\overline{X}$ is a minimax estimator of θ.

The remainder of this section is aimed at extending minimaxity of estimators from smaller models to larger ones.

Lemma 6.1 Suppose that $X \sim P \in \mathcal{M} \equiv \{\text{all } P\text{'s on } \mathcal{X}\}$, and that $\nu : \mathcal{P} \subset \mathcal{M} \mapsto \mathbb{R}$ is a functional (e.g. $\nu(P) = E_P(X) = \int x dP(x)$, $\nu(P) = Var_P(X)$, $\nu(P) = Pf$ for some fixed function f). Suppose that d is a minimax estimator of $\nu(P)$ for $P \in \mathcal{P}_0 \subset \mathcal{P}_1$. If

$$\sup_{P\in\mathcal{P}_0} R(P, d^*) = \sup_{P\in\mathcal{P}_1} R(P, d),$$

then d is also minimax for estimating $\nu(P)$, $P \in \mathcal{P}_1$.

Proof. Suppose that d is not minimax. Then there exists a rule d^* with smaller maximum risk:

$$\sup_{P\in\mathcal{P}_0} R(P, d^*) \le \sup_{P\in\mathcal{P}_1} R(P, d^*) < \sup_{P\in\mathcal{P}_1} R(P, d) = \sup_{P\in\mathcal{P}_0} R(P, d)$$

which contradicts the hypothesis that d is minimax for $\mathcal{P}_0$. $\square$

Example 6.3 This continues the normal mean example 6.2. Suppose that σ^2 is unknown. Thus, with $\Theta = \{(\theta, \sigma^2) : \ \theta \in \mathbb{R}, \ 0 < \sigma^2 < \infty\}$,

$$\sup_{(\theta,\sigma^2)\in\Theta} R((\theta, \sigma^2), \overline{X}_n) = \sup_{(\theta,\sigma^2)\in\Theta} \frac{\sigma^2}{n} = \infty.$$

So, to get something reasonable, we need to restrict σ^2. Let

$$\mathcal{P}_0 = \{N(\theta, M) : \ \theta \in \mathbb{R}\},$$
$$\mathcal{P}_1 = \{N(\theta, \sigma^2) : \ \theta \in \mathbb{R}, \ 0 \le \sigma^2 \le M\}.$$

Then $\overline{X}_n$ is minimax for $\mathcal{P}_0$ by our earlier calculation, $\mathcal{P}_0 \subset \mathcal{P}_1$, and

$$\sup_{\mathcal{P}_1} R(P, \overline{X}_n) = \sup_{\mathcal{P}_0} R(P, \overline{X}_n) = \frac{M}{n}.$$

Thus $\overline{X}_n$ is a minimax estimator of θ for $\mathcal{P}_1$.

Example 6.4 Let $X_1, \ldots, X_n$ be i.i.d. $P \in \mathcal{P}_\mu$ where

$$\mathcal{P}_\mu = \{\text{all probability measures } \ P : \ E_P|X| < \infty\}$$

and consider estimation of $\nu(P) = E_P(X)$ with squared error loss for the families

$$\mathcal{P}_{b\sigma^2} \equiv \{P \in \mathcal{P}_\mu : \ Var_P(X) \le M < \infty\}$$
$$\mathcal{P}_{br} \equiv \{P \in \mathcal{P}_\mu : \ P(a \le X \le B) = 1\} \qquad \text{for some fixed } \ a, b \in \mathbb{R}.$$

Then:
A. $\overline{X}$ is minimax for $\mathcal{P}_{b\sigma^2} = \mathcal{P}_1$ by Example 6.3 since it is minimax for $\mathcal{P}_1 \equiv \{N(\theta, \sigma^2) : \ \theta \in \mathbb{R}, \ 0 < \sigma^2 \le M\}$, and

$$\sup_{P\in\mathcal{P}_0} R(P, \overline{X}) = \sup_{P\in\mathcal{P}_1} R(P, \overline{X}).$$

B. Without loss of generality suppose $a = 0$ and $b = 1$. Let

$$\mathcal{P}_1 = \{P : \ P([0, 1]) = 1\},$$
$$\mathcal{P}_0 = \{P \in \mathcal{P}_1 : \ P(X = 1) = p, \quad P(X = 0) = 1 - p \ \text{ for some } \ 0 < p < 1\}.$$

For $\mathcal{P}_0$ we know that the minimax estimator is

$$d_M(\underline{X}) = \frac{1}{1+\sqrt{n}}\frac{1}{2} + \frac{\sqrt{n}}{1+\sqrt{n}}\overline{X}.$$

Now, with $E_P X \equiv \theta$,

$$\begin{aligned} R(P, d_M) &= \frac{1}{(1+\sqrt{n})^2}\left\{\mathrm{Var}_P(X) + \left(\frac{1}{2} - \theta\right)^2\right\} \\ &\le \frac{1}{(1+\sqrt{n})^2}\{\theta - \theta^2 + (1/4) - \theta + \theta^2\} \\ &\qquad \text{since } \; 0 \le X \le 1 \; \text{ implies } \; E_P X^2 \le E_P X \equiv \theta \\ &= \frac{1/4}{(1+\sqrt{n})^2}. \end{aligned}$$

Thus

$$\sup_{P \in \mathcal{P}_1} R(P, d_M) = \sup_{P \in \mathcal{P}_0} R(P, d_M),$$

and by Lemma 6.1 d_M is minimax.

Remark 6.1 The restriction $\sigma^2 \le M$ in (5) is crucial; note that $\nu(P) = E_P(X)$ is discontinuous at every P in the sense that we can easily have $P_m \to_d P$, but $\nu(P_m)$ fails to converge to $\nu(P)$.

Remark 6.2 If $\Theta = [-M, M] \subset \mathbb{R}$, then $\overline{X}$ is no longer minimax in the normal case; see Bickel (1981). The approximately least favorable densities for $M \to \infty$ are

$$\lambda_M(\theta) = M^{-1}\cos^2((\pi/2)(\theta/M)1_{[-M,M]}(\theta).$$

7 Admissibility and Inadmissibility

Our goal in this section is to establish some basic and simple results about admissibility / inadmissibility of some classical estimators. In particular, we will show that $\overline{X}$ is admissible for the Gaussian location model for $k = 1$, but inadmissible for the Gaussian location model for $k \geq 3$. The fundamental work in this area is that of Stein (1956) and James and Stein (1961). For an interesting expository paper, see Efron and Morris (1977); for more work on admissibility issues see e.g. Eaton (1992), (1997), and Brown (1971).

Theorem 7.1 Any unique Bayes estimator is admissible.

Proof. Suppose that d_Λ is unique Bayes with respect to Λ and is inadmissible. Then there exists an estimator d such that $R(\theta, d) \leq R(\theta, d_\Lambda)$ with strict inequality for some θ, and hence

$$\int_\Theta R(\theta, d)\Lambda(\theta) \leq \int_\Theta R(\theta, d_\Lambda) d\Lambda(\theta)$$

which contradicts uniqueness of d_Λ. Thus d_Λ is admissible. $\Box$

Lemma 7.1 If the loss function $L(\theta, a)$ is squared error (or is convex in a) and, with Q defined by $Q(A) = \int P_\theta(X \in A) d\Lambda(\theta)$, a.e. Q implies a.e. $\mathcal{P}$, then a Bayes rule with finite Bayes risk is unique a.e. $\mathcal{P}$. [a.e. $\mathcal{P}$ means $P(N) = 0$ for all $P \in \mathcal{P}$.]

Proof. See Lehmann and Casella TPE Corollary 4.1.4 page 229. $\Box$

Example 7.1 Consider the Bayes estimator of a normal mean, $d_\Lambda = p_n\mu + (1 - p_n)\overline{X}$, $p_n = (1/\tau^2)/(1/\tau^2 + n/\sigma^2)$. The Bayes risk is finite and a.e. Q implies a.e. $\mathcal{P}$. Hence d_Λ is unique Bayes and admissible.

Theorem 7.2 If X is a random variable with mean θ and variance σ^2, then $aX + b$ is *inadmissible* as an estimator of θ for squared error loss if

(i) $a > 1$
(ii) $a < 0$
(iii) $a = 1,\ b \neq 0$.

Proof. For any a, b the risk of the rule $aX + b$ is

$$R(\theta, aX + b) = a^2\sigma^2 + \{(a - 1)\theta + b\}^2 \equiv \rho(a, b).$$

(i) If $a > 1$

$$\rho(a, b) \geq a^2\sigma^2 > \sigma^2 = \rho(1, 0),$$

so $aX + b$ is dominated by X.
(ii) If $a < 0$, then $(a - 1)^2 > 1$ and

$$\begin{aligned}\rho(a, b) &\geq \{(a - 1)\theta + b\}^2 = (a - 1)^2\{\theta + \frac{b}{a - 1}\}^2 \\ &> \{\theta + \frac{b}{a - 1}\}^2 = \rho(0, -b/(a - 1)).\end{aligned}$$

(iii) If $a = 1$, $b \neq 0$,

$$\rho(1, b) = \sigma^2 + b^2 > \sigma^2 = \rho(1, 0),$$

so $X + b$ is dominated by X. $\square$

Example 7.2 Thus $a\overline{X} + b$ is inadmissible for $a < 0$ or $a > 1$. When $a = 0$, $d = b$ is admissible since it is the only estimator with zero risk at $\theta = b$. When $a = 1$, $b \neq 0$, d is inadmissible. What is left is $\overline{X}$: is $\overline{X}$ admissible as the estimator of a normal mean in $\mathbb{R}$? The following theorem answers this affirmatively.

Theorem 7.3 If $X_1, \ldots, X_n$ are i.i.d. $N(\theta, \sigma^2)$, $\theta \in \Theta = \mathbb{R}$ with σ^2 known, then $\overline{X}$ is an admissible estimator of θ.

Proof. **Limiting Bayes method**. Suppose that $\overline{X}$ is inadmissible (and $\sigma = 1$). Then there is an estimator d^* such that $R(\theta, d^*) \leq 1/n = R(\theta, \overline{X})$ for all θ with risk $< 1/n$ for some θ. Now

$$R(\theta, d) = E_\theta(\theta - d(X))^2$$

is continuous in θ for every d, and hence there exists $\epsilon > 0$ and $\theta_0 < \theta_1$ so that

$$R(\theta, d^*) < \frac{1}{n} - \epsilon \qquad \text{for all } \ \theta_0 < \theta < \theta_1.$$

Let

$$r_\tau^* \equiv \int_\Theta R(\theta, d^*) d\Lambda(\theta)$$

where $\Lambda = N(0, \tau^2)$. Thus

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

so $r_\tau \leq r_\tau^*$. Thus

$$\begin{aligned}
\frac{1/n - r_\tau^*}{1/n - r_\tau} &= \frac{\frac{1}{\sqrt{2\pi}\tau} \int_{-\infty}^{\infty} \{1/n - R(\theta, d^*)\} \exp(-\theta^2/2\tau^2) d\theta}{1/n - \tau^2/(1 + n\tau^2)} \\
&\geq \frac{\frac{1}{\sqrt{2\pi}\tau} \epsilon \int_{\theta_0}^{\theta_1} \exp(-\theta^2/2\tau^2) d\theta}{1/n(1 + n\tau^2)} \\
&= \frac{n(1 + n\tau^2)}{\sqrt{2\pi}\tau} \epsilon \int_{\theta_0}^{\theta_1} \exp(-\theta^2/2\tau^2) d\theta \\
&\to \infty \cdot \epsilon \cdot (\theta_1 - \theta_0) = \infty \qquad \text{as } \ \tau \to \infty.
\end{aligned}$$

Hence

$$\frac{1}{n} - r_\tau^* > \frac{1}{n} - r_\tau$$

for $\tau >$ some τ_0, or, $r_\tau^* < r_\tau$ for some $\tau > \tau_0$, which contradicts d_τ Bayes with respect to Λ_τ with Bayes risk r_τ. Hence $\overline{X}$ is admissible. $\square$

Proof. **(Information inequality method).** The risk of any rule d is

$$\begin{aligned} R(\theta, d) &= E_\theta(d-\theta)^2 = Var_\theta[d(X)] + b^2(\theta) \\ &\geq \frac{[1+b'(\theta)]^2}{n} + b^2(\theta) \end{aligned}$$

since $I(\theta) = 1$. Suppose that d is any estimator satisfying

$$R(\theta, d) \leq 1/n \qquad \text{for all } \theta.$$

then

(a) $$\frac{1}{n} \geq b^2(\theta) + \frac{[1+b'(\theta)]^2}{n}.$$

But (a) implies that:
(i) $|b(\theta)| \leq 1/\sqrt{n}$ for all θ; i.e. b is bounded.
(ii) $b'(\theta) \leq 0$ since $1 + 2b'(\theta) + [b'(\theta)]^2 \leq 1$.
(iii) There exists $\theta_i \to \infty$ such that $b'(\theta_i) \to 0$.
[If $b'(\theta) \leq -\epsilon$ for all $\theta > \theta_0$, then $b(\theta)$ is not bounded.]
Similarly, there exists $\theta_i \to -\infty$ such that $b'(\theta_i) \to 0$. (iv)

$$b^2(\theta) \leq \frac{1}{n} - \frac{[1+b'(\theta)]^2}{n} = -\frac{2b'(\theta) + [b'(\theta)]^2}{n}$$

Hence $b(\theta) = 0$, and $R(\theta, d) \equiv 1/n$. $\square$

Theorem 7.4 (Stein's theorem) If $k \geq 3$, then $\overline{X}$ is inadmissible.

Remark 7.1 The sample mean is admissible when $k = 2$; see Stein (1956), Ferguson (1967), page 170.

Proof. Let $g : \mathbb{R}^k \to \mathbb{R}^k$ have

$$E\left|\frac{\partial}{\partial x_i} g_i(X)\right| < \infty,$$

and consider estimators of the form

$$\widehat{\theta}_n = \overline{X} + n^{-1} g(\overline{X}).$$

Now

$$\begin{aligned} E_\theta|\overline{X}-\theta|^2 - E_\theta|\widehat{\theta}_n - \theta|^2 &= E_\theta|\overline{X}-\theta|^2 - E_\theta|\overline{X}-\theta+n^{-1}g(\overline{X})|^2 \\ \text{(a)} \qquad &= -2n^{-1}E_\theta\langle \overline{X}-\theta, g(\overline{X})\rangle - n^{-2}E_\theta|g(\overline{X})|^2. \end{aligned}$$

$\square$

To proceed further we need an identity due to Stein.

Lemma 7.2 If $X \sim N(\theta, \sigma^2)$, and g is a function with $E|g'(X)| < \infty$, then

(1) $\quad \sigma^2 E g'(X) = E(X-\theta)g(X).$

If $X \sim N_k(\theta, \sigma^2 I)$, and $g : \mathbb{R}^k \to \mathbb{R}^k$ has

$$E|\frac{\partial}{\partial x_i} g_i(X)| < \infty,$$

then:

(2) $\quad \sigma^2 E \frac{\partial}{\partial x_i} g_i(X) = E(X_i - \theta_i) g_i(X), \qquad i = 1, \ldots, k.$

Proof. Without loss of generality, suppose that $\theta = 0$ and $\sigma^2 = 1$. Integration by parts gives

$$\begin{aligned} Eg'(X) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g'(x) \exp(-x^2/2) dx \\ &= -\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(x) \exp(-x^2/2)\{-2x/2\} dx \\ &= EXg(X). \end{aligned}$$

In applying integration by parts here, we have ignored the term $uv|_{-\infty}^{\infty} = g(x)\phi(x)|_{-\infty}^{\infty}$. To prove that this does in fact vanish, it is easiest to apply Fubini's theorem (twice!) in a slightly devious way as follows: let $\phi(t) = (2\pi)^{-1/2}\exp(-t^2/2)$ be the standard normal density. Since $\phi'(t) = -t\phi(t)$, we have both

$$\phi(x) = -\int_x^{\infty} \phi'(t) dt = \int_x^{\infty} t\phi(t) dt$$

and

$$\phi(x) = \int_{-\infty}^{x} \phi'(t) dt = -\int_{-\infty}^{x} t\phi(t) dt.$$

Therefore we can write

$$\begin{aligned} Eg'(X) &= \int_{-\infty}^{\infty} g'(x)\phi(x) dx \\ &= \int_0^{\infty} g'(x) \int_x^{\infty} t\phi(t) dt dx - \int_{-\infty}^{0} g'(x) \int_{-\infty}^{x} t\phi(t) dt dx \\ &= \int_0^{\infty} t\phi(t) \left\{ \int_0^t g'(x) dx \right\} dt - \int_{-\infty}^{0} t\phi(t) \left\{ \int_t^0 g'(x) dx \right\} dt \\ &= \left(\int_0^{\infty} + \int_{-\infty}^{0} \right) \{t\phi(t)\}[g(t) - g(0)]\} dt \\ &= \int_{-\infty}^{\infty} t g(t)\phi(t) dt = EXg(X). \end{aligned}$$

Here the third equality is justified by the hypothesis $E|g'(X)| < \infty$ and Fubini's theorem.]

To prove (2), write $X^{(i)} \equiv (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_k)$. Then

$$\begin{aligned} \sigma^2 E \frac{\partial}{\partial x_i} g_i(X) &= \sigma^2 E E\{\frac{\partial}{\partial x_i} g_i(X) | X^{(i)}\} \\ &= EE\{(X_i - \theta_i) g_i(X) | X^{(i)}\} \qquad \text{by } \ (1) \end{aligned}$$

$\square$

Proof. Now we return to the proof of the theorem. Using (2) in (a) yields

$$\text{(a)} \qquad -2n^{-2}E_\theta \sum_{i=1}^{k} \frac{\partial g_i}{\partial x_i}(\overline{X}) - n^{-2}E_\theta |g(X)|^2.$$

Now let $\psi : \mathbb{R}^k \to \mathbb{R}$ be twice differentiable, and set

$$g(x) = \nabla\{\log \psi(x)\} = \frac{1}{\psi(x)} \left(\frac{\partial \psi(x)}{\partial x_1}, \ldots, \frac{\partial \psi(x)}{\partial x_k} \right).$$

Thus

$$\begin{aligned} \frac{\partial g_i}{\partial x_i}(x) &= \frac{1}{\psi(x)} \frac{\partial^2}{\partial x_i^2}\psi(x) - \frac{1}{\psi^2(x)} \left(\frac{\partial \psi(x)}{\partial x_i} \right)^2 \\ &= \frac{1}{\psi(x)} \frac{\partial^2}{\partial x_i^2}\psi(x) - g_i(x)^2 \end{aligned}$$

and

$$\sum_{i=1}^{k} \frac{\partial g_i}{\partial x_i}(x) = \frac{1}{\psi(x)} \nabla^2 \psi(x) - |g(x)|^2.$$

Hence the right side of (a) is

$$\text{(b)} \qquad = n^{-2}E_\theta |g(\overline{X})|^2 - 2n^{-2}E_\theta \left\{ \frac{1}{\psi(\overline{X})} \nabla^2 \psi(\overline{X}) \right\} > 0$$

if $\psi(x) \geq 0$ and $\nabla^2 \psi \leq 0$, $g \neq 0$ (i.e. ψ is super-harmonic). Here is one example of such a function: A. Suppose that $\psi(x) = |x|^{-(k-2)} = \{x_1^2 + \cdots + x_k^2\}^{-(k-2)/2}$. Then

$$g(x) = \nabla \log \psi(x) = -\frac{k-2}{|x|^2} x$$

and $\nabla^2 \psi(x) = 0$, so ψ is *harmonic*. Thus

$$\hat{\theta}_n = \left(1 - \frac{k-2}{n|\overline{X}|^2} \right) \overline{X}$$

and

$$\begin{aligned} E_\theta |\overline{X} - \theta|^2 - E|\widehat{\theta} - \theta|^2 &= n^{-2} E_\theta |g(\overline{X})|^2 = \left(\frac{k-2}{n} \right)^2 E_\theta |\overline{X}|^{-2} \\ &= \left(\frac{k-2}{\sqrt{n}} \right)^2 E_\theta |\sqrt{n}(\overline{X} - \theta) + \sqrt{n}\theta|^{-2} \\ &= \left(\frac{k-2}{\sqrt{n}} \right)^2 E_0 |X + \sqrt{n}\theta|^{-2} \\ &= \begin{cases} \left(\frac{k-2}{n}\right)^2 E_0 |n^{-1/2}X + \theta|^{-2} = O(n^{-2}), & \theta \neq 0, \\ \frac{(k-2)^2}{n} \frac{1}{k-2} = \frac{k-2}{n}, & \theta = 0 \end{cases} \end{aligned}$$

since $|X|^2 \sim \chi^2_k$ with $E(1/\chi^2_k) = 1/(k-2)$. Hence

$$\frac{E_0|\widehat{\theta}-\theta|^2}{E_0|\overline{X}-\theta|^2} = \frac{2}{k} < 1.$$

For general θ,

$$\begin{aligned} R(\theta,\widehat{\theta}) &= \frac{k}{n} - \frac{(k-2)^2}{n} E_0\left(\frac{1}{|X+\sqrt{n}\theta|^2}\right) \\ &= \frac{k}{n}\left(1 - \frac{k-2}{n} E_0\left(\frac{k-2}{\chi^2_k(\delta)}\right)\right) \end{aligned}$$

since $|X+\sqrt{n}\theta|^2 \sim \chi^2_k(\delta)$ with $\delta = n|\theta|^2/2$. Thus

$$\begin{aligned} \frac{R(\theta,\widehat{\theta})}{R(\theta,\overline{X})} &= \left(1 - \frac{k-2}{n} E_0\left(\frac{k-2}{\chi^2_k(\delta)}\right)\right) \\ &= 2/k \qquad \text{when} \ \ \theta = 0, \\ &\to 1 \qquad \text{as} \ \ n \to \infty \ \ \text{for fixed} \ \ \theta \neq 0, \\ &\to 1 \qquad \text{as} \ \ |\theta| \to \infty \ \ \text{for fixed} \ \ n. \end{aligned}$$

$\square$

Remark 7.2 Note that the James-Stein estimator

$$\widehat{\theta}_n = \left(1 - \frac{k-2}{n|\overline{X}|^2}\right)\overline{X}$$

derived above is not regular at $\theta = 0$: if $\theta_n = tn^{-1/2}$, then $\sqrt{n}(\overline{X}-\theta_n) \stackrel{d}{=} Z \sim N_k(0,\sigma^2 I)$ under P_{θ_n}, so that

$$\begin{aligned} \sqrt{n}(\widehat{\theta}_n - \theta_n) &= \sqrt{n}(\overline{X}-\theta_n) - \frac{k-2}{|\sqrt{n}(\overline{X}-\theta_n)+t|^2}\left\{\sqrt{n}(\overline{X}-\theta_n)+t\right\} \\ &= Z - \frac{k-2}{|Z+t|^2}(Z+t) \end{aligned}$$

which has a distribution dependent on t.

Remark 7.3 It turns out that the James - Stein estimator $\widehat{\theta}_n$ is itself inadmissible; see Lehmann and Casella, TPE pages 356-357, and Section 5.7, pages 376-389.

Remark 7.4 Another interesting function ψ is

$$\psi(x) = \begin{cases} |x|^{-(k-2)}, & |x| \geq \sqrt{k-2}, \\ (k-2)^{-(k-2)/2}\exp([(k-2)-|x|^2]/2), & |x| < \sqrt{k-2}. \end{cases}$$

For this ψ we have

$$g(x) = \nabla \log \psi(x) = \begin{cases} -\frac{k-2}{|x|^2}x, & |x| \geq \sqrt{k-2} \\ -x, & |x| \leq \sqrt{k-2}. \end{cases}$$

Another approach to deriving the James-Stein estimator is via an empirical Bayes approach. We know that the Bayes estimator for squared error loss when $X \sim N_k(\boldsymbol{\theta}, \sigma^2 I)$ and $\boldsymbol{\theta} \sim N_k(0, \tau^2 I) \equiv \Lambda_\tau$ is

$$d_\Lambda(X) = \frac{\tau^2}{\sigma^2 + \tau^2} X,$$

with Bayes risk

$$\mathcal{R}(\Lambda, d_\Lambda) = \frac{k}{1/\sigma^2 + 1/\tau^2} = \frac{k\sigma^2\tau^2}{\sigma^2 + \tau^2}.$$

Regarding τ as unknown and estimating it via the marginal distribution Q of X is a (parametric) empirical Bayes approach. Since the marginal distribution Q of X is $N_k(0, (\sigma^2 + \tau^2)I)$, with density

$$q(x; \tau^2) = \frac{1}{(\sqrt{2\pi(\sigma^2 + \tau^2)})^k} \exp\left(-\frac{\|x\|^2}{2(\sigma^2 + \tau^2)}\right)$$

the MLE of $\sigma^2 + \tau^2$ is $\|X\|^2/k$, and the resulting MLE $\hat{\tau}^2$ of τ^2 is given by

$$\hat{\tau}^2 = \left(\frac{\|X\|^2}{k} - \sigma^2\right)^+.$$

Thus

$$\frac{\hat{\tau}^2}{\sigma^2 + \hat{\tau}^2} = \left(1 - \frac{k\sigma^2}{\|X\|^2}\right)^+,$$

and this leads to the following version of the (positive part) James-Stein estimator:

$$d_{EB,MLE}(X) = \left(1 - \frac{k\sigma^2}{\|X\|^2}\right)^+ X.$$

If, instead of the MLE, we used an unbiased estimator of $\tau^2/(\sigma^2 + \tau^2)$, then we get the positive part James-Stein estimator

$$d_{JS}(X) = \left(1 - \frac{(k-2)\sigma^2}{\|X\|^2}\right)^+ X.$$

8 Asymptotic theory of Bayes estimators

We begin with a basic result for Bayes estimation of a Bernoulli probability θ.

Suppose that $(Y_1, Y_2, \ldots)|\boldsymbol{\theta} = \theta$ are i.i.d. Bernoulli(θ). Suppose that the prior distribution is an arbitrary distribution on $(0,1)$. The Bayes estimator of θ with respect to squared error loss is the posterior mean $d(\underline{Y}) = d(Y_1, \ldots, Y_n) = E(\boldsymbol{\theta}|\underline{Y})$. Now this sequence is a (uniformly integrable) martingale; this type of martingale is sometimes known as a *Doob martingale*. Hence by the martingale convergence theorem,

$$d(\underline{Y}_n) = E(\boldsymbol{\theta}|\underline{Y}_n) \rightarrow_{a.s.} E(\boldsymbol{\theta}|Y_1, Y_2, \ldots).$$

To prove consistency, it remains only to show that $E(\boldsymbol{\theta}|Y_1, Y_2, \ldots) = \boldsymbol{\theta}$.

To see this, we compute conditionally on $\boldsymbol{\theta}$: now

$$E\left(\sum_{i=1}^n Y_i|\boldsymbol{\theta}\right) = n\boldsymbol{\theta}, \qquad Var\left(\sum_{i=1}^n Y_i|\boldsymbol{\theta}\right) = n\boldsymbol{\theta}(1-\boldsymbol{\theta}).$$

Hence, with $\widehat{\theta}_n \equiv \overline{Y}_n$,

$$E(\widehat{\theta}_n - \boldsymbol{\theta})^2 \le \frac{1/4}{n},$$

and by Chebychev's inequality, $\widehat{\theta}_n$ converges in probability to $\boldsymbol{\theta}$. Therefore $\widehat{\theta}_{n_k} \rightarrow \boldsymbol{\theta}$ almost surely for some subsequence n_k. Hence $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$ a.s. where $\tilde{\boldsymbol{\theta}} \equiv \lim_k \widehat{\theta}_{n_k}$ is measurable with respect to $Y_1, Y_2, \ldots$. Thus we have

$$E(\boldsymbol{\theta}|Y_1, Y_2, \ldots) = E(\tilde{\boldsymbol{\theta}}|Y_1, Y_2, \ldots) = \tilde{\boldsymbol{\theta}} = \boldsymbol{\theta} \qquad \text{a.s. } P_\Lambda$$

where

$$P_\Lambda(\underline{Y} \in A,\ \boldsymbol{\theta} \in B) = \int_B P_\theta(\underline{Y} \in A) d\Lambda(\theta).$$

Therefore

$$P_\Lambda(d(\underline{Y}_n) \rightarrow \boldsymbol{\theta}) = \int_{[0,1]} P_\theta(d(\underline{Y}_n) \rightarrow \theta) d\Lambda(\theta) = 1,$$

and this implies that

$$P_\theta(d(\underline{Y}_n) \rightarrow \theta) = 1 \qquad \text{a.e. } \Lambda.$$

Thus the Bayes estimator $d(\underline{Y}_n) \equiv E(\boldsymbol{\theta}|\underline{Y}_n)$ is consistent for Λ a.e. θ.

A general result of this type for smooth, finite - dimensional families was established by Doob (1948), and has been generalized by Breiman, Le Cam, and Schwartz (1964), Freedman (1963), and Schwartz (1965). See van der Vaart (1998), pages 149-151 for the general result. The upshot is that for smooth, finite - dimensional families (θ, Λ) is consistent if an only if θ is in the support of Λ. However the assumption of finite-dimensionality is important: Freedman (1963) gives an example of an inconsistent Bayes rule when the sample space is $\mathcal{X} = \mathbb{Z}^+$ and Θ is the collection of all probabilities on $\mathcal{X}$. The paper by Diaconis and Freedman (1986) on the consistency of Bayes estimates is an outgrowth of Freedman (1963).

Asymptotic Efficiency of Bayes Estimators

Now we consider a more general situation, but still with a real-valued parameter θ. Let $X_1, \ldots, X_n$ be i.i.d. $p_\theta(x)$ with respect to μ where $\theta \in \Theta \subset \mathbb{R}$, and $\theta_0 \in \Theta$ (an open interval) is true; write P for P_{θ_0}.

Assumptions:
B1: (a) - (f) of Theorem 2.6 (TPE, pages 440 - 441) hold. Thus it follows that

$$(1) \qquad l_n(\theta) - l_n(\theta_0) = (\theta - \theta_0)\dot{\mathbf{l}}_\theta(\underline{X}; \theta_0) - \frac{1}{2}(\theta - \theta_0)^2\{nI(\theta_0) + R_n(\theta)\}$$

where $n^{-1}R_n(\theta) \to_p 0$; see problems 6.3.22 and 6.8.3, TPE, pages 503 and 514. We strengthen this to:
B2: for any $\epsilon > 0$ there exists a $\delta > 0$ such that

$$P\left\{\sup_{|\theta-\theta_0|\le\delta} |n^{-1}R_n(\theta)| \ge \epsilon\right\} \to 0 \qquad \text{as } n \to \infty.$$

B3: For any $\delta > 0$ there exists an $\epsilon > 0$ such that

$$P\left\{\sup_{|\theta-\theta_0|\ge\delta} (l_n(\theta) - l_n(\theta_0)) \le -\epsilon\right\} \to 1.$$

B4: The prior density λ of $\boldsymbol{\theta}$ is continuous and > 0 for all $\theta \in \Theta$.
B5: $E_\lambda|\boldsymbol{\theta}| < \infty$.

Theorem 8.1 Suppose that $\lambda^*(t|\underline{X})$ is the posterior density of $\sqrt{n}(\theta - T_n)$ where

$$T_n \equiv \theta_0 + \frac{1}{I(\theta_0)}\{n^{-1}\dot{\mathbf{l}}_\theta(\underline{X}; \theta_0)\}.$$

(i) If B1 - B4 hold, then

$$d_{TV}(\Lambda^*(\cdot|\underline{X}), N(0, 1/I(\theta_0))) = \int |\lambda^*(t|\underline{X}) - \sqrt{I(\theta_0)}\phi(t\sqrt{I(\theta_0)})|dt \to_p 0.$$

(ii) If B1 - B5 hold, then

$$\int (1 + |t|)|\lambda^*(t|\underline{X}) - \sqrt{I(\theta_0)}\phi(t\sqrt{I(\theta_0)})|dt \to_p 0.$$

Theorem 8.2 If B1 - B5 hold and $\tilde{\theta}_n$ is the Bayes estimator with respect to λ for squared error loss, then

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) \to_d N(0, 1/I(\theta_0)),$$

so that $\tilde{\theta}_n$ is $\sqrt{n}$−consistent and asymptotically efficient.

Example 8.1 Suppose that $X_1, \ldots, X_n$ are i.i.d. $N(\theta, \sigma^2)$ and that the prior on θ is $N(\mu, \tau^2)$. Then the posterior distribution is given by $(\boldsymbol{\theta}|\overline{X}) \sim N(p_n\overline{X} + (1 - p_n)\mu, \sigma_n^2/n)$ where

$$p_n \equiv \frac{n/\sigma^2}{n/\sigma^2 + 1/\tau^2}, \qquad \sigma_n^2 = \frac{n}{n/\sigma^2 + 1/\tau^2}$$

so that $p_n \to 1$, $\sqrt{n}(1-p_n) \to 0$, and $\sigma_n^2 \to \sigma^2$. Note that

$$T_n = \theta_0 + \frac{1}{1/\sigma^2}\frac{(\overline{X}-\theta_0)}{\sigma^2} = \overline{X},$$

and the Bayes estimator with squared error loss is

$$E(\boldsymbol{\theta}|\overline{X}) = p_n\overline{X} + (1-p_n)\mu.$$

Thus

$$\begin{aligned}(\sqrt{n}(\boldsymbol{\theta}-\overline{X})|\overline{X}) &\sim N(\sqrt{n}(1-p_n)(\mu-\overline{X}), \sigma_n^2) \\ &\to_d N(0,\sigma^2) = N(0, 1/I(\theta_0)) \qquad \text{a.s.}\end{aligned}$$

in agreement with theorem 8.1, while

$$\begin{aligned}\sqrt{n}(E(\boldsymbol{\theta}|\overline{X})-\theta_0) &= \sqrt{n}(p_n\overline{X}+(1-p_n)\mu-\theta_0) \\ &= p_n\sqrt{n}(\overline{X}-\theta_0)+\sqrt{n}(1-p_n)(\mu-\theta_0) \\ &\stackrel{d}{=} p_nN(0,\sigma^2)+\sqrt{n}(1-p_n)(\mu-\theta_0) \\ &\to_d N(0,\sigma^2)\end{aligned}$$

in agreement with theorem 8.2.

Proof. We first suppose that theorem 8.1 is proved, and show that it implies theorem 8.2. Note that

(a) $\sqrt{n}(\tilde{\theta}_n-\theta_0) = \sqrt{n}(\tilde{\theta}_n-T_n)+\sqrt{n}(T_n-\theta_0)$.

Since

(b) $\sqrt{n}(T_n-\theta_0) \to_d N(0, 1/I(\theta_0))$,

it suffices to show that

(c) $\sqrt{n}(\tilde{\theta}_n-T_n) \to_p 0$.

To prove (c), note that by definition of $\tilde{\theta}_n$

$$\tilde{\theta}_n = \int \theta\lambda(\theta|\underline{X})d\theta = \int \left\{\frac{t}{\sqrt{n}}+T_n\right\}\lambda^*(t|\underline{X})dt,$$

and hence

$$\sqrt{n}(\tilde{\theta}_n-T_n) = \int t\lambda^*(t|\underline{X})dt - \int t\sqrt{I(\theta_0)}\phi(t\sqrt{I(\theta_0)})dt.$$

It folows that

$$\sqrt{n}|\tilde{\theta}_n-T_n| \le \int |t||\lambda^*(t|\underline{X})-\sqrt{I(\theta_0)}\phi(t\sqrt{I(\theta_0)})|dt \to_p 0$$

as $n\to\infty$ by theorem 8.1(ii). Thus (c) holds, and the proof is complete; it remains only to prove theorem 8.1.

Proof of theorem 8.1: (i) By definition of T_n

$$\begin{aligned} \lambda^*(t|\underline{X}) &= \frac{\lambda(T_n + n^{-1/2}t)\exp(l(T_n + n^{-1/2}t))}{\int \lambda(T_n + n^{-1/2}u)\exp(l(T_n + n^{-1/2}u))du} \\ &\equiv e^{\omega(t)}\lambda(T_n + n^{-1/2}t)/C_n \end{aligned}$$

where

(d) $$\omega(t) \equiv l(T_n + n^{-1/2}t) - l(\theta_0) - \frac{1}{2nI(\theta_0)}\{\dot{\mathbf{l}}_\theta(\underline{X};\theta_0)\}^2$$

and

(e) $$C_n \equiv \int e^{\omega(u)}\lambda(T_n + n^{-1/2}u)du.$$

Now if we can show that

(f) $$J_{1n} \equiv \int |e^{\omega(t)}\lambda(T_n + n^{-1/2}t) - \exp(-t^2 I(\theta_0)/2)\lambda(\theta_0)|dt \to_p 0,$$

then

(g) $$C_n \to_p \int \exp(-t^2 I(\theta_0)/2)\lambda(\theta_0)dt = \lambda(\theta_0)\sqrt{2\pi/I(\theta_0)},$$

and the left side of (2) is J_n/C_n where

(h) $$J_n \equiv \int |e^{\omega(t)}\lambda(T_n + n^{-1/2}t) - C_n\sqrt{I(\theta_0)}\phi(t\sqrt{I(\theta_0)})|dt.$$

Furthermore $J_n \le J_{1n} + J_{2n}$ where J_{1n} is defined in (f) and

(i) $$\begin{aligned} J_{2n} &\equiv \int |C_n\sqrt{I(\theta_0)}\phi(t\sqrt{I(\theta_0)}) - \exp(-t^2 I(\theta_0)/2)\lambda(\theta_0)|dt \\ &= \left|\frac{C_n\sqrt{I(\theta_0)}}{\sqrt{2\pi}} - \lambda(\theta_0)\right| \int \exp(-t^2 I(\theta_0)/2)dt \to_p 0 \end{aligned}$$

by (g). Hence it remains only to prove that (f) holds.

(ii) The same proof works with J'_{1n} replacing J_{1n} where J'_{1n} has an additional factor of $(1+|t|)$. □

Lemma 8.1 The following identity holds:

(2) $$\begin{aligned} \omega(t) &\equiv l(T_n + n^{-1/2}t) - l(\theta_0) - \frac{1}{2nI(\theta_0)}\{\dot{\mathbf{l}}_\theta(\underline{X};\theta_0)\}^2 \\ &= -I(\theta_0)\frac{t^2}{2} - \frac{1}{2n}R_n(T_n + n^{-1/2}t)\left\{t + \frac{1}{I(\theta_0)}\frac{\dot{\mathbf{l}}_\theta(\underline{X};\theta_0)}{\sqrt{n}}\right\}^2. \end{aligned}$$

Proof. This follows immediately from the definition of T_n, (1), and algebra. $\square$

Proof. that $J_{1n} \rightarrow_p 0$: Recall the definition of J_{1n} given in (f) of the proof of theorem 8.1. Divide the region of integration into:
(i) $|t| \le M$;
(ii) $M \le |t| \le \delta\sqrt{n}$; and
(iii) $\delta\sqrt{n} < t < \infty$.
For the region (i), the integral is bounded by $2M$ times the supremum of the integrand over the set $|t| \le M$, and by application of lemma 8.1 it is seen that this supremum converges to zero in probability if

$$\text{(a)} \qquad \sup_{|t|\le M} \left| \frac{1}{n} R_n(T_n + n^{-1/2}t) \left(t + \frac{1}{I(\theta_0)} \frac{\dot{\mathbf{l}}_\theta(\underline{X};\theta_0)}{\sqrt{n}} \right)^2 \right| \rightarrow_p 0$$

and

$$\text{(b)} \qquad \sup_{|t|\le M} |\lambda(T_n + n^{-1/2}t) - \lambda(\theta_0)| \rightarrow_p 0.$$

Now λ is continuous by B4 and $T_n \rightarrow_p \theta_0$ by B1, so λ is uniformly continuous in a neighborhood of θ_0 and

$$\text{(c)} \qquad \theta_0 - \delta \le T_n - n^{-1/2}M \le T_n + n^{-1/2}t \le T_n + n^{-1/2}M \le \theta_0 + \delta$$

for $|t| \le M$, so (b) holds. Now $n^{-1/2}\dot{\mathbf{l}}_\theta(\underline{X};\theta_0)$ is bounded in probability (by B1 and the central limit theorem), so it follows from (c) and B2 that (a) holds.

(ii) For the region $M \le |t| \le \delta\sqrt{n}$ it suffices to show that the integrand is bounded by an integrable function with probability close to one, since then the integral can be made small by choosing M sufficiently large. Since the second term of the integrand in J_{1n} is integrable, it suffices to find such an integrable bound for the first term. In particular, it can be shown that for every $\epsilon > 0$ there exists a $C = C(\epsilon, \delta)$ and an integer $N = N(\epsilon, \delta)$ so that, for $n \ge N$,

$$P(\exp(\omega(t)\lambda(T_n + n^{-1/2}t) \le C\exp(-t^2 I(\theta_0)/4) \quad \text{for all} \quad |t| \le \delta\sqrt{n}) \ge 1 - \epsilon;$$

this follows from the assumption B2; see TPE, pages 494-496.

(iii) The integral over the region $|t| \ge \delta\sqrt{n}$ converges to zero by an argument similar to that for the region (ii), but now the assumption B3 comes into play; see TPE pages 495-496.

The proof for J_{1n}' requires only trivial changes using assumption B5. $\square$

Remark 8.1 This proof is from Lehmann and Casella, TPE, pages 489 - 496. This type of theorem apparently dates back to Laplace (1820), and was later rederived by Bernstein (1917), and von Mises (1931). More general versions have been established by Le Cam (1958), Bickel and Yahav (1969), and Ibragimov and Has'minskii (1972). See the discussion on pages 488 and 493 of TPE. For a recent treatment along the lines of Le Cam (1958) that covers the case of $\Theta \subset \mathbb{R}^d$, see chapter 10 of van der Vaart (1998). For a version with the true measure Q generating the data not in $\mathcal{P}$, see Hartigan (1983).

Remark 8.2 See Ibragimov and Has'minskii (1981) and Strasser (1981) for more on the consistency and efficiency of Bayes estimators.

Remark 8.3 Analogues of these theorems for nonparametric and semiparametric problems are currently an active research area. See Ghosal, Ghosh, and van der Vaart (2000), Ghosal and van der Vaart (2001), and Shen (2002).

Now we will continue the development for convex likelihoods started in section 4.7. The notation will be as in section 4.7, and we will make use of lemmas 4.7.1 and 4.7.2. Here, as in section 4.7, we will *not assume* that the distribution P governing the data is a member of the parametric model $\mathcal{P} = \{P_\theta : \ \theta \in \Theta \subset \mathbb{R}^d\}$.

Here are the hypotheses we will impose.

C1. $\lambda(\theta) \le C_1 \exp(C_2|\theta|)$ for all θ for some positive constants C_1, C_2.

C2. λ is continuous at θ_0.

C3. $\log p(x, \theta_0 + t) - \log p(x, \theta_0) = \psi(x)^T t + R(x; t)$ is concave in t.

C4. $E_P \psi(X_1)\psi(X_1)^T \equiv K$, $E_P \psi(X_1) = 0$.

C5. $E_P R(X_1; t) = -t^T Jt/2 + o(|t|^2)$ and $\mathrm{Var}_P(R(X_1; t)) = o(|t|^2)$ where J is symmetric, positive definite.

C6. $X_1, \ldots, X_n$ are i.i.d. P.

Theorem 8.3 Suppose that C1 - C6 hold. Then the maximum likelihood estimator $\hat{\theta}_n$ is $\sqrt{n}-$consistent for $\theta_0 = \theta_0(P)$ and satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d N_d(0, J^{-1}K(J^{-1})^T).$$

Moreover the Bayes estimator $\tilde{\theta}_n = E\{\boldsymbol{\theta}|\underline{X}_n\}$ satisfies

$$\sqrt{n}(\tilde{\theta}_n - \hat{\theta}_n) \to_p 0,$$

and hence also

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) \to_d N_d(0, J^{-1}K(J^{-1})^T).$$

Proof. Let $L_n(\theta) \equiv \prod_{i=1}^n p(X_i, \theta)$ denote the likelihood function. Define random convex functions A_n by

$$\exp(-A_n(t)) = L_n(\widehat{\theta}_n + t/\sqrt{n})/L_n(\widehat{\theta}_n).$$

By definition of the MLE, A_n achieves its minimum value of zero at $t = 0$. Then

$$\begin{aligned} \tilde{\theta}_n &= \frac{\int \theta L_n(\theta)\lambda(\theta)d\theta}{\int L_n(\theta)\lambda(\theta)d\theta} \\ &= \widehat{\theta}_n + \frac{1}{\sqrt{n}} \frac{t\exp(-A_n(t))\lambda(\widehat{\theta}_n + t/\sqrt{n})\exp(-C_2|\widehat{\theta}_n|)dt}{\exp(-A_n(t))\lambda(\widehat{\theta}_n + t/\sqrt{n})\exp(-C_2|\widehat{\theta}_n|)dt} \end{aligned}$$

by the change of variable $\theta = \widehat{\theta}_n + t/\sqrt{n}$.

Claim: The random functions A_n converge in probability uniformly on compact sets to $t^T Jt/2$.

Proof. Let

$$A_n^0(t) = n\mathbb{P}_n\{\log p(x;\theta_0 + t/\sqrt{n}) - \log p(x;\theta_0)\}.$$

From the proof of theorem 4.7.6 (with $h(x;\theta) \equiv -\log p(x;\theta)$ and multiplication by -1),

$$\text{(a)} \qquad A_n^0(t) = U_n^T t + \frac{1}{2}t^T J t + o_p(1),$$

and this holds uniformly in t in compact sets by Lemma 4.x.y. Note that

$$A_n(t) = A_n^0(t + \sqrt{n}(\widehat{\theta}_n - \theta_0)) - A_n^0(\sqrt{n}(\widehat{\theta}_n - \theta_0)).$$

Since $\sqrt{n}(\widehat{\theta}_n - \theta_0) = O_p(1)$ so that for every $\epsilon > 0$ there exists a compact set $K \equiv K_\epsilon$ such that $P(\sqrt{n}(\widehat{\theta}_n - \theta_0) \in K) \geq 1 - \epsilon$. Hence by the uniformity in t in the convergence in (a),

$$\begin{aligned}
A_n(t) &= U_n^T(t + \sqrt{n}(\widehat{\theta}_n - \theta_0)) - U_n^T(\sqrt{n}(\widehat{\theta}_n - \theta_0)) \\
&\qquad + \frac{1}{2}\{(t + \sqrt{n}(\widehat{\theta}_n - \theta_0))^T J (t + \sqrt{n}(\widehat{\theta}_n - \theta_0)) \\
&\qquad - \sqrt{n}(\widehat{\theta}_n - \theta_0)^T J \sqrt{n}(\widehat{\theta}_n - \theta_0)\} + o_p(1) \\
&= U_n^T t + \frac{1}{2}t^T J t + \sqrt{n}(\widehat{\theta}_n - \theta_0)^T J t + o_p(1) \\
&= \frac{1}{2}t^T J t + o_p(1)
\end{aligned}$$

where the last line follows because $\sqrt{n}(\widehat{\theta}_n - \theta_0) = -J^{-1}U_n + o_p(1)$. Since A_n is convex and converges pointwise in probability to $t^T J t/2$, it converges uniformly in probability on compact subsets by lemma 4.7.1. □

Now we return to the main proof. Define $\gamma_n \equiv \inf_{|u|=1} A_n(u)$. It converges in probability to $\gamma_0 = \inf_{|u|=1} u^T J u/2 > 0$. By the same argument used in lemma 4.7.2 it follows that

$$A_n(t) \geq \gamma_n |t| \qquad \text{for} \quad |t| > 1.$$

Now the integrand in the numerator above converges in probability for each fixed t to

$$t\exp(-t^T J t/2)\lambda(\theta_0)\exp(-C_2|\theta_0|),$$

and similarly the integrand of the denominator converges for each fixed t to

$$\exp(-t^T J t/2)\lambda(\theta_0)\exp(-C_2|\theta_0|),$$

Furthermore the domination hypotheses of the following convergence lemma hold for both the numerator and denominator with dominating function

$$D(t) \equiv 2C_1 1\{|t| \leq 1\} + C_1|t|\exp(-\gamma_0|t|/2)1\{|t| > 1\}.$$

Hence the ratio of integrals converges in probability to

$$\frac{\int t\exp(-t^T J t/2)\lambda(\theta_0)\exp(-C_2|\theta_0|)dt}{\int \exp(-t^T J t/2)\lambda(\theta_0)\exp(-C_2|\theta_0|)} = 0.$$

□

Lemma 8.2 Suppose that $X_n(t)$, $Y_n(t)$ are jointly measurable random functions, $X_n(t,\omega)$, $Y_n(t,\omega)$ for $(t,\omega) \in K \times \Omega$ where $K \subset \mathbb{R}^d$ is compact, that λ is a measure on $\mathbb{R}^d$, and :
(i) $Y_n(t) \to_p Y(t)$ and $X_n(t) \to_p X(t)$ for λ almost all $t \in K$.
(ii) $\int_K Y_n(t)d\lambda(t) \to_p \int Y(t)d\lambda(t)$ with $|\int_K Y(t)d\lambda(t)| < \infty$ almost surely.
Then $\int_K X_n(t)d\lambda(t) \to_p \int_K X(t)d\lambda(t)$.

Proof. It suffices to show almost sure convergence for some further subsequence of any given subsequence. By convergence in probability for fixed t, and then dominated convergence

$$\begin{aligned} H_n(\epsilon) &\equiv (P \otimes \Lambda)(\{(\omega,t) : \ |X_n(t,\omega) - X(t,\omega)| > \epsilon) \\ &= \int_K P(\{\omega : \ |X_n(t,\omega) - X(t,\omega)| > \epsilon\})dt \to 0 \end{aligned}$$

by the dominated convergence theorem, and similarly for $\{Y_n\}$. By replacing ϵ by $\epsilon_n \downarrow 0$ and extraction of subsequences we can find a subsequence $\{n'\}$ so that the integrals $\int_K X_n d\lambda$ converge and for some set $N \subset K \times \Omega$ with $P \otimes \lambda(N) = 0$ we get convergence for all $(\omega) \in N^c$ of both $X_{n'}$ and $Y_{n'}$. Then we have $\lambda(\{t : (\omega,t) \in N\}) = 0$ for almost all ω. Hence by Fatou's lemma applied to $Y_{n'} \pm X_{n'}$ we deduce that

$$\int_K X_{n'}(t)d\lambda(t) \to_{a.s.} \int_K X(t)d\lambda(t).$$

□

Corollary 1 Suppose that:
(i) $G_n(t) \to_p G(t)$ for each fixed $t \in \mathbb{R}^d$.
(ii) $P(|G_n(t)| \le D(t)$ for all $t \in \mathbb{R}^d) \to 1$.
(iii) $\int D(t)dt < \infty$.
Then $\int G_n(t)dt \to_p \int G(t)dt$.

Proof. Let $\epsilon > 0$; choose a compact set $K = K_\epsilon$ so that $\int_{K^c} D(t)dt < \epsilon$. This is possible since $\int D(t)dt < \infty$. Then, using $|G(t) \le D(t)$,

$$\begin{aligned} \Big|\int G_n(t)dt - \int G(t)dt\Big| &\le \Big|\int_K G_n(t) - \int_K G(t)dt\Big| + \int_{K^c} |G_n(t)|dt + \int_{K^c} |G(t)|dt \\ &\le \Big|\int_K G_n(t)1_{[|G_n(t)| \le D(t)]}dt - \int_K G(t)dt\Big| \\ &\quad + \int_{K \cap [|G_n(t)| > D(t)]} |G_n(t)|dt + \int_{K^c \cap [|G_n(t)| > D(t)]} |G_n(t)|dt \\ &\quad + 2\int_{K^c} D(t)dt. \end{aligned}$$

Now apply lemma 8.2 with $Y_n(t) \equiv D(t)$ and

$$X_n(t) \equiv G_n(t)1_{[|G_n(t)| \le D(t)]}.$$

Then $X_n(t) \to_p G(t)$ and the second hypothesis of the lemma holds easily. Hence $\int X_n(t)dt \to_p \int G(t)dt$ so that the first term above $\to_p 0$. The second and third terms converge in probability to zero because the set over which the integral is take is empty with arbitrarily high probability; and the third term is bounded by 2ϵ by choice of K. □