Chapter 4: Statistical Decision Theory *

4 Statistical Decision Theory

4.1 Basic Framework and Concepts

- To the usual statistical modeling framework from earlier

$$X, \quad \Theta, \quad \mathcal{P} = \{P_\theta : \theta \in \Theta\}$$

  we add the following elements

  1. some "action space" $\mathcal{A}$ with σ-algebra ϵ,
  2. a suitably measurable "loss function"

$$L(\theta, a) : \Theta \times \mathcal{A} \to [0, \infty),$$

  3. and (non-randomized) decision rules

$$\delta(x) : (\mathcal{X}, \mathcal{B} \to (\mathcal{A}, \epsilon))$$

  For data X, $\delta(x)$ is the action taken based on X.

  To identify "good" devusuib rules δ, we have to average our X, which naturally leads to expectation.

- Risk function The mapping from $\Theta \to [0, \infty)$ given by

$$R(\theta, \delta) \equiv R_\theta L(\theta, \delta(X)) = \int_{\mathcal{X}} L(\theta, \delta(x)) dP_\theta(x)$$

  is call the risk function for θ.

  - δ is *at least as good as* δ' if $R(\theta, \delta) \le R(\theta, \delta')$ for all $\theta \in \Theta$
  - δ is *better than* δ' if $R(\theta, \delta) \le R(\theta, \delta')$ for all $\theta \in \Theta$, and $R(\theta_0, \delta) < R(\theta_0, \delta')$ for some θ_0
  - δ and δ' are *risk equivalent* if $R(\theta, \delta) = R(\theta, \delta')$ for all $\theta \in \Theta$.
  - δ is *best in a class of decision rules* Δ if $\delta \in \Delta$, and δ is at least as good as any other $\delta' \in \Delta$
  - Example: $X \sim N(\theta, 1), \theta \in \mathbb{R}$ with $\Delta =$"the class of all estimators of θ". There is no best element here. Prove by proposing two constant estimators and zero-one loss.

- If there is no best estimator,
  - Try a smaller and appropriate Δ, e.g. unbiased estimators.
  - Reduce the risk function $R(\theta, \delta)$ to a number and compare numbers for different δ's, e.g.: averaging over θ according to some distribution G on Θ is a way to make "Bayes Risk" and look for "Bayes optimal " decision rules.
  - Maximize $R(\theta, \delta)$ over θ and seek to minimize over δ's, i.e. mini-max procedures.

- Inadmissible: δ is inadmissible in Δ if there exists $\delta' \in \Delta$ that is better than δ.

- Admissible: δ is admissible in Δ if it is not inadmissible in Δ.

  Note: One may never want to use an inadmissible rule, but there are decision problems where every rule is inadmissible.

- Behavorial decision rule: If for each $x \in \mathcal{X}, \phi_x$ is a distribution on $(\mathcal{A}, \epsilon)$, then ϕ_x is called a behavorial decision rule.

*STA643: Advanced Theory of Statistical Inference. Instructed by Dr. Daniel Nordman. Arranged by Zhiling Gu

- $\mathcal{D}^* \equiv \{\phi_x\} \equiv$ the class of behavorial decision rules
- $\mathcal{D} \subset \mathcal{D}^*$ where

$$\mathcal{D} \equiv \{\delta(x)\} \equiv \text{the class of non-randomized decision rules } \delta : \mathcal{X} \to \mathcal{A}$$

- The risk function of a behaorial decision rule is defined as

$$R(\theta, \phi) = \int_{\mathcal{X}} \int_{\mathcal{A}} L(\theta, a) d\phi_x(a) dP_\theta(x)$$

- Randomized decision rule: A randomized decision rule ψ is a probability measure on $(\mathcal{D}, \mathcal{F})$ (δ, with a distribution ψ, becomes a random object and we take decision $\delta(X)$.) Notes:
  - Let $\mathcal{D}_* \equiv \{\psi\} \equiv$ the class of randomized decision rules.
  - It's possible to think of

$$\mathcal{D} \subset \mathcal{D}_*$$

  by associating with $\delta \in \mathcal{D}$ a randomized decision rule ψ_δ which places mass 1 on δ (i.e. , $\psi_\delta(\{\delta\}) = 1$)
  - The risk function of a randomized decision rule is defined as

$$R(\theta, \psi) = \int_{\mathcal{D}} R(\theta, \delta) d\psi(\delta) = \int_{\mathcal{D}} \int_{\mathcal{X}} L(\theta, \delta(x)) dP_\theta(x) d\psi(\delta)$$

- Among $(\mathcal{D}, \mathcal{D}^*, \mathcal{D}_*)$, $\mathcal{D}^*$ is perhaps the most natural, while $\mathcal{D}_*$ is the easiest to deal with in some proofs. A natural question is "When are $\mathcal{D}^*$ and $\mathcal{D}_*$ equivalent in termes of generating the same set of risk functions?' It is typically the case under certion space, distribution and loss functions conditions.
  - Example: $(\mathcal{D}, \mathcal{D}^*, \mathcal{D}_*)$ where Behavioural rule and Randomized rule has the same risk function.

$$\begin{aligned}
&X \sim \text{ Bernoulli } (p), \text{ Estimation of } p \in \Theta \equiv [0,1] \equiv \mathcal{A} \\
&\mathcal{X} = \{0,1\}, \quad \mathcal{A} = [0,1], \quad \delta \in D \iff (\delta(0), \delta(1)) \in [0,1] \times [0,1] \equiv \mathcal{A}_0 \\
&\mathcal{D} = \{\delta(x) : \mathcal{X} \to \mathcal{A}\} = \{\delta(x) \mid x = 0,1 \text{ and } \delta(0), \delta(1) \in [0,1]\} \\
&\mathcal{D}^* = \{\phi_x \mid x = 0,1 \text{ and } \phi_0, \phi_1 \text{ are distributions on } \mathcal{A} \equiv [0,1]\} \\
&\mathcal{D}_* = \{\psi \mid \psi \text{ is a probability measure on } (\mathcal{D}, \mathcal{F})\}
\end{aligned}$$

    * $\delta(0) = 0.3$, $\delta(1) = 0.7$ is non-randomized rule
    * $\phi_{X=0} \sim U(0, 0.5), \phi_{X=1} \sim U(0.5, 1)$ then $\phi_X \in D^*$
    * ψ on D, where ψ has a uniform distribution on $(0, 0.5) \times (0.5, 1)$
      Note: if $\tilde{\delta}$ is randomly chosen according to ψ then we observe $X \in \{0,1\}$, we take $\tilde{\delta}(0)$ if $X = 0$, $\tilde{\delta}(1)$ if $X = 1$, so $\psi \in D_*$. That is, first determine the rule, then plug in the observed X.
    * Remark: ϕ_X and ψ in this case are equivalent because

$$\tilde{\delta}(0) \sim U(0, 0.5) \quad \tilde{\delta}(1) \sim U(0.5, 1)$$

- When D^*, D_* contain better tules that are better than those in D? For convex loss functions, rules in D^*, D_* are typically no better.
  - Lemma 51: Suppose that $\mathcal{A}$ is a convex subset of $\mathbb{R}^d$ and ϕ_x is a behavioral decision rule. Define a non-randomized decision rule by

$$\delta(x) = \int_{\mathcal{A}} a d\phi_x(a)$$

  assuming the integral exists. (In the case that $d > 1$, interpret $\delta(x)$ as vector-valued, and the integral as a vector of integrals over d coordinates of $a \in \mathcal{A}$.)
  1. If $L(\theta, \cdot) : \mathcal{A} \to [0, \infty)$ is convex, then

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

  2. If $L(\theta, \cdot) : \mathcal{A} \to [0, \infty)$ is strictly convex, $R(\theta, \phi) < \infty$ and $P_\theta(\{x \mid \phi_x \text{ is non-degenerate }\}) > 0$, then

$$R(\theta, \delta) < R(\theta, \phi)$$

  Prove by Jensen's Inequality. This lemma shows randomization does not hlep in picking the best decisions. Next two lemmas shows averaging out the randomzation will improve convex loss function.

- Corollary 52: Suppose that $\mathcal{A}$ is a convex subset of $\mathbb{R}^d, \phi_x$ is a behavioral decision rule, and

$$\delta(x)=\int_{\mathcal{A}} a d \phi_x(a)$$

assuming the integral exists.

1. If $L(\theta, a): \mathcal{A} \rightarrow[0, \infty)$ is convex in a for all θ, then δ is at least as good as ϕ
2. If $L(\theta, a): \mathcal{A} \rightarrow[0, \infty)$ is convex in a for all θ and, for some θ_0, the function $L\left(\theta_0, a\right): \mathcal{A} \rightarrow[0, \infty)$ is strictly convex in $a, R\left(\theta_0, \phi\right)<\infty$ and $P_{\theta_0}(\{x \mid \phi_x$ is non-degenerate $\})>0$, then δ is better than ϕ

4.2 Finite Dimensional Geometry of Decision Theory

- A helpful device for understanding some of the basics of decision theory is the geometry involved when

$$\Theta=\left\{\theta_1, \ldots, \theta_k\right\}$$

Assume that $R(\theta, \psi)<\infty$ for all $\theta \in \Theta$ and $\psi \in \mathcal{D}_*$. Note that in this case

$$R(\cdot, \psi): \Theta \rightarrow[0, \infty)$$

corresponds to a k -vector in $[0, \infty)^k$

Let $\mathcal{S}=\left\{y_\psi=\left(y_1, y_2, \ldots, y_k\right) \in \mathbb{R}^k \mid y_i=R\left(\theta_i, \psi\right)\right.$ for all i and some $\left.\psi \in \mathcal{D}_*\right\}=$ the set of all randomized risk vectors.

- Theorem 53: $\mathcal{S}$ is a convex set in $\mathbb{R}^k$
- Let $\mathcal{S}^0=\left\{y_\delta=\left(y_1, y_2, \ldots, y_k\right) \in \mathbb{R}^k \mid y_i=R\left(\theta_i, \delta\right)\right.$ for all i and some $\left.\delta \in \mathcal{D}\right\}=$ the set of all non-randomized risk vectors. It turns out that $\mathcal{S}$ is the convex hull of $\mathcal{S}^0$ (or, equivalently, the smallest convex set containing $\mathcal{S}^0$ or the set of all convex combinations of points in $\mathcal{S}^0$ or the intersection of all convex sets containing $\mathcal{S}^0$)
- Lower Quadrant: Definition 54: For $x=\left(x_1, \ldots, x_k\right) \in \mathbb{R}^k$, the lower quadrant of x is

$$Q_x=\left\{z=\left(z_1, \ldots, z_k\right) \in \mathbb{R}^k \mid z_i \leq x_i \text { for all } i=1, \ldots, k\right\}$$

- Theorem 55: $y \in \mathcal{S}$ (or the decision rule giving rise to y) is admissible if and only if

$$Q_y \cap \mathcal{S}=\{y\}$$

- Definition 56: For $\overline{\mathcal{S}}$ the closure of $\mathcal{S}$, the lower boundary of $\mathcal{S}$ is

$$\lambda(\mathcal{S})=\left\{y \in \mathbb{R}^k \mid Q_y \cap \overline{\mathcal{S}}=\{y\}\right\}$$

- Definition 57 : $\mathcal{S}$ is closed from below if $\lambda(\mathcal{S}) \subset \mathcal{S}$ Denote the set of admissible risk points as

$$A(\mathcal{S})=\left\{y \in \mathbb{R}^k \mid Q_y \cap \mathcal{S}=\{y\}\right\}$$

- Theorem 58: If $\mathcal{S}$ is closed (i.e., $\mathcal{S}=\overline{\mathcal{S}}$), then $\lambda(\mathcal{S})=A(\mathcal{S})$
- Theorem 59: If $\mathcal{S}$ is closed from below, then $\lambda(\mathcal{S})=A(\mathcal{S})$.

4.3 Complete Classes of Decision Rules

- Complete Class Definition 60: A class of decision rules $\mathcal{C} \subset \mathcal{D}^*$ is a complete class (for $\mathcal{D}^*$) if, for any given $\phi \notin \mathcal{C}$, there exists $\phi^{\prime} \in \mathcal{C}$ such that $\phi^{\prime}$ is better than ϕ.

  Remark: This indicates $\mathcal{C}$ contains the best rules that we should focus on.

- Essentially Complete Class Definition 61 : $\mathcal{C} \subset \mathcal{D}^*$ is a called an essentially complete class (for $\mathcal{D}^*$) if, for any given $\phi \notin \mathcal{C}$, there exists $\phi^{\prime} \in \mathcal{C}$ such that $\phi^{\prime}$ is at least as good as ϕ.
- Minimal Complete Class Definition 62 : $\mathcal{C} \subset \mathcal{D}^*$ is a minimal complete class for $\mathcal{D}^*$ if $\mathcal{C}$ is complete and is a subset of any other complete class for $\mathcal{D}^*$. Denote the set of admissible rules in $\mathcal{D}^*$ as $A\left(\mathcal{D}^*\right)$ in the following results.
- Theorem 63: If a minimal complete class $\mathcal{C}$ exists, then $\mathcal{C}=A\left(\mathcal{D}^*\right)$
- Theorem 64: If $A\left(\mathcal{D}^*\right)$ is a complete class, then $A\left(\mathcal{D}^*\right)$ is a minimal complete class.

  Note: The statement " $A\left(\mathcal{D}^*\right)$ is a minimal complete class" is, in general, incorrect. Minimimal complete class does not always exists, when $\mathcal{S}$ is not closed and does not contain the minimum Q_y.

4.4 Sufficiency and Decision Theory

- Theorem 65: If T is sufficient for $\mathcal{P}$ and ϕ is a behavioral decision rule, then there exists another behavioral decison rule ϕ' that is a function of T and has the same risk function as ϕ. (Having ϕ' as a function of T means that for $x, y \in \mathcal{X}$ with $T(x)=T(y)$ it must be that ϕ_x' and ϕ_y' are the same distributions on $\mathcal{A}$.

  Think in this fashion: recall $(\mathcal{A}, \mathcal{E})$ is a measure space for actions.

  - Example: Let $X=(X_1, X_2)$ with iid X_1, X_2 as Bernoulli (p).
  - Note that in this example:
    * ϕ_x' is really a behavioral decision rule (for each $x \in \mathcal{X}$, this gives a distribution over $\mathcal{A}$).
    * ϕ_x' is a function of T (if $T(x)=T(y)$ for $x, y \in \mathcal{X}$ then $\phi_x'=\phi_y'$ as distributions on $\mathcal{A}$).
    * Theorem 65 says that ϕ_x and ϕ_x' have the same risk functions (as will be seen in the outline of the proof of the theorem).
    * This construction (by mixing according to the distribution of $X \mid T$) here takes something nonrandomized and produces randomization.

- Lemma 66: Suppose that $\mathcal{A} \subset \mathbb{R}^d$ is convex and δ_1 and δ_2 are two non-randomized decision rules. Then,

$$\delta=\frac{1}{2}\left(\delta_1+\delta_2\right)$$

  is also a non-randomized decision rule. Additionally, for a given θ (i) if $L(\theta, a)$ is convex in a and $R\left(\theta, \delta_1\right)=R\left(\theta, \delta_2\right)$, then

$$R(\theta, \delta) \leq R\left(\theta, \delta_1\right)=R\left(\theta, \delta_2\right)$$

  (ii) if $L(\theta, a)$ is strictly convex in $a, R\left(\theta, \delta_1\right)=R\left(\theta, \delta_2\right)<\infty$ and $P_\theta\left(\delta_1(X) \neq \delta_2(X)\right)>0$, then

$$R(\theta, \delta)<R\left(\theta, \delta_1\right)=R\left(\theta, \delta_2\right)$$

  Proof by Lemma 51.

- Corollary 67: Suppose that $\mathcal{A} \subset \mathbb{R}^d$ is convex and δ_1 and δ_2 are two non-randomized decision rules with identical risk functions. If $L(\theta, a)$ is convex in a for all θ and there exists some θ_0 such that $L\left(\theta_0, a\right)$ is strictly convex in a, $R\left(\theta_0, \delta_1\right)=R\left(\theta_0, \delta_2\right)<\infty$ and $P_{\theta_0}\left(\delta_1(X) \neq \delta_2(X)\right)>0$, then δ_1 and δ_2 are inadmissible (because $\delta=\left(\delta_1+\delta_2\right)/2$ is better)

- Theorem 68 (The Rao-Blackwell Theorem): Suppose that $\mathcal{A} \subset \mathbb{R}^d$ is convex and δ is a non-randomized decision rule with $\mathrm{E}_\theta\|\delta(X)\|<\infty$ for all θ. Suppose further that $T:(\mathcal{X}, \mathcal{B}) \rightarrow(\mathcal{T}, \mathcal{F})$ is sufficient for θ and, with $\mathcal{B}_0=\sigma\langle T\rangle$, let

$$\delta_0(x)=\mathrm{E}_\theta\left(\delta \mid \mathcal{B}_0\right)(x), \quad x \in \mathcal{X}$$

  Then δ_0 is a non-randomized decision rule. Furthermore, for a given θ (i) if $L(\theta, a)$ is convex in a, then

$$R\left(\theta, \delta_0\right) \leq R(\theta, \delta)$$

  (ii) if $L(\theta, a)$ is strictly convex in $a, R(\theta, \delta)<\infty$ and $P_\theta\left(\delta_0(X) \neq \delta(X)\right)>0$, then

$$R\left(\theta, \delta_0\right)<R(\theta, \delta)$$

  Proof (i) by Tower rule and conditional Jensen's Inequality:

$$R(\theta, \delta)=E_\theta[L(\theta, \delta(X))]=E_\theta[E(L(\theta, \delta(X))|B_0))] \geq E_\theta[L(\theta, E_\theta(\delta(X)|B_0))]=R(\theta, \delta_0)$$

  Proof (ii) by Lemma 66: Define $\delta'=\frac{1}{2}(\delta+\delta_0)$ and assume $R(\theta, \delta_0)=R(\theta, \delta)<\infty$. Since $R(\theta, \delta_0)=R(\theta, \delta)<\infty$, $L(\theta, a)$ is strictly convex and $P_\theta(\delta_0(X) \neq \delta(X))>0$, then

$$R(\theta, \delta')<R(\theta, \delta)=R(\theta, \delta_0)$$

  Then, define $\delta''(X)=E_\theta(\delta'(X)|T)=E_\theta(\delta(X)/2+\delta_0(X)/2|T)=\delta_0(X)/2+\delta_0(X)/2=\delta_0(X)$. By Theorem 68(i), $R(\theta, \delta_0)=R(\theta, \delta'') \leq R(\theta, \delta')<R(\theta, \delta_0)$, a contradiction. Therefore the equality does not hold, i.e. under certain constraint, there must be improvement to take the conditional expectation. □

  Note: By sufficiency and $\mathrm{E}_\theta\|\delta(X)\|<\infty, \delta_0(x)=\mathrm{E}_\theta\left(\delta \mid \mathcal{B}_0\right)(x) \equiv \mathrm{E}\left(\delta \mid \mathcal{B}_0\right)(x)$ is free of θ and well-defined. Also, writing $\delta(x)=\left(\delta^{(1)}(x), \ldots, \delta^{(d)}(x)\right) \in \mathcal{A} \subset \mathbb{R}^d$, we may define $\|\delta(x)\|=\sqrt{\left[\delta^{(1)}(x)\right]^2+\cdots+\left[\delta^{(d)}(x)\right]^2}$

- Example: Let $X_1, \ldots, X_n$ be iid $N(\theta, 1), \Theta = \mathbb{R}$. Consider estimation of $\gamma(\theta) = \mathrm{E}_\theta X_1^2 = \theta^2 + 1$ where $\mathcal{A} = \mathbb{R}$ and $L(\theta, a) = (\gamma(\theta) - a)^2$. Note that $T(X) = \sum_{i=1}^n X_i$ is sufficient for θ and consider the moment-based estimator of $\gamma(\theta) = \theta^2 + 1$ given by

$$\frac{1}{n} \sum_{i=1}^n X_i^2 = \frac{1}{n} \sum_{i=1}^n \left(X_i - \bar{X}\right)^2 + (\bar{X})^2$$

4.5 Baye's Decision Rule

The Bayes approach to decision theory is one way of reducing the set of risk functions $\{R(\theta, \delta) : \theta \in \Theta\}$ for a decision rule δ to single numbers so that different decision rules δ and δ' can be compared or "ordered" in a straightforward fashion. Let G denote a distribution on $(\Theta, \mathcal{G})$.

- Definition 69: The Bayes risk of $\phi \in \mathcal{D}^*$ with respect to the prior G is

$$BR(G, \phi) = \int_\Theta R(\theta, \phi) dG(\theta)$$

- The Minimum Bayes risk is

$$BR(G) = \inf_{\phi \in \mathcal{D}^*} BR(G, \phi)$$

- Definition 70 : $\phi \in \mathcal{D}^*$ is said to be a Bayes rule with respect to G (or Bayes with respect to G) if

$$BR(G, \phi) = BR(G)$$

- Definition 71: Let $\epsilon > 0$. Then, $\phi \in \mathcal{D}^*$ is said to be ϵ-Bayes with respect to G if

$$BR(G, \phi) \leq BR(G) + \epsilon$$

  Illustrations and Implications: Consider some finite $\Theta = \{\theta_1, \ldots, \theta_k\}$ geometry (assuming $\mathcal{D}^* = \mathcal{D}_*$) connected with Bayesness. We'll focus on $k = 2$ pictures with risk vectors $y = (y_1, y_2) = (R(\theta_1, \phi), R(\theta_2, \phi))$ and prior probabilities $g = (g_1, g_2)$ on (θ_1, θ_2) with $g_1, g_2 \geq 0, g_1 + g_2 = 1$

  1. Decision rules with the same Bayes risk can be denoted with lines on S =the set of all randomized risk vectors.
  2. A given prior (g) can have more than one Bayes rule (which can be quite different).
  3. Different priors (e.g., g and g') can lead to a rule that is Bayes.
  4. If S is not closed from below, there may not be a rule that is Bayes with respect to a prior g.

- Theorem 72: If $\Theta = \{\theta_1, \theta_2, \ldots,\}$ is countable, G is a prior with $g_i \equiv G(\{\theta_i\}) > 0$ for all $i, BR(G) < \infty$, and $\phi \in \mathcal{D}^*$ is Bayes with respect to G, then ϕ is admissible. Note: One may NOT remove the assumption that $g_i > 0$ for all i in this theorem.

  This suggests that in order to get "Bayesness $\Rightarrow$ admissibility," we need to be sure that the prior G "puts mass everywhere" (see also Theorem 73 to follow).

- Theorem 73: Suppose $\Theta \subset \mathbb{R}^k$ and that every neighborhood of a point $\theta \in \Theta$ has a non-empty intersection with the interior of Θ. Suppose further that, for every $\phi \in \mathcal{D}^*, R(\theta, \phi) < \infty$ is continuous in θ. Let G be a prior distribution that has a non-empty intersection with the interior of Θ. Suppose further that, for every $\phi \in \mathcal{D}^*, R(\theta, \phi) < \infty$ is continuous in θ. Let G be a prior distribution that has support given by Θ in the sense that $G(B) > 0$ holds for every open ball $B \subset \Theta$. Then, if $BR(G) < \infty$ and ϕ is a Bayes rule with respect to G, then ϕ is admissible.

  Notes:

  1. $R(\theta, \phi)$ can be continuous in θ when P_θ varies smoothly as a function of θ.

  2. Basic Idea: If ϕ is inadmissible, then there exists some better rule ϕ' than ϕ where $R(\theta, \phi) \geq R(\theta, \phi')$ holds for all θ. Integrating both sides of this inequality with respect to the prior G gives

$$BR(G, \phi) = \int_\Theta R(\theta, \phi) dG(\theta) \geq \int_\Theta R\left(\theta, \phi'\right) dG(\theta) = BR\left(G, \phi'\right)$$

The problem, though, is that because ϕ' is better than ϕ, then there exists some θ_0 where $R(\theta_0, \phi) > R(\theta_0, \phi')$. And, because $R(\theta, \phi) - R(\theta, \phi')$ is continuous in θ by assumption, there is a neighborhood or ball $B(\theta_0)$ around θ_0 where $R(\theta, \phi) > R(\theta, \phi'), \theta \in B(\theta_0)$, holds and the prior gives mass to this ball $G(B(\theta_0)) > 0$ by assumption. Consequently, the inequality in (1) will become a strict inequality $BR(G, \phi) > BR(G, \phi')$, contradicting that ϕ is Bayes with respect to G.

- Theorem 74 : If every Bayes rule with respect to G has the same risk function $R(\theta, \cdot), \theta \in \Theta$, then all Bayes rules are admissible.

- Corollary 75: If $\phi \in \mathcal{D}^*$ is the only (i.e., unique) Bayes rule with respect to G, then ϕ is admissible.

- Theorem 76: (Separating Hyperplane Theorem) Let S_1 and S_2 be two disjoint convex subsets of $\mathbb{R}^k$. Then, there exists non-zero $p = (p_1, \ldots, p_k) \in \mathbb{R}^k$ such that $\sum_{i=1}^p p_i x_i \leq \sum_{i=1}^p p_i y_i$ for all $x = (x_1, \ldots, x_k) \in S_1$ and $y = (y_1, \ldots, y_k) \in S_2$

- Theorem 77 : If Θ is finite and ϕ is admissible, then ϕ is Bayes with respect to some prior.

  The next result shows that randomized decision rules are not needed for achieving minimum Bayes risk $BR(G)$.

- Theorem 78: Suppose that $\psi \in \mathcal{D}_*$ is Bayes with respect to G and $BR(G) < \infty$. Then, there exists a non-randomized rule $\delta \in \mathcal{D}$ that is also Bayes with respect to G.

  Proof by Fubini's theorem.

  Next we address two remaining questions of 1. When do Bayes rules exist? 2. When they exist, what do they look like?

- Theorem 79: If Θ is finite, $\mathcal{S}$ (the set of risk vectors from randomized decision rules) is closed from below, and G assigns positive probability to each $\theta \in \Theta$ then there exists a decision rule $\delta \in \mathcal{D}$ that is Bayes with respect to G. (See also Theorem 59 for background: Θ is finite, $\lambda(S) \subset S \implies \lambda(S) = \mathcal{A}(S)$.)

  Proof by the property of 'closed from below' and using the seperating hyperplane theorem.

  - Example of Finding Bayes Rule: Let $X \sim N(\theta, 1)$, prior $\theta \sim N\left(0, \tau^2\right)$. Then, posterior

  $$\theta \mid X \sim N\left(X \frac{\tau^2}{1+\tau^2}, \frac{\tau^2}{1+\tau^2}\right)$$

    Consider estimating θ under $L(\theta, a) = (\theta - a)^2$. Then for each X, conditional expected loss given the data $E_{\Theta|X=x}$ is a function of an action $a \in \mathbb{R}$. And $a = X(\frac{\tau^2}{1+\tau^2})$ (the mean of posterior distribution) minimizes the expected loss. Therefore $\delta(X) = X(\frac{\tau^2}{1+\tau^2})$ should be a Baye's Rule.

- In general, the structure of Bayes rules can be described as follows:

  - $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ dominated by σ -finite measure $\mu \& \frac{dP_\theta}{d\mu} = f_\theta$
  - G: distribution on $(\Theta, \mathcal{G})$ (often with G dominated by σ -finite measure ν and $g = \frac{dG}{d\nu}$)
  - π : a joint distribution of (X, θ) (with density $f_\theta(x)g(\theta)$ with respect to $\mu \times \nu$)
  - π^X as the marginal distribution of X from π
  - $\pi^\theta = G($ marginal distribution of θ from $\pi) \& \pi^{X|\theta} = P_\theta$ (conditional distribution of X given θ from π)
  - the posterior distribution $\pi^{\theta|X}$ of θ given X, having a density with respect to ν as

  $$f_{\theta|X}(\theta \mid x) = \frac{f_\theta(x)g(\theta)}{\int_\Theta f_\theta(x)g(\theta)d\nu(\theta)}$$

Then, for a non-randomized decision rule δ, the expected (posterior) loss given $X = x$ is

$$\begin{aligned} \mathrm{E}[L(\theta, \delta(x)) \mid X = x] &= \int_\Theta L(\theta, \delta(x)) \left[\frac{f_\theta(x)}{\int_\Theta f_\theta(x)dG(\theta)}\right] dG(\theta) \\ &\geq \inf_{a \in \mathcal{A}} \int_\Theta L(\theta, a) \left[\frac{f_\theta(x)}{\int_\Theta f_\theta(x)dG(\theta)}\right] dG(\theta) \end{aligned}$$

with equality if and only if $\delta(x)$ minimizes

$$\mathrm{E}[L(\theta, a) \mid X = x] = \int_\Theta L(\theta, a)\left[\frac{f_\theta(x)}{\int_\Theta f_\theta(x) dG(\theta)}\right] dG(\theta)$$

So if, for almost $x\left(\text{according to } \pi^X\right), \delta(x)$ minimizes $\mathrm{E}[L(\theta, a) \mid X = x]$, then $\delta(x)$ will be Bayes with respect to G. This follows from the definition that

$$\begin{aligned}
BR(G, \delta) &= \int_\Theta R(\theta, \delta) dG(\theta) \\
&= \int_\Theta \int_\mathcal{X} L(\theta, \delta(x)) dP_\theta(x) dG(\theta) \\
&= \mathrm{E}_\pi L(\theta, \delta(X)) \\
&= \mathrm{E}_\pi \mathrm{E}[L(\theta, \delta(X)) \mid X] \\
&= \int_\mathcal{X} \int_\Theta L(\theta, \delta(x)) d\pi^{\theta \mid X}(\theta \mid x) d\pi^X(x)
\end{aligned}$$

- Definition 80: A formal non-randomized Bayes rule with respect to a prior G is a rule $\delta(X)$ such that, for each $x \in \mathcal{X}, \delta(x)$ is an $a \in \mathcal{A}$ minimizing

$$\int_\Theta L(\theta, a)\left[\frac{f_\theta(x)}{\int_\Theta f_\theta(x) dG(\theta)}\right] dG(\theta)$$

- Definition 81: If G is a σ-finite measure, a formal non-randomized generalized Bayes rule with respect to G is a rule $\delta(X)$ such that, for each $x \in \mathcal{X}$ $\delta(x)$ is an $a \in \mathcal{A}$ minimizing

$$\int_\Theta L(\theta, a) f_\theta(x) dG(\theta)$$

Note: 1. G is not necessarily probability meansure. 2.There is no normalizing constant.

- Example: $X_1, \ldots, X_n$ are iid $N(\theta, 1)$ random variables. Consider estimating θ under $L(\theta, a) = (\theta - a)^2$. Here $\mathcal{A} = \Theta = \mathbb{R}$ and let G be Lebesgue measure (μ) for θ on $\mathbb{R}$.

$$\begin{aligned}
\int_\mathbb{R} L(\theta, a) f_\theta(x) dG(\theta) &= \int_\mathbb{R} (\theta - a)^2 \Pi_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp(-(x_i - \theta)^2/2) d\mu(\theta) \\
&= (2\pi)^{-n/2} \exp(-(x_i - \bar{X})^2/2) \int_\mathbb{R} (\theta - a)^2 \exp(\theta - \bar{X})^2 d\mu(\theta)
\end{aligned}$$

where the integral part is equal to $E(\theta - a)^2 \sqrt{\frac{2\pi}{n}}$, $\theta \sim N(\bar{X}, \frac{1}{n})$, which is minimized at $a = \bar{X}$ by generalized Baye's rule.

Note:

ϕ Bayes: $BR(G, \phi) = BR(G)$

δ formal Bayes: for each $x, \delta(x)$ minimizes $\mathrm{E}_{\theta \mid X}[L(\theta, a) \mid X = x]$ over a

Hence, formal Bayes with respect to $G \Rightarrow$ Bayes with respect to G

- Standard Results of Formal Bayes Rules
  - Estimation of $\gamma(\theta)$:

    1. For a weighted squared error loss $L(\theta, a) = w(\theta)(\gamma(\theta) - a)^2$, where $w(\theta) > 0$ a Bayes rule with respect to G is

    $$\delta_G(x) = \frac{\mathrm{E}_{\theta \mid X}[w(\theta)\gamma(\theta) \mid X = x]}{\mathrm{E}_{\theta \mid X}[w(\theta) \mid X = x]}$$

    2. For the absolute error loss $L(\theta, a) = |\gamma(\theta) - a|$, a Bayes rule is $\delta_G(x) =$ a median of the conditional distribution of $\gamma(\theta) \mid X = x$
  - "0-1" loss hypothesis testing: For $\Theta = \Theta_0 \cup \Theta_1, \mathcal{A} = \{0, 1\}$, and $L(\theta, a) = I\left[\theta \notin \Theta_a\right]$, a Bayes rule is $\delta_G(x) = I[$ the posterior probability of $\Theta_1 \geq$ the posterior probability of $\Theta_0]$

4.6 Minimax Decision Rules

An alternative to the Bayes reduction of $R(\theta, \phi)$ to a number $BR(G, \phi) = \int_\Theta R(\theta, \phi) dG(\theta)$ is to reduce $R(\theta, \phi)$ to a number $\sup_{\theta \in \Theta} R(\theta, \phi)$. (See pages 349 – 354 of Berger.)

- Definition 82 : A decision rule $\phi \in \mathcal{D}^*$ is said to be minimax if

$$\sup_{\theta \in \Theta} R(\theta, \phi) = \inf_{\phi' \in \mathcal{D}^*} \sup_{\theta \in \Theta} R\left(\theta, \phi'\right)$$

- Definition 83 : If a decision rule $\phi \in \mathcal{D}^*$ has a constant risk function, it is called an equalizer rule.

  Intuitively, if one tries to push down the highest peak in $R(\theta, \phi)$ (as a function of θ to produce a minimax rule, it tends to result in an equalizer rule.

- Theorem 84 : If $\phi \in \mathcal{D}^*$ is an equalizer rule and is admissible, then it is minimax.

  Proof by contradiction of the admissibility.

- Theorem 85: Suppose that $\{\phi_i\}$ is a sequence of decision rules, each ϕ_i being Bayes with respect to G_i. If $BR\left(G_i, \phi_i\right) \to C < \infty$ as $i \to \infty$, and ϕ is a decision rule with $R(\theta, \phi) \leq C$ for all θ, then ϕ is minimax.

  Proof by contradition.

- Corollary 86: If $\phi \in \mathcal{D}^*$ is an equalizer rule and is Bayes with respect to G, then it is minimax.

- Corollary 87: If $\phi \in \mathcal{D}^*$ is Bayes with respect to G and $R(\theta, \phi) \leq BR(G)$ for all θ, then it is minimax.

  Note: Corollary 87 follows from Theorem 85 with ϕ_i and $G_i = G$. Then, Corollary 87 $\Rightarrow$ Corollary 86 when $R(\theta, \phi) = C$ for all θ and $BR(G) = BR(G, \phi)$

  Corollaries 86 & 87 suggest that, for an appropriate G, a Bayes rule with respect to G might be minimax. So, how does guess at or identify such a prior G?

- Definition 88: A prior distribution G is said to be least favorable if

$$BR(G) = \sup_{G'} BR\left(G'\right)$$

  Note: A least favorable prior maximizes Bayes risk over all priors.

- Theorem 89 : If ϕ is Bayes with respect to G and $R(\theta, \phi) \leq BR(G)$ for all θ, then G is least favorable.

  Note: This theorem shows that, in order to use Corollary 87 to prove that a Bayes rule with respect to G is minimax, G must be least favorable. So if we can guess at what would be the least favorable situation for a prior, that may give us insight regarding the minimax rule.

- Example (Composite vs. composite hypothesis testing): Let $X \sim N(\theta, 1), \Theta = \mathbb{R}, \mathcal{A} = \{0, 1\}$, and

$$L(\theta, a) = I[\theta \leq 5] \cdot I[a = 1] + I[\theta > 5] \cdot I[a = 0]$$

  i.e., testing $H_0 : \theta \leq 5$ vs. $H_1 : \theta > 5$ Intuitively, the worst possible prior would have mass 0.5 at $\theta = 5$ and mass 0.5 at $\theta = 5 + \epsilon (\epsilon > 0)$. We can use G_i defined by

$$G_i(\{5\}) = 0.5 \quad G_i\left(\left\{5 + \frac{1}{i}\right\}\right) = 0.5, \quad i = 1, 2, 3, \ldots$$

  along with Theorem 85 to show that $\delta(X) = I[X > 5]$ is minimax.