II Statistical Decision Theory

II.1 Ingredients (frequentist)

- model $\{F_\theta : \theta \in \Theta\}$, quantity of interest $\psi = \Psi(\theta) \in \Psi$
- loss fn $L : \Theta \times \Psi \to [0, \infty)$ s.t. $L(\theta, \psi) = 0$ iff $\psi = \Psi(\theta)$
- observe x
- a decision function δ is s.t. $\delta(x, \cdot)$ is a prob. measure on Ψ $\forall x \in \mathcal{X}$ (a Markov kernel
- after observing x generate decision $\psi \sim \delta(x, \cdot)$
- δ is a nonrandomized decision function if $\exists\, d : \mathcal{X} \to \Psi$ s.t. $\delta(x, \{d(x)\}) = 1$ $\forall x \in \mathcal{X}$, think of d as an estimator
- why consider randomized dec. fns

Lemma ① The class $\mathcal{D}$ of all dec. fns is convex.

Proof: Let $\alpha \in [0,1]$, $\delta_1, \delta_2 \in \mathcal{D}$ then for $x \in \mathfrak{X}$

$\alpha \delta_1(x, \cdot) + (1-\alpha)\delta_2(x, \cdot)$ is a prob. meas. on $\overline{\mathcal{A}}$.

Lemma ② $\overline{\mathcal{A}}$ is convex iff the class D of all nonrandomized dec. fns ~~is convex~~

Proof: $\Rightarrow$] If $\alpha \in [0,1]$, $d_1, d_2 \in D$ then $\alpha d_1 + (1-\alpha) d_2$

$: \mathfrak{X} \to \overline{\mathcal{A}}$ and $\alpha d_1 + (1-\alpha) d_2 \in D$.

$\Leftarrow$] obvious (constants)

<u>estimation problems</u>

- often $\Psi \subseteq \mathbb{R}^k$ is convex
- $L(\theta, \psi) = (\Psi(\theta) - \psi)' A (\Psi(\theta) - \psi)$ = quadratic loss

  where A is p.d. is a convex fn of ψ

  $L(\theta, \psi) = \sum_{i=1}^{k} |\Psi_i(\theta) - \psi_i| = L^1$ loss is a convex fn of ψ

  $L(\theta, \psi) = I_{\{\Psi(\theta)\}^c}(\psi)$ = 0-1 loss is not convex.

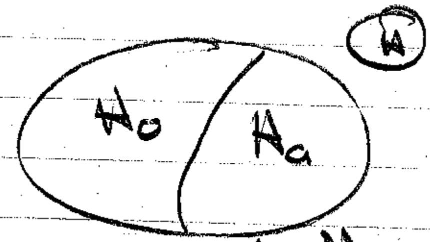
- but what if Ψ is not convex

eg $\Psi = \{a, b, c, d\}$ is not convex. | classification

  0-1 loss seems most appropriate

hypothesis testing problems

- $\Theta = H_0 \cup H_a$
  disjoint



- $\Psi(\theta) = 1 - I_{H_0}(\theta)$ 1 = reject H_0
  0 = accept H_0
- $\Psi = \{0,1\}$ not convex

$\boxed{\text{$H_0$ plays a special role do we need to dichotomize?}}$

- two types of errors:

  type I error $L(\theta,1) = c_0$, $\theta \in H_0$ (reject H_0 when true) false positive

  type II error $L(\theta,0) = c_a$, $\theta \in H_a$ (accept H_0 when false) false negative

- [illegible]

- $L(\theta,\psi) = c_0 I_{H_0 \times \{1\}}(\theta,\psi) + c_a I_{H_a \times \{0\}}(\theta,\psi) \overset{c_0+c_a=1}{=} 1 - I_{\{\Psi(\theta)\}}(\psi)$ 0-1 loss

- point of decision theory: choose a $\delta \in D$ that in some sense minimizes losses.

- for this we look at the distribution of $L(\theta,\psi)$ when $\psi \sim \delta(x,\cdot)$, $x \sim f_\theta$ for each $\theta \in \Theta$ (frequentist inference)

- if dist. of $L(\theta,\psi)$ is "more concentrated about 0 for δ_1 than for δ_2" for every $\theta \in \Theta$ then δ_1 is preferred to δ_2.

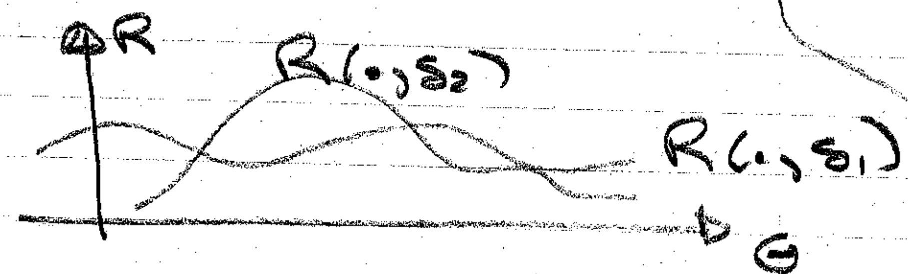
- how do we measure this concentration?
- lots of possibilities but common to use expected loss

Def The __risk function__ of $\delta \in \mathcal{D}$ is given by

$$R(\theta,\delta) = E_\theta\left(E_{\delta(x,\cdot)}(L(\theta,a))\right).$$

- Principle of Minimizing Expected Loss (Maximizing Expected Utility) | the principle
- δ_1 is preferred to δ_2 if $R(\theta,\delta_1) \le R(\theta,\delta_2)$ for __every__ $\theta \in \Theta$

  is this reasonable? why repeated sampling?
- but not a total order



- δ_0 is __optimal__ if $R(\theta,\delta_0) \le R(\theta,\delta)$ $\forall \theta \in \Theta$ and $\forall \delta \in \mathcal{D}$.

Estimation

(A psd)

- Φ convex, $L(\theta, a) = (\Phi(\theta) - a)' A (\Phi(\theta) - a)$
- $R(\theta, d) = \int_{\mathcal{X}} (\Phi(\theta) - d(x))' A (\Phi(\theta) - d(x)) P_\theta(dx)$

= mean-squared error of estimator d at θ.

and assuming $E_\theta(d)$ exists.

$$= \int_{\mathcal{X}} (\Phi(\theta) - E_\theta(d) + E_\theta(d) - d(x))' A (\;) P_\theta(dx)$$

$$= (\Phi(\theta) - E_\theta(d))' A (\Phi(\theta) - E_\theta(d))$$

$$+ (\Phi(\theta) - E_\theta(d))' A \int_{\mathcal{X}} (E_\theta(d) - d(x)) P_\theta(dx) \quad = 0$$

$$+ \int_{\mathcal{X}} (d(x) - E_\theta(d))' A (d(x) - E_\theta(d)) P_\theta(dx)$$

$$= (\text{squared bias})_\theta + \operatorname{tr}(A \operatorname{Var}_\theta(d))$$

- <u>note</u> - $E_\theta(d) \in \bar{\Phi}$ when Φ is convex but not necessarily otherwise (Lecture 19 SP course)

eg hypothesis testing

- $\delta(x,\cdot)$ is a prob. dist on $\mathcal{A} = \{0,1\}$ (0 = accept H_0, 1 = reject H_0)

- $\delta \leftrightarrow \phi \quad \phi : \mathcal{X} \to [0,1]$ a test function where $\phi(x) = \delta(x,\{1\})$ = prob. of rejecting H_0 after observing x.

- so $\delta(x,\cdot) \,|\, x \sim \text{Bernoulli}(\phi(x))$

- $$\int_{\mathcal{A}} L(\theta,a)\,\delta(x,da) = \int_{\{0,1\}} L(\theta,a)\,\delta(x,da)$$
$$= c_0\,\delta(x,\{1\})\,I_{H_0}(\theta) + c_a\,\delta(x,\{0\})\,I_{H_a}(\theta)$$
$$= c_0\,\phi(x)\,I_{H_0}(\theta) + c_a(1-\phi(x))\,I_{H_a}(\theta)$$

- so $R(\theta,\delta) = R(\theta,\phi)$
$$= c_0\,E_\theta(\phi)\,I_{H_0}(\theta) + c_a(1-E_\theta(\phi))\,I_{H_a}(\theta)$$

- $E_\theta(\phi)$ = <u>power function</u> of δ or ϕ.
  = prob. (uncond) of rejecting H_0 when θ is true.

- when $\theta \in H_0$ $R(\theta,\phi) = c_0 E_\theta(\phi)$ and so we want $E_\theta(\phi)$ small, when $\theta \in H_a$ $R(\theta,\phi) = c_a(1-E_\theta(\phi))$ so we want $E_\theta(\phi)$ big

<u>Lemma ③</u> δ is optimal iff $\delta(x,\cdot)$ is degenerate at $\Psi(\theta)$ a.e. P_θ $\forall \theta \in \Theta$

Proof: Let δ_θ be the dec. fn $\delta_\theta(x, \{\Psi(\theta)\}) = 1$ $\forall x \in \mathcal{X}$. Then $\delta_\theta \in \mathcal{D}$ $\forall \theta$ and an optimal δ satisfies $0 \le R(\theta,\delta) \le R(\theta,\delta_\theta) = 0$ and so $0 = \int_{\mathcal{X}} \int_{\Psi} L(\theta, a)\, \delta(x, da)\, P_\theta(dx)$ $\forall \theta$

iff $\int_{\Psi} L(\theta,a)\,\delta(x,da) = 0$ a.e. P_θ $\forall \theta$.

iff $L(\theta,a) = 0$ a.e. $\delta(x,\cdot)$, a.e. P_θ $\forall \theta$

iff $\delta(x,\cdot)$ is (essentially) degenerate at $\Psi(\theta)$ a.e. P_θ $\forall \theta$.

<u>Lemma ④</u> If $\exists \theta_1, \theta_2 \in \Theta$ s.t. $\Psi(\theta_1) \ne \Psi(\theta_2)$ and P_{θ_1}, P_{θ_2} are not concentrated on essentially disjoint sets then $\nexists$ an optimal δ.

Proof: Suppose optimal δ exists. Let

$N_\theta = \{x : \delta(x,\cdot) \text{ is not degenerate at } \Psi(\theta)\}$

Then $P_\theta(N_\theta) = 0$ by Lemma ③ $\forall \theta$. If $x \notin N_{\theta_1} \cup N_{\theta_2}$ then $\delta(x,\cdot)$ is degenerate at $\Psi(\theta_1)$ and $\Psi(\theta_2)$ so $N_{\theta_1} \cup N_{\theta_2} = \mathcal{X}$. Also

$P_{\theta_1}(N_{\theta_1} \cap N_{\theta_2}) = P_{\theta_2}(N_{\theta_1} \cap N_{\theta_2}) = 0.$

and thus P_{θ_1} is concentrated on N_{θ_2} and P_{θ_1} is concentrated on N_{θ_1} and these sets are essentially disjoint ⊗

- so in general the dec. problem doesn't have a solution.