01/25/22

- HW #2 – Q1 : mixture – skip until we cover Ex 3.4.1
  - Q10 : $f(x = \theta - 1 \mid \theta) = f(x = \theta + 1 \mid \theta) = \frac{1}{2}$

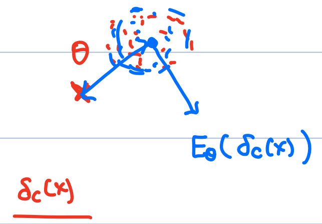
$X_1, X_2$ $\pi(\theta)$ : consider an arbitrary distribution over integers.

- lecture capture – as soon as possible

**JB Example 4**(p10) Assume $X \mid \theta \sim \mathrm{N}(\theta, 1)$. The goal is estimating $\theta$ under squared-error loss, $L(\theta, d) = (\theta - d)^2$. Consider the decision rule $\delta_c(x) = cx$. (SEL)

- Find $R(\theta, \delta_c)$.

$$
\begin{aligned}
R(\theta, \delta_c) &= E_\theta\left( L(\theta, \delta_c) \right) \\
&= E_\theta\left( (\theta - \delta_c(x))^2 \right) \\
&= E_\theta\left[ \left\{ \theta \pm E_\theta(\delta_c(x)) - \delta_c(x) \right\}^2 \right] \\
&= E_\theta\Big[ \left\{ \theta - E_\theta(\delta_c(x)) \right\}^2 \\
&\qquad + 2\left\{ \theta - E_\theta(\delta_c(x)) \right\}\left\{ E_\theta(\delta_c(x)) - \delta_c(x) \right\} \\
&\qquad + \left\{ E_\theta(\delta_c(x)) - \delta_c(x) \right\}^2 \Big]
\end{aligned}
$$




$$= \underbrace{\{\theta - E_\theta(\delta_c(x))\}^2}_{=\text{bias}} + E_\theta\left[\{E_\theta(\delta_c(x)) - \delta_c(x)\}^2\right]$$

$$= \{\theta - E_\theta(\delta_c(x))\}^2 + \underline{Var(\delta_c(x))}$$

Estimator ( decision rule) $\quad \delta_c(x) = cX$

$$E_\theta(\delta_c(x)) = E_\theta(cx) = c\theta$$

$$\Rightarrow \quad R(\theta, \delta_c) = (\theta - c\theta)^2 + Var(cX)$$

$$= (1-c)^2\theta^2 + c^2 \underbrace{Var(X)}_{=1}$$

$$= (1-c)^2\theta^2 + c^2$$

eg) $\quad c = 1/2 \quad \Rightarrow \quad R(\theta, \delta_{1/2}) = \underline{\frac{1}{4}\theta^2 + \frac{1}{4}} \leftarrow$

$\delta_1(x) = x \quad c = 1 \quad \Rightarrow \quad R(\theta, \delta_1) = \underline{1} \leftarrow$

$E_\theta(\delta_1(x)) = E_\theta(x) = \theta \quad c = 2 \quad \Rightarrow \quad R(\theta, \delta_2) = \underline{\theta^2 + 4} \leftarrow$

eg) $\quad X|\theta \sim N(\theta, 1)$, $\quad$ assume $\quad \theta \sim N(0, \tau^2)$

$$\Rightarrow \quad \theta | x \sim N\left(\underbrace{\left(\frac{1}{1} + \frac{1}{\tau^2}\right)^{-1}\left(\frac{x}{1} + \frac{0}{\tau^2}\right)}_{=\left(\frac{\tau^2}{1+\tau^2}\right)x}, \left(\frac{1}{1} + \frac{1}{\tau^2}\right)^{-1}\right)$$

Under the squared error loss function,

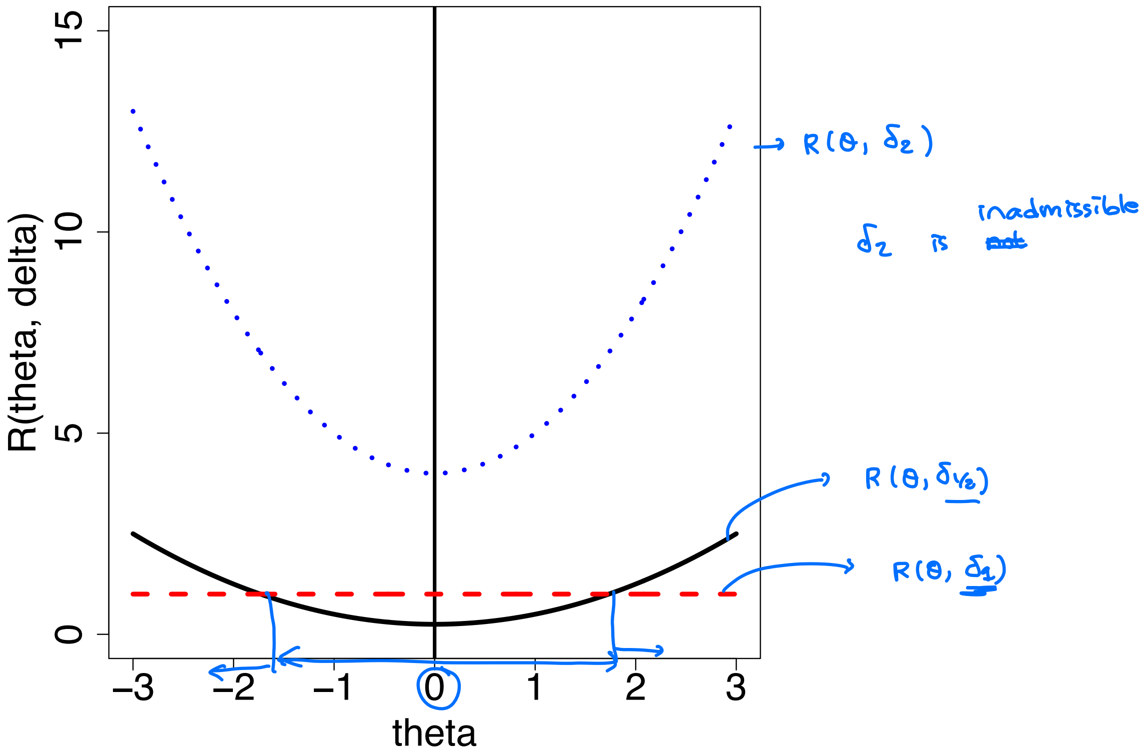
$$d = E^\theta(\theta | x) = \underbrace{\left(\frac{\tau^2}{1+\tau^2}\right)}_{\neq 1} x$$

$$E_\theta\left[\{\theta - \underline{E_\theta(\delta_c(x))}\}\{E_\theta(\delta_c(x)) - \delta_c(x)\}\right]$$

$$= \{\theta - \underline{E_\theta(\delta_c(x))}\} \times E_\theta\{E_\theta(\delta_c(x)) - \delta_c(x)\}$$

$$\text{"} \quad \times \underbrace{\left(E_\theta(\delta_c(x)) - E_\theta(\delta_c(x))\right)}_{=0} = 0$$

## JB Example 4(p10) (contd) Plot of $R(\theta, \delta_c)$



- Difficulties associated with using $R(\theta, \delta)$.

  $\star\star$ For each $\theta \in \Theta$, $R(\theta, \delta)$ is the expected loss based on an average over the random $X \in \mathcal{X}$.

  $\Rightarrow$ *long-run* performance of $\delta(x)$ and **not** directly for the given observation $x$.

  $\star\star$ A function of $\theta \in \Theta$ & $\theta$ is unknown.

  $\Rightarrow$ The frequentist approach $R(\theta, \delta)$ does not induce a total ordering on the set of procedures.

† How can Frequentists choose $\delta$?

- An additional principle must be introduced to select a specific rule for use.

  e.g. $\delta_1$ is preferred to $\delta_3$ under some concept of optimalty.

- Some important frequentist decision principles (CR 2.4 + a lot in JB)

  ⋆⋆ Bayes risk principle

  ⋆⋆ minimax

  ⋆⋆ admissibility

  ⋆⋆ restricted classes: e.g. we only consider unbiased estimators.

- Bayes estimators are *often optimal* for the frequentist concepts of optimality.

## † The Bayes Risk Principle

- The frequentist risk of a decision rule $\delta(x)$ is a function of $\theta$.

$$R(\theta, \delta) = \int_{\mathcal{X}} L(\theta, \delta(x)) f(x \mid \theta) dx.$$

- The *integrated risk* (also called Bayes Risk) is the frequentist risk averaged over $\Theta$ according to their prior $\pi(\theta)$.

$$\begin{aligned} r(\pi, \delta) &= \mathrm{E}^{\pi}\left[R(\theta, \delta)\right] \\ &= \int_{\Theta} \underbrace{\int_{\mathcal{X}} L(\theta, d) f(x \mid \theta) dx}_{= R(\theta, \delta)} \pi(\theta) d\theta. \end{aligned}$$

- $r(\pi, \delta)$ is a real number associated with estimator $\delta$.

$\Rightarrow$ Induces a total ordering on the set of estimators, so allows for the direct comparison of estimators.

† Any connection between $r(\pi, \delta)$ and $\rho(\pi, \delta \mid x)$?

⇒ They lead to the same decision.

- **Th 2.3.2** An estimator minimizing the integrated risk $r(\pi, \delta)$ can be obtained by selecting, for every $x \in \mathcal{X}$, the value $\delta(x)$ which minimizes the posterior expected loss, $\rho(\pi, \delta \mid x)$, since

$$r(\pi, \delta) = \int_{\mathcal{X}} \rho(\theta, \delta(x) \mid x) m(x) dx.$$

$$\underline{r(\pi, \delta)} = E^{\pi}(R(\theta, \delta))$$

$$= \int_{\Theta} R(\theta, \delta)\, \pi(\theta)\, d\theta$$

$$= \int_{\Theta} \int_{\mathcal{X}} L(\theta, \delta) \underbrace{f(x \mid \theta)\, \pi(\theta)}_{= \pi(\theta \mid x) \cdot m(x)}\, dx\, d\theta$$

$$= \int_{\mathcal{X}} \underbrace{\int_{\Theta} L(\theta, \delta)\, \pi(\theta \mid x)\, d\theta}_{= \underline{\rho(\pi, \delta \mid x)}}\, m(x)\, dx$$

- **Def 2.3.3**

  ⋆⋆ **A Bayes estimator** associated with a prior distribution $\pi$ and a loss function $L$ is any estimator $\delta^\pi$, which minimizes $r(\pi, \delta)$.

  ⋆⋆ For every $x \in \mathcal{X}$, it is given by $\delta^\pi(x)$ (**a Bayes action**), argument of $\min_d \rho(\pi, d \mid x)$.

  ⋆⋆ The value $r(\pi) = r(\pi, \delta^\pi)$ is then called **the Bayes risk**.

- **JB Def 9, p160** If $\pi$ is an improper prior, but $\delta^\pi(x)$ is an action which minimizes $\rho(\pi, d \mid x)$ for each $x$ with $m(x) > 0$, then $\delta^\pi$ is called a **generalized Bayes rule.**

$\dagger$ Minimaxity: Minimize the expected loss in the least favorable case ($\Leftrightarrow$ protect against the worst possible state of nature, conservative!)

- **The Minimax Principle.** **JB p18** $\delta_1$ is preferred to $\delta_2$ if

$$\sup_\theta R(\theta, \delta_1) < \sup_\theta R(\theta, \delta_2).$$

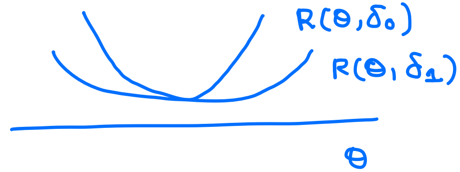
- **Def 2.4.3** The **minimax risk** associated with a loss function $L$ is the value

$$\bar{R} = \inf_{\delta \in \mathcal{D}} \sup_{\theta} R(\theta, \delta) = \inf_{\delta \in \mathcal{D}} \sup_{\theta} \mathrm{E}_{\theta} \{L(\theta, \delta(x))\},$$

and a **minimax estimator** is any (possibly randomized) estimator $\delta_0$ such that

$$\sup_{\theta} R(\theta, \delta_0) = \bar{R}$$

**JB Example 4** (contd) Assume $X \mid \theta \sim \mathrm{N}(\theta, 1)$. The goal is estimating $\theta$ under squared-error loss, $L(\theta, a) = (\theta - a)^2$. Consider the decision rule $\delta_c(x) = cx$. Find the minimax rule.




# † Admissibility

- **Def 2.4.19** An estimator $\delta_0$ is inadmissible if there eixsts an estimator $\delta_1$ which dominates $\delta_0$, that is, such that, for every $\theta$

$$R(\theta, \delta_0) \geq R(\theta, \delta)$$

and, for at least one value $\theta_0$ of the parameter,

$$R(\theta_0, \delta_0) > R(\theta_0, \delta).$$

Otherwise, $\delta_0$ is said to be admissible.

- What is the underlying idea as a criterion?

**Inadmissible estimators should not be considered at all since they can be uniformly improved!**

**JB Example 4** (contd) Assume $X \mid \theta \sim \mathrm{N}(\theta, 1)$. The goal is estimating $\theta$ under squared-error loss, $L(\theta, a) = (\theta - a)^2$. Consider the decision rule $\delta_c(x) = cx$ with $c > 1$. Is the rule admissible?

$*$ Admissibility is related (stronger than minmax) to the Bayesian paradigm.

- Admissibility is automatically satisfied by most Bayes estimators.
- **Prop 2.4.22** If a prior distribution $\pi$ is strictly positive on $\Theta$, with finite Bayes risk and the risk function, $R(\theta, \delta)$, is a continuous function of $\theta$ for every $\delta$, the Bayes estimator $\delta^{\pi}$ is admissible.
- Want to learn more? *Read CR 2 and JB 4.8*

**Example 2.4.6 (Stein Phenomenon)** Suppose a $p$−dimensional vector, $\boldsymbol{X} \sim N_p(\boldsymbol{\theta}, I_p)$ and consider the problem of estimating $\boldsymbol{\theta}$ (a $p$−dim vector). Assume the quadratic loss function
$L(\boldsymbol{\theta}, \delta) = (\boldsymbol{\theta} - \boldsymbol{\delta})'(\boldsymbol{\theta} - \boldsymbol{\delta}) = ||\boldsymbol{\theta} - \boldsymbol{\delta}||^2$. $= \sum_{i=1}^{p} (\theta_i - \delta_i)^2$

$\delta_c(x) = cx$

$\delta_1(x) = x$

- The maximum likelihood estimator $\delta_1(\boldsymbol{X}) = \boldsymbol{X}$

  ⋆⋆ The least squares estimator in standard regression setting

  ⋆⋆ For $p = 1$ or 2, it is admissible and the unique minimax estimator.

$$\delta_{JS} = \underbrace{\left(1 - \frac{(p-2)}{\|x\|}\right)}_{} x \qquad \frac{(p-2)}{\|x\|} > 0$$

$R(\theta, \delta_{JS}) \leq R(\theta, \delta_1)$ for all $\theta$

$R(\theta, \delta_{JS}) < R(\theta, \delta_1)$ for some $\theta$

$\Rightarrow$ dominates $\delta_1 = X$ for $p \geq 3$

$p-2 \leq c \leq 2\cdot(p-2)$

$\leq 2p-4 < 2p-1$

positive part JS

$$\delta_c^+ = \begin{cases} \left(1 - \frac{c}{\|x\|^2}\right) x & \text{if } \|x\|^2 > c \\ 0 & \text{o.w} \end{cases}$$

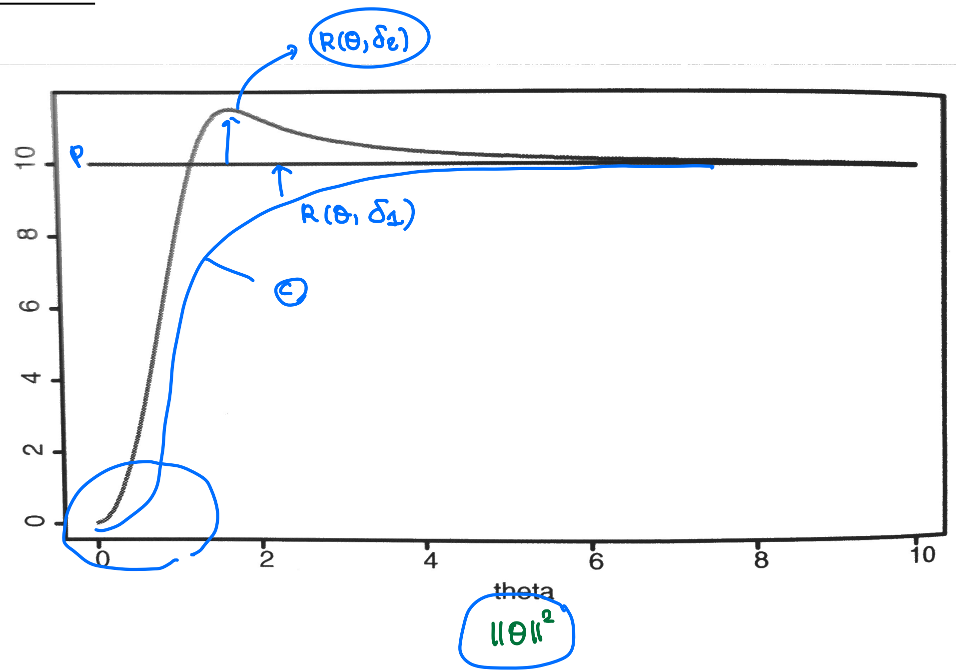
## Example 2.4.6 (contd)

- Consider the positive part James-Stein estimator,

$$\delta_2(\boldsymbol{X}) = \begin{cases} \left(1 - \frac{2p-1}{||\boldsymbol{X}||^2}\right)\boldsymbol{X} & \text{if } ||\boldsymbol{X}||^2 \geq \overset{c=}{2p-1}, \\ \mathbf{0} & \text{ow.} \end{cases}$$

MLE  positive part JS w/ $c=2p-1$

♣ Figure 2.4.1 Comparison of the risks of $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$ for $p = 10$



## Example 2.4.6 (contd)

- ⋆⋆ $\delta_2$ cannot be minimax.
- ⋆⋆ $\delta_2$ is definitely superior on some (the most interesting) part of the parameter space.
- ⋆⋆ "The Stein effect": allows to borrow information from the other components, even when they are independent and deal with totally different estimation problems.
- ⋆⋆ Sometimes the minimax rule is not useful! (or sometimes may not exist)
- ⋆⋆ Following James and Stein, extensive research on this has been done – *Shrinkage estimators*

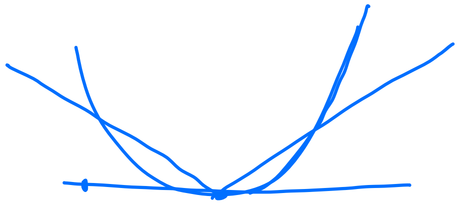
† Usual loss functions (CR Section 2.5)

- Quadratic loss $E^{\pi}(\theta \mid x)$
- Absolute loss ← posterior median
- 0-1 loss ← posterior mode
- Intrinsic loss – entropy distance (Kullback-Leibler divergence), Hellinger loss...

⋆⋆ What are the Bayesian estimators $\delta^{\pi}(x)$ under the classical loss functions?

† The quadratic loss

$$L(\theta, d) = (\theta - d)^2$$



- most common evaluation criterion – simplicity
- penalize large deviations too heavily

• **Prop 2.5.1** The Bayes estimator $\delta^\pi$ associated with the prior distribution $\pi$ and with the quadratic loss $L(\theta, d) = (\theta - d)^2$, is the posterior expectation,

$$\delta^\pi(x) = \mathrm{E}^\pi(\theta \mid x) = \frac{\int_\Theta \theta f(x \mid \theta)\pi(\theta)d\theta}{\int_\Theta f(x \mid \theta)\pi(\theta)d\theta}.$$

- **Cor 2.5.2** The Bayes estimator $\delta^\pi$ associated with $\pi$ and with the weighted quadratic loss $L(\theta, d) = w(\theta)(\theta - d)^2$, where $w(\theta)$ is a nonnegative function, is

$\geq 0$

$$\delta^\pi(x) = \frac{\mathrm{E}^\pi(w(\theta) \cdot \theta \mid x)}{\mathrm{E}^\pi(w(\theta) \mid x)}$$

- **Cor 2.5.3** When $\Theta \in \mathbb{R}^p$, the Bayes estimator $\delta^\pi$ associated with the prior distribution $\pi$ and with the quadratic loss $L(\boldsymbol{\theta}, \boldsymbol{d}) = (\boldsymbol{\theta} - \boldsymbol{d})^t \boldsymbol{Q}(\boldsymbol{\theta} - \boldsymbol{d})$, is the posterior mean $\delta^\pi(x) = \cancel{\mathrm{E}^{\boldsymbol{\theta}|x}(\theta \mid x)}$, for every positive -definite symmetric $p \times p$ matrix $\boldsymbol{Q}$.

$E^\pi(\theta \mid x)$

† The absolute error loss

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

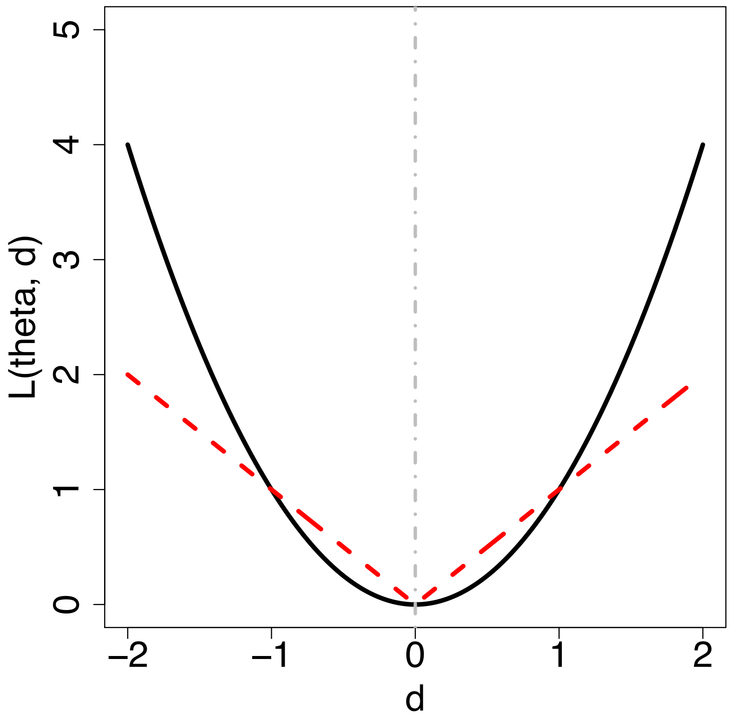
Or multilinear function (more general)

$\theta \geq d$

$$L(\theta, d) = \begin{cases} k_1(\theta - d) & \text{if } \theta - d \geq 0, \\ k_2(d - \theta) & \text{if } \theta - d < 0, \end{cases}$$

- slow down the progression of the quadratic loss for large errors and has a robustifying effect.
- $k_1$ and $k_2$ reflect the relative importance of underestimation or overestimation.
- $k_1 = k_2 \Rightarrow$ the absolute error loss

♣ squared error loss vs absolute error loss



- **Prop 2.5.5** A Bayes estimator associated with the prior distribution $\pi$ and the multilinear loss is a $k_1/(k_1 + k_2)$ fractile of $\pi(\theta \mid x)$.

- If $k_1 = k_2$, the Bayes estimator is the posterior median.

† The 0-1 loss: the penalty associated with an estimator $\delta$ is 0 if the answer is correct and 1 otherwise.

Example 2.5.6 Consider the test of $H0 : \theta \in \Theta_0$ vs $H1 : \theta \notin \Theta_0$. Then $\mathcal{D} = \{0, 1\}$, where 1 stands for acceptance of $H0$ and 0 for rejection. For the 0-1 loss,

if $\underbrace{\theta \in \Theta_0}_{\text{Ho is true}}$ & $d=1$, $0$

$$\rightarrow L(\theta, d) = \begin{cases} 1 - d & \text{if } \theta \in \Theta_0, \\ d & \text{otherwise.} \end{cases}$$

- posterior expected loss

case 1: $d = 1$

$$\rho(\pi, 1 \mid x) = E^{\pi}(L(\theta, 1) \mid x)$$

$$= \int_{\Theta} L(\theta, 1)\, \pi(\theta \mid x)\, d\theta$$

$$= \underbrace{\int_{\Theta_0} \overbrace{L(\theta, 1)}^{=0} \pi(\theta \mid x)\, d\theta}_{=0} + \int_{\Theta_0^c} \underbrace{L(\theta, 1)}_{=1} \pi(\theta \mid x)\, d\theta$$

$$= \int_{\Theta_0^c} \pi(\theta \mid x)\, d\theta$$

$$= P_r(\theta \in \Theta_0^c \mid x)$$

Case 2 : $d = 0$

$$\rho(\pi, 0 \mid x) = P_r(\theta \in \Theta_0 \mid x)$$

1 choose $d = 1$ if $P_r(\theta \in \Theta_0^c \mid x) \leq P_r(\theta \in \Theta_0 \mid x)$

$d = 0$ o.w

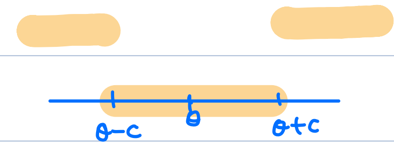
• **Prop 2.5.7** The Bayes estimator associated with the prior distribution $\pi$ and with the 0-1 loss is

$$\delta^{\pi}(x) = \begin{cases} 1 & \text{if } \Pr(\theta \in \Theta_0 \mid x) > \Pr(\theta \notin \Theta_0 \mid x), \\ 0 & \text{otherwise,} \end{cases}$$

i.e., $\delta^{\pi}(x)$ is equal to 1 if and only if $\Pr(\theta \in \Theta \mid x) > 1/2$.

Consider the 0-1 loss function for the estimation problem.

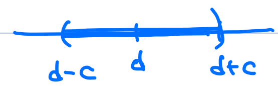
$$L(\theta, d) = \begin{cases} 0 & \text{if } |\theta - d| \le c \\ 1 & \text{o.w} \end{cases}$$

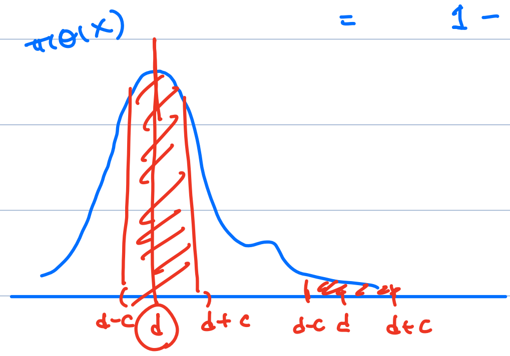



$$\underline{\rho(\pi, d \mid x)} = \int_{\mathbb{R}} L(\theta, d)\, \pi(\theta \mid x)\, d\theta$$

$$= \Pr(\theta \notin (d \pm c) \mid x)$$

$$= 1 - \Pr(\theta \in (d-c, d+c) \mid x)$$




$c \to 0$, $\underline{d}$ = posterior mode

= MAP (maximum a posteriori estimator)