36-705 18/07/2020

Lecture Notes 8 - Minimax Theory - Review

- concerned with best possible estimator $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ of param $\theta \in \Theta$.
- Framework uses a loss function and risk function (exp. value of loss).
- loss fn $L(\theta, \hat{\theta})$ - measure of 'quality' of estimator

- Examples:

$$L(\theta,\hat{\theta}) = (\theta - \hat{\theta})^2 \quad \text{sq. error}$$

$$L(\theta,\hat{\theta}) = |\theta - \hat{\theta}| \quad \text{abs. error}$$

$$L(\theta,\hat{\theta}) = |\theta - \hat{\theta}|^p \quad L_p\text{-loss}$$

$$L(\theta,\hat{\theta}) = \begin{cases} 0 & \text{if } \theta = \hat{\theta} \\ 1 & \text{if } \theta \neq \hat{\theta} \end{cases} \quad \text{zero-one}$$

$$L(\theta,\hat{\theta}) = I(|\hat{\theta} - \theta| > c) \quad \text{large deviation}$$

$$L(\theta,\hat{\theta}) = \int \log\left(\frac{p(x;\theta)}{p(x;\hat{\theta})}\right) p(x;\theta) \quad \text{- KL}$$

(*) vector losses

- $\theta = (\theta_1, \dots, \theta_k)$

$$L(\theta,\hat{\theta}) = \|\theta - \hat{\theta}\|_2^2 = \sum_{j=1}^{k} (\hat{\theta}_j - \theta_j)^2 \quad L_2\text{-loss}$$

$$L(\theta,\hat{\theta}) = \|\theta - \hat{\theta}\|_p = \left(\sum_{j=1}^{k} |\hat{\theta}_j - \theta_j|^p\right)^{1/p} \quad L_p\text{-loss}$$

(*) ML-context:

- prediction (classific) -- $L(Y, h(X)) = I(Y \neq h(X))$

  $Y \in \{0,1\}$

  class. $h(x)$

- Real-valued predic: $L(Y,\hat{Y}) = (Y - \hat{Y})^2$

- Risk fn:

$$R(\theta,\hat{\theta}) = \mathbb{E}_\theta(L(\theta,\hat{\theta})) = \int L(\theta,\hat{\theta}(x_1,\dots,x_n))\, p(x_1,\dots,x_n;\theta)\, dx^n$$

- Minimax: provides optimal estimator wrt to a given loss

(1st): loss depends on data when we evaluate it:-

$L(\theta,\hat{\theta}(x_1,\dots,x_n))$

(2nd): Risk does not, as we integrate the data out, by taking expectations of loss wrt the joint distn $p(x^n;\theta)$ that generated the data, at value of true param; risk is a function of the parmete θ

- L_2/squared error loss:-

$L(\theta,\hat{\theta}) = (\theta-\hat{\theta})^2 \Rightarrow R(\theta,\hat{\theta}) = MSE = \mathbb{E}_\theta[(\theta-\hat{\theta})^2] = (\mathbb{E}_\theta(\hat{\theta})-\theta)^2 + \mathbb{E}[(\hat{\theta}-\mathbb{E}_\theta(\hat{\theta}))^2]$
$= B^2 + V$

(?) unclear as to why we would write $R(\theta,\hat{\theta})$ if ...

(*) Minimax risk definition

(!): feels conceptually tricky.
- let's clarify.
- The definition of minimax risk and minimax estimator involves 'two-stage' minimisation/maximisation over both the estimator $\hat{\theta}$ and parmete θ.
- The comparison of two estimators case is best place to start, generalise from there.
- From Bickel & Doksum (1977):- (adapting not.) (clarity on loss, risk functions)
- $\hat{\theta}$ - estimator/procedure/decision rule
- L - loss function
- $X^n = x^n$ - experiment outcome.
- Loss: $L(\theta,\hat{\theta}(x_1,\dots,x_n))$
- (1st): We do not know the value of the loss, in general, as parmete θ is unknown

- Require our estimator to have desirable properties not just at one particular $x^\wedge$, but over a range of possible $x^\wedge$'s.
- so we use mean/average loss over the sample space
- Regard $L(\theta, \hat{\theta}(X_1, \ldots, X_n))$ as a random variable and introduce the risk function (not just 'risk') as a measure of the performance of the estimator $\hat{\theta}(X_1, \ldots, X_n)$:-

$$R(\theta, \hat{\theta}) = \mathbb{E}_\theta[L(\theta, \hat{\theta}(X_1, \ldots, X_n))]$$

- For each $\hat{\theta}$, R maps (H) to $\mathbb{R}^+$ i.e. $R: (H) \longrightarrow \mathbb{R}^+$ (function of unknown param θ).
- $R(\cdot, \hat{\theta})$ is a priori measure of performance of $\hat{\theta}$.

(W): Should be a lot clearer now; as I was getting confused by construction of risk function

(W): Start intuition from comparison of 2 estimators $\hat{\theta}_i$ and $\hat{\theta}_R$.
- For both estimators; compute their risk functions $R(\theta, \hat{\theta}_i)$, $R(\theta, \hat{\theta}_R)$, which are functions of the unknown parameter θ.
- In the examples; these are functions of the unknown param of our postulated statistical model/distribution family/collection of densities.
- Next bits explain the build-up to minimax risk and minimax est. formulae.
- We compare the risk functions $R_i(\theta, \hat{\theta}_i)$ and $R_R(\theta, \hat{\theta}_R)$ where we have introduced indexing according to estimator used.
- An estimator $\hat{\theta}_i$ is 'better' than $\hat{\theta}_R$ if

$$\max_\theta (R_i(\theta, \hat{\theta}_i)) < \max_\theta (R_R(\theta, \hat{\theta}_R)) \quad (\star)$$

or equivalently $\sup_\theta R_i(\theta, \hat{\theta}_i) < \sup_\theta R_R(\theta, \hat{\theta}_R)$ $(\star)$

Propose:
- An estimator $\hat{\theta}_i$ is 'better' if maximum of corresponding risk function is smaller than another.

- The issue with comparing risk functions; as examples will show:-
  i) How to deal with a set of candidate estimators $(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)$? systematically
  ii) How to deal with issue of inconclusiveness of comparing risk functions (i.e. one does not uniformly dominate another)

- we use maximum risk as we don't know true value of θ. - we assume a worst-case scenario when choosing $\hat{\theta}$ (as higher risk ⇒ higher expected loss)

- When choosing an estimator from a set of estimators, we generalise the 2-estimator case; and choose the estimator $\hat{\theta}$ which has the lowest maximum risk our all estimators $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \ldots$

- This yields minimax risk:

$$R_n = \inf_{\hat{\theta}} \sup_{\theta} R(\theta, \hat{\theta})$$

- For a minimax estimator $\hat{\theta}_{MINIMAX}$; its maximum risk $\left(\sup_{\theta} R(\theta, \hat{\theta}_{MINIMAX})\right)$ (our parameters θ)

will be the same as the

lowest maximum risk $\left(\inf_{\hat{\theta}} \sup_{\theta} R(\theta, \hat{\theta}) = R_n\right)$

(our all estimators $\hat{\theta}$) ↖ (as it is the minimax estimator... by defin.)

- Hence, an estimator $\hat{\theta}$ that is minimax satisfies:-

$$\sup_{\theta} R(\theta, \hat{\theta}) = \inf_{\hat{\theta}} \sup_{\theta} R(\theta, \hat{\theta}).$$

(iv): Conceptual trickiness turned out to be more simple than expected. In particular difficulty with minimax estimator condition.

- That's clarity for you.

1.3. Bayes Estimators

- treat θ as r.v.

- π - prior distn $\pi(\theta)$
- After observing $x^n = (x_1, \dots, x_n)$
- $L(\theta) = p(x^n; \theta)$ (likelihood)

via Bayes, the posterior density:-

$$p(\theta|x^n) = \frac{p(x^n|\theta)\pi(\theta)}{\int p(x^n|\theta)\pi(\theta)\,d\theta} = \frac{p(x^n|\theta)\pi(\theta)}{m(x^n)}$$

or $\pi(\theta|x^n)$

- Note marginal distn of $(x_1, \dots, x_n)$: $m(x^n) = m(x_1, \dots, x_n) = \int p(x^n, \theta)$

(A1) - Posterior risk

- Recall risk of an estimator $\hat{\theta}$ is the expectation of loss wrt the joint distn that generated the data $p(x^n; \theta)$.
- The posterior risk is integral of loss wrt posterior density/distn $\pi(\theta|x^n)$ or $p(\theta|x^n)$.
- Here density and distn used interchangeably.
- Posterior risk: (of an estimator $\hat{\theta}$)

$$r(\hat{\theta}|x^n) = \int L(\theta, \hat{\theta}(x^n)) p(\theta|x^n)\,d\theta$$

(*) compared to Bayes risk we are integrating out parameter θ i.e. taking expectation over posterior rather than prior.

Theorem 6 (AH/comp convenient form for Bayes risk).

$$B_\pi(\hat{\theta}) = \int r(\hat{\theta}|x^n) m(x^n)\,dx^n$$

- Let $\hat{\theta}(x^n)$ be value of θ that minimises $r(\hat{\theta}|x^n)$, $\hat{\theta}$ is Bayes estimator
- i.e. $\hat{\theta}(x^n)$ minimises posterior risk $r(\hat{\theta}|x^n)$ $\longrightarrow$ minimises Bayes risk $B_\pi(\hat{\theta})$ $\longrightarrow$ Bayes est.

(cont): Bayes risk can be viewed as:-

i) Expectation of usual risk $R(\theta,\hat{\theta})$ wrt prior distri $\pi(\theta)$.

ii) Expectation of posterior risk $r(\hat{\theta}|x^n)$ wrt marginal distri $m(x^n)$.

Theorem 7 (Bayes est with various losses)

- under L_2 loss / sq. error loss

i.e. $L(\theta,\hat{\theta}) = (\theta-\hat{\theta})^2$, Bayes est. :-

$$\hat{\theta}(x^n) = \int \theta\, \pi(\theta|x^n)\, d\theta = \mathbb{E}[\theta|X^n = x^n]$$ (i.e. mean of θ wrt posterior density)

- under absolute error loss

$L(\theta,\hat{\theta}) = |\theta-\hat{\theta}|$, Bayes est. is median of posterior $\pi(\theta|x^n)$

- under zero-one loss

$$L(\theta,\hat{\theta}) = \begin{cases} 0 \text{ if } \theta=\hat{\theta} \\ 1 \text{ if } \theta\neq\hat{\theta} \end{cases}$$, Bayes. est is mode of post. $\pi(\theta|x^n)$.

03/10/16 lecture terminates.

- Fill in blanks using notes, Wasserman book.

Example 8

- Let $X_1,\dots,X_n \sim N(\mu,\sigma^2)$, σ^2 known.
- suppose a $N(a,b^2)$ prior is used for μ and $L(\theta,\hat{\theta}) = (\theta-\hat{\theta})^2$ (square error / L_2)
- The Bayes est. with wrt L_2 loss is posterior mean

$$\hat{\theta}(X_1,\dots,X_n) = \frac{b^2}{b^2+\sigma^2/n}\bar{X} + \frac{\sigma^2/n}{b^2+\sigma^2/n}\,a$$

(*) Again, est. is convex combination of MLE and prior mean.

(*) This is exactly same result as in Example 8 under UNF.

(*) However, there we found the posterior mean and variance.
(i.e. compute this to yield point estimator $\longrightarrow$ Bayes estimator)

- still an issue of dealing with minimisation of maximum risk over all possible estimators (could be infinite)

(*) Example 4

- $X_1, \dots, X_n \sim Bern(p)$
- MLE: $\hat{p}_1 = \bar{X}_n$ (sample mean is unbiased) $\Rightarrow B = \mathbb{E}_p[\hat{p}_1] - p = 0$
- As $MSE = B^2 + V$ (or $\mathbb{E}_\theta[(\theta - \hat{\theta})^2] = (\mathbb{E}_\theta[\hat{\theta}] - \theta)^2 + Var_\theta(\hat{\theta})$) and $B = 0$;
- $MSE = \mathbb{E}_p[(p - \hat{p})^2] = Var_p(\hat{p}) = Var(\bar{X}_n)$

$$= \frac{p(1-p)}{n} \quad \text{as } Var(\bar{X}_n) = \frac{\sigma^2}{n}$$

- under L_2 loss i.e. $L(p, \hat{p}) = (p - \hat{p})^2$
- Risk function :- $\mathbb{E}_p[(p - \hat{p})^2]$, which is equal to MSE.

Hence $R(p, \hat{p}_1) = MSE = \frac{p(1-p)}{n}$

- Consider different estimator $\hat{p}_2 = \frac{Y + \alpha}{\alpha + \beta + n}$ where $Y = \sum_{i=1}^{N} X_i$ $\alpha, \beta > 0$ (post. mean using beta (α, β) prior).

(Aim): Find $R(p, \hat{p}_2)$:-

- unlike $\hat{p}_1$ (MLE, sample mean); it is unclear whether $\hat{p}_2$ is unbiased.
- Hence MSE is not necessarily equal to V.
- However equivalence of risk function $R(p, \hat{p})$ and MSE still holds due to assumption of L_2 loss.
- Evaluate MSE :-

$$MSE = B^2 + V = (\mathbb{E}_p[\hat{p}_2] - p)^2 + Var_p(\hat{p})$$

- Have to evaluate each; starting with :-

$$\mathbb{E}_p[\hat{p}_2] = \mathbb{E}\left[\frac{Y + \alpha}{\alpha + \beta + n}\right] = \frac{\mathbb{E}[Y] + \alpha}{\alpha + \beta + n} = \frac{\mathbb{E}[\sum_{i=1}^{N} X_i] + \alpha}{\alpha + \beta + n} = \frac{\sum_{i=1}^{N} \mathbb{E}[X_i] + \alpha}{\alpha + \beta + n}$$

$\Rightarrow \mathbb{E}_p[\hat{p}_2] = \frac{np+\alpha}{(\alpha+\beta+n)}$

- As $X_i \sim Bern(p)$; $\mathbb{E}[X_i] = p$ $Var(X_i) = p(1-p)$ $\mathbb{E}[X_i^2] = p$

- $Y = \sum_{i=1}^{n} X_i$

$\mathbb{E}[Y] = \sum_{i=1}^{n} \mathbb{E}[X_i] = np$ and $Var[Y] = Var\left[\sum_{i=1}^{n} X_i\right] = \sum_{i,j=1}^{n} Cov(X_i, X_j)$

$= \sum_{i=1}^{n} Var(X_i) + \sum_{i \neq j} Cov(X_i, X_j)$

$= \sum_{i=1}^{n} Var(X_i) = np(1-p)$

So $\mathbb{E}[Y^2] = Var[Y] + \mathbb{E}[Y]^2 = np(1-p) + n^2p^2$

- $Var_p(\hat{p}_2) = \mathbb{E}_p[\hat{p}_2^2] - \mathbb{E}_p[\hat{p}_2]^2$

$= \mathbb{E}\left(\frac{(Y+\alpha)^2}{(\alpha+\beta+n)^2}\right) - \frac{(np+\alpha)^2}{(\alpha+\beta+n)^2}$

$= \frac{\mathbb{E}[Y^2] + 2\alpha\mathbb{E}[Y] + \alpha^2}{(\alpha+\beta+n)^2} - \frac{(np+\alpha)^2}{(\alpha+\beta+n)^2}$

$= \frac{(n^2p^2 - np^2 + np + 2\alpha np + \alpha^2) - (n^2p^2 + 2\alpha np + \alpha^2)}{(\alpha+\beta+n)^2}$

$= \frac{np - np^2}{(\alpha+\beta+n)^2}$

$\Rightarrow Var_p(\hat{p}_2) = \frac{np(1-p)}{(\alpha+\beta+n)^2}$

(*) Alternatively: - $Var_p(\hat{p}_2) = Var_p\left(\frac{Y+\alpha}{\alpha+\beta+n}\right) = Var_p\left(\frac{1}{\alpha+\beta+n}(Y+\alpha)\right)$

$= \frac{Var(Y)}{(\alpha+\beta+n)^2} = \frac{np(1-p)}{(\alpha+\beta+n)^2}$

- As $Var(\alpha+X) = Var(X)$, $Var(kX) = k^2 Var(X)$.

Hence:-

$$R(p,\hat{p}_2) = \left(\frac{np+\alpha}{\alpha+\beta+n} - p\right)^2 + \frac{np(1-p)}{(\alpha+\beta+n)^2}$$

set $\alpha=\beta=\sqrt{n}/4 \Rightarrow$

$$\hat{p}_2 = \frac{Y+\sqrt{n}/4}{n+2\sqrt{n}/4} = \frac{Y+\sqrt{n}/4}{n+\sqrt{n}} \qquad R(p,\hat{p}_2) = \frac{n}{4(n+\sqrt{n})^2} \quad (*)$$

(Q1): Evaluate $R(p,\hat{p}_2)$ - computationally heavy; want to move on. to get result

(*) Key point: Comparing risk functions and even one number summaries of risk such as maximum risk is imperfect; often neither risk function uniformly dominates

---

(*) Summaries of risk fn:-

Maximum risk: $\bar{R}(\hat{\theta}) = \sup_\theta R(\theta,\hat{\theta})$

- understand (Q) asked now; student was proposing to use prior + Bayes to get around the pointwise maximisation of $R(\theta,\theta)$

Bayes risk (under prior π): $B_\pi(\hat{\theta}) = \int R(\theta,\hat{\theta})\pi(\theta)\,d\theta$

where $\pi(\theta)$ is a prior on θ.

(Q1): In context of earlier; Bayes risk amounts to putting a prior $\pi(\theta)$ encoding what we believe θ should be (note r.v now), and compute average of risk under this prior.

Example 5

- Consider earlier estimators $\hat{p}_1 = \bar{X}_n$ $\hat{p}_2 = \frac{Y+\alpha}{(\alpha+\beta+n)^2}$ $Y = \sum_{i=1}^n X_i$

- Compute maximum risk of both estimators.

$$\bar{R}(\hat{p}_1) = \max_{0\le p\le 1} \frac{p(1-p)}{n} = \frac{1}{4n} \quad (@\ p=\tfrac{1}{2})$$

$$\bar{R}(\hat{p}_2) = \max_{p} \frac{n}{4(n+\sqrt{n})^2} = \frac{n}{4(n+\sqrt{n})^2}$$

- As denom of $\bar{R}(\hat{p}_2)$ is larger than that of $\bar{R}(\hat{p}_1)$; $\bar{R}(\hat{p}_2) < \bar{R}(\hat{p}_1)$

(*) Now see diagram.

- When n is large; $\bar{R}(\hat{p}_1)$ has smaller risk except for region near $p=\frac{1}{2}$
- So on this basis; could argue for $\hat{p}_1$ over $\hat{p}_2$.

(6m) one number summaries eg. max risk ⟶ imperfect.

- Example (from Wasserman):-
- Integrating
- Define prior $\pi(p) = 1$

$$B_\pi(\hat{p}_1) = \int R(p,\hat{p}_1)\,dp = \int \frac{p(1-p)}{n}\,dp = \frac{1}{n}\left[\frac{1}{2}p^2 - \frac{1}{3}p^3\right]_0^1$$

$$= \frac{1}{6n}$$

$$B_\pi(\hat{p}_2) = \int R(p,\hat{p}_2)\,dp = \int \frac{n}{4(n+\sqrt{n})^2}\,dp = \frac{n}{4(n+\sqrt{n})^2}$$

- $n \ge 20$ $B_\pi(\hat{p}_2) > B_\pi(\hat{p}_1)$ ⇒ $\hat{p}_1$ is better estimator.

(*) Answer depends on prior choice; illustrates why frequentists prefer maximum risk.

(*) Two risk summaries (maximum risk, Bayes risk) yield two methods for deriving estimators (selecting amongst?).

- select $\hat{\theta}$ to minimise max risk $\bar{R}(\hat{\theta})$ ⟶ minimax estimator
- select $\hat{\theta}$ to —"— Bayes risk $B_\pi(\hat{\theta})$ ⟶ Bayes estimator with respect to prior π

- Bayes estimator:

$\hat{\theta}$: $B_n(\hat{\theta}) = \inf_{\hat{\theta}} B_n(\hat{\theta})$

- inf/minimum over all estimators.

- Minimax estimator:

$\hat{\theta}$: $\sup R(\theta, \hat{\theta}) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} R(\theta, \hat{\theta})$ (*)

$= R_n$

$R_n = R_n(\Theta) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} R(\theta, \hat{\theta})$

(*) Not entirely clear: the notation $\theta \in \Theta$

- But otherwise secure

- Think it just means max over entire parameter space (+)

- RHS of above, R_n, is the minimax risk

(+) - statistical decision theory involves:-

I) Determine minimax risk

II) Find an estimator that achieves this risk.

(*) Having found minimax risk ($R_n = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} R(\theta, \hat{\theta})$, we find the minimax estimator that achieves this risk. :- $\sup_{\theta \in \Theta} R(\theta, \hat{\theta}) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} R(\theta, \hat{\theta})$

Asymptotically minimax estimator

$\hat{\theta}$: $\sup_{\theta \in \Theta} R(\theta, \hat{\theta}) \sim \inf_{\hat{\theta}} \sup_{\theta \in \Theta} R(\theta, \hat{\theta})$ $n \to \infty$

where $a_n \sim b_n$ means $\frac{a_n}{b_n} \longrightarrow 1$

(*) may be too difficult, settle for estimator $\hat{\theta}$ that achieves the minimax rate.

Minimax rate estimator

$\hat{\theta}$ $\sup_{\theta \in \Theta} R(\theta, \hat{\theta}) \asymp \inf_{\hat{\theta}} \sup_{\theta \in \Theta} R(\theta, \hat{\theta})$ $n \to \infty$

where $a_n \asymp b_n$ mean that both $\frac{a_n}{b_n}$ and $\frac{b_n}{a_n}$ are both bounded as $n \to \infty$.

(*) Hence; this builds on that result.

- Because we are using T.7 to establish a connection between Bayes estimator as posterior mean under L_2 loss

---

1.4 Minimax Estimators

(07) Bayes estimators with a constant risk function are minimax.

Theorem 9

- Let $\hat{\theta}$ be the Bayes estimator for some prior π
- If / Suppose that

$$R(\theta, \hat{\theta}) \leq B_\pi(\hat{\theta}) \quad \text{for all } \theta$$

then $\hat{\theta}$ is minimax and π is called a least fav. prior

Proof

- suppose $\hat{\theta}$ is not minimax ✓
- then there will be another estimator $\hat{\theta}_0$ such that ✓

$$\sup_\theta R(\theta, \hat{\theta}_0) < \sup_\theta R(\theta, \hat{\theta})$$

(*) As the average of a function is always less than or equal to its maximum, we have ✓

$$B_\pi(\hat{\theta}_0) \leq \sup_\theta R(\theta, \hat{\theta}_0)$$

Hence:

$$B_\pi(\hat{\theta}_0) \leq \sup_\theta R(\theta, \hat{\theta}_0) < \sup_\theta R(\theta, \hat{\theta}) \leq B_\pi(\hat{\theta})$$

(?)

- which is a contradiction.

(*) I am not entirely sure where contradiction is; except it must be in the section I've outlined. NO

(*) comfortable with supposing $\hat{\theta}$ is not minimax → see

(*) Does the contradiction reside in the fact that :-

$$\sup_\theta R(\theta, \hat{\theta}) \leq B_\pi(\hat{\theta}) \qquad \underline{NO}$$

- return to this
- stackexchange (see Wasserman)

(Q5 2) - see later part.

Theorem 10

- Suppose $\hat{\theta}$ is Bayes est wrt prior π, and $\hat{\theta}$ has constant risk $R(\theta, \hat{\theta}) = c$ for some c
- If the risk is constant, $\hat{\theta}$ is minimax.

(?) - Bayes risk or (risk constant?)

Proof:

Bayes risk :- $B_\pi(\hat{\theta}) = \int R(\theta, \hat{\theta}) \pi(\theta) d\theta = c$ (?) 6153

Hence $R(\theta, \hat{\theta}) \leq B_\pi(\hat{\theta})$ for all θ

$\Rightarrow \hat{\theta}$ is minimax via T.9.

(*) inside examples, will help with queries.

(*) Example 11

- see ex 4
- $X_1, \ldots, X_n \sim Bern(p)$ L_2 loss i.e. $L(\theta, \hat{\theta}) = (\theta - \theta)^2$
- showed estimator $\hat{p}(X^n) = \dfrac{\sum_{i=1}^n X_i + \sqrt{n}/4}{n + \sqrt{n}}$ ($\hat{p}_2$ is ex 4).

- $\hat{p}_2$ has a constant risk function

- $R(p, \hat{p}_2) = \dfrac{n}{4(n+\sqrt{n})^2}$ i.e. not a function of parameter p

- This estimator is posterior mean under beta prior (α, β) i.e. $\pi(\alpha, \beta) = Beta(\alpha, \beta)$ (from previous example).
- As we have L_2 loss $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$, via Theorem 7, this estimator is the Bayes estimator (i.e. posterior mean under L_2 loss is Bayes est.)
- That is, minimises the Bayes risk under a beta (a, b) prior with $\alpha = \beta = \sqrt{n}/4$.
- via Theorem 10, as risk function $R(p, \hat{p}_2)$ is constant and $\hat{p}_2$ is Bayes est, $\hat{p}_2$ is minimax

Example 12

- consider Bernoulli $X_1, \ldots, X_n \sim \text{Bern}(p)$
- define following loss:

$$L(p, \hat{p}) = \frac{(p-\hat{p})^2}{p(1-p)}$$

- define following estimator:-

$$\hat{p}(x^n) = \hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Recall r.v. is the estimator
- expectation wrt distr. param by p

- The risk is:

$$R(p, \hat{p}) = \mathbb{E}[L(p, \hat{p})] = \mathbb{E}\left(\frac{(p-\hat{p})^2}{p(1-p)}\right) = \frac{1}{p(1-p)} \underbrace{\mathbb{E}_p[(p-\hat{p})^2]}_{\text{MSE}}$$

$$= \frac{1}{p(1-p)} \text{var}_p(\hat{p})$$

- As $\hat{p}$ is unbiased
- $\text{var}(\bar{X}_n) = \frac{\sigma^2}{n}$

$$= \frac{1}{p(1-p)} \frac{p(1-p)}{n}$$

$$= \frac{1}{n}$$

Hence $R(p, \hat{p})$, as fn of p, is constant

(*) can be shown that for this loss function $\hat{p}(x^n)$ is Bayes estimator under prior $\pi(p) = 1$.

(*) under Theorem 10, $\hat{p}$ is minimax

↳ Bayes risk:

$$B_\pi(\hat{p}) = \int_0^1 R(p, \hat{p}) \pi(p) \, dp = \int_0^1 \frac{1}{n} \cdot 1 \, dp = \frac{1}{n}[p]_0^1 = \frac{1}{n}$$

↑

- (O/S 4) - show this. Don't understand why $\hat{p}$ is a minimax estimator; as I can't see how it is the Bayes estimator i.e. $\hat{p}$ minimises Bayes risk.
- surely any $\hat{p}$ is Bayes est. in this case?

Theorem 9 (redo → Wasserman book helps) (did not understand proof by contra)

- Let $\hat{\theta}$ be the Bayes estimator for some prior π; then $\hat{\theta}$ satisfies
- $B_\pi(\hat{\theta}) = \inf_{\tilde{\theta}} B_\pi(\tilde{\theta})$ (on) (via definition of Bayes est.)
- Suppose that $R(\theta, \hat{\theta}) \leq B_\pi(\hat{\theta})$ for all θ
- Then $\hat{\theta}$ is minimax and π is a least favourable prior

Proof (focus on structure)

- Assume that $\hat{\theta}$ (Bayes estimator wrt prior π) is not minimax
- Contradiction is that we are able to find another estimator $\hat{\theta}_0$ such that $B_\pi(\hat{\theta}_0) \leq B_\pi(\hat{\theta})$; which contradicts our assumption that $\hat{\theta}$ is the Bayes estimator; and minimises the Bayes risk - we have found an estimator with lower Bayes risk $B_\pi(\hat{\theta}_0)$ as consequence of assuming $\hat{\theta}$ is not minimax.
- We therefore conclude that $\hat{\theta}$ is minimax
- On the issue of $\sup_\theta R(\theta, \hat{\theta}) \leq B_\pi(\hat{\theta})$ (which you thought was contradict.)
- We assume that $R(\theta, \hat{\theta}) \leq B_\pi(\hat{\theta})$
- i.e. for our Bayes estimator $\hat{\theta}$, the risk function $R(\theta, \hat{\theta})$ is smaller than the Bayes risk over all values of the parameter θ.

(*) This means that $\sup_\theta R(\theta, \hat{\theta}) \leq B_\pi(\hat{\theta})$ by assumption

$\sup_\theta R(\theta, \hat{\theta})$ is just maximal value of $R(\theta, \hat{\theta})$ evaluated at a particular value of θ (param)

(*) we assume $R(\theta, \hat{\theta}) \leq B_\pi(\hat{\theta}) \ \forall \theta$

Excellent

minimax estimators - normal model

- A function ℓ is bowl-shaped if
- sets $\{x : \ell(x) \leq c\}$ are convex and symmetric about origin
- A loss function L is bowl-shaped if $L(\theta, \hat{\theta}) = \ell(\theta - \hat{\theta})$ for some bowl-shaped function ℓ.

Theorem 13

- suppose r.vec X has a normal distn with mean vector θ and cov matrix Σ.
- If the loss-function is bowl-shaped then X is the unique* minimax estimator of θ.
- up to sets of measure zero.

(*) Param space restricted, above does not apply.

(?) Wasserman states in T. 12.14 for $\bar{X}$ → is there a typo?

Example 14

- Suppose $X \sim N(\theta, 1)$ and $\theta \in [-m, m]$ where $0 < m < 1$ (i.e. restricted param space)
- under L_2 loss (squared error loss), unique minimax est is:-

$$\hat{\theta}(x) = m \tanh(mx) = \frac{e^{mx} - e^{-mx}}{e^{mx} + e^{-mx}}$$

- This is Bayes est wrt prior that puts mass $\frac{1}{2}$ at m and mass $\frac{1}{2}$ at $-m$.
- Risk is not constant but it does satisfy $R(\theta, \hat{\theta}) \leq B_\pi(\hat{\theta})\ \forall \theta$.
- $R(\theta, \hat{\theta})$
- via T. 9. $\hat{\theta}$ is minimax

(**) modern minimax theory → now is param space restricted?

(#) clarity on least favourable priors

WIKI:

- estimator is minimax when it does best in the worst case
- A minimax estimator should be a Bayes estimator w.r.t. least favourable prior distribution of θ.
- Bayes risk of a Bayes estimator $\hat{\theta}$ wrt prior $\pi(\theta)$:-

$$B_\pi(\hat{\theta}) = \int R(\theta, \hat{\theta})\, \pi(\theta)\, d\theta$$

- A prior distri $\pi(\theta)$ is least favourable if for every other prior distri π' the Bayes risk satisfies:-

$$B_\pi(\hat{\theta}) \geq B_{\pi'}(\hat{\theta})$$

(Bayes risk of Bayes est. under least favourable prior is the highest) (hence worst case)

- If $B_\pi(\hat{\theta}) = \sup_\theta R(\theta, \hat{\theta})$

then i) $\hat{\theta}$ is minimax

ii) $\hat{\theta}$ a unique Bayes estimator and unique minimax est.

iii) π is least favourable

(O/S 5): not sure how to reconcile with LN notes.

(*) Proof that

$\bar{X}_n$ is minimax mean squared error loss (fairly involved)

- strategy consists of

i) upper bound minimax risk R_n using maximum risk of arbitrary est

ii) lower bound minimax risk R_n using Bayes risk eval. at Bayes est. under (minimise of Bayes risk).

- computation.

- See notes on justifying how we bound minimax risk R_n
- (Ex) idea:-

$$B_\pi(\hat{\theta}_\pi) \leq \underbrace{\inf_{\hat{\theta}} \sup_{\theta} R(\theta, \hat{\theta})}_{R_n \text{ arb.}} \leq \sup_{\theta} R(\theta, \hat{\theta}_*)$$

- $B_\pi(\hat{\theta}_\pi)$ - Bayes risk eval. at Bayes est. wrt prior π (i.e. minimiser of Bayes risk)
- $\sup_{\theta} R(\theta, \hat{\theta}_*)$ - maximum risk for an arbitrary est. $\hat{\theta}_*$

- computation of upper and lower bounds.
- select sample mean $\bar{X}$ as arbitrary est.
- under L_2 loss, risk of estimator $\hat{\theta} = \bar{X}$ and MSE are the same.

Risk of $\hat{\theta}$: $R(\theta, \hat{\theta}) = \mathbb{E}_\theta[L(\theta, \hat{\theta}(X_1, \ldots, X_n)] = \mathbb{E}_\theta[(\theta - \hat{\theta})^2]$

L_2 loss - $L(\theta, \hat{\theta}(X_1, \ldots, X_n)) = (\theta - \hat{\theta})^2$

MSE of $\hat{\theta}$: $MSE = \mathbb{E}_\theta[(\hat{\theta} - \theta)^2] = B^2 + V$ cont. → pto

---

(Q5 6) - Got confused by L.W. claiming that risk and MSE under L_2 loss are the same; in particular due to sign differences
i.e. $\mathbb{E}_\theta[(\theta - \hat{\theta})^2]$ vs $\mathbb{E}_\theta[(\hat{\theta} - \theta)^2]$

- Guessing this is immaterial.
- This is elementary, but needed to be cleared intuitively and formally
- Intuit: squared differences → order $(\hat{\theta} - \theta)^2$ vs $(\theta - \hat{\theta})^2$ doesn't matter.
- Formally: $(\hat{\theta} - \theta)^2 = \hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2 = (\theta - \hat{\theta})^2$!
- Well done for raising this and clearing → prevents confusion

---

$MSE = B^2 + V$

$= V$ as $\bar{X}$ is unbiased.

$\mathrm{Var}_\theta(\bar{X}) = \frac{1}{n}$

(IS 1): this is elementary, but I can't quite see why $\mathrm{Var}_\theta(\bar{X}) = \frac{1}{n}$. Return to this. ✓

- easy. We assumed that $X_1, \dots, X_n \sim N(\theta, 1)$! "

- so $\mathrm{Var}_\theta(\bar{X}) = \mathrm{Var}_\theta\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\mathrm{Var}_\theta\left(\sum_{i=1}^n X_i\right)$

$= \frac{1}{n^2}\sum_{i=1}^N \mathrm{Var}_\theta(X_i)$ as $X_1, \dots, X_n$ IID

$= \frac{n}{n^2} = \frac{1}{n}$ ☺

Hence $R(\theta, \bar{X}_n) = \frac{1}{n} \quad \forall f(\theta)$

so $\sup_\theta R(\theta, \bar{X}_n) = \frac{1}{n} \Rightarrow R_n \leq \frac{1}{n}$ (upper bound)

- lower bounding R_n. recall $X_1, \dots, X_n \sim N(\theta, 1)$

- Assume conjugate prior $\pi(\theta) = N(0, c^2)$
- Posterior $p(\theta | x_1, \dots, x_n) \sim N(a, b)$ (due to conjugacy).
- can evaluate using same start as eg. 8 in LN7 (via trick of square completion)
- Result from Ex 8:
  - for $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ σ^2 known

  conjugate prior $p(\mu) \sim N(m, \tau^2)$

then mean and variance of posterior $p(\mu | X_1, \dots, X_n)$ is given by:-

$$\mathbb{E}[\mu | X_1, \dots, X_n] = \frac{\tau^2 n}{\tau^2 n + \sigma^2}\bar{x} + \frac{\sigma^2}{\tau^2 n + \sigma^2} m$$

$$\mathrm{var}[\mu | X_1, \dots, X_n] = \frac{\sigma^2 \tau^2 / n}{\tau^2 + \sigma^2/n}$$

- now case; $\mu=0$ $\sigma^2=1$ $m=0$ $\tau^2=c^2$

post.

$\sim p(\theta | X_1=x_1, X_2=x_2, \ldots, X_n=x_n)$

has $\mathbb{E}[\theta | X_1, \ldots, X_n] = \frac{c^2 n}{c^2 n+1} \bar{X} + \frac{1}{c^2 n+1}(0)$

$$= \frac{nc^2 \bar{X}}{nc^2+1}$$

$$\text{Var}[\theta | X_1, \ldots, X_n] = \frac{c^2/n}{c^2+1/n} = \frac{c^2}{nc^2+1}$$

$$\Rightarrow \theta | X^n = x_n \sim N\left(\frac{nc^2\bar{X}}{nc^2+1}, \frac{c^2}{nc^2+1}\right)$$

- Under L_2 loss, Bayes est. is post. mean

$$\Rightarrow \hat{\theta}_\pi = \frac{nc^2\bar{X}}{nc^2+1} \quad \text{(Bayes est.)}$$

(*) Again using fact that risk function and MSE are same under L_2 loss.

$$R(\theta, \hat{\theta}_\pi) = MSE = B^2 + V$$

$$MSE = \mathbb{E}_\theta[(\hat{\theta}_\pi - \theta)^2] = (\mathbb{E}_\theta[\hat{\theta}_\pi] - \theta)^2 + \text{Var}_\theta(\hat{\theta}_\pi)$$

(OIS 8)

Bias: $\mathbb{E}_\theta(\hat{\theta}_\pi) - \theta = \frac{nc^2 \mathbb{E}[\bar{X}]}{nc^2+1} - \theta = \frac{-\theta}{1+nc^2}$ i.e. $O(\frac{1}{n})$

variance: $\text{Var}_\theta(\hat{\theta}_\pi) = \text{Var}_\theta\left(\frac{nc^2\bar{X}}{nc^2+1}\right) = \text{Var}_\theta\left(\frac{nc^2}{nc^2+1}\bar{X}\right)$

$$= \left(\frac{nc^2}{nc^2+1}\right)^2 \text{Var}_\theta(\bar{X}) = \frac{n^2c^4}{(nc^2+1)^2} \cdot \frac{1}{n} = \frac{nc^4}{(1+nc^2)^2}$$

- Not L_2 loss $\longrightarrow$ take exp. of loss to get risk.

$$R(\theta, \hat{\theta}_\pi) = \text{MSE} = B^2 + V = \left(\frac{-\theta}{nc^2+1}\right)^2 + \frac{nc^4}{(nc^2+1)^2}$$

under L_2 loss

$$= \frac{\theta^2 + nc^4}{(1+nc^2)^2}$$

- This is the risk function $R(\theta, \hat{\theta})$.
- We want the Bayes risk of the Bayes est :-

$$B_\pi(\hat{\theta}_\pi) = \int R(\theta, \hat{\theta}_\pi)\pi(\theta)\, d\theta \qquad \pi(\theta) \sim N(0, c^2)$$

$$= \int \frac{\theta^2 + nc^4}{(1+nc^2)^2}\pi(\theta)\, d\theta$$

$$= \mathbb{E}_{\pi(\theta)}\left[\frac{\theta^2 + nc^4}{(1+nc^2)^2}\right] \qquad \mathbb{E}_\pi[\theta^2] = c^2 \text{ as } \mathbb{E}_\pi[\theta] = 0$$

$$= \frac{\mathbb{E}_{\pi(\theta)}[\theta^2] + nc^4}{(1+nc^2)^2} = \frac{c^2 + nc^4}{(1+nc^2)^2} = \frac{c^2}{(1+nc^2)}$$

$$\Rightarrow \quad R_n \geq B_\pi(\hat{\theta}_\pi) = \frac{c^2}{(1+nc^2)}$$

$$\Rightarrow \quad \frac{c^2}{(1+nc^2)} \leq R_n \leq \frac{1}{n} \qquad \forall c$$

c arbitrary

Hence $\lim_{c \to \infty} \left(\frac{c^2}{1+nc^2}\right) = \frac{1}{n}$

$$\Rightarrow \frac{1}{n} \leq R_n \leq \frac{1}{n} \Rightarrow R_n = \frac{1}{n}.$$

- Hence minimax risk R_n is $\frac{1}{n}$
- We now need to find estimator $\hat{\theta}$ that achieves the minimax risk. R_n

- use $\hat{\theta} = \bar{X}$; $\sup_\theta R(\theta, \hat{\theta}) = \frac{1}{n} = R_n$

Hence $\bar{X}_n$ is <u>minimax</u>

(W) A little confused by what context above is w.r.t.

- Notes say $n=1$ and $\sigma^2 = 1$, but I think this means $d=1$ and $\sigma^2 = 1$
i.e. univariate Gaussian, $X_1, \dots, X_n \sim N(\theta, 1)$
$\sigma^2 = 1$ variance
- in more generality, we want the d-dim multivariate Gaussian and σ^2 i.e. $\underline{X}_1, \dots, \underline{X}_n \sim N(\underline{\theta}, \sigma^2 \underline{I}_d)$
where $\underline{X}_1 \in \mathbb{R}^d$ $\underline{\theta} \in \mathbb{R}^d$ $\sigma^2 \underline{I}_d \in \mathbb{R}^{d \times d}$