19.12.2022

YZV 231E

Probability Theory & Stats

Week 13

Gü.

Recap : $\boxed{\text{Bayesian}}$ Estimation :

unknown Θ : r.v. → prior distribution. (prior belief) $p_{\Theta}(\theta)$

Observation : measurement $p_{X|\Theta}(x|\theta)$ → X r.v. (we observe) → Estimator → $\hat{\Theta}$: estimate $= g(X)$

calculate

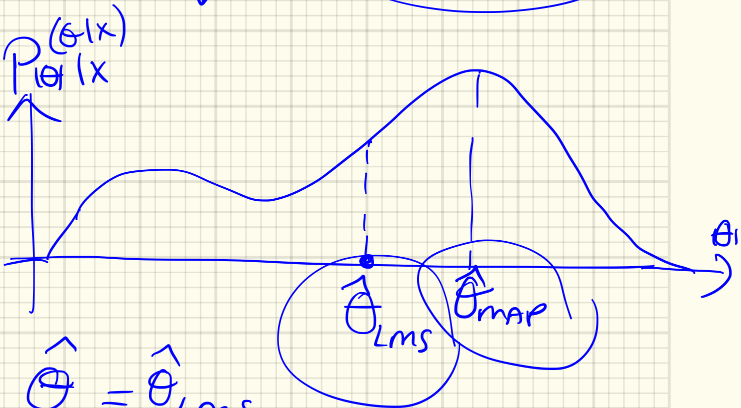
$p_{\Theta|X}(\theta|x)$: posterior distribution via Bayes rule.

1) MAP : $\hat{\theta}_{MAP} = \arg\max_{\theta} p_{\Theta|X}(\theta|x)$

2) LMS : $\hat{\theta}_{LMS} = E_{\Theta|X}[\theta|x]$

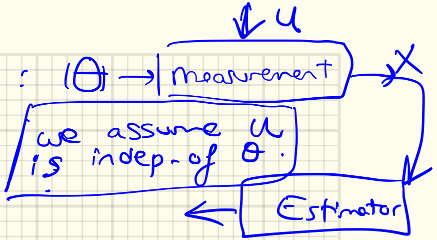
3) Linear LMS :

$g(x) = \hat{\theta} = aX + b$



$\hat{\theta} = \hat{\theta}_{LMS}$.

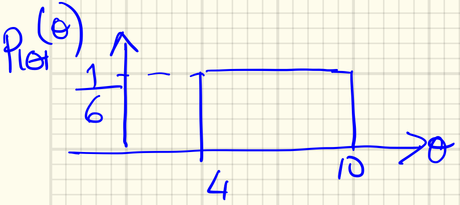
(Ex 8.11 Bertsekas): Suppose $\Theta \sim U[4,10]$:

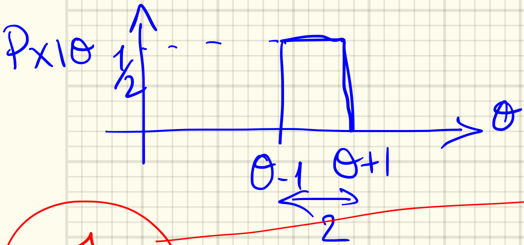


Let's say measurement noise $U \sim [-1,1]$;

Observation r.v. $\boxed{X = \Theta + U}$

Given $\Theta = \theta$ $\quad X|_{\Theta=\theta} \sim U[\theta - 1, \theta + 1]$





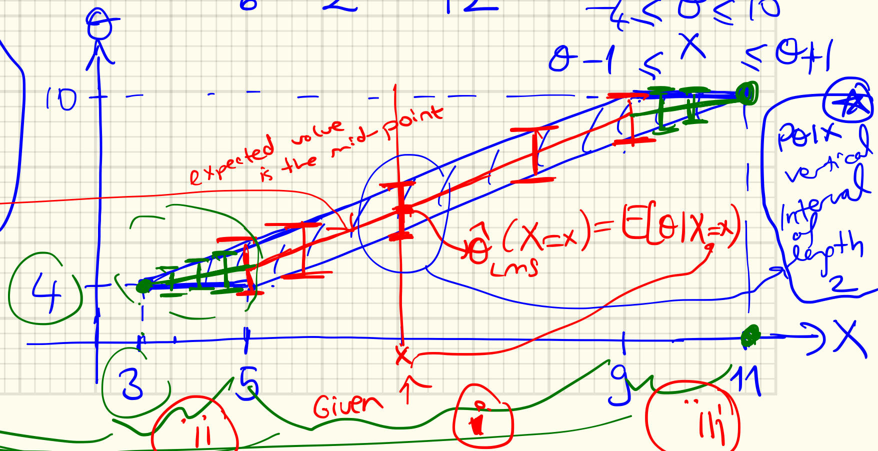
construct the joint density btw Θ, X :

$$p_{X,\Theta} = p_\Theta \cdot p_{X|\Theta}$$ ✓ general formula

$$= \frac{1}{6} \cdot \frac{1}{2} = \frac{1}{12}$$ on support of (X,θ)

$4 \le \theta \le 10$

$\theta - 1 \le x \le \theta + 1$



$\hat{\theta}_{LMS} = E[\Theta|X]$

An Estimator we picked.

3 intervals

Q. How good is this estimate ($\hat{\Theta}_{LMS}$)? Conditional mean-squared Error

$$E\Big[\big(\Theta - \underbrace{E[\Theta|X]}_{\text{we made this estimate } = \hat{\Theta}_{LMS}}\big)^2 \,\Big|\, X=x\Big] = \text{Variance of the posterior distrib.}$$

$\Theta|X$

true value of Θ

$$= \int p_{\Theta|X}(\theta|x)\,\big(\theta - E[\theta|x]\big)^2\, d\theta.$$

Look at Interval (i) (previous plot): previous page ★

$5 \le X \le 9$

$p_{\Theta|X}(\theta|X=x)$ is uniform when $X=x$ fixed on an interval of 2 (vertical)!

Var for uniform distrib of width $=2$

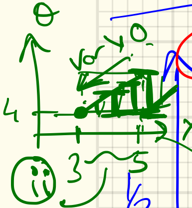
$\frac{1}{b-a}$

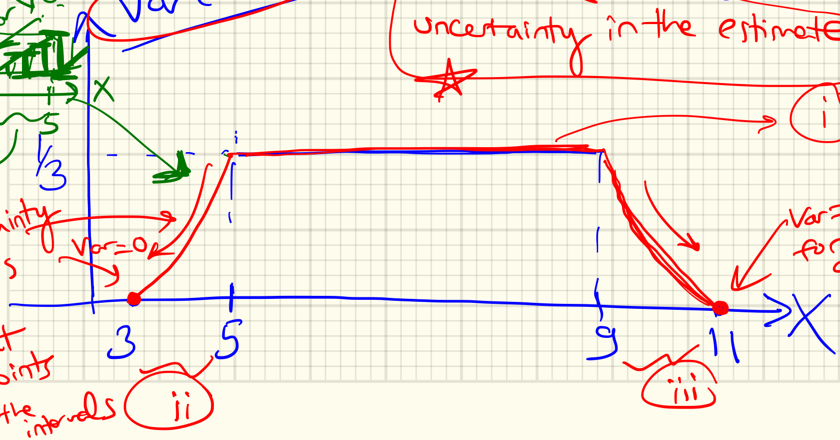
$$\frac{(b-a)^2}{12} = \frac{4}{12} = \frac{1}{3}$$

$\mathrm{Var}(\Theta|X=x)$

Variance $\propto$ uncertainty in the estimate ★

(i) Highest uncertainty.





Some properties of the LMS Estimator: $\hat{\Theta}_{LMS} = E[\Theta | X]$

$\Theta | X$

Estimation Error:

$$\boxed{\underset{r.v.}{\tilde{\Theta}} = \underset{r.v.}{\hat{\Theta}} - \underset{r.v.}{\Theta}}$$

$\hat{\Theta}(X)$ ($X \rightarrow \boxed{\text{Estimator}} \rightarrow \hat{\Theta}(X)$)

$$E[\tilde{\Theta} | X] = E[(\hat{\Theta} - \Theta) | X] = E[\hat{\Theta} | X] - \underbrace{E[\Theta | X]}_{\text{by definition}}$$

$\hat{\Theta}(X) | X$

known given X

$\hat{\Theta}_{LMS}$

$$= \hat{\Theta} - \hat{\Theta} = 0$$

iterate

$$E[\tilde{\Theta} | X] = 0$$

$$E[E[\tilde{\Theta} | X]] = \underbrace{E[\tilde{\Theta}]}_{\ldots} = 0$$

$\Rightarrow$ Estimator $\hat{\Theta}$ is UNBIASED [b/c its error is zero. $E[\tilde{\Theta}] = 0$]

$$Cov(\underset{r.v.}{\tilde{\Theta}}, \underset{r.v.}{\hat{\Theta}}) = ? = E[\tilde{\Theta} \cdot \underset{h(X)}{\underbrace{\hat{\Theta}}}] - E[\tilde{\Theta}] \cdot E[\hat{\Theta}] \quad \text{: Recall}$$

$\longrightarrow$

$$E[\tilde{\theta}\cdot \underbrace{h(X)}\mid X] = ?$$ Given X, $h(X)$ is a number.

$$h(X)\cdot \underbrace{E[\tilde{\theta}\mid X]} = 0$$

$$E[E[\tilde{\theta}\, h(X)\mid X]] = \boxed{E[\tilde{\theta}\cdot h(X)] = 0}$$ for any function $h(X)$

We know $\hat{\theta}(X)$ is a function of X.

$$E[\tilde{\theta}\cdot\hat{\theta}] = 0$$

$$\operatorname{Cov}(\tilde{\theta},\hat{\theta}) = \underbrace{\cancel{E[\tilde{\theta}\cdot\hat{\theta}]}}_{0} - \underbrace{\cancel{E[\tilde{\theta}]}}_{0}\cdot E[\hat{\theta}]$$

$$\operatorname{Cov}(\tilde{\theta},\hat{\theta}) = 0$$

$\therefore$ The $\underbrace{\text{estimation error}}_{\tilde{\theta}}$ is uncorrelated w/ the $\underbrace{\text{estimate}}_{\hat{\theta}}$. $\Longrightarrow$

$\tilde{\Theta}$ & $\hat{\Theta}$ are uncorrelated. $\tilde{\Theta} = \Theta - \hat{\Theta} \rightarrow \boxed{\Theta = \hat{\Theta} + \tilde{\Theta}}$

Variance ~ a measure of uncertainty :

$$\underbrace{\mathrm{Var}(\underset{\hat{\Theta}+\tilde{\Theta}}{\Theta})}_{\text{uncertainty in the r.v. } \Theta} = \underbrace{\mathrm{Var}(\hat{\Theta})}_{\text{uncertainty in the estimate}} + \underbrace{\mathrm{Var}(\tilde{\Theta})}_{\text{uncertainty in the error.}} \cancel{+2\,\mathrm{Cov}(\tilde{\Theta}, \hat{\Theta})}$$

$\mathrm{Cov}(\tilde{\Theta}, \hat{\Theta}) = 0$

Recall Linear LMS Estimator: $\Theta \rightarrow$ measurement $\xrightarrow{X}$ Estimator $\rightarrow \hat{\Theta}_L = g(X)$

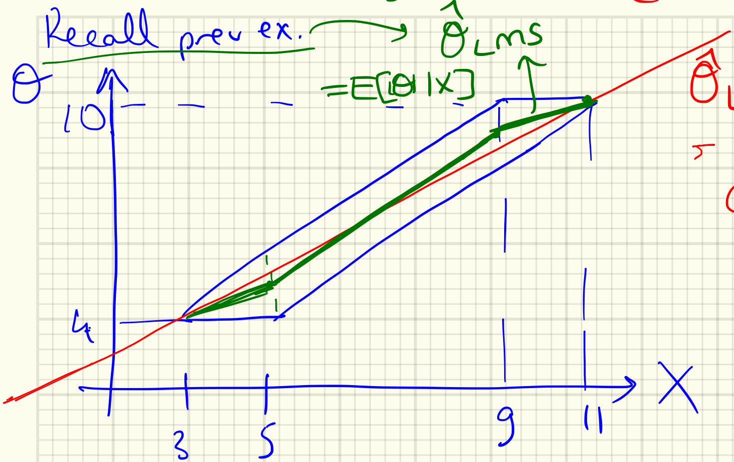
$g(X) = \boxed{a} X + \boxed{b}$: affine mapping of X. (linear + offset) Parameters a & b define the mapping.

$$\min_{a,b} E[(\Theta - (aX+b))^2] \rightarrow \begin{cases} a = \dfrac{\mathrm{cov}(X,\Theta)}{\mathrm{Var}(X)} \\ b = E[\Theta] - \dfrac{\mathrm{cov}(X,\Theta)}{\mathrm{Var}(X)} E(X) \end{cases}$$

exercise
Derive a & b.

"Best" linear estimator (in the MSE sense)

$$\boxed{\hat{\Theta}_L = E[\Theta] + \frac{\mathrm{Cov}(X,\Theta)}{\mathrm{Var}(X)}(X - E[X])}$$



$\mathrm{cov}(X,\Theta) > 0$

linear estimator

$$\hat{\Theta}_L =: E[\Theta] + \underbrace{\frac{\mathrm{cov}(X,\Theta)}{\mathrm{Var}(X)}}_{>0}\left(X - E[X]\right)$$

Q. If $\mathrm{cov}(X,\Theta) = 0$ → knowing X does not help me at all.

$$\hat{\Theta}_L = E[\Theta]$$

Q. If $\mathrm{cov}(X,\Theta) > 0$

If $X > E[X]$

$$\hat{\Theta}_L > E[\Theta]$$

Q. What is the MSE?

$$\boxed{E\left[\left(\hat{\Theta}_L - \Theta\right)^2\right] = (1-\rho^2)\sigma_\Theta^2}$$

exercise: derive this result

$\rho = \dfrac{\mathrm{cov}(X,\Theta)}{\sigma_X \cdot \sigma_\Theta}$: correlation coefficient.

$$\mathrm{MSE}(\hat{\Theta}_L) = (1-\rho^2)\underbrace{\sigma_\Theta^2}_{\text{variance of the original r.v.}}$$

→ Interpret: i) $\sigma_\Theta^2 \nearrow$ → $\mathrm{MSE} \nearrow$

$$\boxed{MSE(\hat{\theta}_L) = (1-\rho^2)\,\sigma_\theta^2}$$

$0 \leq |\rho| \leq 1$.

ii) When X, θ are correlated ($\rho \neq 0$),

MSE $\downarrow$ we improve our estimate $\rightarrow$ its uncertainty is reduced.

iii) $\rho = 1$ $\longrightarrow$ MSE $= 0$

$\rho = -1$

maximal correlation case.

2 r.v.s are linearly related.

iv) $\rho = 0$: $\underbrace{MSE = \sigma_\theta^2}$: measurements X don't help us improve our estimate

2 r.v.s are uncorrelated. $\hookrightarrow$ uncertainty is not reduced.

Linear LMS w/ MULTIPLE DATA : We make several measurements $X_1, \dots, X_n$.

Linear Estimator is of the form:

$$\hat{\Theta} = a_1 X_1 + \dots + a_n X_n + b$$

→ Find the "best" coefficients $a_1, a_2, \dots, b$

"optimal linear LMSE estimator" → in MSE sense,

Recall general LMSE estimator:

$$\hat{\Theta} = E[\Theta \mid X_1, \dots, X_n]$$

conditional expectation

$$\min_{a_1, \dots a_n, b} \underbrace{E[(\hat{\Theta} - \Theta)^2]}_{\text{cost fn} = \text{cost}} = E[(aX_1 + \dots + a_n X_n + b - \Theta)^2]$$

$$\frac{\partial\, \text{Cost}}{a_1} = 0, \quad \frac{\partial\, \text{Cost}}{\partial a_2} = 0 \dots \quad \frac{\partial\, \text{Cost}}{\partial b} = 0$$

$$= a_1^2 E[X_1^2] + 2 a_1 a_2 E[X_1 X_2] + \dots$$

→ System of Linear equations in a_i's & b w/ $E[\cdot]$, $E[X_i^2]$, $E[X_i X_j]$

expectations, covariances, variances,

→ see a textbook.

Compare this to the general estimator that requires the full posterior distrib. $\Theta \mid X_1, X_2, \dots X_n$ to calculate $\hat{\Theta}$! But w/ the Linear LMSE we need to know expectations only! "don't need to know the whole distrib."

Ex: Linear LMS Example: Θ : unknown random parameter,

make multiple measurements $\boxed{X_i = \Theta + W_i}$ ith measurement

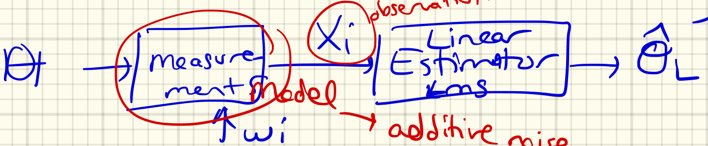
$X_1 = \Theta + W_1$
$X_2 = \Theta + W_2$
$\vdots$
$X_n = \Theta + W_n$

Measurement Noise: $W_i \sim 0, \sigma_i^2$

typically 0 mean w/ known variance.

→ Prior: $\Theta \sim \mu, \sigma_0^2$

Assumption: $\Theta, W_1, \dots, W_n$ are independent.



→ The form of the "optimal" linear estimator has a very neat form:

$$\hat{\Theta}_L = \frac{\mu/\sigma_0^2 + \sum_{i=1}^{n} x_i/\sigma_i^2}{\sum_{i=0}^{n} \frac{1}{\sigma_i^2}}$$

→ sum of all coeffs in numerator

a weighted average of $\underbrace{\mu}_{\text{prior mean}}, \underbrace{X_1, \dots, X_n}_{\text{observations}}$

Weights!! ← pay attention.

$\sigma^2 \approx$ measure of uncertainty

↳ $\frac{1}{\sigma^2} \approx$ reliability.

weights are proportional to inverse uncertainty (≡ reliability) of measurements!

* The Prior mean is treated the same way as the X_i's.

* We did not impose any certain shape of the distributions for W & Θ. We just stated their means & variances.

→ If all r.v.s are normal in this model → "optimal" linear estimator turns out to be equal to the conditional expectation.

For normal r.v.s (Θ, W)

$$\hat{\Theta}_L = E[\Theta \mid X_1, \dots, X_n]$$

"optimal" linear LMSE = "optimal" LMS Estimator (general)

for all Normal r.v.s scenario

→ If your measurement device measures X^2 or X^3 instead of X's : want to estimate θ.

→ Choosing X_i or functions of X_i in the Linear LMS.

$\hat{\theta} = aX + b$ vs $\hat{\theta} = aX^3 + b$ } Form a linear estimator

$$\hat{\theta} = a_1 \underbrace{X}_{X_1} + a_2 \underbrace{X^2}_{X_2} + a_3 \underbrace{X^3}_{X_3} + b + a_4 (\log X)$$

model is still linear.

Data are nonlinear fns of the observations

→ In Linear LMS estimator, which features / functions of (X) observations you want to choose matters. → Depends on your data.

→ For the general LMS estimator: $X, X^2, X^3, \log X$ ⟹ $f(X)$: nonlinear fns of X → carry the same info on X. $\theta | X$ or $\theta | X^2$ are the same.

$E[\theta | X]$ is the same as $E[\theta | X^3]$

∴ Doing nonlinear xformations to your data does not affect the LMS Estimate.

end of

→ Bayesian Estimation methods : Used a prior & Bayes thm to find a posterior distrib.

→ MAP

→ MSE

→ Linear MSE

$\Theta | X$ distrib.

— Standard models eg. $X_i = \Theta + W_i$

— X_i : Uniform $[0, \theta]$; uniform prior on Θ.

— X_i : Bernoulli (p) ; uniform (Beta) prior on p.

— X_i : Gaussian : $\mathcal{N}(\mu, \sigma^2)$: normal prior on Θ

ties to concept of "CONJUGATE PRIOR" ; posterior distrib. has the same functional form as the prior.

→ advanced ML/DL course material,

Classical Estimation :

$$\theta \longrightarrow \boxed{P_X(x;\theta)} \xrightarrow{\text{observation (r.v.)}} \boxed{\text{Estimator}} \longrightarrow \hat{\theta}(X)$$

N : noise

Unknown parameter (but not random) ($\therefore$ no prior distrib assumption !)

measurement process

estimate r.v.

is $\hat{\theta}$ an r.v.? yes

$P_X(x;\theta)$: a family of distributions parameterized by θ.

(eg. a Gaussian w/ mean θ & var = 1.)

not a conditional distribution

Task: Design an estimator $\hat{\theta}$ to keep estimation error small

ie. $\hat{\theta} - \theta$ small

— $\hat{\theta}(X)$: a fn. of an r.v. $\rightarrow$ $\hat{\theta}$ is an r.v.

$\longrightarrow$ Let's check out estimators using $P_X(x;\theta)$

Maximum Likelihood Estimation : (MLE)

- $X \sim p_X(x;\theta)$: a model w/ unknown parameter θ.
- pick θ that "makes the data X we observed most likely to occur".

$$\hat{\theta}_{ML} = \arg\max_{\theta} \boxed{p_X(x;\theta)}$$

model of the measurement process

Recall: Bayesian MAP estimator

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p_{\Theta|X}(\theta|x) = \frac{\boxed{p_{X|\Theta}(x|\theta)} \cdot p_{\Theta}(\theta)}{p_X(x)}$$

Θ was a r.v. $\rightarrow$ we could have a prior p_Θ.

* If the prior is constant (uniform θ) $\rightarrow$ then all θ's are equally likely then ML estimation (classical) takes the same form as the Bayesian MAP estimation.

Ex: Let $X_1, \dots, X_n$ $\boxed{\text{i.i.d}}$ exponential r.v.s, w/ a certain parameter θ.

$X_i \sim \underbrace{\exp(\theta)}_{\text{symbol}}$,

$$X_i \sim \theta . \exp(-\theta . x_i)$$

joint distrib. $\underbrace{p_{x_1,\dots x_n}(x_1,\dots x_n \,;\theta)}$ $= \prod_{i=1}^{n} \theta . e^{-\theta x_i}$

MLE: $\underbrace{\max_{\theta} \prod_{i=1}^{n} \theta . e^{-\theta . x_i}}_{\text{maximize this}}$ $\equiv$ what's the value of θ that makes the observed x's most likely?

$\equiv$ maximizing its log.

take the log

$$\max_{\theta} \left(n \log \theta - \theta \sum_{i=1}^{n} x_i \right) \longrightarrow \frac{\partial(\cdot)}{\partial \theta} = 0 .$$

$$\frac{n}{\theta} - \sum_i x_i = 0$$

$$\Rightarrow \boxed{\hat{\theta}_{ML} = \frac{n}{X_1 + \dots + X_n}}$$

reciprocal of the sample mean: $\dfrac{X_1 + \dots + X_n}{n} \longrightarrow$

Recall: exponential distrib → expected value $= \frac{1}{\theta}$.

∴ $\hat{\theta}_{ML}$ is a reasonable estimate

$$\underbrace{X_1, \dots, X_n} \longrightarrow \boxed{\text{ML estimator}} \longrightarrow \hat{\theta}_n(X_1, \dots, X_n) \text{ : r.v.}$$

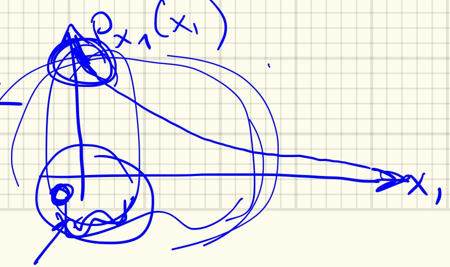
$$\hat{\theta}_{ML} = \frac{n}{\sum_{i=1}^{n} X_i}$$

Desirable Properties of Estimators:

1. Unbiased: $E[\hat{\theta}] = \theta$

- Don't want the estimator to have a $\boxed{\text{systematic error}}$ on the +ve or −ve side of the true parameter θ.

ex:



exponential example w/ $n=1$

$$\hat{\theta}_{ML} = \frac{1}{X_1} \to E\left[\frac{1}{X_1}\right] \to \infty \neq \theta$$

for this example

Biased estimator

Note: ML estimators in general are biased; but can be asymptotically unbiased (ie. as $n \uparrow$ $E[\hat{\theta}_{ML}] \to \theta$)

② Consistent: $\hat{\theta}_n \xrightarrow[\text{in probability}]{} \theta$

prev ex: exponential ex. $X_i \sim \exp(\theta)$.

$$\underbrace{\frac{X_1 + \dots + X_n}{n}}_{\text{sample mean}} \xrightarrow[\text{in prob.}]{\text{WLLN}} \underbrace{E[X]}_{= \frac{\cancel{n} \cdot E[X_i]}{\cancel{n}}} = \frac{1}{\theta}$$

$$\hat{\mu} \xrightarrow{\text{in prob}} \text{true mean } \mu = 1/\theta.$$

$$\hat{\theta}_{ML} = \hat{\theta}_n = \frac{n}{X_1 + \dots + X_n} \xrightarrow{\text{in prob.}} \frac{1}{E[X]} = \theta. \quad \forall \theta$$

$$\hat{\theta} \xrightarrow{\text{in prob}} \theta$$

$\therefore$ MLE is a consistent estimator.

③ "Small" Mean-Squared Error (MSE)

$$E\left[(\hat{\Theta} - \theta)^2\right] = Var(\hat{\theta} - \theta) + E\left[(\hat{\theta} - \theta)\right]^2$$

$\hat{\Theta}$: r.v.

$P_{\hat{\Theta}}$.

for a particular θ value

const.

$Var(Y) = E[Y^2] - \boxed{E(Y)}^2$

$Y = \hat{\theta} - \theta$: Bias

$$MSE = E\left[(\hat{\Theta} - \theta)^2\right] = \underbrace{Var(\hat{\Theta})}_{\text{Uncertainty in the estimate}} + (Bias)^2$$

$Bias = \hat{\theta} - \theta$

we want 0 Bias for an unbiased estimator

Our desire: small $Var(\hat{\theta})$

→ We desire both Small Variance & Small Bias

→ But typically ∃ $\boxed{\text{Bias / Variance Trade-off}}$:

eg. $X \sim \mathcal{N}(\theta, 1)$ → say you have a naive estimator

$\hat{\theta} = 100$, constant output.

For this naive estimator

$\text{Var}(\hat{\theta}) = 0$: smallest 0 variance

→ MSE is huge.

→ $\text{Bias} = \hat{\theta} - \theta$ $\quad \theta^{true}$: 0, 1, 2, ... −1, −2, ... some small value

$(\text{Bias})^2 = (100 - \theta^{true})^2$

$= 10^4 + \ldots$

Huge Bias

Conclusion : You may decrease the variance. → your bias may get huge ↗

→ ∃ a trade-off btw the two !

∴ To come up w/ both Low Bias & Low variance → you have to design more sophisticated estimators ← (advanced class material)

In Classical estimation ; for parameter estimation :

1) MLE ✓

2) Sample Mean Estimator : Get data

$X_1, \ldots, X_n$ i.i.d. , mean θ , variance σ^2.

$X_i = \underbrace{\theta}_{\text{mean}} + W_i$

W_i : i.i.d w/ mean 0, variance σ^2

$\hat{\theta}_n = \dfrac{X_1 + \cdots + X_n}{n}$

sample mean estimator.

Properties of the Sample Mean Estimator :

1) $E[\hat{\theta}_n] = E\left[\dfrac{X_1 + \cdots + X_n}{n}\right] = \theta \rightarrow$ Unbiased : Bias $= 0$ ✓

2) $\hat{\theta}_n \xrightarrow[\text{WLLN}]{} \theta \longrightarrow$ Consistent ✓

3) MSE : $E\left[(\hat{\theta}-\theta)^2\right] = \underbrace{Var(\hat{\theta})}_{\frac{1}{n^2}\cdot n \cdot \sigma^2 = \frac{\sigma^2}{n} \xrightarrow[n\to\infty]{} 0} + \cancel{(bias)^2} \to 0$

$\boxed{\text{exercise}}$ if $X_i \sim N(\theta, \sigma^2)$ i.i.d.

Do MLE

write $p_X(x;\theta) \rightarrow$ maximize w.r.t. θ.

$\rightarrow$ find $\hat{\theta}$ = sample mean.

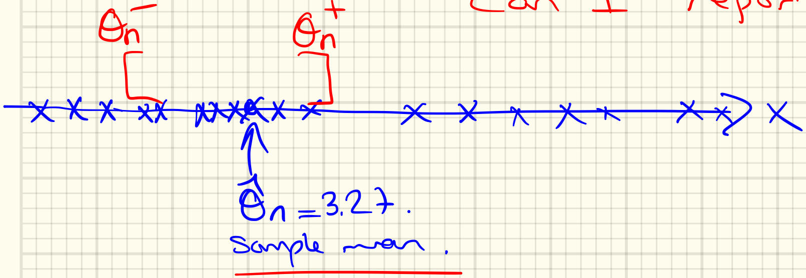
MLE in this case turns out to be the Sample Mean Estimator.

Now:

→ You report your sample mean: $\underbrace{3.27}$ $= \hat{\theta}_n$.

Q. How reliable is that number?

Can I report an interval w/ limits θ_n^- & θ_n^+ where our true parameter may lie?



⇒ CONFIDENCE INTERVALS (CI)

θ (not random) → [ ] $\xrightarrow[X_1, \dots, X_n]{X}$ [Estimator] → $\hat{\theta}$: r.v. : a $\underbrace{\text{point}}$ estimate $\hat{\theta}_n$

→ $[\hat{\theta}_n^-, \hat{\theta}_n^+]$: interval is an r.v.

($\hat{\theta}_n^-$ ↓ r.v., $\hat{\theta}_n^+$ ↓ r.v.)

We pick an α → $\underbrace{(1-\alpha)}$ confidence interval →

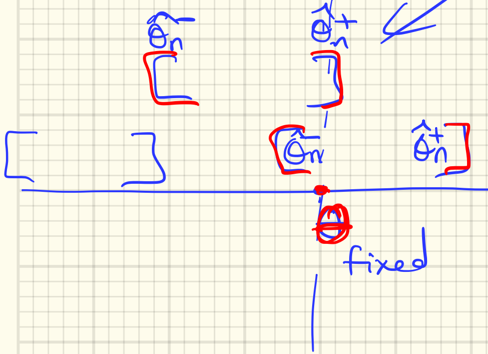
a $(1-\alpha)$ confidence interval $[\hat{\theta}_n^-, \hat{\theta}_n^+]$ s.t

$$P\left(\hat{\theta}_n^- \le \theta \le \hat{\theta}_n^+\right) \ge \underbrace{1-\alpha}_{0.95}, \quad \forall \theta.$$

eg. $\boxed{\alpha = 0.05}$
0.02
0.01

→ 95% confidence Interval Interpretation:

CI is a random interval. We are sampling CI intervals. w/ prob 0.95 (95%), the interval falls on the true value of θ.



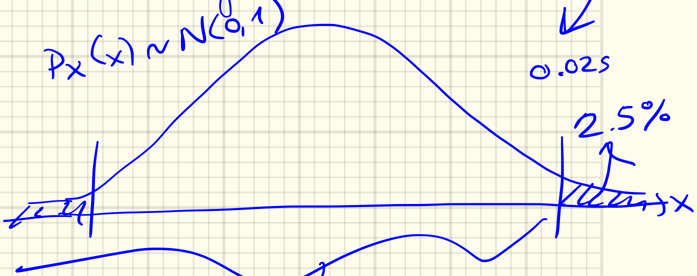
Rather than the statement:
~~w/ prob 95% θ falls in $[\hat{\theta}_n^-, \hat{\theta}_n^+]$~~
b/c θ is not random

Q. How do we construct a 95% or 98% CI?

- CI in estimation of the Mean:

$$\hat{\theta}_n = \frac{X_1 + \cdots + X_n}{n}$$

→ want to calculate $\boxed{95\%}$ confidence interval



From the normal table

$$\Phi(z) = \underline{0.975} = 1 - \frac{0.05}{2}$$

$z = 1.96$

Use CLT; standardize → $\hat{\theta}_n \sim N(0,1)$

$$P\left(\left| \frac{\hat{\theta}_n - \theta}{\sigma/\sqrt{n}} \right| < \underbrace{1.96}_{z} \right) \approx 0.95 \quad \text{(CLT)}$$

rewrite

$$P\left(\underbrace{\hat{\theta}_n - \frac{1.96\sigma}{\sqrt{n}}}_{\hat{\theta}_n^-} \leq \theta \leq \underbrace{\hat{\theta}_n + \frac{1.96\sigma}{\sqrt{n}}}_{\hat{\theta}_n^+} \right) \approx 0.95$$

construct the CI:

$\hat{\theta}_n^-$: lower end of the CI

$\hat{\theta}_n^+$: upper end of the CI.

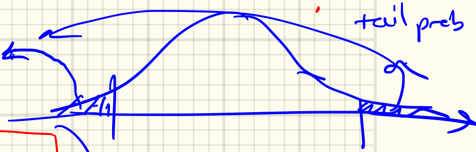
→ 2 observations

(i) as $\underbrace{n \nearrow}_{\text{(more \& more data)}}$; $[\hat{\theta}_n^-, \hat{\theta}_n^+]$: CI ↘ we get more confident that our interval captures the true θ. ✓

(ii) $\sigma \searrow$ (data has lower uncertainty) : CI ↘ ✓

More generally: how to construct the CI?

Let z be s.t. $\Phi(z) = 1 - \frac{\alpha}{2}$



$$P\left(\boxed{\hat{\theta}_n - \frac{z \cdot \sigma}{\sqrt{n}}} \leq \theta \leq \boxed{\hat{\theta}_n + \frac{z\sigma}{\sqrt{n}}}\right) \approx \underbrace{1-\alpha}_{\text{probability}}$$

$$P\left(\hat{\theta}_n^- \leq \theta \leq \hat{\theta}_n^+\right) \approx 1-\alpha$$

Here typically, we know n. ✓ but σ is unknown

For unknown σ; options:

1) Use an upper bound on σ

e.g. $X_i \sim \text{Bernoulli} \rightarrow \underbrace{\sigma \leq \frac{1}{2}}_{\sigma = \frac{1}{2}}$ } CI's are larger than necessary.

$p(1-p)$ $\frac{1}{4}$

2) Estimate σ from the data:

e.g. $X_i \sim \text{Bernoulli}(\theta) \rightarrow \sigma = \sqrt{\theta(1-\theta)}$

$\hat{\theta} \rightarrow \hat{\sigma} \approx \sqrt{\hat{\theta}(1-\hat{\theta})}$: estimate for the standard deviation.

e.g.: sample mean

as $n \uparrow$ $\hat{\theta}$ is a good estimate of θ $\therefore$ $\hat{\sigma}$ is a good estimate of σ.

3) Use a generic estimate of the variance:

Sample Variance: $\sigma^2 = E[(X_i - \theta)^2]$

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \theta)^2 \xrightarrow[\text{WLLN}]{} \sigma^2 \quad \checkmark$$

- We don't know the mean $= \theta$.
- Insert the sample mean estimate $\hat{\theta}_n$:

$$\tilde{S}_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \hat{\theta}_n)^2 \longrightarrow \sigma^2$$

$n \uparrow$ $\hat{\theta}_n \to \theta$

$\tilde{S}_n^2 \to \hat{\sigma}_n^2$

$E[\tilde{S}_n^2] = \sigma^2$: unbiased estimate of the variance.

→ Now use $\sigma = \sqrt{\tilde{S}_n^2} = \tilde{S}_n$ in constructing your Confidence Interval limits.

In constructing the CI's,

2 approximations:

1) We assume the sample mean has a normal distrib.

$$\hat{\theta}_n = \frac{(X_1 + \dots + X_n)}{n}$$ justified by the CLT.

$$\frac{\hat{\theta}_n - \theta}{\sigma/\sqrt{n}} \xrightarrow{CLT} \mathcal{N}(0,1)$$

2) Rather than using the true σ, (we don't know)

we use an approx. of σ (as we just did).