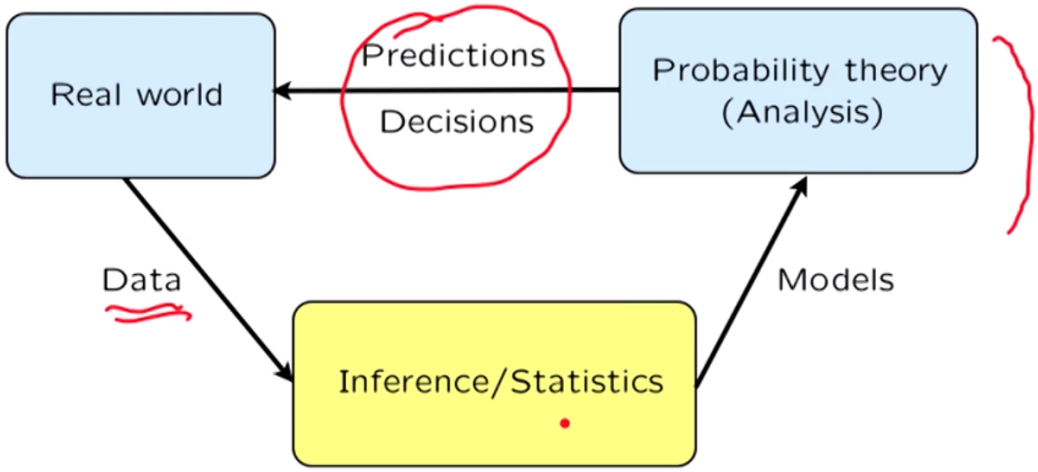
Unit 7 Bayesian Inference

Inference: the big picture

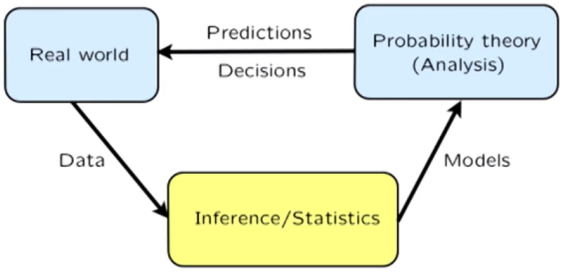


Use data to infer a model that we do analysis on

Model building versus inferring unobserved variables

$X = aS + W$

- Model building:
  - know "signal" S, observe X
  - infer a



- Variable estimation:
  - know a, observe X
  - infer S

W is noise
Signal S
Received X
Finding out a or S is mathematically the same

Hypothesis testing versus estimation

- Hypothesis testing:
  - unknown takes one of few possible values
  - aim at small probability of incorrect decision

  Is it an airplane or a bird?

- Estimation:
  - numerical unknown(s)
  - aim at an estimate that is "close" to the true but unknown value

Exercise: Hypothesis testing versus estimation

4/4 points (graded)

For each one of the following situations, state whether it corresponds to a hypothesis testing or estimation problem.

A grocery store was robbed yesterday morning. The police have determined that the robber was one of the five customers who visited a nearby bank earlier that morning. For those customers, the police know their identity as well as the time that they visited the bank. The police want to:

(a) Guess the time at which the grocery store was robbed.

Estimation ✔

(b) Guess the identity of the robber.

Hypothesis testing ✔

(c) Guess the gender of the robber.

Hypothesis testing ✔

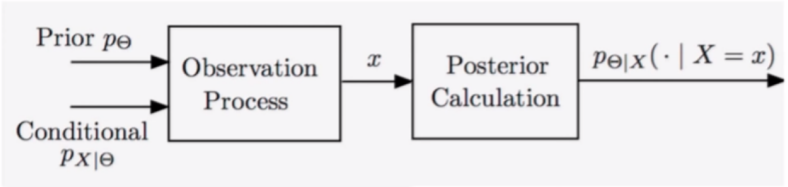
(d) Guess the weight of the robber.

Estimation ✔

The Bayesian inference framework

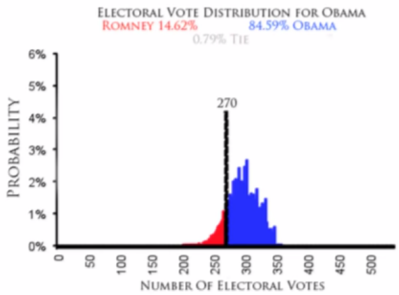
- Unknown Θ
  - treated as a random variable
  - prior distribution p_Θ or f_Θ
- Observation X
  - observation model $p_{X|\Theta}$ or $f_{X|\Theta}$
- Use appropriate version of the Bayes rule to find $p_{\Theta|X}(\cdot \mid X = x)$ or $f_{\Theta|X}(\cdot \mid X = x)$

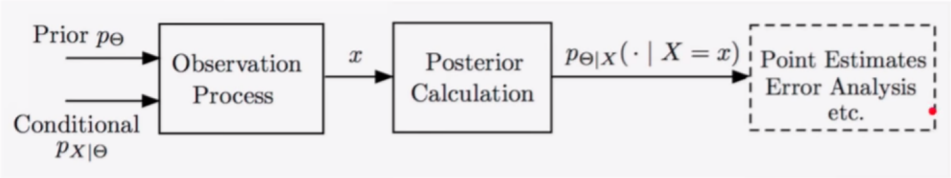
- Where does the prior come from?
  - symmetry
  - known range
  - earlier studies
  - subjective or arbitrary



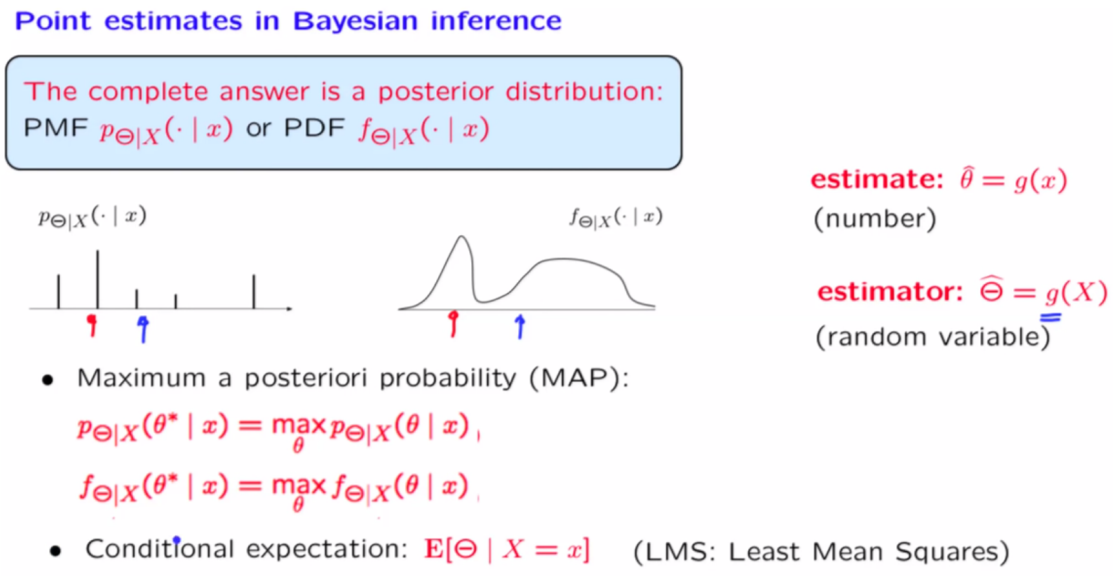
The output of Bayesian inference

The complete answer is a posterior distribution:
PMF $p_{\Theta|X}(\cdot \mid x)$ or PDF $f_{\Theta|X}(\cdot \mid x)$





PMF is full answer of the distribution but may want a single number to represent this



The estimator is a rule that we apply to the data and is a random variable
The estimate is a number and can be a specific value of the estimator

Exercise: Estimates and estimators

3/3 points (graded)

Valerie wants to find an estimator for an unknown random variable Θ. She can observe a random variable X whose distribution satisfies $\mathbf{E}\left[X^2 \mid \Theta\right] = \Theta$. She goes ahead and observes that X took a numerical value of 5. She then estimates Θ as the square of the observed value, namely, 25.

For each of the following questions, choose the most appropriate answer.

1) X^2 is an

Estimator ✔ **Answer:** Estimator

2) 25 is an

Estimate ✔ **Answer:** Estimate

3) $X^3 + 2$ is another (not necessarily good)

Estimator ✔ **Answer:** Estimator

Solution:

In the first and the third cases, we have a random variable $g\left(X\right)$, which is determined as a function of the observation X. Such a random variable is called an estimator.

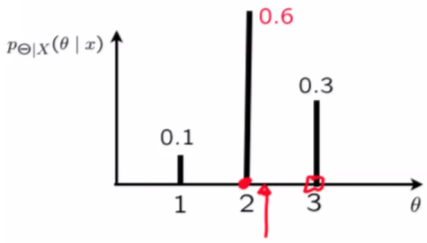
In the second case, we are dealing with the realized numerical value of an estimator, which we call an estimate.

Discrete Θ, discrete X

- values of Θ: alternative hypotheses

$$p_{\Theta|X}(\theta \mid x) = \frac{p_\Theta(\theta)\, p_{X|\Theta}(x \mid \theta)}{p_X(x)}$$

$$p_X(x) = \sum_{\theta'} p_\Theta(\theta')\, p_{X|\Theta}(x \mid \theta')$$



- conditional prob of error:

$P(\hat{\theta} \neq \Theta \mid X = x) = 0.4$

smallest under the MAP rule

- overall probability of error:

$$P(\widehat{\Theta} \neq \Theta) = \sum_x P(\widehat{\Theta} \neq \Theta \mid X = x)\, p_X(x)$$

$$= \sum_\theta P(\widehat{\Theta} \neq \Theta \mid \Theta = \theta)\, p_\Theta(\theta)$$

- MAP rule: $\hat{\theta} = 2$

LMS: $\hat{\Theta} = E[\Theta \mid X = x] = 2.2$

Exercise: Discrete unknowns

2/5 points (graded)

Let Θ_1 and Θ_2 be some unobserved Bernoulli random variables and let X be an observation. Conditional on $X = x$, the posterior joint PMF of Θ_1 and Θ_2 is given by

$$p_{\Theta_1,\Theta_2|X}(\theta_1, \theta_2 \mid x) = \begin{cases} 0.26, & \text{if } \theta_1 = 0, \theta_2 = 0, \\ 0.26, & \text{if } \theta_1 = 0, \theta_2 = 1, \\ 0.21, & \text{if } \theta_1 = 1, \theta_2 = 0, \\ 0.27, & \text{if } \theta_1 = 1, \theta_2 = 1, \\ 0, & \text{otherwise.} \end{cases}$$

We can view this as a hypothesis testing problem where we choose between four alternative hypotheses: the four possible values of (Θ_1, Θ_2).

a) What is the estimate of (Θ_1, Θ_2) provided by the MAP rule?

(1,1) ✔ **Answer:** (1,1)

b) Once you calculate the estimate $(\hat{\theta}_1, \hat{\theta}_2)$ of (Θ_1, Θ_2), you may report the first component, $\hat{\theta}_1$, as your estimate of Θ_1. With this procedure, your estimate of Θ_1 will be

1 ✔ **Answer:** 1

c) What is the probability that Θ_1 is estimated incorrectly (the probability of error) when you use the procedure in part (b)?

0.73 ✖ **Answer:** 0.52

d) What is the MAP estimate of Θ_1 based on X, that is, the one that maximizes $p_{\Theta_1|X}(\theta_1 \mid x)$?

1 ✖ **Answer:** 0

e) The moral of this example is that an estimate of Θ_1 obtained by identifying the maximum of the joint PMF of all unknown random variables is

always the same as ✖ **Answer:** can be different from

the MAP estimate of Θ_1.

Solution:

a) The posterior is largest when $(\theta_1, \theta_2) = (1, 1)$.

b) The corresponding estimate of Θ_1 is the first component of $(1, 1)$, which is 1.

c) The probability of error is the posterior probability that $\Theta_1 = 0$, which is $0.26 + 0.26 = 0.52$.

d) The posterior PMF of Θ_1 is the marginal (posterior) PMF obtained from the joint posterior PMF:

$$\begin{aligned} p_{\Theta_1|X}(0 \mid x) &= 0.26 + 0.26 = 0.52, \\ p_{\Theta_1|X}(1 \mid x) &= 0.21 + 0.27 = 0.48. \end{aligned}$$

Hence, the MAP estimate is $\hat{\theta}_1 = 0$.

e) These can be different, as illustrated by parts (b) and (d).

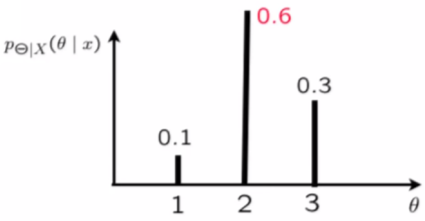
Discrete Θ, continuous X

- Standard example:
  - send signal $\Theta \in \{1, 2, 3\}$

    $X = \Theta + W$

    $W \sim N(0, \sigma^2)$, indep. of Θ

    $f_{X|\Theta}(x \mid \theta) = f_W(x - \theta)$



- MAP rule: $\hat{\theta} =$ 2

$$p_{\Theta|X}(\theta \mid x) = \frac{p_\Theta(\theta)\, f_{X|\Theta}(x \mid \theta)}{f_X(x)}$$

$$f_X(x) = \sum_{\theta'} p_\Theta(\theta') f_{X|\Theta}(x \mid \theta')$$

- conditional prob of error:

  $\mathbf{P}(\hat{\theta} \neq \Theta \mid X = x)$

  smallest under the MAP rule

- overall probability of error:

$$\mathbf{P}(\widehat{\Theta} \neq \Theta) = \int \mathbf{P}(\widehat{\Theta} \neq \Theta \mid X = x) f_X(x)\, dx$$

$$= \sum_\theta \mathbf{P}(\widehat{\Theta} \neq \theta \mid \Theta = \theta)\, p_\Theta(\theta)$$

The MAP rule gives the smallest possible answer for the error
The PDF of W above is being modelled as if it is X shifted by some amount theta (where X is received signal, theta is sent signal and W is noise)

Exercise: Discrete unknown and continuous observation

1/2 points (graded)

Similar to the last example, suppose that $X = \Theta + W$, where Θ is equally likely to take the values -1 and 1, and where W is standard normal noise, independent of Θ. We use the estimator $\widehat{\Theta}$, with $\widehat{\Theta} = 1$ if $X > 0$ and $\widehat{\Theta} = -1$ otherwise. (This is actually the MAP estimator for this problem.)

a) Let us assume that the true value of Θ is 1. In this case, our estimator makes an error if and only if W has a low (negative) value. The conditional probability of error given the true value of Θ is 1, that is, $\mathbf{P}(\widehat{\Theta} \neq 1 \mid \Theta = 1)$, is equal to

- $\Phi(-1)$ ✔
- $\Phi(0)$
- $\Phi(1)$

✖

where Φ is the standard normal CDF.

b) For this problem, the overall probability of error is easiest found using the formula

- $\mathbf{P}(\widehat{\Theta} \neq \Theta) = \int \mathbf{P}(\widehat{\Theta} \neq \Theta \mid X = x) f_X(x)\, dx$
- $\mathbf{P}(\widehat{\Theta} \neq \Theta) = \sum_\theta \mathbf{P}(\widehat{\Theta} \neq \theta \mid \Theta = \theta)\, p_\Theta(\theta)$

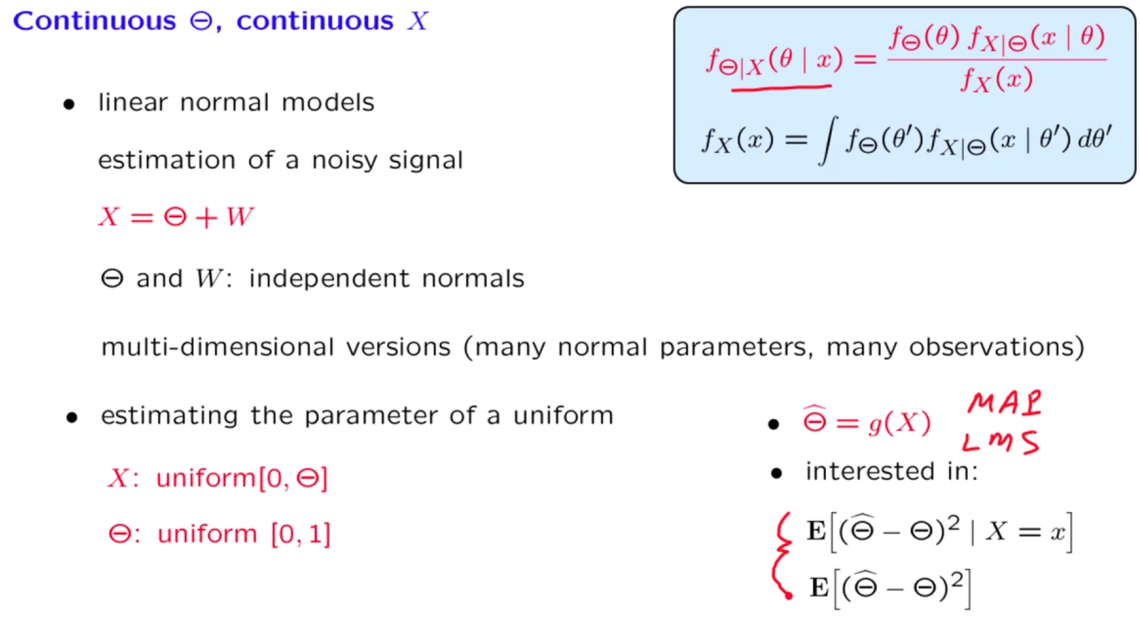
✔

Solution:

a) We have

$$\begin{aligned}\mathbf{P}(\widehat{\Theta} \neq 1 \mid \Theta = 1) &= \mathbf{P}(\Theta + W \leq 0 \mid \Theta = 1) = \mathbf{P}(1 + W \leq 0 \mid \Theta = 1) \\ &= \mathbf{P}(1 + W \leq 0) = \mathbf{P}(W \leq -1) = \Phi(-1)\,.\end{aligned}$$

b) Similar to part (a), $\mathbf{P}(\widehat{\Theta} \neq \theta \mid \Theta = \theta)$ is easy to calculate for either choice of $\theta = -1$ or $\theta = 1$. For this reason, the second formula is easy to implement.



Inferring the unknown bias of a coin and the Beta distribution

- Standard example:
  - coin with bias Θ; prior $f_\Theta(\cdot)$
  - fix n; K =number of heads
- Assume $f_\Theta(\cdot)$ is uniform in $[0,1]$

$$f_{\Theta|K}(\theta \mid k) = \frac{f_\Theta(\theta)\, p_{K|\Theta}(k \mid \theta)}{p_K(k)}$$

$$p_K(k) = \int f_\Theta(\theta') p_{K|\Theta}(k \mid \theta')\, d\theta'$$

$$f_{\Theta|K}(\theta \mid k) = \frac{1 \cdot \binom{n}{k} \theta^k (1-\theta)^{n-k}}{p_K(k)} \qquad \theta \in [0,1]$$

$$= \frac{1}{d(n,k)} \theta^k (1-\theta)^{n-k}$$ "Beta distribution, with parameters $(k+1, n-k+1)$"

- If prior is Beta: $f_\Theta(\theta) = \frac{1}{c}\theta^\alpha (1-\theta)^\beta$ $\qquad \alpha, \beta \geq 0$

$$f_{\Theta|K}(\theta \mid k) = \frac{1}{c} \theta^\alpha (1-\theta)^\beta \binom{n}{k} \theta^k (1-\theta)^{n-k} / p_K(k) \propto \theta^{\alpha+k} (1-\theta)^{\beta+n-k}$$

Split the formula into parts that are and aren't dependent on theta
Beta distribution is dependent on theta^x and (1-theta)^y
If the prior is of a certain form then the posterior will be of a similar form

Exercise: The posterior of a coin's bias

3/3 points (graded)

Let Θ be a continuous random variable that represents the unknown bias (i.e., the probability of Heads) of a coin.

a) The prior PDF f_Θ for the bias of a coin is of the form

$$f_\Theta(\theta) = a\theta^9(1-\theta), \quad \text{for } \theta \in [0,1],$$

where a is a normalizing constant. This indicates a prior belief that the bias Θ of the coin is

High ✔ **Answer:** High

b) We flip the coin 10 times independently and observe 1 Heads and 9 Tails. The posterior PDF of Θ will be of the form $c\theta^m(1-\theta)^n$, where c is a normalizing constant and where

$m =$ 10 ✔ **Answer:** 10

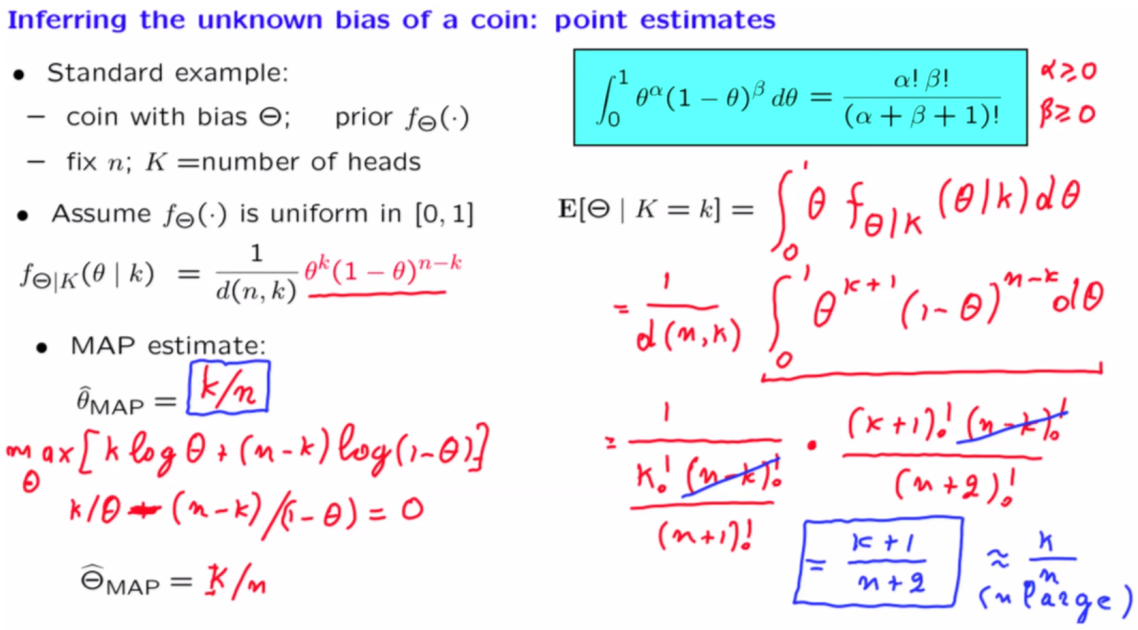
$n =$ 10 ✔ **Answer:** 10

Solution:

a) Because of the high exponent, the term θ^9 is very small when θ is small. This prior, as can also be seen by plotting it, is concentrated on high values of θ and indicates a prior belief in favor of large values.

b) As we saw in the last video, the power to which θ (respectively, $1-\theta$) is raised needs to be incremented by the number of Heads (respectively, Tails) observed, leading to $m = 9+1 = 10$ and $n = 1+9 = 10$. Notice that the resulting posterior is symmetric around 0.5.

This exercise indicates that the strength of the "evidence" incorporated in a prior with $\alpha = 9$ and $\beta = 1$ is exactly counterbalanced by observing 1 Heads and 9 Tails. Differently said, a prior with $\alpha = 9$ and $\beta = 1$ can be thought of as equivalent to prior "evidence" based on 9 Heads and 1 Tails.



Exercise: Moments of the Beta distribution

1/2 points (graded)

Suppose that Θ takes values in $[0, 1]$ and its PDF is of the form

$$f_\Theta(\theta) = a\theta(1-\theta)^2, \quad \text{for } \theta \in [0,1],$$

where a is a normalizing constant.

Use the formula

$$\int_0^1 \theta^\alpha (1-\theta)^\beta \, d\theta = \frac{\alpha!\,\beta!}{(\alpha+\beta+1)!}$$

to find the following:

a) $a =$ 12 ✔ **Answer:** 12

b) $\mathbf{E}[\Theta^2] =$ 1.4548 ✖ **Answer:** 0.2

Solution:

a) Let $\mathrm{I}(\alpha, \beta)$ be the integral in the formula given in the problem statement. The normalizing constant must be equal to $1/\mathrm{I}(1,2)$: this is needed for the PDF to integrate to 1. We have $\mathrm{I}(1,2) = 2!/4! = 1/12$, so that $a = 12$.

b)

$$\mathbf{E}[\Theta^2] = \int_0^1 \theta^2 f_\Theta(\theta)\, d\theta = \int_0^1 a\theta^3(1-\theta)^2\, d\theta = a \cdot \mathrm{I}(3,2) = 12 \cdot \frac{3!\,2!}{6!} = \frac{1}{5}.$$

It common notation for Beta distr. CDF , namely for $\int_0^1 \theta^\alpha (1-\theta)^\beta \, d\theta = \frac{\alpha!\,\beta!}{(\alpha+\beta+1)!}$, so with $\mathrm{I}(\alpha, \beta)$ we designate $\frac{\alpha!\,\beta!}{(\alpha+\beta+1)!}$. And for given $f_\Theta(\theta) = a\theta(1-\theta)^2$ we have $I(1,2)$ as mnemonic for it's CDF value. (Note, $\theta(1-\theta)^2$ is $\theta^1(1-\theta)^2$, $hence I(1,2)$).

Summary

- Problem data: $p_\Theta(\cdot)$, $p_{X|\Theta}(\cdot \mid \cdot)$
- Given the value x of X: find, e.g., $p_{\Theta|X}(\cdot \mid x)$
  - using appropriate version of the Bayes rule (4 choices)
- Estimator $\widehat{\Theta} = g(X)$ Estimate $\hat{\theta} = g(x)$
  - MAP: $\hat{\theta}_{\text{MAP}} = g_{\text{MAP}}(x)$ maximizes $p_{\Theta|X}(\theta \mid x)$
  - LMS: $\hat{\theta}_{\text{LMS}} = g_{\text{LMS}}(x) = \mathbf{E}[\Theta \mid X = x]$
- Performance evaluation of an estimator $\widehat{\Theta}$
  - $\mathbf{P}(\widehat{\Theta} \neq \Theta \mid X = x)$
  - $\mathbf{P}(\widehat{\Theta} \neq \Theta)$
  - $\mathbf{E}\big[(\widehat{\Theta} - \Theta)^2 \mid X = x\big]$
  - $\mathbf{E}\big[(\widehat{\Theta} - \Theta)^2\big]$ total prob/exp thm.

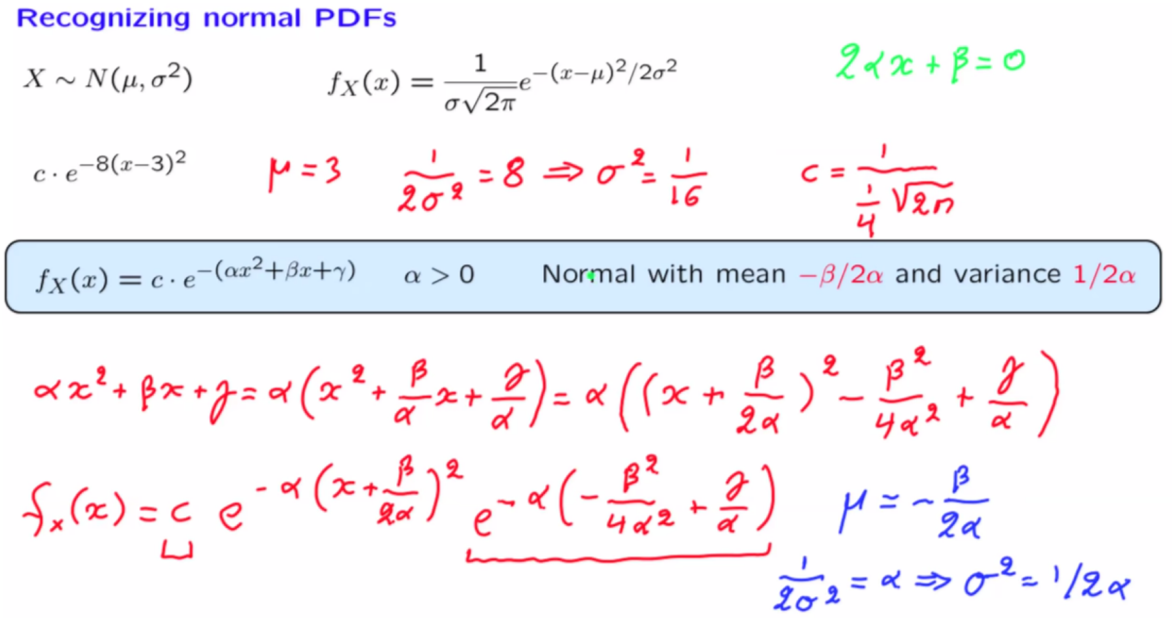
Given a prior and some observed X
Want to get the posterior distribution as a full answer to inference problem
This is the estimator
Applying a given value of X = x, we can then get an estimate from the estimator

Linear Models with normal noise
Most commonly used model as it is a good approximation to many scenarios

Lecture 15: Linear models with normal noise

$X_i = \sum_{j=1}^{m} aij\ \Theta_j + W_i$ W_i, Θ_j : independent, normal

- **Very common and convenient model**
- **Bayes' rule: normal posteriors**
- **MAP and LMS estimates coincide**
  - Simple formulas (linear in the observations)
- **Many nice properties**
- **Trajectory estimation example**



alpha has to be positive because a PDF must integrate to 1 meaning the exponential has to die out as x goes to infinity therefore alpha has to be positive
Used completing the square method to get the exponential in the form (x-something)
In green: differentiate the exponential then know that the mean is taken when the differential is = 0 (at its peak)

Exercise: Recognizing normal PDFs

2/2 points (graded)

The random variable X has a PDF of the form

$$f_X(x) = ce^{-4x^2-24x+30},$$

where c is a normalizing constant. Then,

a) $\mathbf{E}[X] =$ -3 ✔ **Answer:** -3

b) $\text{Var}(X) =$ 1/8 ✔ **Answer:** 0.125

Solution:

a) We recognize this as a normal PDF. The mean is at the peak of the PDF, which is found by setting the derivative of the exponent to zero: $-8x - 24 = 0$, or $x = -3$.

b) The variance is $1/(2\alpha)$, where α is the positive coefficient associated with the term x^2. Thus, the variance is $1/8$.

Estimating a normal random variable in the presence of additive normal noise

$X = \Theta + W \qquad \Theta, W : N(0,1), \quad \text{independent}$

$$f_{\Theta|X}(\theta \mid x) = \frac{f_\Theta(\theta)\, f_{X|\Theta}(x \mid \theta)}{f_X(x)}$$

$$f_X(x) = \int f_\Theta(\theta) f_{X|\Theta}(x \mid \theta)\, d\theta$$

$f_{X|\Theta}(x \mid \theta)$: $X = \theta + W \qquad N(\theta, 1)$

$f_{\Theta|X}(\theta \mid x) = \frac{1}{f_X(x)} \; c\, e^{-\frac{1}{2}\theta^2} \; c\, e^{-\frac{1}{2}(x-\theta)^2} = c(x)\, e^{-\text{quadratic}(\theta)}$

Normal!

Fix $x \qquad \min_\theta \left[\frac{1}{2}\theta^2 + \frac{1}{2}(x-\theta)^2 \right] \qquad \theta + (\theta - x) = 0$

$\hat{\theta}_{\text{MAP}} = \hat{\theta}_{\text{LMS}} = \mathbf{E}[\Theta \mid X = x] = x/2 \qquad \widehat{\Theta}_{\text{MAP}} = \mathbf{E}[\Theta \mid X] = X/2.$

X when X|theta is theta (number) + W (random variable)
Look at min of exponent terms

Estimating a normal parameter in the presence of additive normal noise

$X = \Theta + W \qquad \Theta, W : N(0,1), \quad \text{independent}$

$$f_{\Theta|X}(\theta \mid x) = \frac{f_\Theta(\theta)\, f_{X|\Theta}(x \mid \theta)}{f_X(x)}$$

$$f_X(x) = \int f_\Theta(\theta) f_{X|\Theta}(x \mid \theta)\, d\theta$$

$$\widehat{\Theta}_{\text{MAP}} = \widehat{\Theta}_{\text{LMS}} = \mathbf{E}[\Theta \mid X] = \frac{X}{2}$$

- Even with general means and variances:
  - posterior is normal
  - LMS and MAP estimators coincide
  - these estimators are "linear," of the form $\widehat{\Theta} = aX + b$

Exercise: Normal unknown and additive noise

1/4 points (graded)

As in the last video, let $X = \Theta + W$, where Θ and W are independent normal random variables and W has mean zero.

a) Assume that W has positive variance. Are X and W independent?

No ✔ **Answer:** No

b) Find the MAP estimator of Θ based on X if $\Theta \sim N(1, 1)$ and $W \sim N(0, 1)$, and evaluate the corresponding estimate if $X = 2$.

$\hat{\theta} =$ 0 ✖ **Answer:** 1.5

c) Find the MAP estimator of Θ based on X if $\Theta \sim N(0, 1)$ and $W \sim N(0, 4)$, and evaluate the corresponding estimate if $X = 2$.

$\hat{\theta} =$ 0 ✖ **Answer:** 0.4

d) For this part of the problem, suppose instead that $X = 2\Theta + 3W$, where Θ and W are standard normal random variables. Find the MAP estimator of Θ based on X under this model and evaluate the corresponding estimate if $X = 2$.

$\hat{\theta} =$ 2 ✖ **Answer:** 0.30769

Solution:

a) They are not independent. This is intuitively clear because W has an effect on X. Another way to see it is that we have (by independence of Θ and W) that $\mathbf{E}[\Theta W] = \mathbf{E}[\Theta]\ \mathbf{E}[W] = 0$, which leads to

$$\mathbf{E}[XW] = \mathbf{E}[(\Theta + W)W] = \mathbf{E}[W^2] \neq 0 = \mathbf{E}[X]\ \mathbf{E}[W],$$

which in turn implies that X and W are not independent.

b) If we focus on the terms that involve θ, the posterior is of the form

$$c(x)\, e^{-(\theta-1)^2/2} e^{-(x-\theta)^2/2}.$$

To find the MAP estimate, we set the derivative with respect to θ of the exponent to zero, so that $(\hat{\theta} - 1) + (\hat{\theta} - x) = 0$, or $\hat{\theta} = (1 + x)/2$, which, when $x = 2$, evaluates to $3/2$.

c) If we focus on the terms that involve θ, the posterior is of the form

$$c(x)\, e^{-\theta^2/2} e^{-(x-\theta)^2/(2\cdot 4)}.$$

To find the MAP estimate, we set the derivative with respect to θ of the exponent to zero, so that $\hat{\theta} + (\hat{\theta} - x)/4 = 0$, or $\hat{\theta} = x/5$, which, when $x = 2$, evaluates to $2/5$.

d) Note that conditional on $\Theta = \theta$, the random variable X is normal with mean 2θ and variance 9. If we focus on the terms that involve θ, the posterior is of the form

$$c(x)\, e^{-\theta^2/2} e^{-(x-2\theta)^2/(2\cdot 9)}.$$

To find the MAP estimate, we set the derivative with respect to θ of the exponent to zero, so that $\hat{\theta} + 2(2\hat{\theta} - x)/9 = 0$, or $\hat{\theta} = 2x/13$, which, when $x = 2$, evaluates to $4/13$.

The case of multiple observations

$$X_1 = \Theta + W_1 \qquad \Theta \sim N(x_0, \sigma_0^2) \qquad W_i \sim N(0, \sigma_i^2)$$
$$\vdots$$
$$X_n = \Theta + W_n \qquad \Theta, W_1, \ldots, W_n \text{ independent}$$

$$f_{\Theta|X}(\theta \mid x) = \frac{f_\Theta(\theta)\, f_{X|\Theta}(x \mid \theta)}{f_X(x)}$$

$$f_X(x) = \int f_\Theta(\theta) f_{X|\Theta}(x \mid \theta)\, d\theta$$

$f_{X_i|\Theta}(x_i \mid \theta) = c_i\, e^{-(x_i-\theta)^2/2\sigma_i^2}$

given $\Theta=\theta$: $X_i = \theta + W_i \sim N(\theta, \sigma_i^2)$

$f_{X|\Theta}(x \mid \theta) = f_{X_1,\ldots,X_n|\Theta}(x_1,\ldots,x_n \mid \theta) = \prod_{i=1}^{n} f_{X_i|\Theta}(x_i \mid \theta)$

given $\Theta=\theta$: W_i independent $\Rightarrow$ X_i independent

$f_{\Theta|X}(\theta \mid x) = \frac{1}{f_X(x)} \cdot c_0\, e^{-(\theta-x_0)^2/2\sigma_0^2} \prod_{i=1}^{n} c_i\, e^{-(x_i-\theta)^2/2\sigma_i^2}$ Normal!

X and x are representing a vector of multiple observations here

The case of multiple observations

$$f_{\Theta|X}(\theta \mid x) = c \cdot \exp\Big\{-\text{quad}(\theta)\Big\} \qquad \text{quad}(\theta) = \frac{(\theta - x_0)^2}{2\sigma_0^2} + \frac{(\theta - x_1)^2}{2\sigma_1^2} + \cdots + \frac{(\theta - x_n)^2}{2\sigma_n^2}$$

find peak

$$\frac{d}{d\theta}\,\text{quad}(\theta) = 0: \quad \sum_{i=0}^{n} \frac{(\theta - x_i)}{\sigma_i^2} = 0 \Rightarrow \theta \sum_{i=0}^{n} \frac{1}{\sigma_i^2} = \sum_{i=0}^{n} \frac{x_i}{\sigma_i^2}$$

$$\hat{\theta}_{\text{MAP}} = \hat{\theta}_{\text{LMS}} = \mathbf{E}[\Theta \mid X = x] = \frac{\sum_{i=0}^{n} \frac{x_i}{\sigma_i^2}}{\sum_{i=0}^{n} \frac{1}{\sigma_i^2}}$$

Take derivative of the quadratic sum which gives a sum of some terms

The case of multiple observations

- Key conclusions:
  - posterior is normal
  - LMS and MAP estimates coincide
  - these estimates are "linear," of the form $\hat{\theta} = a_0 + a_1 x_1 + \cdots + a_n x_n$
- Interpretations:
  - estimate $\hat{\theta}$: weighted average of x_0 (prior mean) and x_i (observations)
  - weights determined by variances

$$\hat{\theta}_{\text{MAP}} = \hat{\theta}_{\text{LMS}} = \mathbf{E}[\Theta \mid X = x] = \frac{\sum_{i=0}^{n} \frac{x_i}{\sigma_i^2}}{\sum_{i=0}^{n} \frac{1}{\sigma_i^2}}$$

σ_i^2 large
x_i very noisy
⇒ small weight

The prior mean x0 is being treated just like it's another observation
Each xi is weighted by the variance associated with it

7. Exercise: Multiple observations

Bookmark this page

Exercise: Multiple observations

1/2 points (graded)

Consider a model involving multiple observations of the form $X_i = c_i\Theta + W_i$, $i = 1, 2, \ldots, n$, where $\Theta, W_1, \ldots, W_n$ are independent (not necessarily normal) random variables and the c_i's are known nonzero constants. Assume that Θ has positive variance.

a) Are the random variables X_i, $i = 1, 2, \ldots, n$, independent?

Yes ✖ **Answer:** No

b) Are the random variables X_i, $i = 1, 2, \ldots, n$, conditionally independent given Θ?

Yes ✔ **Answer:** Yes

Solution:

a) The X_i's are dependent because they are all affected by Θ. For a mathematical derivation, you can consider the zero mean case and check that $\mathbf{E}[X_1X_2] = c_1c_2\mathbf{E}[\Theta^2] \neq 0$, whereas $\mathbf{E}[X_1]\,\mathbf{E}[X_2] = 0$.

b) If we are given that $\Theta = \theta$, then $X_i = c_i\theta + W_i$. In the conditional universe, θ is now a number. Furthermore, the W_i's are independent. Thus, the X_i's (which are equal to W_i plus a number) are also (conditionally) independent.

Different when it's an r.v. dependence compared to a number dependence

8. Exercise: Multiple observations, more general model

Bookmark this page

Exercise: Multiple observations, more general model

0/1 point (graded)

Suppose that $X_1 = \Theta + W_1$ and $X_2 = 2\Theta + W_2$, where Θ, W_1, W_2 are independent standard normal random variables. If the values that we observe happen to be $X_1 = -1$ and $X_2 = 1$, then the MAP estimate of Θ is

-0.75 ✖ **Answer:** 0.16667

Solution:

The numerator term of the posterior is equal to a constant times

$$e^{-\theta^2/2}e^{-(x_1-\theta)^2/2}e^{-(x_2-2\theta)^2/2}.$$

To find the MAP estimate, we set x_1 and x_2 to the given values, and set the derivative of the exponent (with respect to θ) to zero. We obtain

$$\theta + (\theta + 1) + 2(2\theta - 1) = 0,$$

which yields $6\theta - 1 = 0$ or $\theta = 1/6$.

The mean squared error

$$f_{\Theta|X}(\theta \mid x) = c \cdot \exp\{-\text{quad}(\theta)\}$$

$$\text{quad}(\theta) = \frac{(\theta - x_0)^2}{2\sigma_0^2} + \frac{(\theta - x_1)^2}{2\sigma_1^2} + \cdots + \frac{(\theta - x_n)^2}{2\sigma_n^2}$$

$X_i = \Theta + W_i$

$$\hat{\theta} = \frac{\sum_{i=0}^{n} \frac{x_i}{\sigma_i^2}}{\sum_{i=0}^{n} \frac{1}{\sigma_i^2}}$$

- Performance measures:

$$\mathbf{E}\left[(\Theta - \widehat{\Theta})^2 \mid X = x\right] = \mathbf{E}\left[(\Theta - \hat{\theta})^2 \mid X = x\right] = \text{var}(\Theta \mid X = x) = 1 \Big/ \sum_{i=0}^{n} \frac{1}{\sigma_i^2}$$

$$\mathbf{E}\left[(\Theta - \widehat{\Theta})^2\right] = \int E\left[(\Theta - \hat{\Theta})^2 \mid X = x\right] f_X(x)\, dx$$

$f_X(x) = c \cdot e^{-(\alpha x^2 + \beta x + \gamma)}$ $\quad \alpha > 0$ Normal with mean $-\beta/2\alpha$ and variance $1/2\alpha$

$\alpha = \frac{1}{2\sigma_0^2} + \cdots + \frac{1}{2\sigma_n^2}$

some σ_i^2 small → MSE small

all σ_i^2 large → MSE large.

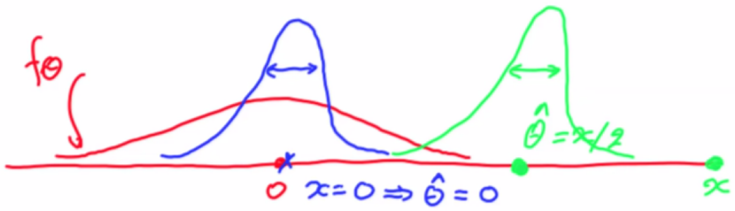
So the expected value having not observed anything is the same as if we had observed something (X=x)
Also when the variance is small the MSE is small
when it is large the MSE is large

The mean squared error

$$\mathbf{E}\left[(\Theta - \widehat{\Theta})^2 \mid X = x\right] = \mathbf{E}\left[(\Theta - \widehat{\Theta})^2\right] = 1 \Big/ \sum_{i=0}^{n} \frac{1}{\sigma_i^2}$$

$$\hat{\theta} = \frac{\sum_{i=0}^{n} \frac{x_i}{\sigma_i^2}}{\sum_{i=0}^{n} \frac{1}{\sigma_i^2}}$$

- Example: $\sigma_0^2 = \sigma_1^2 = \cdots = \sigma_n^2 = \sigma^2$ $\quad \frac{1}{(n+1)\frac{1}{\sigma^2}} = \frac{\sigma^2}{n+1}$
- conditional mean squared error same for all x
- Example: $X = \Theta + W$ $\quad \Theta \sim N(0,1)$, $\quad W \sim N(0,1)$
  independent Θ, W $\quad \widehat{\Theta} = X/2$ $\quad \mathbf{E}\left[(\Theta - \widehat{\Theta})^2 \mid X = x\right] = 1/2$



i.e. the more n's (observations), the smaller the MSE becomes, so it becomes more accurate
conditional MSE same for all x means no observation is any more valuable than another
In red, have a large variation i.e. more uncertainty about theta but once we observe x=0 then the variance decreases and we are more confident about theta being around 0

10. Exercise: The mean-squared error

Bookmark this page

Exercise: The mean-squared error

0/1 point (graded)

In this exercise we want to understand a little better the formula

$$\frac{1}{\sum_{i=0}^{n} \frac{1}{\sigma_i^2}}$$

for the mean squared error by considering two alternative scenarios.

In the first scenario, $\Theta \sim N(0,1)$ and we observe $X = \Theta + W$, where $W \sim N(0,1)$ is independent of Θ.

In the second scenario, the prior information on Θ is extremely inaccurate: $\Theta \sim N(0, \sigma_0^2)$, where σ_0^2 is so large that it can be treated as infinite. But in this second scenario we obtain two observations of the form $X_i = \Theta + W_i$, where the W_i are standard normals, independent of each other and of Θ.

The mean squared error is

- smaller in the first scenario.
- smaller in the second scenario.
- the same in both scenarios. ✔

✖

Solution:

We use the formula for the mean squared error. For the second scenario, we set $\sigma_0^2 = \infty$. In the first scenario, we obtain

$$\frac{1}{\frac{1}{1} + \frac{1}{1}} = \frac{1}{2},$$

and in the second scenario, we obtain the same mean squared error:

$$\frac{1}{\frac{1}{\infty} + \frac{1}{1} + \frac{1}{1}} = \frac{1}{2}.$$

This suggests the following interpretation: the prior information on Θ in the first scenario is, in a loose sense, exactly as informative as having no useful prior information but one more observation, as in the second scenario.

Exercise: The effect of a stronger signal

0/1 point (graded)

For the model $X = \Theta + W$, and under the usual independence and normality assumptions for Θ and W, the mean squared error of the LMS estimator is

$$\frac{1}{(1/\sigma_0^2) + (1/\sigma_1^2)},$$

where σ_0^2 and σ_1^2 are the variances of Θ and W, respectively.

Suppose now that we change the observation model to $Y = 3\Theta + W$. In some sense the "signal" Θ has a stronger presence, relative to the noise term W, and we should expect to obtain a smaller mean squared error. Suppose $\sigma_0^2 = \sigma_1^2 = 1$. The mean squared error of the original model $X = \Theta + W$ is then $1/2$. In contrast, the mean squared error of the new model $Y = 3\Theta + W$ is

1/4

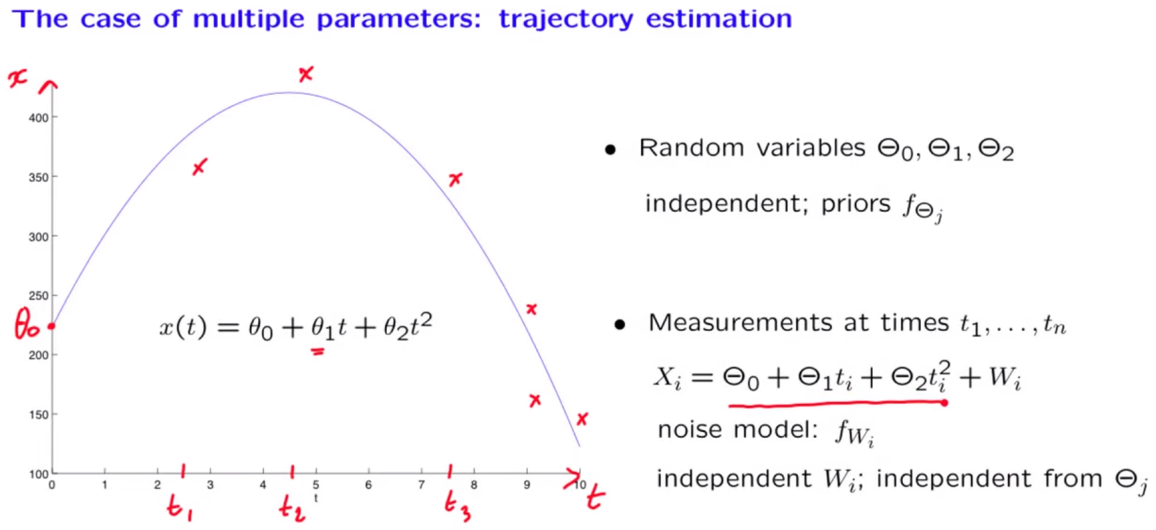
✖ **Answer:** 0.1

Hint: Do not solve the problem from scratch. Think of an alternative observation model in which you observe $Y' = \Theta + (W/3)$.

Solution:

Since Y' is just Y scaled by a factor of $1/3$, Y' carries the same information as Y, so that $\mathbf{E}[\Theta \mid Y] = \mathbf{E}[\Theta \mid Y']$. Thus, the alternative observation model $Y' = \Theta + (W/3)$ will lead to the same estimates and will have the same mean squared error as the unscaled model $Y = 3\Theta + W$. In the equivalent Y' model, we have a noise variance of $1/9$ and therefore the mean squared error is

$$\frac{1}{\frac{1}{1} + \frac{1}{1/9}} = \frac{1}{10}.$$



Trajectory motion from Newton's Law, but assuming we don't accurately know the initial position of the ball when it's thrown, the speed at which it's thrown and the gravitational constant(theta 0,1,2)
Then we start observing the position of the ball at different times, with some noise

A model with normality assumptions

$$X_i = \Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2 + W_i \qquad i = 1, \ldots, n$$

$$f_{\Theta|X}(\theta \mid x) = \frac{f_\Theta(\theta)\, f_{X|\Theta}(x \mid \theta)}{f_X(x)}$$

$$f_X(x) = \int f_\Theta(\theta) f_{X|\Theta}(x \mid \theta)\, d\theta$$

- assume $\Theta_j \sim N(0, \sigma_j^2)$, $W_i \sim N(0, \sigma^2)$; independent
- Given $\Theta = \theta = (\theta_0, \theta_1, \theta_2)$, X_i is: $N(\theta_0 + \theta_1 t_i + \theta_2 t_i^2, \sigma^2)$

$$f_{X_i|\Theta}(x_i \mid \theta) = c \cdot \exp\left\{ -(x_i - \theta_0 - \theta_1 t_i - \theta_2 t_i^2)^2 / 2\sigma^2 \right\}$$

- posterior: $f_{\Theta|X}(\theta \mid x) = \frac{1}{f_X(x)} \prod_{j=0}^{2} f_{\Theta_j}(\theta_j) \prod_{i=1}^{n} f_{X_i|\Theta}(x_i \mid \theta)$

$$c(x) \exp\left\{ -\frac{1}{2}\left(\frac{\theta_0^2}{\sigma_0^2} + \frac{\theta_1^2}{\sigma_1^2} + \frac{\theta_2^2}{\sigma_2^2}\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta_0 - \theta_1 t_i - \theta_2 t_i^2)^2 \right\}$$

Xs without subscripts are again vectors here

A model with normality assumptions

$$f_{\Theta|X}(\theta \mid x) = c(x) \exp\left\{ -\frac{1}{2}\left(\frac{\theta_0^2}{\sigma_0^2} + \frac{\theta_1^2}{\sigma_1^2} + \frac{\theta_2^2}{\sigma_2^2}\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta_0 - \theta_1 t_i - \theta_2 t_i^2)^2 \right\}$$

- MAP estimate: maximize over $(\theta_0, \theta_1, \theta_2)$; (minimize quadratic function)

$\frac{\partial}{\partial \theta_j}(\text{quad}(\theta)) = 0$ — 3 equations, 3 unknowns. ↑ linear

Exercise: Multiple observations and unknowns

4/4 points (graded)

Let Θ_1, Θ_2, W_1, and W_2 be independent standard normal random variables. We obtain two observations,

$$X_1 = \Theta_1 + W_1, \qquad X_2 = \Theta_1 + \Theta_2 + W_2.$$

Find the MAP estimate $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$ of (Θ_1, Θ_2) if we observe that $X_1 = 1$, $X_2 = 3$. (You will have to solve a system of two linear equations.)

$\hat{\theta}_1 =$ 1 ✔ **Answer:** 1

$\hat{\theta}_2 =$ 1 ✔ **Answer:** 1

Solution:

As usual, we focus on the exponential term in the numerator of the expression given by Bayes' rule. The prior contributes a term of the form

$$e^{-\frac{1}{2}(\theta_1^2+\theta_2^2)}.$$

Conditioned on $(\Theta_1, \Theta_2) = (\theta_1, \theta_2)$, the measurements are independent. In the conditional universe, X_1 is normal with mean θ_1, X_2 is normal with mean $\theta_1 + \theta_2$, and both variances are 1. Thus, the term $f_{X_1,X_2|\Theta_1,\Theta_2}$ makes a contribution of the form

$$e^{-\frac{1}{2}(x_1-\theta_1)^2} \cdot e^{-\frac{1}{2}(x_2-\theta_1-\theta_2)^2}.$$

We substitute $x_1 = 1$ and $x_2 = 3$, and in order to find the MAP estimate, we minimize the expression

$$\frac{1}{2}\left(\theta_1^2 + \theta_2^2 + (\theta_1 - 1)^2 + (\theta_1 + \theta_2 - 3)^2\right).$$

Setting the derivatives (with respect to θ_1 and θ_2) to zero, we obtain:

$$\hat{\theta}_1 + (\hat{\theta}_1 - 1) + (\hat{\theta}_1 + \hat{\theta}_2 - 3) = 0, \qquad \hat{\theta}_2 + (\hat{\theta}_1 + \hat{\theta}_2 - 3) = 0,$$

or

$$3\hat{\theta}_1 + \hat{\theta}_2 = 4, \qquad \hat{\theta}_1 + 2\hat{\theta}_2 = 3.$$

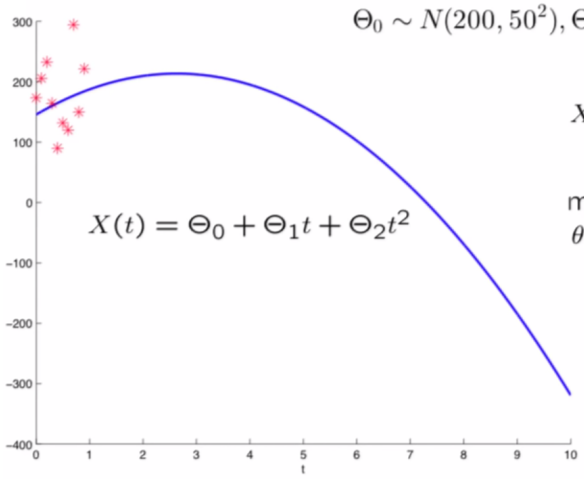
Either by inspection, or by substitution, we obtain the solution $\hat{\theta}_1 = 1, \hat{\theta}_2 = 1$.

Linear normal models.

- Θ_j and X_i and are linear functions of independent normal random variables
- $f_{\Theta|X}(\theta \mid x) = c(x) \exp\left\{ - \text{quadratic}(\theta_1, \ldots, \theta_m)\right\}$ linear regression
- MAP estimate: maximize over $(\theta_1, \ldots \theta_m)$; (minimize quadratic function) linear equations

  $\widehat{\Theta}_{\text{MAP},j}$: linear function of $X = (X_1, \ldots, X_n)$
- Facts:
  - $\widehat{\Theta}_{\text{MAP},j} = \mathbf{E}[\Theta_j \mid X]$
  - marginal posterior PDF of Θ_j: $f_{\Theta_j|X}(\theta_j \mid x)$, is normal
  - MAP estimate based on the joint posterior PDF: same as MAP estimate based on the marginal posterior PDF
  - $\mathbf{E}\left[(\widehat{\Theta}_{i,\text{MAP}} - \Theta_i)^2 \mid X = x\right]$: same for all x

An illustration Estimating the trajectory of a free-falling object



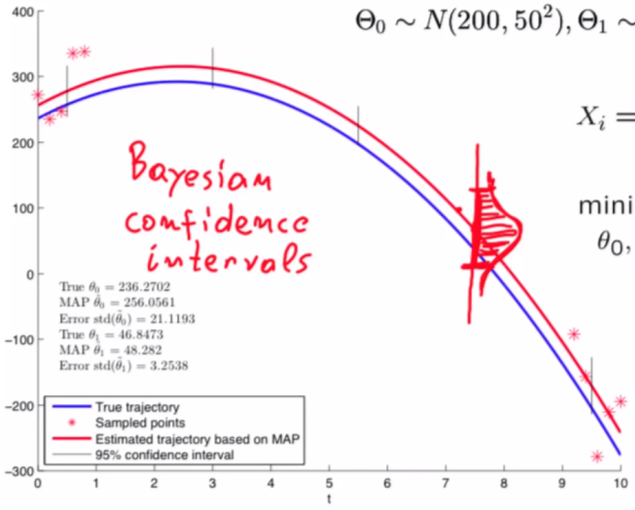
$\Theta_0 \sim N(200, 50^2), \Theta_1 \sim N(50, 50^2), \Theta_2 = -9.81, W_i \sim N(0, 50^2)$

$$X_i = \Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2 + W_i$$

$$\text{minimize}_{\theta_0, \theta_1, \theta_2} \quad \frac{1}{2}\left(\frac{\theta_0^2}{\sigma_0^2} + \frac{\theta_1^2}{\sigma_1^2} + \frac{\theta_2^2}{\sigma_2^2}\right) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta_0 - \theta_1 t_i - \theta_2 t_i^2)^2$$

Taking theta2 as a constant now so its prior disappears

An illustration Estimating the trajectory of a free-falling object



$\Theta_0 \sim N(200, 50^2), \Theta_1 \sim N(50, 50^2), \Theta_2 = -9.81, W_i \sim N(0, 50^2)$

$$X_i = \overbrace{\Theta_0 + \Theta_1 t_i + \Theta_2 t_i^2}^{x(t)} + W_i$$

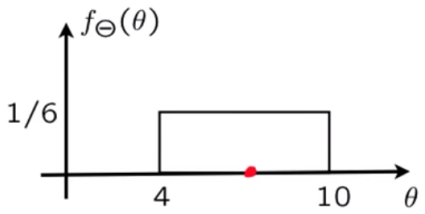
$$\text{minimize}_{\theta_0, \theta_1} \quad (\theta_0 - 200)^2 + (\theta_1 - 50)^2 + \sum_{i=1}^{n}(x_i - \theta_0 - \theta_1 t_i + 9.81 t_i^2)^2$$

$P(x(t) \in \text{interval} \mid \text{data}) = 0.95$

Least Means Squares Estimation

LMS estimation in the absence of observations

- unknown Θ; prior $p_\Theta(\theta)$
  - interested in a point estimate $\hat{\theta}$
  - no observations available
  - MAP rule: any $\hat{\theta} \in [4, 10]$
  - (Conditional) expectation: $\hat{\theta} = 7$



- Criterion: Mean Squared Error (MSE): $\mathbf{E}\left[(\Theta - \hat{\theta})^2\right]$

MAP rule isn't helpful here as the PDF is flat

LMS estimation in the absence of observations

- Least mean squares formulation:

minimize mean squared error (MSE), $\mathbf{E}\left[(\Theta - \hat{\theta})^2\right]$: $\hat{\theta} = \mathbf{E}[\Theta]$.

$E[\Theta^2] - 2E[\Theta]\hat{\theta} + \hat{\theta}^2 \quad \frac{d}{d\hat{\theta}} = 0: \; -2E[\Theta] + 2\hat{\theta} = 0$

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

$\mathrm{Var}(\Theta - \hat{\theta}) + (E[\Theta - \hat{\theta}])^2$ — minimized when $\hat{\theta} = E[\Theta]$

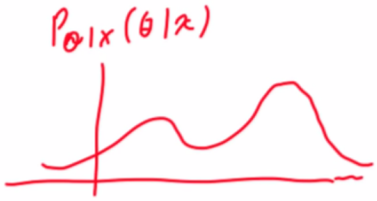
$\mathrm{Var}(\Theta)$

- Optimal mean squared error: $\mathbf{E}\left[(\Theta - \mathbf{E}[\Theta])^2\right] = \mathrm{var}(\Theta)$

theta hat is a number so no expected value
Var of r.v. +/- a constant is same as Var of just r.v. (blue)

LMS estimation of Θ based on X

- unknown Θ; prior $p_\Theta(\theta)$
  - interested in a point estimate $\hat{\theta}$
- observation X; model $p_{X|\Theta}(x\,|\,\theta)$
  - observe that $X = x$



minimize mean squared error (MSE), $\mathbf{E}\left[(\Theta - \hat{\theta})^2\right]$: $\hat{\theta} = \mathbf{E}[\Theta]$

minimize conditional mean squared error, $\mathbf{E}\left[(\Theta - \hat{\theta})^2 \,|\, X = x\right]$: $\hat{\theta} = \mathbf{E}[\Theta \,|\, X = x]$

- LMS estimate: $\hat{\theta} = \mathbf{E}[\Theta \,|\, X = x]$

  estimator: $\widehat{\Theta} = \mathbf{E}[\Theta \,|\, X]$

LMS estimation of Θ based on X

- $\mathbf{E}[\Theta]$ minimizes $\mathbf{E}\left[(\Theta - \hat{\theta})^2\right]$

$$E\left[(\Theta - E[\Theta])^2\right] \le E\left[(\Theta - c)^2\right], \text{ for all } c$$

- $\mathbf{E}[\Theta \,|\, X = x]$ minimizes $\mathbf{E}\left[(\Theta - \hat{\theta})^2 \,|\, X = x\right]$

$$E\left[(\Theta - E[\Theta|X=x])^2 \,|\, X=x\right] \le E\left[(\Theta - g(x))^2 \,|\, X=x\right] \text{ for all } x$$

$$E\left[(\Theta - E[\Theta|X])^2 \,|\, X\right] \le E\left[(\Theta - g(X))^2 \,|\, X\right]$$

$$E\left[(\Theta - E[\Theta|X])^2\right] \le E\left[(\Theta - g(X))^2\right]$$

$\widehat{\Theta}_{\text{LMS}} = \mathbf{E}[\Theta \,|\, X]$ minimizes $\mathbf{E}\left[(\Theta - g(X))^2\right]$, over all estimators $\widehat{\Theta} = g(X)$

The error would be at least as large using some other constant c or function g(x) instead of the expected value
2nd to 3rd line
using Law of iterated expectations to turn a conditional into an unconditional

Exercise: LMS estimation

1/1 point (graded)

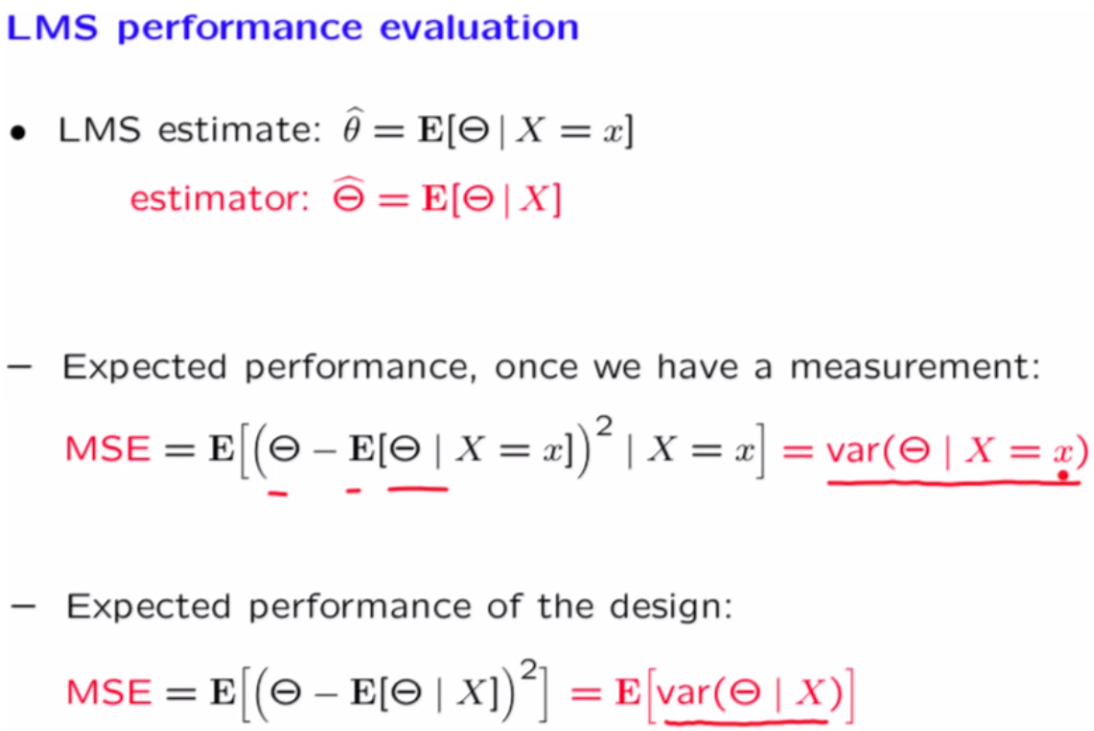
Let Θ be the bias of a coin, i.e., the probability of Heads at each toss. We assume that Θ is uniformly distributed on $[0, 1]$. Let K be the number of Heads in 9 independent tosses.

By performing some fancy and very precise measurements on the structure of that particular coin, we determine that $\Theta = 1/3$. Find the LMS estimate of K based on Θ.

3 ✔ **Answer:** 3

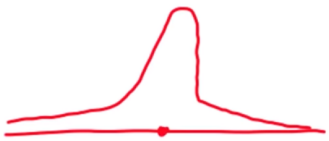
Solution:

Do not be confused by the choice of notation. Here, K is the variable being estimated and Θ is an observation. The posterior in this case is $p_{K|\Theta}$ and is a binomial distribution with parameters 9 and $1/3$. Thus, the LMS estimate is $\mathbf{E}[K \mid \Theta = \theta] = n\theta = 9/3 = 3$.



MSE is the average value of the expected value over all possible values of x
Which is also the variance

LMS estimation of Θ based on X



- LMS relevant to estimation (not hypothesis testing)
- Same as MAP if the posterior is unimodal and symmetric around the mean
  - e.g., when posterior is normal (the case in "linear–normal" models)

6. Exercise: LMS estimation error

Bookmark this page

Exercise: LMS estimation error

0/3 points (graded)

As in the previous exercise, let Θ be the bias of a coin, i.e., the probability of Heads at each toss. We assume that Θ is uniformly distributed on $[0, 1]$. Let K be the number of Heads in 9 independent tosses. We have seen that the LMS estimate of K is $\mathbf{E}[K \mid \Theta = \theta] = n\theta$.

a) Find the conditional mean squared error $\mathbf{E}\left[\left(K - \mathbf{E}[K \mid \Theta = \theta]\right)^2 \mid \Theta = \theta\right]$ if $\theta = 1/3$.

3/4

✖ **Answer:** 2

b) Find the overall mean squared error of this estimation procedure.

3/4

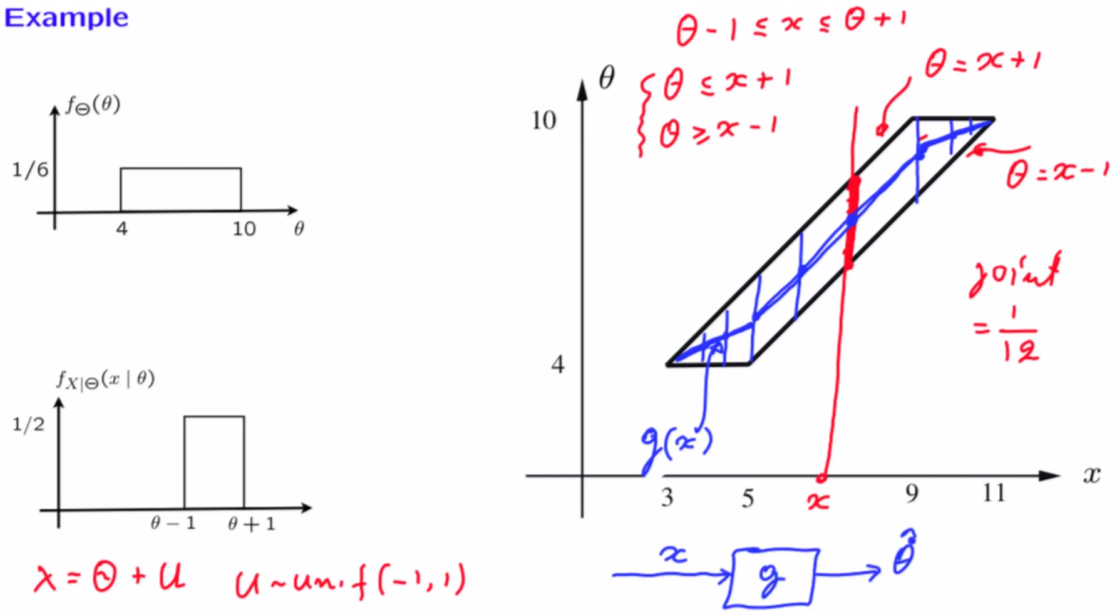
✖ **Answer:** 1.5

Solution:

a) This is the variance of the conditional distribution of K. Since the conditional distribution is binomial with parameters $n = 9$ and $\theta = 1/3$, the conditional variance is $9\,(1/3)\,(2/3) = 2$.

b) This is the average of the conditional variance, averaged over all possible values of the observation Θ, which has a uniform distribution:

$$\begin{aligned}\int_0^1 f_\Theta(\theta)\,\mathrm{Var}(K \mid \Theta = \theta)\,d\theta &= \int_0^1 9\theta(1-\theta)\,d\theta \\ &= \left(9\frac{1}{2}\theta^2 - 9\frac{\theta^3}{3}\right)\Big|_0^1 \\ &= 4.5 - 3 \\ &= 1.5.\end{aligned}$$



the blue line is the midpoint in the joint PDF of theta x and is therefore the estimator of the mean
notice how it is 3 lines of differing slope

8. Exercise: LMS example

Bookmark this page

Exercise: LMS example

1/1 point (graded)

The random variables Θ and X are described by a joint PDF which is uniform on the triangular set defined by the constraints $0 \le x \le 1, 0 \le \theta \le x$. Find the LMS estimate of Θ given that $X = x$, for x in the range $[0, 1]$. Express your answer in terms of x using standard notation.

x/2 ✔ **Answer:** x/2

$\frac{x}{2}$

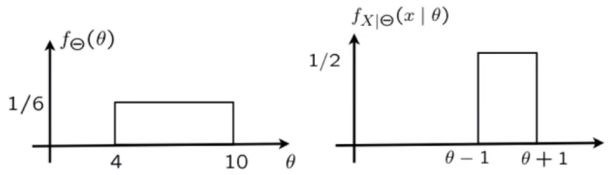
STANDARD NOTATION

Solution:

The conditional PDF of Θ given that $X = x$ is uniform on the set $[0, x]$. Thus, the conditional expectation of Θ given that $X = x$ is equal to $x/2$.

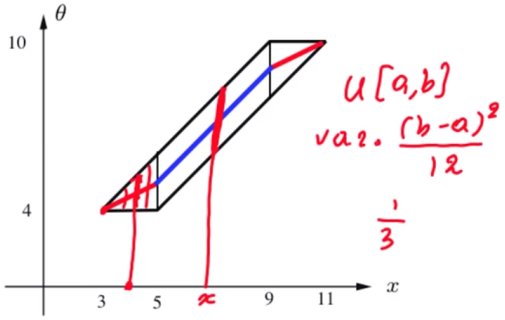
By drawing a triangle where theta = x, the LMS estimator is a slope that goes halfway between this so x/2

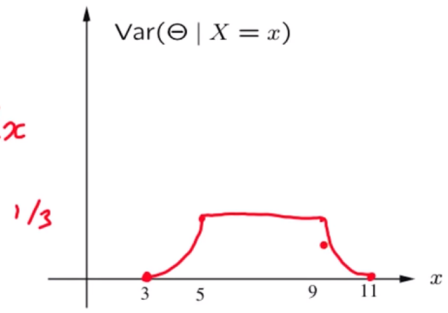
Conditional mean squared error



- $\mathbf{E}[(\Theta - \mathbf{E}[\Theta \mid X = x])^2 \mid X = x]$
  - same as $\mathrm{Var}(\Theta \mid X = x)$: variance of conditional distribution of Θ

$E[Var(\Theta|X)] = \int f_X(x)\, Var(\Theta|X=x)\, dx$





b-a is the width of the interval of theta where x=x (red line)
plotting variance by looking at the changing width of intervals for different values of x
can get the marginal PDF f(x) from the joint that we have and integrating out theta

Exercise: Mean squared error

4/4 points (graded)

As in an earlier exercise, we assume that the random variables Θ and X are described by a joint PDF which is uniform on the triangular set defined by the constraints $0 \le x \le 1, 0 \le \theta \le x$.

a) Find an expression for the conditional mean squared error of the LMS estimator given that $X = x$, valid for $x \in [0, 1]$. Express your answer in terms of x using standard notation.

(x^2)/12 ✔ **Answer:** x^2/12

$$\frac{x^2}{12}$$

b) Find the (unconditional) mean squared error of the LMS estimator.

1/24 ✔ **Answer:** 0.04167

STANDARD NOTATION

Solution:

a) We saw that the conditional PDF of Θ is uniform on the range $[0, x]$. Hence, the conditional variance is $x^2/12$.

b) This is given by the integral of the conditional variance, weighted by the PDF of X. The PDF of X is found using the formula for going from the joint to the marginal, and is $f_X(x) = 2x$, for $x \in [0, 1]$. Thus, the mean squared error is

$$\int_0^1 \frac{x^2}{12} \cdot 2x \, dx = \frac{1}{6} \int_0^1 x^3 \, dx = \frac{1}{24}.$$

x is the width of the triangle at each interval (because theta = x)
and the variance is the width/12 (for a uniform distribution)

LMS estimation with multiple observations or unknowns

- unknown Θ; prior $p_\Theta(\theta)$
  - interested in a point estimate $\hat{\theta}$
- observations $X = (X_1, X_2, \ldots, X_n)$; model $p_{X|\Theta}(x \mid \theta)$
  - observe that $X = x$
  - new universe: condition on $X = x$
- LMS estimate: $\mathbf{E}[\Theta \mid X_1 = x_1, \ldots, X_n = x_n]$

- If Θ is a vector, apply to each component separately

$\Theta = (\Theta_1, \ldots, \Theta_m)$ $\quad \hat{\Theta}_j = E[\Theta_j \mid X_1 = x_1, \ldots, X_n = x_n]$

Some challenges in LMS estimation

$$f_{\Theta|X}(\theta \mid x) = \frac{f_\Theta(\theta)\, f_{X|\Theta}(x \mid \theta)}{f_X(x)}$$

$$f_X(x) = \int f_\Theta(\theta') f_{X|\Theta}(x \mid \theta')\, d\theta'$$

- Full correct model, $f_{X|\Theta}(x \mid \theta)$, may not be available
- Can be hard to compute/implement/analyze

$$E[\Theta_j \mid X = x] = \int\cdots\int \theta_j \, f_{\Theta|X}(\theta \mid x)\, d\theta_1 \cdots d\theta_m$$

Exercise: Multidimensional challenges

0/2 points (graded)

Suppose that f_Θ and $f_{X|\Theta}$ are described by simple closed-form formulas. Suppose that Θ is one-dimensional but X is high-dimensional.

a) Suppose that a specific value x of the random variable X has been observed. Is it true that the calculation of the LMS estimate will always involve only ordinary integrals (integrals with respect to only one variable)?

No ✖ **Answer:** Yes

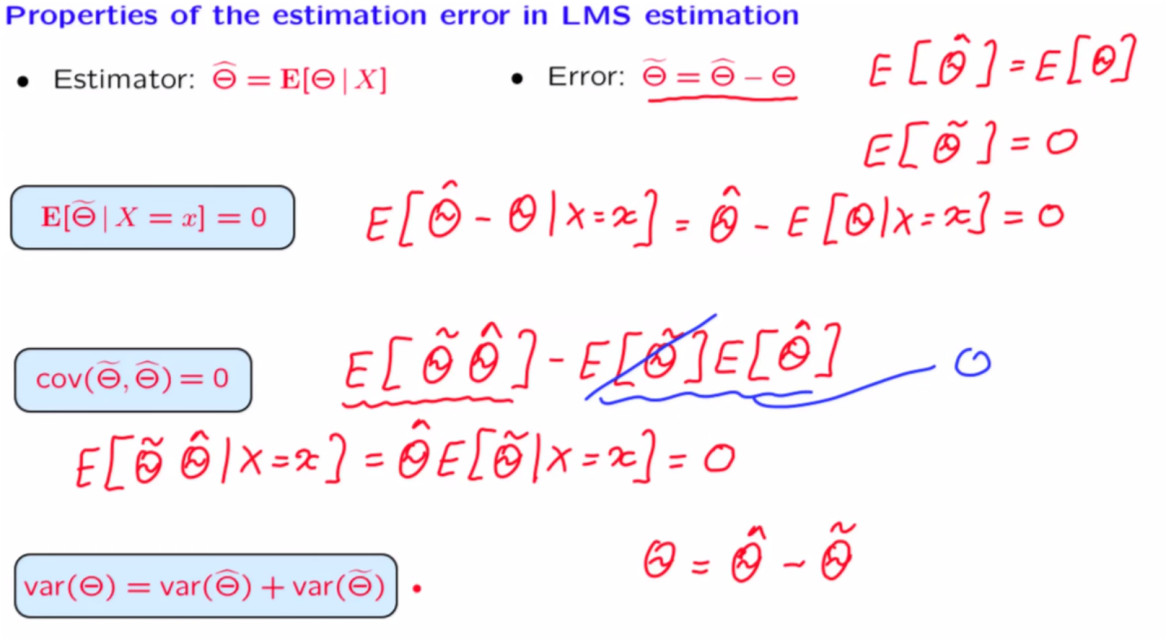
b) Is it true that the calculation of the mean squared error of the LMS estimator will always involve only ordinary integrals (integrals with respect to only one variable)?

Yes ✖ **Answer:** No

Solution:

a) The denominator in Bayes' rule involves an integral with respect to θ. Once the conditional PDF is available, the LMS estimate is calculated by integrating again over the one-dimensional variable θ.

b) In this case, we need to average the conditional variance over all possible values of x, and this will involve a multiple integral.



var property is true when covariances of the estimator and error are 0

14. Exercise: Theoretical properties

Bookmark this page

Exercise: Theoretical properties

1/2 points (graded)

Let $\widehat{\Theta}$ be an estimator of a random variable Θ, and let $\widetilde{\Theta} = \widehat{\Theta} - \Theta$ be the estimation error.

a) In this part of the problem, let $\widehat{\Theta}$ be specifically the LMS estimator of Θ. We have seen that for the case of the LMS estimator, $\mathbf{E}[\widetilde{\Theta} \mid X = x] = 0$ for every x. Is it also true that $\mathbf{E}[\widetilde{\Theta} \mid \Theta = \theta] = 0$ for all θ? Equivalently, is it true that $\mathbf{E}[\widehat{\Theta} \mid \Theta = \theta] = \theta$ for all θ?

Yes ✖ **Answer:** No

b) In this part of the problem, $\widehat{\Theta}$ is no longer necessarily the LMS estimator of Θ. Is the property $\mathsf{Var}(\Theta) = \mathsf{Var}(\widehat{\Theta}) + \mathsf{Var}(\widetilde{\Theta})$ true for every estimator $\widehat{\Theta}$?

No ✔ **Answer:** No

Solution:

a) There is no reason for this relation to be true. For an example, suppose that Θ is a Bernoulli random variable. With a noisy measurement, $\widehat{\Theta}$ will be somewhere in between 0 and 1, and therefore will never be equal to the true value of θ, which is either 0 or 1 exactly.

b) There is no reason for this to be the case. In fact, the variance of $\widehat{\Theta}$, for a poorly chosen estimator, can be larger than the variance of Θ. For an example, consider the usual model of an observation $X = \Theta + W$ and the estimator $\widehat{\Theta} = 100X$.

1. Defective Coin

Bookmark this page

Problem 1. Defective Coin

0/3 points (graded)

A defective coin minting machine produces coins whose probability of Heads is a random variable Q with PDF

$$f_Q(q) = \begin{cases} 5q^4, & \text{if } q \in [0,1], \\ 0, & \text{otherwise.} \end{cases}$$

A coin produced by this machine is tossed repeatedly, with successive tosses assumed to be independent. Let A be the event that the first toss of this coin results in Heads, and let B be the event that the second toss of this coin results in Heads.

1. $\mathbf{P}(A) =$ 0.3125 ✖ **Answer:** 5/6

   (Your answer should be a number.)

2. Find the conditional PDF of Q given event A. Express your answer in terms of q using standard notation.

   For $0 \le q \le 1$, $f_{Q|A}(q) =$ (q^4)/4 ✖ **Answer:** 6*q^5

   $\frac{q^4}{4}$

3. $\mathbf{P}(B \mid A) =$ 0.015625 ✖ **Answer:** 6/7

   (Your answer should be a number.)

Solution:

1. To calculate $\mathbf{P}(A)$, we use the continuous version of the total probability theorem:

$$\mathbf{P}(A) = \int_0^1 \mathbf{P}(A \mid Q = q)\, f_Q(q)\ dq = \int_0^1 q \cdot (5q^4)\ dq = \left[\frac{5}{6}q^6\right]_0^1 = \frac{5}{6}.$$

2. Using Bayes' rule,

$$\begin{aligned} f_{Q|A}(q) &= \frac{\mathbf{P}(A \mid Q = q)\, f_Q(q)}{\mathbf{P}(A)} \\ &= \begin{cases} \dfrac{q \cdot (5q^4)}{5/6}, & \text{if } 0 \le q \le 1, \\ 0, & \text{otherwise,} \end{cases} \\ &= \begin{cases} 6q^5, & \text{if } 0 \le q \le 1, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

3. Again, we use the continuous version of the total probability theorem:

$$\begin{aligned} \mathbf{P}(B \mid A) &= \int_0^1 \mathbf{P}(B \mid A, Q = q)\, f_{Q|A}(q)\ dq \\ &= \int_0^1 \mathbf{P}(B \mid Q = q)\, f_{Q|A}(q)\ dq \\ &= \int_0^1 q\,(6q^5)\ dq \\ &= 6/7. \end{aligned}$$

The second equality holds because for a given value q of Q, the events A and B are (conditionally) independent.

2. Hypothesis test between two coins

Bookmark this page

Problem 2. Hypothesis test between two coins

1/5 points (graded)

Alice has two coins. The probability of Heads for the first coin is $1/4$, and the probability of Heads for the second is $3/4$. Other than this difference, the coins are indistinguishable. Alice chooses one of the coins at random and sends it to Bob. The random selection used by Alice to pick the coin to send to Bob is such that the first coin has a probability p of being selected. Assume that $0 < p < 1$. Bob tries to guess which of the two coins he received by tossing it 3 times in a row and observing the outcome. Assume that for any particular coin, all tosses of that coin are independent.

1. Given that Bob observed k Heads out of the 3 tosses (where $k = 0, 1, 2, 3$), what is the conditional probability that he received the first coin?

- $\dfrac{3^k \cdot p}{3^{3-k} \cdot p + 3^k \cdot (1-p)}$
- $\dfrac{p}{3^{3-k}}$
- $\dfrac{3^{3-k} \cdot p}{3^{3-k} \cdot p + 3^k \cdot (1-p)}$
- $\dfrac{3^{3-k} \cdot (1-p)}{3^{3-k} \cdot p + 3^k \cdot (1-p)}$

✖

2. Hypothesis test between two coins

Bookmark this page

Problem 2. Hypothesis test between two coins

1/5 points (graded)

Alice has two coins. The probability of Heads for the first coin is $1/4$, and the probability of Heads for the second is $3/4$. Other than this difference, the coins are indistinguishable. Alice chooses one of the coins at random and sends it to Bob. The random selection used by Alice to pick the coin to send to Bob is such that the first coin has a probability p of being selected. Assume that $0 < p < 1$. Bob tries to guess which of the two coins he received by tossing it 3 times in a row and observing the outcome. Assume that for any particular coin, all tosses of that coin are independent.

1. Given that Bob observed k Heads out of the 3 tosses (where $k = 0, 1, 2, 3$), what is the conditional probability that he received the first coin?

- $\frac{3^k \cdot p}{3^{3-k} \cdot p + 3^k \cdot (1-p)}$
- $\frac{p}{3^{3-k}}$
- $\frac{3^{3-k} \cdot p}{3^{3-k} \cdot p + 3^k \cdot (1-p)}$
- $\frac{3^{3-k} \cdot (1-p)}{3^{3-k} \cdot p + 3^k \cdot (1-p)}$

✖

3. For this part, assume that $p = 3/4$.

- What is the probability that Bob will guess the coin correctly using the decision rule from part 2?

1/2 ✖

- Suppose instead that Bob tries to guess which coin he received without tossing it. He still guesses the coin in order to minimize the probability of error. What is the probability that Bob will guess the coin correctly under this scenario?

3/4 ✔

4. Bob uses the decision rule of Part 2. If p is small, then Bob will always decide in favor of the second coin, ignoring the results of the three tosses. The range of such p's is $[0, t)$. Find t.

$t =$ ✖ .

1. Let Y be the number of Heads Bob observed in the three tosses. Let C denote the coin that Bob received, so that $C = 1$ if Bob received the first coin, and $C = 2$ if Bob received the second coin. Then $\mathbf{P}(C = 1) = p$ and $\mathbf{P}(C = 2) = 1 - p$. Given the value of C, Y is a binomial random variable.

   We can find the conditional probability that Bob received the first coin given that he observed k Heads using Bayes' rule.

$$\begin{aligned}
\mathbf{P}(C = 1 \mid Y = k) &= \frac{\mathbf{P}(Y = k \mid C = 1)\,\mathbf{P}(C = 1)}{\mathbf{P}(Y = k)} \\
&= \frac{\mathbf{P}(Y = k \mid C = 1)\,\mathbf{P}(C = 1)}{\mathbf{P}(Y = k \mid C = 1)\,\mathbf{P}(C = 1) + \mathbf{P}(Y = k \mid C = 2)\,\mathbf{P}(C = 2)} \\
&= \frac{\binom{3}{k}(1/4)^k(3/4)^{3-k} \cdot p}{\binom{3}{k}(1/4)^k(3/4)^{3-k} + \binom{3}{k}(1/4)^{3-k}(3/4)^k \cdot (1-p)} \\
&= \frac{3^{3-k} \cdot p}{3^{3-k} \cdot p + 3^k \cdot (1-p)}.
\end{aligned}$$

2. Given that Bob observes k Heads, he is to decide whether the first or second coin was used. To minimize the probability of error, he should use the MAP rule, which in this case is to decide on the first coin when $\mathbf{P}(C = 1|Y = k) \geq \mathbf{P}(C = 2|Y = k)$. From symmetry, the second item, namely $\mathbf{P}(C = 2|Y = k)$ is equal to $\frac{3^k \cdot (1-p)}{3^{3-k} \cdot p + 3^k \cdot (1-p)}$. We then have the following equivalent versions of this decision rule:

$$\begin{aligned}
\mathbf{P}(C = 1|Y = k) &\geq \mathbf{P}(C = 2|Y = k) \\
\frac{3^{3-k} \cdot p}{3^{3-k} \cdot p + 3^k \cdot (1-p)} &\geq \frac{3^k \cdot (1-p)}{3^{3-k} \cdot p + 3^k \cdot (1-p)} \\
3^{3-k} \cdot p &\geq 3^k \cdot (1-p) \\
3^{2k-3} &\leq \frac{p}{1-p} \\
2k - 3 &\leq \log_3 \frac{p}{1-p} \\
k &\leq \frac{3}{2} + \frac{1}{2}\log_3 \frac{p}{1-p}.
\end{aligned}$$

3.
   - If $p = 3/4$, the threshold in the rule above is equal to 2. Therefore, Bob will decide that he received the first coin when he observes 0, 1, or 2 Heads, and will decide that he received the second coin when he observes 3 Heads.

     We find the probability of a correct decision using the total probability theorem:

$$\begin{aligned}
\mathbf{P}(\text{Correct}) &= \mathbf{P}(\text{Correct}|C=1)\cdot p + \mathbf{P}(\text{Correct}|C=2)\cdot(1-p) \\
&= \mathbf{P}(Y<3|C=1)\cdot p + \mathbf{P}(Y=3|C=2)\cdot(1-p) \\
&= (1-\mathbf{P}(Y=3|C=1))\cdot p + \mathbf{P}(Y=3|C=2)\cdot(1-p) \\
&= (1-(1/4)^3)(3/4) + (3/4)^3(1/4) \\
&= \frac{216}{256} = \frac{27}{32}.
\end{aligned}$$

   - In the absence of any data, Bob should simply guess that he received whichever coin Alice was more likely to choose, which in this case is the first coin. His decision will be correct if he indeed receives the first coin, which happens with probability $3/4$.

     Note that observing 3 coin tosses increases the probability of making a correct decision from $3/4$ to $27/32$, a difference of approximately 0.09375.

4. Bob will never decide that he received the first coin if the threshold in the decision rule in Part 2 is negative, i.e., when

$$\begin{aligned}
\frac{3}{2} + \frac{1}{2}\log_3\frac{p}{1-p} &< 0 \\
\log_3\frac{p}{1-p} &< -3 \\
\frac{p}{1-p} &< \frac{1}{27} \\
p &< \frac{1}{28}.
\end{aligned}$$

   If $p < 1/28$, the prior probability of receiving the first coin is so low that no amount of evidence from 3 tosses of the coin will make Bob decide he received the first coin.

3. Hypothesis test with a continuous observation

Bookmark this page

Problem 3. Hypothesis test with a continuous observation

3 points possible (graded)

Let Θ be a Bernoulli random variable that indicates which one of two hypotheses is true, and let $\mathbf{P}(\Theta = 1) = p$. Under the hypothesis $\Theta = 0$, the random variable X has a normal distribution with mean 0, and variance 1. Under the alternative hypothesis $\Theta = 1$, X has a normal distribution with mean 2 and variance 1.

Consider the MAP rule for deciding between the two hypotheses, given that $X = x$.

1. Suppose for this part of the problem that $p = 2/3$. The MAP rule can choose in favor of the hypothesis $\Theta = 1$ if and only if $x \geq c_1$. Find the value of c_1.

   $c_1 =$

2. For this part, assume again that $p = 2/3$. Find the conditional probability of error for the MAP decision rule, given that the hypothesis $\Theta = 0$ is true.

   $\mathbf{P}(\text{error}|\Theta = 0) =$

3. Find the overall (unconditional) probability of error associated with the MAP rule for $p = 1/2$.

1. For $0 < p < 1$, we can choose in favor of the hypothesis $\Theta = 1$ if and only if

$$\begin{aligned} f_{X|\Theta}(x \mid 1)\, p_\Theta(1) &\geq f_{X|\Theta}(x|0)\, p_\Theta(0) \\ \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-2)^2\right) \cdot p &\geq \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) \cdot (1-p) \\ \frac{x^2}{2} - \frac{(x-2)^2}{2} &\geq \ln \frac{1-p}{p} \\ x &\geq 1 + \frac{1}{2} \ln \frac{1-p}{p}. \end{aligned}$$

   For $p = 2/3$, this threshold corresponds to $c_1 = 1 - (\ln 2)/2 \approx 0.6534$.

2. Under the hypothesis $\Theta = 0$, an error occurs if we decide $\Theta = 1$. Therefore,

$$\begin{aligned} \mathbf{P}(\text{error}|\Theta = 0) &= \mathbf{P}(X \geq c_1 \mid \Theta = 0) \\ &= 1 - \mathbf{P}(X < c_1 \mid \Theta = 0) \\ &\approx 1 - \Phi(0.65) \\ &\approx 0.2578, \end{aligned}$$

   since under $\Theta = 0$, X is a standard normal random variable.

3. With $p = 1/2$, the threshold becomes 1. Therefore, we decide $\Theta = 1$, whenever $x \geq 1$, and decide $\Theta = 0$, whenever $x < 1$. *f*

$$\begin{aligned} \mathbf{P}(\text{error}) &= \mathbf{P}(\text{error}|\Theta = 0)\, p_\Theta(0) + \mathbf{P}(\text{error}|\Theta = 1)\, p_\Theta(1) \\ &= \mathbf{P}(X \geq 1 \mid \Theta = 0)\, \frac{1}{2} + \mathbf{P}(X < 1 \mid \Theta = 1)\, \frac{1}{2} \\ &= \frac{1 - \Phi(1)}{2} + \frac{\mathbf{P}(X - 2 < -1 \mid \Theta = 1)}{2} \\ &= \frac{1 - \Phi(1)}{2} + \frac{1 - \Phi(1)}{2} \\ &= 1 - \Phi(1) \\ &\approx 1 - 0.8413 = 0.1587. \end{aligned}$$

4. Trajectory estimation

Bookmark this page

Problem 4. Trajectory estimation, Part I

2 points possible (graded)

Note: For this problem, you may find this summary useful. (This is also available at the bottom of Lecture 15, *12. Multiple parameters; trajectory estimation*.)

The vertical coordinate ("height") of an object in free fall is described by an equation of the form

$$x(t) = \theta_0 + \theta_1 t + \theta_2 t^2,$$

where θ_0, θ_1, and θ_2 are some parameters and t stands for time. At certain times $t_1, \ldots, t_n$, we make noisy observations $Y_1, \ldots, Y_n$, respectively, of the height of the object. Based on these observations, we would like to estimate the object's vertical trajectory.

We consider the special case where there is only one unknown parameter. We assume that θ_0 (the height of the object at time zero) is a known constant. We also assume that θ_2 (which is related to the acceleration of the object) is known. We view θ_1 as the realized value of a continuous random variable Θ_1. The observed height at time t_i is $Y_i = \theta_0 + \Theta_1 t_i + \theta_2 t_i^2 + W_i$, $i = 1, \ldots, n$, where W_i models the observation noise. We assume that $\Theta_1 \sim N(0, 1)$, $W_1, \ldots, W_n \sim N(0, \sigma^2)$, and that all these random variables are independent.

In this case, finding the MAP estimate of Θ_1 involves the minimization of

$$\theta_1^2 + \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 t_i - \theta_2 t_i^2)^2$$

with respect to θ_1.

1. Carry out this minimization and choose the correct formula for the MAP estimate, $\hat{\theta}_1$, from the options below.

- $\hat{\theta}_1 = \dfrac{\sum_{i=1}^{n} t_i (y_i - \theta_0 - \theta_2 t_i^2)}{\sigma^2}$
- $\hat{\theta}_1 = \dfrac{\sum_{i=1}^{n} t_i (y_i - \theta_0 - \theta_2 t_i^2)}{\sigma^2 + \sum_{i=1}^{n} t_i^2}$
- $\hat{\theta}_1 = \dfrac{\sum_{i=1}^{n} t_i (y_i - \theta_0 - \theta_2 t_i^2)}{\sigma^2 + \sum_{i=1}^{n} \theta_2 t_i^2}$
- none of the above

2. The formula for $\hat{\theta}_1$ can be used to define the MAP estimator, $\hat{\Theta}_1$ (a random variable), as a function of $t_1, \ldots, t_n$ and the random variables $Y_1, \ldots, Y_n$. Identify whether the following statement is true:

The MAP estimator $\hat{\Theta}_1$ has a normal distribution.

Select an option

Solution:

1. Setting the partial derivative with respect to θ_1 equal to zero, we obtain

$$\theta_1 - \frac{1}{\sigma^2} \sum_{i=1}^{n} t_i \left(y_i - \theta_0 - \theta_1 t_i - \theta_2 t_i^2\right) = 0,$$

yielding the MAP estimate

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{n} t_i \left(y_i - \theta_0 - \theta_2 t_i^2\right)}{\sigma^2 + \sum_{i=1}^{n} t_i^2}.$$

2. We have

$$\hat{\Theta}_1 = \frac{\sum_{i=1}^{n} t_i \left(Y_i - \theta_0 - \theta_2 t_i^2\right)}{\sigma^2 + \sum_{i=1}^{n} t_i^2}.$$

Recall that the observation model is $Y_i = \theta_0 + \Theta_1 t_i + \theta_2 t_i^2 + W_i$, and so we can rewrite the estimator as

$$\begin{aligned} \hat{\Theta}_1 &= \frac{\sum_{i=1}^{n} t_i \left(\Theta_1 t_i + W_i\right)}{\sigma^2 + \sum_{i=1}^{n} t_i^2} \\ &= \frac{\Theta_1 \sum_{i=1}^{n} t_i^2 + \sum_{i=1}^{n} t_i W_i}{\sigma^2 + \sum_{i=1}^{n} t_i^2}. \end{aligned}$$

We see that $\hat{\Theta}_1$ is a linear function of Θ_1 and $W_1, \ldots, W_n$, which are all normal and independent. Since a linear function of independent normal random variables is normal, it follows that $\hat{\Theta}_1$ is normal.

Problem 4. Trajectory estimation, Part II

3 points possible (graded)

1. Let $\sigma = 1$ and consider the special case of only two observations ($n = 2$). Write down a formula for the mean squared error $\mathbb{E}\left[(\hat{\Theta}_1 - \Theta_1)^2\right]$, as a function of t_1 and t_2. Enter **t_1** for t_1 and **t_2** for t_2.

$\mathbb{E}\left[(\hat{\Theta}_1 - \Theta_1)^2\right] =$

2. Consider the "experimental design" problem of choosing when to make measurements. Under the assumptions of the previous part, and under the constraints $0 \le t_1, t_2 \le 10$, find the values of t_1 and t_2 that minimize the mean squared error associated with the MAP estimator.

$t_1 =$

$t_2 =$

5. Hypothesis test between two normals

Bookmark this page

Problem 5. Hypothesis test between two normals

2 points possible (graded)

Conditioned on the result of an unbiased coin flip, the random variables $T_1, T_2, \ldots, T_n$ are independent and identically distributed, each drawn from a common normal distribution with mean zero. If the result of the coin flip is Heads, this normal distribution has variance 1; otherwise, it has variance 4. Based on the observed values $t_1, t_2, \ldots, t_n$, we use the MAP rule to decide whether the normal distribution from which they were drawn has variance 1 or variance 4. The MAP rule decides that the underlying normal distribution has variance 1 if and only if

$$\left| c_1 \sum_{i=1}^{n} t_i^2 + c_2 \sum_{i=1}^{n} t_i \right| < 1.$$

Find the values of $c_1 \geq 0$ and $c_2 \geq 0$ such that this is true. Express your answer in terms of n, and use "ln" to denote the natural logarithm function, as in "ln(3)".

$c_1 =$

$c_2 =$

Solution:

Let $\Theta = 0$ denote that the observations $t_1, t_2, \ldots, t_n$ were generated from a normal distribution with variance 1, and let $\Theta = 1$ denote that they were generated from a normal distribution with variance 4. For simplicity, let us use the notation $N(t_1, \ldots, t_n; 0, \sigma^2)$ to denote the joint PDF of n i.i.d. normal random variables with mean 0 and variance σ^2, evaluated at $t_1, \ldots, t_n$.

Therefore, given the observations $t_1, \ldots, t_n$, the posterior probability that the samples are generated from a normal distribution with variance 1 is

$$\mathbf{P}(\Theta = 0 \mid T_1 = t_1, \ldots, T_n = t_n) = \frac{(1/2) \cdot N(t_1, \ldots, t_n; 0, 1)}{(1/2) \cdot N(t_1, \ldots, t_n; 0, 1) + (1/2) \cdot N(t_1, \ldots, t_n; 0, 4)}.$$

Similarly, the probability that the samples are generated from a normal distribution with variance 4 is given by

$$\mathbf{P}(\Theta = 1 \mid T_1 = t_1, \ldots, T_n = t_n) = \frac{(1/2) \cdot N(t_1, \ldots, t_n; 0, 4)}{(1/2) \cdot N(t_1, \ldots, t_n; 0, 1) + (1/2) \cdot N(t_1, \ldots, t_n; 0, 4)}.$$

The MAP rule favors $\Theta = 0$ if the following inequality holds:

$$\mathbf{P}(\Theta = 0 \mid T_1 = t_1, \ldots, T_n = t_n) > \mathbf{P}(\Theta = 1 \mid T_1 = t_1, \ldots, T_n = t_n)$$

We notice that the denominators in the expressions for $\mathbf{P}(\Theta = 0 \mid \ldots)$ and $\mathbf{P}(\Theta = 1 \mid \ldots)$ are the same, so it suffices to compare the numerators. Therefore, the MAP rule favors $\Theta = 0$ if the following inequality holds:

$$\begin{aligned} N(t_1, \ldots, t_n; 0, 1) &> N(t_1, \ldots, t_n; 0, 4) \\ \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi \cdot 1}} e^{-\frac{t_i^2}{2 \cdot 1}} &> \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi \cdot 4}} e^{-\frac{t_i^2}{2 \cdot 4}}. \end{aligned}$$

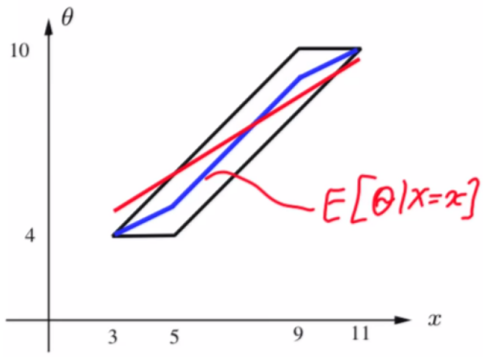
With a little bit of algebra, we obtain

$$\left| \frac{3}{8} \sum_{i=1}^{n} t_i^2 \right| < n \cdot \ln(2).$$

Note: If the means under the two hypotheses were different, a similar answer would be obtained but with a nonzero coefficient c_2.

LLMS formulation

- Unknown Θ; observation X
- Minimize $\mathbf{E}\left[(\widehat{\Theta} - \Theta)^2\right]$
- Estimators $\widehat{\Theta} = g(X) \longrightarrow \widehat{\Theta}_{\text{LMS}} = \mathbf{E}[\Theta \mid X]$
- Consider estimators of Θ, of the form $\widehat{\Theta} = aX + b$
- Minimize $\mathbf{E}\left[(\Theta - aX - b)^2\right]$, w.r.t. a, b
- If $\mathbf{E}[\Theta \mid X]$ is linear in X, then $\widehat{\Theta}_{\text{LMS}} = \widehat{\Theta}_{\text{LLMS}}$



in other words want to choose a line such that the difference between the line and the estimator is as small as possible
This is just an optimisation of 2 numbers a and b

3. Exercise: LMS and LLMS

Bookmark this page

Exercise: LMS and LLMS

2/2 points (graded)

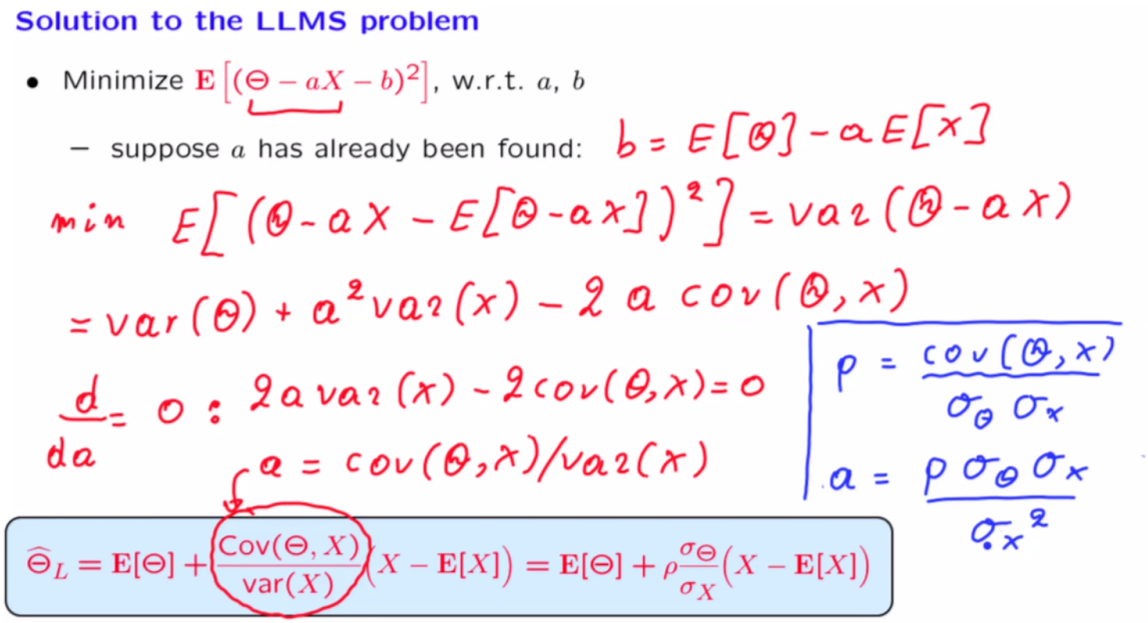
Suppose that the random variables Θ and X are not independent, but $\mathbf{E}\left[\Theta \mid X = x\right] = 3$ for all x. Then the LLMS estimator of Θ based on X is of the form $aX + b$, with

$a =$ 0 ✔ **Answer:** 0

$b =$ 3 ✔ **Answer:** 3

Solution:

The LMS estimator of Θ based on X is of the form $\mathbf{E}\left[\Theta \mid X\right] = 3$. This is already linear in X (with $a = 0$ and $b = 3$), and therefore it is also the LLMS estimator.



if we know a then that's how b should be chosen

5. Exercise: LLMS without a constant term

Bookmark this page

Exercise: LLMS without a constant term

2/2 points (graded)

Suppose that instead of estimators of the form $aX + e$, we consider estimators of the form $\widehat{\Theta} = aX$ and ask for the value of a that minimizes the mean squared error. Mimic the derivation you have just seen and find the optimal value of a. Your answer should be an algebraic expression involving some of the constants b, c, d, where $b = \mathbf{E}[\Theta^2]$, $c = \mathbf{E}[\Theta X]$, $d = \mathbf{E}[X^2]$.

c/d

✔ **Answer:** c/d

$\frac{c}{d}$

Solution:

The mean squared error is

$$\mathbf{E}\left[(\Theta - aX)^2\right] = \mathbf{E}[\Theta^2] - 2a\mathbf{E}[\Theta X] + a^2\mathbf{E}[X^2].$$

By setting to zero the derivative with respect to a, we find that

$$a = \frac{\mathbf{E}[\Theta X]}{\mathbf{E}[X^2]} = \frac{c}{d}.$$

Remarks on the solution and on the error variance

$$\widehat{\Theta}_L = \mathbf{E}[\Theta] + \frac{\mathrm{Cov}(\Theta, X)}{\mathrm{var}(X)}\big(X - \mathbf{E}[X]\big) = \mathbf{E}[\Theta] + \rho\frac{\sigma_\Theta}{\sigma_X}\big(X - \mathbf{E}[X]\big)$$

- Only means, variances, covariances matter
- $\rho > 0$: $X > E[X] \Rightarrow \hat{\Theta}_L > E[\Theta]$
- $\rho = 0$: $\hat{\Theta}_L = E[\Theta]$

$|\rho| = 1$ $\quad \hat{\Theta}_L = \Theta$

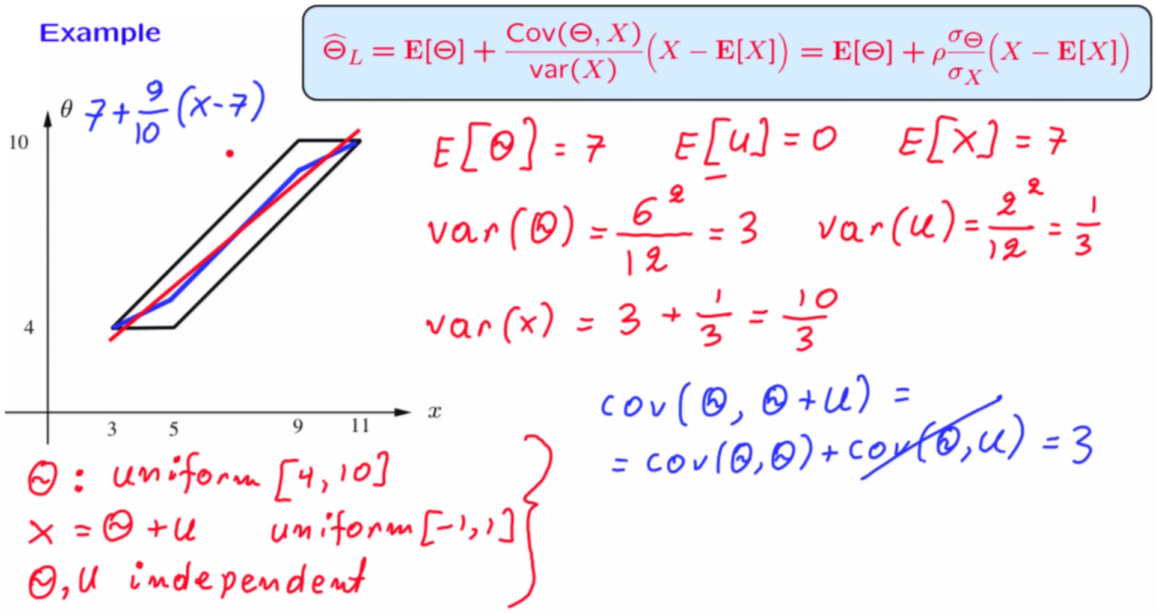
assume $E[\Theta] = E[X] = 0$

$$\mathbf{E}\left[(\widehat{\Theta}_L - \Theta)^2\right] = (1 - \rho^2)\,\mathrm{var}(\Theta)$$

$$E\left[\left(\Theta - \rho\frac{\sigma_\Theta}{\sigma_X}X\right)^2\right] = \sigma_\Theta^2 - 2\rho\frac{\sigma_\Theta}{\sigma_X}\rho\sigma_\Theta\sigma_X + \rho^2\frac{\sigma_\Theta^2}{\sigma_X^2}\sigma_X^2$$

(positive correlation, roe)
vice versa for negative correlation
bottom formula is the top but replacing expected values with 0

expected value of the theta squared is its standard deviation squared (variance)
same happens for x
the derived formula is still valid even if theta and x have non zero means
if the r.v.s are perfectly correlated then the error on the linear estimator becomes 0 and we can estimate the unknown variable using a linear function



cov(theta, theta) = var(theta)
cov(theta, u) = 0 as they are independent
the constant 9/10 indicates a positive correlation between theta and x but is 9/10 as the line is slightly slanted in comparison to the diagram

8. Exercise: LLMS drill

Bookmark this page

Exercise: LLMS drill

2/2 points (graded)

Suppose that Θ and W are independent, both with variance 1, and that $X = \Theta + W$. Furthermore, $\mathbf{E}[\Theta] = 1$ and $\mathbf{E}[W] = 2$. The LLMS estimator $\widehat{\Theta} = aX + b$ has

$a =$ 0.5 ✔ **Answer:** 0.5

$b =$ -0.5 ✔ **Answer:** -0.5

Hint: Remember the formula $\mathsf{Cov}(X + Y, Z) = \mathsf{Cov}(X, Z) + \mathsf{Cov}(Y, Z)$.

Solution:

We have $\mathbf{E}[X] = \mathbf{E}[\Theta] + \mathbf{E}[W] = 3$ and $\mathsf{Var}(X) = \mathsf{Var}(\Theta) + \mathsf{Var}(W) = 2$. Also,

$$\mathsf{Cov}(X, \Theta) = \mathsf{Cov}(\Theta, \Theta) + \mathsf{Cov}(\Theta, W) = \mathsf{Var}(\Theta) + 0 = 1.$$

Therefore, the LLMS estimator is

$$\widehat{\Theta} = 1 + \frac{1}{2}(X - 3) = \frac{1}{2}X - \frac{1}{2}.$$

9. Exercise: Possible values of the estimates

Bookmark this page

Exercise: Possible values of the estimates

1/2 points (graded)

Suppose that the random variable Θ takes values in the interval $[0, 1]$.

a) Is it true that the LMS estimator is guaranteed to take values only in the interval $[0, 1]$?

Yes ✔ **Answer:** Yes

b) Is it true that the LLMS estimator is guaranteed to take values only in the interval $[0, 1]$?

Yes ✖ **Answer:** No

Solution:

a) The conditional expectation $\mathbf{E}[\Theta \mid X = x]$ is a weighted average of the values of Θ, weighted according to the posterior PDF. A weighted average of values in $[0, 1]$ must lie in $[0, 1]$.

b) On the other hand, there is no such guarantee for the LLMS estimator. You can see this from the picture in the last example. Or you may consider the example where $X = \Theta + W$, where W can take any real value. Then, the term aX can take any real value, and can therefore fall outside the range $[0, 1]$.

LLMS for inferring the parameter of a coin

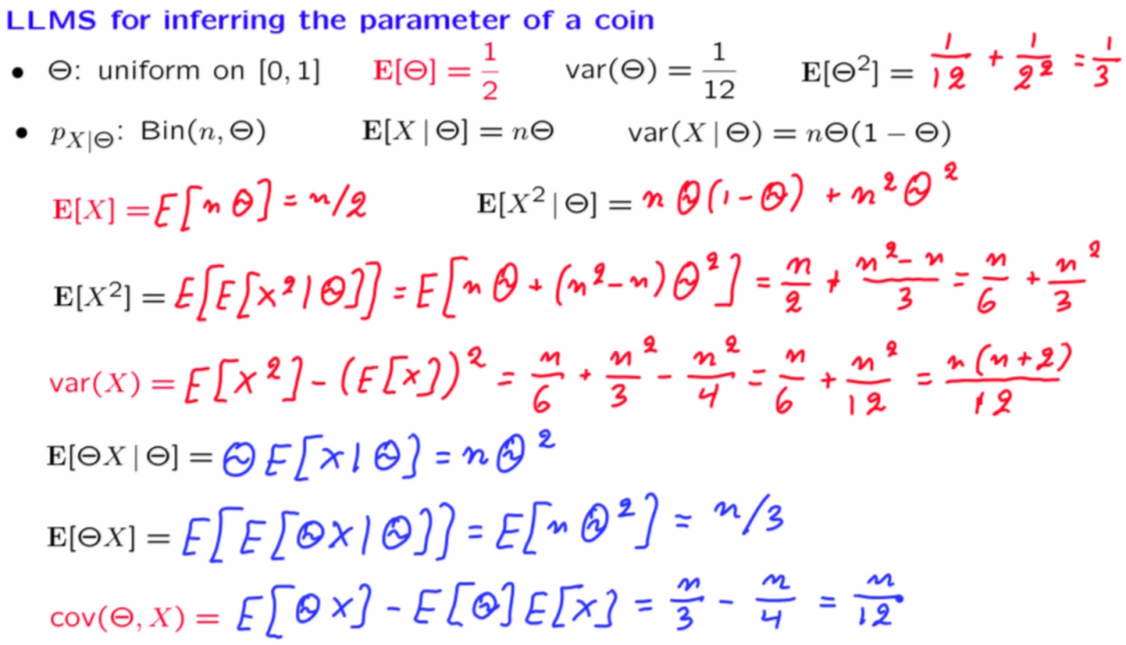
- Standard example:
  - coin with bias Θ; prior $f_\Theta(\cdot)$
  - fix n; X =number of heads
- Assume $f_\Theta(\cdot)$ is uniform in $[0, 1]$

$$\widehat{\Theta}_{\text{LMS}} = \frac{X+1}{n+2} = \widehat{\Theta}_{\text{LLMS}}$$

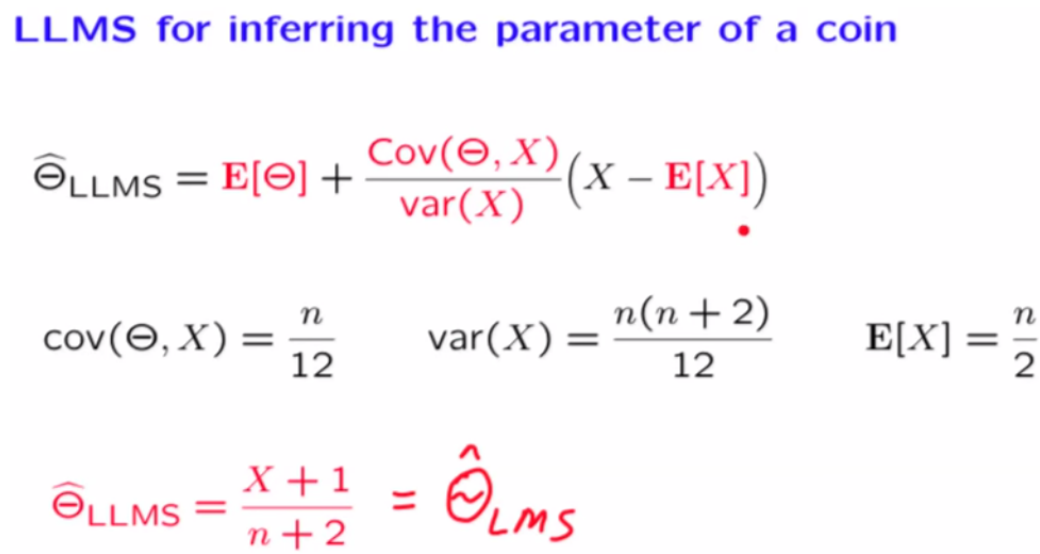
$$\widehat{\Theta}_{\text{LLMS}} = \mathbf{E}[\Theta] + \frac{\text{Cov}(\Theta, X)}{\text{var}(X)}\big(X - \mathbf{E}[X]\big)$$

If we know the estimator is linear in X then LMS estimator = LLMS estimator but going to go through the calculation for this anyway

good exercise for covering many concepts



(bin is binomial)
calculating all the different terms needed for the LLMS



Exercise: Comparison for the coin problem

1/1 point (graded)

Recall that the MAP estimator for the problem of estimating the bias of a coin is X/n, which is different from the LLMS estimator $(X+1)/(n+2)$. How do they compare in terms of mean squared error (MSE)?

- MAP has a smaller MSE.
- LLMS has a smaller MSE. ✔
- They have the same MSE.

Solution:

The LLMS estimator coincides with the LMS estimator and therefore achieves the smallest possible mean squared error.

LLMS with multiple observations

- Unknown Θ; observations $X = (X_1, \ldots, X_n)$
- Consider estimators of the form: $\widehat{\Theta} = a_1 X_1 + \cdots + a_n X_n + b$
- Find best choices of $a_1, \ldots, a_n, b$

  minimize: $\mathbf{E}\left[(a_1 X_1 + \cdots + a_n X_n + b - \Theta)^2\right]$ $= a_1^2 E[X_1^2] + 2a_1 a_2 E[X_1 X_2] + \cdots + a_1 E[X_1 \Theta] + \cdots$
- If $\mathbf{E}[\Theta \mid X]$ is linear in X, then $\widehat{\Theta}_{\text{LMS}} = \widehat{\Theta}_{\text{LLMS}}$
- Solve linear system in b and the a_i
- Only means, variances, covariances matter
- If multiple unknown Θ_j, apply to each one, separately

13. Exercise: LLMS with multiple observations

Bookmark this page

Exercise: LLMS with multiple observations

1/3 points (graded)

Suppose that Θ, X_1, and X_2 have zero means. Furthermore,

$$\text{Var}(X_1) = \text{Var}(X_2) = \text{Var}(\Theta) = 4,$$

and

$$\text{Cov}(\Theta, X_1) = \text{Cov}(\Theta, X_2) = \text{Cov}(X_1, X_2) = 1.$$

The LLMS estimator of Θ based on X_1 and X_2 is of the form $\widehat{\Theta} = a_1 X_1 + a_2 X_2 + b$. Find the coefficients a_1, a_2, and b. *Hint:* To find b, recall the argument we used for the case of a single observation.

$a_1 =$ 1/2 ✖ **Answer:** 0.2

$a_2 =$ 1/2 ✖ **Answer:** 0.2

$b =$ 0 ✔ **Answer:** 0

Solution:

By the same argument as in the case of a single observation, we will have $b = \mathbf{E}[\Theta - a_1 X_1 - a_2 X_2] = 0$. Using the variance and covariance information we are given, the expression we want to minimize is

$$\mathbf{E}\left[(a_1 X_1 + a_2 X_2 - \Theta)^2\right] = 4a_1^2 + 4a_2^2 + 4 + 2a_1 a_2 - 2a_1 - 2a_2.$$

Because of symmetry, we see that the optimal solution will satisfy $a_1 = a_2 = a$, so the expression is of the form $8a^2 + 4 + 2a^2 - 4a$. By setting the derivative to zero, we find that $20a = 4$, or $a = 1/5$.

E[x]^2 = var because E[x] = 0

$E[(a_1 X_1 + a_2 X_2 - \theta)^2]$

$a_1^2 E[X_1^2] + a_1 a_2 E[X_1 X_2] - a_1 E[X_1 \theta]$

$+ a_1 a_2 E[X_1 X_2] + a_2^2 E[X_2^2] - a_2 E[X_2 \theta]$

$- a_1 E[X_1 \theta] - a_2 E[X_2 \theta] + E[\theta^2]$

$var(X_1) = E[X^2] - (E[X])^2$

$E[X_1^2] = var = 4$

$4a_1^2 + 4a_2^2 + 2a_1 a_2 + 4 - 2a_1 - 2a_2$

$E[X_1 X_2]$ $E[X_1 \theta]$ $E[X_2 \theta]$

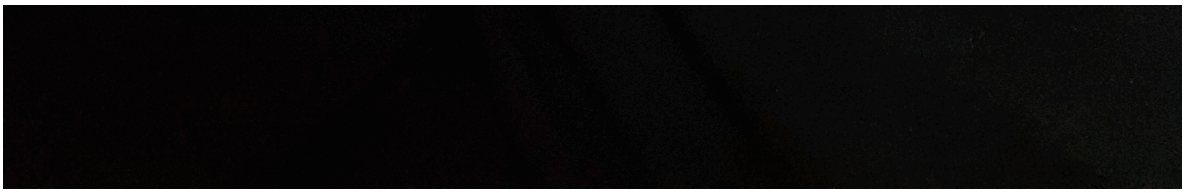
terms $= 1$

because

$cov(X,Y) = E[XY] - E[X]E[Y]$

↑ $= 0$ in this case

so $E[XY] = cov(X,Y) = 1$



The simplest LLMS example with multiple observations

$$X_1 = \Theta + W_1 \qquad \Theta \sim x_0,\ \sigma_0^2 \qquad W_i \sim 0,\ \sigma_i^2$$
$$\vdots$$
$$X_n = \Theta + W_n \qquad \Theta, W_1, \ldots, W_n \text{ uncorrelated}$$

- Suppose $\Theta, W_1, \ldots, W_n$ are independent normal

$$\hat{\theta}_{\text{LMS}} = \mathbf{E}[\Theta \mid X = x] = \frac{\sum_{i=0}^{n} \frac{x_i}{\sigma_i^2}}{\sum_{i=0}^{n} \frac{1}{\sigma_i^2}} \qquad \widehat{\Theta}_{\text{LMS}} = \mathbf{E}[\Theta \mid X] = \frac{\frac{x_0}{\sigma_0^2} + \sum_{i=1}^{n} \frac{X_i}{\sigma_i^2}}{\sum_{i=0}^{n} \frac{1}{\sigma_i^2}} = \widehat{\Theta}_{\text{LLMS}}$$

- Suppose general (not normal) distributions, but same means, variances, as in normal example
  - all covariances also the same
  - solution must be the same

don't need the variables to be independent, it's enough to assume that they're uncorrelated

The representation of the data matters in LLMS

- Estimation based on X versus X^3
  - LMS: $\mathbf{E}[\Theta \mid X]$ is the same as $\mathbf{E}[\Theta \mid X^3]$
  - LLMS is different: estimator $\widehat{\Theta} = aX + b$ versus $\widehat{\Theta} = aX^3 + b$

$$\text{cov}(\Theta, X^3) \quad \text{var}(X^3)$$

  - can also consider $\widehat{\Theta} = a_1 X + a_2 X^2 + a_3 X^3 + b$
  - can also consider $\widehat{\Theta} = a_1 X + a_2 e^X + a_3 \log X + b$

would be forming a different LLMS estimator
still considered a linear estimator as it's the factors of a and b that are important

can consider x^1 x^2 x^3 as different observations of x
all these combinations are possible but of varying complexity

16. Exercise: Choice of representations

Bookmark this page

Exercise: Choice of representations

0/1 point (graded)

We wish to estimate an unknown quantity Θ. Our measuring equipment produces an observation of the form $X = \Theta^3 + W$, where W is a noise term which is small relative to the range of Θ. Which type of linear estimator is preferable in such a situation?

- $\widehat{\Theta} = aX + b$
- $\widehat{\Theta} = aX^3 + b$
- $\widehat{\Theta} = aX^{1/3} + b$ ✔

✖

Solution:

If the noise W were completely absent, we would estimate Θ by letting $\widehat{\Theta} = X^{1/3}$. In the presence of small noise, our estimator should again have a similar form, which argues in favor of the third option.