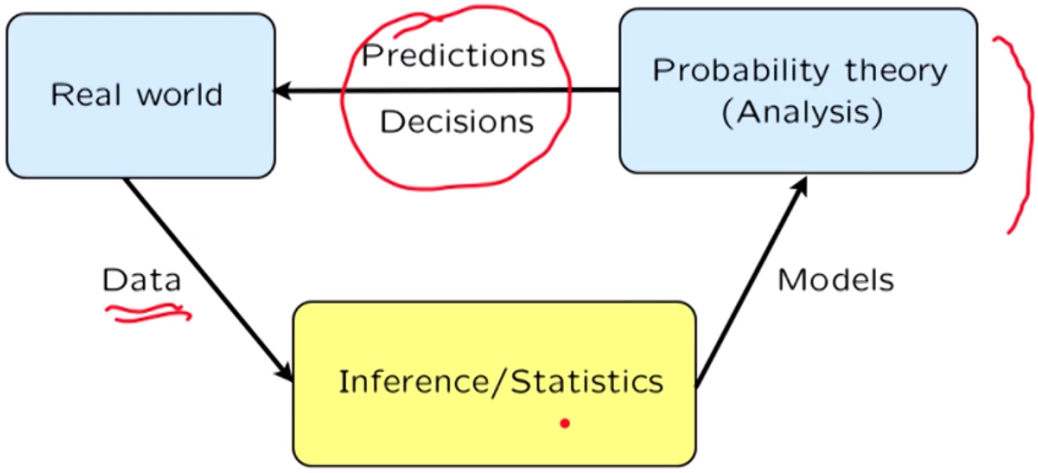
# Unit 7 Bayesian Inference

## Inference: the big picture



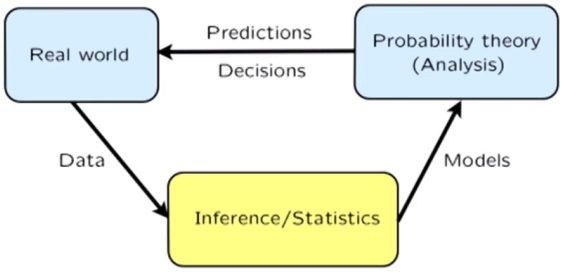

Use data to infer a model that we do analysis on

## 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 ✔

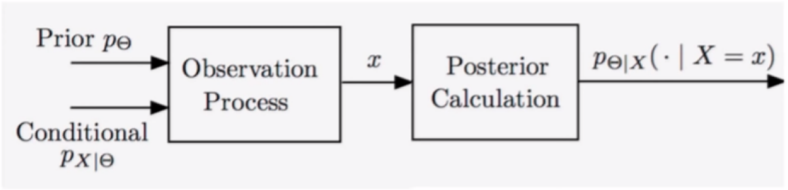
(d) Guess the weight of the robber.

Estimation ✔

## The Bayesian inference framework

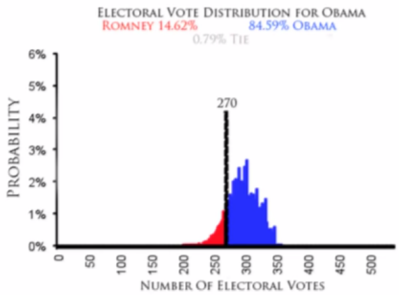
- Unknown $\Theta$
  - treated as a random variable
  - prior distribution $p_\Theta$ or $f_\Theta$
- Observation $X$
  - observation model $p_{X|\Theta}$ or $f_{X|\Theta}$
- Use appropriate version of the Bayes rule to find $p_{\Theta|X}(\cdot \mid X = x)$ or $f_{\Theta|X}(\cdot \mid X = x)$

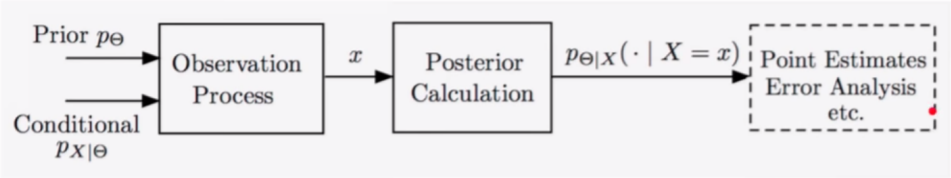
- Where does the prior come from?
  - symmetry
  - known range
  - earlier studies
  - subjective or arbitrary



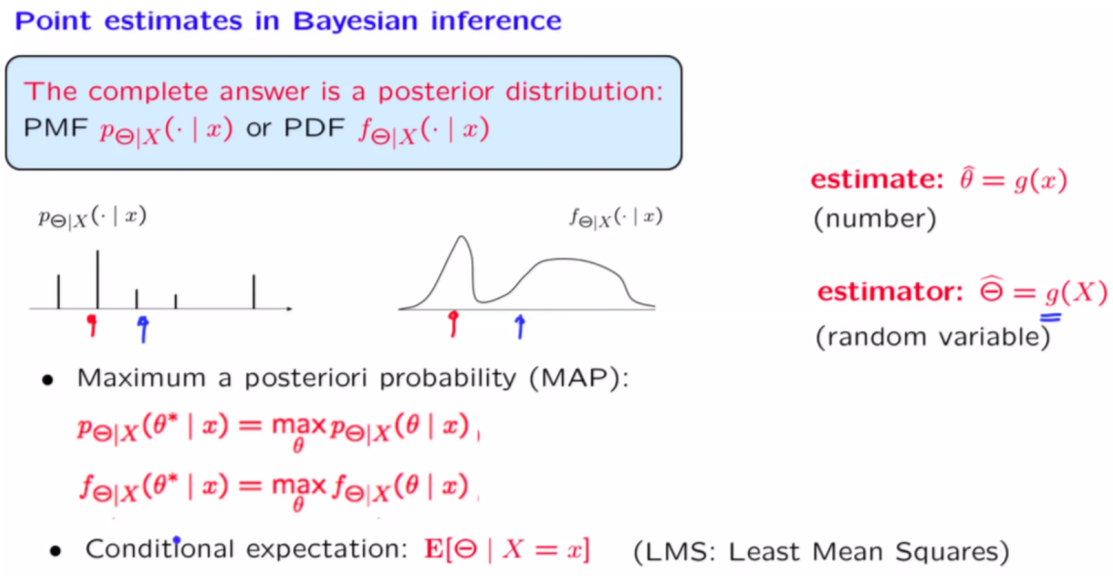

## The output of Bayesian inference

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 $\Theta$. 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 $\Theta$ 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.

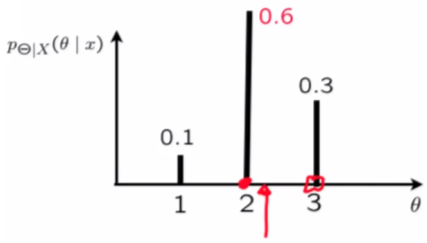
In the second case, we are dealing with the realized numerical value of an estimator, which we call an estimate.

**Discrete $\Theta$, discrete $X$**

- values of $\Theta$: 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 $\Theta_1$ and $\Theta_2$ be some unobserved Bernoulli random variables and let $X$ be an observation. Conditional on $X = x$, the posterior joint PMF of $\Theta_1$ and $\Theta_2$ 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 $(\Theta_1, \Theta_2)$.

a) What is the estimate of $(\Theta_1, \Theta_2)$ provided by the MAP rule?

(1,1) ✔ **Answer:** (1,1)

b) Once you calculate the estimate $(\hat{\theta}_1, \hat{\theta}_2)$ of $(\Theta_1, \Theta_2)$, you may report the first component, $\hat{\theta}_1$, as your estimate of $\Theta_1$. With this procedure, your estimate of $\Theta_1$ will be

1 ✔ **Answer:** 1

c) What is the probability that $\Theta_1$ 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 $\Theta_1$ 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 $\Theta_1$ 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 $\Theta_1$.

**Solution:**

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

b) The corresponding estimate of $\Theta_1$ 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 $\Theta_1$ 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).

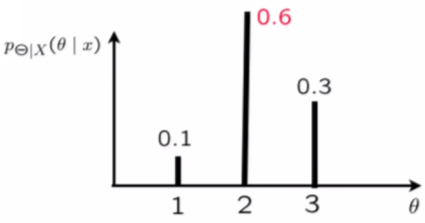
**Discrete $\Theta$, continuous $X$**

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

    $X = \Theta + W$

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

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




- 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 $\Theta$ is equally likely to take the values $-1$ and $1$, and where $W$ is standard normal noise, independent of $\Theta$. 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 $\Theta$ 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 $\Theta$ is $1$, that is, $\mathbf{P}(\widehat{\Theta} \neq 1 \mid \Theta = 1)$, is equal to

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

✖

where $\Phi$ is the standard normal CDF.

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

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

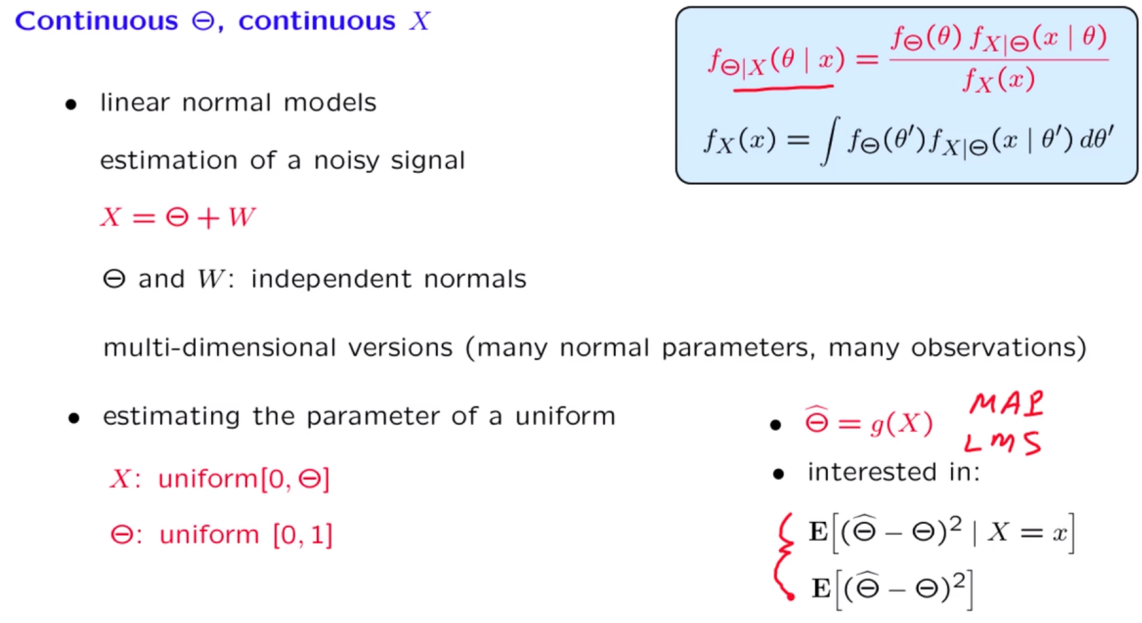
✔

**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 $\Theta$; 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 $\Theta$ be a continuous random variable that represents the unknown bias (i.e., the probability of Heads) of a coin.

a) The prior PDF $f_\Theta$ 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 $\Theta$ 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 $\Theta$ will be of the form $c\theta^m(1-\theta)^n$, where $c$ is a normalizing constant and where

$m =$ 10 ✔ **Answer:** 10

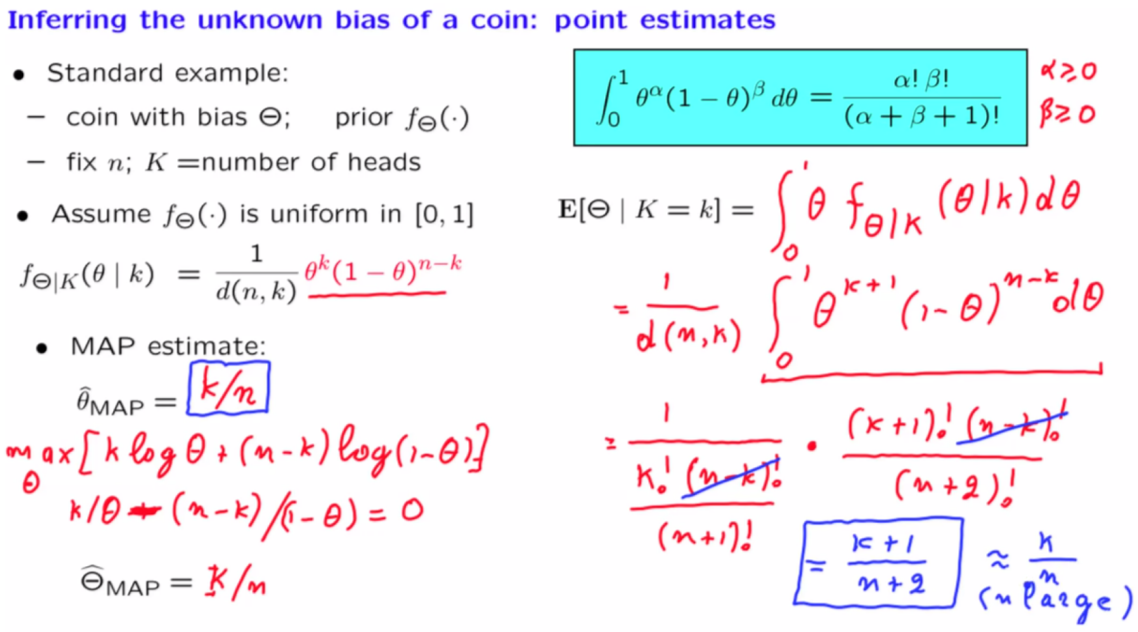
$n =$ 10 ✔ **Answer:** 10

**Solution:**

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

b) As we saw in the last video, the power to which $\theta$ (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 $\Theta$ 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.

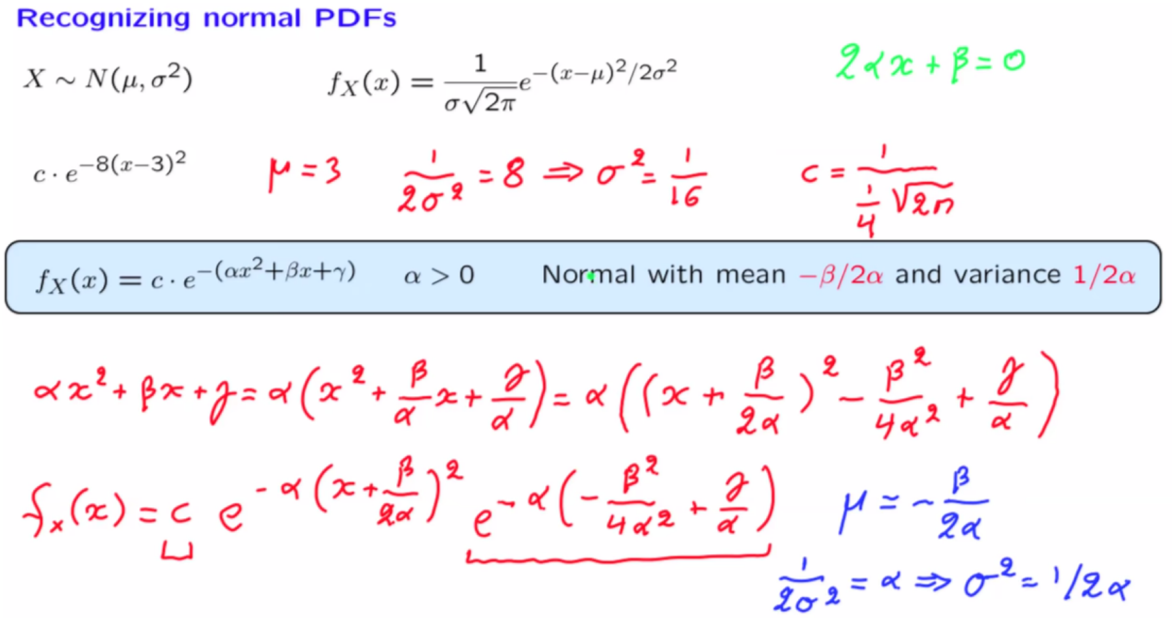
Given a prior and some observed X
Want to get the posterior distribution as a full answer to inference problem
This is the estimator
Applying a given value of X = x, we can then get an estimate from the estimator

**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

$X_i = \sum_{j=1}^{m} aij\ \Theta_j + W_i$ $W_i$, $\Theta_j$ : 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) 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 $\alpha$ 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 $\Theta$ 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 $\Theta$ 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 $\Theta$ 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 $\Theta$ and $W$ are standard normal random variables. Find the MAP estimator of $\Theta$ 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 $\Theta$ 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 $\theta$, 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 $\theta$ 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 $\theta$, 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 $\theta$ 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 $2\theta$ and variance $9$. If we focus on the terms that involve $\theta$, 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 $\theta$ 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

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

$\sigma_i^2$ large
$X_i$ very noisy
$\Rightarrow$ 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 $\Theta$ 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 $\Theta$?

Yes ✔ **Answer:** Yes

**Solution:**

a) The $X_i$'s are dependent because they are all affected by $\Theta$. 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, $\theta$ 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 $\Theta, W_1, W_2$ 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 $\Theta$ 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 $\theta$) 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 $\sigma_i^2$ small $\rightarrow$ MSE small

all $\sigma_i^2$ large $\rightarrow$ MSE large.

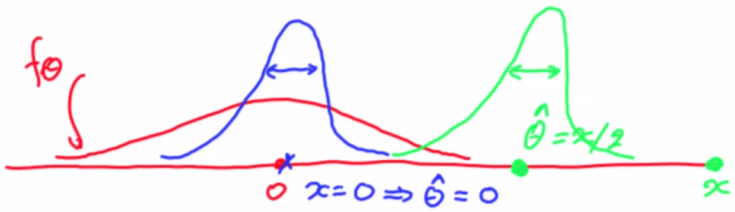
So the expected value having not observed anything is the same as if we had observed something (X=x)
Also when the variance is small the MSE is small
when it is large the MSE is large

## 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 $\Theta, W$ $\quad \widehat{\Theta} = X/2$ $\quad \mathbf{E}\left[(\Theta - \widehat{\Theta})^2 \mid X = x\right] = 1/2$




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 $\Theta$.

In the second scenario, the prior information on $\Theta$ is extremely inaccurate: $\Theta \sim N(0, \sigma_0^2)$, where $\sigma_0^2$ 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 $\Theta$.

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 $\Theta$ 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 $\Theta$ and $W$, the mean squared error of the LMS estimator is

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

where $\sigma_0^2$ and $\sigma_1^2$ are the variances of $\Theta$ and $W$, respectively.

Suppose now that we change the observation model to $Y = 3\Theta + W$. In some sense the "signal" $\Theta$ 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

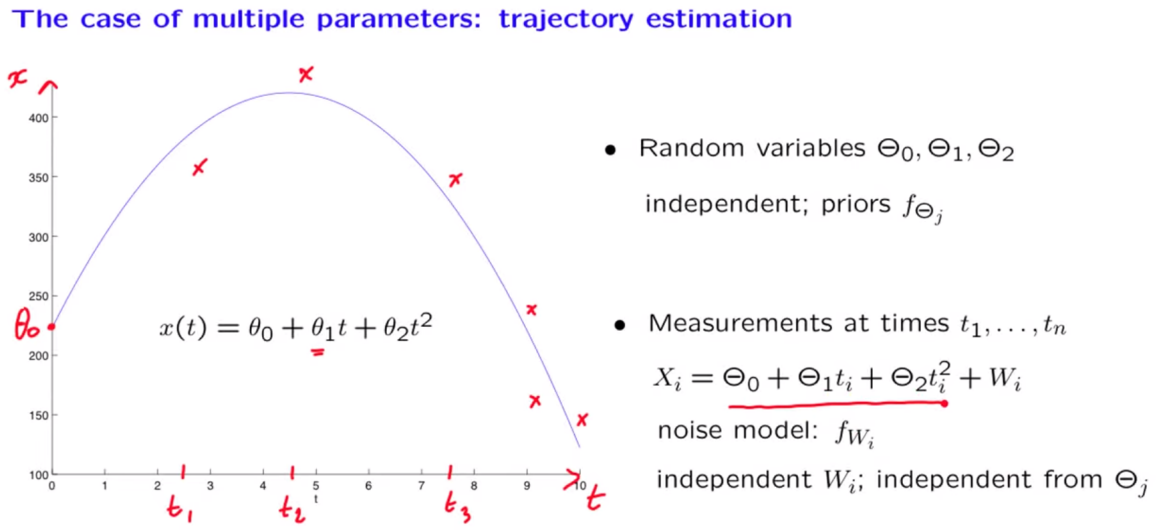
1/4 ✖ **Answer:** 0.1

*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 $\Theta_1$, $\Theta_2$, $W_1$, 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 $(\Theta_1, \Theta_2)$ 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 $\theta_1$, $X_2$ 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 $\theta_1$ and $\theta_2$) 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

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

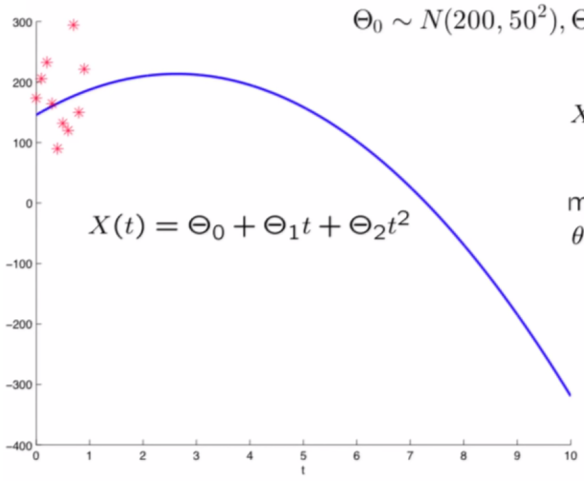
Either by inspection, or by substitution, we obtain the solution $\hat{\theta}_1 = 1, \hat{\theta}_2 = 1$.

## Linear normal models.

- $\Theta_j$ 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 $\Theta_j$: $f_{\Theta_j|X}(\theta_j \mid x)$, 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



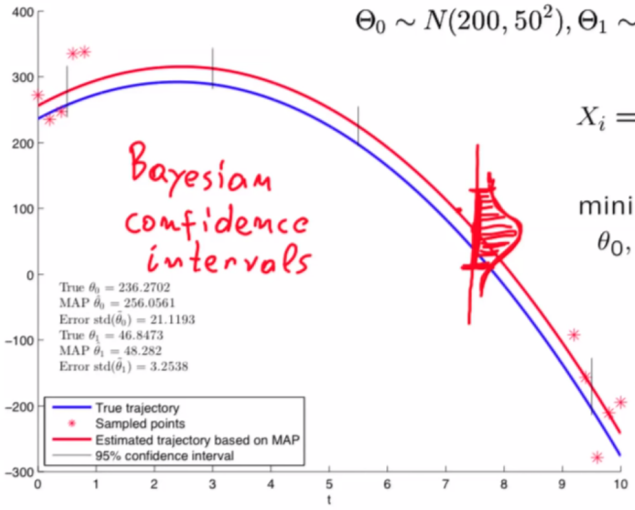

$$\Theta_0 \sim N(200, 50^2), \Theta_1 \sim N(50, 50^2), \Theta_2 = -9.81, W_i \sim N(0, 50^2)$$

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

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

Taking theta2 as a constant now so its prior disappears

**An illustration** Estimating the trajectory of a free-falling object




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

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

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

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