CSE 312

Foundations of Computing II

Lecture 19: Maximum Likelihood Estimation

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Aleks Jovcic

Slide Credit: Based on Stefano Tessaro's slides for 312 19au
incorporating ideas from Anna Karlin, Alex Tsun, Rachel Lin, Hunter Schafer & myself ☺

Final Pset

- Slightly longer, slightly harder, less time to work
- Released Tuesday, August 16th at 11:59pm PST
- Due Friday, August 19th at 11:59pm PST
  - **No late days can be spent!**
  - If something comes up, please let me know as soon as possible
- Individual, but working and studying together is encouraged
  - **No office hours during this time**
  - Prepare and go to office hours ahead of time
  - There will be a form to find classmates to work with as needed
  - Remember that you do not need to typeset if that will take up too much time
- TA-led review session, in-class review, TBA

Agenda

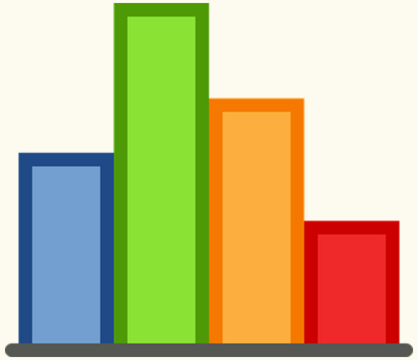
- Idea: Estimation
- Maximum Likelihood Estimation
- MLE with continuous random variables
- General Steps

Probability vs statistics

$Ber(p = 0.5)$

Probability
given model, predict data

$P(THHTHH)$

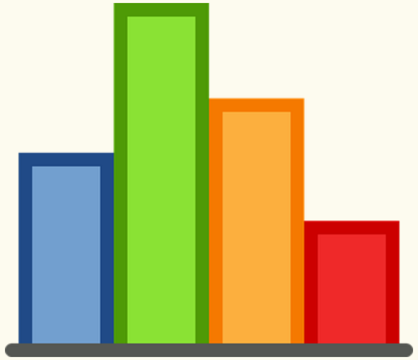


Probability vs statistics

$Ber(p = 0.5)$

Probability
given model, predict data

$P(THHTHH)$



$Ber(p = ???)$

Statistics
given data, predict model

$THHTHH$

Probability: Viewpoint up to Now

Distribution
$\mathbb{P}(x; \theta)$ $\rightarrow$ Independent samples $x_1, \ldots, x_n$ from $\mathbb{P}(x\,; \theta)$

θ = <u>known</u> parameter

θ tells us how samples are distributed.
$\mathbb{P}(x\,; \theta)$ viewed as a function of x (fixed θ)

Statistics: Parameter Estimation – Workflow

Parameter estimate

Distribution $\mathbb{P}(x;\theta)$ → Independent samples $x_1, \dots, x_n$ from $\mathbb{P}(x;\theta)$ → **Algorithm** → $\hat{\theta}$

$\hat{\theta}$

θ = unknown parameter

Don’t know how samples are distributed.

Example

$\sim Ber(\theta)$

Suppose we have a mystery coin with some probability p of coming up heads. We flip the coin 8 times, independent of other flips and see the following sequence. of flips

$$TTHTHTTH$$

Given this data, what would you estimate p is?

$\underbrace{\frac{3}{8}}$ vs. $\frac{4}{8}$

vs. $\frac{5}{8}$

Poll:

a. 1/2
b. 5/8
c. 3/8
d. 1/4

Agenda

- Idea: Estimation
- Maximum Likelihood Estimation
- MLE with continuous random variables
- General Steps

Likelihood

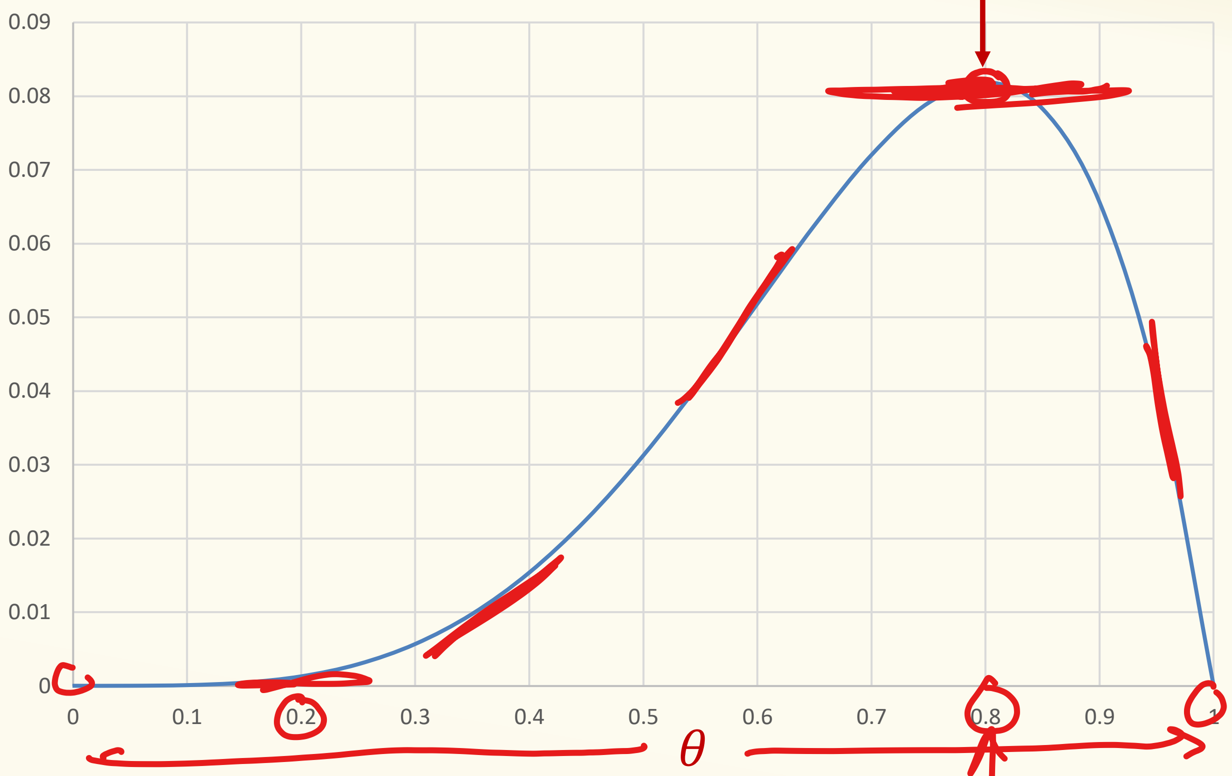
Say we see outcome HHTHH.

You tell me your best guess about the value of the unknown parameter θ (aka p) is 4/5. Is there some way that you can argue "objectively" that this is the best estimate?

Likelihood

$l(x \mid \theta)$

Max Prob of seeing HHTHH



$x_1\ x_2\ x_3$

Say we see outcome HHTHH.

You tell me your best guess about the value of the unknown parameter θ (aka p) is 4/5. Is there some way that you can argue "objectively" that this is the best estimate?

$$\mathcal{L}(HHTHH \mid \theta) = \theta^4(1-\theta)$$

$\theta \cdot \theta \cdot (1-\theta) \cdot \theta \cdot \theta$

θ

$\hat{\theta} = 0.8$

Likelihood of Different Observations

(Discrete case)

$\mathcal{L}$

Definition. The **likelihood** of independent observations $x_1, \dots, x_n$ is

$$\mathcal{L}(x_1, \dots, x_n | \theta) = \prod_{i=1}^{n} \mathbb{P}(x_i; \theta)$$

$= \mathbb{P}(x_1 | \theta) \cdot \mathbb{P}(x_2; \theta) \cdots \cdot \mathbb{P}(x_n; \theta)$

Likelihood of Different Observations

(Discrete case)

Definition. The **likelihood** of independent observations $x_1, \dots, x_n$ is

$$\mathcal{L}(x_1, \dots, x_n | \theta) = \prod_{i=1}^{n} \mathbb{P}(x_i; \theta)$$

Maximum Likelihood Estimation (MLE). Given data $x_1, \dots, x_n$, find $\hat{\theta}$ ("the MLE") of model such that $L\left(x_1, \dots, x_n \middle| \hat{\theta}\right)$ is maximized!

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \, \mathcal{L}(x_1, \dots, x_n | \theta)$$

Likelihood vs. Probability

A **probability function** $\Pr(x\,;\theta)$ is a function with input being an event x for some fixed probability model (w/ param θ).

$$\sum_x \Pr(x\,;\theta) = 1$$

A **likelihood function** $\mathcal{L}(x\,|\theta)$ is a function with input being θ (the param of the prob. Model) for some fixed dataset x.

These notions are very closely connected, but answer different questions. We are trying to find the θ that maximizes likelihood, thus we are looking for the **maximum likelihood estimator.**

Example – Coin Flips

Observe: Coin-flip outcomes $x_1, \ldots, x_n$, with n_H heads, n_T tails

– I.e., $n_H + n_T = n$

Goal: estimate θ = prob. heads.

$= \prod_{i=1}^{n} \mathbb{P}(x_i; \theta) = \prod_{i=1}^{n} \ldots = \theta \cdot (1-\theta)$

$$L(x_1, \ldots, x_n | \theta) = \theta^{n_H}(1-\theta)^{n_T}$$

$$\frac{\partial}{\partial \theta} L(x_1, \ldots, x_n | \theta) = ???$$

While it is not difficult to compute this derivative, we make our lives easier by observing that we are always taking a derivative of a product….

Log-Likelihood

We can save some work if we work with the **log-likelihood** instead of the likelihood directly.

Definition. The **log-likelihood** of independent observations $x_1, \ldots, x_n$ is

$$\mathcal{LL}(x_1, \ldots, x_n|\theta) = \ln \mathcal{L}(x_1, \ldots, x_n|\theta)$$

$$= \ln \prod_{i=1}^{n} \mathbb{P}(x_i; \theta) = \sum_{i=1}^{n} \ln \mathbb{P}(x_i; \theta)$$

Useful log properties

$$\log(ab) = \log(a) + \log(b)$$
$$\log(a/b) = \log(a) - \log(b)$$
$$\log\left(a^b\right) = b\log(a)$$

Example – Coin Flips

Observe: Coin-flip outcomes $x_1, \ldots, x_n$, with n_H heads, n_T tails

– I.e., $n_H + n_T = n$

Goal: estimate θ = prob. heads.

$$\mathcal{L}(x_1, \ldots., x_n | \theta) = \theta^{n_H}(1-\theta)^{n_T}$$

$$\ln \mathcal{L}(x_1, \ldots., x_n | \theta) =$$

Example – Coin Flips

Observe: Coin-flip outcomes $x_1, \ldots, x_n$, with n_H heads, n_T tails

– I.e., $n_H + n_T = n$

Goal: estimate θ = prob. heads.

$$\mathcal{L}(x_1, \ldots, x_n | \theta) = \theta^{n_H}(1-\theta)^{n_T}$$

$$\ln \mathcal{L}(x_1, \ldots, x_n | \theta) = n_H \ln \theta + n_T \ln(1-\theta)$$

$$\frac{\partial}{\partial \theta} \ln \mathcal{L}(x_1, \ldots, x_n | \theta) = n_H \cdot \frac{1}{\theta} - n_T \cdot \frac{1}{1-\theta}$$

Solve $n_H \cdot \frac{1}{\hat{\theta}} - n_T \cdot \frac{1}{1-\hat{\theta}} = 0$

$$\boxed{\hat{\theta} = \frac{n_H}{n}}$$

Agenda

- Idea: Estimation
- Maximum Likelihood Estimation
- MLE with continuous random variables
- General Steps

The Continuous Case

Given n samples $x_1, \ldots, x_n$ from a Gaussian $\mathcal{N}(\mu, \sigma^2)$, estimate $\theta = (\mu, \sigma^2)$

Definition. The **likelihood** of independent observations $x_1, \ldots, x_n$ is

$$\mathcal{L}(x_1, \ldots, x_n | \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

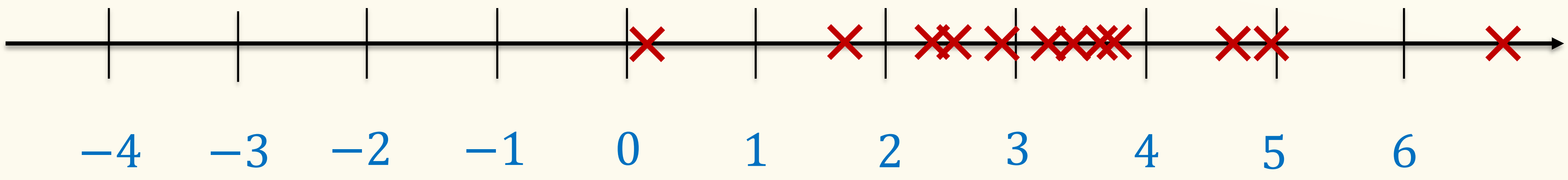
Density function! (Why?)

Why density?

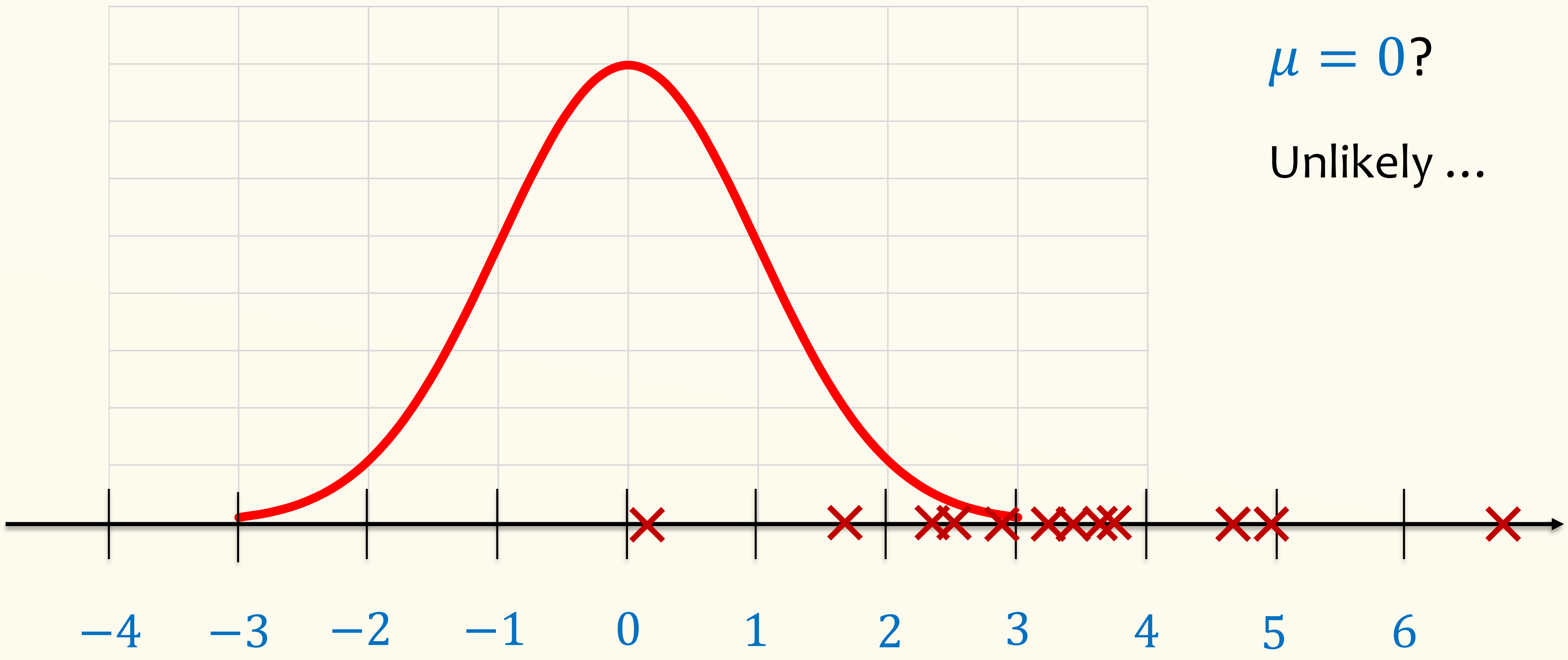
- Density $\neq$ probability, but:
  - For maximizing likelihood, we really only care about relative likelihoods, and density captures that
  - has desired property that likelihood increases with better fit to the model

n samples $x_1, \dots, x_n \in \mathbb{R}$ from Gaussian $\mathcal{N}(\mu, 1)$. Most likely μ?

[i.e., we are given the promise that the variance is one]



n samples $x_1, \dots, x_n \in \mathbb{R}$ from Gaussian $\mathcal{N}(\mu, 1)$. Most likely μ?



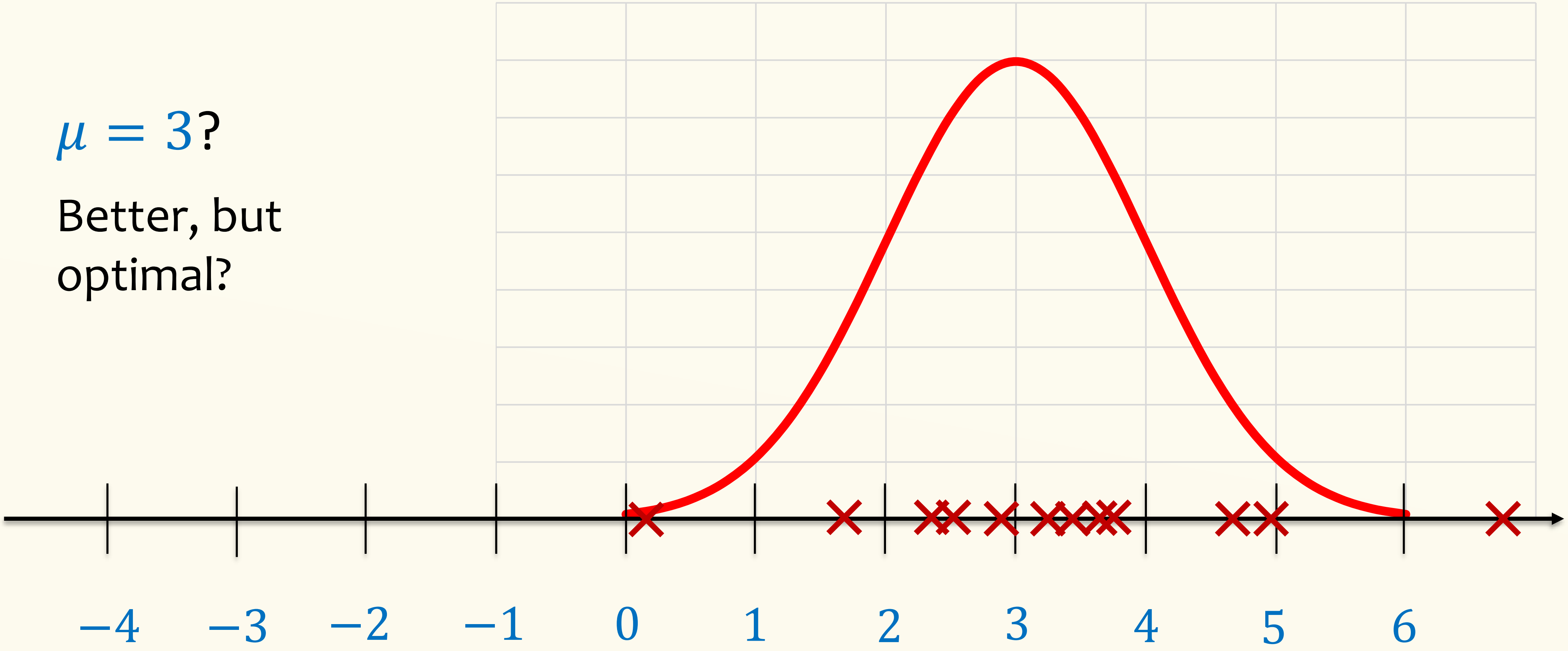
$\mu = 0$?

Unlikely …

n samples $x_1, \dots, x_n \in \mathbb{R}$ from Gaussian $\mathcal{N}(\mu, 1)$. Most likely μ?

$\mu = 3$?

Better, but optimal?



Example – Gaussian Parameters

Normal outcomes $x_1, \ldots, x_n$, known variance $\sigma^2 = 1$

Goal: estimate θ expectation

$$\mathcal{L}(x_1, \ldots, x_n | \theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\theta)^2}{2}} =$$

$$\log(ab) = \log(a) + \log(b)$$
$$\log(a/b) = \log(a) - \log(b)$$
$$\log\left(a^b\right) = b\log(a)$$

Example – Gaussian Parameters

Normal outcomes $x_1, \ldots, x_n$, known variance $\sigma^2 = 1$

Goal: estimate θ expectation

$$\mathcal{L}(x_1, \ldots, x_n | \theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\theta)^2}{2}} = \left(\frac{1}{\sqrt{2\pi}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\theta)^2}{2}}$$

$$\ln \mathcal{L}(x_1, \ldots, x_n | \theta) = -n \frac{\ln 2\pi}{2} - \sum_{i=1}^{n} \frac{(x_i - \theta)^2}{2}$$

Example – Gaussian Parameters

Goal: estimate θ= expectation

Normal outcomes $x_1, \dots, x_n$, known variance $\sigma^2 = 1$

$$\ln \mathcal{L}(x_1, \dots, x_n | \theta) = -n \frac{\ln 2\pi}{2} - \sum_{i=1}^{n} \frac{(x_i - \theta)^2}{2}$$

Example – Gaussian Parameters

Goal: estimate θ= expectation

Normal outcomes $x_1, \ldots, x_n$, known variance $\sigma^2 = 1$

$$\ln \mathcal{L}(x_1, \ldots., x_n|\theta) = -n\frac{\ln 2\pi}{2} - \sum_{i=1}^{n} \frac{(x_i-\theta)^2}{2}$$

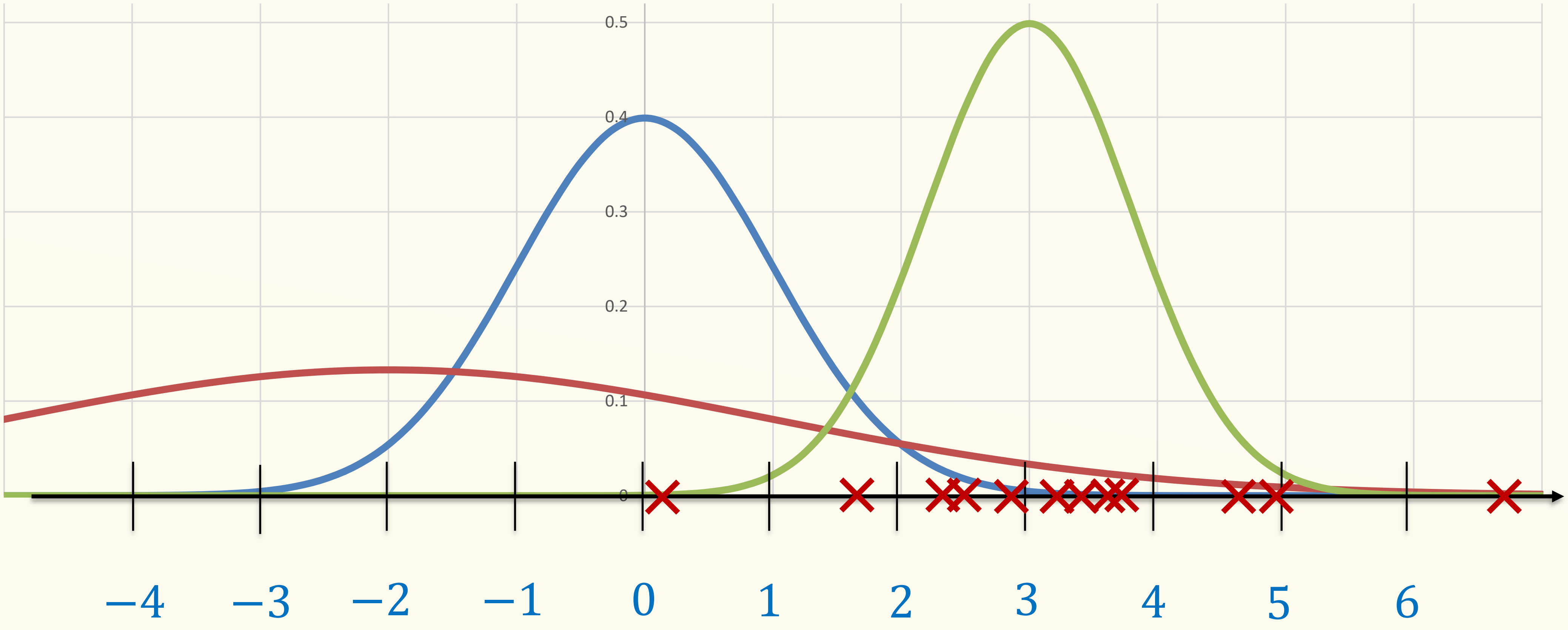
Note: $\frac{\partial}{\partial\theta}\frac{(x_i-\theta)^2}{2} = \frac{1}{2}\cdot 2 \cdot (x_i - \theta)\cdot(-1) = \theta - x_i$

$$\frac{\partial}{\partial\theta}\ln \mathcal{L}(x_1, \ldots., x_n|\theta) = \sum_{i=1}^{n}(x_i - \theta) = \sum_{i=1}^{n} x_i - n\theta = 0$$

$$\hat{\theta} = \frac{\sum_i^n x_i}{n}$$

In other words, MLE is the *sample mean* of the data.

Next steps: n samples $x_1, \ldots, x_n \in \mathbb{R}$ from Gaussian $\mathcal{N}(\mu, \sigma^2)$. Most likely μ and σ^2?



Agenda

- Idea: Estimation
- Maximum Likelihood Estimation
- MLE with continuous random variables
- General Steps

General Recipe

1. **Input** Given n iid samples $x_1, \dots, x_n$ from parametric model with parameters θ.
2. **Likelihood** Define your likelihood $\mathcal{L}(x_1, \dots., x_n | \theta)$.
   - For discrete $\mathcal{L}(x_1, \dots., x_n | \theta) = \prod_{i=1}^{n} \Pr(x_i ; \theta)$
   - For continuous $\mathcal{L}(x_1, \dots., x_n | \theta) = \prod_{i=1}^{n} f(x_i ; \theta)$
3. **Log** Compute $\ln \mathcal{L}(x_1, \dots., x_n | \theta)$
4. **Differentiate** Compute $\frac{\partial}{\partial \theta} \ln \mathcal{L}(x_1, \dots., x_n | \theta)$
5. **Solve for** $\hat{\theta}$ by setting derivative to 0 and solving for max.

Generally, you need to do a second derivative test to verify it is a maximum, but we won't ask you to do that in CSE 312.