CSE 312

Foundations of Computing II

Lecture 22: Loose Ends and Maximum Likelihood Estimation

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Anna R. Karlin

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

Feedback

- I'm going to fast for some of you.
  - I'll pause more to give you a chance to ask questions.
  - You ask more questions.
  - Read the section or watch videos before class.
  - Come to next class with questions about previous class.
- "Examples in class are too complex. "
  - I can't seem to please all of the people all of the time!
- Which material is in the book?
  - Pretty much everything.
- Grades/quizzes/etc – don't worry!
- "There needs to be more commenting on the python code to explain new syntax, like for calling objects from other classes."
  - Please send a message on edstem pointing out places where you think more comments are needed and we can try to add some. [For both past and future psets]

Law of Total Probability and Law of Total Expectation

Law of Total Probability. Let E be an event and let Y be a discrete random variable that takes values $\{1,2,\ldots,n\}$. Then,

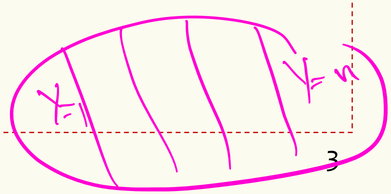
$$\Pr[E] = \sum_{i=1}^{n} Pr[E|Y = i]\Pr(Y = i)$$

Law of Total Expectation. Let X be a random variable and let Y be a discrete random variable that takes values $\{1,2,\ldots,n\}$. Then,

$$E[X] = \sum_{i=1}^{n} E[X|Y = i]\Pr(Y = i)$$

$$\sum_{x \in \Omega_X} x \Pr(X = x \mid Y = i)$$

cond pmf of X cond Y=i



Law of Total Probability

Law of Total Probability (discrete). Let E be an event and let Y be a discrete random variable that takes values $\{1,2,\ldots,n\}$. Then,

$$\Pr[E] = \sum_{i=1}^{n} Pr[E|Y = i]\Pr(Y = i)$$

Law of Total Probability (cont). Let E be an event and let Y be a continuous random variable. Then,

$$\Pr[E] = \int_{-\infty}^{+\infty} Pr[E|Y = y]\, f_Y(y)\, \mathrm{d}y$$

$Y \in [y, y+dy]$

$X \sim \text{Poisson}(10)$ $\quad Pr(X=2) = e^{-10}\frac{10^2}{2!}$

Example: Number of accidents a random person has in a year is Poisson(Y) where Y itself is a random variable. What is the probability that a random person has two accidents? $\quad \lambda e^{-\lambda x}$

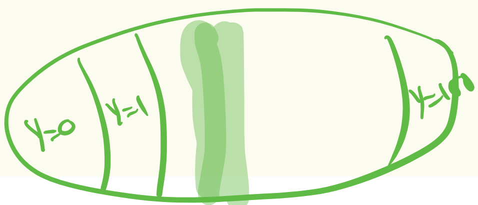
Discrete example:

Y is Binomial $(100, 0.3)$.

$\Omega_Y = \{0, 1, \ldots, 100\}$

$$Pr(X=2) = \sum_{k=0}^{100} Pr(X=2 \mid Y=k)\, Pr(Y=k)$$

$$= \sum_{k=0}^{100} e^{-k}\frac{k^2}{2!} \binom{100}{k} 0.3^k\, 0.7^{100-k}$$



Continuous example:

Y is exponential with parameter 1

Ω_Y

$$Pr(X=2) = \int_0^\infty Pr(X=2 \mid Y=y)\, f_Y(y)\, dy$$

$$= \int_0^\infty e^{-y}\frac{y^2}{2!} e^{-y}\, dy$$

Law of Total Expectation

Law of Total Expectation (discrete. Let X be a random variable and Y be a discrete random variable that takes values $\{1,2,\ldots,n\}$. Then,

$$E[X] = \sum_{i=1}^{n} E[X|Y = i]\Pr(Y = i)$$

Law of Total Expectation (cont). Let X be a random variable and let Y be a continuous random variable. Then,

$$E[X] = \int_{-\infty}^{+\infty} E[X|Y = y] f_Y(y)\, \mathrm{d}y$$

Example:

X is discrete uniform on $\{0, \ldots 10\}$.
Y is discrete uniform on $\{0, \ldots X\}$.
What is $\mathrm{E}(Y)$?

Example:

X is continuous uniform on $(0,10)$. Y is continuous uniform on $(0, X)$. What is $\mathrm{E}(Y)$?

Agenda

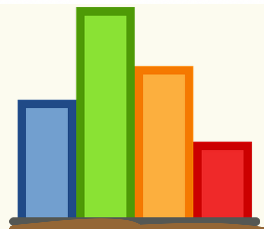
- Idea: Estimation
- Maximum Likelihood Estimation (example: mystery coin)
- Continuous random variables
- General Steps

Probability vs statistics

$Ber(p = 0.5)$

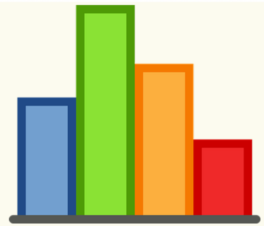
Probability
given model, predict data

$P(THHTHH)$



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

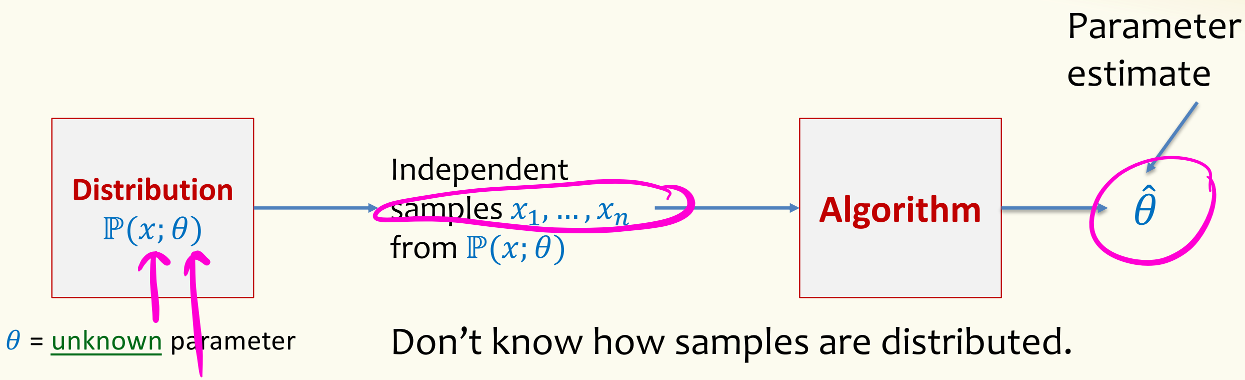
Distribution
$\mathbb{P}(x;\theta)$

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

θ = known parameter

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

Statistics: Parameter Estimation – Workflow



Don’t know how samples are distributed.

Statistics: Parameter Estimation – Workflow

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

Parameter estimate

θ = unknown parameter

Don't know how samples are distributed.

$\mathcal{L}(x|\theta)$ viewed as a function of θ (fixed x)

Example: $\mathcal{L}(x|\theta)$ = coin flip distribution with unknown θ = probability of heads

Observation: HTTHHHTHTHTTTTHTHTTTTTTHT

Goal: Estimate θ from data

Example

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

$$TTHTHTTH$$

3 H's / 8 total

Given this data, what would you estimate p is?

Poll: https://pollev.com/ annakarlin185

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

Agenda

- Idea: Estimation
- Maximum Likelihood Estimation (example: mystery coin)
- 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?

For what param value is outcome HHTHH most likely to happen

$P(H;\theta)\ P(H;\theta)\ P(T;\theta)\ P(H;\theta)\ P(H;\theta)$

$\mathcal{L}(HHTHH;\theta) = \theta \cdot \theta (1-\theta) \theta \cdot \theta$

$= \theta^4 (1-\theta)$

Find θ that maximizes this fn.

$\frac{d}{d\theta}(\theta^4 - \theta^5) = 4\theta^3 - 5\theta^4 = \theta^3(4 - 5\theta)$

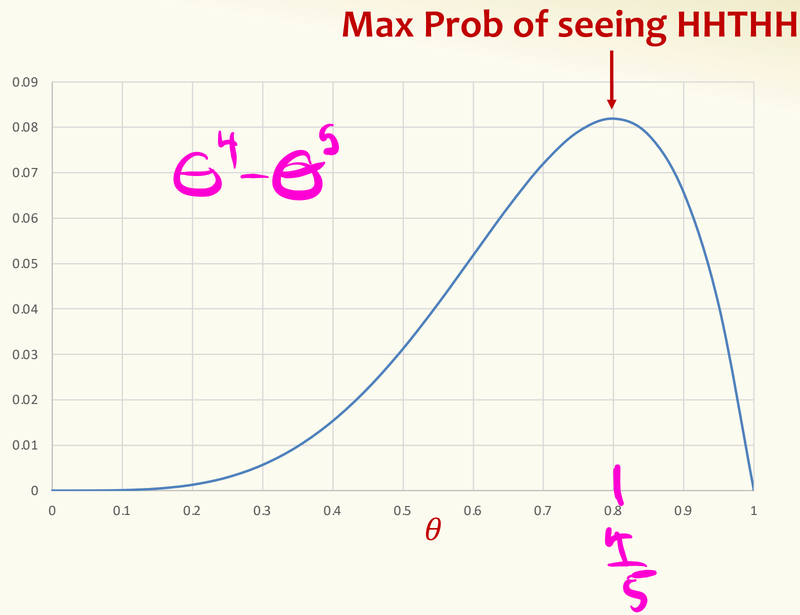
$\theta^3 (4-5\theta) = 0$

$\theta = \frac{4}{5}$

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?



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

Likelihood fn: prob of seeing this outcome if the param was θ

Likelihood of Different Observations

(Discrete case)

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

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

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

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

Usually: Solve $\frac{\partial L(x_1, \ldots, x_n | \theta)}{\partial \theta} = 0$ or $\frac{\partial \ln L(x_1, \ldots, x_n | \theta)}{\partial \theta} = 0$ [+check it's a max!]

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 $x_1, \dots, x_n$, with n_H heads, n_T tails

– I.e., $n_H + n_T = n$

Goal: estimate θ = prob. heads.

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

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

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

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

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 $x_1, \dots, x_n$, with n_H heads, n_T tails

– I.e., $n_H + n_T = n$

Goal: estimate θ = prob. heads.

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

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

Example – Coin Flips

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

– I.e., $n_H + n_T = n$

Goal: estimate θ = prob. heads.

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

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

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

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

Brain Break

Agenda

- Idea: Estimation
- Maximum Likelihood Estimation (example: mystery coin)
- 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

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

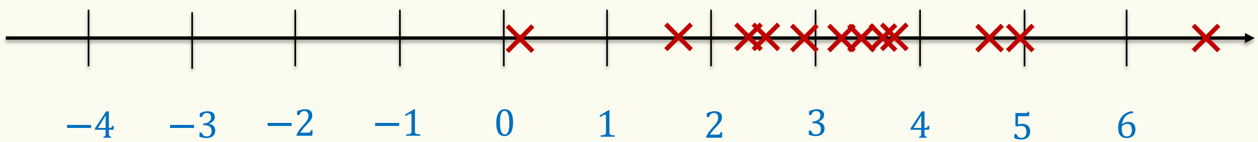
Density function! (Why?)

Why density?

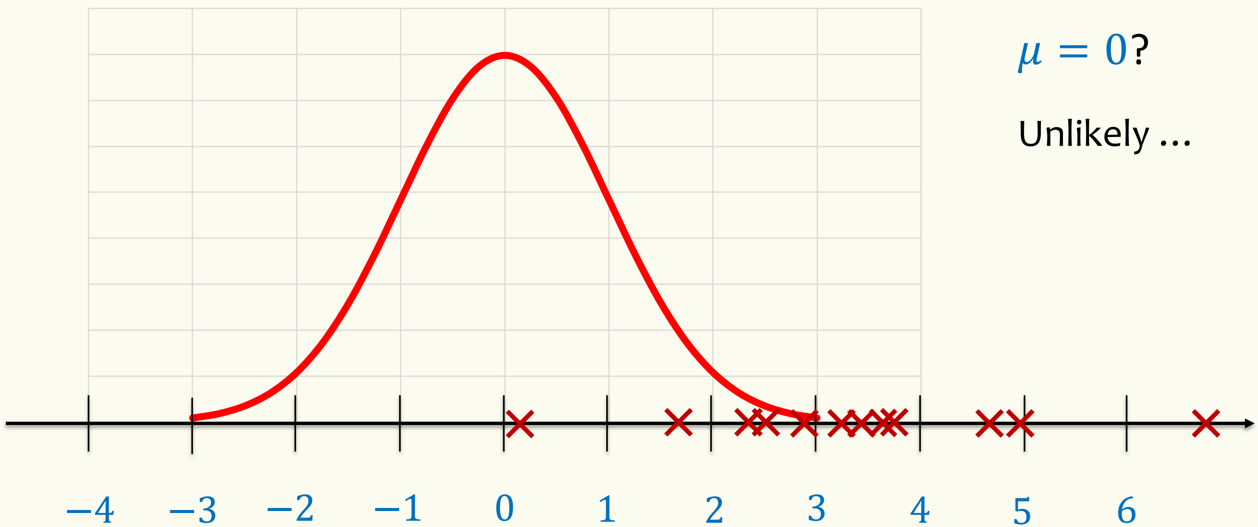
- Density $\neq$ probability, but:
  - For maximizing likelihood, we really only care about relative likelihoods, and density captures that
  - has desired property that likelihood increases with better fit to the model

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



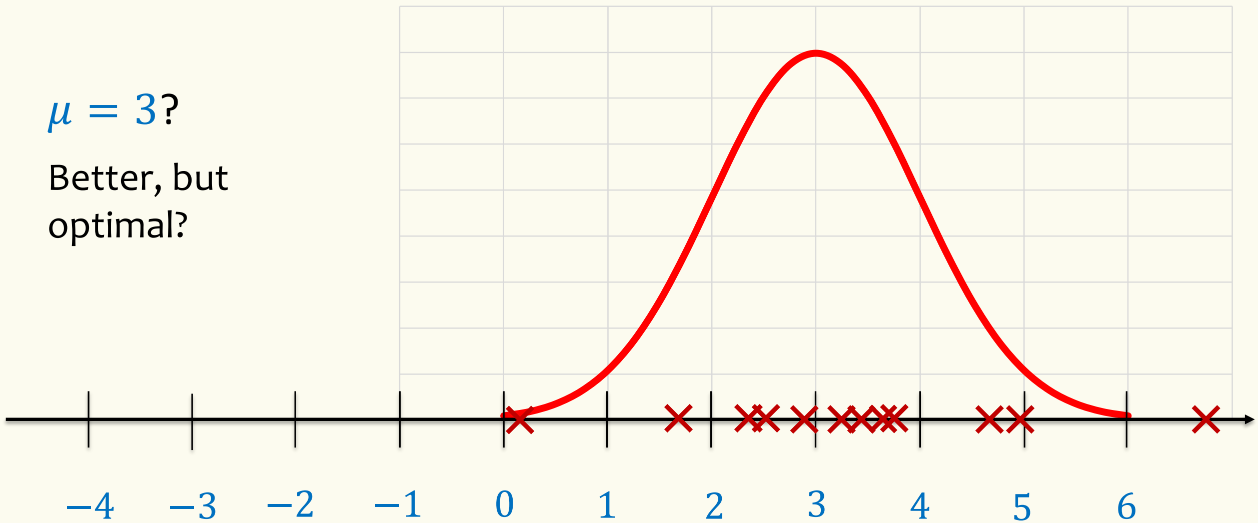
$\mu = 0$?

Unlikely …

n samples $x_1, \ldots, x_n \in \mathbb{R}$ 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

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

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

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

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

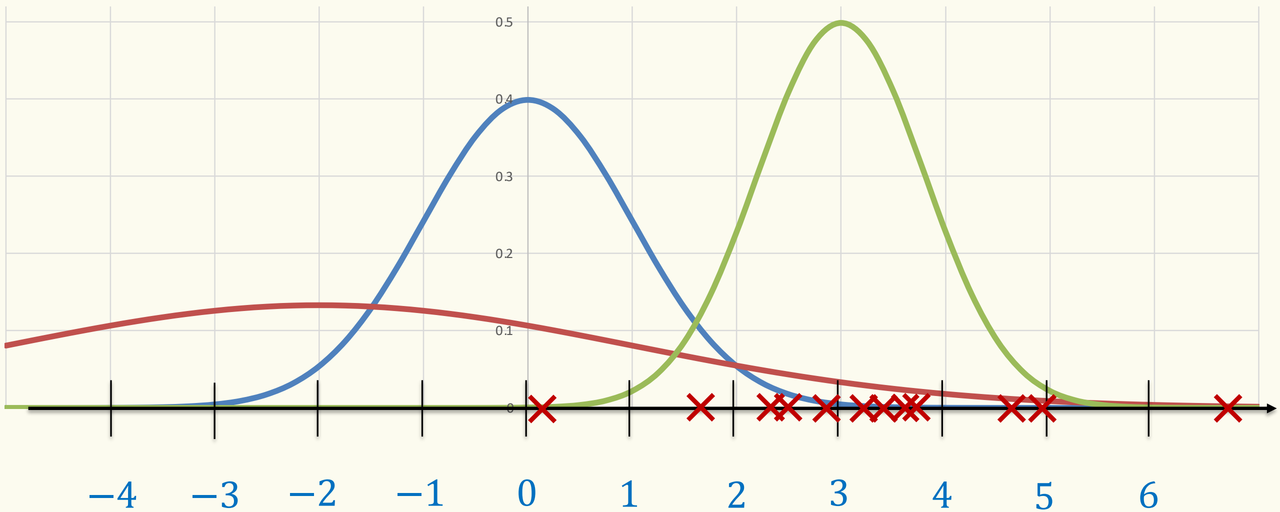
Note: $\frac{\partial}{\partial \theta} \frac{(x_i - \theta)^2}{2} = \frac{1}{2} \cdot 2 \cdot (x_i - \theta) \cdot (-1) = \theta - x_i$

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

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

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

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



Agenda

- Idea: Estimation
- Maximum Likelihood Estimation (example: mystery coin)
- Continuous random variables
- General Steps

General Recipe

1. **Input** Given n iid samples $x_1, \ldots, x_n$ from parametric model with parameters θ.
2. **Likelihood** Define your likelihood $\mathcal{L}(x_1, \ldots., x_n|\theta)$.
   - For discrete $\mathcal{L}(x_1, \ldots., x_n|\theta) = \prod_{i=1}^{n} \Pr(x_i\,; \theta)$
   - For continuous $\mathcal{L}(x_1, \ldots., x_n|\theta) = \prod_{i=1}^{n} f(x_i\,; \theta)$
3. **Log** Compute $\ln \mathcal{L}(x_1, \ldots., x_n|\theta)$
4. **Differentiate** Compute $\frac{\partial}{\partial\theta} \ln \mathcal{L}(x_1, \ldots., x_n|\theta)$
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.

Another example of continuous law of total probability

X and Y are independent, where X has CDF $\mathrm{F_X}(x)$ and Y has pdf $\mathrm{f_Y}(y)$. What is $\mathrm{P}(X > 5Y)$?

Law of Total Probability (cont). Let E be an event and let Y be a continuous random variable. Then,

$$\Pr[E] = \int_{-\infty}^{+\infty} Pr[E|Y = y]\, f_Y(y)\, \mathrm{d}y$$