21: Parameters and MLE

David Varodayan
January 26, 2020
Adapted from slides by Lisa Yan

Rejection sampling algorithm

Review

Inference question: What is $P(F_{lu} = 1 | U = 1, T = 1)$?

```
def rejection_sampling(event, observation):
    samples = sample_a_ton()
    samples_observation =
        reject_inconsistent(samples, observation)
    samples_event =
        reject_inconsistent(samples_observation, event)
    return len(samples_event)/len(samples_observation)
```

[flu, und, fev, tir]

```
Sampling...
[0, 1, 0, 1]
[0, 1, 0, 1]
[0, 1, 0, 1]
[0, 0, 0, 0]
[0, 1, 0, 1]
[0, 1, 1, 1]
[0, 1, 0, 0]
[1, 1, 1, 1]
[0, 0, 1, 1]
...
[0, 1, 0, 1]
Finished sampling
```

Rejection sampling

If you can sample enough from the joint distribution, you can answer any probability inference question.

With enough samples, you can correctly compute:

- Probability estimates
- Conditional probability estimates
- Expectation estimates

Because your samples are a representation of the joint distribution!

[flu, und, fev, tir]

```
Sampling...
[0, 1, 0, 1]
[0, 1, 0, 1]
[0, 1, 0, 1]
[0, 0, 0, 0]
[0, 1, 0, 1]
[0, 1, 1, 1]
[0, 1, 0, 0]
[1, 1, 1, 1]
[0, 0, 1, 1]
...
[0, 1, 0, 1]
Finished sampling
```

P(has flu | undergrad and is tired) = 0.122

Disadvantages of rejection sampling

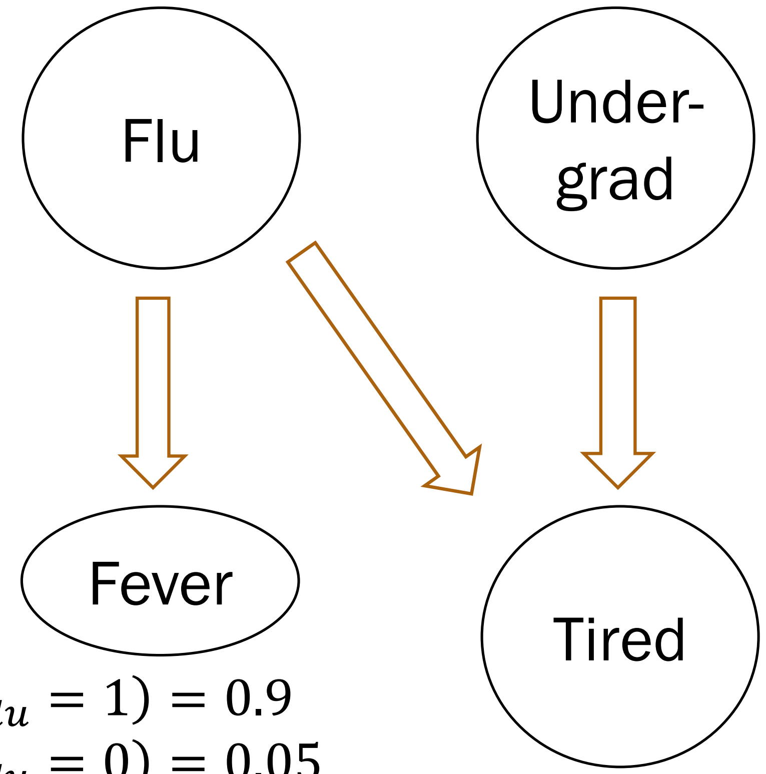
$$P(F_{lu} = 1 | F_{ev} = 1)?$$

What if we never encounter some samples?

[flu=0, und, fev=1, tir]

$P(F_{lu} = 1) = 0.1$

$P(U = 1) = 0.8$



$P(F_{ev} = 1 | F_{lu} = 1) = 0.9$
$P(F_{ev} = 1 | F_{lu} = 0) = 0.05$

$P(T = 1 | F_{lu} = 0, U = 0) = 0.1$
$P(T = 1 | F_{lu} = 0, U = 1) = 0.8$
$P(T = 1 | F_{lu} = 1, U = 0) = 0.9$
$P(T = 1 | F_{lu} = 1, U = 1) = 1.0$

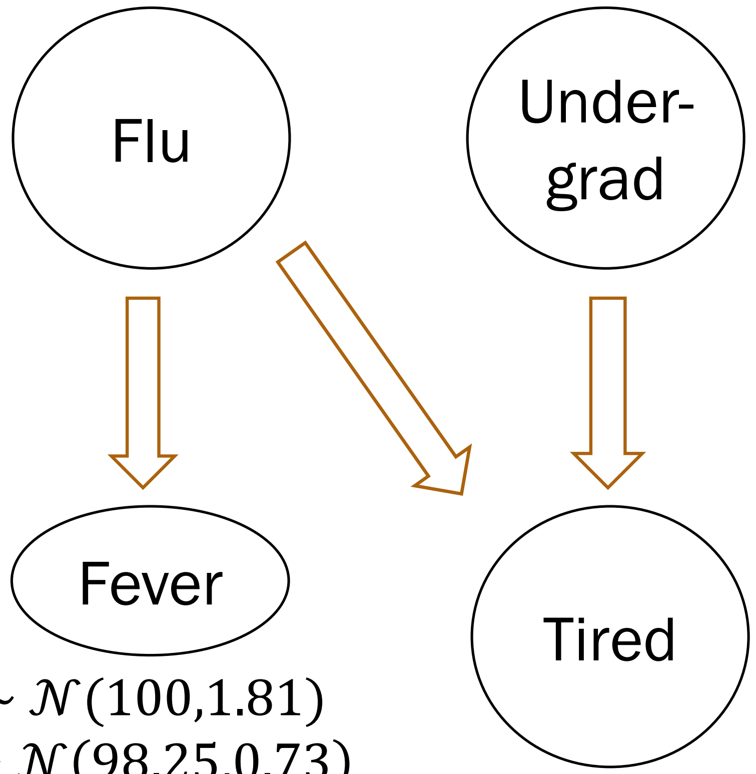
Disadvantages of rejection sampling

$$P(F_{lu} = 1 | F_{ev} = 99.4)?$$

What if we never encounter some samples?

What if random variables are continuous?

$P(F_{lu} = 1) = 0.1$ $\quad$ $P(U = 1) = 0.8$



$F_{ev} | F_{lu} = 1 \sim \mathcal{N}(100, 1.81)$

$F_{ev} | F_{lu} = 0 \sim \mathcal{N}(98.25, 0.73)$

$P(T = 1 | F_{lu} = 0, U = 0) = 0.1$

$P(T = 1 | F_{lu} = 0, U = 1) = 0.8$

$P(T = 1 | F_{lu} = 1, U = 0) = 0.9$

$P(T = 1 | F_{lu} = 1, U = 1) = 1.0$

Gibbs Sampling (not covered)

Basic idea:

- Fix all observed events
- Incrementally sample a new value for each random variable
- Difficulty: More coding for computing different posterior probabilities

Learn in extra notebook!

(or by taking CS228/CS238)

Announcements

Problem Set 5

Due: Friday 2/28
Covers: Up to Lecture 19

Late Day Reminder

No late days permitted past last day of the quarter, 3/13

Autograded Coding Problems

Run your code in the command line, not just in a Jupyter notebook cell

CS109 Contest

Due: Monday 3/9 11:59pm

Today's plan

Inference:

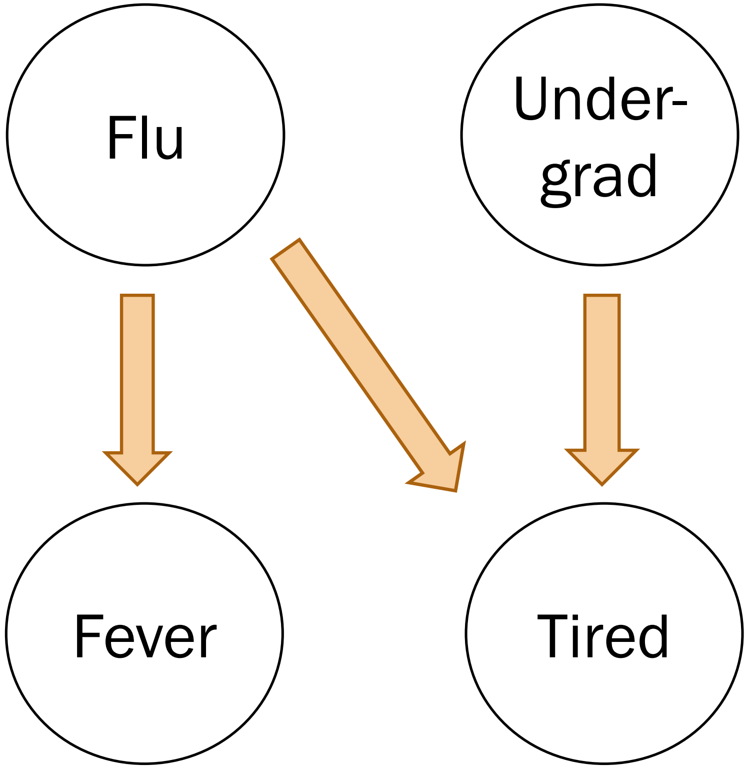
1. Math
2. Rejection sampling ("joint" sampling)
3. Optional: Gibbs sampling (MCMC algorithm) (extra notebook)

Intro to Parameter Estimation

Maximum Likelihood Estimation (MLE)

Where do the numbers come from?

$P(F_{lu} = 1) = 0.1$ $\quad$ $P(U = 1) = 0.8$



$P(F_{ev} = 1 | F_{lu} = 1) = 0.9$
$P(F_{ev} = 1 | F_{lu} = 0) = 0.05$

$P(T = 1 | F_{lu} = 0, U = 0) = 0.1$
$P(T = 1 | F_{lu} = 0, U = 1) = 0.8$
$P(T = 1 | F_{lu} = 1, U = 0) = 0.9$
$P(T = 1 | F_{lu} = 1, U = 1) = 1.0$

Given experiment data, how do we come up with a reasonable probabilistic **model**?

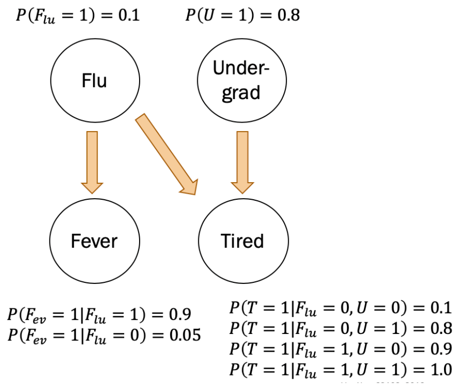
Story so far

At this point:

If you are given a **model** with all the necessary probabilities, you can make predictions.

$Y \sim \text{Poi}(5)$

$X_1, \dots, X_n$ i.i.d.
$X \sim \text{Ber}(0.2)$,
$X = \sum_{i=1}^{n} X_i$

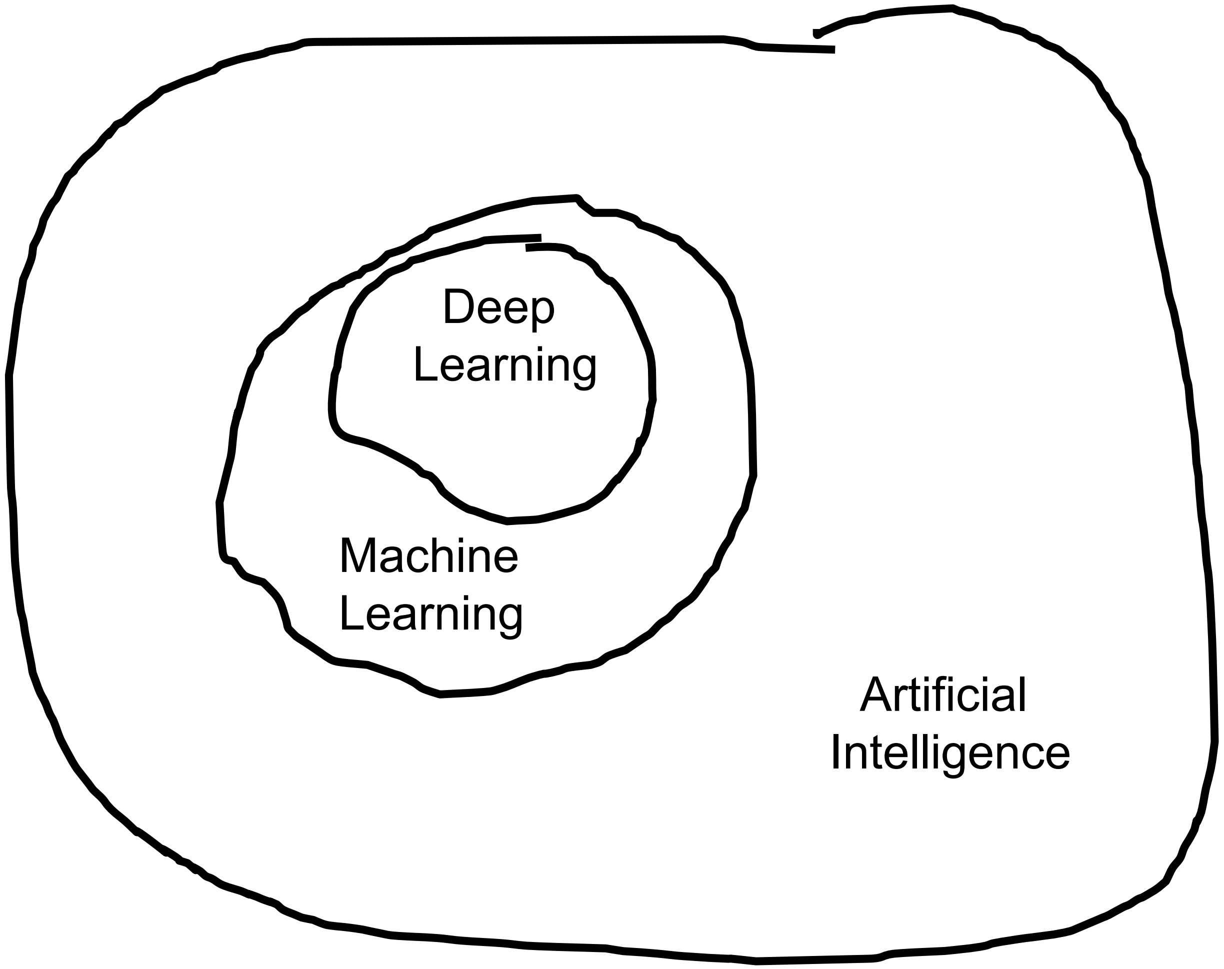


But what if you want to **learn** the probabilities in the model?

What if you want to learn the **structure** of the model, too?

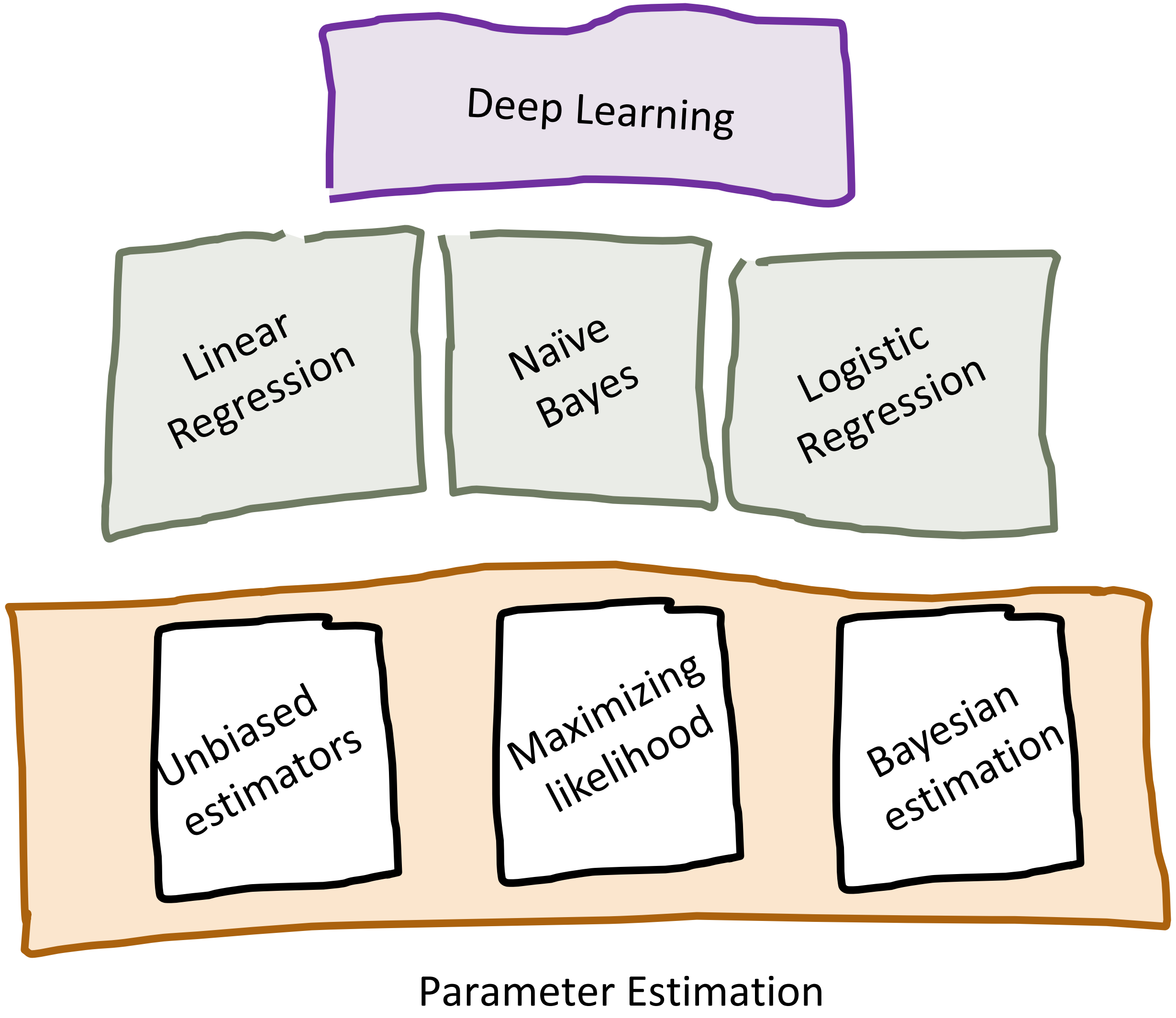
Machine Learning

AI and Machine Learning



ML: Rooted in probability theory

Our path from here



- Understand the theory to help you debug.
- Understand the theory to push on the grander challenges.

What are parameters?

def Many random variables we have learned so far are **parametric models**:

$$\text{Distribution} = \text{model} + \text{parameter } \theta$$

ex The distribution Ber(0.2) = Bernoulli model, parameter $\theta = 0.2$.

For each of the distributions below, what is the parameter θ?

1. Ber(p) $\theta = p$
2. Poi(λ)
3. Uni(α, β)
4. $\mathcal{N}(\mu, \sigma^2)$
5. $Y = mX + b$

What are parameters?

<u>def</u> Many random variables we have learned so far are **parametric models**:

$$\text{Distribution} = \text{model} + \text{parameter } \theta$$

<u>ex</u> The distribution $\text{Ber}(0.2)$ = Bernoulli model, parameter $\theta = 0.2$.

For each of the distributions below, what is the parameter θ?

1. $\text{Ber}(p)$ $\quad \theta = p$
2. $\text{Poi}(\lambda)$ $\quad \theta = \lambda$
3. $\text{Uni}(\alpha, \beta)$ $\quad \theta = (\alpha, \beta)$
4. $\mathcal{N}(\mu, \sigma^2)$ $\quad \theta = (\mu, \sigma^2)$
5. $Y = mX + b$ $\quad \theta = (m, b)$

θ is the parameter of a distribution.
θ can be a vector of parameters!

Why do we care?

In real world, we don't know the "true" parameters.

- But we do get to **observe data**: (# times coin comes up heads, lifetimes of disk drives produced, # visitors to website per day, etc.)

def **estimator** $\hat{\theta}$: random variable estimating parameter θ from data.

In parameter estimation,

We use the **point estimate** of parameter estimate (best single value):

- Better understanding of the process producing data
- Future **predictions** based on model
- Simulation of future processes

Today's plan

Inference:

1. Math
2. Rejection sampling ("joint" sampling)
3. Optional: Gibbs sampling (MCMC algorithm)

Intro to Parameter Estimation

Maximum Likelihood Estimation (MLE)

Recall some estimators

Consider n i.i.d. random variables $X_1, X_2, \ldots, X_n$.

- The sequence $X_1, X_2, \ldots, X_n$ is a **sample** from distribution F.
- X_i have distribution F with $E[X_i] = \mu, \text{Var}(X_i) = \sigma^2$.

Sample mean:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

unbiased **estimate** of μ

$$E[\bar{X}] = \mu$$

Sample variance:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})^2$$

unbiased estimate of σ^2

$$E[S^2] = \sigma^2$$

Estimating a Bernoulli parameter

Consider n i.i.d. random variables $X_1, X_2, \dots, X_n$.

- The sequence $X_1, X_2, \dots, X_n$ is a **sample** from distribution F.
- X_i have distribution F with $E[X_i] = \mu, \text{Var}(X_i) = \sigma^2$.

- Suppose distribution $F = \text{Ber}(\theta)$ with unknown parameter θ.
- Say you have three estimates $\hat{\theta}$: $\hat{\theta} = 0.5$, $\hat{\theta} = 0.8$, or $\hat{\theta} = 1$

Which estimate is most likely to give you the following sample ($n = 10$)?

$$[0, 0, 1, 1, 1, 1, 1, 1, 1, 1]$$

Estimating a Bernoulli parameter

Consider n i.i.d. random variables $X_1, X_2, \ldots, X_n$.

- The sequence $X_1, X_2, \ldots, X_n$ is a **sample** from distribution F.
- X_i have distribution F with $E[X_i] = \mu$, $\text{Var}(X_i) = \sigma^2$.

- Suppose distribution $F = \text{Ber}(\theta)$ with unknown parameter θ.
- Say you have three estimates $\hat{\theta}$: $\hat{\theta} = 0.5$, $\hat{\theta} = 0.8$, or $\hat{\theta} = 1$

Which estimate is most likely to give you the following sample ($n = 10$)?

$$[0, 0, 1, 1, 1, 1, 1, 1, 1, 1]$$

$P(\text{sample}|\theta = 0.5) = (0.5)^2(0.5)^8 = 0.00097$

$P(\text{sample}|\theta = 0.8) = (0.2)^2(0.8)^8 = 0.00671$ ⇐ Estimate $\hat{\theta} = 0.8$

$P(\text{sample}|\theta = 1.0) = (0)^2(1.0)^8 = 0$

Defining the likelihood of data

Consider a sample of n i.i.d. random variables $X_1, X_2, \ldots, X_n$.

- X_i was drawn from a distribution with density function $f(X_i|\theta)$.
- Observed data: $(x_1, x_2, \ldots, x_n)$

Note: now explicitly specify parameter θ of distribution

Likelihood question:

How likely is the observed data $(x_1, x_2, \ldots, x_n)$ given parameter θ?

Likelihood function, $L(\theta)$:

$$L(\theta) = \prod_{i=1}^{n} f(X_i|\theta)$$

This is just a product, since X_i are i.i.d.

Maximum Likelihood Estimator

Consider a sample of n i.i.d. random variables $X_1, X_2, \ldots, X_n$.

<u>def</u> The **Maximum Likelihood Estimator (MLE)** of θ is the value of θ that maximizes $L(\theta)$.

$$\theta_{MLE} = \arg\max_{\theta} L(\theta)$$

Maximum Likelihood Estimator

Consider a sample of n i.i.d. random variables $X_1, X_2, \dots, X_n$.

def The **Maximum Likelihood Estimator (MLE)** of θ is the value of θ that maximizes $L(\theta)$.

$$\theta_{MLE} = \arg\max_{\theta} L(\theta)$$

Likelihood Function

$$L(\theta) = \prod_{i=1}^{n} f(X_i|\theta)$$

For continuous X_i, $f(X_i|\theta)$ is PDF; for discrete X_i, $f(X_i|\theta)$ is PMF

Maximum Likelihood Estimator

Consider a sample of n i.i.d. random variables $X_1, X_2, \dots, X_n$.

def The **Maximum Likelihood Estimator (MLE)** of θ is the value of θ that maximizes $L(\theta)$.

$$\theta_{MLE} = \underset{\theta}{\arg\max}\ L(\theta)$$

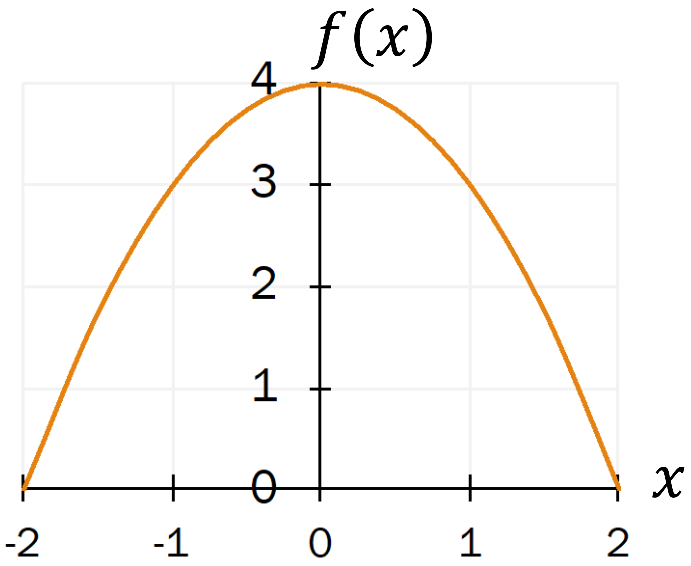
The argument θ
that maximizes $L(\theta)$

New function: arg max

$$\arg\max_x f(x)$$

The x that maximizes the function $f(x)$.

Let $f(x) = -x^2 + 4$, where $-2 < x < 2$.



1. $\max_x f(x)$?
2. $\arg\max_x f(x)$?

Argmax properties

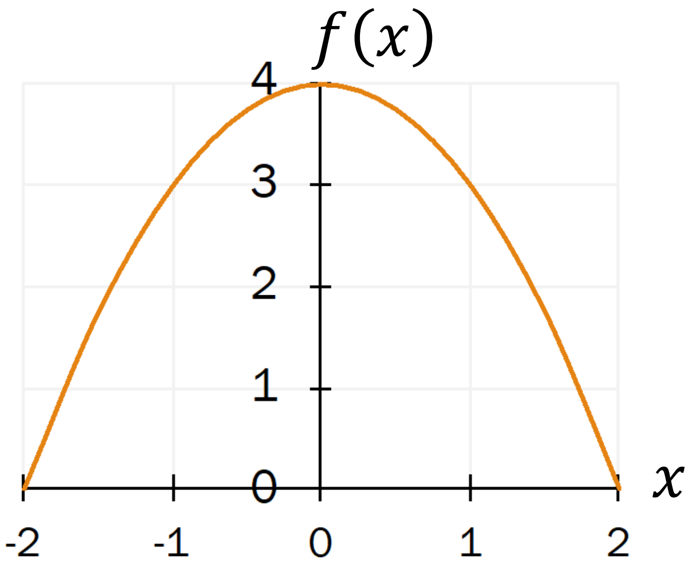
$$\arg\max_x f(x)$$

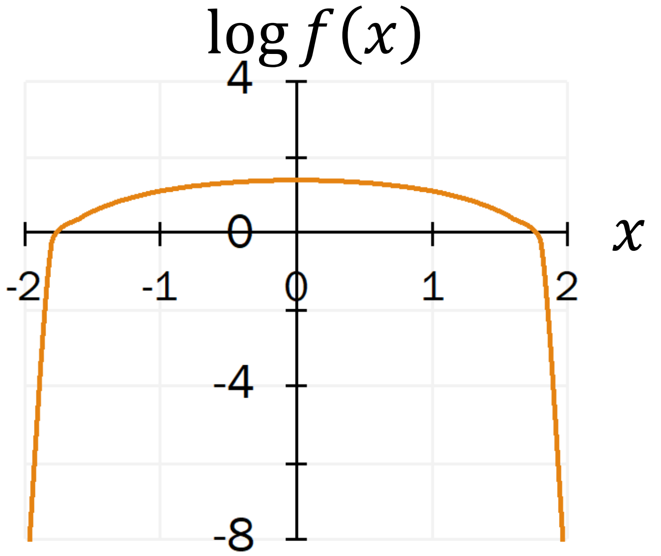
The x that maximizes the function $f(x)$.

$$= \arg\max_x \log f(x)$$

Let $f(x) = -x^2 + 4$, where $-2 < x < 2$.

$$\arg\max_x f(x) = 0$$





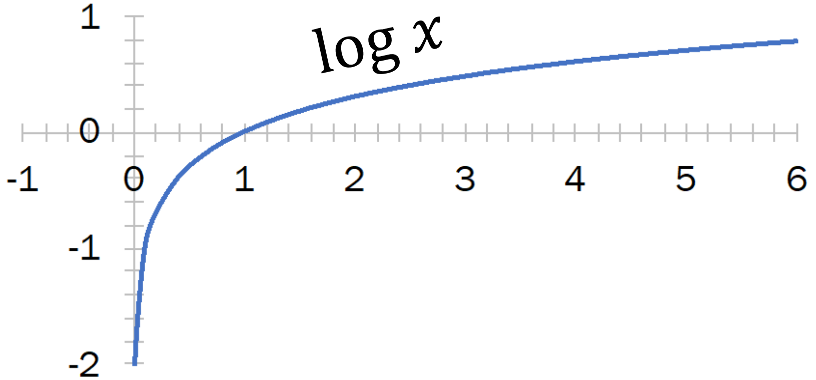
Argmax properties

$$\arg\max_x f(x)$$

The x that maximizes the function $f(x)$.

$$= \arg\max_x \log f(x)$$

- Log is monotonic:
  $x \leq y \Leftrightarrow \log x \leq \log y$



- Log of product = sum of logs:

$$\log(ab) = \log a + \log b$$

Argmax properties

$$\arg\max_x f(x)$$

The x that maximizes the function $f(x)$.

$$= \arg\max_x \log f(x)$$

(log is monotonic: $x \leq y \Leftrightarrow \log x \leq \log y$)

$$= \arg\max_x (c \log f(x))$$

for any positive constant c

($x \leq y \Leftrightarrow c \log x \leq c \log y$)

Maximum Likelihood Estimator

Consider a sample of n i.i.d. random variables $X_1, X_2, \dots, X_n$.

<u>def</u> The Maximum Likelihood Estimator (MLE) of θ is the value of θ that maximizes $L(\theta)$.

$$\theta_{MLE} = \arg\max_{\theta} L(\theta)$$

θ_{MLE} also maximizes the log-likelihood function $LL(\theta)$:

$$LL(\theta) = \log L(\theta) = \log\left(\prod_{i=1}^{n} f(X_i|\theta)\right) = \sum_{i=1}^{n} \log f(X_i|\theta)$$

$$\theta_{MLE} = \arg\max_{\theta} LL(\theta)$$

(log is monotonic)

Story so far

- We want to estimate a parameter θ for a density $f(X_i|\theta)$.

- Consider a sample of n i.i.d. random variables $X_1, X_2, \ldots, X_n$.

Likelihood $L(\theta) = \prod_{i=1}^{n} f(X_i|\theta)$ Log-likelihood $LL(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta)$

- We can choose θ by finding the argmax of the log-likelihood of data:

$$\theta_{MLE} = \arg\max_{\theta} LL(\theta) = \arg\max_{\theta} \sum_{i=1}^{n} \log f(X_i|\theta)$$

Computing the MLE

$\theta_{MLE} = \arg\max_{\theta} LL(\theta)$

General approach for finding θ_{MLE} , the MLE of θ:

1. Determine formula for $LL(\theta)$

$$LL(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta)$$

2. Differentiate $LL(\theta)$ w.r.t. (each) θ

$$\frac{\partial LL(\theta)}{\partial \theta}$$

To maximize: $\frac{\partial LL(\theta)}{\partial \theta} = 0$

3. Solve resulting (simultaneous) equations

(algebra or computer)

4. Make sure derived $\hat{\theta}_{MLE}$ is a maximum
   - Check $LL(\theta_{MLE} \pm \epsilon) < LL(\theta_{MLE})$
   - Often ignored in expository derivations
   - We'll ignore it here too (and won't require it in class)

Maximum Likelihood with Bernoulli

Consider a sample of n i.i.d. random variables $X_1, X_2, \ldots, X_n$.

- Let $X_i \sim \text{Ber}(p)$.

What is $\theta_{MLE} = p_{MLE}$?

1. Determine formula for $LL(\theta)$
2. Differentiate $LL(\theta)$ w.r.t. (each) θ, set to 0
3. Solve resulting (simultaneous) equations

$$LL(\theta) = \sum_{i=1}^{n} \log f(X_i|p)$$

What is the PMF $f(X_i|p)$?

A. p

B. $1 - p$

C. $\begin{cases} p & \text{if } X_i = 1 \\ 1 - p & \text{if } X_i = 0 \end{cases}$

D. $p^{X_i}(1-p)^{1-X_i}$ where $X_i \in \{0,1\}$

Maximum Likelihood with Bernoulli

Consider a sample of n i.i.d. random variables $X_1, X_2, \ldots, X_n$.

- Let $X_i \sim \text{Ber}(p)$.

What is $\theta_{MLE} = p_{MLE}$?

1. Determine formula for $LL(\theta)$
2. Differentiate $LL(\theta)$ w.r.t. (each) θ, set to 0
3. Solve resulting equations

$$LL(\theta) = \sum_{i=1}^{n} \log f(X_i|p)$$

- Is differentiable
- Valid PMF over discrete domain

What is the PMF $f(X_i|p)$?

A. p

B. $1 - p$

C. $\begin{cases} p & \text{if } X_i = 1 \\ 1 - p & \text{if } X_i = 0 \end{cases}$

D. $p^{X_i}(1-p)^{1-X_i}$ where $X_i \in \{0,1\}$

Maximum Likelihood with Bernoulli

Consider a sample of n i.i.d. random variables $X_1, X_2, \ldots, X_n$.

- Let $X_i \sim \text{Ber}(p)$.
- $f(X_i|p) = p^{X_i}(1-p)^{1-X_i}$ where $X_i \in \{0,1\}$

What is $\theta_{MLE} = p_{MLE}$?

1. Determine formula for $LL(\theta)$
2. Differentiate $LL(\theta)$ w.r.t. (each) θ, set to 0
3. Solve resulting equations

$$LL(\theta) = \sum_{i=1}^{n} \log f(X_i|p)$$

$$= \sum_{i=1}^{n} \log\left(p^{X_i}(1-p_i)^{1-X_i}\right) = \sum_{i=1}^{n} [X_i \log p + (1-X_i)\log(1-p)]$$

$$= Y(\log p) + (n-Y)\log(1-p), \text{ where } Y = \sum_{i=1}^{n} X_i$$

Maximum Likelihood with Bernoulli

Consider a sample of n i.i.d. random variables $X_1, X_2, \ldots, X_n$.

- Let $X_i \sim \text{Ber}(p)$.
- $f(X_i|p) = p^{X_i}(1-p)^{1-X_i}$ where $X_i \in \{0,1\}$

What is $\theta_{MLE} = p_{MLE}$?

1. Determine formula for $LL(\theta)$
2. Differentiate $LL(\theta)$ w.r.t. (each) θ, set to 0
3. Solve resulting equations

$$LL(\theta) = \sum_{i=1}^{n} [X_i \log p + (1 - X_i) \log(1-p)] = Y(\log p) + (n - Y)\log(1-p)$$

where $Y = \sum_{i=1}^{n} X_i$

$$\frac{\partial LL(\theta)}{\partial p} = Y\frac{1}{p} + (n - Y)\frac{-1}{1-p} = 0$$

Maximum Likelihood with Bernoulli

Consider a sample of n i.i.d. random variables $X_1, X_2, \ldots, X_n$.

- Let $X_i \sim \text{Ber}(p)$.
- $f(X_i|p) = p^{X_i}(1-p)^{1-X_i}$ where $X_i \in \{0,1\}$

What is $\theta_{MLE} = p_{MLE}$?

1. Determine formula for $LL(\theta)$
2. Differentiate $LL(\theta)$ w.r.t. (each) θ, set to 0
3. Solve resulting equations

$$LL(\theta) = \sum_{i=1}^{n} [X_i \log p + (1 - X_i) \log(1-p)] = Y(\log p) + (n - Y)\log(1-p)$$

where $Y = \sum_{i=1}^{n} X_i$

$$\frac{\partial LL(\theta)}{\partial p} = Y\frac{1}{p} + (n-Y)\frac{-1}{1-p} = 0$$

$$p_{MLE} = \frac{1}{n}Y = \frac{1}{n}\sum_{i=1}^{n} X_i$$

MLE of the Bernoulli parameter, p_{MLE}, is the unbiased estimate of the mean, $\bar{X}$ (sample mean)

Quick check

- You draw n i.i.d. random variables $X_1, X_2, \ldots, X_n$ from the distribution F, yielding the following sample:

$$[0, 0, 1, 1, 1, 1, 1, 1, 1, 1] \qquad (n = 10)$$

- Suppose distribution $F = \text{Ber}(p)$ with unknown parameter p.

1. What is p_{MLE}, the MLE of the parameter p?

   A. 1.0
   B. 0.5
   C. 0.8
   D. 0.2
   E. None/other

Quick check

- You draw n i.i.d. random variables $X_1, X_2, \ldots, X_n$ from the distribution F, yielding the following sample:

$$[0, 0, 1, 1, 1, 1, 1, 1, 1, 1] \qquad (n = 10)$$

- Suppose distribution $F = \text{Ber}(p)$ with unknown parameter p.

1. What is p_{MLE}, the MLE of the parameter p?
2. What is the likelihood $L(\theta)$ of this particular sample?

Quick check

- You draw n i.i.d. random variables $X_1, X_2, \dots, X_n$ from the distribution F, yielding the following sample:

$$[0, 0, 1, 1, 1, 1, 1, 1, 1, 1] \qquad (n = 10)$$

- Suppose distribution $F = \text{Ber}(p)$ with unknown parameter p.

1. What is p_{MLE}, the MLE of the parameter p?
2. What is the likelihood $L(\theta)$ of this particular sample?

$f(X_i|p) = p^{X_i}(1-p)^{1-X_i}$ where $X_i \in \{0,1\}$

$$L(\theta) = \prod_{i=1}^{n} f(X_i|p) \quad \text{where } \theta = p$$

$$= p^8(1-p)^2$$

Maximum Likelihood Algorithm

1. Decide on a model for the distribution of your samples. Define the PMF/PDF for the distribution.

$$f(X_i|p)$$

2. Write out the log-likelihood function.

$$LL(\theta) = \sum_{i=1}^{n} \log f(X_i|p)$$

3. State that the optimal parameters are the argmax of the log-likelihood function.

$$\theta_{MLE} = \arg\max_{\theta} LL(\theta)$$

4. Use an optimization algorithm to calculate argmax:
   - Differentiate $LL(\theta)$ w.r.t (each) θ, set to 0
   - Solve resulting (simultaneous) equations

Maximum Likelihood with Poisson

Consider a sample of n i.i.d. random variables $X_1, X_2, \ldots, X_n$.

- Let $X_i \sim \text{Poi}(\lambda)$.
- PMF: $f(X_i|\lambda) = \frac{e^{-\lambda}\lambda^{X_i}}{X_i!}$

What is $\theta_{MLE} = \lambda_{MLE}$?

Maximum Likelihood with Poisson

Consider a sample of n i.i.d. random variables $X_1, X_2, \ldots, X_n$.

- Let $X_i \sim \text{Poi}(\lambda)$.
- PMF: $f(X_i|\lambda) = \frac{e^{-\lambda}\lambda^{X_i}}{X_i!}$

What is $\theta_{MLE} = \lambda_{MLE}$?

1. Determine formula for $LL(\theta)$
2. Differentiate $LL(\theta)$ w.r.t. (each) θ, set to 0
3. Solve resulting equations

$$LL(\theta) = \sum_{i=1}^{n} \log\left(\frac{e^{-\lambda}\lambda^{X_i}}{X_i!}\right) = \sum_{i=1}^{n} -\lambda \log e + X_i \log \lambda - \log X_i!$$

$$= -n\lambda + \log(\lambda)\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \log(X_i!)$$ (using natural log, $\ln e = 1$)

Maximum Likelihood with Poisson

Consider a sample of n i.i.d. random variables $X_1, X_2, \dots, X_n$.

- Let $X_i \sim \text{Poi}(\lambda)$.
- PMF: $f(X_i|\lambda) = \frac{e^{-\lambda}\lambda^{X_i}}{X_i!}$

What is $\theta_{MLE} = \lambda_{MLE}$?

1. Determine formula for $LL(\theta)$
2. Differentiate $LL(\theta)$ w.r.t. (each) θ, set to 0
3. Solve resulting equations

$$LL(\theta) = -n\lambda + \log(\lambda)\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \log(X_i!)$$

$$\frac{\partial LL(\theta)}{\partial \lambda} = -n + \frac{1}{\lambda}\sum_{i=1}^{n} X_i = 0$$

($\sum_{i=1}^{n} \log(X_i!)$ is a constant w.r.t λ)

Maximum Likelihood with Poisson

Consider a sample of n i.i.d. random variables $X_1, X_2, \dots, X_n$.

- Let $X_i \sim \text{Poi}(\lambda)$.
- PMF: $f(X_i|\lambda) = \dfrac{e^{-\lambda}\lambda^{X_i}}{X_i!}$

What is $\theta_{MLE} = \lambda_{MLE}$?

1. Determine formula for $LL(\theta)$
2. Differentiate $LL(\theta)$ w.r.t. (each) θ, set to 0
3. Solve resulting equations

$$LL(\theta) = -n\lambda + \log(\lambda)\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \log(X_i!)$$

$$\frac{\partial LL(\theta)}{\partial \lambda} = -n + \frac{1}{\lambda}\sum_{i=1}^{n} X_i = 0 \quad \Rightarrow \quad \lambda_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

MLE of the Poisson parameter, λ_{MLE}, is the unbiased estimate of the mean, $\bar{X}$ (sample mean)

Maximum Likelihood with Uniform

Consider a sample of n i.i.d. random variables $X_1, X_2, \dots, X_n$.

Let $X_i \sim \text{Uni}(\alpha, \beta)$.

$$f(X_i|\alpha,\beta) = \begin{cases} \dfrac{1}{\beta - \alpha} & \text{if } \alpha \le X_i \le \beta \\ 0 & \text{otherwise} \end{cases}$$

1. Determine formula for $L(\theta)$

Likelihood:

$$L(\theta) = \begin{cases} \left(\dfrac{1}{\beta - \alpha}\right)^n & \text{if } \alpha \le X_1, X_2, \dots, X_n \le \beta \\ 0 & \text{otherwise} \end{cases}$$

2. Differentiate $LL(\theta)$ w.r.t. (each) θ, set to 0

A. Great, let's do it
B. Differentiation is hard
C. Constraint $\alpha \le X_1, X_2, \dots, X_n \le \beta$ makes differentiation hard

Example sample from a Uniform

Consider a sample of n i.i.d. random variables $X_1, X_2, \ldots, X_n$.

Let $X_i \sim \text{Uni}(\alpha, \beta)$.

$$L(\theta) = \begin{cases} \left(\frac{1}{\beta - \alpha}\right)^n & \text{if } \alpha \leq X_1, X_2, \ldots, X_n \leq \beta \\ 0 & \text{otherwise} \end{cases}$$

Suppose $X_i \sim \text{Uni}(0,1)$.
You observe data:

$$[0.15, 0.20, 0.30, 0.40, 0.65, 0.70, 0.75]$$

Which parameters would give you maximum $L(\theta)$?

A. $\text{Uni}(\alpha = 0, \beta = 1)$

B. $\text{Uni}(\alpha = 0.15, \beta = 0.75)$

C. $\text{Uni}(\alpha = 0.15, \beta = 0.70)$

Example sample from a Uniform

Consider a sample of n i.i.d. random variables $X_1, X_2, \dots, X_n$.

Let $X_i \sim \text{Uni}(\alpha, \beta)$.

$$L(\theta) = \begin{cases} \left(\frac{1}{\beta - \alpha}\right)^n & \text{if } \alpha \le X_1, X_2, \dots, X_n \le \beta \\ 0 & \text{otherwise} \end{cases}$$

Suppose $X_i \sim \text{Uni}(0,1)$.
You observe data:

$$[0.15, 0.20, 0.30, 0.40, 0.65, 0.70, 0.75]$$

Which parameters would give you maximum $L(\theta)$?

A. $\text{Uni}(\alpha = 0 \quad, \beta = 1 \quad)$ $\quad (1)^7 = 1$

B. $\text{Uni}(\alpha = 0.15, \beta = 0.75)$ $\quad \left(\frac{1}{0.6}\right)^7 = 35.7$

C. $\text{Uni}(\alpha = 0.15, \beta = 0.70)$ $\quad \left(\frac{1}{0.55}\right)^6 \cdot 0 = 0$

Original parameters may not yield maximum likelihood.

Maximum Likelihood with Uniform

Consider a sample of n i.i.d. random variables $X_1, X_2, \ldots, X_n$.

Let $X_i \sim \text{Uni}(\alpha, \beta)$.

$$L(\theta) = \begin{cases} \left(\dfrac{1}{\beta - \alpha}\right)^n & \text{if } \alpha \le X_1, X_2, \ldots, X_n \le \beta \\ 0 & \text{otherwise} \end{cases}$$

θ_{MLE}: $\alpha_{MLE} = \min(x_1, x_2, \ldots, x_n)$ $\quad\quad$ $\beta_{MLE} = \max(x_1, x_2, \ldots, x_n)$

Intuition:

- Want interval size $(\beta - \alpha)$ to be as small as possible to maximize likelihood function per datapoint
- Need to make sure all observed data is in interval (if not, then $L(\theta) = 0$)

(demo)

Small samples = problems with MLE

Maximum Likelihood Estimator θ_{MLE} :

- Best explains data we have seen
- Does not attempt to generalize to unseen data.

$$\theta_{MLE} = \arg\max_{\theta} L(\theta)$$
$$= \arg\max_{\theta} LL(\theta)$$

In many cases, $\mu_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i$ Sample mean (MLE for Bernoulli p, Poisson λ, Normal μ)

- Unbiased ($E[\mu_{MLE}] = \mu$ regardless of size of sample, n)

For some cases, like Uniform: $\alpha_{MLE} \geq \alpha$, $\beta_{MLE} \leq \beta$

- Biased. Problematic for small sample size
- Example: If $n = 1$ then $\alpha = \beta$, yielding an invalid distribution

Properties of MLE

Maximum Likelihood Estimator:

- Best explains data we have seen
- Does not attempt to generalize to unseen data.

$$\theta_{MLE} = \arg\max_{\theta} L(\theta)$$
$$= \arg\max_{\theta} LL(\theta)$$

- Often used when sample size n is large relative to parameter space
- Potentially biased (though asymptotically less so, as $n \to \infty$)
- Consistent: $\lim_{n\to\infty} P\left(\left|\hat{\theta} - \theta\right| < \varepsilon\right) = 1$ where $\varepsilon > 0$

  As $n \to \infty$ (i.e., more data), probability that $\hat{\theta}$ significantly differs from θ is zero