23: Maximum A Posteriori (MAP)

Lisa Yan

November 13, 2019

Review: Maximum Likelihood Algorithm

1. Decide on a model for the distribution of your samples. Define the PMF/PDF for the distribution. $f(X_i|\theta)$

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

3. State: $$\theta_{MLE} = \arg\max_{\theta} LL(\theta)$$

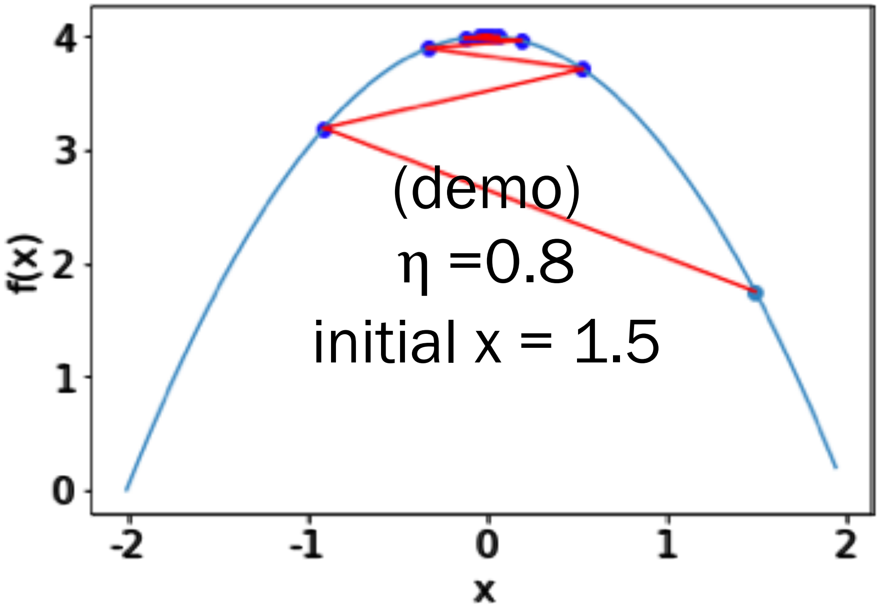
4. Use an optimization algorithm to calculate argmax:

Option 1: Optimization with	Option 2: Optimization with
Math	Gradient Ascent

Gradient ascent algorithm

Walk uphill and you will find a local maxima
(if your step is small enough).

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



1. $$\frac{df}{dx} = -2x$$ Gradient at x

2. Gradient ascent algorithm:

```
initialize x
repeat many times:
  compute gradient
  x += η * gradient
```

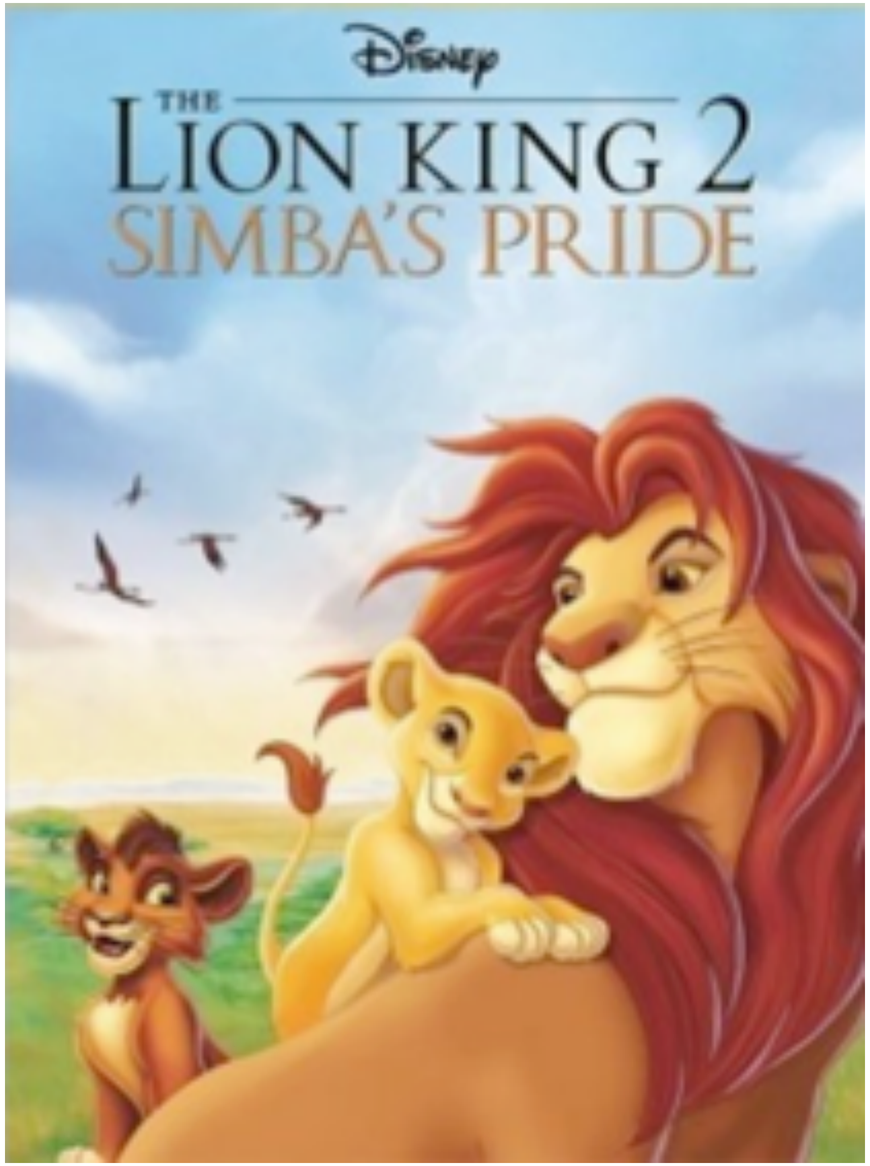
"learning rate" η

Today's plan

Gradient Ascent

- MLE for Linear Regression lite

Maximum A Posteriori


Disney
THE
LION KING 2
SIMBA'S PRIDE

Linear Regression Lite

Let X = CO_2 level (ppm) change from 1980,
Y = Average Land-Ocean Temperature (°C).

You observe n datapoints:

$$\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \dots, \left(x^{(n)}, y^{(n)}\right)$$

Example:
$(0, 0.26), (1.2, 0.32), \dots, (368.58, 0.85)$

New notation!

$\left(x^{(i)}, y^{(i)}\right)$: the i-th datapoint in our sample of size n
has density function $f\left(x^{(i)}, y^{(i)} | \theta\right)$

Linear Regression Lite

Let X = CO_2 level (ppm) change from 1980,
Y = Average Land-Ocean Temperature (°C).

You observe n datapoints:

$$\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \dots, \left(x^{(n)}, y^{(n)}\right)$$

Example:
$(0, 0.26), (1.2, 0.32), \dots, (368.58, 0.85)$

Linear Regression Model:

- $Y = \theta X + Z$ (linear relationship)
- $Z \sim \mathcal{N}(0, \sigma^2)$ (error normally distributed)
- $\Rightarrow Y|X \sim \mathcal{N}(\theta X, \sigma^2)$

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

Gradient ascent with Linear Regression Lite

Model:

$$Y|X \sim \mathcal{N}(\theta X, \sigma^2)$$

n datapoints in sample:

$$\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)$$

1. Calculate likelihood of data, $L(\theta)$.

$$L(\theta) = \prod_{i=1}^{n} f\left(x^{(i)}, y^{(i)} | \theta\right)$$

(chain rule)

$$= \prod_{i=1}^{n} f\left(x^{(i)} | \theta\right) f\left(y^{(i)} | x^{(i)}, \theta\right)$$

($x^{(i)}$ indep. of θ)

$$= \prod_{i=1}^{n} f\left(x^{(i)}\right) f\left(y^{(i)} | x^{(i)}, \theta\right)$$

$$= \prod_{i=1}^{n} f\left(x^{(i)}\right) \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(y^{(i)} - \theta x^{(i)}\right)^2 / (2\sigma^2)}$$

$f\left(y^{(i)} | x^{(i)}, \theta\right)$ is PDF of $\mathcal{N}\left(\theta x^{(i)}, \sigma^2\right)$

Gradient ascent with Linear Regression Lite

Model:

$$Y|X \sim \mathcal{N}(\theta X, \sigma^2)$$

n datapoints in sample:

$$\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \dots, \left(x^{(n)}, y^{(n)}\right)$$

$$L(\theta) = \prod_{i=1}^{n} f\left(x^{(i)}\right) \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(y^{(i)} - \theta x^{(i)}\right)^2 / (2\sigma^2)}$$

2. Calculate log-likelihood of data, $LL(\theta)$.

$$LL(\theta) = \log L(\theta) \quad = \log\left[\prod_{i=1}^{n} f\left(x^{(i)}\right) \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(y^{(i)} - \theta x^{(i)}\right)^2 / (2\sigma^2)}\right]$$

$$= \sum_{i=1}^{n} \log\left[f\left(x^{(i)}\right) \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(y^{(i)} - \theta x^{(i)}\right)^2 / (2\sigma^2)}\right]$$

log(prod) = sum(log)

$$= \sum_{i=1}^{n} \log f\left(x^{(i)}\right) - \sum_{i=1}^{n} \log \sqrt{2\pi}\sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y^{(i)} - \theta x^{(i)}\right)^2$$

log(prod) = sum(log)
+ using natural log

Gradient ascent with Linear Regression Lite

Model:

$$Y|X \sim \mathcal{N}(\theta X, \sigma^2)$$

n datapoints in sample:

$$\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)$$

$$LL(\theta) = \sum_{i=1}^{n} \log f\left(x^{(i)}\right) - \sum_{i=1}^{n} \log \sqrt{2\pi}\sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y^{(i)} - \theta x^{(i)}\right)^2$$

3. State MLE as optimization objective.

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

$$= \arg\max_{\theta} \left[\sum_{i=1}^{n} \log f\left(x^{(i)}\right) - \sum_{i=1}^{n} \log \sqrt{2\pi}\sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y^{(i)} - \theta x^{(i)}\right)^2 \right]$$

🎉 Celebrate!

$$= \arg\max_{\theta} \left[- \sum_{i=1}^{n} \left(y^{(i)} - \theta x^{(i)}\right)^2 \right]$$

(eliminate constants w.r.t. $\arg\max_{\theta}$)

Gradient ascent with Linear Regression Lite

Model:

$$Y|X \sim \mathcal{N}(\theta X, \sigma^2)$$

n datapoints in sample:

$$(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(n)}, y^{(n)})$$

Goal: $$\theta_{MLE} = \arg\max_{\theta} \left[-\sum_{i=1}^{n} \left(y^{(i)} - \theta x^{(i)}\right)^2 \right]$$

4. Compute gradient w.r.t. θ.

$$\frac{\partial}{\partial \theta} \left[-\sum_{i=1}^{n} \left(y^{(i)} - \theta x^{(i)}\right)^2 \right]$$

$$= -\sum_{i=1}^{n} \frac{\partial}{\partial \theta} \left(y^{(i)} - \theta x^{(i)}\right)^2$$

$$= -\sum_{i=1}^{n} 2\left(y^{(i)} - \theta x^{(i)}\right)\left(-x^{(i)}\right) = \sum_{i=1}^{n} 2\left(y^{(i)} - \theta x^{(i)}\right)\left(x^{(i)}\right)$$

Gradient ascent with Linear Regression Lite

Model:

$$Y|X \sim \mathcal{N}(\theta X, \sigma^2)$$

n datapoints in sample:

$$\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)$$

Goal: $$\theta_{MLE} = \arg\max_{\theta} \left[-\sum_{i=1}^{n} \left(y^{(i)} - \theta x^{(i)}\right)^2 \right]$$

Gradient: $$\frac{\partial LL(\theta)}{\partial \theta} = \sum_{i=1}^{n} 2\left(y^{(i)} - \theta x^{(i)}\right)\left(x^{(i)}\right)$$

5. Optimize.

```
initialize θ
repeat many times:
  compute gradient
  θ += η * gradient
```

(demo)

Gradient ascent with multiple parameters

Preview

If $\theta = (\theta_0, \theta_1, \theta_2, \ldots, \theta_j, \ldots, \theta_m)$, what is $\theta_{MLE} = \arg\max_{\theta} LL(\theta)$?

Gradient update step for the j-th parameter, θ_j) :

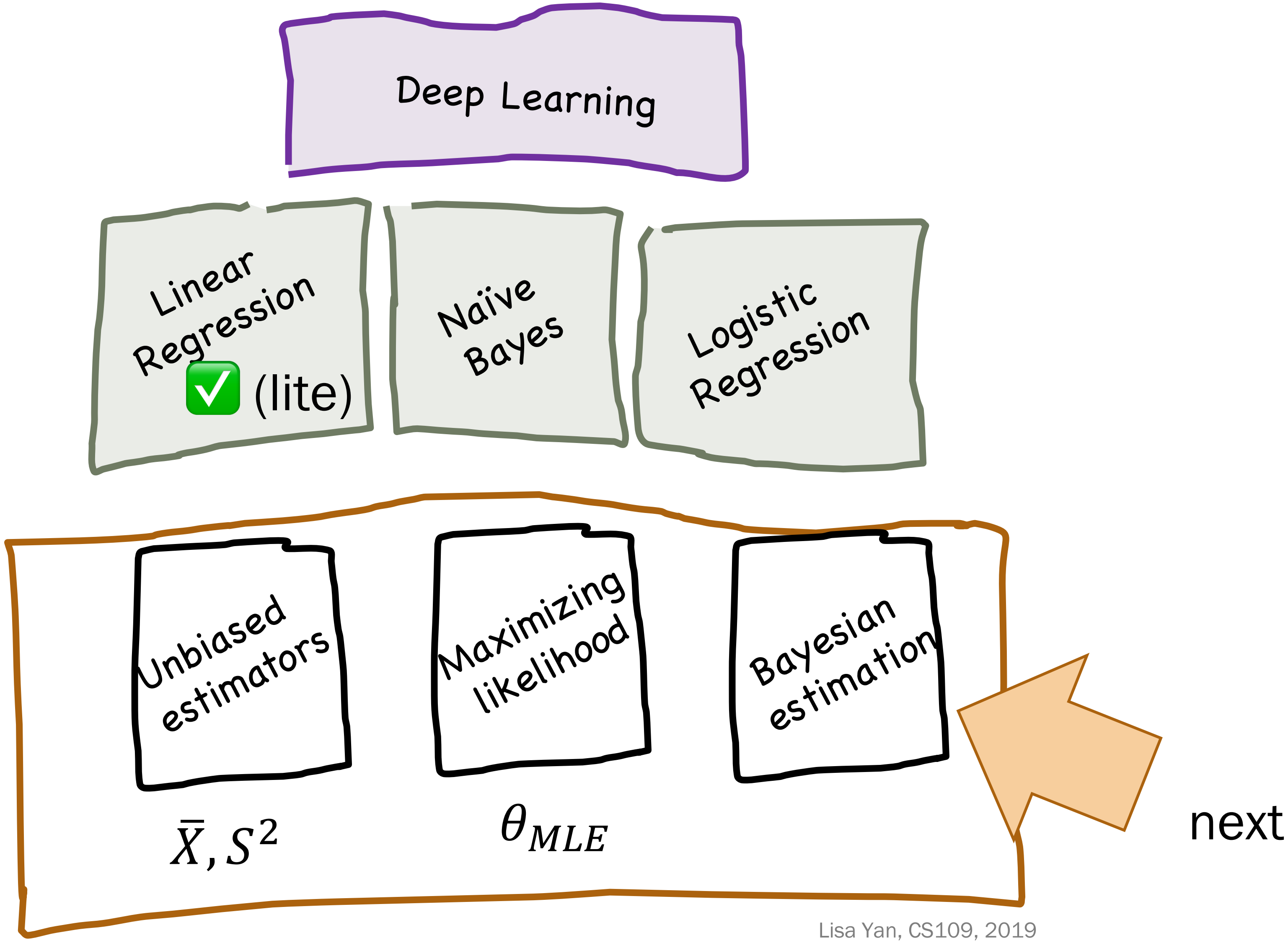
$$\theta_j^{\text{new}} = \theta_j^{\text{old}} + \eta \cdot \frac{\partial LL(\theta^{\text{old}})}{\partial \theta_j^{\text{old}}}$$

```
initialize θj = 0 for 0 ≤ j ≤ m
repeat many times:
    gradient[j] = 0 for all 0 ≤ j ≤ m
    compute all gradient[j] for all 0 ≤ j ≤ m

    θj += η * gradient[j] for all 0 ≤ j ≤ m
```

Compute all of gradient based on current θ before updating θ.

Our path



Deep Learning
Linear Regression ✅ (lite)
Naïve Bayes
Logistic Regression
Unbiased estimators
Maximizing likelihood
Bayesian estimation
$\bar{X}, S^2$
θ_{MLE}
next

Today's plan

Gradient Ascent

- MLE for Linear Regression lite

Maximum A Posteriori

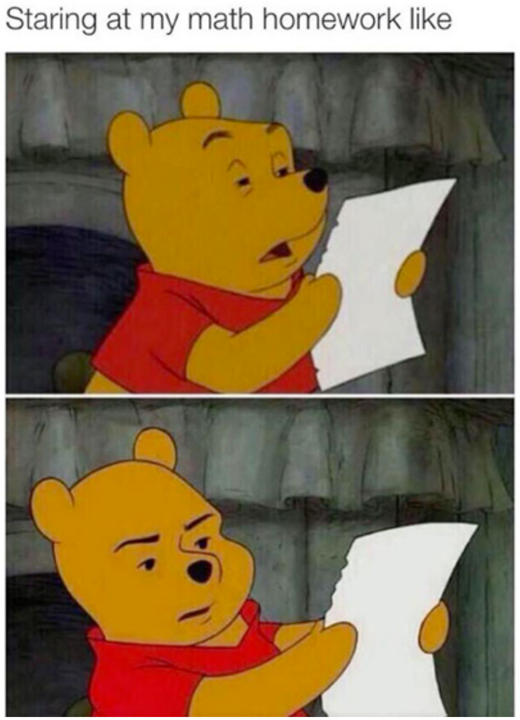
- Picking a conjugate distribution as your prior
- Laplace smoothing

Okay, just one more MLE with the Multinomial

Consider a sample of n i.i.d. random variables $Y_1, Y_2, \dots, Y_n$.

- Let $Y_k \sim \text{Multinomial}(p_1, p_2, \dots, p_m)$, where $\sum_{i=1}^{m} p_i = 1$
- Let X_i = # of trials with outcome i, where $\sum_{i=1}^{m} X_i = n$

Let's give an example!



Okay, just one more MLE with the Multinomial

Consider a sample of n i.i.d. random variables $Y_1, Y_2, \ldots, Y_n$.

- Let $Y_k \sim \text{Multinomial}(p_1, p_2, \ldots, p_m)$, where $\sum_{i=1}^{m} p_i = 1$
- Let X_i = # of trials with outcome i, where $\sum_{i=1}^{m} X_i = n$

Example: Suppose Y_k = outcome of 6-sided die.

$m = 6, \sum_{i=1}^{6} p_i = 1$

- Roll the dice $n = 12$ times.
- Observe data: 3 ones, 2 twos, 0 threes, 3 fours, 1 fives, 3 sixes

$$X_1 = 3, X_2 = 2, X_3 = 0,$$
$$X_4 = 3, X_5 = 1, X_6 = 3$$

Check: $X_1 + X_2 + \cdots + X_6 = 12$

Okay, just one more MLE with the Multinomial

Consider a sample of n i.i.d. random variables $Y_1, Y_2, \ldots, Y_n$.

- Let $Y_k \sim$Multinomial$(p_1, p_2, \ldots, p_m)$, where $\sum_{i=1}^{m} p_i = 1$
- Let X_i = # of trials with outcome i, where $\sum_{i=1}^{m} X_i = n$

Joint PDF $f(X_1, X_2, \ldots, X_m | p_1, p_2, \ldots, p_m)$:

Likelihood $L(\theta)$
of observing the sample (size n)
$(X_1, X_2, \ldots, X_m)$

A. $$\frac{n!}{x_1!\, x_2! \cdots x_m!} p_1^{x_1} p_2^{x_2} \cdots p_m^{x_m}$$

B. $$p_1^{x_1} p_2^{x_2} \cdots p_m^{x_m}$$

C. $$\frac{n!}{x_1!\, x_2! \cdots x_m!} x_1^{p_1} x_2^{p_2} \cdots x_m^{p_m}$$

Okay, just one more MLE with the Multinomial

Consider a sample of n i.i.d. random variables $Y_1, Y_2, \dots, Y_n$.

- Let $Y_k \sim \text{Multinomial}(p_1, p_2, \dots, p_m)$, where $\sum_{i=1}^{m} p_i = 1$
- Let X_i = # of trials with outcome i, where $\sum_{i=1}^{m} X_i = n$

Joint PDF $f(X_1, X_2, \dots, X_m | p_1, p_2, \dots, p_m)$: 👉 Likelihood $L(\theta)$ of observing the sample (size n) $(X_1, X_2, \dots, X_m)$

A. $$\frac{n!}{x_1!\, x_2! \cdots x_m!} p_1^{x_1} p_2^{x_2} \cdots p_m^{x_m}$$

B. $$p_1^{x_1} p_2^{x_2} \cdots p_m^{x_m}$$

C. $$\frac{n!}{x_1!\, x_2! \cdots x_m!} x_1^{p_1} x_2^{p_2} \cdots x_m^{p_m}$$

🤔

Okay, just one more MLE with the Multinomial

Consider a sample of n i.i.d. random variables $Y_1, Y_2, \dots, Y_n$.

- Let $Y_k \sim \text{Multinomial}(p_1, p_2, \dots, p_m)$, where $\sum_{i=1}^{m} p_i = 1$
- Let X_i = # of trials with outcome i, where $\sum_{i=1}^{m} X_i = n$

Joint PDF $f(X_1, X_2, \dots, X_m | p_1, p_2, \dots, p_m) = \dfrac{n!}{x_1!\, x_2! \cdots x_m!} p_1^{x_1} p_2^{x_2} \cdots p_m^{x_m} = L(\theta)$

Log-likelihood:

$$LL(\theta) = \log(n!) - \sum_{i=1}^{m} \log(X_i!) + \sum_{i}^{m} X_i \log(p_i), \text{ such that } \sum_{i=1}^{m} p_i = 1$$

Optimize with Lagrange multipliers in extra slides → θ_{MLE}: $p_i = \dfrac{X_i}{n}$

Intuitively, probability p_i = proportion of outcomes

When MLEs attack!

MLE for Multinomial: $p_i = \frac{X_i}{n}$

Consider a 6-sided die.

- Roll the dice $n = 12$ times.
- Observe: 3 ones, 2 twos, 0 threes, 3 fours, 1 fives, 3 sixes

What is θ_{MLE}? (select all that apply)

A. $p_1 = 3/12$
B. $p_2 = 2/12$
C. $p_3 = 0/12$
D. $p_4 = 3/12$
E. $p_5 = 1/12$
F. $p_6 = 3/12$
G. Other

When MLEs attack!

MLE for Multinomial: $p_i = \frac{X_i}{n}$

Consider a 6-sided die.

- Roll the dice $n = 12$ times.
- Observe: 3 ones, 2 twos, 0 threes, 3 fours, 1 fives, 3 sixes

θ_{MLE}:

$$p_1 = 3/12$$
$$p_2 = 2/12$$
$$p_3 = 0/12$$
$$p_4 = 3/12$$
$$p_5 = 1/12$$
$$p_6 = 3/12$$

- MLE: you'll **never…EVER…** roll a three.
- Do you really believe that?

Frequentist:

Roll more!
Prob. = frequency in limit

Bayesian:

Have prior beliefs of probability, even before any rolls!

Estimating our parameter directly

(our focus so far)

Maximum Likelihood Estimator (MLE)

What is the parameter θ that maximizes the likelihood of our observed data $(x_1, x_2, \dots, x_n)$?

$$L(\theta) = f(X_1, X_2, \dots, X_n | \theta)$$
$$= \prod_{i=1}^{n} f(X_i | \theta)$$

$$\theta_{MLE} = \arg\max_{\theta} f(X_1, X_2, \dots, X_n | \theta)$$

likelihood of data

Observations:

- MLE maximizes probability of observing data given a parameter θ.
- If we are estimating θ, shouldn't we maximize the probability of θ directly?

Break for jokes/ announcements

Announcements

Problem Set 5

Released:	yes
Due:	Friday 11/15
Covers:	Up to today (inference)
Note:	Errata, updated today 10am

Late day reminder: No late days permitted past last day of the quarter, 12/6 (Friday)

CS109 Contest

Due:	Monday 12/2 11:59pm

Estimating our parameter directly

(our focus so far)

Maximum Likelihood Estimator (MLE)

What is the parameter θ that **maximizes the likelihood** of our observed data $(x_1, x_2, \ldots, x_n)$?

$$L(\theta) = f(X_1, X_2, \ldots, X_n|\theta) = \prod_{i=1}^{n} f(X_i|\theta)$$

$$\theta_{MLE} = \arg\max_{\theta} f(X_1, X_2, \ldots, X_n|\theta)$$

likelihood of data

(our focus today)

Maximum a Posteriori (MAP) Estimator

Given our observed data $(x_1, x_2, \ldots, x_n)$, what is the **most likely parameter** θ?

$$\theta_{MAP} = \arg\max_{\theta} f(\theta|X_1, X_2, \ldots, X_n)$$

posterior distribution of θ

Maximum A Posterior (MAP) Estimator

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

<u>def</u> The **Maximum a Posteriori (MAP) Estimator** of θ is the value of θ that maximizes the posterior distribution of θ.

$$\theta_{MAP} = \arg\max_{\theta} f(\theta | X_1, X_2, \ldots, X_n)$$

Intuition with Bayes' Theorem:

$L(\theta)$, probability of data given parameter θ

After seeing data, posterior belief of θ

Before seeing data, prior belief of θ

posterior, likelihood, prior:

$$P(\theta | \text{data}) = \frac{P(\text{data} | \theta) P(\theta)}{P(\text{data})}$$

Solving for θ_{MAP}

posterior likelihood prior

$$P(\theta|\text{data}) = \frac{P(\text{data}|\theta)P(\theta)}{P(\text{data})}$$

- Observe data: $X_1, X_2, \dots, X_n$, all i.i.d.
- Let likelihood be same as MLE: $f(X_1, X_2, \dots, X_n|\theta) = \prod_{i=1}^{n} f(X_i|\,\theta)$
- Let the prior distribution of θ be $g(\theta)$.

$$\theta_{MAP} = \arg\max_{\theta} f(\theta|X_1, X_2, \dots, X_n) = \arg\max_{\theta} \frac{f(X_1, X_2, \dots, X_n|\theta)g(\theta)}{h(X_1, X_2, \dots, X_n)}$$

(Bayes' Theorem)

$$= \arg\max_{\theta} \frac{g(\theta)\prod_{i=1}^{n} f(X_i|\,\theta)}{h(X_1, X_2, \dots, X_n)}$$

(independence)

$$= \arg\max_{\theta} g(\theta) \prod_{i=1}^{n} f(X_i|\,\theta)$$

($1/h(X_1, X_2, \dots, X_n)$ is a positive constant w.r.t. θ)

$$= \arg\max_{\theta} \left(\log g(\theta) + \sum_{i=1}^{n} \log f(X_i|\,\theta) \right)$$

(start of most important slide of today)

Maximum A Posterior (MAP) Estimator

The MAP estimator has 2 interpretations:

$$\theta_{MAP} = \arg\max_{\theta} f(\theta | X_1, X_2, \ldots, X_n)$$

The mode of the posterior distribution of θ

$$= \arg\max_{\theta} \left(\log g(\theta) + \sum_{i=1}^{n} \log f(X_i | \theta) \right)$$

The θ that maximizes log prior + log-likelihood

In both cases, you must specify your prior, $g(\theta)$.

Key to MAP estimator: You should pick a prior $g(\theta)$ that makes computing the mode of the posterior distribution is **easy**.

(in this class) Use a conjugate distribution.

(end of most important slide of today)

Today's plan

Gradient Ascent

- MLE for Linear Regression lite

Maximum A Posteriori

- Picking a conjugate distribution as your prior
- Laplace smoothing

Beta distribution refresher

We have seen one conjugate distribution so far:

$X \sim \text{Beta}(a, b)$

$a > 0, b > 0$

Support of X: $(0, 1)$

PDF $\quad f(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$

where $B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1} dx$, normalizing constant

- Beta is the **conjugate distribution** for Bernoulli, meaning:

Prior	$\text{Beta}(a = n_{imag} + 1, b = m_{imag} + 1)$
Experiment	Observe n successes and m failures
Posterior	$\text{Beta}\left(a = n_{imag} + n + 1, b = m_{imag} + m + 1\right)$

- Mode of $\text{Beta}(a, b)$: $\frac{a-1}{a+b-2}$

MAP estimator for Bernoulli

Suppose you observe data D: n heads, m tails

1. Decide model.
   Bernoulli with parameter p
2. Decide prior distribution of parameter θ, $g(\theta)$.
   $\theta \sim \text{Beta}(a+1, b+1)$
3. Compute θ_{MAP}
   (below)

Solution:

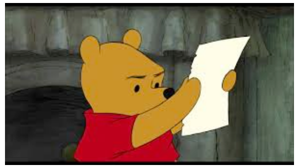
- Beta is a conjugate distribution for Bernoulli.
- If prior $\theta \sim \text{Beta}(a+1, b+1)$ and data = $\{n$ heads, m tails$\}$, then posterior distribution

$$\theta|\text{data} \sim \text{Beta}(a+n+1, b+m+1)$$

- θ_{MAP} is mode of posterior distribution

$$\theta_{MAP} = \frac{a+n}{a+n+b+m}$$

(mode of $\text{Beta}(a+n+1, b+m+1)$)



MAP estimator for Bernoulli, from first principles

Suppose you observe data D: n heads, m tails

1. Decide model.
   Bernoulli with parameter p
2. Decide prior distribution of parameter θ, $g(\theta)$.
   $\theta \sim \text{Beta}(a+1, b+1)$
3. Compute θ_{MAP}
   (below)

$$\theta_{MAP} = \arg\max_{\theta} (\log g(\theta) + \log f(X_1, X_2, \ldots, X_n | \theta))$$

(θ_{MAP} = argmax of log prior + log-likelihood)

$$= \arg\max_{p} \left(\log \left(\frac{1}{\beta} p^{a+1-1} (1-p)^{b+1-1} \right) + \log \left(\binom{n+m}{n} p^n (1-p)^m \right) \right)$$

(PDF of Beta, likelihood of n heads, m tails)

$$= \arg\max_{p} \left(\log \frac{1}{\beta} + a \log(p) + b \log(1-p) + \log \binom{n+m}{n} + n \log p + m \log(1-p) \right)$$

$$= \arg\max_{p} ((a+n) \log(p) + (b+m) \log(1-p))$$

(eliminate constants w.r.t. $\arg\max_{p}$)

MAP estimator for Bernoulli, from first principles

Suppose you observe data D: n heads, m tails

1. Decide model.
   Bernoulli with parameter p
2. Decide prior distribution of parameter θ, $g(\theta)$.
   $\theta \sim \text{Beta}(a+1, b+1)$
3. Compute θ_{MAP}
   (below)

$$\theta_{MAP} = p_{MAP} = \arg\max_p \big((a+n)\log(p) + (b+m)\log(1-p)\big)$$

Differentiate w.r.t. p and set to 0:

$$\frac{(a+n)}{p} - \frac{(b+m)}{1-p} = 0$$

Solve for p:

$$(a+n)(1-p) = (b+m)p$$
$$(a+n) - (a+n)p = (b+m)p$$
$$p(a+n+b+m) = a+n$$

$$p_{MAP} = \frac{a+n}{a+n+b+m}$$

MAP estimator so far

MAP estimator: $\theta_{MAP} = \arg\max_{\theta} f(\theta | X_1, X_2, \ldots, X_n)$

The mode of the posterior distribution of θ

You should pick a prior $g(\theta)$ that makes computing the mode of the posterior distribution **easy**.

Use a conjugate distribution.

The conjugate for Bernoulli is Beta.

- Our prior (subjective) belief:
  $\theta \sim \text{Beta}(a+1, b+1)$
  (Saw $a+b=$ imaginary trials; of those, a were successes.)
- Posterior distribution:
  $(\theta | n \text{ heads, } m \text{ tails}) \sim \text{Beta}(a+n+1, b+m+1)$

Conjugate distributions

MAP estimator:

$$\theta_{MAP} = \arg\max_{\theta} f(\theta | X_1, X_2, \ldots, X_n)$$

The mode of the posterior distribution of θ

Distribution parameter	Prior distribution for parameter
Bernoulli p	Beta
Binomial p	Beta
Multinomial p_i	Dirichlet
Poisson λ	Gamma
Exponential λ	Gamma
Normal μ	Normal
Normal σ^2	Inverse Gamma

Don't need to know Inverse Gamma…
but it will know you ☺

Multinomial is Multiple times the fun

Dirichlet$(a_1, a_2, \ldots, a_m)$ is the conjugate for Multinomial.

- Generalizes Beta in the same way Multinomial generalizes Bernoulli/Binomial:

$$f(x_1, x_2, \ldots, x_m) = \frac{1}{B(a_1, a_2, \ldots, a_m)} \prod_{i=1}^{m} x_i^{a_i - 1}$$

Prior Dirichlet$(a_1 + 1, a_2 + 1, \ldots, a_m + 1)$
Saw $\sum_{i=1}^{m} a_i$ imaginary trials, a_i of outcome i

Experiment Observe $n_1 + n_2 + \cdots + n_m$ new trials, with n_i of outcome i

Posterior Dirichlet$(a_1 + n_1 + 1, a_2 + n_2 + 1, \ldots, a_m + n_m + 1)$

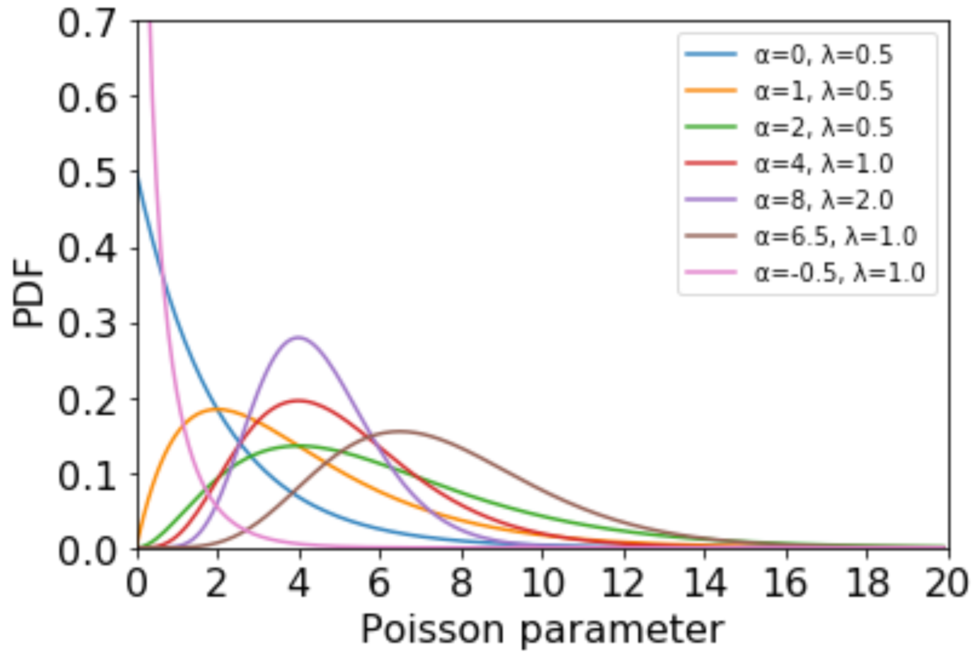
MAP:

$$p_i = \frac{a_i + n_i}{\sum_{i=1}^{m} a_i + \sum_{i=1}^{m} n_i}$$

Good times with Gamma

Gamma(α, λ) is the conjugate for Poisson.

- Also conjugate for Exponential, but we won't delve into that
- Mode of gamma: α/λ



Prior $\theta \sim \text{Gamma}(\alpha, \lambda)$
Saw α total imaginary events during λ prior time periods

Experiment Observe n events during next k time periods

Posterior $(\theta | n \text{ events in } k \text{ periods}) \sim \text{Gamma}(\alpha + n, \lambda + k)$

$$\theta_{MAP} = \frac{\alpha + n}{\lambda + k}$$

MAP for Poisson

Gamma(α, λ) is conjugate for Poisson

Mode: α/λ

Let λ be the average # of successes in a time period.

1. What does it mean to have a prior of $\theta \sim \text{Gamma}(10,5)$?

MAP for Poisson

Gamma(α, λ) is conjugate for Poisson

Mode: α/λ

Let λ be the average # of successes in a time period.

1. What does it mean to have a prior of $\theta \sim \text{Gamma}(10,5)$?

Observe 10 imaginary events in 5 time periods, i.e., observe at Poisson rate = 2

Now perform the experiment and see 11 events in next 2 time periods.

2. Given your prior, what is the posterior distribution?

3. What is θ_{MAP}?

MAP for Poisson

Gamma(α, λ) is conjugate for Poisson

Mode: α/λ

Let λ be the average # of successes in a time period.

1. What does it mean to have a prior of $\theta \sim \text{Gamma}(10,5)$?

   Observe 10 imaginary events in 5 time periods, i.e., observe at Poisson rate = 2

Now perform the experiment and see 11 events in next 2 time periods.

2. Given your prior, what is the posterior distribution?

   $(\theta | n$ events in k periods$) \sim \text{Gamma}(21, 7)$

3. What is θ_{MAP}?

   $\theta_{MAP} = 3$, the updated Poisson rate

Today's plan

Gradient Ascent

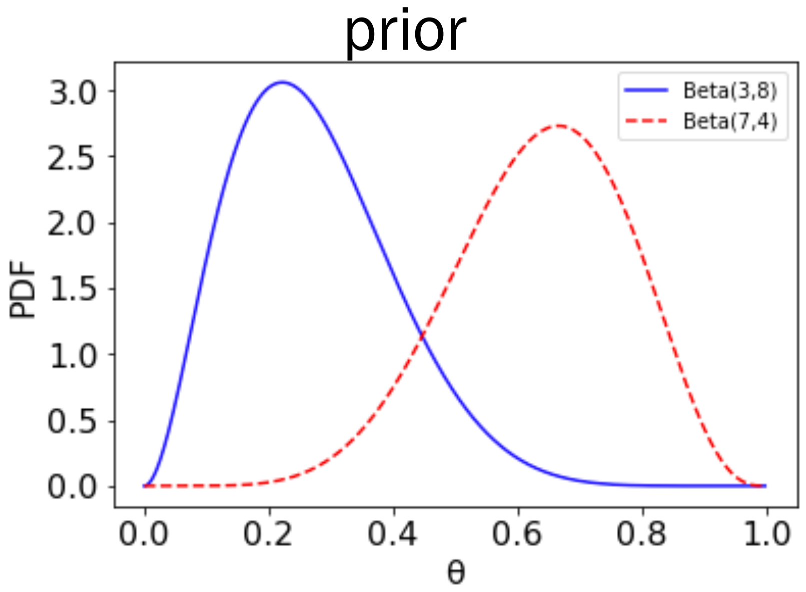
- MLE for Linear Regression lite

Maximum A Posteriori

- Picking a conjugate distribution as your prior
- Laplace smoothing

Where'd you get them priors?

- Let θ be the probability a coin turns up heads.
- Model θ with 2 different priors:
  - Prior 1: Beta(3,8): 2 imaginary heads, 7 imaginary tails — mode: $\frac{2}{9}$
  - Prior 2: Beta(7,4): 6 imaginary heads, 3 imaginary tails — mode: $\frac{6}{9}$

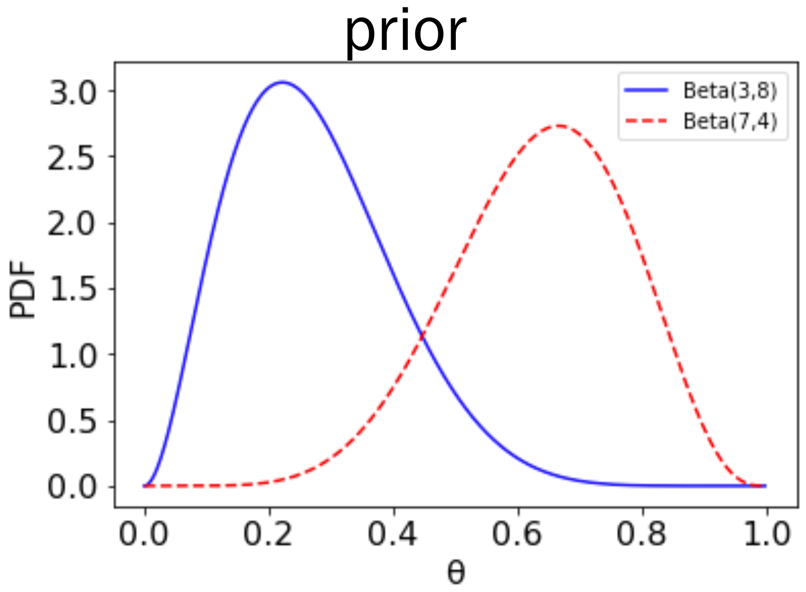


Now flip 100 coins and get 58 heads and 42 tails.

1. What are the two posterior distributions?
2. What are the modes of the two posterior distributions?

Where'd you get them priors?

- Let θ be the probability a coin turns up heads.
- Model θ with 2 different priors:
  - Prior 1: Beta(3,8): 2 imaginary heads, 7 imaginary tails — mode: $\frac{2}{9}$
  - Prior 2: Beta(7,4): 6 imaginary heads, 3 imaginary tails — mode: $\frac{6}{9}$



Now flip 100 coins and get 58 heads and 42 tails.

1. What are the two posterior distributions?
2. What are the modes of the two posterior distributions?

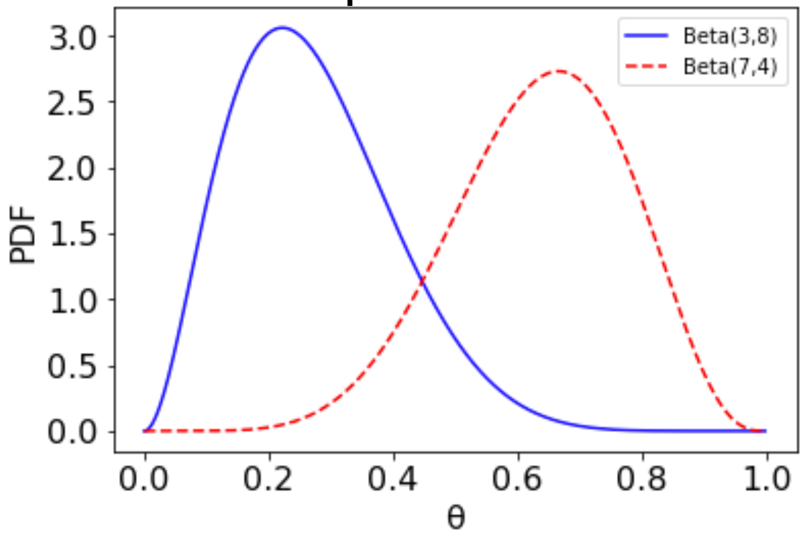
Posterior 1: Beta(61,50) mode: $\frac{60}{109}$

Posterior 2: Beta(65,46) mode: $\frac{64}{109}$

Where'd you get them priors?

- Let θ be the probability a coin turns up heads.
- Model θ with 2 different priors:
  - Prior 1: Beta(3,8): 2 imaginary heads, 7 imaginary tails — mode: $\frac{2}{9}$
  - Prior 2: Beta(7,4): 6 imaginary heads, 3 imaginary tails — mode: $\frac{6}{9}$

prior



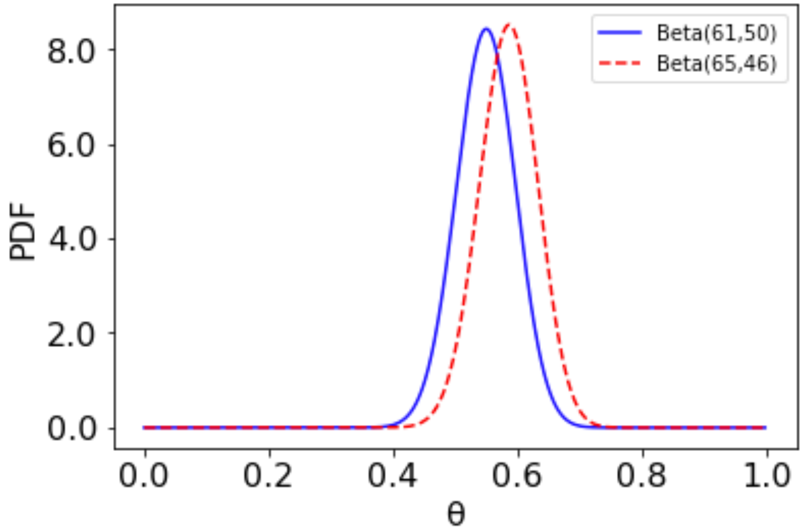
Now flip 100 coins and get 58 heads and 42 tails.

Posterior 1: Beta(61,50) — mode: $\frac{60}{109}$

Posterior 2: Beta(65,46) — mode: $\frac{64}{109}$

As long as we collect enough data, posteriors will converge to the true value.

posterior



Laplace smoothing

MAP with **Laplace smoothing**: a prior which represents **one** imagined observation of each outcome.

Consider our previous 6-sided die.

- Roll the dice $n = 12$ times.
- Observe: 3 ones, 2 twos, 0 threes, 3 fours, 1 fives, 3 sixes

Recall θ_{MLE}:

⚠️

$p_1 = 3/12, p_2 = 2/12, p_3 = 0/12,$
$p_4 = 3/12, p_5 = 1/12, p_6 = 3/12$

θ_{MAP} with Laplace smoothing:

- Assume Dirichlet prior where each outcome seen $k = 1$ times.
- **Laplace estimate**:

$$p_i = \frac{X_i + 1}{n + m}$$

$p_1 = 4/18, p_2 = 3/18, p_3 = 1/18,$
$p_4 = 4/18, p_5 = 2/18, p_6 = 4/18$

👉 Laplace smoothing avoids the case where you estimate a parameter of 0.

Extra slides

Finding the MLE for Multinomial

MLE for Multinomial

Consider a sample of n i.i.d. random variables $Y_1, Y_2, \ldots, Y_n$.

- Let $Y_k \sim \text{Multinomial}(p_1, p_2, \ldots, p_m)$, where $\sum_{i=1}^{m} p_i = 1$
- Let X_i = # of trials with outcome i, where $\sum_{i=1}^{m} X_i = n$

Joint PDF $f(X_1, X_2, \ldots, X_m | p_1, p_2, \ldots, p_m) = \dfrac{n!}{x_1! \, x_2! \cdots x_m!} p_1^{x_1} p_2^{x_2} \cdots p_m^{x_m} = L(\theta)$

Log-likelihood:

$$LL(\theta) = \log(n!) - \sum_{i=1}^{m} \log(X_i!) + \sum_{i}^{m} X_i \log(p_i), \quad \text{such that } \sum_{i=1}^{m} p_i = 1$$

Optimize with Lagrange multipliers in extra slides → θ_{MLE}: $p_i = \dfrac{X_i}{n}$

Intuitively, probability p_i = proportion of outcomes

Optimizing MLE for Multinomial

$$\theta = (p_1, p_2, \dots, p_m)$$

$$\theta_{MLE} = \arg\max_{\theta} LL(\theta)$$, where $\sum_{i=1}^{m} p_i = 1$

Use Lagrange multipliers to account for constraint

Lagrange multipliers:

$$A(\theta) = LL(\theta) + \lambda\left(\sum_{i=1}^{m} p_i - 1\right) = \sum_{i} X_i \log(p_i) + \lambda\left(\sum_{i=1}^{m} p_i - 1\right)$$

(drop non-p_i terms)

Differentiate w.r.t. each p_i, in turn:

$$\frac{\partial A(\theta)}{\partial p_i} = X_i \frac{1}{p_i} + \lambda = 0 \quad \Rightarrow \quad p_i = -\frac{X_i}{\lambda}$$

Solve for λ, noting $\sum_{i=1}^{m} X_i = n, \sum_{i=1}^{m} p_i = 1$:

$$\sum_{i=1}^{m} p_i = \sum_{i=1}^{m} -\frac{X_i}{\lambda} = 1 \quad \Rightarrow 1 = -\frac{n}{\lambda} \quad \Rightarrow \lambda = -n$$

Substitute λ into p_i

$$p_i = \frac{X_i}{n}$$