# 22: MAP

Lisa Yan and Jerry Cain

November 2, 2020

# Quick slide reference

# Maximum a Posteriori Estimator

# Maximum Likelihood Estimator

Review

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

Maximum Likelihood Estimator (MLE)

What is the parameter $\theta$ that **maximizes the likelihood** of our observed data $(X_1, X_2, \ldots, X_n)$?

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

$$\theta_{MLE} = \arg\max_{\theta} \underbrace{f(X_1, X_2, \ldots, X_n | \theta)}_{\text{likelihood of data}}$$

Observations:

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

Today: **Bayesian estimation** using the Bayesian definition of probability!

# Maximum A Posteriori (MAP) Estimator

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

**Maximum Likelihood Estimator (MLE)**

What is the parameter $\theta$ that **maximizes the likelihood** of our observed data $(X_1, X_2, \ldots, X_n)$?

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

$$\theta_{MLE} = \arg\max_{\theta} \underbrace{f(X_1, X_2, \ldots, X_n | \theta)}_{\text{likelihood of data}}$$

**Maximum a Posteriori (MAP) Estimator**

Given our observed data $(X_1, X_2, \ldots, X_n)$, what is the **most likely parameter** $\theta$?

$$\theta_{MAP} = \arg\max_{\theta} \underbrace{f(\theta | X_1, X_2, \ldots, X_n)}_{\text{posterior distribution of } \theta}$$

# Maximum A Posteriori (MAP) Estimator

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

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

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

Intuition with Bayes' Theorem:

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

After seeing data, posterior belief of $\theta$

likelihood prior

posterior

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

Before seeing data, prior belief of $\theta$

# Solving for $\theta_{MAP}$

- 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 $\theta$ 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. $\theta$)

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

# $\theta_{MAP}$: Interpretation 1

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

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

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

$$= \arg\max_{\theta} g(\theta) \prod_{i=1}^{n} f(X_i|\,\theta)$$ ($1/h(X_1, X_2, \ldots, X_n)$ is a positive constant w.r.t. $\theta$)

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

$\theta_{MAP}$ maximizes
log prior + log-likelihood

# $\theta_{MAP}$: Interpretation 2

- 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 $\theta$ be $g(\theta)$.

$$\theta_{MAP} = \arg\max_{\theta} f(\theta|X_1, X_2, \dots, X_n) \quad = \text{ar}$$

The mode of the posterior distribution of $\theta$

(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. $\theta$)

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

$\theta_{MAP}$ maximizes
log prior + log-likelihood

# Mode: A statistic of a random variable

The **mode** of a random variable $X$ is defined as:

(X discrete, PMF $p(x)$) $$\arg\max_x p(x)$$

$$\arg\max_x f(x)$$ ($X$ continuous, PDF $f(x)$)

- Intuitively: The value of $X$ that is "most likely."
- Note that some distributions may not have a unique mode (e.g., Uniform distribution, or Bernoulli(0.5))

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

$\theta_{MAP}$ is the most likely $\theta$ given the data $X_1, X_2, \ldots, X_n$.

# Bernoulli MAP: Choosing a prior

# How does MAP work? (for Bernoulli)

Observe data — $n$ heads, $m$ tails

Choose model — Bernoulli($p$)

Choose prior on $\theta$ — (some $g(\theta)$)

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

maximize
log prior + log-likelihood

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

- Differentiate, set to 0
- Solve

A lot of our effort in MAP depends on the $g(\theta)$ we choose.

# MAP for Bernoulli

- Flip a coin 8 times. Observe $n = 7$ heads and $m = 1$ tail.
- Choose a prior on $\theta$. What is $\theta_{MAP}$?

Suppose we pick a prior $\theta \sim \mathcal{N}(0.5, 1^2)$. $g(\theta) = \frac{1}{\sqrt{2\pi}} e^{-(p-0.5)^2/2}$

1. Determine log prior + log likelihood

$$\log g(\theta) + \log f(X_1, X_2, \ldots, X_n|\theta)$$

$$= \log\left(\frac{1}{\sqrt{2\pi}} e^{-(p-0.5)^2/2}\right) + \log\left(\binom{n+m}{n} p^n (1-p)^m\right)$$

$$= -\log\left(\sqrt{2\pi}\right) - (p-0.5)^2/2 + \log\binom{n+m}{n} + n\log p + m\log(1-p)$$

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

$$-(p-0.5) + \frac{n}{p} - \frac{m}{1-p} = 0$$

3. Solve resulting equations

cubic equations why

We should choose an "easier" prior. This one is hard!

# A better approach: Use conjugate distributions

Observe data — $n$ heads, $m$ tails

Choose model — Bernoulli$(p)$

Choose prior on $\theta$ — (some $g(\theta)$)

(choose conjugate distribution)

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

maximize
log prior + log-likelihood

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

- Differentiate, set to 0
- Solve

Up next: Conjugate priors are great for MAP!

# Bernoulli MAP: Conjugate prior

# Beta is a conjugate distribution for Bernoulli



Beta is a conjugate distribution for Bernoulli, meaning:

- Prior and posterior parametric forms are the same
- Practically, conjugate means easy update:
  Add numbers of "successes" and "failures" seen to Beta parameters.
- You can set the prior to reflect how fair/biased you think the experiment is apriori.

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

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

(we'll prove this in a few minutes)

Beta parameters $a$, $b$ are called **hyperparameters.**
Interpret Beta$(a, b)$: $a + b - 2$ trials,
of which $a - 1$ are successes

# How does MAP work? (for Bernoulli)

Observe data — $n$ heads, $m$ tails

Choose model — Bernoulli$(p)$

Choose prior on $\theta$ — (some $g(\theta)$) — (choose conjugate distribution)

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

maximize
log prior + log-likelihood

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

- Differentiate, set to 0
- Solve

Mode of posterior distribution of $\theta$

(posterior is also conjugate)

# Conjugate strategy: MAP for Bernoulli

- Flip a coin 8 times. Observe $n = 7$ heads and $m = 1$ tail.
- Choose a prior on $\theta$. What is $\theta_{MAP}$?

Define as data, $D$

1. Choose a prior — Suppose we pick a prior $\theta \sim \text{Beta}(a, b)$.
2. Determine posterior — Because Beta is a conjugate distribution for Bernoulli, the posterior distribution is $\theta | D \sim \text{Beta}(a + n, b + m)$
3. Compute MAP

$$\theta_{MAP} = \frac{a + n - 1}{a + n + b + m - 2}$$

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

# MAP in practice

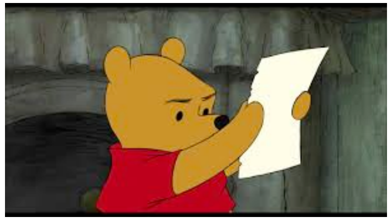
- Flip a coin 8 times. Observe $n = 7$ heads and $m = 1$ tail.
- What is the MAP estimator of the Bernoulli parameter $p$, if we assume a prior on $p$ of $\text{Beta}(2, 2)$?

# MAP in practice

- Flip a coin 8 times. Observe $n = 7$ heads and $m = 1$ tail.
- What is the MAP estimator of the Bernoulli parameter $p$, if we assume a prior on $p$ of $\text{Beta}(2, 2)$?

1. Choose a prior

$\theta \sim \text{Beta}(2,2)$.



Before flipping the coin, we imagined 2 trials: 1 imaginary head, 1 imaginary tail.

2. Determine posterior

Posterior distribution of $\theta$ given observed data is $\text{Beta}(9, 3)$

3. Compute MAP

$$\theta_{MAP} = \frac{8}{10}$$

After the coin, we saw 10 trials:
8 heads (imaginary and real),
2 tails (imaginary and real).

# Proving the mode of Beta

Observe data — $n$ heads, $m$ tails

Choose model — Bernoulli$(p)$

Choose prior on $\theta$ — (some $g(\theta)$) | (choose conjugate) Beta$(a, b)$

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

maximize
log prior + log-likelihood

Mode of posterior distribution of $\theta$

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

- Differentiate, set to 0
- Solve

(posterior is also conjugate)

These are equivalent interpretations of $\theta_{MAP}$. We'll use this equivalence to prove the mode of Beta.

# From first principles: MAP for Bernoulli, conjugate prior

- Flip a coin 8 times. Observe $n = 7$ heads and $m = 1$ tail.
- Choose a prior on $\theta$. What is $\theta_{MAP}$?

Suppose we pick a prior $\theta \sim \text{Beta}(a, b)$. $g(\theta = p) = \frac{1}{\beta} p^{a-1}(1-p)^{b-1}$

normalizing constant, $\beta$

1. Determine log prior + log likelihood

$$\log g(\theta) + \log f(X_1, X_2, \dots, X_n | \theta) = \log\left(\frac{1}{\beta} p^{a-1}(1-p)^{b-1}\right) + \log\left(\binom{n+m}{n} p^n (1-p)^m\right)$$

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

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

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

3. Solve

(next slide)

# From first principles: MAP for Bernoulli, conjugate prior

- Flip a coin 8 times. Observe $n = 7$ heads and $m = 1$ tail.
- Choose a prior $\theta$. What is $\theta_{MAP}$?

Suppose we pick a prior $\theta \sim \text{Beta}(a, b)$. $g(\theta) = \frac{1}{\beta} p^{a-1}(1-p)^{b-1}$

normalizing constant, $\beta$

3. Solve for $p$

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

(from previous slide)

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

$$\theta_{MAP} = \frac{a+n-1}{a+n+b+m-2}$$ ✅

The mode of the posterior, $\text{Beta}(a+n, b+m)$!

If we choose a conjugate prior, we avoid calculus with MAP: just report mode of posterior.

# (live)

# 22: MAP

Lisa Yan and Jerry Cain
November 2, 2020

# Maximum A Posteriori (MAP) Estimator

Review

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

**Maximum Likelihood Estimator (MLE)**

What is the parameter $\theta$ that **maximizes the likelihood** of our observed data $(X_1, X_2, \dots, X_n)$?

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

$$\theta_{MLE} = \arg\max_{\theta} \underbrace{f(X_1, X_2, \dots, X_n|\theta)}_{\text{likelihood of data}}$$

**Maximum a Posteriori (MAP) Estimator**

Given our observed data $(X_1, X_2, \dots, X_n)$, what is the **most likely parameter** $\theta$?

$$\theta_{MAP} = \arg\max_{\theta} \underbrace{f(\theta|X_1, X_2, \dots, X_n)}_{\text{posterior distribution of } \theta}$$

# How does MAP work?



Observe data

Choose model with parameter $\theta$

Choose prior on $\theta$

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

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

Two valid approaches to computing $\theta_{MAP}$

Mode of posterior distribution of $\theta$

or

maximize log prior + log-likelihood

If we choose a conjugate prior, we avoid calculus with MAP: just report mode of posterior.

# Conjugate distributions

# Quick MAP for Bernoulli and Binomial



Beta$(a, b)$ is a conjugate prior for the probability of success in Bernoulli and Binomial distributions.

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

**Prior** Beta$(a, b)$
Saw $a + b - 2$ imaginary trials: $a - 1$ successes, $b - 1$ failures

**Experiment** Observe $n + m$ new trials: $n$ successes, $m$ failures

**Posterior** Beta$(a + n, b + m)$

MAP:

$$p = \frac{a + n - 1}{a + b + n + m - 2}$$

# Conjugate distributions

MAP estimator:

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

The mode of the posterior distribution of $\theta$

| Distribution parameter | Conjugate distribution |
|---|---|
| Bernoulli $p$ | Beta |
| Binomial $p$ | Beta |
| Multinomial $p_i$ | Dirichlet |
| Poisson $\lambda$ | Gamma |
| Exponential $\lambda$ | Gamma |
| Normal $\mu$ | Normal |
| Normal $\sigma^2$ | Inverse Gamma |

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

CS109: We'll only focus on MAP for Bernoulli/Binomial $p$, Multinomial $p_i$, and Poisson $\lambda$.

# Multinomial is Multiple times the fun

Dirichlet$(a_1, a_2, \ldots, a_m)$ is a 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$(a_1, a_2, \ldots, a_m)$
Saw $\left(\sum_{i=1}^{m} a_i\right) - m$ imaginary trials, with $a_i - 1$ of outcome $i$

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

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

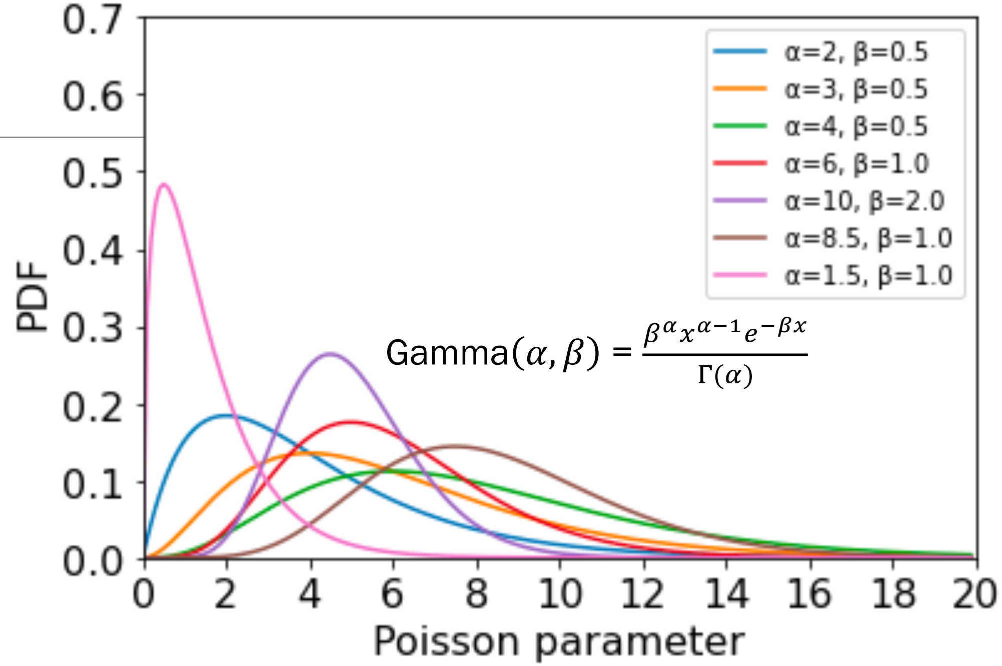
MAP:

$$p_i = \frac{a_i + n_i - 1}{\left(\sum_{i=1}^{m} a_i\right) + \left(\sum_{i=1}^{m} n_i\right) - m}$$

# Good times with Gamma

Gamma$(\alpha, \beta)$ is a conjugate for Poisson.

- Also conjugate for Exponential, but we won't delve into that
- Mode of gamma: $(\alpha - 1)/\beta$




**Prior** $\theta \sim \text{Gamma}(\alpha, \beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}$

Saw $\alpha - 1$ total imaginary events during $\beta$ prior time periods

**Experiment** Observe $n$ events during next $k$ time periods

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

MAP: $$\theta_{MAP} = \frac{a + n - 1}{\beta + k}$$

# MAP for Poisson

$\text{Gamma}(\alpha, \beta)$ is conjugate for Poisson

Mode: $\frac{\alpha-1}{\beta}$

Let $\lambda$ be the average # of successes in a time period.

1. What does it mean to have a prior of $\theta \sim \text{Gamma}(11,5)$?

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 $\theta_{MAP}$?

# MAP for Poisson

$\text{Gamma}(\alpha, \beta)$ is conjugate for Poisson  Mode: $\frac{\alpha-1}{\beta}$

Let $\lambda$ be the average # of successes in a time period.

1. What does it mean to have a prior of $\theta \sim \text{Gamma}(11,5)$?

   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$) \sim \text{Gamma}(22, 7)$

3. What is $\theta_{MAP}$?

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

# Interlude for jokes/announcements

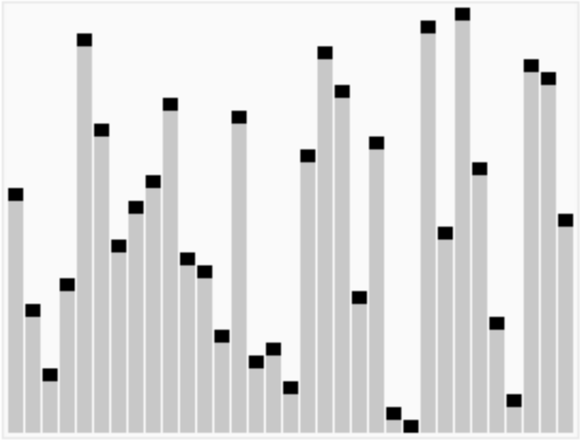
# Announcements

Quiz 2 Grades Released Soon

Wednesday's Lecture: Optional

No Discussion Section This Week!

Lisa and I Still Have Wednesday OH!

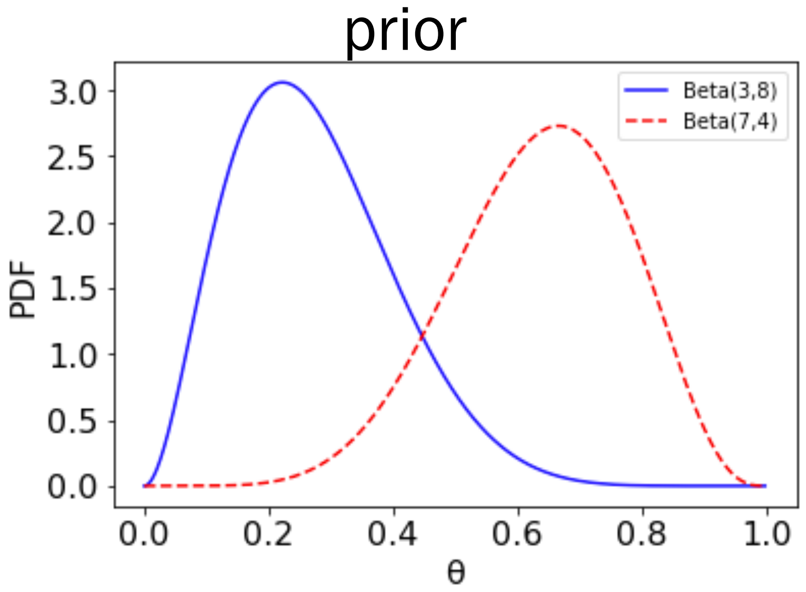


https://en.wikipedia.org/wiki/Quicksort

# Choosing hyperparameters for conjugate prior

# Where'd you get them priors?

- Let $\theta$ be the probability a coin turns up heads.
- Model $\theta$ 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: $\frac{6}{9}$

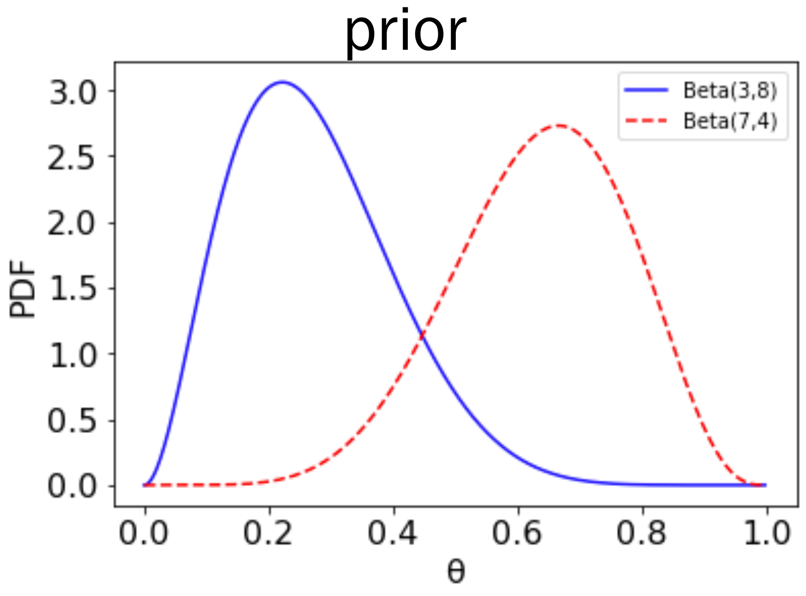



Now flip 100 coins and get 58 heads and 42 tails.

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

# Where'd you get them priors?

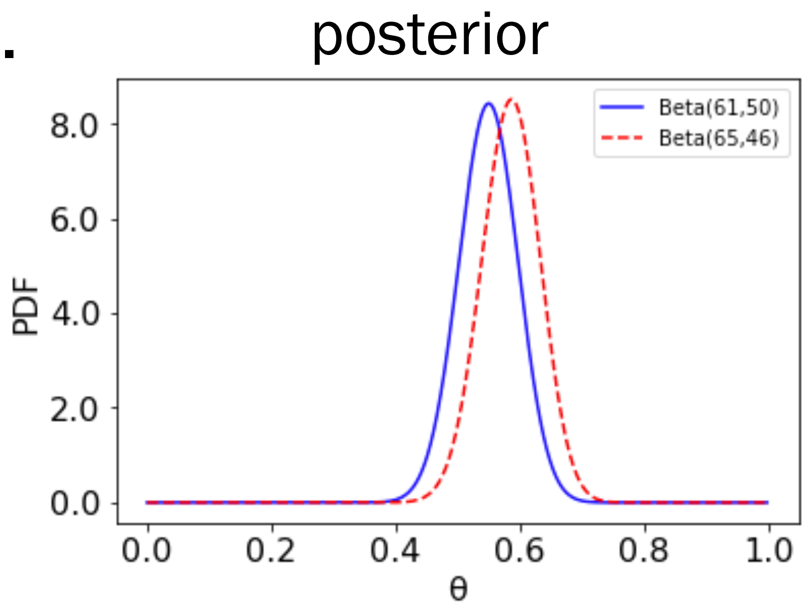
- Let $\theta$ be the probability a coin turns up heads.
- Model $\theta$ with 2 different priors:
  - Prior 1: Beta(3,8): 2 imaginary heads, 7 imaginary tails  mode: $\frac{2}{9}$
  - Prior 2: Beta(7,4): 6 imaginary heads, 3 imaginary tails  mode: $\frac{6}{9}$




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

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

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




Provided we collect enough data, posteriors will converge to the true value.

# Laplace smoothing

MAP with **Laplace smoothing**: a prior which represents $k$ imagined observations of each outcome.

- Categorical data (i.e., Multinomial, Bernoulli/Binomial)
- Also known as additive smoothing

**Laplace estimate** Imagine $k = 1$ of each outcome (follows from Laplace's "law of succession")

Example: Laplace estimate for coin probabilities from aforementioned experiment (100 coins: 58 heads, 42 tails)

heads $\frac{59}{102}$ tails $\frac{43}{102}$

Laplace smoothing:
- Easy to implement/remember

# Back to our happy Laplace

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 $\theta_{MLE}$: $p_1 = 3/12, p_2 = 2/12, p_3 = 0/12,$ ⚠️
$p_4 = 3/12, p_5 = 1/12, p_6 = 3/12$

What are your Laplace estimates for each roll outcome?

🤔

# Back to our happy Laplace

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 $\theta_{MLE}$: $p_1 = 3/12, p_2 = 2/12, p_3 = 0/12,$ ⚠️
$p_4 = 3/12, p_5 = 1/12, p_6 = 3/12$

What are your Laplace estimates for each roll outcome?

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

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

Laplace smoothing:
- Easy to implement/remember
- **Avoids estimating a parameter of 0**

LIVE

# Bayesian Envelope Demo

# Two envelopes

Two envelopes: One contains $\$X$, the other contains $\$2X$.

- Select an envelope and open it.
- Before opening the envelope, think either equally good.

Is the following reasoning valid?

- Let $Y = \$$ in envelope you selected.
- Let $Z = \$$ in other envelope.

$$E[Z|Y] = \frac{1}{2} \cdot \frac{Y}{2} + \frac{1}{2} \cdot 2Y = \frac{5}{4} Y$$

Follow-up: What happened by opening the envelope?

# Two envelopes

Two envelopes: One contains $\$X$, the other contains $\$2X$.

- Select an envelope and open it.
- Before opening the envelope, think either equally good.

Is the following reasoning valid?

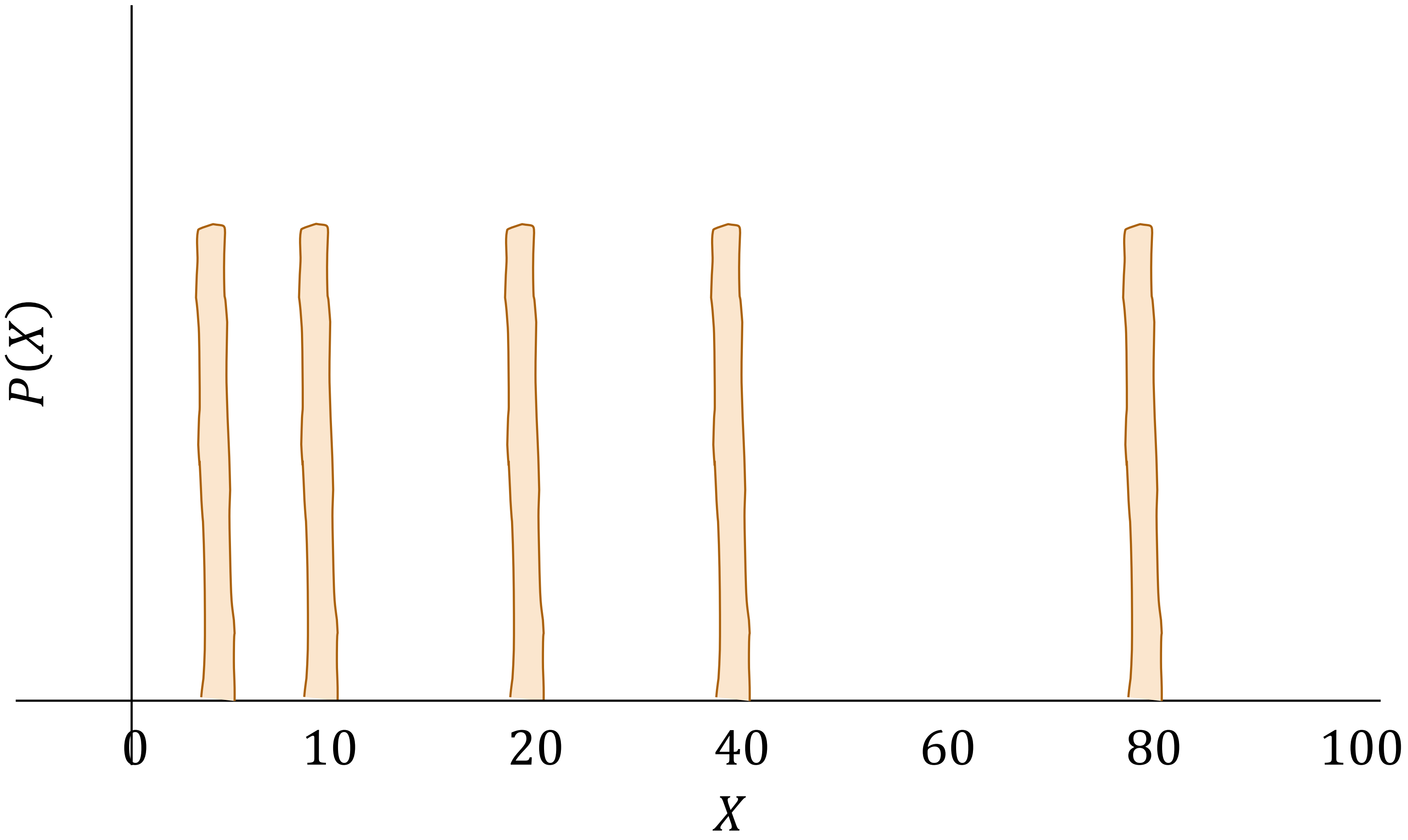
- Let $Y$ = \$ in envelope you selected.
- Let $Z$ = \$ in other envelope.

$$E[Z|Y] = \frac{1}{2} \cdot \frac{Y}{2} + \frac{1}{2} \cdot 2Y = \frac{5}{4} Y$$

- Assumes all values of $X$ (where $0 < X < \infty$) equally likely
- Infinitely many values of $X$
- So not a true probability distribution over $X$ (does not integrate to 1)

Follow-up: What happened by opening the envelope?

# Are all values equally likely?




Infinite powers of two times 10

# Two envelopes: The subjectivity of probability

Your belief about the content of envelopes:

- Since implied distribution over $X$ is not a true probability distribution, what *is* our distribution over $X$?

## Frequentist

Play game infinitely many times, see how often different values come up

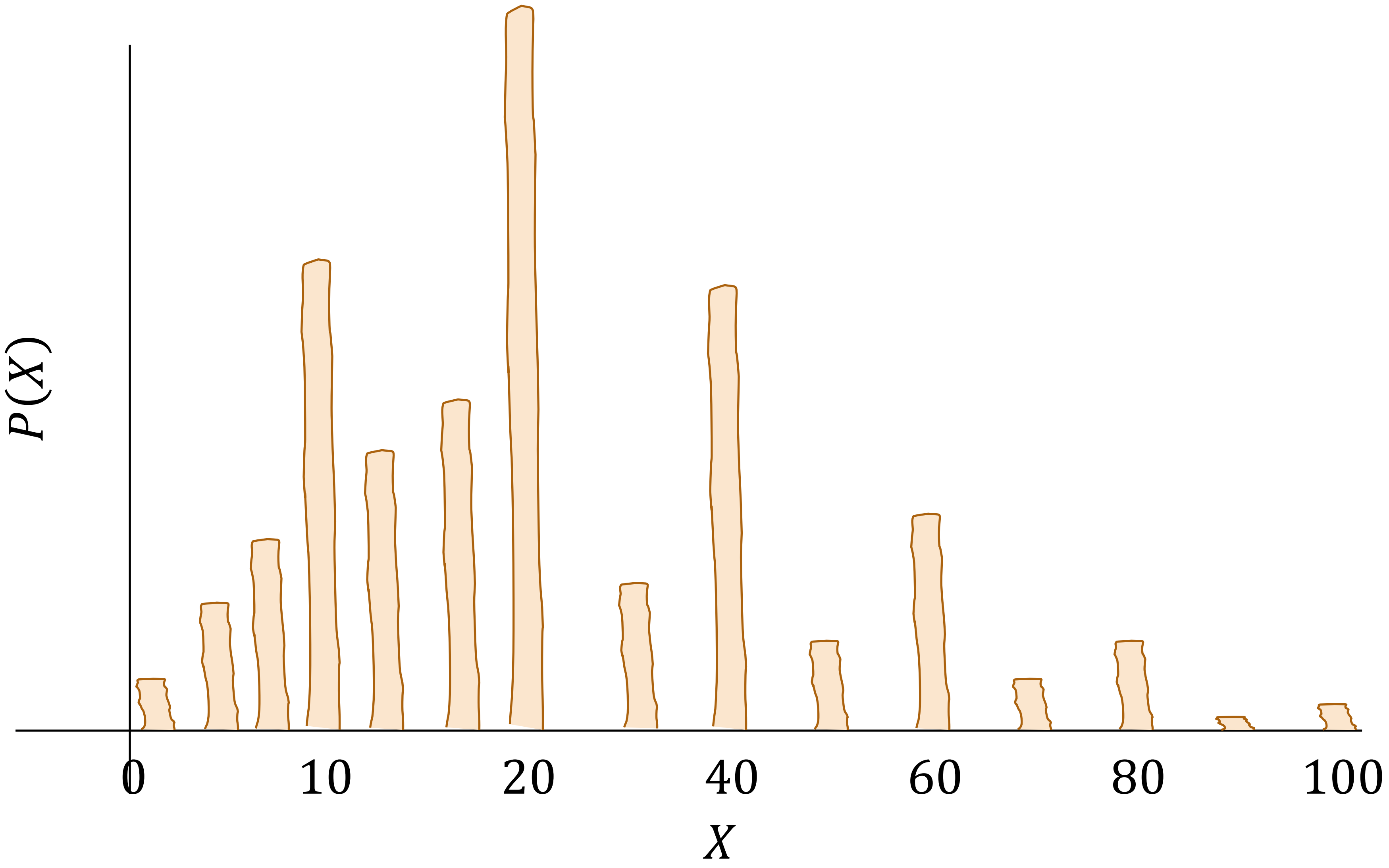
Dilemma: You can only play the game once!

## Bayesian

Have prior belief of distribution of $X$

- Prior belief is a *subjective* probability
- Allows us to answer questions with limited data, or even no data at all
- As we run more experiments, all prior beliefs are eclipsed by data

# Two envelopes: The subjectivity of probability



# The envelope, please

Bayesian: Have a prior distribution over $X, P(X)$

- Let $Y$ = \$ in envelope you selected. Open envelope to determine $Y$.
- Let $Z$ = \$ in other envelope.

If $Y > E[Z|Y]$, keep your envelope, otherwise switch. No inconsistency!!

- Opening envelope provides data to compute $P(X|Y)$
- ...which allows you to compute $E[Z|Y]$

Of course, need to think about your prior distribution over $X$, but...

Bayesian probability: It doesn't matter how you construct your prior, but you **must** have one (whatever it is)

Imagine if envelope you opened contained $\$20.01$. Should you switch?

# How much is a half cent?

Have a wonderful Monday!