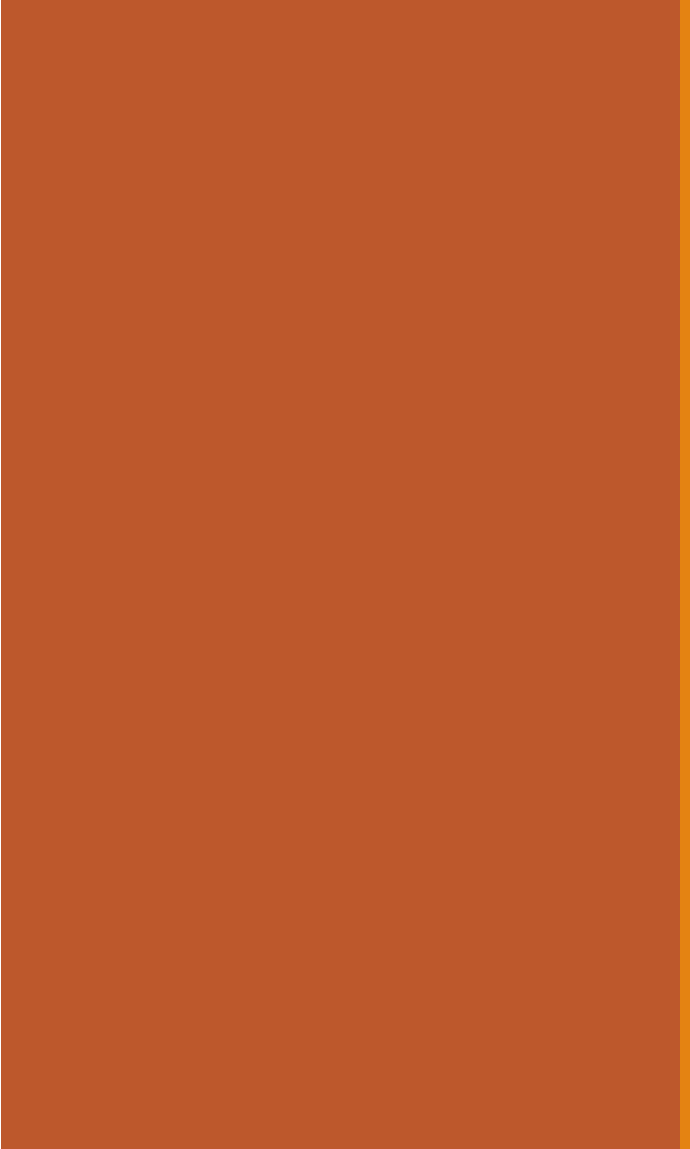




# **CS109: Maximum A Posteriori**



# Maximum a Posteriori Estimator

# Maximum Likelihood 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  $(X_1, X_2, \dots, X_n)$ ?

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

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

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  $X_1, X_2, \dots, X_n$  (data).

Maximum  
Likelihood  
Estimator  
(MLE)

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

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

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

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**  $\theta$ ?

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

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).

def 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:

Diagram illustrating Bayes' Theorem with callouts:

- Callout (blue oval): After seeing data, posterior belief of  $\theta$
- Callout (orange oval):  $L(\theta)$ , probability of data given parameter  $\theta$
- Callout (red oval): Before seeing data, prior belief of  $\theta$

Bayes' Theorem equation:

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

# 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  $g(\theta)$ .

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

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

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

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



## $\theta_{MAP}$ : Interpretation 1

- Observe data:  $X_1, X_2, \dots, X_n$ , 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  $g(\theta)$ .

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

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

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

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

$\theta_{MAP}$  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  $g(\theta)$ .

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

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:

$$\begin{array}{ll} (X \text{ discrete,} & \arg \max_x p(x) \\ \text{PMF } p(x)) & \end{array} \qquad \begin{array}{ll} \arg \max_x f(x) & (X \text{ continuous,} \\ \text{PDF } f(x)) & \end{array}$$

- 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))

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

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



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

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

$$\begin{aligned} \log g(\theta) + \log f(X_1, X_2, \dots, X_n | \theta) \\ &= \log \left( \frac{1}{\sqrt{2\pi}} e^{-(p-0.5)^2/2} \right) + \log \left( \binom{n+m}{n} p^n (1-p)^m \right) \\ &= -\log(\sqrt{2\pi}) - (p-0.5)^2/2 + \log \binom{n+m}{n} + n \log p + m \log(1-p) \end{aligned}$$

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 equation, not cute

We should choose a prior belief distribution that plays nice.

# A better approach: Use conjugate distributions

Observe data

Choose model

Choose **prior on  $\theta$**

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

$n$  heads,  $m$  tails

Bernoulli( $p$ )

(some  $g(\theta)$ )

(choose **conjugate**  
distribution)

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

Review

Beta is a **conjugate distribution** for Bernoulli, meaning:

- Parametric forms of prior and posterior 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( $a = n_{imag} + n + 1, b = m_{imag} + m + 1$ )

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

(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

Choose model

Choose **prior on  $\theta$**

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

$n$  heads,  $m$  tails

Bernoulli( $p$ )

(some  $g(\theta)$ )

(choose conjugate distribution)

maximize  
log prior + log-likelihood

Mode of posterior  
distribution of  $\theta$

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

(posterior is also conjugate)

- Differentiate, set to 0
- Solve



# Conjugate strategy: MAP for Bernoulli

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

## 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

$$\theta_{MAP} = \frac{a + n - 1}{a + n + b + m - 2} \quad (\text{mode of } \text{Beta}(a + n, b + m))$$



# 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 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 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 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

Choose model

Choose prior on  $\theta$

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

These are **equivalent interpretations** of  $\theta_{MAP}$ .

We'll use this equivalence to prove the mode of Beta.

$n$  heads,  $m$  tails

Bernoulli( $p$ )

(some  $g(\theta)$ )

(choose conjugate)

Beta( $a, b$ )

maximize  
**log prior + log-likelihood**

**Mode of posterior distribution of  $\theta$**

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

(posterior is also conjugate)

- Differentiate, set to 0
- Solve

# From first principles: MAP for Bernoulli, conjugate prior

- Flip a coin  $n + m$  times. Observe  $n$  heads and  $m$  tails.
- 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

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

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  $n + m$  times. Observe  $n$  heads and  $m$  tails.
- Choose a prior on  $\theta$ . 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  $p$  
$$\frac{a-1}{p} + \frac{n}{p} - \frac{b-1}{1-p} - \frac{m}{1-p} = 0 \quad (\text{from previous slide})$$

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

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



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.

# 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  
( $X_1, X_2, \dots, X_n$ )?

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

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

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**  $\theta$ ?

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

posterior distribution  
of  $\theta$

# How does MAP work?

Review

Observe data

Choose model with parameter  $\theta$

Choose **prior on  $\theta$**

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

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

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

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.



A large orange rectangle with a thin yellow border on the right side, positioned on the left side of the slide.

# Conjugate distributions

# Quick MAP for Bernoulli and Binomial

Review

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( $a_1, a_2, \dots, a_m$ ) is a conjugate for Multinomial.

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

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

**Prior**

Dirichlet( $a_1, a_2, \dots, a_m$ )

Saw  $(\sum_{i=1}^m a_i) - m$  imaginary trials, with  $a_i - 1$  of outcome  $i$

**Experiment**

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

**Posterior**

Dirichlet( $a_1 + n_1, a_2 + n_2, \dots, a_m + n_m$ )

MAP:

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

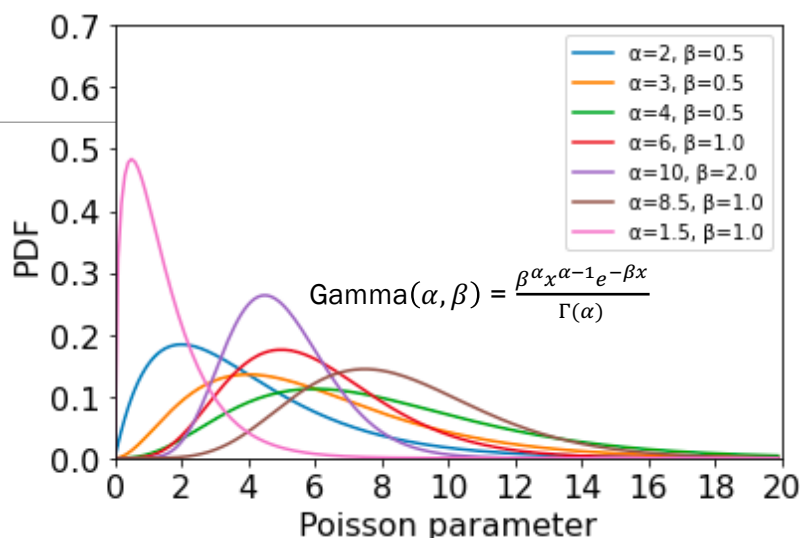
Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain CS109, Winter 2021



# 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:

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

# MAP for Poisson

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

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?

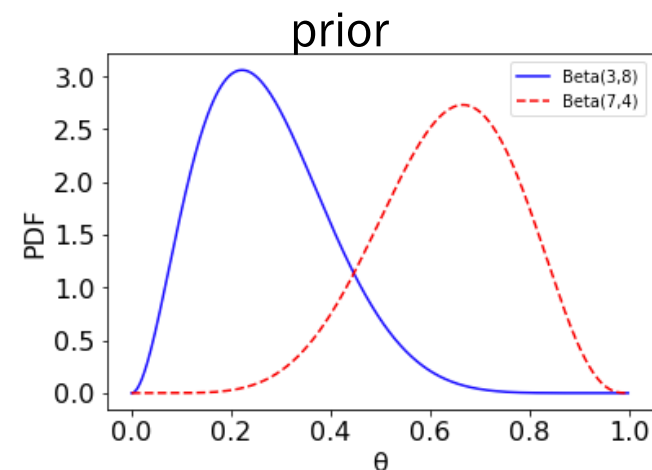
$(\theta | n \text{ events in } k \text{ periods}) \sim \text{Gamma}(22, 7)$

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

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

# 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.

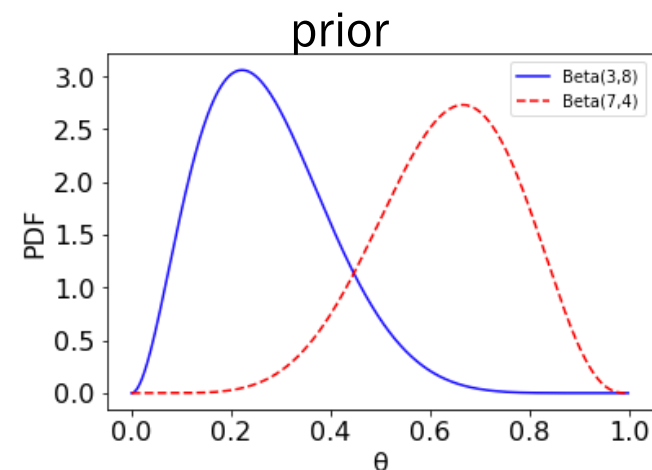
1. What are the two posterior distributions?
2. What are the modes of the two posterior distributions?





# Where'd you get them priors?

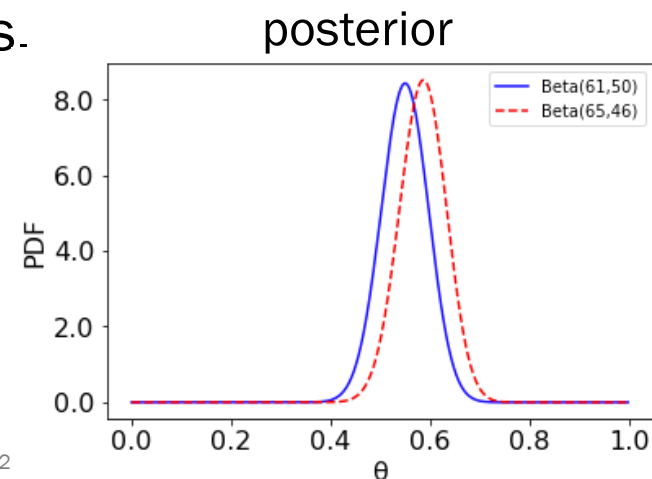
- Let  $\theta$  be the probability a coin turns up heads.
- Model  $\theta$  with 2 different priors:
  - Prior 1: **Beta(3,8)**: 2 imaginary heads, 7 imaginary tails      mode:  $\frac{2}{9}$
  - Prior 2: **Beta(7,4)**: 6 imaginary heads, 3 imaginary tails      mode:  $\frac{6}{9}$



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

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

What are your Laplace estimates for each roll outcome?

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

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

Laplace smoothing:

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