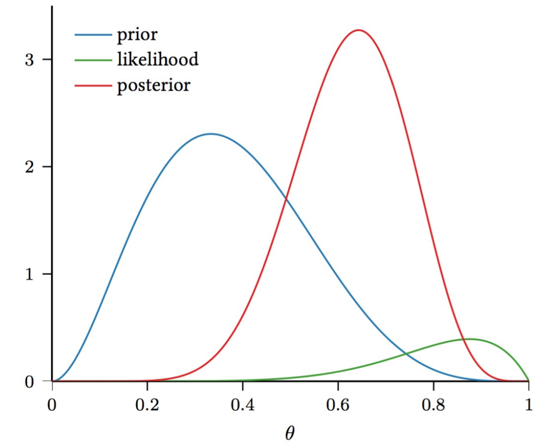
Table of Contents



22: MAP

Jerry Cain

March 1, 2024

[Lecture Discussion on Ed](#)

Maximum a Posteriori Estimator

Maximum Likelihood Estimator

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

Maximum Likelihood Estimator (MLE)

What parameter θ **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 determines θ value that maximizes the probability of observing the sample.
- If we're estimating θ, couldn't we just maximize the probability of θ?

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

Maximum A Posteriori (MAP) Estimator

Not Review! New!

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

Maximum Likelihood Estimator (MLE)

What parameter θ **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 the sample data $(X_1, X_2, \dots, X_n)$, what is the **most probable parameter** θ?

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

Maximum A Posteriori (MAP) Estimator

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

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

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

notice that both the prior and the posterior focus on θ as primary variables.

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

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

Solving for θ_{MAP}

- Observe data: $X_1, X_2, \ldots, X_n$, all iid
- 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 θ be $g(\theta)$. intentionally generic at this point

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

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

θ_{MAP}: Interpretation 1

- Observe data: $X_1, X_2, \dots, X_n$, all iid
- 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)

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

θ_{MAP} maximizes
log prior + log-likelihood

θ_{MAP}: Interpretation 2

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

that is what this really is

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

The mode of the posterior distribution of θ

(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)

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

θ_{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))

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

this is another maximization problem that generally requires we set derivatives to 0 and solve.

θ_{MAP} is the most likely θ 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 θ (some $g(\theta)$)

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

maximize
log prior + log-likelihood

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

- Differentiate, set to 0
- Solve

MAP depends on what $g(\theta)$ we choose.

MAP for Bernoulli

not unreasonable! it's centered around 0.5 but allows for the actual probability to be slightly higher or lower.

- Flip a coin 8 times. Observe $n = 7$ heads and $m = 1$ tail.
- Choose a prior on θ. What is θ_{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

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

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

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

2. Differentiate wrt (each) θ, set to 0

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

We should choose a prior that's easier to deal with. This one is hard!

reasonable counterargument: can we justify a decision to choose $g(\theta)$ that's easier to deal with?

3. Solve resulting equations

cubic equations, nope not going to do it

A better approach: Use conjugate distributions

Observe data — n heads, m tails

Choose model — Bernoulli(p)

if the choice is really up to us, choose something that's easy to manipulate while still doing a good job of modelling the prior.

Choose prior on θ — (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

Mostly Review

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 a priori.

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 θ — (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 θ

(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 θ. What is θ_{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 $\text{Beta}(2, 2)$?

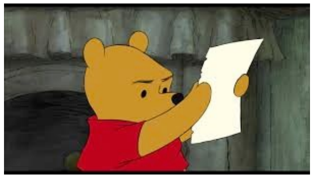
MAP in practice

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

mode of prior Beta(2,2)

$$\frac{2-1}{2+2-2} = \frac{1}{2}$$



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

9 is really $2 + 7$

3 is really $2 + 1$

2. Determine posterior Posterior distribution of θ given observed data is $\text{Beta}(9, 3)$

3. Compute MAP $\theta_{MAP} = \frac{8}{10}$

$$\text{mode} = \frac{9-1}{9+3-2} = \frac{8}{10}$$

After the experiment, 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 θ — (some arbitrary $g(\theta)$) — (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

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

- Differentiate, set to 0
- Solve

Mode of posterior distribution of θ

(posterior is also conjugate)

These are equivalent interpretations of θ_{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 θ. What is θ_{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, β

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) θ, 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 θ. What is θ_{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, β

3. Solve for p

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

(from previous slide)

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

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

$$\Longrightarrow p(a+n+b+m-2) = a+n-1$$

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

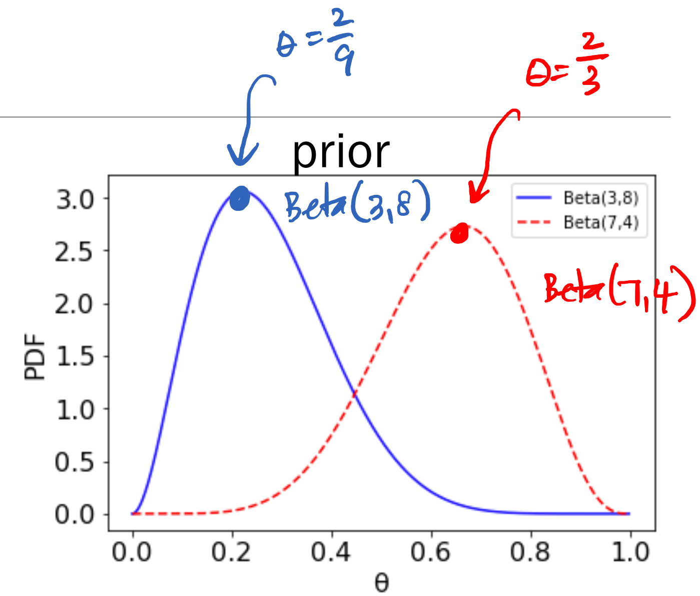
The mode of the posterior, $\text{Beta}(a+n, b+m)$!

If we choose a conjugate prior, we avoid calculus with MAP, and we can simply report mode of posterior.

Choosing hyperparameters for conjugate prior

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: $\frac{6}{9}$

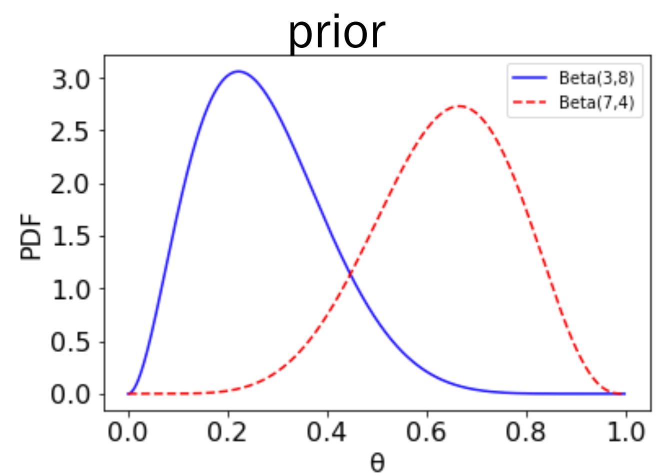


Now flip 100 coins and get 58 heads and 42 tails.

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

Where'd you get them priors?

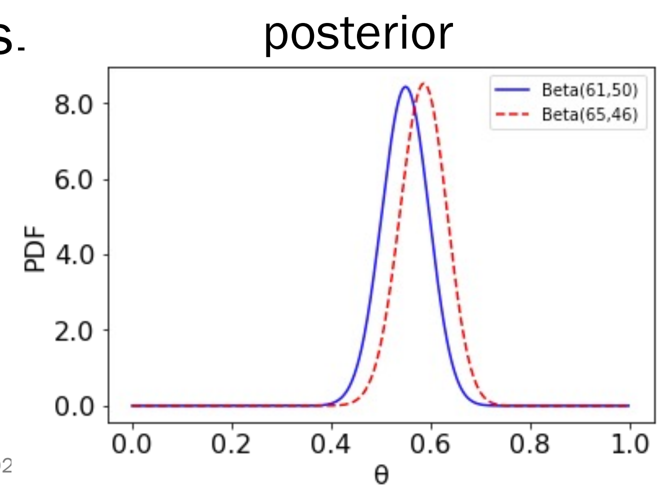
- Let θ be the probability a coin turns up heads.
- Model θ with 2 different priors:
    - Prior 1: Beta(3,8): 2 imaginary heads, 7 imaginary tails — mode: $\frac{2}{9}$
    - Prior 2: Beta(7,4): 6 imaginary heads, 3 imaginary tails — mode: $\frac{6}{9}$



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

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 and choice of priors will matter less and less.

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 probabilities from previously mentioned experiment (100 coins: 58 heads, 42 tails)

heads $\frac{59}{102} = \frac{58+1}{100+2}$ 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 θ_{MLE}: $p_1 = 3/12, p_2 = 2/12, p_3 = 0/12,$ $p_4 = 3/12, p_5 = 1/12, p_6 = 3/12$

the zero here is ruinous, because it prevents p3 from becoming anything else.

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

What are your Laplace estimates for each roll outcome?

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

$X_3 = 0 \Rightarrow \frac{0+1}{12+6}$

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

Laplace smoothing:

- Easy to implement/remember
- **Avoids parameter estimation of 0**

often important

Extra: Other Conjugates

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 θ

ideal model for prior and posterior

Distribution parameter	Conjugate distribution
Bernoulli p	Beta
Binomial p	Beta
Multinomial p_i	Dirichlet → generalization of Beta
Poisson λ	Gamma
Exponential λ	Gamma
Normal μ	Normal
Normal σ^2	Inverse Gamma

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

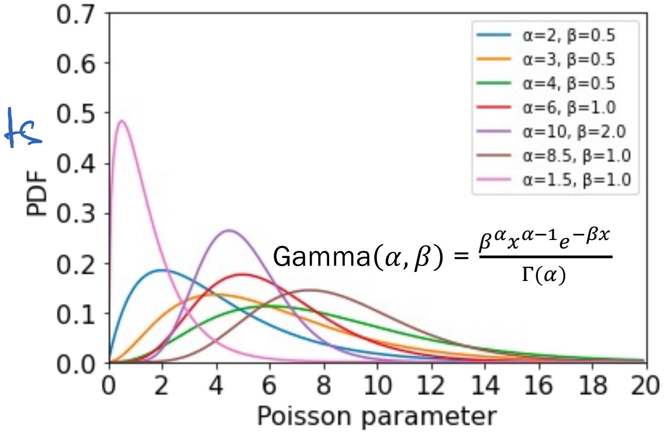
MAP:

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

Good times with Gamma

Gamma(α, β) is a conjugate for Poisson.

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



α : count events

β: tracks the time that's passed

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

Saw $\alpha - 1$ total imaginary events during β 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:

$$\theta_{MAP} = \frac{\overbrace{\alpha + n}^{\text{the new } \alpha} - 1}{\underbrace{\beta + k}_{\text{the new } \beta}}$$

MAP for Poisson

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

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

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

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

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 θ_{MAP}?

MAP for Poisson

Gamma(α, β) is conjugate for Poisson

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

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

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

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 θ_{MAP}?

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