22: Maximum a Posteriori

Jerry Cain
May 20, 2024

Lecture Discussion on Ed
Maximum a Posteriori Estimator
Maximum A Posteriori (MAP) Estimator

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

Maximum Likelihood Estimator (MLE)

What parameter $\theta$ maximizes the likelihood of our observed data $(X_1, X_2, ..., X_n)$?

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

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

Maximum a Posteriori (MAP) Estimator

Given the sample data $(X_1, X_2, ..., X_n)$, what is the most probable parameter $\theta$?

$$\theta_{MAP} = \arg \max_{\theta} f(\theta | X_1, X_2, ..., X_n)$$
Maximum A Posteriori (MAP) Estimator

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

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

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

Intuition with Bayes’ Theorem:

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

- $L(\theta)$, probability of data given parameter $\theta$
- **likelihood**
- **prior**
- Before seeing data, prior belief of $\theta$
- After seeing data, posterior belief of $\theta$
Solving for \( \theta_{MAP} \)

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

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

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

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

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

\( f(X_1, X_2, \ldots, X_n | \theta)g(\theta) \) is a positive constant w.r.t. \( \theta \)
\( \theta_{\text{MAP}} \): Interpretation 1

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

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

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

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

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

\( \theta_{\text{MAP}} \) maximizes \( \log \text{prior} + \log \text{-likelihood} \)
**θ_{MAP}: Interpretation 2**

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

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

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

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

- The mode of the posterior distribution of $\theta$

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

\[
\text{The mode of the posterior distribution of } \theta
\]

\[
\text{\(\theta_{MAP}\) maximizes log prior + log-likelihood}
\]
Mode: A statistic of a random variable

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

$$
\text{arg max } \ p(x) \quad \quad \quad \quad \quad \text{arg max } \ f(x)
$$

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

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

$\theta_{MAP}$ is the most probable $\theta$ 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 $\theta_{\text{MAP}} = \arg \max_{\theta} f(\theta | X_1, X_2, ..., X_n)$

maximize
$\log \text{prior} + \log\text{-likelihood}$

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

- Differentiate, set to 0
- Solve

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

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

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

1. Determine log prior + log likelihood
   \[
   \log g(\theta) + \log f(X_1, X_2, ..., X_n | \theta) = \log \left( \frac{1}{\sqrt{2\pi}} e^{-\frac{(p-0.5)^2}{2}} \right) + \log \left( \binom{n+m}{n} p^n (1-p)^m \right) 
   \]
   \[
   = -\log(\sqrt{2\pi}) - (p - 0.5)^2/2 + \log \left( \binom{n+m}{n} \right) + n \log p + m \log(1 - p) 
   \]

2. Differentiate wrt (each) \( \theta \), 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!

3. Solve resulting equations
   cubic equations, nope not going to do it
A better approach: Use conjugate distributions

Observe data

Choose model

Choose prior on $\theta$

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

$n$ heads, $m$ tails

Bernoulli($p$)

(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!

Lisa Yan, Chris Piech, Mehran Sahami, and Jerry Cain, CS109, Spring 2024

Stanford University
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 a priori.

<table>
<thead>
<tr>
<th>Prior</th>
<th>Beta(a = n_{imag} + 1, b = m_{imag} + 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment</td>
<td>Observe (n) successes and (m) failures</td>
</tr>
<tr>
<td>Posterior</td>
<td>Beta(a = n_{imag} + n + 1, b = m_{imag} + m + 1)</td>
</tr>
</tbody>
</table>

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,
with \(a - 1\) successes and with \(b - 1\) failures
How does MAP work? (for Bernoulli)

1. **Choose model**
   - Choose the Bernoulli model, $\text{Bernoulli}(p)$

2. **Observe data**
   - Observe the data, $n$ heads, $m$ tails

3. **Choose prior on $\theta$**
   - Choose the prior distribution, $g(\theta)$

4. **Find $\theta_{MAP}$**
   - Find $\theta_{MAP} = \arg \max_{\theta} f(\theta|X_1, X_2, ..., X_n)$
     - maximize $\log \text{prior} + \log \text{-likelihood}$
     - $\log g(\theta) + \sum_{i=1}^{n} \log f(X_i|\theta)$
     - Differentiate, set to 0
     - Solve

5. **Mode of posterior distribution of $\theta$**
   - Mode of posterior distribution of $\theta$
   - (posterior is also conjugate)

(choose conjugate distribution)
**Conjugate strategy: MAP for Bernoulli**

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

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

2. Determine posterior
   
   Posterior distribution of $\theta$ given observed data is Beta($9$, $3$)

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

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

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 $\theta$
(some arbitrary $g(\theta)$)

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

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

maximize
$$\log \text{prior} + \log\text{-likelihood}$$

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

- Differentiate, set to 0
- Solve

Mode of posterior distribution of $\theta$
(posterior is also conjugate)
MAP for Bernoulli, conjugate prior (from first principles)

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

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

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

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

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)
MAP for Bernoulli, conjugate prior (from first principles)

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

3. Solve for $p$

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

(from previous slide)

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

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

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

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

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

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

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

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

Provided we collect enough data, posteriors will converge to the true value and choice of prior will matter 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)

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Count</th>
<th>Laplace Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heads</td>
<td>59</td>
<td>$\frac{59}{102}$</td>
</tr>
<tr>
<td>Tails</td>
<td>43</td>
<td>$\frac{43}{102}$</td>
</tr>
</tbody>
</table>

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 \( \theta_{MLE} \):

\[
p_1 = \frac{3}{12}, p_2 = \frac{2}{12}, p_3 = \frac{0}{12}, \quad \text{⚠}
p_4 = \frac{3}{12}, p_5 = \frac{1}{12}, p_6 = \frac{3}{12}
\]

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 \( \theta_{MLE} \):
\[
p_1 = \frac{3}{12}, p_2 = \frac{2}{12}, p_3 = \frac{0}{12}, p_4 = \frac{3}{12}, p_5 = \frac{1}{12}, p_6 = \frac{3}{12}
\]

What are your Laplace estimates for each roll outcome?
\[
p_i = \frac{X_i + 1}{n + m}
\]
\[
p_1 = \frac{4}{18}, p_2 = \frac{3}{18}, p_3 = \frac{1}{18}, p_4 = \frac{4}{18}, p_5 = \frac{2}{18}, p_6 = \frac{4}{18}
\]

Laplace smoothing:
- Easy to implement/remember
- Avoids parameter estimation of 0
Extra: Other Conjugates
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 \( \theta \)

<table>
<thead>
<tr>
<th>Distribution parameter</th>
<th>Conjugate distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bernoulli ( p )</td>
<td>Beta</td>
</tr>
<tr>
<td>Binomial ( p )</td>
<td>Beta</td>
</tr>
<tr>
<td>Multinomial ( p_i )</td>
<td>Dirichlet</td>
</tr>
<tr>
<td>Poisson ( \lambda )</td>
<td>Gamma</td>
</tr>
<tr>
<td>Exponential ( \lambda )</td>
<td>Gamma</td>
</tr>
<tr>
<td>Normal ( \mu )</td>
<td>Normal</td>
</tr>
<tr>
<td>Normal ( \sigma^2 )</td>
<td>Inverse Gamma</td>
</tr>
</tbody>
</table>
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 \((\sum_{i=1}^{m} a_i) - m\) 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:**

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

\[
\theta_{MAP} = \frac{\alpha + n - 1}{\beta + k}
\]
MAP for Poisson

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

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

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

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

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.
MAP for Poisson

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

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

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

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

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

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

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

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