Gradient Ascent
Noah Arthurs
CS109, Stanford University
Review
Supervised Learning

Real World Problem

Model the problem

Formal Model $\theta$

Learning Algorithm

Testing Data

Prediction Function $\theta^*$

Training Data

Evaluation score
Modelling

Real World Problem

Model the problem

Formal Model $\theta$

Learning Algorithm

Prediction Function $\theta^*$

Testing Data

Evaluation score

Training Data
Testing Real World Problem
- Model the problem
  - Formal Model $\theta$
    - Learning Algorithm
      - Testing Data
        - Prediction Function $\theta^*$
          - Evaluation score
• Consider \( n \) I.I.D. random variables \( X_1, X_2, \ldots, X_n \)
  
  ▪ \( X_i \) is a sample from density function \( f(X_i \mid \theta) \)
    ○ Note: now explicitly specify parameter \( \theta \) of distribution
  
  ▪ We want to determine how “likely” the observed data \((x_1, x_2, \ldots, x_n)\) is based on density \( f(X_i \mid \theta) \)
  
  ▪ Define the **Likelihood function**, \( L(\theta) \):

    \[
    L(\theta) = \prod_{i=1}^{n} f(X_i \mid \theta)
    \]

    ○ This is just a product since \( X_i \) are I.I.D.

  ▪ Intuitively: what is probability of observed data using density function \( f(X_i \mid \theta) \), for some choice of \( \theta \)
The **Maximum Likelihood Estimator** (MLE) of \( \theta \), is the value of \( \theta \) that maximizes \( L(\theta) \)

- More formally: \( \theta_{MLE} = \arg \max_{\theta} L(\theta) \)
- More convenient to use **log-likelihood function**, \( LL(\theta) \):

\[
LL(\theta) = \log L(\theta) = \log \prod_{i=1}^{n} f(X_i \mid \theta) = \sum_{i=1}^{n} \log f(X_i \mid \theta)
\]

- \( \theta \) that maximizes \( LL(\theta) \) also maximizes \( L(\theta) \)
  - Formally: \( \arg \max_{\theta} LL(\theta) = \arg \max_{\theta} L(\theta) \)
  - Similarly, for any positive constant \( c \) (not dependent on \( \theta \)):

\[
\arg \max_{\theta} (c \cdot LL(\theta)) = \arg \max_{\theta} LL(\theta) = \arg \max_{\theta} L(\theta)
\]
Maximum Likelihood

\[ L(\theta) = \prod_{i=1}^{n} f(X_i \mid \theta) \]

\[ LL(\theta) = \sum_{i=1}^{n} \log f(X_i \mid \theta) \]

\[ \hat{\theta} = \arg\max_{\theta} LL(\theta) \]
Option #1: Straight optimization
Computing the MLE

- General approach for finding MLE of $\theta$
  - Determine formula for $LL(\theta)$
  - Differentiate $LL(\theta)$ w.r.t. (each) $\theta$: $\frac{\partial LL(\theta)}{\partial \theta}$
  - To maximize, set $\frac{\partial LL(\theta)}{\partial \theta} = 0$
  - Solve resulting (simultaneous) equations to get $\theta_{MLE}$
    - Make sure that derived $\hat{\theta}_{MLE}$ is actually a maximum (and not a minimum or saddle point). E.g., check $LL(\theta_{MLE} \pm \varepsilon) < LL(\theta_{MLE})$
      - This step often ignored in expository derivations
      - So, we'll ignore it here too (and won't require it in this class)
Bernoulli PMF

\[ X \sim \text{Ber}(p) \]

\[ f(X = x|p) = p^x(1 - p)^{1-x} \]
Consider I.I.D. random variables $X_1, X_2, \ldots, X_n$

- $X_i \sim \text{Ber}(p)$
- Probability mass function, $f(X_i | p)$, can be written as:
  $$f(X_i | p) = p^{x_i}(1-p)^{1-x_i} \quad \text{where } x_i = 0 \text{ or } 1$$
- Likelihood:
  $$L(\theta) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$
- Log-likelihood:
  $$LL(\theta) = \sum_{i=1}^{n} \log(p^{x_i}(1-p)^{1-x_i}) = \sum_{i=1}^{n} \left[ X_i \log p + (1-X_i) \log(1-p) \right]$$
  $$= Y \log p + (n-Y) \log(1-p) \quad \text{where } Y = \sum_{i=1}^{n} X_i$$
- Differentiate w.r.t. $p$, and set to 0:
  $$\frac{\partial LL(p)}{\partial p} = Y \frac{1}{p} + (n-Y) \frac{-1}{1-p} = 0 \quad \Rightarrow \quad p_{MLE} = \frac{Y}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Maximizing Likelihood with Bernoulli
Maximizing Likelihood with Poisson

- Consider I.I.D. random variables $X_1, X_2, \ldots, X_n$
  - $X_i \sim \text{Poi}(\lambda)$
  - PMF: $f(X_i \mid \lambda) = \frac{e^{-\lambda} \lambda^{X_i}}{X_i!}$
  - Likelihood: $L(\theta) = \prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^{X_i}}{X_i!}$
  - Log-likelihood:
    
    $$LL(\theta) = \sum_{i=1}^{n} \log\left(\frac{e^{-\lambda} \lambda^{X_i}}{X_i!}\right) = \sum_{i=1}^{n} \left[ -\lambda \log(e) + X_i \log(\lambda) - \log(X_i!) \right]$$
    
    $$= -n\lambda + \log(\lambda) \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \log(X_i!)$$
  - Differentiate w.r.t. $\lambda$, and set to 0:
    
    $$\frac{\partial LL(\lambda)}{\partial \lambda} = -n + \frac{1}{\lambda} \sum_{i=1}^{n} X_i = 0 \quad \Rightarrow \quad \lambda_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} X_i$$
Consider I.I.D. random variables $X_1, X_2, ..., X_n$

- $X_i \sim N(\mu, \sigma^2)$
- PDF: $f(X_i | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(X_i-\mu)^2/(2\sigma^2)}$

- Log-likelihood:

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

- First, differentiate w.r.t. $\mu$, and set to 0:

$$\frac{\partial LL(\mu, \sigma^2)}{\partial \mu} = \sum_{i=1}^{n} 2(X_i - \mu) / (2\sigma^2) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu) = 0$$

- Then, differentiate w.r.t. $\sigma$, and set to 0:

$$\frac{\partial LL(\mu, \sigma^2)}{\partial \sigma} = \sum_{i=1}^{n} - \frac{1}{\sigma} + 2(X_i - \mu)^2 / (2\sigma^3) = -\frac{n}{\sigma} + \sum_{i=1}^{n} (X_i - \mu)^2 / (\sigma^3) = 0$$
Now have two equations, two unknowns:

\[ \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu) = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} X_i = n \mu \quad \Rightarrow \quad \mu_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} X_i \]

First, solve for \( \mu_{\text{MLE}} \):

\[ \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu) = 0 \]

Then, solve for \( \sigma^2_{\text{MLE}} \):

\[ -\frac{n}{\sigma} + \frac{n}{\sigma^3} \sum_{i=1}^{n} (X_i - \mu)^2 = 0 \quad \Rightarrow \quad n \sigma^2 = \sum_{i=1}^{n} (X_i - \mu)^2 \]

\[ \sigma^2_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_{\text{MLE}})^2 \]

Note: \( \mu_{\text{MLE}} \) unbiased, but \( \sigma^2_{\text{MLE}} \) biased
• Consider I.I.D. random variables $X_1, X_2, \ldots, X_n$
  - $X_i \sim \text{Uni}(\alpha, \beta)$
  - PDF: $f(X_i \mid \alpha, \beta) = \begin{cases} \frac{1}{\beta - \alpha} & \alpha \leq x_i \leq \beta \\ 0 & \text{otherwise} \end{cases}$
  - Likelihood: $L(\theta) = \begin{cases} \left(\frac{1}{\beta - \alpha}\right)^n & \alpha \leq x_1, x_2, \ldots, x_n \leq \beta \\ 0 & \text{otherwise} \end{cases}$
    - Constraint $\alpha \leq x_1, x_2, \ldots, x_n \leq \beta$ makes differentiation tricky
    - Intuition: want interval size $(\beta - \alpha)$ to be as small as possible to maximize likelihood function for each data point
    - But need to make sure all observed data contained in interval
      - If all observed data not in interval, then $L(\theta) = 0$
  - Solution: $\alpha_{\text{MLE}} = \min(x_1, \ldots, x_n)$, $\beta_{\text{MLE}} = \max(x_1, \ldots, x_n)$
How do small samples affect MLE?

- In many cases, \( \mu_{MLE} = \frac{1}{n} \sum_{i=1}^{n} X_i \) = sample mean
  - Unbiased. Not too shabby…

- As seen with Normal, \( \sigma^2_{MLE} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_{MLE})^2 \)
  - Biased. Underestimates for small \( n \) (e.g., 0 for \( n = 1 \))

- As seen with Uniform, \( \alpha_{MLE} \geq \alpha \) and \( \beta_{MLE} \leq \beta \)
  - Biased. Problematic for small \( n \) (e.g., \( \alpha = \beta \) when \( n = 1 \))

- Small sample phenomena intuitively make sense:
  - Maximum likelihood ⇒ best explain data we’ve seen
  - Does not attempt to generalize to unseen data
Argmax

Option #2: Gradient Ascent
Gradient Ascent

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

Especially good if function is convex
Gradient Ascent

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

Repeat many times

This is some profound life philosophy

Walk uphill and you will find a local maxima
(if your step size is small enough)
End Review
Review: Maximum Likelihood Algorithm

1. Decide on a model for the likelihood of your samples. This is often using a PMF or PDF.

2. Write out the log likelihood function.

3. State that the optimal parameters are the argmax of the log likelihood function.

4. Use an optimization algorithm to calculate argmax.
1. Decide on a model for the likelihood of your samples. This is often using a PMF or PDF.

2. Write out the log likelihood function.

3. State that the optimal parameters are the \text{argmax} of the log likelihood function.

4. Calculate the derivative of LL with respect to theta.

5. Use an optimization algorithm to calculate argmax.
Gradient Ascent

Initialize: $\theta_j = 0$ for all $0 \leq j \leq m$

Calculate all $\theta_j$
Gradient Ascent

Initialize: $\theta_j = 0$ for all $0 \leq j \leq m$

Repeat many times:

- $\text{gradient}[j] = 0$ for all $0 \leq j \leq m$

  *Calculate all gradient[j]'s based on data*

- $\theta_j += \eta \times \text{gradient}[j]$ for all $0 \leq j \leq m$
Linear Regression Lite
Predicting CO$_2$

$X_1 = \text{Temperature}$

$X_2 = \text{Elevation}$

$X_3 = \text{CO}_2 \text{ level yesterday}$

$X_4 = \text{GDP of region}$

$X_5 = \text{Acres of forest growth}$

$Y = \text{CO}_2 \text{ levels}$
Predicting CO\textsubscript{2} (simple)

\[ X = \text{CO}_2 \text{ level} \]

\[ Y = \text{Average Global Temperature} \]

\[ \text{N training datapoints} \]

\[ (x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots (x^{(n)}, y^{(n)}) \]

\textbf{Linear Regression Lite Model}

\[ Y = \theta \cdot X + Z \quad Z \sim N(0, \sigma^2) \quad Y|X \sim N(\theta X, \sigma^2) \]
1) Write Likelihood Fn

\[ L(\theta) = \prod_{i=1}^{n} f(y^{(i)}, x^{(i)} | \theta) \]

Shorthand for:

\[ f(Y = y^{(i)}, X = x^{(i)} | \theta) \]
1) Write Likelihood Funcation

N training datapoints

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

Model

\(Y | X \sim N(\theta X, \sigma^2)\)

First, calculate Likelihood of the data

\[
L(\theta) = \prod_{i=1}^{n} f(y^{(i)}, x^{(i)} | \theta)
\]

Let's break up this joint

\[
= \prod_{i=1}^{n} f(y^{(i)} | x^{(i)}, \theta) f(x^{(i)})
\]

\[
= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y^{(i)} - \theta x^{(i)})^2}{2\sigma^2}} f(x^{(i)})
\]

\(f(x^{(i)})\) is independent of \(\theta\)

Definition of \(f(y^{(i)} | x^{(i)})\)
2) Write Log Likelihood Fn

N training datapoints: \((x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots (x^{(n)}, y^{(n)})\)

Likelihood function: 

\[
L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y^{(i)} - \theta x^{(i)})^2}{2\sigma^2}} f(x^{(i)})
\]

Second, calculate Log Likelihood of the data

\[
LL(\theta) = \log L(\theta)
\]

\[
= \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y^{(i)} - \theta x^{(i)})^2}{2\sigma^2}} f(X^{(i)})
\]

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

\[
= n \log \frac{1}{\sqrt{2\pi}} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y^{(i)} - \theta x^{(i)})^2 + \sum_{i=1}^{n} \log f(x^{(i)})
\]
3) State MLE as Optimization

N training datapoints: \((x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots (x^{(n)}, y^{(n)})\)

Log Likelihood: \[
LL(\theta) = n \log \frac{1}{\sqrt{2\pi}} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y^{(i)} - \theta x^{(i)})^2 + \sum_{i=1}^{n} \log f(x^{(i)})
\]

Third, celebrate!

\[
\hat{\theta} = \arg\max_{\theta} - \sum_{i=1}^{n} (y^{(i)} - \theta x^{(i)})^2
\]
4) Find derivative

N training datapoints: \((x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots (x^{(n)}, y^{(n)})\)

Goal: \(\hat{\theta} = \arg\max_\theta - \sum_{i=1}^{n} (y^{(i)} - \theta x^{(i)})^2\)

Fourth, optimize!

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

\[
= - \sum_{i=1}^{n} \frac{\partial}{\partial \theta} (y^{(i)} - \theta X^{(i)})^2
\]

\[
= - \sum_{i=1}^{n} 2(y^{(i)} - \theta x^{(i)})(-x^{(i)})
\]

\[
= \sum_{i=1}^{n} 2(y^{(i)} - \theta x^{(i)})(x^{(i)})
\]
5) Run optimization code

N training datapoints: \((x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(n)}, y^{(n)})\)

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

\[
\frac{\partial LL(\theta)}{\partial \theta} = \sum_{i=1}^{n} 2(y^{(i)} - \theta x^{(i)})(x^{(i)})
\]
Gradient Ascent

Initialize: $\theta_j = 0$ for all $0 \leq j \leq m$

Repeat many times:

$\text{gradient}[j] = 0$ for all $0 \leq j \leq m$

Calculate all gradient$[j]$’s based on data and current setting of theta

$\theta_j += \eta \times \text{gradient}[j]$ for all $0 \leq j \leq m$
Initialize: $\theta = 0$

Repeat many times:

$\text{gradient} = 0$

Calculate gradient based on data

$\theta \ += \ \eta \ * \ \text{gradient}$
Linear Regression (simple)

Initialize: $\theta = 0$

Repeat many times:

gradient = 0

For each training example $(x, y)$:

Update gradient for current training example

$\theta += \eta \times \text{gradient}$
Initialize: $\theta = 0$

Repeat many times:

- $\text{gradient} = 0$
- For each training example $(x, y)$:
  - $\text{gradient} += 2(y - \theta x) x$
- $\theta += \eta \ast \text{gradient}$
Linear Regression
Predicting CO\(_2\)

\[ X_1 = \text{Temperature} \]
\[ X_2 = \text{Elevation} \]
\[ X_3 = \text{CO}_2 \text{ level yesterday} \]
\[ X_4 = \text{GDP of region} \]

\[ Y = \text{CO}_2 \text{ levels} \]
Linear Regression

Problem: Predict real value $Y$ based on observing variable $X$

Model: Linear weight every feature

$$\hat{Y} = \theta_1 X_1 + \cdots + \theta_m X_m + \theta_{m+1}$$

$$= \theta^T X$$

Training: Gradient ascent to chose the best thetas to describe your data

$$\hat{\theta}_{MLE} = \arg\max_{\theta} - \sum_{i=1}^{n} (Y^{(i)} - \theta^T x^{(i)})^2$$
**Linear Regression**

Initialize: $\theta_j = 0$ for all $0 \leq j \leq m$

Repeat many times:

- $\text{gradient}[j] = 0$ for all $0 \leq j \leq m$

For each training example $(x, y)$:

- For each parameter $j$:
  - $\text{gradient}[j] += (y - \theta^T x)(-x[j])$

- $\theta_j += \eta * \text{gradient}[j]$ for all $0 \leq j \leq m$
Predicting CO₂

\[ Y = \text{CO}_2 \text{ levels} \]
\[ \hat{Y} = \theta_1 X_1 + \cdots + \theta_m X_m + \theta_{m+1} \]
\[ = \theta^T X \]

\[ X_1 = \text{Temperature} \]
\[ \theta_1 = -2.3 \]

\[ X_2 = \text{Elevation} \]
\[ \theta_2 = +1.2 \]

\[ X_3 = \text{CO}_2 \text{ level yesterday} \]
\[ \theta_3 = +10.2 \]

\[ X_4 = \text{GDP of region} \]
\[ \theta_4 = +3.3 \]
\[ \theta_5 = +95.4 \]
Training data: set of \( N \) pre-classified data instances

- \( N \) training pairs: \((x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(n)}, y^{(n)})\)
  - Use superscripts to denote \( i \)-th training instance

Learning algorithm: method for determining \( g(X) \)

- Given a new input observation of \( x = x_1, x_2, \ldots, x_m \)
- Use \( g(x) \) to compute a corresponding output (prediction)
Stretch!
Maximum A Posteriori
Noah Arthurs
CS109, Stanford University
Our Path

Neural Networks

Linear Regression
Naïve Bayes
Logistic Regression

Unbiased estimators
Maximizing likelihood
Bayesian estimation
Something rotten in the world of MLE
So good to see you again!
• I have two envelopes, will allow you to have one
  ▪ One contains $X$, the other contains $2X$
  ▪ Select an envelope
    ○ Open it!
  ▪ Now, would you like to switch for other envelope?
  ▪ To help you decide, compute $E[\text{\$ in other envelope}]$
    ○ Let $Y = \text\$ in envelope you selected
    \[E[\text{\$ in other envelope}] = \frac{1}{2} \cdot \frac{Y}{2} + \frac{1}{2} \cdot 2Y = \frac{5}{4} Y\]
  ▪ Before opening envelope, think either equally good
  ▪ So, what happened by opening envelope?
    ○ And does it really make sense to switch?
The "two envelopes" problem set-up

- Two envelopes: one contains $X$, other contains $2X$
- You select an envelope and open it
  - Let $Y =$ $\text{in envelope you selected}$
  - Let $Z =$ $\text{in other envelope}$

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

- $E[Z \mid Y]$ above assumes all values $X$ (where $0 < X < \infty$) are equally likely
  - Note: there are infinitely many values of $X$
  - So, not true probability distribution over $X$ (doesn’t integrate to 1)
All Values are Equally Likely?

Infinite powers of two...
Subjectivity of Probability

• Belief about contents of envelopes
  ▪ Since implied distribution over X is not a true probability distribution, what is our distribution over X?
    o **Frequentist**: play game infinitely many times and see how often different values come up.
    o **Problem**: I only allow you to play the game *once*

• Bayesian probability
  o Have *prior* belief of distribution for X (or anything for that matter)
  o Prior belief is a *subjective* probability
    • By extension, *all* probabilities are subjective
  o Allows us to answer question when we have no/limited data
    • E.g., probability a coin you’ve never flipped lands on heads
  o As we get more data, prior belief is “swamped” by data
Subjectivity of Probability

The graph represents the probability distribution function (p(X)) for a random variable X. The x-axis represents the values of X ranging from 0 to 100, and the y-axis represents p(X) for each value of X.
Bayesian: have prior distribution over X, \( P(X) \)
- Let \( Y = $ \) in envelope you selected
- Let \( Z = $ \) in other envelope
- Open your envelope to determine \( Y \)
- If \( Y > \mathbb{E}[Z \mid Y] \), keep your envelope, otherwise switch
  - No inconsistency!
- Opening envelope provides data to compute \( P(X \mid Y) \) and thereby compute \( \mathbb{E}[Z \mid Y] \)
- Of course, there’s the issue of how you determined your prior distribution over X…
  - Bayesian: Doesn’t matter how you determined prior, but you must have one (whatever it is)
  - Imagine if envelope you opened contained $20.01
Envelope Summary:
Probabilities are beliefs.
Incorporating prior beliefs is useful.
Priors for Parameter Estimation?
Bayes’ Theorem ($\theta = \text{model parameters}, D = \text{data}$):

- **Prior**: before seeing any data, what is belief about model
  - I.e., what is distribution over parameters $\theta$

- **Likelihood**: you’ve seen this before (in context of MLE)
  - Probability of data given probability model (parameter $\theta$)

- **Posterior**: after seeing data, what is belief about model
  - After data $D$ observed, have posterior distribution $p(\theta \mid D)$ over parameters $\theta$ conditioned on data. Use this to predict new data.

$$P(\theta \mid D) = \frac{P(D \mid \theta) P(\theta)}{P(D)}$$
MLE vs MAP

Data: \(x^{(1)}, \ldots, x^{(n)}\)

Maximum Likelihood Estimation

\[ \hat{\theta}_{MLE} = \arg\max_{\theta} f(X^{(1)} = x^{(1)}, \ldots, X^{(n)} = x^{(n)} | \theta) \]

\[ = \arg\max_{\theta} \left( \sum_{i} \log f(X^{(i)} = x^{(i)} | \theta) \right) \]

Maximum A Posteriori

\[ \hat{\theta}_{MAP} = \arg\max_{\theta} f(\Theta = \theta | X^{(1)} = x^{(1)}, \ldots, X^{(n)} = x^{(n)}) \]
Notation Shorthand

**MAP, without shorthand**

\[
\hat{\theta}_{MAP} = \operatorname{argmax}_\theta f(\Theta = \theta|X(1) = x(1), \ldots, X(n) = x(n))
\]

**Our shorthand notation**

- \(\theta\) is shorthand for the event: \(\Theta = \theta\)
- \(x^{(i)}\) is shorthand for the event: \(X^{(i)} = x^{(i)}\)

**MAP, now with shorthand**

\[
\hat{\theta}_{MAP} = \operatorname{argmax}_\theta f(\theta|x^{(1)}, \ldots, x^{(n)})
\]
MLE vs MAP

Data: \( x^{(1)}, \ldots, x^{(n)} \)

Maximum Likelihood Estimation

\[
\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} f(x^{(1)}, \ldots, x^{(n)} | \theta) \\
= \arg\max_{\theta} \left( \sum_{i} \log f(x^{(i)} | \theta) \right)
\]

Maximum A Posteriori

\[
\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} f(\theta | x^{(1)}, \ldots, x^{(n)})
\]
Most important slide of today
Maximum A Posteriori

data: \( x^{(1)}, \ldots, x^{(n)} \)  

\[ \hat{\theta}_{MAP} = \arg \max_{\theta} f(\theta|x^{(1)}, \ldots, x^{(n)}) \]
Maximum A Posteriori

\[ \hat{\theta}_{MAP} = \arg\max_{\theta} \ f(\theta | x^{(1)}, \ldots, x^{(n)}) \]

\[ \hat{\theta}_{MAP} = \arg\max_{\theta} \frac{g(\theta) f(x^{(1)}, x^{(2)}, \ldots, x^{(n)} | \theta)}{h(x^{(1)}, x^{(2)}, \ldots, x^{(n)})} \]

\[ = \arg\max_{\theta} \frac{g(\theta) \prod_{i=1}^{n} f(x^{(i)} | \theta)}{h(x^{(1)}, x^{(2)}, \ldots, x^{(n)})} \]

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

\[ = \arg\max_{\theta} \left( \log(g(\theta)) + \sum_{i=1}^{n} \log(f(x^{(i)} | \theta)) \right) \]
Choose the value of $\theta$ that maximizes:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \left( \log(g(\theta)) + \sum_{i=1}^{n} \log(f(x^{(i)}|\theta)) \right)$$
MLE vs MAP

Data: \( x^{(1)}, \ldots, x^{(n)} \)

Maximum Likelihood Estimation

\[
\hat{\theta}_{MLE} = \arg \max_{\theta} f(x^{(1)}, \ldots, x^{(n)} | \theta) \\
= \arg \max_{\theta} \left( \sum_i \log f(x^{(i)} | \theta) \right)
\]

Maximum A Posteriori

\[
\hat{\theta}_{MAP} = \arg \max_{\theta} f(\theta | x^{(1)}, \ldots, x^{(n)}) \\
= \arg \max_{\theta} \left( \log(g(\theta)) + \sum_{i=1}^{n} \log(f(x^{(i)} | \theta)) \right)
\]
Gotta get that intuition


**P(θ | D)**  

For Bernoulli

- Prior: \( \theta \sim \text{Beta}(a, b) \); data = \{n heads, m tails\}
- Estimate \( p, \text{aka} \ \theta \)

\[
\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} f(\theta|\text{data}) = \arg\max_{\theta} f(\text{data}|\theta)g(\theta)
\]

\[
= \arg\max_{\theta} \log g(\theta) + \log f(\text{data}|\theta)
\]

This is the beta PDF

This is ???
$P(\theta \mid D)$ For Bernoulli

- Prior: $\theta \sim \text{Beta}(a, b)$; data = \{n heads, m tails\}
- Estimate $p$, aka $\theta$

\[
\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} f(\theta \mid \text{data}) = \arg\max_{\theta} f(\text{data} \mid \theta)g(\theta)
\]

\[
= \arg\max_{\theta} \log g(\theta) + \log f(\text{data} \mid \theta)
\]

\[
= \arg\max_{\theta} \log \left[ \frac{1}{\beta} \theta^{a-1}(1 - \theta)^{b-1} \right] + n \log f(\text{heads} \mid \theta) + m \log f(\text{tails} \mid \theta)
\]

\[
= \arg\max_{\theta} \log \frac{1}{\beta} + (a - 1) \log \theta + (b - 1) \log(1 - \theta) + n \log \theta + m \log(1 - \theta)
\]

\[
= \arg\max_{\theta} (a - 1 + n) \log \theta + (b - 1 + m) \log(1 - \theta)
\]
Prior: $\theta \sim \text{Beta}(a, b)$; $D = \{n \text{ heads}, m \text{ tails}\}$

Estimate $p$, aka $\theta$

$$\hat{\theta}_{MAP} = \arg\max_{\theta} f(\theta|\text{data})$$

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

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

That’s the mode of the updated beta
Hyper Parameters

Hyperparameter

$a, b$ are fixed

Prior

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

Data distribution

$X_i \sim \text{Bern}(p)$

MAP will estimate the most likely value of $p$ for this model
Where’d Ya Get Them $P(\theta)$?

- $\theta$ is the probability a coin turns up heads
- Model $\theta$ with 2 different priors:
  - $P_1(\theta)$ is Beta(3,8) (blue)
  - $P_2(\theta)$ is Beta(7,4) (red)
- They look pretty different!

- Now flip 100 coins; get 58 heads and 42 tails
  - What do posteriors look like?
• As long as we collect enough data, posteriors will converge to the true value!