22: Gradient Ascent

David Varodayan
February 27, 2020
Adapted from slides by Lisa Yan
Maximum Likelihood Algorithm

1. Decide on a model for the distribution of your samples. Define the PMF/PDF for the distribution.

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:
   
   Option 1: Optimization with Math
   - Differentiate $LL(\theta)$, set to 0
   - Solve resulting simultaneous equations
Maximum Likelihood with Poisson

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

- Let $X_i \sim \text{Poi}(\lambda)$.
- PMF: $f(X_i | \lambda) = \frac{e^{-\lambda} \lambda^{X_i}}{X_i!}$

What is $\theta_{MLE} = \lambda_{MLE}$?

1. Determine formula for $LL(\theta)$

$$ LL(\theta) = \sum_{i=1}^{n} \log \left( \frac{e^{-\lambda} \lambda^{X_i}}{X_i!} \right) = \sum_{i=1}^{n} (-\lambda \log e + X_i \log \lambda - \log X_i!) $$

$$ = -n\lambda + \log(\lambda) \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \log(X_i!) \quad \text{(using natural log, } \ln e = 1) $$
Maximum Likelihood with Poisson

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

- Let \( X_i \sim \text{Poi}(\lambda) \).
- PMF: \( f(X_i | \lambda) = \frac{e^{-\lambda} \lambda^{X_i}}{X_i!} \)

What is \( \theta_{\text{MLE}} = \lambda_{\text{MLE}} \)?

1. Determine formula for \( LL(\theta) \)
2. Differentiate \( LL(\theta) \) w.r.t. (each) \( \theta \), set to 0
3. Solve resulting equations

\[
LL(\theta) = -n\lambda + \log(\lambda) \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \log(X_i!)
\]

\[
\frac{\partial LL(\theta)}{\partial \lambda} = -n + \frac{1}{\lambda} \sum_{i=1}^{n} X_i = 0 \quad \Rightarrow \quad \lambda_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} X_i
\]

MLE of the Poisson parameter, \( \lambda_{\text{MLE}} \), is the unbiased estimate of the mean, \( \bar{X} \) (sample mean)
Today’s plan

Maximum Likelihood Estimation (MLE)
• MLE of Bernoulli, Poisson, Normal, Uniform
• MLE and small samples

Gradient Ascent
• Linear Regression lite
Maximum Likelihood with Uniform

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

Let \( X_i \sim \text{Uni}(\alpha, \beta) \).

\[
f(X_i | \alpha, \beta) = \begin{cases} \frac{1}{\beta - \alpha} & \text{if } \alpha \leq x_i \leq \beta \\ 0 & \text{otherwise} \end{cases}
\]

1. Determine formula for \( L(\theta) \)

Likelihood:

\[
L(\theta) = \begin{cases} \left( \frac{1}{\beta - \alpha} \right)^n & \text{if } \alpha \leq x_1, x_2, \ldots, x_n \leq \beta \\ 0 & \text{otherwise} \end{cases}
\]

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

A. Great, let’s do it
B. Differentiation is hard
C. Constraint \( \alpha \leq x_1, x_2, \ldots, x_n \leq \beta \) makes differentiation hard
Example sample from a Uniform

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

Let \( X_i \sim \text{Uni}(\alpha, \beta) \).

\[
L(\theta) = \begin{cases} 
\left( \frac{1}{\beta - \alpha} \right)^n & \text{if } \alpha \leq x_1, x_2, ..., x_n \leq \beta \\
0 & \text{otherwise}
\end{cases}
\]

Suppose \( X_i \sim \text{Uni}(0, 1) \). [0.15, 0.20, 0.30, 0.40, 0.65, 0.70, 0.75]

You observe data:

Which parameters would give you maximum \( L(\theta) \)?

A. \( \text{Uni}(\alpha = 0, \beta = 1) \)
B. \( \text{Uni}(\alpha = 0.15, \beta = 0.75) \)
C. \( \text{Uni}(\alpha = 0.15, \beta = 0.70) \)
Example sample from a Uniform

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

Let $X_i \sim \text{Uni}(\alpha, \beta)$.

$$L(\theta) = \begin{cases} \left(\frac{1}{\beta - \alpha}\right)^n & \text{if } \alpha \leq x_1, x_2, \ldots, x_n \leq \beta \\ 0 & \text{otherwise} \end{cases}$$

Suppose $X_i \sim \text{Uni}(0,1)$.

You observe data: $[0.15, 0.20, 0.30, 0.40, 0.65, 0.70, 0.75]$

Which parameters would give you maximum $L(\theta)$?

A. Uni($\alpha = 0, \beta = 1$) \hspace{1cm} (1)^7 = 1

B. Uni($\alpha = 0.15, \beta = 0.75$) \hspace{1cm} \left(\frac{1}{0.6}\right)^7 = 35.7

C. Uni($\alpha = 0.15, \beta = 0.70$) \hspace{1cm} \left(\frac{1}{0.55}\right)^6 \cdot 0 = 0

Original parameters may not yield maximum likelihood.
Maximum Likelihood with Uniform

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

Let $X_i \sim \text{Uni}(\alpha, \beta)$.

$$L(\theta) = \begin{cases} 
\left(\frac{1}{\beta - \alpha}\right)^n & \text{if } \alpha \leq x_1, x_2, \ldots, x_n \leq \beta \\
0 & \text{otherwise}
\end{cases}$$

$\theta_{\text{MLE}}$: $\alpha_{\text{MLE}} = \min(x_1, x_2, \ldots, x_n)$ \quad $\beta_{\text{MLE}} = \max(x_1, x_2, \ldots, x_n)$

Intuition:

- Want interval size $(\beta - \alpha)$ to be as small as possible to maximize likelihood function per datapoint

(demo)

- Need to make sure all observed data is in interval (if not, then $L(\theta) = 0$)
Today’s plan

Maximum Likelihood Estimation (MLE)
• MLE of Bernoulli, Poisson, Normal, Uniform
• MLE and small samples

Gradient Ascent
• Linear Regression lite
Small samples = problems with MLE

Maximum Likelihood Estimator $\theta_{MLE}$:
- Best explains data we have seen
- Does not attempt to generalize to unseen data.

In many cases, $\mu_{MLE} = \frac{1}{n} \sum_{i=1}^{n} X_i$ Sample mean
- Unbiased ($E[\mu_{MLE}] = \mu$ regardless of size of sample, $n$)

For some cases, like Uniform: $\alpha_{MLE} \geq \alpha$, $\beta_{MLE} \leq \beta$
- Biased. Problematic for small sample size
- Example: If $n = 1$ then $\alpha = \beta$, yielding an invalid distribution

$\theta_{MLE} = \arg \max_{\theta} L(\theta) = \arg \max_{\theta} LL(\theta)$ (MLE for Bernoulli $p$, Poisson $\lambda$, Normal $\mu$)
Properties of MLE

Maximum Likelihood Estimator:

• Best explains data we have seen
• Does not attempt to generalize to unseen data.

$$\theta_{MLE} = \arg \max_{\theta} L(\theta)$$

= arg max $$LL(\theta)$$

• Often used when sample size $n$ is large relative to parameter space
• Potentially biased (though asymptotically less so, as $n \to \infty$)

• Consistent: $$\lim_{n \to \infty} P(|\hat{\theta} - \theta| < \varepsilon) = 1$$ where $\varepsilon > 0$

As $n \to \infty$ (i.e., more data), probability that $\hat{\theta}$ significantly differs from $\theta$ is zero
Announcements

Problem Set 6
Due: Wednesday 3/11
Covers: Up to Lecture 25

Late Day Reminder
No late days permitted past last day of the quarter, 3/13

Autograded Coding Problems
Run your code in the command line, not just in a Jupyter notebook cell

CS109 Contest
Due: Monday 3/9 11:59pm
Today’s plan

Maximum Likelihood Estimation (MLE)
• MLE of Bernoulli, Poisson, Normal, Uniform
• MLE and small samples

Gradient Ascent
• Linear Regression lite
Maximum Likelihood Algorithm

1. Decide on a model for the distribution of your samples. Define the PMF/PDF for the distribution.

2. Compute:

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

3. State:

\[ \theta_{MLE} = \arg \max_{\theta} LL(\theta) \]

4. Use an optimization algorithm to calculate argmax:

   Option 1: Optimization with Math
   Option 2: Optimization with Gradient Ascent
Multiple ways to calculate argmax

Let \( f(x) = -x^2 + 4 \), where \(-2 < x < 2\).

What is \( \text{arg max } f(x) \)?

1. Graph and guess

2. Differentiate, set to 0, and solve

\[
\frac{df}{dx} = -2x = 0
\]

\[ x = 0 \]

3. Gradient ascent: educated guess & check
Gradient Ascent

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

If your function is convex, Local maxima = global maxima
Gradient ascent algorithm

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

Let \( f(x) = -x^2 + 4 \), where \(-2 < x < 2\).

1. \( \frac{df}{dx} = -2x \)  
   Gradient at \( x \)

2. Gradient ascent algorithm:
   
   initialize \( x \)
   repeat many times:
   compute gradient
   \( x += \eta \) * gradient

(demo)
Linear Regression Lite

Let $X = \text{CO}_2$ level (ppm) change from 1980, $Y = \text{Average Land-Ocean Temperature (°C)}$.

You observe $n$ datapoints:

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

Example:

$(0, 0.26), (1.2, 0.32), \ldots, (68.58, 0.85)$

New notation!

$(x^{(i)}, y^{(i)})$: the $i$-th datapoint in our sample of size $n$ has density function $f(x^{(i)}, y^{(i)} | \theta)$
Linear Regression Lite

Let $X = \text{CO}_2$ level (ppm) change from 1980, $Y = \text{Average Land-Ocean Temperature (°C)}$.

You observe $n$ datapoints:

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

Example:

$(0, 0.26), (1.2, 0.32), \ldots, (368.58, 0.85)$

Linear Regression Model:

- $Y = \theta X + Z$ (linear relationship)
- $Z \sim \mathcal{N}(0, \sigma^2)$ (error normally distributed)

$\Rightarrow Y|X \sim \mathcal{N}(\theta X, \sigma^2)$

What is $\theta_{MLE} = \arg\max_{\theta} LL(\theta)$?
Gradient ascent with Linear Regression Lite

Model:

\[ Y | X \sim \mathcal{N}(\theta X, \sigma^2) \]

\( n \) datapoints in sample:

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

1. Calculate likelihood of data, \( L(\theta) \).

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

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

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

\( f(y^{(i)}|x^{(i)}, \theta) \) is PDF of \( \mathcal{N}(\theta x^{(i)}, \sigma^2) \).
Gradient ascent with Linear Regression Lite

Model:

\[ Y|X \sim \mathcal{N}(\theta X, \sigma^2) \]

\( n \) datapoints in sample:

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

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

2. Calculate log-likelihood of data, \( LL(\theta) \).

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

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

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

log(prod) = sum(log)

log(prod) = sum(log) + using natural log
Gradient ascent with Linear Regression Lite

Model:

\[ Y | X \sim \mathcal{N}(\theta X, \sigma^2) \]

\( n \) datapoints in sample:

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

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

3. State MLE as optimization objective.

\[ \theta_{MLE} = \arg \max_{\theta} LL(\theta) \]

\[ = \arg \max_{\theta} \left[ \sum_{i=1}^{n} \log f(x^{(i)}) - \sum_{i=1}^{n} \log \sqrt{2\pi\sigma} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y^{(i)} - \theta x^{(i)})^2 \right] \]

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

(eliminate constants w.r.t. \( \arg \max \))
Gradient ascent with Linear Regression Lite

Model:
\[ Y|X \sim \mathcal{N}(\theta X, \sigma^2) \]

\( n \) datapoints in sample:
\( (x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(n)}, y^{(n)}) \)

Goal:
\[ \theta_{MLE} = \arg \max_{\theta} \left[ -\sum_{i=1}^{n} (y^{(i)} - \theta x^{(i)})^2 \right] \]

4. Compute gradient w.r.t. \( \theta \).
\[
\frac{\partial}{\partial \theta} \left[ -\sum_{i=1}^{n} (y^{(i)} - \theta x^{(i)})^2 \right] = -\sum_{i=1}^{n} \frac{\partial}{\partial \theta} (y^{(i)} - \theta x^{(i)})^2 \\
= -\sum_{i=1}^{n} 2(y^{(i)} - \theta x^{(i)})(-x^{(i)}) = \sum_{i=1}^{n} 2(y^{(i)} - \theta x^{(i)})(x^{(i)})
\]
Gradient ascent with Linear Regression Lite

**Model:**

\[ Y|X \sim \mathcal{N}(\theta X, \sigma^2) \]

**Goal:**

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

**Gradient:**

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

5. Optimize.

- initialize \( \theta \)
- repeat many times:
  - compute gradient
  - \( \theta += \eta \times \text{gradient} \)

---

**n datpoints in sample:**

\((x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(n)}, y^{(n)})\)
Gradient ascent with multiple parameters

If $\theta = (\theta_0, \theta_1, \theta_2, ..., \theta_j, ..., \theta_m)$, what is $\theta_{MLE} = \arg \max_{\theta} LL(\theta)$?

Gradient update step for the $j$-th parameter, $\theta_j$:

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

initialize $\theta_j = 0$ for $0 \leq j \leq m$
repeat many times:

gradient[j] = 0 for all $0 \leq j \leq m$
compute all gradient[j] for all $0 \leq j \leq m$

$\theta_j += \eta \cdot \text{gradient}[j]$ for all $0 \leq j \leq m$
Maximum Likelihood Estimation (MLE)
  • MLE of Bernoulli, Poisson, Normal, Uniform
  • MLE and small samples

Gradient Ascent
  • Linear Regression Lite
Maximum Likelihood with Normal

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

- Let $X_i \sim \mathcal{N}(\mu, \sigma^2)$.

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

What is $\theta_{MLE} = (\mu_{MLE}, \sigma_{MLE}^2)$?

1. Determine formula for $LL(\theta)$
2. Differentiate $LL(\theta)$ w.r.t. (each) $\theta$, set to 0
3. Solve resulting equations

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

(using natural log)

$$= -\sum_{i=1}^{n} \log(\sqrt{2\pi\sigma}) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}$$
Consider a sample of \( n \) i.i.d. random variables \( X_1, X_2, \ldots, X_n \).

- Let \( X_i \sim \mathcal{N}(\mu, \sigma^2) \).

What is \( \theta_{MLE} = (\mu_{MLE}, \sigma^2_{MLE}) \)?

1. Determine formula for \( LL(\theta) \)

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

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

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

3. Solve resulting equations

\[
f(X_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}}
\]
Consider a sample of \( n \) i.i.d. random variables \( X_1, X_2, \ldots, X_n \).

- Let \( X_i \sim \mathcal{N}(\mu, \sigma^2) \).

What is \( \theta_{MLE} = (\mu_{MLE}, \sigma_{MLE}^2) \)?

1. Determine formula for \( LL(\theta) \) with respect to \( \mu \)

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

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

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

\[
\frac{\partial LL(\theta)}{\partial \sigma} = - \sum_{i=1}^{n} \frac{1}{\sigma} + \sum_{i=1}^{n} \frac{2(X_i - \mu)^2}{2\sigma^3} = - \frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (X_i - \mu)^2 = 0
\]
Maximum Likelihood with Normal

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

- Let $X_i \sim \mathcal{N}(\mu, \sigma^2)$.

What is $\theta_{MLE} = (\mu_{MLE}, \sigma_{MLE}^2)$?

3. Solve resulting equations

Two equations, two unknowns:

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

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

First, solve for $\mu_{MLE}$:

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} X_i - \frac{1}{\sigma^2} \sum_{i=1}^{n} \mu = 0 \Rightarrow \sum_{i=1}^{n} X_i = n\mu \Rightarrow \mu_{MLE} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

unbiased
Maximum Likelihood with Normal

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

- Let \( X_i \sim \mathcal{N}(\mu, \sigma^2) \).

\[
f(X_i | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(X_i - \mu)^2}{2\sigma^2}}
\]

What is \( \theta_{MLE} = (\mu_{MLE}, \sigma_{MLE}^2) \)?

3. Solve resulting equations

Two equations, two unknowns:

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

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

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

unbiased

Next, solve for \( \sigma_{MLE} \):

\[
\frac{1}{\sigma^3} \sum_{i=1}^{n} (X_i - \mu)^2 = \frac{n}{\sigma} \Rightarrow \sum_{i=1}^{n} (X_i - \mu)^2 = \sigma^2 n \Rightarrow \sigma_{MLE}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_{MLE})^2
\]

biased