24: Linear Regression and Gradient Ascent

Lisa Yan

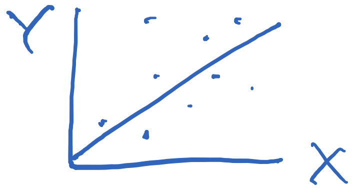
June 1, 2020

Quick slide reference

Linear Regression

Today's goals

We are going to learn linear regression.



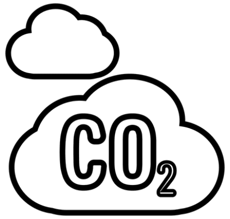
- Also known as "fit a straight line to data"
- However, linear models are too simple for more complex datasets.
- Furthermore, many tasks in CS deal with classification (categorical data), not regression.

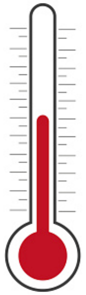
The reason we cover this topic is to teach us important skills that will help us design and understand more complicated ML algorithms:

1. How to model likelihood of training data $(\boldsymbol{x}^{(i)}, y^{(i)})$
2. What rules of argmax/calculus are important to remember
3. What gradient ascent is and why it is useful

Regression: Predicting real numbers

Training data: $(\boldsymbol{x}^{(1)}, y^{(1)}), (\boldsymbol{x}^{(2)}, y^{(2)}), \ldots, (\boldsymbol{x}^{(n)}, y^{(n)})$





Global Land-Ocean temperature

	CO2 levels	Output
Year 1	338.8	0.26
Year 2	340.0	0.32
…		⋮
Year n	340.76	0.14

$\boldsymbol{X} = (X_1)$
(assume one feature)

$Y \in \mathbb{R}$

Model:

prediction $\hat{Y} = g(\boldsymbol{X})$,

for some parametric function g

Linear Regression

Assume linear model (and $\boldsymbol{X}$ is 1-D): $\boldsymbol{X} = \langle X_1 \rangle = X$

$$\hat{Y} = g(\boldsymbol{X}) = aX + b$$

Training

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

Learn parameters $\theta = (a, b)$

Two approaches:

- Analytical solution via mean squared error
- Iterative solution via MLE and gradient ascent

Linear Regression: MSE

Mean Squared Error (MSE)

For regression tasks, we usually want a $g(X)$ that minimizes MSE:

$$\theta_{MSE} = \arg\min_{\theta} E\left[\left(Y - \hat{Y}\right)^2\right] = \arg\min_{\theta} E\left[\left(Y - g(X)\right)^2\right]$$

- Y and $\hat{Y} = g(X)$ are both random variables
- Intuitively: Choose parameter θ that minimizes the expected squared deviation ("error") of your prediction $\hat{Y}$ from the true Y

For linear regression, where $\theta = (a, b)$ and $\hat{Y} = aX + b$:

$$E[(Y - aX - b)^2]$$

Don't make me get non-linear!

$$\theta_{MSE} = \arg\min_{\theta=(a,b)} E[(Y - aX - b)^2]$$

$$a_{MSE} = \rho(X,Y)\frac{\sigma_Y}{\sigma_X}, \qquad b_{MSE} = \mu_Y - a_{MSE}\,\mu_X$$

(Derivation included at the end of this lecture)

Can we find these statistics on X and Y from our training data?

Training data: $\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)$

Not exactly, but *we can estimate* them!

Don't make me get non-linear!

$$\theta_{MSE} = \arg\min_{\theta=(a,b)} E[(Y - aX - b)^2]$$

$$a_{MSE} = \rho(X,Y)\frac{\sigma_Y}{\sigma_X}, \qquad b_{MSE} = \mu_Y - a_{MSE}\,\mu_X$$

(Derivation included at the end of this lecture)

Can we find these statistics on X and Y from our training data?

Training data: $\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)$

<u>Estimate</u> parameters based on observed training data:

$$\hat{a}_{MSE} = \hat{\rho}(X,Y)\frac{S_Y}{S_X}, \qquad \hat{b}_{MSE} = \bar{Y} - \hat{a}_{MSE}\bar{X}$$

$\hat{\rho}(X,Y)$: Sample correlation (Wikipedia)

Linear Regression

Assume linear model (and $\boldsymbol{X}$ is 1-D):

$$\hat{Y} = g(\boldsymbol{X}) = aX + b$$

Training

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

Learn parameters $\theta = (a, b)$

If we want to minimize the mean squared error of our prediction,

$$\hat{a}_{MSE} = \hat{\rho}(X, Y)\frac{S_Y}{S_X}, \qquad \hat{b}_{MSE} = \bar{Y} - \hat{a}_{MSE}\bar{X}$$

Linear Regression: MLE

Linear Regression

Assume linear model (and $\boldsymbol{X}$ is 1-D):

$$\text{predictor } \hat{Y} = g(\boldsymbol{X}) = aX + b$$

Training

Learn parameters $\theta = (a, b)$

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

We've seen which parameters minimize mean squared error. $E[(Y-\hat{Y})^2]$

What if we want parameters that maximize the likelihood of the training data?

Note: Maximizing likelihood is typically an objective for classification models.

Likelihood, it's been a minute

Review

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

$X_i \sim \text{Poi}(\lambda)$

- X_i was drawn from a distribution with density function $f(X_i|\theta)$.
  or mass
- Observed data: $(X_1, X_2, \ldots, X_n)$

Likelihood question:

How likely is the observed data $(X_1, X_2, \ldots, X_n)$ given parameter θ?

Likelihood function, $L(\theta)$:

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

This is just a product, since X_i are i.i.d.

Likelihood of the training data

Training data (n datapoints): (shorthand)

- $\left(x^{(i)}, y^{(i)}\right)$ drawn i.i.d. from a distribution $f\left(X = x^{(i)}, Y = y^{(i)} | \theta\right) = f\left(x^{(i)}, y^{(i)} | \theta\right)$
- $\hat{Y} = g(X)$, where $g(\cdot)$ is a function with parameter θ

$\theta_{MLE} = \arg\max_{\theta} \prod_{i=1}^{n} f(x^{(i)}, y^{(i)} | \theta)$

We can show that θ_{MLE} maximizes the log conditional likelihood function:

$$\theta_{MLE} = \arg\max_{\theta} \sum_{i=1}^{n} \log f\left(y^{(i)} \middle| \, x^{(i)}, \theta\right)$$

(This derivation is included at the end of this video)

(difficult)

Linear Regression, MLE

1. Assume linear model (and $\boldsymbol{X}$ is 1-D):

predictor $\hat{Y} = g(\boldsymbol{X}) = aX + b$

2. Define maximum likelihood estimator:

$$\theta_{MLE} = \arg\max_{\theta} \underbrace{\sum_{i=1}^{n} \log f\left(y^{(i)} \mid x^{(i)}, \theta\right)}$$

what is the conditional distribution of $Y \mid X, \theta$?

⚠️ Issue: We have a model of the <u>prediction</u> $\hat{Y}$ (and not Y)

- Remember MSE approach, where we minimize the squared **error** between $\hat{Y}$ and Y?

$\mathbb{E}[(Y - \hat{Y})^2]$

- Now, we **model this error** directly!

$$Y = \hat{Y} + Z$$
$$= aX + b + Z$$

error/noise
(also random)

Comparison: MSE vs MLE

$$\hat{Y} = g(\boldsymbol{X}) = aX + b$$

Minimum Mean Squared Error

$$\theta_{MSE} = \arg\min_{\theta} E\left[\left(Y - g(X)\right)^2\right]$$

- Do not directly model Y (nor error)
- Parameters are estimates of statistics from training data:

$$\hat{a}_{MSE} = \hat{\rho}(X, Y)\frac{S_Y}{S_X}$$

$$\hat{b}_{MSE} = \bar{Y} - \hat{a}_{MSE}\bar{X}$$

Maximum Likelihood Estimation

$$\theta_{MLE} = \arg\max_{\theta} \sum_{i=1}^{n} \log f\left(y^{(i)} \mid x^{(i)}, \theta\right)$$

- Directly model error between predicted $\hat{Y}$ and Y

$$Y = \hat{Y} + Z = aX + b + Z$$

If we assume error $Z \sim \mathcal{N}(0, \sigma^2)$, then these two estimators are **equivalent.**

$$\theta_{MSE} = \theta_{MLE}!$$

Linear Regression, MLE (next steps)

1. Assume linear model (and $\boldsymbol{X}$ is 1-D):

$$\hat{Y} = g(\boldsymbol{X}) = aX + b$$

2. Define maximum likelihood estimator:

$$\theta_{MLE} = \arg\max_{\theta} \sum_{i=1}^{n} \log f\left(y^{(i)} \mid x^{(i)}, \theta\right)$$

3. Model error, Z:

$$Y = aX + b + Z, \text{ where } Z \sim \mathcal{N}(0, \sigma^2)$$

4. Pick $\theta = (a, b)$ that maximizes likelihood of training data

We will not analytically find a solution. Instead, we are going to use **gradient ascent**, an **iterative optimization algorithm.**

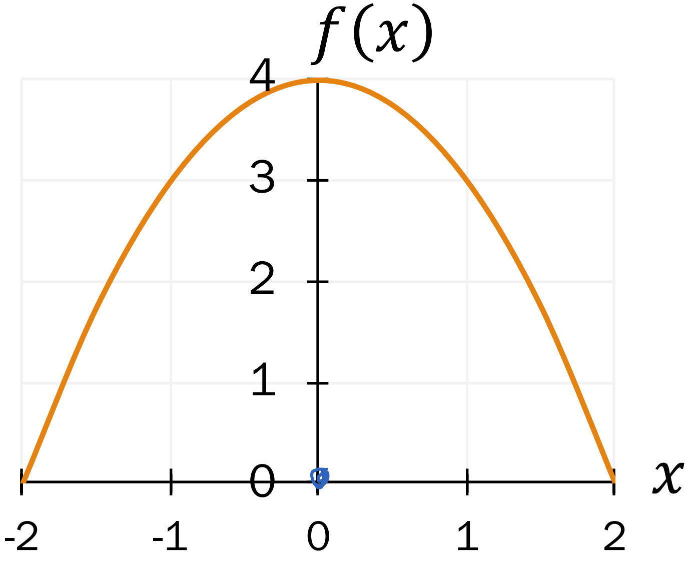
Gradient Ascent

Multiple ways to calculate argmax

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

What is $\arg\max_x \underbrace{f(x)}_{\text{objective function}}$?

A. Graph and guess

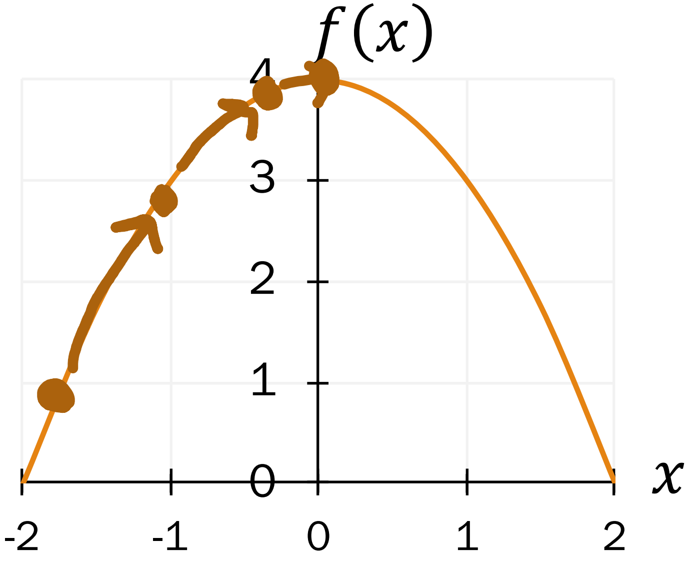


B. Differentiate, set to 0, and solve

$$\frac{df}{dx} = -2x = 0$$

$$x = 0$$

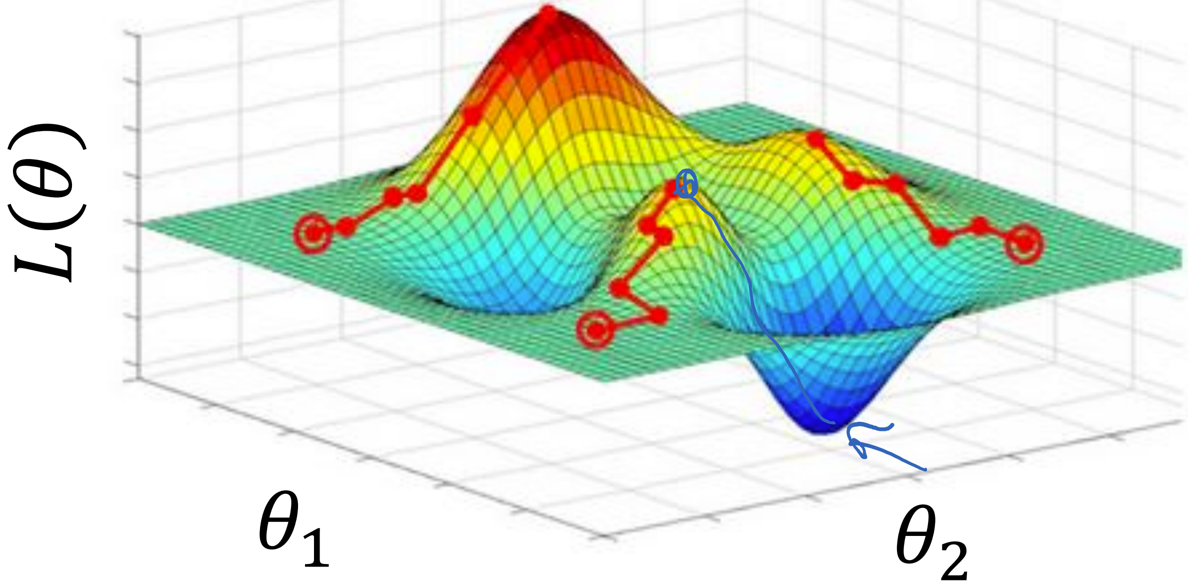
C. Gradient ascent: educated guess & check

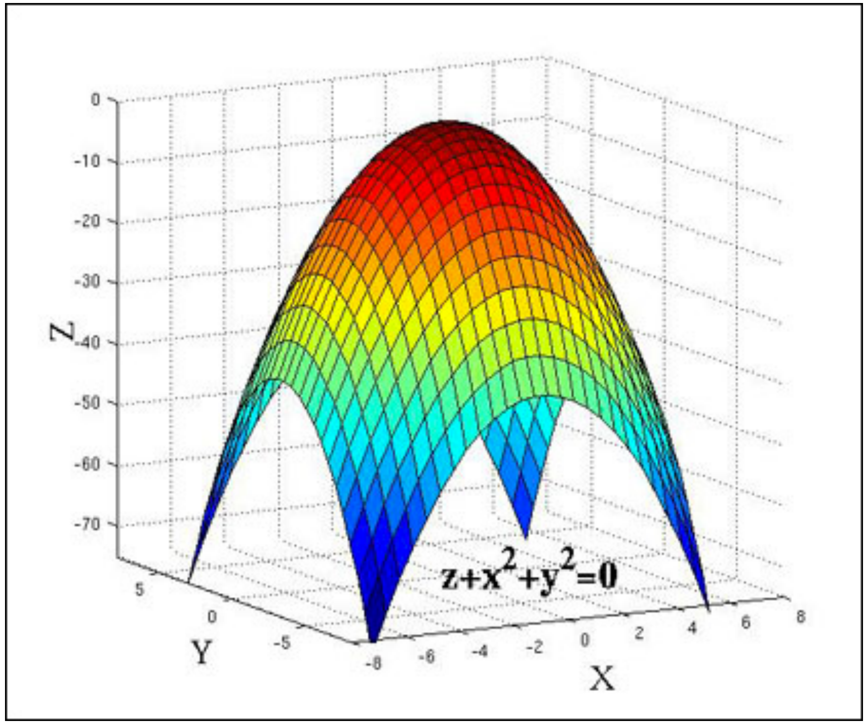


Gradient ascent

CS 109
- $L(\theta)$ are concave
- $LL(\theta)$ are also concave

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



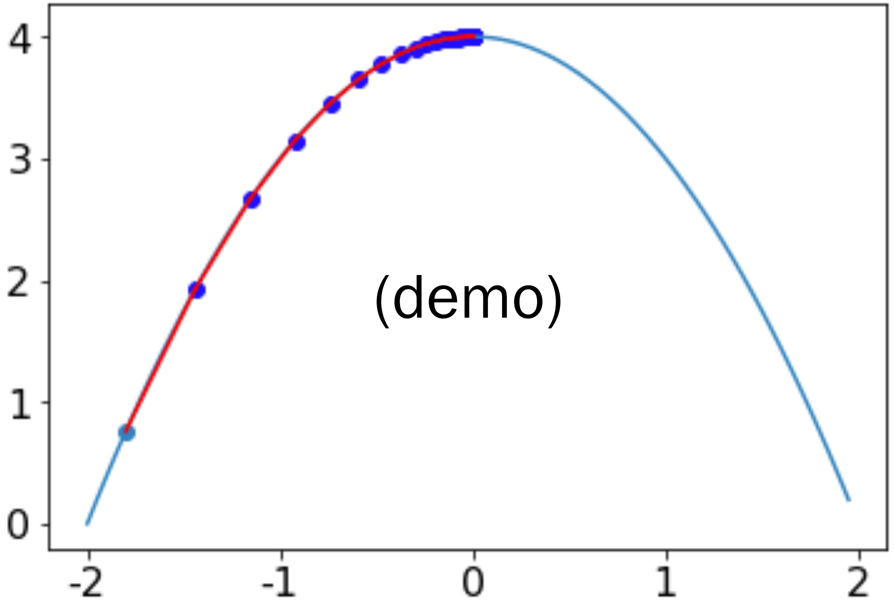


If your function is concave,
Local maxima = global maxima

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 += η * gradient
```

learning rate

Computing the MLE

Review

General approach for finding $\theta_{MLE} = \arg\max_{\theta} LL(\theta)$:

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

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

2. Differentiate $LL(\theta)$ w.r.t. (each) θ

$$\frac{\partial LL(\theta)}{\partial \theta}$$

To maximize:

$$\frac{\partial LL(\theta)}{\partial \theta} = 0$$

3. Solve resulting (simultaneous) equations

(algebra or **computer**)

If algebra is intractable, we can still find a maximum using gradient ascent!

(live) 24: Linear Regression and Gradient Ascent

Lisa Yan

June 1, 2020

Three goals today

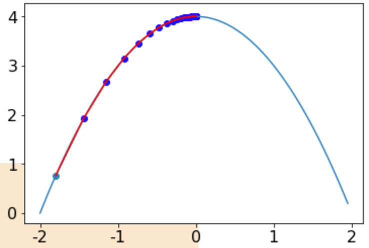
1. How to model likelihood of training data $\left(\boldsymbol{x}^{(i)}, y^{(i)}\right)$

$$\theta_{MLE} = \arg\max_{\theta} \sum_{i=1}^{n} \log f\left(y^{(i)} \mid x^{(i)}, \theta\right)$$

(θ_{MLE} maximizes log conditional likelihood)

2. What rules of argmax/calculus are important to remember

3. What gradient ascent is, why it is useful, and how to use it



1. Compute gradient.
2.
```
initialize x
repeat many times:
  compute gradient
   x += η * gradient
```

Linear Regression, MLE (so far)

1. Assume linear model (and $\boldsymbol{X}$ is 1-D):

$$\hat{Y} = g(\boldsymbol{X}) = aX + b$$

2. Define maximum likelihood estimator:

$$\theta_{MLE} = \arg\max_{\theta} \sum_{i=1}^{n} \log f\left(y^{(i)} \middle| x^{(i)}, \theta\right)$$

3. Model error, Z:

error / noise

$$Y = aX + b + Z, \text{ where } Z \sim \mathcal{N}(0, \sigma^2)$$

4. Pick $\theta = (a, b)$ that maximize likelihood of training data

Let's get started!

Computing the MLE with gradient ascent

General approach for finding θ_{MLE} , the MLE of θ:

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

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

$$\sum_{i=1}^{n} \log f\left(y^{(i)} | \, x^{(i)}, \theta\right)$$

Now: optimize log conditional likelihood

2. Differentiate ~~$LL(\theta)$~~ w.r.t. (each) θ

$$\cancel{\frac{\partial LL(\theta)}{\partial \theta}}$$

To maximize:
$$\frac{\partial LL(\theta)}{\partial \theta} = 0$$

$$\frac{\partial}{\partial \theta_j} \sum_{i=1}^{n} \log f\left(y^{(i)} | \, x^{(i)}, \theta\right)$$

$\theta_1 \quad \theta_2$

$\theta = (a, b)$

3. Solve resulting (simultaneous) equations

(algebra or computer)

(computer)
Gradient Ascent

1. Determine formula for log conditional likelihood

Model: $\theta = (a, b)$
$Y = aX + b + Z$
$Z \sim \mathcal{N}(0, \sigma^2)$

$\hat{Y} = aX + b$

Optimization problem: $\arg\max_{\theta} \sum_{i=1}^{n} \underbrace{\log f(y^{(i)} | x^{(i)}, \theta)}_{1} = \theta_{MLE}$

1. What is the conditional distribution, $Y | X, \theta$?

2. Rewrite the objective:

$$\arg\max_{\theta} \sum_{i=1}^{n} \log f(y^{(i)} | x^{(i)}, \theta)$$

1. Determine formula for log conditional likelihood

Model: $\theta = (a, b)$

$Y = aX + b + Z$

$Z \sim \mathcal{N}(0, \sigma^2)$

Optimization problem: $\arg\max_\theta \sum_{i=1}^{n} \log f\left(y^{(i)} \mid x^{(i)}, \theta\right)$

---

1. What is the conditional distribution, $Y|X,\theta$?

$Y|X=x, \theta=(a,b)$

$Y = ax + b + Z$

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

$$f\left(y^{(i)} \mid x^{(i)}, \theta\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(y^{(i)} - \left(ax^{(i)} + b\right)\right)^2 / (2\sigma^2)}$$

2. Rewrite the objective:

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

using natural log

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

1. Determine formula for log conditional likelihood

Model: $\theta = (a, b)$
$Y = aX + b + Z$
$Z \sim \mathcal{N}(0, \sigma^2)$

Optimization problem: $\arg\max_{\theta} \sum_{i=1}^{n} \log f\left(y^{(i)} \mid x^{(i)}, \theta\right)$

3. Use argmax properties to get rid of constants

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

(from previous slide)

$$= \arg\max_{\theta} \left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y^{(i)} - ax^{(i)} - b\right)^2\right]$$

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

Argmax refresher #1:
Invariant to additive constants

Argmax refresher #2: $1/2\sigma^2 > 0$
Invariant to positive constant scalars

1. Determine formula for log conditional likelihood

Model: $\theta = (a, b)$
$Y = aX + b + Z$
$Z \sim \mathcal{N}(0, \sigma^2)$

Optimization problem: $\arg\max_{\theta} \sum_{i=1}^{n} \log f\left(y^{(i)} \mid x^{(i)}, \theta\right)$

4. Celebrate!

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

2. Compute gradient

Model: $\theta = (a, b)$

$Y = aX + b + Z$

$Z \sim \mathcal{N}(0, \sigma^2)$

Optimization problem: $\arg\max_{\theta} \sum_{i=1}^{n} \log f(y^{(i)} | x^{(i)}, \theta)$

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

1. What is the derivative of the objective function w.r.t. a? (w.r.t. – "with respect to")

$$\frac{\partial}{\partial a}\left[-\sum_{i=1}^{n}(y^{(i)} - ax^{(i)} - b)^2\right] = -\sum_{i=1}^{n}\frac{\partial}{\partial a}(y^{(i)} - ax^{(i)} - b)^2$$

$$= -\sum_{i=1}^{n} 2(y^{(i)} - ax^{(i)} - b)(-x^{(i)})$$

$$= \sum_{i=1}^{n} 2(y^{(i)} - ax^{(i)} - b)(x^{(i)})$$

(rewrite)

Calculus refresher #1:
Derivative(sum) = sum(derivative)

Calculus refresher #2:
Chain rule 🌟 🌟 🌟

$\frac{d}{da} f(g(a)) = \frac{df(z)}{dz} \frac{dg(a)}{da}$

2. Compute gradient

Model: $\theta = (a, b)$

$Y = aX + b + Z$

$Z \sim \mathcal{N}(0, \sigma^2)$

Optimization problem:

$$\arg\max_{\theta} \sum_{i=1}^{n} \log f\left(y^{(i)} \mid x^{(i)}, \theta\right)$$

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

1. What is the derivative of the objective function w.r.t. a?

$$\sum_{i=1}^{n} 2\left(y^{(i)} - ax^{(i)} - b\right)\left(x^{(i)}\right)$$

2. What is the derivative of the objective function w.r.t. b?

2. Compute gradient

Model: $\theta = (a, b)$
$Y = aX + b + Z$
$Z \sim \mathcal{N}(0, \sigma^2)$

Optimization problem:

$$\arg\max_{\theta} \sum_{i=1}^{n} \log f\left(y^{(i)} \mid x^{(i)}, \theta\right)$$

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

1. What is the derivative of the objective function w.r.t. a?

$$\sum_{i=1}^{n} 2\left(y^{(i)} - ax^{(i)} - b\right)\left(x^{(i)}\right)$$

2. What is the derivative of the objective function w.r.t. b?

$$\frac{\partial}{\partial b}\left[-\sum_{i=1}^{n}\left(y^{(i)} - ax^{(i)} - b\right)^2\right]$$

$$= -\sum_{i=1}^{n} \frac{\partial}{\partial b}\left(y^{(i)} - ax^{(i)} - b\right)^2$$

$$= -\sum_{i=1}^{n} 2\left(y^{(i)} - ax^{(i)} - b\right) \cdot \frac{\partial}{\partial b}\left(y^{(i)} - ax^{(i)} - b\right)$$

$$= -\sum_{i=1}^{n} 2\left(y^{(i)} - ax^{(i)} - b\right)(-1)$$

2. Compute gradient

Model: $\theta = (a, b)$
$Y = aX + b + Z$
$Z \sim \mathcal{N}(0, \sigma^2)$

Optimization problem:

$$\arg\max_{\theta} \sum_{i=1}^{n} \log f\left(y^{(i)} \mid x^{(i)}, \theta\right)$$

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

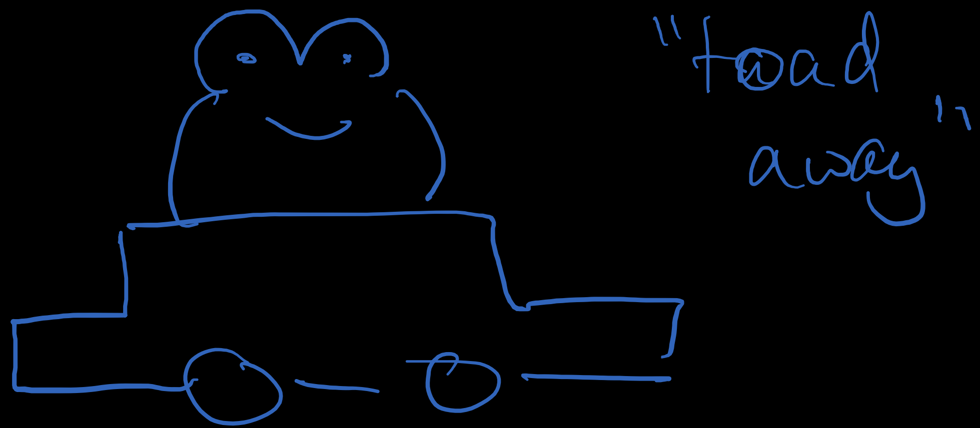
1. What is the derivative of the objective function w.r.t. a?

$$\sum_{i=1}^{n} 2\left(y^{(i)} - ax^{(i)} - b\right)\left(x^{(i)}\right)$$

2. What is the derivative of the objective function w.r.t. b?

$$\sum_{i=1}^{n} 2\left(y^{(i)} - ax^{(i)} - b\right)$$

If we set to 0 and solve, we will get an **analytical solution** for a_{MLE}, b_{MLE}.
We will reach the same solution with **gradient ascent.**



Interlude for jokes/announcements

Announcements

Problem Set 6

Out: ~~later today~~

Due: Wednesday 6/10

Covers: through ~~next~~ this Wed.

No late days or on-time bonus

READ THE README.PDF IN PSET6_CODE.ZIP

What topics do you want to see this week?
https://forms.gle/AZy7R7CNkNsLZKq2A

End of Quarter

https://us.edstem.org/courses/109/discussion/74470

Interesting probability news

Astronomer Uses Bayesian Statistics to Weigh Likelihood of Complex Life and Intelligence beyond Earth

"In Bayesian inference, prior probability distributions always need to be selected," [the astronomer] said.

"But a key result here is that when one compares the rare-life versus common-life scenarios, the common-life scenario is always at least nine times more likely than the rare one."

http://www.sci-news.com/astronomy/bayesian-statistics-likelihood-extraterrestrial-life-intelligence-08443.html

[CS109 Current Events Spreadsheet]()

3. Gradient ascent with multiple parameters

Optimization problem: $\arg\max_{\theta} \left[\underline{- \sum_{i=1}^{n} \left(y^{(i)} - a x^{(i)} - b \right)^2} \right]$

$= \arg\max_{\theta} h(\theta)$

Gradient: $\dfrac{\partial h(\theta)}{\partial a} = \sum_{i=1}^{n} 2\left(y^{(i)} - a x^{(i)} - b \right)\left(x^{(i)} \right)$

$\dfrac{\partial h(\theta)}{\partial b} = \sum_{i=1}^{n} 2\left(y^{(i)} - a x^{(i)} - b \right)$

```
initialize θ = (a, b)
repeat many times:
  compute gradient
  θ += η * gradient
```

How does this work for multiple parameters?

3. Gradient ascent with multiple parameters

Optimization problem: $\arg\max_{\theta} \left[-\sum_{i=1}^{n} \left(y^{(i)} - a x^{(i)} - b \right)^2 \right]$

$= \arg\max_{\theta} h(\theta)$

Gradient:

gradient_a

$$\boxed{\frac{\partial h(\theta)}{\partial a}} = \sum_{i=1}^{n} 2\left(y^{(i)} - a x^{(i)} - b\right)\left(x^{(i)}\right)$$

gradient_b

$$\boxed{\frac{\partial h(\theta)}{\partial b}} = \sum_{i=1}^{n} 2\left(y^{(i)} - a x^{(i)} - b\right)$$

```
a, b = 0, 0                         # initialize θ
repeat many times:
   gradient_a, gradient_b = 0, 0
   # TODO: fill in



   a += η * gradient_a       # θ += η * gradient
   b += η * gradient_b
```

How do we pseudocode the gradient computation? 🤔

$\eta = 10^{-6}$

3. Gradient ascent with multiple parameters

Optimization problem: $\arg\max_{\theta} \left[-\sum_{i=1}^{n} \left(y^{(i)} - a x^{(i)} - b \right)^2 \right]$

$$= \arg\max_{\theta} h(\theta)$$

Gradient: $\dfrac{\partial h(\theta)}{\partial a} = \sum_{i=1}^{n} 2\left(y^{(i)} - a x^{(i)} - b\right)\left(x^{(i)}\right)$

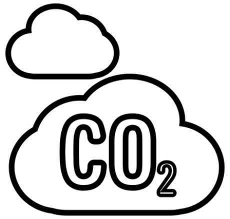
$$\frac{\partial h(\theta)}{\partial b} = \underbrace{\sum_{i=1}^{n} 2\underbrace{\left(y^{(i)} - a x^{(i)} - b\right)}}$$

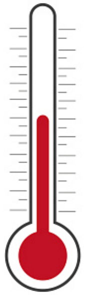
```
a, b = 0, 0                        # initialize θ
repeat many times:
  gradient_a, gradient_b = 0, 0
  for each training example (x, y):
    diff = y − (a ∗ x + b)
    gradient_a += 2 ∗ diff ∗ x
    gradient_b += 2 ∗ diff
  a += η ∗ gradient_a        # θ += η ∗ gradient
  b += η ∗ gradient_b
```

Finish computing gradient before updating any part of θ.

Global land-ocean temperature prediction

Training data: $(\boldsymbol{x}^{(1)}, y^{(1)}), (\boldsymbol{x}^{(2)}, y^{(2)}), \ldots, (\boldsymbol{x}^{(n)}, y^{(n)})$





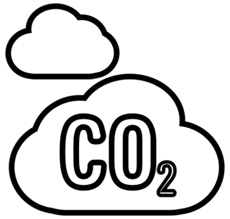
	CO2 levels	Output
Year 1	338.8	0.26
Year 2	340.0	0.32
…		⋮
Year n	340.76	0.14

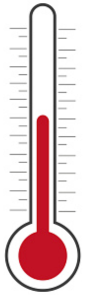
$\boldsymbol{X} = (X_1)$
(assume one feature)

$Y \in \mathbb{R}$

Global land-ocean temperature prediction

Training data: $(\boldsymbol{x}^{(1)}, y^{(1)}), (\boldsymbol{x}^{(2)}, y^{(2)}), \ldots, (\boldsymbol{x}^{(n)}, y^{(n)})$





	CO2 levels	Output
Year 1	338.8	0.26
Year 2	340.0	0.32
…		⋮
Year n	340.76	0.14

$\boldsymbol{X} = (X_1)$
(assume one feature)

$Y \in \mathbb{R}$

Minimizing
Mean Square Error

Review

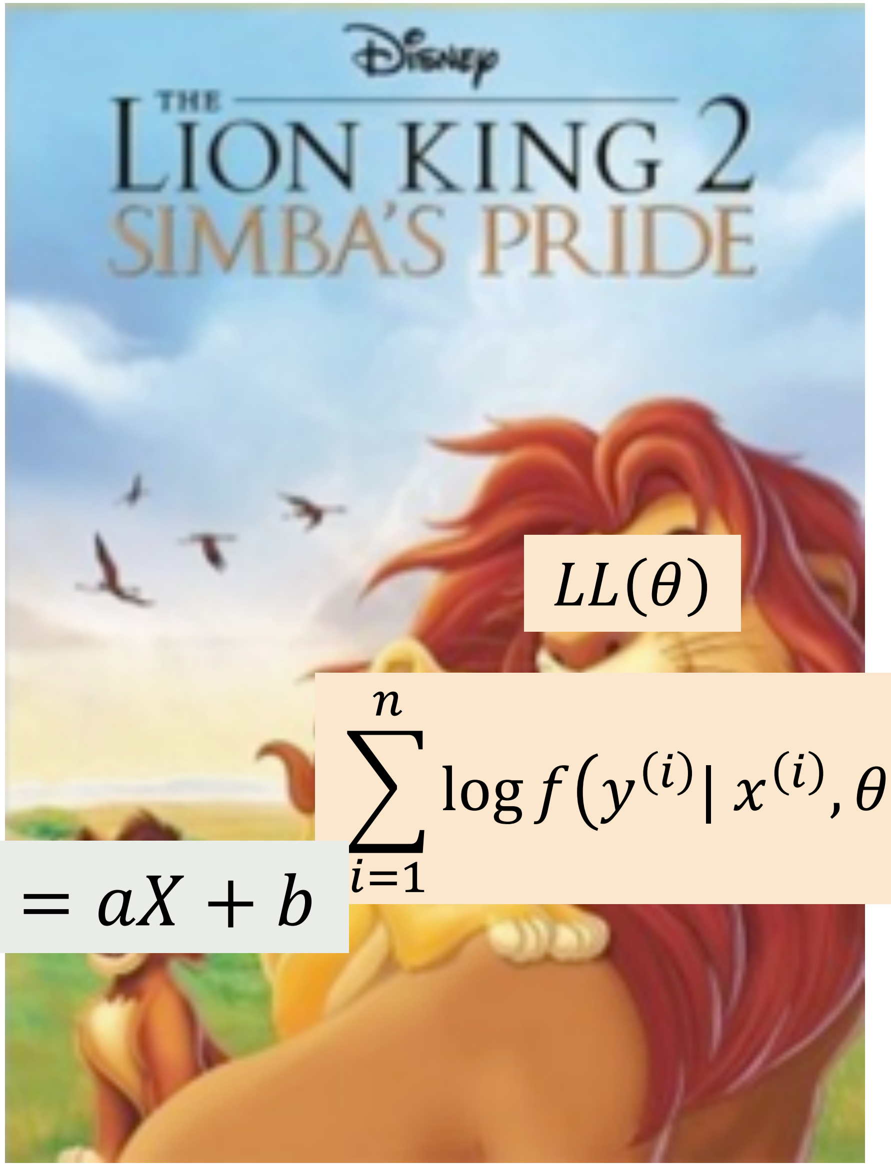
$$\theta_{MSE} = \arg\min_{\theta} E\left[(Y - g(X))^2\right]$$

$$\hat{Y} = \hat{\rho}(X,Y)\frac{S_Y}{S_X}(X - \bar{X}) + \bar{Y}$$

52

$$a_{MSE} = 0.01405$$
$$b_{MSE} = 0.17511$$

Let's try it out



$LL(\theta)$

$$\sum_{i=1}^{n} \log f\left(y^{(i)} \mid x^{(i)}, \theta\right)$$

$$\hat{Y} = g(\boldsymbol{X}) = aX + b$$

(demo)

3b. Interpret

Optimization problem: $\arg\max_{\theta}\left[-\sum_{i=1}^{n}\left(y^{(i)}-ax^{(i)}-b\right)^2\right]$

$$= \arg\max_{\theta} h(\theta)$$

Gradient: $\frac{\partial h(\theta)}{\partial a}=\sum_{i=1}^{n}2\left(y^{(i)}-ax^{(i)}-b\right)\left(x^{(i)}\right)$

$$\frac{\partial h(\theta)}{\partial b}=\sum_{i=1}^{n}2\left(y^{(i)}-ax^{(i)}-b\right)$$

```
a, b = 0, 0                    # initialize θ
repeat many times:
  gradient_a, gradient_b = 0, 0
  for each training example (x, y):
    diff = y – (a * x + b)
    gradient_a += 2 * diff * x
    gradient_b += 2 * diff
  a += η * gradient_a         # θ += η * gradient
  b += η * gradient_b
```

Updates to a and b should include information from all n training datapoints

3b. Interpret

Optimization problem:

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

$$=\arg\max_{\theta} h(\theta)$$

Gradient:

$$\frac{\partial h(\theta)}{\partial a}=\sum_{i=1}^{n}2\left(y^{(i)}-ax^{(i)}-b\right)\left(x^{(i)}\right)$$

$$\frac{\partial h(\theta)}{\partial b}=\sum_{i=1}^{n}2\left(y^{(i)}-ax^{(i)}-b\right)$$

```
a, b = 0, 0                    # initialize θ
repeat many times:
  gradient_a, gradient_b = 0, 0
  for each training example (x, y):
    diff = y – (a ∗ x + b)
    gradient_a += 2 ∗ diff ∗ x
    gradient_b += 2 ∗ diff
  a += η ∗ gradient_a         # θ += η ∗ gradient
  b += η ∗ gradient_b
```

How do we interpret the contribution of the i-th training datapoint?

3b. Interpret

Optimization problem: $\arg\max_{\theta}\left[-\sum_{i=1}^{n}\left(y^{(i)}-ax^{(i)}-b\right)^2\right]$

$= \arg\max_{\theta} h(\theta)$

Gradient:

$$\frac{\partial h(\theta)}{\partial a}=\sum_{i=1}^{n}2\left(y^{(i)}-ax^{(i)}-b\right)\left(x^{(i)}\right)$$

$$\frac{\partial h(\theta)}{\partial b}=\sum_{i=1}^{n}2\left(y^{(i)}-ax^{(i)}-b\right)$$

```
a, b = 0, 0                    # initialize θ
repeat many times:
  gradient_a, gradient_b = 0, 0
  for each training example (x, y):
    diff = y − (a ∗ x + b)
    gradient_a += 2 ∗ diff ∗ x
    gradient_b += 2 ∗ diff
  a += η ∗ gradient_a          # θ += η ∗ gradient
  b += η ∗ gradient_b
```

$\hat{y}^{(i)} = ax^{(i)} + b$

Prediction error!

$y^{(i)} - \hat{y}^{(i)}$

3b. Interpret

Optimization problem: $$\arg\max_{\theta}\left[-\sum_{i=1}^{n}\left(y^{(i)}-ax^{(i)}-b\right)^2\right]$$

$$=\arg\max_{\theta} h(\theta)$$

Gradient: $$\frac{\partial h(\theta)}{\partial a}=\sum_{i=1}^{n}2\left(y^{(i)}-ax^{(i)}-b\right)\left(x^{(i)}\right)$$

$$\frac{\partial h(\theta)}{\partial b}=\sum_{i=1}^{n}2\left(y^{(i)}-ax^{(i)}-b\right)$$

```
a, b = 0, 0                  # initialize θ
repeat many times:
  gradient_a, gradient_b = 0, 0
  for each training example (x, y):
    prediction_error = y − (a * x + b)
    gradient_a += 2 * prediction_error * x
    gradient_b += 2 * prediction_error
  a += η * gradient_a         # θ += η * gradient
  b += η * gradient_b
```

3b. Interpret

Optimization problem: $\arg\max_{\theta} \left[-\sum_{i=1}^{n} \left(y^{(i)} - ax^{(i)} - b\right)^2 \right]$

$= \arg\max_{\theta} h(\theta)$

Gradient: $\frac{\partial h(\theta)}{\partial a} = \sum_{i=1}^{n} 2\left(y^{(i)} - ax^{(i)} - b\right)\left(x^{(i)}\right)$

$\frac{\partial h(\theta)}{\partial b} = \sum_{i=1}^{n} 2\left(y^{(i)} - ax^{(i)} - b\right)$

```
a, b = 0, 0                  # initialize θ
repeat many times:
  gradient_a, gradient_b = 0, 0
  for each training example (x, y):
    prediction_error = y − (a ∗ x + b)
    gradient_a += 2 ∗ prediction_error ∗ x
    gradient_b += 2 ∗ prediction_error
  a += η ∗ gradient_a       # θ += η ∗ gradient
  b += η ∗ gradient_b
```

$\hat{Y} = aX + b$, so update to a should also scale by $x^{(i)}$

3b. Interpret

Optimization problem: $\arg\max_{\theta}\left[-\sum_{i=1}^{n}\left(y^{(i)}-ax^{(i)}-b\right)^2\right]$

$= \arg\max_{\theta} h(\theta)$

Gradient: $\dfrac{\partial h(\theta)}{\partial a}=\sum_{i=1}^{n}2\left(y^{(i)}-ax^{(i)}-b\right)\left(x^{(i)}\right)$

$\dfrac{\partial h(\theta)}{\partial b}=\sum_{i=1}^{n}2\left(y^{(i)}-ax^{(i)}-b\right)$

```
a, b = 0, 0                   # initialize θ
repeat many times:
  gradient_a, gradient_b = 0, 0
  for each training example (x, y):
    prediction_error = y − (a ∗ x + b)
    gradient_a += 2 ∗ prediction_error ∗ x
    gradient_b += 2 ∗ prediction_error ∗ 1
  a += η ∗ gradient_a         # θ += η ∗ gradient
  b += η ∗ gradient_b
```

$\hat{Y} = aX + b$, so update to b just scales by 1, not $x^{(i)}$

1

Reflecting on today

We did a lot today!

- Learned gradient ascent
- Modeled likelihood of training dataset
- Thanked argmax for its convenience
- Remembered calculus
- Implemented gradient ascent with multiple parameters to optimize for

Next up, we will use all these skills and more to tackle the final prediction model of CS109:

Logistic Regression

Extra: Derivations

$\hat{Y} = aX + b$

- $(a_{MSE}, b_{MSE}) = \theta_{MSE}$
  $= \arg\min_{\theta} \mathbb{E}[(Y - \hat{Y})^2]$
- $\theta_{MLE} = \arg\max_{\theta} L(\theta)$

Don't make me get non-linear!

$$\theta_{MSE} = \arg\min_{\theta=(a,b)} E[(Y - aX - b)^2]$$

$\frac{d}{da}(f(a))^2 = 2f(a)\frac{df}{da}$

1. Differentiate w.r.t. (each) θ, set to 0

$$\frac{\partial}{\partial a}E[(Y - aX - b)^2] = E\left[\frac{\partial}{\partial a}(Y - aX - b)^2\right]$$

($E[\cdot]$ is a linear function w.r.t. a)

$2(Y-aX-b)(-X)$

$$= E[-2(Y - aX - b)X]$$

$$= -2E[XY] + 2aE[X^2] + 2bE[X] = 0$$

$$\frac{\partial}{\partial b}E[(Y - aX - b)^2] = E[-2(Y - aX - b)]$$

$$= -2E[Y] + 2aE[X] + 2b = 0$$

2. Solve resulting simultaneous equations

$$a_{MSE} = \frac{E[XY] - E[X]E[Y]}{E[X^2] - (E[X])^2} = \frac{\text{Cov}(X,Y)}{\text{Var}(X)} = \rho(X,Y)\frac{\sigma_Y}{\sigma_X}$$

$\frac{\text{Cov}(X,Y)}{\text{Var}(X)} = \frac{\rho(X,Y)\sigma_X\sigma_Y}{\sigma_X\sigma_X}$

$$b_{MSE} = E[Y] - a_{MSE}E[X] = \mu_Y - \rho(X,Y)\frac{\sigma_Y}{\sigma_X}\mu_X = \mu_Y - a_{MSE}\mu_X$$

Log conditional likelihood, a derivation

$\hat{Y} = g(X)$, where $g(\cdot)$ is a function with parameter θ

Show that θ_{MLE} maximizes the log conditional likelihood function:

$$\theta_{MLE} = \arg\max_{\theta} \sum_{i=1}^{n} \log f\left(y^{(i)} \mid x^{(i)}, \theta\right)$$

<u>Proof</u>:

$$\theta_{MLE} = \arg\max_{\theta} \prod_{i=1}^{n} f\left(x^{(i)}, y^{(i)} \mid \theta\right) \qquad = \arg\max_{\theta} \sum_{i=1}^{n} \log \underbrace{f\left(x^{(i)}, y^{(i)} \mid \theta\right)}_{f(x^{(i)}\mid\theta) f(y^{(i)}\mid x^{(i)}, \theta)}$$

(θ_{MLE} also maximizes $LL(\theta)$)

$$= \arg\max_{\theta} \sum_{i=1}^{n} \log \underline{f\left(x^{(i)} \mid \theta\right)} + \sum_{i=1}^{n} \log f\left(y^{(i)} \mid x^{(i)}, \theta\right)$$

(chain rule, log of product = sum of logs)

$$= \arg\max_{\theta} \sum_{i=1}^{n} \log f\left(x^{(i)}\right) + \sum_{i=1}^{n} \log f\left(y^{(i)} \mid x^{(i)}, \theta\right)$$

($x^{(i)}$ indep. of θ)

$$= \arg\max_{\theta} \sum_{i=1}^{n} \log f\left(y^{(i)} \mid x^{(i)}, \theta\right)$$

($f\left(x^{(i)}\right)$ constant w.r.t. θ)

Solving MSE equations (office hours)

$$\frac{\partial}{\partial a} E[(Y - aX - b)^2] = -2E[XY] + 2aE[X^2] + \boxed{2bE[X]} = 0 \quad (1)$$

$$\frac{\partial}{\partial b} E[(Y - aX - b)^2] = -2E[Y] + 2aE[X] + 2b = 0 \quad (2)$$

Scale (2) by $\mathbb{E}[X]$: $-2\mathbb{E}[X]\mathbb{E}[Y] + 2a(\mathbb{E}[X])^2 + \boxed{2b\mathbb{E}[X]} = 0 \quad (3)$

(1) subtract (3) : $-2\mathbb{E}[XY] + 2\mathbb{E}[X]\mathbb{E}[Y] + 2a\mathbb{E}[X^2] - 2a(\mathbb{E}[X])^2 = 0$

Rewrite, solve for a : $-2(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]) = -2a[\mathbb{E}[X^2] - (\mathbb{E}[X])^2]$

$$a_{MSE} = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[X^2] - (\mathbb{E}[X])^2}$$

Rewrite (2), solve for b : $-2\mathbb{E}[Y] + 2a\mathbb{E}[X] = -2b$

$$b_{MSE} = \mathbb{E}[Y] - a_{MSE}\mathbb{E}[X]$$

$$a_{MSE} = \frac{E[XY] - E[X]E[Y]}{E[X^2] - (E[X])^2} \qquad b_{MSE} = E[Y] - \underbrace{a}_{a_{MSE}}E[X]$$