DS-UA 112
Introduction to Data Science

Week 11: Lecture 2

Gradient Descent – Determining Parameters

If we cannot solve for the parameters in a model, then how can we make guesses?

DS-UA 112
Introduction to Data Science

Week 11: Lecture 2

Gradient Descent – Determining Parameters

Adapted from Nolan, Speed, Gonzalez, Lau

Announcements

- ▶ Please check Week 11 agenda on NYU Classes
  - ▶ Lab 10
    - ▶ Due on Friday April 10 at 11:59PM EST
  - ▶ Homework 4
    - ▶ Due on Saturday April 18 at 11:59PM EST
  - ▶ Survey
    - ▶ Link to Qualtrics

Review

▶ If we have qualitative data, then we must transform it to quantitative data. However we should be careful with the encoding of the categories.

▶ We can add another independent variable for each category. The additional variables take the value 0 or 1. We call it a one-hot encoding

origin
usa
usa
europe
...
usa
japan
japan

origin=usa	origin=europe	origin=japan
1	0	0
1	0	0
0	1	0
...	...	...
1	0	0
0	0	1
0	0	1

Review

- Remember the prediction in a linear model takes the form

  $$c_0 + c_1 x_1 + \ldots + c_n x_n$$

  where c_0 is the intercept and c_i is a slope.

- However we could treat the intercept like a slope by adding another feature to the model with constant value 1

  $$c_0 = c_0 \ 1 = c_0 x_0$$

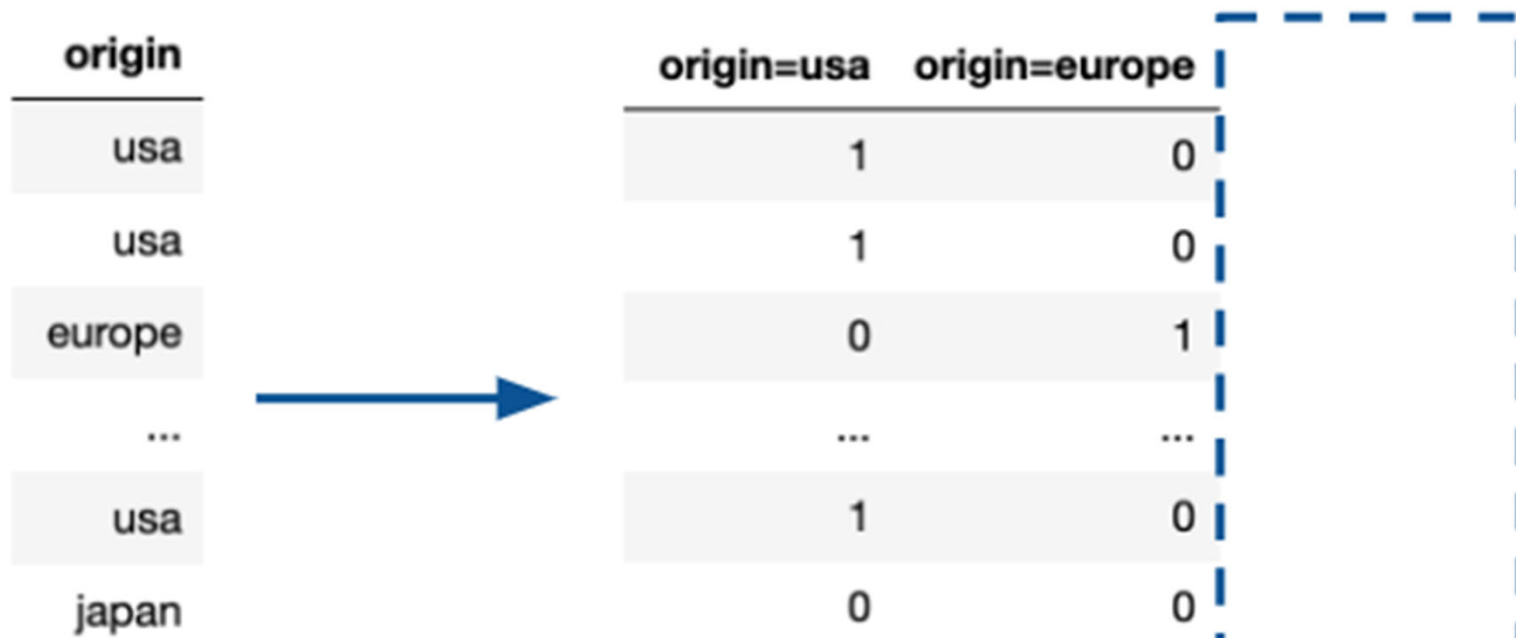| intercept		origin=usa	origin=europe	origin=japan
1		1	0	0
1		1	0	0
1		0	1	0
...	=	... +	... +	...
1		1	0	0
1		0	0	1
1		0	0	1

Review

► Note the redundancy in linear models with an intercept and one-hot encoding of categorical data. Different combinations of parameters can yield the same model.

0	3	3	3
intercept	origin=usa	origin=europe	origin=japan
1	1	0	0
1	1	0	0
1	0	1	0
...	...	...	...
1	1	0	0

=

3	0	0	0
intercept	origin=usa	origin=europe	origin=japan
1	1	0	0
1	1	0	0
1	0	1	0
...	...	...	...
1	1	0	0

Review
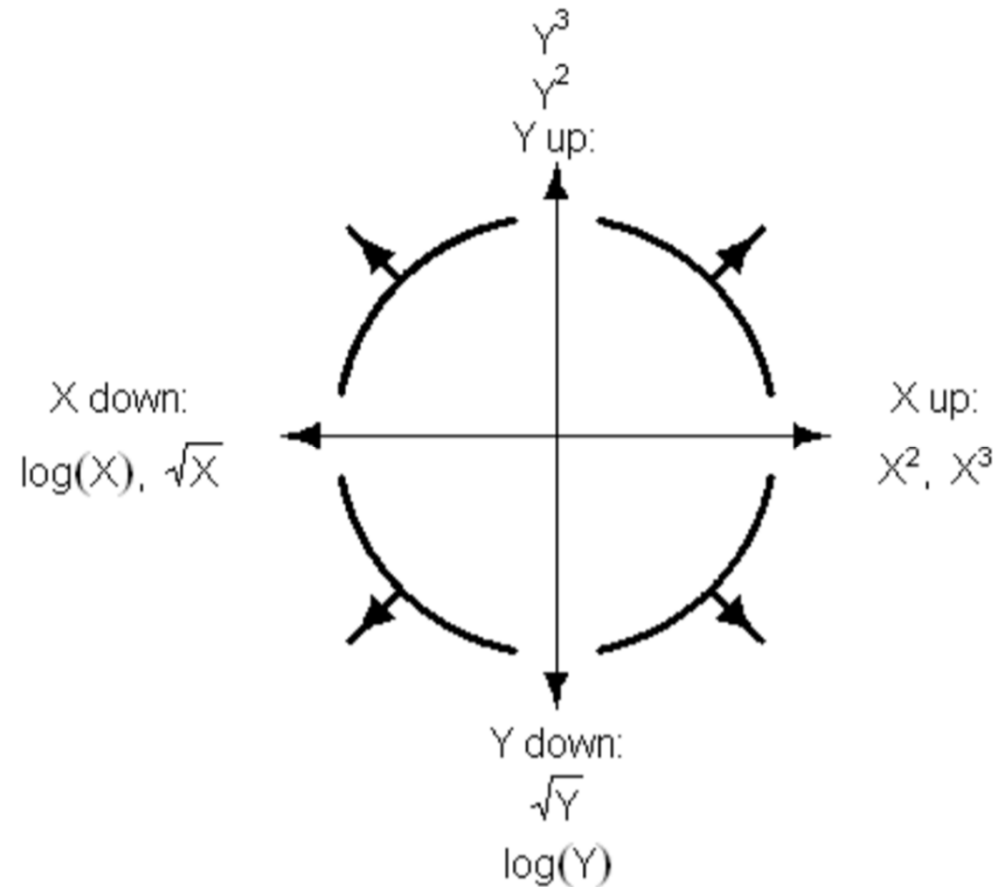
▶ So we can eliminate a feature in the one-hot encoding by dropping a column
Set the argument `drop_first` equal to True in the pandas function
`get_dummies` to omit a column from the one-hot encoding.

origin
usa
usa
europe
...
usa
japan

origin=usa	origin=europe
1	0
1	0
0	1
...	...
1	0
0	0

Review

▶ We want to use linear models to predict the dependent variable from the independent variable. However, the independent variable and dependent variable might not have a linear relationship.

▶ If a scatter-plot of the data does not cluster around a line, then we can try a transformation of the form

  ▶ Replace x with x^p

  ▶ Replace y with y^q

  ▶ Replace x with log(x)

  ▶ Replace y with log(y)

Y up: Y^3, Y^2

X down: log(X), $\sqrt{X}$

X up: X^2, X^3

Y down: $\sqrt{Y}$, log(Y)

Logarithmic Transformations

▶ Remember that logarithmic transformations help us with visualization. We can transform a large range of numbers to a small range of numbers suitable for a chart.
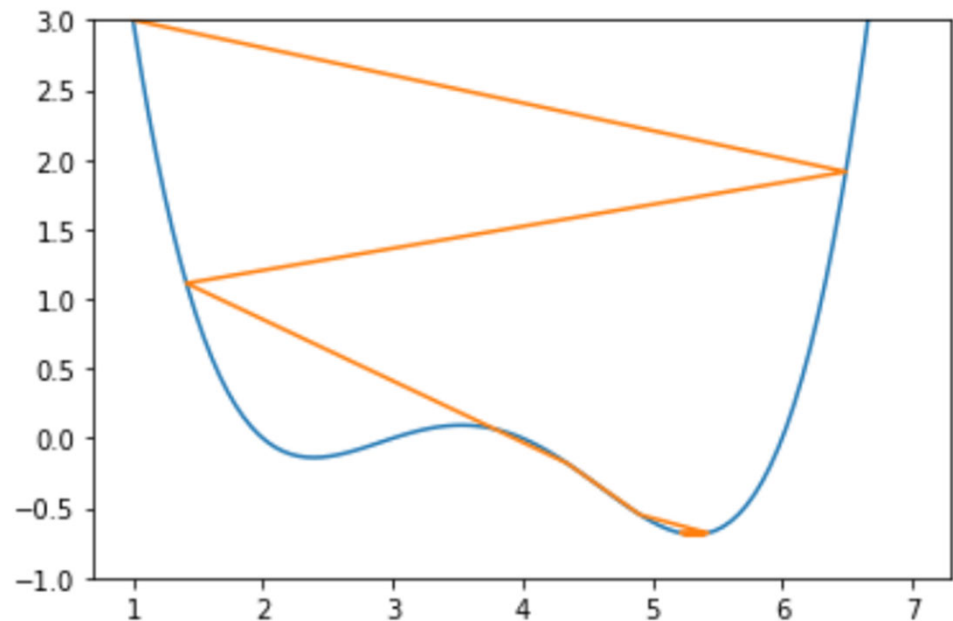
$$\log_b(a) = c \iff b^c = a$$

▶ If the independent variable and dependent variable have different scales, then we can apply logarithms to straighten out the data.


$y = \log_2(x)$

9
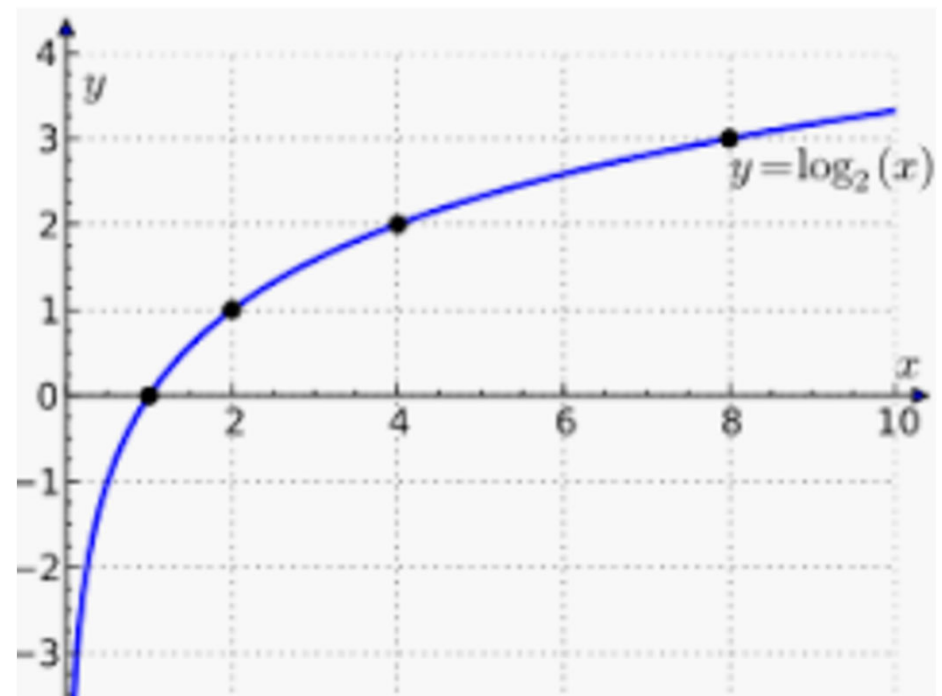
Agenda

▶ Features
  ▶ Logarithmic Transformations
▶ Gradient Descent
  ▶ Initial Guess
 ▶ Update Rule
  ▶ Learning Rate

Determining Parameters

▶ Remember that the parameters are the missing pieces in the model. We need to solve for parameters that fit the model to the data.

▶ We call the average loss between observations and prediction the empirical risk function. We want to determine parameters that minimize the empirical risk function.

▶ For example, how can we minimize the function

$f(x) = x^4 - 15x^3 + 80x^2 - 180x + 144$
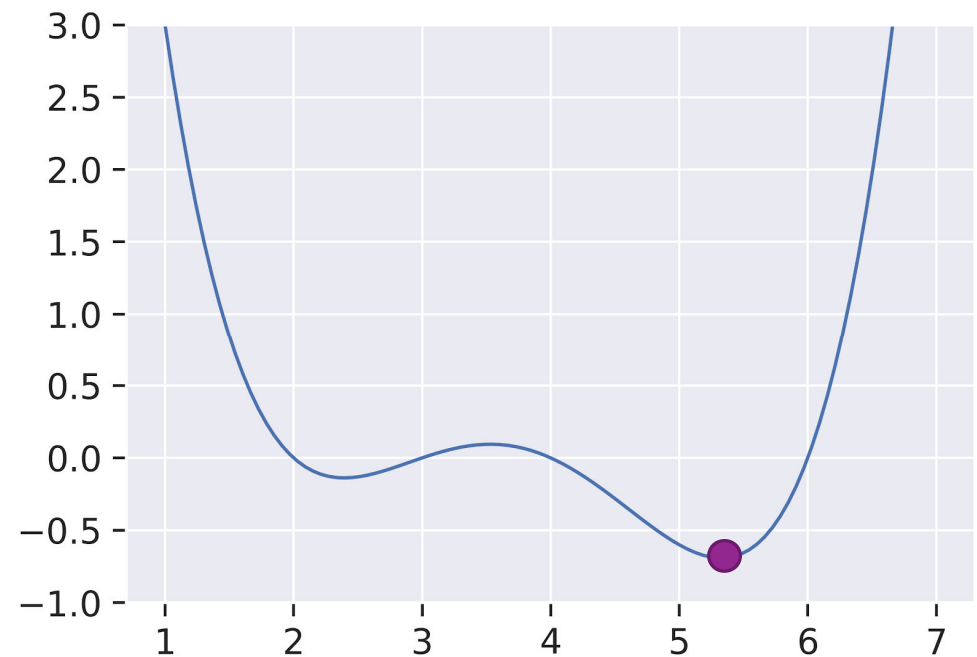
Derivatives

▶ We use derivatives to study the rate of change of a functions. We denote the derivative of f by $\frac{df}{dx}$.

▶ At a point x the value of $\frac{df}{dx}$ is approximately
$$\frac{f(x+h) - f(x)}{(x+h) - x}$$
where h is a small number.

▶ If the derivative is negative then increasing the input will decrease the output. If the derivative is positive then increasing the input will decrease the output.


12
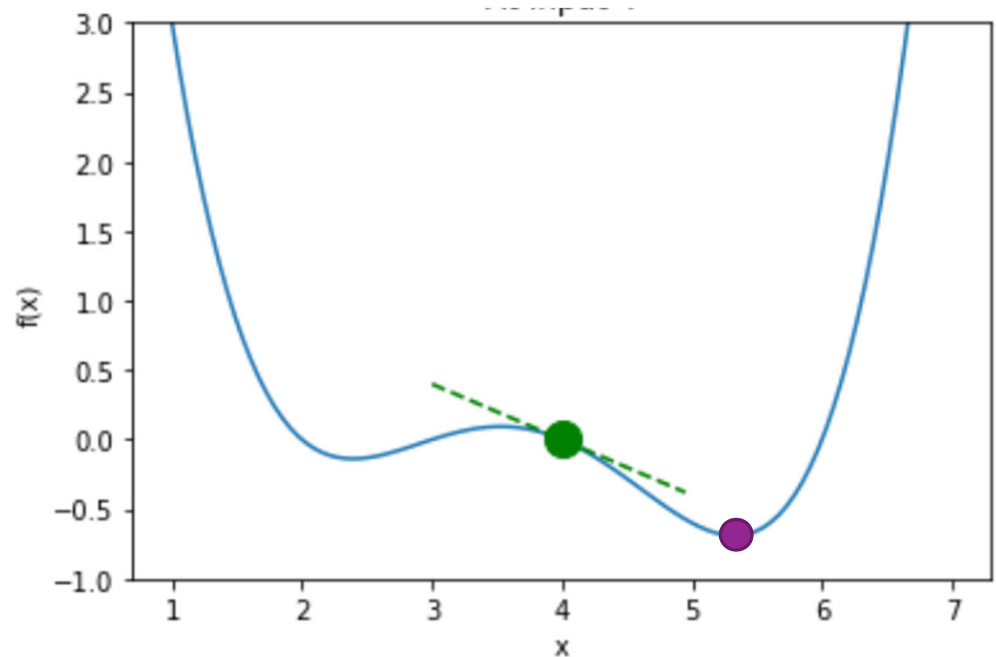
Gradient Descent

▶ We could try to guess many different inputs and check the outputs to minimize the function. However that approach is inefficient and inaccurate.

▶ Instead we will just make one guess and use the derivative to update the guess. Since we call the derivative of functions with multiple inputs the gradient, we call the approach to minimizing the function gradient descent.


13
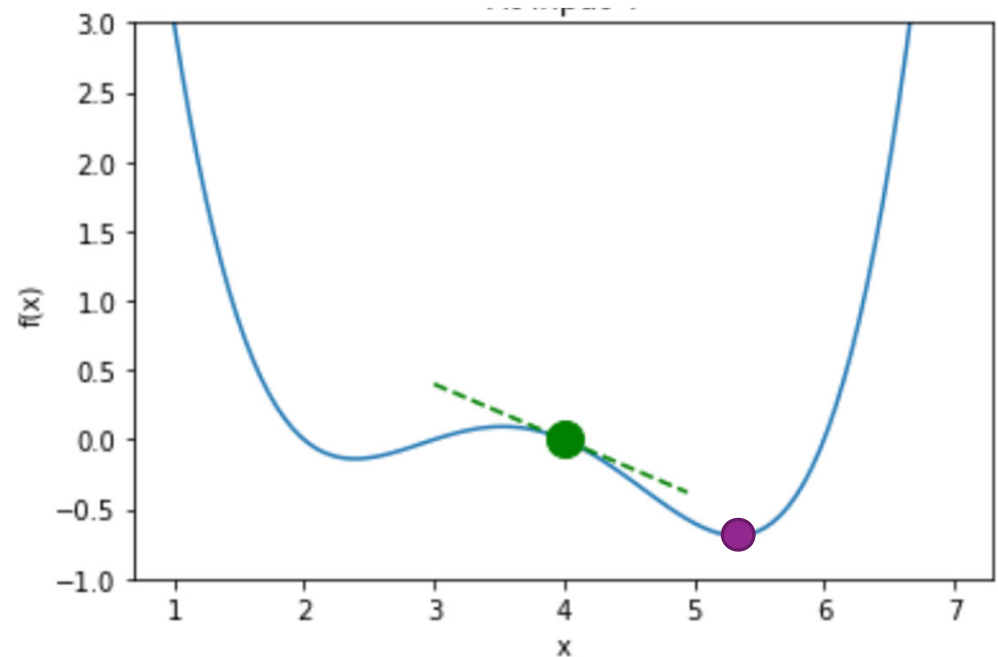
Gradient Descent

▶ To the left of a minimum the derivative is negative. So we update the guess by moving it to the right.

▶ To the right of a minimum the derivative is positive. So we update the guess by moving it to the left

▶ The derivative indicates the direction and amount to change the input. However, we may need to repeat the updates many times to find the minimum
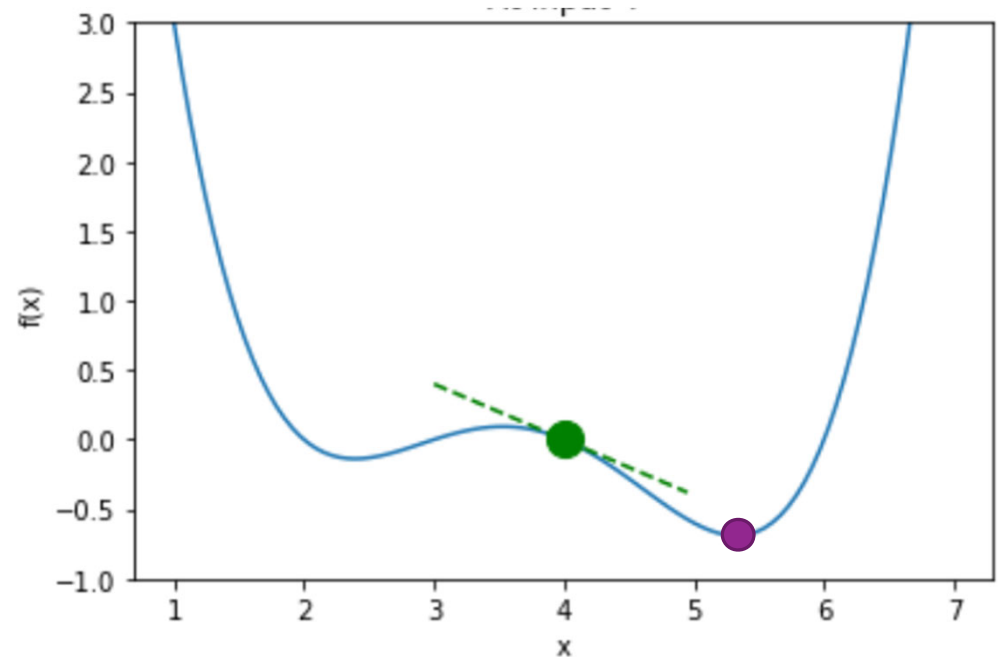
Missing the Minimum

▶ We want to find the minimum output for the function among all inputs. We nickname the value the global minimum.

▶ Some inputs might resemble a minimum because nearby inputs have larger outputs. We nickname these values local minima.

▶ Gradient descent can get stuck at a local minimum. So we need to be careful about the implementation of gradient descent for minimizing functions with local minima


Local minimum

Global minimum

15
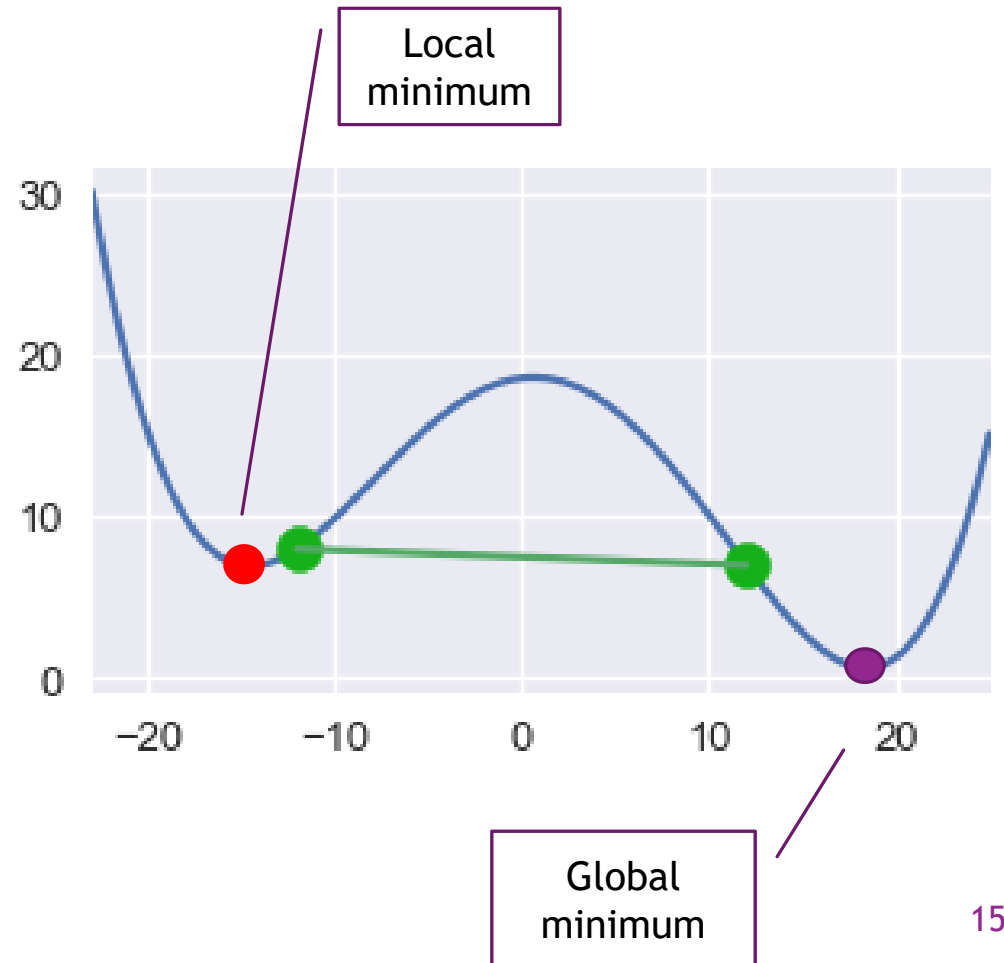
Secant Line

▶ We can identify functions with local minima through inspection of the graph.

   ▶ Between two points on the graph we draw a line called the secant line.

   ▶ If the secant line lies below the graph then the function has at least one local minimum

▶ If the secant line lies above the graph for any two points then we call the function convex. Gradient descent works well for convex functions
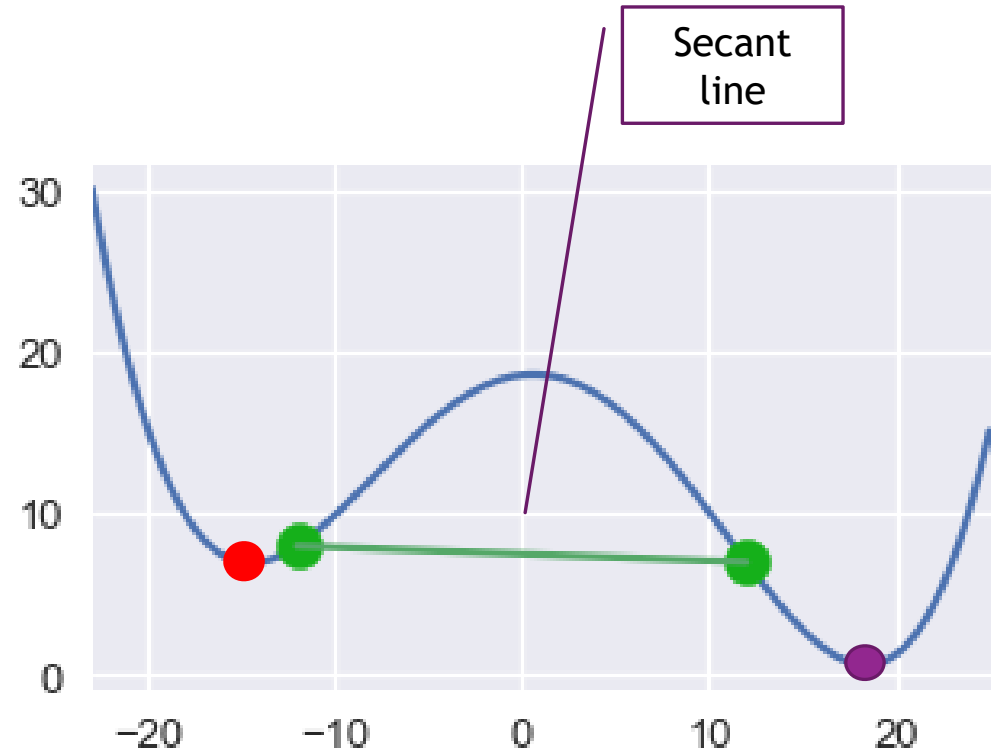
Secant line

Convex Function

- We can identify functions with local minima through inspection of the graph.

  - Between two points on the graph we draw a line called the secant line.

  - If the secant line lies below the graph then the function has at least one local minimum

- If the secant line lies above the graph for any two points then we call the function convex. Gradient descent works well for convex functions

Secant line
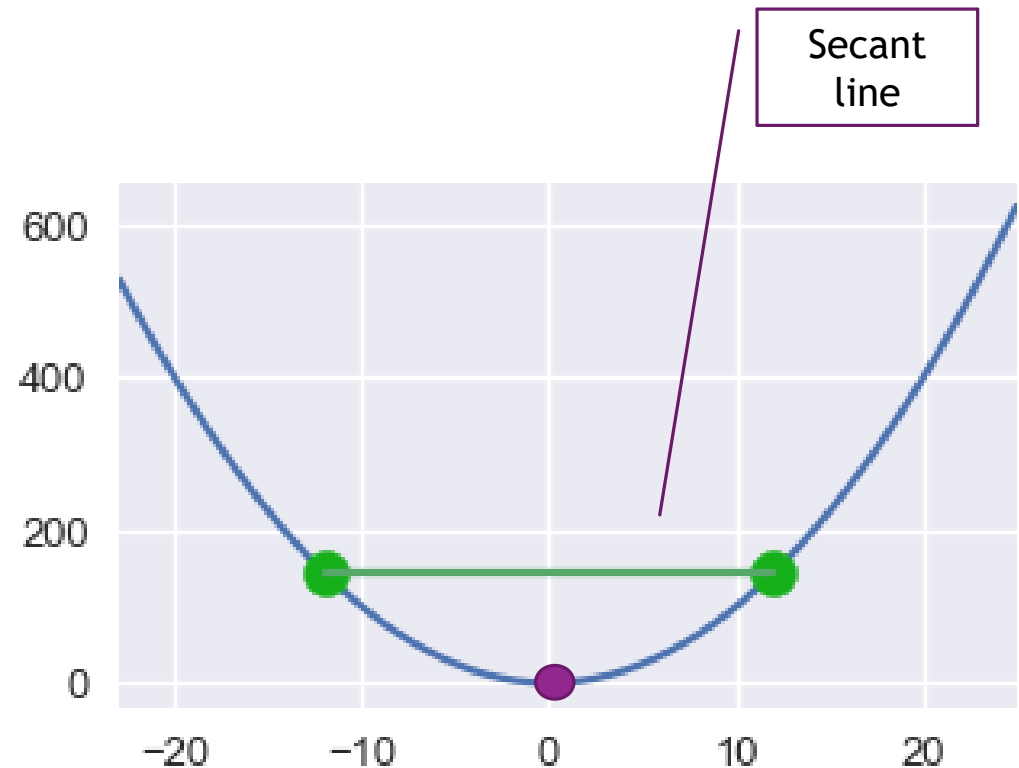
Summary

- **Features**
  - Logarithmic Transformations
- **Gradient Descent**
  - Initial Guess
  - Update Rule
  - Learning Rate

Goals

- Apply logarithm transformations to straighten out the relationship between independent variable and dependent variable

- Use gradient descent to update an initial guess through many iteration. Understand the need for a learning rate

Questions

- **Questions on Piazza?**
  - Please provide your feedback along with questions

- **Question for You!**

Why are logarithms useful in models for exponential growth?

IDEAS | EVERYDAY MATH

When a Virus Spreads Exponentially

The key to stopping the Covid-19 pandemic lies in lowering the rate at which infections multiply.

By *Eugenia Cheng*
April 2, 2020 2:01 pm ET

PRINT   AA TEXT

Fighting a pandemic like Covid-19 requires experts in many fields: epidemiologists who study the spread of disease, doctors who treat the sick, scientists who work on finding a vaccine. There is math involved in all of these specialties, but math can also help us to make sense of the barrage of information that we're receiving daily.

The starting point is the math of exponential growth. The word "exponential" is sometimes