GENERAL ASSEMBLY

Regression Metrics, Loss Functions, and Gradient Descent

Patrick D. Smith

LEARNING OBJECTIVES

- Learn how to utilize MAE and RMSE as regression metrics
- Learn the importance of and how to utilize loss functions
- Learn about and utilize gradient descent as a means to minimize loss functions

OPENING

Think back to regression - what are we trying to predict?

Think back to regression - what are we trying to predict?

The average “y” value associated with an “x” value

Loss Functions (AKA Error Functions, AAKA Cost Functions)

- Regressions model the relationship between predictors and dependent variables. But what is the relationship they are measuring, exactly, and how does it "fit" the model?

Loss Functions (AKA Error Functions, AAKA Cost Functions)

Say we have a dependent variable Y and an independent variable (predictor) X.

Standard linear regression solves for the mean value of Y predicted using X and an intercept:

$$y = \beta_0 + \beta_1 x_1$$

It solves this by minimizing the sum of squared errors:

$$\sum_i (\hat{y}_i - y_i)^2$$

Regression Metrics, Loss Functions, and Gradient Descent

Loss Functions (AKA Error Functions, AAKA Cost Functions)

This is called a loss function. The "loss" is considered the increasing sum of squared errors, which indicate a bad fit between predictors and outcome. We minimize the loss by finding the smallest sum.

$$\sum_i (\hat{y}_i - y_i)^2$$

Loss Functions (AKA Error Functions, AAKA Cost Functions)

- In this lesson we're going to dig deeper into loss functions and their applications. Different loss functions are useful in different scenarios and there are two very popular loss functions that are used in conjunction with regression. In this case they are sometimes referred to as regression metrics.

Root Mean Squared Error (RMSE), Mean Absolute Error (MAE)

Loss Functions (AKA Error Functions, AAKA Cost Functions)

- In this lesson we're going to dig deeper into loss functions and their applications. Different loss functions are useful in different scenarios and there are two very popular loss functions that are used in conjunction with regression. In this case they are sometimes referred to as regression metrics.

Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE)

Loss Functions (AKA Error Functions, AAKA Cost Functions)

First let's try a very simplified statistics problem. Given a dataset, how can we summarize it with a single number? Do you know any ways?

- This is equivalent to fitting a constant model to the data. It turns out that the **mean minimizes the RMSE** and the **median minimizes the MAE.**
- By analogy, when fitting a model, MAE is more tolerant to outliers. In other words, the degree of error of an outlier has a large impact when using RMSE versus the MAE.
- Since the choice of loss function affects model fit, it's important to consider how you want errors to impact your models.

Loss Functions (AKA Error Functions, AAKA Cost Functions)

Summary

- Use MAE when how far off an error is makes little difference
- Use RMSE when more extreme errors should have a large impact

Finally, note that linear regressions with MAE instead of RMSE are called least absolute deviation regressions rather than least squares regressions.

Regression Metrics: RMSE

Root Mean Square Error (RMSE)

- Squaring the residuals, averaging the squares, and taking the square root gives us the RMSE
- We can use the RMSE as a measure of the spread of the y values about the predicted predicted y value
- We can expect 68% of the y values to be within one RMSE, and 95% to be within two RMSE
- Also called the ***residual variation***

$$\text{RMSE} = \sqrt{\frac{\sum_i \left(\hat{y}_i - y_i\right)^2}{n}}$$

Regression Metrics: MAE

Mean Absolute Error (MAE)

- Squaring the residuals, averaging the squares, and taking the square root gives us the RMSE
- We can use the RMSE as a measure of the spread of the y values about the predicted predicted y value
- We can expect 68% of the y values to be within one RMSE, and 95% to be within two RMSE
- Also called the ***residual variation***

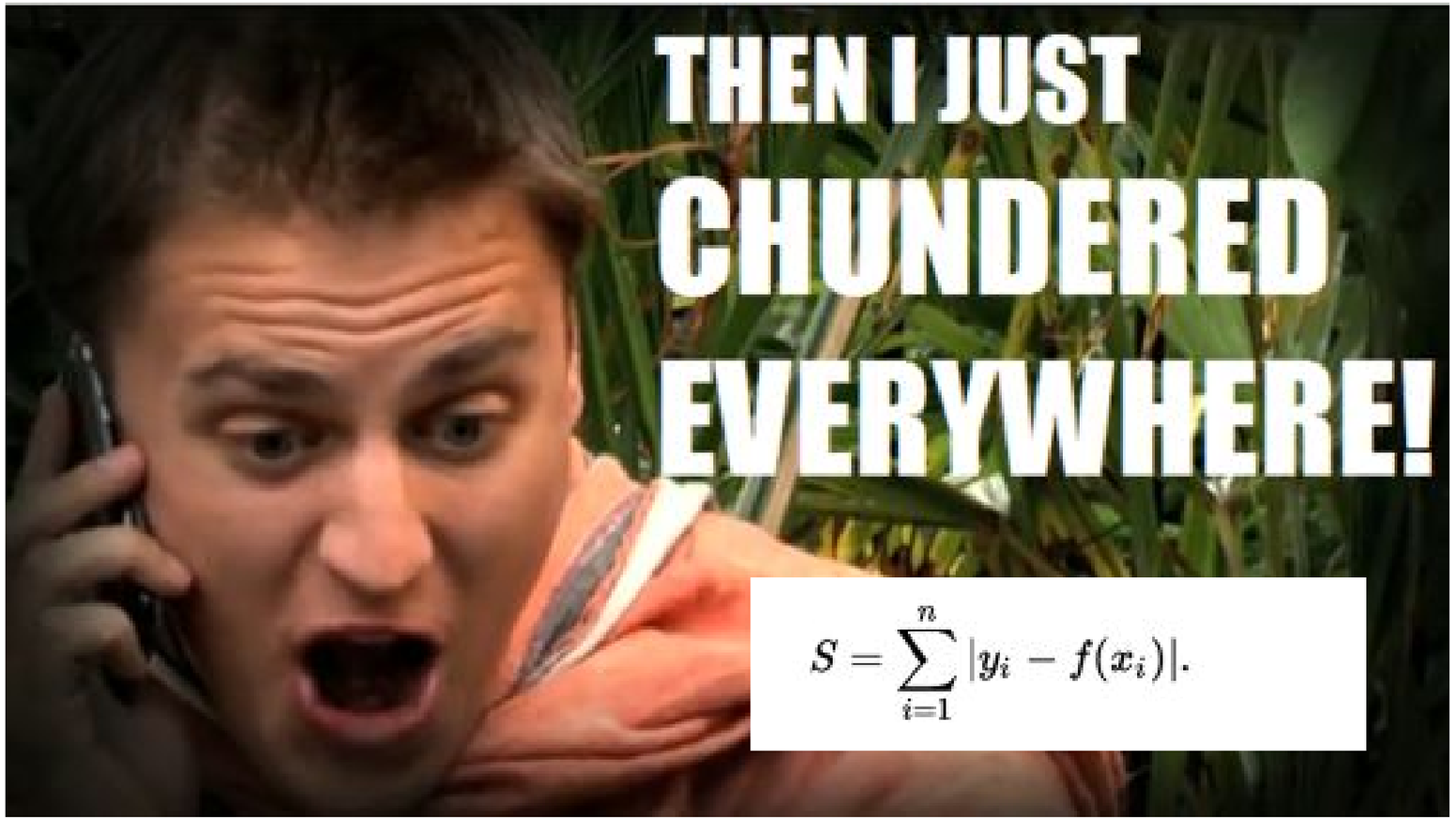
$$\text{MAE} = \frac{\sum_i |\hat{y}_i - y_i|}{n}\Bigg|$$

Loss Functions and Regression Metrics in Python

Regression Metrics, Loss Functions, and Gradient Descent

Side Deviation (Pun Intended)

LAD - and alternative to OLS



Regression Metrics: MAE

LAD

Least absolute deviation (LAD) (also known as least absolute residuals, least absolute errors and least absolute value) is an alternative method to the conventional ordinary least squares (OLS), for building regression models.

Instead of estimating the coefficients that minimise the **sum of squared residuals,** LAD estimates the coefficients that minimises the sum of the **absolute residuals.**

Regression Metrics: MAE

LAD

Compared to OLS, LAD has the advantage of being resistant to outliers and robust to departures from the normality assumption.

Regression Metrics: MAE

LAD

Compared to OLS, LAD has the advantage of being resistant to outliers and robust to departures from the normality assumption.

However, it's highly computationally expensive

Gradient Descent

Gradient Descent

- Gradient descent is in essence an algorithm designed to minimize functions. It is popular in machine learning and statistics for use in minimizing loss functions such as least squares.
- The gradient descent algorithm uses the derivative of the loss function to move in the direction where the loss function is *"descending"*.

Gradient Descent

- Gradient descent is in essence an algorithm designed to minimize functions. It is popular in machine learning and statistics for use in minimizing loss functions such as least squares.
- The gradient descent algorithm uses the derivative of the loss function to move in the direction where the loss function is *"descending"*.

Let’s derive gradient descent

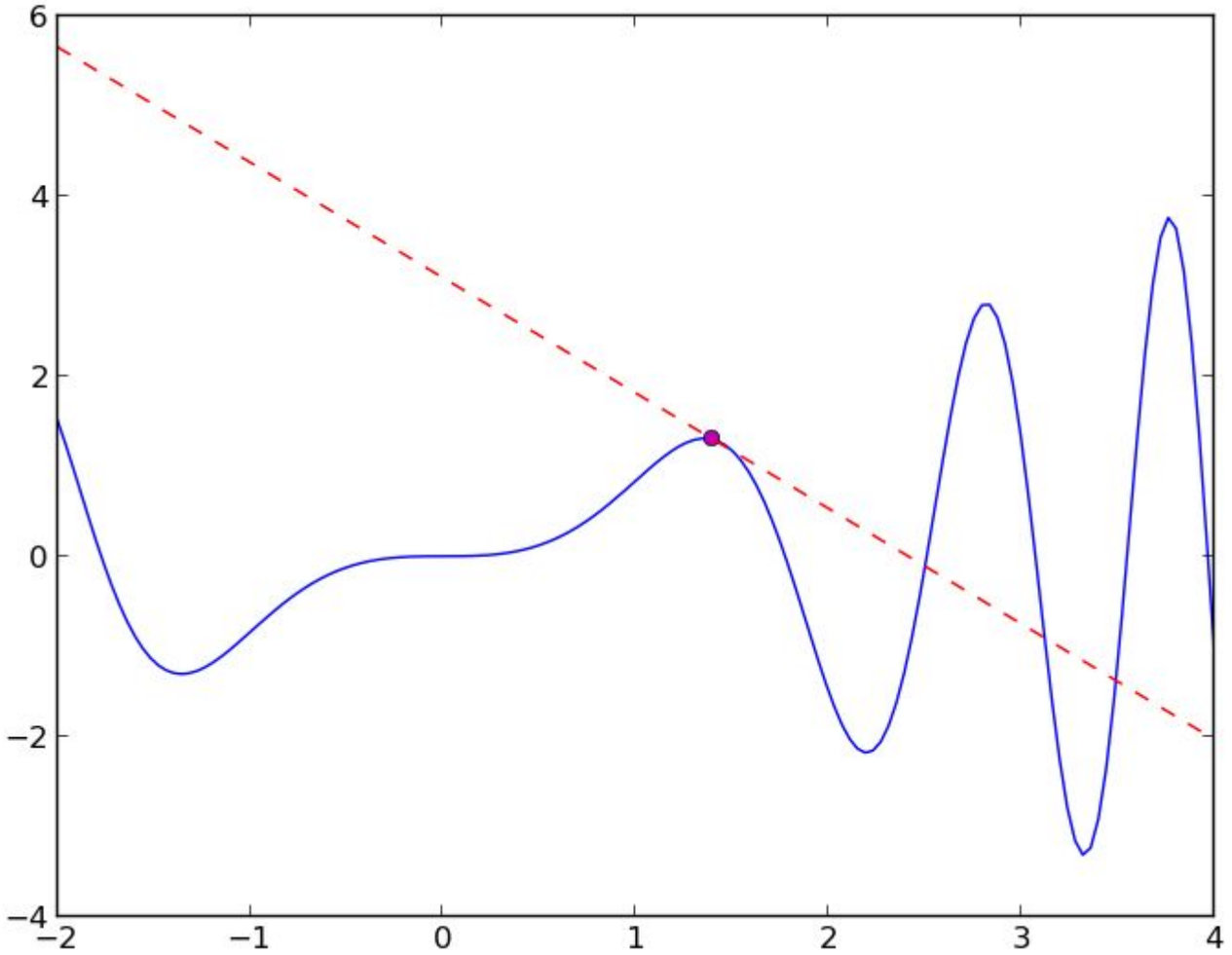
Gradient Descent: Derivatives

Review: The derivative of a function measures the rate of change of the values of the function with respect to another quantity.

- We are not going to cover the calculus of derivatives today, but will give examples through explaining their use in gradient descent.

Gradient Descent: Derivatives

The derivative of a function measures the rate of change of the values of the function with respect to another quantity.



Gradient Descent: Derivatives

A derivative of a function indicates whether the function is increasing or decreasing based on the value of the derivative.

- If the function is not changing (the tangent line is flat), the derivative is 0
- If the function is increasing (the tangent slope is positive), the derivative is positive
- If the function is decreasing (the tangent slope is negative), the derivative is negative

Gradient Descent: Derivatives

To minimize a loss function, we therefore can utilize gradient descent

Recall: The Least Square Loss

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Recall: Linear Regression with a single predictor value

$$y = \beta_0 + \beta_1 x_1$$

We can combine the regression equation and the least square into one loss function

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - (\beta_0 + \beta_1 x_i))^2$$

Gradient Descent: Derivatives

We are going to calculate the two partial derivatives of the loss function. Partial derivatives are derivatives with respect to one variable while keeping the other variables constant. Our partial derivatives will be:

- The derivative of the loss function with respect to beta0 (the intercept)
- The derivative of the loss function with respect to beta1 (the slope/coefficient for x1)

This is because the error function is defined by these two parameters. In other words, the value of the error function depends on the changes in beta0 and beta1.

What about x and y? Those variables affect the calculation of the loss, but they are not changing.

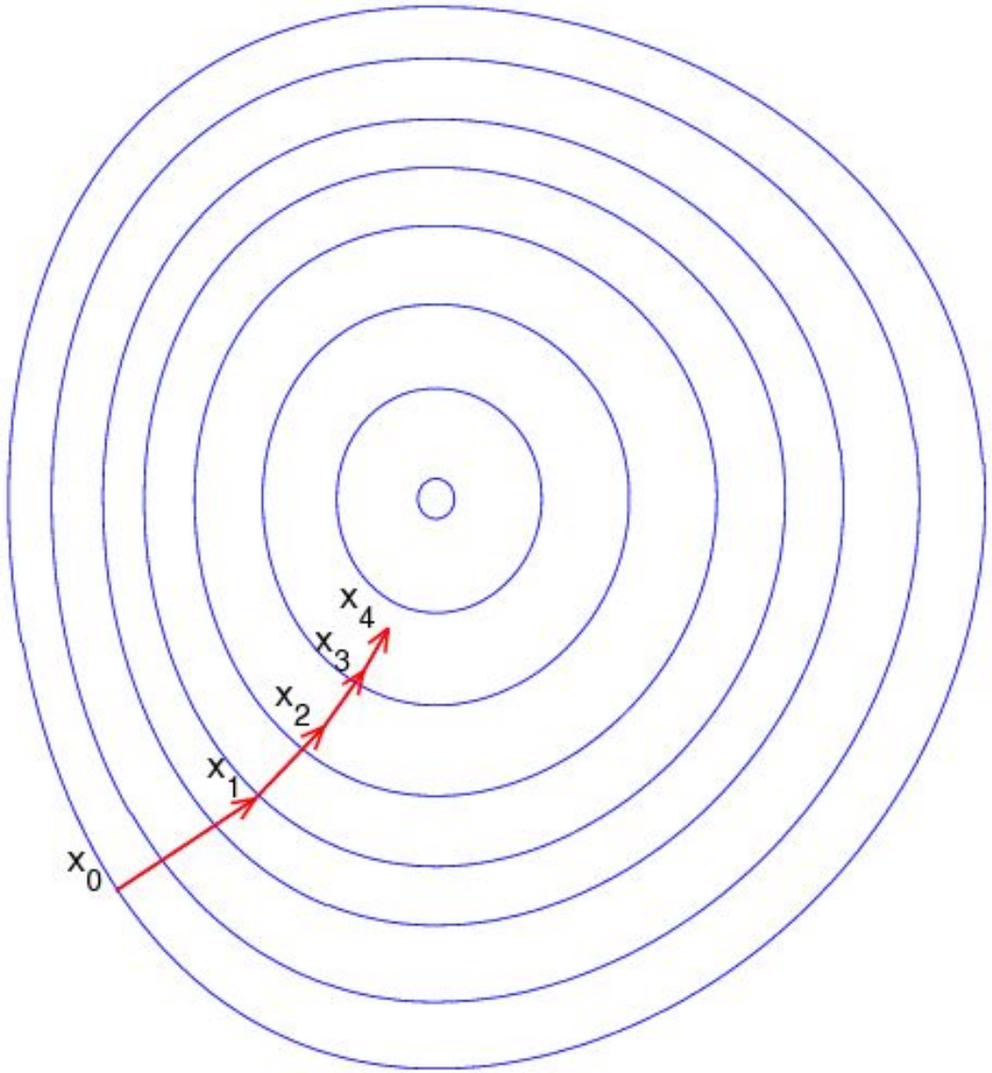
Gradient Descent: Derivatives

Recall that a positive derivative indicates an increasing function and a negative derivative indicates a decreasing function.

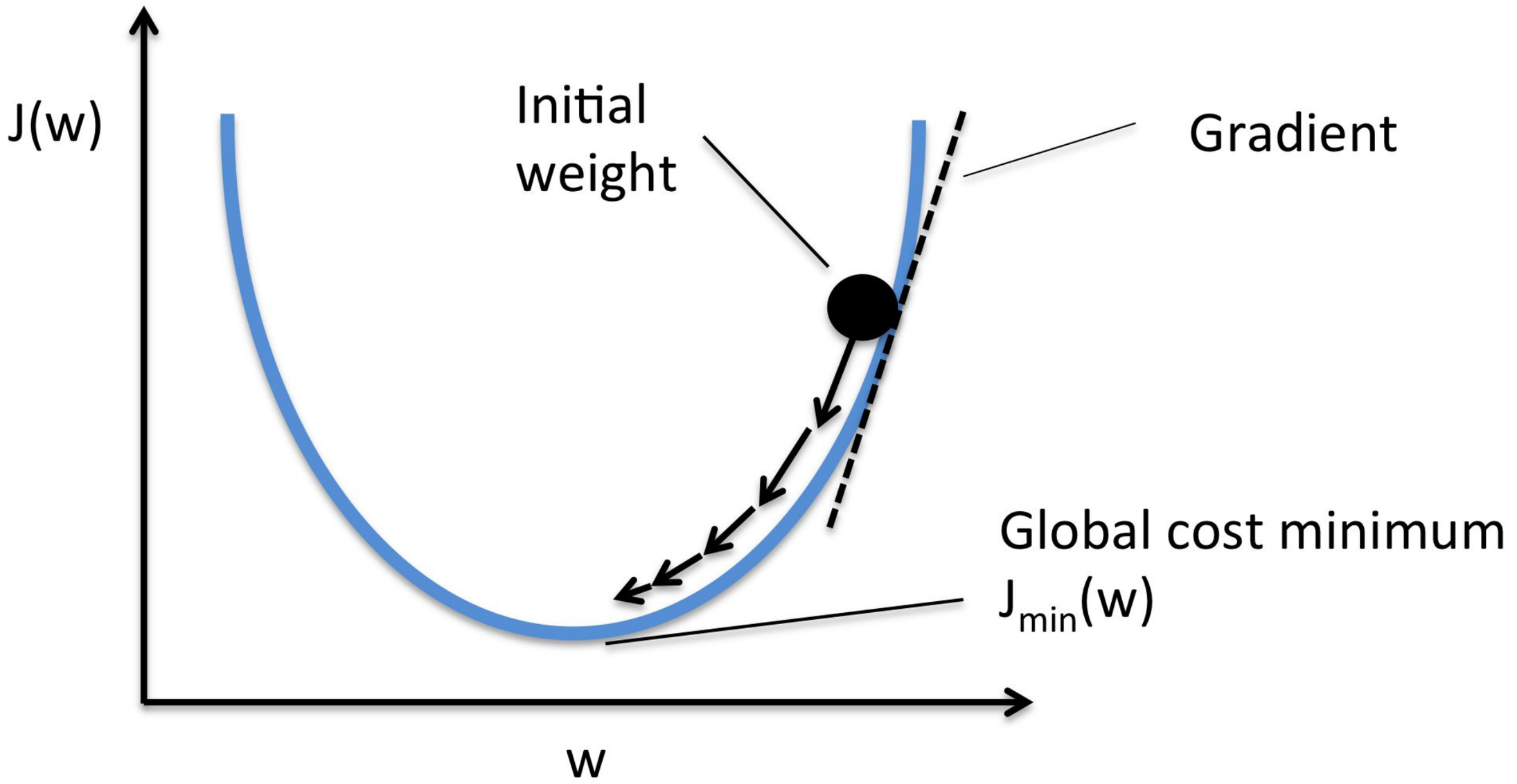
If we subtract a fraction of the partial derivative of beta1 from beta1, and subtract a fraction of the partial derivative of beta0 from beta0, we will modify beta1 and beta0 such that the value of the error function shrinks!

We can repeat this incremental process until we reach the minimum of the function.This is called ***gradient descent*** because we are iteratively moving down the gradient of the error function to its minimum.

Gradient Descent



Gradient Descent



Gradient Descent: Derivatives

Let’s practice by deriving gradient descent by hand for another loss function, **Mean Squared Error (MSE)**

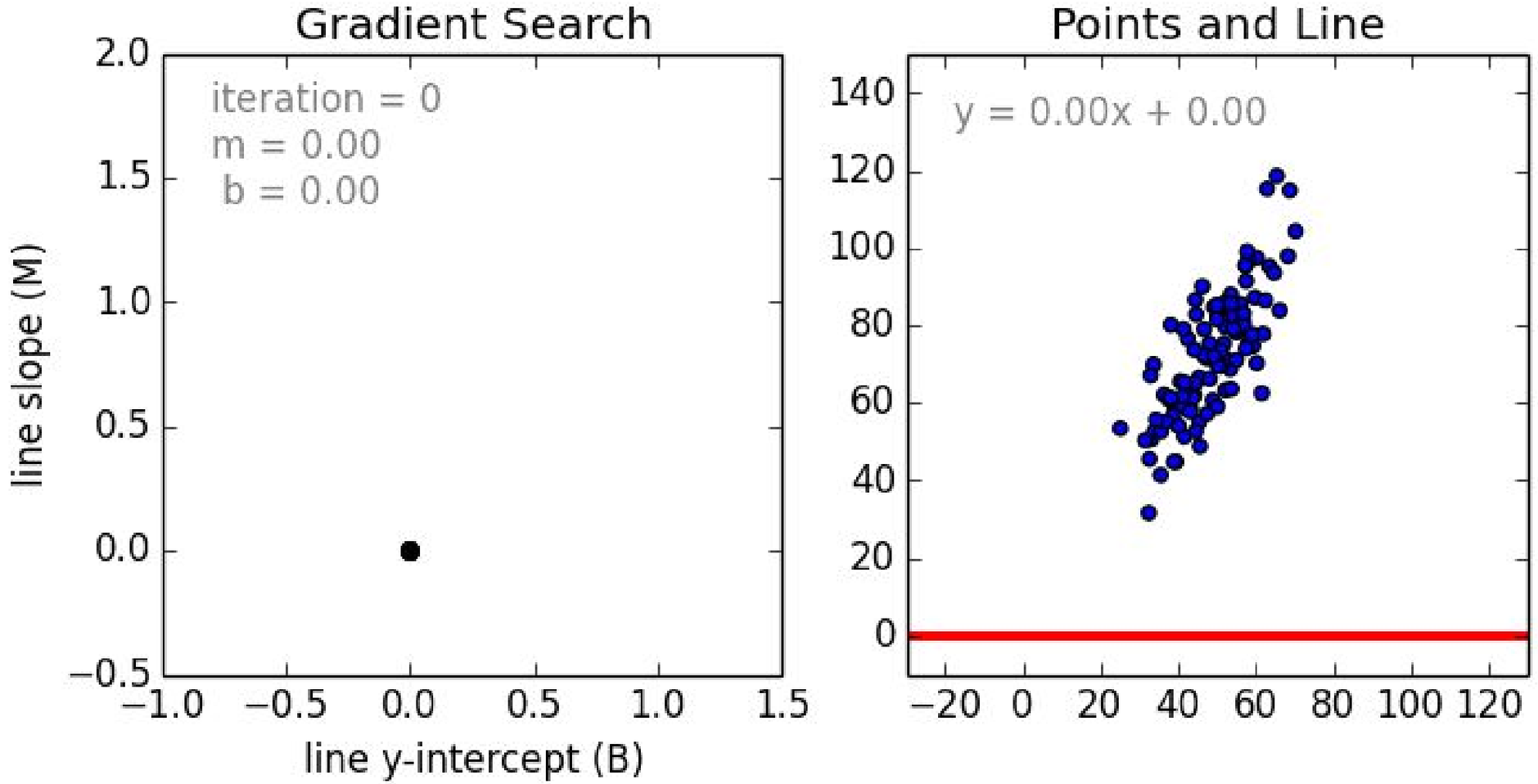
Gradient Descent in Python

Gradient Descent

We’re going to solve a typical linear regression problem of fitting a line to a set of points.

The line model is defined by two parameters - the line's slope m, and y-intercept b. Gradient descent attempts to find the best values for these parameters, subject to an error function

Gradient Descent



Potential Downsides of Gradient Descent

- One of the most fickle things about gradient descent is the step size (also known as learning rate). If this is not tuned properly, the algorithm may never converge and in fact explode into extreme values.
- Gradient descent also only works where there is a gradient to follow. Here is a toy example of a function where gradient descent will fail:

Notes on the Learning Rate

- Selecting the right learning rate is critical. If the learning rate is too large, you can overstep the minimum and even diverge.
- The only concern with using too small of a learning rate is that you will need to run more iterations of gradient descent, increasing your training time.

BEFORE NEXT CLASS

EXIT TICKET

DON’T FORGET TO FILL OUT YOUR EXIT TICKET