From:
DS100 UC Berkeley

Gradient Descent for Linear Regression (Simple and Multiple)

Introduction to Data Science
Spring 1403

Yadollah Yaghoobzadeh

Our goals for this lecture

- Optimizing complex models – how do we select parameters when the loss function is "tricky"?
  - Identifying cases where straight calculus or geometric arguments won't work
  - Introducing an alternative technique – gradient descent

Agenda

- Optimization: where are we?
- Minimizing an arbitrary 1D function
- Gradient descent on a 1D model
- Gradient descent on high-dimensional models
- Batch, mini-batch, and stochastic gradient descent

Agenda

- **Optimization: where are we?**
- Minimizing an arbitrary 1D function
- Gradient descent on a 1D model
- Gradient descent on high-dimensional models
- Batch, mini-batch, and stochastic gradient descent

What We've Done

- Takeaways from the past lecture:
  - Choose a model
  - Choose a loss function
  - Optimize parameters – choose the values of that minimize the model's loss

- How have we optimized?
  - Use calculus to solve for θ

  Take derivatives, set equal to 0, solve.

Where We're Going

We made some big assumptions

- assumed that the loss function was differentiable at all points and that the algebra was manageable

To design more complex models with different loss functions, we need a new optimization technique: **gradient descent**.

Big Idea: use an algorithm instead of solving for an exact answer

Our Roadmap

Big Idea: use an algorithm instead of solving for an exact answer

Structure for today's lecture:

1. Use a simple example (some arbitrary function) to build the intuition for our algorithm
2. Apply the algorithm to a *simple model* to see it in action
3. Formalize the algorithm to be applied to *any model*

Remember our goal: find the parameters that **minimize the model's loss**

Let's do it.

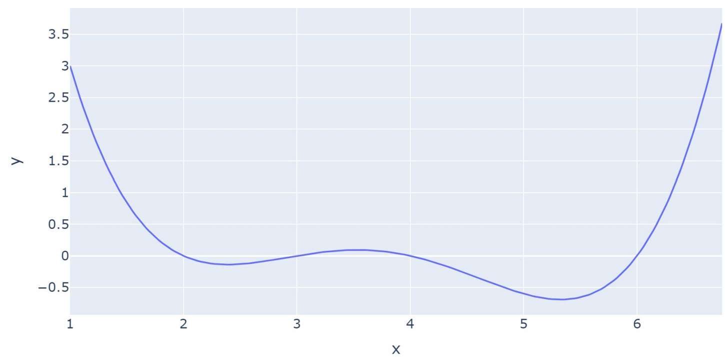
Minimizing an arbitrary 1D function

- Optimization: where are we?
- **Minimizing an arbitrary 1D function**
- Gradient descent on a 1D model
- Gradient descent on high-dimensional models
- Batch, mini-batch, and stochastic gradient descent

An Arbitrary Function

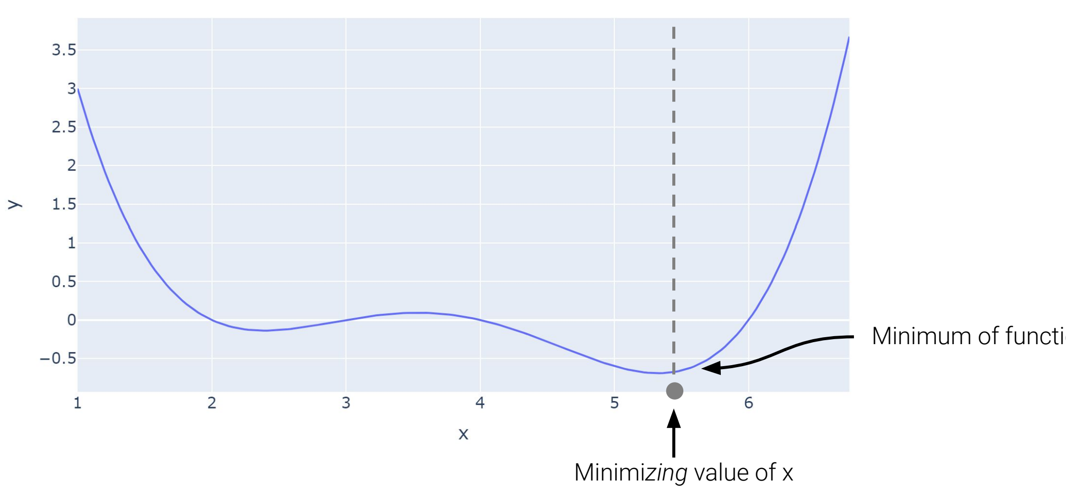
```python
def arbitrary(x):
    return (x**4 - 15*x**3 + 80*x**2 - 180*x + 144)/10

x = np.linspace(1, 6.75, 200)
fig = px.line(y = arbitrary(x), x = x)
```



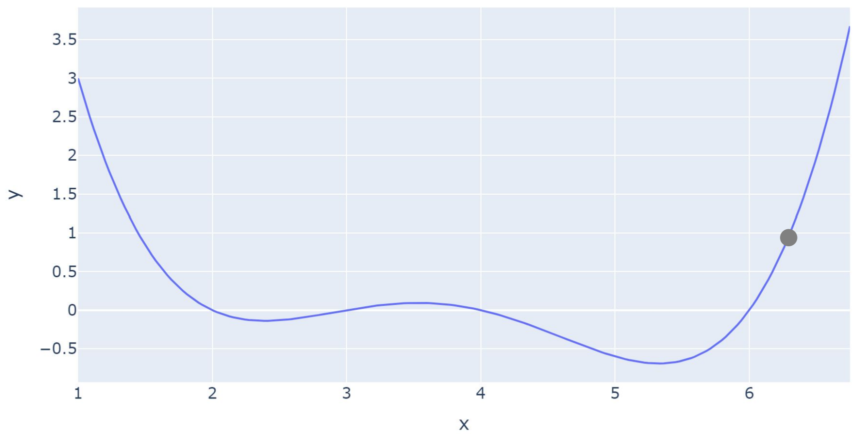
An Arbitrary Function

Our goal is to find the value of x that minimizes our function.



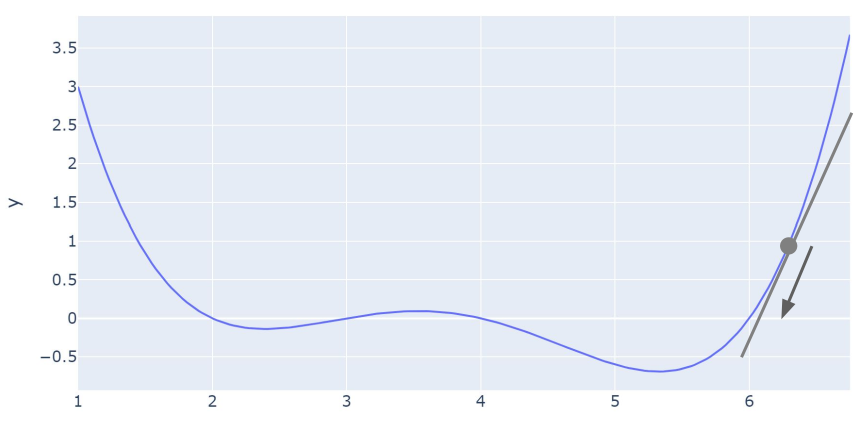
Finding the Minimum

We could start with a random guess.



Finding the Minimum

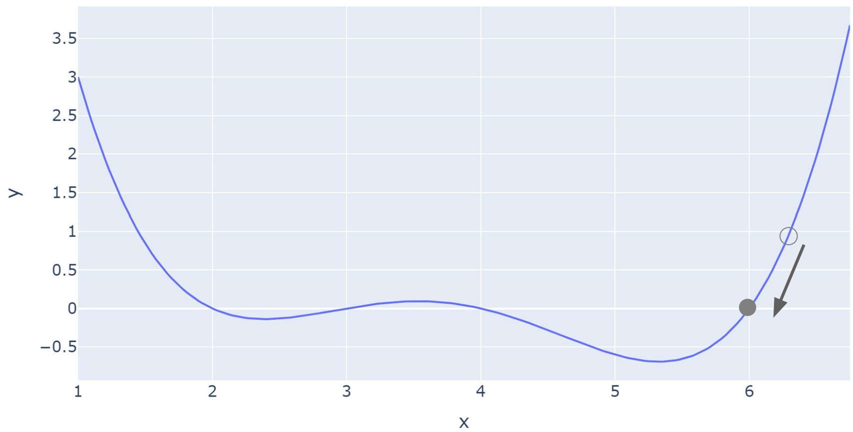
Where do we go next? We “step” downhill.



Follow the slope of the line down to the minimum.

Finding the Minimum

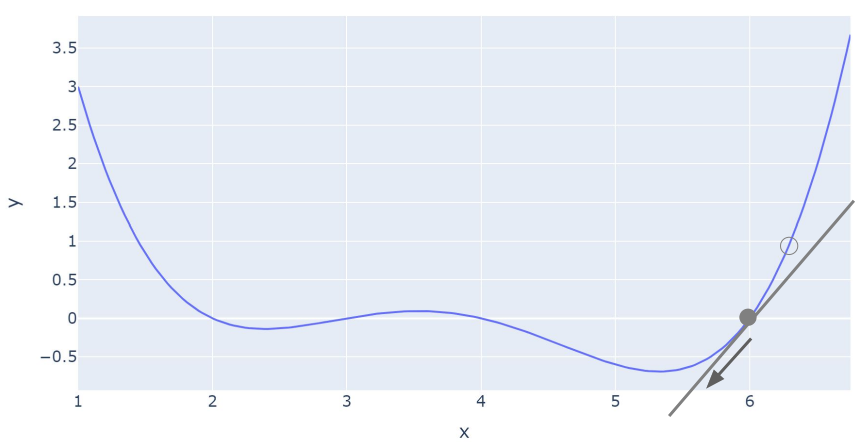
We arrive closer to the minimum.



Positive slope ⟶ step to the left

Finding the Minimum

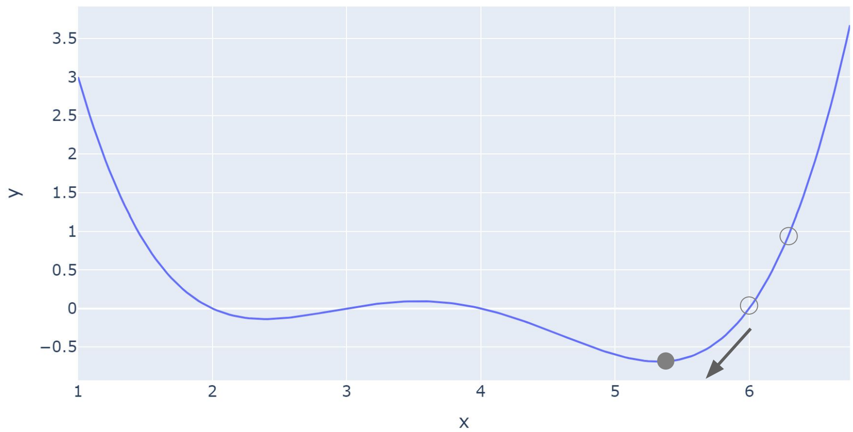
Do this again: follow the slope downwards towards the minimum



Positive slope ⟶ step to the left

Finding the Minimum

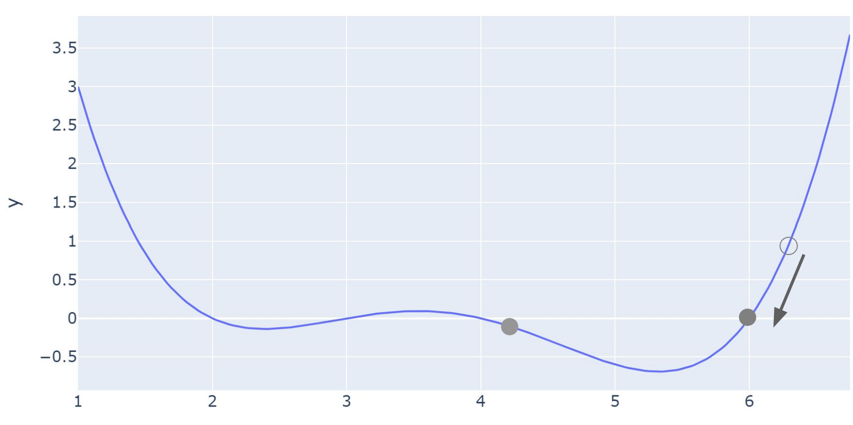
Do this again: follow the slope downwards towards the minimum



Positive slope ⟶ step to the left

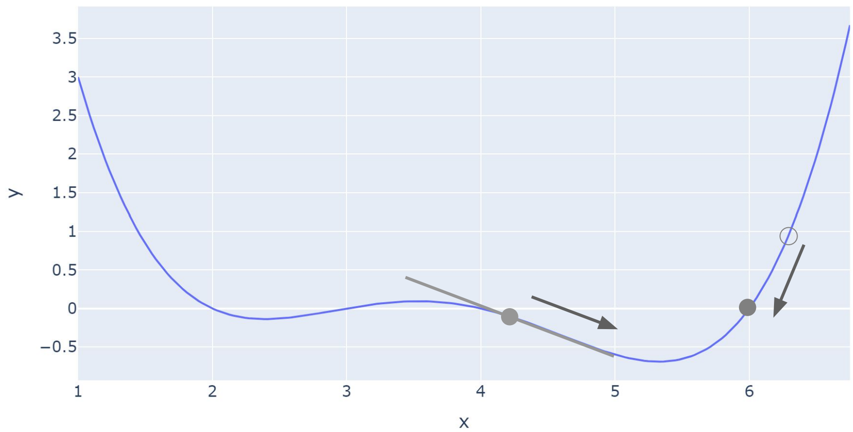
Finding the Minimum

What if we had started elsewhere?



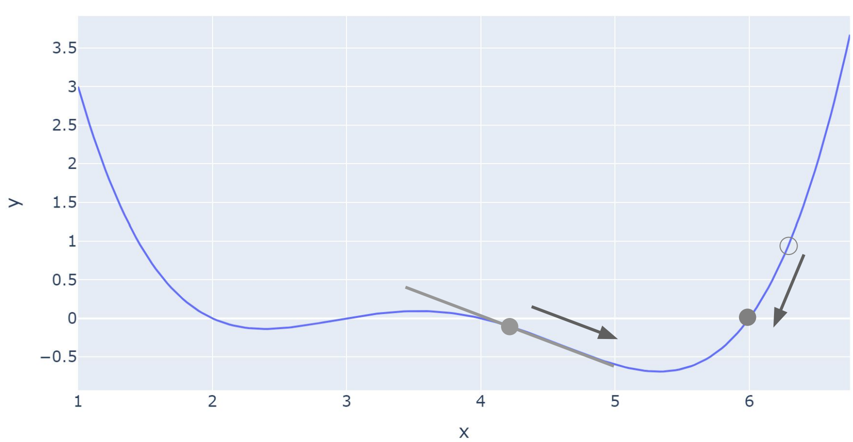
Finding the Minimum

What if we had started elsewhere?



Finding the Minimum

What if we had started elsewhere?

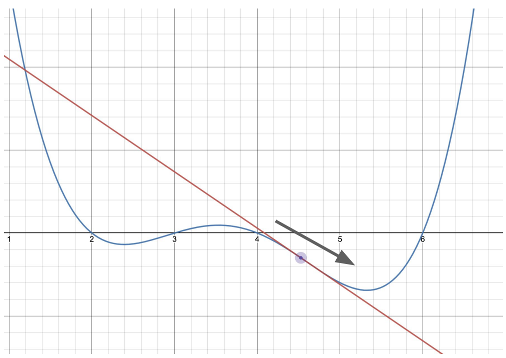


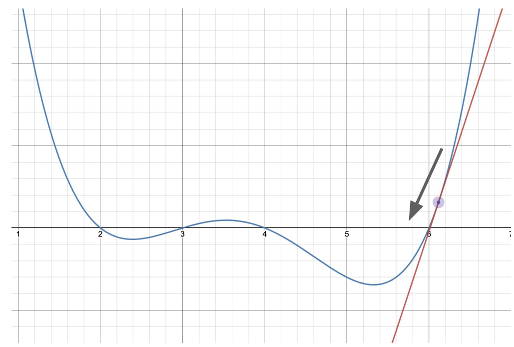
Negative slope → step to the right

Slopes Tell Us Where to Go

Negative slope → step to the right
Move in the *positive* direction

Positive slope → step to the left
Move in the *negative* direction





The derivative of the function at each point tells us the direction of our next guess.
Demo link: https://www.desmos.com/calculator/twpnylu4lr

Slopes Tell Us Where to Go

The derivative of the function at each point tells us the direction of our next guess.

Negative slope → step to the right
Move x in the *positive* direction

Positive slope → step to the left
Move x in the *negative* direction

Our first attempt at making an algorithm: step in the opposite direction to the slope

$$x^{(t+1)} = x^{(t)} - \frac{d}{dx} f\left(x^{(t)}\right)$$

Our next guess for the minimizing x

...starts at the previous guess

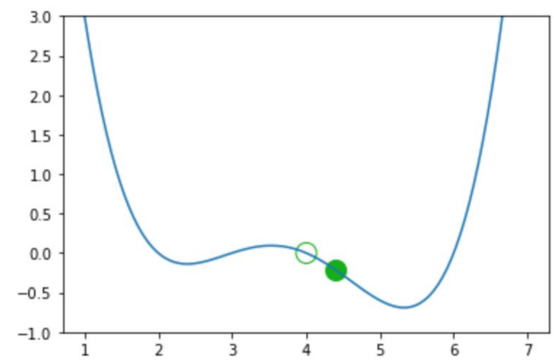
...and moves opposite to the slope

Algorithm Attempt #1

```python
def plot_one_step(x):
    new_x = x - derivative_arbitrary(x)
    plot_arbitrary()
    plot_x_on_f(arbitrary, new_x)
    plot_x_on_f_empty(arbitrary, x)
    print(f'old x: {x}')
    print(f'new x: {new_x}')
```

```python
plot_one_step(4)
```

```
old x: 4
new x: 4.4
```

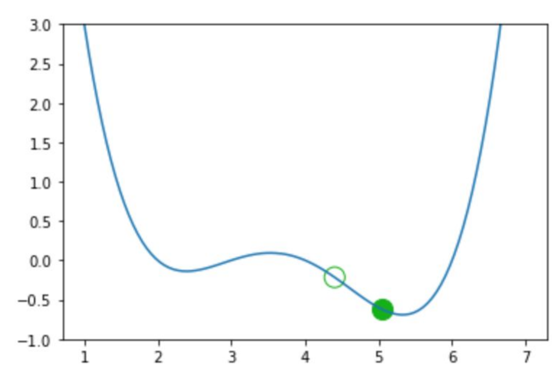


Algorithm Attempt #1

```python
def plot_one_step(x):
    new_x = x - derivative_arbitrary(x)
    plot_arbitrary()
    plot_x_on_f(arbitrary, new_x)
    plot_x_on_f_empty(arbitrary, x)
    print(f'old x: {x}')
    print(f'new x: {new_x}')
```

```python
plot_one_step(4.4)
```

```
old x: 4.4
new x: 5.0464000000000055
```

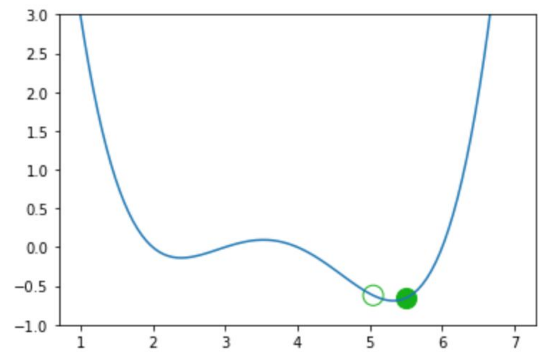


Algorithm Attempt #1

```python
def plot_one_step(x):
    new_x = x - derivative_arbitrary(x)
    plot_arbitrary()
    plot_x_on_f(arbitrary, new_x)
    plot_x_on_f_empty(arbitrary, x)
    print(f'old x: {x}')
    print(f'new x: {new_x}')
```

```python
plot_one_step(5.0464)
```

```
old x: 5.0464
new x: 5.49673060106241
```

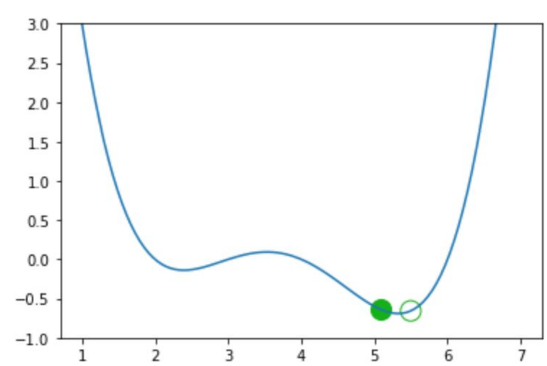


Algorithm Attempt #1

```python
def plot_one_step(x):
    new_x = x - derivative_arbitrary(x)
    plot_arbitrary()
    plot_x_on_f(arbitrary, new_x)
    plot_x_on_f_empty(arbitrary, x)
    print(f'old x: {x}')
    print(f'new x: {new_x}')
```

```python
plot_one_step(5.4967)
```

```
old x: 5.4967
new x: 5.080917145374805
```

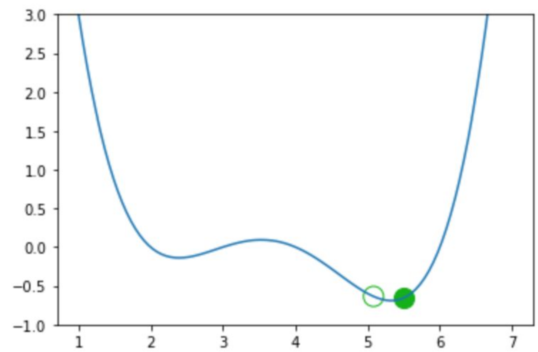


Algorithm Attempt #1

```python
def plot_one_step(x):
    new_x = x - derivative_arbitrary(x)
    plot_arbitrary()
    plot_x_on_f(arbitrary, new_x)
    plot_x_on_f_empty(arbitrary, x)
    print(f'old x: {x}')
    print(f'new x: {new_x}')
```

```python
plot_one_step(5.080917145374805)
```

```
old x: 5.080917145374805
new x: 5.489966698640582
```



Algorithm Attempt #1

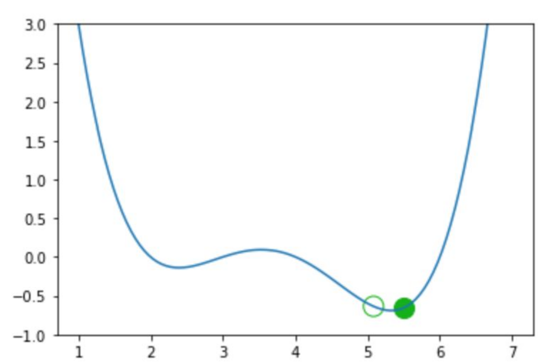
```python
def plot_one_step(x):
    new_x = x - derivative_arbitrary(x)
    plot_arbitrary()
    plot_x_on_f(arbitrary, new_x)
    plot_x_on_f_empty(arbitrary, x)
    print(f'old x: {x}')
    print(f'new x: {new_x}')
```

We appear to be bouncing back and forth. Turns out we are stuck!

- Any suggestions for how we can avoid this issue?

```python
plot_one_step(5.080917145374805)
```

```
old x: 5.080917145374805
new x: 5.489966698640582
```



Introducing a Learning Rate

Problem: each step is too big, so we overshoot the minimizing x

Solution: decrease the size of each step

Updated algorithm: α represents a **learning rate** that we choose. It controls the size of each step.

$$x^{(t+1)} = x^{(t)} - \alpha \frac{d}{dx} f\left(x^{(t)}\right)$$

Our next guess for the minimizing x

...starts at the previous guess

...and moves opposite to the slope

Let's try $\alpha = 0.3$

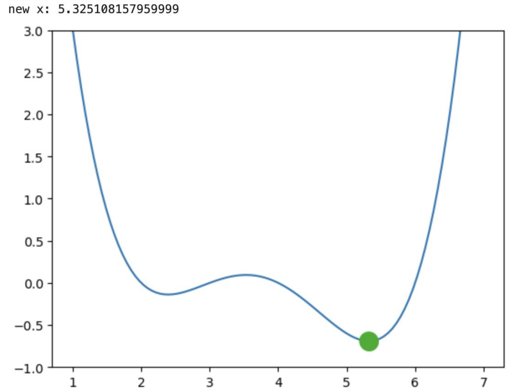
Algorithm Attempt #2

```python
def plot_one_step_lr(x):
    # Implement our new algorithm with a learning rate
    new_x = x - 0.3 * derivative_arbitrary(x)

    # Plot the updated guesses
    plot_arbitrary()
    plot_x_on_f(arbitrary, new_x)
    plot_x_on_f_empty(arbitrary, x)
    print(f'old x: {x}')
    print(f'new x: {new_x}')
```

```
plot_one_step_lr(5.323)
```

```
old x: 5.323
new x: 5.325108157959999
```



When do we stop updating?

Some options:

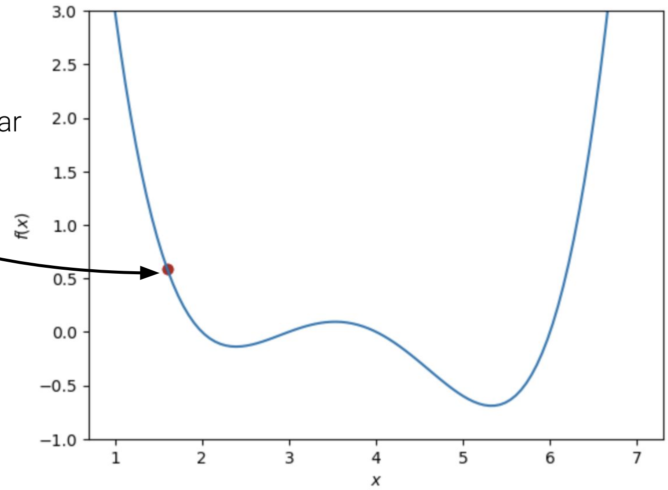
- After a fixed number of updates
- Subsequent update doesn't change "much"

Convergence: GD settles on a solution and stops updating significantly (or at all)

Convexity

What if our initial guess had been elsewhere?

Starting guess near local minimum

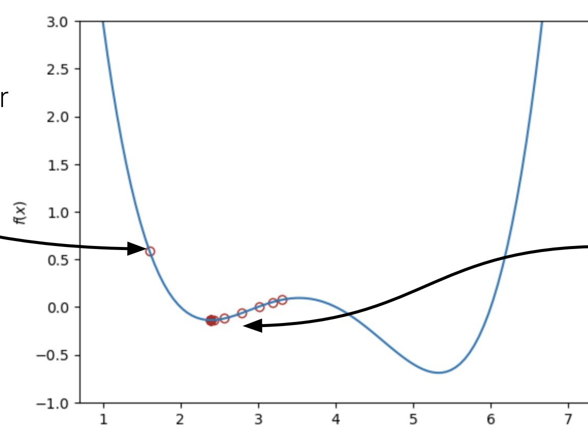


Convexity

What if our initial guess had been elsewhere?

The algorithm may have gotten “stuck” in a local minimum.

Starting guess near local minimum



Final guess fails to find the true global minimum

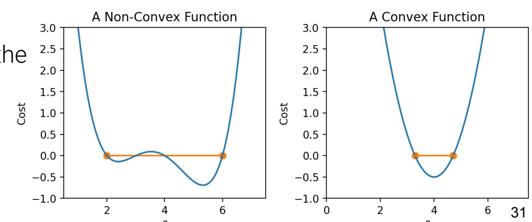
Convexity

For a **convex** function, any local minimum is a global minimum – we avoid the situation where the algorithm converges on some critical point that is not the minimum of the function.

- Our arbitrary function is non-convex

Algorithm is only guaranteed to converge (given enough iterations and an appropriate step size) for convex functions.

if I draw a line between any two points on the curve, all values on the curve must be at or below the line.



Agenda

- Optimization: where are we?
- Minimizing an arbitrary 1D function
- **Gradient descent on a 1D model**
- Gradient descent on high-dimensional models
- Batch, mini-batch, and stochastic gradient descent

From Arbitrary Functions to Loss Functions

In a modeling context, we aim to minimize a *loss function* by choosing the minimizing model *parameters.*

Terminology clarification:

- In past lectures, we have used “loss” to refer to the error incurred on a *single* datapoint
- In applications, we usually care more about the average error across *all* datapoints

Going forward, we will take the “model’s loss” to mean the model’s average error across the dataset. This is sometimes also known as the empirical risk, cost function, or objective function.

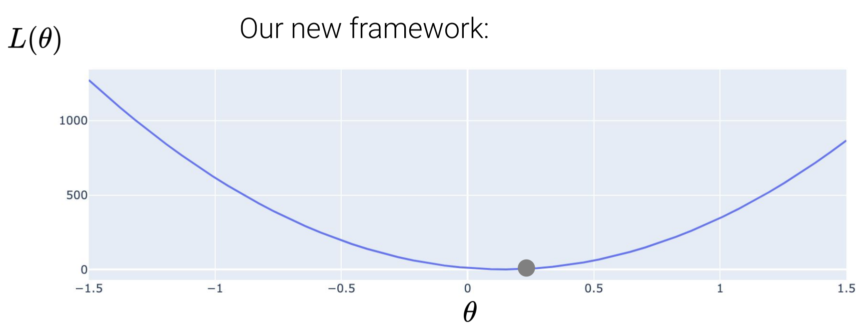
$$L(\theta) = R(\theta) = \frac{1}{n} \sum_{i=1}^{n} l(y,\, \hat{y})$$

From Arbitrary Functions to Loss Functions

In a modeling context, we aim to minimize a *loss function* by choosing the minimizing model *parameters.*

Goal: choose the value of θ that minimizes $L(\theta)$, the model’s loss on the dataset

Our new framework:



Compute the model’s loss for this choice of parameter

Test several values of the parameter θ

From Arbitrary Functions to Loss Functions

Goal: choose the value of θ that minimizes $L(\theta)$, the model's loss on the dataset

The **1D gradient descent** algorithm:

$$\theta^{(t+1)} = \theta^{(t)} - \alpha \frac{d}{d\theta} L\left(\theta^{(t)}\right)$$

Take our algorithm from before, replace x with θ and f with L.

Gradient Descent on the `tips` Dataset

We want to predict the tip (y) given the price of a meal (x). To do this:

- Choose a model: $\hat{y} = \theta_1 x$
- Choose a loss function: $L(\theta) = MSE(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \theta_1 x_i\right)^2$

Gradient Descent on the `tips` Dataset

We want to predict the tip (y) given the price of a meal (x). To do this:

- Choose a model: $\hat{y} = \theta_1 x$
- Choose a loss function: $L(\theta) = MSE(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta_1 x_i)^2$
- Optimize the model parameter – apply gradient descent

$$\theta^{(t+1)} = \theta^{(t)} - \alpha \frac{d}{d\theta} L\left(\theta^{(t)}\right)$$

Gradient Descent on the `tips` Dataset

- Optimize the model parameter – apply gradient descent

$$\theta^{(t+1)} = \theta^{(t)} - \alpha \frac{d}{d\theta} L\left(\theta^{(t)}\right)$$

Our loss function

$$L(\theta) = MSE(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta_1 x_i)^2$$

Gradient Descent on the `tips` Dataset

- Optimize the model parameter – apply gradient descent

$$\theta^{(t+1)} = \theta^{(t)} - \alpha \frac{d}{d\theta} L\left(\theta^{(t)}\right)$$

Our loss function

$$L(\theta) = MSE(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta_1 x_i)^2$$

Take the derivative wrt θ_1

$$\frac{d}{d\theta_1} L\left(\theta_1^{(t)}\right) = \frac{-2}{n}\sum_{i=1}^{n}\left(y_i - \theta_1^{(t)} x_i\right)x_i$$

$\longrightarrow$

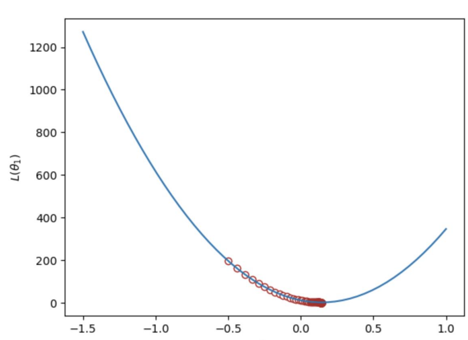
The gradient descent update rule

$$\theta_1^{(t+1)} = \theta_1^{(t)} - \alpha\frac{-2}{n}\sum_{i=1}^{n}\left(y_i - \theta_1^{(t)} x_i\right)x_i$$

Gradient Descent on the `tips` Dataset

Loss function: $MSE(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta_1 x_i)^2$

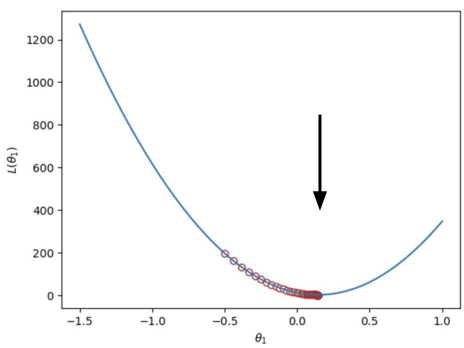
GD update rule: $\theta_1^{(t+1)} = \theta_1^{(t)} - \alpha\frac{-2}{n}\sum_{i=1}^{n}\left(y_i - \theta_1^{(t)} x_i\right)x_i$



MSE is minimized when we set $\theta_1 = 0.1437$

MSE is Convex!

When we visualized the MSE loss on the `tips` data, there was a single global minimum



This is one reason why the MSE is a popular choice of loss function: it behaves "nicely" for optimization

Agenda

- ❑ Optimization: where are we?
- ❑ Minimizing an arbitrary 1D function
- ❑ Gradient descent on a 1D model
- ❑ **Gradient descent on high-dimensional models**
- ❑ Batch, mini-batch, and stochastic gradient descent

Models in 2D or Higher

Usually, models will have more than one parameter that needs to be optimized.

Simple linear regression: $\hat{y} = \theta_0 + \theta_1 x$

Multiple linear regression: $\hat{\mathbb{Y}} = \theta_0 + \theta_1 \mathbb{X}_{:,1} + \theta_2 \mathbb{X}_{:,\,2} \ldots + \theta_p \mathbb{X}_{:,\,p}$

Idea: expand gradient descent so we can update our guesses for *all* model parameters, all in one go

Multiple Linear Regression

Define the **multiple linear regression** model:

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p$$

Parameters are $\theta = [\theta_0, \theta_1, \ldots, \theta_p]$

Is this linear in θ?

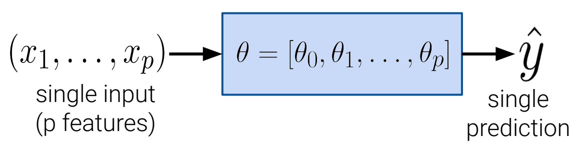
A. no
B. yes
C. maybe

Multiple Linear Regression

Define the **multiple linear regression** model:

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p$$

Parameters are $\theta = [\theta_0, \theta_1, \ldots, \theta_p]$

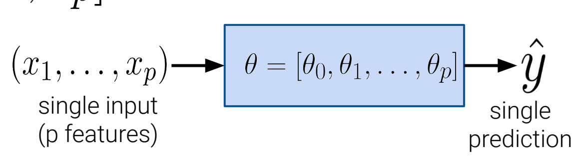


Multiple Linear Regression

Define the **multiple linear regression** model:

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p$$

Parameters are $\theta = [\theta_0, \theta_1, \ldots, \theta_p]$



From SLR to Multiple linear regression

x	y
x_1	y_1
x_2	y_2
$\vdots$	$\vdots$
x_n	y_n

	FG	PTS
1	1.8	5.3
2	0.4	1.7
3	1.1	3.2
4	6.0	13.9
5	3.4	8.9
...	...	...

$x_{:,1}$	$x_{:,2}$		$x_{:,p}$	y
x_{11}	x_{12}	$\ldots$	x_{1p}	y_1
x_{21}	x_{22}	$\ldots$	x_{2p}	y_2
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_{n1}	x_{n2}	$\ldots$	x_{np}	y_n

	FG	AST	3PA	PTS
1	1.8	0.6	4.1	5.3
2	0.4	0.8	1.5	1.7
3	1.1	1.9	2.2	3.2
4	6.0	1.6	0.0	13.9
5	3.4	2.2	0.2	8.9
...	...	...	...	...

Multiple Linear Regression

Define the **multiple linear regression** model:

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p$$

$$\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} \longrightarrow \boxed{\begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{bmatrix}} \longrightarrow \hat{y}$$

single **observation**
(p features)

single
prediction

Vector Notation

$$\hat{y} = \theta_0 + \underbrace{\theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p}_{\text{This part looks a little like a dot product...}}$$

We want to collect all the θ_i's into a single vector.

$$= \boxed{\theta_0} + \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} = \theta_0 \cdot 1 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p$$

😤What about this one???

Vector Notation

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p$$
$$= \theta_0 \cdot 1 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p$$

We want to collect all the θ_i's into a single vector.

$$= \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{bmatrix} \cdot \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} = x^\top \theta$$

bias term, intercept term

Matrix Notation

$$\begin{cases} \hat{y}_1 = \theta_0 + \theta_1 x_{11} + \theta_2 x_{12} + \cdots + \theta_p x_{1p} \\ \hat{y}_2 = \theta_0 + \theta_1 x_{21} + \theta_2 x_{22} + \cdots + \theta_p x_{2p} \\ \vdots \\ \hat{y}_n = \theta_0 + \theta_1 x_{n1} + \theta_2 x_{n2} + \cdots + \theta_p x_{np} \end{cases}$$

$$\begin{cases} \hat{y}_1 = x_1^\top \theta & \text{where } x_1^\top = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \end{bmatrix} \text{ is datapoint/observation 1} \\ \hat{y}_2 = x_2^\top \theta & \text{where } x_2^\top = \begin{bmatrix} 1 & x_{21} & x_{22} & \dots & x_{2p} \end{bmatrix} \text{ is datapoint/observation 2} \\ \vdots \\ \hat{y}_n = x_n^\top \theta & \text{where } x_n^\top = \begin{bmatrix} 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \text{ is datapoint/observation n} \end{cases}$$

Matrix Notation

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \boxed{\begin{matrix} \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \end{bmatrix} \\ \begin{bmatrix} 1 & x_{21} & x_{22} & \dots & x_{2p} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \end{matrix}} \; \theta$$

n row vectors, each with dimension **(p+1)**

Vectorize predictions and parameters to encapsulate all n equations into a single matrix equation.

Matrix Notation

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \mathbb{X} \; \theta$$

Design matrix with dimensions n x (p + 1)

The Multiple Linear Regression Model Using Matrix Notation

We can express our linear model on our entire dataset as follows:

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ 1 & x_{31} & x_{32} & \dots & x_{3p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{bmatrix}$$

$$\boxed{\hat{\mathbb{Y}} = \mathbb{X}\theta}$$

Prediction vector $\mathbb{R}^n$

Design matrix $\mathbb{R}^{n \times (p+1)}$

Parameter vector $\mathbb{R}^{(p+1)}$

Note that our **true output** is also a vector: $\mathbb{Y} \in \mathbb{R}^n$

Mean Squared Error with L2 Norms

We can rewrite mean squared error as a squared L2 norm:

$$R(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$
$$= \frac{1}{n}||\mathbb{Y} - \hat{\mathbb{Y}}||_2^2$$

With our linear model $\hat{\mathbb{Y}} = \mathbb{X}\theta$:

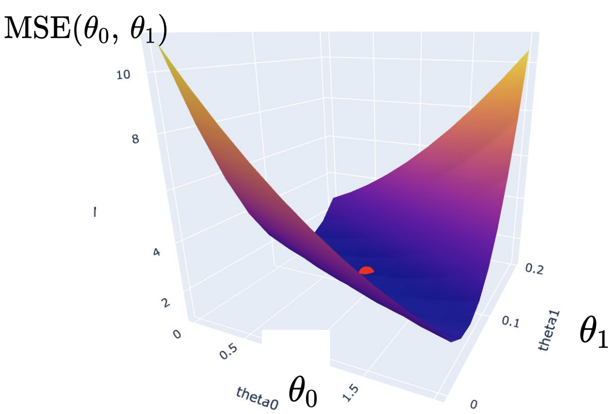
$$R(\theta) = \frac{1}{n}||\mathbb{Y} - \mathbb{X}\theta||_2^2$$

Back to GD

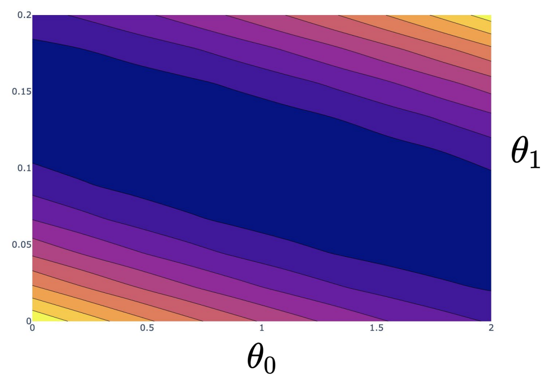
Models in 2D or Higher

With multiple parameters to optimize, we consider a **loss surface**

- What is the model's loss for a particular *combination* of possible parameter values?



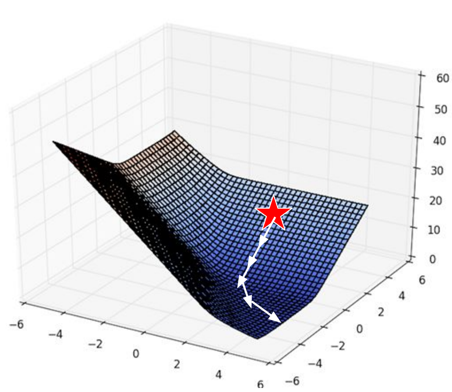
A bird's eye view from above:



The Gradient Vector

As before, the derivative of the loss function tells us the best way towards the minimum value

On a 2D (or higher) surface, the best way to go down (gradient) is described by a *vector*



For the *vector* of parameter values $\vec{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$

Take the *partial derivative* of loss with respect to each parameter

A Math Aside: Partial Derivatives

For an equation with multiple variables, we take a **partial derivative** by differentiating with respect to just one variable at a time.

Intuitively: how does the function change if we vary one variable, while holding the others constant?

$$f(x, y) = 3x^2 + y$$

Take the partial derivative wrt x: treat y as a constant $\quad \dfrac{\partial f}{\partial x} = 6x$

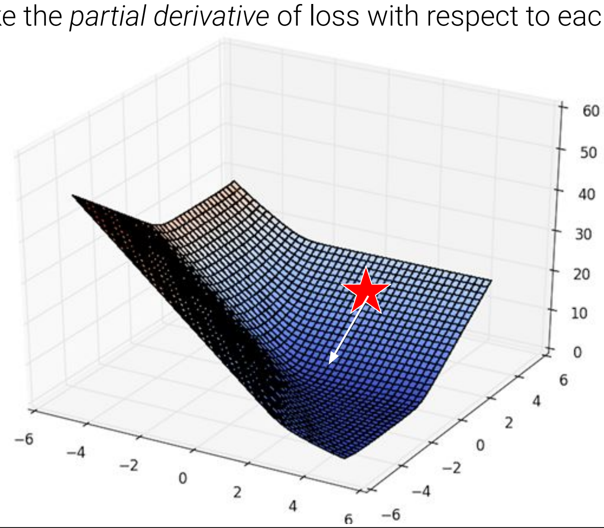
Take the partial derivative wrt y: treat x as a constant $\quad \dfrac{\partial f}{\partial y} = 1$

This symbol means "partial derivative"

The Gradient Vector

For the *vector* of parameter values $\vec{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$

Take the *partial derivative* of loss with respect to each parameter: $\dfrac{\partial L}{\partial \theta_0'} \quad \dfrac{\partial L}{\partial \theta_1}$



The **gradient vector** is

$$\nabla_{\vec{\theta}} L = \begin{bmatrix} \frac{\partial L}{\partial \theta_0} \\ \frac{\partial L}{\partial \theta_1} \end{bmatrix}$$

$-\nabla_{\vec{\theta}} L$ always points in the downhill direction of the surface.

Gradient Descent in Multiple Dimensions

Recall our 1D update rule: $$\theta^{(t+1)} = \theta^{(t)} - \alpha \frac{d}{d\theta} L\left(\theta^{(t)}\right)$$

Now, for models with multiple parameters, we work in terms of vectors:

$$\begin{bmatrix} \theta_0^{(t+1)} \\ \theta_1^{(t+1)} \\ \vdots \end{bmatrix} = \begin{bmatrix} \theta_0^{(t)} \\ \theta_1^{(t)} \\ \vdots \end{bmatrix} - \alpha \begin{bmatrix} \frac{\partial L}{\partial \theta_0} \\ \frac{\partial L}{\partial \theta_1} \\ \vdots \end{bmatrix}$$

Written in a more compact form:

$$\vec{\theta}^{(t+1)} = \vec{\theta}^{(t)} - \alpha \nabla_{\vec{\theta}} L\left(\vec{\theta}^{(t)}\right)$$

Gradient Descent Update Rule

Gradient descent algorithm: nudge θ in negative gradient direction until θ converges.

For a model with multiple parameters:

$$\vec{\theta}^{(t+1)} = \vec{\theta}^{(t)} - \alpha \underbrace{\nabla_{\vec{\theta}} L\left(\vec{\theta}^{(t)}\right)}_{\text{gradient of the loss function evaluated at current } \theta}$$

Next value for θ

θ: Model weights   **L**: loss function

α: Learning rate (ours is constant; other techniques have α decrease over time)

62

Agenda

- Optimization: where are we?
- Minimizing an arbitrary 1D function
- Gradient descent on a 1D model
- Gradient descent on high-dimensional models
- **Batch, mini-batch, and stochastic gradient descent**

Batch Gradient Descent

We have just derived **batch gradient descent.**

- We used our *entire* dataset (as one big batch) to compute gradients
- Recall the derivative of MSE for our 1D model – involves working with *all* n datapoints

$$\frac{d}{d\theta_1} L\left(\theta_1^{(t)}\right) = \frac{-2}{n} \sum_{i=1}^{n} \left(y_i - \theta_1^{(t)} x_i\right) x_i$$

Using all datapoints is often impractical when our dataset is large.

Computing each gradient will take a long time; gradient descent will converge slowly because each individual update is slow.

Mini-batch Gradient Descent

An alternative: use only a *subset* of the full dataset at each update.

Estimate the true gradient of the loss surface using just this subset of the data.

Batch size: the number of datapoints to use in each subset

In mini-batch GD:

- Compute the gradient on the first x% of the data. Update the parameter guesses.
- Compute the gradient on the next x% of the data. Update the parameter guesses.
- …
- Compute the gradient on the last x% of the data. Update the parameter guesses.

Training Epoch

Mini-batch Gradient Descent

In mini-batch GD:

- Compute the gradient on the first x% of the data. Update the parameter guesses.
- Compute the gradient on the next x% of the data. Update the parameter guesses.
- …
- Compute the gradient on the last x% of the data. Update the parameter guesses.

Training Epoch

In a single training epoch, we use every datapoint in the data once.

We then perform several training epochs until we are satisfied.

Stochastic Gradient Descent

- In the most extreme case, we may perform gradient descent with a batch size of just one datapoint – this is called stochastic gradient descent.
- Works surprisingly well in practice! Averaging across several epochs gives a similar result as directly computing the true gradient on all the data.

In stochastic GD:

- Compute the gradient on the first datapoint. Update the parameter guesses.
- Compute the gradient on the next datapoint. Update the parameter guesses.
- …
- Compute the gradient on the last datapoint. Update the parameter guesses.

Training Epoch

Comparing GD Techniques

Batch gradient descent:

- Computes the true gradient
- Always descends towards the true minimum of loss

Mini-batch/stochastic gradient descent:

- *Approximates* the true gradient
- May not descend towards the true minimum with each update