

Least Squares Optimization and Gradient Descent Algorithm

Example

- Single Variable Linear Regression

estimate $\hat{y}_i = \theta_0 + \theta_1 x_i$

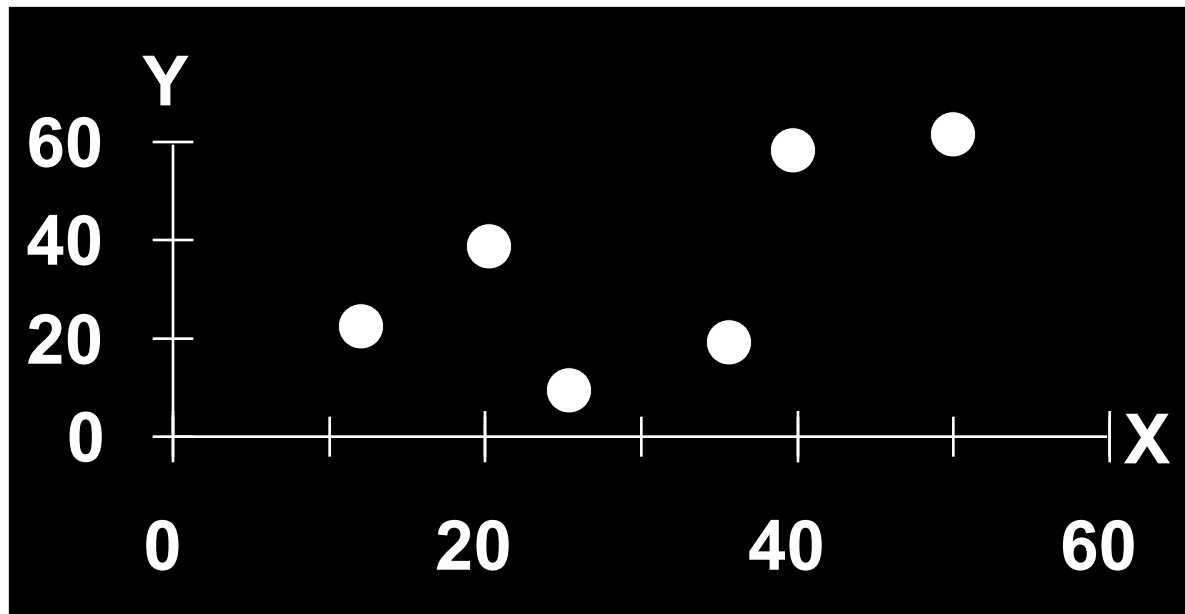
x	y
12	20
20	40
25	10
35	20
40	60
50	65



**ESTIMATING PARAMETERS:
LEAST SQUARES METHOD**

SCATTER PLOT

Plot all (X_i, Y_i) pairs, and plot your learned model

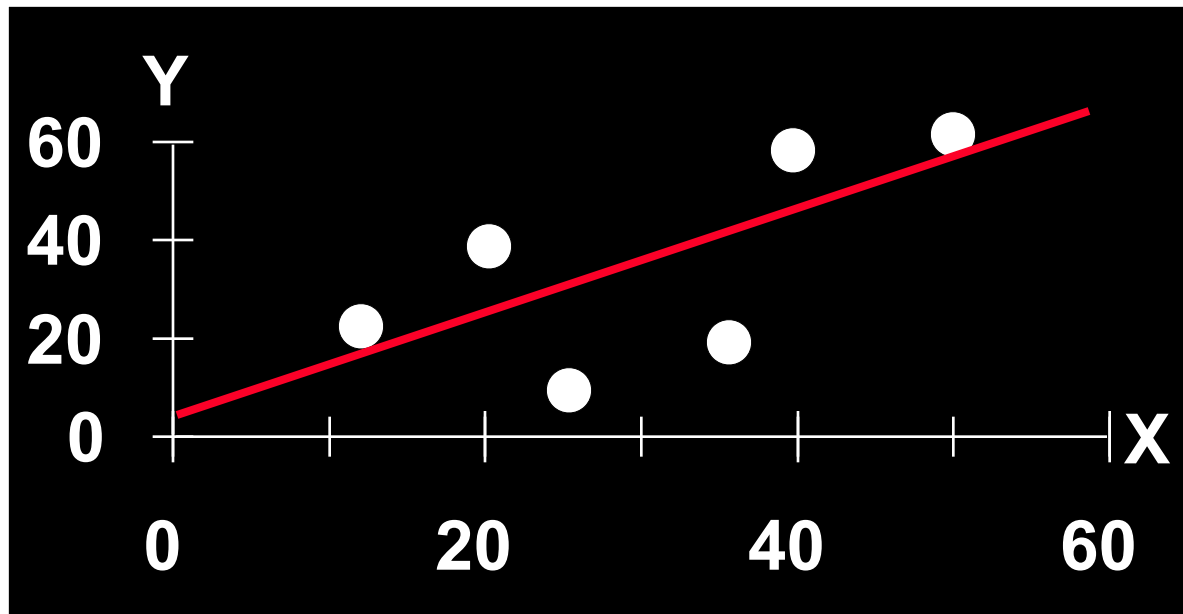


QUESTION

How would you draw a line through the points?

How do you determine which line “fits the best” ...?

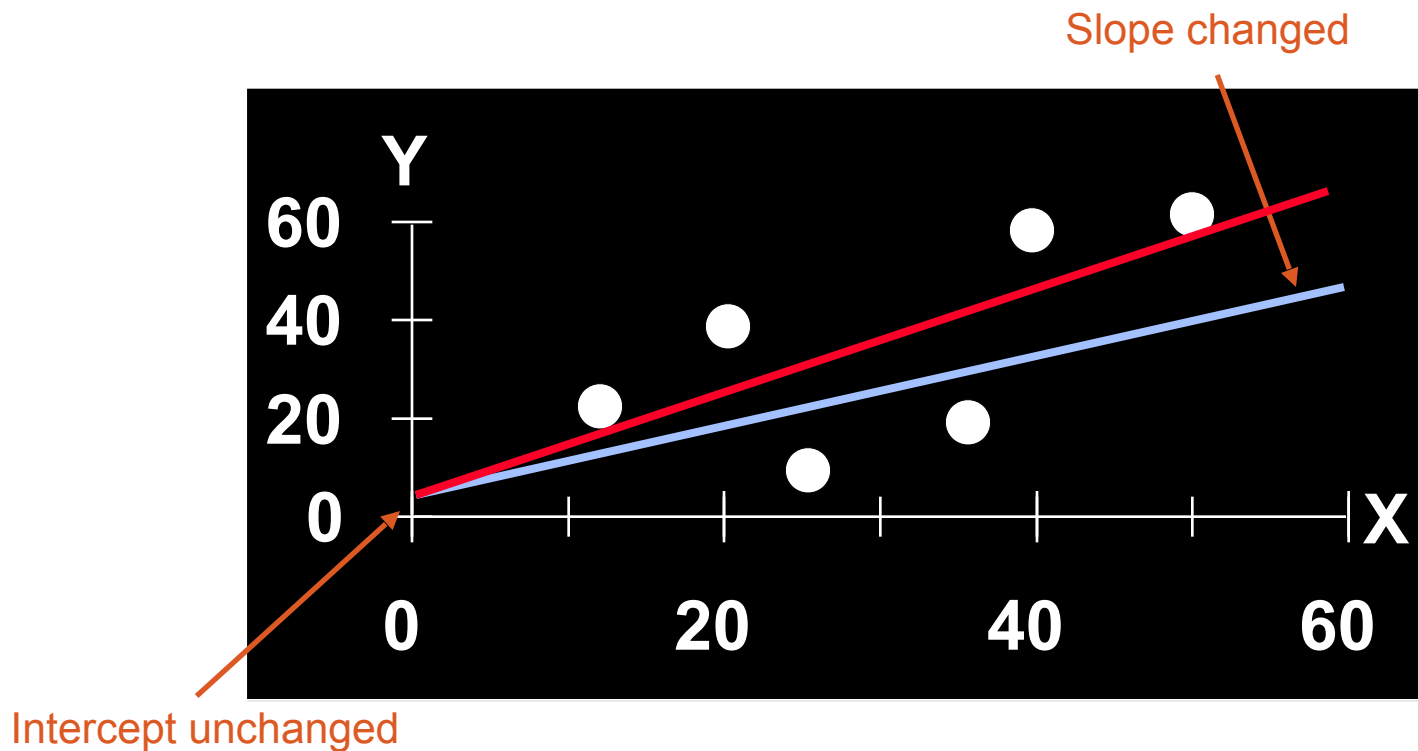
??????????



QUESTION

How would you draw a line through the points?

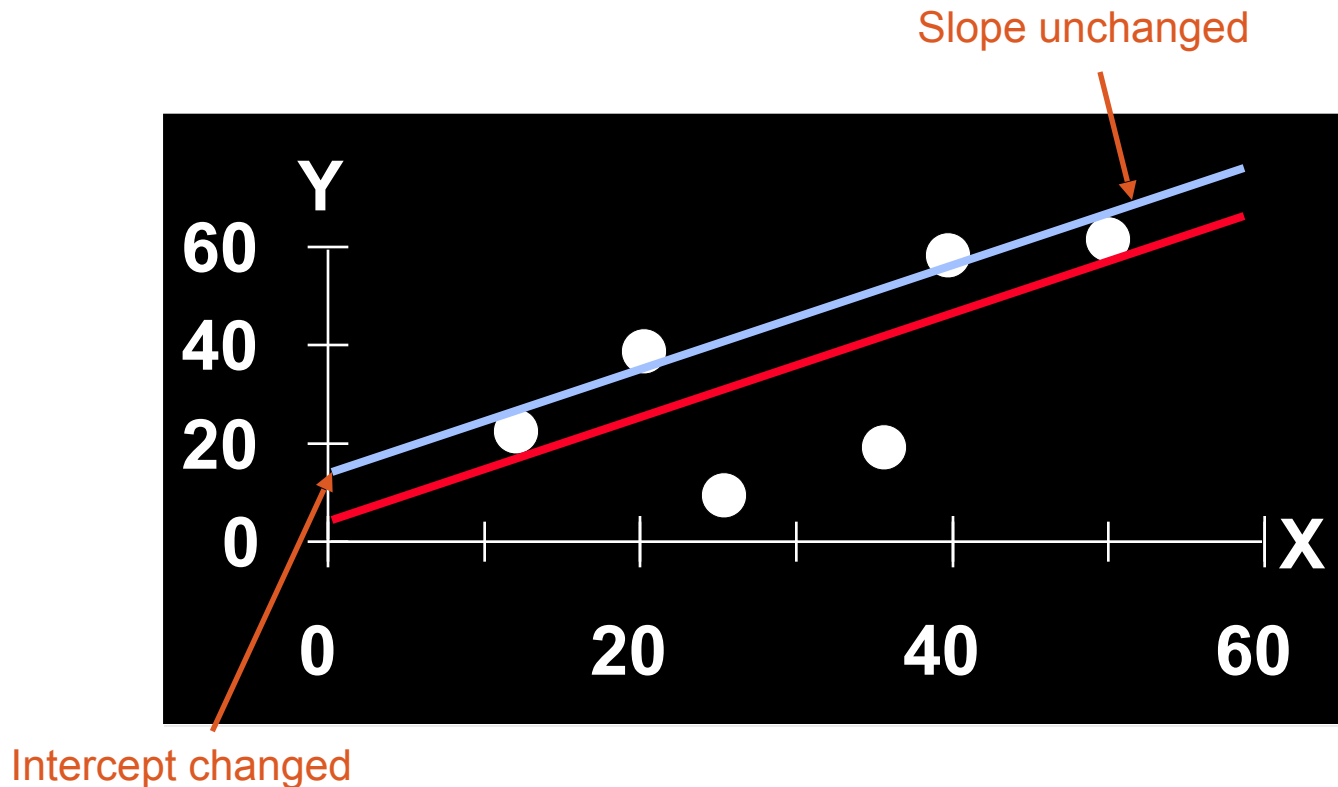
How do you determine which line “fits the best” ??????????



QUESTION

How would you draw a line through the points?

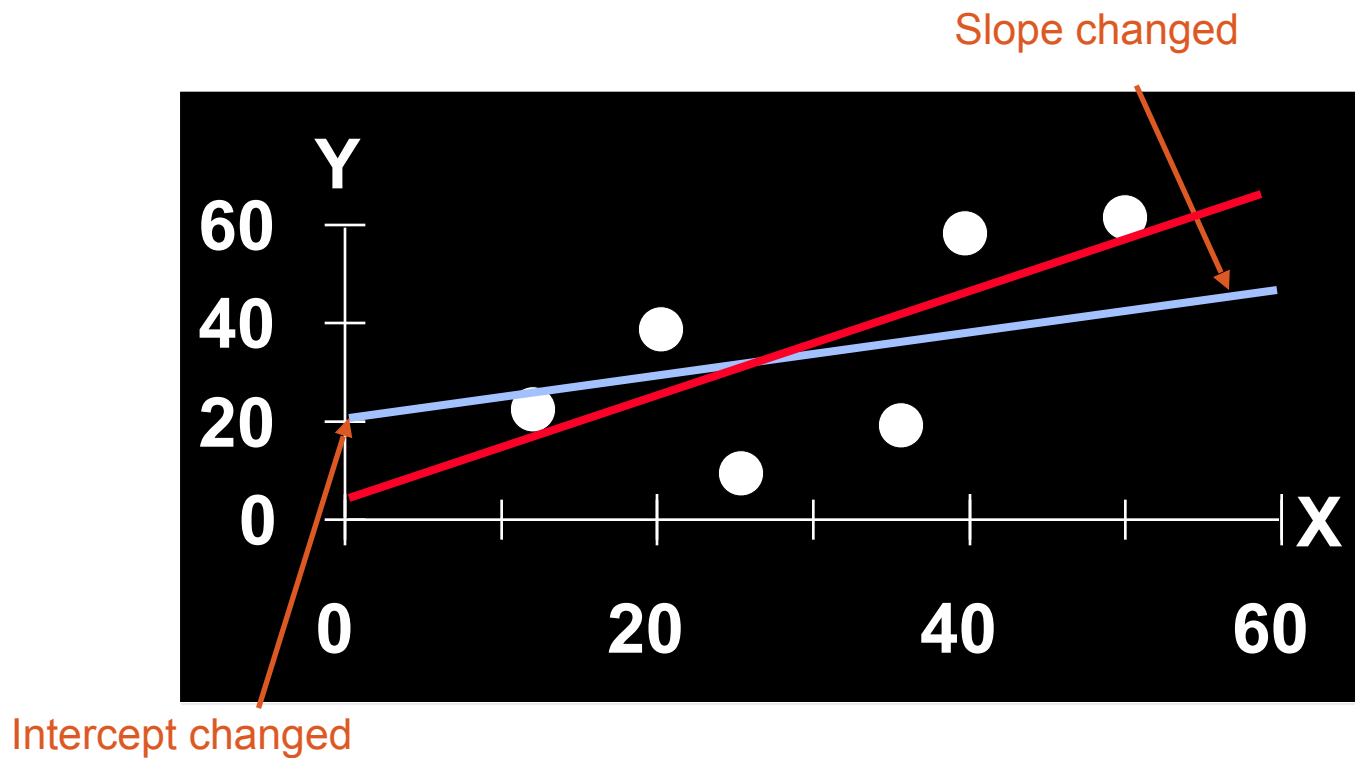
How do you determine which line “fits the best” ??????????



QUESTION

How would you draw a line through the points?

How do you determine which line “fits the best” ??????????



LEAST SQUARES

Best fit: difference between the true (observed) Y-values and the estimated Y-values is minimized:

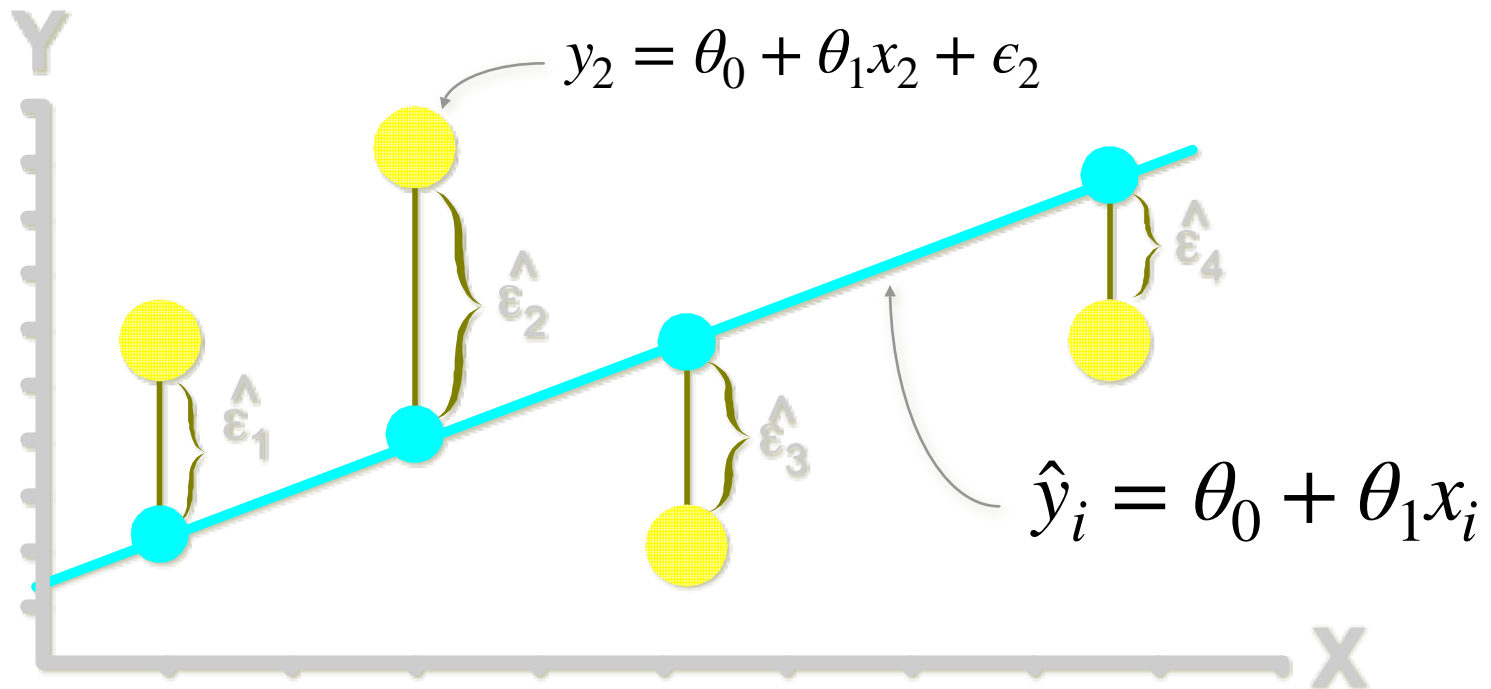
- Positive errors offset negative errors ...
- ... square the error!

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

Least squares minimizes the sum of the squared errors

LEAST SQUARES, GRAPHICALLY

LS Minimizes $\sum_{i=1}^n \epsilon_i^2 = \epsilon_1^2 + \epsilon_2^2 + \dots + \epsilon_n^2$



Example

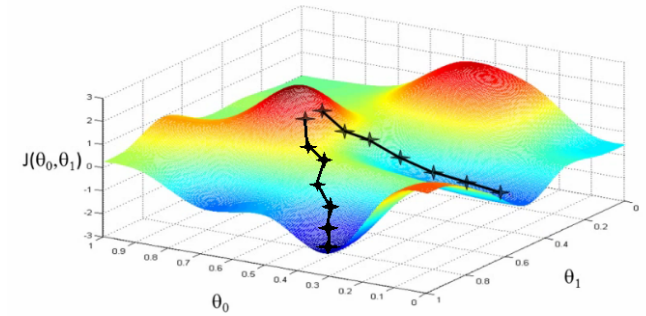
- Single Variable Linear Regression

estimate $\hat{y}_i = \theta_0 + \theta_1 x_i$

x	y
Area(sq. ft.)	Price (in 1000\$)
1600	220
1400	180
2100	350
...	...
....	
2400	500

Multivariate Regression

- Multi Linear Regression



$$\hat{y}_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \dots + \theta_m x_{im}$$

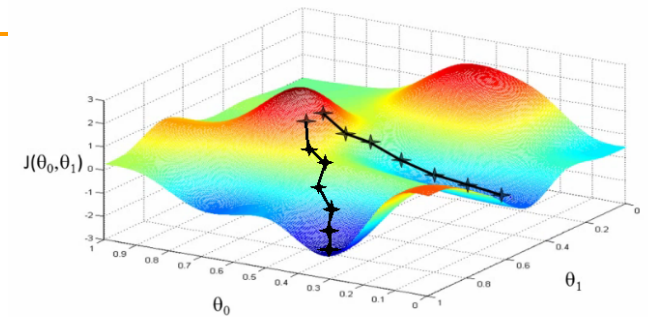
y Price (in 1000\$)	x_1 Area(sq. ft.)	x_2 # Bathrooms	x_3 # Bedrooms
220	1600	2.5	3
180	1400	1.5	3
350	2100	3.5	4
...	...	...	...
....		...	...
500	2400	4	5

$$\hat{y}_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \dots + \theta_m x_{im}$$

Price (in 1000\$)	Area(sq. ft.)	# Bathrooms	# Bedrooms
220	1600	2.5	3
180	1400	1.5	3
350	2100	3.5	4
...	...	...	...
500	2400	4	5

1400	x_{i1}
1.5	x_{i2}
3	x_{i3}

Multivariate Regression



Multi Linear Regression

$$y_i = \theta_0 x_{i0} + \theta_1 x_{i1} + \theta_2 x_{i2} + \dots + \theta_m x_{im}$$

y	x_0	x_1	x_2	x_3	
Price (in 1000\$)		Area(sq. ft.)	# Bathrooms	# Bedrooms	
220	1	1600	2.5	3	
y_i 180	1	1400	1.5	3	
350	1	2100	3.5	4	
...	...	...	...	...	
....			...	...	
500	1	2400	4	5	

	x_i
1	x_{i0}
1400	x_{i1}
1.5	x_{i2}
3	x_{i3}

Multivariate Regression Model

- Model:

$$\hat{y}_i = \theta_0 x_{i0} + \theta_1 x_{i1} + \theta_2 x_{i2} + \dots + \theta_m x_{im}$$

$$\hat{y}_i = \sum_{j=0}^m \theta_{ij} x_{ij}$$

feature 1 = x_0 (constant, 1)

feature 2 = x_1 (area, sq. ft.)

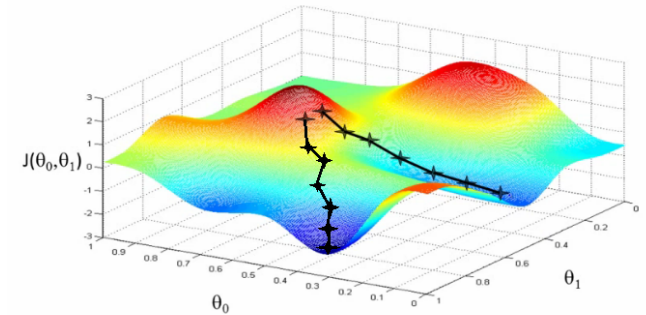
feature 3 = x_2 (# of bedrooms)

feature 4 = x_3 (# of bathrooms)

....

....

feature m = x_m



One Observation Model

- Matrix Notation
For observation i

$$\hat{y}_i = \sum_{j=0}^m \theta_{ij} x_{ij}$$

$y_i =$

--	--	--	--	--	--	--

 $x_{i0} \quad x_{i1} \quad x_{i2} \quad \dots \quad x_{im}$

 $\theta_0 \quad \theta_1 \quad \theta_2 \quad \dots \quad \theta_m$

$$y_i = X_i^T \theta$$

All Observation Model

- Matrix Notation
For all observations

x_{10}	x_{11}	x_{12}	..	..	x_{im}
x_{20}	x_{21}	x_{22}	..	..	x_{2m}
x_{30}	x_{31}	x_{32}	..	..	x_{3m}
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
x_{n0}	x_{n1}	x_{n2}	..	..	x_{nm}

θ_0
θ_1
θ_2
.
θ_m

 $=$

y_0
y_1
y_2
.
.
.
y_n

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

LEAST SQUARES OPTIMIZATION

Rewrite inputs:

Each row is a feature vector paired with a label for a single input

$$X = \begin{bmatrix} (x^{(1)})^T \\ (x^{(2)})^T \\ \dots \\ (x^{(n)})^T \end{bmatrix} \in \mathbb{R}^{n \times m}, \quad y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \dots \\ y^{(n)} \end{bmatrix} \in \mathbb{R}^n$$

n labeled inputs

m features

Rewrite optimization problem:

$$\text{minimize}_{\theta} \frac{1}{2} \|X\theta - y\|_2^2$$

*Recall $\|z\|_2^2 = z^T z = \sum z_i^2$

LEAST SQUARES OPTIMIZATION

Rewrite inputs:

Each row is a feature vector paired with a label for a single input

$$X = \begin{bmatrix} (x^{(1)})^T \\ (x^{(2)})^T \\ \dots \\ (x^{(n)})^T \end{bmatrix} \in \mathbb{R}^{n \times m}, y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \dots \\ y^{(n)} \end{bmatrix} \in \mathbb{R}^n$$

n labeled inputs

m features

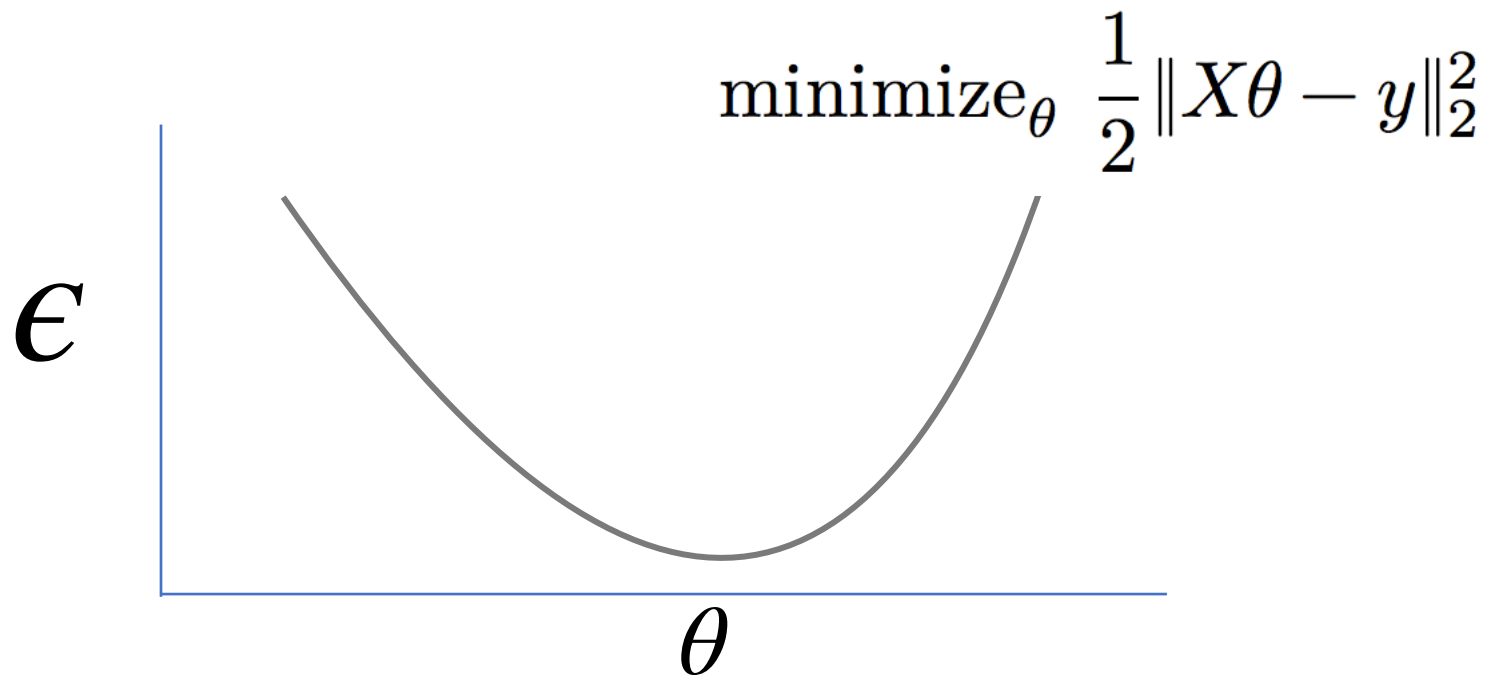
Rewrite optimization problem:

$$\text{minimize}_{\theta} \frac{1}{2} \|X\theta - y\|_2^2$$

$$\Rightarrow \text{minimize} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \epsilon_i^2$$

*Recall $\|z\|_2^2 = z^T z = \sum z_i^2$

ERROR FUNCTION



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

GRADIENTS

Minimizing a multivariate function involves finding a point where the gradient is zero:

$$\nabla_{\theta} f(\theta) = 0 \text{ (the vector of zeros)}$$

Points where the gradient is zero are **local** minima

- If the function is convex, also a **global** minimum

Let's solve the least squares problem!

We'll use the multivariate generalizations of some concepts from MATH141/142 ...

- Chain rule: $\nabla_{\theta} f(X\theta) = X^T \nabla_{X\theta} f(X\theta)$
- Gradient of squared ℓ^2 norm: $\nabla_{\theta} \|\theta - z\|_2^2 = 2(\theta - z)$

LEAST SQUARES

Recall the least squares optimization problem:

$$\text{minimize}_{\theta} \frac{1}{2} \|X\theta - y\|_2^2$$

What is the gradient of the optimization objective ????????

$$\nabla_{\theta} \frac{1}{2} \|X\theta - y\|_2^2 =$$

Chain rule:

$$\nabla_{\theta} f(X\theta) = X^T \nabla_{X\theta} f(X\theta)$$

$$X^T \nabla_{X\theta} \frac{1}{2} \|X\theta - y\|_2^2 =$$

Gradient of norm:

$$\nabla_{\theta} \|\theta - z\|_2^2 = 2(\theta - z)$$

$$\nabla_{\theta} \frac{1}{2} \|X\theta - y\|_2^2 = X^T (X\theta - y)$$

LEAST SQUARES

Recall: points where the gradient **equals zero** are minima.

$$\nabla_{\theta} \frac{1}{2} \|X\theta - y\|_2^2 = X^T (X\theta - y)$$

So where do we go from here?????????

$$X^T (X\theta - y) = 0$$

Solve for model
parameters θ

$$X^T X\theta - X^T y = 0 \Rightarrow X^T X\theta = X^T y$$

$$(X^T X)^{-1} X^T X\theta = (X^T X)^{-1} X^T y$$

$$\theta = (X^T X)^{-1} X^T y$$

LINEAR REGRESSION AS OPTIMIZATION PROBLEM

Let's consider linear regression that minimizes the sum of squared error, i.e., least squares ...

1. Hypothesis function: ????????

- **Linear** hypothesis function $h_{\theta}(x) = \theta^T x$

2. Loss function: ????????

- **Squared** error loss $\ell(\hat{y}, y) = \frac{1}{2} (\hat{y} - y)^2$

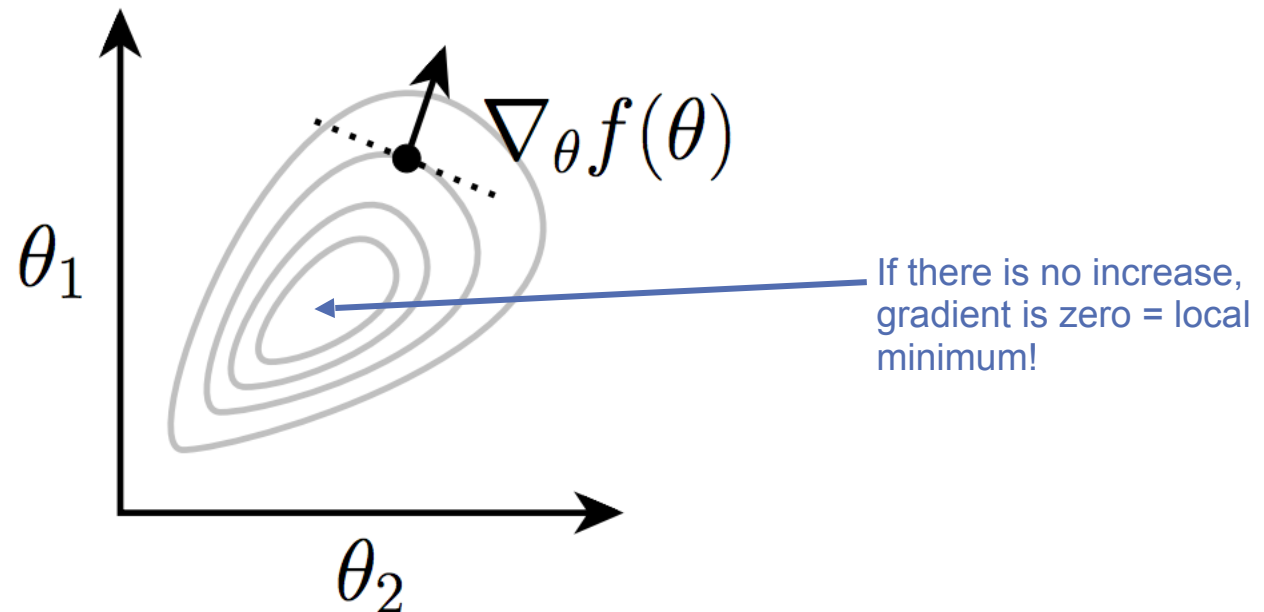
4. Optimization problem: ????????

$$\min_{\theta} \sum_{i=1}^n (\theta^T x^{(i)} - y^{(i)})^2$$

GRADIENT DESCENT

We used the gradient as a condition for optimality

It also gives the local **direction of steepest increase** for a function:



Intuitive idea: take small steps **against** the gradient.

GRADIENT DESCENT

Algorithm for any* hypothesis function $h_\theta: \mathbb{R}^n \rightarrow \mathcal{Y}$, loss function $\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+$, step size α :

Initialize the parameter vector:

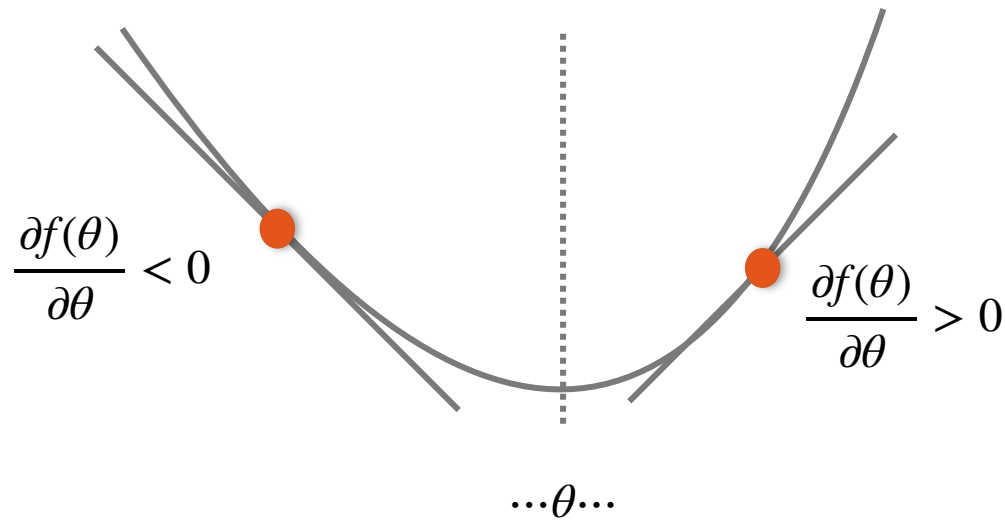
- $\theta \leftarrow 0$

Repeat until satisfied (e.g., exact or approximate convergence):

- Compute gradient: $g \leftarrow \sum_{i=1}^n \nabla_{\theta} \ell(h_{\theta}(x^{(i)}), y^{(i)})$
- Update parameters: $\theta \leftarrow \theta - \alpha \cdot g$

*must be reasonably well behaved

GRADIENT DESCENT



$$\theta := \theta - \alpha \frac{\partial f(\theta)}{\partial \theta}$$

EXAMPLE

Function: $f(x,y) = x^2 + 2y^2$

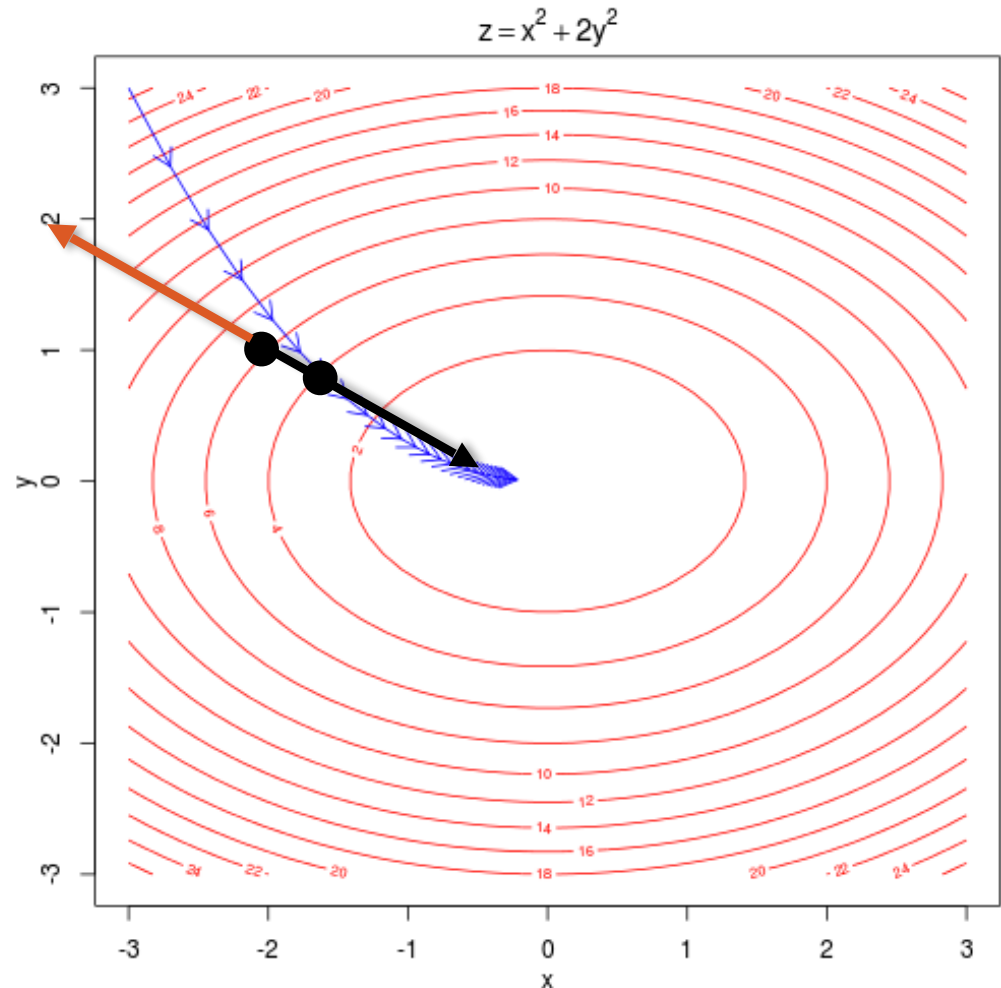
Gradient: ???????????

$$\nabla f(x,y) = \begin{bmatrix} 2x \\ 4y \end{bmatrix}$$

Let's take a gradient step
from $(-2, +1)$:

$$\nabla f(-2,1) = \begin{bmatrix} -4 \\ 1 \end{bmatrix}$$

Step in the direction $(0.04, -0.01)$, scaled by step size
Repeat until no movement



GRADIENT DESCENT

$$\hat{y} = \theta_0 + \theta_1 x$$

GRADIENT DESCENT

$$\hat{y} = \theta_0 + \theta_1 x$$

$$\theta_0 = 0.1 \quad \theta_1 = 0.1$$

GRADIENT DESCENT

$$\hat{y} = \theta_0 + \theta_1 x$$

$$\theta_0 = 0.1 \quad \theta_1 = 0.1$$

x	y
0.2	0.44
0.31	0.123
0.45	0.75
0.26	0.39

GRADIENT DESCENT

$$\hat{y} = \theta_0 + \theta_1 x$$

$$\theta_0 = 0.1 \quad \theta_1 = 0.1$$

x	y	$\hat{y} = \theta_0 + \theta_1 x$
0.2	0.44	0.12
0.31	0.123	0.131
0.45	0.75	0.145
0.26	0.39	0.175

GRADIENT DESCENT

$$\hat{y} = \theta_0 + \theta_1 x$$

$$\theta_0 = 0.1 \quad \theta_1 = 0.1$$

x	y	$\hat{y} = \theta_0 + \theta_1 x$	SSE
			$\frac{1}{2}(\hat{y} - y)^2$
0.2	0.44	0.12	0.0512
0.31	0.123	0.131	0.000032
0.45	0.75	0.145	0.183
0.26	0.39	0.175	0.0231

GRADIENT DESCENT

$$\hat{y} = \theta_0 + \theta_1 x$$

$$\theta_0 = 0.1 \quad \theta_1 = 0.1$$

x	y	$\hat{y} = \theta_0 + \theta_1 x$	SSE	$\frac{\partial(SSE)}{\partial \theta_0}$
			$\frac{1}{2}(\hat{y} - y)^2$	$\hat{y} - y$
0.2	0.44	0.12	0.0512	-0.32
0.31	0.123	0.131	0.000032	0.008
0.45	0.75	0.145	0.183	-0.605
0.26	0.39	0.175	0.0231	-0.215

GRADIENT DESCENT

$$\hat{y} = \theta_0 + \theta_1 x$$

$$\theta_0 = 0.1 \quad \theta_1 = 0.1$$

x	y	$\hat{y} = \theta_0 + \theta_1 x$	SSE	$\frac{\partial(SSE)}{\partial \theta_0}$	$\frac{\partial(SSE)}{\partial \theta_1}$
			$\frac{1}{2}(\hat{y} - y)^2$	$\hat{y} - y$	$(\hat{y} - y)x$
0.2	0.44	0.12	0.0512	-0.32	-0.064
0.31	0.123	0.131	0.000032	0.008	0.00248
0.45	0.75	0.145	0.183	-0.605	-0.27225
0.26	0.39	0.175	0.0231	-0.215	-0.16125

GRADIENT DESCENT

$$\hat{y} = \theta_0 + \theta_1 x$$

$$\theta_0 = 0.1 \quad \theta_1 = 0.1$$

x	y	$\hat{y} = \theta_0 + \theta_1 x$	SSE	$\frac{\partial(SSE)}{\partial \theta_0}$	$\frac{\partial(SSE)}{\partial \theta_1}$
			$\frac{1}{2}(\hat{y} - y)^2$	$\hat{y} - y$	$(\hat{y} - y)x$
0.2	0.44	0.12	0.0512	-0.32	-0.064
0.31	0.123	0.131	0.000032	0.008	0.00248
0.45	0.75	0.145	0.183	-0.605	-0.27225
0.26	0.39	0.175	0.0231	-0.215	-0.16125
				-1.132	-0.495

GRADIENT DESCENT

$$\hat{y} = \theta_0 + \theta_1 x$$

$$\theta_0 = 0.1 \quad \theta_1 = 0.1$$

x	y	$\hat{y} = \theta_0 + \theta_1 x$	SSE $\frac{1}{2}(\hat{y} - y)^2$	$\frac{\partial(SSE)}{\partial\theta_0}$ $\hat{y} - y$	$\frac{\partial(SSE)}{\partial\theta_1}$ $(\hat{y} - y)x$
0.2	0.44	0.12	0.0512	-0.32	-0.064
0.31	0.123	0.131	0.000032	0.008	0.00248
0.45	0.75	0.145	0.183	-0.605	-0.27225
0.26	0.39	0.175	0.0231	-0.215	-0.16125
				-1.132	-0.495

$$\theta_0 := \theta_0 - \alpha \sum_{i=1}^n \frac{\partial(SSE)}{\partial\theta_0}$$

$$\theta_1 := \theta_1 - \alpha \sum_{i=1}^n \frac{\partial(SSE)}{\partial\theta_1}$$

GRADIENT DESCENT

$$\hat{y} = \theta_0 + \theta_1 x$$

$$\theta_0 = 0.1 \quad \theta_1 = 0.1$$

x	y	$\hat{y} = \theta_0 + \theta_1 x$	SSE $\frac{1}{2}(\hat{y} - y)^2$	$\frac{\partial(SSE)}{\partial\theta_0}$ $\hat{y} - y$	$\frac{\partial(SSE)}{\partial\theta_1}$ $(\hat{y} - y)x$
0.2	0.44	0.12	0.0512	-0.32	-0.064
0.31	0.123	0.131	0.000032	0.008	0.00248
0.45	0.75	0.145	0.183	-0.605	-0.27225
0.26	0.39	0.175	0.0231	-0.215	-0.16125
				-1.132	-0.495

$$\theta_0 := \theta_0 - \alpha \sum_{i=1}^n \frac{\partial(SSE)}{\partial\theta_0}$$

$$\theta_1 := \theta_1 - \alpha \sum_{i=1}^n \frac{\partial(SSE)}{\partial\theta_1}$$

$$\alpha = 0.01$$

GRADIENT DESCENT

$$\hat{y} = \theta_0 + \theta_1 x$$

$$\theta_0 = 0.1 \quad \theta_1 = 0.1$$

x	y	$\hat{y} = \theta_0 + \theta_1 x$	SSE $\frac{1}{2}(\hat{y} - y)^2$	$\frac{\partial(SSE)}{\partial\theta_0}$ $\hat{y} - y$	$\frac{\partial(SSE)}{\partial\theta_1}$ $(\hat{y} - y)x$
0.2	0.44	0.12	0.0512	-0.32	-0.064
0.31	0.123	0.131	0.000032	0.008	0.00248
0.45	0.75	0.145	0.183	-0.605	-0.27225
0.26	0.39	0.175	0.0231	-0.215	-0.16125
				-1.132	-0.495

$$\theta_0 := \theta_0 - \alpha \sum_{i=1}^n \frac{\partial(SSE)}{\partial\theta_0}$$

$$\theta_1 := \theta_1 - \alpha \sum_{i=1}^n \frac{\partial(SSE)}{\partial\theta_1}$$

$$\alpha = 0.01$$

$$\theta_0 = 0.1 - 0.01 \times (-1.132)$$

GRADIENT DESCENT

$$\hat{y} = \theta_0 + \theta_1 x$$

$$\theta_0 = 0.1 \quad \theta_1 = 0.1$$

x	y	$\hat{y} = \theta_0 + \theta_1 x$	SSE $\frac{1}{2}(\hat{y} - y)^2$	$\frac{\partial(SSE)}{\partial\theta_0}$ $\hat{y} - y$	$\frac{\partial(SSE)}{\partial\theta_1}$ $(\hat{y} - y)x$
0.2	0.44	0.12	0.0512	-0.32	-0.064
0.31	0.123	0.131	0.000032	0.008	0.00248
0.45	0.75	0.145	0.183	-0.605	-0.27225
0.26	0.39	0.175	0.0231	-0.215	-0.16125
				-1.132	-0.495

$$\theta_0 := \theta_0 - \alpha \sum_{i=1}^n \frac{\partial(SSE)}{\partial\theta_0}$$

$$\theta_1 := \theta_1 - \alpha \sum_{i=1}^n \frac{\partial(SSE)}{\partial\theta_1}$$

$$\alpha = 0.01$$

$$\theta_0 = 0.1 - 0.01 \times (-1.132)$$

$$\theta_0 = 0.11132$$

GRADIENT DESCENT

$$\hat{y} = \theta_0 + \theta_1 x$$

$$\theta_0 = 0.1 \quad \theta_1 = 0.1$$

x	y	$\hat{y} = \theta_0 + \theta_1 x$	SSE $\frac{1}{2}(\hat{y} - y)^2$	$\frac{\partial(SSE)}{\partial\theta_0}$ $\hat{y} - y$	$\frac{\partial(SSE)}{\partial\theta_1}$ $(\hat{y} - y)x$
0.2	0.44	0.12	0.0512	-0.32	-0.064
0.31	0.123	0.131	0.000032	0.008	0.00248
0.45	0.75	0.145	0.183	-0.605	-0.27225
0.26	0.39	0.175	0.0231	-0.215	-0.16125
				-1.132	-0.495

$$\theta_0 := \theta_0 - \alpha \sum_{i=1}^n \frac{\partial(SSE)}{\partial\theta_0}$$

$$\theta_0 = 0.1 - 0.01 \times (-1.132)$$

$$\theta_0 = 0.11132$$

$$\theta_1 := \theta_1 - \alpha \sum_{i=1}^n \frac{\partial(SSE)}{\partial\theta_1}$$

$$\theta_1 = 0.1 - 0.01 \times (-0.495)$$

$$\theta_1 = 0.10495$$

$$\alpha = 0.01$$

GRADIENT DESCENT

Algorithm for any* hypothesis function $h_\theta: \mathbb{R}^n \rightarrow \mathcal{Y}$, loss function $\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+$, step size α :

Initialize the parameter vector:

- $\theta \leftarrow 0$

Repeat until satisfied (e.g., exact or approximate convergence):

- Compute gradient: $g \leftarrow \sum_{i=1}^m \nabla_\theta \ell(h_\theta(x^{(i)}), y^{(i)})$
- Update parameters: $\theta \leftarrow \theta - \alpha \cdot g$


$$\theta_0 := \theta_0 - \alpha \sum_{i=1}^m \frac{\partial(SSE)}{\partial \theta_0}$$

$$\theta_1 := \theta_1 - \alpha \sum_{i=1}^m \frac{\partial(SSE)}{\partial \theta_1}$$

*must be reasonably well behaved

GRADIENT DESCENT - MULTIVARIATE

$$\theta \leftarrow \theta - \alpha \cdot g$$



$$\begin{aligned}\theta_0 &:= \theta_0 - \alpha \sum_{i=1}^n \frac{\partial(SSE)}{\partial \theta_0} \\ \theta_1 &:= \theta_1 - \alpha \sum_{i=1}^n \frac{\partial(SSE)}{\partial \theta_1} \\ \theta_2 &:= \theta_2 - \alpha \sum_{i=1}^n \frac{\partial(SSE)}{\partial \theta_2} \\ &\vdots \\ \theta_m &:= \theta_m - \alpha \sum_{i=1}^n \frac{\partial(SSE)}{\partial \theta_m}\end{aligned}$$

GRADIENT DESCENT - MULTIVARIATE

$$\theta = 0$$

Repeat{

obtain all partial derivatives w.r.t θ 's first

$$\theta_j := \theta_j - \alpha \frac{1}{n} \sum_{i=1}^n (h(\theta(x_i)) - y_i) x_{ji}$$

(update θ_j for all $j = 1 \dots m$ simultaneously

}

$$\frac{1}{n} \sum_{i=1}^n (h(\theta(x_i)) - y_i) x_{ji} = \frac{\partial}{\partial \theta_j} f(\theta)$$

$$\theta_0 := \theta_0 - \alpha \frac{1}{n} \sum_{i=1}^n (h(\theta(x_i)) - y_i) x_{0i}$$

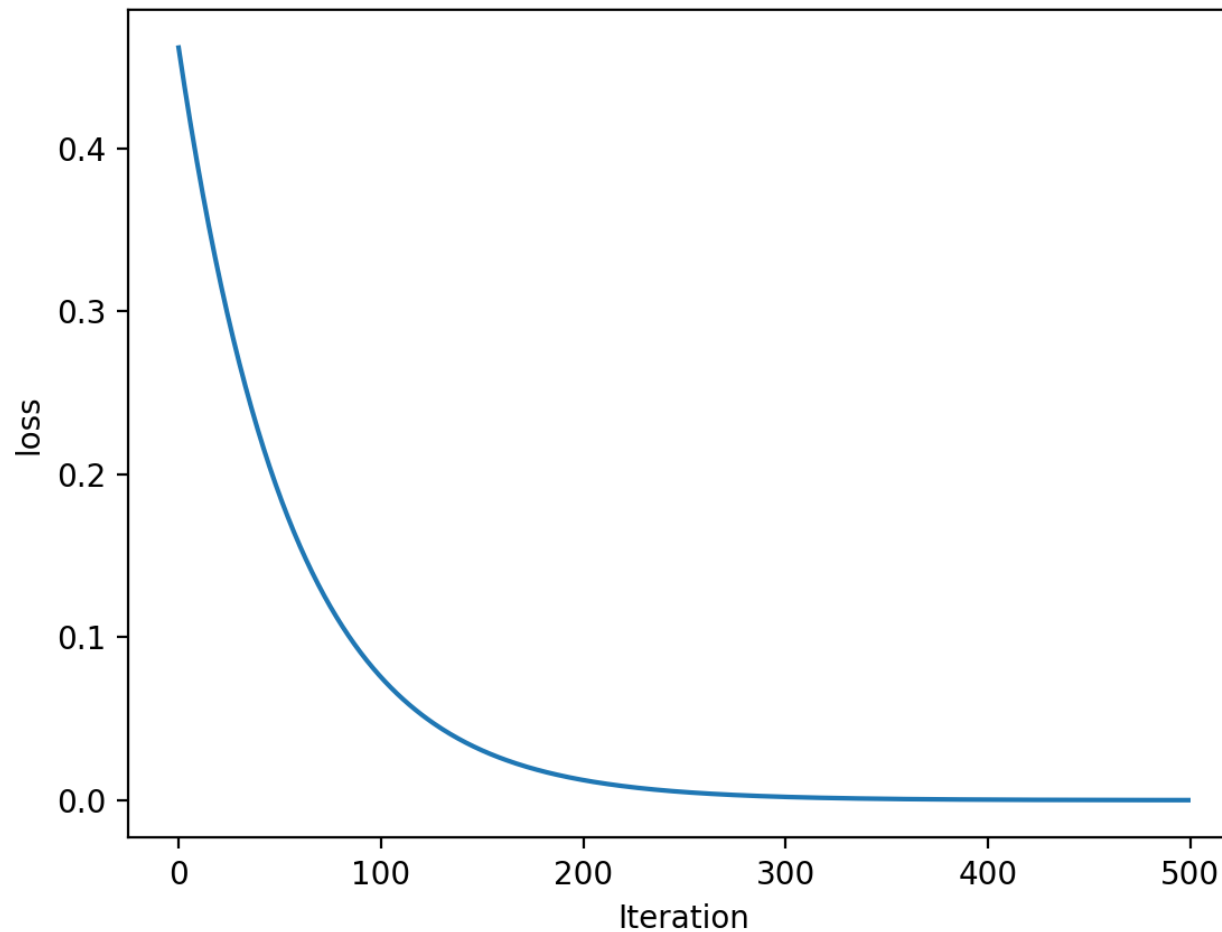
$$\theta_1 := \theta_1 - \alpha \frac{1}{n} \sum_{i=1}^n (h(\theta(x_i)) - y_i) x_{1i}$$

$$\theta_2 := \theta_2 - \alpha \frac{1}{n} \sum_{i=1}^n (h(\theta(x_i)) - y_i) x_{2i}$$

•
•
•

$$\theta_m := \theta_m - \alpha \frac{1}{n} \sum_{i=1}^n (h(\theta(x_i)) - y_i) x_{mi}$$

PLOTTING LOSS OVER TIME



STOCHASTIC GRADIENT DESCENT

- MULTIVARIATE

$$\theta = 0$$

Repeat{

obtain all partial derivatives w.r.t θ 's first

$$\theta_j := \theta_j - \alpha \frac{1}{n} \sum_{i=1}^n (h(\theta(x_i)) - y_i) x_{ji}$$

(update θ_j for all $j = 1 \dots m$ simultaneously

}

$$\frac{1}{n} \sum_{i=1}^n (h(\theta(x_i)) - y_i) x_{ji} = \frac{\partial}{\partial \theta_j} f(\theta)$$


STOCHASTIC GRADIENT DESCENT

$$\theta = 0$$

Repeat{

i = random index between 1 and m

$$\theta_j := \theta_j - \alpha (h(\theta(x_i)) - y_i) x_{ji}$$

(update θ_j for all $j = 1 \dots n$

}

GRADIENT DESCENT

$$\hat{y} = \theta_0 + \theta_1 x$$

$$\theta_0 = 0.1 \quad \theta_1 = 0.1$$

x	y	$\hat{y} = \theta_0 + \theta_1 x$	SSE	$\frac{\partial(SSE)}{\partial \theta_0}$	$\frac{\partial(SSE)}{\partial \theta_1}$
			$\frac{1}{2}(\hat{y} - y)^2$	$\hat{y} - y$	$(\hat{y} - y)x$
0.2	0.44	0.12	0.0512	-0.32	-0.064
0.31	0.123	0.131	0.000032	0.008	0.00248
0.45	0.75	0.145	0.183	-0.605	-0.27225
0.26	0.39	0.175	0.0231	-0.215	-0.16125
				-1.132	-0.495

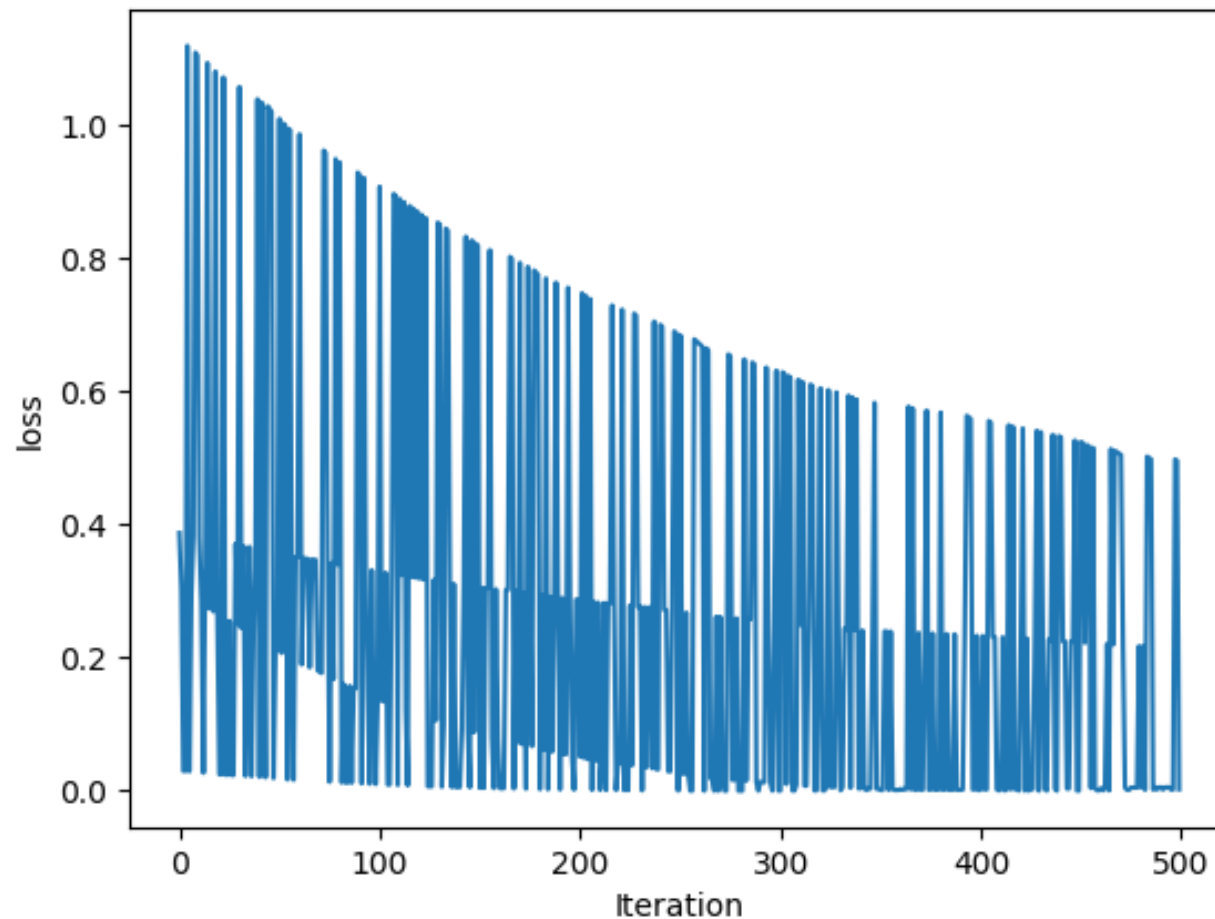
GRADIENT DESCENT

$$\hat{y} = \theta_0 + \theta_1 x$$

$$\theta_0 = 0.1 \quad \theta_1 = 0.1$$

x	y	$\hat{y} = \theta_0 + \theta_1 x$	SSE $\frac{1}{2}(\hat{y} - y)^2$	$\frac{\partial(SSE)}{\partial\theta_0}$ $\hat{y} - y$	$\frac{\partial(SSE)}{\partial\theta_1}$ $(\hat{y} - y)x$	
0.31	0.123	0.131	0.000032	0.008	0.00248	

STOCHASTIC GRADIENT DESCENT



STOCHASTIC GRADIENT DESCENT

- MINI BATCH

$$\theta = 0$$

Repeat {

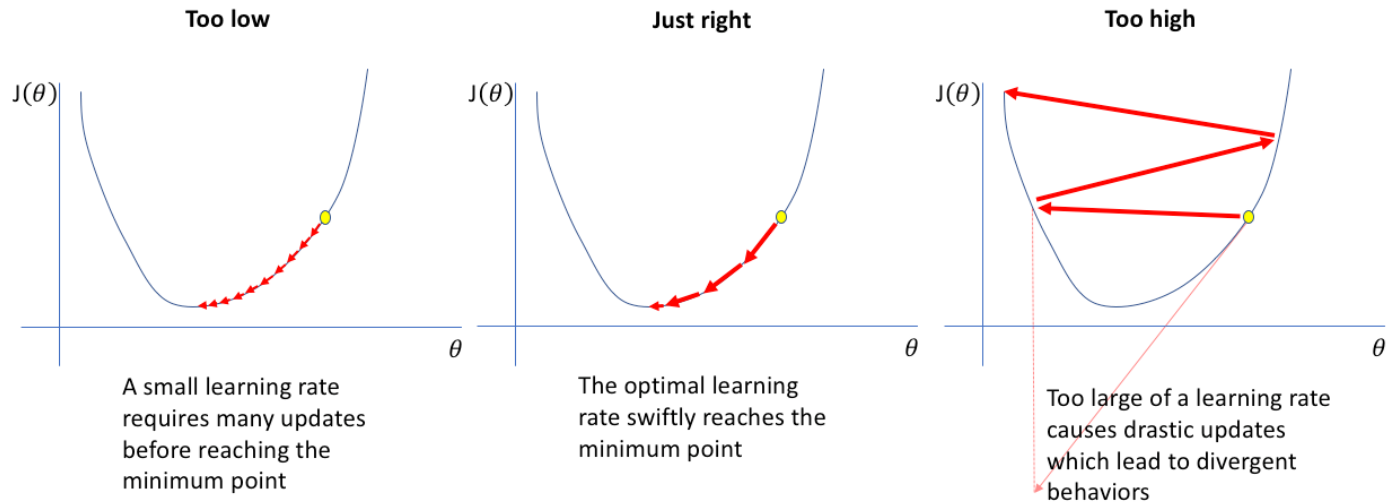
$i_1, \dots, i_l$ = random index between 1 and m

$$\theta_j := \theta_j - \alpha \frac{1}{l} \sum_{i=1}^l (h(\theta(x_i)) - y_i) x_{ji}$$

(update θ_j for all $j = 1 \dots n$

}

Gradient Descent



GRADIENT DESCENT IN PURE(-ISH) PYTHON

```
# Training data (X, y), T time steps, alpha step
def grad_descent(X, y, T, alpha):
    m, n = X.shape          # m = #examples, n = #features
    theta = np.zeros(n)     # initialize parameters
    f = np.zeros(T)         # track loss over time
```

```
    for i in range(T):
        # loss for current parameter vector theta
        f[i] = 0.5*np.linalg.norm(X.dot(theta) - y)**2
        # compute steepest ascent at f(theta)
        g = np.transpose(X).dot(X.dot(theta) - y)
        # step down the gradient
        theta = theta - alpha*g
    return theta, f
```

Implicitly using squared loss and linear hypothesis function above; drop in your favorite gradient for kicks!