Linear Regression

Aleksanyan Lida

Matrix

A **matrix** is a rectangular arrangement of numbers into rows and columns.

For example, matrix A has two **rows** and three **columns**.

$$A = \begin{array}{c} \text{3 columns} \\ \downarrow \quad \downarrow \quad \downarrow \\ \begin{bmatrix} -2 & 5 & 6 \\ 5 & 2 & 7 \end{bmatrix} \end{array} \begin{array}{l} \\ \\ \leftarrow \\ \leftarrow \end{array} \text{2 rows}$$

Matrix

Matrix dimensions

The **dimensions** of a matrix tells its size: the number of rows and columns of the matrix, *in that order*.

Since matrix A has two rows and three columns, we write its dimensions as 2×3, pronounced "two by three".

In contrast, matrix B has three rows and two columns, so it is a 3×2 matrix.

$$B = \begin{bmatrix} -8 & -4 \\ 23 & 12 \\ 18 & 10 \end{bmatrix}$$

1) What are the dimensions of matrix D?

$$D = \begin{bmatrix} -7 & 24 & 2 \\ 1 & 15 & 11 \\ -9 & 12 & 0 \\ 8 & -3 & -1 \end{bmatrix}$$

Matrix

Matrix elements

A **matrix element** is simply a matrix entry. Each element in a matrix is identified by naming the row and column in which it appears.

For example, consider matrix G:

$$G = \begin{bmatrix} 4 & 14 & -7 \\ 18 & 5 & 13 \\ -20 & 4 & 22 \end{bmatrix}$$

The element $g_{2,1}$ is the entry in the second row and the first column.

$$G = \begin{bmatrix} 4 & 14 & -7 \\ 18 & 5 & 13 \\ -20 & 4 & 22 \end{bmatrix}$$

In this case $g_{2,1} = 18$.

In general, the element in row i and column j of matrix A is denoted as $a_{i,j}$.

$$A = \begin{bmatrix} 2 & -4 & 8 \\ 1 & 5 & -5 \\ -2 & 6 & 2 \end{bmatrix}$$

$a_{1,3} =$ ______

Matrix C is a 2×3 matrix with $c_{1,2} = 6$.

Which could be matrix C?

Choose 1 answer:

(A) $\begin{bmatrix} 1 & 2 \\ 6 & 4 \\ 5 & -2 \end{bmatrix}$

(B) $\begin{bmatrix} -9 & 6 \\ 7 & -3 \\ -3 & 5 \end{bmatrix}$

(C) $\begin{bmatrix} 2 & 6 & 8 \\ 7 & -3 & 1 \end{bmatrix}$

(D) $\begin{bmatrix} 2 & 10 & 8 \\ 6 & -3 & 1 \end{bmatrix}$

Matrix addition

Adding matrices

Given $A = \begin{bmatrix} 4 & 8 \\ 3 & 7 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 0 \\ 5 & 2 \end{bmatrix}$, let's find $A + B$.

We can find the sum simply by adding the corresponding entries in matrices A and B. This is shown below.

$$A + B = \begin{bmatrix} 4 & 8 \\ 3 & 7 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 5 & 2 \end{bmatrix} = \begin{bmatrix} 4+1 & 8+0 \\ 3+5 & 7+2 \end{bmatrix} = \begin{bmatrix} 5 & 8 \\ 8 & 9 \end{bmatrix}$$

Matrix subtraction

Subtracting matrices

Similarly, to subtract matrices, we subtract the corresponding entries.

For example, let's consider $C = \begin{bmatrix} 2 & 8 \\ 0 & 9 \end{bmatrix}$ and $D = \begin{bmatrix} 5 & 6 \\ 11 & 3 \end{bmatrix}$.

We can find $C - D$ by subtracting the corresponding entries in matrices C and D. This is shown below.

$$C - D = \begin{bmatrix} 2 & 8 \\ 0 & 9 \end{bmatrix} - \begin{bmatrix} 5 & 6 \\ 11 & 3 \end{bmatrix} = \begin{bmatrix} 2-5 & 8-6 \\ 0-11 & 9-3 \end{bmatrix} = \begin{bmatrix} -3 & 2 \\ -11 & 6 \end{bmatrix}$$

Scalars and scalar multiplication

When we work with matrices, we refer to real numbers as **scalars**.

The term **scalar multiplication** refers to the product of a real number and a matrix. In scalar multiplication, each entry in the matrix is multiplied by the given scalar.

For example, given that $A = \begin{bmatrix} 10 & 6 \\ 4 & 3 \end{bmatrix}$, let's find $2A$.

To find $2A$, simply multiply each matrix entry by 2:

$$2A = 2 \cdot \begin{bmatrix} 10 & 6 \\ 4 & 3 \end{bmatrix} = \begin{bmatrix} 2 \cdot 10 & 2 \cdot 6 \\ 2 \cdot 4 & 2 \cdot 3 \end{bmatrix} = \begin{bmatrix} 20 & 12 \\ 8 & 6 \end{bmatrix}$$

Matrix multiplication

We are now ready to look at an example of matrix multiplication.

Given $A = \begin{bmatrix} 1 & 7 \\ 2 & 4 \end{bmatrix}$ and $B = \begin{bmatrix} 3 & 3 \\ 5 & 2 \end{bmatrix}$, let's find matrix $C = AB$.

$$\begin{array}{c} \vec{a_1} \rightarrow \\ \vec{a_2} \rightarrow \end{array} \underbrace{\begin{bmatrix} 1 & 7 \\ 2 & 4 \end{bmatrix}}_{A} \cdot \underbrace{\begin{array}{cc} \vec{b_1} & \vec{b_2} \\ \downarrow & \downarrow \\ \end{array}\begin{bmatrix} 3 & 3 \\ 5 & 2 \end{bmatrix}}_{B} = \underbrace{\begin{bmatrix} \vec{a_1} \cdot \vec{b_1} & \vec{a_1} \cdot \vec{b_2} \\ \vec{a_2} \cdot \vec{b_1} & \vec{a_2} \cdot \vec{b_2} \end{bmatrix}}_{C} = \begin{bmatrix} 38 & 17 \\ 26 & 14 \end{bmatrix}$$

$$\underset{M}{\begin{bmatrix} M_{00} & M_{01} & M_{02} & M_{03} \\ M_{10} & M_{11} & M_{12} & M_{13} \\ M_{20} & M_{21} & M_{22} & M_{23} \\ M_{30} & M_{31} & M_{32} & M_{33} \end{bmatrix}} \times \underset{X}{\begin{bmatrix} X_0 \\ X_1 \\ X_2 \\ X_3 \end{bmatrix}} = \underset{C}{\begin{bmatrix} C_0 \\ C_1 \\ C_2 \\ C_3 \end{bmatrix}}$$

Matrix transpose

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}^T = \begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \\ a_{13} & a_{23} & a_{33} \end{bmatrix}$$

Transpose of a Matrix

$$A = \begin{bmatrix} 1 & 3 & 4 \\ 2 & 3 & 2 \end{bmatrix}$$

$$A^T = ?$$

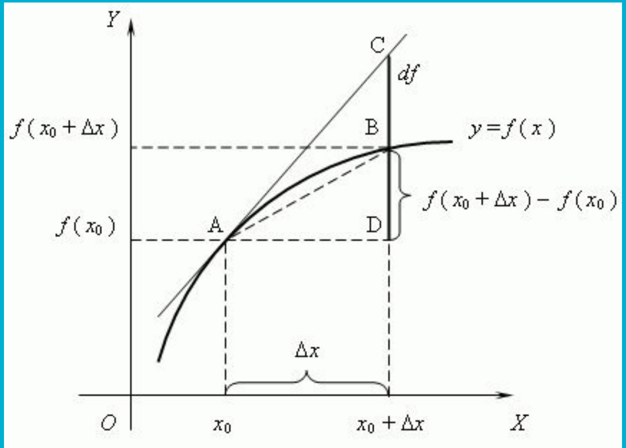
Functions derivatives

Lagrange's notation: f'

Leibniz's notation: $\frac{dy}{dx}$

Newton's notation: $\dot{y}$

In Leibniz's notation, the derivative of f is expressed as $\frac{d}{dx}f(x)$. When we have an equation $y = f(x)$ we can express the derivative as $\frac{dy}{dx}$.



$$f(x)=k\in\mathbb{R}\Rightarrow f'(x)=0$$

$$f(x)=x\Rightarrow f'(x)=1$$

$$f(x)=x^k\Rightarrow f'(x)=kx^{k-1}$$

$$f(x)=\frac{1}{x}\Rightarrow f'(x)=-\frac{1}{x^2}$$

$$f(x)=\sqrt{x}\Rightarrow f'(x)=\frac{1}{2\sqrt{x}}$$

$$f(x)=\ln x\Rightarrow f'(x)=\frac{1}{x}$$

$$f(x)=\log_a x\Rightarrow f'(x)=\frac{1}{x\ln a}$$

$$f(x)=e^x\Rightarrow f'(x)=e^x$$

$$f(x)=a^x\Rightarrow f'(x)=a^x\ln a$$

$$f(x)=\sin x\Rightarrow f'(x)=\cos x$$

$$f(x)=\cos x\Rightarrow f'(x)=-\sin x$$

$$f(x)=\tan x\Rightarrow f'(x)=\sec^2 x=1+\tan^2 x$$

$$f(x)=\arcsin x\Rightarrow f'(x)=\frac{1}{\sqrt{1-x^2}}$$

$$f(x)=\arctan x\Rightarrow f'(x)=\frac{1}{1+x^2}$$

Functions derivatives

$$(f \cdot g)' = f' \cdot g + f \cdot g'$$
$$(f \cdot g \cdot h)' = f' \cdot g \cdot h + f \cdot g' \cdot h + f \cdot g \cdot h'$$

$$f(x) = \frac{g(x)}{h(x)}$$

$$f'(x) = \frac{g'(x)h(x) - g(x)h'(x)}{[h(x)]^2}$$

Chain Rule

If f and g are both differentiable and $F(x)$ is the composite function defined by $F(x) = f(g(x))$ then F is differentiable and F' is given by the product

$$F'(x) = f'(g(x))\, g'(x)$$

Differentiate outer function

Differentiate inner function

- For a multivariable function, like $f(x, y) = x^2 y$, computing partial derivatives looks something like this:

$$\frac{\partial f}{\partial x} = \underbrace{\frac{\partial}{\partial x} x^2 y}_{\substack{\text{Treat } y \text{ as constant;} \\ \text{take derivative.}}} = 2xy$$

$$\frac{\partial f}{\partial y} = \underbrace{\frac{\partial}{\partial y} x^2 y}_{\substack{\text{Treat } x \text{ as constant;} \\ \text{take derivative.}}} = x^2 \cdot 1$$

These are called **second partial derivatives**, and the notation is analogous to the $\frac{d^2 f}{dx^2}$ notation for the ordinary second derivative in single-variable calculus:

$$\frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right) = \frac{\partial^2 f}{\partial x^2}$$

$$\frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right) = \frac{\partial^2 f}{\partial x \partial y}$$

$$\frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right) = \frac{\partial^2 f}{\partial y \partial x}$$

$$\frac{\partial}{\partial y}\left(\frac{\partial f}{\partial y}\right) = \frac{\partial^2 f}{\partial y^2}$$

Using the f_x notation for the partial derivative (in this case with respect to x), you might also see these second partial derivatives written like this:

$$(f_x)_x = f_{xx}$$

$$(f_y)_x = f_{yx}$$

$$(f_x)_y = f_{xy}$$

$$(f_y)_y = f_{yy}$$

Higher order derivatives

$$\frac{\partial}{\partial x}\frac{\partial}{\partial y}\frac{\partial}{\partial z}\frac{\partial}{\partial y}\frac{\partial f}{\partial z} = \frac{\partial^5 f}{\underbrace{\partial x}_{5^{th}}\,\underbrace{\partial y}_{4^{th}}\,\underbrace{\partial z}_{3^{rd}}\,\underbrace{\partial y}_{2^{nd}}\,\underbrace{\partial z}_{1^{st}}}$$

Gradient

- The gradient of a scalar-valued multivariable function $f(x, y, \dots)$, denoted ∇f, packages all its partial derivative information into a vector:

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \\ \vdots \end{bmatrix}$$

Example 1: Two dimensions

If $f(x, y) = x^2 - xy$, which of the following represents ∇f?

Choose 1 answer:

(A) $\begin{bmatrix} 2x - x \\ x^2 - y \end{bmatrix}$

(B) $\begin{bmatrix} 2x - y \\ -x \end{bmatrix}$

Example 2: Three dimensions

What is the gradient of $f(x, y, z) = x - xy + z^2$?

Choose 1 answer:

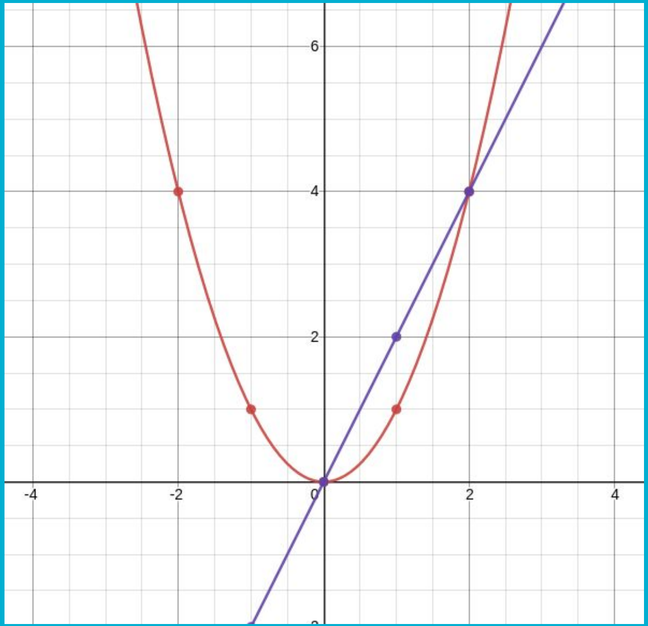
(A) $\nabla f(x, y, z) = \begin{bmatrix} 1 - y \\ -x \\ 2z \end{bmatrix}$

(B) $\nabla f(x, y, z) = \begin{bmatrix} 1 - y + z^2 \\ x - x + z^2 \\ x - xy + 2z \end{bmatrix}$

Gradient

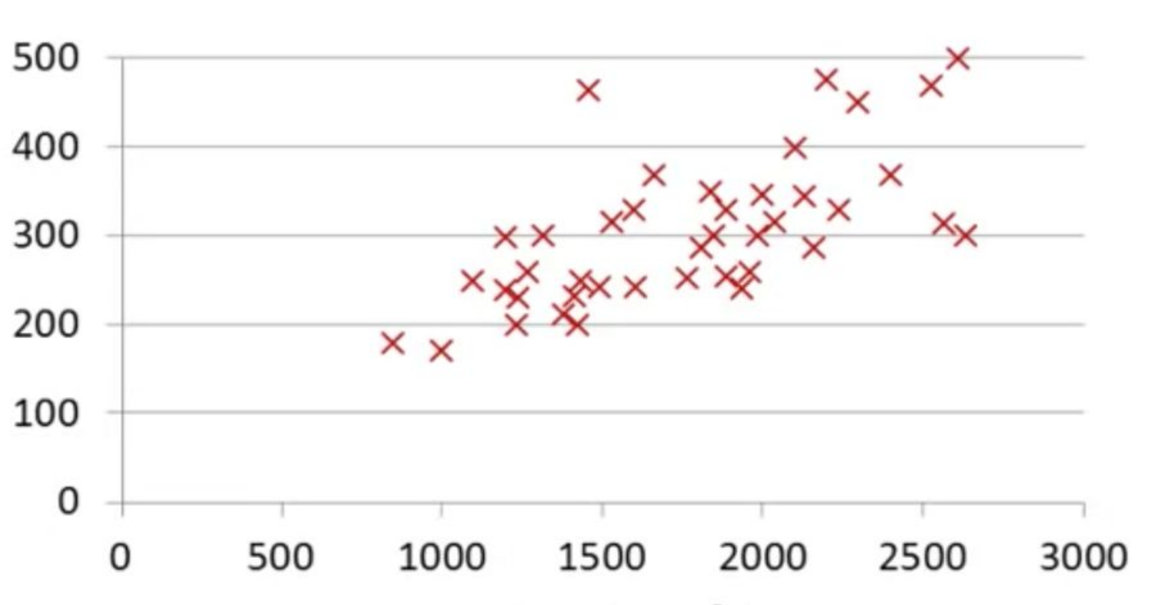
The most important thing to remember about the gradient: The gradient of f, if evaluated at an input (x_0, y_0), points in the direction of steepest ascent.

So, if you walk in the direction of the gradient, you will be going straight up the hill. Similarly, *the magnitude of the vector* $\nabla f(x_0, y_0)$ *tells you what the slope of the hill is* in that direction.



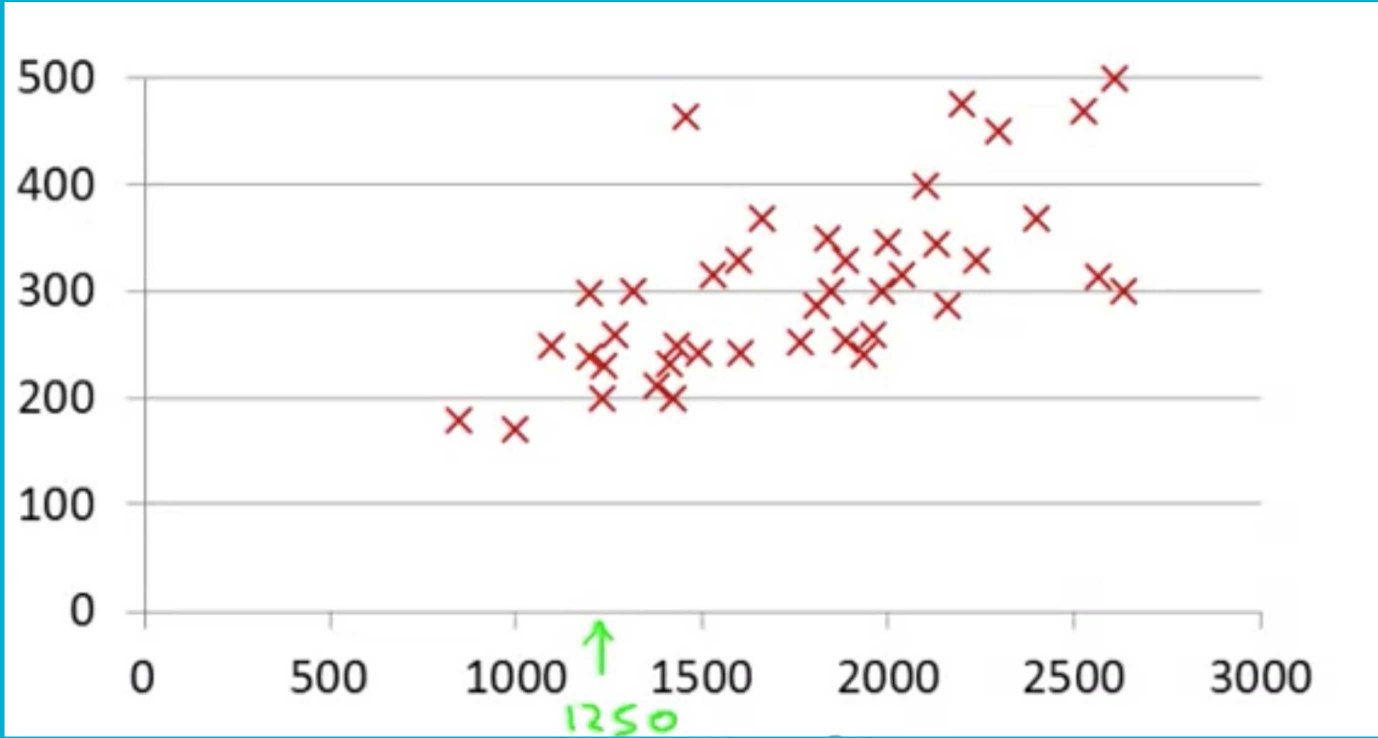
Name	Symbol	Example
Derivative	$\frac{d}{dx}$	$\frac{d}{dx}(x^2) = 2x$
Partial derivative	$\frac{\partial}{\partial x}$	$\frac{\partial}{\partial x}(x^2 - xy) = 2x - y$
Gradient	∇	$\nabla(x^2 - xy) = \begin{bmatrix} 2x - y \\ -x \end{bmatrix}$

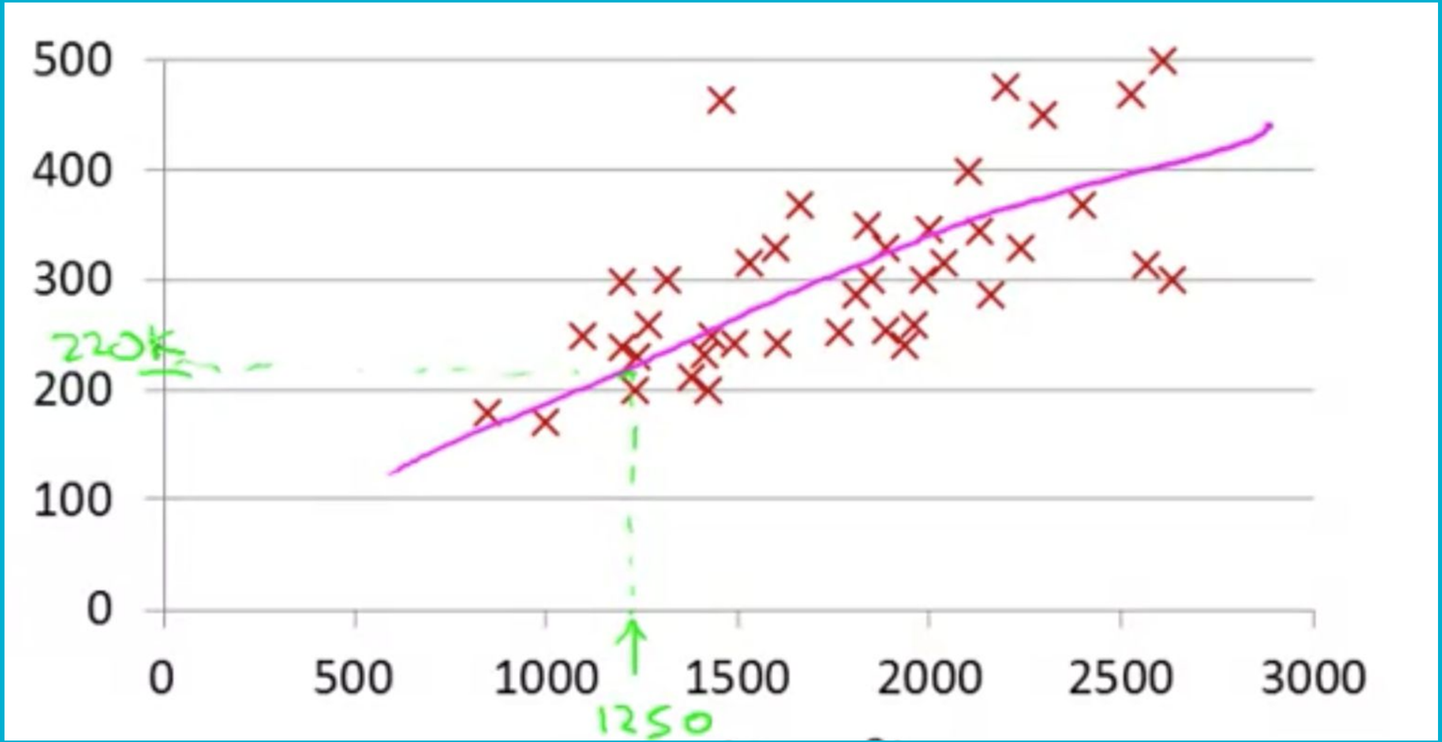
Model Representation



- **Supervised learning**: as given “right” answer for each example in the data

- **Regression problem**: as we’re going to predict real-valued output(not discret)


500
400
300
200
100
0
0
500
1000
1500
2000
2500
3000
1250



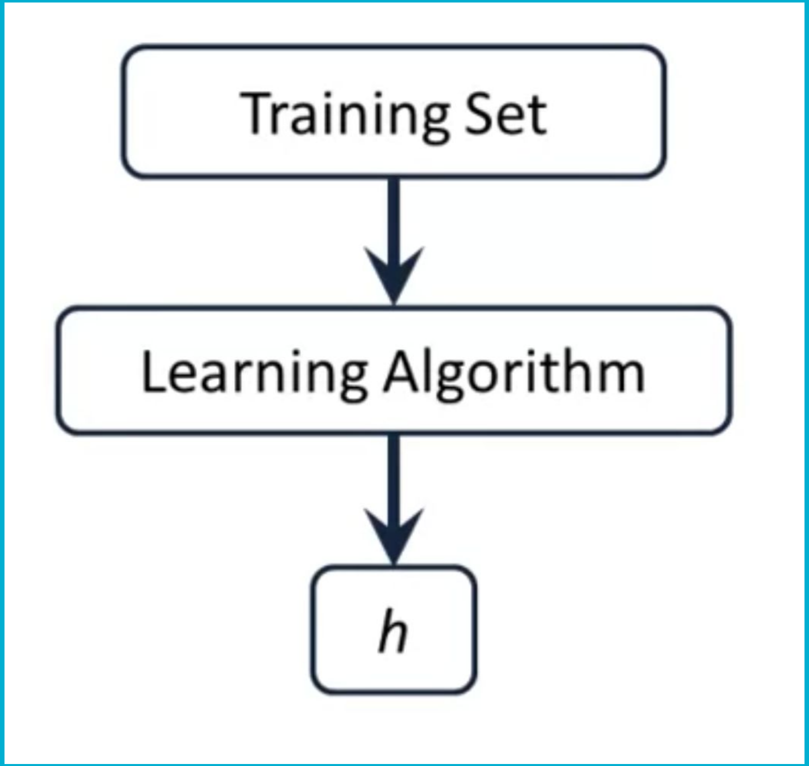
500
400
300
200
100
0
0
500
1000
1500
2000
2500
3000
220K
1250

Training set of housing prices (Portland, OR)

Size in feet^2 (x)	Price ($) in 1000's (y)
2104	460
1416	232
1534	315
852	178
...	...

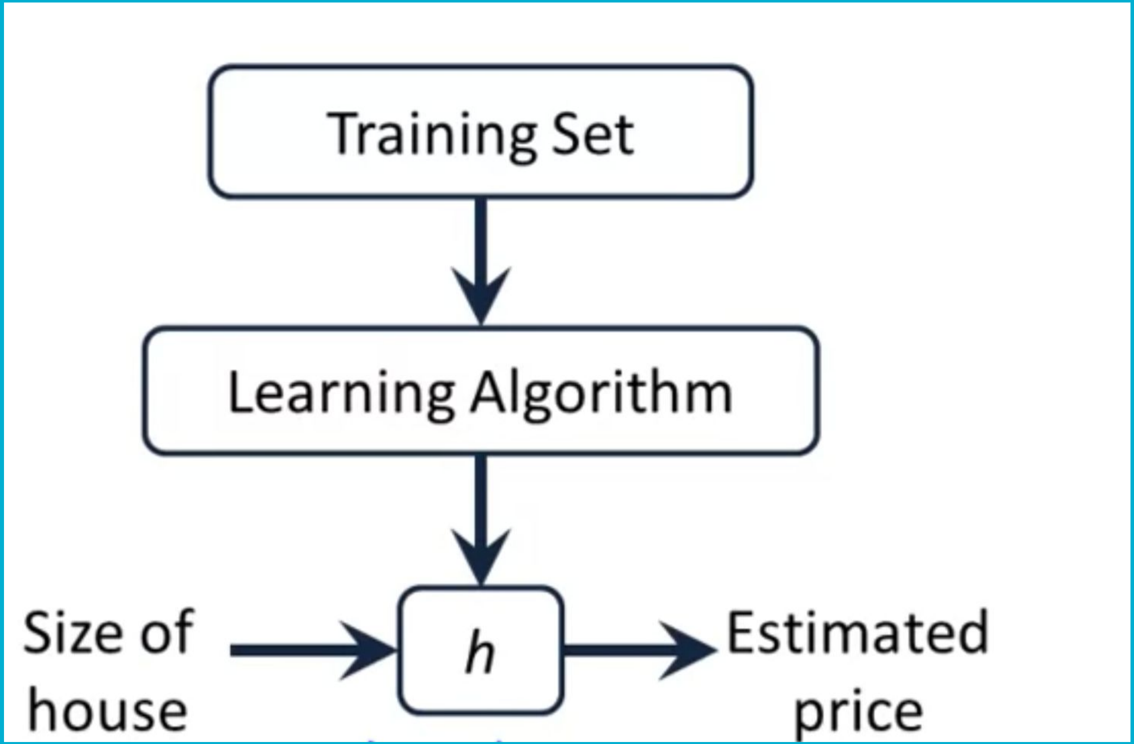
Notation:

m = Number of training examples
x's = "input" variable / features
y's = "output" variable / "target" variable



h is the hypothesis

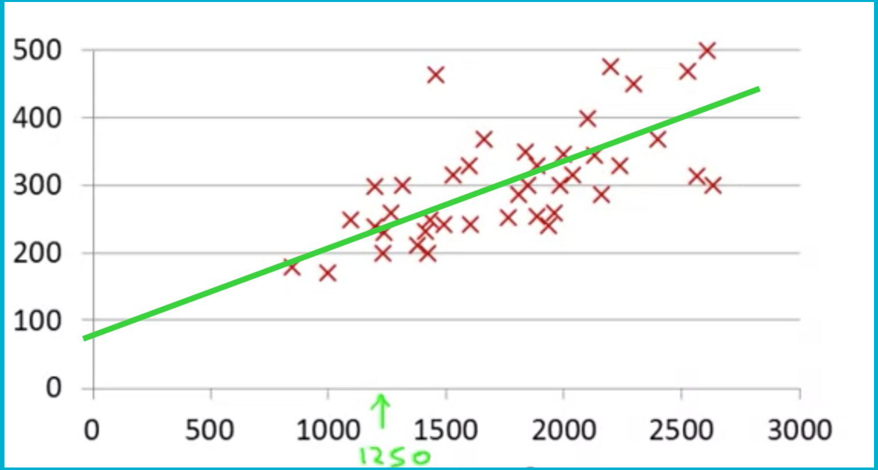
What are the inputs and outputs of h in our example?



Training Set
Learning Algorithm
Size of house
h
Estimated price

How to represent h?

$$h_{\theta}(X) = \theta_0 + \theta_1(X)$$

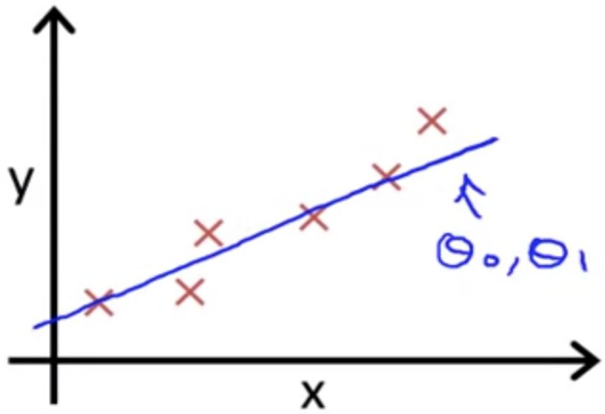


Linear regression with one variable is called **univariate** linear regression

$$h_{\theta}(X) = \theta_0 + \theta_1(X)$$

θ_0, θ_1 are the parameters.

The problem is - **How to find the parameters?**



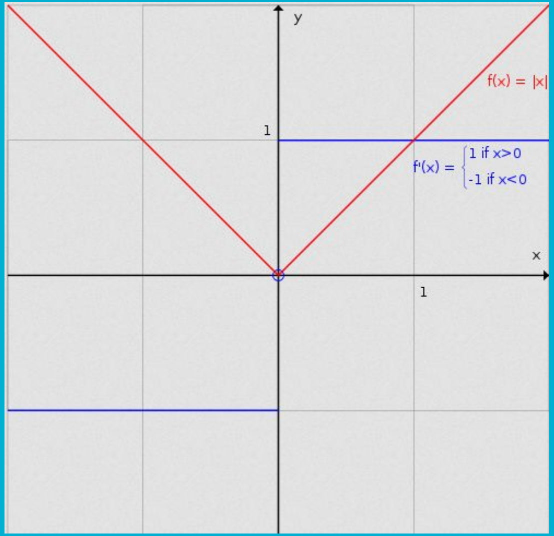
Idea: Choose θ_0, θ_1 so that
$h_\theta(x)$ is close to y for our
training examples (x, y)

Cost function

1. $$\sum_{i=1}^{n} Y_i - b_0 - b_1 X_i$$

2. $$\sum_{i=1}^{n} \left| Y_i - b_0 - b_1 X_i \right|$$

3. $$\sum_{i=1}^{n} \left(Y_i - b_0 - b_1 X_i \right)^2$$



So what we have?

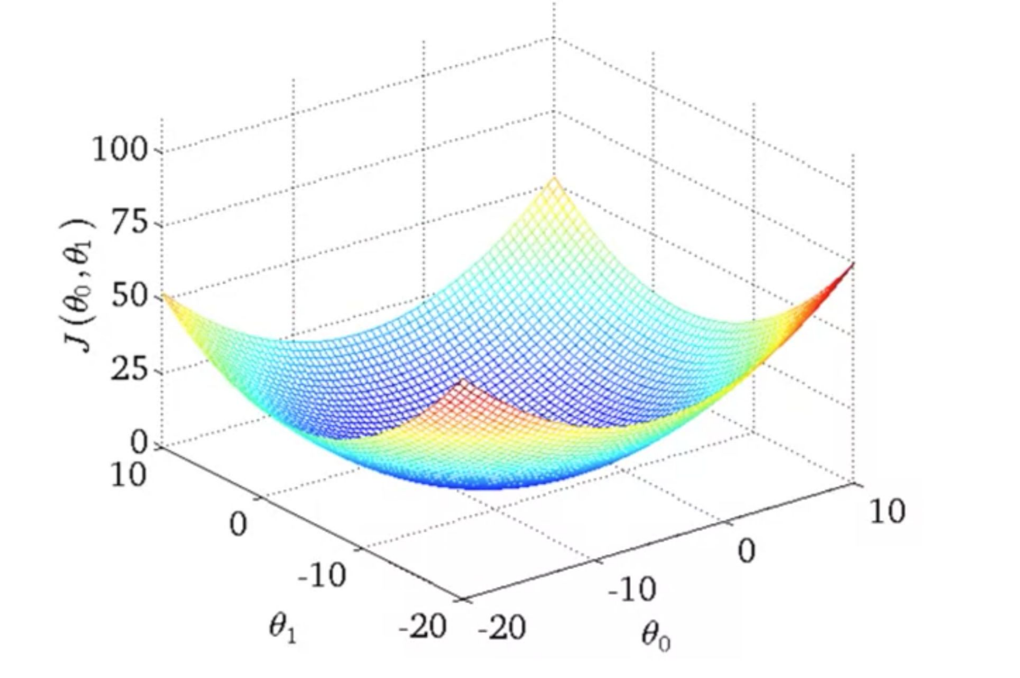
Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x$

Parameters: θ_0, θ_1

Cost Function: $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$

Goal: $\underset{\theta_0, \theta_1}{\text{minimize}}\ J(\theta_0, \theta_1)$

Minimization with two parameters



Have some function $J(\theta_0, \theta_1)$

Want $\min\limits_{\theta_0, \theta_1} J(\theta_0, \theta_1)$

Outline:

- Start with some θ_0, θ_1
- Keep changing θ_0, θ_1 to reduce $J(\theta_0, \theta_1)$

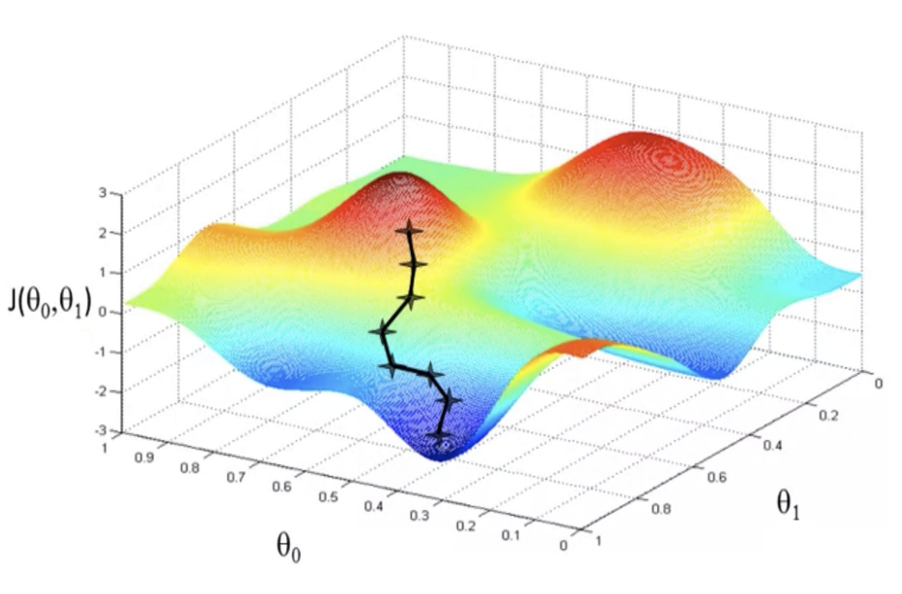
  until we hopefully end up at a minimum

Gradient Descent Algorithm

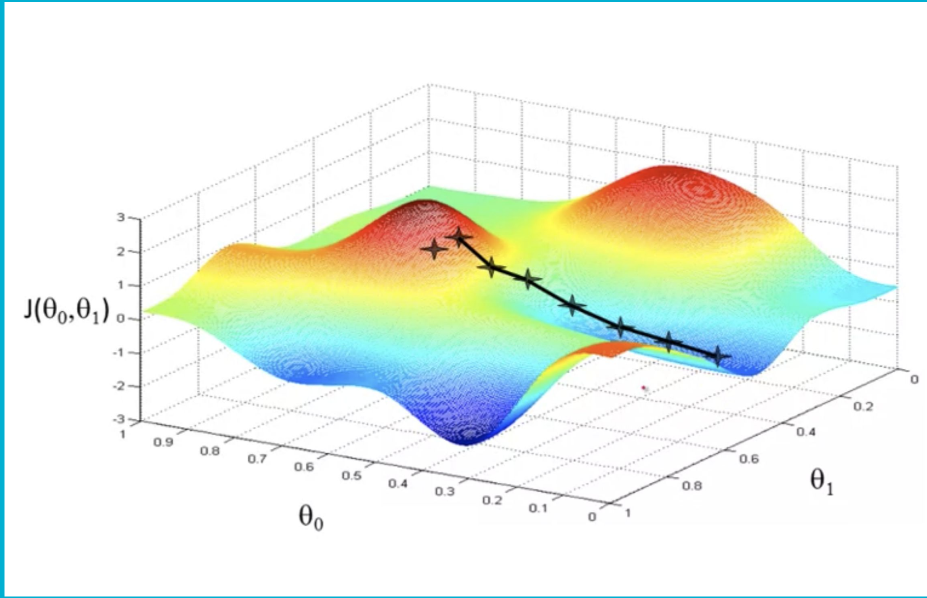
repeat until convergence {

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \qquad (\text{for } j = 0 \text{ and } j = 1)$$

}



J(θ0,θ1)
3
2
1
0
-1
-2
-3
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
θ0
0
0.2
0.4
0.6
0.8
1
θ1



J(θ0,θ1)
3
2
1
0
-1
-2
-3
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
θ0
0
0.2
0.4
0.6
0.8
1
θ1

Correct: Simultaneous update

$$\text{temp0} := \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$$
$$\text{temp1} := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1)$$
$$\theta_0 := \text{temp0}$$
$$\theta_1 := \text{temp1}$$

$$\frac{\delta}{\delta\theta_j}(J(\theta_0, \theta_1)) = \frac{\delta}{\delta\theta}\left(\frac{1}{2m}\sum_{i=0}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2\right) =$$

$$= \frac{\delta}{\delta\theta_j}\left(\frac{1}{2m}\sum_{i=0}^{m}(\theta_0 + \theta_1 x^{(i)} - y^{(i)})^2\right)$$

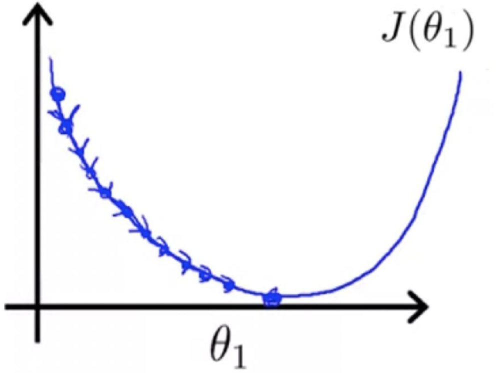
$$\begin{aligned}
&\text{repeat until convergence } \{ \\
&\quad \theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right) \\
&\quad \theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right) \cdot x^{(i)} \\
&\}
\end{aligned}$$

Here, **α** is called the ***learning rate***. This is a very natural algorithm that repeatedly takes a step in the direction of steepest decrease of **J**.

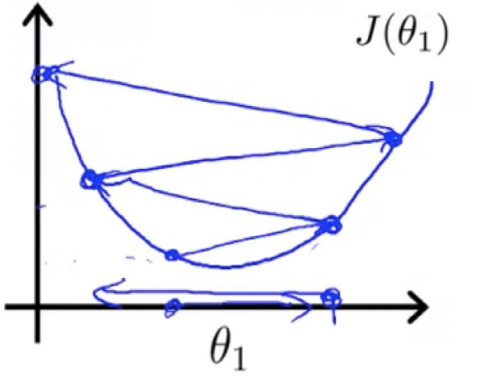
Learning rate

$$\theta_1 := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_1)$$

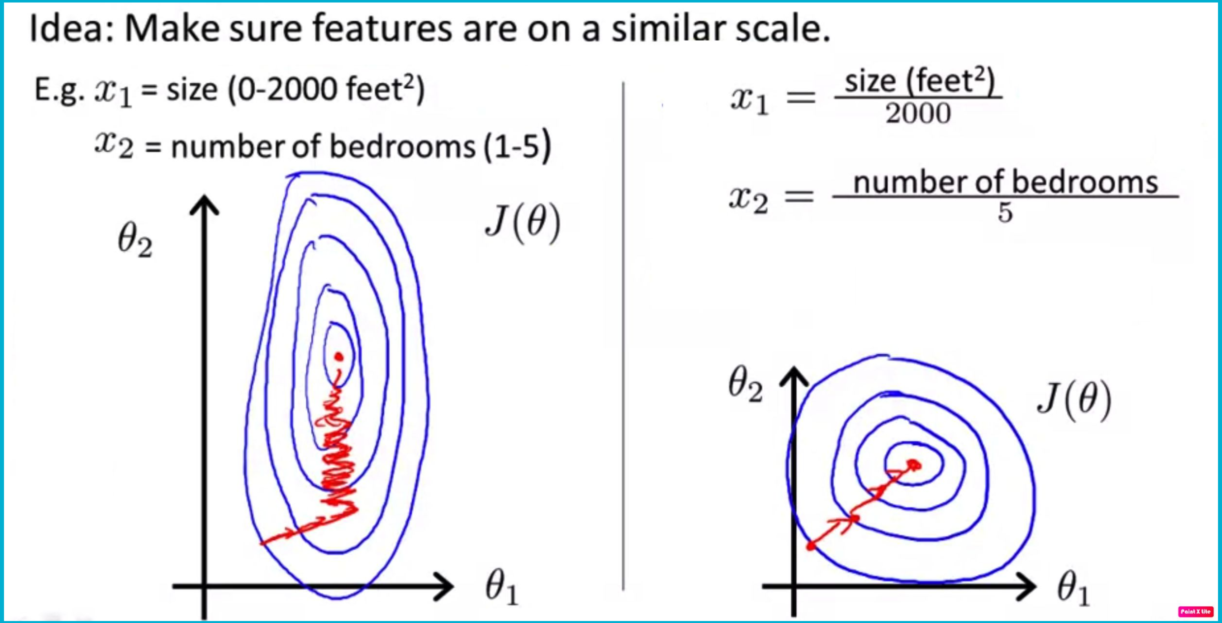
If α is too small, gradient descent can be slow.



If α is too large, gradient descent can overshoot the minimum. It may fail to converge, or even diverge.



Feature Scaling



Mean Normalization

Get every feature into approximately a $-1 \leq x_i \leq 1$ range.

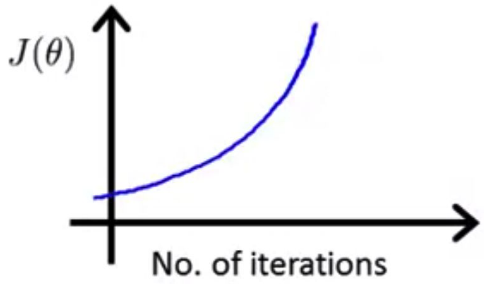
Replace x_i with $x_i - \mu_i$ to make features have approximately zero mean (Do not apply to $x_0 = 1$).

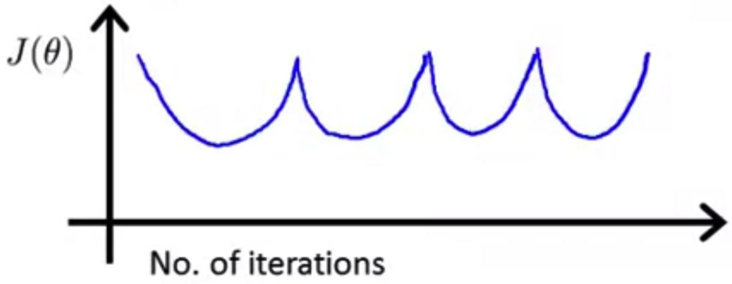
E.g. $$x_1 = \frac{size - 1000}{2000}$$

$$x_2 = \frac{\#bedrooms - 2}{5}$$

$$-0.5 \leq x_1 \leq 0.5, -0.5 \leq x_2 \leq 0.5$$

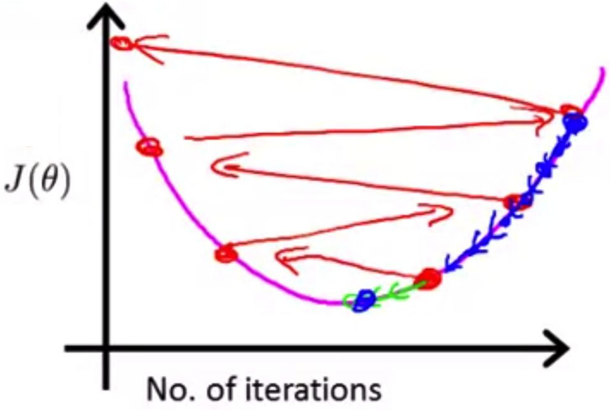
Learning rate selection





Gradient descent not working.

Use smaller α.



- For sufficiently small α, $J(\theta)$ should decrease on every iteration.
- But if α is too small, gradient descent can be slow to converge.

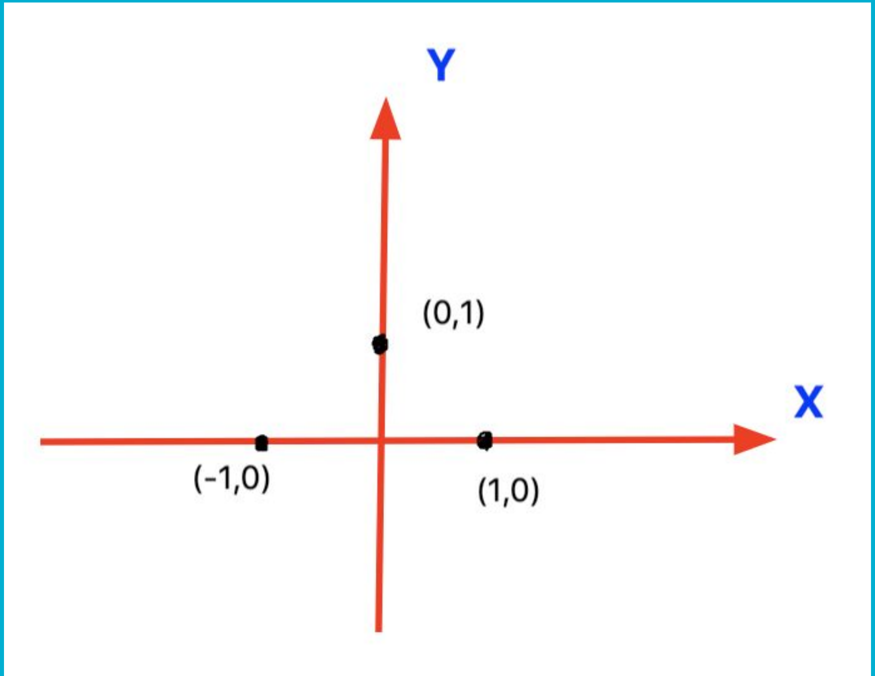
Summary

- If α is too small: slow convergence.
- If α is too large: $J(\theta)$ may not decrease on every iteration; may not converge.

To choose α, try

$$\dots, 0.001, \qquad , 0.01, \qquad , 0.1, \qquad , 1, \dots$$

→ Find hypothesis function for this training set:



Multiple features

Size ($feet^2$)	Number of bedrooms	Number of floors	Age of home (years)	Price ($1000)
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178
...	...	...	...	...

Notations

$x^{(i)}$ $-$ i^{th} *example*

$x_j^{(i)}$ $-$ *value of feature j in* i^{th} *example*

Hypothesis

- **Then :** $h_\theta(x) = \theta_0 + \theta_1 x$
- **Now :** $h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots$

For convenience let's define x0 = 1

- **Finally :** $h_\theta(X) = \theta^T X$

New gradient descent

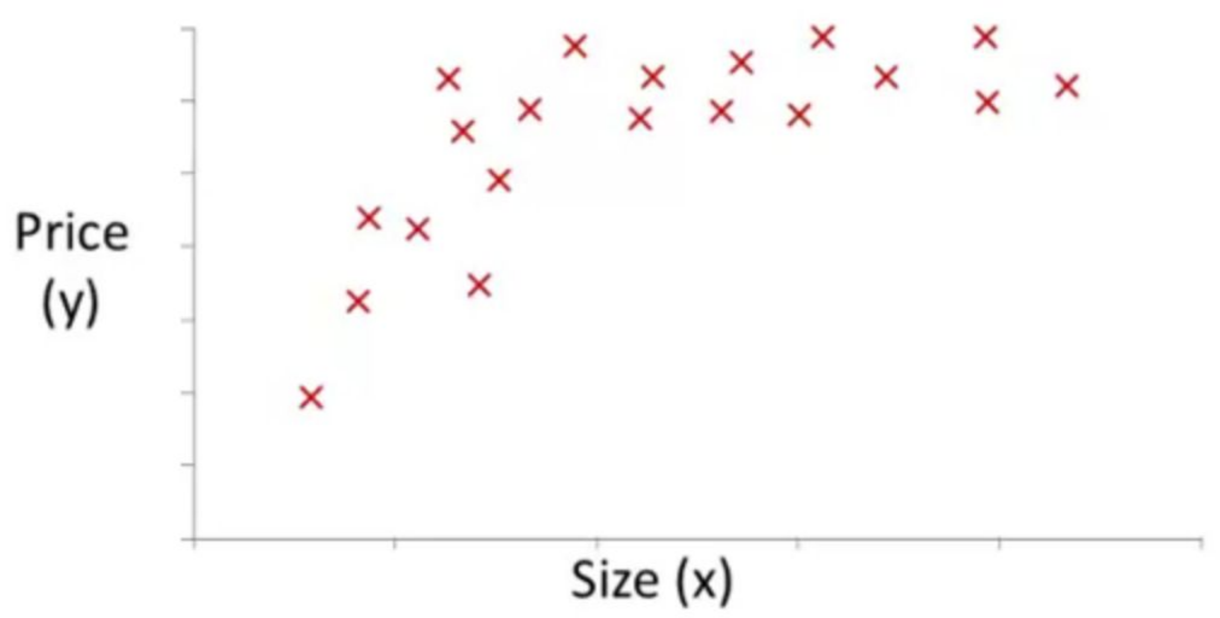
Repeat {

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

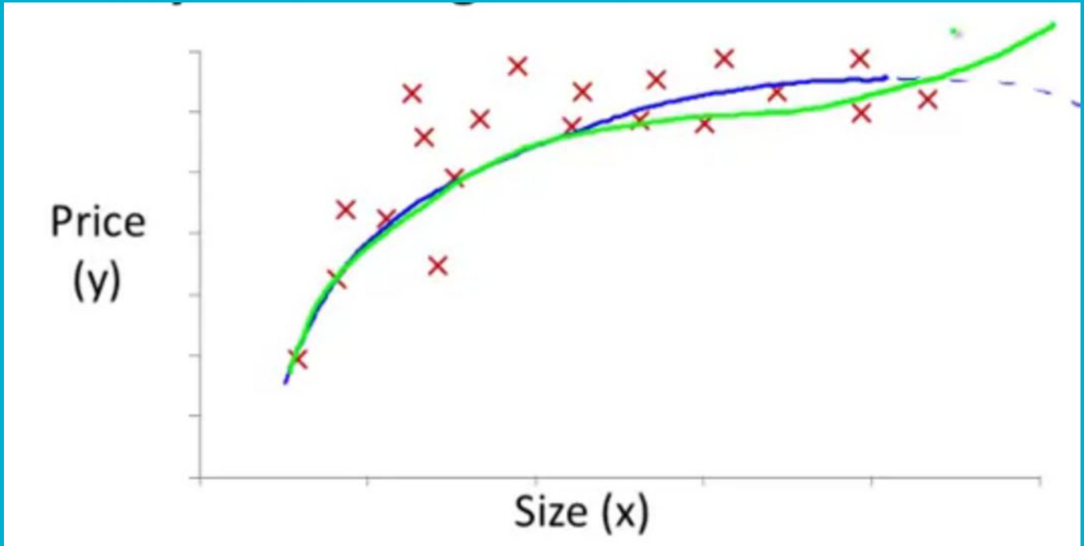
(simultaneously update θ_j for $j = 0, \ldots, n$)

}

Polynomial Regression



Polynomial Regression



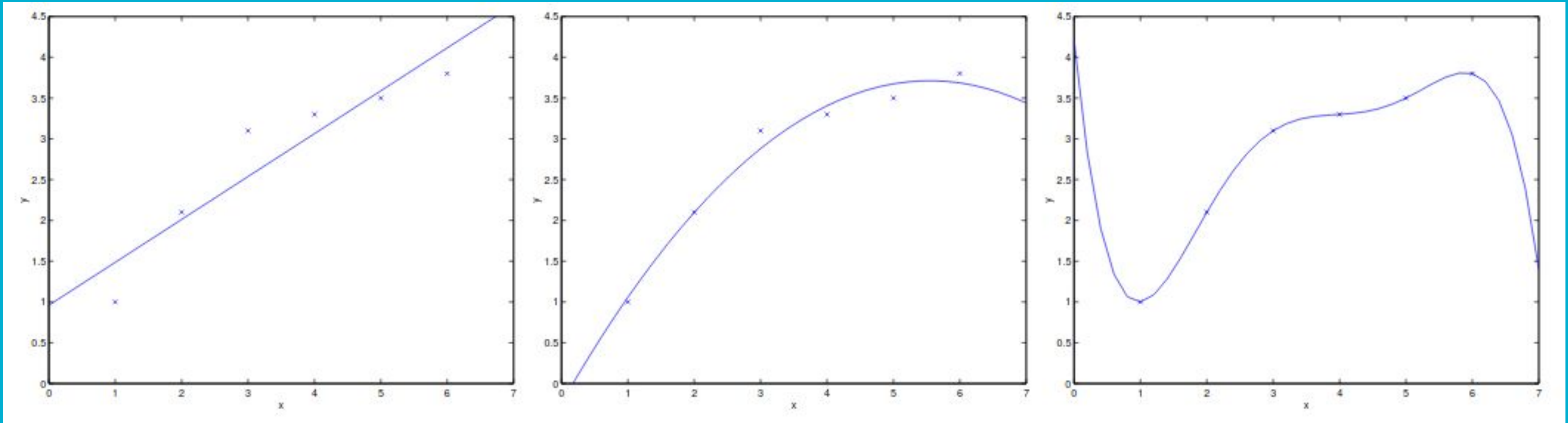
$$\theta_0 + \theta_1 x + \theta_2 x^2$$

$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$$

$$\begin{aligned} h_\theta(x) &= \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 \\ &= \theta_0 + \theta_1 (size) + \theta_2 (size)^2 + \theta_3 (size)^3 \end{aligned}$$

$$x_1 = (size)$$
$$x_2 = (size)^2$$
$$x_3 = (size)^3$$

What problem do you see here?



Thank you