SUPERVISED LEARNING AND LINEAR REGRESSION

DR. FARHAD RAZAVI

OUTLINE

- Machine Learning as a new paradigm
- Different types of Machine Learning
- Regression Model
- Cost Function

CONCEPT OF A FUNCTION

- A function in its simplest definition takes an input and after some manipulation will generate a unique output.
- For instance, it might take a number x and return square root of it $\sqrt{x}$.
- It can take time of the day t and return the position of Jupiter in the sky (x, y, z).
- It can take a color image and return its grayscale image.
- It can take an email and place it in the inbox or spam (1,0)!

$$x \mapsto f(x)$$

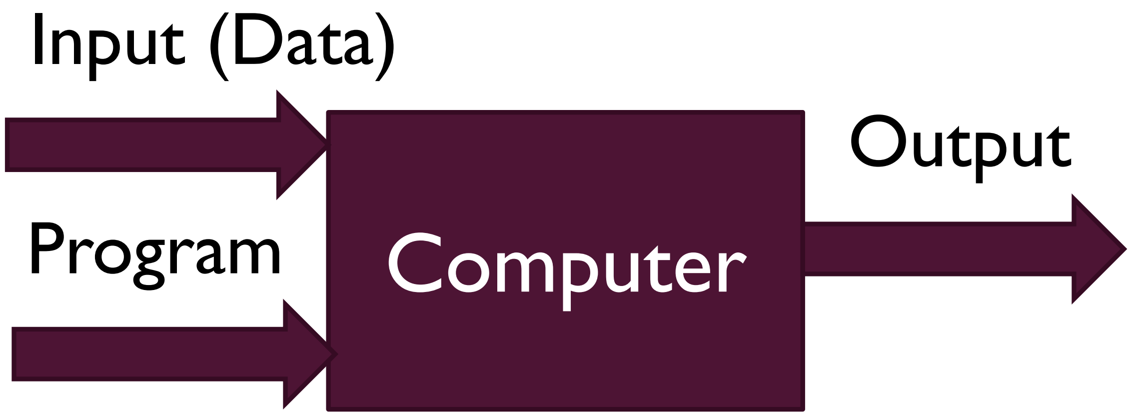
$$f(x) = \sqrt{x}$$

CALCULATING FUNCTIONS

- For the case of square root of x, we need to find a root finding algorithm such as Newton-Raphson.
- Then we need to program it using a computer language!

```python
def newton_method(number, number_iters = 500):
    a = float(number) # number to get square root of
    for i in range(number_iters): # iteration number
        number = 0.5 * (number + a / number) # update
    return number
```

- In general, all the computer programs take some inputs and generate some corresponding outputs.

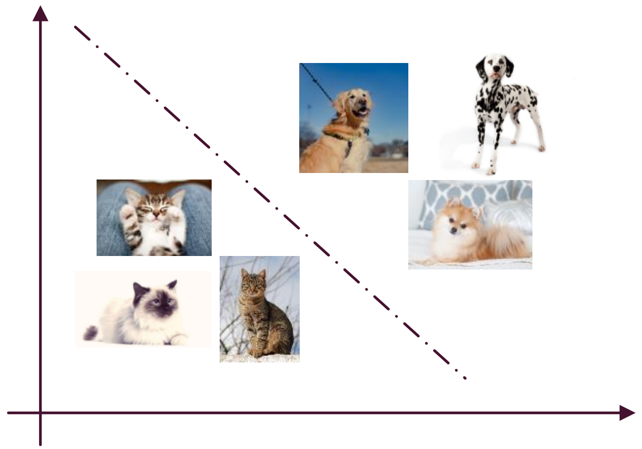


CHANGING THE PARADIGM

- Data can be seen as the input to our model (program).
- Train the model to fit the data to its output.
- If we succeed, we have trained computers to do new things!
- This is the first step in intelligence!
- Program that can infer useful information from implicit data patterns.

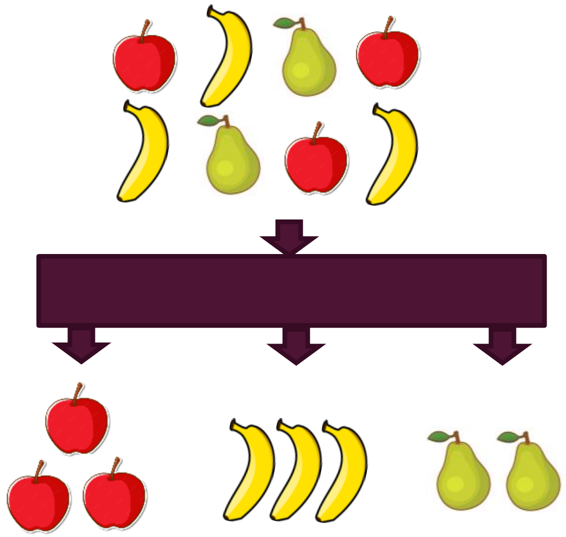
DIFFERENT TYPES OF MACHINE LEARNING

Supervised Learning



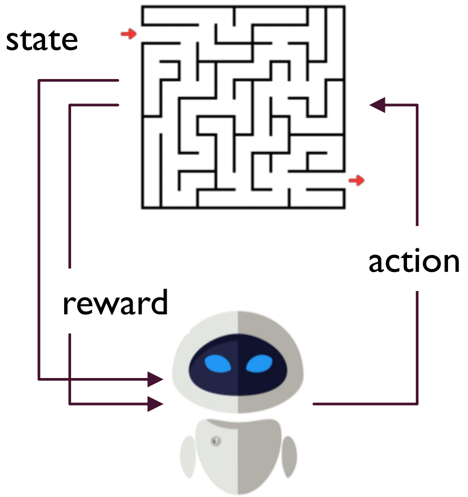
Task driven
(Classification/Regression)

Unsupervised Learning



Data driven
Clustering

Reinforcement Learning

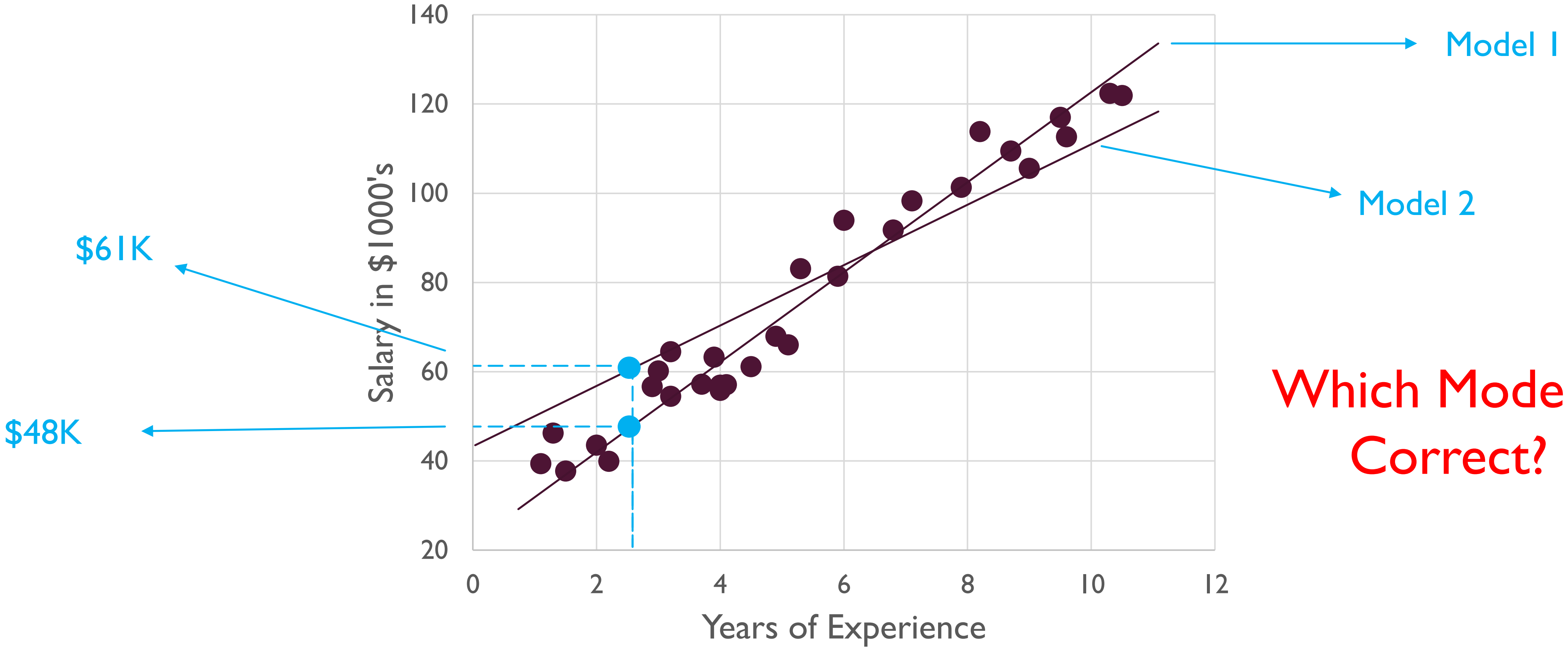


Algorithm learn to react
to an environment

SUPERVISED LEARNING

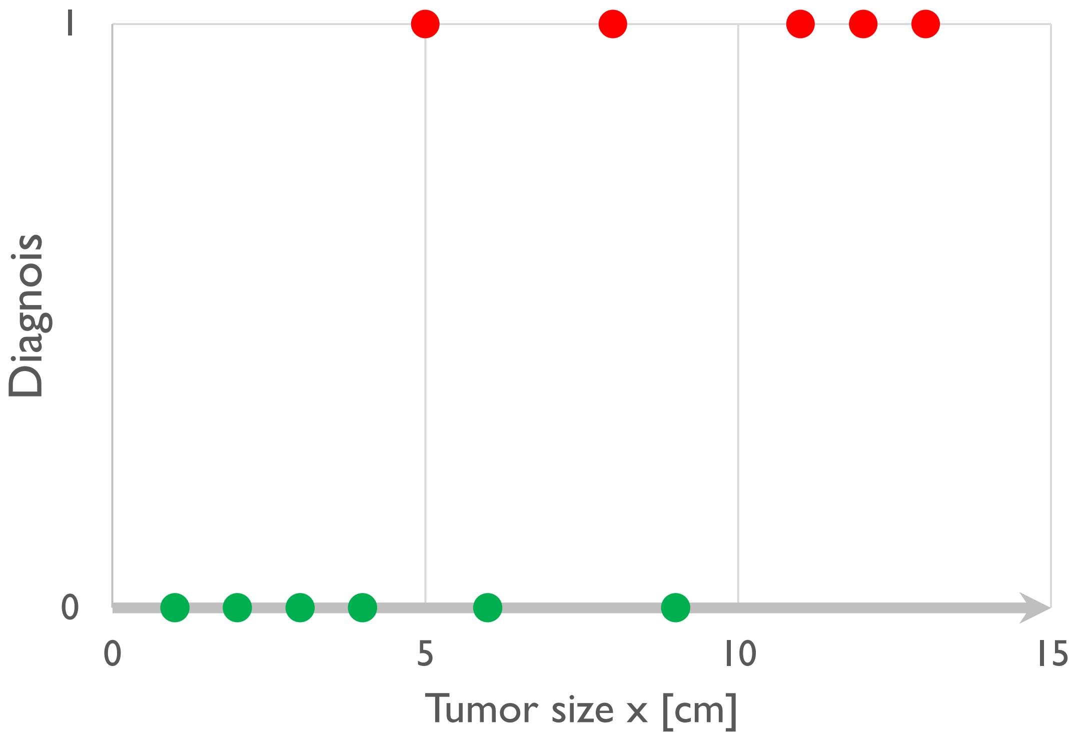
Input (X)	Output (Y)	Application
email	Spam? (0,1)	spam filtering
audio	text transcripts	Speech recognition
Image of cell	Cancerous? (0,1)	Machine translation
Ad, user info	Click? (0,1)	Online advertising
Image, radar info	Position of other cars	Self-driving car
Image of phone	Defect? (0,1)	Visual inspection

PREDICTING SALARY (REGRESSION MODEL)



Which Model is Correct?

CLASSIFICATION



Size	Diagnosis	Designated Value
1	Benign	0
2	Benign	0
3	Benign	0
4	Benign	0
5	Malignant	1
6	Benign	0
8	Malignant	1
9	Benign	0
11	Malignant	1
12	Malignant	1
13	Malignant	1

LET'S RECAP

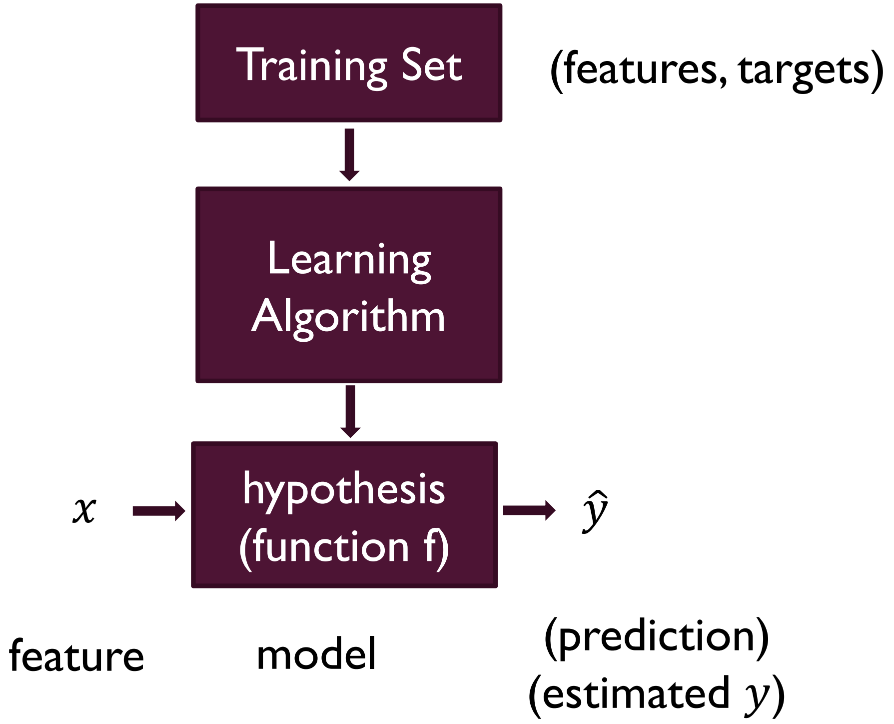
- There are two very important types of models in supervised learning.
  - Regression Models
  - Classification Models
- Regression Model:
  - We are trying to predict an outcome from infinitely (or practically infinitely) many possible numbers.
  - The output can be modeled as a real number
  - Linear regression is one of the most fundamental types of regression.
    - Many Nonlinear models can be reduced to a linear regression (more on this later).
- Classification Model:
  - There are only a small number of possible output values (mostly 0, 1 but it could be other finite sets).
  - It is also called "categorization" model.

TERMINOLOGY REVIEW

Item #	Years of Experience x	Salary in \$1000's y
(1)	1.1	39.343
(2)	1.3	46.205
(3)	1.5	37.731
(4)	2	43.525
…	…	…
(40)	9	105.231

- Notation:
- x = "input" variable also known as "feature"
- y = "output" variable also known as "target"
- m = number of training examples
- m = single training example
- $(x^{(i)}, y^{(i)})$ = i^{th} training example (1^{st}, 2^{nd}, 3^{rd}, …)
- $x^{(i)} \neq x^i$ (this is not a power notation)

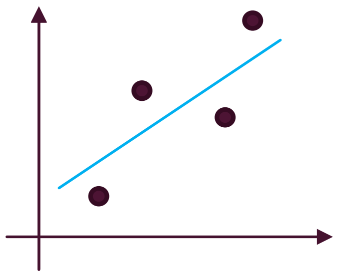
STEPS INVOLVED IN MACHINE LEARNING



How to pick a specific function?

$$f_{w,b}(x) = wx + b$$

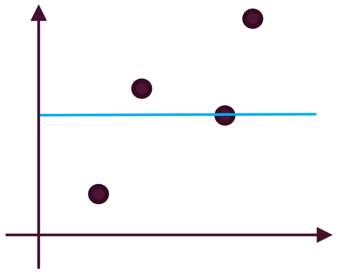
$$f(x) = wx + b$$



- w and b are the parameters of our model
- This model is called Linear regression with one variable or Univariate linear regression

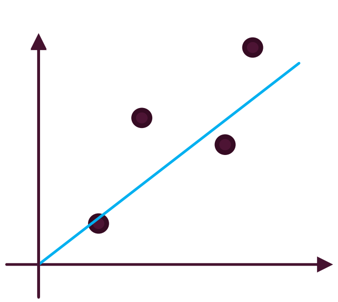
FINDING THE PARAMETERS OF THE MODEL

$f(x) = wx + b$



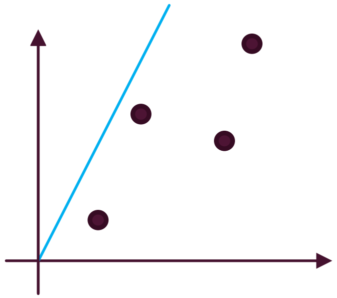
$(w, b) = (0,1)$

$f(x) = wx + b$



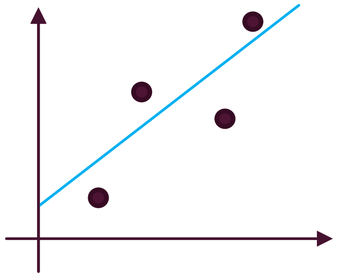
$(w, b) = (1,0)$

$f(x) = wx + b$



$(w, b) = (2,0)$

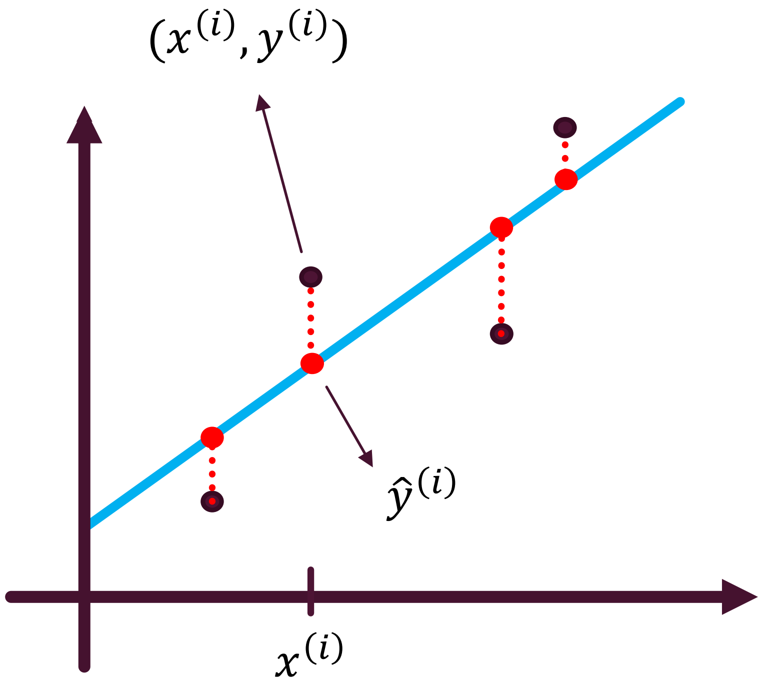
$f(x) = wx + b$



$(w, b) = (1,0.5)$

- We need to define a metric for how well each of these parameters estimate our data.
- We call this metric cost function

COST FUNCTION



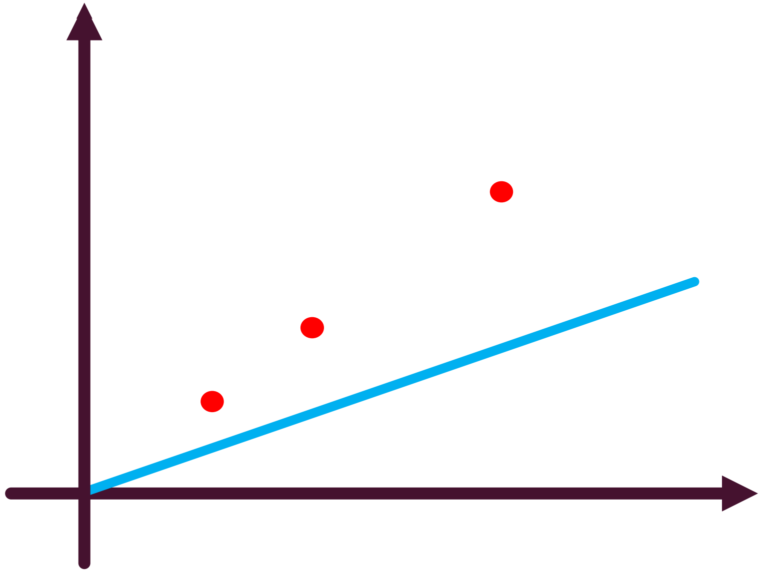
- We want the distance from our prediction $\hat{y}$ to our target y to be minimum.

- $\left(\hat{y}^{(i)} - y^{(i)}\right)$ — error at i^th point.
- $\left(\hat{y}^{(i)} - y^{(i)}\right)^2$ — make sure error term is positive.
- $\sum_{i=1}^{m}\left(\hat{y}^{(i)} - y^{(i)}\right)^2$ — sum over all the error terms
- $\frac{1}{m}\sum_{i=1}^{m}\left(\hat{y}^{(i)} - y^{(i)}\right)^2$ — normalize the summation
- $\frac{1}{2m}\sum_{i=1}^{m}\left(\hat{y}^{(i)} - y^{(i)}\right)^2$ — helps to simplifies the maths
- $J(w, b) = \frac{1}{2m}\sum_{i=1}^{m}\left(f_{w,b}\left(x^{(i)}\right) - y^{(i)}\right)^2$ — squared error cost function

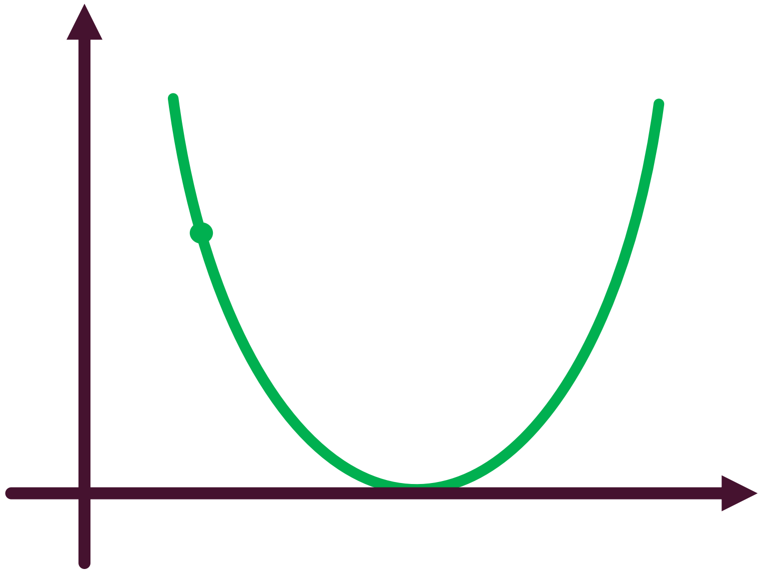
COST FUNCTION EXPLORATION

$b = 0$

$f_{w,b}(x) = wx + b = wx$



$$J(w,b) = \frac{1}{6}\sum_{i=1}^{3}\left(f_{w,b}\left(x^{(i)}\right) - y^{(i)}\right)^2$$



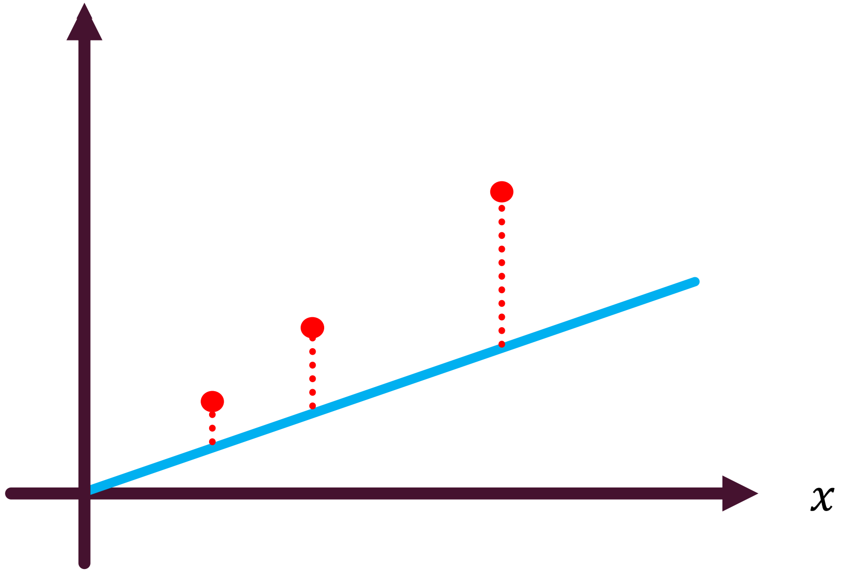
$$6\,J(w) = \left(wx^{(1)} - y^{(1)}\right)^2 + \left(wx^{(2)} - y^{(2)}\right)^2 + \left(wx^{(3)} - y^{(3)}\right)^2$$

$$J(w) = \mathrm{A}w^2 + \mathrm{B}w + C$$

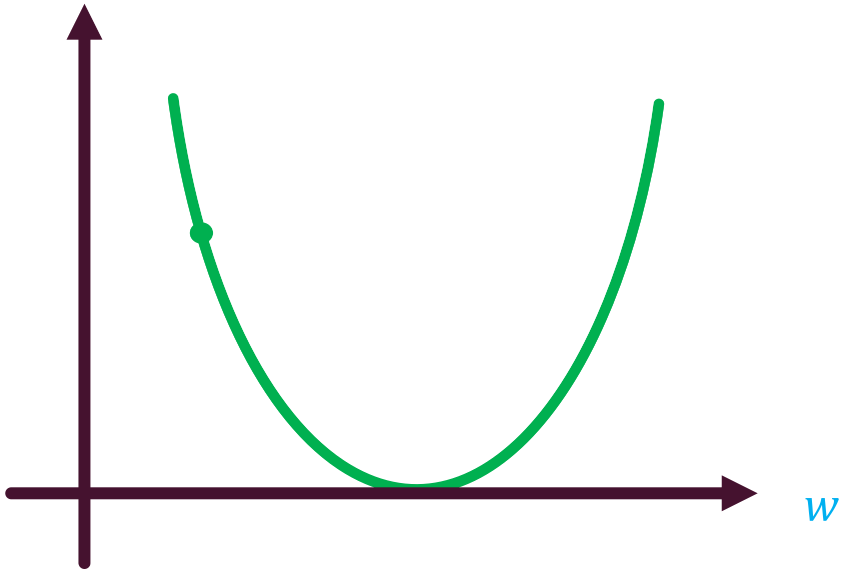
COST FUNCTION EXPLORATION

$b = 0$

$f_{w,b}(x) = wx + b = wx$



$$J(w,b) = \frac{1}{6}\sum_{i=1}^{3}\left(f_{w,b}\left(x^{(i)}\right) - y^{(i)}\right)^2$$

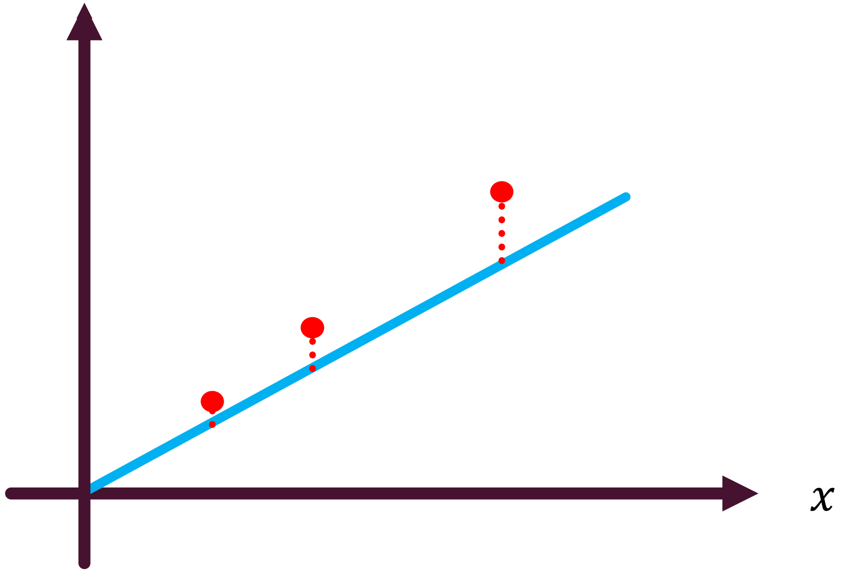


$J(w) = Aw^2 + Bw + C$

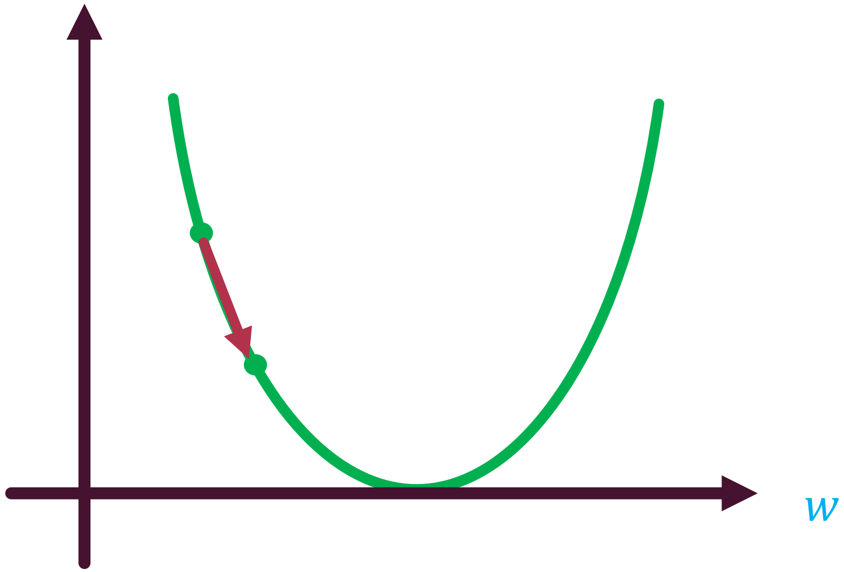
COST FUNCTION EXPLORATION

$b = 0$

$f_{w,b}(x) = wx + b = wx$



$$J(w,b) = \frac{1}{6}\sum_{i=1}^{3}\left(f_{w,b}\left(x^{(i)}\right) - y^{(i)}\right)^2$$

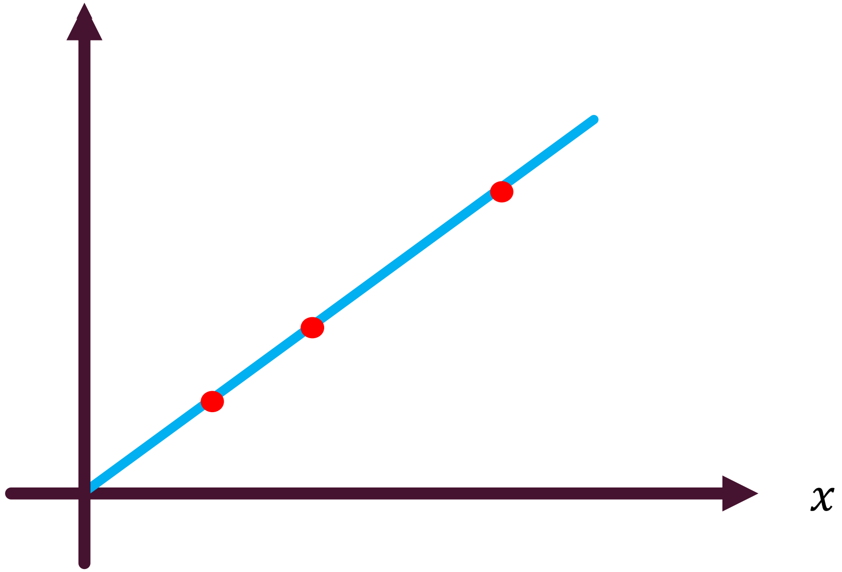


$J(w) = \mathrm{A}w^2 + \mathrm{B}w + C$

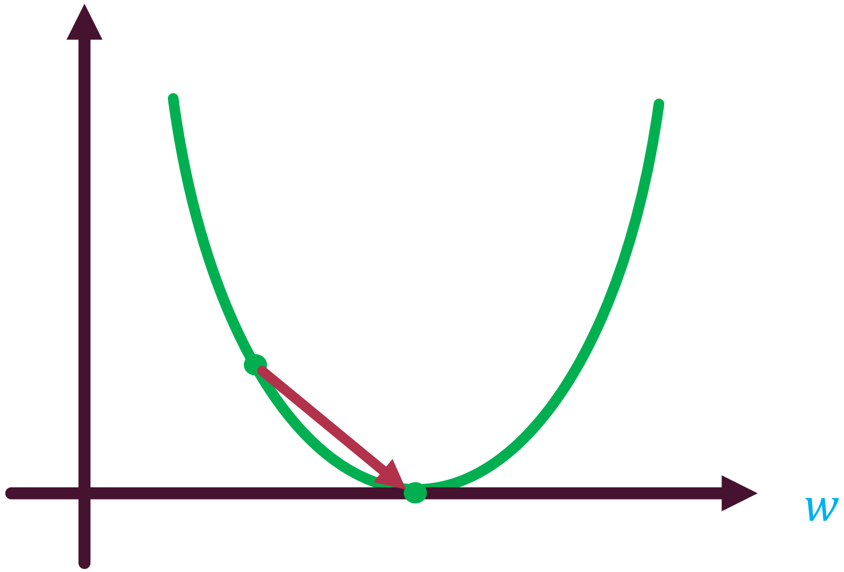
COST FUNCTION EXPLORATION

$b = 0$

$f_{w,b}(x) = wx + b = wx$



$$J(w,b) = \frac{1}{6}\sum_{i=1}^{3}\left(f_{w,b}\left(x^{(i)}\right) - y^{(i)}\right)^2$$

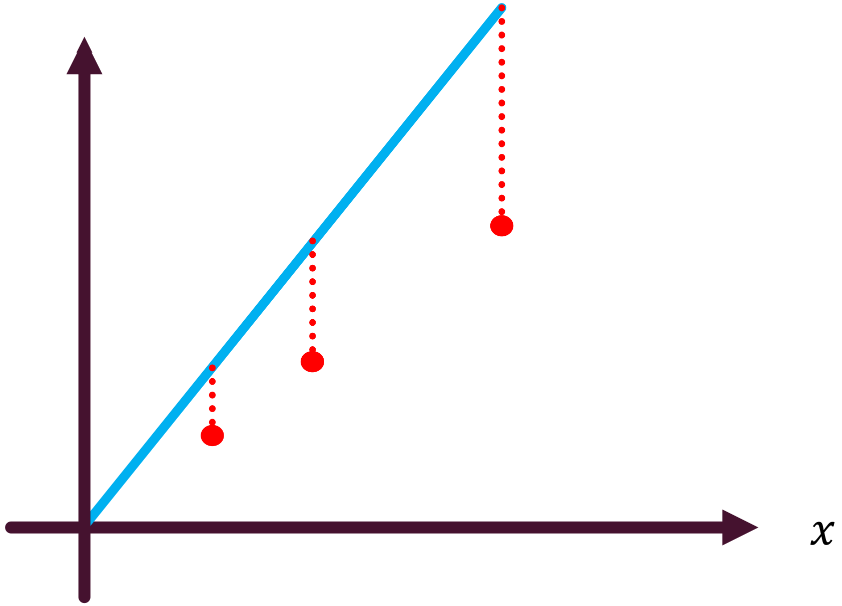


$J(w)$= Aw^2 + Bw + C

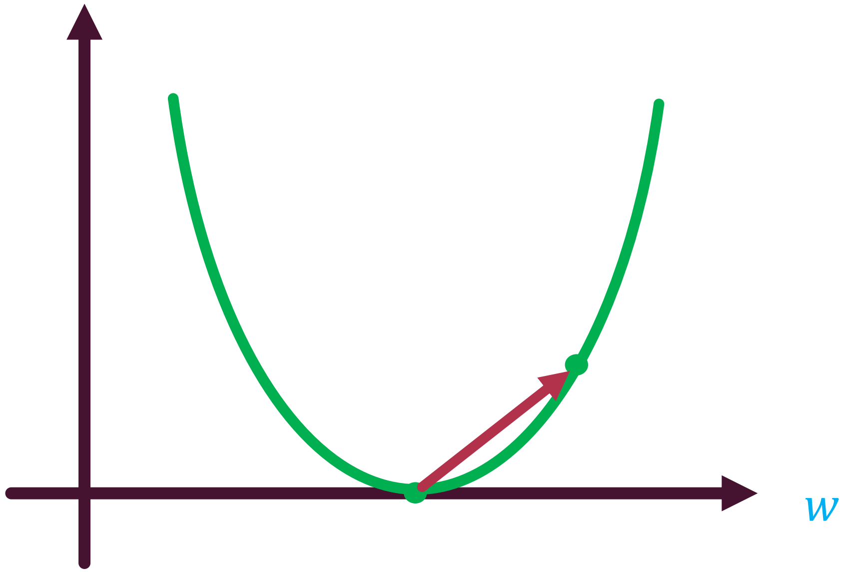
COST FUNCTION EXPLORATION

$b = 0$

$f_{w,b}(x) = wx + b = wx$



$$J(w,b) = \frac{1}{6}\sum_{i=1}^{3}\left(f_{w,b}\left(x^{(i)}\right) - y^{(i)}\right)^2$$

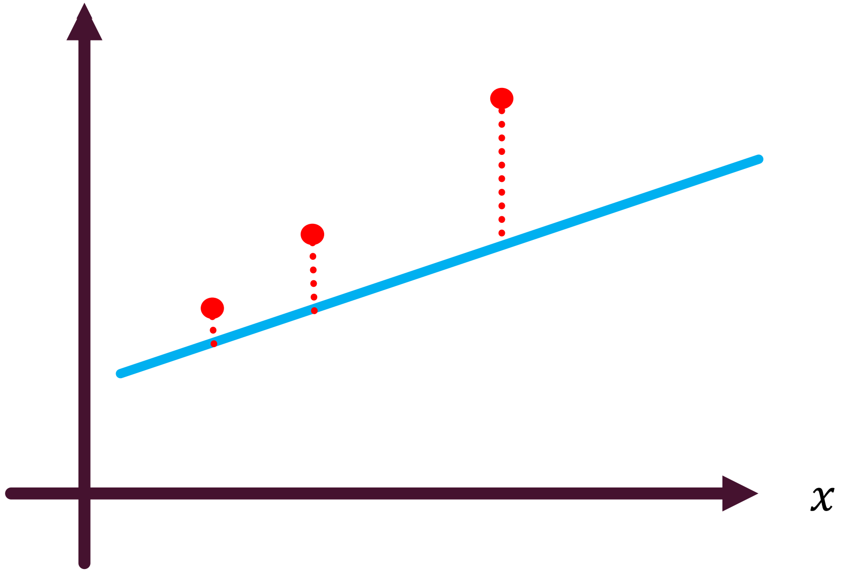


$J(w) = \mathrm{A}w^2 + \mathrm{B}w + C$

COST FUNCTION

$b \neq 0$

$f_{w,b}(x) = wx + b$

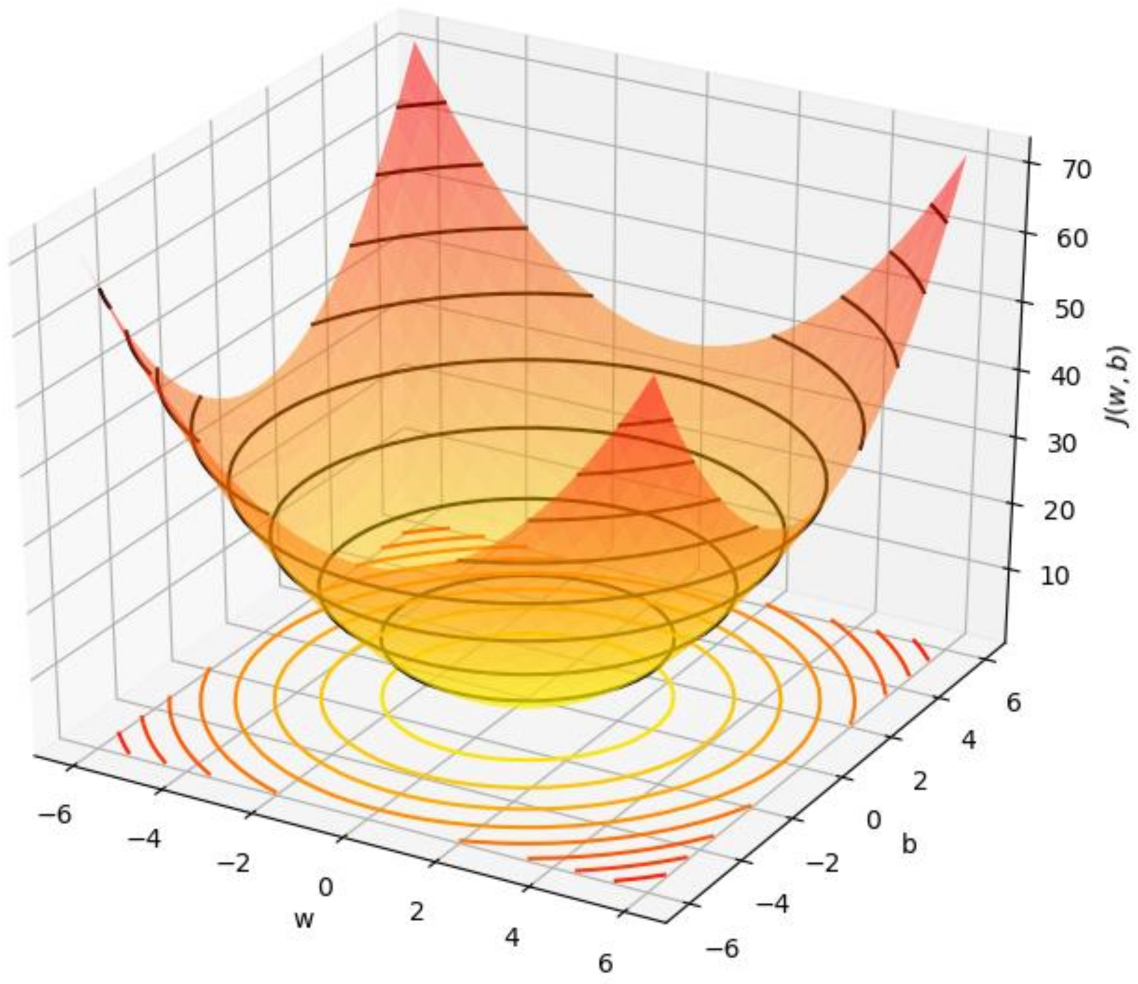


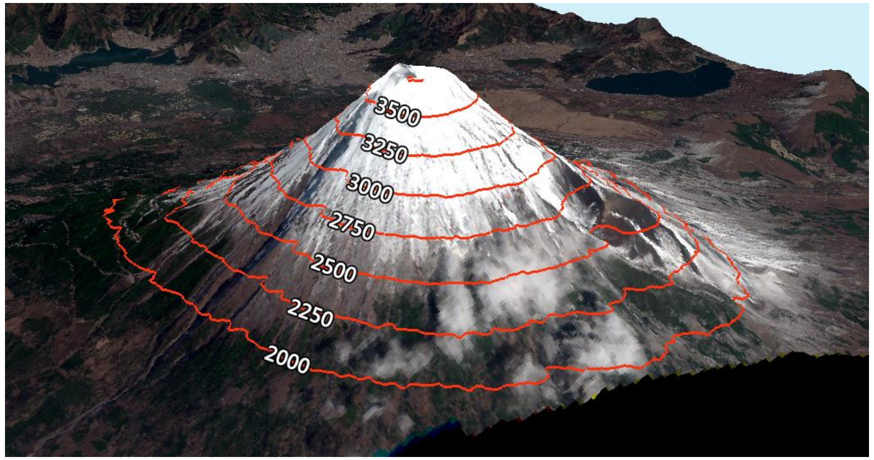
$$J(w,b) = \frac{1}{6}\sum_{i=1}^{3}\left(f_{w,b}\left(x^{(i)}\right) - y^{(i)}\right)^2$$

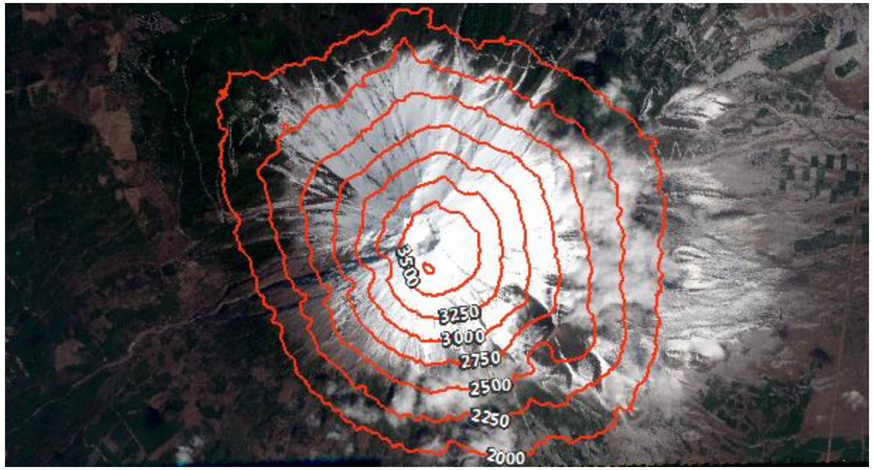
$$J(w,b) = \frac{1}{6}\sum_{i=1}^{3}\left(wx^{(i)} + b - y^{(i)}\right)^2$$

$$J(w,b) = \mathrm{A}w^2 + \mathrm{B}b^2 + \mathrm{C}wb + \mathrm{D}w + \mathrm{E}b + F$$

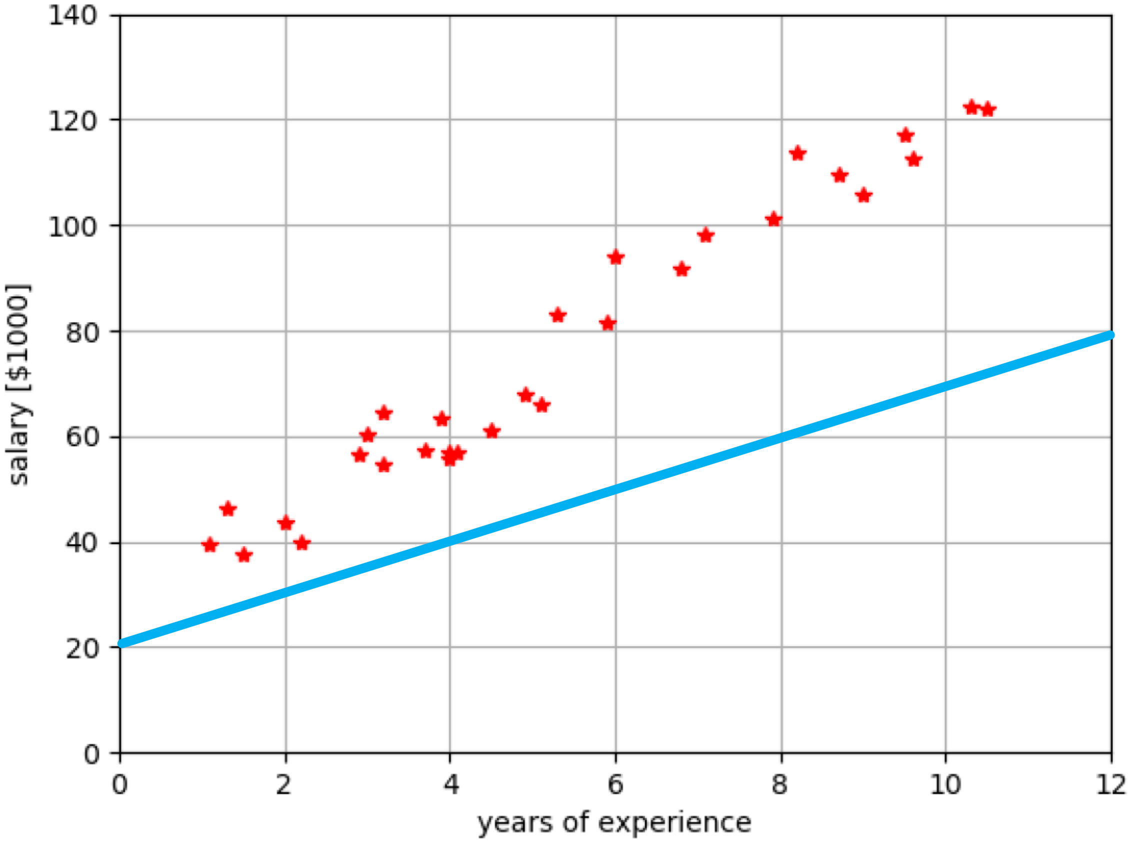
COST FUNCTION WITH TWO PARAMETERS

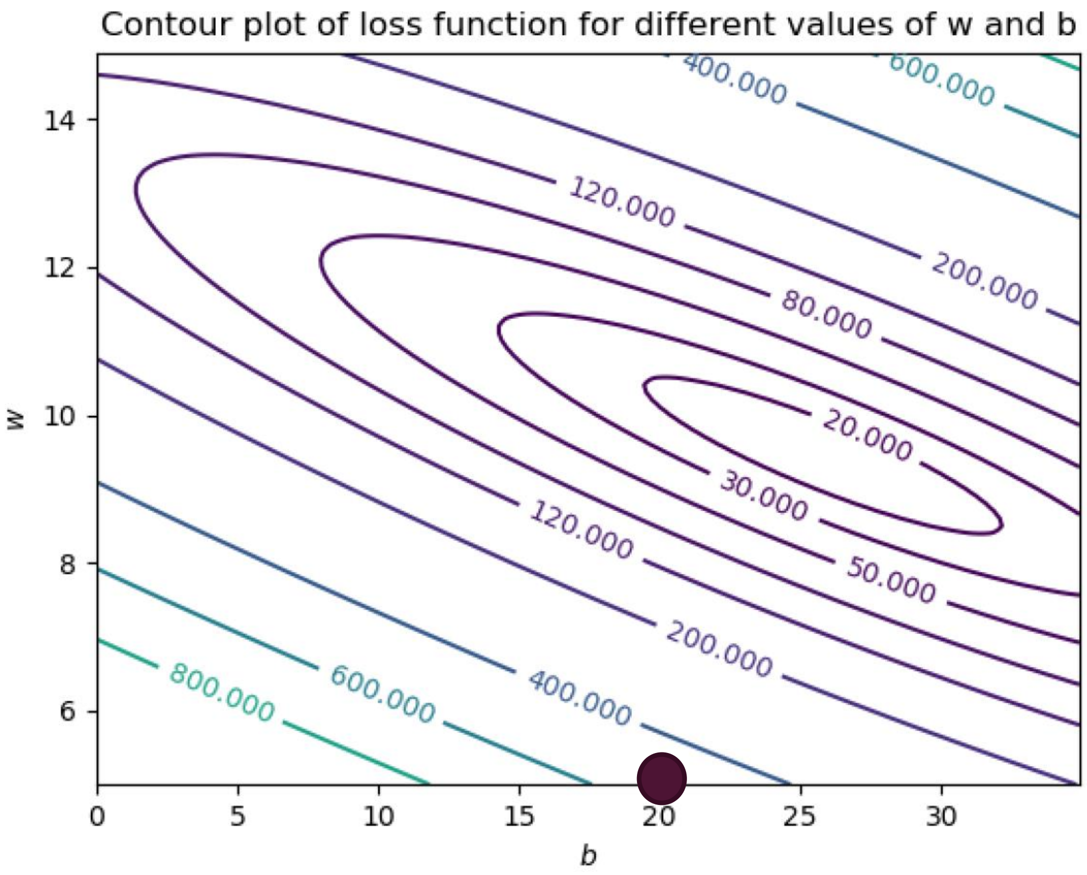






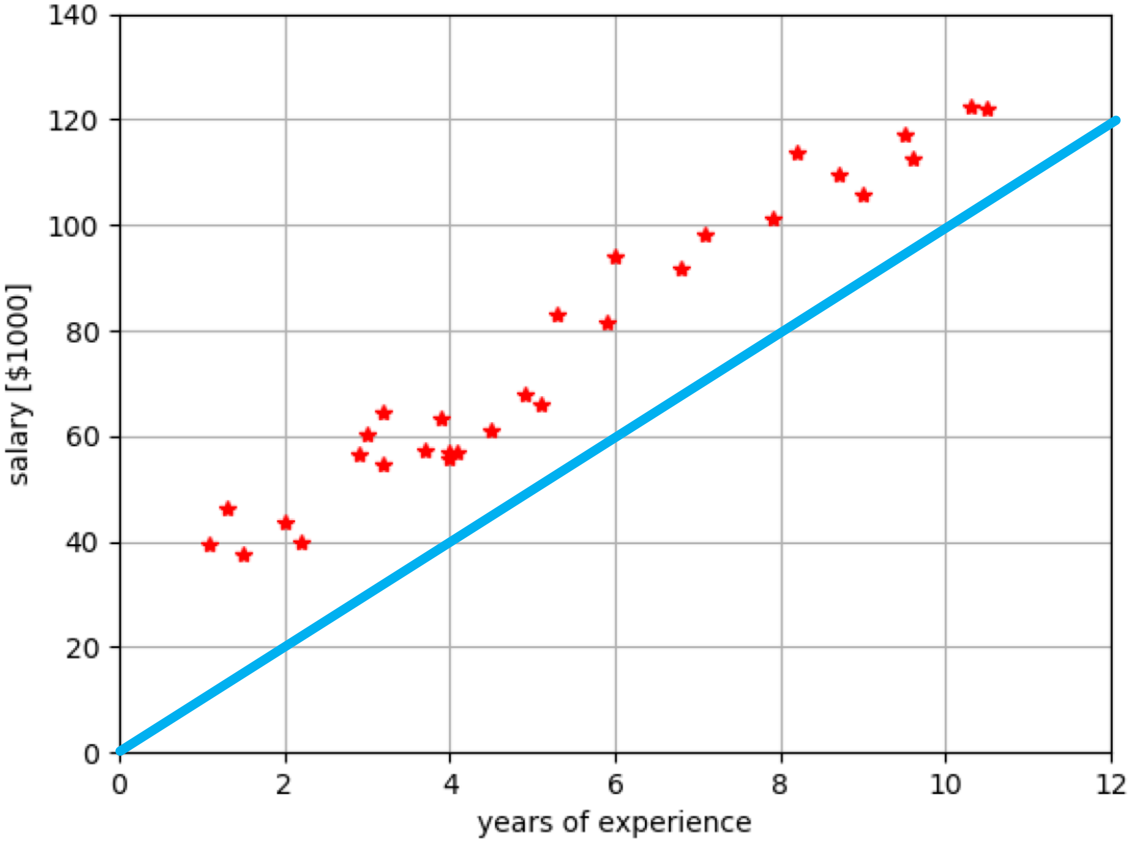
COST FUNCTION WITH TWO PARAMETERS

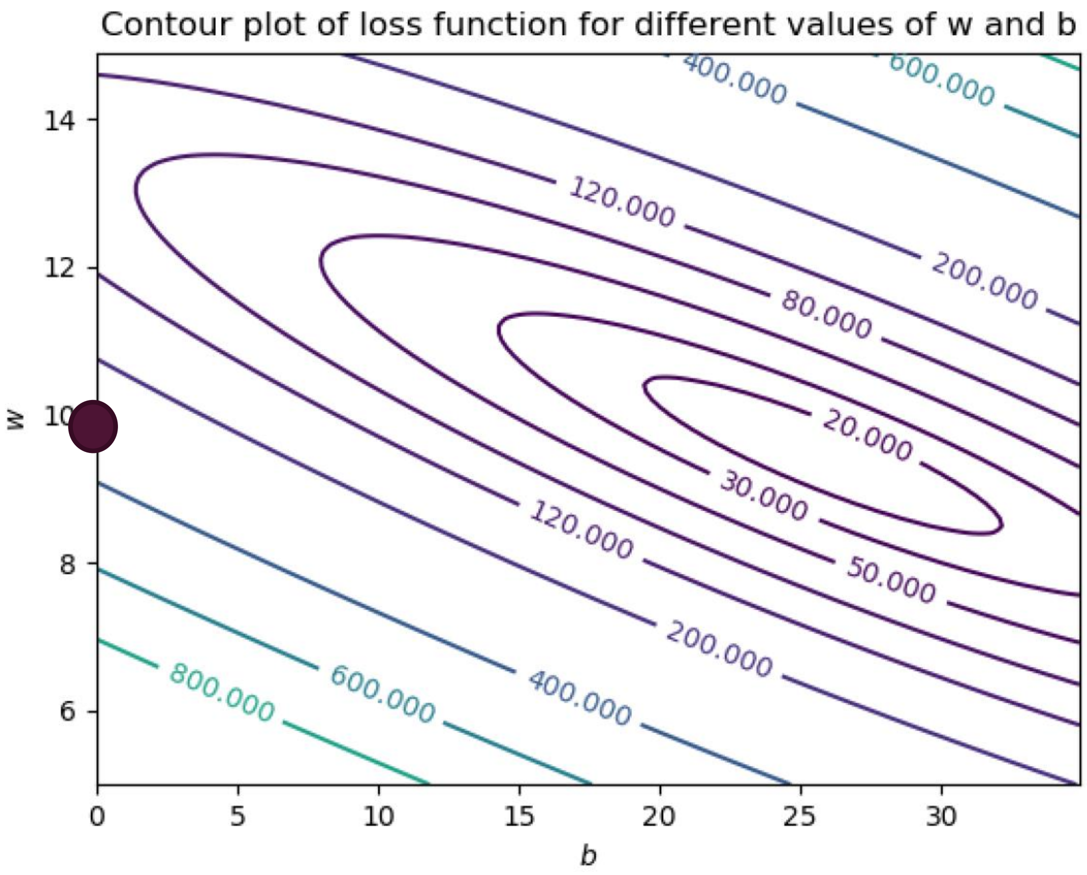




$$f_{w,b}(x) = wx + b = 5x + 20$$

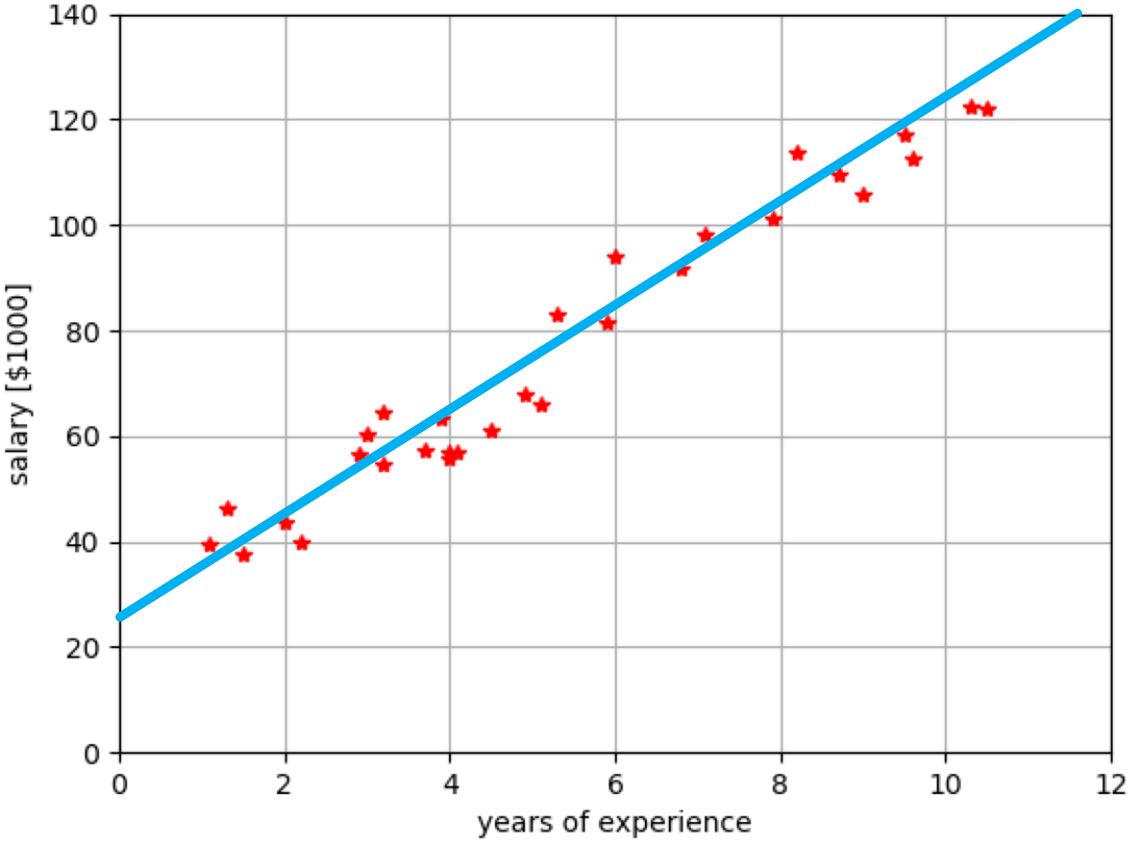
COST FUNCTION WITH TWO PARAMETERS

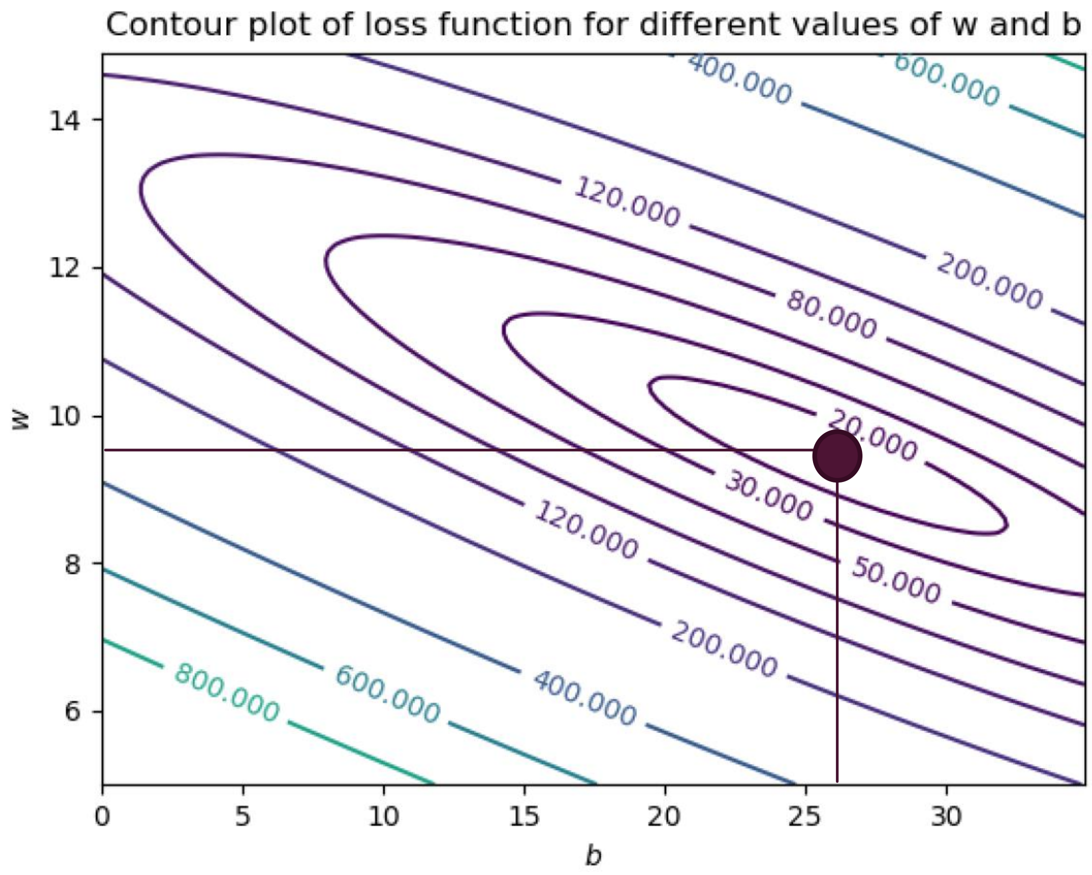




$$f_{w,b}(x) = wx + b = 10x + 0$$

COST FUNCTION WITH TWO PARAMETERS





$$f_{w,b}(x) = wx + b = 9.5x + 26$$

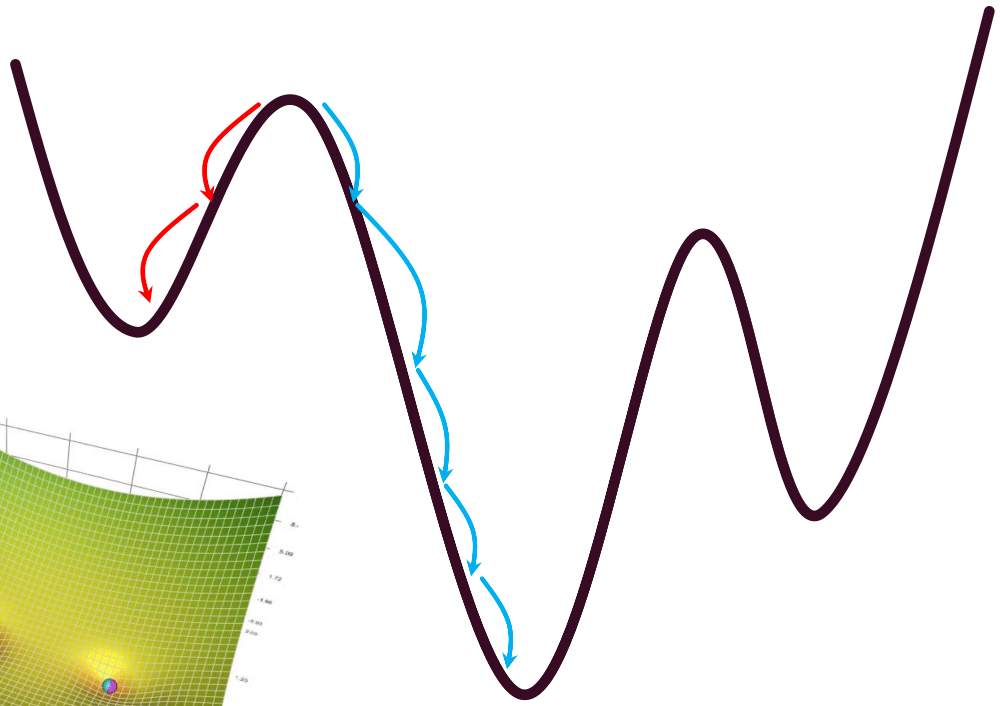
GRADIENT DESCENT

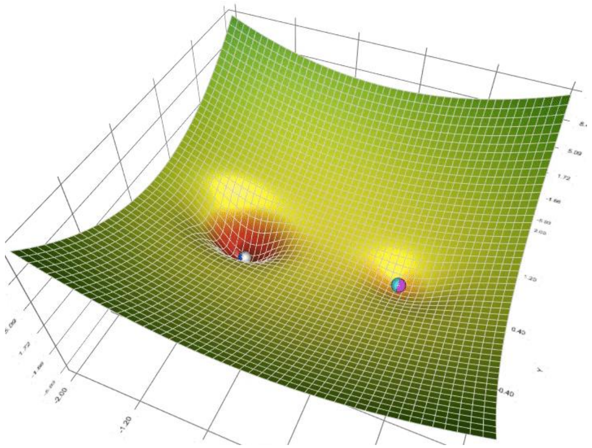
- Let's say we have a cost function $J(w, b)$.
- Gradient Descent is an algorithm which will minimize the cost function over its parameters:
  - $\min_{w,b} J(w, b)$
- It is a very general algorithm and can be applied to any machine learning technique.
- It is used even in deep learning algorithms.
- It can be applied to cost functions with many parameters.
  - $J(w_1, w_2, \dots, w_n, b)$

GRADIENT DESCENT INTUITION

- Gradient Descent Algorithm:
  - Start with some w and b.
  - Keep changing w, and b to reduce $J(w, b)$.
  - Continue until we settle at or near a minimum.

- Drawbacks:
  - It could get stuck in a local minima.



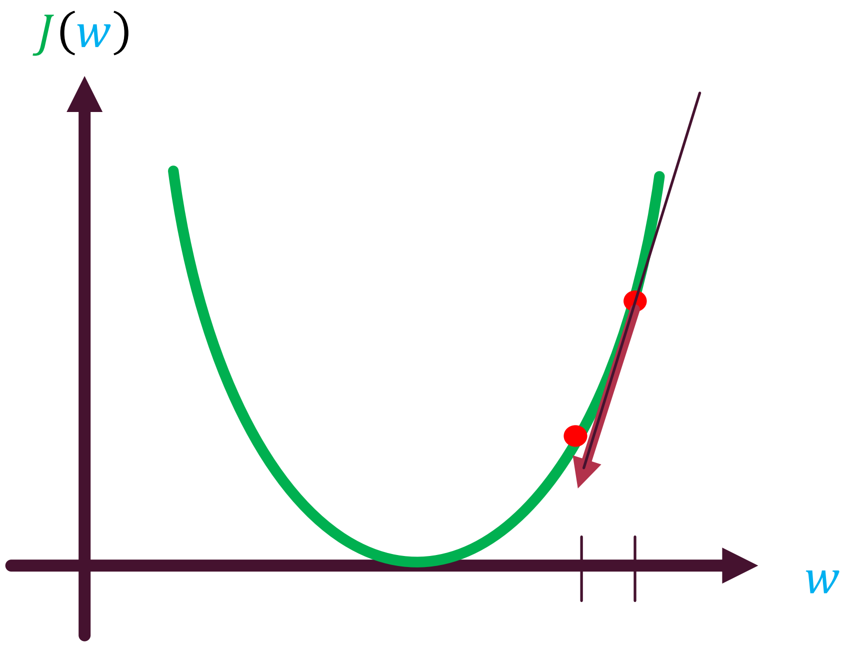


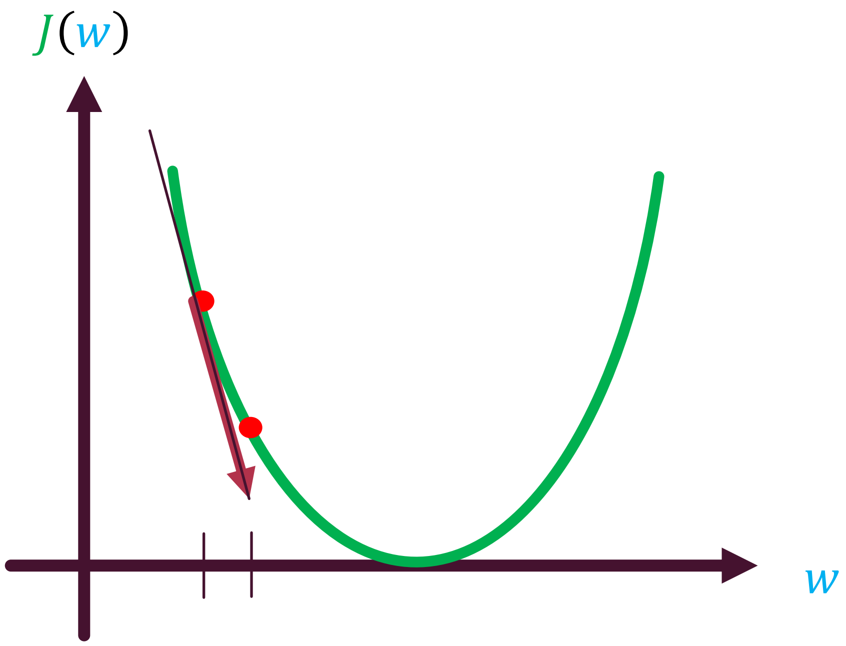
BATCH GRADIENT DESCENT

- At each iteration do the following:
- Calculate the next value of w as $w - \alpha \frac{\partial}{\partial w} J(w, b)$
  - α: learning rate (always positive $\alpha > 0$)
  - $\frac{\partial}{\partial w} J(w, b)$: Derivative of function $J(w, b)$ with respect to w
- Calculate the next value of b as $b - \alpha \frac{\partial}{\partial b} J(w, b)$
  - $\frac{\partial}{\partial b} J(w, b)$: Derivative of function $J(w, b)$ with respect to b
- Make sure w and b are updated simultaneously.
- Repeat until converged.

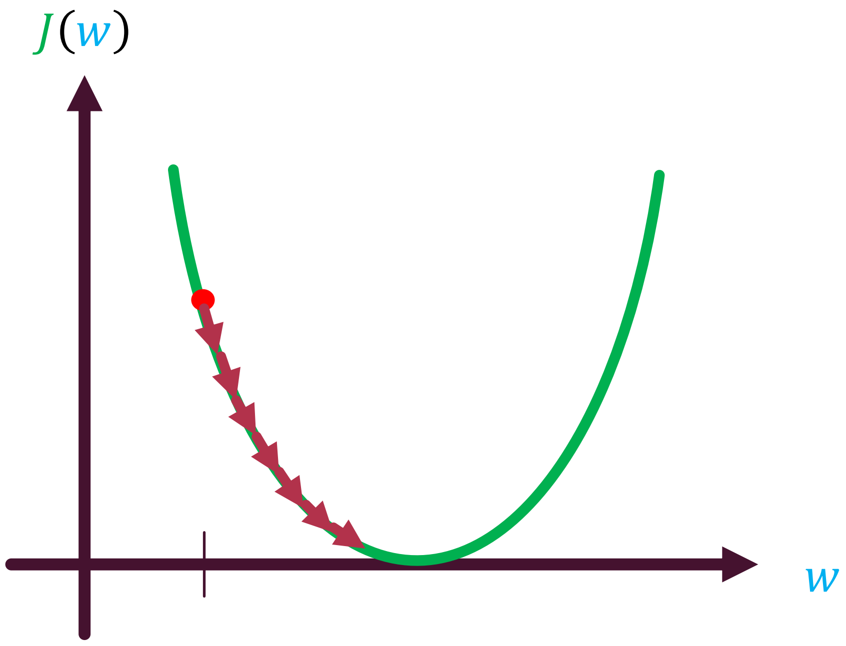
DIRECTION OF MOVEMENT

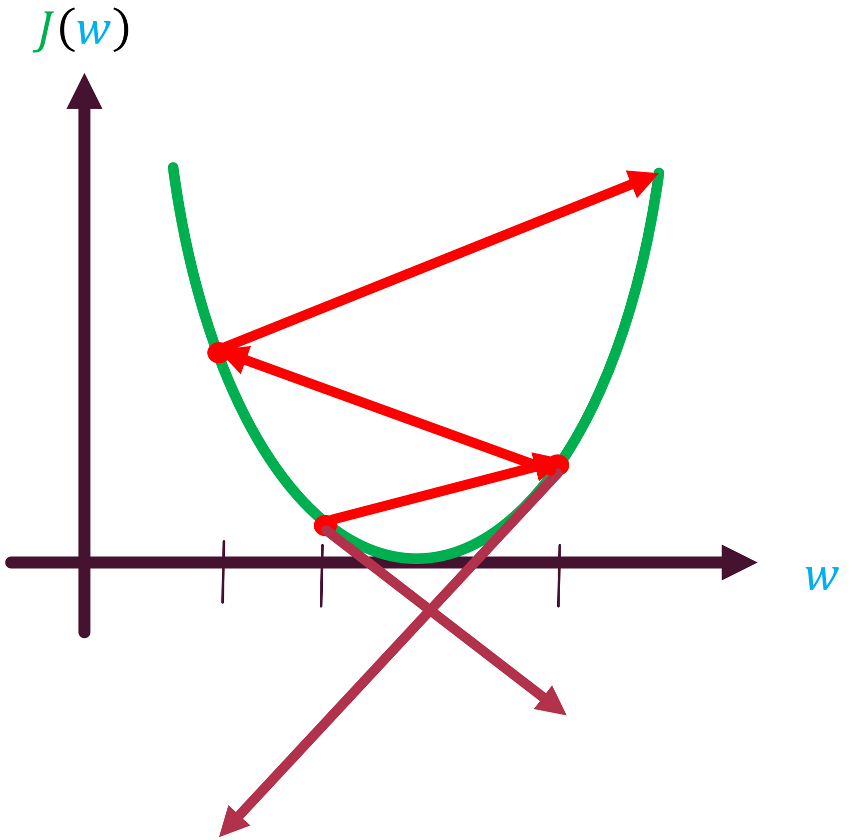
$$w \to w - \alpha \frac{\partial}{\partial w} J(w, b = 0) = w - \alpha \frac{d}{dw} J(w)$$



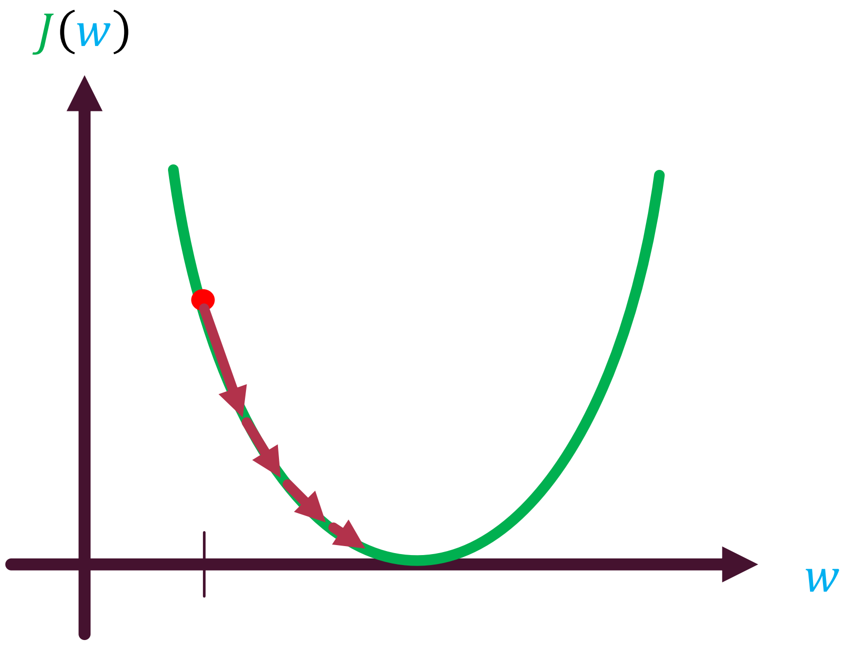


CHOICE OF α





FIXED LEARNING RATE



- Gradient Descent can always reach to a local minimum with a fixed learning rate.
- $w \to w - \alpha \frac{\partial}{\partial w} J(w, b)$
- Near a local minimum
  - Derivative becomes smaller
  - Update steps become smaller
  - There is no need to decrease the learning rate

GRADIENT DESCENT FOR LINEAR REGRESSION

- Linear Regression:
  - $f_{w,b}(x) = wx + b$
  - $J(w, b) = \frac{1}{2m} \sum_{i=1}^{m} \left(f_{w,b}\left(x^{(i)}\right) - y^{(i)}\right)^2$
- Gradient Descent
  - Repeat until convergence {

$$w \to w - \alpha \frac{\partial}{\partial w} J(w, b) \; ; \qquad \frac{\partial}{\partial w} J(w, b) = \frac{1}{m} \sum_{i=1}^{m} \left(f_{w,b}\left(x^{(i)}\right) - y^{(i)}\right) x^{(i)}$$

$$b \to b - \alpha \frac{\partial}{\partial b} J(w, b); \qquad \frac{\partial}{\partial b} J(w, b) = \frac{1}{m} \sum_{i=1}^{m} \left(f_{w,b}\left(x^{(i)}\right) - y^{(i)}\right)$$

}

- Cost function of Linear Regression is quadratic. This means it has only one minimum value, so we are guaranteed to reach global minimum.