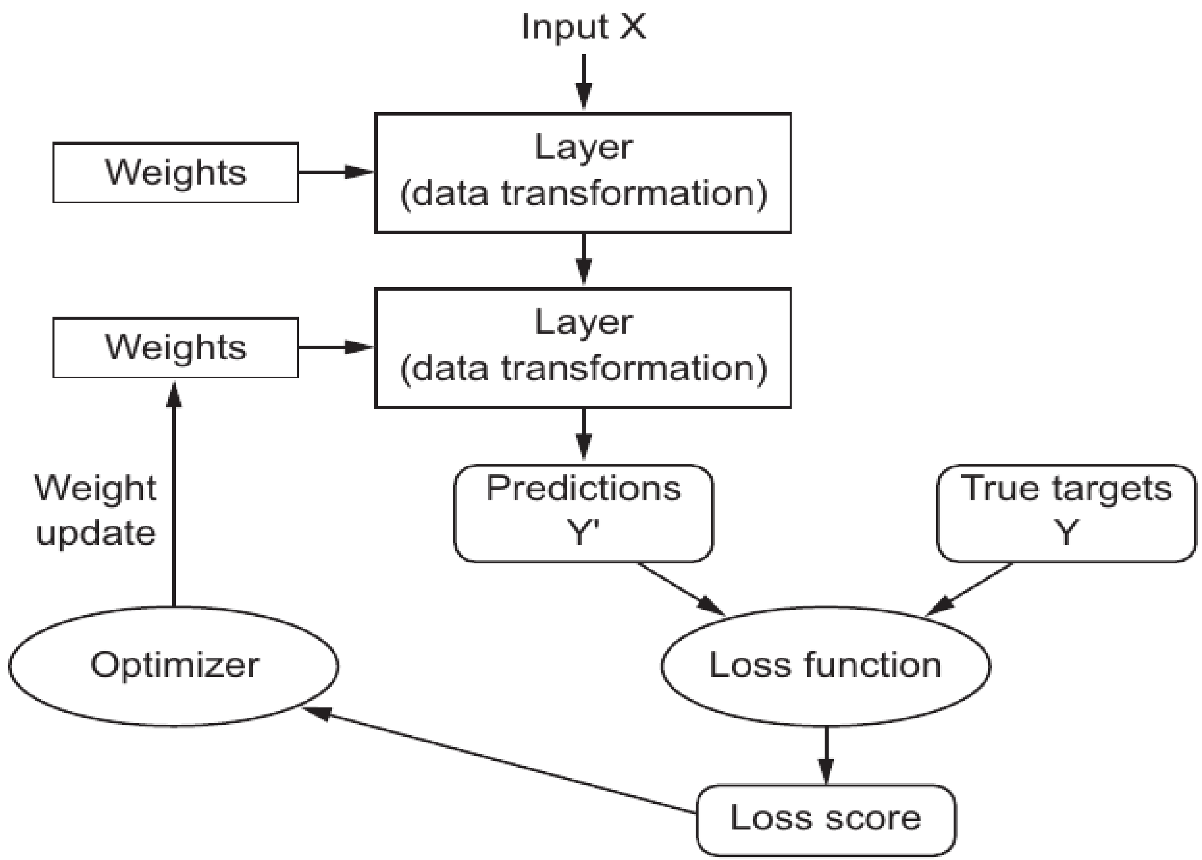
CORK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA CHORCAÍ

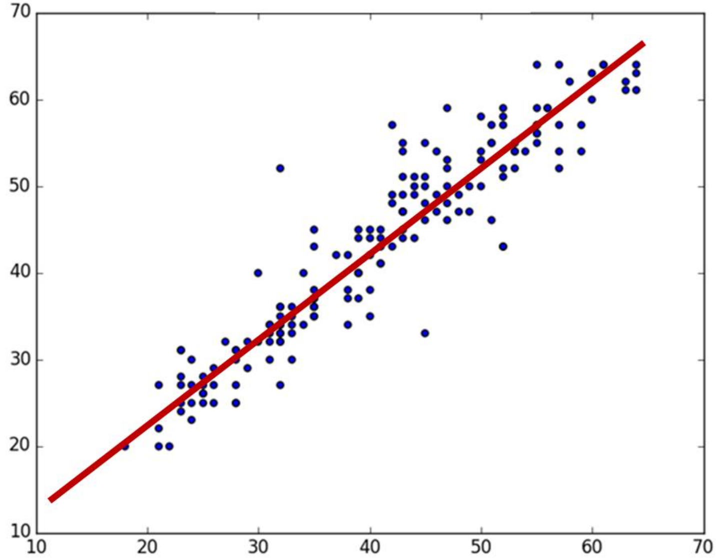
Machine Learning

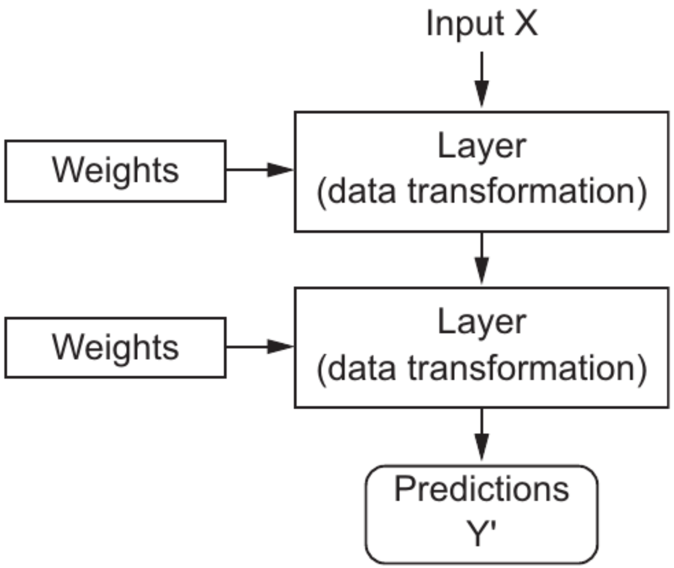
Machine Learning

Lecture: Linear Regression

Ted Scully

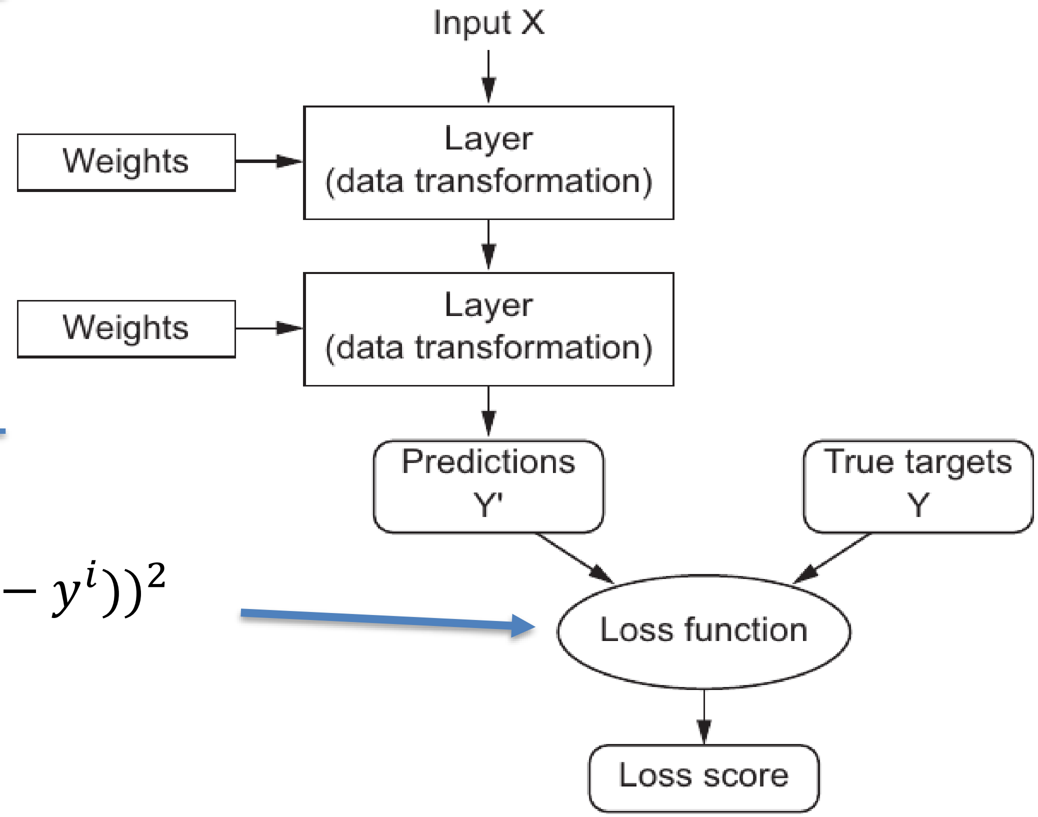

Input X
Weights
Layer
(data transformation)
Weights
Layer
(data transformation)
Weight
update
Predictions
Y'
True targets
Y
Optimizer
Loss function
Loss score



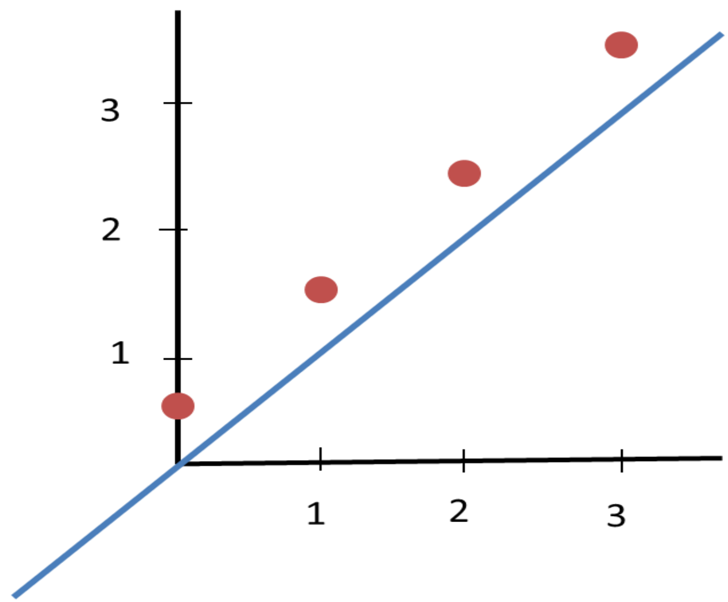


- $h(x) = \lambda_1 x + b$

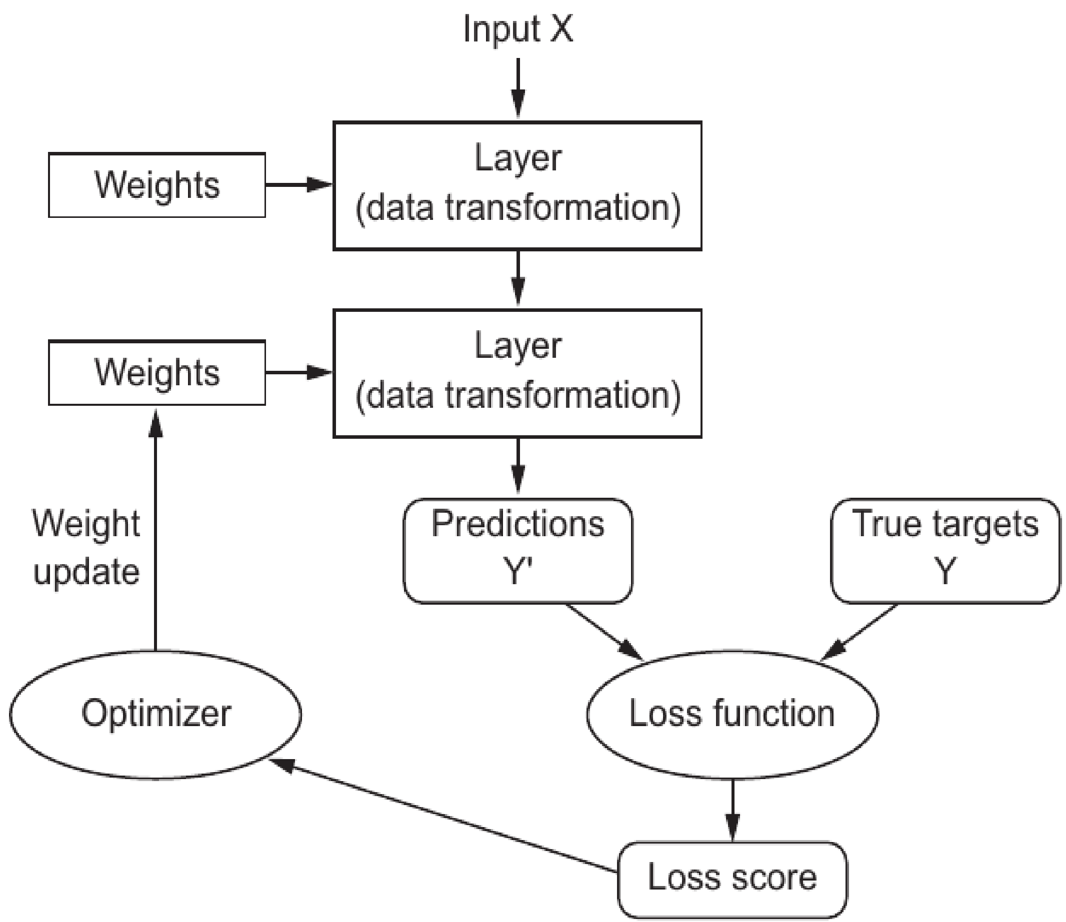
$$h(x) = \lambda_1 x + b$$



$$C(\lambda_1, b) = \frac{1}{2m}\sum_{i=0}^{m}((h(x^i) - y^i))^2$$



$$h(x) = \lambda_1 x + b$$



Gradient Descent Optimizer

$$C(\lambda_1, b) = \frac{1}{2m}\sum_{i=0}^{m}((h(x^i) - y^i))^2$$

Gradient Descent and Linear Regression

- To understand and visualize the application of gradient descent we are going to **simplify the problem even further**.
- We are going to assume that the y-intercept (*b*) is always 0. Therefore, we only have one variable to worry about now (λ_1)

- ***h(x) =*** $\boldsymbol{\lambda_1 x}$

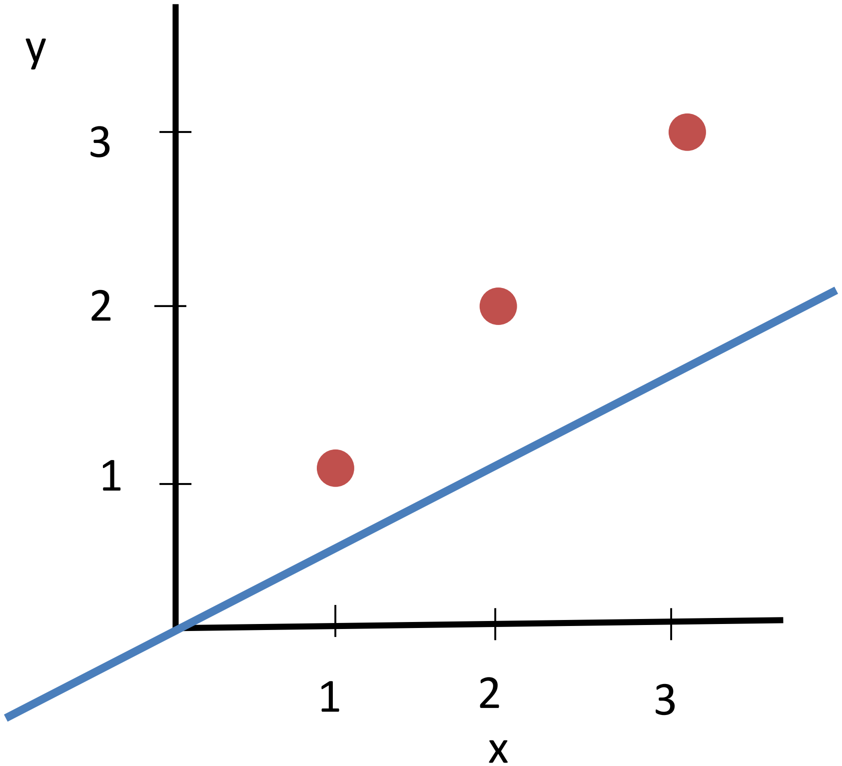
$$c(\lambda_1) = \frac{1}{2m}\sum_{i=0}^{m}((h(x^i) - y^i))^2$$

$$minimise_{\lambda_1}\ C(\lambda_1)$$

Over the next few slides we will pick a few different value for λ_1 and we will monitor the corresponding value of the cost function. The following is our simple dataset:

X	Y
1	1
2	2
3	3

Function h(x)



$$h(x) = \lambda_1 x$$

X	Y
1	1
2	2
3	3

Let's examine what happens when I give $\boldsymbol{\lambda_1}$ **a value of 0.5**.

We subsequent calculate the associated cost:

$$c(\lambda_1) = \frac{1}{2m}\sum_{i=0}^{m}((h(x^i) - y^i))^2$$

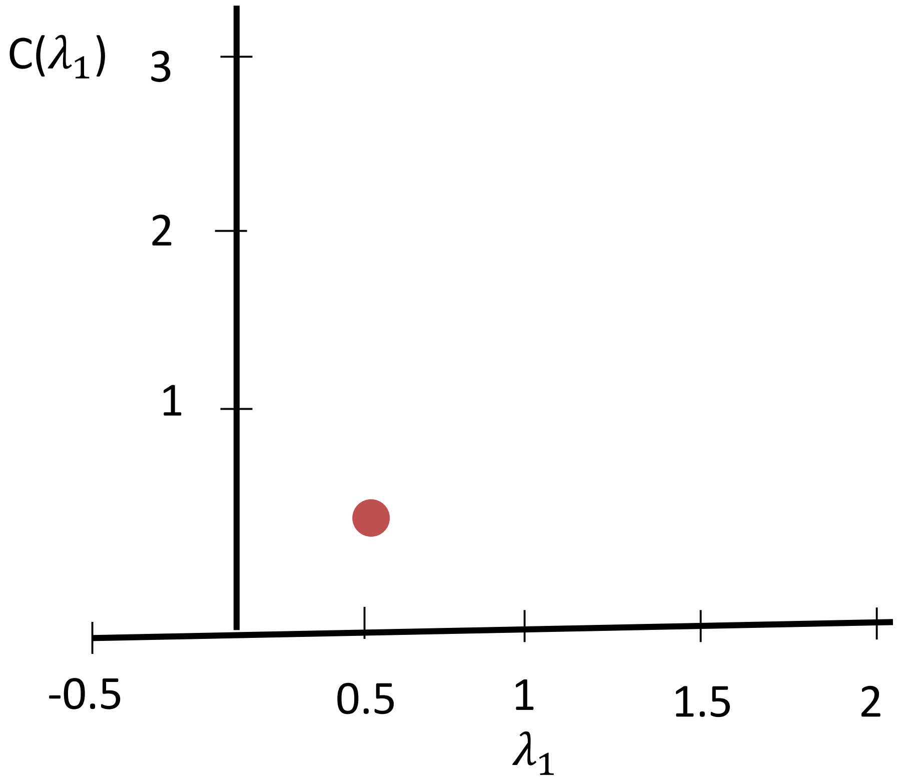
$$\frac{1}{6}(\,(0.5-1)^2 + (1-2)^2 + (1.5-3)\,^2)$$

$$\frac{1}{6}(\,0.25 + 1 + 2.25)$$

$$= 0.58$$

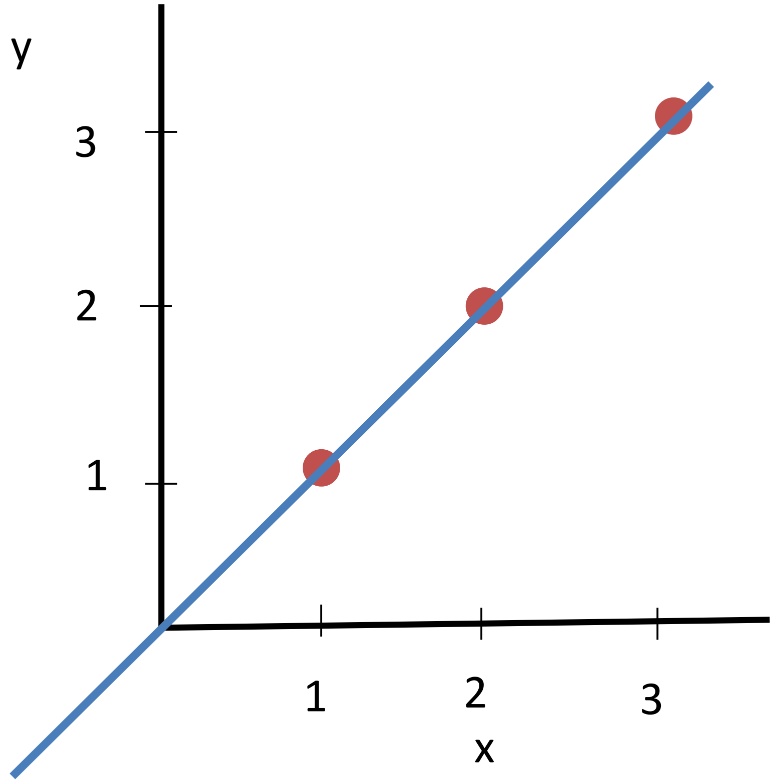
Notice we define our line only using a single parameter λ_1 . Controls the slope and must pass through the origin.

Function $\mathbf{C(\lambda_1)}$



We are going to pick multiple values for λ_1 and map the relationship between $C(\lambda_1)$ and λ_1 . When $\lambda_1 = 0.5$ then $C(\lambda_1)$ = 0.58

Function h(x)



$$h(x) = \lambda_1 x$$

X	Y
1	1
2	2
3	3

Now we examine what happens when I give $\boldsymbol{\lambda_1}$ **a value of 1**.

This provides an excellent fit to the data.

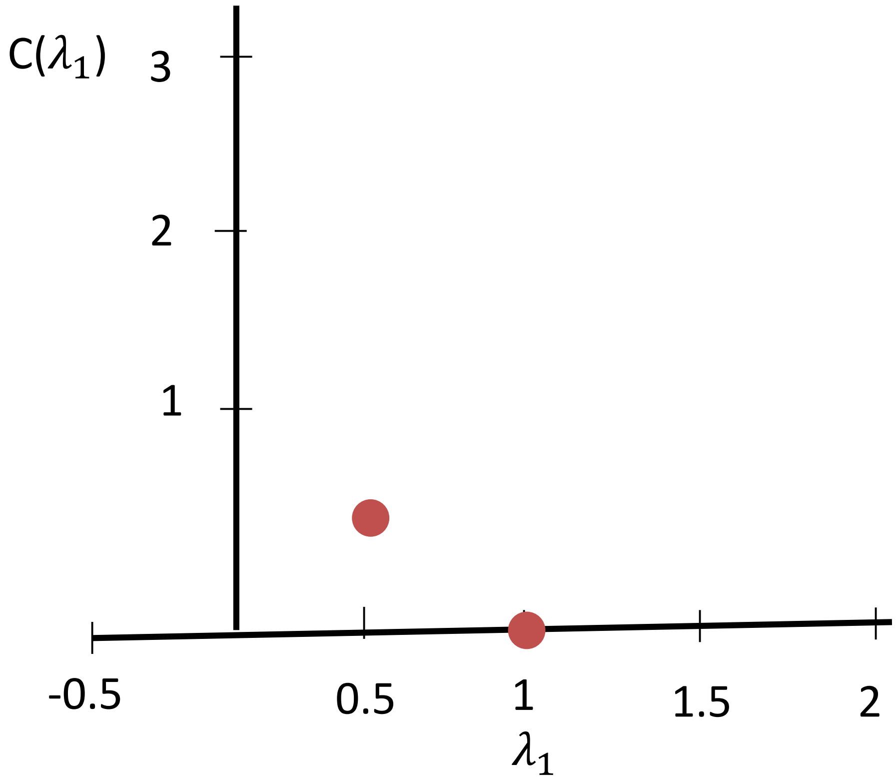
$$c(\lambda_1) = \frac{1}{2m}\sum_{i=0}^{m}((h(x^i) - y^i))^2$$

$$\frac{1}{6}(\ (1-1)^2 + (2-2)^2 + (3-3)\ ^2)$$

$$\frac{1}{6}(\ 0)$$

$$= 0$$

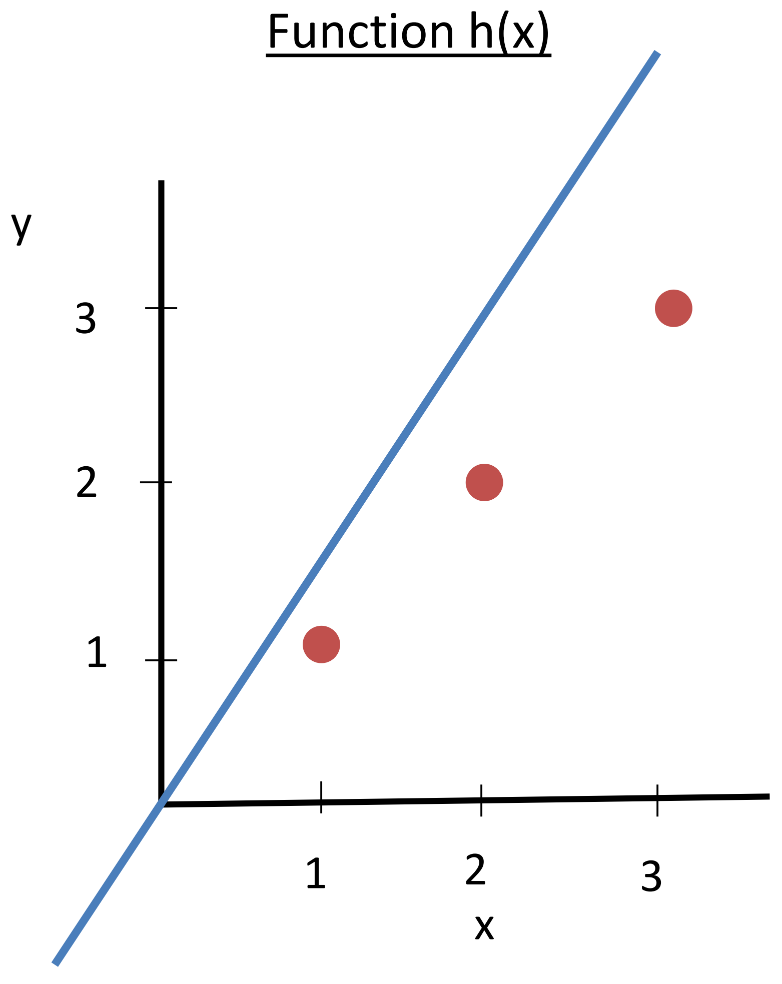
Function **C(λ_1)**



We are going to pick multiple values for λ_1 and map the relationship between C(λ_1) and λ_1 . When $\lambda_1 = 1$ then C(λ_1) = 0

$$h(x) = \lambda_1 x$$

X	Y
1	1
2	2
3	3



Let's examine what happens when I give $\boldsymbol{\lambda_1}$ **a value of 1.5**.

We subsequent calculate the associated cost:

$$c(\lambda_1) = \frac{1}{2m}\sum_{i=0}^{m}((h(x^i) - y^i))^2$$

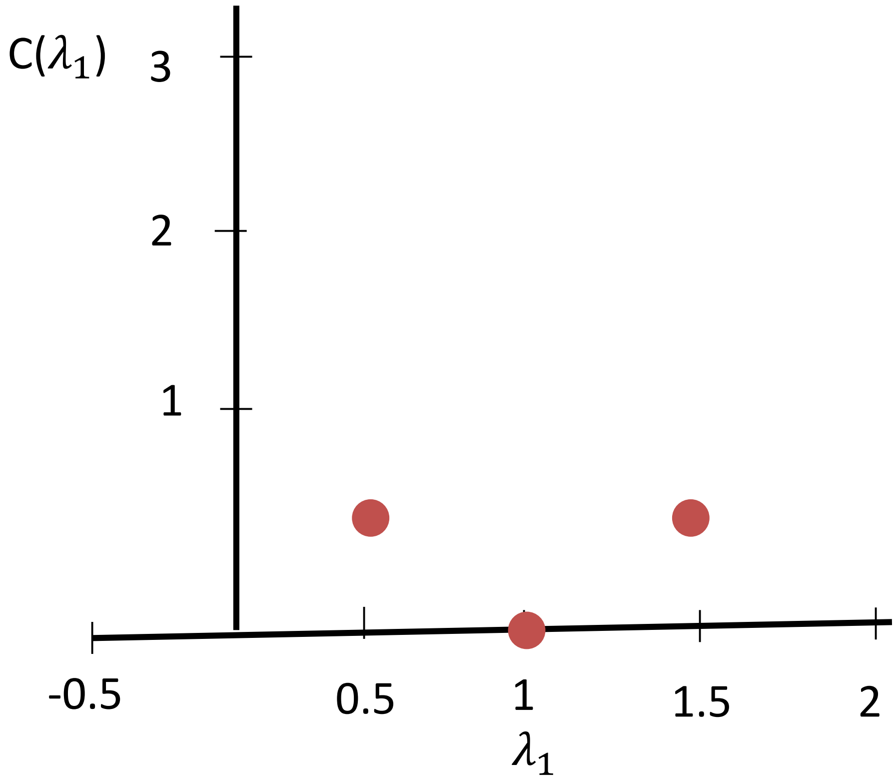
$$\frac{1}{6}(\,(1.5-1)^2 + (3-2)^2 + (4.5-3)^{\,2})$$

$$\frac{1}{6}(\,0.25 + 1 + 2.25)$$

$$= 0.58$$

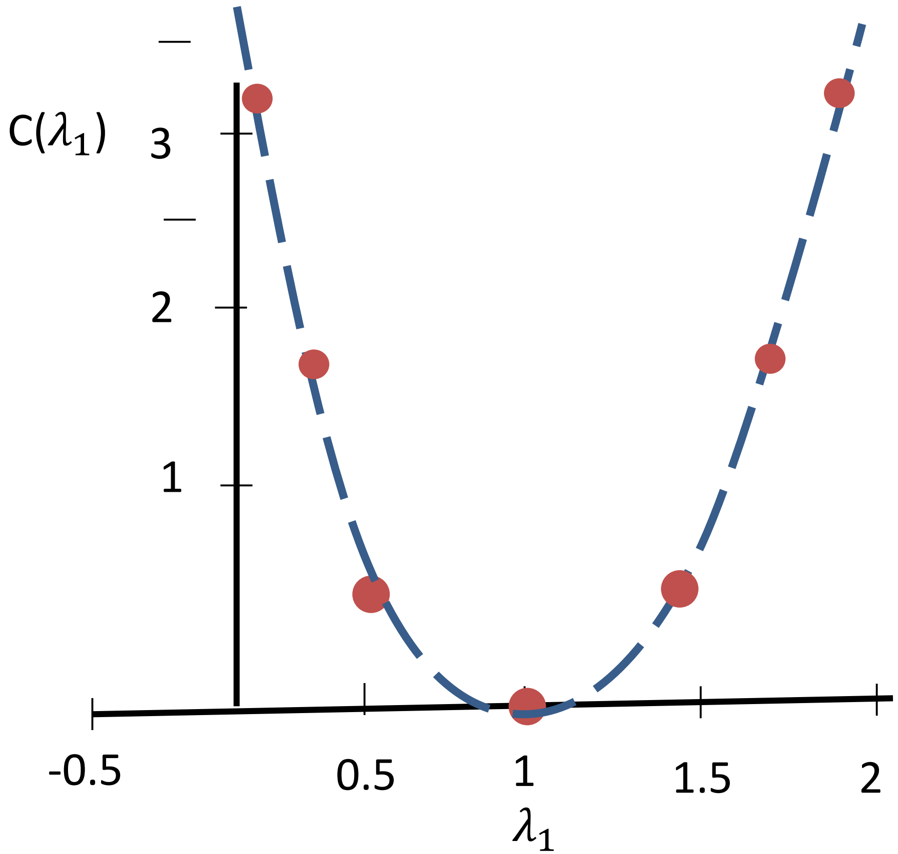
Notice we define our line only using a single parameter λ_1 . Controls the slope and must pass through the origin.

Function $\mathbf{C(\lambda_1)}$



We are going to pick multiple values for λ_1 and map the relationship between $C(\lambda_1)$ and λ_1 . When $\lambda_1 = 1$ then $C(\lambda_1)$ = 0

Function $\mathbf{C(\lambda_1)}$

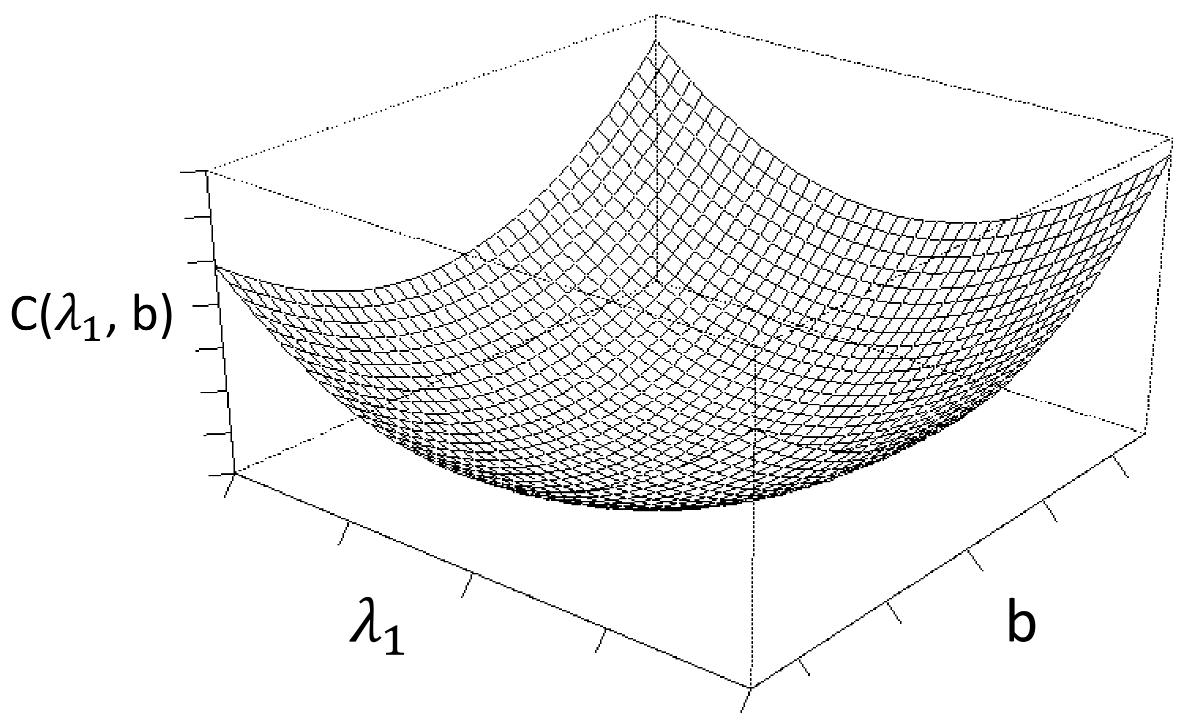


If we were to continue this process and map the shape of $C(\lambda_1)$ then it would exhibit a regular **convex** shape as shown above.

$h(x) = \lambda_1 x + b$

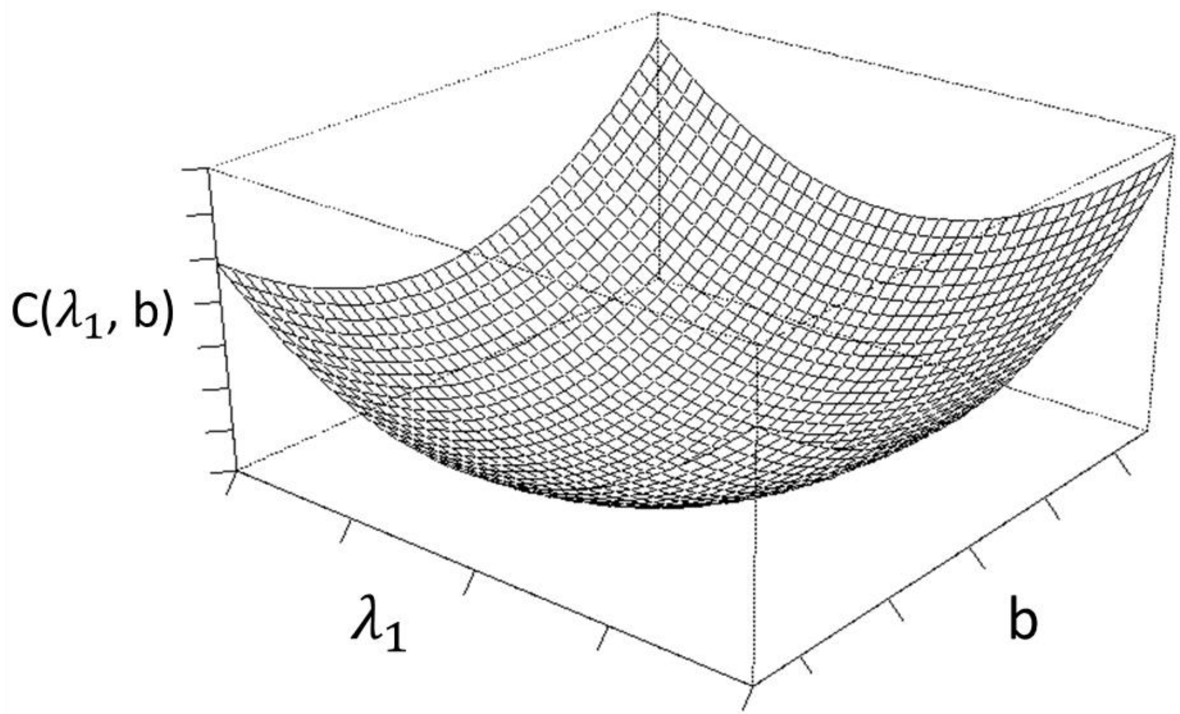
Linear Regression – A Search Problem

- Now let's add the our additional parameter *b* back into our linear equation
- Below is an example **depiction of** $\mathrm{c}(\lambda_1)$ when we have two parameters ($\underline{\lambda_1}$ and *b*)



Gradient Decent – An Optimization Algorithm

- We have a function **C($\boldsymbol{\lambda_1}$, $\boldsymbol{b}$)** and our objective is to determine the values of λ_1 and b that will give the **minimum** value of C(λ_1, b)
- Gradient Decent (Overview)
  - Start with a **random** value of λ_1 and b
  - Alter the value of $\lambda_1$and b in order to continually reduce the value of C(λ_1, b)
  - Continue until we reach a minimum



Gradient Decent – Algorithm

repeat {

$$\lambda_1 = \lambda_1 - \alpha \frac{\partial}{\partial \lambda_1} \mathrm{C}(\lambda_1, b)$$

$$b = b - \alpha \frac{\partial}{\partial b} \mathrm{C}(\lambda_1, b)$$

}

Where

- α is a numerical value that is called the **<u>learning rate</u>**. It controls the size of the decent that we make during each iteration. The larger the value of the greater the value of α
- $\frac{\partial}{\partial \lambda_1}$ and $\frac{\partial}{\partial b}$ are **<u>derivative term</u>** allowing us to calculate the slope of any point for the function

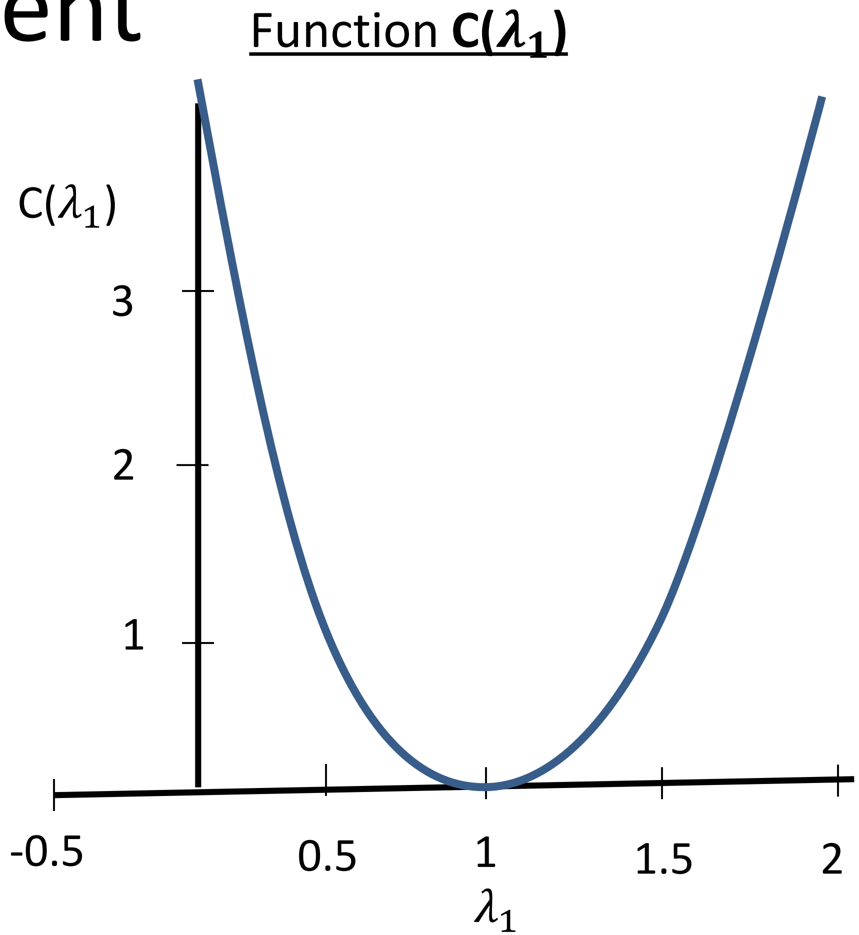
More Detail on Parameters of Gradient Decent

- Lets' return to our simple example from earlier where we have

  h(x) = λ_1 *x*

- And our objective is:

$$minimise_{\lambda_1}\ C(\lambda_1)$$



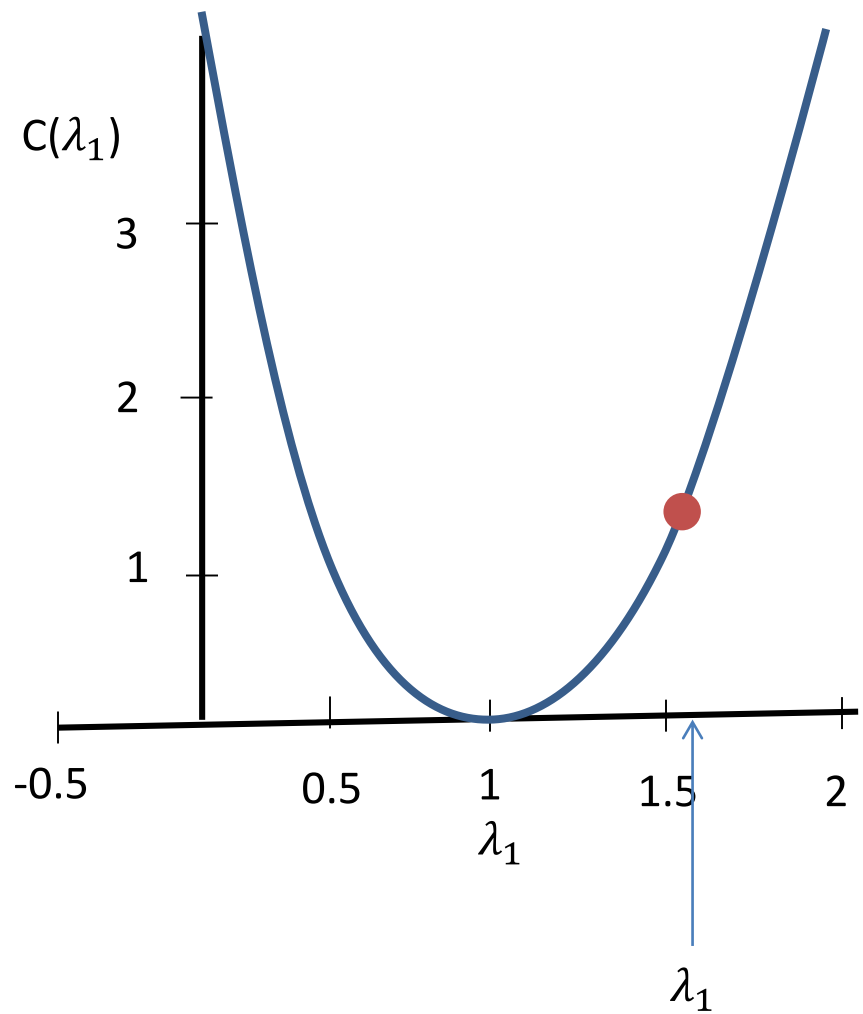
More Detail on Parameters of Gradient Decent

- Our rule updating λ_1 according to the Grad. Dec. algorithm is:

$$\lambda_1 = \lambda_1 - \alpha \frac{\partial}{\partial \lambda_1} C(\lambda_1)$$

- We randomly pick an initial value for λ_1 as seen in the graph.

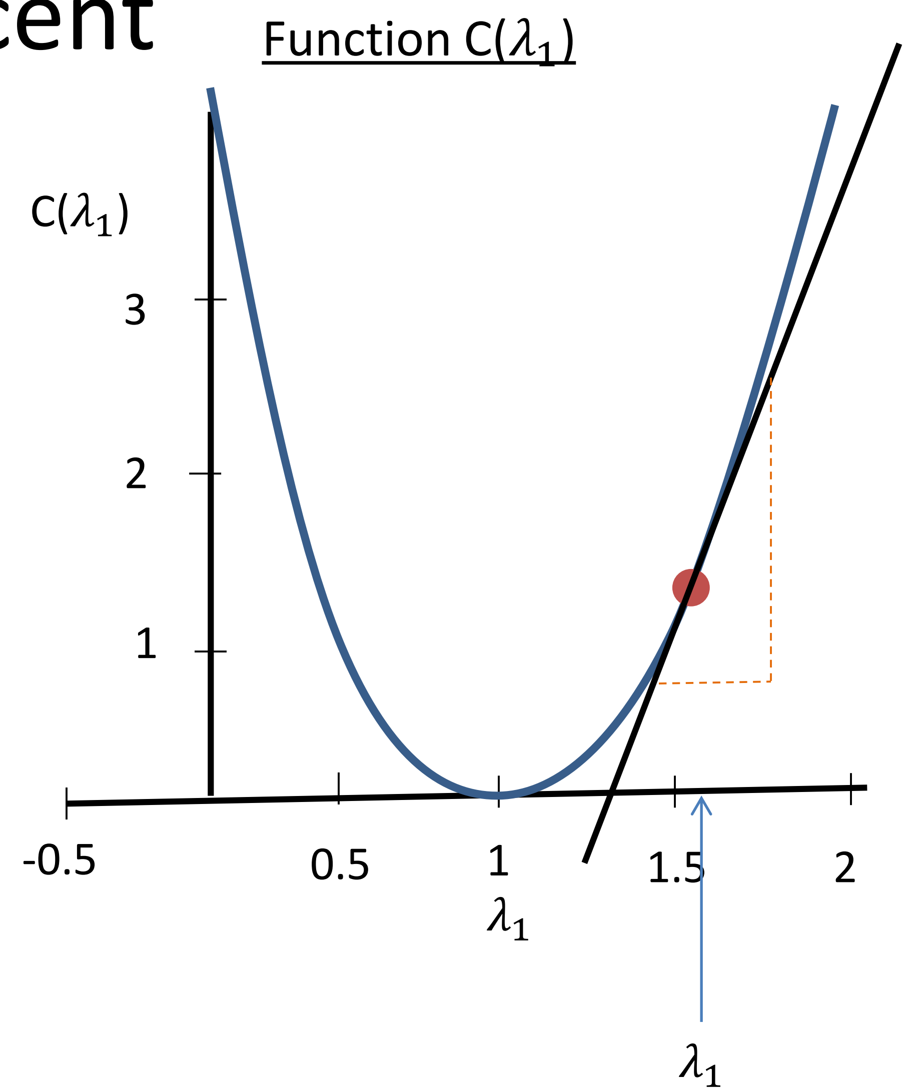
Function $C(\lambda_1)$



More Detail on Parameters of Gradient Decent

$$\lambda_1 = \lambda_1 - \alpha \boxed{\frac{\partial}{\partial \lambda_1} C(\lambda_1)}$$

This derivative term allows us to calculate **the slope of a line that forms a tangent to the point** we have selected. We can see the slope of this line is a positive number (slope obviously being height divided by horizontal – dashed lines)

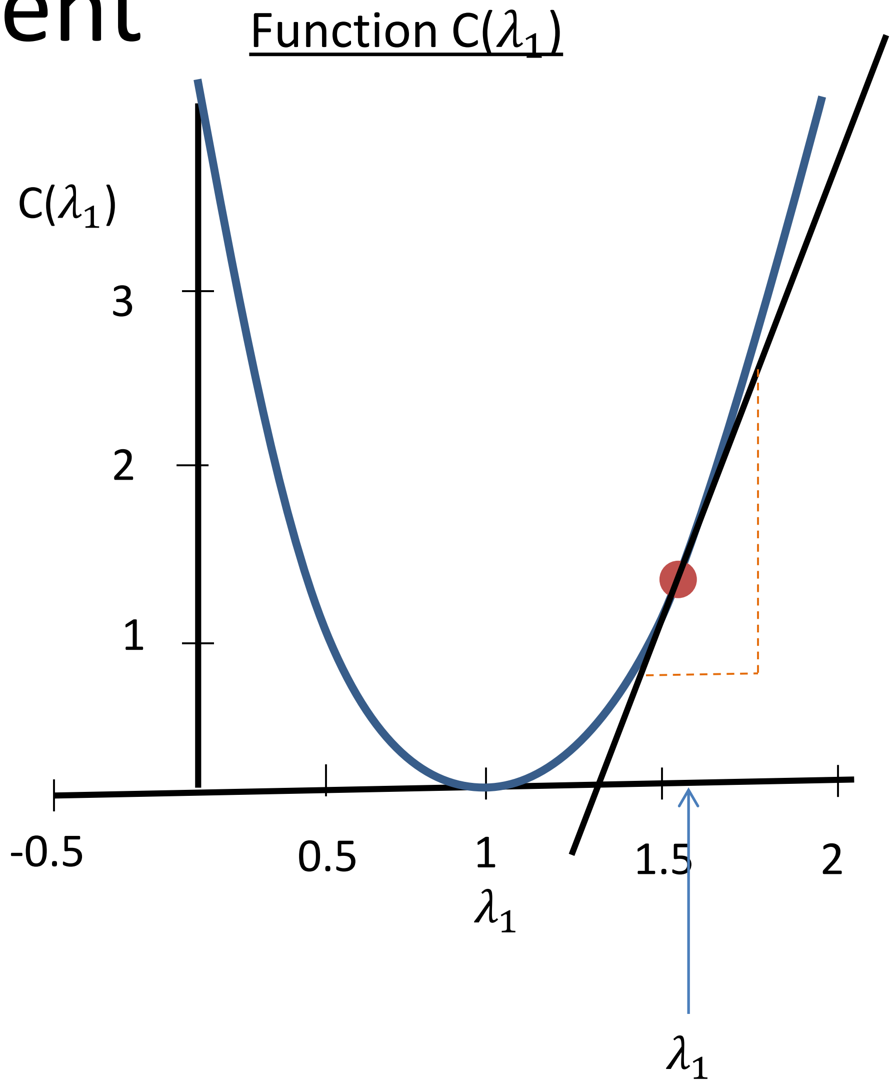


More Detail on Parameters of Gradient Decent

$$\lambda_1 = \lambda_1 - \alpha \boxed{\frac{\partial}{\partial \lambda_1} C(\lambda_1)}$$

The derivative value in this case will be a **positive number**, which is then multiplied by the α (the learning rate).

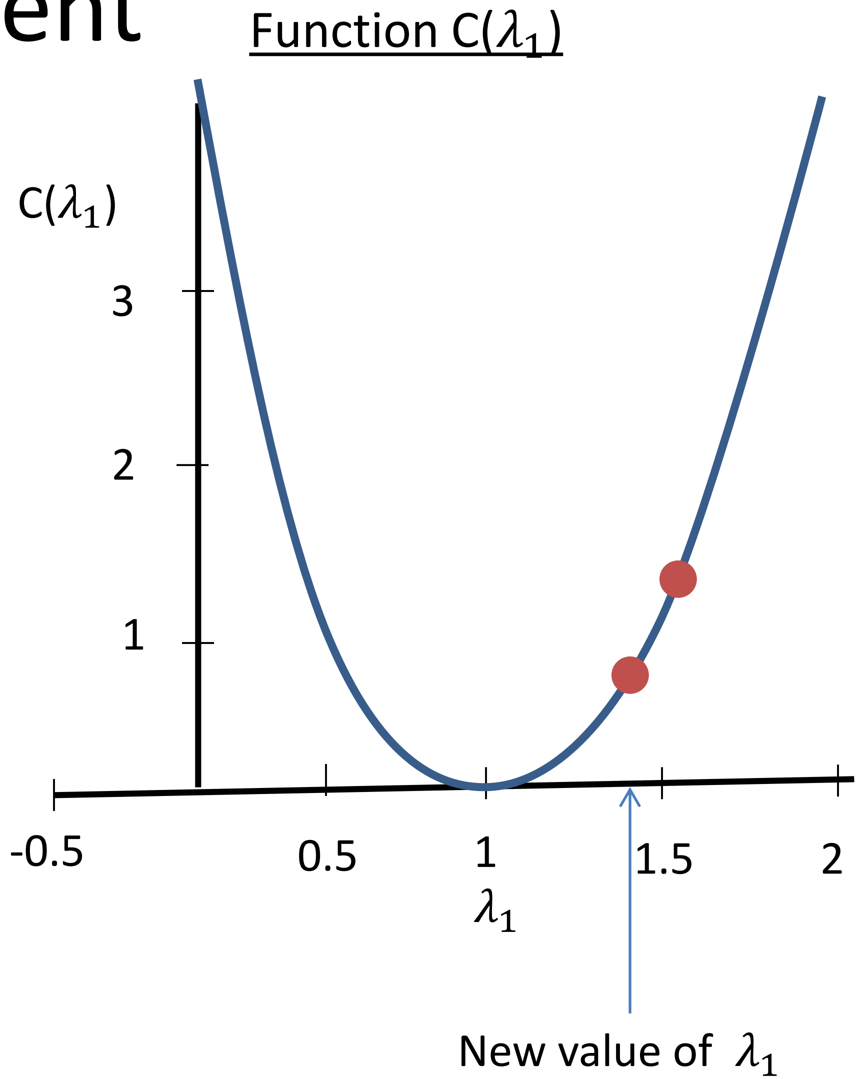
Therefore, this ultimately reduces the value of λ_1 as we are subtracting a small positive number.



More Detail on Parameters of Gradient Decent

$$\lambda_1 = \lambda_1 - \alpha \boxed{\frac{\partial}{\partial \lambda_1} C(\lambda_1)}$$

Notice the value of λ_1 is reduced

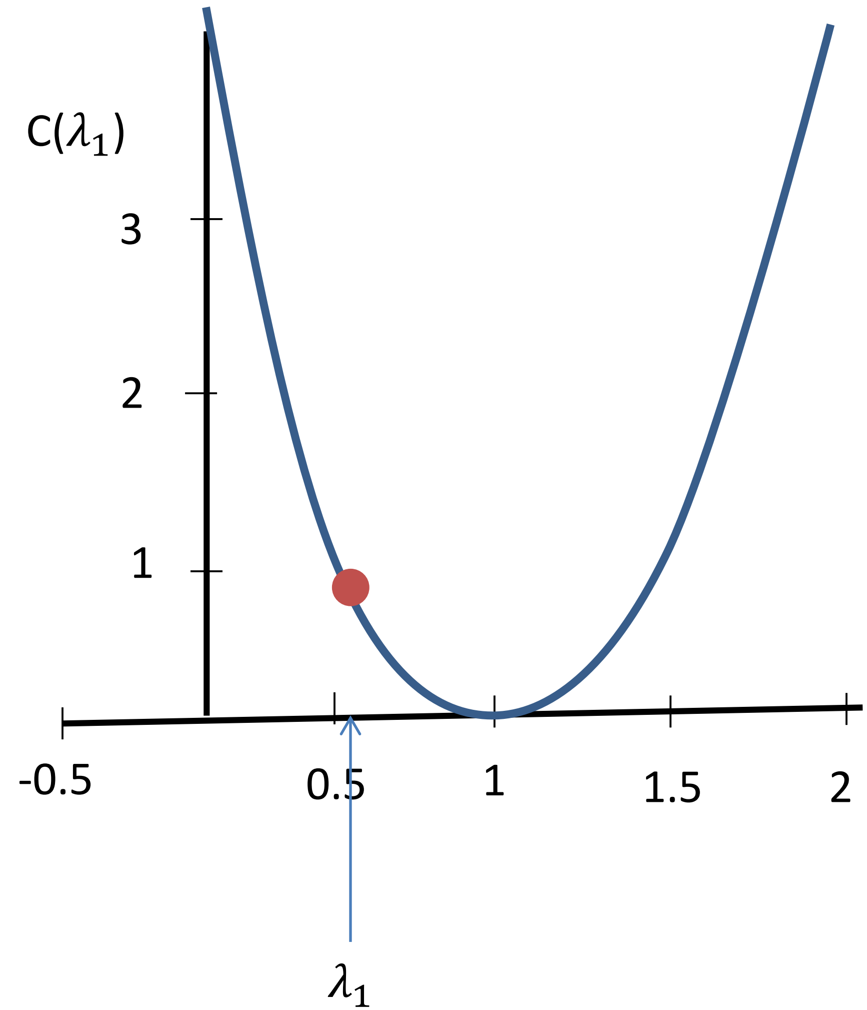


More Detail on Parameters of Gradient Decent

$$\lambda_1 = \lambda_1 - \alpha \boxed{\frac{\partial}{\partial \lambda_1} C(\lambda_1)}$$

What happens to the value of λ_1 if we select an initial value of λ_1 as shown in the graph?

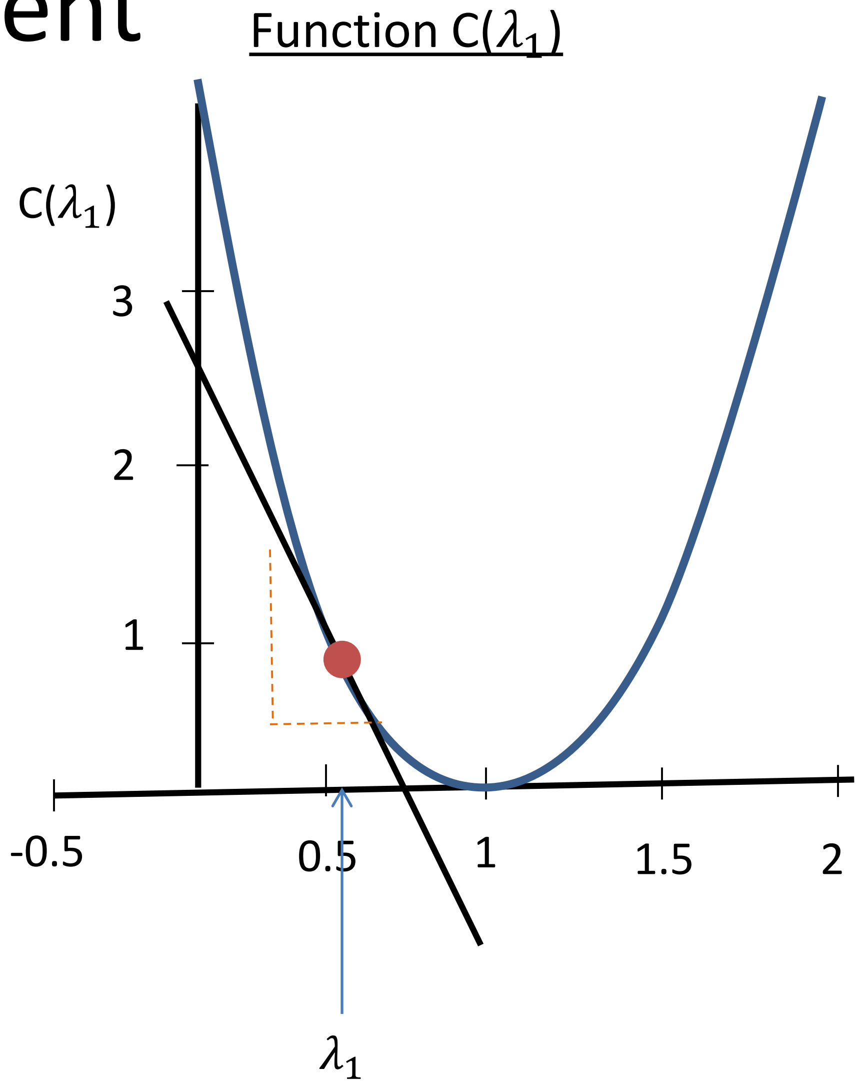
Function $C(\lambda_1)$



More Detail on Parameters of Gradient Decent

$$\lambda_1 = \lambda_1 - \alpha \boxed{\frac{\partial}{\partial \lambda_1} C(\lambda_1)}$$

What happens to the value of λ_1 if we select an initial value of λ_1 as shown in the graph?

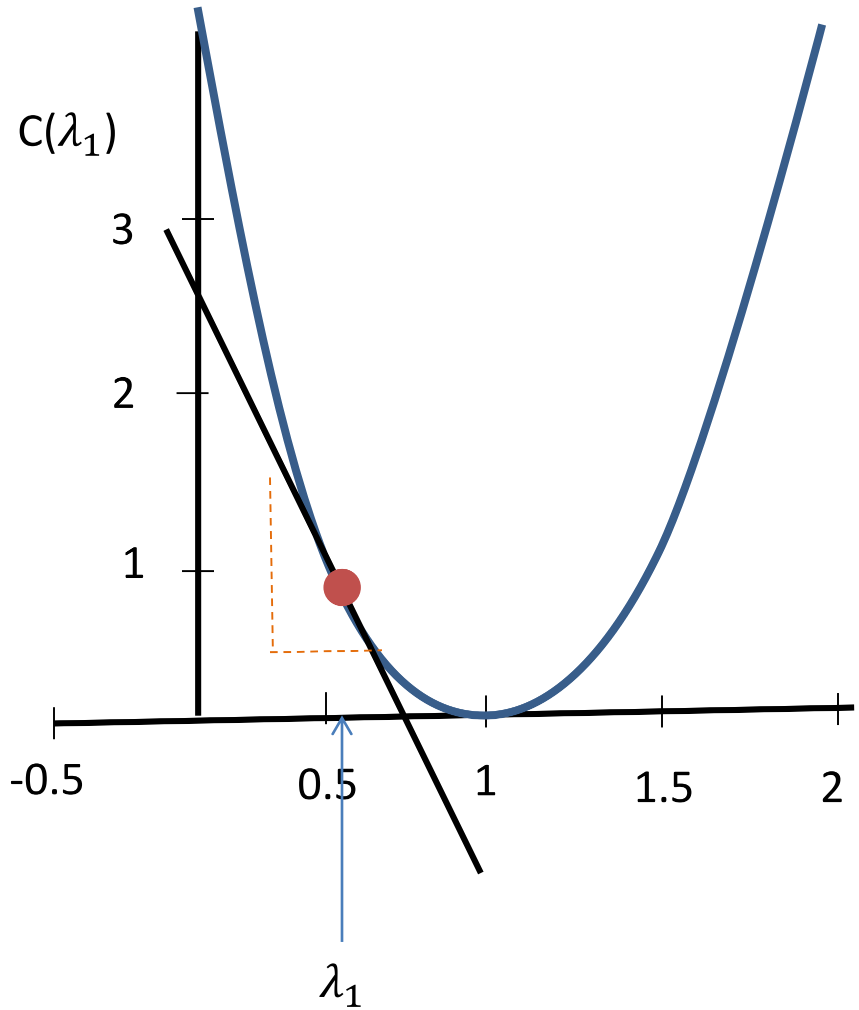


More Detail on Parameters of Gradient Decent

$$\lambda_1 = \lambda_1 - \alpha \boxed{\frac{\partial}{\partial \lambda_1} C(\lambda_1)}$$

The derivative of this line is negative. It is then multiplied by α and subtracted from λ_1. This has the effect of increasing the value of λ_1

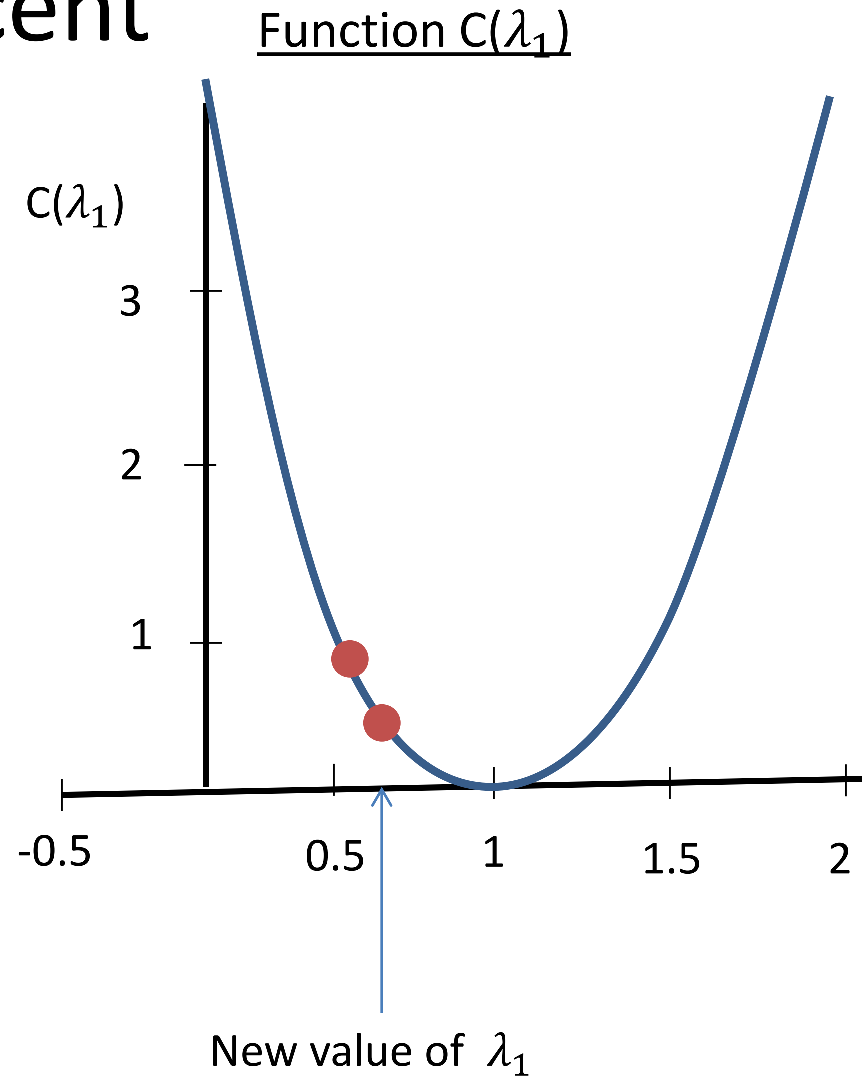
Function $C(\lambda_1)$



More Detail on Parameters of Gradient Decent

$$\lambda_1 = \lambda_1 - \alpha \boxed{\frac{\partial}{\partial \lambda_1} C(\lambda_1)}$$

The value of λ_1 is increased and is making it ways towards the minimum. Notice that we continue to move towards the minimum value of λ_1

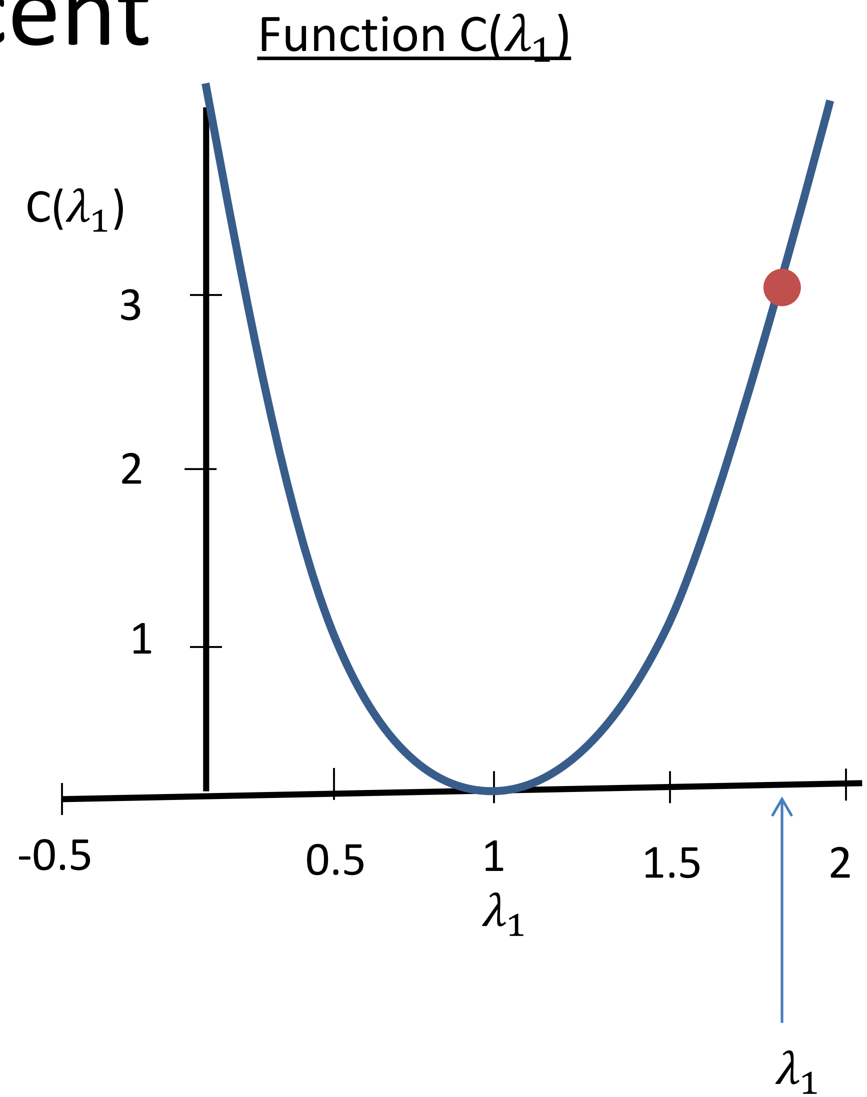


More Detail on Parameters of Gradient Decent

- As we mentioned before **the value of $\boldsymbol{\alpha}$** controls the rate of decent.

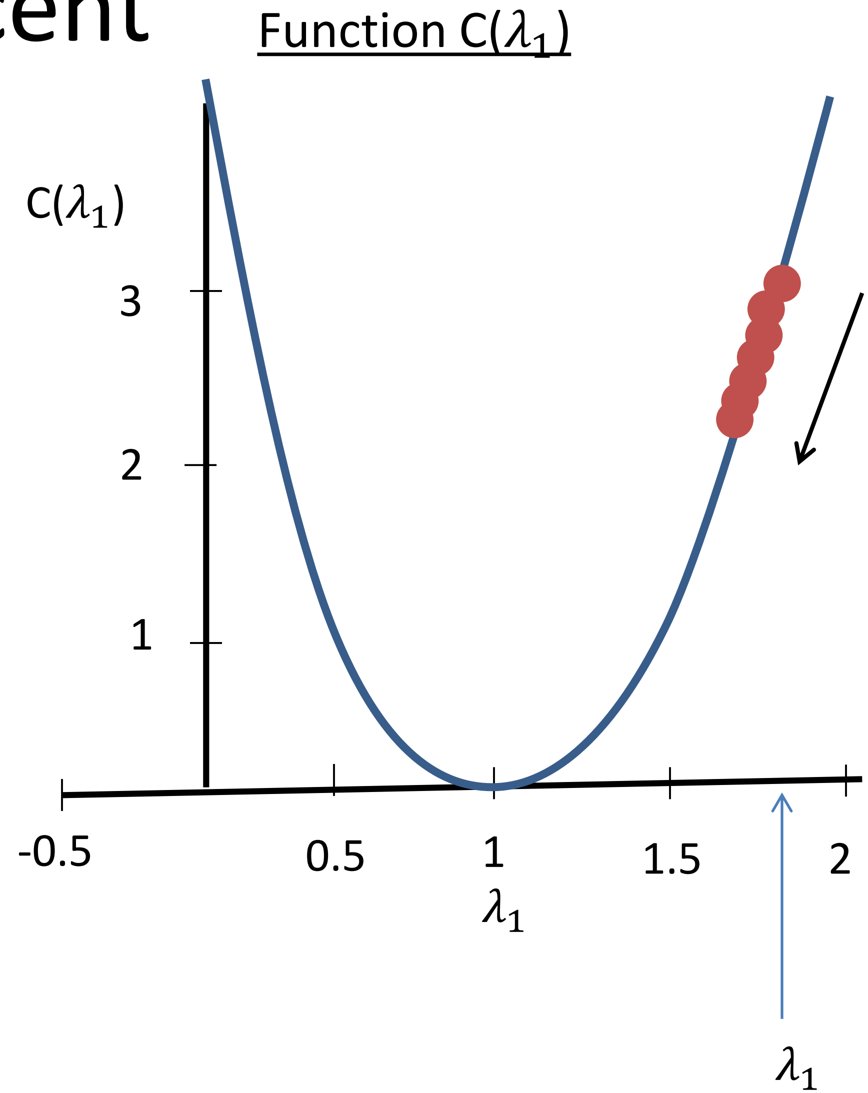
$$\lambda_1 = \lambda_1 - \alpha \frac{\partial}{\partial \lambda_1} C(\lambda_1)$$

- Consider the scenario where the value of α is very small. What effect do you think it would have on the update of λ_1



More Detail on Parameters of Gradient Decent

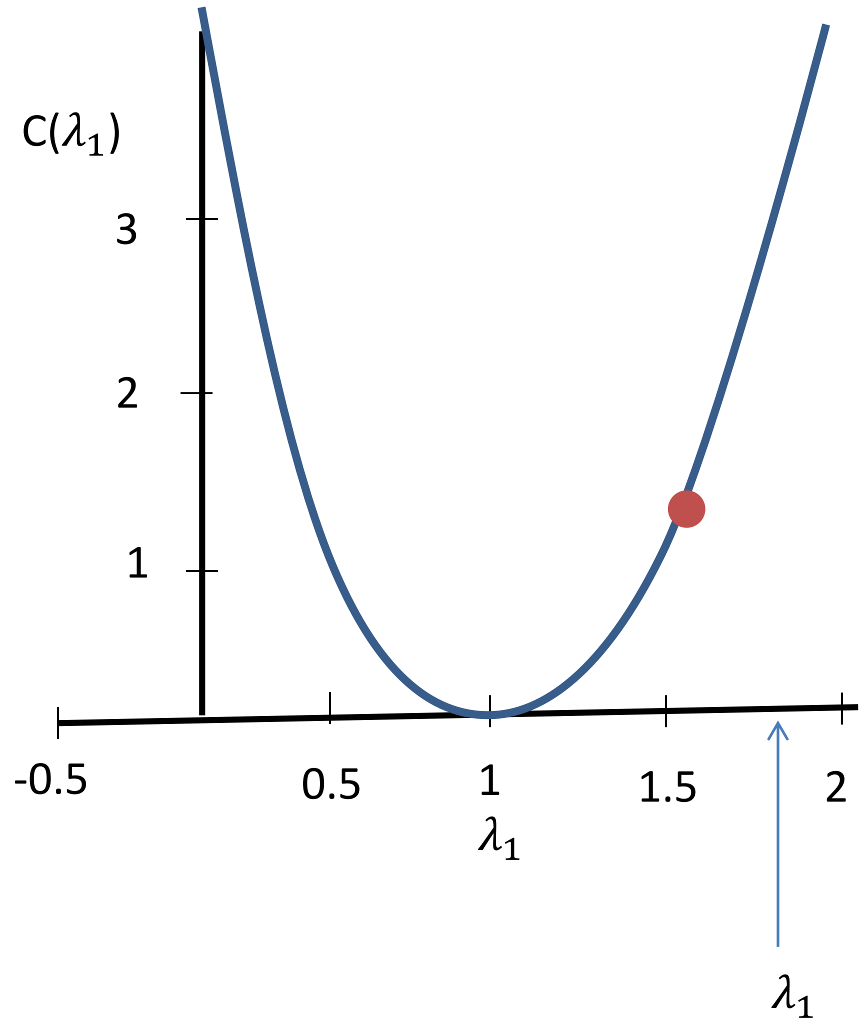
- This would result in very small movement in the value of λ_1 and would cause a long time until we reach convergence.



More Detail on Parameters of Gradient Decent

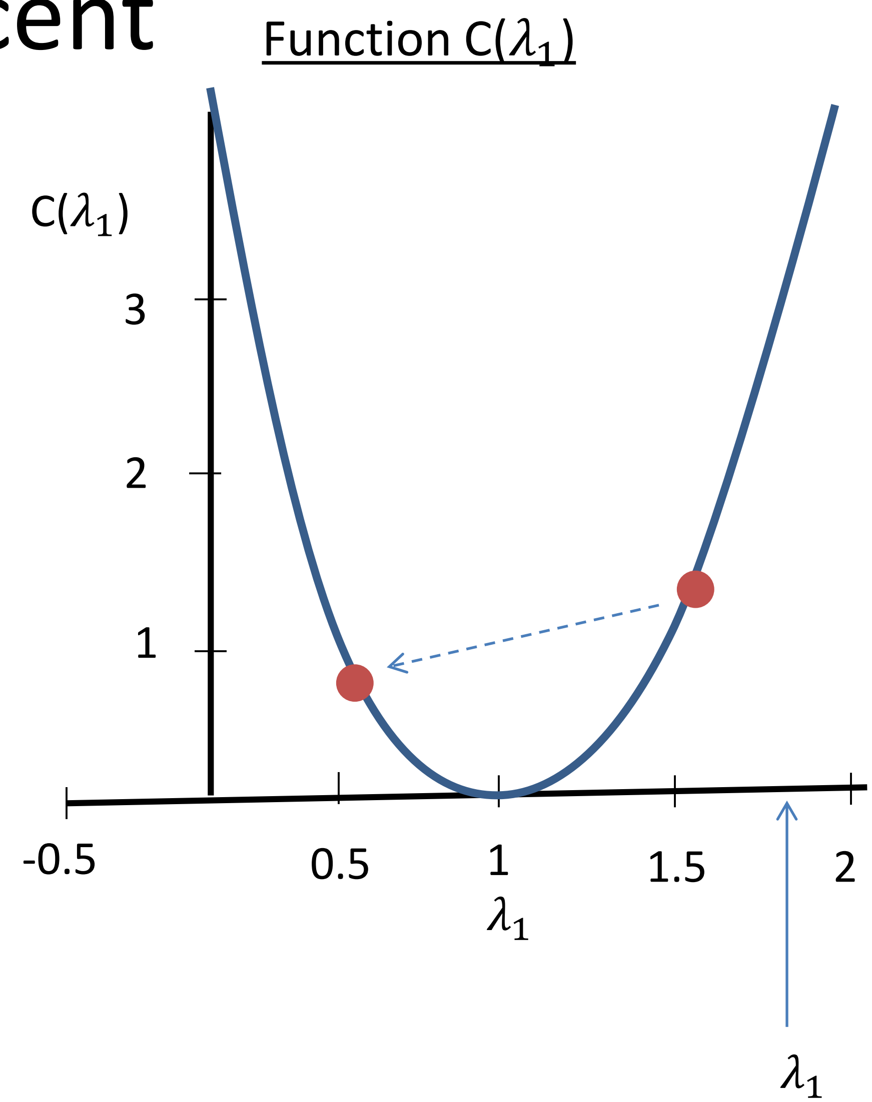
- Conversely if the value of α is very large then the change in value of λ_1 will be very substantial. Why would this be a problem?

Function $C(\lambda_1)$



More Detail on Parameters of Gradient Decent

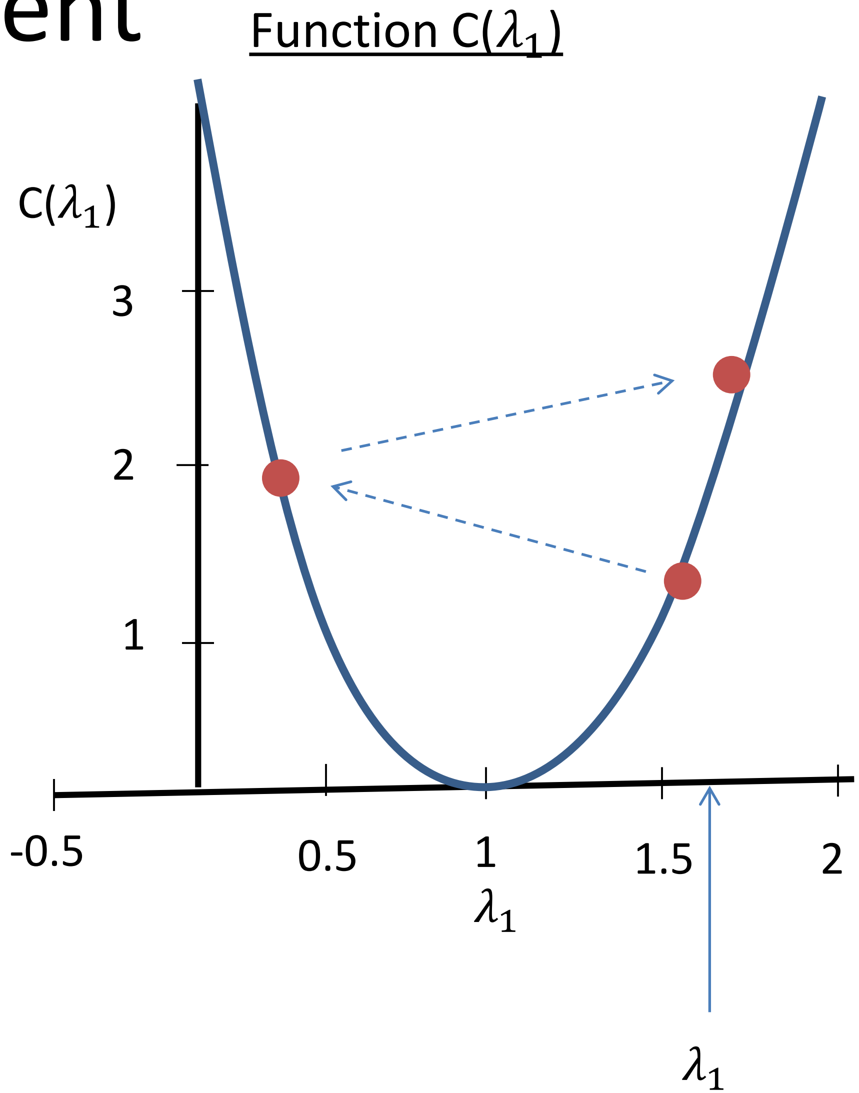
- It can cause very large jumps in the value of λ_1 and we could potentially miss the minimum as shown on the graph.



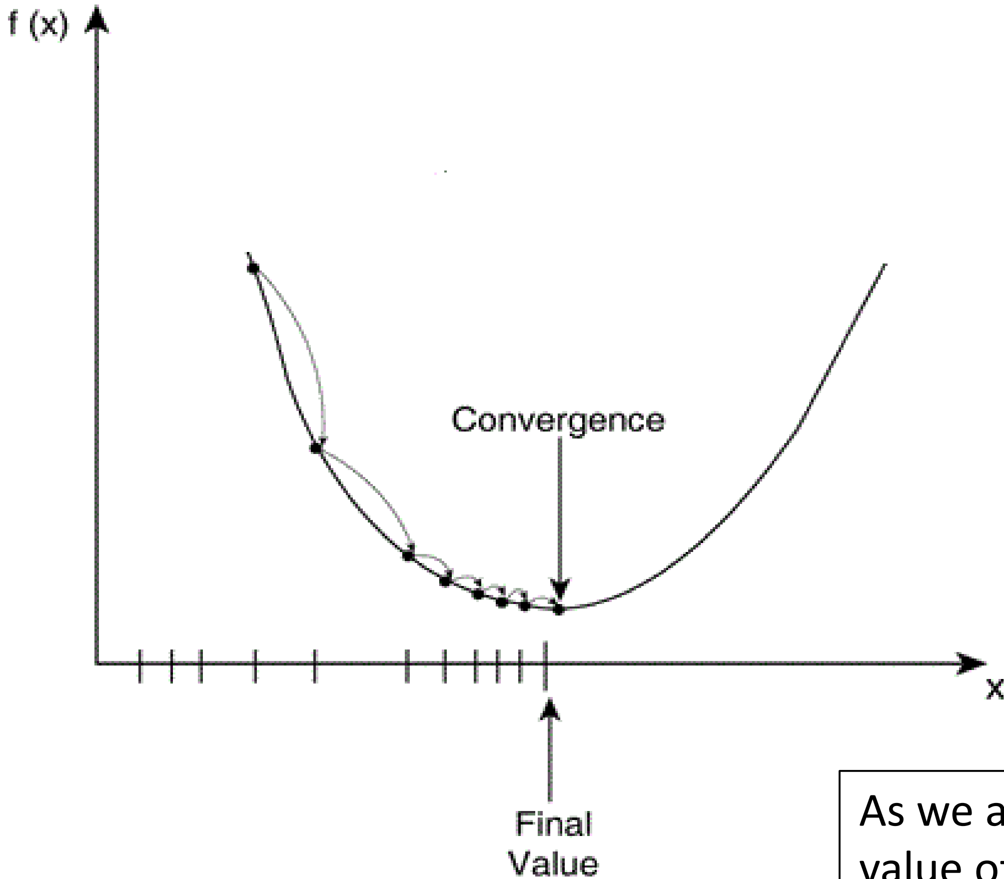
More Detail on Parameters of Gradient Decent

- It can cause very large jumps in the value of λ_1 and we could potentially miss the minimum as shown on the graph.

- **What do you think will happen if we pick λ_1 such that it has the minimum value?**



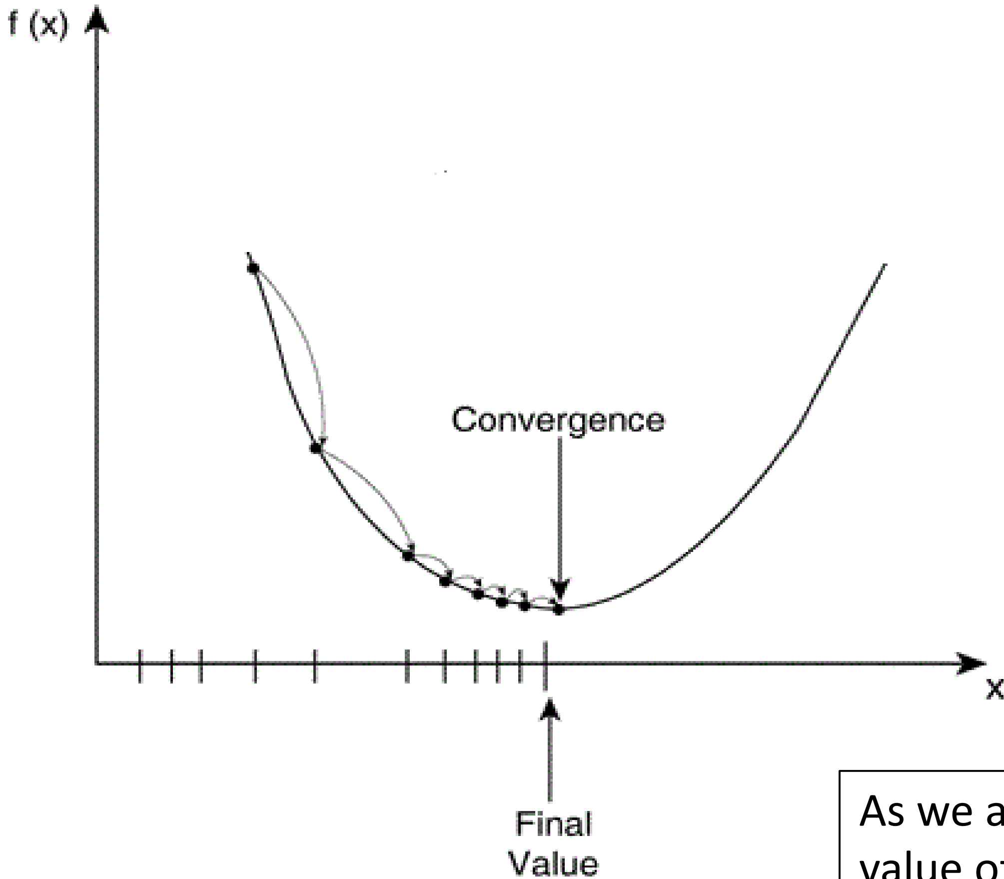
More Detail on Parameters of Gradient Decent



Can you explain why the size of the steps taken by gradient decent reduce as we move towards the minimum value in this example?

As we approach the local minimum the value of the derivative gets smaller and smaller (change in x with respect to y is smaller therefore the slope value is smaller)

More Detail on Parameters of Gradient Decent



Can you explain why the size of the steps taken by gradient decent reduce as we move towards the minimum value in this example?

As we approach the local minimum the value of the derivative gets smaller and smaller (change in x with respect to y is smaller therefore the slope value is smaller)

Derivatives for Linear Regression

$$\lambda_1 = \lambda_1 - \alpha \boxed{\frac{\partial}{\partial \lambda_1} C(\lambda_1, b)}$$

$$b = b - \alpha \boxed{\frac{\partial}{\partial b} C(\lambda_1, b)}$$

$$\frac{\partial}{\partial \lambda_1} C(\lambda_1,\ b) = \frac{1}{m} \sum_{i=0}^{m} ((h(x^i) - y^i))(x^i)$$

$$\frac{\partial}{\partial b} C(\lambda_1,\ b) = \frac{1}{m} \sum_{i=0}^{n} ((h(x^i) - y^i))$$

Gradient Decent Algorithm

repeat {

$$\lambda_1 = \lambda_1 - \alpha \frac{\partial}{\partial \lambda_1} \mathrm{C}(\lambda_1, b)$$

$$b = b - \alpha \frac{\partial}{\partial b} \mathrm{C}(\lambda_1, b)$$

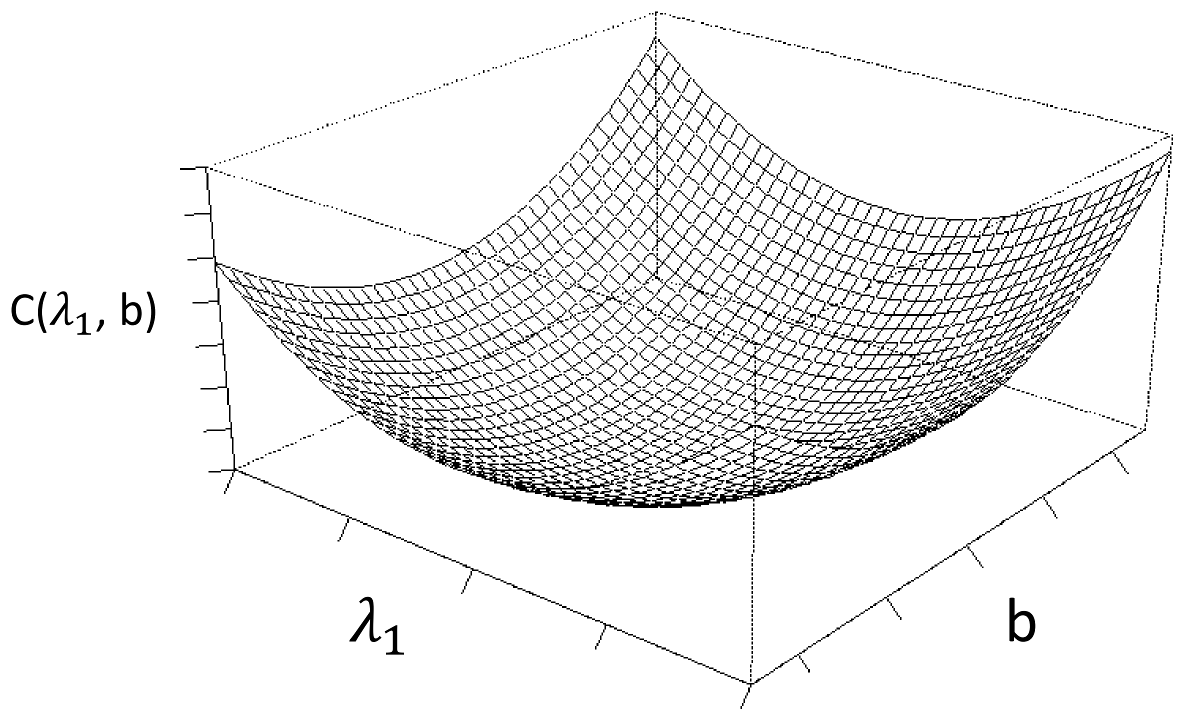
}

repeat {

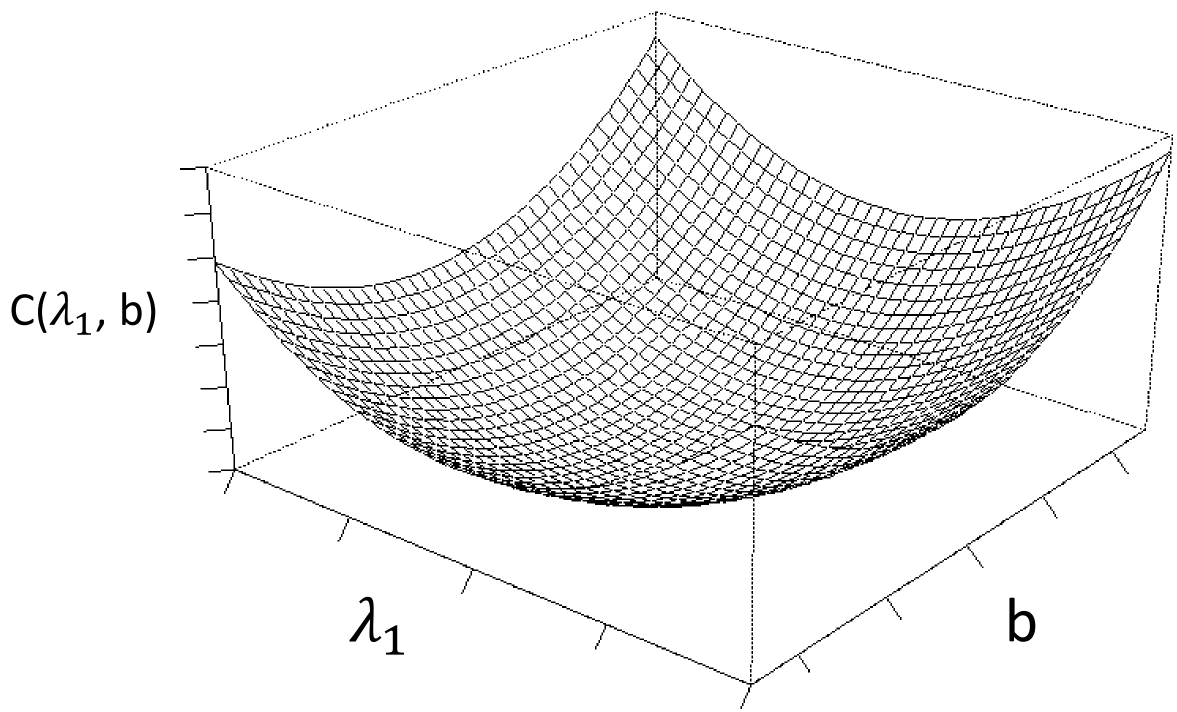
$$\lambda_1 = \lambda_1 - \alpha \frac{1}{m} \sum_{i=0}^{m} ((h(x^i) - y^i))(x^i)$$

$$b = b - \alpha \frac{1}{m} \sum_{i=0}^{m} ((h(x^i) - y^i))$$

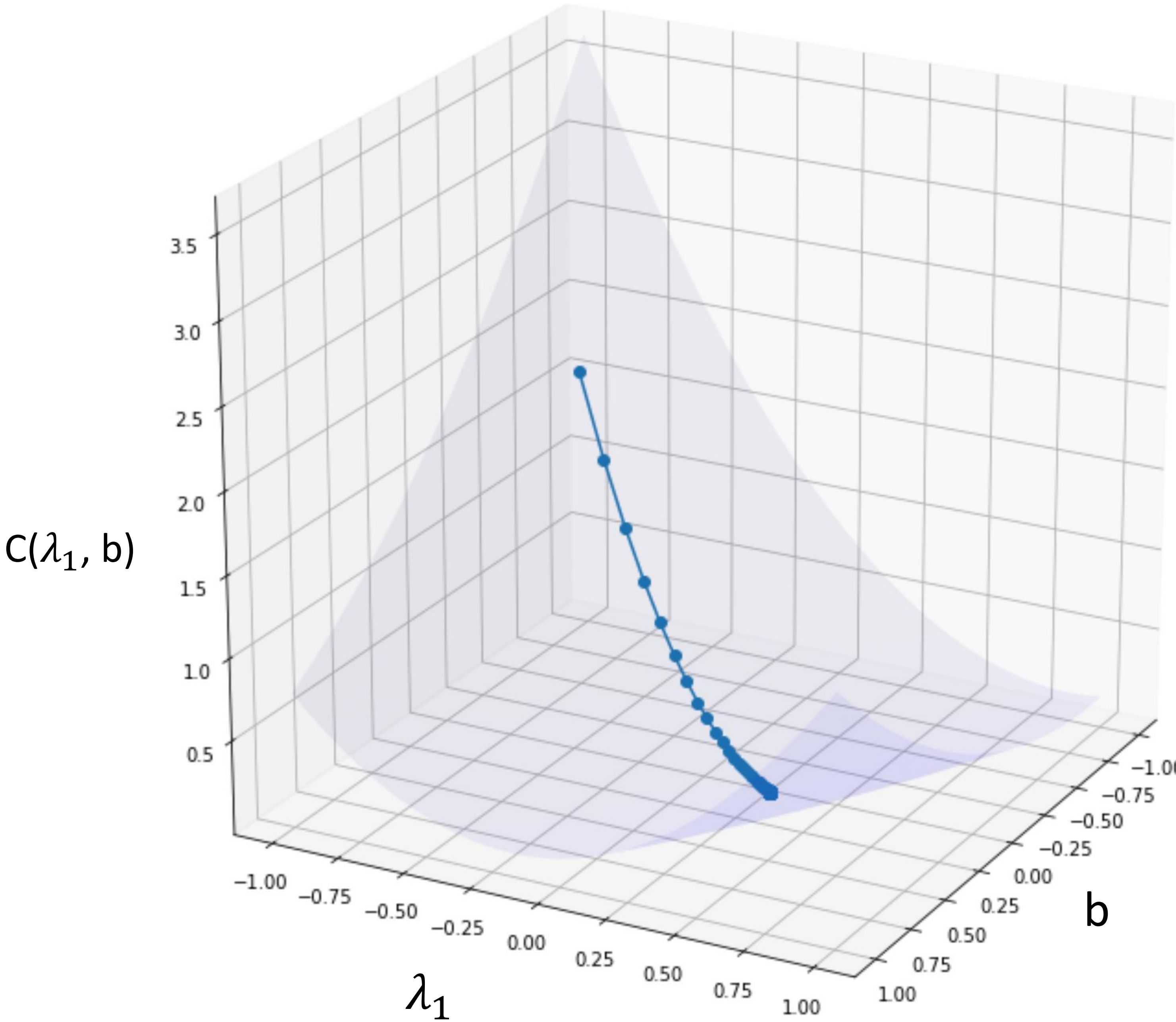
}



C(λ_1, b)
λ_1
b



C(λ_1, b)
λ_1
b



$C(\lambda_1, b)$
3.5
3.0
2.5
2.0
1.5
1.0
0.5
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
λ_1
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
b