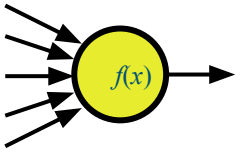
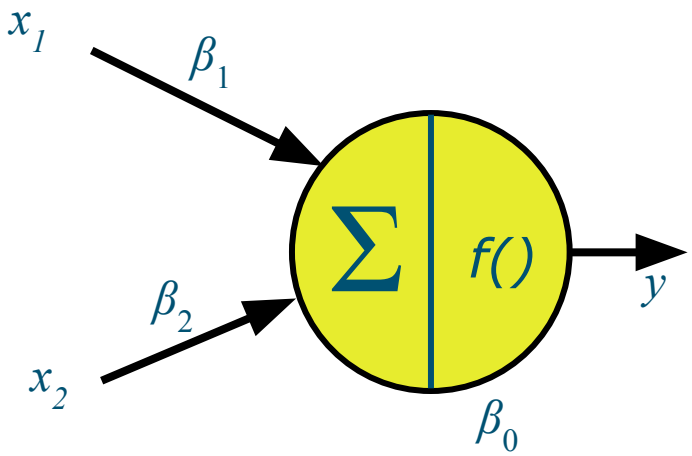
Intro to Deep Learning

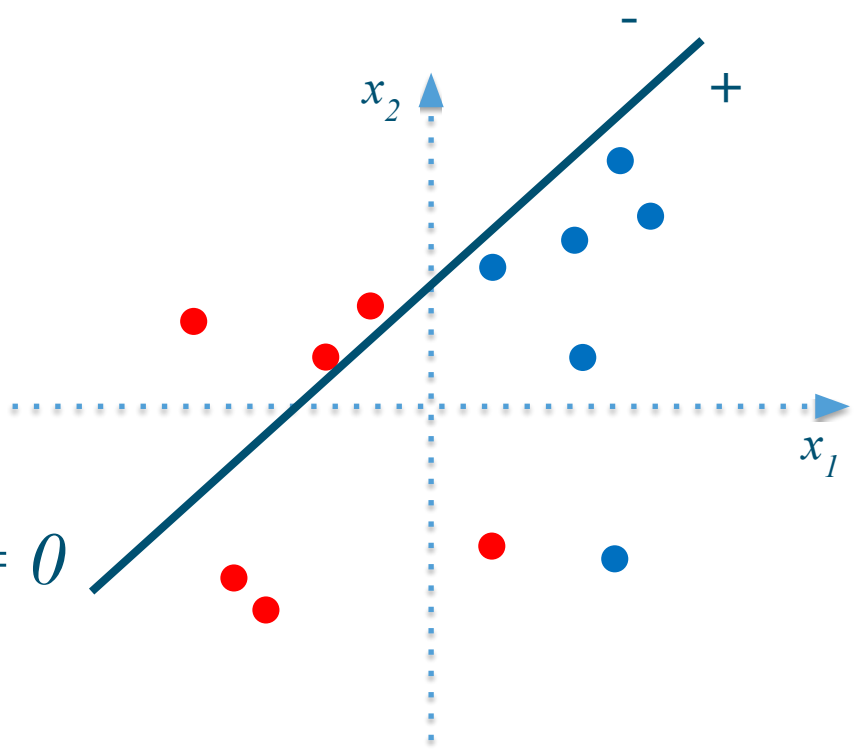
1) The neuron model
2) **Gradient descent**
3) Architectures (mostly for vision)
4) Model evaluation

Learning the weights of a single neuron



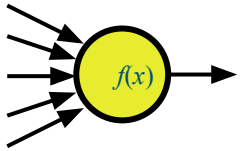
$$y = f(\beta_1 x_1 + \beta_2 x_2 + \beta_0)$$



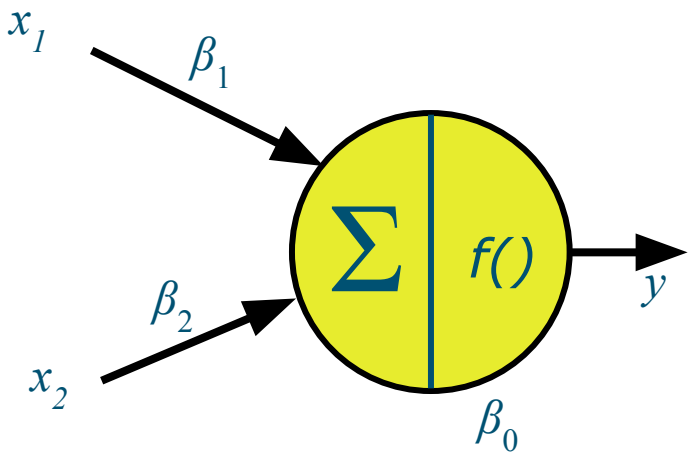


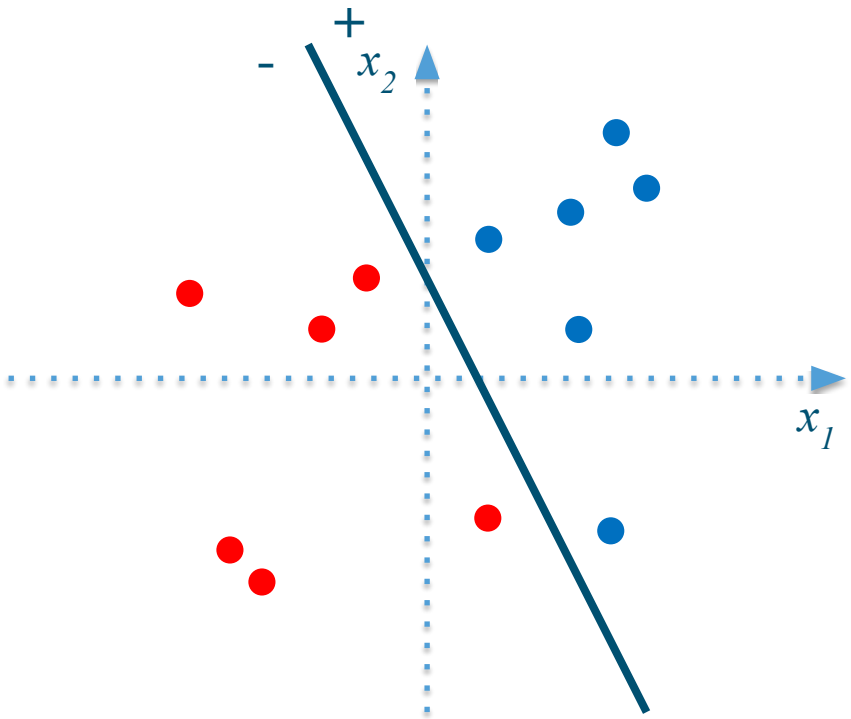
$$\beta_1 x_1 + \beta_2 x_2 + \beta_0 = 0$$

Learning the weights of a single neuron



$$y = f(\beta_1 x_1 + \beta_2 x_2 + \beta_0)$$





The big question about the weights

- A single neuron can solve any linear problem, given the right weights β.
- But how to find the right weights?
- We will use **gradient descent**

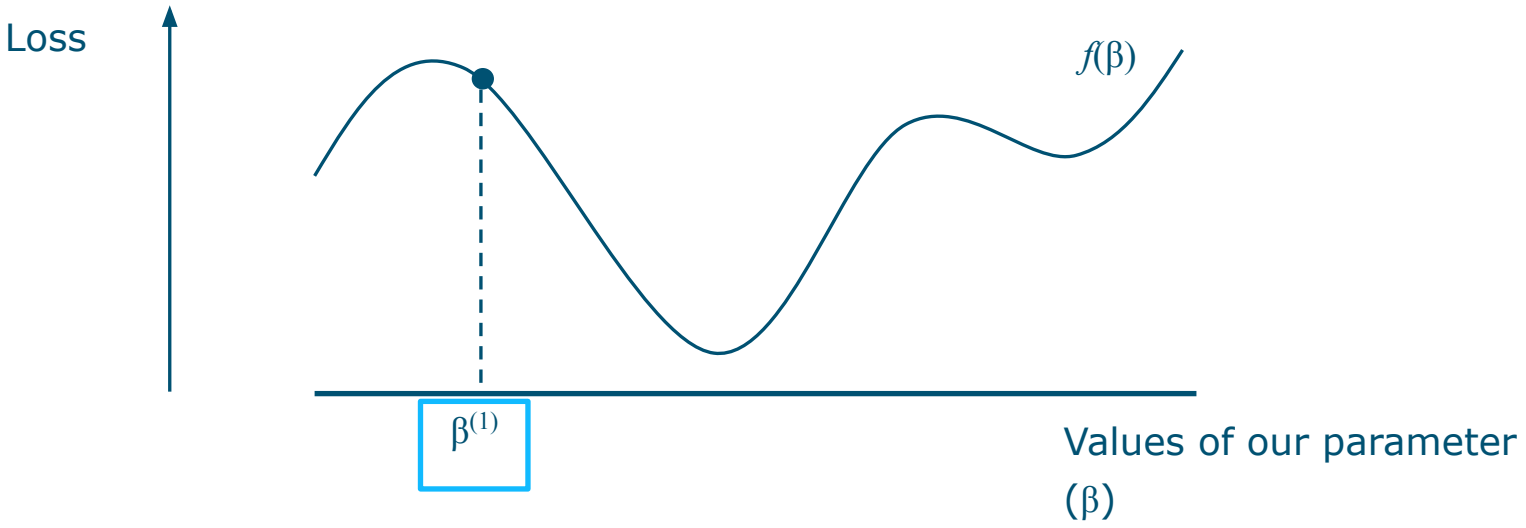
Gradient descent

We want to update a weight β to make the model better

$$\beta^{(r+1)} = \beta^{(r)} + ?$$

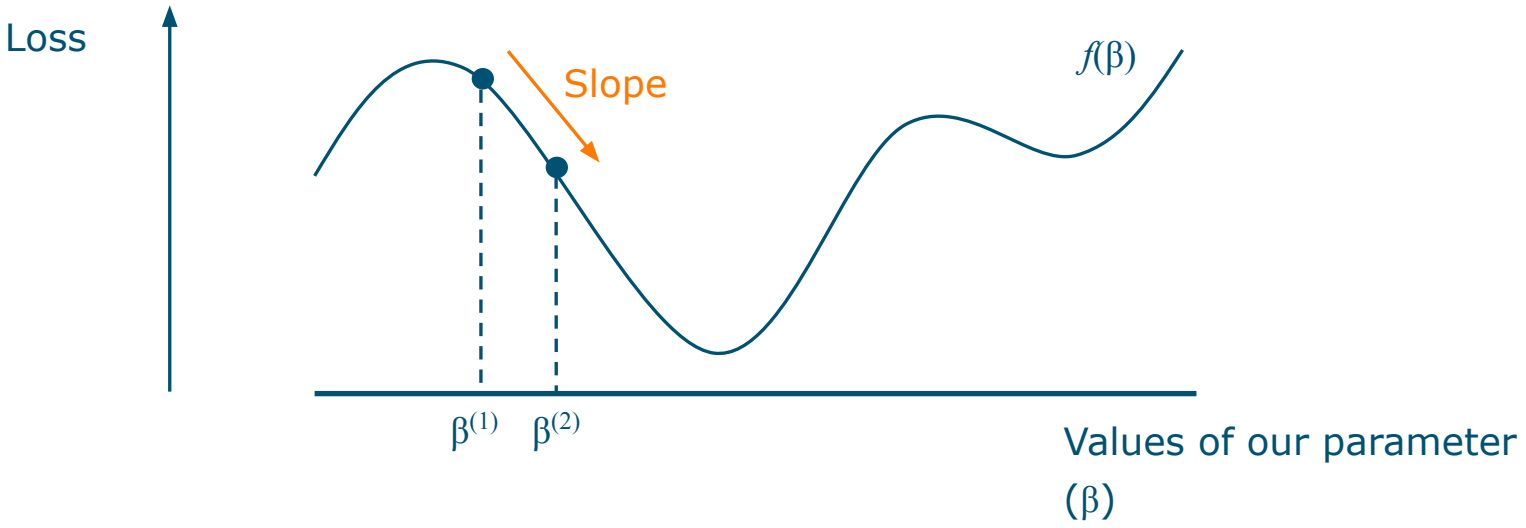
Gradient descent

- We want to find the minimum of an error function (or loss)
- We start with an initial guess of the parameter β



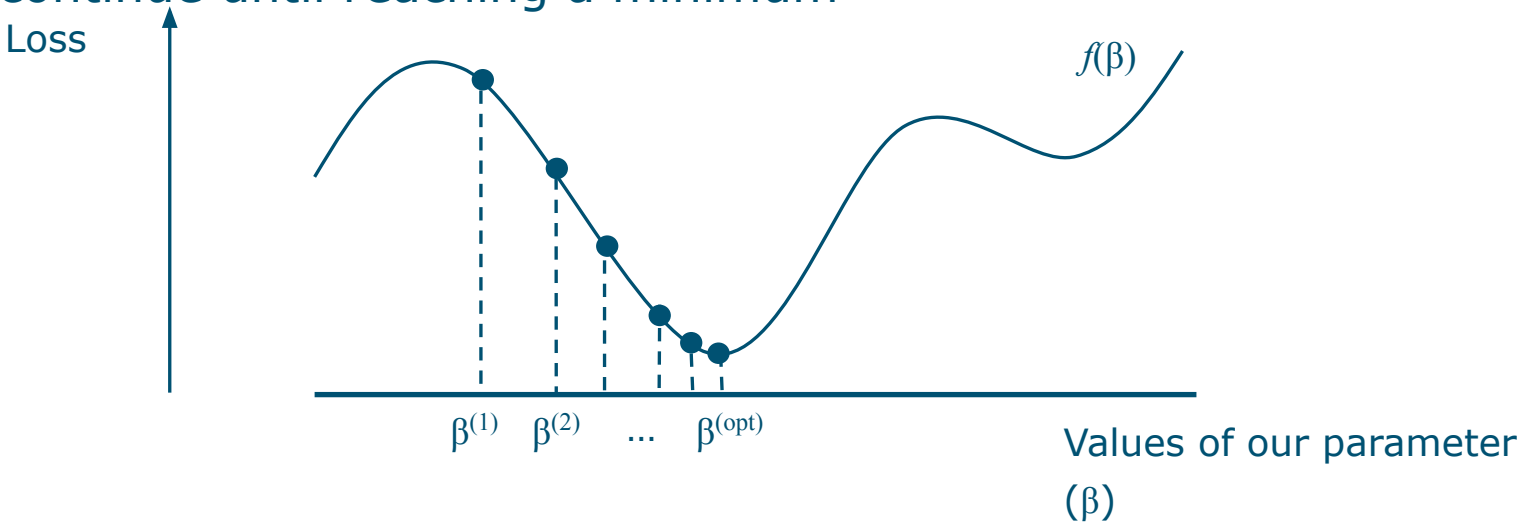
Gradient descent

- We want to find the minimum of an error function
- We start with an initial guess of the parameter β
- We change its value in the direction of maximal slope



Gradient descent

- We want to find the minimum of an error function
- We start with an initial guess of the parameter β
- We change its value in the direction of maximal slope
- We continue until reaching a minimum



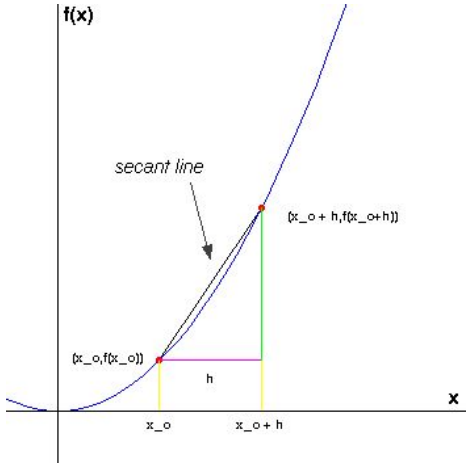
Gradient descent - how

- We want to find the minimum of an error function
- We start with an initial guess of the parameter β
- We change its value in the direction of maximal slope
- We until reaching a minimum

The derivative measures the steepness of the graph of a function at some particular point on the graph.
Thus, the derivative is a slope.

$$\frac{\partial f(x)}{\partial x}|x_0 = f'(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}$$

Partial derivative notation



It is also a function, it varies according to x

Source: https://www.ugrad.math.ubc.ca/coursedoc/math100/notes/derivs/deriv5.html

Derivative rules

Basic Derivatives Rules

Constant Rule: $\frac{d}{dx}(c) = 0$

Constant Multiple Rule: $\frac{d}{dx}[cf(x)] = cf'(x)$

Power Rule: $\frac{d}{dx}(x^n) = nx^{n-1}$

Sum Rule: $\frac{d}{dx}[f(x)+g(x)] = f'(x)+g'(x)$

Difference Rule: $\frac{d}{dx}[f(x)-g(x)] = f'(x)-g'(x)$

Product Rule: $\frac{d}{dx}[f(x)g(x)] = f(x)g'(x)+g(x)f'(x)$

Quotient Rule: $\frac{d}{dx}\left[\frac{f(x)}{g(x)}\right] = \frac{g(x)f'(x)-f(x)g'(x)}{[g(x)]^2}$

Chain Rule: $\frac{d}{dx}f(g(x)) = f'(g(x))g'(x)$

Derivative Rules

Exponential Functions

$\frac{d}{dx}(e^x) = e^x$

$\frac{d}{dx}(a^x) = a^x \ln a$

$\frac{d}{dx}(e^{g(x)}) = e^{g(x)}g'(x)$

$\frac{d}{dx}(a^{g(x)}) = \ln(a)\, a^{g(x)}\, g'(x)$

Logarithmic Functions

$\frac{d}{dx}(\ln x) = \frac{1}{x}, x > 0$

$\frac{d}{dx}\ln(g(x)) = \frac{g'(x)}{g(x)}$

$\frac{d}{dx}(\log_a x) = \frac{1}{x \ln a}, x > 0$

$\frac{d}{dx}(\log_a g(x)) = \frac{g'(x)}{g(x)\ln a}$

Trigonometric Functions

$\frac{d}{dx}(\sin x) = \cos x$

$\frac{d}{dx}(\cos x) = -\sin x$

$\frac{d}{dx}(\tan x) = \sec^2 x$

$\frac{d}{dx}(\csc x) = -\csc x \cot x$

$\frac{d}{dx}(\sec x) = \sec x \tan x$

$\frac{d}{dx}(\cot x) = -\csc^2 x$

Inverse Trigonometric Functions

$\frac{d}{dx}(\sin^{-1} x) = \frac{1}{\sqrt{1-x^2}}, x \neq \pm 1$

$\frac{d}{dx}(\cos^{-1} x) = \frac{-1}{\sqrt{1-x^2}}, x \neq \pm 1$

$\frac{d}{dx}(\tan^{-1} x) = \frac{1}{1+x^2}$

$\frac{d}{dx}(\cot^{-1} x) = \frac{-1}{1+x^2}$

$\frac{d}{dx}(\sec^{-1} x) = \frac{1}{x\sqrt{x^2-1}}, x \neq \pm 1, 0$

$\frac{d}{dx}(\csc^{-1} x) = \frac{-1}{x\sqrt{x^2-1}}, x \neq \pm 1, 0$

Hyperbolic Functions

$\frac{d}{dx}(\sinh x) = \cosh x$

$\frac{d}{dx}(\cosh x) = \sinh x$

$\frac{d}{dx}(\tanh x) = \operatorname{sech}^2 x$

$\frac{d}{dx}(\operatorname{csch} x) = -\operatorname{csch} x \coth x$

$\frac{d}{dx}(\operatorname{sech} x) = -\operatorname{sech} x \tanh x$

$\frac{d}{dx}(\coth x) = -\operatorname{csch} x$

Inverse Hyperbolic Functions

$\frac{d}{dx}(\sinh^{-1} x) = \frac{1}{\sqrt{1+x^2}}$

$\frac{d}{dx}(\cosh^{-1} x) = \frac{1}{\sqrt{x^2-1}}, x > 1$

$\frac{d}{dx}(\tanh^{-1} x) = \frac{1}{1-x^2}, |x| < 1$

$\frac{d}{dx}(\operatorname{csch}^{-1} x) = \frac{-1}{|x|\sqrt{1-x^2}}, x \neq 0$

$\frac{d}{dx}(\operatorname{sech}^{-1} x) = \frac{-1}{x\sqrt{1-x^2}}, 0 < x < 1$

$\frac{d}{dx}(\coth^{-1} x) = \frac{1}{1-x^2}, |x| > 1$

https://www.onlinemathlearning.com/derivative-rules.html

What is a loss in practice?

It's a number that measures how bad a model is (we want to make the loss small).

For regression problems (example, a counting problem), a typical loss is the Mean Squared Error (MSE):

$$\mathcal{L}_{\text{MSE}} = \sum_i (\hat{y}_i - y_i)^2$$

For binary classification problems, a typical loss is the Binary Cross Entropy (note that 0 < y < 1):

$$\mathcal{L}_{\text{CE}} = \sum_i y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)$$

What is a loss in practice?

Let's use MSE as an example.

$$\mathcal{L}_{\text{MSE}} = \sum_i (\hat{y}_i - y_i)^2$$

What would be the derivative of the loss and what does it mean?

$$\frac{\partial \mathcal{L}_{\text{MSE}}}{\partial \hat{y}_i}$$

Gradient descent - how

To update a weight β, we remove to its value the derivative

$$\beta^{(r+1)} = \beta^{(r)} - \alpha \frac{\partial f(\beta)}{\partial \beta}$$

α *is called a learning rate*

(r) is the current iteration, (r+1) is the next iteration

Gradient descent - how

To update a weight β, we remove to its value the derivative

$$\beta^{(r+1)} = \beta^{(r)} - \alpha \frac{\partial f(\beta)}{\partial \beta}$$

α *is called a learning rate*

(r) is the current iteration, (r+1) is the next iteration

Why the minus sign?

Gradient descent - how

To update a weight β, we remove to its value the derivative

$$\beta^{(r+1)} = \beta^{(r)} - \alpha \frac{\partial f(\beta)}{\partial \beta}$$

α *is called a learning rate*

(r) is the current iteration, (r+1) is the next iteration

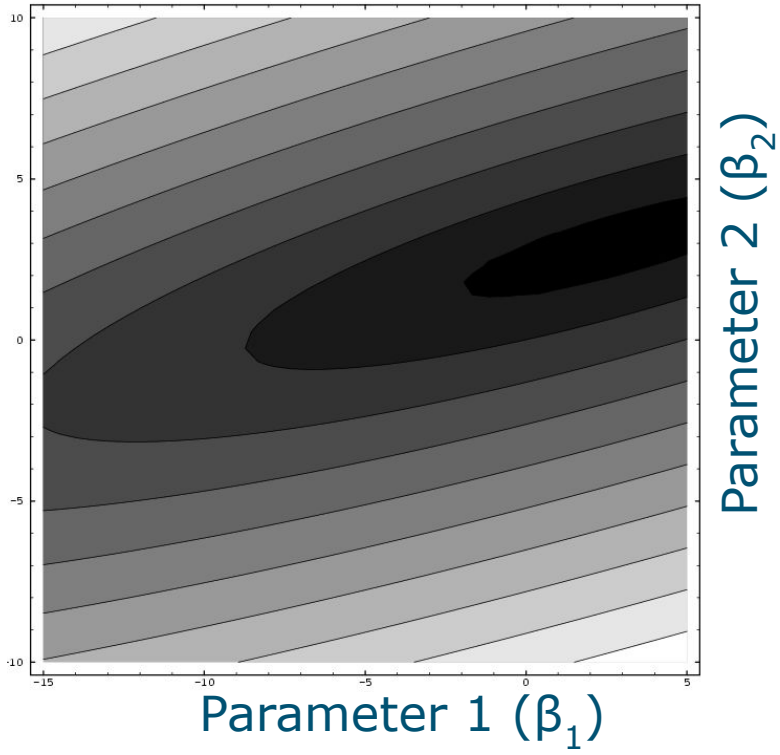
Why the minus sign?

Because we want to move towards a minimum!

Some intuition

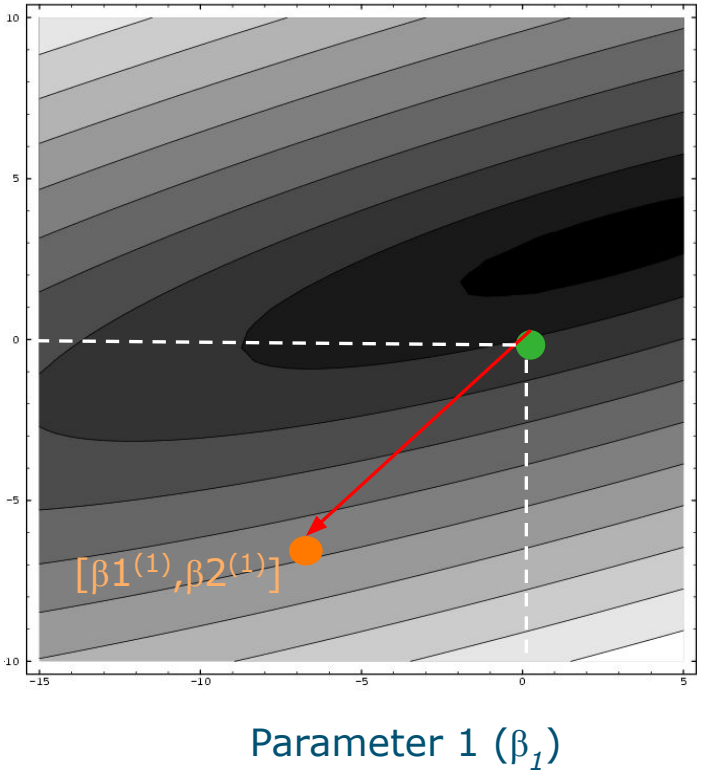
- In the 2D case this is simpler to visualize

- (the grey tone of the plot is the error, the darker the smaller)



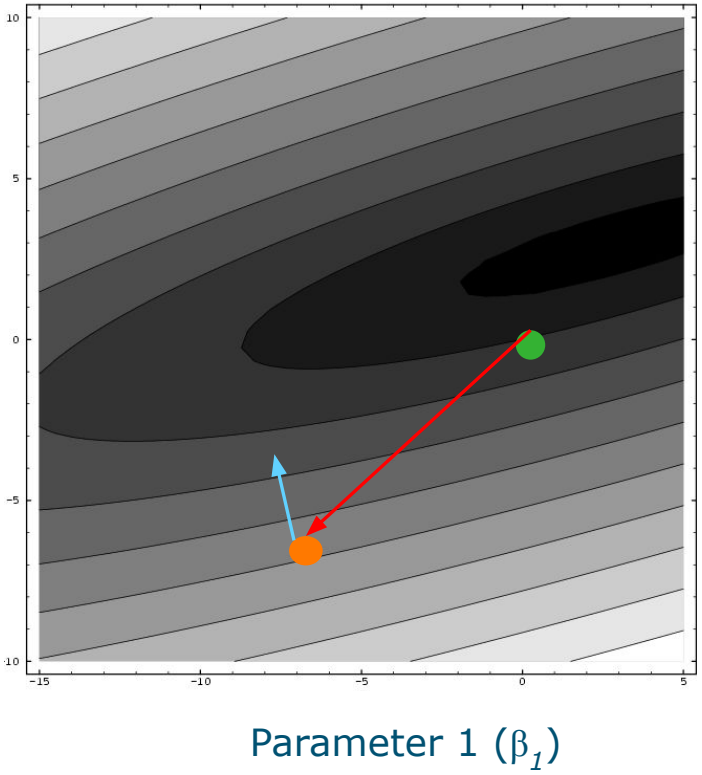
Some intuition

- In the 2D case this is simpler to visualize
- Your initial parameter $[\beta_1^{(1)},\beta_2^{(1)}]$ is a 2D vector from the origin [0,0]



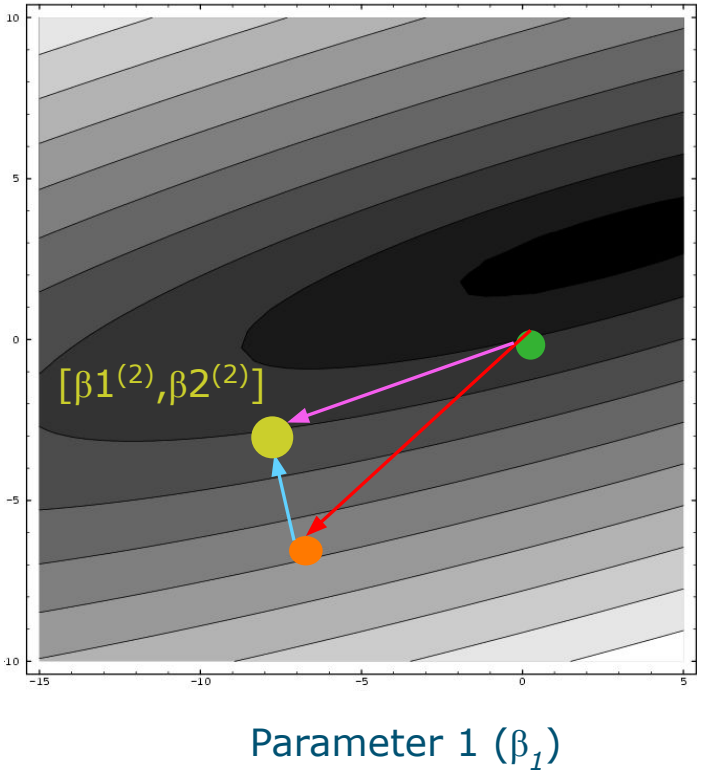
Some intuition

- In the 2D case this is simpler to visualize
- Your initial parameter $[\beta_1^{(1)},\beta_2^{(1)}]$ is a 2D vector from the origin [0,0]
- We compute the direction of maximal slope



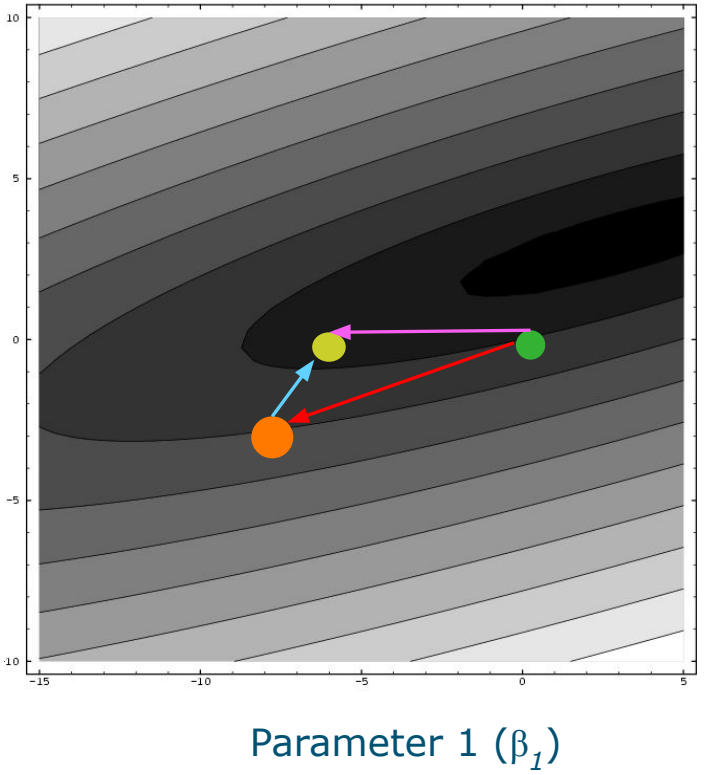
Some intuition

- In the 2D case this is simpler to visualize
- Your initial parameter $[\beta_1^{(1)},\beta_2^{(1)}]$ is a 2D vector from the origin [0,0]
- We compute the direction of maximal slope
- The new parameter $[\beta_1^{(2)},\beta_2^{(2)}]$ is the sum of the two vectors

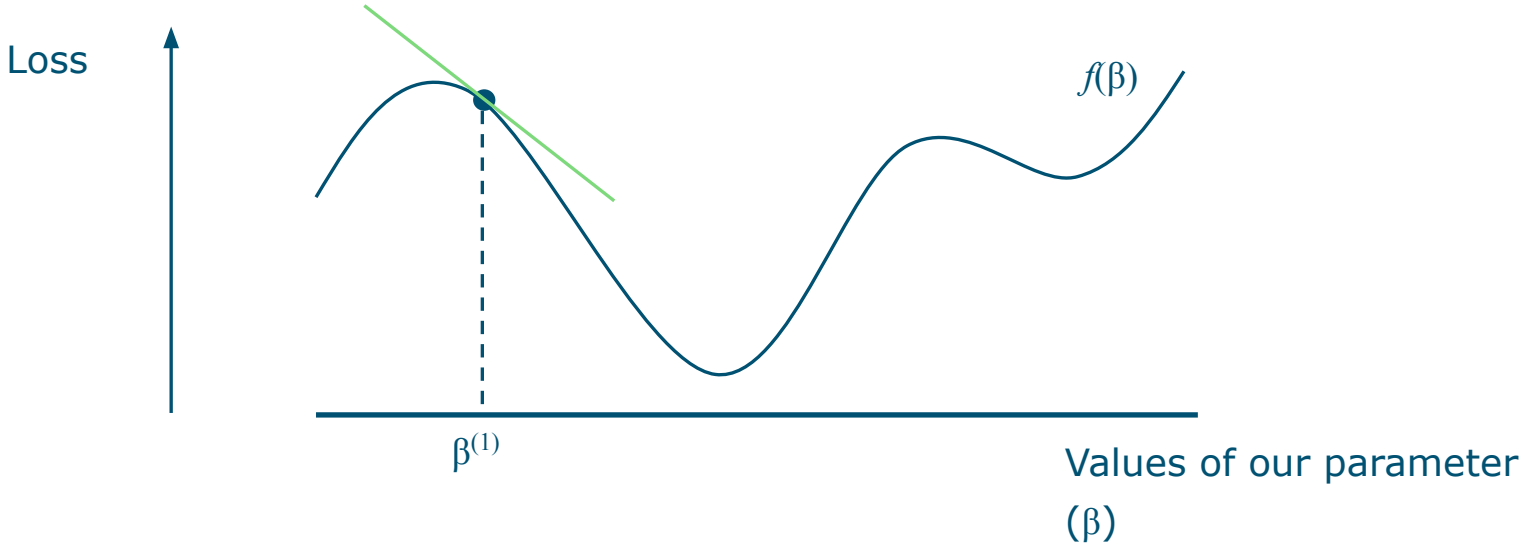


… and then again

- In the 2D case this is simpler to visualize
- Your initial parameter $[\beta_1^{(2)},\beta_2^{(2)}]$ is a 2D vector from the origin [0,0]
- We compute the direction of maximal slope
- The new parameter $[\beta_1^{(3)},\beta_2^{(3)}]$ is the sum of the two vectors

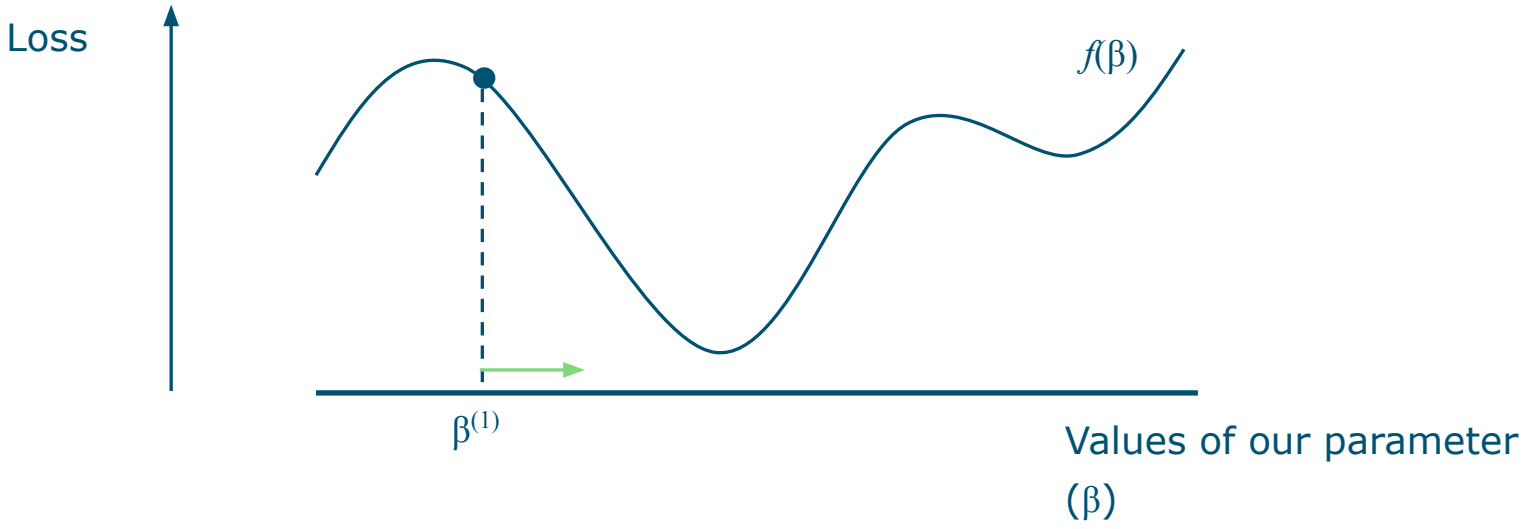


In images – 1 calculate the derivative



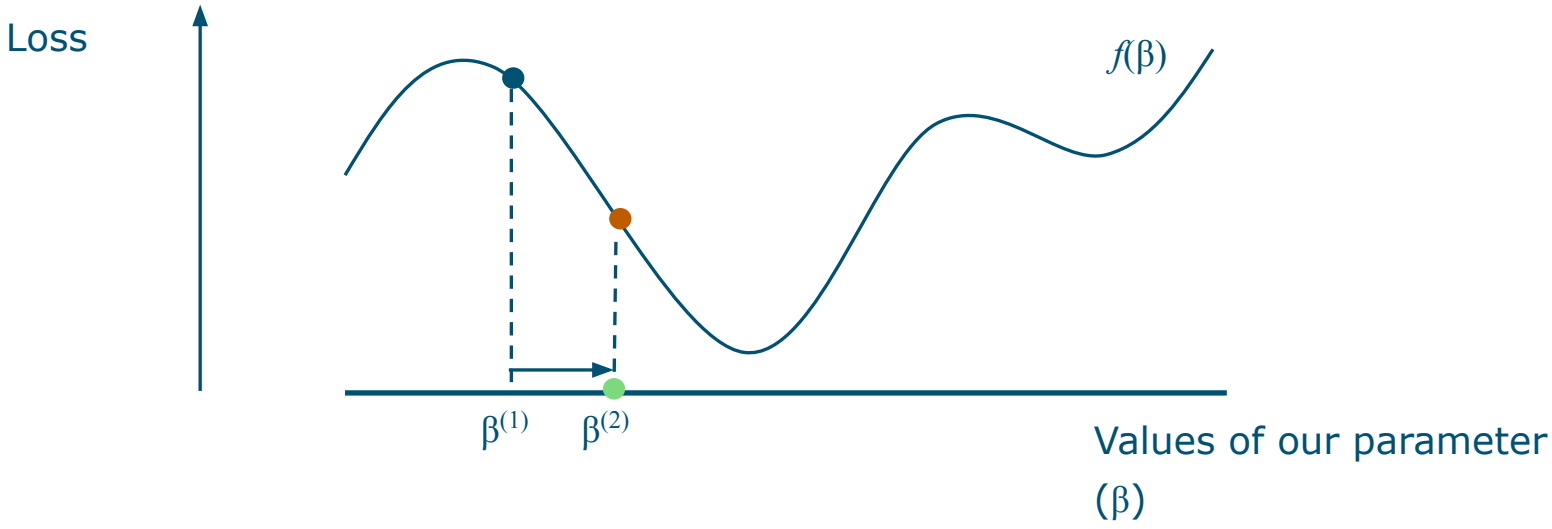
$$\beta^{(r+1)} = \beta^{(r)} - \alpha \frac{\partial f(\beta)}{\partial \beta}$$

In images – 2 this gives you the step vector



$$\beta^{(r+1)} = \beta^{(r)} - \alpha \frac{\partial f(\beta)}{\partial \beta}$$

In images – 3 which gives you the new value of β and its error



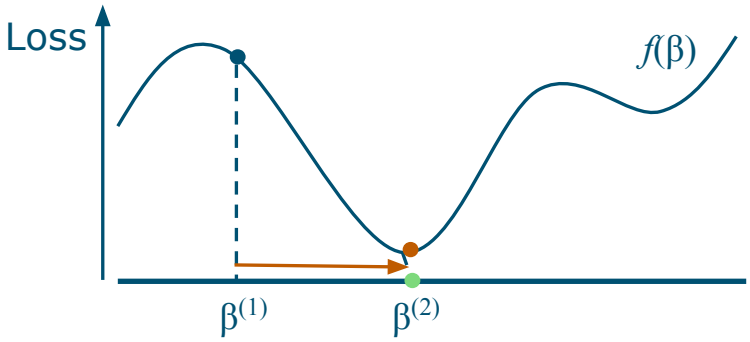
$$\beta^{(r+1)} = \beta^{(r)} - \alpha \frac{\partial f(\beta)}{\partial \beta}$$

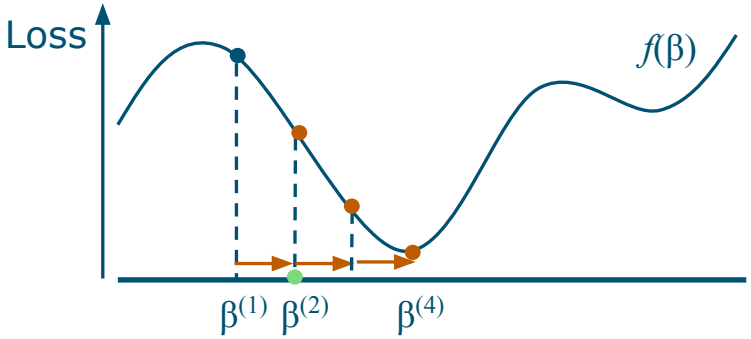
The learning rate

$$\beta^{(r+1)} = \beta^{(r)} - \alpha \frac{\partial f(\beta)}{\partial \beta}$$

- It decides by how much you multiply the step vector

Large α can lead to faster convergence than small



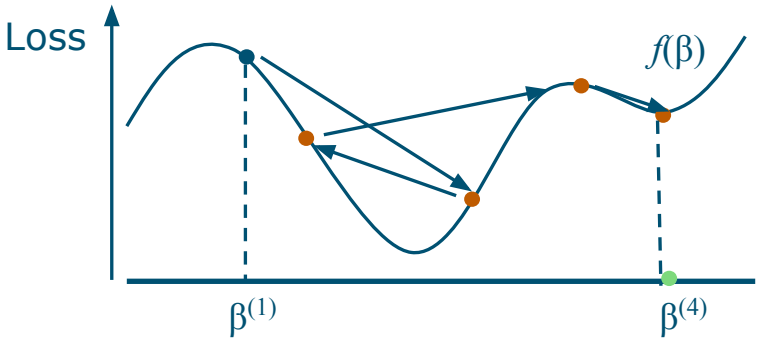


The learning rate

$$\beta^{(r+1)} = \beta^{(r)} - \alpha \frac{\partial f(\beta)}{\partial \beta}$$

- It decides by how much you multiply the step vector

Too large α can lead to disaster



Momentum

- The last trick is called momentum
- Momentum discourages sharp changes of direction in the descent. $\Delta\beta^{(r)} = \beta^{(r)} - \beta^{(r-1)}$
- It basically adds a term going in the direction of the previous step

Gradient descent like before

momentum

$$\beta^{(r+1)} = \beta^{(r)} - \alpha\frac{\partial f(\beta)}{\partial\beta} + \eta\Delta\beta$$

https://distill.pub/2017/momentum/

Pen and paper - Exercise 1

We try to minimize the function $f(\beta)=5\beta^2-3\beta+2$ with respect to β (this means by modifying only β) using gradient descent. Let us initialize as $\beta=4$.

1. What is the value of β after 3 gradient descent steps, using a learning rate of 1?
2. What is the value of β after 3 gradient descent steps, using a learning rate of 0.01?
3. What is the value of β after 3 gradient descent steps, using a learning rate of 0.1?
4. In which of these cases could momentum help and why?

Tip: What is the derivative of $f(\beta)$?

Pen and paper - Exercise 1

We try to minimize the function $f(\beta)=5\beta^2-3\beta+2$ with respect to β (this means by modifying only β) using gradient descent. Let us initialize as $\beta=4$.

1. What is the value of β after 3 gradient descent steps, using a learning rate of 1?
2. What is the value of β after 3 gradient descent steps, using a learning rate of 0.01?
3. What is the value of β after 3 gradient descent steps, using a learning rate of 0.1?
4. In which of these cases could momentum help and why?

Tip: What is the derivative of $f(\beta)$?
It's $df(\beta)/d\beta = 10\beta - 3$

$$\beta^{(r+1)} = \beta^{(r)} - \alpha \frac{\partial f(\beta)}{\partial \beta}$$

$$f(x) = 5x^2 - 3x + 2$$

$$f'(x) = 10x - 3$$

a. Starting at $x_0 = 4$: $x_1 = x_0 - \alpha \left.\frac{\partial f(x)}{\partial x}\right|_{x=x_0}$

$$= x_0 - \alpha f'(x_0)$$

$f'(x_0) = 40 - 3 = 37$, $\alpha = 1$

$$x_1 = 4 - 37 = 33$$

At $x_1 = 33$: $x_2 = x_1 - \alpha f'(x_1)$

$$= 33 - (330 - 3) = -300$$

At $x_2 = -300$: $x_3 = x_2 - \alpha f'(x_2) = -300 - (-3000 - 3)$

$$= 2703$$

$f(\beta)=5\beta^2-3\beta+2$

It's the same but the variable is called x instead of β

$$f(x) = 5x^2 - 3x + 2$$

$$f'(x) = 10x - 3$$

b. Starting at $x_0 = 4$: $x_1 = x_0 - a \left.\frac{\partial f(x)}{\partial x}\right|_{x=x_0}$

$$= x_0 - a f'(x_0)$$

$f'(x_0) = 40 - 3 = 37$, $a = 0.01$

$$x_1 = 4 - 0.01 \cdot 37 = 4 - 0.37 = 3.63$$

At $x_1 = 33$: $x_2 = x_1 - a f'(x_1)$

$$= 3{,}63 - 0.01\,(36{,}3 - 3) = 3{,}297$$

At $x_2 = -300$: $x_3 = x_2 - a f'(x_2)$

$$= 3{,}297 - 0.01\,(32{,}97 - 3) = 2{,}99$$