Week 2

[Week 1 we learnt about functions with 1 variable]

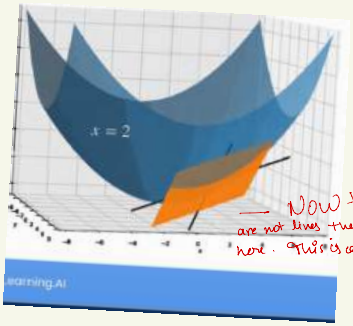
Now 2 variables or more.

Optimizing tangent planes on 2-D is hard similarly optimizing two or more functions can get very complicated even for a computer

To speed this up

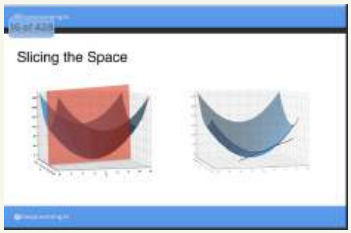
1) Gradient descent.

Functions of two Variables



Now tangents are not lines they are planes like here. This is called a tangent plane

$F(x,y) = x^2 + y^2$ 2 inputs 1 output. So 3-D space



So how do we get a tangent plane. We cut the space into planes then calculate tangent on the planes

Say we cut by $y = 4$

$F(x,4) = x^2 + 4^2$

$\frac{\partial (F(x,4))}{\partial x} = 2x$

Fix $x = 2$

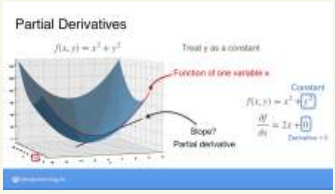
$F(2,y) = 2^2 + y^2$

$\frac{\partial}{\partial y}(F(2,y)) = 2y$

2 the lines $2x$, $2y$ form the tangent plane.

this plane. so you slice space in both directions.

Partial Derivatives



The slope of the parabola we get when fixing y as a constant.

& " " " x " " " " " " x " " "

$x^2 + y^2$

treat y as a constant → $2x$

treat x as a constant → $2y$

$F(x,y)$

$f_x = \frac{\partial f}{\partial x}$ Partial derivative of F with respect to x

$f_y = \frac{\partial F}{\partial y}$ partial derivative of F with respect to y

$F(x,y) = x^2 + y^2$ → $\frac{\partial F}{\partial x} = 2x$ (in relation to x y is constant)

$\frac{\partial F}{\partial y} = 2y$ (" y " y x " ")

Q) $F(x) = 3x^2 y^3$ So if we take in relation to x we treat y as a constant.

$\frac{\partial F(x)}{\partial x} = ?$

$3 \cdot y^3 \cdot \frac{\partial (x^2)}{\partial x} = 3y^3 \cdot 2x = \underline{6xy^3}$

Now $\frac{\partial(3x^2y^3)}{\partial y}$ → x is now a constant. so $\frac{\partial(y^3)}{\partial y} \cdot 3x^2 = 3y^2 \cdot 3x^2 = 9x^2y^2$

Gradient → $F(x^2+y^2)$ → You can slice in two ways each gives you a partial derivative

The gradient is simply the vector containing these partial derivatives.

so for $F(x,y)$ → gradient = $\begin{bmatrix} 2x \\ 2y \end{bmatrix}$ or $\nabla F = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix}$

This is because $F(x,y)$ has 2 variables

if a function has 17 variables we will have 17 partial derivatives.

So these gradients $\begin{bmatrix} 2x \\ 2y \end{bmatrix}$ is a pretty good description of the tangent plane, because it gives the slopes of the two lines that form the tangent plane.

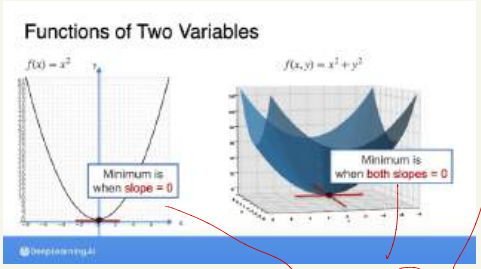
Q) $F(x,y) = x^2+y^2$ calculate the gradient F, ∇F at $(2,3)$

gradient = $\begin{bmatrix} \frac{\partial F}{\partial x} \\ \frac{\partial F}{\partial y} \end{bmatrix}$ $\frac{df}{dx} = 2x$ $\frac{df}{dy} = 2y$.

$\Rightarrow \begin{bmatrix} 2x \\ 2y \end{bmatrix}$ at $(2,3)$ → $\begin{bmatrix} 4 \\ 6 \end{bmatrix}$

The gradient is useful to minimize / maximize a function of two or more variable in the same way a derivative for 1 variable.

Functions of two variables



imp

on the left $f'(x) = 0$

$2x = 0 \quad x = 0$

on the right

$\frac{\partial f}{\partial x} = 0$ & $\frac{\partial f}{\partial y} = 0$

$2x = 0 \qquad 2y = 0$

same as $(x, y) = (0, 0)$

So basically to find the min or max you have to set the partial derivatives as zero & solve that linear equation.

Motivation for optimization in two variables

say you're in a sauna & you can move in any direction. You want to find the best place. You want to take steps to find the coldest place in the room.

HOW? → 1) You look for the place where if you move in any direction from there you feel an increase in temp.

Another way to look at it is if you were to take a tangent plane to a function, this tangent plane should be parallel to the floor.

The two partial derivatives should be zero.

Exercise: $T = F(x,y) = 85 - \frac{1}{90}x^2(x-6)y^2(y-6)$

Proactive saving:

given this function find $\partial F/\partial x$ & $\partial F/\partial y$

$$F(x,y) = 85 - \frac{1}{90}x^3y^3 + \frac{1}{15}x^3y^2 + \frac{1}{10}x^2y^3 - \frac{2}{5}x^2y^2$$

$$\rightarrow \frac{\partial F}{\partial x} = \frac{-1}{30}x^2y^3 + \frac{1}{5}x^2y^2 + \frac{2}{15}xy^3 - \frac{4}{5}xy^2$$

$$\rightarrow \frac{\partial F}{\partial y} = \frac{-1}{30}x^3y^2 + \frac{2}{15}x^3y + \frac{1}{5}x^2y^2 - \frac{4}{5}x^2y$$

$\frac{\partial F}{\partial x} = 0$ & $\frac{\partial F}{\partial y} = 0$

Then we get a lot of values then we go through all pairs

$$\frac{\partial F}{\partial x} = \frac{-xy^2}{30}(xy - 6x - 4y + 24) \rightarrow \left[\frac{-xy^2}{30}(x-4)(y-6)\right]$$

$$\frac{\partial F}{\partial x} = \frac{-xy^2}{30}(x(y-6) - 4(y-6))$$

Now if $\frac{\partial F}{\partial x} = 0$ & $\frac{\partial F}{\partial y} = 0$ so 3 products = 0 that means.

$$\frac{\partial F}{\partial x} = \frac{-xy^2}{30}(x-4)(y-6) \quad (1)$$

! Next page!

$$\frac{\partial F}{\partial y} = \frac{-1}{30}x^2y(xy - 4x - 6y + 24)$$

$$\frac{\partial F}{\partial y} = \frac{-1}{30}x^2y(y-4)(x-6) \quad (2)$$

① $-\frac{xy^2}{30}(x-4)(y-6)$ ② $-\frac{x^2y}{30}(y-4)(x-6)$

so $xy^2=0$, $x-4=0$, $y-6=0$ $\quad x^2y=0$, $y-4=0$ $x-6=0$

$x=0,y=0$, $x=4,y=6$ — ③ $\quad x=0,y=0$, $y=4$, $x=6$ — ④

From ③ & ④ we get the following candidates:

$x=0$, $y=0$ $\quad x=0,y=4$ $\quad x=4,y=0$ $\quad x=6,y=0$ $\quad x=0,y=6$ $\quad x=4,y=6$ $\quad x=6,y=6$

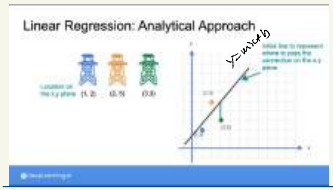
red out of bound

As you can see sauna is 0 to Fineitne direction

From the candidates you see $x=4, y=4$ is minimal.

Let's use the power line model again (Linear regression)

But 2-D.



The goal here is to find the optimal place for a fiber line connection that goes in a straight line in such a way that you reduce the total cost of connecting to the 3 power lines. This can be done by connecting wires to the power lines but the wires must be parallel to the y axis.

Now if we focus on the blue power line (1, y) if we make parallel to the y-axis & connect we connect at point (1, mx1+b)

cost of connection = square of length of wires

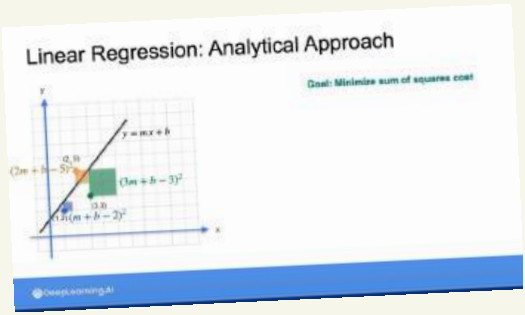
so $y=mx+b$ is the fiber line we need to find m & b so that distance is minimized

---

goal Minimize sum of squares cost.

blue → (1,2) green → (3,3) orange → (2,5)

So you can see the distance



For the blue line: $(m+b-2)^2$

" green " : $(3m+b-3)^2$

" orange " : $(2m+b-5)^2$

why squared → function is x^2

total cost → $(m+b-2)^2+(3m+b-3)^2+(2m+b-5)^2$

$\Rightarrow 14m^2+3b^2+38+12mb-42m-20b = E(m,b)$

$\frac{\partial E}{\partial m}=0$, $\frac{\partial E}{\partial b}=0$

$\frac{\partial E}{\partial m} = 28m+0+12b+42$

$\rightarrow 28m+12b+42=0$

$\frac{\partial E}{\partial b} = 6b+12m-20=0$

$6b+6-20=0$

$6b-14=0$

$[b=7/3]$

$28m+12b-42$

$-24m-12b+40$

$4m-2=0$

$[m=1/2]$

$m = 1/2 \quad b = 7/3 \quad E(m = 1/2, b = 7/3) \approx 4.167$

To solve the linear Equations can we do something? Yes Gradient descent

Quiz 1) $F(x,y) = x^2y + 3x^2 \quad \frac{\partial F}{\partial x} = 2xy + 6x$ 2) $F(x,y) = xy^2 + 2x + 3y$ gradient $F(x,y) = ?$ 0, -6

gradient $= \begin{bmatrix} \partial F/\partial x \\ \partial F/\partial y \end{bmatrix} \rightarrow \begin{matrix} y^2+2 \\ 2xy+3 \end{matrix}$

3) $P(x,y) = x^2 + 2y^2 + 8y$ min of F is:

$\frac{\partial F}{\partial x} = 2x = 0 \quad \frac{\partial F}{\partial y} = 4y + 8$

$(x = 0) \quad y = -8/4 \quad (y = -2)$

So if we put $F(x=0, y=-2) = 4 - 11 = -12$

Q) gradient of $x^2 + 2xyz + z^2$

$\frac{\partial}{\partial x} = 2x + 2yz \quad \frac{\partial}{\partial y} = 2xz \quad \frac{\partial}{\partial z} = 2xy + 2z$

$2x + 2yz = 0$

$2xz = 0$

$2xy + 2z = 0$

Gradient Descent

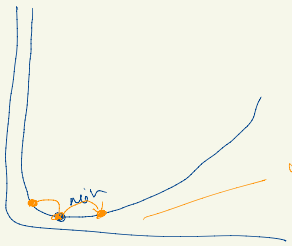
When we try to solve the gradient or derivatives of functions especially ones with many variables it is very difficult

So we use Gradient Descent. - iterative way to find min/max

eg) $P(x) = e^x - \log(x)$ minimum!

$P'(x) = e^x - \frac{1}{x} = 0$

$e^x = 1/x$



we move in both directions, find which is smaller & move there

— This is the learning rate.

Now is there any information. yes. the slope of any current point is

if

If slope -ve move to right

" " +ve " " left

Gradient Descent

1) Define learning rate α & choose a starting point x_0.
2) Update $x_k = x_{k-1} - \alpha f'(x_{k-1})$
3) Repeat step2 until you are close to the true minimum x^*

If new point x_1 $\quad x_1 = x_0 - f'(x_0)$ ← old point.

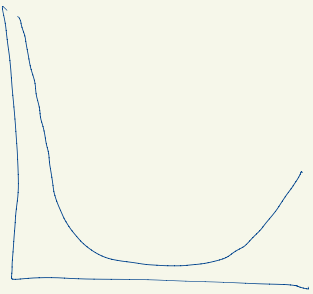
↳ slope of x_0

Big steps can be chaotic we need small steps.

$$x_1 = x_0 - 0.01\, f'(x_0)$$

0.01 ← learning rate

You can choose any learning rate you want.



$$x_1 = x_0 - \alpha f'(x_0)$$

↳ learning rate

$f(x) = e^x - \log(x)$ $\quad f'(x) = e^x - 1/x$

Start $x = 0.05$ $\quad$ Rate: $\alpha = 0.005$

$f'(0.05) = -18.9$.

move by $-0.005\, f'(0.05)$

Now let's try gradient descent

1) start $x = 0.05$, $\alpha = 0.005$

$f'(0.05) = -18.9$

move by $-0.005 \cdot f'(0.05)$ $\quad x_1 = 0.1447$

Fine

$f'(0.1447) = -5.7552$

move by $-0.005 f'(0.1447)$

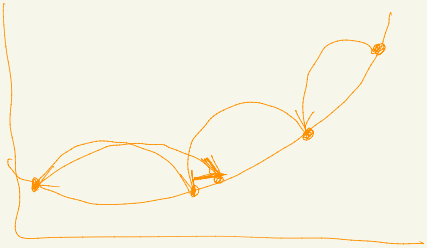
$x \to 0.1735$

Notice we never needed to solve derivative = 0

we only need to know the derivative & apply it to the upcoming step

What is a Good Learning Rate.

Too large



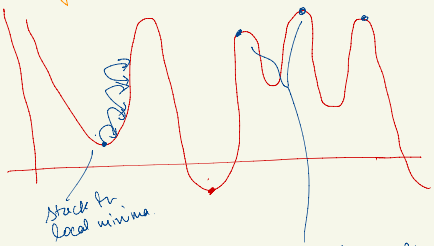
You may never reach the minimum because the steps are too big.

Too small

takes too long, or you might never reach it.

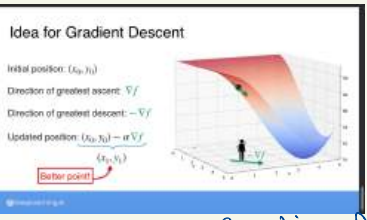
Very hard to get the best learning rate

Draw Backs



To fix this you can take multiple initial starting points for gradient descent

Gradient descent with Real Example. [More than 1 variable]



So in this the gradient = ∇f

If you move in ∇f = greatest ascent

so you move with ∇f direction.

In one variable we used derivative in two

we use gradient.

Let's see how it works → $T = f(x,y) = 85 - \frac{1}{90}x^2(x-6)y^2(y-6)$

$$\nabla f = \begin{bmatrix} \partial f/\partial x \\ \partial f/\partial y \end{bmatrix} \qquad \nabla f = \begin{bmatrix} -\frac{1}{90}x(3x-12)y^2(y-6) \\ -\frac{1}{90}x^2(x-6)y(3y-12) \end{bmatrix}$$

Start $x=0.5$, $y=0.6$

$\nabla f(0.5, 0.6) = \begin{bmatrix} -0.1134 \\ -0.0935 \end{bmatrix}$ move by $-0.05 \nabla f(0.5, 0.6)$

Now $x = 0.5057$

$y = 0.6047$

Now run with new x, y.

Find $\nabla f(0.5057, 0.6047) = \begin{bmatrix} -0.1162 \\ -0.0961 \end{bmatrix}$ move by $-\alpha \nabla f(0.5057, 0.6047)$

If you repeat this many times you get to the min $\begin{bmatrix} x \to 0.5115 \\ y \to 0.6095 \end{bmatrix}$

Functionality → $f(x,y)$ Goal: Finding minimum of $f(x,y)$

Step 1: Define a learning rate α, choose a starting point (x_0, y_0)

Step 2: Update: $\begin{bmatrix} x_k \\ y_k \end{bmatrix} = \begin{bmatrix} x_{k-1} \\ y_{k-1} \end{bmatrix} - \alpha \nabla f(x_{k-1}, y_{k-1})$

Step 3 Repeat step 2 until you are close enough to true minimum.

You know when the true min is happening, because you move really really slowly or not at all.

Note: This still has drawbacks of local / global minimum you can avoid this by using different places.

We solved a power lines problem

let's solve using gradient descent

We had the following:

(2,5), (1,2), (3,3)

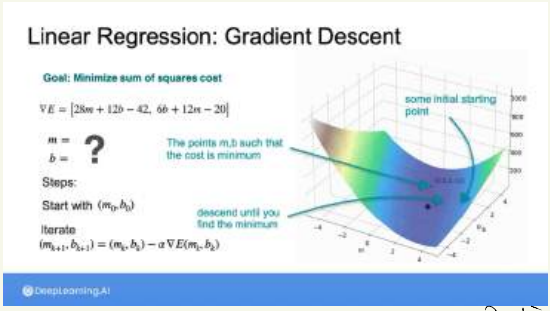
The problem became

$E(m,b) = 14m^2 + 3b^2 + 38 + 12mb - 42m - 20b$

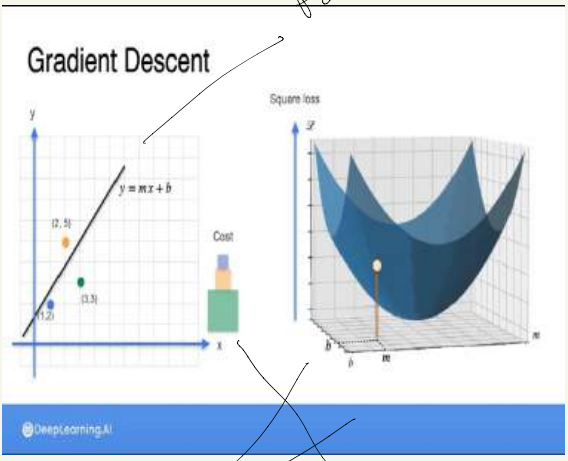
we solved $E(m,b)$ by finding $\frac{E}{\partial m} = 0$ & $\frac{E}{\partial b} = 0$

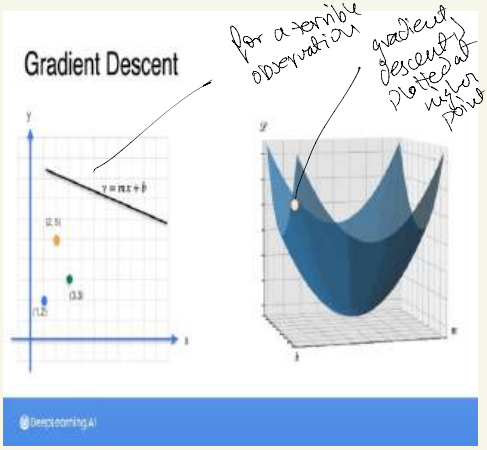
Now let's try to find the gradient descent.

The gradient was $\begin{bmatrix} 28m + 12b - 42 \\ 6b + 12m - 20 \end{bmatrix}$



for a good line



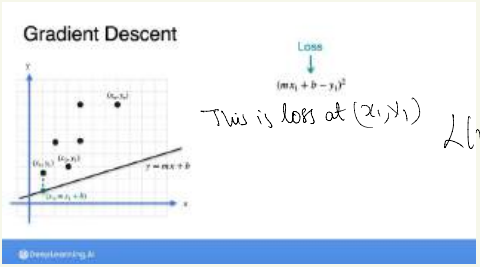


for a terrible observation

gradient descent plotted at wider point

We map m, b in the bottom plane

we are stacking up the cost function on y-axis or vertical axis.



Now we add add all losses of all points & divide by n to get average

$$L(m,b) = \frac{1}{2n}\left[(mx_1 + b - y_1)^2 + \ldots + (mx_n + b - y_n)^2\right]$$

↳ This extra 2 is that when we take derivative exponent multiplies by 2. It doesn't really matter but it'll

Now we create a descent.

we start at a random $\begin{bmatrix} m_0 \\ b_0 \end{bmatrix} \rightarrow \begin{bmatrix} m_1 \\ b_1 \end{bmatrix}$

$$\begin{bmatrix} m_1 \\ b_1 \end{bmatrix} = \begin{bmatrix} m_0 \\ b_0 \end{bmatrix} - \alpha \nabla L_1(m_0, b_0)$$

$$\begin{bmatrix} m_n \\ b_n \end{bmatrix} = \begin{bmatrix} m_{n-1} \\ b_{n-1} \end{bmatrix} - \alpha \nabla L_2(m_{n-1}, b_{n-1})$$

formal version of gradient descent