Gradient Descent

x	y
1	1
2	2
3	3

$$\hat{y} = \theta_0 + \theta_1 x$$

$$\varepsilon_i = y_i - \hat{y}_i = y_i - (\theta_0 + \theta_1 x_i)$$

$$\varepsilon_1 = 1 - \theta_0 - \theta_1 \qquad \varepsilon_1^2 = (1 - \theta_0 - \theta_1)^2$$

$$\varepsilon_2 = 2 - \theta_0 - 2\theta_1$$

$$\varepsilon_3 = 3 - \theta_0 - 3\theta_1$$

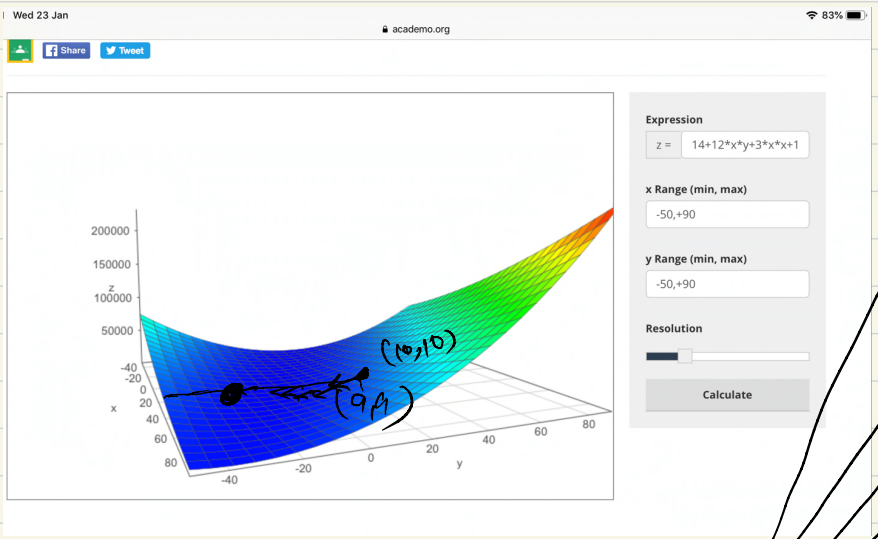
$$\sum \varepsilon_i^2 = f(\theta_0, \theta_1) = \left(1 + \theta_0^2 + \theta_1^2 - 2\theta_0 - 2\theta_1 + 2\theta_0\theta_1\right) + \left(4 + \theta_0^2 + 4\theta_1^2 - 4\theta_0 + 4\theta_0\theta_1 - 8\theta_1\right) + \left(9 + \theta_0^2 + 9\theta_1^2 - 6\theta_0 - 18\theta_1 + 6\theta_0\theta_1\right)$$

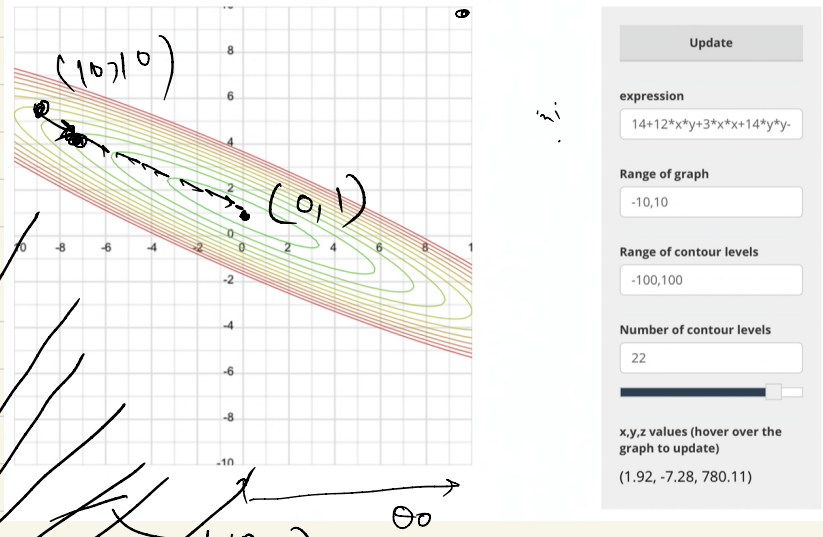
$$= 14 + 3\theta_0^2 + 14\theta_1^2 - 12\theta_0 - 28\theta_1 + 12\theta_0\theta_1$$

$3x^2 + 14y^2 - 2x - 2y + 12xy + 14$

3D Plot

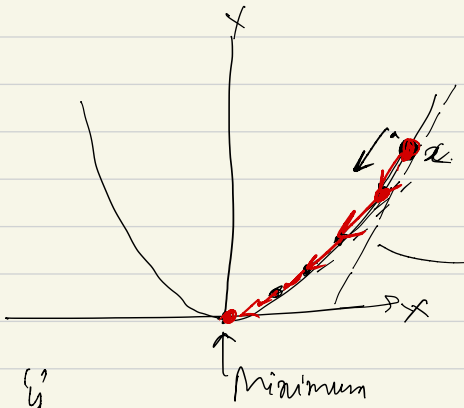
CONTOUR PLOT





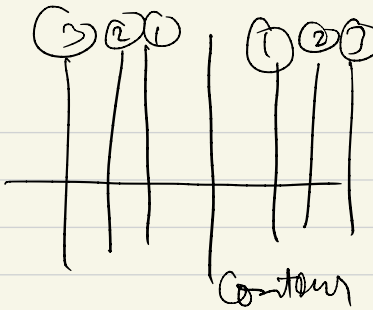
GRADIENT DESCENT

* General optimisation technique



$y = x^2$

Steepest descent $= -\frac{\Delta y}{\Delta x}\Big|_{x=x_1}$

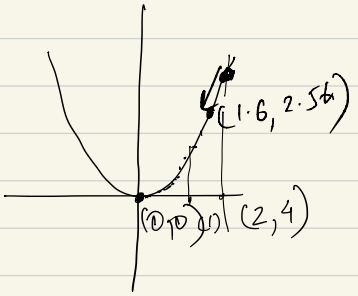


Question: Find minimum y

1. Start with some (x_0)
2. Move from (x_0) to x_1 s.t. $y_1 < y_0$.
3. Repeat 2

$$x_2 = x_1 - \alpha \frac{\Delta y}{\Delta x}\Big|_{x_1}$$


local minima.



$\alpha = 0.1$

$x_0 = 2$

$y_0 = 4$

$\frac{\Delta y}{\Delta x} = 2x$

$\therefore y = x^2$

Solving by Δ.

$x_1 = x_0 - \alpha \frac{\Delta y}{\Delta x}\Big|_{x_0} = x_0 - .2x_0$

$= .8x_0$

$x_1 = .8 \times 2 = 1.6$

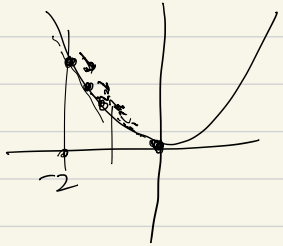
$x_2 = x_1 - .2x_1 = .8x_1 = 1.28$

$\vdots$

$x_N = (.8)^N x_0$

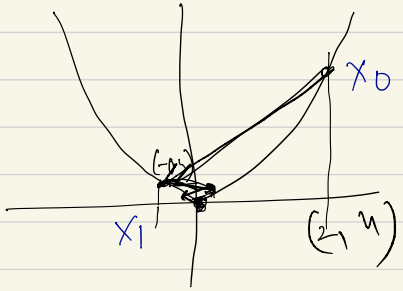
$x_N \to 0$ for high N.

OK, what if we start from $x_0 = -2$



$x_1 = x_0 - \alpha \frac{\Delta y}{\Delta x}\Big|_{x_0} = .8x_0 = -1.1$

Q: what if 'α' is large



$\alpha = 0.6$

$x_0 = 2$

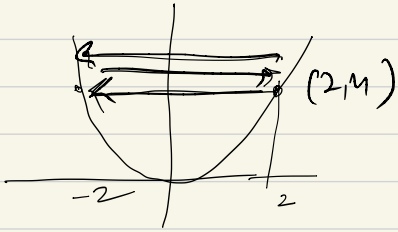
$x_1 = x_0 - \alpha \frac{\Delta y}{\Delta x}\Big|_{x_0} = x_0 - (0.6)(2x)\Big|_{x_0} = -0.2x_0$

$x_2 = -.2x_1 = (-0.4)(-0.2) = .08$

what if $\alpha = 1$

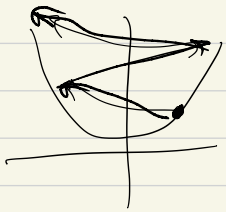
$x_1 = x_0 - (1)(2x)\Big|_{x_0} = -x_0$

$x_2 = -x_1$



$\alpha = 2$

$x_1 = x_0 - (2)(2x_0)\big|_{x_0} = -3x_0$



DIVERGING

$\alpha = 10^{-3}$ (α TOO SMALL?)

$x_1 = x_0 - (10^{-3})(2x_0)\big|_{x_0} \approx x_0$

TOO MANY STEPS

α TOO SMALL → SLOW
α TOO LARGE → DIVERGE

Many implementations

$\alpha = \frac{K}{t}, \frac{K}{\sqrt{t}}$

$\frac{K}{\sqrt{t+k_0}}$

G.D. for linear reg.

$$\hat{y}_i = \theta_0 + \theta_1 x_i$$

$$e_i = y_i - \hat{y}_i = y_i - (\theta_0 + \theta_1 x_i)$$

$$\Sigma e_i^2 = \Sigma (y_i - (\theta_0 + \theta_1 x_i))^2$$

① Start with random values of θ_0 & θ_1

Till convergence:

$$\theta_0 = \theta_0 - \alpha \frac{\partial}{\partial \theta_0} (\Sigma e_i^2)$$

$$\theta_1 = \theta_1 - \alpha \frac{\partial}{\partial \theta_1} (\Sigma e_i^2)$$

} Simultaneously

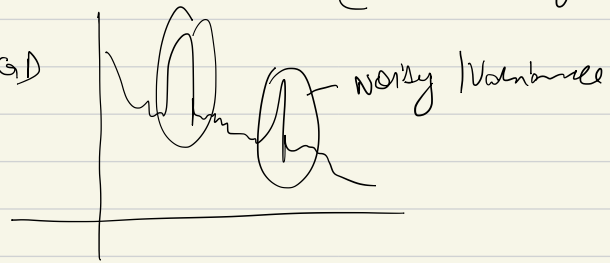
Now $\frac{\partial}{\partial \theta_0} (\Sigma e_i^2) = 2 \Sigma (y_i - (\theta_0 + \theta_1 x_i)) (-1)$ & $\frac{\partial}{\partial \theta_1} (\Sigma e_i^2) = 2 \Sigma (y_i - (\theta_0 + \theta_1 x_i)) (-x_i)$

G. Descent Variants

(Slower update)

① Vanilla (or Batch) : Update params after going through all data

② S.G.D : " " for each data point (more noisy)



Mini batch (Best of both worlds)

Gradient Descent

① Good for online setting (more data).

② Good for large data

Normal Equations

① Good for simple

② No need to worry about learning rate, etc.

③ Non trivial to solve (impossible...

Projected Gradient Descent

a) Learn a fit where $\theta_0, \ldots \theta_d$ are all ≥ 0

$$\theta_i = \max\left(\theta_i - \alpha \frac{\Delta \xi(\theta_1, \ldots \theta_i, \ldots \theta_d)}{\Delta \theta_i}, 0\right)$$