ANNOUNCEMENTS

1. HW2 late due tomorrow, 5pm (NOT 11:59 pm)
2. Prelim conflict declaration out on Ed (fill by 03/04)

TURN YOUR NON-NOTE-TAKING DEVICES OFF NOW!

---

while you wait, here's an icebreaker – Did you go sledding on the slope this semester?

Did starting at different parts of the slope affect your speed?

TODAY - OPTIMIZATION

logistic regression cost function

$$J(\theta) = \sum_{j=1}^{n} \log\left(1 + e^{-y^{(j)} \theta^T x^{(j)}}\right)$$

→ take the derivate $\overset{\text{set}}{=} 0$ → solve for θ

GOAL - instead of compute the "analytical" derivative, instead we seek to find iterates

$\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(k)}$, starting from some guess $\theta^{(0)}$

Idea is that, come up with some iteration, G such that

iteration

$$\theta^{(k+1)} = G(\theta^{(k)})$$

next estimate of θ

current estimate

Eshan says — $G(\theta^*) = \theta^*$ — "fixed point" of G

IDEA-1 : GRADIENT DESCENT

given $\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(K)}$

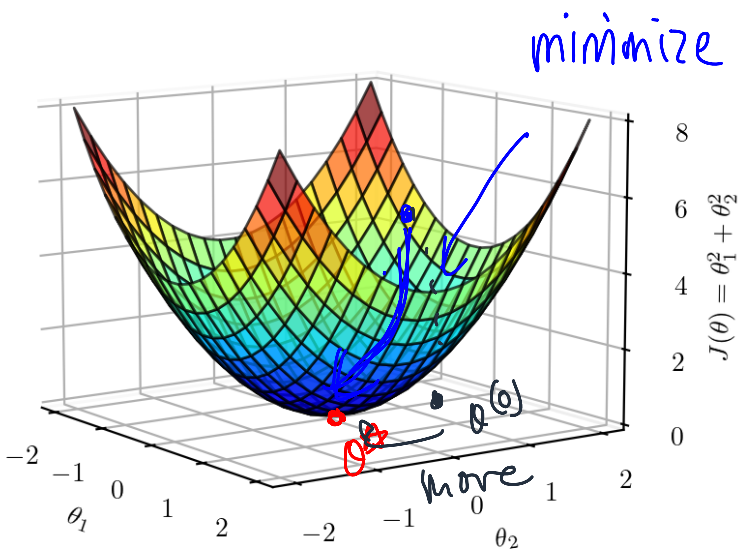
Example

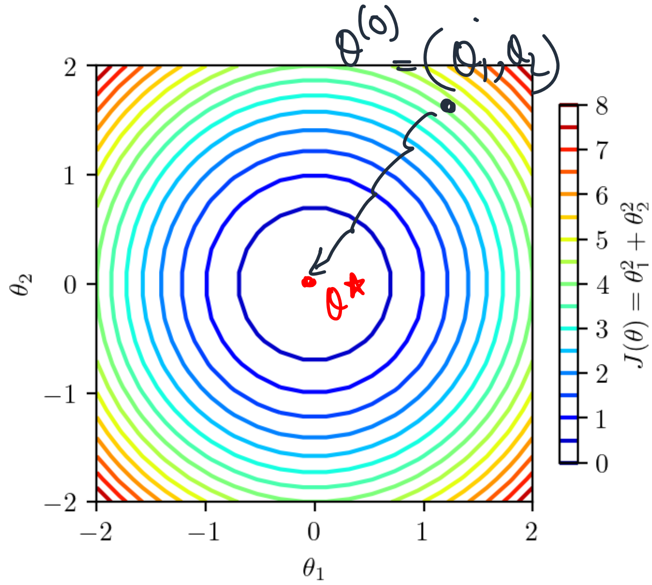
$J(\theta) = \theta_1^2 + \theta_2^2$

$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$

Specifically,

$\theta^* = \arg\min_{\theta} J(\theta) \longrightarrow$ what is θ^*? $= \begin{bmatrix} 0 \\ 0 \end{bmatrix}$





"TOP" view of 3D plot

DERIVATIVES

$$f'(x^{(0)}) \approx \frac{f(x^{(0)}+h) - f(x^{(0)})}{h} \qquad h = O(10^{-5})$$

when dealing with vectors, take partials!

$$J(\theta_1, \theta_2) = \theta_1^2 + \theta_2^2$$

$$\frac{\partial J}{\partial \theta_1} = 2\theta_1 \qquad \frac{\partial J}{\partial \theta_2} = 2\theta_2$$

evaluate at $(0,2)$

$$\frac{\partial J}{\partial \theta_1} = 2\theta_1 = 0 \qquad \frac{\partial J}{\partial \theta_2} = 2\theta_2 = 4$$

$$f(\theta_1 + h) = f(\theta_1) + \frac{\partial f}{\partial \theta_1} h$$

$$= f(\theta_1) + 0$$

increasing "h" doesn't affect "f" so long as small 'h'

$$f(\theta_2 + h) = f(\theta_2) + \frac{\partial f}{\partial \theta_2} h$$

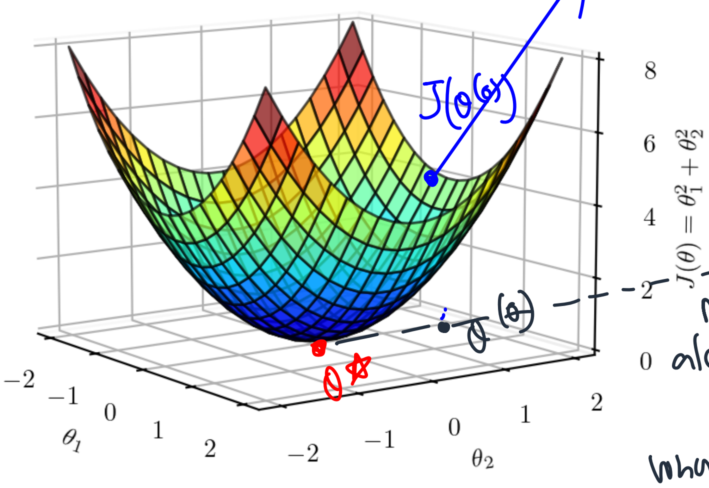
$$f(\theta_2 + h) = f(\theta_2) + 4h$$

increase θ_2 by "h" amplies f by "4h"

<u>OBSERVATION</u> :   this sensitivity measure can help navigate
"$\underline{J}$"
maximal impact can be understood!

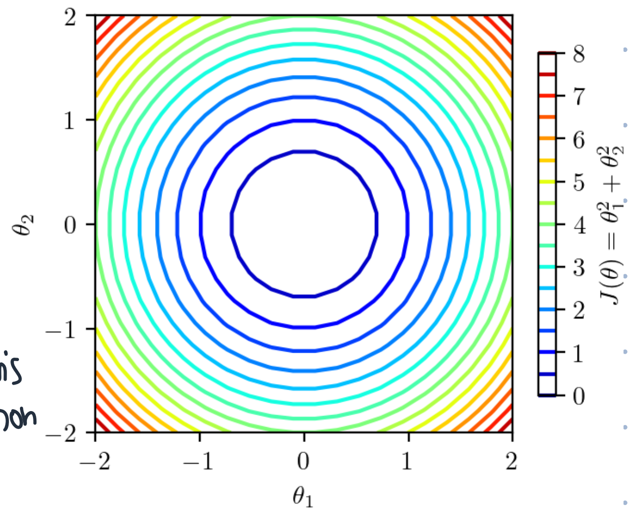
"J" moves so!

$J(\theta^{(0)})$



$\theta^{(0)}$ - - -> move along this direction

θ^*



why?

$$\frac{\partial J}{\partial \theta_1} = 0 \;, \quad \frac{\partial J}{\partial \theta_2} = 4$$

"Convenience" notation

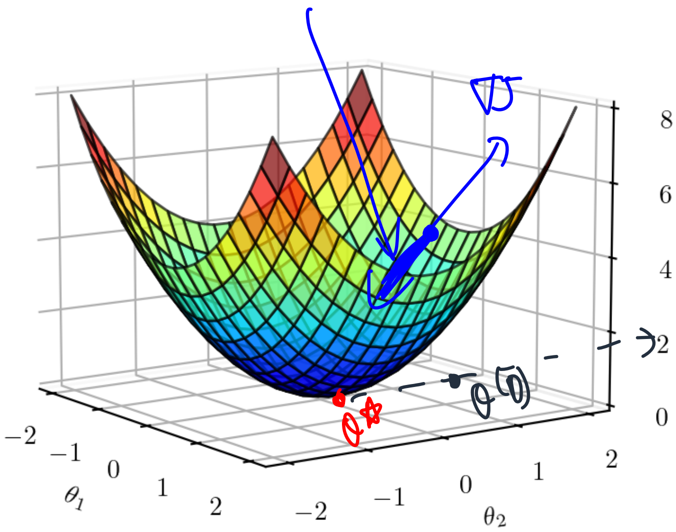
$$\nabla J(\theta^{(0)}) = \begin{bmatrix} \partial J / \partial \theta_1 \\ \partial J / \partial \theta_2 \end{bmatrix}_{\theta = \theta^{(0)}} = \begin{bmatrix} 0 \\ 4 \end{bmatrix}$$

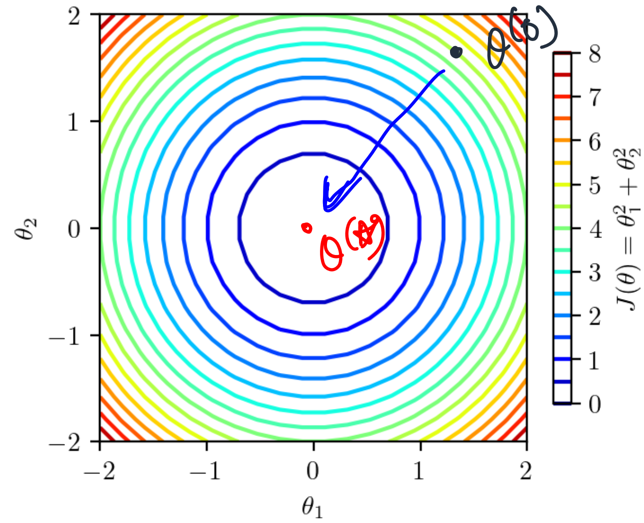
the "GRADIENT" vector

evaluated at $\theta^{(0)}$

IDEA : GRADIENT tells us "steepest" ascent,
so, move in the opposite direction to gradient

movement direction





ITERATION OF STEEPEST "DESCENT"

given, $\theta^{(0)}$ — start

we want $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(K)}$

$$\boxed{G(\theta^*) = \theta^*}$$

$$\theta^{(K+1)} = \theta^{(K)} - \alpha \nabla J(\theta^{(K)})$$

'α' — step size — hyperparameter

let's understand behavior at θ^*

$$G(\theta^{*+1}) = G(\theta^*) - \alpha \nabla J(\theta^*)$$

At θ^*, we have $\nabla J(\theta^*) = 0$

$$\Rightarrow G(\theta^{*+1}) = G(\theta^*) = \theta^*$$

IMPLICIT ASSUMPTION — "J" was once-differentiable,
i.e., ∇J is "possible"

ALPHA CONTROLS HOW FAST!

$0.5 < \alpha < 1 \rightarrow$ "good" convergence rate

$\alpha < 0.1 \rightarrow$ # steps = too much — too long to converge

$\alpha > 1 \rightarrow$ can't get to the minimum — overshoots minimum

$\alpha = 1 \rightarrow$ oscillates!

GOAL: Find "sufficiently small" alpha that gets you the fastest convergence!

CONVERGENCE GUARANTEES

$J(\theta) = \theta_1^2 + \theta_2^2$ — ask what convergence guarantees can we solve?

General form of $J(\theta)$

$$J(\theta) = \frac{1}{2}\theta^T A \theta + b^T\theta + c$$

$A = 2I,\ b = \vec{0},\ c = 0$

$J(\theta) = \theta^T\theta$

Gradient of $J(\theta)$

$$\nabla J = \frac{1}{2}(A + A^T)\theta + b$$

Assume '$\underline{A}$' is symmetric and $J(\theta)$ is actually strictly convex

$A = A^T$

$$\nabla J = A\theta + b$$

$$\theta^{(k+1)} = \theta^{(k)} - \alpha(A\theta^{(k)} + b)$$

$$-\quad \theta^* = \theta^{(*+1)} = \theta^{(*)} - \alpha(A\theta^* + b)$$

$$\underbrace{\theta^{(k+1)} - \theta^*}_{\varepsilon^{(k+1)}} = (\theta^{(k)} - \theta^*) - \alpha(A(\theta^{(k)} - \theta^*))$$

$$= \left[I - \alpha A\right]\underbrace{\theta^{(k)} - \theta^{(*)}}_{\varepsilon^{(k)}}$$

$$\varepsilon^{(k+1)} = \underbrace{[\;\cdots\;]}_{\text{happens to be } < 1} \varepsilon^{(k)}$$

happens to be < 1 → guarantee convergence!

Q. What governs the amount of "stretch" a matrix applies to a vector?

⇒ The EIGENVALUES! or the largest eigenvalue!

⇒ the LARGEST e-v — λ_{max} determines convergence!

$$\boxed{\alpha < \frac{2}{\lambda_{max}}}$$ — convergence guaranteed!

$$\varepsilon^{(k+1)} = [\;\cdots\;]\,\varepsilon^{(k)}$$

↳ (linear convergence!)

For $J(\theta) = \theta_1^2 + \theta_2^2$

$\alpha < 1$ — converge

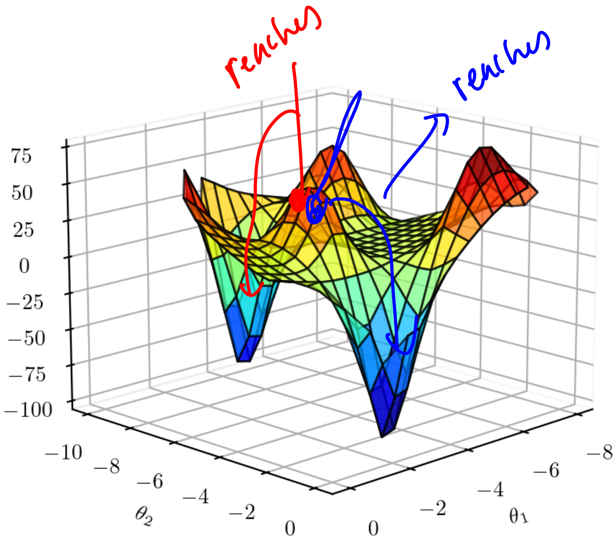
$\alpha = 1$ — oscillate

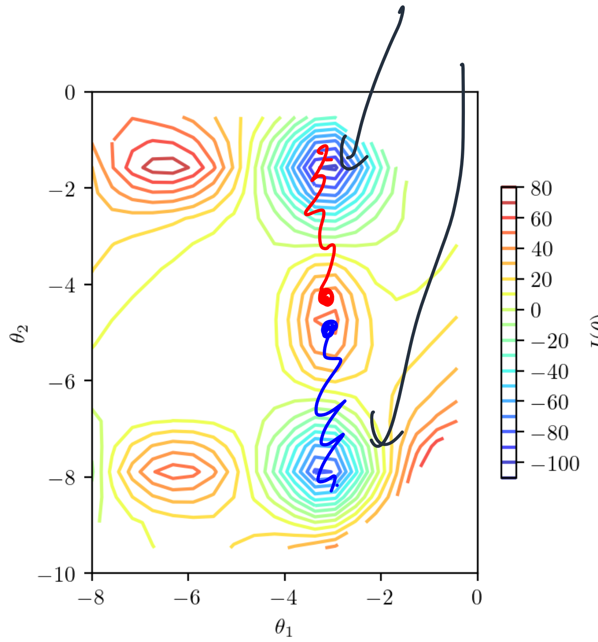
$\alpha > 1$ — diverge

Q. How to choose starting point?

Pradhi – 'yes" – reach same optima
Cindy – "no" – local / not global

both are fixed points of "G"





EN ROUTE TO NEWTON

GRADIENT DESCENT — $\theta^{(1)}, \ldots, \theta^{(k)}$ such that

$$\theta^{*+1} = G(\theta^{*}) = \theta^{*}$$

what happens at global optima?

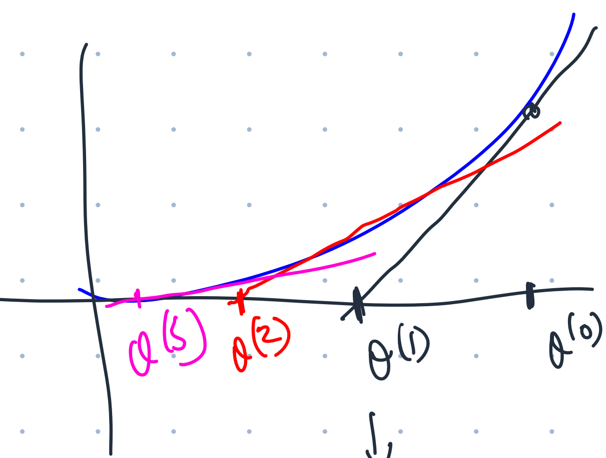
$$\nabla J(\theta^{*}) = 0$$

IDEA — If I have some f, how do find the roots of that function

θ, such that $f(\theta) = 0$

We want to find roots "iteratively"

find solution as follows, starting from $\theta^{(0)}$



→ fit a tangent as my approximation to the function

↓ root of my tangent approximation is the next estimate