ANNOUNCEMENTS

5780 : Quiz-3 out on Canvas, due soon after prelim (03/13)

3780/5780 : HW3 out soon! — due next Fr, 11:59pm (late due : Sun, 11:59pm)

TURN YOUR NON-NOTE-TAKING DEVICES OFF NOW! — lots of fun stuff to cover!

POWERFUL FRAMEWORK

iterative approaches to optize cost $J(\theta)$

given $\theta^{(0)}$, form: $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(k)}$

$$G(\theta^*) = \theta^*$$

GRADIENT DESCENT

move downhill in steepest descent direction

$$\theta^{(k+1)} = \theta^{(k)} + \alpha \left(-\nabla J(\theta^{(k)})\right)$$

how big a step

NO CONVERGENCE PROOFS ON THE EXAM, BUT →

$$\|\varepsilon^{(k+1)}\| = \|(I - \alpha A)\,\varepsilon^{(k)}\|$$

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

strictly convex

$$\alpha < \frac{2}{\lambda_{max}}$$

→ "linear convergence!"

Alex pressed "step" 100 times!

NEWTON'S METHOD

$$G(\theta^*) = \theta^* \quad \text{— iteration @ convergence}$$

At θ^*, $\boxed{\nabla J = 0}$

Reformulation — If we had a func "f", where does $f(x) = 0$ occur?

Approximate function —

$$f(\theta) = f(\theta^k) + f'(\theta^k)(\theta - \theta^k) + \frac{f''(\theta^k)}{2}(\theta - \theta^k)^2 + \cdots$$

around $\theta^{(k)}$

If $\theta \to \theta^k \Rightarrow (\theta - \theta^k)^2$ is small

$$f(\theta) \approx f(\theta^k) + f'(\theta^k)(\theta - \theta^k)$$

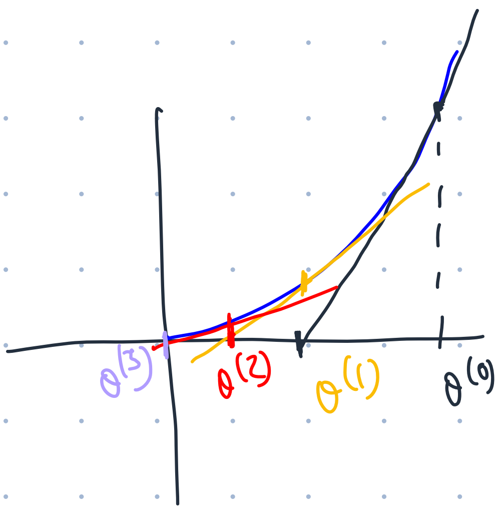
next iterate,

$$f(\theta^{k+1}) \approx f(\theta^k) + f'(\theta^k)(\theta^{k+1} - \theta^k)$$

GOAL : Find θ, such that $f(\theta) = 0$

$$f(\theta^{k+1}) = 0 \Rightarrow -f(\theta^k) = f'(\theta^k)(\theta^{k+1} - \theta^k)$$

$$\Rightarrow \boxed{\theta^{k+1} = \theta^k - \frac{f(\theta^k)}{f'(\theta^k)}}$$



WHAT ARE WE TRYINT TO FIND ROOTS OF?

$$\nabla J = 0$$

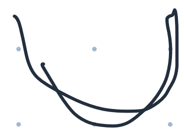
$$\theta^{(k+1)} = \theta^k - \frac{f(\theta^k)}{f'(\theta^k)}$$

for ∇J:

$$\theta^{(k+1)} = \theta^k - \frac{\nabla J(\theta^k)}{\nabla^2 J(\theta^k)} \rightarrow H_J(\theta^{k}) \qquad = \theta^k - H_J(\theta^k)^{-1} \nabla J(\theta^k)$$

$$H_J(\theta^k) = \begin{bmatrix} \frac{\partial^2 J}{\partial \theta_1^2} & \frac{\partial J}{\partial \theta_1 \partial \theta_2} \\ \frac{\partial J}{\partial \theta_2 \partial \theta_1} & \frac{\partial J}{\partial \theta_2^2} \end{bmatrix}$$

$$\|\varepsilon^{(k+1)}\| = C\|\varepsilon^{(k)}\|^2$$

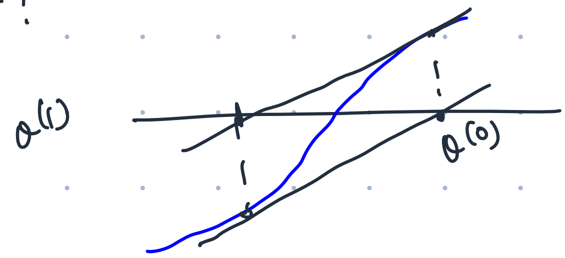


$$J(\theta) = \theta_1^2 + \theta_2^2$$

REASONS WHY NEWTON IS <u>BAD</u>!

1. Inverting H_J is computationally expensive!

2. Can oscillate without convergence!



3. Since Newton isn't a "minimizing" method and more of root-finding, you could also end up in a local maxima!

GRADIENT DESCENT

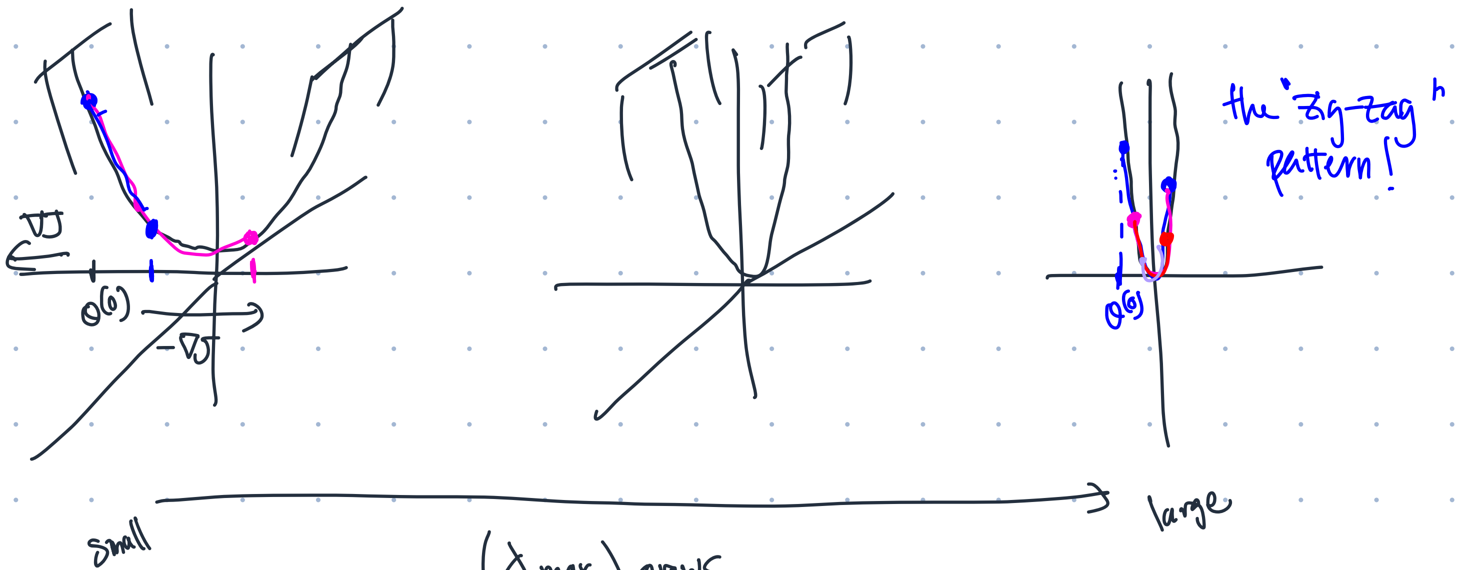
$J(\theta) = \frac{1}{2}\theta^T A \theta$

$\| \varepsilon^{(k+1)} \| = \| (I - \alpha A)\, \varepsilon^{(k)} \|$

$\alpha < 2/\lambda_{max}$ — guarantees convergence!

max convergence @ $\boxed{\alpha = \frac{2}{\lambda_{min} + \lambda_{max}}}$

RATE OF CONVERGENCE $= \frac{\lambda_{max}/\lambda_{min} - 1}{\lambda_{max}/\lambda_{min} + 1}$



$\left(\frac{\lambda_{max}}{\lambda_{min}}\right)$ grows

OPT $\alpha = \frac{2}{\lambda_{min} + \lambda_{max}}$

Guaranteed convergence is NOT good enough

How to get faster, hopefully more stable convergence?

No ZIG-ZAG

IDEA: we're going to model a "heavier ball"!

dampens at each step.

$$v^{(k)} = \beta v^{(k-1)} + \nabla J(\theta^k), \quad \beta < 1 \text{ — dampen}$$

starting with $v^{(0)} = \vec{0}$

"dampens": $\beta v^{(2)} + \beta^2 (v^{(1)}) + \ldots$

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

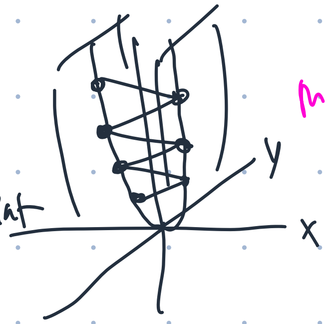
Convergence rate:

$$\frac{\frac{\lambda_{max}}{\lambda_{min}} - 1}{\frac{\lambda_{max}}{\lambda_{min}} + 1} \longrightarrow \frac{\sqrt{\frac{\lambda_{max}}{\lambda_{min}}} - 1}{\sqrt{\frac{\lambda_{max}}{\lambda_{min}}} + 1}$$

10 iterations of GD $\longrightarrow$ 1 iteration with momentum!

"Almost" flat surfaces.

we make little to no progress upon reaching a seemingly flat surface!



moved a "lot" along x, but barely along "y"

$$g_i^{(k)} = g_i^{(k-1)} + \left(\nabla J(\theta^k)[i]\right)^2 \rightarrow$$ collects how much update made along "i"

$\begin{bmatrix} 100 \\ 0.001 \end{bmatrix}$ g_1 really high $= 100^2$

$$\alpha_i \rightarrow \frac{\alpha}{\sqrt{g_i^{(k)}}} \Leftarrow \underline{\underline{\text{AdaGrad}}}$$

Coordinate-wise "α"

$$\theta_i^{k+1} = \theta_i^k - \alpha_i \nabla J(\theta^k)$$
$$= \theta_i^k - \frac{\alpha}{\sqrt{g_i^{(k)}}} \nabla J(\theta^k)$$