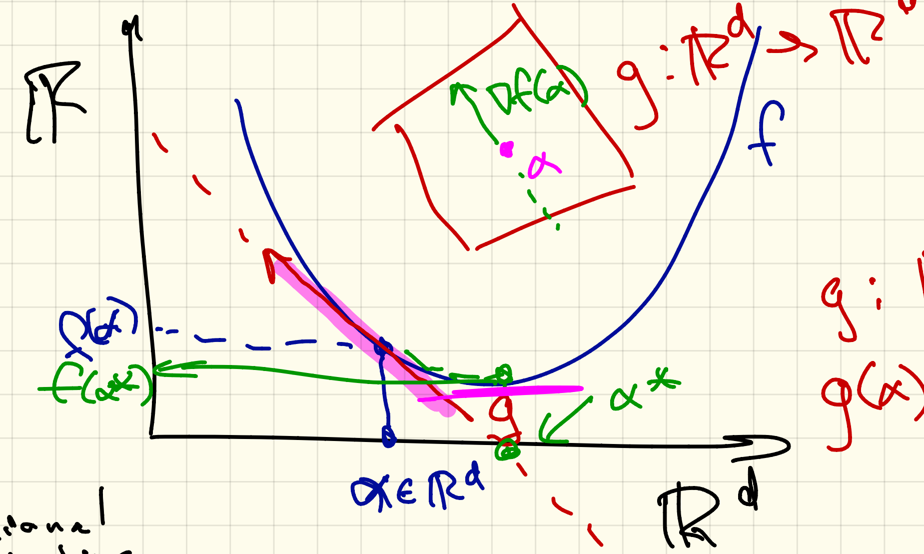
FoDA L16 : Gradient Descent : Algorithms & Convergence

function: $f: \mathbb{R}^d \to \mathbb{R}$

gradient $\nabla f: \mathbb{R}^d \to \mathbb{R}^d$



$g: \mathbb{R}^d \to \mathbb{R}^d$

$g(\alpha) = \langle \nabla f(\alpha), v \rangle$

directional derivative

$$\nabla_v f(\alpha) = \lim_{h \to 0} \frac{f(\alpha + hv) - f(\alpha)}{h} = \langle \nabla f(\alpha), v \rangle$$

Gradient Descent

given function $f : \mathbb{R}^d \to \mathbb{R}$

∇f

Goal

$$\min_{\alpha \in \mathbb{R}^d} f(\alpha) \quad \text{and/or} \quad \alpha^* = \operatorname*{argmin}_{\alpha \in \mathbb{R}^d} f(\alpha)$$

0. Initialize $\alpha^{(0)} = \alpha_{start} \in \mathbb{R}^d$ $\quad k = 0$

1. repeat

gradient @ $\alpha^{(k)}$

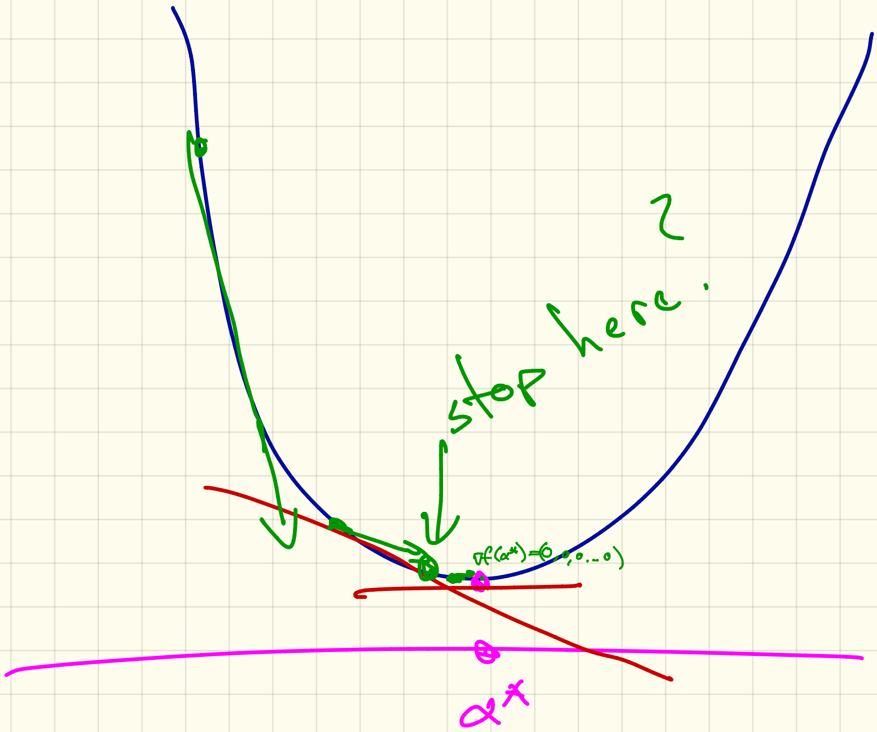
learning rate

$$\alpha^{(k+1)} = \alpha^{(k)} - \gamma_k \nabla f(\alpha^{(k)}) \qquad k = k+1$$

until $\left(\| \nabla f(\alpha^{(k)}) \| \le \tau \quad \text{or} \quad k = T \right)$

2. return $\alpha^{(k)}$ or $f(\alpha^{(k)})$

stopping condition


stop here?
∇f(α*) = (0, 0, 0 ... 0)
α*

Learning Rate

$$\alpha = \alpha - \gamma \nabla f(\alpha)$$

learning rate

L-Lipschitz bound (on gradient $g = \nabla f$)

$g : \mathbb{R}^d \to \mathbb{R}^k \qquad \forall\, p, q \in \mathbb{R}^d$

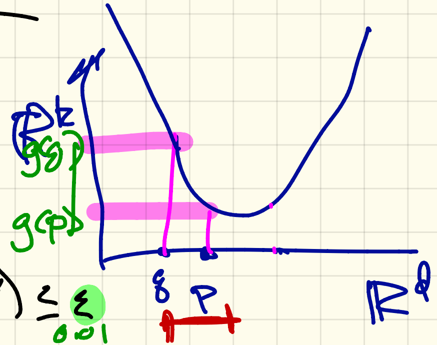
$$\| g(p) - g(q) \| \leq L \| p - q \|$$

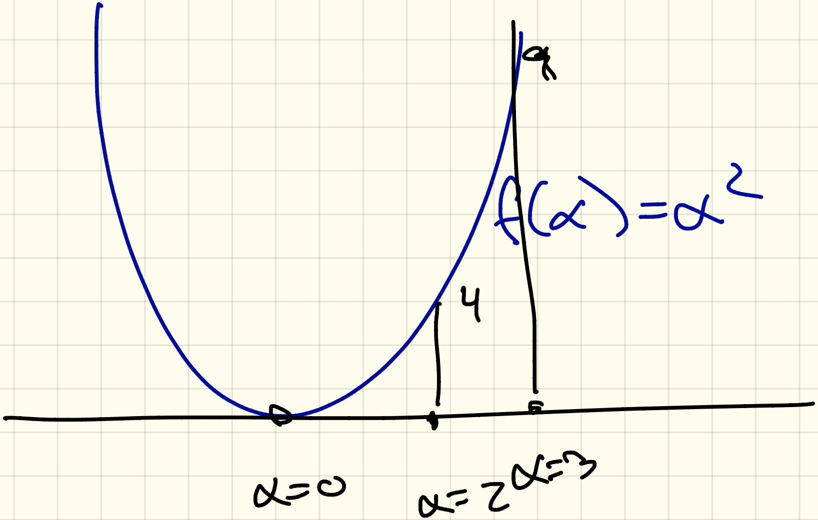
depend on $\alpha^{(0)}$

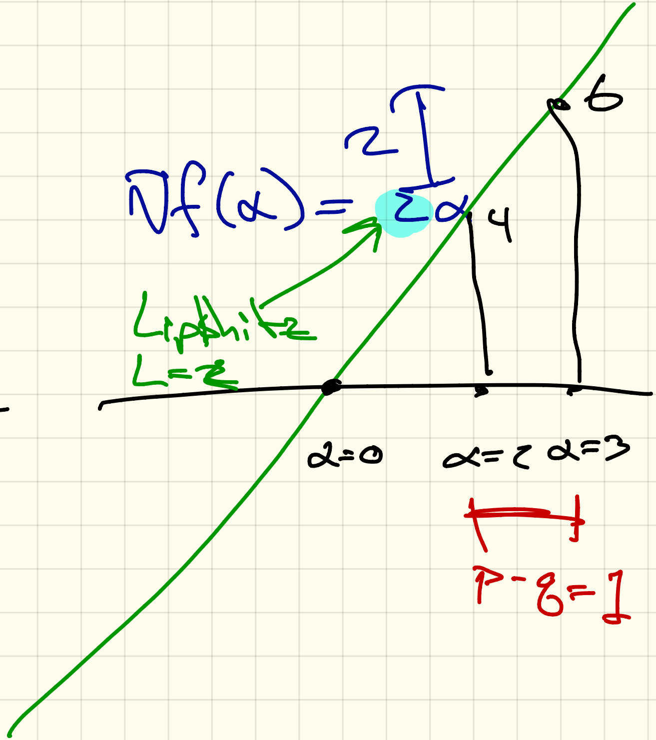
If $g = \nabla f$ is L-Lipschitz

- set $\gamma \leq \frac{1}{L} \Rightarrow$ GD converge to stationary point
- if f convex $\Rightarrow$ converge to α^* after $t = c/\varepsilon$ steps $f(\alpha^{(t)}) - f(\alpha^*) \leq \varepsilon$

$>100 \qquad 0.01$







$$f(\alpha) = 7\alpha^2$$

$$\nabla f(\alpha) = 14\alpha \implies L = 14$$

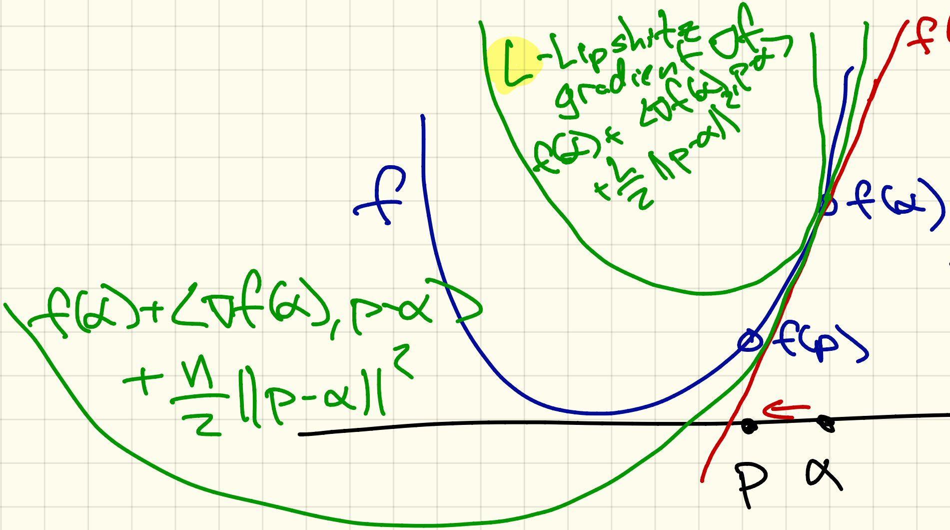
Strongly Convex Functions

$f: \mathbb{R}^d \to \mathbb{R}$ $\quad \eta$-strongly convex

eta

$\forall \alpha, p \in \mathbb{R}^d$

$$f(p) \geq f(\alpha) + \langle \nabla f(\alpha), p-\alpha \rangle + \frac{\eta}{2}\|p-\alpha\|^2$$



and L-Lipschitz gradient ∇f

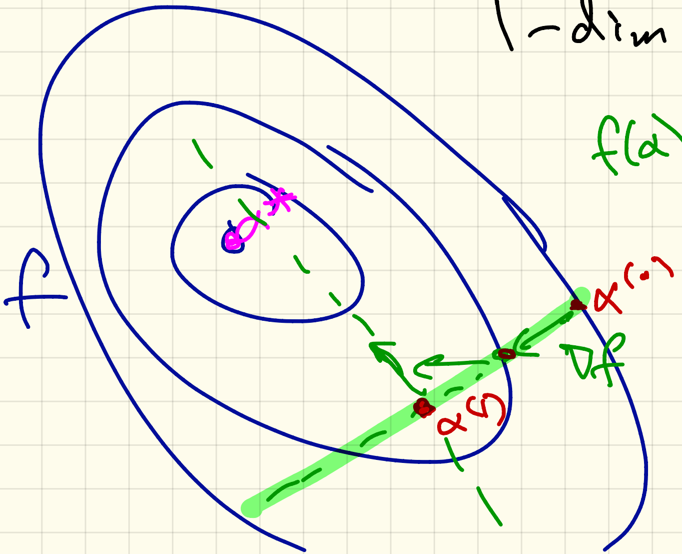
Set $\gamma = \frac{2}{\eta + L}$

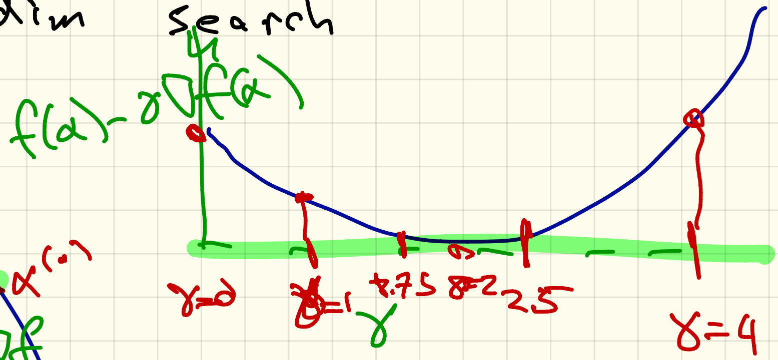
$k = C \cdot \log \frac{1}{\varepsilon}$ steps

$f(\alpha^{(k)}) - f(\alpha^*) \leq \varepsilon$

How to set Learning Rate?

Line Search: after choosing ∇f
reduced d-dim search to
1-dim search





Adjustable Rate

back-tracking line search

$\beta \in (0.1, 0.8)$ $\quad \beta = 0.75$

$\gamma_0 = $ large $\quad \gamma_{i+1} = \gamma_i \cdot \beta$ sometimes

when $f(\alpha - \gamma \nabla f(\alpha)) \geq f(\alpha) - \frac{\gamma}{2}\|\nabla f(\alpha)\|^2$