FoDA L16

Gradient Descent

#2

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

$\nabla f : \mathbb{R}^d \to \mathbb{R}^d$ $\qquad f(\alpha) \quad \alpha = (\alpha_1 \ldots \alpha_d)$

$$\nabla f = \left(\frac{\partial f}{\partial \alpha_1}, \frac{\partial f}{\partial \alpha_2}, \ldots, \frac{\partial f}{\partial \alpha_d} \right)$$

$u = (u_1, u_2 \ldots u_d)$

$(u_1 \quad u_2, \ldots \quad u_d)$

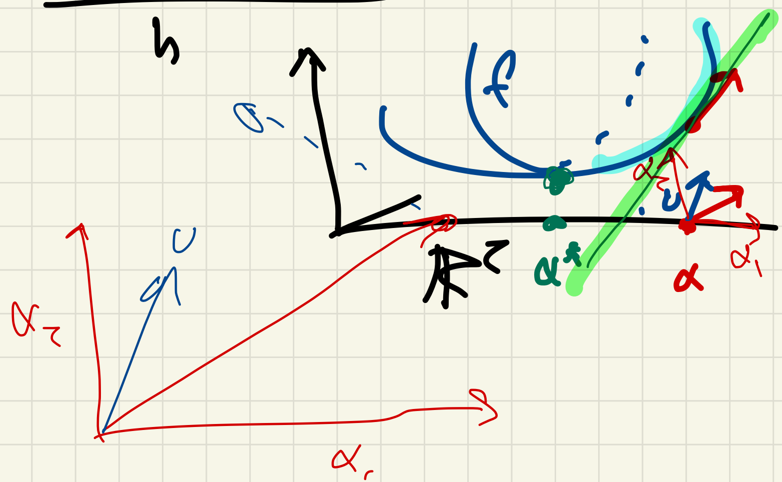
directional derivative $\nabla_u f(\alpha) = \lim_{h \to 0} \frac{f(\alpha + hu) - f(\alpha)}{h}$

$\|u\| = 1$

$$\nabla_u f(\alpha) = \langle \nabla f(\alpha), u \rangle$$

$\max_{\substack{u \\ \|u\|=1}} \langle \nabla f(\alpha), u \rangle$

$\hookrightarrow u = \frac{\nabla f(\alpha)}{\|\nabla f(\alpha)\|}$



Gradient Descent Algo

Goal: $\min_{\alpha \in \mathbb{R}^d} f(\alpha)$ or $\operatorname{argmin}_{\alpha \in \mathbb{R}^d} f(\alpha)$

guess

Initialize $\alpha^{(0)} = \alpha_{start}$, $k = 0$

repeat

learning rate

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

$k = k+1$

until $\left(\|\nabla f(\alpha^{(k)})\| < \tau \right)$

threshold

(or T steps)

Stopping condition

\# steps T (time constrained)

$\|\nabla f(\alpha)\| < T$

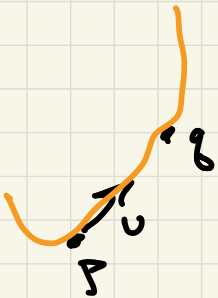
at optimal α^*

$\nabla f(\alpha^*) = (0, 0, \ldots 0)$

$\|\nabla f(\alpha^*)\| = 0$

Learning Rate γ how to choose?

Lipschitz function $g : \mathbb{R}^d \to \mathbb{R}^k$ is L-Lipschitz



if $\forall\, p, q \in \mathbb{R}^d$ $\quad \|g(p) - g(q)\| \le L \|\overbrace{p-q}^{h}\|$

$u = \dfrac{p-q}{\|p-q\|}$ $\quad q = p + uh$ $\quad \dfrac{\|g(p) - g(p+uh)\|}{h} \le L$

Let $\nabla f = g$

Assume ∇f is L-Lipschitz

Set $\gamma < \frac{1}{L}$ then

GD will converge if convex f

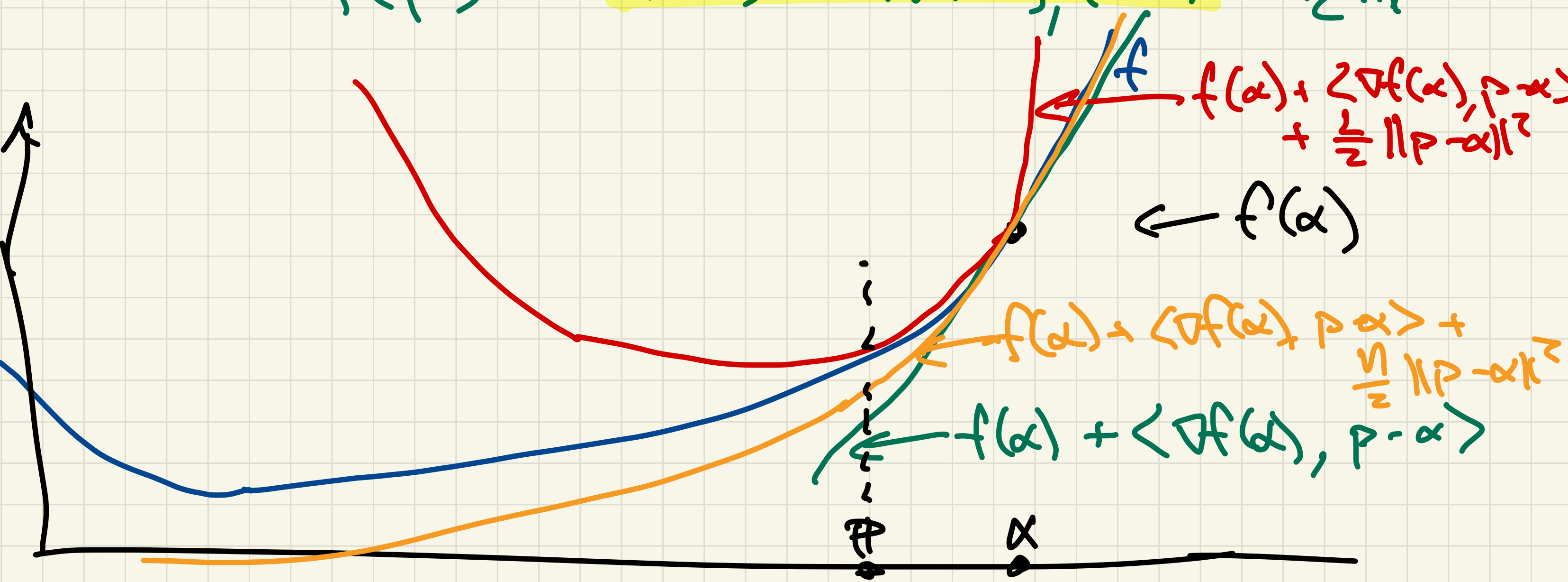
after $k = c/\varepsilon$ steps

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

Strongly convex function (strict)

f η-strongly convex $\quad \forall \alpha, p \in \mathbb{R}^d$

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



If f is η-strongly convex, ∇f L-Lipschitz,

set $\gamma \le \frac{2}{(L+\eta)}$ after $k > C \cdot \log \frac{1}{\varepsilon}$

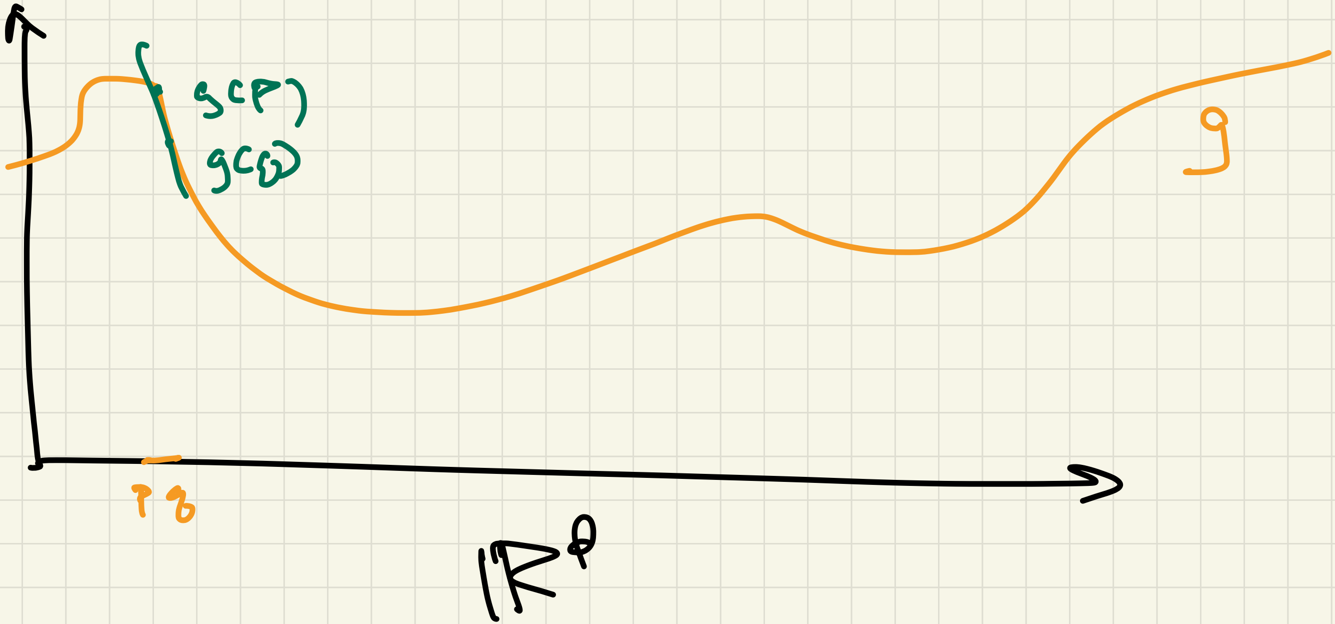
$$f(x^{(k)}) - f(x^*) \le \varepsilon$$

Imagine $C = 1$, $\log = \log_{10}$

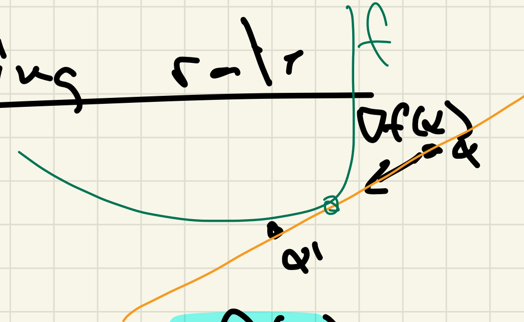
error $\varepsilon = \overbrace{0.00 \ldots 0}^{k} xxx$

Lipschitz

g



Adjusting γ ← learning rate



Line Search

Every step $\alpha' = \alpha - \gamma \nabla f(\alpha)$

goal minimize $f(\alpha')$

$$\gamma^* = \underset{\gamma \in \mathbb{R}}{\arg\min}\, f(\alpha - \gamma \nabla f(\alpha))$$

binary search

Often too slow

Adjustable Rate

Start w/ $\gamma = \gamma_0$ a bit larger than guess

sometimes $\gamma = \gamma\beta$ $\quad \beta \in (0.1, 0.8)$

$\beta = 0.75$

Condition

if $\boxed{f(\alpha - \gamma \nabla f(\alpha)) > f(\alpha) - \frac{\gamma}{2}\|\nabla f(\alpha)\|^2}$

then $\gamma = \beta\gamma$