L15: Gradient Descent

Mar 2, 2026

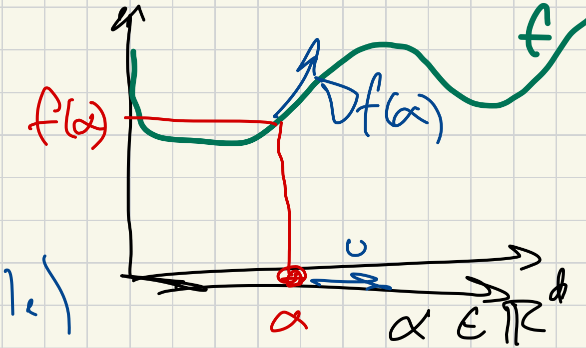
FoDA

Jeff M. Phillips

Gradients

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

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_d}\right)$$



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

direction (and scale) of steepest increase

H v = direction of steepest increase $\|v\| = 1$

$$\nabla_v f(x) = \langle \nabla f, v \rangle$$

$$\to \nabla_v f(x) = \|\nabla f\| \quad \text{amount increase}$$

$$v = \frac{\nabla f}{\|\nabla f\|}$$

Example

$$\alpha = (\alpha_0 = x,\ \alpha_1 = y,\ \alpha_2 = z) \in \mathbb{R}^3$$

$$f(\overbrace{x, y, z}^{\alpha}) = 3x^2 - 2y^3 - 2xe^z$$

$$\nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \right) = \left(6x - 2e^z, -6y^2, -2xe^z \right)$$

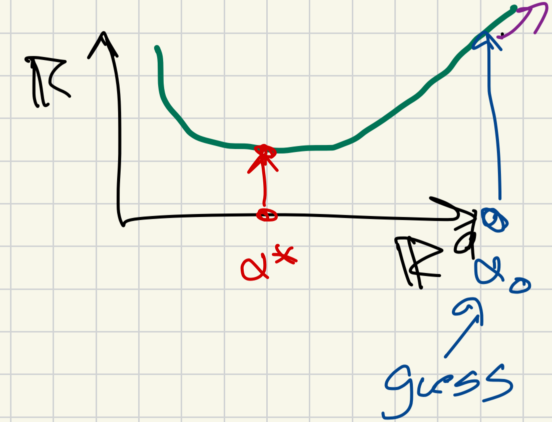
$$\nabla f(3, -2, 1) = (18 - 2e, -24, -6e)$$

$$\alpha = (3, -2, 1)$$

Gradient Descent

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

Goal $\min\limits_{\alpha \in \mathbb{R}^d} f(\alpha)$



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

1. **repeat**

learning rate

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

until $\Big(\underbrace{k = T}_{\text{enough steps}}$ or $\underbrace{\|\nabla f(\alpha^{(k)})\| < \tau}_{\text{improvement small enough}} \Big)$

2. $\boxed{\text{return } \alpha^{(k)}}$

Simple GD

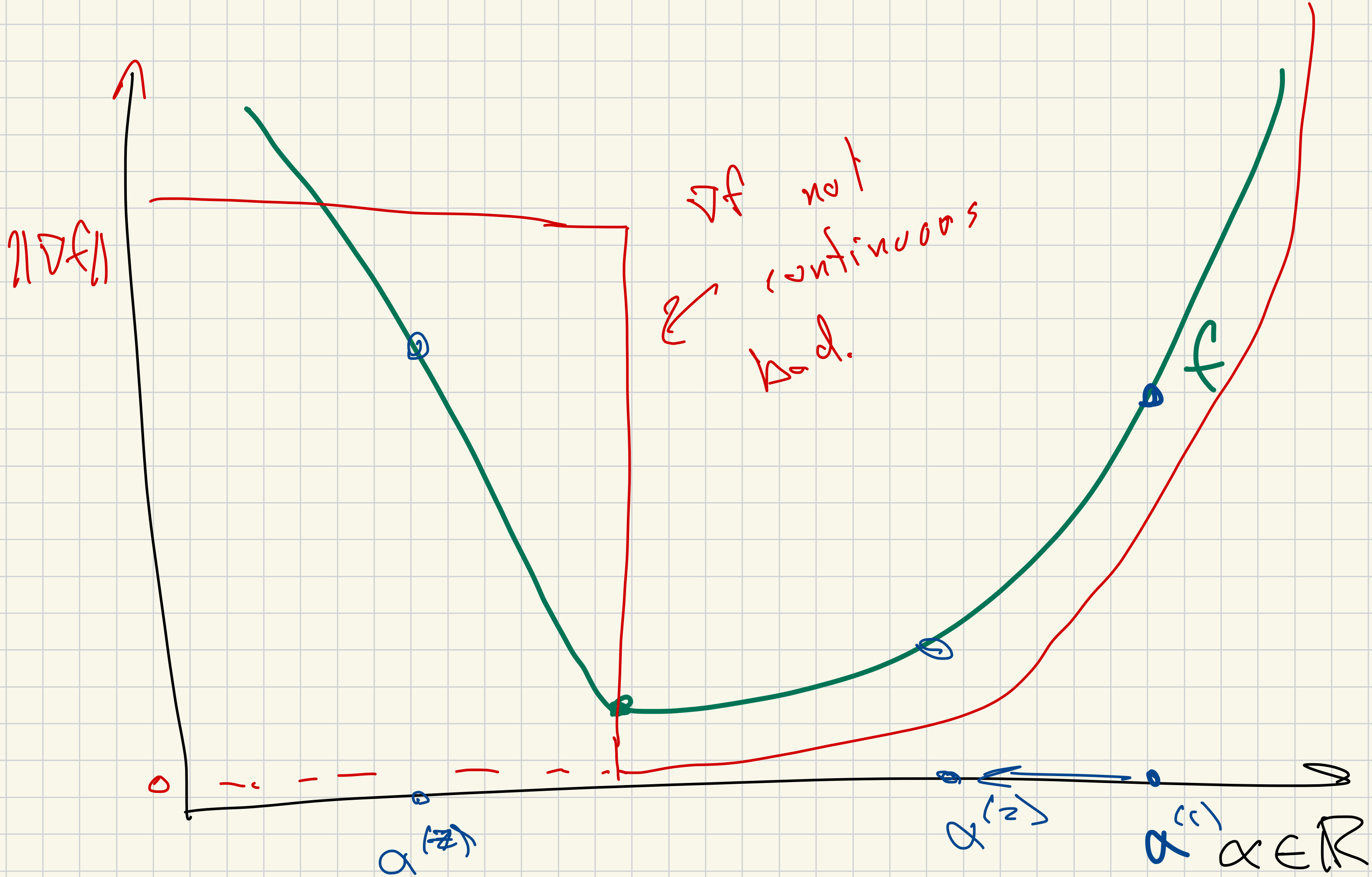
$\alpha = (0, 0 \ldots 0) \in \mathbb{R}^d$

repeat

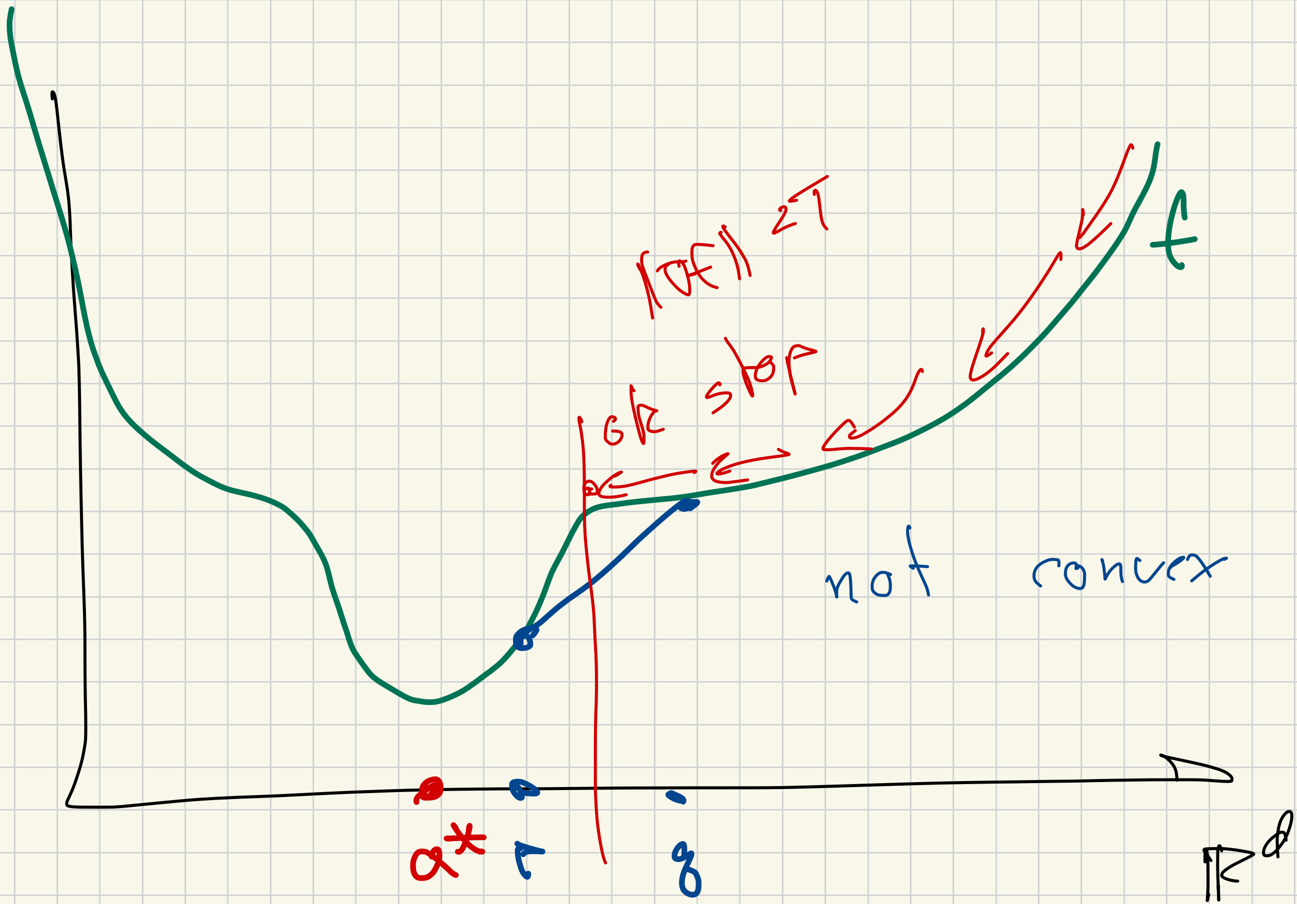
$\alpha = \alpha - \gamma \nabla f(\alpha)$

until $(\|\nabla f(\alpha)\| \leq T)$

return α



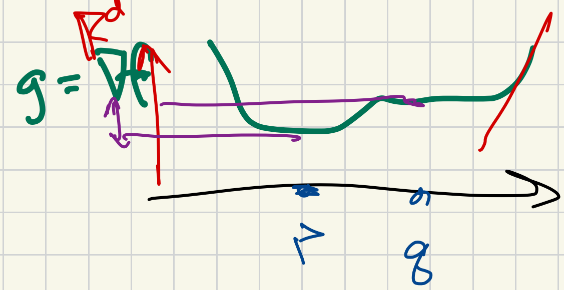
‖∇f‖
∇f not continuous and.
f
0
α^(2)
α^(1)
α ∈ ℝ



‖∇f‖ ≥ T
GR stop
f
not convex
a*
r
q
ℝ^d

if convex? $\rightarrow$ eventually get to minimum
if Learning rate is
not too high!

Learning Rate?



L-Lipschitz function

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

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

---

if $g = \nabla f$ is L-Lipschitz set LR $\gamma < \frac{1}{L}$

↳ GD on f, converge to stationary point.

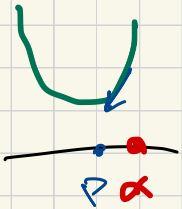
if f is also convex → global min.

run for $k = C/\varepsilon$ steps then $f(x^{(k)}) - f(x^*) \le \varepsilon$

const.

Strongly convex functions

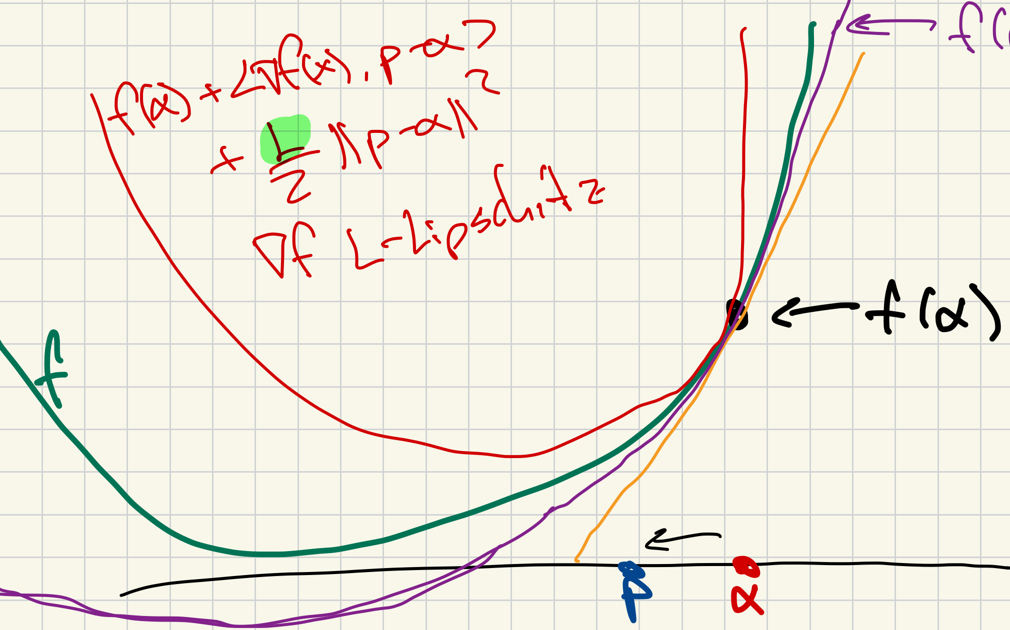
$f : \mathbb{R}^d \to \mathbb{R}$ η-strongly convex if



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

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

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



∇f L-Lipschitz

f - η-strongly convex

$$\gamma \leq \frac{2}{L+\eta}$$

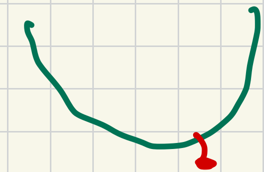
after $k = C \cdot \log(1/\varepsilon)$ steps

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

linear convergence.

How to <u>auto-adjust</u> the learning rate

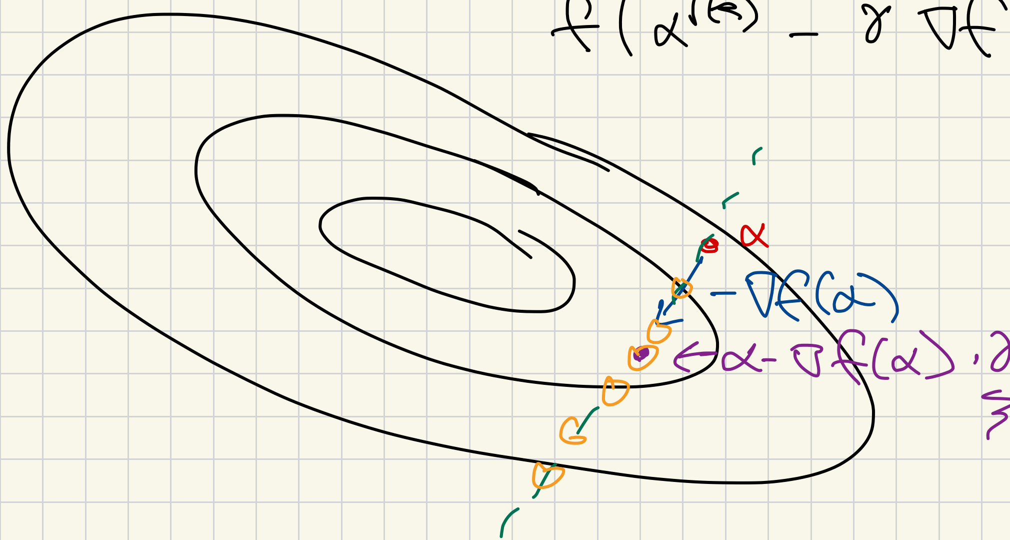
$$\alpha^{(k+1)} = \alpha^{(k)} - \gamma_k \nabla f(\alpha^{(k)})$$



(1) line search

at step k: choose γ so

$$f\left(\alpha^{(k)} - \gamma \nabla f(\alpha^{(k)})\right)$$ small as possible



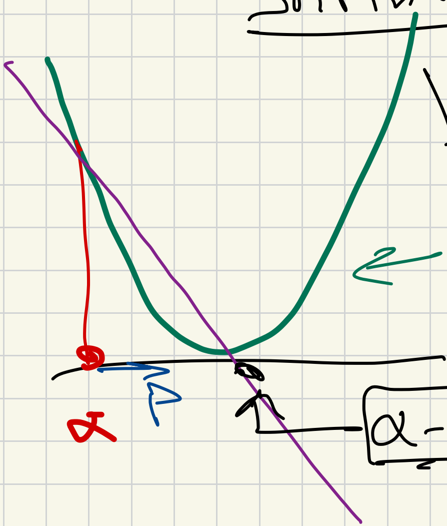
② backtracking

start γ too large $\quad \beta = 0.75$

Repeatedly $\gamma = \beta \gamma$

shrink if

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



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

$$\boxed{\alpha - \gamma \nabla f(\alpha)}$$