L1. Minimization and saddle point problem

- Non-convex minimization

$$\min_{x \in \mathbb{R}^d} f(x)$$

$O(\varepsilon^{-2})$ steps to find ε-stationary pt

assump. on f : "nothing but C^1 "

> Reference: Section 1.2.3
> [N2018]Nesterov. *Introductory lectures on convex optimization*. Vol. 137. Berlin: Springer International Publishing, 2018.

(A1) f is C^1 on $\mathbb{R}^d$, l-smooth
i.e. ∇f is l-Lipschitz

(A2) $\min_x f(x) = f^*$ finite, i.e. f is lower bounded below on $\mathbb{R}^d$.

<u>GD</u> (Gradient Descent)

$$x_{k+1} = x_k - s \nabla f(x_k)$$

No other information, f may have local minima, and generally GD only finds a stationary pt.

Rk. Without (A2), GD may diverge

e.g. $f(x,y) = xy$, along $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$, $f \to -\infty$.

define $\Delta_f := f(x_0) - f^* \geq 0$.

Thm If $0 < s < \frac{2}{l}$, then for $N = O(\frac{1}{\varepsilon^2})$, $\exists k \leq N$

s.t. $\|\nabla f(x_k)\| \leq \varepsilon$.

pf. idea: make $f(x_k) \Downarrow$, then bcz f has a lower bound, it can not always make progress.

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{l}{2} \|x_{k+1} - x_k\|^2$$

$$f(x_{k+1}) \le f(x_k) + \langle \nabla f(x_k), x_{k+1}-x_k \rangle + \frac{l}{2}\|x_{k+1}-x_k\|^2$$

by Taylor expansion and l-smoothness of f

$$x_{k+1} - x_k = -s\nabla f(x_k)$$

$$\Rightarrow f(x_{k+1}) - f(x_k) \le -s\|\nabla f(x_k)\|^2 + \frac{l}{2}s^2\|\nabla f(x_k)\|^2$$

$$= -s\underbrace{\left(1-\frac{ls}{2}\right)}_{>0 \text{ if } \frac{ls}{2}<1}\|\nabla f(x_k)\|^2$$

suppose $s = \frac{2}{l}\alpha$, $0<\alpha<1$, then

$$f(x_{k+1}) - f(x_k) \le -\frac{2\alpha(1-\alpha)}{l}\|\nabla f(x_k)\|^2$$

telescopic sum $k=0,\cdots,N$

$$\frac{2\alpha(1-\alpha)}{l}\sum_{k=0}^{N}\|\nabla f(x_k)\|^2 \le f(x_0) - f(x_{N+1}) \le \Delta_f$$

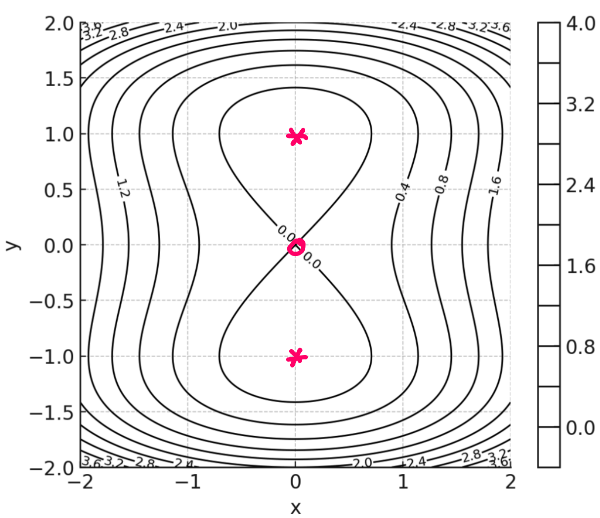
$\Rightarrow \exists\ k \le N$ s.t.

$$\|\nabla f(x_k)\|^2 \le \frac{l\Delta_f}{2\alpha(1-\alpha)}\cdot\frac{1}{N+1} \sim \frac{1}{N} \qquad \#$$

Rk. The $O(\varepsilon^{-2})$ iteration complexity is tight, to find ε-stationary pt is $\Omega(\varepsilon^{-2})$ grad evaluations.

[CDHS2017] Carmon, Duchi, Hinder, Sidford. Lower bounds for finding stationary points. arXiv:1710.11606, 2017.

Eg. $f(x,y) = \frac{1}{2}x^2 + \frac{1}{4}y^4 - \frac{1}{2}y^2$

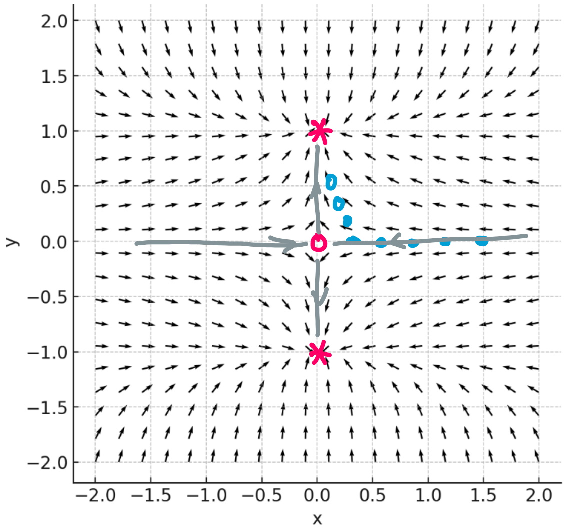


Gradient Descent Directions for f(x,y)

3 stationary pts

$(0,\pm 1)$ local min

$(0,0)$ saddle pt



Starting GD from

$x_0 = (1,0)$

in theory $x_k \to (0,0)$

Rk. finding saddle pt is not easy in practical computation

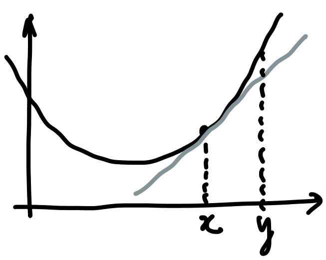
- convex minimization

$$\min_x f(x) \qquad f \text{ is } C^1 \text{ on } \mathbb{R}^d$$

$f(x)$ is convex on $\mathbb{R}^d$

<u>fact</u> If f is C^1 and convex, then

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \ \forall x, y \in \mathbb{R}^d.$$



"above the tangent line"

<u>lemma</u> If f is C^1 and convex, $\nabla f(x^*) = 0$, then

$$f(x^*) = \min_x f(x),$$

i.e. x^* is a global minimum. (may not unique)

pf. $\forall x$, $f(x) \geq f(x^*) + \langle \cancel{\nabla f(x^*)}^{\,0}, x - x^* \rangle$ #.

<u>Prop</u> f is C^1, then f is convex iff.

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0, \ \forall x, y.$$

Rk. The field $F(x) = \nabla f(x)$ is called <u>monotone</u>.

pf. "$\Rightarrow$" $\left.\begin{array}{l} f(x) - f(y) \geq \langle \nabla f(y), x-y \rangle \\ f(y) - f(x) \geq \langle \nabla f(x), y-x \rangle \end{array}\right\} \Rightarrow$ adding the two.

"$\Leftarrow$" ☺

$$F(t) := f(x+t(y-x)), \quad F(1) = f(y), \; F(0) = f(x)$$

$$f(y) - f(x) = \int_0^1 F'(t)\, dt$$

$$= \int_0^1 \langle \underbrace{\nabla f(\overbrace{x+t(y-x)}^{x(t)})}_{\nabla f(x) + \nabla f(x(t)) - \nabla f(x(0))}, y-x \rangle\, dt$$

$$= \langle \nabla f(x), y-x \rangle$$

$$+ \int_0^1 \underbrace{\langle \nabla f(x(t)) - \nabla f(x(0)), \frac{x(t)-x(0)}{t} \rangle}_{\geq 0} dt$$

#

Prop. f is C^2 on $\mathbb{R}^d$. f is convex iff $\nabla^2 f \succeq 0$.

pf (Ex) i.e. $\nabla^2 f(x) \succeq 0, \forall x$.

- strongly convexity

def f is C^1 on $\mathbb{R}^d$, f is μ-strongly convex if ($\mu > 0$, <u>μ-S.C.</u>)

$$f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle + \frac{\mu}{2} \|y-x\|^2, \; \forall x, y.$$

Prop. f is C^1 and μ-S.C., then

$$\langle \nabla f(x) - \nabla f(y), x-y \rangle \geq \mu \|x-y\|^2, \quad \forall x, y.$$

Rk. $F(x) = \nabla f(x)$ is <u>μ-strongly monotone</u>.

Prop f is C^2 on $\mathbb{R}^d$, f is μ-S.C. iff $\nabla^2 f \succeq \mu I_d$.

lemma f is C^1 and μ-S.C. on $\mathbb{R}^d$, then $\min_x f(x)$ is attained at a unique (global) minimizer.

pf. First use "quadratic growth" to show that level set is compact, then $\min_x f(x)$ is attained within a compact set. (Q.G.)

$$QG: \quad f(y) \geq f(x_0) + \langle \nabla f(x_0), y - x_0 \rangle + \frac{\mu}{2} \|y - x_0\|^2$$

$$\sim \|y - x_0\|^2 \quad \text{as } \|y\| \to \infty.$$

Then use μ-S.C. to prove uniqueness of x^* #

Thm If f is C^1, l-smooth, μ-S.C, then GD converges exponentially fast (or "linear convergence"), i.e (Rk. $l \geq \mu$)

1) $f(x_k) \leq f^* + \varepsilon$ (function value convergence)

2) $\|\nabla f(x_k)\| \leq \varepsilon$ (first-order optimality)

3) $\|x_k - x^*\| \leq \varepsilon$ (variable convergence) "identification of x^* parameters"

within $k = O(\log \frac{1}{\varepsilon})$ steps, assuming $\delta < \frac{2\mu}{l^2} = \frac{1}{\kappa}\frac{2}{l}$

Rk $\kappa := \frac{l}{\mu}$ is called condition number, $\kappa \geq 1$.

pf. x^* $\exists!$ by S.C.

3) $\Rightarrow$ 2): $\|\nabla f(x_k) - \cancel{\nabla f(x^*)}^{0}\| \leq l \|x_k - x^*\|$ by l-smoothness.

3) $\Rightarrow$ 1): $\underset{f^*}{f(x^*)} \geq f(x_k) + \underbrace{\langle \nabla f(x_k), x^* - x_k \rangle}$

$$\leq \underbrace{\|\nabla f(x_k)\|}_{\text{by 2)} \to \leq \varepsilon} \underbrace{\|x_k - x^*\|}_{\leq \varepsilon} \leq \varepsilon^2$$

function value actually achieves $O(\varepsilon^2)$ approximation.

We thus only prove 3).

Recall that $\quad x_{k+1} = x_k - s\nabla f(x_k)$.

$$x_{k+1} - x^* = x_k - x^* - s\nabla f(x_k)$$

$$\Rightarrow \|x_{k+1} - x^*\|^2 = \|x_k - x^*\|^2 - 2s\underbrace{\langle \nabla f(x_k), x_k - x^* \rangle}_{\text{by monotonicity}} + s^2 \underbrace{\|\nabla f(x_k)\|^2}_{\text{bdd.}}$$

$$\langle \nabla f(x_k) - \cancel{\nabla f(x^*)}^{0}, x_k - x^* \rangle \geq \mu \|x_k - x^*\|^2$$

$$\|\nabla f(x_k) - \cancel{\nabla f(x^*)}^{0}\| \leq \ell \|x_k - x^*\|$$

$$\leq \|x_k - x^*\|^2 - 2s\mu \|x_k - x^*\|^2 + s^2 \ell^2 \|x_k - x^*\|^2$$

$$= \underbrace{(1 - 2s\mu + s^2\ell^2)}_{<1\,?} \|x_k - x^*\|^2$$

when $0 < 2s\mu - s^2\ell^2$

$\Leftrightarrow s < \frac{2\mu}{\ell^2}$ #

Rk. The proof only uses that $\nabla f(x)$ is strongly-monotone (in addition to ℓ-smoothness of f). See L2.