L5. NC-PL: two-timescale GDA

$$\min_x \max_y L(x,y)$$

NC-$\underset{\uparrow}{\underline{SC}}$: $\overset{\text{weakly convex}}{\text{noncovex}}$ in x, strongly concave in y

more generally can be PL

- GDA with properly set (two scales of) time steps

  $\underbrace{\text{converges}}_{\downarrow}$ in $O(\frac{1}{\varepsilon^2})$ steps

  find a "stionary pt" with $O(\varepsilon)$ gradient

Q. Why $O(\frac{1}{\varepsilon^2})$?

Recall that this is the iteration complexity of vanilla GD

on nonconvex minimization, and it is tight (see L1)

Thus, suppose we use that $\max_y$ is SC and can be

solved fast (in $O(\log \frac{1}{\varepsilon})$ inner-loop steps) to "eliminate"

the variable, then we have the minimization problem

$$\min_x \Phi(x), \quad \Phi(x) := \max_y L(x,y),$$

and this still needs $O(\frac{1}{\varepsilon^2})$ outer-loop steps.

In this sense, the $O(\frac{1}{\varepsilon^2})$ iteration of GDA for the

minimax problem is also "tight".

- Below, we first prove the $O(\frac{1}{\varepsilon^2})$ convergence of GDA for

  the NC-PL type. The variants:

  $\{$ A-GDA   Alternative-GDA

1 S-GDA Smoothed-GDA (with proximal step)

[LJJ 2020] Lin, Jin, Jordan. On gradient descent ascent for nonconvex-concave minimax problems. ICML 2020.
[YOLH 2022] Yang, Orvieto, Lucchi, He. Faster single-loop algorithms for minimax optimization without strong concavity. AISTATS 2022.

- PL condition (Polyak-Łojasiewicz)

$$\min_x h(x) = h^* \quad . \ h \ C^1 \text{ on } \mathbb{R}^d.$$

def h is μ-PL if

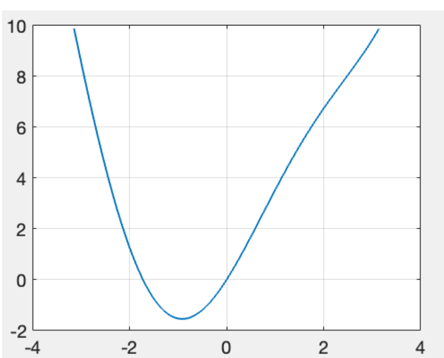
$$\frac{1}{2}\|\nabla h(x)\|^2 \geq \mu(h(x) - h^*), \quad \forall x \in \mathbb{R}^d.$$

lemma h is μ-$\underbrace{\text{strongly convex}}_{\text{S.C.}} \Rightarrow \mu$-PL

fact. h can be μ-PL without μ-S.C.

Eg. non-convex but PL.

$$h(x) = x^2 + 3\sin x$$

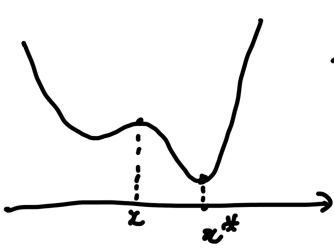


$$h' = 2x + 3\cos x$$
$$x^* \approx -0.9148$$

$$h'' = 2 - 3\sin x, \text{ can } < 0.$$

$$\frac{\frac{1}{2}|h'(x)|^2}{h(x) - h^*} \geq \mu \approx 0.416$$

Q. What does NOT satisfy PL?



$h'(x) = 0, \ h(x) - h^* > 0$

Eg. Moreau envelope

$$\ell(x) = -\frac{\rho}{2} x_2^2 , \qquad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\nabla^2 \ell = \begin{bmatrix} 0 & 0 \\ 0 & -\rho \end{bmatrix} \succeq -\rho I. \qquad \rho\text{-weakly convex}$$

$$\ell_\eta(x) = \max_z \underbrace{\ell(z) + \frac{\eta}{2}\|z - x\|^2}_{L_{\eta,x}(z)} = L(z)$$

$\underline{Q}$ Is $L_{\eta,x}(z)$ PL ?

- if $\eta > \rho$, $L(z)$ is $(\eta-\rho)$- S.C.

$$L(z) = -\frac{\rho}{2} z_2^2 + \frac{\eta}{2}(z_1 - x_1)^2 + \frac{\eta}{2}(z_2 - x_2)^2$$

- if $\eta < \rho$, no minimiser on $\mathbb{R}^2$

$\eta = \rho$, let $x_2 = 0$,

$$L(z) = \frac{\eta}{2}(z_1 - x_1)^2 , \qquad \nabla^2 L = \begin{bmatrix} \eta & 0 \\ 0 & 0 \end{bmatrix}$$

convex but not S.C.

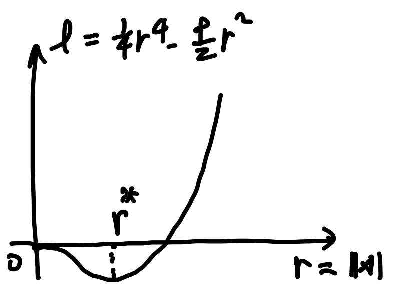
$$\nabla L(z) = \begin{bmatrix} \eta(z_1 - x_1) \\ 0 \end{bmatrix}, \quad L^* = 0$$

$$\frac{\frac{1}{2}\|\nabla L(z)\|^2}{L(z) - L^*} = \frac{\frac{1}{2}\eta^2 (z_1 - x_1)^2}{\frac{\eta}{2}(z_1 - x_1)^2} = \eta > 0 . \quad \forall z$$

Thus, L is convex, non- S.C., but μ-PL.

Another example:

$$\ell(x) = \frac{1}{4}\|x\|^4 - \frac{\rho}{2}\|x\|^2 \qquad \rho\text{- weakly convex}$$



$$L(z) = \frac{1}{4}\|z\|^4 - \frac{\rho}{2}\|z\|^2 + \frac{\eta}{2}\|z - x\|^2 , \quad x \text{ fixed}$$

$$\nabla L(z) = (\|z\|^2 - \rho) z + \eta (z - x)$$

suppose $x = 0$, then $L(z) = \frac{1}{4}\|z\|^4 - \frac{1}{2}(\rho-\eta)\|z\|^2$.

$$L(z) - L^* = \frac{1}{4}\left(\|z\|^2 - (\rho-\eta)\right)^2.$$

$$\nabla L(z) = (\|z\|^2 - (\rho-\eta))\|z\|$$

$$\Rightarrow \frac{\frac{1}{2}\|\nabla L(z)\|^2}{L(z) - L^*} = 2\|z\|^2 \to 0+ \text{ when } z \to 0$$

locally PL (on a nbh) but not globally

- PL $\Rightarrow$ linear convergence in minimization

[KNS 2016] Karimi, Nutini, Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. (2016)

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

(PL) $\quad \frac{1}{2}\|\nabla h(x)\|^2 \geq \mu (h(x) - h^*), \ \forall x$

(EB) $\quad \|\nabla h(x)\| \geq \mu \|x - x^*\|, \ \forall x$

error bound

generally $\|x - x_p\|$, x_p is the projection of x to the minimizer set $\mathcal{X}^*$

(QG) $\quad h(x) - h^* \geq \frac{\mu}{2}\|x - x^*\|^2, \ \forall x$

quadratic growth

Thm. If h is C^1 and l-smooth on $\mathbb{R}^d$, then

$$(SC) \Rightarrow (EB) \equiv (PL) \Rightarrow (QG)$$

If furtherly h is convex, then $(EB) \equiv (PL) \equiv (QG)$

ref. Thm 2 of [KNS 2016].

Thm Suppose h is C^1 and l-smooth, $h^* = h(x^*)$, $x^* \exists !$,

h satisfies (PL) with $\mu > 0$ then GD

$$x_{k+1} \leftarrow x_k - s \nabla h(x_k)$$

with $s < \frac{2}{\ell}$ converges linearly, i.e.

$h(x_k) - h^* \le \varepsilon^2$ in $k = O(\log \frac{1}{\varepsilon})$ steps.

Rk By (QG), $h(x_k) - h^* \ge \frac{\mu}{2} \|x_k - x^*\|^2$, so objective value convergence to $O(\varepsilon^2)$ implies $\|x_k - x^*\| = O(\varepsilon)$, also $\|\nabla h(x_k) - \cancel{\nabla h(x^*)}^{\,0}\| \le \ell \|x_k - x^*\| = O(\varepsilon)$.

pf. descent of $h(x_k)$: by ℓ-smoothness of h,

$$h(x_{k+1}) \le h(x_k) + \langle \nabla h(x_k), \underbrace{x_{k+1} - x_k}_{= -s \nabla h(x_k)} \rangle + \frac{\ell}{2} \|x_{k+1} - x_k\|^2$$

$$= h(x_k) - s \|\nabla h(x_k)\|^2 + \frac{\ell}{2} s^2 \|\nabla h(x_k)\|^2$$

$$= h(x_k) - s(\underbrace{1 - \tfrac{\ell}{2} s}_{0 < 1 - \frac{\ell}{2}s < 1 \text{ if } \frac{\ell}{2}s < 1}) \|\nabla h(x_k)\|^2$$

set $s = \frac{1}{\ell}$

$$\downarrow$$

$$= h(x_k) - \frac{s}{2} \|\nabla h(x_k)\|^2$$

This is $h(x_k) - h(x_{k+1}) \ge \frac{s}{2} \|\nabla h(x_k)\|^2$

so far no difference from the "NC case" proof in L1.

By PL, $\|\nabla h(x_k)\|^2 \ge 2\mu (h(x_k) - h^*)$,

$$\Rightarrow \quad h(x_{k+1}) \le h(x_k) - s\mu (h(x_k) - h^*)$$

$$\Rightarrow \quad (h(x_{k+1}) - h^*) \le \underbrace{(1 - s\mu)}_{= 1 - \frac{\mu}{\ell} = 1 - \frac{1}{\kappa}} (h(x_k) - h^*)$$

$\kappa := \frac{\ell}{\mu}$ condition number, $\kappa \ge 1$

This shows $h(x_k) - h^* \le (1 - \frac{\mu}{\ell})^k (h(x_0) - h^*)$ ...

Rk. (PL) is a relaxed assump. from (SC) but still ensures exponential convergence of GD, or GD-like schemes.

- NC-PL minimax problem

$$\min_x \max_y L(x,y)$$

(A1) L is C^1 on $\mathbb{R}^n \times \mathbb{R}^m$, ℓ-smooth

(A2) $L(x,\cdot)$ is μ-PL in y and $y^*(x)$ $\exists !$, $\forall x$

Intuitively, y is the "fast variable" to be solved.

A-GDA (τ, η) two-timescale

$$\begin{cases} x_{k+1} \leftarrow x_k - \tau \nabla_x L(x_k, y_k) \\ y_{k+1} \leftarrow y_k + \eta \nabla_y L(\underline{x_{k+1}}, y_k) \end{cases}$$

$\uparrow x_k$ if without alternating (just GDA)

We will set $\tau \sim \frac{1}{\ell}\frac{1}{\kappa^2}$, $\eta \sim \frac{1}{\ell}$, $\kappa := \frac{\ell}{\mu} \geq 1$

τ is smaller, so the update of x_k is slower.