

L6. NC-PL: convergence of two-timescale GDA

$$\min_{x \in \mathbb{R}^n} \quad \max_{y \in \mathbb{R}^m} \quad L(x, y) \quad K := \frac{1}{\mu} \geq 1$$

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

(A2) $\forall x, L(x, \cdot)$ is μ -PL in y and $y^*(x) \in \mathbb{I}$

$$\text{Def. } \hat{y}(x) = \arg \max_y L(x, y) \quad \exists !, \quad \hat{\Phi}(x) := \max_y L(x, y)$$

$$\frac{1}{2} \| \nabla_y L(x, y) \|^2 \geq \mu (\bar{L}(x) - L(x, y)), \quad \forall y$$

$B \in (PL) \equiv (PB)$, we also have

$$(EB) \quad \|\nabla_y L(x)w\| \geq \|(y, w)\|_{\mathcal{X}^* \times \mathcal{X}}$$

AGDA (2, η) two-time scale

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

Thm For $\epsilon \sim \frac{1}{\ell K^2}$, $\eta \sim \frac{1}{\ell}$ to be specified in proof. Suppose

$\mathcal{E}(x) = \max_y L(x, y)$ has finite lower bound. Then, AGDA

finds $\|\nabla \tilde{L}(x_k)\| \leq \varepsilon$ in $k \leq T = O\left(\frac{k^2 \eta}{\varepsilon^2}\right)$ steps, and

x_k is an $(\sqrt{2}\varepsilon, \varepsilon_k)$ -stationary pt of L .

Rk. We say $(\hat{\Sigma}, \hat{g})$ is an $(\varepsilon, \varepsilon')$ -stationary pt of L if

$$\|\nabla_x L(\hat{x}, \hat{y})\| \leq \varepsilon, \quad \|\nabla_y L(\hat{x}, \hat{y})\| \leq \varepsilon'.$$

To prove the Thm., we first establish Lip-const. of \mathbb{E}

$$\underline{\text{Lena}} \text{ (Lip of } \mathbb{E}) \quad \kappa = \frac{1}{\mu}$$

$$i) \|\bar{y}(x_1) - \bar{y}(x_2)\| \leq K \|x_1 - x_2\|, \quad \forall x_1, x_2$$

$$ii) \bar{\Phi} \text{ is } L\text{-smooth, } L = l(1+K), \quad \nabla \bar{\Phi}(x) = \nabla_x L(x, y^*(x)).$$

$$\text{pf: i) } \forall x_1, x_2, \quad \nabla_y L(x_i, y_i^*) = 0, \quad i=1,2, \quad y_i^* = y^*(x_i).$$

Because $L(x_1, \cdot)$ is μ -PL in y ,

$$\begin{aligned} \mu \|y_2^* - y_1^*\| &\leq \|\nabla_y L(x_1, y_2^*)\| \quad L\text{-smoothness} \\ (\text{EB}) \quad &\| \nabla_y L(x_1, y_2^*) - \nabla_y L(x_2, y_2^*) \| \leq l \|x_1 - x_2\| \end{aligned}$$

ii) The expression of $\nabla \bar{\Phi}$ is by Danskin's. $\forall x_1, x_2,$

$$\begin{aligned} \|\nabla \bar{\Phi}(x_1) - \nabla \bar{\Phi}(x_2)\| &= \|\nabla_x L(x_1, y_1^*) - \nabla_x L(x_2, y_2^*)\| \\ &\leq l (\|x_1 - x_2\| + \underbrace{\|y_1^* - y_2^*\|}_{\substack{\text{L-smoothness} \\ \leq K \|x_1 - x_2\| \text{ by i)}}) \quad \# \end{aligned}$$

pf of Thm:

- descent of $\bar{\Phi}$:
 L smoothness of $\bar{\Phi}$

$$\bar{\Phi}(x_{k+1}) \leq \bar{\Phi}(x_k) + \underbrace{\langle \nabla \bar{\Phi}(x_k), x_{k+1} - x_k \rangle}_{\substack{\parallel \\ -2 \nabla_x L(x_k, y_k)}}$$

$$= \bar{\Phi}(x_k) - \tau \underbrace{\langle \nabla \bar{\Phi}(x_k), \nabla_x L(x_k, y_k) \rangle}_{\substack{\parallel \\ \nabla \bar{\Phi}(x_k) + \delta_k}} + \frac{1}{2} \tau^2 \|\nabla_x L(x_k, y_k)\|^2$$

$$\delta_k = \nabla_x L(x_k, y_k) - \underbrace{\nabla_x L(x_k, y^*(x_k))}_{\nabla_x L(x_k, y^*(x_k))}$$

$$\Rightarrow \|\delta_k\| \leq \lambda \underbrace{\|y_k - y^*(x_k)\|}_{\substack{\parallel \\ \text{by (EB)}}} \leq K \|\nabla_y L(x_k, y_k)\| \quad \dots \quad ④$$

$$\leq \bar{\Phi}(x_k) - \tau \|\nabla \bar{\Phi}(x_k)\|^2 - \tau \langle \nabla \bar{\Phi}(x_k), \delta_k \rangle + \frac{\tau}{2} \|\nabla \bar{\Phi}(x_k) + \delta_k\|^2$$

$$= \bar{\Phi}(x_k) - \frac{\tau}{2} \|\nabla \bar{\Phi}(x_k)\|^2 + \frac{\tau}{2} \|\delta_k\|^2 \quad \text{Assume, } L \leq 1$$

$$\tau \leq \frac{1}{L(K, K)} \leq \frac{1}{2L}$$

- descent of $L(x_k, y_k)$

in the y -step, ℓ -smoothness $\eta \|\nabla_y L(x_{k+1}, y_k)\|$

$$L(x_{k+1}, y_{k+1}) - L(x_{k+1}, y_k) \geq \langle \nabla_y L(x_{k+1}, y_k), \underbrace{y_{k+1} - y_k}_{-\frac{\ell}{2} \|y_{k+1} - y_k\|^2} \rangle$$

$$= \eta \|\nabla_y L(x_{k+1}, y_k)\|^2 - \frac{1}{2} \eta^2 \|\nabla_y L(x_{k+1}, y_k)\|^2$$

$$\geq \frac{1}{2} \|\nabla_y L(x_{k+1}, y_k)\|^2 \quad \text{Assume: } \ell \eta \leq 1$$

in the x -step,

$$L(x_{k+1}, y_k) - L(x_k, y_k) \geq \langle \nabla_x L(x_k, y_k), \underbrace{x_{k+1} - x_k}_{-\frac{\ell}{2} \|x_{k+1} - x_k\|^2} \rangle$$

$$= -\zeta \|\nabla_x L(x_k, y_k)\|^2 - \frac{1}{2} \zeta^2 \|\nabla_x L(x_k, y_k)\|^2$$

already have $\ell \zeta \leq \frac{1}{2}$

$$\geq -\frac{5}{4} \zeta \|\nabla_x L(x_k, y_k)\|^2 \quad \dots \textcircled{2}$$

①, ② \Rightarrow

$$L(x_{k+1}, y_{k+1}) - L(x_k, y_k) \geq \frac{1}{2} \|\nabla_y L(x_{k+1}, y_k)\|^2 - \frac{5}{4} \zeta \|\nabla_x L(x_k, y_k)\|^2$$

- Design the Lyapunov function $\lambda > 0$ TBD

$$V_k = V(x_k, y_k) := \bar{\varphi}(x_k) + \alpha (\bar{\varphi}(x_k) - L(x_k, y_k))$$

$$\begin{aligned} V_k - V_{k+1} &= (1+\alpha) (\bar{\varphi}(x_k) - \bar{\varphi}(x_{k+1})) - \alpha (L(x_k, y_k) - L(x_{k+1}, y_{k+1})) \\ &\geq (1+\alpha) \left(\frac{\zeta}{2} \|\nabla \bar{\varphi}(x_k)\|^2 - \frac{\zeta}{2} \|\bar{\delta}_k\|^2 \right) \\ &\quad + \alpha \left(\frac{1}{2} \|\nabla_y L(x_{k+1}, y_k)\|^2 - \frac{5}{4} \zeta \underbrace{\|\nabla_x L(x_k, y_k)\|^2}_{\nabla \bar{\varphi}(x_k) + \delta_k} \right) \end{aligned}$$

The two negative terms are trouble

- $\nabla_x L(x_k, y_k) \approx \nabla \bar{\varphi}(x_k)$ up to $\bar{\delta}_k$, so handled by $\|\nabla \bar{\varphi}(x_k)\|$;
- $\|\bar{\delta}_k\| \leq K \|\nabla_y L(x_k, y_k)\|$ by ④, not quite $\nabla_y L(x_{k+1}, y_k)$

$$\underbrace{\|\nabla_y L(x_{k+1}, y_k) - \nabla_y L(x_k, y_k)\|}_{=: e_k} \text{ but maybe close to.}$$

$$= \lambda \underbrace{\|\nabla_x L(x_k, y_k)\|}_{\text{again controlled by a).}} \quad \text{--- (5)}$$

$$\text{For a: } \|\nabla_x L(x_k, y_k)\|^2 = \|\nabla \bar{L}(x_k) + \bar{\delta}_k\|^2 \leq 2 \left(\|\nabla \bar{L}(x_k)\|^2 + \|\bar{\delta}_k\|^2 \right) \quad \text{--- (3)}$$

$$\begin{aligned} \text{For b: } \|\nabla_y L(x_{k+1}, y_k)\|^2 &= \|\nabla_y L(x_k, y_k) + e_k\|^2 \\ &\geq \frac{1}{2} \|\nabla_y L(x_k, y_k)\|^2 - \|e_k\|^2 \\ &\quad (\|a+b\|^2 \geq \frac{1}{2} \|a\|^2 - \|b\|^2) \end{aligned}$$

$$\begin{aligned} \Rightarrow V_k - V_{k+1} &\geq -\frac{1+\alpha}{2} \|\nabla \bar{L}(x_k)\|^2 - \frac{1+\alpha}{2} \|\bar{\delta}_k\|^2 - \frac{5\alpha}{4} \|\nabla_x L(x_k, y_k)\|^2 \\ &\quad + \eta \frac{\alpha}{4} \|\nabla_y L(x_k, y_k)\|^2 - \eta \frac{\alpha}{2} \|e_k\|^2 \\ &\quad \text{by (5) } \leq \lambda^2 \|\nabla_x L(x_k, y_k)\|^2 \\ &\quad \eta \lambda^2 \varepsilon^2 \leq \frac{\eta}{2} \quad (\eta \leq 1, \lambda \leq \frac{1}{2}) \\ &\geq -\frac{1+\alpha}{2} \|\nabla \bar{L}(x_k)\|^2 - \frac{1+\alpha}{2} \|\bar{\delta}_k\|^2 - \frac{3\alpha}{2} \|\nabla_x L(x_k, y_k)\|^2 \\ &\quad + \eta \frac{\alpha}{4} \|\nabla_y L(x_k, y_k)\|^2 \\ &\geq -\left(\frac{1+\alpha}{2} - 3\alpha\right) \|\nabla \bar{L}(x_k)\|^2 - \left(\frac{1+\alpha}{2} + 3\alpha\right) \|\bar{\delta}_k\|^2 \\ &\quad + \eta \frac{\alpha}{4} \|\nabla_y L(x_k, y_k)\|^2 \end{aligned}$$

$$= -\frac{1-5\alpha}{2} \|\nabla \bar{L}(x_k)\|^2 - \underbrace{\left(-\frac{1+7\alpha}{2} \|\bar{\delta}_k\|^2 + \eta \frac{\alpha}{4} \|\nabla_y L(x_k, y_k)\|^2\right)}_{\text{by (3) } \geq 0}$$

$$\geq \left(\eta \frac{\alpha}{4} - \frac{1+7\alpha}{2} k^2\right) \|\nabla_y L(x_k, y_k)\|^2$$

$$\text{if } \eta \frac{\alpha}{4} - \frac{1+7\alpha}{2} k^2 > 0, \text{ then } \geq 0$$

$$\text{let } \alpha = \frac{1}{8}, \quad \eta \frac{1}{32} \geq \frac{1}{2} k^2 \quad \text{Assume.}$$

$$> \frac{\eta}{8} \|\nabla \bar{L}(x_k)\|^2 + \left(\frac{1}{32} - \frac{1}{2} k^2\right) \|\nabla_y L(x_k, y_k)\|^2 \quad (6)$$

$$\text{We have shown } \frac{\eta}{8} \|\nabla \bar{L}(x_k)\|^2 \leq V_k - V_{k+1}$$

$$\dots \geq \frac{\eta}{8} \|\nabla \bar{L}(x_1)\|^2 \dots$$

By telescoping, $\overline{V} \leq \sum_{k=0}^T \|\nabla \underline{L}(x_k)\| = V_0 - V_T$

$$V_T = \underline{L}(x_T) + \alpha \left(\overbrace{\underline{L}(x_T) - L(x_T, y_T)}^{20} \right) \geq \underline{L}(x_T) \geq \underline{L}^*$$

$$\underline{L}^* = \min_{\underline{x}} \underline{L}(\underline{x}) \text{ assumed to be finite}$$

$$\Rightarrow V_0 - V_T \leq V_0 - \underline{L}^* = \underline{L}(x_0) - \underline{L}^* + \frac{1}{\alpha} (\underline{L}(x_0) - L(x_0, y_0))$$

$\Delta = \Delta \text{ const. depending on initial value } (x_0, y_0)$

$$\Rightarrow \sum_{k=0}^T \|\nabla \underline{L}(x_k)\|^2 \leq 8 \frac{\Delta}{\alpha}$$

$$\Rightarrow \exists k \leq T \text{ s.t. } \|\nabla \underline{L}(x_k)\|^2 \leq 8 \frac{\Delta}{\alpha T}$$

Overall, we have assumed

$$\begin{cases} \eta \leq 1 \\ \epsilon \leq \epsilon(HK) \leq 1 \\ \epsilon K^2 \leq \frac{\eta}{32} \end{cases}$$

This can be fulfilled by setting $\eta = \frac{1}{2}$, $\epsilon = \frac{1}{32} \frac{1}{\epsilon K^2}$

$$\text{Then, } \|\nabla \underline{L}(x_k)\|^2 \leq \frac{\Delta}{\alpha T} \sim \frac{\epsilon K^2 \Delta}{T}.$$

One can use the refined estimate (*) to show that

$$\|\nabla_x L(x_k, y_k)\| \leq \sqrt{2} \epsilon, \quad \|\nabla_y L(x_k, y_k)\| \leq \epsilon/K.$$

$$\begin{aligned} V_k - V_{k+1} &\geq \frac{\epsilon}{8} \|\nabla \underline{L}(x_k)\|^2 + \underbrace{\left(\frac{\eta}{32} - \epsilon K^2 \right) \|\nabla_y L(x_k, y_k)\|^2}_{= \frac{\epsilon}{4} K^2 \text{ if } \frac{\eta}{32} = \frac{1}{4} \epsilon K^2} \\ &= \frac{\epsilon}{8} \left(\|\nabla \underline{L}(x_k)\|^2 + 2K^2 \|\nabla_y L(x_k, y_k)\|^2 \right) \end{aligned}$$

By summing $k=0, \dots, K-1$, $\exists k \leq T \sim \frac{\epsilon K^2 \Delta}{T}$ s.t.

$$\|\nabla \underline{L}(x_k)\|^2 + 2K^2 \|\nabla_y L(x_k, y_k)\|^2 \leq \epsilon^2 \quad (*)$$

This x_k satisfies that $\|\nabla \underline{L}(x_k)\| \leq \epsilon$.

We now derive an upper bound of $\|\nabla_x L(x_k, y_k)\|$ for x, y ,

$$\|\nabla_x L(x, y) - \nabla_x L(x, y^*(x))\| \leq \ell \|y - y^*(x)\| \leq K \|\nabla_y L(x, y)\|$$

then $\nabla \bar{\Phi}(x)$

$$\begin{aligned}\|\nabla_x L(x_k, y_k)\| &\leq \underbrace{\|\nabla_x L(x_k, y_k) - \nabla \bar{\Phi}(x_k)\|}_{\leq K \|\nabla_y L(x_k, y_k)\|} + \|\nabla \bar{\Phi}(x_k)\|\\ &\leq K \|\nabla_y L(x_k, y_k)\|\end{aligned}$$

$$\begin{aligned}\Rightarrow \|\nabla_x L(x_k, y_k)\|^2 &\leq (\|\nabla \bar{\Phi}(x_k)\| + K \|\nabla_y L(x_k, y_k)\|)^2 \\ &\leq 2 (\|\nabla \bar{\Phi}(x_k)\|^2 + K^2 \|\nabla_y L(x_k, y_k)\|^2)\end{aligned}$$

Back to (2), we have

$$\begin{aligned}\varepsilon^2 &\geq \|\nabla \bar{\Phi}(x_k)\|^2 + 2K^2 \|\nabla_y L(x_k, y_k)\|^2 \\ &\geq K^2 \|\nabla_y L(x_k, y_k)\|^2 + \frac{1}{2} \|\nabla_x L(x_k, y_k)\|^2\end{aligned}$$

$$\text{Then, } K \|\nabla_y L(x_k, y_k)\|, \sqrt{\frac{1}{2} \|\nabla_x L(x_k, y_k)\|^2} \leq \varepsilon. \quad \square$$

KK. The proof for GDA is similar.

• Proximal Step S-GDA

$$\left\{ \begin{array}{l} x_{k+1} \leftarrow x_k - \tau \left(\underbrace{\nabla_x L(x_k, y_k)}_{\partial_x K(x_k, y_k; w_k)} + \underbrace{p(x_k - w_k)}_{\partial_w K(x_k, y_k; w_k)} \right) \\ y_{k+1} \leftarrow y_k + \eta \underbrace{\nabla_y L(x_{k+1}, y_k)}_{\partial_y K(x_{k+1}, y_k; w_k)} + \underbrace{p(x_{k+1} - w_k)}_{-\frac{p}{\mu} \partial_w K(x_{k+1}, y_k; w_k)} \\ w_{k+1} \leftarrow w_k + \underbrace{p(x_{k+1} - w_k)}_{K(x, y; w)} \end{array} \right.$$

$$\min_x \bar{\Phi}(x) + \frac{\rho}{2} \|x - w\|^2 = \min_x \max_y L(x, y) + \frac{\rho}{2} \|x - w\|^2$$

$$\text{Set } p = 2\mu, \rho \sim \eta \mu \Rightarrow \frac{\rho}{p} \sim \frac{\eta \mu}{\mu} = \frac{\eta}{\kappa}$$

Thm. For $\tau \sim \frac{1}{2}, \eta \sim \frac{1}{2}, \rho = 2\mu, \beta \sim \eta \mu$, S-GDA finds

an $(\varepsilon, \frac{2}{\sqrt{K}})$ -stationary pt of L in $\mathcal{O}(\frac{\ell K}{\varepsilon^2})$ steps.

ref. Thm 4.1 [YOLH 2022]

KK: Comparing to GD, the step size is less restricted by condition number κ , and ite. complexity is also better, i.e. $\sim \kappa$ instead of κ^2 .

Generally, proximal step improves "stability" and allows for larger step size in GD.

$$\min_x f(x) \quad f_s(x) = \min_u f(u) + \frac{1}{2s} \|u - x\|^2$$

$$\text{GD: } x_{k+1} = x_k - \delta \nabla f(x_k) \quad (\text{fwd Euler})$$

$$\begin{aligned} \text{PPM: } x_{k+1} &= \text{Prox}_{S,f}(x_k) \quad (\text{bwd Euler}) \\ &= x_k - \delta \nabla f_s(x_k) \text{ as if GD on } f_s \end{aligned}$$

Suppose f is ℓ -smooth, L-W.C., $\kappa = \frac{\ell}{p} \geq 1$

- GD needs $\delta < \frac{2}{\ell}$, typically $\delta = \frac{1}{\ell}$, then L1 shows

$$\frac{1}{T} \sum_{k=0}^{T-1} \|\nabla f(x_k)\|^2 \leq \frac{4\Delta}{\ell}, \quad \Delta = f(x_0) - f^*$$

- f_s need $\delta < \frac{1}{p}$, typically $\delta = \frac{1}{2p}$,
then f_s is $\bar{\ell}$ -smooth, $\bar{\ell} = \max \left\{ \frac{1}{\delta}, \frac{1}{p} - \delta \right\} = 2p$.

for GD on f_s , need $\delta < \frac{2}{\bar{\ell}} = \frac{1}{p}$ satisfied.

$$\Rightarrow \frac{1}{T} \sum_{k=0}^{T-1} \underbrace{\|\nabla f_s(x_k)\|_1^2}_{\nabla f_s(x_k)_+} \leq \frac{2\bar{\Delta}}{T}, \quad \bar{\Delta} = f_s(x_0) - f_s^* \leq \Delta$$

Thus, in GD, $s \sim \frac{1}{\ell}$, $T \sim \frac{\ell}{\epsilon^2}$,

in PPM, $s \sim \frac{1}{p}$, $T \sim \frac{p}{\epsilon^2}$,

both are $O(\kappa)$ better.