L4. NC-NC: interaction dominance, PPM and EGM

Road map: - SC-SC of L_s under Inter. Dom. condition

- GDA on L_s = damped PPM on L

- PPM computed approximately by EGM.

[GLWM 2023] Grimmer, Lu, Worah, Mirrokni (2023). The landscape of the proximal point method for nonconvex–nonconcave minimax optimization. *Mathematical Programming*, *201*(1), 373-407.

[HLG 2024] Hajizadeh, Lu, Grimmer (2024). On the linear convergence of extragradient methods for nonconvex–nonconcave minimax problems. INFORMS Journal on Optimization, 6(1), pp.19-31.

$$\min_x \max_y L(x,y) \qquad L: \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$$

L is C^2, ℓ-smooth, ρ-WC-WC, $\frac{1}{s} > \rho$

Recall the saddle envelope

$$L_s(x,y) = \operatorname*{argmin}_u \max_v \; L(u,v) + \frac{1}{2s}\|u-x\|^2 - \frac{1}{2s}\|v-y\|^2$$

$$z = \begin{bmatrix} x \\ y \end{bmatrix}, \quad z_+ = \begin{bmatrix} x_+ \\ y_+ \end{bmatrix} = \mathrm{Prox}_{s,L}(z)$$

$$\partial = \begin{bmatrix} \nabla_x \\ -\nabla_y \end{bmatrix}, \quad \partial^2 = \begin{bmatrix} \nabla^2_{xx} & \nabla^2_{xy} \\ -\nabla^2_{yx} & -\nabla^2_{yy} \end{bmatrix}$$

$$\begin{cases} \partial L_s(z) = \frac{1}{s}(z - z_+) = \partial L(z_+) \\ \partial^2 L_s(z) = \frac{1}{s}\left(I - (I + s\,\partial^2 L(z_+))^{-1}\right) \end{cases}$$

The Hessian expression:

$$\left.\begin{bmatrix} \underline{\nabla^2_{xx} L_s} & \nabla^2_{xy} L_s \\ -\nabla^2_{yx} L_s & \underline{-\nabla^2_{yy} L_s} \end{bmatrix}\right|_z = \frac{1}{s}\left(I - \left(I + s \left.\begin{bmatrix} \nabla^2_{xx} L & \nabla^2_{xy} L \\ -\nabla^2_{yx} L & -\nabla^2_{yy} L \end{bmatrix}\right|_{z_+}\right)^{-1}\right)$$

By inverse of block matrix,

$$\nabla^2_{xx} L_s(z) = \frac{1}{s}\left(I - \left(I + s\left[\underline{\nabla^2_{xx}L + \nabla^2_{xy}L\left(\tfrac{1}{s}I - \nabla^2_{yy}L\right)^{-1}\nabla^2_{yx}L}\right]\Big|_{z_+}\right)^{-1}\right)$$

$$-\nabla^2_{yy} L_s(z) = \frac{1}{s}\left(I - \left(I + s\left[\underline{-\nabla^2_{yy}L + \nabla^2_{yx}L\left(\tfrac{1}{s}I + \nabla^2_{xx}L\right)^{-1}\nabla^2_{xy}L}\right]\Big|_{z_+}\right)^{-1}\right)$$

Consider

$$[\text{matrix}] = \nabla^2_{xx}L + \nabla^2_{xy}L\underbrace{\left(\tfrac{1}{s}I - \nabla^2_{yy}L\right)^{-1}}_{\succeq\left(\frac{1}{s}-\rho\right)I}\nabla^2_{yx}L \succeq \nabla^2_{xx}L$$

Thus if $\nabla^2_{xx}L(z_+) \succeq 0$, then $[\text{matrix}] \succeq 0$, and then $\nabla^2_{xx}L_s(z) \succeq 0$. But it is also possible for $[\text{matrix}] \succeq 0$ even if $\nabla^2_{xx}L$ N.C., if $\nabla^2_{xy}L$ is large.

(A3) <u>Interaction dominance</u>

Given $s>0$, $\exists\, \alpha(s) > 0$ s.t.

$$\left[\nabla^2_{xx}L + \nabla^2_{xy}L\left(\tfrac{1}{s}I - \nabla^2_{yy}L\right)^{-1}\nabla^2_{yx}L\right](z) \succeq \alpha I, \quad \forall z \in \mathbb{R}^n \times \mathbb{R}^m$$

α-Inter. Dom. in x

$$\left[-\nabla^2_{yy}L + \nabla^2_{yx}L\left(\tfrac{1}{s}I + \nabla^2_{xx}L\right)^{-1}\nabla^2_{xy}L\right](z) \succeq \alpha I, \quad \forall z$$

α-Inter. Dom. in y

Eg. L has α-Inter. Dom. in x if

$$\frac{\nabla^2_{xy}L(z)\,\nabla^2_{yx}L(z)}{\frac{1}{s} + \ell} \succeq -\nabla^2_{xx}L(z) + \alpha I, \quad \forall z$$

pf. $\left(\frac{1}{s} - \nabla^2_{yy}L\right) \preceq \left(\frac{1}{s} + \ell\right) I$ by that $\nabla_y L$ is ℓ-Lip. #

Prop (SC-SC of L_s under Inter. Dom.)

Suppose L has α-Inter. Dom. in x, then

$$\frac{1}{s} I \succeq \nabla^2_{xx} L_s(z) \succeq \frac{1}{s+1/\alpha} I, \quad \forall z.$$

If L has α-Inter. Dom. in y, then

$$\frac{1}{s} I \succeq -\nabla^2_{yy} L_s(z) \succeq \frac{1}{s+1/\alpha} I, \quad \forall z.$$

pf. $\nabla^2_{xx} L_s = \frac{1}{s}\left(I - (I + s\,[\text{matrix}])^{-1}\right)$

[matrix] $\succeq \alpha I$ by Inter. Dom. $\Rightarrow \frac{1}{s+1/\alpha}$ lower bound

For the upper bound, use that $(I + s[\text{matrix}])^{-1} \succeq 0$. #

Now we have that L_s is $\bar{\mu}$-S.C-S.C, $\bar{\mu} = \frac{1}{s+1/\alpha}$,

under (A1)(A2)(A3). L_s also share stationary pts with L.

Idea: <u>GDA</u> applied to L_s will converge exponentially fast

to the (unique) stationary pt z^* of L_s, by the

$\Downarrow$

the unique stationary pt of L (saddle pt)

theory in L2.

We still need to show the $\bar{l}$-smoothness of L_s.

Q. How to compute GDA of L_s?

Key observation: GDA of L_s = damped PPM of L.

<u>damped PPM</u> $\quad z_k = \begin{bmatrix} x_k \\ y_k \end{bmatrix}, \quad 0 < \lambda \le 1$

$$\begin{cases} z_+ \leftarrow \operatorname{Prox}_{s,L}(z_k) \\ z_{k+1} \leftarrow (1-\lambda) z_k + \lambda z_+ \end{cases}$$

It is vanilla PPM when $\lambda = 1$.

Prop damped PPM of L with $0 < \lambda \le 1$ is the same as GDA of L_s with step size $\bar{s} = \lambda s$.

pf. GDA of L_s:

$$z_{k+1} \leftarrow z_k - \bar{s} \underbrace{\partial L_s(z_k)}_{= \frac{1}{s}(z_k - z_+)}$$

$$= z_k - \lambda (z_k - z_+). \qquad \#$$

This means that damped PPM can converge exponentially fast to the saddle pt, as long as s, λ properly chosen.

$$s < 1/\rho, \qquad 0 < \bar{s} < \frac{2\bar{\mu}}{\bar{\ell}^2}.$$

- L-smoothness of saddle envelope

let's revisit the smoothness (regularity) of Moreau envelope.

$$f_s(x) = \min_u f(u) + \frac{1}{2s}\|u - x\|^2, \quad f \text{ } \ell\text{-smooth, } \rho\text{-W.C.}$$

$$x_+ = \operatorname{argmin}_u, \quad \exists! \text{ when } \frac{1}{s} > \rho.$$

$$\nabla f_s(x) = \frac{1}{s}(x - x_+) = \nabla f(x_+)$$

We want a Lip-constant of ∇f_s, is $x_+(x)$ Lipschitz?

$$\nabla^2 f_s(x) = \frac{1}{s}\left[I - (I + s\nabla^2 f(x_+))^{-1}\right]$$

$$0 \prec (I + s\nabla^2 f(x_+))^{-1} \preceq \frac{1}{1 - s\rho} I$$

$$\Rightarrow -\frac{1}{\frac{1}{\rho} - s} I \preceq \nabla^2 f_s(x) \prec \frac{1}{s} I$$

Thus ∇f_s has a Lip-constant $\bar{\ell} = \max\left\{\frac{1}{s}, \frac{1}{\frac{1}{\rho} - s}\right\}$

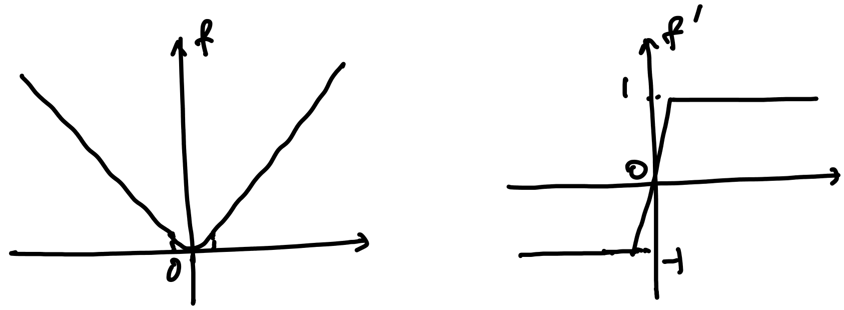
Q when f_s "smoother" f? $\bar{\ell} \le \ell$?

$\ell \geq \rho$. Eg. $\ell = 3\rho$, $\frac{1}{s} = 2\rho$, $\frac{1}{\frac{1}{\rho}-s} = \frac{\rho}{1-\rho s} = 2\rho$,

$$\Rightarrow \bar{\ell} = 2\rho < 3\rho.$$

Intuitively, $\bar{\ell}$ only "see" the interplay btw s and ρ, and not the ℓ of f.

Eg. $f(x)$ close to $|x|$, ℓ is large, but f is O-W.C.
$\Rightarrow \rho > 0$ can be any small positive number



Prop L_s is $\bar{\ell}$-smooth with $\bar{\ell} = \max\left\{\frac{1}{s}, \frac{1}{|s - 1/\rho|}\right\}$.

ref. Prop 2.9 of [GLWM 2023]

- Approximate PPM by extra gradient method (EGM)

The solving of z_+ involves inner-loop,

Recall that $z_k - z_+ = s\,\partial L(z_+)$

$$z_+ = z_k - s\,\underbrace{\partial L(z_+)}_{\partial L(z_k)} \leftarrow \text{unknown}$$

EG: approx z_+ by $\hat{z}$ that can be computed explicitly

formally, $\|z_k - z_+\| = O(s)$,

$$\Rightarrow \|\partial L(z_+) - \partial L(z_k)\| \leq \ell \|z_k - z_+\| = O(s)$$

Then, let
$$\left.\begin{aligned} z' &= z_k - s\,\partial L(z_k) \\ z_+ &= z_k - s\,\partial L(z_+) \end{aligned}\right\} \Rightarrow \|z' - z_+\| = O(s^2)$$

idea: z_k is an $O(s)$ approx. to z_+, yet the vanilla gradient step will give z' an $O(s^2)$ approx of z_+. GD (or GDA) will just use z', but the extra grad will use z' to evaluate gradient again

$$\left.\begin{aligned} \hat{z} &= z_k - s\,\partial L(z') \\ z_+ &= z_k - s\,\partial L(z_+) \end{aligned}\right\} \Rightarrow \|\hat{z} - z_+\| = O(s^3)$$

Then $\hat{z}$ is an $O(s^3)$ approx of z_+.

In many convergence analysis, by choosing small (but finite) step size s, then effect of $O(s^3)$ error can be bounded and then EGM enjoys the same convergence rate as (damped) PPM.

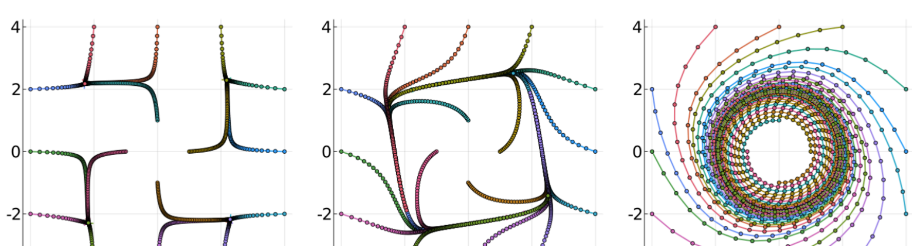
ref. [HLG 2024]

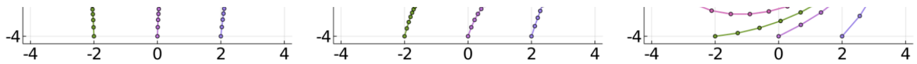
EGM of damped PPM:

$$\begin{cases} z' \leftarrow z_k - s\,\partial L(z_k) \\ \hat{z} \leftarrow z_k - s\,\partial L(z') \qquad \hat{z} \simeq z_+ \\ z_{k+1} \leftarrow (1-\lambda) z_k + \lambda \hat{z} \end{cases}$$

$$\begin{aligned} z_{k+1} &= (1-\lambda) z_k + \lambda (z_k - s\,\partial L(z')) \\ &= z_k - \lambda s\,\partial L(z') \end{aligned}$$

$$\min_x \max_y L(x,y) = f(x) + Axy - f(y), \qquad \text{with} \quad f(x) = (x-1)(x+1)(x-3)(x+3),$$



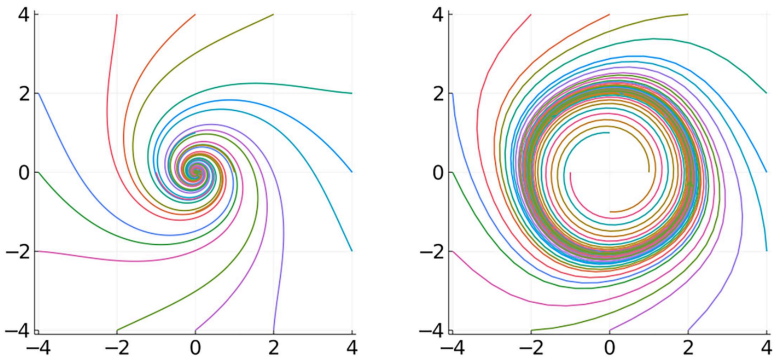


(e) EGM w/ A=1 (f) EGM w/ A=10 (g) EGM w/ A=100

$\lambda = 0.01$ $\lambda = 1$

(a) Convergence of damped EGM (b) Cycling of vanilla EGM

A=100



Q. What if Inter. Dom. only holds in one of x or y?

$$\begin{cases} \alpha\text{-Inter. Dom. in both } x \text{ and } y \Rightarrow \text{SC-SC of } L_s \\ \alpha\text{-Inter. Dom. only in } x \text{ or } y \Rightarrow \text{NC-SC of } L_s \end{cases}$$

The convergence speed degenerate from $O(\log \frac{1}{\varepsilon})$ to $O(\frac{1}{\varepsilon^2})$

as for the GDA convergence rate for NC-PL type. (L5)