Convergence of First-Order Methods for *(some)* Nonconvex-Nonconcave Minimax Optimization

Benjamin Grimmer (Johns Hopkins)
with Saeed Hajizadeh (UIC), Haihao Lu (UChicago), Pratik Worah (Google), Vahab Mirrokni (Google)

Johns Hopkins University

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

(5 minutes) Minimax problems in learning.

(10 minutes) Difficulties in nonconvex-nonconcave regimes.

(20 minutes) One (optimizer’s) path for avoiding these difficulties.

(5 minutes) Extensions and other paths forward.

Minimax Optimization in Machine Learning

Many machine learning problems fit in our general minimax form

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

This structure come up consistently throughout the week.

A few examples where difficult minimax problems arise…

(i) **Robust Training,**

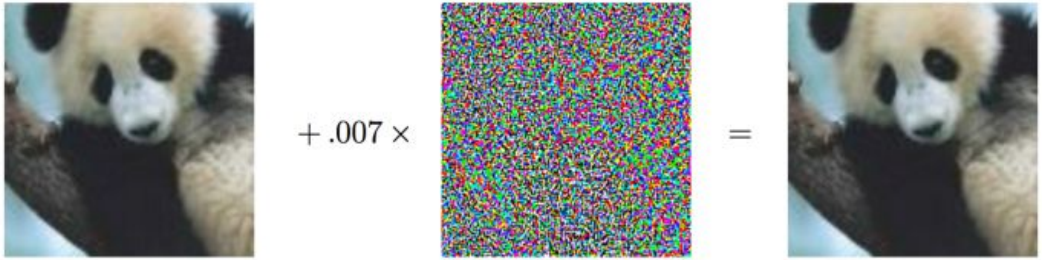
(ii) **Generative Adversarial Nets (GANs),**

(iii) **Reinforcement Learning.**

Robust Training

Consider learning to map features *u* onto labels *v* with a model *x*:

$$\min_{x} \mathbb{E}_{(u,v)} \left[\ell(u, v, x) \right]$$



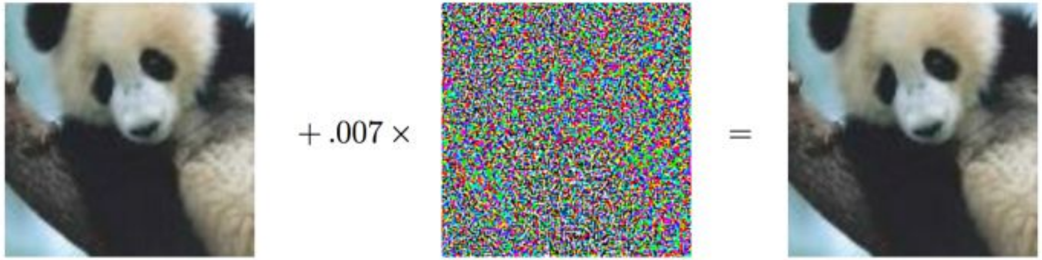
``panda'' 57.7% confidence — perturbation — ``gibbon'' 99.3% confidence

[Goodfellow et al., 2015]

Robust Training

Consider learning to map features *u* onto labels *v* with a model *x*:

$$\min_x \mathbb{E}_{(u,v)} \left[\ell(u, v, x)\right] \implies \min_x \mathbb{E}_{(u,v)} \left[\max_{y \in S} \ell(u + y, v, x)\right]$$



``panda'' 57.7% confidence

perturbation

``gibbon'' 99.3% confidence

[Goodfellow et al., 2015]

[Madry et al., 2018]
[Wang et al., 2019]
etc.

Generative Adversarial Nets (GANs)

$$\min_G \max_D \mathbb{E}_{s \sim p_{data}} [\log D(s)] + \mathbb{E}_{e \sim p_{latent}} [\log(1 - D(G(e)))]$$

G is a network generating fake data from noise.

D is a network discriminating data from fakes.

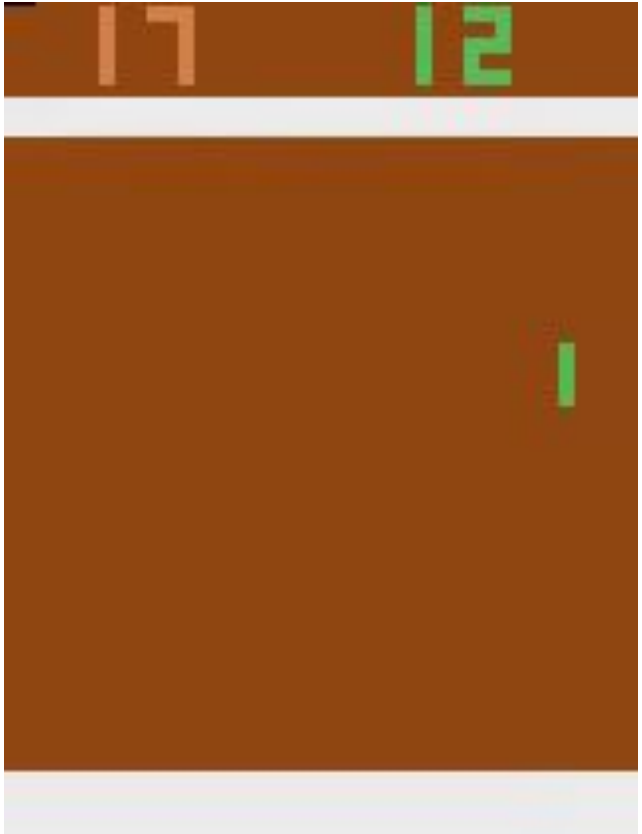
[Goodfellow et al., 2014]

Reinforcement Learning

Given state space S and actions A,

we seek a policy π maximizing reward

$$\max_{\pi:\mathcal{S}\times\mathcal{A}\to[0,1]} \mathbb{E}_{s_0}\mathbb{E}_{\pi}\left[\sum_{i=1}^{\infty}\gamma^i R(s_i,a_i)\right]$$



[Minh et al., 2013]

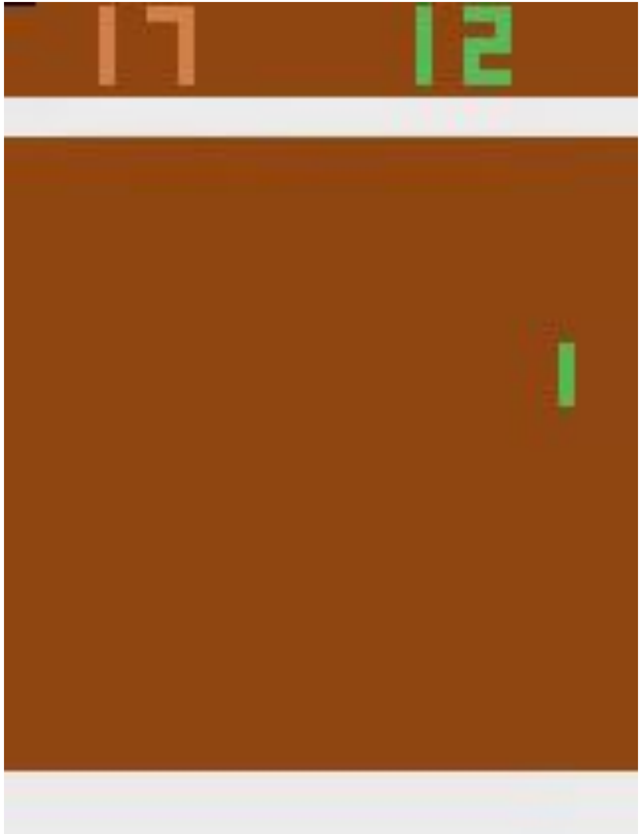
Reinforcement Learning

Given state space *S* and actions *A*,
we seek a policy *π* maximizing reward

$$\max_{\pi:\mathcal{S}\times\mathcal{A}\to[0,1]} \mathbb{E}_{s_0}\mathbb{E}_{\pi}\left[\sum_{i=1}^{\infty}\gamma^i R(s_i,a_i)\right]$$

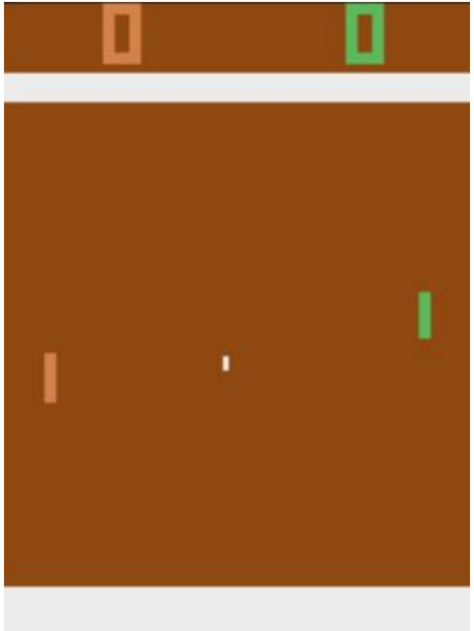
Dually, we can seek values *V(s)*
satisfying the Bellman equation

$$V(s) = \max_a \left\{R(s,a) + \gamma \mathbb{E}_{s'|s,a} V(s')\right\}$$



[Minh et al., 2013]

Reinforcement Learning



A minimax approach can merge these two ideas

$$\min_{V} \max_{\alpha,\pi} (1-\gamma)\mathbb{E}_{s\sim\mu}[V(s)] + \sum_{a,s} \alpha(s)\pi(a|s)\Delta[V](s,a)$$

where $\Delta[V](s,a) = R(s,a) + \gamma \mathbb{E}_{s'|s,a}[V(s')] - V(s)$.

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

Two natural ways to view minimax problems:

-A *sequential game* where *x* plays first and then *y* follows,

-A *simultaneous game* with *x* and *y* competing.

For convex-concave objectives**, these perspectives are the same as

$$\min_x \max_y L(x, y) = \max_y \min_x L(x, y) \; .$$

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

Two natural ways to view minimax problems:

-A *sequential game* where *x* plays first and then *y* follows,

-A *simultaneous game* with *x* and *y* competing.

For convex-concave objectives**, these perspectives are the same as

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

However, our motivating examples are not convex-concave!

Existing Theory - Sequentially Handling Nonconvexities

Globally, Approximately Solve the ``max`` Subproblem.

If we can solve over *y*, the problem reduces to nonconvex minimization

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

[Rafique et al, 2018] [Lin et al, 2019, 2020] [Thekumparampil et al, 2019]...

Locally, Approximately Solve the ``max`` Subproblem.

[Heusel et al, 2017] [Mangoubi and Vishnoi, 2021]...

Exploring notions of what minimax stationary means.

[Daskalakis and Panageas, 2018][Jin et al, 2020][Mazumdar et al, 2020]

Our focus - Simultaneous Game Perspective

Our goal: Find ``First-Order Nash Equilibrium``
(stationary points)

$$F(x, y) := \begin{bmatrix} \nabla_x L(x, y) \\ -\nabla_y L(x, y) \end{bmatrix} = 0$$

Our focus - Simultaneous Game Perspective

Our goal: Find ``First-Order Nash Equilibrium``
(stationary points)

$$F(x,y) := \begin{bmatrix} \nabla_x L(x,y) \\ -\nabla_y L(x,y) \end{bmatrix} = 0$$

Common/basic first-order algorithms:

Gradient Descent Ascent ***(GDA)*** $z_{k+1} = z_k - \alpha_k F(z_k)$

Alternating GDA *(AGDA)* $\begin{cases} x_{k+1} = x_k - \alpha_k \nabla_x L(x_k, y_k) \\ y_{k+1} = y_k + \alpha_k \nabla_y L(x_{k+1}, y_k) \end{cases}$

Proximal Point Method *(PPM)* $z_{k+1} = z_k - \alpha_k F(z_{k+1})$

Extragradient Method *(EGM)* $\begin{cases} \hat{z}_{k+1} = z_k - \alpha_k F(z_k) \\ z_{k+1} = z_k - \alpha_k F(\hat{z}_{k+1}) \end{cases}$

Our focus - Simultaneous Game Perspective

Our goal: Find ``First-Order Nash Equilibrium``
(stationary points)

$$F(x,y) := \begin{bmatrix} \nabla_x L(x,y) \\ -\nabla_y L(x,y) \end{bmatrix} = 0$$

Common/basic first-order algorithms:

Gradient Descent Ascent (*GDA*) $z_{k+1} = z_k - \alpha_k F(z_k)$

Alternating GDA (*AGDA*) $\begin{cases} x_{k+1} = x_k - \alpha_k \nabla_x L(x_k, y_k) \\ y_{k+1} = y_k + \alpha_k \nabla_y L(x_{k+1}, y_k) \end{cases}$

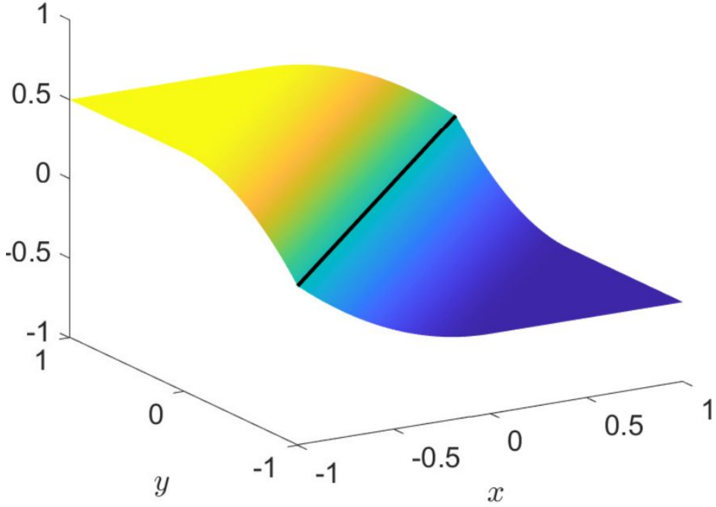
Proximal Point Method (*PPM*) $z_{k+1} = z_k - \alpha_k F(z_{k+1})$

Extragradient Method (*EGM*) $\begin{cases} \hat{z}_{k+1} = z_k - \alpha_k F(z_k) \\ z_{k+1} = z_k - \alpha_k F(\hat{z}_{k+1}) \end{cases}$

Minimax Difficulty Ex 1 (of 2)

Minimax Difficulty Ex 1 (of 2) [Lee and Kim, 2021]

$$f(x,y) = \begin{cases} \frac{R}{2} & \textit{for } x < y - \sqrt{\frac{R}{L}} \\ -\frac{L}{2}(x-y)^2 - \sqrt{LR}(x-y) & \textit{for } y - \sqrt{\frac{R}{L}} \leq x < y \\ \frac{L}{2}(x-y)^2 - \sqrt{LR}(x-y) & \textit{for } y \leq x < y + \sqrt{\frac{R}{L}} \\ -\frac{R}{2} & \textit{for } y + \sqrt{\frac{R}{L}} < x. \end{cases}$$



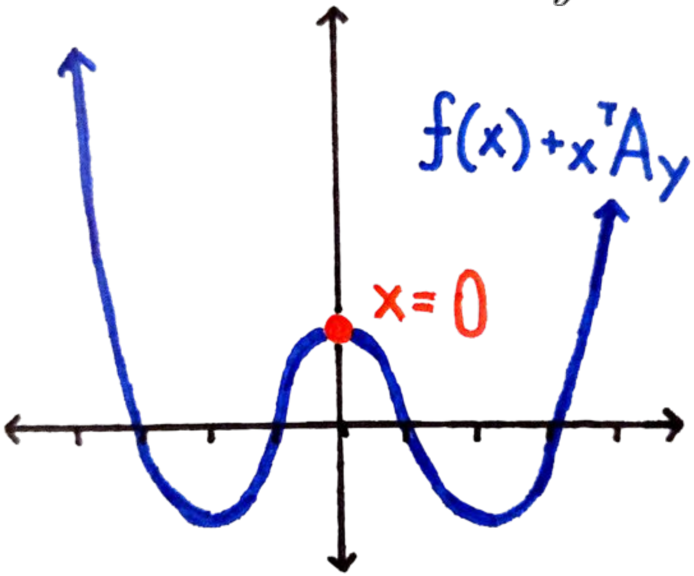
Every point on the line x=y has gradient operator point in this line!

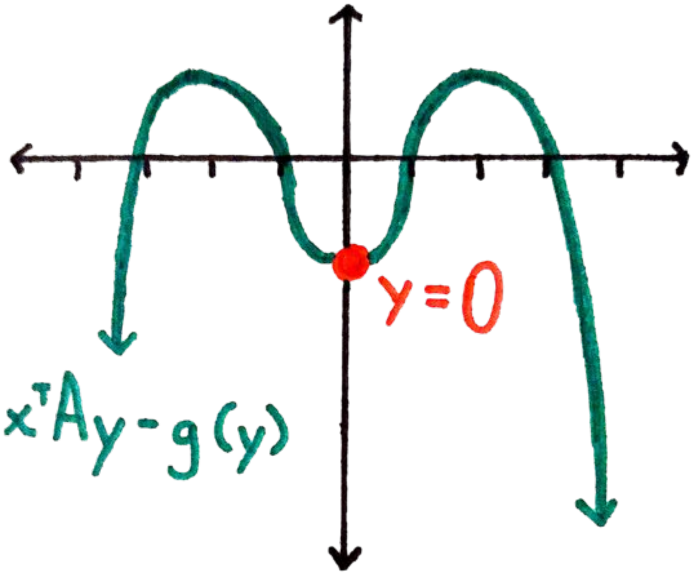
No method moving in the span of these will ever escape this line!

Minimax Difficulty Ex 2 (of 2)

Minimax Difficulty Ex 2 (of 2)

$$\min_x \max_y f(x) + x^T A y - g(y)$$

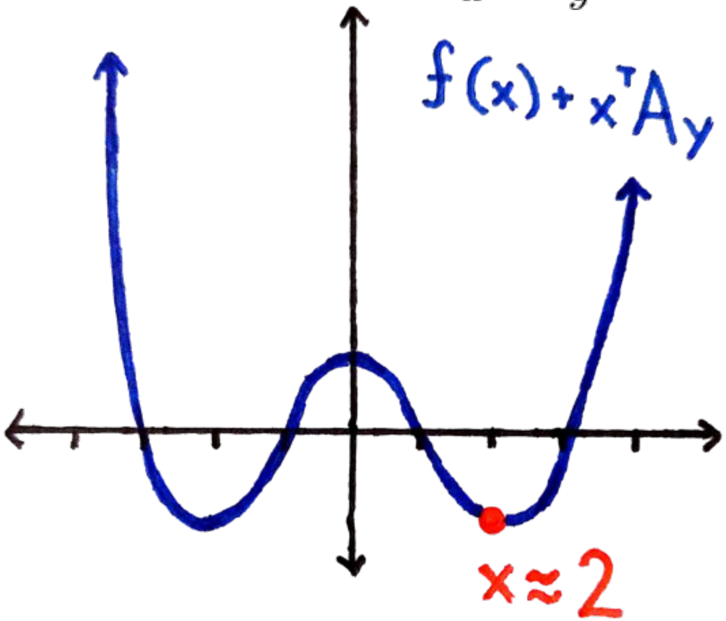


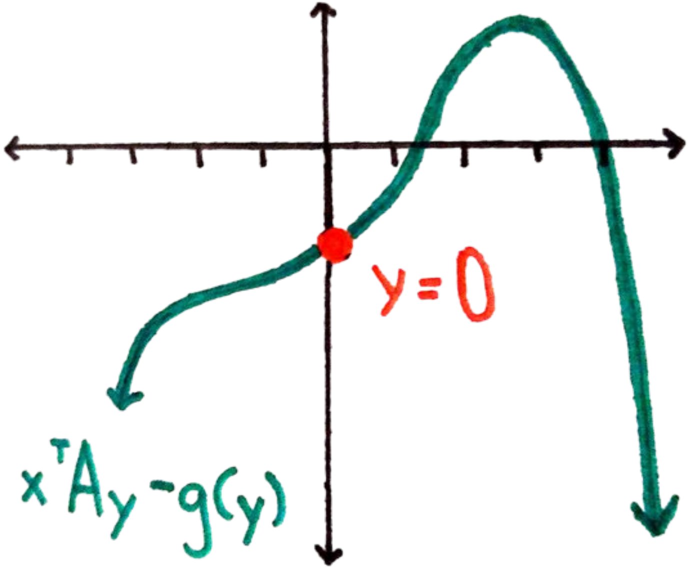


$$f(x) = g(x) = (x-3)(x-1)(x+1)(x+3)$$

Minimax Difficulty Ex 2 (of 2)

$$\min_x \max_y f(x) + x^T Ay - g(y)$$

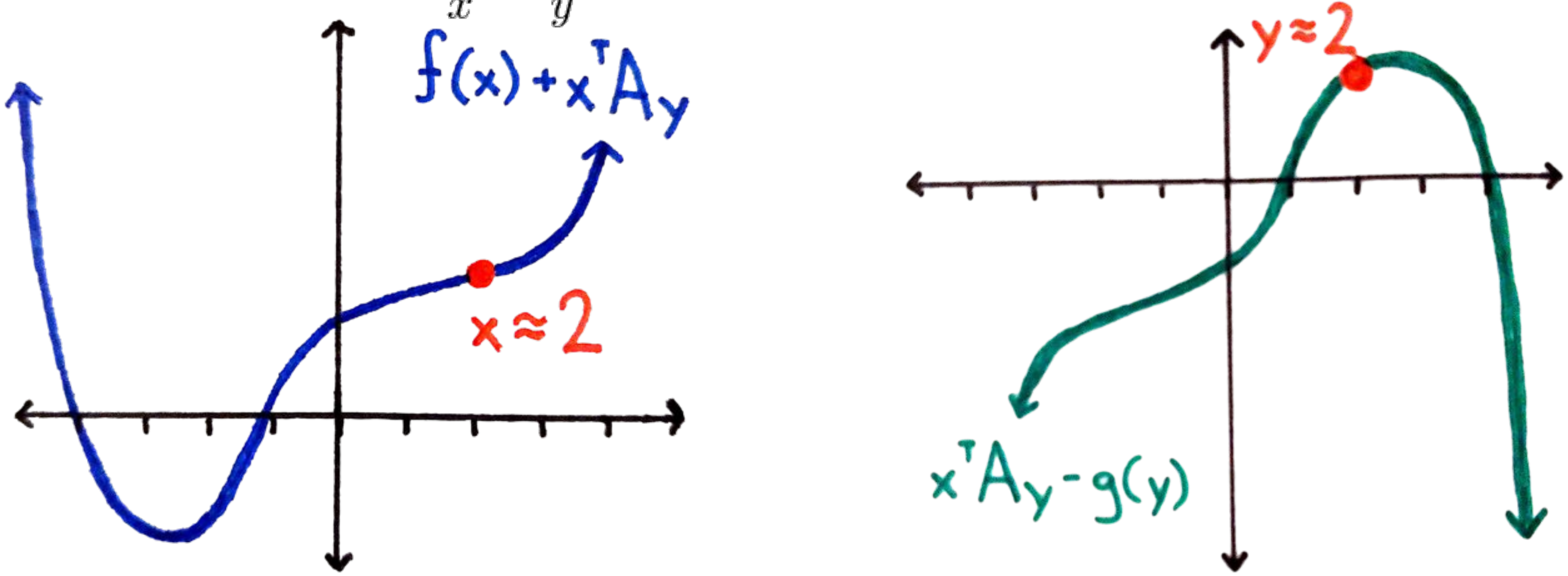




$$f(x) = g(x) = (x-3)(x-1)(x+1)(x+3)$$

Minimax Difficulty Ex 2 (of 2)

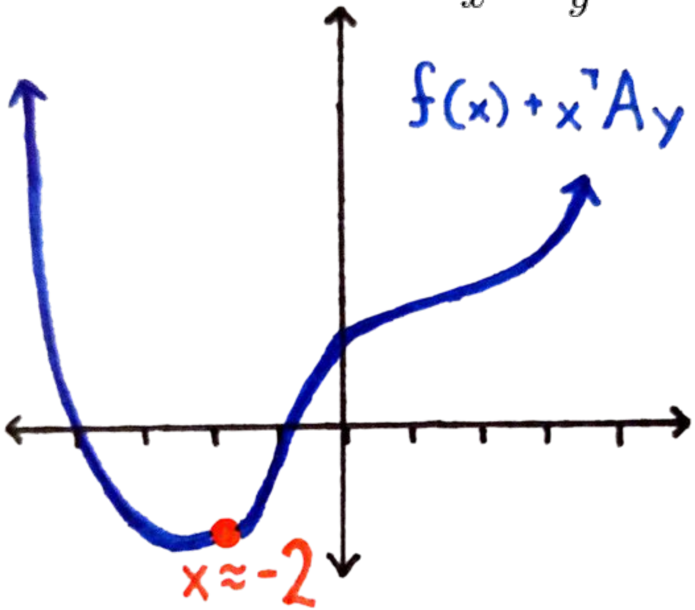
$$\min_x \max_y f(x) + x^T Ay - g(y)$$

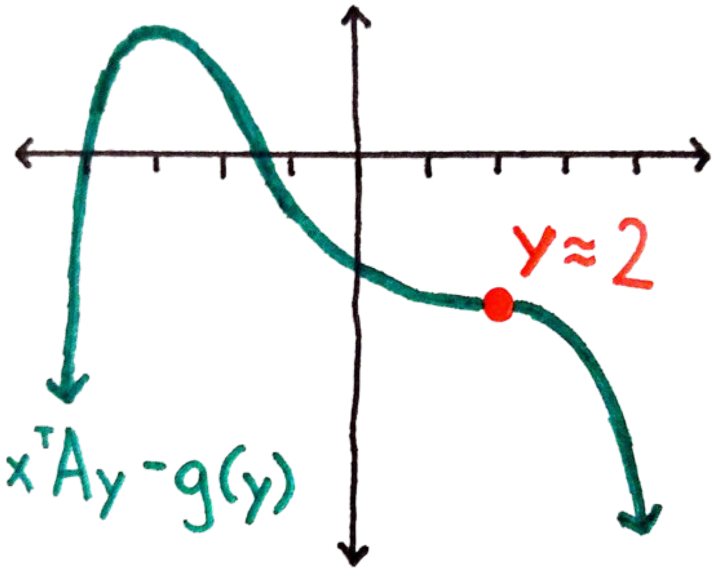


$$f(x) = g(x) = (x-3)(x-1)(x+1)(x+3)$$

Minimax Difficulty Ex 2 (of 2)

$$\min_x \max_y f(x) + x^T A y - g(y)$$

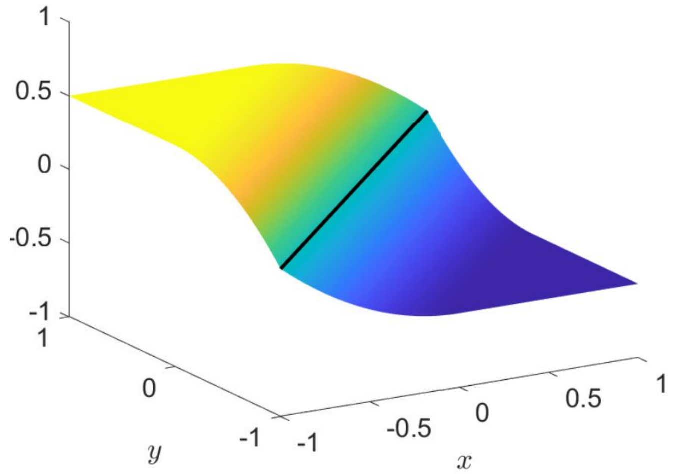


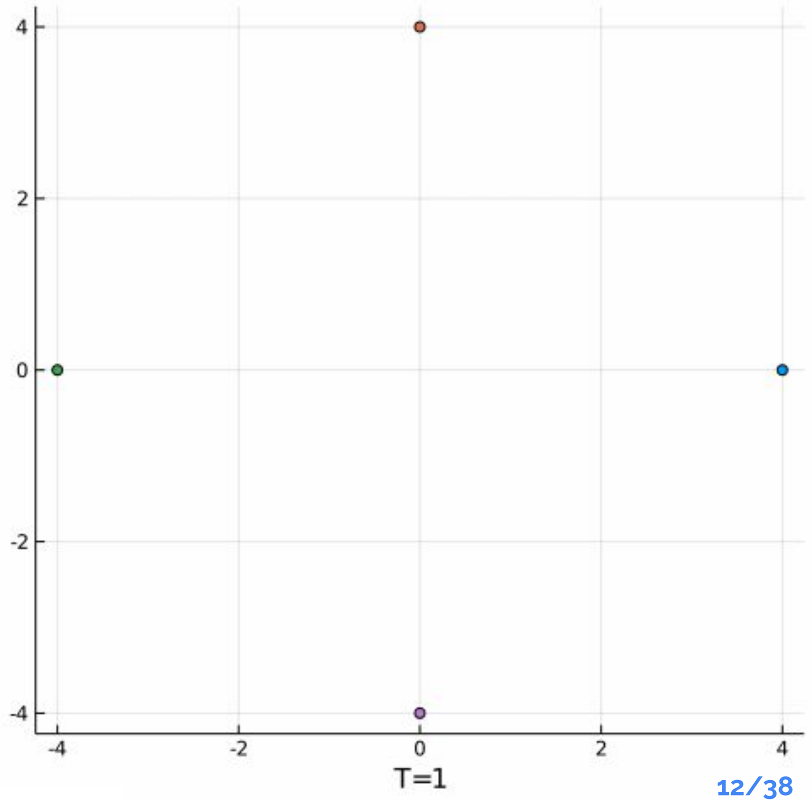


$$f(x) = g(x) = (x-3)(x-1)(x+1)(x+3)$$

Minimax Difficulties

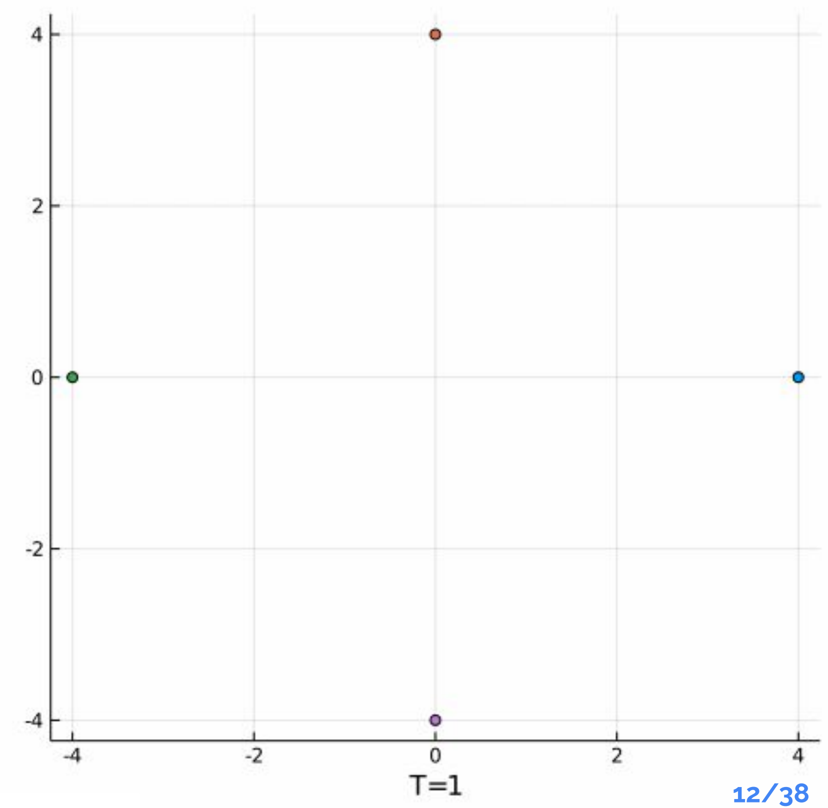
First-order updating can get stuck in a subspace or be attracted into a cycle (right Proximal Point Method shown).





The Question

When do standard algorithms converge despite nonconvexities and nonconcavities?



Existing Theory for Handling Nonconvexities

Strong Structural Assumptions.

-[Lui et al, 2020] assumes global solution to a Variational Inequality,

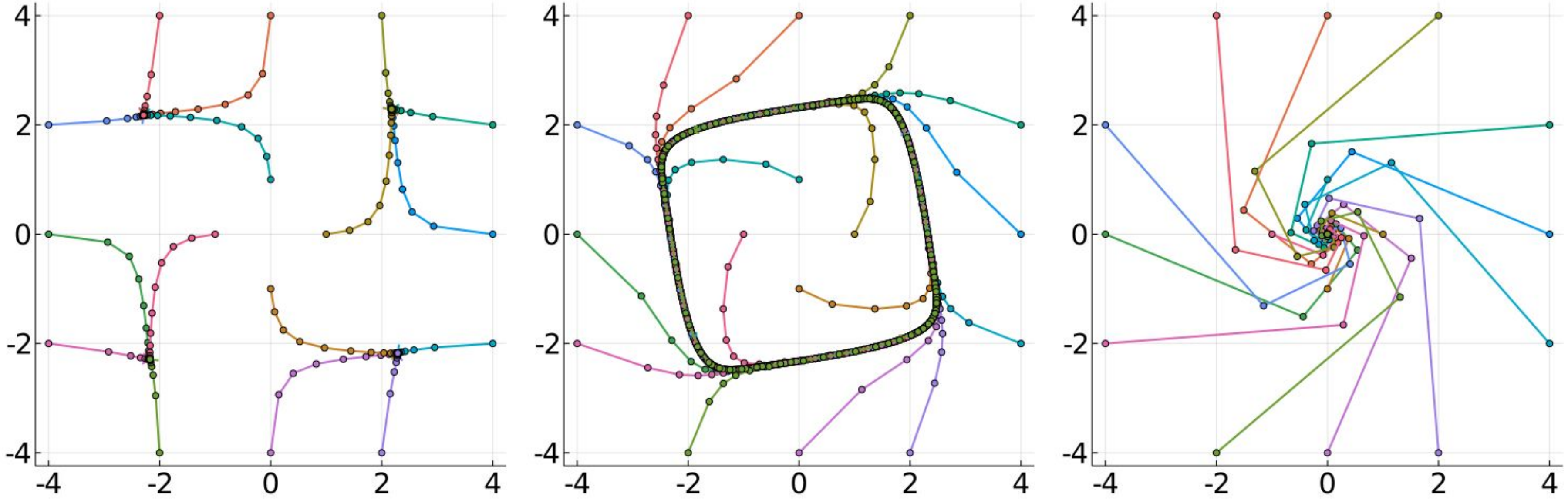
-[Nouiehed et al, 2019] [Yang et al, 2020] assume a PL condition,

-[Bauschke et al, 2020] assumes smoothness and bounded negative cocoercivity,

-[Ostrovskii et al,2021] assumes very small domain for maximizing variable,

An Observation about our Toy Example

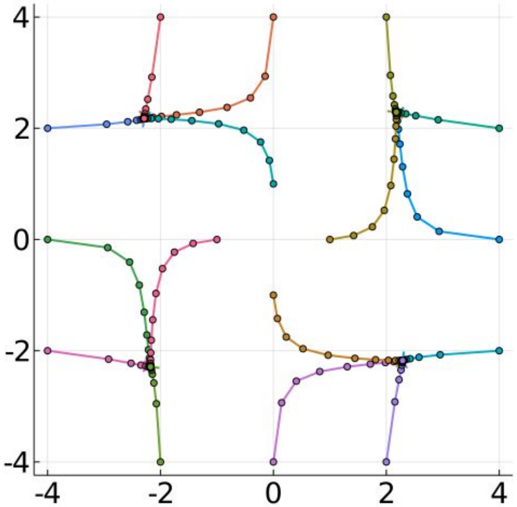
The interaction between *x* and *y* controls the algorithmic behavior.



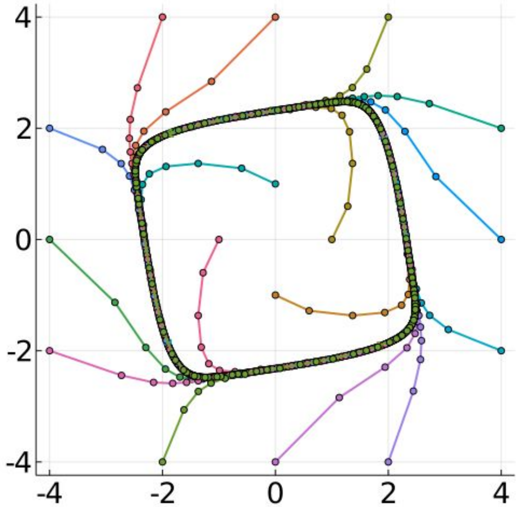
$A = 1$ $A = 10$ $A = 100$

This Convergence Landscape Holds in General!

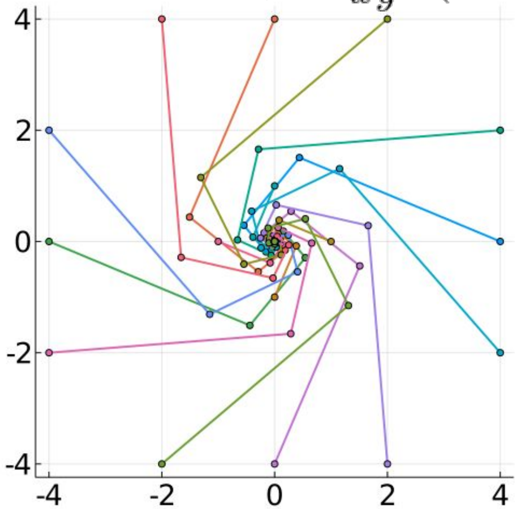
Convergence for generic minimax problems is controlled by $\nabla^2_{xy} L(x, y)$.



Interaction Weak Regime:
Local convergence occurs when $\nabla^2_{xy} L(x, y)$ is sufficiently bounded and Lipschitz.

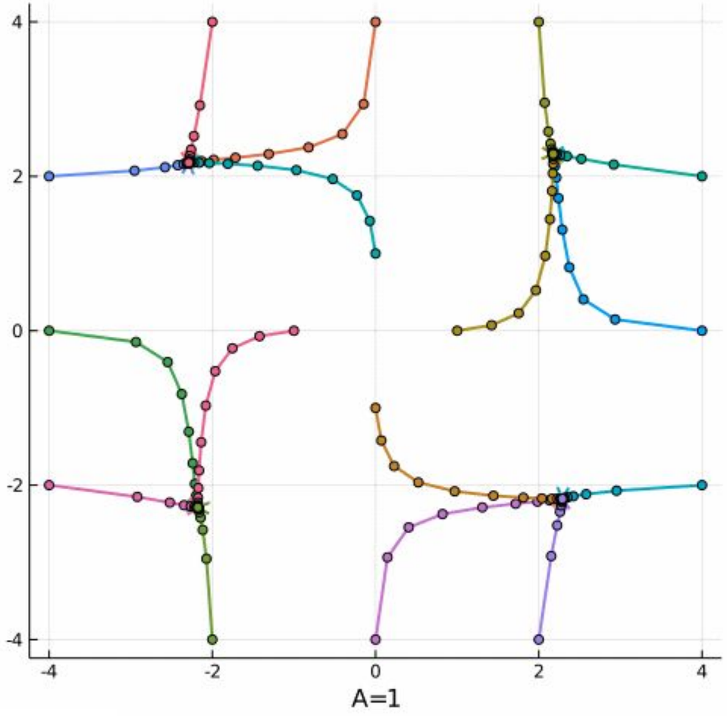


Interaction Moderate Regime:
Cycling and divergence can occur, preventing guarantees.



Interaction Dominate Regime:
Global convergence occurs when $\nabla^2_{xy} L(x, y)$ dominates any negative curvature in $\nabla^2_{xx} L(x, y), -\nabla^2_{yy} L(x, y)$

Formalizing our Landscape Picture



Formalizing our Landscape Picture

We consider unconstrained problems

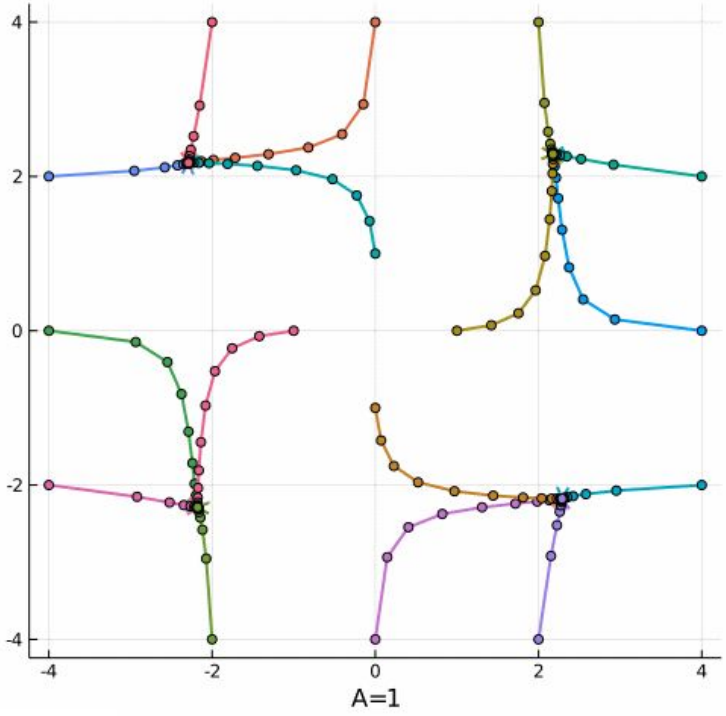
$$\min_{x \in \mathbb{R}^n} \max_{y \in \mathbb{R}^m} L(x, y)$$

with a twice differentiable objective,

and apply the Proximal Point Method

$$(x_{k+1}, y_{k+1}) = \text{prox}_\eta(x_k, y_k)$$

$$:= \arg \min_{u \in \mathbb{R}^n} \max_{v \in \mathbb{R}^m} L(u, v) + \frac{\eta}{2}\|u - x_k\|^2 - \frac{\eta}{2}\|v - y_k\|^2 \ .$$



Classic Convergence Review

Classically, an objective function is ***β*-smooth** if

$$\|\nabla L(z) - \nabla L(z')\| \leq \beta \|z - z'\|$$

and ***μ>0*-strongly convex-strongly concave** if

$$\nabla^2_{xx} L(z) \succeq \mu I \;, \quad -\nabla^2_{yy} L(z) \succeq \mu I \;.$$

Theorem. Under these two conditions, Gradient Descent Ascent (GDA)

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} - s \begin{bmatrix} \nabla_x L(x_k, y_k) \\ -\nabla_y L(x_k, y_k) \end{bmatrix}$$

linearly converges to the unique minimax solution for small enough *s*.

Our Convergence Assumptions

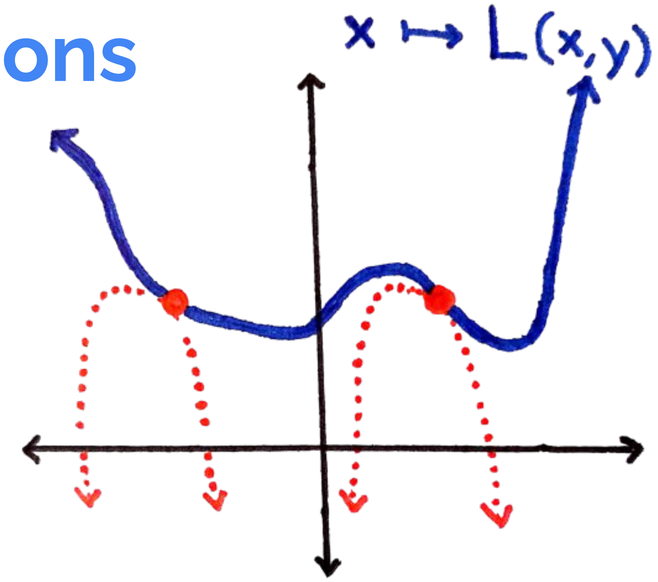
We avoid both of these strong assumptions.

We only assume ***ρ*-weak convexity** in x

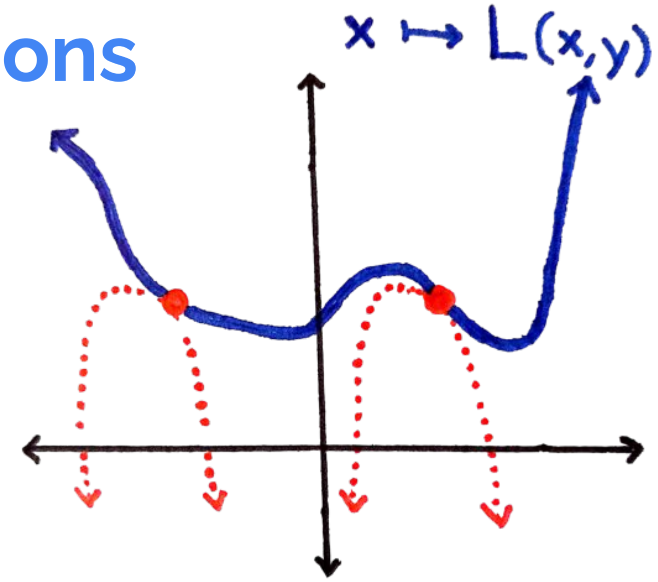
$$\nabla^2_{xx} L(z) \succeq -\rho I \ ,$$

and symmetrically, ***ρ*-weak concavity** in y

$$-\nabla^2_{yy} L(z) \succeq -\rho I \ .$$



Our Convergence Assumptions



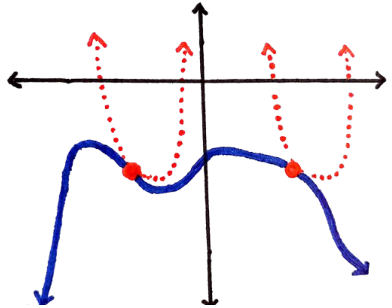
We avoid both of these strong assumptions.

We only assume ***ρ*-weak convexity** in x

$$\nabla^2_{xx} L(z) \succeq -\rho I \ ,$$

and symmetrically, ***ρ*-weak concavity** in y

$$-\nabla^2_{yy} L(z) \succeq -\rho I \ .$$



Core Tool

We consider the **saddle envelope** of [Attouch and Wets, 1983]

$$L_\eta(x, y) := \min_{u \in \mathbb{R}^n} \max_{v \in \mathbb{R}^m} L(u, v) + \frac{\eta}{2}\|u - x\|^2 - \frac{\eta}{2}\|v - y\|^2 .$$

(This generalizes the Moreau envelope, of which I am a huge fan.)

Core Tool

We consider the **saddle envelope** of [Attouch and Wets, 1983]

$$L_\eta(x, y) := \min_{u \in \mathbb{R}^n} \max_{v \in \mathbb{R}^m} L(u, v) + \frac{\eta}{2}\|u - x\|^2 - \frac{\eta}{2}\|v - y\|^2 .$$

(This generalizes the Moreau envelope, of which I am a huge fan.)

My dog ``*Moreau*``

Core Tool

We consider the **saddle envelope** of [Attouch and Wets, 1983]

$$L_\eta(x,y) := \min_{u\in\mathbb{R}^n} \max_{v\in\mathbb{R}^m} L(u,v) + \frac{\eta}{2}\|u-x\|^2 - \frac{\eta}{2}\|v-y\|^2 \ .$$

(This generalizes the Moreau envelope, of which I am a huge fan.)

Insights for Nonconvex-Nonconcave Objectives.

(i) The saddle envelope closely follows *L(x,y)*.

(ii) The saddle envelope is β-smooth, even if *L(x,y)* isn't.

(iii) The saddle envelope can be convex-concave, even if *L(x,y)* isn't.

Gradients of the Saddle Envelope

Proposition. The gradient of the saddle envelope is given by

$$\begin{bmatrix} \nabla_x L_\eta(x,y) \\ \nabla_y L_\eta(x,y) \end{bmatrix} = \begin{bmatrix} \eta(x - x_+) \\ \eta(y_+ - y) \end{bmatrix} = \begin{bmatrix} \nabla_x L(x_+, y_+) \\ \nabla_y L(x_+, y_+) \end{bmatrix}$$

where $(x_+, y_+) = \operatorname{prox}_\eta(x, y)$.

Gradients of the Saddle Envelope

Proposition. The gradient of the saddle envelope is given by

$$\begin{bmatrix} \nabla_x L_\eta(x,y) \\ \nabla_y L_\eta(x,y) \end{bmatrix} = \begin{bmatrix} \eta(x - x_+) \\ \eta(y_+ - y) \end{bmatrix} = \begin{bmatrix} \nabla_x L(x_+, y_+) \\ \nabla_y L(x_+, y_+) \end{bmatrix}$$

where $(x_+, y_+) = \mathrm{prox}_\eta(x, y)$.

Corollary 1. The saddle envelope preserves stationary points

$$\nabla L(x,y) = 0 \iff \nabla L_\eta(x,y) = 0\ .$$

Gradients of the Saddle Envelope

Proposition. The gradient of the saddle envelope is given by

$$\begin{bmatrix} \nabla_x L_\eta(x,y) \\ \nabla_y L_\eta(x,y) \end{bmatrix} = \begin{bmatrix} \eta(x - x_+) \\ \eta(y_+ - y) \end{bmatrix} = \begin{bmatrix} \nabla_x L(x_+, y_+) \\ \nabla_y L(x_+, y_+) \end{bmatrix}$$

where $(x_+, y_+) = \mathrm{prox}_\eta(x, y)$.

Corollary 2. Applying GDA on the saddle envelope with step-size *s=λ/η*

is equivalent to the following damped PPM on *L(x,y)*

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = (1 - \lambda) \begin{bmatrix} x_k \\ y_k \end{bmatrix} + \lambda\, \mathrm{prox}_\eta(x_k, y_k)\ .$$

Hessians of the Saddle Envelope

Proposition. The Hessian of the saddle envelope is given by

$$\begin{bmatrix} \nabla^2_{xx} L_\eta(z) & \nabla^2_{xy} L_\eta(z) \\ -\nabla^2_{yx} L_\eta(z) & -\nabla^2_{yy} L_\eta(z) \end{bmatrix} = \eta I - \eta^2 \left(\eta I + \begin{bmatrix} \nabla^2_{xx} L(z_+) & \nabla^2_{xy} L(z_+) \\ -\nabla^2_{yx} L(z_+) & -\nabla^2_{yy} L(z_+) \end{bmatrix} \right)^{-1}$$

where $z_+ = \mathrm{prox}_\eta(z)$.

Hessians of the Saddle Envelope

Proposition. The Hessian of the saddle envelope is given by

$$\begin{bmatrix} \nabla^2_{xx} L_\eta(z) & \nabla^2_{xy} L_\eta(z) \\ -\nabla^2_{yx} L_\eta(z) & -\nabla^2_{yy} L_\eta(z) \end{bmatrix} = \eta I - \eta^2 \left(\eta I + \begin{bmatrix} \nabla^2_{xx} L(z_+) & \nabla^2_{xy} L(z_+) \\ -\nabla^2_{yx} L(z_+) & -\nabla^2_{yy} L(z_+) \end{bmatrix} \right)^{-1}$$

where $z_+ = \mathrm{prox}_\eta(z)$.

Corollary 3. The saddle envelope is smooth with constant

$$\max\{\eta, |\eta^{-1} - \rho^{-1}|^{-1}\} \ .$$

For convex-concave problems, this simplifies to η-smoothness.

Hessians of the Saddle Envelope

Proposition. The Hessian of the saddle envelope is given by

$$\begin{bmatrix} \nabla^2_{xx} L_\eta(z) & \nabla^2_{xy} L_\eta(z) \\ -\nabla^2_{yx} L_\eta(z) & -\nabla^2_{yy} L_\eta(z) \end{bmatrix} = \eta I - \eta^2 \left(\eta I + \begin{bmatrix} \nabla^2_{xx} L(z_+) & \nabla^2_{xy} L(z_+) \\ -\nabla^2_{yx} L(z_+) & -\nabla^2_{yy} L(z_+) \end{bmatrix} \right)^{-1}$$

where $z_+ = \mathrm{prox}_\eta(z)$.

Corollary 4. The saddle envelope is strongly convex in *x* whenever

$$\nabla^2_{xx} L(z) + \nabla^2_{xy} L(z)(\eta I - \nabla^2_{yy} L(z))^{-1} \nabla^2_{yx} L(z) \succeq \alpha I$$

and strongly concave in *y* whenever

$$-\nabla^2_{yy} L(z) + \nabla^2_{yx} L(z)(\eta I + \nabla^2_{xx} L(z))^{-1} \nabla^2_{xy} L(z) \succeq \alpha I \ .$$

The Saddle Envelope Convexifies!

For example, the saddle envelope is convex-concave whenever

$$\frac{\nabla^2_{xy}L(z)\nabla^2_{yx}L(z)}{\eta+\beta} \succeq -\nabla^2_{xx}L(z)\,, \quad \frac{\nabla^2_{yx}L(z)\nabla^2_{xy}L(z)}{\eta+\beta} \succeq \nabla^2_{yy}L(z)$$

provided the objective has β-Lipschitz gradient in x and y separately.

The Saddle Envelope Convexifies!

For example, the saddle envelope is convex-concave whenever

$$\frac{\nabla^2_{xy}L(z)\nabla^2_{yx}L(z)}{\eta+\beta} \succeq -\nabla^2_{xx}L(z)\,, \quad \frac{\nabla^2_{yx}L(z)\nabla^2_{xy}L(z)}{\eta+\beta} \succeq \nabla^2_{yy}L(z)$$

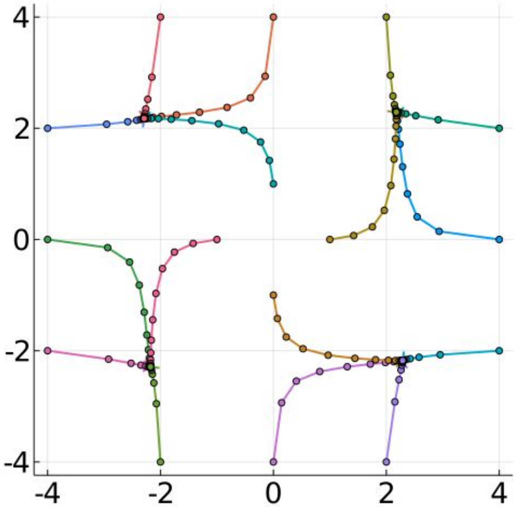
provided the objective has β-Lipschitz gradient in x and y separately.

Definition. We say an objective is ***α*-interaction dominant** if

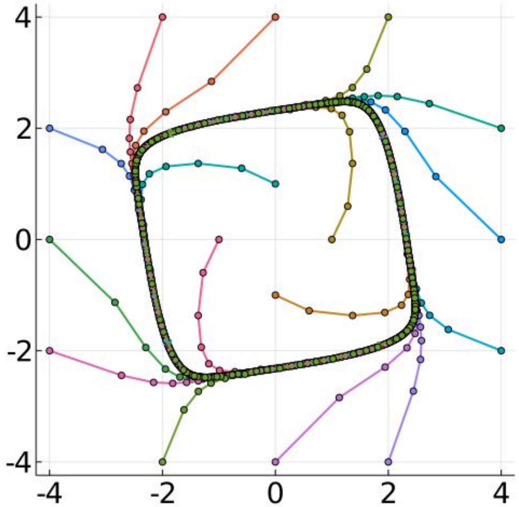
$$\nabla^2_{xx}L(z) + \nabla^2_{xy}L(z)(\eta I - \nabla^2_{yy}L(z))^{-1}\nabla^2_{yx}L(z) \succeq \alpha I$$
$$-\nabla^2_{yy}L(z) + \nabla^2_{yx}L(z)(\eta I + \nabla^2_{xx}L(z))^{-1}\nabla^2_{xy}L(z) \succeq \alpha I\ .$$

This Convergence Landscape Holds in General!

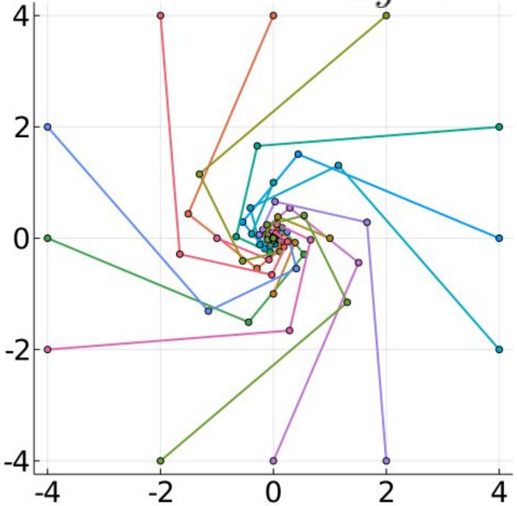
Convergence for generic minimax problems is controlled by $\nabla^2_{xy} L(x, y)$.



Interaction Weak Regime:
Local convergence occurs when $\nabla^2_{xy} L(x, y)$ is sufficiently bounded and Lipschitz.

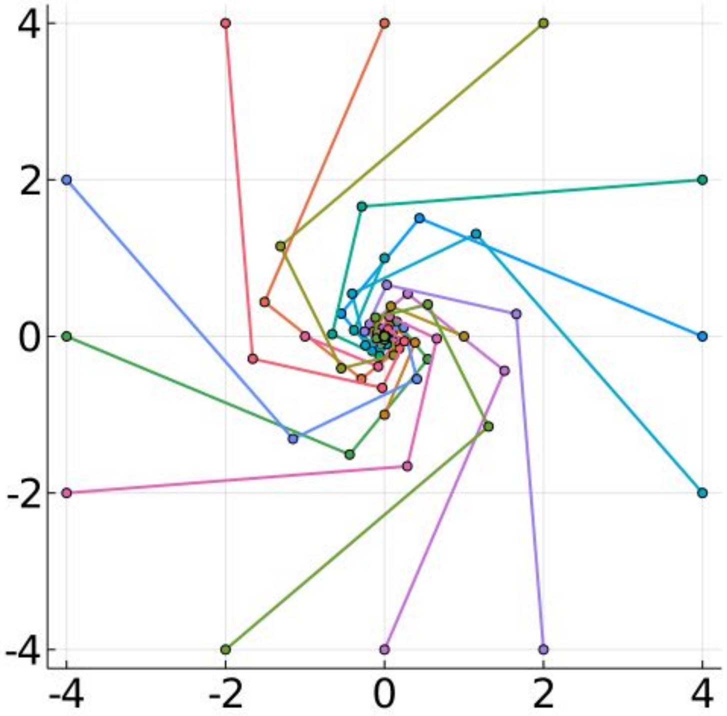


Interaction Moderate Regime:
Cycling and divergence can occur, preventing guarantees.



Interaction Dominate Regime:
Global convergence occurs when $\nabla^2_{xy} L(x, y)$ dominates any negative curvature in $\nabla^2_{xx} L(x, y), -\nabla^2_{yy} L(x, y)$

Interaction Dominant Convergence

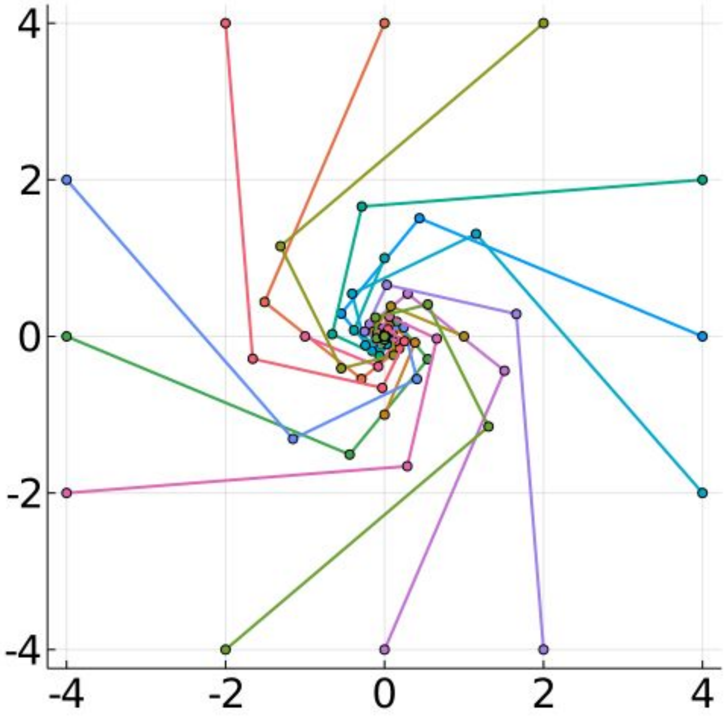


$$\nabla^2_{xy} L(x, y) = 100$$

Interaction Dominant Convergence

Theorem.
If $\alpha>0$-interaction dominant holds in x and y, the damped PPM with $\eta=2\rho$ and $\lambda=(1+\eta/\alpha)^{-1}$ converges to a stationary point with

$$\left\|\begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix}\right\|^2 \leq \left(1 - \frac{1}{(2\rho/\alpha+1)^2}\right)^k \left\|\begin{bmatrix} x_0 - x^* \\ y_0 - y^* \end{bmatrix}\right\|^2 .$$



$\nabla^2_{xy} L(x,y) = 100$

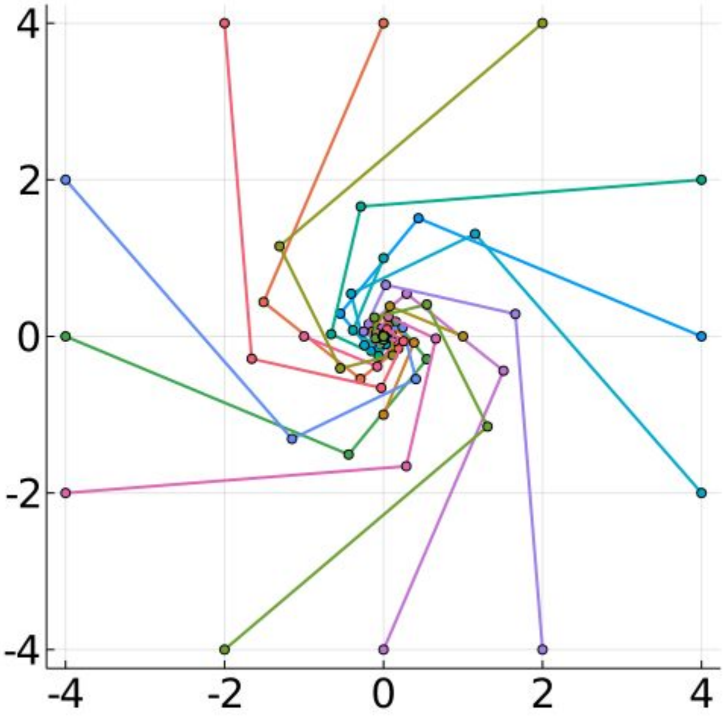
Interaction Dominant Convergence

Theorem.
If $\alpha>0$-interaction dominant holds in x and y, the damped PPM with $\eta=2\rho$ and $\lambda=(1+\eta/\alpha)^{-1}$ converges to a stationary point with

$$\left\|\begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix}\right\|^2 \leq \left(1 - \frac{1}{(2\rho/\alpha+1)^2}\right)^k \left\|\begin{bmatrix} x_0 - x^* \\ y_0 - y^* \end{bmatrix}\right\|^2 .$$

Proof Ingredients.

1. The saddle envelope is very structured.
2. GDA converges on the saddle envelope.
3. Equivalently, PPM converges on $L(x,y)$.



$\nabla^2_{xy} L(x,y) = 100$

One-Sided Interaction Dominant Convergence

Theorem. If α-interaction dominance holds in x or y, a PPM variant has

$$T \geq O(\varepsilon^{-2}) \implies \min_{k \leq T} \|\nabla L(z_k)\| \leq \varepsilon \ .$$

One-Sided Interaction Dominant Convergence

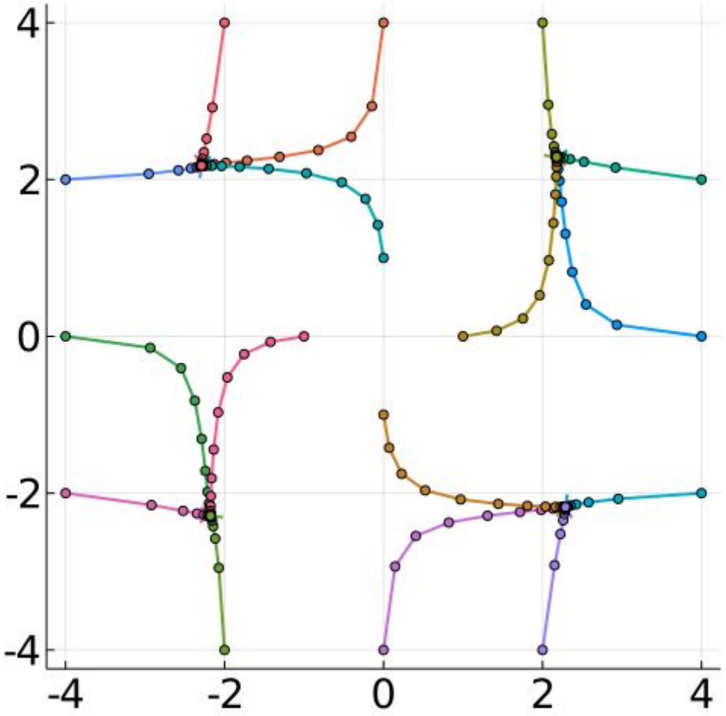
Theorem. If α-interaction dominance holds in x or y, a PPM variant has

$$T \geq O(\varepsilon^{-2}) \implies \min_{k \leq T} \|\nabla L(z_k)\| \leq \varepsilon\ .$$

Proof Ingredients.

1. The saddle envelope will still be smooth and nonconvex-concave.
2. [Lin et al., 2019] give a GDA variant for nonconvex-concave problems.
3. Thus a PPM variant works for such nonconvex-nonconcave problems.

Interaction Weak Convergence



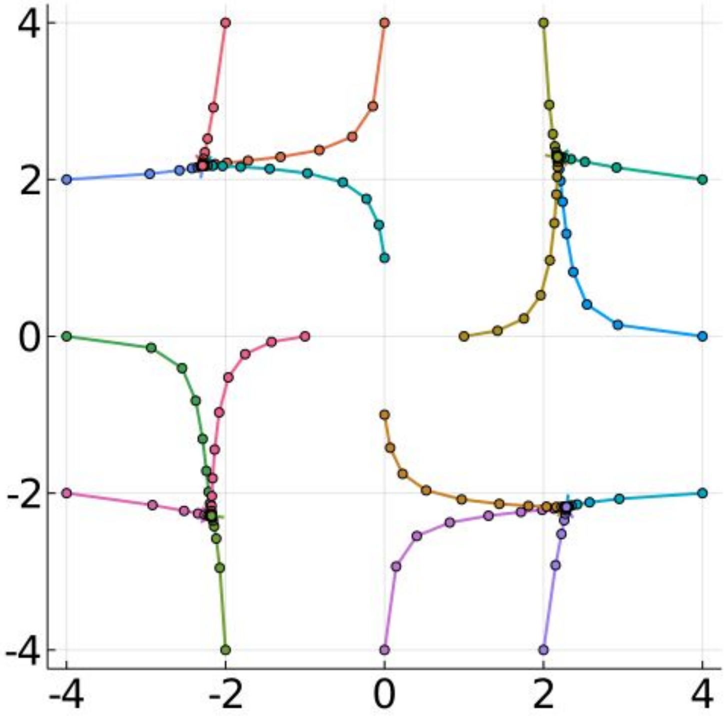
$$\nabla^2_{xy} L(x, y) = 1$$

Interaction Weak Convergence

If there was no interaction: $\nabla^2_{xy} L(z) = 0$

then a stationary point follows from solving

$$\begin{cases} x^* = \text{a local minimizer of } \min_u L(u, y') \\ y^* = \text{a local maximizer of } \max_v L(x', v). \end{cases}$$



$\nabla^2_{xy} L(x, y) = 1$

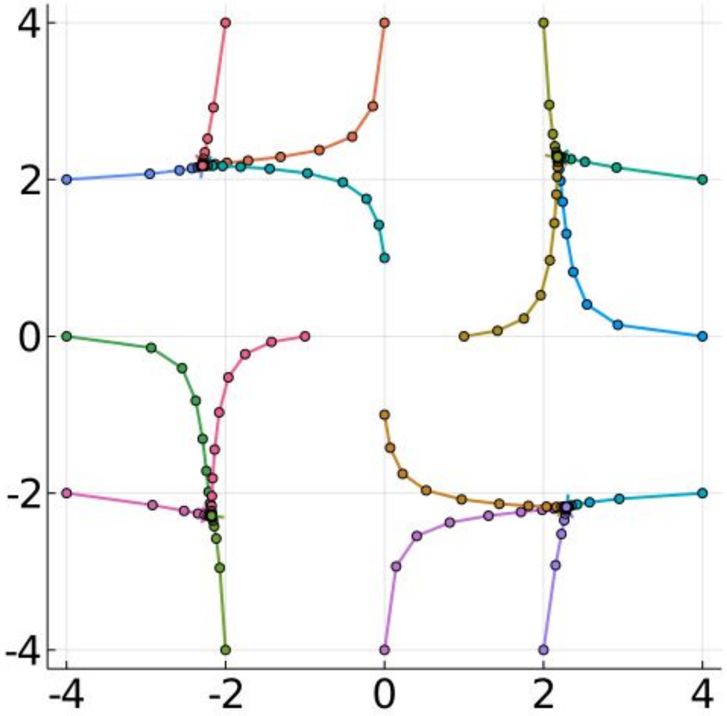
Interaction Weak Convergence

If there was no interaction: $\nabla^2_{xy} L(z) = 0$
then a stationary point follows from solving

$$\begin{cases} x^* = \text{a local minimizer of } \min_u L(u, y') \\ y^* = \text{a local maximizer of } \max_v L(x', v). \end{cases}$$

If interaction is small, we initialize PPM with

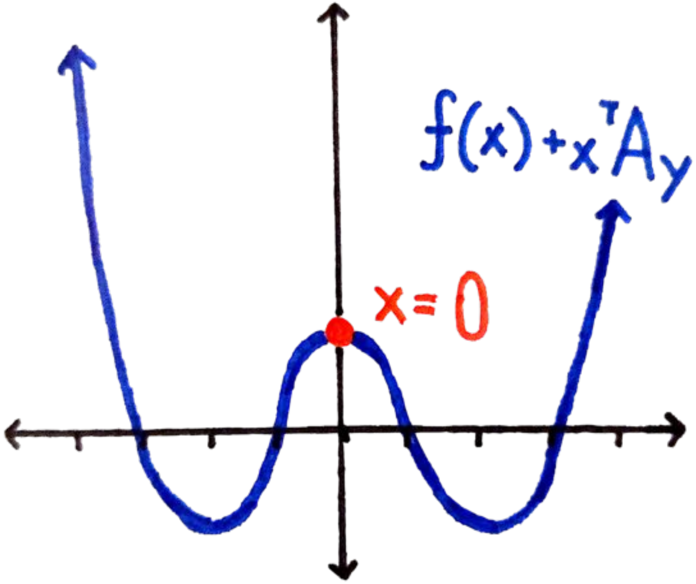
$$\begin{cases} x_0 = \text{a local minimizer of } \min_u L(u, y') \\ y_0 = \text{a local maximizer of } \max_v L(x', v) \end{cases}.$$

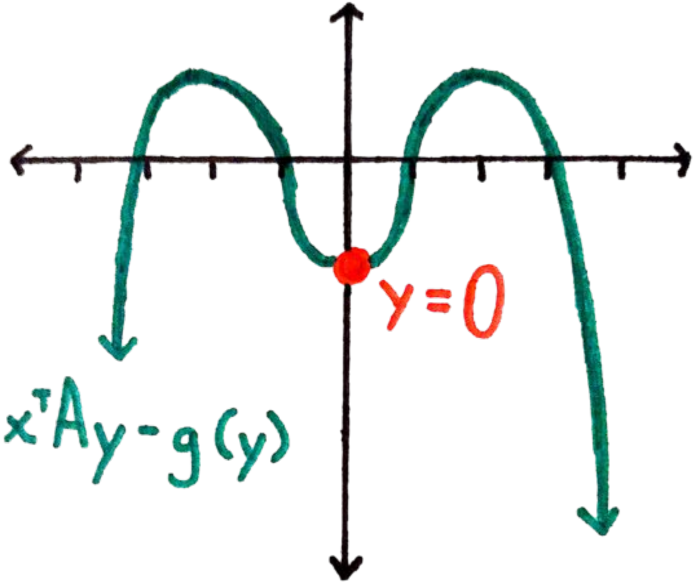


$\nabla^2_{xy} L(x, y) = 1$

Example Initialization when A is small

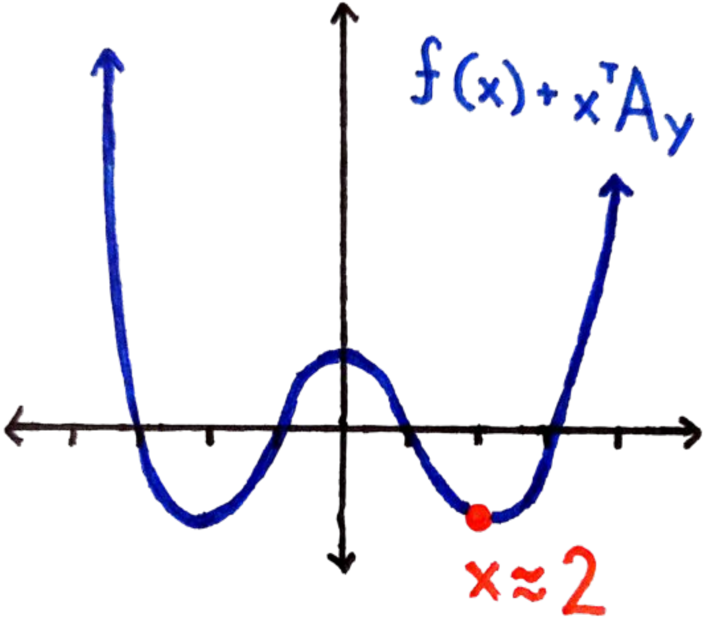
$$\begin{cases} x_0 = \text{a local minimizer of } \min_u L(u,0) \\ y_0 = \text{a local maximizer of } \max_v L(0,v) \end{cases}$$

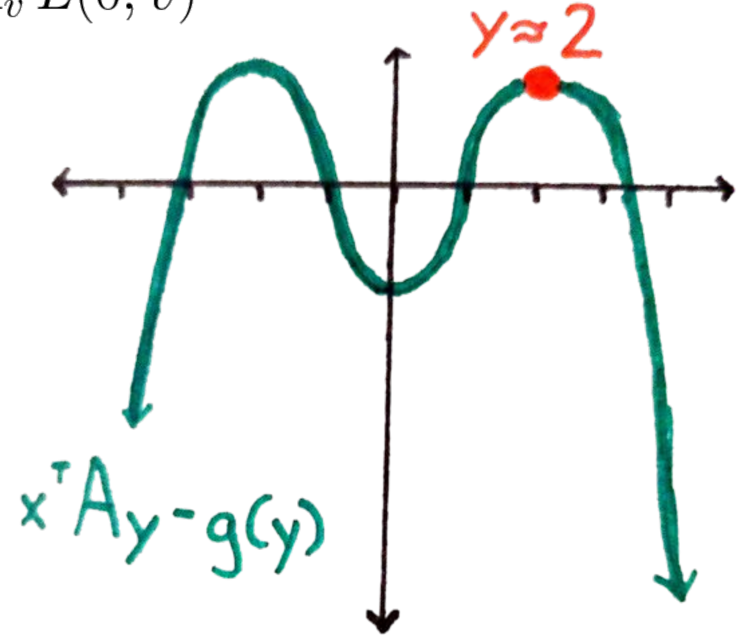




Example Initialization when A is small

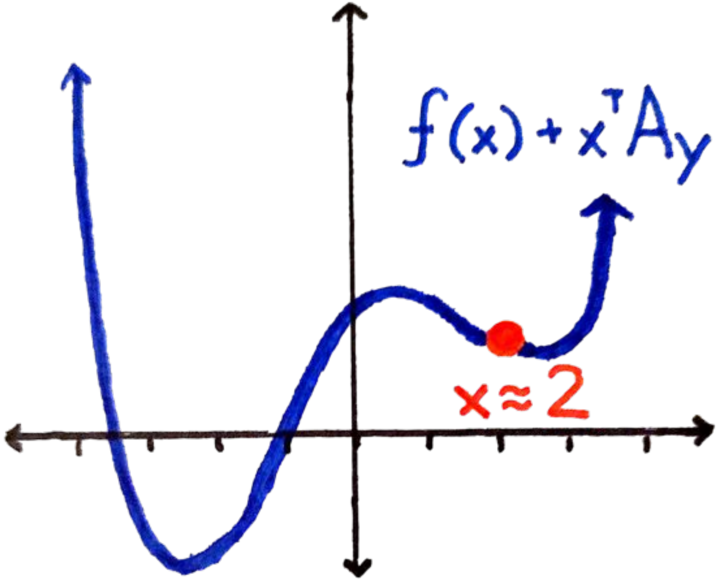
$$\begin{cases} x_0 = \text{a local minimizer of } \min_u L(u,0) \\ y_0 = \text{a local maximizer of } \max_v L(0,v) \end{cases}$$

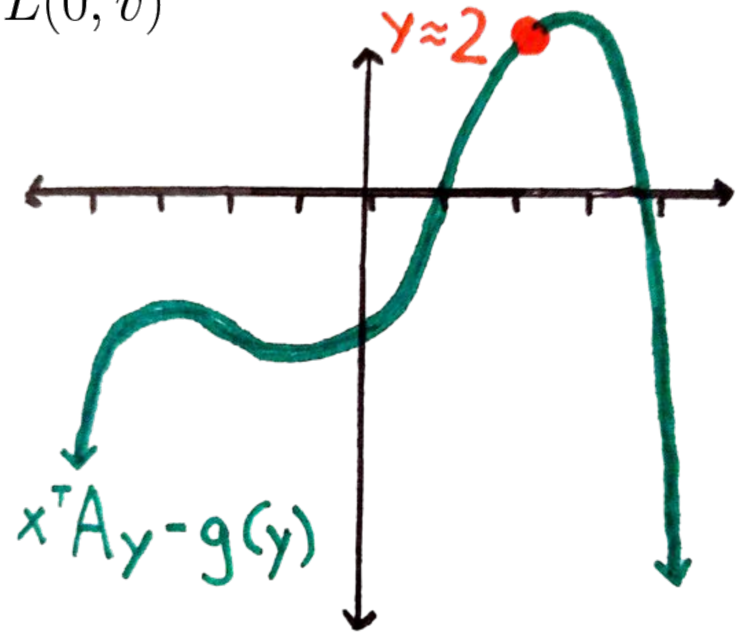




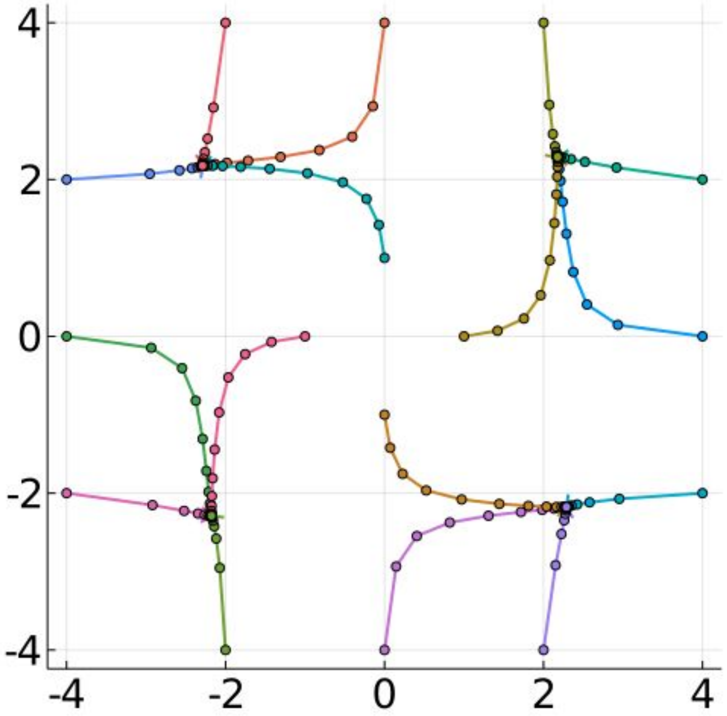
Example Initialization when A is small

$$\begin{cases} x_0 = \text{a local minimizer of } \min_u L(u,0) \\ y_0 = \text{a local maximizer of } \max_v L(0,v) \end{cases}$$





Interaction Weak Convergence



$$\nabla^2_{xy} L(x, y) = 1$$

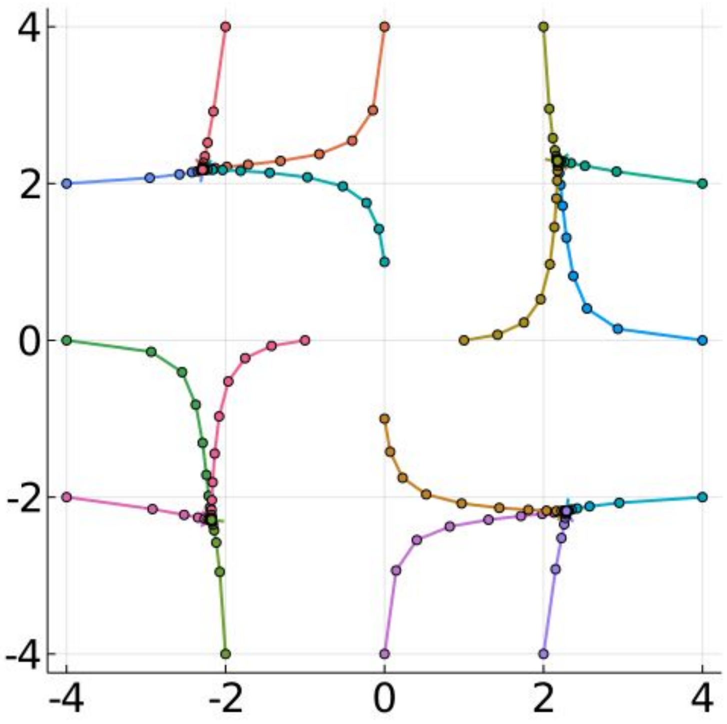
Interaction Weak Convergence

Theorem.

The damped PPM with $\eta=2\rho$ and $\lambda=(1+2\eta/\alpha_0)^{-1}$ converges to a stationary point with

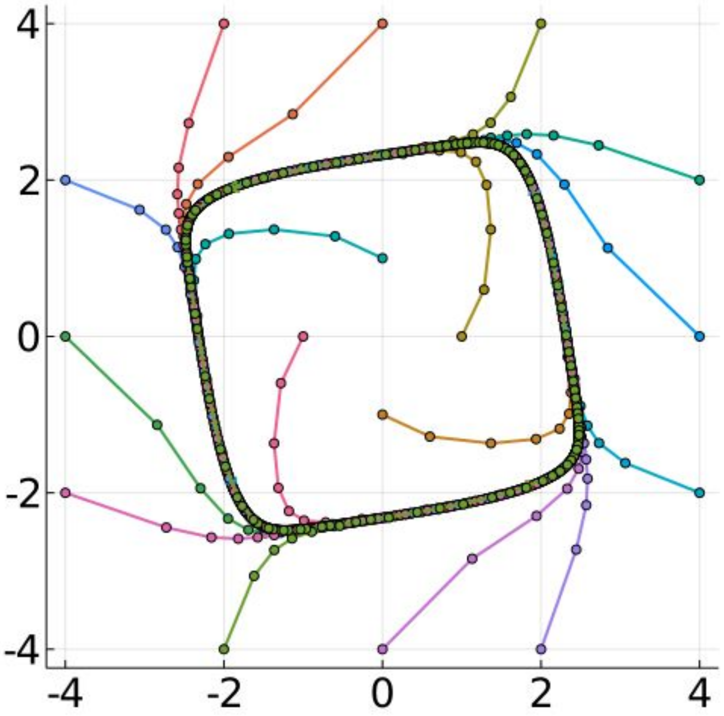
$$\left\|\begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix}\right\|^2 \le \left(1 - \frac{1}{(4\rho/\alpha_0+1)^2}\right)^k \left\|\begin{bmatrix} x_0 - x^* \\ y_0 - y^* \end{bmatrix}\right\|^2$$

provided $\|(x_0, y_0) - (x', y')\|$ is sufficiently small and the Hessian's interaction term is sufficiently small and Lipschitz.



$\nabla^2_{xy} L(x, y) = 1$

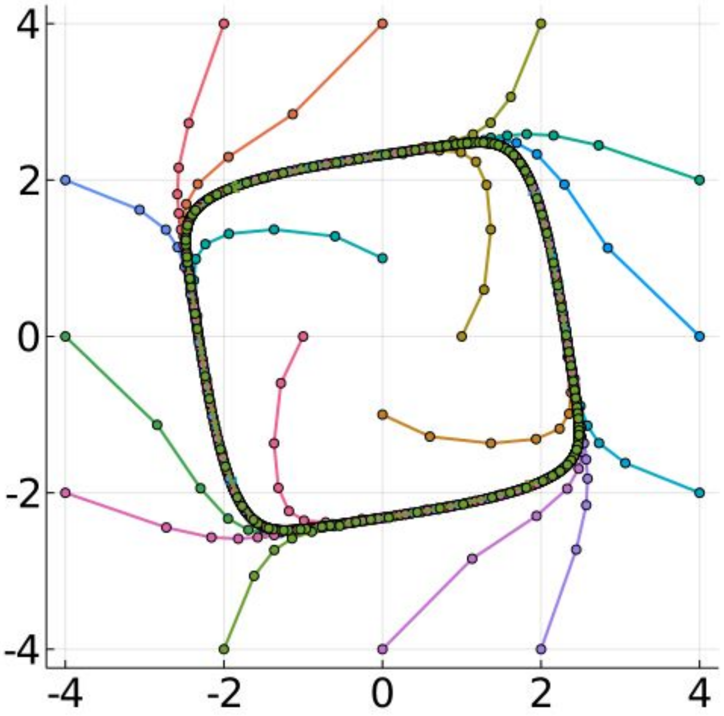
Interaction Moderate Cycling and Divergence



$$\nabla^2_{xy} L(x, y) = 10$$

Interaction Moderate Cycling and Divergence

Cycling. Our running example shows that globally attractive limit cycles can form.



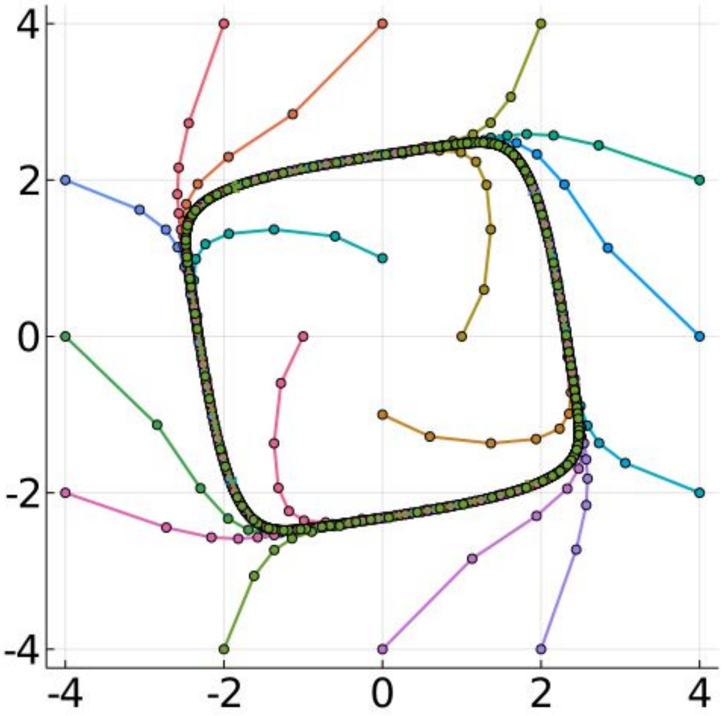
$$\nabla^2_{xy} L(x, y) = 10$$

Interaction Moderate Cycling and Divergence

Cycling. Our running example shows that globally attractive limit cycles can form.

Divergence. The boundary of our interaction dominant regime is tight. (For any *α≤0*, we can construct a diverging *α*-interaction dominant problem).

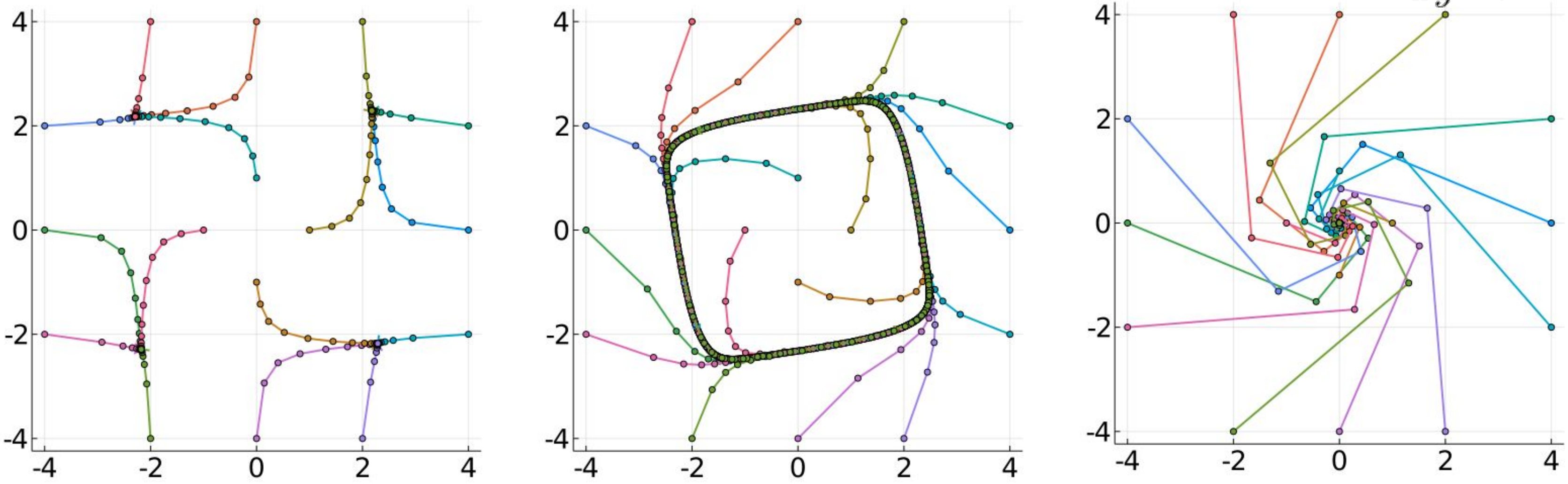
*See [**G.**, Lu, Worah, Mirrokni, 2020] for full details and some limited theory.*



$$\nabla^2_{xy} L(x,y) = 10$$

This Convergence Landscape Holds in General!

Convergence for generic minimax problems is controlled by $\nabla^2_{xy} L(x, y)$.



Interaction Weak Regime:
Local convergence occurs when $\nabla^2_{xy} L(x, y)$ is sufficiently bounded and Lipschitz.

Interaction Moderate Regime:
Cycling and divergence can occur, preventing guarantees.

Interaction Dominate Regime:
Global convergence occurs when $\nabla^2_{xy} L(x, y)$ dominates any negative curvature in $\nabla^2_{xx} L(x, y), -\nabla^2_{yy} L(x, y)$

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

(5 minutes) Minimax problems in learning.

(10 minutes) Difficulties in nonconvex-nonconcave regimes.

(20 minutes) One (optimizer's) path for avoiding these difficulties.

(5 minutes) Extensions and other paths forward.

Extension to Nonsmooth/Constrained Settings

We can no longer use second-order characterizations.

Extension to Nonsmooth/Constrained Settings

We can no longer use second-order characterizations.

Instead, we use a first-order operator characterization:

$$F(x, y) = \partial_x L(x, y) \times -\partial_y L(x, y)$$

***ρ*-weak convexity-weak concavity** becomes ``negative monotonicity``

$$\langle F(z) - F(z'), z - z' \rangle \geq -\rho \|z - z'\|^2$$

***0*-interaction dominance** becomes ``negative comonotonicity``

$$\langle F(z) - F(z'), z - z' \rangle \geq -\eta \|F(z) - F(z')\|^2$$

Extension to Other Algorithms?

Extension to Other Algorithms?

Positive Results:

If we additionally assume smoothness of $L(x,y)$,

the Extra-gradient Method (EGM) converges similarly.

No-So Positive Results:

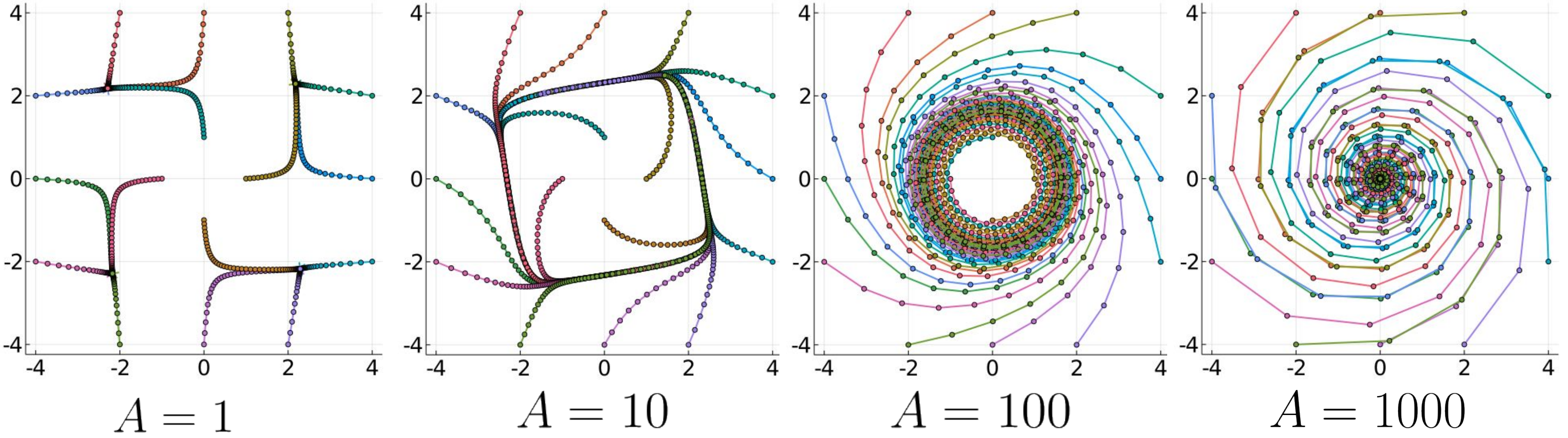
(Alternating) Gradient Descent Ascent follows a different landscape.

ODE tools can still give us some insights.

Landscape of the Extragradient Method

$$\begin{bmatrix} \hat{x} \\ \hat{y} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} + s \begin{bmatrix} -\nabla_x L(x_k, y_k) \\ \nabla_y L(x_k, y_k) \end{bmatrix}$$

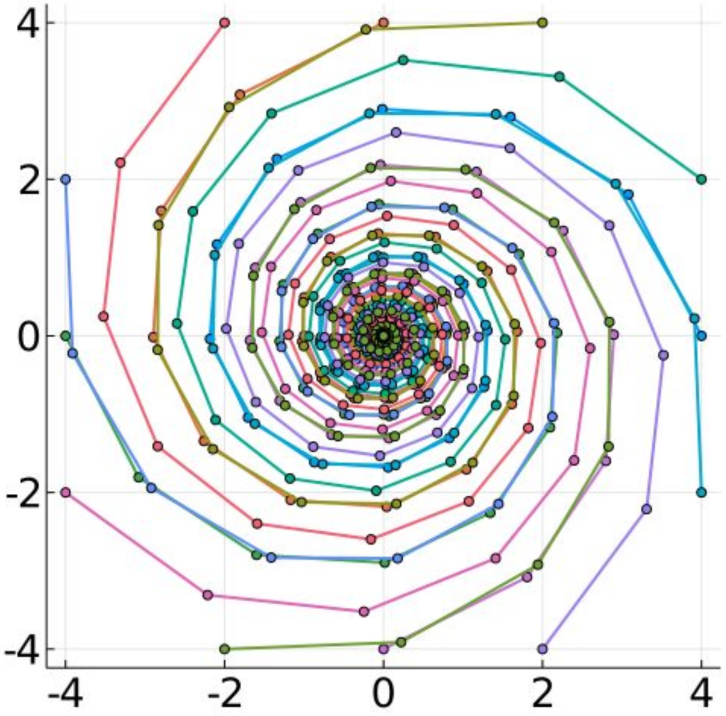
$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} + s \begin{bmatrix} -\nabla_x L(\hat{x}, \hat{y}) \\ \nabla_y L(\hat{x}, \hat{y}) \end{bmatrix}$$



Interaction Dominant Convergence for EGM

Theorem.
If the objective function *L(x,y)* is
(i) $\alpha>0$-interaction dominant in *x* and *y*,
(ii) sufficiently β-smooth,
(iii) stepsizes are chosen carefully,
then a damped EGM converges linearly.

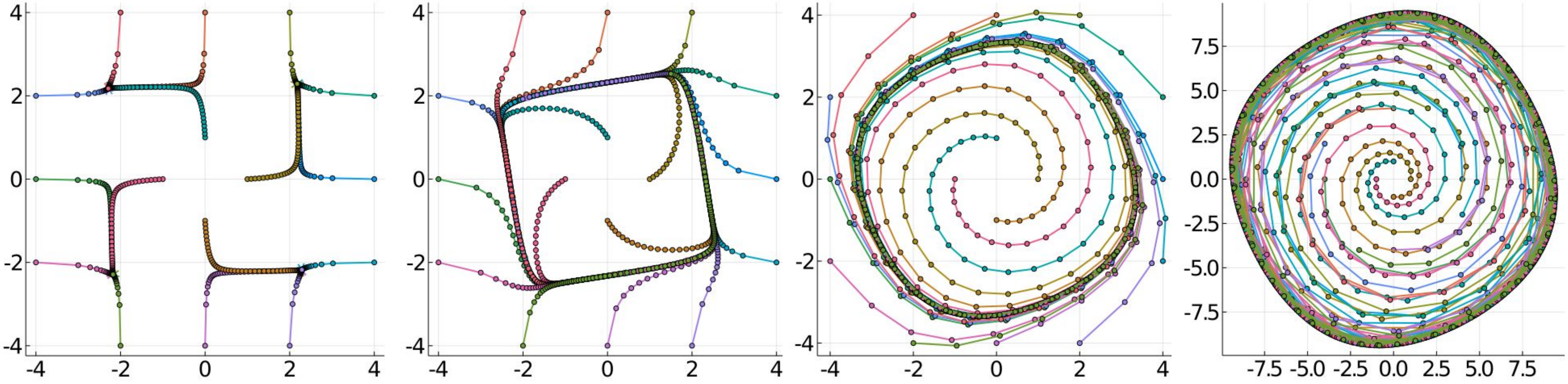
*See [Hajizadeh, Lu, **G.**, 2022] for full details.*



$$\nabla^2_{xy} L(x,y) = 1000$$

Landscape of Gradient Descent Ascent

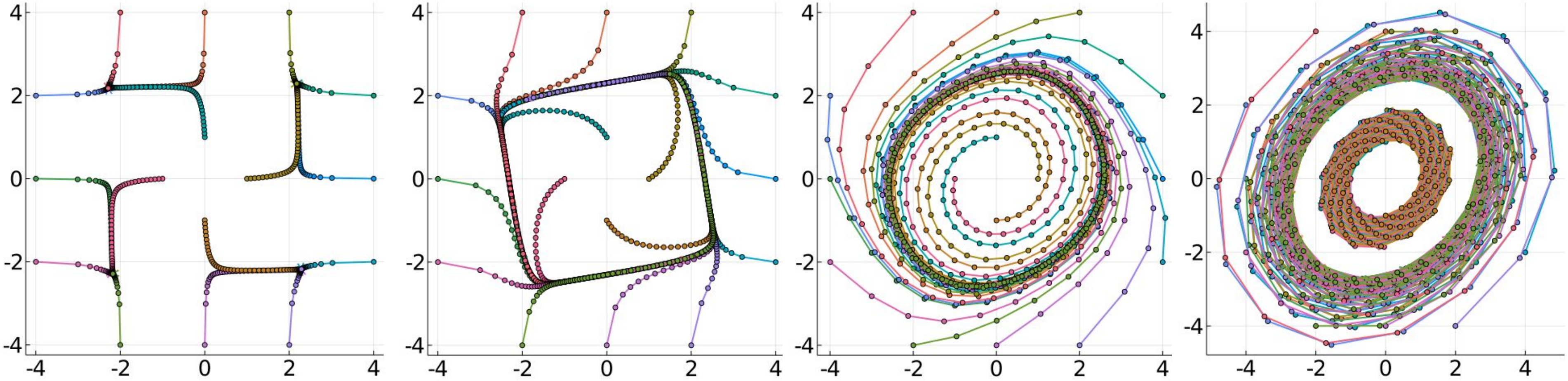
$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} + s \begin{bmatrix} -\nabla_x L(x_k, y_k) \\ \nabla_y L(x_k, y_k) \end{bmatrix}$$



$A = 1$ $\quad$ $A = 10$ $\quad$ $A = 100$ $\quad$ $A = 1000$

Landscape of Alternating Gradient Descent Ascent

$$x_{k+1} = x_k - s\nabla_x L(x_k, y_k)$$
$$y_{k+1} = y_k + s\nabla_y L(x_{k+1}, y_k)$$



$A = 1$ $A = 10$ $A = 100$ $A = 1000$

Failure to Extend to GDA and AGDA

We could study the ODE given as the *stepsize* goes to zero:

[Ratliff et al., 2014] [Nagarajan and Kolter, 2017]
[Mazumdar and Ratliff, 2019] [Vlatakis-Gkaragkounis et al., 2019], etc...

Alas, GDA, AGDA, and PPM all have the same ODE limit:

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -\nabla_x L(x,y) \\ \nabla_y L(x,y) \end{bmatrix}$$

and so, this ODE cannot describe differences in their behaviors.

Higher Order O(s)-ODE Approximations [Lu, 2020] [Shi et al, 2018]

Gradient Descent Ascent ODE

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -\nabla_x L(x,y) \\ \nabla_y L(x,y) \end{bmatrix} + \frac{s}{2} \begin{bmatrix} \nabla^2_{xx} L(x,y) & \nabla^2_{xy} L(x,y) \\ -\nabla^2_{yx} L(x,y) & -\nabla^2_{yy} L(x,y) \end{bmatrix} \begin{bmatrix} -\nabla_x L(x,y) \\ \nabla_y L(x,y) \end{bmatrix}$$

Proximal Point Method ODE

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -\nabla_x L(x,y) \\ \nabla_y L(x,y) \end{bmatrix} - \frac{s}{2} \begin{bmatrix} \nabla^2_{xx} L(x,y) & \nabla^2_{xy} L(x,y) \\ -\nabla^2_{yx} L(x,y) & -\nabla^2_{yy} L(x,y) \end{bmatrix} \begin{bmatrix} -\nabla_x L(x,y) \\ \nabla_y L(x,y) \end{bmatrix}$$

Alternating Gradient Descent Ascent ODE

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -\nabla_x L(x,y) \\ \nabla_y L(x,y) \end{bmatrix} + \frac{s}{2} \begin{bmatrix} \nabla^2_{xx} L(x,y) & \nabla^2_{xy} L(x,y) \\ \nabla^2_{yx} L(x,y) & -\nabla^2_{yy} L(x,y) \end{bmatrix} \begin{bmatrix} -\nabla_x L(x,y) \\ \nabla_y L(x,y) \end{bmatrix}$$

AGDA's ODE Convergence to Limit Points

Let $A = \nabla^2_{xx} L(x,y), B = \nabla^2_{xy} L(x,y), C = -\nabla^2_{yy} L(x,y)$.

Theorem. The AGDA ODE converges linearly in the norm $P = \begin{bmatrix} I & \frac{1}{2}sB^T \\ \frac{1}{2}sB & I \end{bmatrix}$ whenever (and not if the condition is strictly violated)

$$\begin{bmatrix} A + \frac{s}{2}A^2 + \frac{s^2}{4}(AB^TB + B^TBA) & \frac{s}{2}(AB^T + B^TC) + \frac{s^2}{4}(A^2B^T + B^TC^2) \\ \frac{s}{2}(B^TA + CB) + \frac{s^2}{4}(BA^2 + C^2B) & C + \frac{s}{2}C^2 + \frac{s^2}{4}(CBB^T + BB^TC) \end{bmatrix} \succ 0.$$

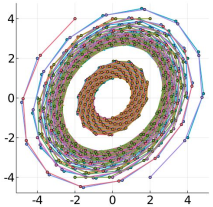
AGDA's ODE Convergence to Limit Points

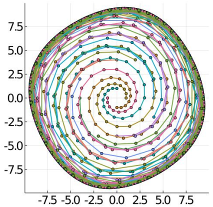
Let $A = \nabla^2_{xx} L(x,y), B = \nabla^2_{xy} L(x,y), C = -\nabla^2_{yy} L(x,y)$.

Theorem. The AGDA ODE converges linearly in the norm $P = \begin{bmatrix} I & \frac{1}{2} s B^T \\ \frac{1}{2} s B & I \end{bmatrix}$ whenever (and not if the condition is strictly violated)

$$\begin{bmatrix} A + \frac{s}{2}A^2 + \frac{s^2}{4}(AB^TB + B^TBA) & \frac{s}{2}(AB^T + B^TC) + \frac{s^2}{4}(A^2B^T + B^TC^2) \\ \frac{s}{2}(B^TA + CB) + \frac{s^2}{4}(BA^2 + C^2B) & C + \frac{s}{2}C^2 + \frac{s^2}{4}(CBB^T + BB^TC) \end{bmatrix} \succ 0.$$

This alternative norm needed for AGDA aligns with numerical observations:

AGDA 

vs GDA 

*[**G.**, Lu, Worah, Mirrokni, 2022]*

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

(5 minutes) Minimax problems in learning.

(10 minutes) Difficulties in nonconvex-nonconcave regimes.

(20 minutes) One (optimizer's) path for avoiding these difficulties.

(5 minutes) Extensions and other paths forward.

Thank You All for the *Fantastic* Workshop!

Questions?