Lecture 15: Extragradient

Start Recording!

Reminders

- Office Hours tomorrow with the TAs(11-12)
- Scribes notes of Lecture 7 and 9 are available.

References for this lecture:

1. Gidel, Gauthier, et al. "A variational inequality perspective on generative adversarial networks." ICLR 2019
2. Mokhtari, Aryan, Asuman Ozdaglar, and Sarath Pattathil. "A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach." *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020.
3. Azizian, Waïss, et al. "A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games." *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020.

Last Time:

GOAL:

$$\min_{\theta} \max_{\phi} \mathcal{L}(\theta, \phi)$$

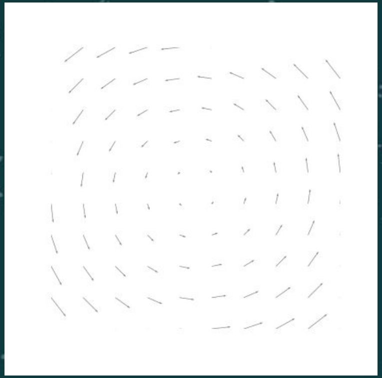
Where the payoff is convex-concave.

Example:

$$\min_{\theta} \max_{\phi} (\theta - \theta^*)^\top A(\phi - \phi^*)$$

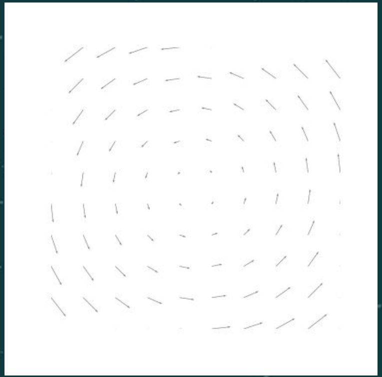
Summary

Simultaneous Gradient Descent-Ascent (Sim-GDA):



$$\begin{cases} \theta_{t+1} = \theta_t - \eta\phi_t \\ \phi_{t+1} = \phi_t + \eta\theta_t \end{cases}$$

Alternated Gradient Descent-Ascent (Alt-GDA)

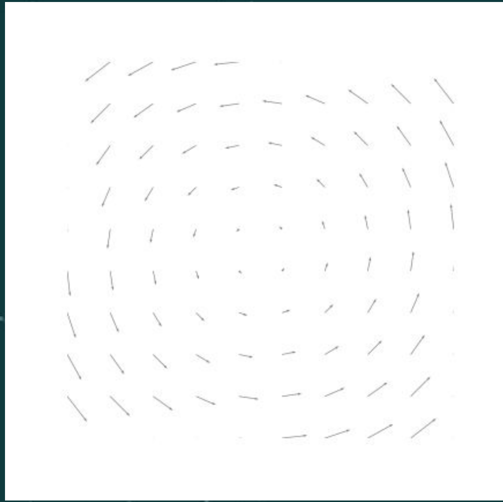


t+1 here????

$$\begin{cases} \theta_{t+1} = \theta_t - \eta\phi_t \\ \phi_{t+1} = \phi_t + \eta\theta_{t+1} \end{cases}$$

Proximal Point Method:

$$\begin{cases} \theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L}(\theta_{t+1}, \phi_{t+1}) \\ \phi_{t+1} = \phi_t + \eta \nabla_\phi \mathcal{L}(\theta_{t+1}, \phi_{t+1}) \end{cases}$$



Implicit Update: we need to solve a non-linear System

Variational Inequality Perspective

We only 'care' about the gradient-based updates, i.e., the vector field:

$$F(\theta_t, \phi_t) := \begin{pmatrix} \nabla_\theta \mathcal{L}(\theta_t, \varphi_t) \\ -\nabla_\phi \mathcal{L}(\theta_t, \varphi_t) \end{pmatrix}$$

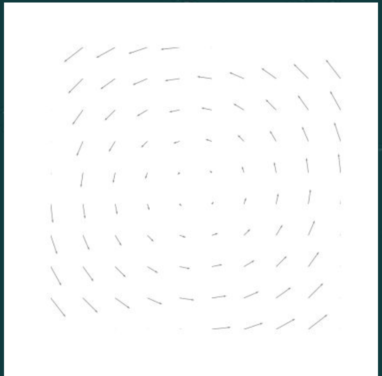
Previous plots. We represented the joint space (θ_t, ϕ_t)

More compact formalism:

$$\omega_t := (\theta_t, \phi_t)$$

Summary of the VIP

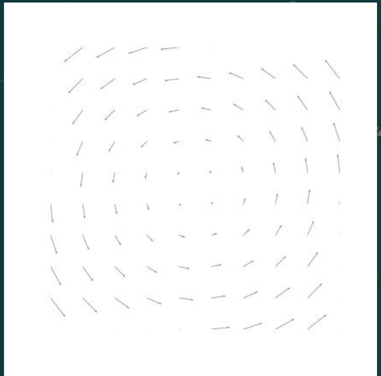
Sim-GDA:



$$\begin{cases}\theta_{t+1} = \theta_t - \eta\phi_t \\ \phi_{t+1} = \phi_t + \eta\theta_t\end{cases}$$

$$\omega_{t+1} = \omega_t - \eta F(\omega_t)$$

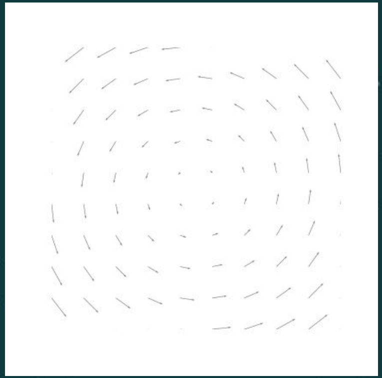
Prox-Point:



$$\begin{cases}\theta_{t+1} = \theta_t - \eta\nabla_\theta\mathcal{L}\boxed{(\theta_{t+1}, \phi_{t+1})} \\ \phi_{t+1} = \phi_t + \eta\nabla_\phi\mathcal{L}\boxed{(\theta_{t+1}, \phi_{t+1})}\end{cases}$$

$$\omega_{t+1} = \omega_t - \eta F(\boxed{\omega_{t+1}})$$

Alt-GDA

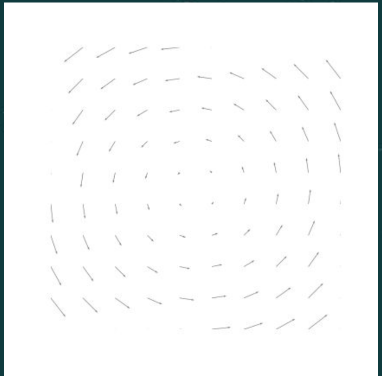


$$\begin{cases}\theta_{t+1} = \theta_t - \eta\phi_t \\ \phi_{t+1} = \phi_t + \eta\theta_{t+1}\end{cases}$$

???????

Summary of the VIP

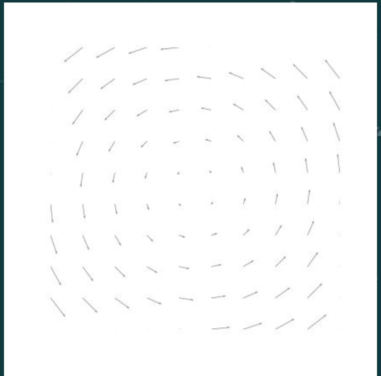
Sim-GDA:



$$\begin{cases}\theta_{t+1} = \theta_t - \eta\phi_t \\ \phi_{t+1} = \phi_t + \eta\theta_t\end{cases}$$

$$\omega_{t+1} = \omega_t - \eta F(\omega_t)$$

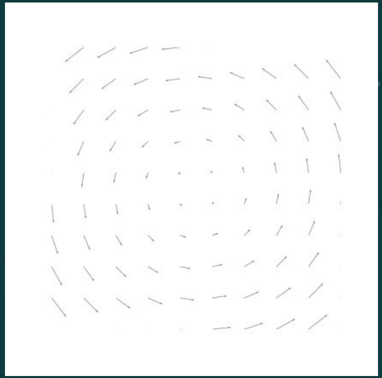
Prox-Point:



$$\begin{cases}\theta_{t+1} = \theta_t - \eta\nabla_\theta\mathcal{L}(\theta_{t+1}, \phi_{t+1}) \\ \phi_{t+1} = \phi_t + \eta\nabla_\phi\mathcal{L}(\theta_{t+1}, \phi_{t+1})\end{cases}$$

$$\omega_{t+1} = \omega_t - \eta F(\boxed{\omega_{t+1}})$$

Alt-GDA



$$\begin{cases}\theta_{t+1} = \theta_t - \eta\phi_t \\ \phi_{t+1} = \phi_t + \eta\theta_{t+1}\end{cases}$$

Not the right framework

Variational Inequality Perspective

<u>Goal:</u> Find a stationary point of the vector field:

$$F(\omega^*) = 0$$

<u>In zero sum game:</u> Equivalent to find a point with 0 gradient for each player

<u>If the game is convex concave:</u> equivalent to find a Nash!

Extragradient

Proximal Point method:

$$\omega_{t+1} = \omega_t - \eta F(\omega_{t+1})$$

Idea: Approximate ω_{t+1} with a gradient step

$$\omega_{t+1/2} = \omega_t - \eta F(\omega_t)$$

Extragradient

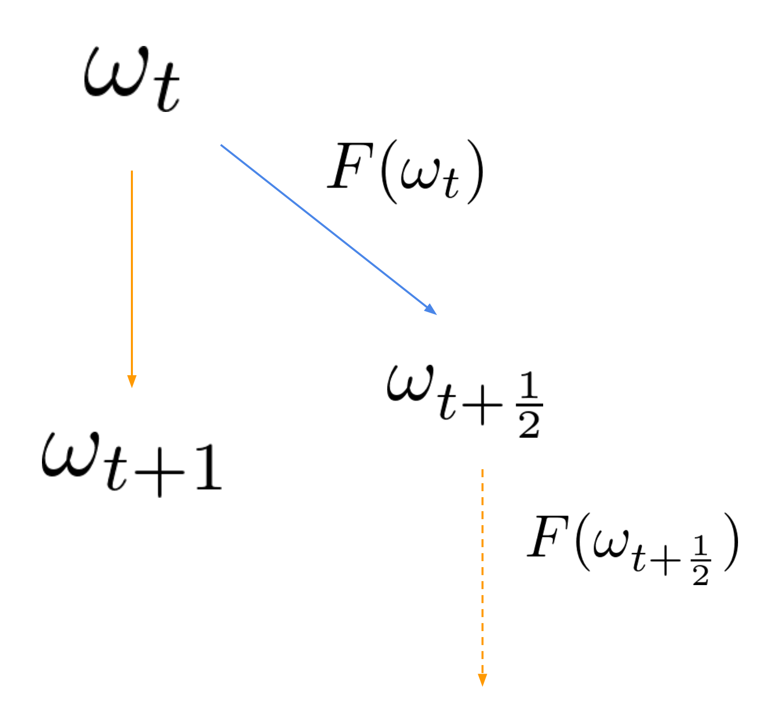
Extragradient:

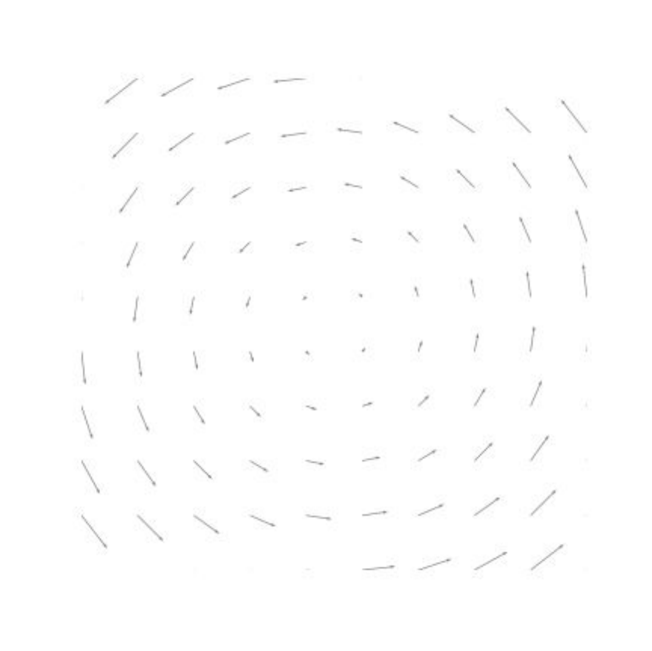
$$\omega_{t+1} = \omega_t - \eta F(\omega_{t+1/2})$$

$$\omega_{t+1/2} = \omega_t - \eta F(\omega_t)$$

Remark: Now the method is **explicit**!!!

ExtraGradient





Warm-up: Bilinear example

Exercice 1: Write the updates rules for EG for the following case

$$\min_{\theta} \max_{\phi} \phi \cdot \theta$$

Exercice 2: Show that for a small enough step-size:

$$\theta_t^2 + \phi_t^2 \leq \rho^t(\theta_0^2 + \phi_0^2) \quad \text{where} \quad 0 < \rho < 1$$

Standard Assumption

Definition: Monotone operator

$$\langle F(\omega) - F(\omega'), \omega - \omega' \rangle \geq 0\,, \quad \forall \omega, \omega'$$

Intuition: Generalization of convexity.

Exercice1: Show that f is a convex function then ∇f is a monotone operator.

Exercice2: For $\min_\theta \max_\phi \theta^\top A \phi$ we have $F(\theta, \phi) = \begin{pmatrix} A\phi \\ -A^\top \theta \end{pmatrix}$

Show that F is monotone

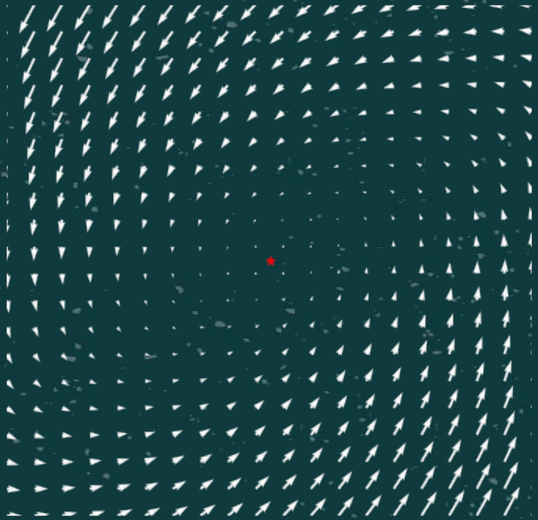
Intuition

Monotonicity implies: $\langle F(\omega), \omega^* - \omega_t \rangle \geq 0$

Examples

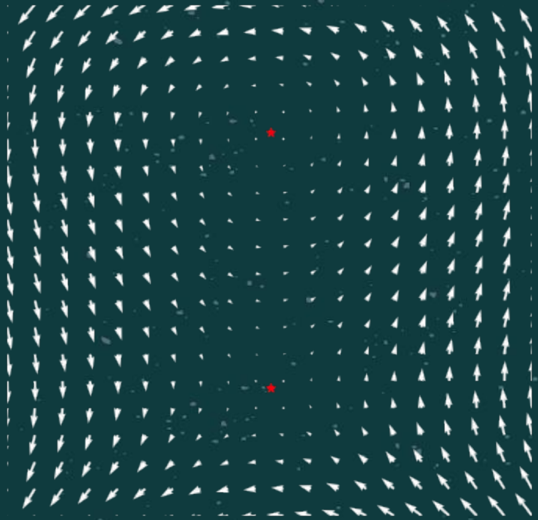
Example 1:

$$F(x,y) = \begin{pmatrix} -y \\ x - y \end{pmatrix}$$



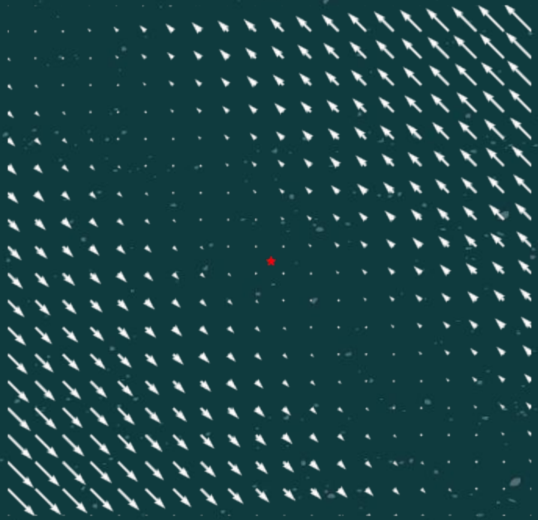
Example 2:

$$F(x,y) = \begin{pmatrix} (y - .5)(y + .5) \\ -x \end{pmatrix}$$



Example 3:

$$F(x,y) = \begin{pmatrix} -y - x \\ y + x \end{pmatrix}$$



Convergence of Extra Gradient (General case)

Lemma for Gradient Descent:

$$\|\theta_{t+1} - \theta^*\|_2^2 = \|\theta_t - \theta^*\|_2^2 \underbrace{- 2\eta\, g(\theta_t)^\top(\theta_t - \theta^*)}_{\text{Local Progress thanks to monotonicity}} + \overbrace{\|\theta_{t+1} - \theta_t\|_2^2}^{\text{Error due to discretization}}$$

Lemma for EG:

Distance to the optimum

$$\|\omega_{t+1} - \omega^*\|_2^2 = \|\omega_t - \omega^*\|_2^2 \overbrace{- 2\eta F(\omega_{t+1/2})^\top(\omega_{t+1/2} - \omega^*)}^{\text{Local Progress thanks to monotonicity}}$$

$$\underbrace{+\eta^2\|F(\omega_{t+1/2}) - F(\omega_t)\|_2^2 - \|\omega_{t+1/2} - \omega_t\|_2^2}_{\text{Error due to discretization}}$$

Convergence of Extra Gradient (General case)

Lemma for Gradient Descent:

$$\|\theta_{t+1} - \theta^*\|_2^2 = \|\theta_t - \theta^*\|_2^2 \underbrace{- 2\eta\, g(\theta_t)^\top (\theta_t - \theta^*)} + \overbrace{\|\theta_{t+1} - \theta_t\|_2^2}^{\text{Error due to discretization}}$$

Lemma for

Question: (Hattie) Can we tell where does Gradient Fail?

$$\|\omega_{t+1} - \omega^*\|_2^2 = \|\omega_t - \omega^*\|_2^2 \overbrace{- 2\eta F(\omega_{t+1/2})^\top (\omega_{t+1/2} - \omega^*)}$$

$$\underbrace{+\eta^2 \|F(\omega_{t+1/2}) - F(\omega_t)\|_2^2 - \|\omega_{t+1/2} - \omega_t\|_2^2}_{\text{Error due to discretization}}$$

Need to control the Error Term

Error Term:

$$+\boxed{\eta^2}\underbrace{\|F(\omega_{t+1/2}) - F(\omega_t)\|_2^2}_{\boxed{\text{Want this to be not too big}}} - \underbrace{\|\omega_{t+1/2} - \omega_t\|_2^2}_{\boxed{\text{Negative term!!!!}}}$$

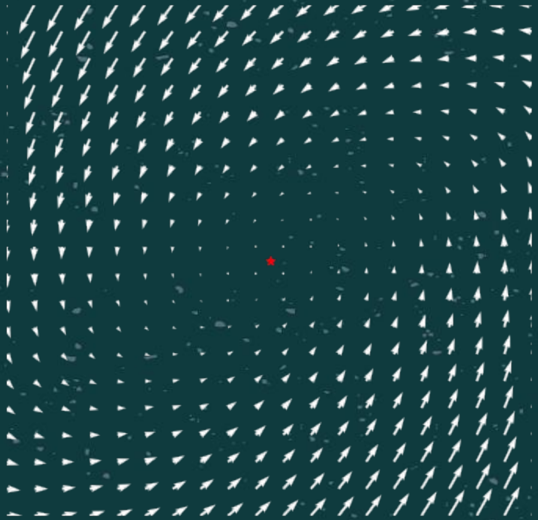
Lipschitz Operator:

$$\|F(\omega) - F(\omega')\| \leq L\|\omega - \omega'\|$$

Examples

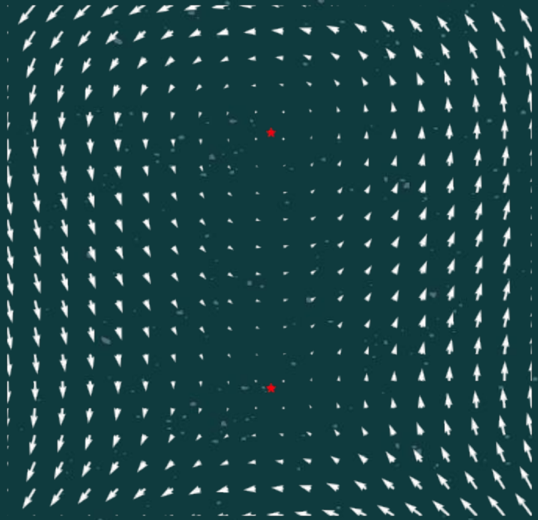
Example 1:

$$F(x,y) = \begin{pmatrix} -y \\ x - y \end{pmatrix}$$



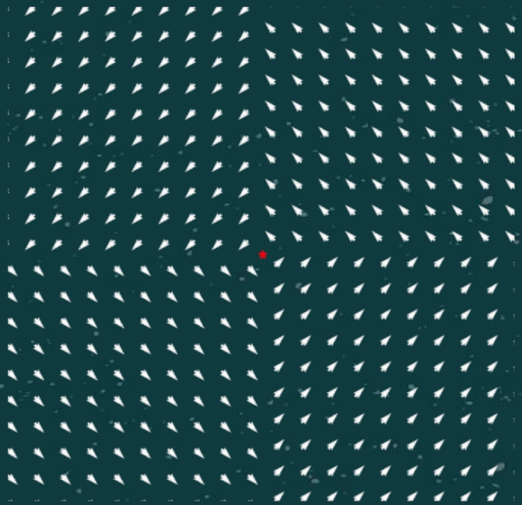
Example 2:

$$F(x,y) = \begin{pmatrix} (y-.5)(y+.5) \\ -x \end{pmatrix}$$



Example 3:

$$F(x,y) = \begin{pmatrix} -sign(y) \\ sign(x) \end{pmatrix}$$



Examples

Example 1:

$$F(x,y) = \begin{pmatrix} -y \\ x - y \end{pmatrix}$$

Example 2:

$$F(x,y) = \begin{pmatrix} (y - .5)(y + .5) \\ -x \end{pmatrix}$$

Example 3:

$$F(x,y) = \begin{pmatrix} -sign(y) \\ sign(x) \end{pmatrix}$$

Question: (Hattie) How do we check Lipschitzness in practice?

Convergence ExtraGradient

Lemma for EG + Monotonicity:

Distance to the optimum

Local Progress thanks to monotonicity

$$\|\omega_{t+1} - \omega^*\|_2^2 = \|\omega_t - \omega^*\|_2^2 \overbrace{- 2\eta F(\omega_{t+1/2})^\top (\omega_{t+1/2} - \omega^*)}$$

$$\left.+(\eta^2 L^2 - 1)\|\omega_{t+1/2} - \omega_t\|_2^2\right\}$$

$$< \|\omega_t - \omega^*\|_2^2$$

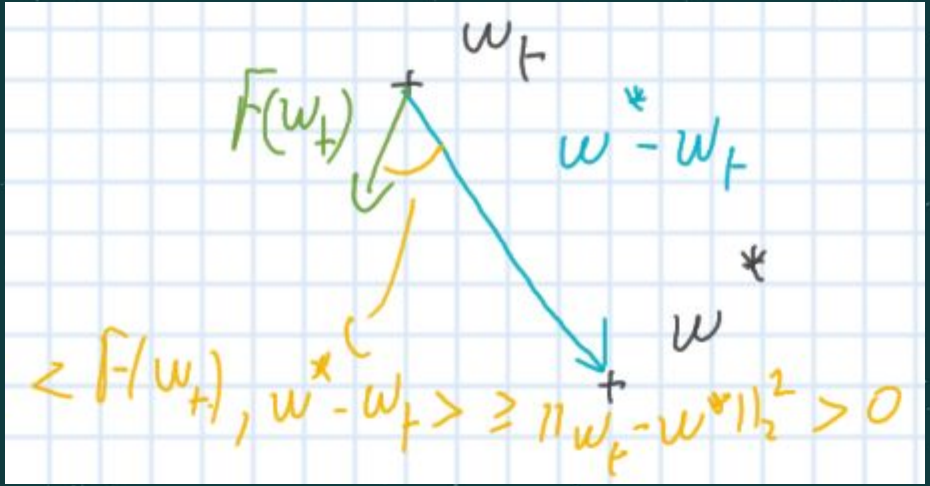
Error due to discretization controlled thanks to Lipschitzness

With small enough step-size

Strongly Monotone Operator

Definition: Strongly Monotone operator (generalization of strongly convex functions)

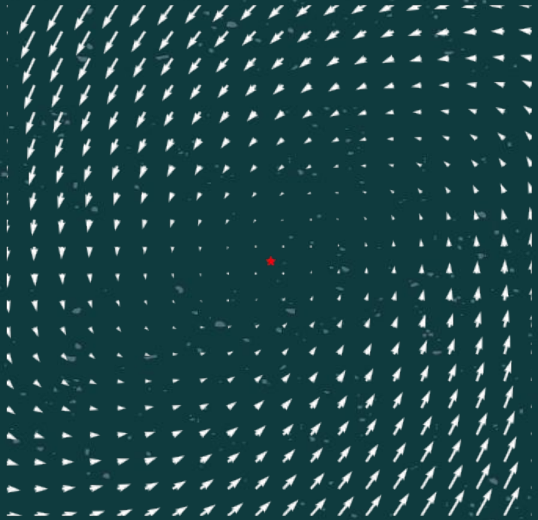
$$\langle F(\omega) - F(\omega'), \omega - \omega' \rangle \geq \mu \|\omega - \omega'\|_2^2$$



Examples

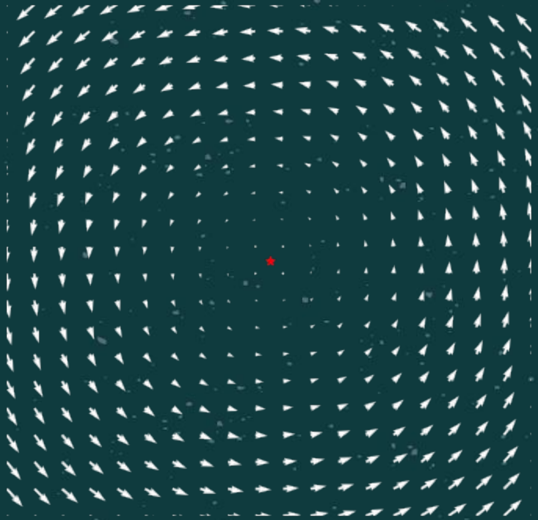
Example 1:

$$F(x,y) = \begin{pmatrix} -y \\ x-y \end{pmatrix}$$



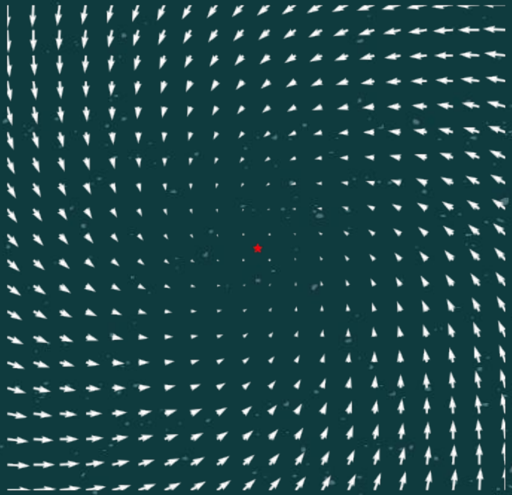
Example 2:

$$F(x,y) = \begin{pmatrix} -y \\ x \end{pmatrix}$$



Example 3:

$$F(x,y) = \begin{pmatrix} -y-x \\ x-y \end{pmatrix}$$



Convergence Result

Theorem: L-Lipchitz operator

1. If The operator is strongly monotone: (for \eta = 1/4L)

$$\|\omega_t - \omega^*\|_2^2 \leq (1 - \frac{\mu}{4L})^t \|\omega_0 - \omega^*\|_2^2$$

2. If the Operator is monotone: (be can get better than this)

$$\|\omega_t - \omega^*\|_2^2 \to 0$$

Convergence Result

Theorem: L-Lipchitz operator

1. If The operator is strongly monotone: (for \eta = 1/4L)

$$\|\omega_t - \omega^*\|_2^2 \leq (1 - \frac{\mu}{4L})^t \|\omega_0 - \omega^*\|_2^2$$

2. If

Question: (Safwen) Why ¼ ???

$$\|\omega_t - \omega^*\|_2^2 \to 0$$

Optimistic Method

Extragradient:

$$\omega_{t+1/2} = \omega_t - \eta F(\omega_t)$$
$$\omega_{t+1} = \omega_t - \eta F(\omega_{t+1/2})$$

Idea: Since $\omega_t \approx \omega_{t-1/2}$ and we have already computed $F(\omega_{t-1/2})$

Optimistic method:

$$\omega_{t+1/2} = \omega_t - \eta F(\omega_{t-1/2})$$
$$\omega_{t+1} = \omega_t - \eta F(\omega_{t+1/2})$$

Optimistic method

Optimistic method:

$$\omega_{t+1/2} = \omega_t - \eta F(\omega_{t-1/2})$$
$$\omega_{t+1} = \omega_t - \eta F(\omega_{t+1/2})$$

Equivalent formulation: (standard)

$$\omega_{t+1/2} = \omega_{t-1/2} - 2\eta F(\omega_{t-1/2}) + \eta F(\omega_{t-3/2})$$

Optimistic method

Opt

Take home message:

Optimistic method and Extragradient are very similar and have convergence results that are relatively equivalents.

Equ

Questions:

- (Mathieu and Carl) What if we do several extrapolations steps?
- (Andjela) What happens if the vector field is **not** monotone?
- (Pierluca) Can we go beyond saddle point games?

$\omega_t \quad\quad {}_2)$

Useful Links and refs:

- Mescheder, Lars, Andreas Geiger, and Sebastian Nowozin. "Which training methods for GANs do actually converge?." *International conference on machine learning*. PMLR, 2018.
- Mescheder, Lars, Sebastian Nowozin, and Andreas Geiger. "The numerics of gans." NeurIPS (2017).
- Azizian, Waïss, et al. "A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games." *AISTATS*, 2020.
- Sion, Maurice. "On general minimax theorems." *Pacific Journal of mathematics* 8.1 (1958): 171-176.