A Unified Analysis of Gradient-Based Methods for a Whole Spectrum of Differentiable Games

Waïss Azizian Ioannis Mitliagkas Simon Lacoste-Julien
Gauthier Gidel

AISTATS 2020

Joint work with...

Waïss Azizian[1]

Ioannis Mitliagkas[2]

Simon Lacoste-Julien[2]

Gauthier Gidel[2]

[1]École Normale Supérieure, Paris [2]Mila & DIRO, Université de Montréal

Overview

Motivation

- More and more ML frameworks formulated as games [Goodfellow et al., 2014; Madry et al., 2018].
- However, new challenges arise in game optimization, such as cycles [Balduzzi et al., 2018; Gidel et al., 2019b]

$\Rightarrow$ Some classes of games still poorly understood...

(Partial and biased) landscape of game optimization

Cooperative games: strongly monotone games

- Standard setting for last-iterate convergence guarantees
- Reasonable methods converge linearly (such as the gradient method [Rockafellar, 1976], extragradient [Tseng, 1995]...)

Bilinear example: Particular "adversarial" game

$$\min_{x \in \mathbb{R}^{d_1}} \max_{y \in \mathbb{R}^{d_2}} x^T A y + b^T x + c^T y$$

- Same cyclic behavior as in GAN training [Mescheder et al., 2017]: gradient method diverges! [Balduzzi et al., 2018; Gidel et al., 2019b]
- Variants have been introduced,
  - extragradient [Liang and Stokes, 2018; Gidel et al., 2019a]
  - optimistic gradient [Daskalakis et al., 2018]
  - consensus optimization [Mescheder et al., 2017], ...

$\Rightarrow$ Converge linearly on this particular example

Problems

- No unified analysis of the variants of the gradient method for both cooperative games and the bilinear example.

Problems

- No unified analysis of the variants of the gradient method for both cooperative games and the bilinear example.
- What happens for general adversarial games, i.e. games with no strong monotonicity ?

Problems

- No unified analysis of the variants of the gradient method for both cooperative games and the bilinear example.
- What happens for general adversarial games, i.e. games with no strong monotonicity ?
- What happens “in between”, i.e. for games with both a cooperative and an adversarial component ?

Contributions

For unconstrained n-player games,

- First local and global *unified* analysis of extragradient,

Contributions

For unconstrained n-player games,

- First local and global *unified* analysis of extragradient,
  - Unifies the results on cooperative games and the bilinear example.

Contributions

For unconstrained n-player games,

- First local and global *unified* analysis of extragradient,
  - Unifies the results on cooperative games and the bilinear example.
  - Linear convergence rate for non-strongly monotone adversarial games.

Contributions

For unconstrained n-player games,

- First local and global *unified* analysis of extragradient,
  - Unifies the results on cooperative games and the bilinear example.
  - Linear convergence rate for non-strongly monotone adversarial games.
  - In between extrgradient enjoys the *best of both world*.

Contributions

For unconstrained n-player games,

- First local and global *unified* analysis of extragradient,
  - Unifies the results on cooperative games and the bilinear example.
  - Linear convergence rate for non-strongly monotone adversarial games.
  - In between extrgradient enjoys the *best of both world*.
- Extend this analysis to optimistic gradient and consensus optimization.

Contributions

For unconstrained n-player games,

- First local and global *unified* analysis of extragradient,
  - Unifies the results on cooperative games and the bilinear example.
  - Linear convergence rate for non-strongly monotone adversarial games.
  - In between extrgradient enjoys the *best of both world*.
- Extend this analysis to optimistic gradient and consensus optimization.
- Lower bounds which show that extragradient is optimal among general extrapolation methods (without momentum).

Classes of games and local analysis of extragradient

Unconstrained two-player games

Player 1:
Parameter $\omega_1 \in \mathbb{R}^{d_1}$,
Goal: minimize loss $\ell_1(\omega_1, \omega_2)$

Player 2:
Parameter $\omega_2 \in \mathbb{R}^{d_2}$,
Goal: minimize loss $\ell_2(\omega_1, \omega_2)$

We want a *Nash equilibrium*: $(\omega_1^*, \omega_2^*) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ s.t.

$$\begin{cases} \omega_1^* \in \arg\min_{\omega_1 \in \mathbb{R}^{d_1}} \ell_1(\omega_1, \omega_2^*) \\ \omega_2^* \in \arg\min_{\omega_2 \in \mathbb{R}^{d_2}} \ell_2(\omega_1^*, \omega_2) \end{cases}$$

Gradient vector field

First-order condition: If $\ell_1(\cdot, \omega_2)$ and $\ell_2(\omega_1, \cdot)$ convex $\forall \omega_1, \omega_2$,

$$\begin{cases} \omega_1^* \in \arg\min_{\omega_1 \in \mathbb{R}^{d_1}} \ell_1(\omega_1, \omega_2^*) \\ \omega_2^* \in \arg\min_{\omega_2 \in \mathbb{R}^{d_2}} \ell_2(\omega_1^*, \omega_2) \end{cases} \iff \begin{cases} \nabla_{\omega_1} \ell_1(\omega^*) = 0 \\ \nabla_{\omega_2} \ell_2(\omega^*) = 0 \end{cases}$$

Gradient method:

$$\begin{cases} \omega_1^{t+1} = \omega_1^t - \eta \nabla_{\omega_1} \ell_1(\omega_1^t, \omega_2^t) \\ \omega_2^{t+1} = \omega_2^t - \eta \nabla_{\omega_2} \ell_2(\omega_1^t, \omega_2^t) \end{cases}$$

Gradient vector field

First-order condition: If $\ell_1(\cdot, \omega_2)$ and $\ell_2(\omega_1, \cdot)$ convex $\forall \omega_1, \omega_2$,

$$\begin{cases} \omega_1^* \in \arg\min_{\omega_1 \in \mathbb{R}^{d_1}} \ell_1(\omega_1, \omega_2^*) \\ \omega_2^* \in \arg\min_{\omega_2 \in \mathbb{R}^{d_2}} \ell_2(\omega_1^*, \omega_2) \end{cases} \iff \begin{cases} \nabla_{\omega_1} \ell_1(\omega^*) = 0 \\ \nabla_{\omega_2} \ell_2(\omega^*) = 0 \end{cases}$$

Gradient method:

$$\begin{cases} \omega_1^{t+1} = \omega_1^t - \eta \nabla_{\omega_1} \ell_1(\omega_1^t, \omega_2^t) \\ \omega_2^{t+1} = \omega_2^t - \eta \nabla_{\omega_2} \ell_2(\omega_1^t, \omega_2^t) \end{cases}$$

Can be rewritten using the *gradient vector field*:

$$v(\omega) = v(\omega_1, \omega_2) = \begin{pmatrix} \nabla_{\omega_1} \ell_1(\omega_1, \omega_2) \\ \nabla_{\omega_2} \ell_2(\omega_1, \omega_2) \end{pmatrix}$$

Gradient vector field

First-order condition: If $\ell_1(\cdot, \omega_2)$ and $\ell_2(\omega_1, \cdot)$ convex $\forall \omega_1, \omega_2$,

$$\begin{cases} \omega_1^* \in \underset{\omega_1 \in \mathbb{R}^{d_1}}{\arg\min}\, \ell_1(\omega_1, \omega_2^*) \\ \omega_2^* \in \underset{\omega_2 \in \mathbb{R}^{d_2}}{\arg\min}\, \ell_2(\omega_1^*, \omega_2) \end{cases} \iff \begin{cases} \nabla_{\omega_1} \ell_1(\omega^*) = 0 \\ \nabla_{\omega_2} \ell_2(\omega^*) = 0 \end{cases} \iff v(\omega^*) = 0$$

Gradient method:

$$\begin{cases} \omega_1^{t+1} = \omega_1^t - \eta \nabla_{\omega_1} \ell_1(\omega_1^t, \omega_2^t) \\ \omega_2^{t+1} = \omega_2^t - \eta \nabla_{\omega_2} \ell_2(\omega_1^t, \omega_2^t) \end{cases}$$

Can be rewritten using the *gradient vector field*:

$$v(\omega) = v(\omega_1, \omega_2) = \begin{pmatrix} \nabla_{\omega_1} \ell_1(\omega_1, \omega_2) \\ \nabla_{\omega_2} \ell_2(\omega_1, \omega_2) \end{pmatrix}$$

Gradient vector field

First-order condition: If $\ell_1(\cdot, \omega_2)$ and $\ell_2(\omega_1, \cdot)$ convex $\forall \omega_1, \omega_2$,

$$\begin{cases} \omega_1^* \in \underset{\omega_1 \in \mathbb{R}^{d_1}}{\arg\min}\, \ell_1(\omega_1, \omega_2^*) \\ \omega_2^* \in \underset{\omega_2 \in \mathbb{R}^{d_2}}{\arg\min}\, \ell_2(\omega_1^*, \omega_2) \end{cases} \iff \begin{cases} \nabla_{\omega_1} \ell_1(\omega^*) = 0 \\ \nabla_{\omega_2} \ell_2(\omega^*) = 0 \end{cases} \iff v(\omega^*) = 0$$

Gradient method:

$$\begin{cases} \omega_1^{t+1} = \omega_1^t - \eta \nabla_{\omega_1} \ell_1(\omega_1^t, \omega_2^t) \\ \omega_2^{t+1} = \omega_2^t - \eta \nabla_{\omega_2} \ell_2(\omega_1^t, \omega_2^t) \end{cases} \iff \omega^{t+1} = \omega^t - \eta v(\omega^t)$$

Can be rewritten using the *gradient vector field*:

$$v(\omega) = v(\omega_1, \omega_2) = \begin{pmatrix} \nabla_{\omega_1} \ell_1(\omega_1, \omega_2) \\ \nabla_{\omega_2} \ell_2(\omega_1, \omega_2) \end{pmatrix}$$

Problem: Given a vector field v, find ω^* s.t. $v(\omega^*) = 0$.

Spectral properties govern local behaviour

Around ω^*:

$$v(\omega) \approx \underbrace{v(\omega^*)}_{=0} + \nabla v(\omega^*)(\omega - \omega^*)$$

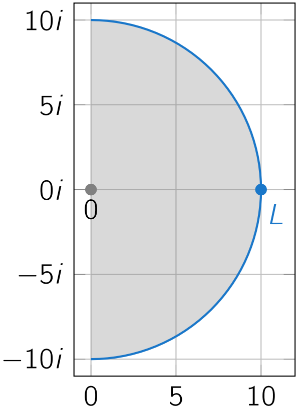
Main idea:
Local behavior of a method $\longleftrightarrow$ Properties of $\operatorname{Sp} \nabla v(\omega^*)$.

Cooperative games

Assumptions:

- v Lipschitz ≈

$$|\lambda| \leq L, \quad \forall \lambda \in \mathrm{Sp}\, \nabla v(x^*)$$



Cooperative games

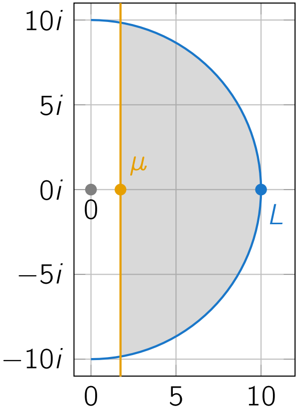
Assumptions:

- v Lipschitz ≈

$$|\lambda| \leq L, \quad \forall \lambda \in \mathrm{Sp}\, \nabla v(x^*)$$

- Strong monotonicity ≈

$$\Re\lambda \geq \mu, \quad \forall \lambda \in \mathrm{Sp}\, \nabla v(x^*)$$



Cooperative games

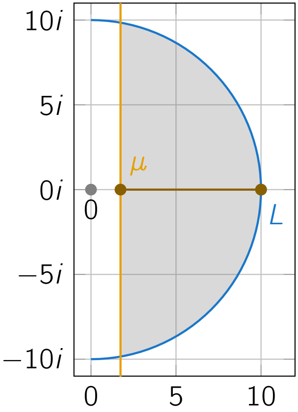
Assumptions:

- v Lipschitz ≈

$$|\lambda| \leq L, \quad \forall \lambda \in \operatorname{Sp} \nabla v(x^*)$$

- Strong monotonicity ≈

$$\Re\lambda \geq \mu, \quad \forall \lambda \in \operatorname{Sp} \nabla v(x^*)$$



For strongly convex optimization:

$$\operatorname{Sp} \nabla v(\omega^*) = \operatorname{Sp} \nabla^2 f(\omega^*) \subset [\mu, L] \quad \text{with} \quad \mu > 0$$

Cooperative games

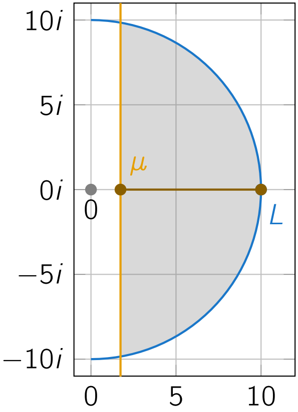
Assumptions:

- v Lipschitz $\approx$

$$|\lambda| \leq L, \quad \forall \lambda \in \operatorname{Sp} \nabla v(x^*)$$

- Strong monotonicity $\approx$

$$\Re\lambda \geq \mu, \quad \forall \lambda \in \operatorname{Sp} \nabla v(x^*)$$



Lemma (Bertsekas [1999]; Gidel et al. [2019b])

Gradient method converges linearly at ω^* iff

$$\forall \lambda \in \operatorname{Sp} \nabla v(\omega^*), \ \Re\lambda > 0$$

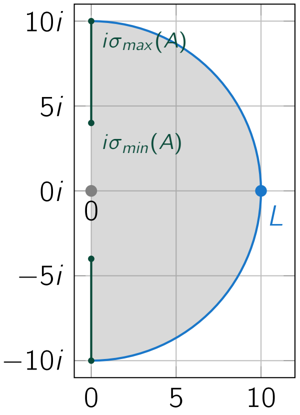
Bilinear game

For $A \in \mathbb{R}^{m \times m}$, $b, c \in \mathbb{R}^m$,

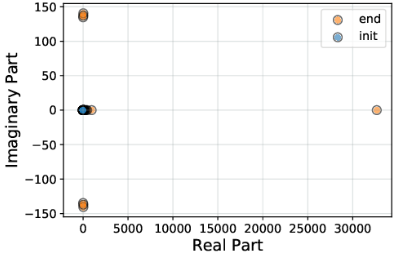
$$\min_{x \in \mathbb{R}^m} \max_{y \in \mathbb{R}^m} x^T A y + b^T x + c^T y$$

Spectrum:

$$\operatorname{Sp} \nabla v(\omega^*) = \{\pm i\sigma \,|\, \sigma^2 \in \operatorname{Sp} AA^T\}$$

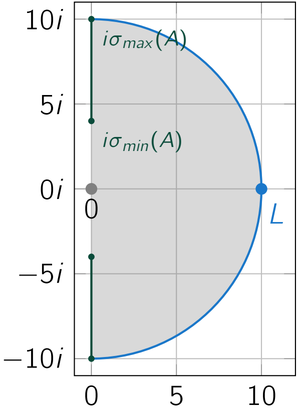


Bilinear game

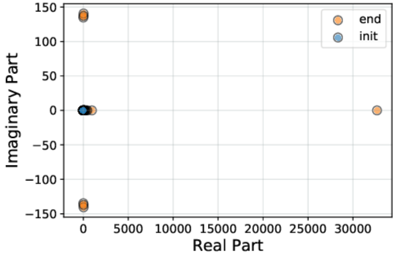


(c) NSGAN on MNIST

From Berard et al. [2020]

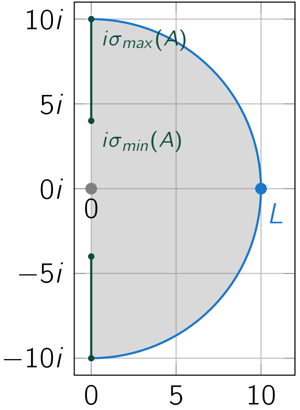


Bilinear game



(c) NSGAN on MNIST

From Berard et al. [2020]

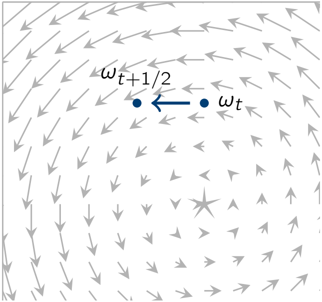


⇒ Bilinear games as limiting example of GANs [Mescheder et al., 2017]

Extragradient

Extragradient method [Korpelevich, 1976]:

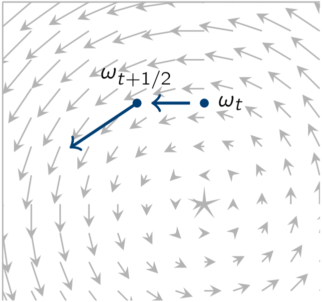
$$\omega^{t+1} = \omega^{t} - \eta v(\underbrace{\omega^{t} - \eta v(\omega^{t})}_{\omega_{t+1/2}})$$



Extragradient

Extragradient method [Korpelevich, 1976]:

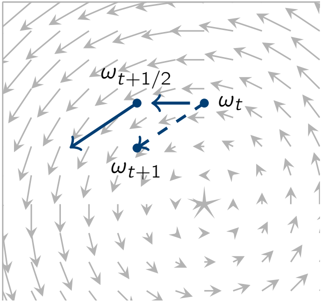
$$\omega^{t+1} = \omega^t - \eta v(\underbrace{\omega^t - \eta v(\omega^t)}_{\omega_{t+1/2}})$$



Extragradient

Extragradient method [Korpelevich, 1976]:

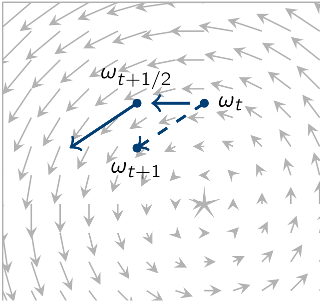
$$\omega^{t+1} = \omega^t - \eta v(\underbrace{\omega^t - \eta v(\omega^t)}_{\omega_{t+1/2}})$$



Extragradient

Extragradient method [Korpelevich, 1976]:

$$\omega^{t+1} = \omega^t - \eta v(\underbrace{\omega^t - \eta v(\omega^t)}_{\omega_{t+1/2}})$$



Theorem (See Mokhtari et al. [2019])

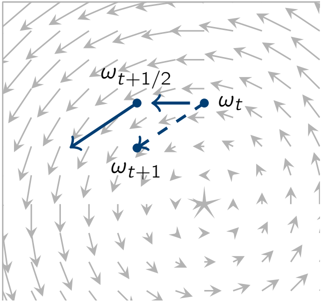
If v is μ-strongly monotone and L-Lipschitz,

$$\|\omega^t - \omega^*\|^2 \leq \left(1 - \frac{\mu}{4L}\right)^t \|\omega^0 - \omega^*\|^2$$

Extragradient

Extragradient method [Korpelevich, 1976]:

$$\omega^{t+1} = \omega^t - \eta v(\underbrace{\omega^t - \eta v(\omega^t)}_{\omega_{t+1/2}})$$



Lemma (Tseng [1995])

On the bilinear game,

$$\|\omega^t - \omega^*\|^2 \leq \left(1 - \frac{1}{2}\frac{\sigma_{min}(A)^2}{\sigma_{max}(A)^2}\right)^t \|\omega^0 - \omega^*\|^2$$

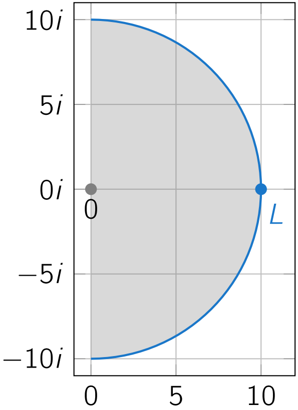
Unifying local analysis of extragradient

Theorem

If, $\forall \lambda \in \operatorname{Sp} \nabla v(x^*)$,

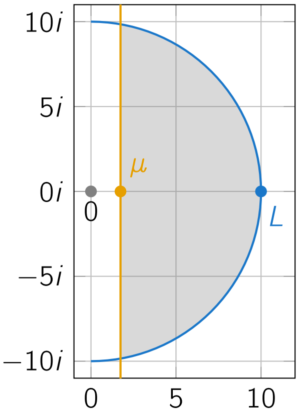
- $|\lambda| \leq L$

then,



$$\|x_t - x^*\| \lesssim \left(1 - \frac{1}{4}\left(\qquad \right)\right)^t \|x_0 - x^*\|$$

Unifying local analysis of extragradient



Theorem

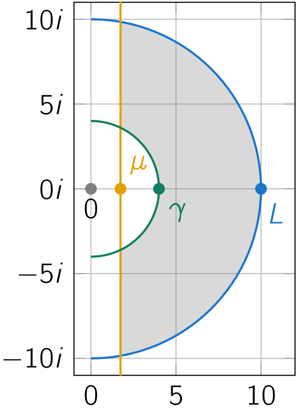
If, $\forall \lambda \in \operatorname{Sp} \nabla v(x^*)$,

- $|\lambda| \leq L$
- $\Re \lambda \geq \mu \geq 0$

then,

$$\|x_t - x^*\| \lesssim \left(1 - \frac{1}{4}\left(\frac{\mu}{L} \qquad\qquad \right)\right)^t \|x_0 - x^*\|$$

Unifying local analysis of extragradient



Theorem

If, $\forall \lambda \in \operatorname{Sp} \nabla v(x^*)$,

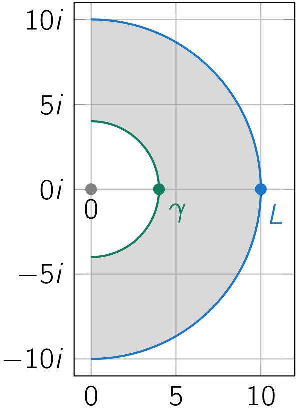
- $|\lambda| \leq L$
- $\Re\lambda \geq \mu \geq 0$
- $|\lambda| \geq \gamma \geq 0$

then,

$$\|x_t - x^*\| \lesssim \left(1 - \frac{1}{4}\left(\frac{\mu}{L} + \frac{1}{16}\frac{\gamma^2}{L^2}\right)\right)^t \|x_0 - x^*\|$$

$\Rightarrow$ recovers the standard rate with μ

Linear convergence without strong monotonicity



Corollary

If, $\forall \lambda \in \operatorname{Sp} \nabla v(x^*)$,

- $|\lambda| \leq L$
- $\Re\lambda \geq \mu = 0$
- $|\lambda| \geq \gamma \geq 0$

then,

$$\|x_t - x^*\| \lesssim \left(1 - \frac{1}{64}\frac{\gamma^2}{L^2}\right)^t \|x_0 - x^*\|$$

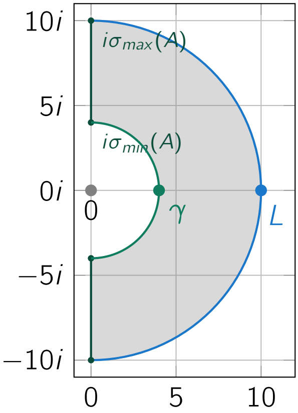
Linear convergence without strong monotonicity

Corollary

If, $\forall \lambda \in \operatorname{Sp} \nabla v(x^*)$,

- $|\lambda| \leq L$
- $\Re\lambda \geq \mu = 0$
- $|\lambda| \geq \gamma \geq 0$

then,



$$\|x_t - x^*\| \lesssim \left(1 - \frac{1}{64}\frac{\gamma^2}{L^2}\right)^t \|x_0 - x^*\|$$

$\Rightarrow$ recovers the bilinear case $\gamma = \sigma_{\min}(A)$.

Best of both worlds in between

For $\epsilon > 0$ small,

$$\min_{x \in \mathbb{R}} \max_{y \in \mathbb{R}} \frac{\epsilon}{2}(x^2 - y^2) + xy$$

Best of both worlds in between

For $\epsilon > 0$ small,

$$\min_{x \in \mathbb{R}} \max_{y \in \mathbb{R}} \frac{\epsilon}{2}(x^2 - y^2) + xy$$

- $\mu = \epsilon$

Best of both worlds in between

For $\epsilon > 0$ small,

$$\min_{x \in \mathbb{R}} \max_{y \in \mathbb{R}} \frac{\epsilon}{2}(x^2 - y^2) + xy$$

- $\mu = \epsilon$
- $\gamma = 1 - \epsilon$

Best of both worlds in between

For $\epsilon > 0$ small,

$$\min_{x \in \mathbb{R}} \max_{y \in \mathbb{R}} \frac{\epsilon}{2}(x^2 - y^2) + xy$$

- $\mu = \epsilon$
- $\gamma = 1 - \epsilon$
- $L = 1 + \epsilon$

Best of both worlds in between

For $\epsilon > 0$ small,

$$\min_{x \in \mathbb{R}} \max_{y \in \mathbb{R}} \frac{\epsilon}{2}(x^2 - y^2) + xy$$

- $\mu = \epsilon$
- $\gamma = 1 - \epsilon$
- $L = 1 + \epsilon$

$$\frac{\gamma^2}{L^2} \approx 1 - 2\epsilon \gg \frac{\mu}{L} \approx \epsilon$$

Global analysis

Local Assumptions.

- $|\lambda| \leq L$

Global Assumptions.

- L-Lipschitz,

$\forall \lambda \in \operatorname{Sp} \nabla v(x^*)$,

Global analysis

Local Assumptions.

- $|\lambda| \leq L$
- $\Re\lambda \geq \mu \geq 0$

$\forall \lambda \in \operatorname{Sp} \nabla v(x^*)$,

Global Assumptions.

- L-Lipschitz,
- μ-strongly monotone (or only monotone),

Global analysis

Local Assumptions.

- $|\lambda| \leq L$
- $\Re\lambda \geq \mu \geq 0$
- $|\lambda| \geq \gamma > 0$

$\forall \lambda \in \operatorname{Sp} \nabla v(x^*)$,

Global Assumptions.

- L-Lipschitz,
- μ-strongly monotone (or only monotone),
- Error bound of Tseng [1995] $\approx$

$$\sigma_{min}(\nabla v) \geq \gamma$$

Global analysis

Local Assumptions.

- $|\lambda| \leq L$
- $\Re\lambda \geq \mu \geq 0$
- $|\lambda| \geq \gamma > 0$

$\forall \lambda \in \mathrm{Sp}\, \nabla v(x^*)$,

Global Assumptions.

- L-Lipschitz,
- μ-strongly monotone (or only monotone),
- Error bound of Tseng [1995] $\approx$

$$\sigma_{min}(\nabla v) \geq \gamma$$

$\Rightarrow$ Global unifying guarantees for extragradient, optimistic gradient descent and consensus optimization ! (see paper for details)

Conclusion and perspectives

Takeaway: Local and global *unified* analysis of extragradient,

Conclusion and perspectives

Takeaway: Local and global *unified* analysis of extragradient,

- Unifies the results on cooperative games and the bilinear example.

Conclusion and perspectives

Takeaway: Local and global *unified* analysis of extragradient,

- Unifies the results on cooperative games and the bilinear example.
- Linear convergence rate for non-strongly monotone adversarial games.

Conclusion and perspectives

Takeaway: Local and global *unified* analysis of extragradient,

- Unifies the results on cooperative games and the bilinear example.
- Linear convergence rate for non-strongly monotone adversarial games.
- In between extrgradient enjoys the *best of both world*.

Conclusion and perspectives

Takeaway: Local and global *unified* analysis of extragradient,

- Unifies the results on cooperative games and the bilinear example.
- Linear convergence rate for non-strongly monotone adversarial games.
- In between extrgradient enjoys the *best of both world*.

Not discussed: Lower bounds, comparison with gradient descent, link to proximal method, consensus optimization and optimistic method... See the paper !

Conclusion and perspectives

Takeaway: Local and global *unified* analysis of extragradient,

- Unifies the results on cooperative games and the bilinear example.
- Linear convergence rate for non-strongly monotone adversarial games.
- In between extrgradient enjoys the *best of both world*.

Not discussed: Lower bounds, comparison with gradient descent, link to proximal method, consensus optimization and optimistic method... See the paper !

Perspectives: Now that we have convergence for a broad class of games, can we have faster convergence with the same unfiying properties ?

Bibliography I

D. Balduzzi, S. Racaniere, J. Martens, J. Foerster, K. Tuyls, and T. Graepel. The Mechanics of n-Player Differentiable Games. In *ICML*, 2018.

H. Berard, G. Gidel, A. Almahairi, P. Vincent, and S. Lacoste-Julien. A closer look at the optimization landscapes of generative adversarial networks. In *ICLR*, 2020.

D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. Training gans with optimism. In *ICLR*, 2018.

G. Gidel, H. Berard, P. Vincent, and S. Lacoste-Julien. A variational inequality perspective on generative adversarial nets. In *ICLR*, 2019a.

G. Gidel, R. A. Hemmat, M. Pezeshki, R. L. Priol, G. Huang, S. Lacoste-Julien, and I. Mitliagkas. Negative momentum for improved game dynamics. In *AISTATS*, 2019b.

Bibliography II

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In *NIPS*, 2014.

G. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 1976.

T. Liang and J. Stokes. Interaction Matters: A Note on Non-asymptotic Local Convergence of Generative Adversarial Networks. *arXiv:1802.06132 [cs, stat]*, Feb. 2018. arXiv: 1802.06132.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

L. Mescheder, S. Nowozin, and A. Geiger. The Numerics of GANs. In *NIPS*, 2017.

Bibliography III

A. Mokhtari, A. Ozdaglar, and S. Pattathil. A Unified Analysis of Extra-gradient and Optimistic Gradient Methods for Saddle Point Problems: Proximal Point Approach. *arXiv:1901.08511*, 2019.

R. T. Rockafellar. Monotone Operators and the Proximal Point Algorithm. *SIAM Journal on Control and Optimization*, 1976.

P. Tseng. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 1995.