Understanding the Role of Optimism in Minimax Optimization: A Proximal Point Approach

Aryan Mokhtari (UT Austin)

joint work with

Sarath Pattathil (MIT) & Asu Ozdaglar (MIT)

Workshop on Bridging Game Theory & Deep Learning
NeurIPS 2019
Vancouver, Canada, December 14, 2019

Introduction

- We focus on minimax optimization (known as saddle-point problem)

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

Introduction

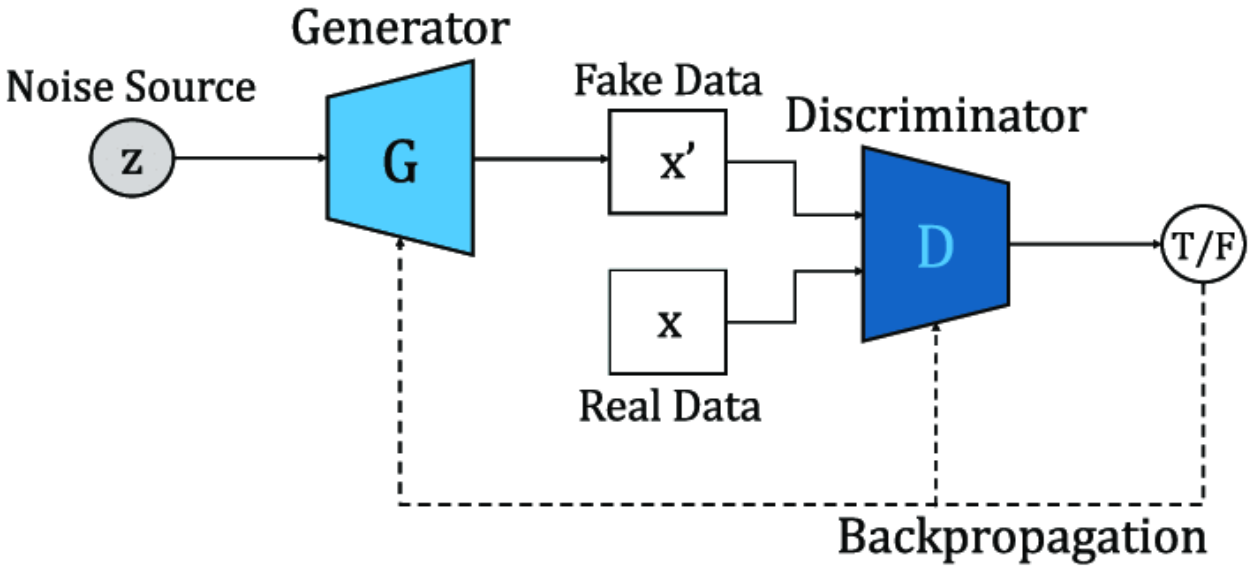
- We focus on minimax optimization (known as saddle-point problem)

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

- Minimax optimization has been studied for a very long time:
  - Game Theory: [Basar & Oldser, '99]
  - Control Theory: [Hast et al. '13]
  - Robust Optimization: [Ben-Tal et al., '09]

- Recently popularized to train GANs - [Goodfellow et al., '14].

Training GANs



- A game between the discriminator and generator
  - Can be formulated as a minimax optimization problem
  - Nonconvex-nonconcave since both discriminator and generator are neural networks.

Algorithms

- A simple algorithm to use is the Gradient Descent Ascent (GDA)

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k), \qquad \mathbf{y}_{k+1} = \mathbf{y}_k + \eta \nabla_{\mathbf{y}} f(\mathbf{x}_k, \mathbf{y}_k).$$

Algorithms

- A simple algorithm to use is the Gradient Descent Ascent (GDA)

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k), \qquad \mathbf{y}_{k+1} = \mathbf{y}_k + \eta \nabla_{\mathbf{y}} f(\mathbf{x}_k, \mathbf{y}_k).$$

- The Optimistic Gradient Descent Ascent (OGDA) method is

$$\begin{aligned}
\mathbf{x}_{k+1} &= \mathbf{x}_k - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k) - \eta(\nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k) - \nabla_{\mathbf{x}} f(\mathbf{x}_{k-1}, \mathbf{y}_{k-1})) \\
\mathbf{y}_{k+1} &= \mathbf{y}_k + \eta \nabla_{\mathbf{y}} f(\mathbf{x}_k, \mathbf{y}_k) + \eta(\nabla_{\mathbf{y}} f(\mathbf{x}_k, \mathbf{y}_k) - \nabla_{\mathbf{y}} f(\mathbf{x}_{k-1}, \mathbf{y}_{k-1}))
\end{aligned}$$

Algorithms

- A simple algorithm to use is the Gradient Descent Ascent (GDA)

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k), \qquad \mathbf{y}_{k+1} = \mathbf{y}_k + \eta \nabla_{\mathbf{y}} f(\mathbf{x}_k, \mathbf{y}_k).$$

- The Optimistic Gradient Descent Ascent (OGDA) method is

$$\begin{aligned}\mathbf{x}_{k+1} &= \mathbf{x}_k - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k) - \eta(\nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k) - \nabla_{\mathbf{x}} f(\mathbf{x}_{k-1}, \mathbf{y}_{k-1})) \\ \mathbf{y}_{k+1} &= \mathbf{y}_k + \eta \nabla_{\mathbf{y}} f(\mathbf{x}_k, \mathbf{y}_k) + \eta(\nabla_{\mathbf{y}} f(\mathbf{x}_k, \mathbf{y}_k) - \nabla_{\mathbf{y}} f(\mathbf{x}_{k-1}, \mathbf{y}_{k-1}))\end{aligned}$$

- It has been observed that "negative momentum" or "optimism" helps in training GANs. ([Gidel et al., '18] , [Daskalakis et al., '18])

Impact of Optimism

Consider the following images generated by GANs (from [Daskalakis et al. '18])

Figure: Adam

Figure: Optimistic Adam

- Adam - Similar and inferior quality images.
- Optimistic Adam - Diverse and higher quality images

History of Optimism

- Clear that optimism is helping in faster convergence.

History of Optimism

- Clear that optimism is helping in faster convergence.
- There are several results which study Optimistic Methods

History of Optimism

- Clear that optimism is helping in faster convergence.
- There are several results which study Optimistic Methods

- Introduced in [Popov '80] as a variant of Extra-Gradient (EG) method
- In the context of Online Learning in [Rakhlin et al. '13]
- Variant of Forward-Backward algorithm [Malitsky & Tam, '19]

History of Optimism

- Clear that optimism is helping in faster convergence.
- There are several results which study Optimistic Methods

- Introduced in [Popov '80] as a variant of Extra-Gradient (EG) method
- In the context of Online Learning in [Rakhlin et al. '13]
- Variant of Forward-Backward algorithm [Malitsky & Tam, '19]
- Several works analyzing OGDA in special settings [Daskalakis et al. '18], [Gidel et al. '19], [Liang & Stokes '19], [Mertikopoulos et al. '19]

History of Optimism

- Clear that optimism is helping in faster convergence.
- There are several results which study Optimistic Methods

- Introduced in [Popov '80] as a variant of Extra-Gradient (EG) method
- In the context of Online Learning in [Rakhlin et al. '13]
- Variant of Forward-Backward algorithm [Malitsky & Tam, '19]
- Several works analyzing OGDA in special settings [Daskalakis et al. '18], [Gidel et al. '19], [Liang & Stokes '19], [Mertikopoulos et al. '19]

- We show that optimism or negative momentum helps to approximate proximal point method more accurately!
    - OGDA can be considered as an approximation of Proximal Point

Simple Example

- Consider the following bilinear problem:

$$\min_{\mathbf{x}\in\mathbb{R}^d} \max_{\mathbf{y}\in\mathbb{R}^d} f(\mathbf{x},\mathbf{y}) = \mathbf{x}^\top \mathbf{y}$$

The solution is $(\mathbf{x}^*, \mathbf{y}^*) = (\mathbf{0}, \mathbf{0})$.

Simple Example

- Consider the following bilinear problem:

$$\min_{\mathbf{x}\in\mathbb{R}^d} \max_{\mathbf{y}\in\mathbb{R}^d} f(\mathbf{x},\mathbf{y}) = \mathbf{x}^\top \mathbf{y}$$

  The solution is $(\mathbf{x}^*, \mathbf{y}^*) = (\mathbf{0}, \mathbf{0})$.

- The Gradient Descent Ascent (GDA) updates for this problem:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \underbrace{\mathbf{y}_k}_{\nabla_{\mathbf{x}} f(\mathbf{x}_k,\mathbf{y}_k)}, \qquad \mathbf{y}_{k+1} = \mathbf{y}_k + \eta \underbrace{\mathbf{x}_k}_{\nabla_{\mathbf{y}} f(\mathbf{x}_k,\mathbf{y}_k)}$$

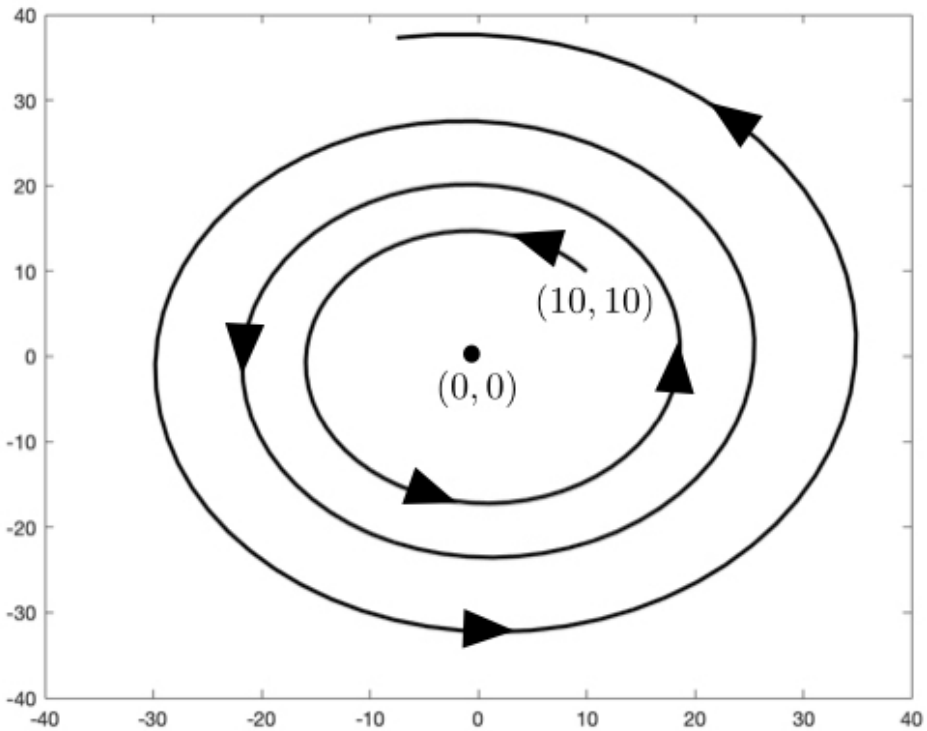
  where η is the stepsize.

GDA

- On running GDA, after k iterations we have:

$$\|\mathbf{x}_{k+1}\|^2 + \|\mathbf{y}_{k+1}\|^2 = (1+\eta^2)(\|\mathbf{x}_k\|^2 + \|\mathbf{y}_k\|^2)$$

- GDA diverges as $(1+\eta^2) > 1$

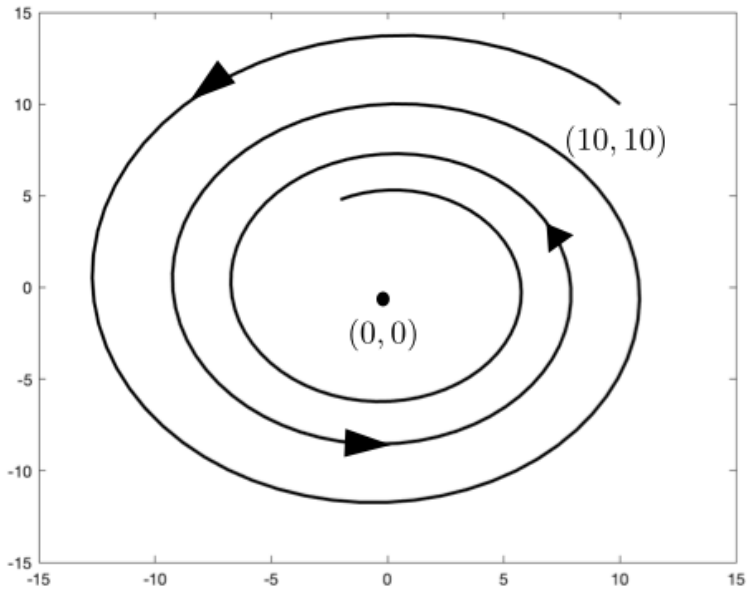


OGDA

- OGDA updates for the bilinear problem

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta\Big(\underbrace{2\mathbf{y}_k - \mathbf{y}_{k-1}}_{2\nabla_{\mathbf{x}} f(\mathbf{x}_k,\mathbf{y}_k) - \nabla_{\mathbf{x}} f(\mathbf{x}_{k-1},\mathbf{y}_{k-1})}\Big)$$

$$\mathbf{y}_{k+1} = \mathbf{y}_k + \eta\Big(\underbrace{2\mathbf{x}_k - \mathbf{x}_{k-1}}_{2\nabla_{\mathbf{y}} f(\mathbf{x}_k,\mathbf{y}_k) - \nabla_{\mathbf{y}} f(\mathbf{x}_{k-1},\mathbf{y}_{k-1})}\Big)$$



Proximal Point

- The Proximal Point (PP) updates for the same problem:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \underbrace{\mathbf{y}_{k+1}}_{\nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})} \qquad \mathbf{y}_{k+1} = \mathbf{y}_k + \eta \underbrace{\mathbf{x}_{k+1}}_{\nabla_{\mathbf{y}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})}$$

  where η is the stepsize.

Proximal Point

- The Proximal Point (PP) updates for the same problem:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \underbrace{\mathbf{y}_{k+1}}_{\nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})} \qquad \mathbf{y}_{k+1} = \mathbf{y}_k + \eta \underbrace{\mathbf{x}_{k+1}}_{\nabla_{\mathbf{y}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})}$$

  where η is the stepsize.

- The difference from GDA is that the gradient at the iterate $(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})$ is used for the update instead of the gradient at $(\mathbf{x}_k, \mathbf{y}_k)$.

Proximal Point

- The Proximal Point (PP) updates for the same problem:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \underbrace{\mathbf{y}_{k+1}}_{\nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})} \qquad \mathbf{y}_{k+1} = \mathbf{y}_k + \eta \underbrace{\mathbf{x}_{k+1}}_{\nabla_{\mathbf{y}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})}$$

  where η is the stepsize.

- The difference from GDA is that the gradient at the iterate $(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})$ is used for the update instead of the gradient at $(\mathbf{x}_k, \mathbf{y}_k)$.

- Although for this problem it takes a simple form

$$\mathbf{x}_{k+1} = \frac{1}{1+\eta^2}\left(\mathbf{x}_k - \eta \mathbf{y}_k\right), \qquad \mathbf{y}_{k+1} = \frac{1}{1+\eta^2}\left(\mathbf{y}_k + \eta \mathbf{x}_k\right)$$

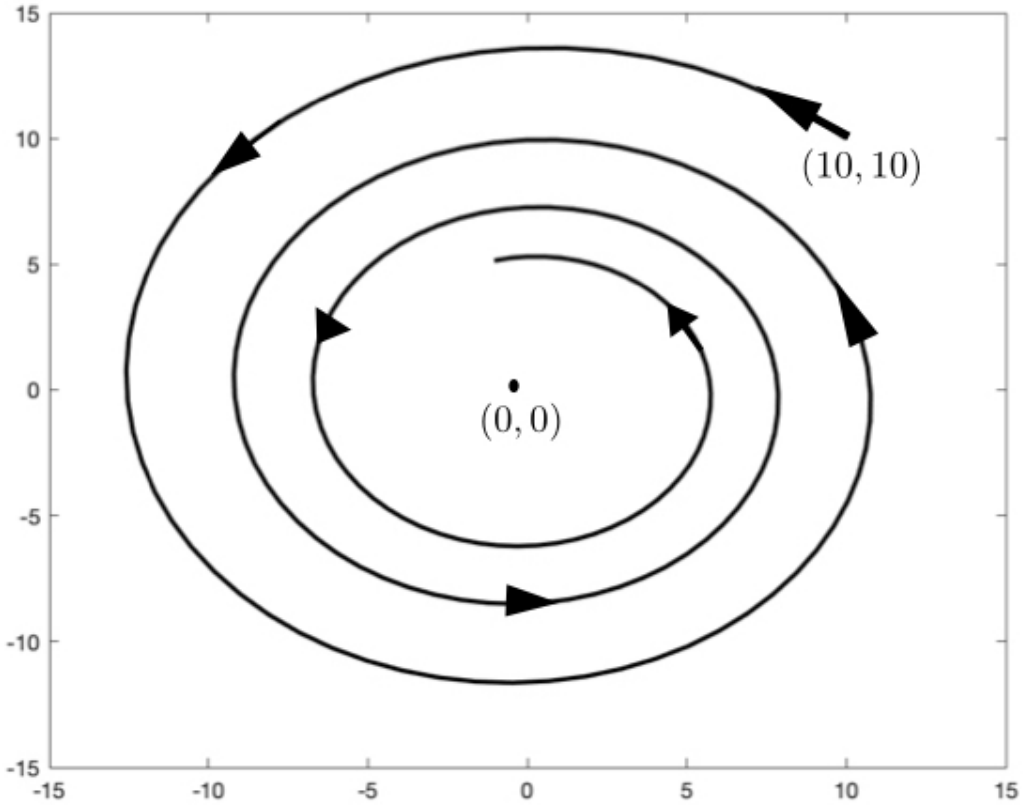
  Proximal Point method in general involves operator inversion and is not easy to implement.

Proximal Point

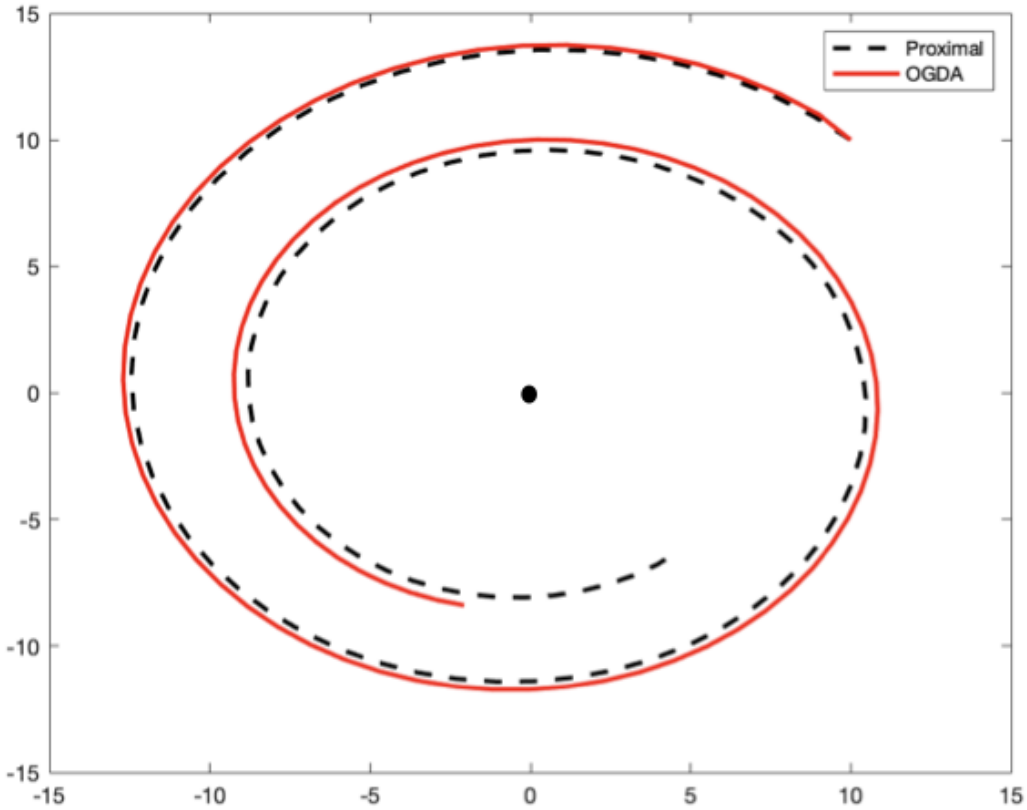
- On running PP, after k iterations we have:

$$\|\mathbf{x}_{k+1}\|^2 + \|\mathbf{y}_{k+1}\|^2 = \frac{1}{1+\eta^2}(\|\mathbf{x}_k\|^2 + \|\mathbf{y}_k\|^2)$$

- PP converges as $1/(1+\eta^2) < 1$



OGDA vs Proximal Point



- It seems like OGDA approximates Proximal Point method!
- Their convergence paths are very similar

Contributions & Outline

- View OGDA as approximations of PP for finding a saddle point

Contributions & Outline

- View OGDA as approximations of PP for finding a saddle point
- We show the iterates of OGDA are $o(\eta^2)$ approx. of PP iterates
- We focus on the convex-concave setting in this talk

Contributions & Outline

- View OGDA as approximations of PP for finding a saddle point
- We show the iterates of OGDA are $o(\eta^2)$ approx. of PP iterates
- We focus on the convex-concave setting in this talk
- We use the PP approximation viewpoint to show that for OGDA
  - The iterates are bounded
  - Function value converges at a rate of $\mathcal{O}(1/k)$

Contributions & Outline

- View OGDA as approximations of PP for finding a saddle point
- We show the iterates of OGDA are $o(\eta^2)$ approx. of PP iterates
- We focus on the convex-concave setting in this talk
- We use the PP approximation viewpoint to show that for OGDA
  - The iterates are bounded
  - Function value converges at a rate of $\mathcal{O}(1/k)$
- At the end, we also provide convergence rate results for
  - Bilinear problems
  - Strongly convex-strongly concave problems

Contributions & Outline

- View OGDA as approximations of PP for finding a saddle point
- We show the iterates of OGDA are $o(\eta^2)$ approx. of PP iterates
- We focus on the convex-concave setting in this talk
- We use the PP approximation viewpoint to show that for OGDA
  - The iterates are bounded
  - Function value converges at a rate of $\mathcal{O}(1/k)$
- At the end, we also provide convergence rate results for
  - Bilinear problems
  - Strongly convex-strongly concave problems
- We revisit the Extra-gradient (EG) method using the same approach

Problem

- We consider finding the saddle point of the problem:

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

Problem

- We consider finding the saddle point of the problem:

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

- f is convex in $\mathbf{x}$ and concave in $\mathbf{y}$.
- $f(\mathbf{x},\mathbf{y})$ is continuously differentiable in $\mathbf{x}$ and $\mathbf{y}$.
- $\nabla_{\mathbf{x}} f$ and $\nabla_{\mathbf{y}} f$ are Lipschitz in $\mathbf{x}$ and $\mathbf{y}$. L denotes the Lipschitz constant.

Problem

- We consider finding the saddle point of the problem:

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

- f is convex in $\mathbf{x}$ and concave in $\mathbf{y}$.
- $f(\mathbf{x},\mathbf{y})$ is continuously differentiable in $\mathbf{x}$ and $\mathbf{y}$.
- $\nabla_{\mathbf{x}} f$ and $\nabla_{\mathbf{y}} f$ are Lipschitz in $\mathbf{x}$ and $\mathbf{y}$. L denotes the Lipschitz constant.
- $(\mathbf{x}^*,\mathbf{y}^*) \in \mathbb{R}^m \times \mathbb{R}^n$ is a saddle point if it satisfies:

$$f(\mathbf{x}^*,\mathbf{y}) \leq f(\mathbf{x}^*,\mathbf{y}^*) \leq f(\mathbf{x},\mathbf{y}^*),$$

for all $\mathbf{x} \in \mathbb{R}^m, \mathbf{y} \in \mathbb{R}^n$.

Proximal Point

- The PP method at each step solves the following:

$$(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}) = \arg \min_{\mathbf{x}\in\mathbb{R}^m} \max_{\mathbf{y}\in\mathbb{R}^n} \left\{ f(\mathbf{x}, \mathbf{y}) + \frac{1}{2\eta}\|\mathbf{x} - \mathbf{x}_k\|^2 - \frac{1}{2\eta}\|\mathbf{y} - \mathbf{y}_k\|^2 \right\}.$$

Proximal Point

- The PP method at each step solves the following:

$$(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}) = \arg \min_{\mathbf{x}\in\mathbb{R}^m} \max_{\mathbf{y}\in\mathbb{R}^n} \left\{ f(\mathbf{x}, \mathbf{y}) + \frac{1}{2\eta}\|\mathbf{x} - \mathbf{x}_k\|^2 - \frac{1}{2\eta}\|\mathbf{y} - \mathbf{y}_k\|^2 \right\}.$$

- Using the first order optimality conditions leads to the following update:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}), \qquad \mathbf{y}_{k+1} = \mathbf{y}_k + \eta \nabla_{\mathbf{y}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}).$$

Proximal Point

- The PP method at each step solves the following:

$$(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}) = \arg \min_{\mathbf{x}\in\mathbb{R}^m} \max_{\mathbf{y}\in\mathbb{R}^n} \left\{ f(\mathbf{x}, \mathbf{y}) + \frac{1}{2\eta}\|\mathbf{x} - \mathbf{x}_k\|^2 - \frac{1}{2\eta}\|\mathbf{y} - \mathbf{y}_k\|^2 \right\}.$$

- Using the first order optimality conditions leads to the following update:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta\nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}), \qquad \mathbf{y}_{k+1} = \mathbf{y}_k + \eta\nabla_{\mathbf{y}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}).$$

Theorem (Convergence of Proximal Point)

The iterates generated by Proximal Point satisfy

$$\left[\max_{\mathbf{y}\in\mathcal{D}} f(\hat{\mathbf{x}}_k, \mathbf{y}) - f^\star\right] + \left[f^\star - \min_{\mathbf{x}\in\mathcal{D}} f(\mathbf{x}, \hat{\mathbf{y}}_k)\right] \leq \frac{\|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \|\mathbf{y}_0 - \mathbf{y}^*\|^2}{\eta k}.$$

where $\mathcal{D} := \{(\mathbf{x}, \mathbf{y}) \mid \|\mathbf{x} - \mathbf{x}^*\|^2 + \|\mathbf{y} - \mathbf{y}^*\|^2 \leq \|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \|\mathbf{y}_0 - \mathbf{y}^*\|^2\}$.

OGDA updates - How prediction takes place

- One way of approximating the Proximal Point update is as follows

$$\nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})) \approx \nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k) + (\nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k) - \nabla_{\mathbf{x}} f(\mathbf{x}_{k-1}, \mathbf{y}_{k-1}))$$
$$\nabla_{\mathbf{y}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})) \approx \nabla_{\mathbf{y}} f(\mathbf{x}_k, \mathbf{y}_k) + (\nabla_{\mathbf{y}} f(\mathbf{x}_k, \mathbf{y}_k) - \nabla_{\mathbf{y}} f(\mathbf{x}_{k-1}, \mathbf{y}_{k-1}))$$

OGDA updates - How prediction takes place

- One way of approximating the Proximal Point update is as follows

$$\nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})) \approx \nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k) + (\nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k) - \nabla_{\mathbf{x}} f(\mathbf{x}_{k-1}, \mathbf{y}_{k-1}))$$
$$\nabla_{\mathbf{y}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})) \approx \nabla_{\mathbf{y}} f(\mathbf{x}_k, \mathbf{y}_k) + (\nabla_{\mathbf{y}} f(\mathbf{x}_k, \mathbf{y}_k) - \nabla_{\mathbf{y}} f(\mathbf{x}_{k-1}, \mathbf{y}_{k-1}))$$

- This leads to the OGDA update

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k) - \eta(\nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k) - \nabla_{\mathbf{x}} f(\mathbf{x}_{k-1}, \mathbf{y}_{k-1}))$$
$$\mathbf{y}_{k+1} = \mathbf{y}_k + \eta \nabla_{\mathbf{y}} f(\mathbf{x}_k, \mathbf{y}_k) + \eta(\nabla_{\mathbf{y}} f(\mathbf{x}_k, \mathbf{y}_k) - \nabla_{\mathbf{y}} f(\mathbf{x}_{k-1}, \mathbf{y}_{k-1}))$$

OGDA vs PP

- Given a point $(\mathbf{x}_k, \mathbf{y}_k)$, let
  - $(\hat{\mathbf{x}}_{k+1}, \hat{\mathbf{y}}_{k+1})$ be the point obtained by performing PP on $(\mathbf{x}_k, \mathbf{y}_k)$.
  - $(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})$ be the point obtained by performing OGDA on $(\mathbf{x}_k, \mathbf{y}_k)$

OGDA vs PP

- Given a point $(\mathbf{x}_k, \mathbf{y}_k)$, let
  - $(\hat{\mathbf{x}}_{k+1}, \hat{\mathbf{y}}_{k+1})$ be the point obtained by performing PP on $(\mathbf{x}_k, \mathbf{y}_k)$.
  - $(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})$ be the point obtained by performing OGDA on $(\mathbf{x}_k, \mathbf{y}_k)$
- For a given stepsize $\eta > 0$ we have

$$\|\mathbf{x}_{k+1} - \hat{\mathbf{x}}_{k+1}\| \leq o(\eta^2), \qquad \|\mathbf{y}_{k+1} - \hat{\mathbf{y}}_{k+1}\| \leq o(\eta^2).$$

- Proof follows from Taylor's expansion of $\nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})$ and $\nabla_{\mathbf{y}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})$ at the point $(\mathbf{x}_k, \mathbf{y}_k)$.

Convergence rates

Theorem (Convex-Concave case)

Let the stepsize η satisfies the condition $0 < \eta \leq 1/2L$, then the iterates generated by OGDA satisfy

$$[\max_{\mathbf{y}\in\mathcal{D}} f(\hat{\mathbf{x}}_k, \mathbf{y}) - f^\star] + [f^\star - \min_{\mathbf{x}\in\mathcal{D}} f(\mathbf{x}, \hat{\mathbf{y}}_k)] \leq \frac{(\|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \|\mathbf{y}_0 - \mathbf{y}^*\|^2)(8L + \frac{1}{2\eta})}{k}$$

where $\mathcal{D} := \{(\mathbf{x}, \mathbf{y}) \mid \|\mathbf{x} - \mathbf{x}^*\|^2 + \|\mathbf{y} - \mathbf{y}^*\|^2 \leq 2\left(\|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \|\mathbf{y}_0 - \mathbf{y}^*\|^2\right)\}$.

- OGDA has an iteration complexity of $\mathcal{O}(1/k)$ (same as PP)
- First convergence guarantee for OGDA

Proximal Point with error

- We study 'inexact' versions of the Proximal Point method, given by

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}) + \varepsilon_k^x$$
$$\mathbf{y}_{k+1} = \mathbf{y}_k + \eta \nabla_{\mathbf{y}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}) - \varepsilon_k^y$$

Proximal Point with error

- We study 'inexact' versions of the Proximal Point method, given by

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}) + \varepsilon_k^x$$
$$\mathbf{y}_{k+1} = \mathbf{y}_k + \eta \nabla_{\mathbf{y}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}) - \varepsilon_k^y$$

- The 'error' $\varepsilon_k^x, \varepsilon_k^y$ should be such that
  - Algorithm should be easy to implement.
  - Retains convergence properties of PP.

Proximal Point with error

- We study 'inexact' versions of the Proximal Point method, given by

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}) + \varepsilon_k^x$$
$$\mathbf{y}_{k+1} = \mathbf{y}_k + \eta \nabla_{\mathbf{y}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}) - \varepsilon_k^y$$

- The 'error' $\varepsilon_k^x, \varepsilon_k^y$ should be such that
  - Algorithm should be easy to implement.
  - Retains convergence properties of PP.

Lemma (3 point equality for Proximal Point with error)

$$F(\mathbf{z}_{k+1})^\top (\mathbf{z}_{k+1} - \mathbf{z})$$
$$= \frac{1}{2\eta}\|\mathbf{z}_k - \mathbf{z}\|^2 - \frac{1}{2\eta}\|\mathbf{z}_{k+1} - \mathbf{z}\|^2 - \frac{1}{2\eta}\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 + \frac{1}{\eta}\varepsilon_k{}^T(\mathbf{z}_{k+1} - \mathbf{z}),$$

Here $\mathbf{z} = [\mathbf{x}; \mathbf{y}]$, $F(\mathbf{z}) = [\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}); -\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})]$ *and* $\varepsilon_k = [\varepsilon_k^x; -\varepsilon_k^y]$.

Convergence rates

Proof Sketch: (Here $\mathbf{z} = [\mathbf{x}; \mathbf{y}]$ and $F(\mathbf{z}) = [\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}); -\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})]$.)

- For OGDA, we have: $\varepsilon_k = \eta[F(\mathbf{z}_{k+1}) - 2F(\mathbf{z}_k) + F(\mathbf{z}_{k-1})]$.
- Substitute this error in the 3 point lemma for PP with error to get:

$$\sum_{k=0}^{N-1} F(\mathbf{z}_{k+1})^\top (\mathbf{z}_{k+1} - \mathbf{z}) \leq \frac{1}{2\eta}\|\mathbf{z}_0 - \mathbf{z}\|^2 - \frac{1}{2\eta}\|\mathbf{z}_N - \mathbf{z}\|^2 - \frac{L}{2}\|\mathbf{z}_N - \mathbf{z}_{N-1}\|^2 + (F(\mathbf{z}_N) - F(\mathbf{z}_{N-1}))^\top (\mathbf{z}_N - \mathbf{z}).$$

Convergence rates

Proof Sketch: (Here $\mathbf{z} = [\mathbf{x}; \mathbf{y}]$ and $F(\mathbf{z}) = [\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}); -\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})]$.)

- For OGDA, we have: $\varepsilon_k = \eta[F(\mathbf{z}_{k+1}) - 2F(\mathbf{z}_k) + F(\mathbf{z}_{k-1})]$.
- Substitute this error in the 3 point lemma for PP with error to get:

$$\sum_{k=0}^{N-1} F(\mathbf{z}_{k+1})^\top (\mathbf{z}_{k+1} - \mathbf{z}) \leq \frac{1}{2\eta} \|\mathbf{z}_0 - \mathbf{z}\|^2 - \left(\frac{1}{2\eta} - \frac{L}{2}\right) \|\mathbf{z}_N - \mathbf{z}\|^2$$

Convergence rates

Proof Sketch: (Here $\mathbf{z} = [\mathbf{x}; \mathbf{y}]$ and $F(\mathbf{z}) = [\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}); -\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})]$.)

- For OGDA, we have: $\varepsilon_k = \eta[F(\mathbf{z}_{k+1}) - 2F(\mathbf{z}_k) + F(\mathbf{z}_{k-1})]$.
- Substitute this error in the 3 point lemma for PP with error to get:

$$\sum_{k=0}^{N-1} F(\mathbf{z}_{k+1})^\top (\mathbf{z}_{k+1} - \mathbf{z}) \leq \frac{1}{2\eta} \|\mathbf{z}_0 - \mathbf{z}\|^2 - \left(\frac{1}{2\eta} - \frac{L}{2}\right) \|\mathbf{z}_N - \mathbf{z}\|^2$$

- $F(\mathbf{z})^\top (\mathbf{z} - \mathbf{z}^*) \geq 0$ for all $\mathbf{z}$ $\Rightarrow$ The iterates are bounded

$$\|\mathbf{z}_N - \mathbf{z}^*\|^2 \leq c \|\mathbf{z}_0 - \mathbf{z}^*\|^2$$

- Using convexity-concavity of $f(\mathbf{x}, \mathbf{y})$ we have:

$$\begin{aligned}
[\max_{\mathbf{y} \in \mathcal{D}} f(\hat{\mathbf{x}}_N, \mathbf{y}) - f^\star] + [f^\star - \min_{\mathbf{x} \in \mathcal{D}} f(\mathbf{x}, \hat{\mathbf{y}}_N)] &\leq \frac{1}{N} \sum_{k=0}^{N-1} F(\mathbf{z}_k)^\top (\mathbf{z}_k - \mathbf{z}) \\
&\leq \frac{C \|\mathbf{z}_0 - \mathbf{z}^*\|^2}{N}
\end{aligned}$$

Other Approximation of Proximal Point

- The Proximal Point updates are:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}), \qquad \mathbf{y}_{k+1} = \mathbf{y}_k + \eta \nabla_{\mathbf{y}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}).$$

Other Approximation of Proximal Point

- The Proximal Point updates are:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}), \qquad \mathbf{y}_{k+1} = \mathbf{y}_k + \eta \nabla_{\mathbf{y}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}).$$

- This can also be written as:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_k - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}), \mathbf{y}_k + \eta \nabla_{\mathbf{y}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})),$$
$$\mathbf{y}_{k+1} = \mathbf{y}_k + \eta \nabla_{\mathbf{y}} f(\mathbf{x}_k - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}), \mathbf{y}_k + \eta \nabla_{\mathbf{y}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})).$$

Other Approximation of Proximal Point

- The Proximal Point updates are:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}), \qquad \mathbf{y}_{k+1} = \mathbf{y}_k + \eta \nabla_{\mathbf{y}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}).$$

- This can also be written as:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_k - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}), \mathbf{y}_k + \eta \nabla_{\mathbf{y}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})),$$
$$\mathbf{y}_{k+1} = \mathbf{y}_k + \eta \nabla_{\mathbf{y}} f(\mathbf{x}_k - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}), \mathbf{y}_k + \eta \nabla_{\mathbf{y}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})).$$

- Approximate the inner gradient at $(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})$ by the gradient at $(\mathbf{x}_k, \mathbf{y}_k)$:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_k - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k), \mathbf{y}_k + \eta \nabla_{\mathbf{y}} f(\mathbf{x}_k, \mathbf{y}_k)),$$
$$\mathbf{y}_{k+1} = \mathbf{y}_k + \eta \nabla_{\mathbf{y}} f(\mathbf{x}_k - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k), \mathbf{y}_k + \eta \nabla_{\mathbf{y}} f(\mathbf{x}_k, \mathbf{y}_k)).$$

- which is nothing but the Extra-gradient Algorithm!

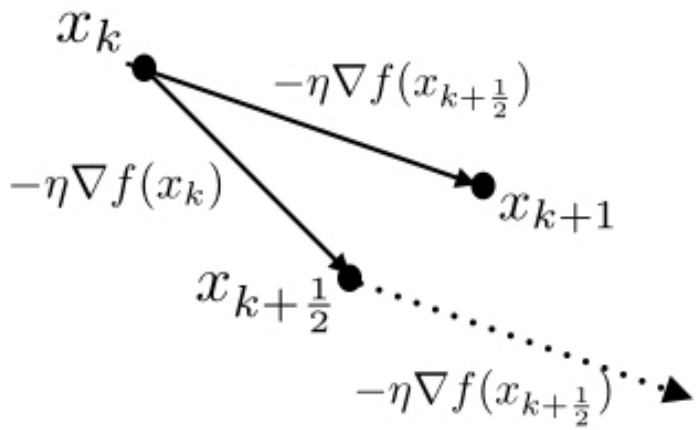
EG updates - How prediction takes place

- The updates of EG

$$\mathbf{x}_{k+1/2} = \mathbf{x}_k - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k), \qquad \mathbf{y}_{k+1/2} = \mathbf{y}_k + \eta \nabla_{\mathbf{y}} f(\mathbf{x}_k, \mathbf{y}_k).$$

The gradients evaluated at the midpoints $\mathbf{x}_{k+1/2}$ and $\mathbf{y}_{k+1/2}$ are used to compute the new iterates $\mathbf{x}_{k+1}$ and $\mathbf{y}_{k+1}$ by performing the updates

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_{k+1/2}, \mathbf{y}_{k+1/2}),$$
$$\mathbf{y}_{k+1} = \mathbf{y}_k + \eta \nabla_{\mathbf{y}} f(\mathbf{x}_{k+1/2}, \mathbf{y}_{k+1/2}).$$



EG vs PP

- Given a point $(\mathbf{x}_k, \mathbf{y}_k)$, let
  - $(\hat{\mathbf{x}}_{k+1}, \hat{\mathbf{y}}_{k+1})$ be the point we obtain by performing PP on $(\mathbf{x}_k, \mathbf{y}_k)$.
  - $(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})$ be the point we obtain by performing EG on $(\mathbf{x}_k, \mathbf{y}_k)$

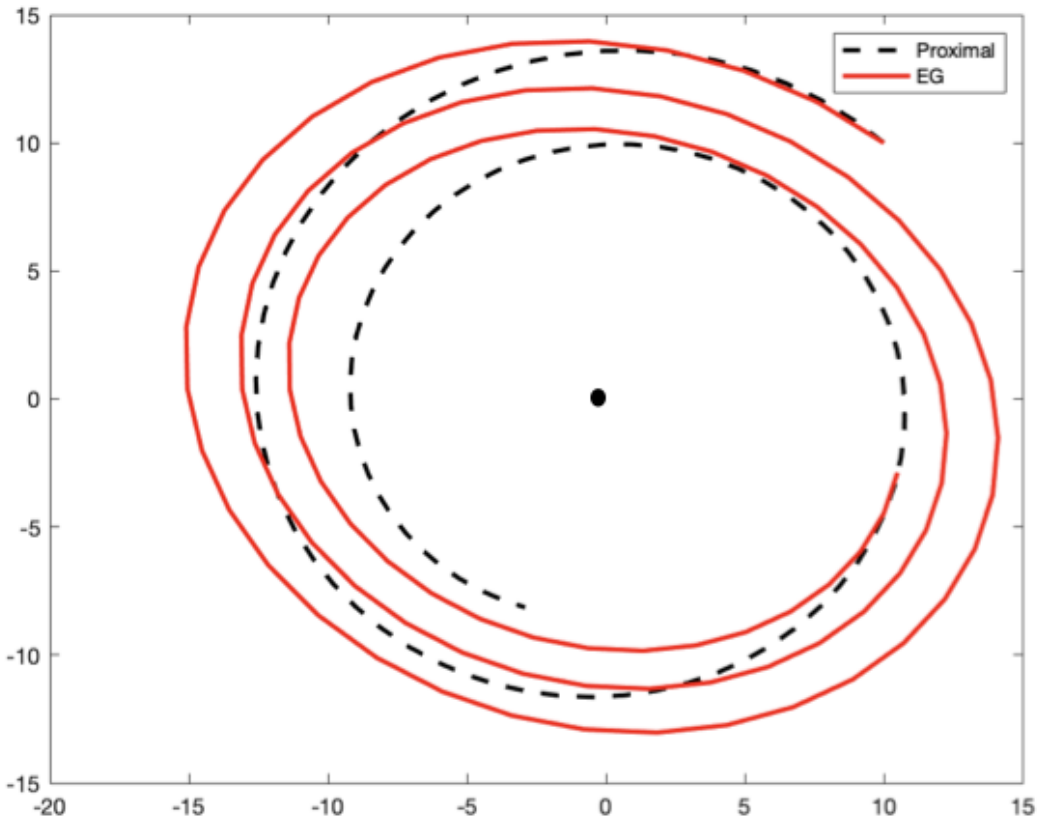
EG vs PP

- Given a point $(\mathbf{x}_k, \mathbf{y}_k)$, let
  - $(\hat{\mathbf{x}}_{k+1}, \hat{\mathbf{y}}_{k+1})$ be the point we obtain by performing PP on $(\mathbf{x}_k, \mathbf{y}_k)$.
  - $(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})$ be the point we obtain by performing EG on $(\mathbf{x}_k, \mathbf{y}_k)$
- Then, for a given stepsize $\eta > 0$ we have

$$\|\mathbf{x}_{k+1} - \hat{\mathbf{x}}_{k+1}\| \leq o(\eta^2), \qquad \|\mathbf{y}_{k+1} - \hat{\mathbf{y}}_{k+1}\| \leq o(\eta^2).$$

- Proof follows from Taylor's expansion of $\nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})$ and $\nabla_{\mathbf{y}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})$, as well as $\nabla_{\mathbf{x}} f(\mathbf{x}_{k+1/2}, \mathbf{y}_{k+1/2})$ and $\nabla_{\mathbf{y}} f(\mathbf{x}_{k+1/2}, \mathbf{y}_{k+1/2})$ about the point $(\mathbf{x}_k, \mathbf{y}_k)$.

EG vs PP



- EG also does a good job in approximating Proximal Point method!

Convergence rates

Theorem (Convex-Concave case)

For $\eta = \frac{\sigma}{L}$, where $\sigma \in (0,1)$, the iterates generated by the EG method satisfy

$$\left[\max_{\mathbf{y}\in\mathcal{D}} f(\hat{\mathbf{x}}_k, \mathbf{y}) - f^\star\right] + \left[f^\star - \min_{\mathbf{x}\in\mathcal{D}} f(\mathbf{x}, \hat{\mathbf{y}}_k)\right] \leq \frac{DL\left(16 + \frac{33}{2(1-\sigma^2)}\right)}{k}$$

where $D := \|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \|\mathbf{y}_0 - \mathbf{y}^*\|^2$ *and*
$\mathcal{D} := \{(\mathbf{x}, \mathbf{y}) \mid \|\mathbf{x} - \mathbf{x}^*\|^2 + \|\mathbf{y} - \mathbf{y}^*\|^2 \leq (2 + \frac{2}{1-\eta^2 L^2})D\}$.

Convergence rates

Theorem (Convex-Concave case)

For $\eta = \frac{\sigma}{L}$, where $\sigma \in (0,1)$, the iterates generated by the EG method satisfy

$$\left[\max_{\mathbf{y}\in\mathcal{D}} f(\hat{\mathbf{x}}_k, \mathbf{y}) - f^\star\right] + \left[f^\star - \min_{\mathbf{x}\in\mathcal{D}} f(\mathbf{x}, \hat{\mathbf{y}}_k)\right] \leq \frac{DL\left(16 + \frac{33}{2(1-\sigma^2)}\right)}{k}$$

where $D := \|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \|\mathbf{y}_0 - \mathbf{y}^*\|^2$ *and*
$\mathcal{D} := \{(\mathbf{x}, \mathbf{y}) \mid \|\mathbf{x} - \mathbf{x}^*\|^2 + \|\mathbf{y} - \mathbf{y}^*\|^2 \leq (2 + \frac{2}{1-\eta^2 L^2})D\}$.

- EG has an iteration complexity of $\mathcal{O}(1/k)$ (same as Proximal Point)

Convergence rates

Theorem (Convex-Concave case)

For $\eta = \frac{\sigma}{L}$, where $\sigma \in (0,1)$, the iterates generated by the EG method satisfy

$$\left[\max_{\mathbf{y}\in\mathcal{D}} f(\hat{\mathbf{x}}_k, \mathbf{y}) - f^\star\right] + \left[f^\star - \min_{\mathbf{x}\in\mathcal{D}} f(\mathbf{x}, \hat{\mathbf{y}}_k)\right] \leq \frac{DL\left(16 + \frac{33}{2(1-\sigma^2)}\right)}{k}$$

where $D := \|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \|\mathbf{y}_0 - \mathbf{y}^*\|^2$ *and*
$\mathcal{D} := \{(\mathbf{x}, \mathbf{y}) \mid \|\mathbf{x} - \mathbf{x}^*\|^2 + \|\mathbf{y} - \mathbf{y}^*\|^2 \leq (2 + \frac{2}{1-\eta^2 L^2})D\}$.

- EG has an iteration complexity of $\mathcal{O}(1/k)$ (same as Proximal Point)
- $\mathcal{O}(1/k)$ rate for the convex-concave case was shown in [Nemirovski '04] when the feasible set is compact.
- [Monteiro & Svaiter '10] extended to unbounded sets using a different termination criterion.

Convergence rates

Theorem (Convex-Concave case)

For $\eta = \frac{\sigma}{L}$, where $\sigma \in (0,1)$, the iterates generated by the EG method satisfy

$$\left[\max_{\mathbf{y}\in\mathcal{D}} f(\hat{\mathbf{x}}_k, \mathbf{y}) - f^\star\right] + \left[f^\star - \min_{\mathbf{x}\in\mathcal{D}} f(\mathbf{x}, \hat{\mathbf{y}}_k)\right] \leq \frac{DL\left(16 + \frac{33}{2(1-\sigma^2)}\right)}{k}$$

where $D := \|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \|\mathbf{y}_0 - \mathbf{y}^*\|^2$ *and*
$\mathcal{D} := \{(\mathbf{x}, \mathbf{y}) \mid \|\mathbf{x} - \mathbf{x}^*\|^2 + \|\mathbf{y} - \mathbf{y}^*\|^2 \leq (2 + \frac{2}{1-\eta^2 L^2})D\}$.

- EG has an iteration complexity of $\mathcal{O}(1/k)$ (same as Proximal Point)
- $\mathcal{O}(1/k)$ rate for the convex-concave case was shown in [Nemirovski '04] when the feasible set is compact.
- [Monteiro & Svaiter '10] extended to unbounded sets using a different termination criterion.
- Our result shows
  - a convergence rate of $\mathcal{O}(1/k)$ in terms of function value
  - without assuming compactness

Extensions

- **Bilinear:**
  - $f(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{B}\mathbf{y}$
  - $\mathbf{B} \in \mathbb{R}^{d \times d}$ is a square full-rank matrix.
  - $(\mathbf{x}^*, \mathbf{y}^*) = (\mathbf{0}, \mathbf{0})$ is the unique saddle point.
  - The condition number is defined as $\kappa := \frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})}$

Extensions

- **Bilinear:**
  - $f(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{B}\mathbf{y}$
  - $\mathbf{B} \in \mathbb{R}^{d \times d}$ is a square full-rank matrix.
  - $(\mathbf{x}^*, \mathbf{y}^*) = (\mathbf{0}, \mathbf{0})$ is the unique saddle point.
  - The condition number is defined as $\kappa := \frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})}$

- **Strongly Convex - Strongly Concave:**
  - $f(\mathbf{x}, \mathbf{y})$ is continuously differentiable in $\mathbf{x}$ and $\mathbf{y}$.
  - f is μ_x-strongly convex in $\mathbf{x}$ and μ_y-strongly concave in $\mathbf{y}$. We define $\mu := \min\{\mu_x, \mu_y\}$.
  - The condition number is defined as $\kappa := \frac{L}{\mu}$

Convergence rates

Theorem (Bilinear case (OGDA))

Let the stepsize $\eta = (1/40\sqrt{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})})$*, then the iterates generated by the OGDA method satisfy*

$$\|\mathbf{x}_{k+1}\|^2 + \|\mathbf{y}_{k+1}\|^2 \leq \left(1 - \frac{1}{c\kappa}\right)^k (\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2),$$

where c *is a positive constant independent of the problem parameters.*

Convergence rates

Theorem (Bilinear case (OGDA))

Let the stepsize $\eta = (1/40\sqrt{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})})$*, then the iterates generated by the OGDA method satisfy*

$$\|\mathbf{x}_{k+1}\|^2 + \|\mathbf{y}_{k+1}\|^2 \leq \left(1 - \frac{1}{c\kappa}\right)^k (\|\mathbf{x}_0\|^2 + \|\mathbf{y}_0\|^2),$$

where c *is a positive constant independent of the problem parameters.*

Theorem (Bilinear case (EG))

Let the stepsize $\eta = 1/(2\sqrt{2\lambda_{\max}(\mathbf{B}^\top \mathbf{B})})$*, then the iterates generated by the EG method satisfy*

$$\|\mathbf{x}_{k+1}\|^2 + \|\mathbf{y}_{k+1}\|^2 \leq \left(1 - \frac{1}{c\kappa}\right) (\|\mathbf{x}_k\|^2 + \|\mathbf{y}_k\|^2),$$

where c *is a positive constant independent of the problem parameters.*

Convergence rates

Theorem (Strongly convex-Strongly concave case (OGDA))

Let the stepsize $\eta = (1/(4L))$, then the iterates generated by OGDA satisfy

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 + \|\mathbf{y}_{k+1} - \mathbf{x}^*\|^2 \leq \left(1 - \frac{1}{c\kappa}\right)^k (\|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \|\mathbf{y}_0 - \mathbf{y}^*\|^2),$$

where c is a positive constant independent of the problem parameters.

Convergence rates

Theorem (Strongly convex-Strongly concave case (OGDA))

Let the stepsize $\eta = (1/(4L))$, then the iterates generated by OGDA satisfy

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 + \|\mathbf{y}_{k+1} - \mathbf{x}^*\|^2 \leq \left(1 - \frac{1}{c\kappa}\right)^k (\|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \|\mathbf{y}_0 - \mathbf{y}^*\|^2),$$

where c is a positive constant independent of the problem parameters.

Theorem (Strongly convex-Strongly concave case (EG))

Let the stepsize $\eta = 1/(4L)$, then the iterates generated by the EG method satisfy

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 + \|\mathbf{y}_{k+1} - \mathbf{x}^*\|^2 \leq \left(1 - \frac{1}{c\kappa}\right) (\|\mathbf{x}_k - \mathbf{x}^*\|^2 + \|\mathbf{y}_k - \mathbf{x}^*\|^2),$$

where c is a positive constant independent of the problem parameters.

Convergence rates

- OGDA and EG have an iteration complexity of $\mathcal{O}(\kappa \log(1/\epsilon))$
  - This is the same iteration complexity as the PP method.
  - Similar rates obtained in [Tseng 95], [Liang & Stokes 19], [Gidel et al. 19]

- Performance of GDA in these settings
  - **Bilinear:** GDA may not converge.
  - **Strongly convex-Strongly concave:** GDA has an iteration complexity of $\mathcal{O}(\kappa^2 \log(1/\epsilon))$

Generalized OGDA

- Using the interpretation that OGDA is an approximation of PP, we can extend this algorithm to a more general set of parameters.

Generalized OGDA

- Using the interpretation that OGDA is an approximation of PP, we can extend this algorithm to a more general set of parameters.
- The Generalized OGDA updates are:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k) - \beta \big(\nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k) - \nabla_{\mathbf{x}} f(\mathbf{x}_{k-1}, \mathbf{y}_{k-1}) \big)$$
$$\mathbf{y}_{k+1} = \mathbf{y}_k + \alpha \nabla_{\mathbf{y}} f(\mathbf{x}_k, \mathbf{y}_k) + \beta \big(\nabla_{\mathbf{y}} f(\mathbf{x}_k, \mathbf{y}_k) - \nabla_{\mathbf{y}} f(\mathbf{x}_{k-1}, \mathbf{y}_{k-1}) \big)$$

Generalized OGDA

- Using the interpretation that OGDA is an approximation of PP, we can extend this algorithm to a more general set of parameters.
- The Generalized OGDA updates are:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha\nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k) - \beta\big(\nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k) - \nabla_{\mathbf{x}} f(\mathbf{x}_{k-1}, \mathbf{y}_{k-1})\big)$$
$$\mathbf{y}_{k+1} = \mathbf{y}_k + \alpha\nabla_{\mathbf{y}} f(\mathbf{x}_k, \mathbf{y}_k) + \beta\big(\nabla_{\mathbf{y}} f(\mathbf{x}_k, \mathbf{y}_k) - \nabla_{\mathbf{y}} f(\mathbf{x}_{k-1}, \mathbf{y}_{k-1})\big)$$

- Note that
  - $\beta = 0$, this reduces to the GDA updates.
  - $\beta = \alpha$, this reduces to the standard OGDA updates.

Conclusions and Future Work

- Studied Optimism through the lens of Proximal Point method
- Analyzed OGDA as an inexact version of Proximal Point method
  - Convex-concave, bilinear, strongly convex-strongly concave
- Revisited EG as an approximation of the Proximal Point method
- This interpretation also aided in the design of Generalized OGDA

Conclusions and Future Work

- Studied Optimism through the lens of Proximal Point method
- Analyzed OGDA as an inexact version of Proximal Point method
  - Convex-concave, bilinear, strongly convex-strongly concave
- Revisited EG as an approximation of the Proximal Point method
- This interpretation also aided in the design of Generalized OGDA

- Extension to more general settings:
  - nonconvex-concave
  - weakly convex-weakly concave

Thanks!

- Convergence Rate of O(1/k) for Optimistic Gradient and Extra-gradient Methods in Smooth Convex-Concave Saddle Point Problems
  https://arxiv.org/pdf/1906.01115.pdf
- A Unified Analysis of Extra-gradient and Optimistic Gradient Methods for Saddle Point Problems: Proximal Point Approach
  https://arxiv.org/pdf/1901.08511.pdf