A Unified Analysis of Extra-gradient and Optimistic Gradient Methods for Saddle Point Problems: Proximal Point Approach

Aryan Mokhtari
UT Austin

Asuman Ozdaglar
MIT

Sarath Pattathil
MIT

Abstract

In this paper we consider solving saddle point problems using two variants of Gradient Descent-Ascent algorithms, Extra-gradient (EG) and Optimistic Gradient Descent Ascent (OGDA) methods. We show that both of these algorithms admit a unified analysis as approximations of the classical proximal point method for solving saddle point problems. This viewpoint enables us to develop a new framework for analyzing EG and OGDA for bilinear and strongly convex-strongly concave settings. Moreover, we use the proximal point approximation interpretation to generalize the results for OGDA for a wide range of parameters.

1 Introduction

In this paper, we study the following saddle point problem

$$\min_{\mathbf{x}\in\mathbb{R}^m} \max_{\mathbf{y}\in\mathbb{R}^n} f(\mathbf{x},\mathbf{y}), \tag{1}$$

where the function $f:\mathbb{R}^m\times\mathbb{R}^n\to\mathbb{R}$ is a convex-concave function, i.e., $f(\cdot,\mathbf{y})$ is convex for all $\mathbf{y}\in\mathbb{R}^n$ and $f(\mathbf{x},\cdot)$ is concave for all $\mathbf{x}\in\mathbb{R}^m$. We are interested in computing a saddle point of problem (1) defined as a pair $(\mathbf{x}^*,\mathbf{y}^*)\in\mathbb{R}^m\times\mathbb{R}^n$ that satisfies the condition

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

for all $\mathbf{x}\in\mathbb{R}^m, \mathbf{y}\in\mathbb{R}^n$. This problem formulation appears in several areas, including zero-sum games (Basar and Olsder, 1999), robust optimization (Ben-Tal et al., 2009), robust control (Hast et al., 2013) and more recently in machine learning in the context of Generative Adversarial Networks (GANs); see (Goodfellow et al., 2014) for an introduction to GANs and (Arjovsky et al., 2017) for the formulation of Wasserstein GANs.

Motivated by the interest in computational methods for solving the minmax problem in (1), in this paper we consider convergence rate analysis of discrete-time gradient based optimization algorithms for finding a saddle point of problem (1). We focus on Extra-gradient (EG) and Optimistic Gradient Descent Ascent (OGDA) methods, which have attracted much attention in the recent literature because of their superior empirical performance in GAN training (see (Liang and Stokes, 2019), (Daskalakis et al., 2017)). EG is a classical method which was introduced by Korpelevich (1976). Its linear rate of convergence for smooth and strongly convex-strongly concave functions $f(\mathbf{x},\mathbf{y})$ and bilinear functions, i.e., $f(\mathbf{x},\mathbf{y})=\mathbf{x}^\top\mathbf{A}\mathbf{y}$, was established in the variational inequality literature (see (Facchinei and Pang, 2007) and (Tseng, 1995)). The convergence properties of OGDA were recently studied in (Daskalakis et al., 2017), which showed the convergence of the iterates to a neighborhood of the solution when the objective function is bilinear. The recent paper (Liang and Stokes, 2019) used a dynamical system approach to prove the linear convergence of the OGDA and EG methods for the special case when $f(\mathbf{x},\mathbf{y})=\mathbf{x}^\top\mathbf{A}\mathbf{y}$ and the matrix $\mathbf{A}$ is square and full rank. It also presented a linear convergence rate of the vanilla Gradient Ascent Descent (GDA) method when the objective function $f(\mathbf{x},\mathbf{y})$ is strongly convex-strongly concave. In a recent paper (Gidel et al., 2019), a variant of the EG method is considered, relating it to OGDA updates, and show the linear convergence of the corresponding EG iterates in the case where $f(\mathbf{x},\mathbf{y})$ is strongly convex-strongly concave[1] (though without showing the convergence rate for the OGDA iterates).

The previous works use disparate approaches to analyze EG and OGDA methods, obtaining results in several different settings and making it difficult to see the connections and unifying principles between these iterative

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

[1] $f(\mathbf{x},\mathbf{y})$ is strongly convex-strongly concave when it is strongly convex with respect to $\mathbf{x}$ and strongly concave with respect to $\mathbf{y}$.

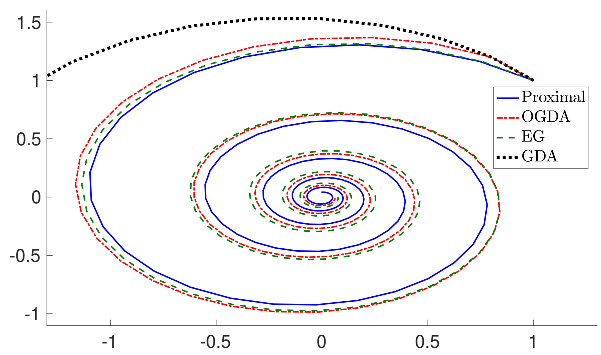

Figure 1: Convergence trajectories of proximal point (PP), extra-gradient (EG), optimistic gradient descent ascent (OGDA), and gradient descent ascent (GDA) for $\min_x \max_y xy$. The proximal point method has the fastest convergence. EG and OGDA approximate the trajectory of PP and both converge to the optimal solution. The GDA method is the only method that diverges.

methods. In this paper, we show that the update of EG and OGDA can be interpreted as approximations of the Proximal Point (PP) method, introduced in (Martinet, 1970) and studied in (Rockafellar, 1976b). This viewpoint allows us to understand why EG and OGDA are convergent for a bilinear problem. It also enables us to generalize OGDA (in terms of parameters) and obtain new convergence rate results for these generalized algorithms for the bilinear case. Our results recover the linear convergence rate results of Tseng (1995) for EG and the linear rate results of Liang and Stokes (2019) for the bilinear case of OGDA. We obtain new linear convergence rate estimates for OGDA for the strongly convex-strongly concave case as well as linear convergence rates for the generalized OGDA method.

Related Work. The result in (Tseng, 1995) showed convergence of EG method to an ϵ optimal solution with iteration complexity of $\mathcal{O}(\kappa \log(1/\epsilon))$ (see Assumption 1 and Remark 1 for the definition of κ) , when the function $f(\mathbf{x}, \mathbf{y})$ is smooth and strongly convex-strongly concave and when f is bilinear. A variational inequality perspective of saddle point problems was used in proving these results. More recently, Liang and Stokes (2019) analyzed EG and OGDA for the case when f is bilinear, using a dynamical system perspective. The authors showed a complexity of $\mathcal{O}(\kappa \log(1/\epsilon))$ for OGDA and a complexity of $\mathcal{O}(\kappa^2 \log(1/\epsilon))$ for EG, without anlyzing the general strongly convex-strongly concave setting. In another recent independent work, Gidel et al. (2019) analyzed the convergence of the OGDA method using the interpretation that OGDA is a variant of EG using extrapolation from the past. In this connection, the OGDA iterates are the "midpoints" whereas Gidel et al. (2019) provides a convergence of the original points (not the OGDA iterates) to an error of ϵ in $\mathcal{O}(\kappa \log(1/\epsilon))$. In this paper, we establish an overall complexity of $\mathcal{O}(\kappa \log(1/\epsilon))$ for both OGDA and EG in bilinear and strongly convex-strongly concave settings by interpreting these methods as approximations of the proximal point method. The results of our paper are compared with existing results in Table 1. Apart from the algorithms summarized in Table 1, we also propose a generalized version of OGDA which extends the classical OGDA algorithm to a wider range of stepsize parameters and show its convergence for bilinear case.

There are several papers that study the convergence rate of algorithms for solving saddle point problems *over a compact set*. Nemirovski (2004) showed $\mathcal{O}(1/k)$ convergence rate for the mirror-prox algorithm (a special case of which is the EG method) in convex-concave saddle point problems over compact sets. This result was extended to unbounded sets by Monteiro and Svaiter (2010) where a different error criterion was used. Nedić and Ozdaglar (2009) analyzed the (sub)Gradient Descent Ascent (GDA) algorithm for convex-concave saddle point problems when the (sub)gradients are bounded over the constraint set.

Several papers study the special case of Problem (1) when the objective function is of the form $f(\mathbf{x}, \mathbf{y}) = g(\mathbf{x}) + \mathbf{x}^\top \mathbf{A}\mathbf{y} - h(\mathbf{y})$, i.e., the cross term is bilinear. For this case, when the functions g and h are strongly convex, primal-dual gradient-type methods converge linearly (Chen and Rockafellar, 1997; Bauschke et al., 2011). Further, Du and Hu (2019) showed that GDA achieves a linear convergence rate when g is convex and h is strongly convex. Chambolle and Pock (2011) introduced a primal-dual variant of the proximal point method that converges to a saddle point at a sublinear rate when g and h are convex and at a linear rate when g and h are strongly convex.

For the case when $f(\mathbf{x}, \mathbf{y})$ is strongly concave with respect to $\mathbf{y}$, but possibly nonconvex with respect to $\mathbf{x}$, Sanjabi et al. (2018a) provided convergence to a first-order stationary point using an algorithm that requires running multiple updates with respect to $\mathbf{y}$ at each step. Recently, Sanjabi et al. (2018b) extended this result to the setting when f is Polyak-Lojasiewicz with respect to $\mathbf{y}$.

There are several papers which solve stochastic version of Problem (1), i.e., the case where one does not have access to the exact gradients of the function, but in fact an unbiased estimate of it. Papers including (Nemirovski et al., 2009; Juditsky et al., 2011; Chen et al., 2014) solve this problem in the case where the objec-

Reference	Assumptions on $f(\mathbf{x}, \mathbf{y})$	Rate (EG)	Rate (OGDA)
Liang and Stokes (2019)	Bilinear	$\mathcal{O}(\kappa^2 \log(1/\epsilon))$	$\mathcal{O}(\kappa \log(1/\epsilon))$
Liang and Stokes (2019)	Strongly Convex-Strongly Concave	✗	✗
Gidel et al. (2019)	Bilinear	✗	✗
Gidel et al. (2019)	Strongly Convex-Strongly Concave	✗	$\mathcal{O}(\kappa \log(1/\epsilon))^{\star}$
Tseng (1995)	Bilinear	$\mathcal{O}(\kappa \log(1/\epsilon))$	✗
Tseng (1995)	Strongly Convex-Strongly Concave	$\mathcal{O}(\kappa \log(1/\epsilon))$	✗
This paper	Bilinear	$\mathcal{O}(\kappa \log(1/\epsilon))$	$\mathcal{O}(\kappa \log(1/\epsilon))$
This paper	Strongly Convex-Strongly Concave	$\mathcal{O}(\kappa \log(1/\epsilon))$	$\mathcal{O}(\kappa \log(1/\epsilon))$

Table 1: Comparison of rates in different papers
($\star$- Gidel et al. (2019) shows the convergence of the half points and not the original OGDA iterates.)

tive function is convex in $\mathbf{x}$ and concave in $\mathbf{y}$. More recently Palaniappan and Bach (2016) uses a variance reduced version of the proximal gradient method and Chavdarova et al. (2019) uses a variance reduced version of the EG method to solve Problem (1) when the function is strongly convex in $\mathbf{x}$ and strongly concave in $\mathbf{y}$ and the function has a finite sum structure.

Optimistic gradient methods have also been studied in the context of convex online learning. In particular, Rakhlin and Sridharan (2013a,b) introduced the general version of the Optimistic Mirror Descent algorithm in the framework of online optimization. Prior to this work, a special case of Optimistic Mirror descent was analyzed by Chiang et al. (2012), again in the context of online learning.

Outline. The rest of the paper is organized as follows. We start the paper by presenting some definitions and preliminaries required for presenting our results in Section 2. Then, we revisit the Proximal Point (PP) point method in Section 3 and present its convergence properties for bilinear (Theorem 1) and general strongly convex-strongly concave (Theorem 2) problems. In Section 4, we show that the Optimistic Gradient Descent Ascent (OGDA) is an approximation of PP (Proposition 1) and prove its linear convergence rate for bilinear (Theorem 3) and strongly convex-strongly concave (Theorem 4) problems. We generalize the OGDA method in terms of its parameters and show the convergence of the generalized OGDA method for the bilinear case (Theorem 5). In Section 5, we recap the update of Extra-gradient (EG) method for solving a saddle point problem. Then, we show that EG can be interpreted as an approximation of PP (Proposition 2) and use this interpretation to study the convergence properties of EG in bilinear problems (Theorem 6) and general strongly convex-strongly concave problems (Theorem 7). In Section 6, we present our numerical results, comparing the performance of PP, EG, and OGDA for solving both a bilinear problem and a quadratic program. We close the paper with concluding remarks.

Notation. Lowercase boldface $\mathbf{v}$ denotes a vector and uppercase boldface $\mathbf{A}$ denotes a matrix. We use $\|\mathbf{v}\|$ to denote the Euclidean norm of vector $\mathbf{v}$. Given a multi-input function $f(\mathbf{x}, \mathbf{y})$, its gradient with respect to $\mathbf{x}$ and $\mathbf{y}$ at $(\mathbf{x}_0, \mathbf{y}_0)$ are denoted by $\nabla_{\mathbf{x}} f(\mathbf{x}_0, \mathbf{y}_0)$ and $\nabla_{\mathbf{y}} f(\mathbf{x}_0, \mathbf{y}_0)$, respectively. We refer to the largest and smallest eigenvalues of a matrix $\mathbf{A}$ by $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$, respectively.

2 Preliminaries

In this section we present properties and notations used in our results.

Definition 1. *A function $\phi : \mathbb{R}^n \to \mathbb{R}$ is L-smooth if it has L-Lipschitz continuous gradients on $\mathbb{R}^n$, i.e., for any $\mathbf{x}, \hat{\mathbf{x}} \in \mathbb{R}^n$, we have $||\nabla\phi(\mathbf{x}) - \nabla\phi(\hat{\mathbf{x}})|| \le L||\mathbf{x} - \hat{\mathbf{x}}||$.*

Definition 2. *A continuously differentiable function $\phi : \mathbb{R}^n \to \mathbb{R}$ is μ-strongly convex on $\mathbb{R}^n$ if for any $\mathbf{x}, \hat{\mathbf{x}} \in \mathbb{R}^n$, we have $\phi(\hat{\mathbf{x}}) \ge \phi(\mathbf{x}) + \nabla\phi(\mathbf{x})^T(\hat{\mathbf{x}} - \mathbf{x}) + \frac{\mu}{2}||\hat{\mathbf{x}} - \mathbf{x}||^2$. Further, $\phi(\mathbf{x})$ is μ-strongly concave if $-\phi(\mathbf{x})$ is μ-strongly convex. If we set $\mu = 0$, then we recover the definition of convexity for a continuous differentiable function.*

Definition 3. *The pair $(\mathbf{x}^*, \mathbf{y}^*)$ is a saddle point of a convex-concave function $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, if for any $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$, we have $f(\mathbf{x}^*, \mathbf{y}) \le f(\mathbf{x}^*, \mathbf{y}^*) \le f(\mathbf{x}, \mathbf{y}^*)$.*

Throughout the paper, we consider two specific cases of Problem (1) stated in the next set of assumptions.

Assumption 1. *The function $f(\mathbf{x}, \mathbf{y})$ is a bilinear function of the form $f(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{B} \mathbf{y}$, where $\mathbf{B} \in \mathbb{R}^{d \times d}$ is a square full-rank matrix. The point $(\mathbf{x}^*, \mathbf{y}^*) = (\mathbf{0}, \mathbf{0})$ is the unique saddle point. In this case, we define the condition number of the problem as $\kappa := \frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\lambda_{\min}(\mathbf{B}^\top\mathbf{B})}$.*

Assumption 2. *The function $f(\mathbf{x}, \mathbf{y})$ is continuously differentiable in $\mathbf{x}$ and $\mathbf{y}$. Further, f is μ_x-strongly convex in $\mathbf{x}$ and μ_y-strongly concave in $\mathbf{y}$. The unique saddle point of $f(\mathbf{x}, \mathbf{y})$ is denoted by $(\mathbf{x}^*, \mathbf{y}^*)$. We define $\mu = \min\{\mu_x, \mu_y\}$.*

Assumption 3. *The gradient $\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y})$, is L_x-Lipschitz in $\mathbf{x}$ and L_{xy}-Lipschitz in $\mathbf{y}$, i.e.,*

$$\|\nabla_{\mathbf{x}} f(\mathbf{x}_1,\mathbf{y}) - \nabla_{\mathbf{x}} f(\mathbf{x}_2,\mathbf{y})\| \leq L_x \|\mathbf{x}_1 - \mathbf{x}_2\| \quad \text{for all } \mathbf{y},$$
$$\|\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y}_1) - \nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y}_2)\| \leq L_{xy} \|\mathbf{y}_1 - \mathbf{y}_2\| \quad \text{for all } \mathbf{x}.$$

Moreover, the gradient $\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y})$, is L_y-Lipschitz in $\mathbf{y}$ and L_{yx}-Lipschitz in $\mathbf{x}$, i.e.,

$$\|\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y}_1) - \nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y}_2)\| \leq L_y \|\mathbf{y}_1 - \mathbf{y}_2\| \quad \text{for all } \mathbf{x},$$
$$\|\nabla_{\mathbf{y}} f(\mathbf{x}_1,\mathbf{y}) - \nabla_{\mathbf{x}} f(\mathbf{x}_2,\mathbf{y})\| \leq L_{yx} \|\mathbf{x}_1 - \mathbf{x}_2\| \quad \text{for all } \mathbf{y}.$$

We define $L = \max\{L_x, L_{xy}, L_y, L_{yx}\}$.

Remark 1. *Under Assumptions 2 and 3, we define the condition number of the problem as $\kappa := L/\mu$.*

In the following sections, we present and analyze three different iterative algorithms for solving the saddle point problem introduced in (1). The k-th iterates of any of these algorithms are denoted by $(\mathbf{x}_k, \mathbf{y}_k)$. We denote

$$r_k = \|\mathbf{x}_k - \mathbf{x}^*\|^2 + \|\mathbf{y}_k - \mathbf{y}^*\|^2, \tag{2}$$

as the distance to the saddle point $(\mathbf{x}^*, \mathbf{y}^*)$ at iteration k.

3 Proximal Point method

We start our analysis by Proximal Point (PP) method, which will serve as a benchmark for the analysis of Extra-gradient and Optimistic Gradient Descent Ascent methods. The update of PP method for minimizing a convex function h is defined as

$$\mathbf{x}_{k+1} = \text{prox}_{\frac{1}{\eta},h}(\mathbf{x}_k) = \text{argmin}\left\{h(\mathbf{x}) + \frac{1}{2\eta}\|\mathbf{x} - \mathbf{x}_k\|^2\right\}, \tag{3}$$

where η is a positive scalar (Bertsekas, 1999; Beck, 2017). Using the optimality condition of the update in (3), one can also write the update of the PP method as $\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla h(\mathbf{x}_{k+1})$. This expression shows that the PP method is an *implicit algorithm*. Convergence properties of PP for convex minimization have been extensively studied (Rockafellar, 1976a; Güler, 1991; Ferris, 1991; Eckstein and Bertsekas, 1992; Parikh et al., 2014; Beck, 2017). The extension of PP for solving saddle point problems has been also studied in Rockafellar (1976b). Here, we recap the update of PP for solving the min-max problem in (1). To do so, we define the iterates $\{\mathbf{x}_{k+1}, \mathbf{y}_{k+1}\}$ as the unique solution to the saddle point problem

$$\min_{\mathbf{x}\in\mathbb{R}^m} \max_{\mathbf{y}\in\mathbb{R}^n} \left\{f(\mathbf{x},\mathbf{y}) + \frac{1}{2\eta}\|\mathbf{x} - \mathbf{x}_k\|^2 - \frac{1}{2\eta}\|\mathbf{y} - \mathbf{y}_k\|^2\right\}. \tag{4}$$

Using the optimality conditions of (4) (which are necessary and sufficient since the problem in (4) is convex), the update of the PP method for the saddle point problem in (1) can be written as

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}),$$
$$\mathbf{y}_{k+1} = \mathbf{y}_k + \eta \nabla_{\mathbf{y}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}). \tag{5}$$

Note that implementing the system of updates in (5) requires computing the operators $(\mathbf{I} + \eta \nabla_{\mathbf{x}} f)^{-1}$ and $(\mathbf{I} + \eta \nabla_{\mathbf{y}} f)^{-1}$, and, therefore, may not be computationally affordable for any general function f.

In the following theorem, we show that the PP method converges linearly to $(\mathbf{x}^*, \mathbf{y}^*) = (\mathbf{0}, \mathbf{0})$ which is the unique solution of the problem $\min_{\mathbf{x}} \max_{\mathbf{y}} \mathbf{x}^\top \mathbf{B} \mathbf{y}$. This result was established in Theorem 2 of (Rockafellar, 1976b) and we mention it here for completeness and we later use it as a benchmark.

Theorem 1. *Consider the saddle point problem in (1) under Assumption 1 and the proximal point method in (5). Further, recall the definition of r_k in (2). Then, for any $\eta > 0$, the iterates $\{\mathbf{x}_k, \mathbf{y}_k\}_{k\geq 0}$ generated by the proximal point method satisfy*

$$r_{k+1} \leq \frac{1}{1+\eta^2 \lambda_{min}(\mathbf{B}^\top \mathbf{B})} r_k.$$

In the following theorem, we characterize the convergence rate of PP for a function $f(\mathbf{x},\mathbf{y})$ that is strongly convex with respect to $\mathbf{x}$ and strongly concave with respect to $\mathbf{y}$. Once again, this result was established in (Rockafellar, 1976b) and we mention it here for completeness and we later use it as a benchmark.

Theorem 2. *Consider the saddle point problem in (1) under Assumption 2 and the proximal point method in (5). Further, recall the definition of r_k in (2). Then, for any $\eta > 0$, the iterates $\{\mathbf{x}_k, \mathbf{y}_k\}_{k\geq 0}$ generated by the proximal point method satisfy*

$$r_{k+1} \leq \frac{1}{1+\eta\mu} r_k.$$

Theorem 2 states that for the general saddle point problem in (1), if the function is strongly convex-strongly concave, the iterates generated by the PP method converge linearly to the optimal solution.

4 Optimistic Gradient Descent Ascent method

In this section, we study the Optimistic Gradient Descent Ascent (OGDA) method for solving saddle point problems. We first show that OGDA can be considered as an approximation of the proximal point method. Then, we use this interpretation to analyze its convergence properties for bilinear and strongly convex-

Algorithm 1 OGDA method for saddle point problems

Require: $\eta > 0$, vectors $\mathbf{x}_{-1}, \mathbf{y}_{-1}, \mathbf{x}_0, \mathbf{y}_0 \in \mathbb{R}^d$

for $k = 1, 2, \dots$ **do**

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

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

end for

strongly concave settings. The proximal point approximation approach also allows us to generalize the update of OGDA as we discuss in detail in Section 4.2.

4.1 Convergence rate of the OGDA Method

The main idea behind the updates of the OGDA method is the addition of a “negative-momentum" term to the updates which can be clearly seen when we write the iterations as follows:

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

The last term in parenthesis for each of the updates can be interpreted as a “negative-momentum", differentiating the OGDA method from vanilla Gradient Descent Ascent (GDA). It can be seen that OGDA is approximating the proximal point update direction using linear extrapolation of the previous gradients, i.e., $\nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})) \approx \nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k) + (\nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k) - \nabla_{\mathbf{x}} f(\mathbf{x}_{k-1}, \mathbf{y}_{k-1}))$ and similarly for $\nabla_{\mathbf{y}} f(\cdot, \cdot)$.

We analyze the OGDA method as an approximation of the Proximal Point (PP) method presented in Section 3. We first focus on the bilinear case (Assumption 1) for which the OGDA updates are

$$\begin{aligned}
\mathbf{x}_{k+1} &= \mathbf{x}_k - 2\eta \mathbf{B}\mathbf{y}_k + \eta \mathbf{B}\mathbf{y}_{k-1}, \\
\mathbf{y}_{k+1} &= \mathbf{x}_k + 2\eta \mathbf{B}^\top \mathbf{x}_k + \eta \mathbf{B}^\top \mathbf{x}_{k-1}.
\end{aligned}$$

Note that the update of the PP method for the variable $\mathbf{x}$ in the considered bilinear problem is

$$\begin{aligned}
\mathbf{x}_{k+1} &= (I + \eta^2 \mathbf{B}\mathbf{B}^\top)^{-1}(\mathbf{x}_k - \eta \mathbf{B}\mathbf{y}_k) \\
&= (I - \eta^2 \mathbf{B}\mathbf{B}^\top + o(\eta^2))(\mathbf{x}_k - \eta \mathbf{B}\mathbf{y}_k) \\
&= \mathbf{x}_k - \eta \mathbf{B}\mathbf{y}_k - \eta \mathbf{B}(\eta \mathbf{B}^\top \mathbf{x}_k - \eta^2 \mathbf{B}^\top \mathbf{B}\mathbf{y}_k) + o(\eta^2),
\end{aligned}$$

where we used the fact that $I - \eta^2 \mathbf{B}\mathbf{B}^\top$ is an approximation of $(I + \eta^2 \mathbf{B}\mathbf{B}^\top)^{-1}$ with an error of $o(\eta^2)$. Regrouping the terms and using the updates of the PP method yield

$$\begin{aligned}
\mathbf{x}_{k+1} &= \mathbf{x}_k - 2\eta \mathbf{B}\mathbf{y}_k \\
&\qquad - \eta \mathbf{B}(\eta \mathbf{B}^\top \mathbf{x}_k - (1 + \eta^2 \mathbf{B}^\top \mathbf{B})\mathbf{y}_k) + o(\eta^2) \\
&= \mathbf{x}_k - 2\eta \mathbf{B}\mathbf{y}_k \\
&\qquad - \eta \mathbf{B}(\eta \mathbf{B}^\top \mathbf{x}_k - \mathbf{y}_{k-1} - \eta \mathbf{B}^\top \mathbf{x}_{k-1}) + o(\eta^2) \\
&= \mathbf{x}_k - 2\eta \mathbf{B}\mathbf{y}_k + \eta \mathbf{B}\mathbf{y}_{k-1} + o(\eta^2),
\end{aligned}$$

where the last expression is the OGDA update for variable $\mathbf{x}$ plus an additional error of $o(\eta^2)$. A similar derivation can be done for the update of variable $\mathbf{y}$ to show that OGDA is an approximation of the PP method up to $o(\eta^2)$. In the following proposition, we show that this observation can be generalized for any general smooth (possibly nonconvex) function $f(\mathbf{x}, \mathbf{y})$.

Proposition 1. *Consider the saddle point problem in* (1). *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 the PP update on* $(\mathbf{x}_k, \mathbf{y}_k)$, *and let* $(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})$ *be the point we obtain by performing the OGDA update on* $(\mathbf{x}_k, \mathbf{y}_k)$. *Then, for a given stepsize* $\eta > 0$ *we have*

$$\begin{aligned}
\|\mathbf{x}_{k+1} - \hat{\mathbf{x}}_{k+1}\| &\leq o(\eta^2), \\
\|\mathbf{y}_{k+1} - \hat{\mathbf{y}}_{k+1}\| &\leq o(\eta^2). \qquad (6)
\end{aligned}$$

To analyze the convergence of OGDA, we view it as a proximal point algorithm with an additional error term. In the following theorem, we characterize the convergence rate of the OGDA method for the bilinear saddle point problem defined in Assumption 1.

Theorem 3 (Bilinear case)**.** *Consider the saddle point problem in* (1) *under Assumption 1 and the OGDA method outlined in Algorithm 1. Further, recall the definition of* r_k *in* (2). *If we set* $\eta = (1/40\sqrt{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})})$, *then the iterates* $\{\mathbf{x}_k, \mathbf{y}_k\}_{k \geq 0}$ *generated by the OGDA method satisfy*

$$r_{k+1} \leq \left(1 - c\kappa^{-1}\right)^k \hat{r}_0,$$

where $\hat{r}_0 = \max\{r_2, r_1, r_0\}$ *and* c *is a positive constant independent of the problem parameters.*

The result in Theorem 3 shows linear convergence of OGDA in a bilinear problem of the form $f(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{B}\mathbf{y}$ where matrix $\mathbf{B}$ is square and full rank. It further shows that the overall number of iterations to obtain an ϵ-accurate solution is of $\mathcal{O}(\kappa \log(1/\epsilon))$, where κ is the problem condition number as defined in Assumption 1. We would like to mention that this result is similar to the one shown in (Liang and Stokes, 2019), except here we analyze OGDA as an approximation of PP.

In the following theorem, we again use the proximal point approximation interpretation of OGDA to provide a convergence rate estimate for this algorithm when it is used for solving a general strongly convex-strongly concave saddle point problem.

Theorem 4 (Strongly convex-strongly concave case)**.** *Consider the saddle point problem in* (1) *under Assumptions 2 and 3 and the OGDA method outlined in Algorithm 1. Further, recall the definition of r_k in* (2)*. If we set $\eta = (1/(8L))$, then the iterates $\{\mathbf{x}_k, \mathbf{y}_k\}_{k\geq 0}$ generated by OGDA satisfy*

$$r_{k+1} \leq \left(1 - c\kappa^{-1}\right)^k \hat{r}_0,$$

where $\hat{r}_0 = c_1\kappa^2 r_0$ and c, c_1 are positive constant independent of the problem parameters.

The result in Theorem 4 shows that OGDA converges linearly to the optimal solution under the assumptions that f is smooth and strongly convex-strongly concave. In other words, it shows that to achieve a point with error $r_k \leq \epsilon$, we need to run at most $\mathcal{O}(\kappa \log(1/\epsilon))$ iterations of OGDA. This result can be compared with the results in (Gidel et al., 2019), a recent independent work which derived the OGDA updates as a variant of the EG updates (interpreting OGDA as extrapolation from the past). In this connection, the OGDA iterates are the "midpoints" whereas (Gidel et al., 2019) provides a convergence of the original points (not the OGDA iterates) to an error of ϵ in $\mathcal{O}(\kappa \log(1/\epsilon))$.

4.2 Generalized OGDA method

The update of OGDA both in theory and practice is only studied for the case that the coefficients of both $\nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k)$ and $\nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k) - \nabla_{\mathbf{x}} f(\mathbf{x}_{k-1}, \mathbf{y}_{k-1})$ are η. This implies that in the OGDA update at step k, the coefficient of the current gradient, i.e., $\nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k)$, should be exactly twice the coefficient of the negative of the previous gradient, i.e., $-\nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k)$. It has been an open question to see if different stepsizes can be used for these terms. In this section, we generalize OGDA where the coefficients for the gradient descent and the negative momentum terms are not necessary equal to each other. We consider the following OGDA dynamics with general stepsize parameters $\alpha, \beta > 0$:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - (\alpha + \beta)\nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k) + \beta \nabla_{\mathbf{x}} f(\mathbf{x}_{k-1}, \mathbf{y}_{k-1}), \tag{7}$$

$$\mathbf{y}_{k+1} = \mathbf{y}_k + (\alpha + \beta)\nabla_{\mathbf{y}} f(\mathbf{x}_k, \mathbf{y}_k) - \beta \nabla_{\mathbf{y}} f(\mathbf{x}_{k-1}, \mathbf{y}_{k-1}). \tag{8}$$

Note that for $\alpha = \beta$, we recover the original OGDA method. Our goal is to show that OGDA is convergent even if α and β are not equal to each other, as long as their difference is sufficiently small. In the following theorem, we formally state our result for the generalized OGDA method described in (7) and (8) when the objective function f has a bilinear form of $f(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{B} \mathbf{y}$.

Theorem 5 (Generalized bilinear case)**.** *Consider the saddle point problem in* (1) *under Assumption 1 and the generalized OGDA method in* (7)*-*(8)*. Further, recall the definition of r_k in* (2)*. If we set $\alpha = 1/(40\sqrt{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})})$ and α and β satisfy the conditions $0 < \alpha - K\alpha^2 \leq \beta \leq \alpha$ for some constant $K > 0$, then the iterates $\{\mathbf{x}_k, \mathbf{y}_k\}_{k\geq 0}$ generated by the generalized OGDA method satisfy*

$$r_{k+1} \leq \left(1 - c\kappa^{-1}\right)^k \ \hat{r}_0,$$

where $\hat{r}_0 = \max\{r_2, r_1, r_0\}$ and c is a positive constant independent of the problem parameters.

Theorem 5 shows that it is not necessary to use a factor of 2 in the OGDA update to have a linearly convergent method and for a wide range of parameters this result holds. A result similar to Theorem 5 can be established when $\beta > \alpha$. We do not state the results here due to space limitations.

5 Extra-gradient method

In this section, we study the Extra-gradient (EG) method for solving the general saddle point problem in (1) and provide linear rates of convergence for the bilinear and the strongly convex-strongly concave case by interpreting this algorithm as an approximation of the proximal point method. [2]

The main idea of the EG method is to use the gradient at the current point to find a mid-point, and then use the gradient at that mid-point to find the next iterate. To be more precise, given a stepsize $\eta > 0$, the update of EG at step k for solving the saddle point problem in (1) has two steps. First, we find mid-point iterates $\mathbf{x}_{k+1/2}$ and $\mathbf{y}_{k+1/2}$ by performing a primal-dual gradient update as

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

Then, 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

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

The steps of the EG method for solving saddle point problems are outlined in Algorithm 2.

Note that in the update of the EG method, as the name suggests, requires evaluation of extra

[2] Note that while Nemirovski (2004) shows that EG is related to 'conceptual mirror prox' (similar to PP), unlike our analysis, Nemirovski (2004) does not analyze EG as an approximation of PP.

Algorithm 2 Extra-gradient method for saddle point problem

Require: Stepsize $\eta > 0$, initial vectors $\mathbf{x}_0, \mathbf{y}_0 \in \mathbb{R}^d$
1: **for** $k = 1, 2, \ldots$ **do**
2: Compute $\mathbf{x}_{k+1/2} = \mathbf{x}_k - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k)$ and $\mathbf{y}_{k+1/2} = \mathbf{y}_k + \eta \nabla_{\mathbf{y}} f(\mathbf{x}_k, \mathbf{y}_k)$;
3: Update $\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_{k+1/2}, \mathbf{y}_{k+1/2})$ and $\mathbf{y}_{k+1} = \mathbf{y}_k + \eta \nabla_{\mathbf{y}} f(\mathbf{x}_{k+1/2}, \mathbf{y}_{k+1/2})$;
4: **end for**

gradients at the midpoints $\mathbf{x}_{k+1/2}$ and $\mathbf{y}_{k+1/2}$ which doubles the computational complexity of EG compared to the vanilla Gradient Descent Ascent (GDA) method. The midpoints of EG evolve as $\mathbf{x}_{k+1/2} = \mathbf{x}_{k-1/2} - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_{k-1/2}, \mathbf{y}_{k-1/2}) - \eta \left(\nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k) - \nabla_{\mathbf{x}} f(\mathbf{x}_{k-1}, \mathbf{y}_{k-1})\right)$ (and similarly for $\mathbf{y}$), which shows that the EG method approximates the PP update direction as $\nabla_{\mathbf{x}} f(\mathbf{x}_{k+1/2}, \mathbf{y}_{k+1/2})) \approx \nabla_{\mathbf{x}} f(\mathbf{x}_{k-1/2}, \mathbf{y}_{k-1/2}) + (\nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k) - \nabla_{\mathbf{x}} f(\mathbf{x}_{k-1}, \mathbf{y}_{k-1}))$ We show next EG approximates the Proximal Point (PP) method more accurately, as compared to the GDA method. Consider the bilinear saddle point problem defined in Assumption 1. By following the update of PP in Section 3 and simplifying the expressions, the PP update for the bilinear problem under Assumption 1 can be written as

$$\mathbf{x}_{k+1} = (\mathbf{I} + \eta^2 \mathbf{B}\mathbf{B}^\top)^{-1}(\mathbf{x}_k - \eta \mathbf{B}\mathbf{y}_k),$$
$$\mathbf{y}_{k+1} = (\mathbf{I} + \eta^2 \mathbf{B}^\top \mathbf{B})^{-1}(\mathbf{y}_k + \eta \mathbf{B}^\top \mathbf{x}_k).$$

As the computation of the inverse $(\mathbf{I}+\eta^2\mathbf{B}^\top\mathbf{B})^{-1}$ could be costly, one can use $\mathbf{I}$ instead with an error of $o(\eta)$. This approximation retrieves the update of GDA which is known to possibly diverge for bilinear saddle point problems (see Daskalakis et al. (2017)). If we use the more accurate approximation $(I + \eta^2 \mathbf{B}\mathbf{B}^\top)^{-1} \approx (\mathbf{I} - \eta^2 \mathbf{B}\mathbf{B}^\top)$ which has an error of $o(\eta^2)$, we obtain

$$\begin{aligned}\mathbf{x}_{k+1} &= (\mathbf{I} - \eta^2 \mathbf{B}\mathbf{B}^\top + o(\eta^2))(\mathbf{x}_k - \eta \mathbf{B}\mathbf{y}_k) \\ &= \mathbf{x}_k - \eta \mathbf{B}\mathbf{y}_k - \eta^2 \mathbf{B}\mathbf{B}^\top \mathbf{x}_k + o(\eta^2), \quad (9)\end{aligned}$$
$$\begin{aligned}\mathbf{y}_{k+1} &= (\mathbf{I} - \eta^2 \mathbf{B}^\top \mathbf{B} + o(\eta^2))(\mathbf{y}_k + \eta \mathbf{B}^\top \mathbf{x}_k) \\ &= \mathbf{y}_k + \eta \mathbf{B}^\top \mathbf{x}_k - \eta^2 \mathbf{B}^\top \mathbf{B}\mathbf{y}_k + o(\eta^2). \quad (10)\end{aligned}$$

If we ignore the extra terms in (9)-(10) which are of $o(\eta^2)$, we recover the update of the EG method for the bilinear saddle point problem defined in Assumption 1. Therefore, in the bilinear problem, the EG method can be interpreted as an approximation of the PP method with an error of $o(\eta^2)$. In the following proposition, we extend this result and show that for any general smooth (possibly nonconvex) function $f(\mathbf{x}, \mathbf{y})$, EG is an $o(\eta^2)$ approximation of PP.

Proposition 2. *Consider the saddle point problem in* (1). *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 the PP update on* $(\mathbf{x}_k, \mathbf{y}_k)$, *and let* $(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})$ *be the point we obtain by performing the EG update on* $(\mathbf{x}_k, \mathbf{y}_k)$. *Then, for a given stepsize* $\eta > 0$ *we have*

$$\|\mathbf{y}_{k+1} - \hat{\mathbf{y}}_{k+1}\| \le o(\eta^2),$$
$$\|\mathbf{x}_{k+1} - \hat{\mathbf{x}}_{k+1}\| \le o(\eta^2). \quad (11)$$

The next theorem views the EG method as the PP method with an error and properly bounds the error to provide convergence rate estimates for the EG method in the bilinear case.

Theorem 6 (Bilinear case)**.** *Consider the saddle point problem in* (1) *under Assumption 1 and the extra-gradient (EG) method outlined in Algorithm 2. Further, recall the definition of* r_k *in* (2). *If we set* $\eta = 1/(2\sqrt{2\lambda_{\max}(\mathbf{B}^\top\mathbf{B})})$, *then the iterates* $\{\mathbf{x}_k, \mathbf{y}_k\}_{k\ge0}$ *generated by the EG method satisfy*

$$r_{k+1} \le \left(1 - c\kappa^{-1}\right)\, r_k, \quad (12)$$

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

The result in Theorem 6 shows linear convergence of the iterates generated by the EG method for a bilinear problem of the form $f(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{B}\mathbf{y}$ where the matrix $\mathbf{B}$ is square and full rank. In other words, we obtain that the overall number of iterations to reach a point satisfying $\|\mathbf{x}_k\|^2 + \|\mathbf{y}_k\|^2 \le \epsilon$ is at most $\mathcal{O}(\kappa \log(1/\epsilon))$ which matches the rate achieved in Tseng (1995) for bilinear problems. It is worth mentioning that we obtain this optimal complexity of $\mathcal{O}(\kappa \log(1/\epsilon))$ for EG in bilinear problems by analyzing this algorithm as an approximation of the PP method which differs from the approach used in (Tseng, 1995) that directly analyzes EG using a variational inequality approach. We further would like to add that this result improves the $\mathcal{O}(\kappa^2 \log(1/\epsilon))$ of (Liang and Stokes, 2019) for EG in bilinear problems.

The following theorem characterizes the convergence rate of the EG method when $f(\mathbf{x}, \mathbf{y})$ is strongly convex-strongly concave.

Theorem 7 (Strongly convex-strongly concave case)**.** *Consider the saddle point problem in* (1) *under Assumptions 2 and 3 and the extra-gradient (EG) method outlined in Algorithm 2. Further, recall the definition of the condition number* κ *in Remark 1 and the definition of* r_k *in* (2). *If we set* $\eta = 1/(8L)$, *then the iterates* $\{\mathbf{x}_k, \mathbf{y}_k\}_{k\ge0}$ *generated by the EG method satisfy*

$$r_{k+1} \le \left(1 - c\kappa^{-1}\right) r_k,$$

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

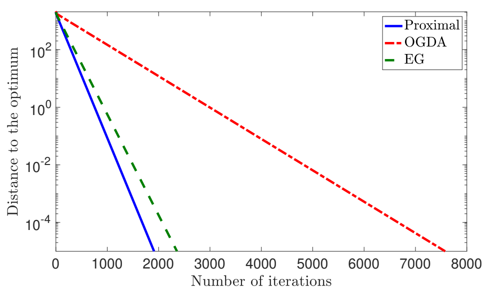

Figure 2: Convergence of proximal point (PP), extra-gradient (EG), and optimistic gradient descent ascent (OGDA) in terms of number of iterations for the bilinear problem in (13). All algorithms converge linearly, and the proximal point method has the best performance. Stepsizes of EG and OGDA were tuned for best performance.

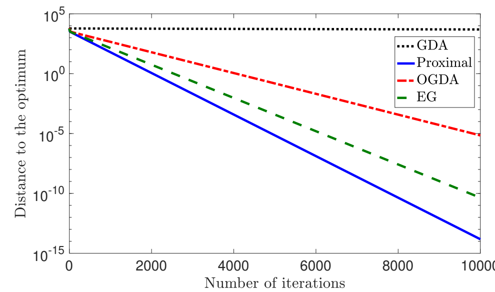

Figure 3: Convergence of proximal point (PP), extra-gradient (EG), optimistic gradient descent ascent (OGDA), and gradient descent ascent (GDA) in terms of number of iterations for the quadratic problem in (14). Stepsizes of EG, OGDA and GDA were tuned for best performance.

The result in Theorem 7 characterizes a linear convergence rate for the EG algorithm in a general smooth and strongly convex-strongly concave case. Similar to the bilinear case, our proof relies on interpreting EG as an approximation of the PP method. Theorem 7 further shows that the computational complexity of EG to achieve an ϵ-suboptimal solution, i.e., $\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 + \|\mathbf{y}_{k+1} - \mathbf{y}^*\|^2 \leq \epsilon$, is $\mathcal{O}(\kappa \log(1/\epsilon))$, where $\kappa = L/\mu$ is the condition number of the number. Note that this complexity bound can also be obtained from the results in (Tseng, 1995).

6 Numerical Experiments

In this section, we compare the performance of the Proximal Point (PP) method with the Extra–Gradient (EG), Gradient Descent Ascent (GDA), and Optimistic Gradient Descent Ascent (OGDA) methods.

We first focus on the following bilinear problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} \max_{\mathbf{y}\in\mathbb{R}^d} \mathbf{x}^\top \mathbf{B}\mathbf{y}. \tag{13}$$

where we set $\mathbf{B} \in \mathbb{R}^{d\times d}$ to be a diagonal matrix with a condition number of 100, and we set the dimension of the problem to $d = 10$. The iterates are initialized at $\mathbf{x}_0 = \mathbf{10}$ and $\mathbf{y}_0 = \mathbf{10}$, where $\mathbf{10}$ is a d dimensional vector with all elements equal to 10. Figure 2 demonstrates the errors of PP, OGDA, and EG versus number of iterations for this bilinear problem. Note that in this figure we do not show the error of GDA since it diverges for this problem, as illustrated in Figure 1 (For more details check (Daskalakis et al., 2017)). We can observe that all the three considered algorithms converge linearly to the optimal solution $(\mathbf{x}^*, \mathbf{y}^*) = (\mathbf{0}, \mathbf{0})$.

We proceed to study the performance of PP, EG, GDA, and OGDA for solving the following strongly convex-strongly concave saddle point problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} \max_{\mathbf{y}\in\mathbb{R}^n} \frac{1}{n}\left[-\frac{1}{2}\|\mathbf{y}\|^2 - \mathbf{b}^\top\mathbf{y} + \mathbf{y}^\top \mathbf{A}\mathbf{x}\right] + \frac{\lambda}{2}\|\mathbf{x}\|^2 \tag{14}$$

This is the saddle point reformulation of the linear regression

$$\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2n}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \tag{15}$$

with an L_2 regularization, as shown in Du and Hu (2019). As done in Du and Hu (2019), we generate the rows of the matrix $\mathbf{A}$ according to a Gaussian distribution $\mathcal{N}(0, I_d)$. Here, we set $d = 50$ and $n = 10$, and assume $\mathbf{b} = \mathbf{0}$. We also set the regularizer to $\lambda = 1/n$. Figure 3 illustrates the distance to the optimal solution of the considered algorithms versus the number of iterations. As we can see, EG and OGDA perform better than GDA and their convergence paths are closer to the one for PP which has the fastest rate. This observation matches our theoretical claim that EG and OGDA are more accurate approximations of PP relative to GDA.

7 Conclusions

We considered discrete time gradient based methods for solving saddle point problems, with a focus on the Extra-gradient (EG) and the Optimistic Gradient Descent Ascent (OGDA) methods. We showed that EG and OGDA can be seen as approximations of the classical Proximal Point (PP) method and used this interpretation to establish linear convergence rate for both of these algorithms in the bilinear and strongly convex-strongly concave settings. We further introduced a generalized version of OGDA and established its convergence guarantees for a wide range of parameters. We also compared the performance of EG, OGDA and PP for a strongly-convex strongly-concave saddle point problem.

References

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 214–223.

Basar, T. and Olsder, G. J. (1999). *Dynamic noncooperative game theory*, volume 23. SIAM.

Bauschke, H. H., Combettes, P. L., et al. (2011). *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer.

Beck, A. (2017). *First-Order Methods in Optimization*, volume 25. SIAM.

Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. (2009). *Robust optimization*, volume 28. Princeton University Press.

Bertsekas, D. P. (1999). *Nonlinear programming*. Athena Scientific Belmont.

Chambolle, A. and Pock, T. (2011). A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145.

Chavdarova, T., Gidel, G., Fleuret, F., and Lacoste-Julien, S. (2019). Reducing noise in GAN training with variance reduced extragradient. *arXiv preprint arXiv:1904.08598*.

Chen, G. H. and Rockafellar, R. T. (1997). Convergence rates in forward–backward splitting. *SIAM Journal on Optimization*, 7(2):421–444.

Chen, Y., Lan, G., and Ouyang, Y. (2014). Optimal primal-dual methods for a class of saddle point problems. *SIAM Journal on Optimization*, 24(4):1779–1814.

Chiang, C., Yang, T., Lee, C., Mahdavi, M., Lu, C., Jin, R., and Zhu, S. (2012). Online optimization with gradual variations. In *COLT 2012 - The 25th Annual Conference on Learning Theory, June 25-27, 2012, Edinburgh, Scotland*, pages 6.1–6.20.

Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. (2017). Training GANs with optimism. *CoRR*, abs/1711.00141.

Du, S. S. and Hu, W. (2019). Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, pages 196–205.

Eckstein, J. and Bertsekas, D. P. (1992). On the douglas rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1-3):293–318.

Facchinei, F. and Pang, J.-S. (2007). *Finite-dimensional variational inequalities and complementarity problems*. Springer Science & Business Media.

Ferris, M. C. (1991). Finite termination of the proximal point algorithm. *Mathematical Programming*, 50(1-3):359–366.

Gidel, G., Berard, H., Vignoud, G., Vincent, P., and Lacoste-Julien, S. (2019). A variational inequality perspective on generative adversarial networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Güler, O. (1991). On the convergence of the proximal point algorithm for convex minimization. *SIAM Journal on Control and Optimization*, 29(2):403–419.

Hast, M., Astrom, K., Bernhardsson, B., and Boyd, S. (2013). Pid design by convex-concave optimization. In *Control Conference (ECC), 2013 European*, pages 4460–4465. Citeseer.

Juditsky, A., Nemirovski, A., and Tauvel, C. (2011). Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58.

Korpelevich, G. (1976). The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756.

Liang, T. and Stokes, J. (2019). Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, pages 907–915.

Martinet, B. (1970). Brève communication. régularisation d'inéquations variationnelles par approximations successives. *Revue française d'informatique et de recherche opérationnelle. Série rouge*, 4(R3):154–158.

Monteiro, R. D. and Svaiter, B. F. (2010). On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. *SIAM Journal on Optimization*, 20(6):2755–2787.

Nedić, A. and Ozdaglar, A. (2009). Subgradient methods for saddle-point problems. *Journal of optimization theory and applications*, 142(1):205–228.

Nemirovski, A. (2004). Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609.

Palaniappan, B. and Bach, F. (2016). Stochastic variance reduction methods for saddle-point problems. In *Advances in Neural Information Processing Systems*, pages 1416–1424.

Parikh, N., Boyd, S., et al. (2014). Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239.

Rakhlin, A. and Sridharan, K. (2013a). Online learning with predictable sequences. In *COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA*, pages 993–1019.

Rakhlin, S. and Sridharan, K. (2013b). Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems*, pages 3066–3074.

Rockafellar, R. T. (1976a). Augmented lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of operations research*, 1(2):97–116.

Rockafellar, R. T. (1976b). Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898.

Sanjabi, M., Ba, J., Razaviyayn, M., and Lee, J. D. (2018a). On the convergence and robustness of training gans with regularized optimal transport. In *Advances in Neural Information Processing Systems*, pages 7091–7101.

Sanjabi, M., Razaviyayn, M., and Lee, J. D. (2018b). Solving non-convex non-concave min-max games under Polyak-Lojasiewicz condition. *arXiv preprint arXiv:1812.02878*.

Tseng, P. (1995). On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 60(1-2):237–252.