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-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

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

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

$$f(x^*, y) \leq f(x^*, y^*) \leq f(x, y^*),$$

for all $x \in \mathbb{R}^m, y \in \mathbb{R}^n$. 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 $f(x, y)$ and bilinear functions, i.e., $f(x, y) = x^T A y$, 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 $f(x, y) = x^T A y$ and the matrix $A$ is square and full rank. It also presented a linear convergence rate of the vanilla Gradient Ascent Descent (GDA) method when the objective function $f(x, y)$ 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 $f(x, y)$ 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

$^1f(x, y)$ is strongly convex-strongly concave when it is strongly convex with respect to $x$ and strongly concave with respect to $y$. 

---

Aryan Mokhtari  
UT Austin  
Asuman Ozdaglar  
MIT  
Sarath Pattathil  
MIT
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 Marinet (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 $O(\kappa \log(1/\epsilon))$ (see Assumption 4 and Remark 4 for the definition of $\kappa$), when the function $f(x, y)$ 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 $O(\kappa \log(1/\epsilon))$ for OGDA and a complexity of $O(\kappa^2 \log(1/\epsilon))$ for EG, without analyzing 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 $\epsilon$ in $O(\kappa \log(1/\epsilon))$. In this paper, we establish an overall complexity of $O(\kappa \log(1/\epsilon))$ 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 $O(1/k)$ 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 $f(x, y) = g(x) + x^\top A y - h(y)$, 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 $f(x, y)$ is strongly concave with respect to $y$, but possibly nonconvex with respect to $x$, Sanjabi et al. (2018a) provided convergence to a first-order stationary point using an algorithm that requires running multiple updates with respect to $y$ at each step. Recently, Sanjabi et al. (2018b) extended this result to the setting when $f$ is Polyak-Lojasiewicz with respect to $y$.

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-
We start the paper by presenting some definitions and work, a special case of Optimistic Mirror descent was version of the Optimistic Mirror Descent algorithm in interpretation to study the convergence properties of EG problem. Then, we show that EG can be interpreted as case (Theorem 5). In Section 5, we recap the update of the performance of PP, EG, and OGDA for solving bilinear problems (Theorem 6) and general strongly convex-strongly concave (Theorem 2) problems. In Section 6, we present our numerical results, comparing in bilinear problems (Theorem 1) and general strongly convex-strongly concave (Theorem 3) and strongly convex-strongly concave (Theorem 4) problems. 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 $x, x' \in \mathbb{R}^n$, we have $\|\nabla \phi(x) - \nabla \phi(x')\| \leq L\|x - x'\|$.  

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

**Definition 3.** The pair $(x^*, y^*)$ is a saddle point of a convex-concave function $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, if for any $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$, we have $f(x^*, y) \leq f(x, y^*) \leq f(x, y')$. Throughout the paper, we consider two specific cases of Problem (1) stated in the next set of assumptions.

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

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

### Table 1: Comparison of rates in different papers

<table>
<thead>
<tr>
<th>Reference</th>
<th>Assumptions on $f(x, y)$</th>
<th>Rate (EG)</th>
<th>Rate (OGDA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liang and Stokes [2019]</td>
<td>Bilinear</td>
<td>$O(\kappa^2 \log(1/\epsilon))$</td>
<td>$O(\kappa \log(1/\epsilon))$</td>
</tr>
<tr>
<td>Liang and Stokes [2019]</td>
<td>Strongly Convex-Strongly Concave</td>
<td>$\times$</td>
<td>$\times$</td>
</tr>
<tr>
<td>Gidel et al. [2019]</td>
<td>Bilinear</td>
<td>$\times$</td>
<td>$\times$</td>
</tr>
<tr>
<td>Gidel et al. [2019]</td>
<td>Strongly Convex-Strongly Concave</td>
<td>$\times$</td>
<td>$O(\kappa \log(1/\epsilon))^*$</td>
</tr>
<tr>
<td>Tseng [1995]</td>
<td>Bilinear</td>
<td>$O(\kappa \log(1/\epsilon))$</td>
<td>$\times$</td>
</tr>
<tr>
<td>Tseng [1995]</td>
<td>Strongly Convex-Strongly Concave</td>
<td>$O(\kappa \log(1/\epsilon))$</td>
<td>$\times$</td>
</tr>
<tr>
<td>This paper</td>
<td>Strongly Convex-Strongly Concave</td>
<td>$O(\kappa \log(1/\epsilon))$</td>
<td>$O(\kappa \log(1/\epsilon))$</td>
</tr>
</tbody>
</table>

(*- Gidel et al. [2019] shows the convergence of the half points and not the original OGDA iterates.)
Assumption 3. The gradient $\nabla_x f(x, y)$, is $L_x$-Lipschitz in $x$ and $L_{xy}$-Lipschitz in $y$, i.e.,

\[
\|\nabla_x f(x_1, y) - \nabla_x f(x_2, y)\| \leq L_x \|x_1 - x_2\| \quad \text{for all } y,
\]

\[
\|\nabla_x f(x, y_1) - \nabla_x f(x, y_2)\| \leq L_{xy}\|y_1 - y_2\| \quad \text{for all } x.
\]

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

\[
\|\nabla_y f(x_1, y) - \nabla_y f(x_2, y)\| \leq L_y \|y_1 - y_2\| \quad \text{for all } x,
\]

\[
\|\nabla_y f(x, y_1) - \nabla_y f(x, y_2)\| \leq L_{yx}\|x_1 - x_2\| \quad \text{for all } y.
\]

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 $(x_k, y_k)$. We denote $r_k = \|x_k - x^*\|^2 + \|y_k - y^*\|^2$, as the distance to the saddle point $(x^*, y^*)$ 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

\[
 x_{k+1} = \text{prox}_{\frac{1}{2\eta}h}(x_k) = \text{argmin} \left\{ h(x) + \frac{1}{2\eta} \|x - x_k\|^2 \right\},
\]

where $\eta$ 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 $x_{k+1} = x_k - \eta \nabla h(x_k)$. 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 $\{x_{k+1}, y_{k+1}\}$ as the unique solution to the saddle point problem

\[
\min_{x \in \mathbb{R}^m} \max_{y \in \mathbb{R}^n} \left\{ f(x, y) + \frac{1}{2\eta} \|x - x_k\|^2 - \frac{1}{2\eta} \|y - y_k\|^2 \right\},
\]

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

\[
x_{k+1} = x_k - \eta \nabla_x f(x_{k+1}, y_{k+1}),
\]

\[
y_{k+1} = y_k + \eta \nabla_y f(x_{k+1}, y_{k+1}).
\]

Note that implementing the system of updates in (5) requires computing the operators $(I + \eta \nabla f)^{-1}$ and $(I + \eta \nabla f)^{-1}$, 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 $(x^*, y^*) = (0, 0)$ which is the unique solution of the problem $\min_x \max_y x^\top By$. 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 3 and the proximal point method in (5). Further, recall the definition of $r_k$ in (2). Then, for any $\eta > 0$, the iterates $\{x_k, y_k\}_{k \geq 0}$ generated by the proximal point method satisfy

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

In the following theorem, we characterize the convergence rate of PP for a function $f(x, y)$ that is strongly convex with respect to $x$ and strongly concave with respect to $y$. 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 3 and the proximal point method in (5). Further, recall the definition of $r_k$ in (2). Then, for any $\eta > 0$, the iterates $\{x_k, y_k\}_{k \geq 0}$ 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-
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:

\[
x_{k+1} = x_k - \eta \nabla_x f(x_k, y_k) - \eta \nabla_x f(x_{k-1}, y_{k-1}),
\]

\[
y_{k+1} = y_k + 2\eta \nabla_y f(x_k, y_k) - \eta \nabla_y f(x_{k-1}, y_{k-1}).
\]

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., \(\nabla_x f(x_{k+1}, y_{k+1}) \approx \nabla_x f(x_k, y_k) + (\nabla_x f(x_{k-1}, y_{k-1}) - \nabla_x f(x_{k-1}, y_{k-1}))\) and similarly for \(\nabla_y f(\cdot, \cdot)\).

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

\[
x_{k+1} = x_k - 2\eta B y_k + \eta B y_{k-1},
\]

\[
y_{k+1} = y_k + 2\eta B^\top x_k + \eta B^\top x_{k-1}.
\]

Note that the update of the PP method for the variable \(x\) in the considered bilinear problem is

\[
x_{k+1} = (I + \eta^2 B B^\top)^{-1}(x_k - \eta B y_k)
\]

\[
= (I - \eta^2 B B^\top + o(\eta^2))(x_k - \eta B y_k)
\]

\[
= x_k - \eta B y_k - \eta B(\eta B^\top x_k - \eta^2 B^\top B y_k) + o(\eta^2),
\]

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

\[
x_{k+1} = x_k - 2\eta B y_k
\]

\[
- \eta B(\eta B^\top x_k - (1 + \eta^2 B^\top B)y_k) + o(\eta^2)
\]

\[
= x_k - 2\eta B y_k
\]

\[
- \eta B(\eta B^\top x_k - y_{k-1} - \eta B^\top x_{k-1}) + o(\eta^2)
\]

\[
= x_k - 2\eta B y_k + \eta B y_{k-1} + o(\eta^2),
\]

where the last expression is the OGDA update for variable \(x\) plus an additional error of \(o(\eta^2)\). A similar derivation can be done for the update of variable \(y\) 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 \(f(x, y)\).

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

\[
\|x_{k+1} - \hat{x}_{k+1}\| \leq o(\eta^2),
\]

\[
\|y_{k+1} - \hat{y}_{k+1}\| \leq o(\eta^2).
\]

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 \(\eta = (1/40\sqrt{\lambda_{\text{max}}(B^\top B)})\), then the iterates \(\{x_k, y_k\}_{k \geq 0}\) generated by the OGDA method satisfy

\[
r_{k+1} \leq (1 - c\kappa^{-1})^k r_0,
\]

where \(r_0 = \max\{r_2, r_1, r_0\}\) 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 \(f(x, y) = x^\top B y\) where matrix \(B\) is square and full rank. It further shows that the overall number of iterations to obtain an \(\epsilon\)-accurate solution is of \(O(\kappa \log(1/\epsilon))\), where \(\kappa\) 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 1 and 2 and the OGD method outlined in Algorithm 1. Further, recall the definition of \( r_k \) in (2). If we set \( \eta = (1/(8L)) \), then the iterates \( \{x_k, y_k\}_{k \geq 0} \) generated by OGD satisfy

\[
r_{k+1} \leq (1 - c_k^{-1})^k r_0,
\]

where \( r_0 = c_1 \kappa^2 r_0 \) and \( c, c_1 \) are positive constant independent of the problem parameters.

The result in Theorem 4 shows that OGD 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 within error \( r_k \leq \epsilon \), we need to run at most \( O(\kappa \log(1/\epsilon)) \) iterations of OGD. This result can be compared with the results in (Gidel et al., 2019), a recent independent work which derived the OGD updates as a variant of the EG update (interpreting OGD as extrapolation from the past). In this connection, the OGD iterates are the “midpoints” whereas (Gidel et al., 2019) provides a convergence of the original points (not the OGD iterates) to an error of \( \epsilon \) in \( O(\kappa \log(1/\epsilon)) \).

### 4.2 Generalized OGD method

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

\[
x_{k+1} = x_k - (\alpha + \beta)\nabla_x f(x_k, y_k) + \beta \nabla_x f(x_{k-1}, y_{k-1}),
\]

\[
y_{k+1} = y_k + (\alpha + \beta)\nabla_y f(x_k, y_k) - \beta \nabla_y f(x_{k-1}, y_{k-1}).
\]

Note that for \( \alpha = \beta \), we recover the original OGD method. Our goal is to show that OGD is convergent even if \( \alpha \) and \( \beta \) 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 OGD method described in (7) and (8) when the objective function \( f \) has a bilinear form of \( f(x, y) = x^\top B y \).

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

\[
r_{k+1} \leq (1 - c_k^{-1})^k r_0,
\]

where \( r_0 = \max\{r_2, r_1, r_0\} \) 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 OGD 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.

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 \( x_{k+1/2} \) and \( y_{k+1/2} \) by performing a primal-dual gradient update as

\[
x_{k+1/2} = x_k - \eta \nabla_x f(x_k, y_k),
\]

\[
y_{k+1/2} = y_k + \eta \nabla_y f(x_k, y_k).
\]

Then, the gradients evaluated at the midpoints \( x_{k+1/2} \) and \( y_{k+1/2} \) are used to compute the new iterates \( x_{k+1} \) and \( y_{k+1} \) by performing the updates

\[
x_{k+1} = x_k - \eta \nabla_x f(x_{k+1/2}, y_{k+1/2}),
\]

\[
y_{k+1} = y_k + \eta \nabla_y f(x_{k+1/2}, y_{k+1/2}).
\]

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

---

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 $x_0, y_0 \in \mathbb{R}^d$

1: for $k = 1, 2, \ldots$ do
2: Compute $x_{k+1} = x_k - \eta \nabla_x f(x_k, y_k)$ and $y_{k+1} = y_k + \eta \nabla_y f(x_k, y_k)$;
3: Update $x_{k+1} = x_k - \eta \nabla_x f(x_{k+1}/2, y_{k+1}/2)$ and $y_{k+1} = y_k + \eta \nabla_y f(x_{k+1}/2, y_{k+1}/2)$;
4: end for

Proposition 2. Consider the saddle point problem in (1). Given a point $(x_k, y_k)$, let $(x_{k+1}, y_{k+1})$ be the point we obtain by performing the PP update on $(x_k, y_k)$, and let $(x_{k+1}, y_{k+1})$ be the point we obtain by performing the EG update on $(x_k, y_k)$. Then, for a given stepsize $\eta > 0$ we have

$$
\|y_{k+1} - y_{k+1}\| \leq o(\eta^2),
$$

$$
\|x_{k+1} - x_{k+1}\| \leq o(\eta^2).
$$

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 4 and the extra-gradient (EG) method outlined in Algorithm 2. Further, recall the definition of $r_k$ in (2). If we set $\eta = 1/(2\sqrt{2\lambda_{\max}(B^\top B)})$, then the iterates $\{x_k, y_k\}_{k \geq 0}$ generated by the EG method satisfy

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

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 $f(x, y) = x^\top By$ where the matrix $B$ is square and full rank. In other words, we obtain that the overall number of iterations to reach a point satisfying $\|x_k\|^2 + \|y_k\|^2 \leq \epsilon$ is at most $O(\kappa \log(1/\epsilon))$ which matches the rate achieved in [Tseng (1995)] for bilinear problems. It is worth mentioning that we obtain this optimal complexity of $O(\kappa \log(1/\epsilon))$ 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 $O(\kappa^2 \log(1/\epsilon))$ of [Liang and Stokes (2019)] for EG in bilinear problems.

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

Theorem 7 (Strongly convex-strongly concave case). Consider the saddle point problem in (1) under Assumptions 3 and 4 and the extra-gradient (EG) method outlined in Algorithm 2. Further, recall the definition of the condition number $\kappa$ in Remark 1 and the definition of $r_k$ in (2). If we set $\eta = 1/(8L)$, then the iterates $\{x_k, y_k\}_{k \geq 0}$ generated by the EG method satisfy

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

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. All algorithms converge linearly, and the proximal method has the best performance. Stepsizes of EG and OGDA were tuned for best performance.

The result in Theorem 4 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 5 further shows that the computational complexity of EG to achieve an $\epsilon$-suboptimal solution, i.e., $\|x_{k+1} - x^*\|^2 + \|y_{k+1} - y^*\|^2 \leq \epsilon$, is $O(\kappa \log(1/\epsilon))$, where $\kappa = \lambda/\mu$ 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

$$\min_{x \in \mathbb{R}^d} \max_{y \in \mathbb{R}^d} x^T By.$$ (13)

where we set $B \in \mathbb{R}^{d \times d}$ 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 $x_0 = 10$ and $y_0 = 10$, where 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 $(x^*, y^*) = (0, 0)$.

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

$$\min_{x \in \mathbb{R}^d} \max_{y \in \mathbb{R}^d} \frac{1}{n} \left[ -\frac{1}{2} \|y\|^2 - b^T y + y^T Ax \right] + \frac{\lambda}{2} \|x\|^2$$ (14)

This is the saddle point reformulation of the linear regression

$$\min_{x \in \mathbb{R}^d} \frac{1}{2n} \|Ax - b\|^2$$ (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 $A$ according to a Gaussian distribution $N(0, I_d)$. Here, we set $d = 50$ and $n = 10$, and assume $b = 0$. 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


