# Convergence Rate of $\mathcal{O}(1/k)$ for Optimistic Gradient and Extra-gradient Methods in Smooth Convex-Concave Saddle Point Problems

Aryan Mokhtari*†, Asuman Ozdaglar*‡, Sarath Pattathil*§

**Abstract**

We study the iteration complexity of the optimistic gradient descent-ascent (OGDA) method and the extra-gradient (EG) method for finding a saddle point of a convex-concave unconstrained min-max problem. To do so, we first show that both OGDA and EG can be interpreted as approximate variants of the proximal point method. This is similar to the approach taken in [Nemirovski, 2004] which analyzes EG as an approximation of the 'conceptual mirror prox'. In this paper, we highlight how gradients used in OGDA and EG try to approximate the gradient of the Proximal Point method. We then exploit this interpretation to show that both algorithms produce iterates that remain within a bounded set. We further show that the primal dual gap of the averaged iterates generated by both of these algorithms converge with a rate of $\mathcal{O}(1/k)$. Our theoretical analysis is of interest as it provides a the first convergence rate estimate for OGDA in the general convex-concave setting. Moreover, it provides a simple convergence analysis for the EG algorithm in terms of function value without using compactness assumption.

## 1 Introduction

Given a function $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$, we consider finding a saddle point of the problem

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

where a saddle point of Problem (1) is defined as a pair $(\mathbf{x}^*, \mathbf{y}^*) \in \mathbb{R}^m \times \mathbb{R}^n$ that satisfies

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

for all $\mathbf{x} \in \mathbb{R}^m, \mathbf{y} \in \mathbb{R}^n$. Throughout the paper, we assume that the function $f(\mathbf{x}, \mathbf{y})$ is *convex-concave*, i.e., for any $\mathbf{y} \in \mathbb{R}^n$, the function $f(\mathbf{x}, \mathbf{y})$ is a convex function of $\mathbf{x}$ and for any $\mathbf{x} \in \mathbb{R}^m$, the function $f(\mathbf{x}, \mathbf{y})$ is a concave function of $\mathbf{y}$. This formulation arises in several areas, including zero-sum games [Basar & 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).

Our goal in this paper is to analyze the convergence rate of some discrete-time gradient based optimization algorithms for finding a saddle point of Problem (1) in the convex-concave case. In particular, we focus on Extra-gradient (EG) and Optimistic Gradient Descent Ascent (OGDA) methods because of their widespread use for training GANs (see [Daskalakis et al., 2018; Liang & Stokes, 2019]). EG method is a classical algorithm for solving saddle point problems introduced by Korpelevich [1976]. Its linear rate of convergence for smooth and strongly convex-strongly concave functions $f(\mathbf{x}, \mathbf{y})$ [1] and bilinear functions, i.e., $f(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{A}\mathbf{y}$ (where $\mathbf{A}$ is a square, full

*The authors are in alphabetical order.

†Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX. mokhtari@austin.utexas.edu

‡Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA. asuman@mit.edu.

§Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA. sarathp@mit.edu.

[1] Note that when we state that $f(\mathbf{x}, \mathbf{y})$ is strongly convex-strongly concave, it means that $f(\cdot, \mathbf{y})$ is strongly convex for all $\mathbf{y} \in \mathbb{R}^n$ and $f(\mathbf{x}, \cdot)$ is strongly concave for all $\mathbf{x} \in \mathbb{R}^m$.

rank matrix), was established in Korpelevich [1976] as well as the variational inequality literature (see [Tseng, 1995] and [Facchinei & Pang, 2007]). Its $\mathcal{O}(1/k)$ convergence rate for the constrained convex-concave setting was first established by Nemirovski [2004] under the assumption that the feasible set is convex and compact.[2] Monteiro & Svaiter [2010] established a similar $\mathcal{O}(1/k)$ convergence rate for EG without assuming compactness of the feasible set by using a new termination criterion that relies on enlargement of the operator of the VI reformulation of the saddle point problem defined in [Burachik et al., 1997]. OGDA was introduced by Popov [1980], as a variant of the Extragradient method, and has gained popularity recently due to its performance in training GANs (see [Daskalakis et al., 2018]). To the best of our knowledge, iteration complexity of OGDA for the convex-concave case has not been studied before.

In this paper, we provide a unified convergence analysis for establishing a sublinear convergence rate of $\mathcal{O}(1/k)$ in terms of the function value difference of the averaged iterates and a saddle point for both OGDA and EG for convex-concave saddle point problems. Our analysis holds for unconstrained problems and does not require boundedness of the feasible set, and it establishes rate results using the function value differences as used in [Nemirovski, 2004] (suitably redefined for an unconstrained feasible set, see Section 5). Therefore, we get convergence of the EG method in unconstrained spaces without using the modified termination (error) criterion proposed in [Monteiro & Svaiter, 2010]. The key idea of our approach is to view both OGDA and EG iterates as approximations of the iterates of the proximal point method that was first introduced by Martinet [1970] and later studied by Rockafellar [1976]. We would like to add that the idea of interpreting EG as an approximation of the Proximal Point method was first studied in [Nemirovski, 2004]. He considers the conceptual mirror prox, which is similar to the proximal point method, and shows that the mirror prox algorithm (of which EG is a special case) provides a good implementable approximation to this method. Further, Monteiro & Svaiter [2010] use a similar interpretation and propose the Hybrid Proximal Extragradient method to establish the convergence of EG in unbounded settings using a different convergence criteria. More recently, Mokhtari et al. [2020] study both OGDA and EG as approximations of proximal point method and analyze these algorithms for bilinear and strongly convex-strongly concave problems.

More specifically, we first consider a proximal point method with error and establish some key properties of its iterates. We then focus on OGDA as an approximation of proximal point method and use this connection to show that the iterates of OGDA remain in a compact set. We incorporate this result to prove a sublinear convergence rate of $\mathcal{O}(1/k)$ for the primal-dual gap of the averaged iterates generated by the OGDA update. We next consider EG where two gradient pairs are used in each iteration, one to compute a midpoint and other to find the new iterate using the gradient of the midpoint. Our first step again is to show boundedness of the iterates generated by EG. We then approximate the evolution of the midpoints using a proximal point method and use this approximation to establish $\mathcal{O}(1/k)$ convergence rate for the function value of the averaged iterates generated by EG. As the convergence results of EG have already been established in papers including [Nemirovski, 2004] and Monteiro & Svaiter [2010], we relegate the proofs of Lemmas and Theorems corresponding to EG to the Appendix.

## Related Work

Several recent papers have studied the convergence rate of OGDA and EG for the case when the objective function is bilinear or strongly convex-strongly concave. Daskalakis et al. [2018] showed the convergence of the OGDA iterates to a neighborhood of the solution when the objective function is bilinear. Liang & Stokes [2019] used a dynamical system approach to prove the linear convergence of the OGDA method 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. They 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. Gidel et al. [2018] considered a variant of the EG method, relating it to OGDA updates, and showed the linear convergence of the corresponding EG iterates in the case where $f(\mathbf{x}, \mathbf{y})$ is strongly convex-strongly concave (though without showing the convergence rate for the OGDA iterates). Optimistic gradient methods have also been studied in the context of convex online learning [Chiang et al., 2012; Rakhlin & Sridharan, 2013a,b].

[2] The result in [Nemirovski, 2004] shows a $\mathcal{O}(1/k)$ convergence rate for the mirror-prox algorithm which specializes to the EG method for the Euclidean case.

Nedić & 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, showing a convergence rate of $\mathcal{O}(1/\sqrt{k})$ in terms of the function value difference of the averaged iterates and a saddle point.

Chambolle & Pock [2011] focused on a particular case of the saddle point problem where the coupling term in the objective function is bilinear, i.e., $f(\mathbf{x},\mathbf{y}) = G(\mathbf{x}) + \mathbf{x}^\top \mathbf{K}\mathbf{y} - H(\mathbf{y})$ with $G$ and $H$ convex functions. They proposed a proximal point based algorithm which converges at a rate $\mathcal{O}(1/k)$ and further showed linear convergence when the functions $G$ and $H$ are strongly convex. Chen et al. [2014] proposed an accelerated variant of this algorithm when $G$ is smooth and showed an optimal rate of $(\frac{L_G}{k^2} + \frac{L_K}{k})$, where $L_G$ and $L_K$ are the smoothness parameters of $G$ and the norm of the linear operator $K$ respectively. When the functions $G$ and $H$ are strongly convex, primal-dual gradient-type methods converge linearly, as shown in Chen & Rockafellar [1997]; Bauschke et al. [2011]. Further, Du & Hu [2019] showed that GDA achieves a linear convergence rate in this linearly coupled setting when $G$ is convex and $H$ is strongly convex.

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

**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 points $(\mathbf{x}_0,\mathbf{y}_0)$ 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 convex on $\mathbb{R}^n$ if for any $\mathbf{x},\hat{\mathbf{x}} \in \mathbb{R}^n$, we have*

$$\phi(\hat{\mathbf{x}}) \ge \phi(\mathbf{x}) + \nabla\phi(\mathbf{x})^T(\hat{\mathbf{x}} - \mathbf{x}).$$

*Further, $\phi(\mathbf{x})$ is concave if $-\phi(\mathbf{x})$ is convex.*

**Definition 3.** *The pair $(\mathbf{x}^*,\mathbf{y}^*)$ is a saddle point of a convex-concave function $f(\mathbf{x},\mathbf{y})$, 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 will assume that the following conditions are satisfied.

**Assumption 1.** *The function $f(\mathbf{x},\mathbf{y})$ is continuously differentiable in $\mathbf{x}$ and $\mathbf{y}$. Further, for any $\mathbf{y} \in \mathbb{R}^n$, the function $f(\mathbf{x},\mathbf{y})$ is a convex function of $\mathbf{x}$ and for any $\mathbf{x} \in \mathbb{R}^m$, the function $f(\mathbf{x},\mathbf{y})$ is a concave function of $\mathbf{y}$.*

**Assumption 2.** *The gradient $\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y})$, is $L_{xx}$-Lipschitz with respect to $\mathbf{x}$ and $L_{xy}$-Lipschitz with respect to $\mathbf{y}$ and the gradient $\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y})$, is $L_{yy}$-Lipschitz with respect to $\mathbf{y}$ and $L_{yx}$-Lipschitz with respect to $\mathbf{x}$, i.e.,*

$$\begin{aligned}
\|\nabla_{\mathbf{x}} f(\mathbf{x}_1,\mathbf{y}) - \nabla_{\mathbf{x}} f(\mathbf{x}_2,\mathbf{y})\| &\le L_{xx}\|\mathbf{x}_1 - \mathbf{x}_2\| \quad \text{for all } \mathbf{y},\\
\|\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y}_1) - \nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y}_2)\| &\le L_{xy}\|\mathbf{y}_1 - \mathbf{y}_2\| \quad \text{for all } \mathbf{x},\\
\|\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y}_1) - \nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y}_2)\| &\le L_{yy}\|\mathbf{y}_1 - \mathbf{y}_2\| \quad \text{for all } \mathbf{x},\\
\|\nabla_{\mathbf{y}} f(\mathbf{x}_1,\mathbf{y}) - \nabla_{\mathbf{y}} f(\mathbf{x}_2,\mathbf{y})\| &\le L_{yx}\|\mathbf{x}_1 - \mathbf{x}_2\| \quad \text{for all } \mathbf{y}.
\end{aligned}$$

*We define $L := 2 \times \max\{L_{xx}, L_{xy}, L_{yx}, L_{yy}\}$.* [3]

[3] In this definition we need an additional factor of 2 because in the analysis we use $L$ as the Lipschitz continuity of the operator $F(\cdot) = [\nabla_{\mathbf{x}} f(\cdot); -\nabla_{\mathbf{y}} f(\cdot)]$.

**Assumption 3.** *The solution set $\mathcal{Z}^*$ defined as*

$$\mathcal{Z}^* := \{[\mathbf{x};\mathbf{y}] \in \mathbb{R}^{n+m} : (\mathbf{x},\mathbf{y}) \text{ is a saddle point of Problem (1)}\}, \tag{2}$$

*is nonempty.*

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 these algorithms are denoted by $(\mathbf{x}_k, \mathbf{y}_k)$. We denote the averaged (ergodic) iterates by $\hat{\mathbf{x}}_k, \hat{\mathbf{y}}_k$, defined as follows:

$$\hat{\mathbf{x}}_k = \frac{1}{k}\sum_{i=1}^{k} \mathbf{x}_i, \qquad \hat{\mathbf{y}}_k = \frac{1}{k}\sum_{i=1}^{k} \mathbf{y}_i. \tag{3}$$

In our convergence analysis, we use a variational inequality approach in which we define the vector $\mathbf{z} = [\mathbf{x};\mathbf{y}] \in \mathbb{R}^{n+m}$ as our decision variable and define the operator $F : \mathbb{R}^{m+n} \to \mathbb{R}^{m+n}$ as

$$F(\mathbf{z}) = [\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y}); -\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y})]. \tag{4}$$

In the following lemma we characterize the properties of operator $F$ in (4) when the conditions in Assumptions 1 and 2 are satisfied. We would like to emphasize that the following lemma is well-known – see, e.g., Nemirovski [2004] – and we state it for completeness.

**Lemma 1.** *Let $F(\cdot)$ be defined as in Equation (4). Suppose Assumptions 1 and 2 hold. Then*
*(a) $F$ is a monotone operator, i.e., for any $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^{m+n}$, we have*

$$\langle F(\mathbf{z}_1) - F(\mathbf{z}_2), \mathbf{z}_1 - \mathbf{z}_2\rangle \geq 0.$$

*(b) $F$ is an $L$-Lipschitz continuous operator, i.e., for any $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^{m+n}$, we have*

$$\|F(\mathbf{z}_1) - F(\mathbf{z}_2)\| \leq L\|\mathbf{z}_1 - \mathbf{z}_2\|.$$

*(c) For all $\mathbf{z}^* \in \mathcal{Z}^*$, we have $F(\mathbf{z}^*) = 0$.*

According to Lemma 1, when $f$ is convex-concave and smooth, the operator $F$ defined in (4) is monotone and Lipschitz. The third result in Lemma 1 shows that any saddle point of problem (1) satisfies the first-order optimality condition, i.e for all $(\mathbf{x}^*, \mathbf{y}^*) \in \mathcal{Z}^*$, we have:

$$\nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*) = 0 \qquad \nabla_{\mathbf{y}} f(\mathbf{x}^*, \mathbf{y}^*) \tag{5}$$

Before presenting our main results, we state the following well known result (see for example Nemirovski [2004]) which will be used later in the analysis of OGDA and EG. We present the proof here for completeness.

**Proposition 1.** *Recall the definition of the operator $F(\cdot)$ in (4) and the points $\hat{\mathbf{x}}_k, \hat{\mathbf{y}}_k$ in (3). Suppose Assumptions 1 and 3 hold. Then for any $\mathbf{z} = [\mathbf{x};\mathbf{y}] \in \mathbb{R}^{m+n}$, we have*

$$f(\hat{\mathbf{x}}_N, \mathbf{y}) - f(\mathbf{x}, \hat{\mathbf{y}}_N) \leq \frac{1}{N}\sum_{k=1}^{N} F(\mathbf{z}_k)^\top (\mathbf{z}_k - \mathbf{z}) \tag{6}$$

*Proof.* Using the definition of the operator $F$, we can write

$$\begin{aligned}
\frac{1}{N}\sum_{k=1}^{N} F(\mathbf{z}_k)^\top (\mathbf{z}_k - \mathbf{z}) &= \frac{1}{N}\sum_{k=1}^{N} [\nabla_{\mathbf{x}} f(\mathbf{x}_k,\mathbf{y}_k)^\top (\mathbf{x}_k - \mathbf{x}) + \nabla_{\mathbf{y}} f(\mathbf{x}_k,\mathbf{y}_k)^\top (\mathbf{y} - \mathbf{y}_k)] \\
&\geq \frac{1}{N}\sum_{k=1}^{N} [f(\mathbf{x}_k,\mathbf{y}_k) - f(\mathbf{x},\mathbf{y}_k) + f(\mathbf{x}_k,\mathbf{y}) - f(\mathbf{x}_k,\mathbf{y}_k)] \\
&= \frac{1}{N}\sum_{k=1}^{N} [f(\mathbf{x}_k,\mathbf{y}) - f(\mathbf{x},\mathbf{y}_k)],
\end{aligned} \tag{7}$$

where the inequality holds due to the fact that $f$ is convex-concave. Using convexity of $f$ with respect to $\mathbf{x}$ and concavity of $f$ with respect to $\mathbf{y}$, we have

$$\frac{1}{N}\sum_{k=1}^{N} f(\mathbf{x}_k, \mathbf{y}) \geq f(\hat{\mathbf{x}}_N, \mathbf{y}), \qquad \frac{1}{N}\sum_{k=1}^{N} f(\mathbf{x}, \mathbf{y}_k) \leq f(\mathbf{x}, \hat{\mathbf{y}}_N). \tag{8}$$

Combining inequalities (7) and (8) yields

$$\frac{1}{N}\sum_{k=1}^{N} F(\mathbf{z}_k)^\top (\mathbf{z}_k - \mathbf{z}) \geq f(\hat{\mathbf{x}}_N, \mathbf{y}) - f(\mathbf{x}, \hat{\mathbf{y}}_N),$$

completing the proof. □

# 3 Proximal point method with error

One of the classical algorithms studied for solving the saddle point problem in (1) is the Proximal Point (PP) method, introduced in Martinet [1970] and studied in Rockafellar [1976]. The PP method generates the iterate $\{\mathbf{x}_{k+1}, \mathbf{y}_{k+1}\}$ which is defined as the unique solution to the saddle point problem[4]

$$\min_{\mathbf{x}\in\mathbb{R}^m} \max_{\mathbf{y}\in\mathbb{R}^n} \left\{ f(\mathbf{x}, \mathbf{y}) + \frac{1}{2\eta}\|\mathbf{x} - \mathbf{x}_k\|^2 - \frac{1}{2\eta}\|\mathbf{y} - \mathbf{y}_k\|^2 \right\}. \tag{9}$$

It can be verified that if the pair $\{\mathbf{x}_{k+1}, \mathbf{y}_{k+1}\}$ is the solution of problem (9), then $\mathbf{x}_{k+1}$ and $\mathbf{y}_{k+1}$ satisfy

$$\mathbf{x}_{k+1} = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^m} \left\{ f(\mathbf{x}, \mathbf{y}_{k+1}) + \frac{1}{2\eta}\|\mathbf{x} - \mathbf{x}_k\|^2 \right\}, \tag{10}$$

$$\mathbf{y}_{k+1} = \operatorname*{argmax}_{\mathbf{y}\in\mathbb{R}^n} \left\{ f(\mathbf{x}_{k+1}, \mathbf{y}) - \frac{1}{2\eta}\|\mathbf{y} - \mathbf{y}_k\|^2 \right\}. \tag{11}$$

Using the optimality conditions of the updates in (10) and (11) (which are necessary and sufficient since the problems in (10) and (11) are strongly convex and strongly concave, respectively), the update of the PP method for the saddle point problem in (1) can be written as

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{x}_k - \eta \nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}), \\ \mathbf{y}_{k+1} &= \mathbf{y}_k + \eta \nabla_{\mathbf{y}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}). \end{aligned} \tag{12}$$

It is well-known that the proximal point method achieves a sublinear rate of $\mathcal{O}(1/k)$ when $k$ is the number of iterations for convex minimization and for solving monotone variational inequalities (see Güler [1991, 1992]; Bruck Jr [1977]; Teboulle [1997]; Nemirovski [2004]). Note that Nemirovski [2004] in fact analyzed the conceptual mirror prox (the proximal point method) as a building block to analyze the mirror-prox algorithm. For completeness, we present the convergence rate of the proximal point method for convex-concave saddle point problems in the following theorem (see Appendix A for the proof).

**Theorem 1.** *Suppose Assumption 1 holds. Let $\{\mathbf{x}_k, \mathbf{y}_k\}$ be the iterates generated by the updates in (12). Consider the definition of the averaged iterates $\hat{\mathbf{x}}_k, \hat{\mathbf{y}}_k$ in (3). Then for all $k \geq 1$, we have*

$$|f(\hat{\mathbf{x}}_k, \hat{\mathbf{y}}_k) - f(\mathbf{x}^*, \mathbf{y}^*)| \leq \frac{\|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \|\mathbf{y}_0 - \mathbf{y}^*\|^2}{\eta k}. \tag{13}$$

The result in Theorem 1 shows that by following the update of proximal point method the gap between the function value for the averaged iterates $(\hat{\mathbf{x}}_k, \hat{\mathbf{y}}_k)$ and the function value for a saddle point $(\mathbf{x}^*, \mathbf{y}^*)$ of the problem (1) approaches zero at a sublinear rate of $\mathcal{O}(1/k)$.

[4]Again $\{\mathbf{x}_{k+1}, \mathbf{y}_{k+1}\}$ is unique since the objective function of problem (9) is strongly convex in $\mathbf{x}$ and strongly concave in $\mathbf{y}$

Our goal is to provide similar convergence rate estimates for OGDA and EG using the fact that these two methods can be interpreted as approximate versions of the proximal point method. To do so, let us first rewrite the update of the proximal point method given in (12) as

$$\mathbf{z}_{k+1} = \mathbf{z}_k - \eta F(\mathbf{z}_{k+1}), \tag{14}$$

where $\mathbf{z} = [\mathbf{x}; \mathbf{y}] \in \mathbb{R}^{m+n}$ and the operator $F$ is defined in (4). In the following proposition, we establish a relation for the iterates of a proximal point method with error. This relation will be used later for our analysis of OGDA and EG methods.

**Proposition 2.** *Consider the sequence of iterates $\{\mathbf{z}_k\} \in \mathbb{R}^{n+m}$ generated by the following update*

$$\mathbf{z}_{k+1} = \mathbf{z}_k - \eta F(\mathbf{z}_{k+1}) + \boldsymbol{\varepsilon}_k, \tag{15}$$

*where $F : \mathbb{R}^{n+m} \to \mathbb{R}^{n+m}$ is a monotone and Lipschitz continuous operator, $\boldsymbol{\varepsilon}_k \in \mathbb{R}^{n+m}$ is an arbitrary vector, and $\eta$ is a positive constant. Then for any $\mathbf{z} \in \mathbb{R}^{n+m}$ and for each $k \geq 1$ we have*

$$\begin{aligned} &F(\mathbf{z}_{k+1})^\top (\mathbf{z}_{k+1} - \mathbf{z}) \\ &\quad = \frac{1}{2\eta}\|\mathbf{z}_k - \mathbf{z}\|^2 - \frac{1}{2\eta}\|\mathbf{z}_{k+1} - \mathbf{z}\|^2 - \frac{1}{2\eta}\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 + \frac{1}{\eta}{\boldsymbol{\varepsilon}_k}^\top (\mathbf{z}_{k+1} - \mathbf{z}). \end{aligned} \tag{16}$$

*Proof.* According to the update in (15), we can show that for any $\mathbf{z} \in \mathbb{R}^{m+n}$ we have

$$\begin{aligned} \|\mathbf{z}_{k+1} - \mathbf{z}\|^2 = \|\mathbf{z}_k - \mathbf{z}\|^2 - 2\eta(\mathbf{z}_k - \mathbf{z})^\top F(\mathbf{z}_{k+1}) + \eta^2\|F(\mathbf{z}_{k+1})\|^2 + \|\boldsymbol{\varepsilon}_k\|^2 \\ + 2{\boldsymbol{\varepsilon}_k}^\top (\mathbf{z}_k - \mathbf{z} - \eta F(\mathbf{z}_{k+1})). \end{aligned} \tag{17}$$

We add and subtract the inner product $2\eta\mathbf{z}_{k+1}^\top F(\mathbf{z}_{k+1})$ to the right hand side and regroup the terms to obtain

$$\begin{aligned} \|\mathbf{z}_{k+1} - \mathbf{z}\|^2 = \|\mathbf{z}_k - \mathbf{z}\|^2 - 2\eta(\mathbf{z}_{k+1} - \mathbf{z})^\top F(\mathbf{z}_{k+1}) - 2\eta(\mathbf{x}_k - \mathbf{x}_{k+1})^\top F(\mathbf{z}_{k+1}) \\ + \eta^2\|F(\mathbf{z}_{k+1})\|^2 + \|\boldsymbol{\varepsilon}_k\|^2 + 2{\boldsymbol{\varepsilon}_k}^T (\mathbf{z}_k - \mathbf{z} - \eta F(\mathbf{z}_{k+1})). \end{aligned} \tag{18}$$

Replacing $F(\mathbf{z}_{k+1})$ with $(1/\eta)(-\mathbf{z}_{k+1} + \mathbf{z}_k + \boldsymbol{\varepsilon}_k)$, we obtain

$$\begin{aligned} &\|\mathbf{z}_{k+1} - \mathbf{z}\|^2 \\ &= \|\mathbf{z}_k - \mathbf{z}\|^2 - 2\eta(\mathbf{z}_{k+1} - \mathbf{z})^\top F(\mathbf{z}_{k+1}) + 2(\mathbf{z}_k - \mathbf{z}_{k+1})^\top (\mathbf{z}_{k+1} - \mathbf{z}_k - \boldsymbol{\varepsilon}_k) \\ &\quad + \|\mathbf{z}_{k+1} - \mathbf{z}_k - \boldsymbol{\varepsilon}_k\|^2 + \|\boldsymbol{\varepsilon}_k\|^2 + 2{\boldsymbol{\varepsilon}_k}^T (\mathbf{z}_{k+1} - \mathbf{z} - \boldsymbol{\varepsilon}_k) \\ &= \|\mathbf{z}_k - \mathbf{z}\|^2 - 2\eta(\mathbf{z}_{k+1} - \mathbf{z})^\top F(\mathbf{z}_{k+1}) - \|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 + 2{\boldsymbol{\varepsilon}_k}^T (\mathbf{z}_{k+1} - \mathbf{z}). \end{aligned} \tag{19}$$

On rearranging the terms, we obtain the following inequality:

$$\begin{aligned} &F(\mathbf{z}_{k+1})^\top (\mathbf{z}_{k+1} - \mathbf{z}) \\ &\quad = \frac{1}{2\eta}\|\mathbf{z}_k - \mathbf{z}\|^2 - \frac{1}{2\eta}\|\mathbf{z}_{k+1} - \mathbf{z}\|^2 - \frac{1}{2\eta}\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 + \frac{1}{\eta}{\boldsymbol{\varepsilon}_k}^T (\mathbf{z}_{k+1} - \mathbf{z}), \end{aligned} \tag{20}$$

and the proof is complete. □

# 4 Optimistic Gradient Descent Ascent

In this section, we focus on analyzing the performance of optimistic gradient descent ascent (OGDA) for finding a saddle point of a general smooth convex-concave function. It has been shown that the OGDA method achieves the same iteration complexity as the proximal point method for both strongly convex-strongly concave and bilinear problems; see Liang & Stokes [2019], Gidel et al. [2018], Mokhtari et al. [2020]. However, its iteration complexity for a general smooth convex-concave case has not been established to the best of our knowledge. In this section, we show that the function value of the averaged iterate generated by the OGDA method converges to the function value at a saddle point at a rate of $\mathcal{O}(1/k)$, which matches the convergence rate of the proximal point method shown in Theorem 1.

Given a stepsize $\eta > 0$, the OGDA method updates the iterates $\mathbf{x}_k$ and $\mathbf{y}_k$ for each $k \geq 0$ as

$$\begin{aligned}\mathbf{x}_{k+1} &= \mathbf{x}_k - 2\eta \nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k) + \eta \nabla_{\mathbf{x}} f(\mathbf{x}_{k-1}, \mathbf{y}_{k-1}), \\ \mathbf{y}_{k+1} &= \mathbf{y}_k + 2\eta \nabla_{\mathbf{y}} f(\mathbf{x}_k, \mathbf{y}_k) - \eta \nabla_{\mathbf{y}} f(\mathbf{x}_{k-1}, \mathbf{y}_{k-1})\end{aligned} \tag{21}$$

with the initial conditions $\mathbf{x}_0 = \mathbf{x}_{-1}$ and $\mathbf{y}_0 = \mathbf{y}_{-1}$. The main difference between the updates of OGDA in (21) and the gradient descent ascent (GDA) method is in the additional "momentum" terms $-\eta(\nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k) - \nabla_{\mathbf{x}} f(\mathbf{x}_{k-1}, \mathbf{y}_{k-1}))$ and $\eta(\nabla_{\mathbf{y}} f(\mathbf{x}_k, \mathbf{y}_k) - \nabla_{\mathbf{y}} f(\mathbf{x}_{k-1}, \mathbf{y}_{k-1}))$. This additional term makes the update of OGDA a better approximation to the update of the proximal point method compared to the update of the GDA; for more details we refer readers to Proposition 1 in Mokhtari et al. [2020].

To establish the convergence rate of OGDA for convex-concave problems, we first illustrate the connection between the updates of proximal point method and OGDA. Note that using the definitions of the vector $\mathbf{z} = [\mathbf{x}; \mathbf{y}] \in \mathbb{R}^{n+m}$ and the operator $F(\mathbf{z}) = [\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}); -\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})] \in \mathbb{R}^{n+m}$, we can rewrite the update of the OGDA algorithm at iteration $k$ as

$$\mathbf{z}_{k+1} = \mathbf{z}_k - 2\eta F(\mathbf{z}_k) + \eta F(\mathbf{z}_{k-1}). \tag{22}$$

Considering this expression, we can also write the update of OGDA as an approximation of the proximal point update, i.e.,

$$\mathbf{z}_{k+1} = \mathbf{z}_k - \eta F(\mathbf{z}_{k+1}) + \boldsymbol{\varepsilon}_k, \tag{23}$$

where the error vector $\boldsymbol{\varepsilon}_k$ is given by

$$\boldsymbol{\varepsilon}_k = \eta[(F(\mathbf{z}_{k+1}) - F(\mathbf{z}_k)) - (F(\mathbf{z}_k) - F(\mathbf{z}_{k-1}))]. \tag{24}$$

To derive the convergence rate of OGDA for the unconstrained problem in (1), we first use the result in Proposition 2 to derive a result for the specific case of OGDA updates. We then show that the iterates generated by the OGDA method remain in a bounded set. This is done in the following lemma (Note that boundedness of OGDA iterates can be deduced from [Popov, 1980], whereas a result similar to Lemma 2(b) was shown in a recent independent paper by Malitsky & Tam [2018]).

**Lemma 2.** *Let $\{\mathbf{z}_k\}$ be the iterates generated by the optimistic gradient descent ascent (OGDA) method introduced in (22) with the initial conditions $\mathbf{x}_0 = \mathbf{x}_{-1}$ and $\mathbf{y}_0 = \mathbf{y}_{-1}$ (i.e. $\mathbf{z}_0 = \mathbf{z}_{-1}$). If Assumptions 1, 2, and 3 hold and the stepsize $\eta$ satisfies the condition $0 < \eta \leq \frac{1}{2L}$, then:*
*(a) The iterates $\{\mathbf{z}_k\}$ satisfy the following relation:*

$$\begin{aligned}&F(\mathbf{z}_{k+1})^\top (\mathbf{z}_{k+1} - \mathbf{z}) \\ &\leq \frac{1}{2\eta}\|\mathbf{z}_k - \mathbf{z}\|^2 - \frac{1}{2\eta}\|\mathbf{z}_{k+1} - \mathbf{z}\|^2 - \frac{L}{2}\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 + \frac{L}{2}\|\mathbf{z}_k - \mathbf{z}_{k-1}\|^2 \\ &\quad + (F(\mathbf{z}_{k+1}) - F(\mathbf{z}_k))^\top (\mathbf{z}_{k+1} - \mathbf{z}) - (F(\mathbf{z}_k) - F(\mathbf{z}_{k-1}))^\top (\mathbf{z}_k - \mathbf{z}).\end{aligned} \tag{25}$$

*(b) The iterates $\{\mathbf{z}_k\}$ stay within the compact set $\mathcal{D}$ defined as*

$$\mathcal{D} := \{(\mathbf{x}, \mathbf{y}) \mid \|\mathbf{x} - \mathbf{x}^*\|^2 + \|\mathbf{y} - \mathbf{y}^*\|^2 \leq 2\left(\|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \|\mathbf{y}_0 - \mathbf{y}^*\|^2\right)\}, \tag{26}$$

*where $(\mathbf{x}^*, \mathbf{y}^*) = \mathbf{z}^* \in \mathcal{Z}^*$ is a saddle point of the problem defined in (1).*

*Proof.* Since OGDA iterates satisfy Equation (23) with the error vector $\boldsymbol{\varepsilon}_k$ given in Equation (24), using Proposition 2 with this error vector $\boldsymbol{\varepsilon}_k$ leads to

$$\begin{aligned}&F(\mathbf{z}_{k+1})^\top (\mathbf{z}_{k+1} - \mathbf{z}) \\ &= \frac{1}{2\eta}\|\mathbf{z}_k - \mathbf{z}\|^2 - \frac{1}{2\eta}\|\mathbf{z}_{k+1} - \mathbf{z}\|^2 - \frac{1}{2\eta}\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 \\ &\quad + (F(\mathbf{z}_{k+1}) - F(\mathbf{z}_k))^\top (\mathbf{z}_{k+1} - \mathbf{z}) - (F(\mathbf{z}_k) - F(\mathbf{z}_{k-1}))^\top (\mathbf{z}_{k+1} - \mathbf{z}).\end{aligned} \tag{27}$$

We add and subtract the inner product $(F(\mathbf{z}_k) - F(\mathbf{z}_{k-1}))^\top (\mathbf{z}_k - \mathbf{z})$ to the right hand side of the preceding relation to obtain

$$\begin{aligned}&F(\mathbf{z}_{k+1})^\top (\mathbf{z}_{k+1} - \mathbf{z}) \\ &= \frac{1}{2\eta}\|\mathbf{z}_k - \mathbf{z}\|^2 - \frac{1}{2\eta}\|\mathbf{z}_{k+1} - \mathbf{z}\|^2 - \frac{1}{2\eta}\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 \\ &\quad + (F(\mathbf{z}_{k+1}) - F(\mathbf{z}_k))^\top (\mathbf{z}_{k+1} - \mathbf{z}) - (F(\mathbf{z}_k) - F(\mathbf{z}_{k-1}))^\top (\mathbf{z}_k - \mathbf{z}) \\ &\quad + (F(\mathbf{z}_k) - F(\mathbf{z}_{k-1}))^\top (\mathbf{z}_k - \mathbf{z}_{k+1}).\end{aligned} \tag{28}$$

Note that $(F(\mathbf{z}_k) - F(\mathbf{z}_{k-1}))^\top(\mathbf{z}_k - \mathbf{z}_{k+1})$ can be upper bounded by

$$\begin{aligned}(F(\mathbf{z}_k) - F(\mathbf{z}_{k-1}))^\top(\mathbf{z}_k - \mathbf{z}_{k+1}) &\leq \|F(\mathbf{z}_k) - F(\mathbf{z}_{k-1})\|\|\mathbf{z}_k - \mathbf{z}_{k+1}\| \\ &\leq L\|\mathbf{z}_k - \mathbf{z}_{k-1}\|\|\mathbf{z}_k - \mathbf{z}_{k+1}\| \\ &\leq \frac{L}{2}\|\mathbf{z}_k - \mathbf{z}_{k-1}\|^2 + \frac{L}{2}\|\mathbf{z}_k - \mathbf{z}_{k+1}\|^2, \end{aligned} \tag{29}$$

where the second inequality holds due to Lipschitz continuity of the operator $F$ (Lemma 1(b)) and the last inequality holds due to Young's inequality.[5] Replacing $(F(\mathbf{z}_k) - F(\mathbf{z}_{k-1}))^\top(\mathbf{z}_k - \mathbf{z}_{k+1})$ in (28) by its upper bound in (29) yields

$$\begin{aligned}&F(\mathbf{z}_{k+1})^\top(\mathbf{z}_{k+1} - \mathbf{z}) \\ &\leq \frac{1}{2\eta}\|\mathbf{z}_k - \mathbf{z}\|^2 - \frac{1}{2\eta}\|\mathbf{z}_{k+1} - \mathbf{z}\|^2 - \frac{1}{2\eta}\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 \\ &\quad + (F(\mathbf{z}_{k+1}) - F(\mathbf{z}_k))^\top(\mathbf{z}_{k+1} - \mathbf{z}) - (F(\mathbf{z}_k) - F(\mathbf{z}_{k-1}))^\top(\mathbf{z}_k - \mathbf{z}) \\ &\quad + \frac{L}{2}\|\mathbf{z}_k - \mathbf{z}_{k-1}\|^2 + \frac{L}{2}\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 \\ &\leq \frac{1}{2\eta}\|\mathbf{z}_k - \mathbf{z}\|^2 - \frac{1}{2\eta}\|\mathbf{z}_{k+1} - \mathbf{z}\|^2 - \frac{L}{2}\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 + \frac{L}{2}\|\mathbf{z}_k - \mathbf{z}_{k-1}\|^2 \\ &\quad + (F(\mathbf{z}_{k+1}) - F(\mathbf{z}_k))^\top(\mathbf{z}_{k+1} - \mathbf{z}) - (F(\mathbf{z}_k) - F(\mathbf{z}_{k-1}))^\top(\mathbf{z}_k - \mathbf{z}), \end{aligned} \tag{30}$$

where the second inequality follows as $\eta \leq 1/2L$ and therefore $-\frac{1}{2\eta}\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 \leq -L\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2$. This completes the proof of Part (a) of the lemma. Now, taking the sum of the preceding relation from $k = 0, \cdots, N-1$, we obtain

$$\begin{aligned}&\sum_{k=0}^{N-1} F(\mathbf{z}_{k+1})^\top(\mathbf{z}_{k+1} - \mathbf{z}) \\ &\leq \frac{1}{2\eta}\|\mathbf{z}_0 - \mathbf{z}\|^2 - \frac{1}{2\eta}\|\mathbf{z}_N - \mathbf{z}\|^2 - \frac{L}{2}\|\mathbf{z}_N - \mathbf{z}_{N-1}\|^2 + \frac{L}{2}\|\mathbf{z}_0 - \mathbf{z}_{-1}\|^2 \\ &\quad + (F(\mathbf{z}_N) - F(\mathbf{z}_{N-1}))^\top(\mathbf{z}_N - \mathbf{z}) - (F(\mathbf{z}_0) - F(\mathbf{z}_{-1}))^\top(\mathbf{z}_0 - \mathbf{z}). \end{aligned} \tag{31}$$

Now set $\mathbf{z} = \mathbf{z}^*$, where $\mathbf{z}^* \in \mathcal{Z}^*$, to obtain

$$\begin{aligned}&\sum_{k=0}^{N-1} F(\mathbf{z}_{k+1})^\top(\mathbf{z}_{k+1} - \mathbf{z}^*) \\ &\leq \frac{1}{2\eta}\|\mathbf{z}_0 - \mathbf{z}^*\|^2 - \frac{1}{2\eta}\|\mathbf{z}_N - \mathbf{z}^*\|^2 - \frac{L}{2}\|\mathbf{z}_N - \mathbf{z}_{N-1}\|^2 + \frac{L}{2}\|\mathbf{z}_0 - \mathbf{z}_{-1}\|^2 \\ &\quad + (F(\mathbf{z}_N) - F(\mathbf{z}_{N-1}))^\top(\mathbf{z}_N - \mathbf{z}^*) - (F(\mathbf{z}_0) - F(\mathbf{z}_{-1}))^\top(\mathbf{z}_0 - \mathbf{z}^*). \end{aligned} \tag{32}$$

Note that each term of the summand in the sum in the left is nonnegative due to monotonicity of $F$ and therefore the sum is also nonnegative. Further, we know that $\mathbf{z}_0 = \mathbf{z}_{-1}$. Using these observations we can write

$$\begin{aligned}0 \leq \frac{1}{2\eta}\|\mathbf{z}_0 - \mathbf{z}^*\|^2 - \frac{1}{2\eta}\|\mathbf{z}_N - \mathbf{z}^*\|^2 - \frac{L}{2}\|\mathbf{z}_N - \mathbf{z}_{N-1}\|^2 \\ + (F(\mathbf{z}_N) - F(\mathbf{z}_{N-1}))^\top(\mathbf{z}_N - \mathbf{z}^*). \end{aligned} \tag{33}$$

Using Lipschitz continuity of the operator $F(\cdot)$ (Lemma 1(b)) and Young's inequality in the pre-

[5] We use the following form of Young's inequality throughout the paper:

$$\mathbf{a}^\top\mathbf{b} \leq \frac{\|\mathbf{a}\|^2}{2} + \frac{\|\mathbf{b}\|^2}{2}$$

ceding relation, we have

$$\begin{aligned}0 &\leq \frac{1}{2\eta}\|\mathbf{z}_0 - \mathbf{z}^*\|^2 - \frac{1}{2\eta}\|\mathbf{z}_N - \mathbf{z}^*\|^2 - \frac{L}{2}\|\mathbf{z}_N - \mathbf{z}_{N-1}\|^2 \\ &\quad + L\|\mathbf{z}_N - \mathbf{z}_{N-1}\|\|\mathbf{z}_N - \mathbf{z}^*\| \\ &\leq \frac{1}{2\eta}\|\mathbf{z}_0 - \mathbf{z}^*\|^2 - \frac{1}{2\eta}\|\mathbf{z}_N - \mathbf{z}^*\|^2 - \frac{L}{2}\|\mathbf{z}_N - \mathbf{z}_{N-1}\|^2 \\ &\quad + \frac{L}{2}\|\mathbf{z}_N - \mathbf{z}_{N-1}\|^2 + \frac{L}{2}\|\mathbf{z}_N - \mathbf{z}^*\|^2 \\ &\leq \frac{1}{2\eta}\|\mathbf{z}_0 - \mathbf{z}^*\|^2 - \frac{1}{2\eta}\|\mathbf{z}_N - \mathbf{z}^*\|^2 + \frac{L}{2}\|\mathbf{z}_N - \mathbf{z}^*\|^2 \end{aligned} \tag{34}$$

Regrouping the terms gives us

$$\|\mathbf{z}_N - \mathbf{z}^*\|^2 \leq \frac{1}{(1-\eta L)}\|\mathbf{z}_0 - \mathbf{z}^*\|^2. \tag{35}$$

Using the condition $\eta \leq 1/2L$, it follows that for any iterate $N$ we have

$$\|\mathbf{z}_N - \mathbf{z}^*\|^2 \leq 2\|\mathbf{z}_0 - \mathbf{z}^*\|^2, \tag{36}$$

and the claim in Part (b) follows. □

According to Lemma 2, the sequence of iterates $\{\mathbf{x}_k, \mathbf{y}_k\}$ generated by OGDA method stays within a closed and bounded convex set. We use this result to prove a sublinear convergence rate of $\mathcal{O}(1/k)$ for the function value of the averaged iterates generated by OGDA to the function value at a saddle point, for smooth and convex-concave functions in the following theorem.

**Theorem 2.** *Suppose Assumptions 1, 2 and 3 hold. Let $\{\mathbf{x}_k, \mathbf{y}_k\}$ be the iterates generated by the OGDA updates in (21). Let the initial conditions satisfy $\mathbf{x}_0 = \mathbf{x}_{-1}$ and $\mathbf{y}_0 = \mathbf{y}_{-1}$. Consider the definition of the averaged iterates $\hat{\mathbf{x}}_N, \hat{\mathbf{y}}_N$ in (3) and the compact convex set $\mathcal{D}$ in (26). If the stepsize $\eta$ satisfies the condition $0 < \eta \leq 1/2L$, then for all $N \geq 1$, we have*

$$\left[\max_{\mathbf{y}:(\hat{\mathbf{x}}_N, \mathbf{y}) \in \mathcal{D}} f(\hat{\mathbf{x}}_N, \mathbf{y}) - f^\star\right] + \left[f^\star - \min_{\mathbf{x}:(\mathbf{x}, \hat{\mathbf{y}}_N) \in \mathcal{D}} f(\mathbf{x}, \hat{\mathbf{y}}_N)\right] \leq \frac{D(8L + \frac{1}{2\eta})}{N}, \tag{37}$$

*where $f^\star = f(\mathbf{x}^*, \mathbf{y}^*)$ and $D = \|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \|\mathbf{y}_0 - \mathbf{y}^*\|^2$.*

*Proof.* From Lemma 2(a), we have that the iterates generated by the OGDA method satisfy Equation (25). On taking the sum of this relation from $k = 0, \cdots, N-1$, we obtain for any $\mathbf{z}$

$$\begin{aligned}&\sum_{k=0}^{N-1} F(\mathbf{z}_{k+1})^\top (\mathbf{z}_{k+1} - \mathbf{z}) \\ &\leq \frac{1}{2\eta}\|\mathbf{z}_0 - \mathbf{z}\|^2 - \frac{1}{2\eta}\|\mathbf{z}_N - \mathbf{z}\|^2 - \frac{L}{2}\|\mathbf{z}_N - \mathbf{z}_{N-1}\|^2 + \frac{L}{2}\|\mathbf{z}_0 - \mathbf{z}_{-1}\|^2 \\ &\quad + (F(\mathbf{z}_N) - F(\mathbf{z}_{N-1}))^\top (\mathbf{z}_N - \mathbf{z}) - (F(\mathbf{z}_0) - F(\mathbf{z}_{-1}))^\top (\mathbf{z}_0 - \mathbf{z}). \end{aligned} \tag{38}$$

Note that for any $\mathbf{z}_1, \mathbf{z}_2 \in \mathcal{D}$, we have:

$$\begin{aligned}\|\mathbf{z}_1 - \mathbf{z}_2\|^2 &\leq 2\|\mathbf{z}_1 - \mathbf{z}^*\|^2 + 2\|\mathbf{z}_2 - \mathbf{z}^*\|^2 \\ &\leq 4\|\mathbf{z}_0 - \mathbf{z}^*\|^2 + 4\|\mathbf{z}_0 - \mathbf{z}^*\|^2 \\ &\leq 8D \end{aligned} \tag{39}$$

where we have used the fact that $\|\mathbf{z} - \mathbf{z}^*\|^2 \leq 2\|\mathbf{z}_0 - \mathbf{z}^*\|^2$ for all $\mathbf{z} \in \mathcal{D}$ along with the fact that for all $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, $\|\mathbf{a} + \mathbf{b}\|^2 \leq 2\|\mathbf{a}\|^2 + 2\|\mathbf{b}\|^2$. As $\mathbf{z}_{-1} = \mathbf{z}_0$ and $\eta \leq 1/2L$, for any $\mathbf{z} \in \mathcal{D}$ we have

$$\begin{aligned}\frac{1}{N}\sum_{k=0}^{N-1} F(\mathbf{z}_{k+1})^\top (\mathbf{z}_{k+1} - \mathbf{z}) &\leq \frac{\frac{1}{2\eta}\|\mathbf{z}_0 - \mathbf{z}\|^2 + (F(\mathbf{z}_N) - F(\mathbf{z}_{N-1}))^\top (\mathbf{z}_N - \mathbf{z})}{N} \\ &\leq \frac{D(8L + \frac{1}{2\eta})}{N}. \end{aligned} \tag{40}$$

This inequality follows since:

$$\begin{aligned}(F(\mathbf{z}_N) - F(\mathbf{z}_{N-1}))^\top (\mathbf{z}_N - \mathbf{z}) &\leq \|F(\mathbf{z}_N) - F(\mathbf{z}_{N-1})\| \|\mathbf{z}_N - \mathbf{z}\| \\ &\leq L\|\mathbf{z}_N - \mathbf{z}_{N-1}\| \|\mathbf{z}_N - \mathbf{z}\| \end{aligned} \tag{41}$$

and for any $\mathbf{x}, \mathbf{y} \in \mathcal{D}$, we have:

$$\begin{aligned}\|\mathbf{x} - \mathbf{y}\| &\leq \|\mathbf{x} - \mathbf{z}^*\| + \|\mathbf{y} - \mathbf{z}^*\| \\ &\leq 2\sqrt{2D} \end{aligned} \tag{42}$$

Therefore, we have:

$$(F(\mathbf{z}_N) - F(\mathbf{z}_{N-1}))^\top (\mathbf{z}_N - \mathbf{z}) \leq 8LD \tag{43}$$

which immediately gives us Inequality (40). Combining relation (40) with Proposition 1 we have that for all $\mathbf{x}, \mathbf{y} \in \mathcal{D}$

$$f(\hat{\mathbf{x}}_N, \mathbf{y}) - f(\mathbf{x}, \hat{\mathbf{y}}_N) \leq \frac{D(8L + \frac{1}{2\eta})}{N}. \tag{44}$$

which gives us the following convergence rate estimate:

$$\left[\max_{\mathbf{y}:(\hat{\mathbf{x}}_N, \mathbf{y}) \in \mathcal{D}} f(\hat{\mathbf{x}}_N, \mathbf{y}) - f^\star\right] + \left[f^\star - \min_{\mathbf{x}:(\mathbf{x}, \hat{\mathbf{y}}_N) \in \mathcal{D}} f(\mathbf{x}, \hat{\mathbf{y}}_N)\right] \leq \frac{D(8L + \frac{1}{2\eta})}{N},$$

where $f^\star = f(\mathbf{x}^*, \mathbf{y}^*)$. □

Note that convergence in Theorem 2 is shown in terms of the Primal-Dual gap $\max_{\mathbf{y}:(\hat{\mathbf{x}}_N, \mathbf{y}) \in \mathcal{D}} f(\hat{\mathbf{x}}_N, \mathbf{y}) - \min_{\mathbf{x}:(\mathbf{x}, \hat{\mathbf{y}}_N) \in \mathcal{D}} f(\mathbf{x}, \hat{\mathbf{y}}_N)$ which is a common measure to capture closeness to the solution in convex-concave setting (see [Nemirovski, 2004]). Indeed, the duality gap is zero if and only if $(\hat{\mathbf{x}}_N, \hat{\mathbf{y}}_N)$ is a saddle point of the problem. The primal-dual gap also has the following game theoretic interpretation. If Player $\mathbf{x}$ is playing $\hat{\mathbf{x}}_N$, then $\max_{\mathbf{y}:(\hat{\mathbf{x}}_N, \mathbf{y}) \in \mathcal{D}} f(\hat{\mathbf{x}}_N, \mathbf{y})$ quantifies how much Player $\mathbf{y}$ can gain by playing an action in the set $\mathcal{D}$. Similarly, if Player $\mathbf{y}$ is playing $\hat{\mathbf{y}}_N$, then $-\min_{\mathbf{x}:(\mathbf{x}, \hat{\mathbf{y}}_N) \in \mathcal{D}} f(\mathbf{x}, \hat{\mathbf{y}}_N)$ quantifies how much Player $\mathbf{x}$ can gain by playing an action in $\mathcal{D}$. Therefore, the quantity $\max_{\mathbf{y}:(\hat{\mathbf{x}}_N, \mathbf{y}) \in \mathcal{D}} f(\hat{\mathbf{x}}_N, \mathbf{y}) - \min_{\mathbf{x}:(\mathbf{x}, \hat{\mathbf{y}}_N) \in \mathcal{D}} f(\mathbf{x}, \hat{\mathbf{y}}_N)$ is a measure of the sum of how much each player can gain if they unilaterally deviate from the strategy $(\hat{\mathbf{x}}_N, \hat{\mathbf{y}}_N)$. This goes to zero at the Nash Equilibrium (saddle point), where no player can gain by unilaterally deviating from the equilibrium strategy.

Also, note that the result in Theorem 2 also implies that $|f(\hat{\mathbf{x}}_N, \hat{\mathbf{y}}_N) - f^*| \leq 9LD/N$ as we show in the following corollary.

**Corollary 1.** *Suppose Assumptions 1, 2 and 3 hold. Let $\{\mathbf{x}_k, \mathbf{y}_k\}$ be the iterates generated by the OGDA updates in* (21)*. Consider the definition of the averaged iterates $\hat{\mathbf{x}}_N, \hat{\mathbf{y}}_N$ in* (3)*. If the stepsize $\eta$ satisfies the condition $0 < \eta \leq 1/2L$, then for all $N \geq 1$, we have*

$$|f(\hat{\mathbf{x}}_N, \hat{\mathbf{y}}_N) - f^\star| \leq \frac{D(8L + \frac{1}{2\eta})}{N},$$

*where $f^\star = f(\mathbf{x}^*, \mathbf{y}^*)$.*

*Proof.* Note that $[\max_{\mathbf{y}:(\hat{\mathbf{x}}_N, \mathbf{y}) \in \mathcal{D}} f(\hat{\mathbf{x}}_N, \mathbf{y}) - f^\star]$ and $[f^\star - \min_{\mathbf{x}:(\mathbf{x}, \hat{\mathbf{y}}_N) \in \mathcal{D}} f(\mathbf{x}, \hat{\mathbf{y}}_N)]$ are both non-negative. To verify note that

$$\max_{\mathbf{y}:(\hat{\mathbf{x}}_N, \mathbf{y}) \in \mathcal{D}} f(\hat{\mathbf{x}}_N, \mathbf{y}) \geq f(\hat{\mathbf{x}}_N, \mathbf{y}^*) \geq f(\mathbf{x}^*, \mathbf{y}^*)$$

and

$$\min_{\mathbf{x}:(\mathbf{x}, \hat{\mathbf{y}}_N) \in \mathcal{D}} f(\mathbf{x}, \hat{\mathbf{y}}_N) \leq f(\mathbf{x}^*, \hat{\mathbf{y}}_N) \leq f(\mathbf{x}^*, \mathbf{y}^*)$$

(since $(\mathbf{x}^*, \mathbf{y}^*) \in \mathcal{D}$). Further, note that $(\hat{\mathbf{x}}_N, \hat{\mathbf{y}}_N)$ belongs to the set $\mathcal{D}$. Hence, it yields

$$f(\hat{\mathbf{x}}_N, \hat{\mathbf{y}}_N) - f^\star \leq \max_{\mathbf{y}:(\hat{\mathbf{x}}_N, \mathbf{y}) \in \mathcal{D}} f(\hat{\mathbf{x}}_N, \mathbf{y}) - f^\star \leq \frac{D(8L + \frac{1}{2\eta})}{N}.$$

Also, we can show that

$$f^\star - f(\hat{\mathbf{x}}_N, \hat{\mathbf{y}}_N) \leq f^\star - \min_{\mathbf{x}:(\mathbf{x},\hat{\mathbf{y}}_N)\in\mathcal{D}} f(\mathbf{x}, \hat{\mathbf{y}}_N) \leq \frac{D(8L + \frac{1}{2\eta})}{N}.$$

Therefore, $|f(\hat{\mathbf{x}}_N, \hat{\mathbf{y}}_N) - f^\star| \leq \frac{D(8L+\frac{1}{2\eta})}{N}$. □

The result in Corollary 1 shows that the function value of the averaged iterates generated by OGDA converges to the function value at a saddle point of problem (1) at a sublinear rate of $\mathcal{O}(1/k)$ when the function is smooth and convex-concave. To the best of our knowledge, this is the first non-asymptotic complexity bound for OGDA for the convex-concave setting. Moreover, note that without computing any extra gradient evaluation, i.e., computing only one gradient per iteration with respect to $\mathbf{x}$ and $\mathbf{y}$, OGDA recovers the convergence rate of proximal point method.

# 5 Extragradient Method

In this section, we consider finding a saddle point of a general smooth convex-concave function using the Extra-gradient (EG) method. Similar to our analysis of the OGDA method, we show that by interpreting the EG method as an approximation of the proximal point method it is possible to establish a convergence rate of $\mathcal{O}(1/k)$ through a simple analysis.

Consider the update of EG in which we first compute a set of mid-point iterates $\{\mathbf{x}_{k+\frac{1}{2}}, \mathbf{y}_{k+\frac{1}{2}}\}$ using the gradients with respect to $\mathbf{x}$ and $\mathbf{y}$ at the current iterate

$$\begin{aligned}\mathbf{x}_{k+\frac{1}{2}} &= \mathbf{x}_k - \eta\nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k),\\ \mathbf{y}_{k+\frac{1}{2}} &= \mathbf{y}_k + \eta\nabla_{\mathbf{y}} f(\mathbf{x}_k, \mathbf{y}_k).\end{aligned} \tag{45}$$

Then, we compute the next iterates of the EG method $\{\mathbf{x}_{k+1}, \mathbf{y}_{k+1}\}$ using the gradients at the mid-points $\{\mathbf{x}_{k+\frac{1}{2}}, \mathbf{y}_{k+\frac{1}{2}}\}$, i.e.,

$$\begin{aligned}\mathbf{x}_{k+1} &= \mathbf{x}_k - \eta\nabla_{\mathbf{x}} f(\mathbf{x}_{k+\frac{1}{2}}, \mathbf{y}_{k+\frac{1}{2}}),\\ \mathbf{y}_{k+1} &= \mathbf{y}_k + \eta\nabla_{\mathbf{y}} f(\mathbf{x}_{k+\frac{1}{2}}, \mathbf{y}_{k+\frac{1}{2}}).\end{aligned} \tag{46}$$

We aim to show that EG, similar to OGDA, can be analyzed for convex-concave problems by considering it as an approximation of the proximal point. To do so, let us use the notation $\mathbf{z} = [\mathbf{x}; \mathbf{y}] \in \mathbb{R}^{n+m}$ and $F(\mathbf{z}) = [\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}); -\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})] \in \mathbb{R}^{n+m}$ to write the update of EG as

$$\begin{aligned}\mathbf{z}_{k+\frac{1}{2}} &= \mathbf{z}_k - \eta F(\mathbf{z}_k),\\ \mathbf{z}_{k+1} &= \mathbf{z}_k - \eta F(\mathbf{z}_{k+\frac{1}{2}}).\end{aligned} \tag{47}$$

To better highlight the connection between proximal point and EG, let us focus on the expression for the update of the mid-point iterates in EG. Considering the updates in (47), we have

$$\begin{aligned}\mathbf{z}_{k+\frac{1}{2}} &= \mathbf{z}_k - \eta F(\mathbf{z}_k),\\ &= \mathbf{z}_{k-1} - \eta F(\mathbf{z}_{k-\frac{1}{2}}) - \eta F(\mathbf{z}_k)\\ &= \mathbf{z}_{k-\frac{1}{2}} + \eta F(\mathbf{z}_{k-1}) - \eta F(\mathbf{z}_{k-\frac{1}{2}}) - \eta F(\mathbf{z}_k)\end{aligned}$$

where the second equality follows by replacing $\mathbf{z}_k$ by its update $\mathbf{z}_{k-1} - \eta F(\mathbf{z}_{k-\frac{1}{2}})$, and the second equality follows by considering the update $\mathbf{z}_{k-\frac{1}{2}} = \mathbf{z}_{k-1} - \eta F(\mathbf{z}_{k-1})$. Therefore, rearranging this equation, we can rewrite the updates as

$$\mathbf{z}_{k+\frac{1}{2}} = \mathbf{z}_{k-\frac{1}{2}} - \eta F(\mathbf{z}_{k-\frac{1}{2}}) - \eta(F(\mathbf{z}_k) - F(\mathbf{z}_{k-1})). \tag{48}$$

One can consider the expression $F(\mathbf{z}_k) - F(\mathbf{z}_{k-1})$ as an approximation of the variation $F(\mathbf{z}_{k+\frac{1}{2}}) - F(\mathbf{z}_{k-\frac{1}{2}})$. To be more precise, if we assume that the variations $F(\mathbf{z}_k) - F(\mathbf{z}_{k-1})$ and $F(\mathbf{z}_{k+\frac{1}{2}}) - F(\mathbf{z}_{k-\frac{1}{2}})$ are close to each other, i.e., $F(\mathbf{z}_{k+\frac{1}{2}}) - F(\mathbf{z}_{k-\frac{1}{2}}) \approx F(\mathbf{z}_k) - F(\mathbf{z}_{k-1})$, then the update in (48) behaves like the proximal point update with. respect to the mid-point iterates, i.e.,

$$\mathbf{z}_{k+\frac{1}{2}} = \mathbf{z}_{k-\frac{1}{2}} - \eta F(\mathbf{z}_{k+\frac{1}{2}}). \tag{49}$$

We first derive a result similar to Proposition 2 for the specific case of EG iterates (Lemma 3(a)). We then show the the boundedness of the EG iterates in Lemma 3(b) (Note that the boundedness of the EG updates can also be deduced from the convergence results of Korpelevich [1976] and Monteiro & Svaiter [2010]).

**Lemma 3.** *Let $\{\mathbf{z}_k\}, \{\mathbf{z}_{k+\frac{1}{2}}\}$ be the iterates generated by the extra-gradient (EG) method introduced in (47). If Assumptions 1, 2 and 3 hold and the stepsize $\eta$ satisfies the condition $0 < \eta < 1/L$, then:*
*(a) The iterates $\{\mathbf{z}_k\}, \{\mathbf{z}_{k+\frac{1}{2}}\}$ satisfy the following relation:*

$$\begin{aligned} &F(\mathbf{z}_{k+\frac{1}{2}})^\top(\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}) \\ &\quad\leq \frac{1}{2\eta}\|\mathbf{z}_{k-\frac{1}{2}} - \mathbf{z}\|^2 - \frac{1}{2\eta}\|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}\|^2 + \frac{L}{2}\|\mathbf{z}_{k-\frac{1}{2}} - \mathbf{z}_{k-1}\|^2 \\ &\qquad + (F(\mathbf{z}_{k+\frac{1}{2}}) - F(\mathbf{z}_k))^\top(\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}) - (F(\mathbf{z}_{k-\frac{1}{2}}) - F(\mathbf{z}_{k-1}))^\top(\mathbf{z}_{k-\frac{1}{2}} - \mathbf{z}). \end{aligned} \tag{50}$$

*(b) The iterates $\{\mathbf{z}_k\}, \{\mathbf{z}_{k+\frac{1}{2}}\}$ stay within the compact set $\mathcal{D}$ defined as*

$$\mathcal{D} := \{(\mathbf{x}, \mathbf{y}) \mid \|\mathbf{x} - \mathbf{x}^*\|^2 + \|\mathbf{y} - \mathbf{y}^*\|^2 \leq \left(2 + \frac{2}{1 - \eta^2 L^2}\right)(\|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \|\mathbf{y}_0 - \mathbf{y}^*\|^2)\}, \tag{51}$$

*where $(\mathbf{x}^*, \mathbf{y}^*) = \mathbf{z}^* \in \mathcal{Z}^*$ is a saddle point of the problem defined in (1). Moreover, the sum $\sum_{k=0}^{\infty} \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2$ is bounded above by*

$$\sum_{k=0}^{\infty} \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2 \leq \frac{\|\mathbf{z}_0 - \mathbf{z}^*\|^2}{1 - \eta^2 L^2}. \tag{52}$$

*Proof.* See Appendix B for the proof. □

The result in Lemma 3 shows that the iterates generated by the update of EG belong to a bounded and closed set. Now we use this result to show that the function value of the averaged iterates converges at a sublinear rate of $\mathcal{O}(1/k)$ to the function value at a saddle point for the EG method in the following theorem.

**Theorem 3.** *Suppose Assumptions 1, 2 and 3 hold. Let $\{\mathbf{x}_{k+1/2}, \mathbf{y}_{k+1/2}\}$ be the iterates generated by the EG updates in (45)-(46). Let the initial conditions satisfy $\mathbf{x}_0 = \mathbf{x}_{-1/2}$ and $\mathbf{y}_0 = \mathbf{y}_{-1/2}$ Consider the definition of the averaged iterates $\hat{\mathbf{x}}_N, \hat{\mathbf{y}}_N$ in (3) and the compact convex set $\mathcal{D}$ in (51). If the stepsize $\eta$ satisfies the condition $\eta = \frac{\sigma}{L}$ for any $\sigma \in (0, 1)$, then for all $N \geq 1$, we have*

$$\left[\max_{\mathbf{y}:(\hat{\mathbf{x}}_N, \mathbf{y}) \in \mathcal{D}} f(\hat{\mathbf{x}}_N, \mathbf{y}) - f^\star\right] + \left[f^\star - \min_{\mathbf{x}:(\mathbf{x}, \hat{\mathbf{y}}_N) \in \mathcal{D}} f(\mathbf{x}, \hat{\mathbf{y}}_N)\right] \leq \frac{DL\left(16 + \frac{33}{2(1-\sigma^2)}\right)}{N}, \tag{53}$$

*where $f^\star = f(\mathbf{x}^*, \mathbf{y}^*)$ and $D = \|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \|\mathbf{y}_0 - \mathbf{y}^*\|^2$.*

*Proof.* See Appendix C for the proof. □

Now, similar to Corollary 1, we have:

**Corollary 2.** *Suppose Assumptions 1, 2 and 3 hold. Let $\{\mathbf{x}_{k+1/2}, \mathbf{y}_{k+1/2}\}$ be the iterates generated by the EG updates in (45)-(46). Let the initial conditions satisfy $\mathbf{x}_0 = \mathbf{x}_{-1/2}$ and $\mathbf{y}_0 = \mathbf{y}_{-1/2}$ Consider the definition of the averaged iterates $\hat{\mathbf{x}}_N, \hat{\mathbf{y}}_N$ in (3). If the stepsize $\eta$ satisfies the condition $\eta = \frac{\sigma}{L}$ for any $\sigma \in (0, 1)$, then for all $N \geq 1$, we have*

$$|f(\hat{\mathbf{x}}_N, \hat{\mathbf{y}}_N) - f^\star| \leq \frac{DL\left(16 + \frac{33}{2(1-\sigma^2)}\right)}{N},$$

*where $f^\star = f(\mathbf{x}^*, \mathbf{y}^*)$.*

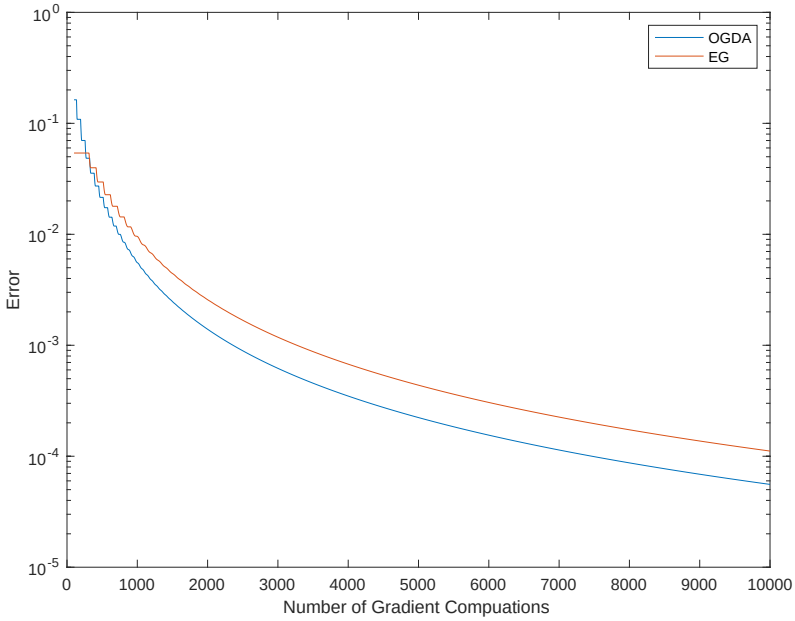


Figure 1: Number of Gradient computations required ($x$-axis) to reach any error level ($y$-axis) for both OGDA and EG for the problem in Equation (54)

# 6 Discussion and Numerical Experiments

The main message of this work is that the OGDA algorithm obtains the same convergence rate of $\mathcal{O}(1/k)$, the best achievable rate (see [Nemirovski, 2004]), also achieved by EG. However, the advantage of OGDA is that we need only one gradient computation at each step, as opposed to two gradient computations needed in EG. This shows the computational advantage that OGDA has over EG.

We compare the performance of OGDA and EG in terms of gradient computations, on the bilinear minimax games considered in [Nemirovski, 2004], without any constraint. In particular, we consider the following minimax problem:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \max_{\mathbf{y}\in\mathbb{R}^n} \mathbf{x}^\top \mathbf{B}\mathbf{y}, \tag{54}$$

where $\mathbf{B} \in \mathbb{R}^{n\times n}$ is a sparse random matrix generated as follows. Each element is nonzero independently with probability $p$. If an element is chosen to be non-zero, it is chosen uniformly from $[-1, 1]$. We compare the number of gradient computations required to reach a desired accuracy level for this problem in Figure 1. As we observe, both EG and OGDA converge to the saddle point of the bilinear problem at a sublinear rate of $\mathcal{O}(1/k)$, but OGDA slightly outperforms EG in terms of number of gradient evaluations. Once again, this is due to the fact that for both descent and ascent updates of OGDA requires only one gradient computation each, while EG requires two gradient computations for both updates at each iteration.

Note that the Lipschitz constants for the considered problem can be estimated from data using standard line search techniques. In particular, Beck & Teboulle [2009] discuss a backward tracking algorithm (ISTA with backtracking) which can be used to estimate the Lipschitz constants, in particular $L_{xx}$ and $L_{yy}$. Several variants of this algorithm, including the Lipschitz line-search algorithm (Algorithm 2) in [Schmidt et al., 2015], can also be used to estimate the Lipschitz constants $L_{xx}$ and $L_{yy}$. For the specific case of saddle point problems, a recent paper [Hamedani & Aybat, 2018] proposes a line search algorithm, to estimate the Lipschitz constant $L_{xx}, L_{xy}, L_{yx}$ and $L_{yy}$. They propose an algorithm - Accelerated Primal Dual with backtracking (Algorithm 2.3) which uses a backtracking procedure, similar to [Malitsky & Pock, 2018], to locally estimate the Lipschitz constants of the problem. Also, regarding the initial error, we would like to highlight that in the analysis of convex minimization problems or convex-concave saddle point problems, we often have a term of the form $\|\mathbf{x}_0 - \mathbf{x}^*\|^2$ in the upper bound (for instance see [Nesterov, 2013; Monteiro & Svaiter, 2010]) which shows the effect of initial error. This parameter is hard to es-

timate in general but can be upper bounded in specific cases. For example, if we are looking at for mixed strategies in zero-sum games, we know that we are looking for a solution lies in the probability simplex, so we can bound the initial error simply by the diameter of the simplex. In general, if we know that our iterates of the algorithm are going to lie in some compact set, we can upper bound the initial distance to the solution simply by the diameter of the compact set.

# 7 Conclusions

In this paper, we established convergence guarantees of the optimistic gradient ascent-descent (OGDA) and Extra-gradient (EG) methods for unconstrained, smooth, and convex-concave saddle point problems. In particular, we showed a sublinear convergence rate of $\mathcal{O}(1/k)$ in terms of function value error for both OGDA and EG by interpreting them as approximate variants of the proximal point method. This result leads to the first theoretical guarantee for OGDA in convex-concave saddle point problems. Moreover, it provides a simple and short proof for the convergence rate of EG in convex-concave saddle point problems when we measure optimality gap in terms of function value.

# A Proof of Theorem 1

The update of the proximal point method can be written as:

$$\mathbf{z}_{k+1} = \mathbf{z}_k - \eta F(\mathbf{z}_{k+1}) \tag{55}$$

According to this update we can show that

$$\|\mathbf{z}_{k+1} - \mathbf{z}\|^2 = \|\mathbf{z}_k - \mathbf{z}\|^2 - 2\eta(\mathbf{z}_k - \mathbf{z})^\top F(\mathbf{z}_{k+1}) + \eta^2 \|F(\mathbf{z}_{k+1})\|^2 \tag{56}$$

Now add and subtract the inner product $2\eta \mathbf{z}_{k+1}^\top F(\mathbf{z}_{k+1})$ to the right hand side and regroup the terms to obtain

$$\begin{aligned}\|\mathbf{z}_{k+1} - \mathbf{z}\|^2 = \|\mathbf{z}_k - \mathbf{z}\|^2 &- 2\eta(\mathbf{z}_{k+1} - \mathbf{z})^\top F(\mathbf{z}_{k+1}) - 2\eta(\mathbf{x}_k - \mathbf{x}_{k+1})^\top F(\mathbf{z}_{k+1}) \\ &+ \eta^2 \|F(\mathbf{z}_{k+1})\|^2.\end{aligned} \tag{57}$$

Replace $F(\mathbf{z}_{k+1})$ with $(1/\eta)(-\mathbf{z}_{k+1} + \mathbf{z}_k)$ to obtain

$$\begin{aligned}&\|\mathbf{z}_{k+1} - \mathbf{z}\|^2 \\ &= \|\mathbf{z}_k - \mathbf{z}\|^2 - 2\eta(\mathbf{z}_{k+1} - \mathbf{z})^\top F(\mathbf{z}_{k+1}) + 2(\mathbf{z}_k - \mathbf{z}_{k+1})^\top (\mathbf{z}_{k+1} - \mathbf{z}_k) \\ &\quad + \|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 \\ &= \|\mathbf{z}_k - \mathbf{z}\|^2 - 2\eta(\mathbf{z}_{k+1} - \mathbf{z})^\top F(\mathbf{z}_{k+1}) - \|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2.\end{aligned} \tag{58}$$

On rearranging the terms, we get the following

$$F(\mathbf{z}_{k+1})^\top (\mathbf{z}_{k+1} - \mathbf{z}) = \frac{1}{2\eta}\|\mathbf{z}_k - \mathbf{z}\|^2 - \frac{1}{2\eta}\|\mathbf{z}_{k+1} - \mathbf{z}\|^2 - \frac{1}{2\eta}\|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2, \tag{59}$$

Now, on substituting $\mathbf{z} = \mathbf{z}^*$, and noting that $F(\mathbf{z}_{k+1})^\top (\mathbf{z}_{k+1} - \mathbf{z}^*) \geq 0$, we have:

$$\|\mathbf{z}_{k+1} - \mathbf{z}^*\|^2 \leq \|\mathbf{z}_k - \mathbf{z}^*\|^2 - \|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 \tag{60}$$

and the proof of boundedness is complete.

On adding Equation (59) from $k = 0, \cdots N-1$ and diving by $N$, we get:

$$\frac{1}{N}\sum_{k=1}^{N} F(\mathbf{z}_k)^\top (\mathbf{z}_k - \mathbf{z}) \leq \frac{\|\mathbf{z}_0 - \mathbf{z}\|^2}{\eta N} \tag{61}$$

Now, using Proposition 1 we can write

$$|f(\hat{\mathbf{x}}_N, \hat{\mathbf{y}}_N) - f^\star| \leq \frac{\|\mathbf{x}_0 - \mathbf{x}\|^2 + \|\mathbf{y}_0 - \mathbf{y}\|^2}{\eta N}, \tag{62}$$

and the proof is complete.

# B Proof of Lemma 3

(a) Considering the updates in (48) and (49) we can write the update of mid-points in EG as

$$\mathbf{z}_{k+\frac{1}{2}} = \mathbf{z}_{k-\frac{1}{2}} - \eta F(\mathbf{z}_{k+\frac{1}{2}}) + \boldsymbol{\varepsilon}_k, \tag{63}$$

where,

$$\boldsymbol{\varepsilon}_k = \eta \left[ (F(\mathbf{z}_{k+\frac{1}{2}}) - F(\mathbf{z}_{k-\frac{1}{2}})) - (F(\mathbf{z}_k) - F(\mathbf{z}_{k-1})) \right]. \tag{64}$$

Therefore, we can simplify the last term in Equation (16) of Proposition 2 as follows:

$$\begin{aligned}
&\frac{1}{\eta}\boldsymbol{\varepsilon}_k{}^\top(\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}) \\
&= \frac{1}{\eta} \times [(\eta F(\mathbf{z}_{k+\frac{1}{2}}) - \eta F(\mathbf{z}_k)) - (\eta F(\mathbf{z}_{k-\frac{1}{2}}) - \eta F(\mathbf{z}_{k-1}))]^\top(\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}) \\
&= (F(\mathbf{z}_{k+\frac{1}{2}}) - F(\mathbf{z}_k))^\top(\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}) - (F(\mathbf{z}_{k-\frac{1}{2}}) - F(\mathbf{z}_{k-1}))^\top(\mathbf{z}_{k-\frac{1}{2}} - \mathbf{z}) \\
&\qquad - (F(\mathbf{z}_{k-\frac{1}{2}}) - F(\mathbf{z}_{k-1}))^\top(\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}).
\end{aligned} \tag{65}$$

Using Lipschitz continuity of the operator $F$ (Lemma 1(b)) and Young's inequality, we have

$$\begin{aligned}
&\frac{1}{\eta}\boldsymbol{\varepsilon}_k{}^\top(\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}) \\
&\quad \le F(\mathbf{z}_{k+\frac{1}{2}}) - F(\mathbf{z}_k))^\top(\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}) - (F(\mathbf{z}_{k-\frac{1}{2}}) - F(\mathbf{z}_{k-1}))^\top(\mathbf{z}_{k-\frac{1}{2}} - \mathbf{z}) \\
&\qquad + L\|\mathbf{z}_{k-\frac{1}{2}} - \mathbf{z}_{k-1}\|\|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}\| \\
&\quad \le F(\mathbf{z}_{k+\frac{1}{2}}) - F(\mathbf{z}_k))^\top(\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}) - (F(\mathbf{z}_{k-\frac{1}{2}}) - F(\mathbf{z}_{k-1}))^\top(\mathbf{z}_{k-\frac{1}{2}} - \mathbf{z}) \\
&\qquad + \frac{L}{2}\|\mathbf{z}_{k-\frac{1}{2}} - \mathbf{z}_{k-1}\|^2 + \frac{L}{2}\|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}\|^2
\end{aligned} \tag{66}$$

Substituting the upper bound in (66) into Equation (16) of Proposition 2, implies that

$$\begin{aligned}
&F(\mathbf{z}_{k+\frac{1}{2}})^\top(\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}) \\
&\quad \le \frac{1}{2\eta}\|\mathbf{z}_{k-\frac{1}{2}} - \mathbf{z}\|^2 - \frac{1}{2\eta}\|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}\|^2 - \frac{1}{2\eta}\|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}\|^2 \\
&\qquad + (F(\mathbf{z}_{k+\frac{1}{2}}) - F(\mathbf{z}_k))^\top(\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}) - (F(\mathbf{z}_{k-\frac{1}{2}}) - F(\mathbf{z}_{k-1}))^\top(\mathbf{z}_{k-\frac{1}{2}} - \mathbf{z}) \\
&\qquad + \frac{L}{2}\|\mathbf{z}_{k-\frac{1}{2}} - \mathbf{z}_{k-1}\|^2 + \frac{L}{2}\|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}\|^2.
\end{aligned} \tag{67}$$

Since $\eta < 1/L$, we have $-\frac{1}{2\eta}\|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}\|^2 + \frac{L}{2}\|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}\|^2 \le 0$ and therefore

$$\begin{aligned}
&F(\mathbf{z}_{k+\frac{1}{2}})^\top(\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}) \\
&\le \frac{1}{2\eta}\|\mathbf{z}_{k-\frac{1}{2}} - \mathbf{z}\|^2 - \frac{1}{2\eta}\|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}\|^2 + \frac{L}{2}\|\mathbf{z}_{k-\frac{1}{2}} - \mathbf{z}_{k-1}\|^2 \\
&\quad + (F(\mathbf{z}_{k+\frac{1}{2}}) - F(\mathbf{z}_k))^\top(\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}) - (F(\mathbf{z}_{k-\frac{1}{2}}) - F(\mathbf{z}_{k-1}))^\top(\mathbf{z}_{k-\frac{1}{2}} - \mathbf{z}).
\end{aligned} \tag{68}$$

which completes the proof of Part (a).
(b) Based on the update of EG in (47), we can write

$$\begin{aligned}
&\|\mathbf{z}_k - \mathbf{z}\|^2 \\
&= \|\mathbf{z}_k - \mathbf{z}_{k+1} + \mathbf{z}_{k+1} - \mathbf{z}\|^2 \\
&= \|\mathbf{z}_{k+1} - \mathbf{z}\|^2 + 2(\mathbf{z} - \mathbf{z}_{k+1})^\top(\mathbf{z}_{k+1} - \mathbf{z}_k) + \|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 \\
&= \|\mathbf{z}_{k+1} - \mathbf{z}\|^2 + 2(\mathbf{z} - \mathbf{z}_{k+\frac{1}{2}})^\top(\mathbf{z}_{k+1} - \mathbf{z}_k) \\
&\qquad + 2(\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k+1})^\top(\mathbf{z}_{k+1} - \mathbf{z}_k) + \|\mathbf{z}_{k+1} - \mathbf{z}_k\|^2 \\
&= \|\mathbf{z}_{k+1} - \mathbf{z}\|^2 + 2(\mathbf{z} - \mathbf{z}_{k+\frac{1}{2}})^\top(\mathbf{z}_{k+1} - \mathbf{z}_k) + \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2 - \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k+1}\|^2.
\end{aligned} \tag{69}$$

Now we proceed to bound the difference $\|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k+1}\|^2$. Using the fact that the operator $F$ is $L$-Lipschitz (Lemma 1(b)), we have

$$\begin{aligned}\|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k+1}\|^2 &= \eta^2 \|F(\mathbf{z}_{k+\frac{1}{2}} - F(\mathbf{z}_k)\|^2 \\ &\leq \eta^2 L^2 \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2. \end{aligned} \tag{70}$$

Substituting this upper bound back into (69) and taking $\mathbf{z} = \mathbf{z}^*$ implies

$$\begin{aligned}&\|\mathbf{z}_k - \mathbf{z}^*\|^2 \\ &\quad \geq \|\mathbf{z}_{k+1} - \mathbf{z}^*\|^2 + 2(\mathbf{z}^* - \mathbf{z}_{k+\frac{1}{2}})^\top (\mathbf{z}_{k+1} - \mathbf{z}_k) + (1 - \eta^2 L^2)\|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2. \end{aligned} \tag{71}$$

Further, since the operator $F$ is monotone, we have

$$\begin{aligned}(\mathbf{z}^* - \mathbf{z}_{k+\frac{1}{2}})^\top (\mathbf{z}_{k+1} - \mathbf{z}_k) &= \eta (F(\mathbf{z}_{k+\frac{1}{2}}))^\top (\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}^*) \\ &\geq \eta (F(\mathbf{z}_{k+\frac{1}{2}}) - F(\mathbf{z}^*))^\top (\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}^*) \\ &\geq 0, \end{aligned} \tag{72}$$

where in the first inequality we used the fact that $F(\mathbf{z}^*) = \mathbf{0}$ (Lemma 1(c)), and the last inequality holds due to monotonicity of $F$ (Lemma 1(a)). Therefore, we can replace the inner product $2(\mathbf{z}^* - \mathbf{z}_{k+\frac{1}{2}})^\top (\mathbf{z}_{k+1} - \mathbf{z}_k)$ in (71) by its lower bound 0 to obtain

$$\|\mathbf{z}_k - \mathbf{z}^*\|^2 \geq \|\mathbf{z}_{k+1} - \mathbf{z}^*\|^2 + (1 - \eta^2 L^2)\|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2 \tag{73}$$

The result in (73) shows that the seqeunce $\|\mathbf{z}_k - \mathbf{z}^*\|^2$ is non-increasing. Therefore, for any iterate $k$, it holds that

$$\|\mathbf{z}_k - \mathbf{z}^*\|^2 \leq \|\mathbf{z}_0 - \mathbf{z}^*\|^2. \tag{74}$$

Now, for all $k \geq 0$, we have:

$$\begin{aligned}\|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}^*\|^2 &\leq 2\|\mathbf{z}_k - \mathbf{z}^*\|^2 + 2\|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2 \\ &\leq \left(2 + \frac{2}{1 - \eta^2 L^2}\right) \|\mathbf{z}_k - \mathbf{z}^*\|^2 \\ &\leq \left(2 + \frac{2}{1 - \eta^2 L^2}\right) \|\mathbf{z}_0 - \mathbf{z}^*\|^2 \end{aligned} \tag{75}$$

where the first inequality follows from the fact that for all $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, $\|\mathbf{a} + \mathbf{b}\|^2 \leq 2\|\mathbf{a}\|^2 + 2\|\mathbf{b}\|^2$, the second inequality follows from (73) and the third inequality follows from (74). Therefore from (74) and (75), since $0 < 1 - \eta^2 L^2 < 1$, we see that the iterates $\{\mathbf{z}_k\}, \{\mathbf{z}_{k+\frac{1}{2}}\}$ belong to the compact set $\mathcal{D}$ defined in (51).

Now by summing both sides of (73) for $k = 0, \ldots, \infty$, we obtain

$$(1 - \eta^2 L^2) \sum_{k=0}^{\infty} \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2 \leq \|\mathbf{z}_0 - \mathbf{z}^*\|^2 \tag{76}$$

Therefore, by regrouping the terms we obtain

$$\sum_{k=0}^{\infty} \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2 \leq \frac{\|\mathbf{z}_0 - \mathbf{z}^*\|^2}{1 - \eta^2 L^2}, \tag{77}$$

and the claim in (52) follows.

# C Proof of Theorem 3

Using Equation (50) of Lemma 3(a), summing it from $k = 0, \cdots, N-1$ and dividing by $N$, we

obtain

$$\frac{1}{N}\sum_{k=0}^{N-1} F(\mathbf{z}_{k+\frac{1}{2}})^\top(\mathbf{z}_{k+\frac{1}{2}}-\mathbf{z}) \\ \leq \frac{\frac{1}{2\eta}\|\mathbf{z}_0-\mathbf{z}\|^2+(F(\mathbf{z}_{N-\frac{1}{2}})-F(\mathbf{z}_{N-1}))^\top(\mathbf{z}_{N-\frac{1}{2}}-\mathbf{z})}{N}+\frac{L}{2N}\sum_{k=0}^{N-1}\|\mathbf{z}_{k-\frac{1}{2}}-\mathbf{z}_{k-1}\|^2. \tag{78}$$

The bound in Equation (52) from Lemma 3(b) yields

$$\begin{aligned}\frac{1}{N}\sum_{k=0}^{N-1} & F(\mathbf{z}_{k+\frac{1}{2}})^\top(\mathbf{z}_{k+\frac{1}{2}}-\mathbf{z}) \\ &\leq \frac{L\|\mathbf{z}_0-\mathbf{z}\|^2+(F(\mathbf{z}_{N-\frac{1}{2}})-F(\mathbf{z}_{N-1}))^\top(\mathbf{z}_{N-\frac{1}{2}}-\mathbf{z})}{N}+\frac{L\|\mathbf{z}_0-\mathbf{z}^*\|^2}{2(1-\eta^2L^2)N} \\ &\leq \frac{L\|\mathbf{z}_0-\mathbf{z}\|^2+L\|\mathbf{z}_{N-\frac{1}{2}}-\mathbf{z}_{N-1}\|\|\mathbf{z}_{N-\frac{1}{2}}-\mathbf{z}\|+\frac{L}{2(1-\sigma^2)}\|\mathbf{z}_0-\mathbf{z}^*\|^2}{N},\end{aligned} \tag{79}$$

where in the last inequality we use Lipschitz continuity of the operator $F$ (Lemma 1(b)) and the fact that $\eta=\frac{\sigma}{L}$. Note that for any $\mathbf{z}_1,\mathbf{z}_2\in\mathcal{D}$, we have:

$$\begin{aligned}\|\mathbf{z}_1-\mathbf{z}_2\| &\leq \|\mathbf{z}_1-\mathbf{z}^*\|+\|\mathbf{z}_2-\mathbf{z}^*\| \\ &\leq \sqrt{\left(2+\frac{2}{1-\eta^2L^2}\right)}\|\mathbf{z}_0-\mathbf{z}^*\|+\sqrt{\left(2+\frac{2}{1-\eta^2L^2}\right)}\|\mathbf{z}_0-\mathbf{z}^*\| \\ &\leq 2\sqrt{D\left(2+\frac{2}{1-\sigma^2}\right)}.\end{aligned} \tag{80}$$

Therefore, for any point $\mathbf{z}$ in the set $\mathcal{D}$, we can substitute the preceding relation in Equation (79) to get

$$\frac{1}{N}\sum_{k=0}^{N-1} F(\mathbf{z}_{k+\frac{1}{2}})^\top(\mathbf{z}_{k+\frac{1}{2}}-\mathbf{z}) \leq \frac{DL\left(16+\frac{33}{2(1-\sigma^2)}\right)}{N}. \tag{81}$$

Now, using Proposition 1 we have that for all $\mathbf{x},\mathbf{y}\in\mathcal{D}$:

$$f(\hat{\mathbf{x}}_N,\mathbf{y})-f(\mathbf{x},\hat{\mathbf{y}}_N)\leq\frac{DL\left(16+\frac{33}{2(1-\sigma^2)}\right)}{N}, \tag{82}$$

where $\hat{\mathbf{x}}_N=\frac{1}{N}\sum_{k=0}^{N-1}\mathbf{x}_{k+1/2}$ and $\hat{\mathbf{y}}_N=\frac{1}{N}\sum_{k=0}^{N-1}\mathbf{y}_{k+1/2}$ which gives us the following convergence result:

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

where $f^\star=f(\mathbf{x}^*,\mathbf{y}^*)$.

# References


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

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

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

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

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

Bruck Jr, R. E. On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in hilbert space. *Journal of Mathematical Analysis and Applications*, 61(1):159–164, 1977.

Burachik, R. S., Iusem, A. N., and Svaiter, B. F. Enlargement of monotone operators with applications to variational inequalities. *Set-Valued Analysis*, 5(2):159–180, 1997.

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

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

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

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

Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. Training gans with optimism. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

Du, S. S. and Hu, W. 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*, pp. 196–205, 2019.

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

Gidel, G., Berard, H., Vincent, P., and Lacoste-Julien, S. A variational inequality perspective on generative adversarial nets. *arXiv preprint arXiv:1802.10551*, 2018.

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

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

Güler, O. New proximal point algorithms for convex minimization. *SIAM Journal on Optimization*, 2(4):649–664, 1992.

Hamedani, E. Y. and Aybat, N. S. A primal-dual algorithm for general convex-concave saddle point problems. *arXiv preprint arXiv:1803.01401*, 2018.

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

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

Liang, T. and Stokes, J. 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*, pp. 907–915, 2019.

Malitsky, Y. and Pock, T. A first-order primal-dual algorithm with linesearch. *SIAM Journal on Optimization*, 28(1):411–432, 2018.

Malitsky, Y. and Tam, M. K. A forward-backward splitting method for monotone inclusions without cocoercivity. *arXiv preprint arXiv:1808.04162*, 2018.

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

Mokhtari, A., Ozdaglar, A., and Pattathil, S. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. pp. 1497–1507, 2020.

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

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

Nemirovski, A. 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, 2004.

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Popov, L. D. A modification of the arrow-hurwicz method for search of saddle points. *Mathematical Notes*, 28(5):845–848, 1980.

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

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

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

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

Schmidt, M., Babanezhad, R., Ahmed, M., Defazio, A., Clifton, A., and Sarkar, A. Non-uniform stochastic average gradient method for training conditional random fields. In *artificial intelligence and statistics*, pp. 819–828, 2015.

Teboulle, M. Convergence of proximal-like algorithms. *SIAM Journal on Optimization*, 7(4): 1069–1083, 1997.

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