Min-Max Optimization under Delays

Arman Adibi, Aritra Mitra, and Hamed Hassani

***Abstract*— Delays and asynchrony are inevitable in large-scale machine-learning problems where communication plays a key role. As such, several works have extensively analyzed stochastic optimization with delayed gradients. However, as far as we are aware, no analogous theory is available for min-max optimization, a topic that has gained recent popularity due to applications in adversarial robustness, game theory, and reinforcement learning. Motivated by this gap, we examine the performance of standard min-max optimization algorithms with delayed gradient updates. First, we show (empirically) that even small delays can cause prominent algorithms like Extra-gradient (`EG`) to diverge on simple instances for which `EG` guarantees convergence in the absence of delays. Our empirical study thus suggests the need for a careful analysis of delayed versions of min-max optimization algorithms. Accordingly, under suitable technical assumptions, we prove that Gradient Descent-Ascent (`GDA`) and `EG` with delayed updates continue to guarantee convergence to saddle points for convex-concave and strongly convex-strongly concave settings. Our complexity bounds reveal, in a transparent manner, the slow-down in convergence caused by delays.**

I. Introduction

Min-max optimization is a fundamental problem with applications in various fields, including game theory [1], machine learning [2], robust optimization [3], and more recently, adversarial robustness [4]. As such, the convergence analysis of various min-max optimization algorithms has received considerable attention over the years [5]–[8]. While this has resulted in a rich literature that provides non-asymptotic guarantees for the vanilla versions of these algorithms, not much is known about their *robustness* to different types of perturbations that show up in practice. In particular, for large-scale machine learning problems involving communication between multiple servers and agents, such perturbations get manifested in the form of (unavoidable) delays and asynchrony. Consequently, several works have extensively studied stochastic optimization with delayed gradients; since the literature on this topic is vast, we refer the reader to [9]–[13] and the references therein. However, to our knowledge, there is no analogous theory for min-max optimization. Motivated by this gap, the goal of our paper is to build an understanding of the effect of delays on the convergence of common min-max optimization algorithms like Gradient Descent-Ascent (`GDA`) and Extra-Gradient (`EG`). Our main contributions in this regard are as follows.

A. Adibi and H. Hassani are with the Department of Electrical and Systems Engineering, University of Pennsylvania. Email: {aadibi, hassani}@seas.upenn.edu. A. Mitra is with the Department of Electrical and Computer Engineering, North Carolina State University. Email: amitra2@ncsu.edu. This work was supported by NSF Award 1837253, NSF CAREER award CIF 1943064, and the Air Force Office of Scientific Research Young Investigator Program (AFOSR-YIP) under award FA9550-20-1-0111.

A. Summary of Main Results

• We start with a result that is perhaps surprising. In Section II-A, we empirically examine the effect of delays on the behavior of the Extra-Gradient algorithm due to Korpelevich [5]. We observe that even with the smallest possible delay, i.e., a unit delay, `EG` diverges on a simple convex-concave function; see Fig. 1.[1] Notably, in the absence of delays, `EG` provably guarantees convergence to a saddle-point for this function. This observation, although empirical, suggests that *delays can have non-trivial effects on the convergence of popular min-max optimization algorithms*.

• Our empirical study conveys the message that technical assumptions that are typically not required to study vanilla `EG` might, in fact, turn out to be needed to ensure convergence under delays. Accordingly, in Section III, we study `DEG` - a version of `EG` with updates based on delayed gradients - for smooth, convex-concave functions over a *bounded* domain. In Theorem 1, we show that `DEG` guarantees convergence to a saddle-point at a rate $O(\sqrt{\tau_{\max}}/\sqrt{T})$, where T is the number of iterations, and $\tau_{\max}$ is a uniform bound on the delays. Our proof of this result is based on a connection to adversarial perturbations on statistical min-max learning problems in the recent work [14].

In the absence of delays, the convergence rates of `EG` and Gradient Descent-Ascent (`GDA`) are $O(1/T)$ [15] and $O(1/\sqrt{T})$ [6], respectively. Our empirical divergence result (see Footnote 1) and Theorem 1 collectively suggest *that under delays, the behavior of `EG` is similar to that of `GDA`.*

• To further investigate the above point, we turn our attention to the behavior of `GDA` under delays in Section IV; we refer to this delayed version as `DGDA`. For smooth, convex-concave functions with bounded gradients, we prove that `DGDA` exhibits a convergence rate of $O(\sqrt{\tau_{\max}}/\sqrt{T})$ - exactly like `DEG`; see Theorem 2. However, unlike the analysis for `DEG`, we do not assume a bounded domain. Instead, we provide a careful analysis to argue that with suitable step-sizes, the iterates of `DGDA` remain bounded.

• All our results above pertain to scenarios where there is some underlying assumption of boundedness (either on the gradients or on the domain). Thus, one may ask: *Can min-max optimization algorithms under delays converge in the absence of such boundedness assumptions?* In Section V, we answer this question in the affirmative by studying `DGDA` for smooth, strongly convex-strongly concave functions. We prove that `DGDA` guarantees *linear convergence to the saddle point* at a rate of $O(\exp(-T/\tau_{\max}^3))$; see Theorem 3.

[1] The Gradient Descent-Ascent (`GDA`) algorithm diverges on this instance even in the absence of delays [7].

As far as we are aware, our results above are novel and provide the first steps toward theoretically understanding the robustness of min-max optimization algorithms to delay-induced perturbations. Our results are summarized in Table I.

II. Problem Setting

In this section, we start by describing the basic setup of a min-max optimization problem. Next, we show empirically how EG can diverge with even one-step delays. Finally, we conclude the section by outlining some technical assumptions that will be made for the majority of the paper to ensure boundedness and convergence of iterates.

The basic min-max optimization setup. Let $\mathcal{X}$ and $\mathcal{Y}$ be nonempty, convex subsets of $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively.[2] Given a mapping of the form $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, we are interested in solving the following optimization problem:

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x, y). \tag{1}$$

Throughout the paper, we will assume that $f(x, y)$ is *continuously differentiable* in x and y, and *convex-concave* over $\mathcal{X} \times \mathcal{Y}$. Specifically, $f(\cdot, y) : \mathcal{X} \to \mathbb{R}$ is convex for every $y \in \mathcal{Y}$, and $f(x, \cdot) : \mathcal{Y} \to \mathbb{R}$ is concave for every $x \in \mathcal{X}$. Our goal is to find a saddle point (x^*, y^*) of $f(x, y)$ over the set $\mathcal{X} \times \mathcal{Y}$, where a saddle point is defined as a vector pair $(x^*, y^*) \in \mathcal{X} \times \mathcal{Y}$ that satisfies

$$f(x^*, y) \le f(x^*, y^*) \le f(x, y^*), \forall x \in \mathcal{X}, y \in \mathcal{Y}. \tag{2}$$

For any $\bar{x} \in \mathcal{X}$ and $\bar{y} \in \mathcal{Y}$, let $\nabla_x f(\bar{x}, \bar{y})$ and $\nabla_y f(\bar{x}, \bar{y})$ denote the partial gradients of $f(x, y)$ with respect to x and y, respectively, at $(\bar{x}, \bar{y})$. Typical first-order iterative min-max optimization algorithms such as GDA, EG, and Optimistic Gradient Descent-Ascent (OGDA) aim to solve for (x^*, y^*) based on an oracle that provides partial gradients of $f(x, y)$ evaluated at the most recent iterates of the algorithm.

The delay model. Not much, however, is known about scenarios where the oracle is *imperfect*. To that end, we studied the effect of adversarial perturbations on the partial gradients of $f(x, y)$ in our recent work [14]. In this work, we take a different stance. Instead of considering *arbitrary* adversarial perturbations, we will focus on *structured* perturbations induced by delays. As mentioned earlier in the Introduction, the source of such delays could be communication latencies or system-level computational challenges such as stragglers, both of which are prevalent in distributed systems. In this work, given an iterative min-max optimization algorithm that generates a sequence of iterates $\{(x_k, y_k)\}$, we assume that at iteration k, we only have access to partial gradients of $f(x, y)$ computed at a *stale* iterate $(x_{k-\tau_k}, y_{k-\tau_k})$, i.e., we have access to $\nabla_x f(x_{k-\tau_k}, y_{k-\tau_k})$ and $\nabla_y f(x_{k-\tau_k}, y_{k-\tau_k})$, where τ_k is the delay at iteration k. While we allow the delays to be time-varying, throughout the paper, we will work under the running assumption that all delays are uniformly bounded, i.e., there exists some positive integer $\tau_{\max}$ such that $\tau_k \le \tau_{\max}, \forall k$.

[2]While we will assume that $\mathcal{X}$ and $\mathcal{Y}$ are bounded sets in Section III, this assumption will be later relaxed in Sections IV and V.

Our goal is to understand what happens, when for computing the next iterate (x_{k+1}, y_{k+1}), one uses these delayed gradients as opposed to $\nabla_x f(x_k, y_k)$ and $\nabla_y f(x_k, y_k)$. Specifically, we ask:

- Can we hope for convergence to saddle points using delayed versions of algorithms like GDA and EG?
- If so, for different classes of functions, how do the convergence rates get affected by $\tau_{\max}$?

In the next subsection, we demonstrate (empirically) that the answers to such questions are more nuanced than what one might initially expect.

A. Divergence of Extra-Gradient Algorithm under Delay

Let us start by quickly reviewing how the Extra-gradient (EG) algorithm for finding saddle-points operates in an unconstrained setting. EG first computes a set of mid-points $(\hat{x}_k, \hat{y}_k)$ by using partial gradients evaluated at the current iterate (x_k, y_k):

$$\begin{aligned} \hat{x}_k &\leftarrow x_k - \alpha \nabla_x f(x_k, y_k) \\ \hat{y}_k &\leftarrow y_k + \alpha \nabla_y f(x_k, y_k), \end{aligned} \tag{3}$$

where α is a suitable step-size. Next, using gradients evaluated at the mid-points, EG computes the next iterates as

$$\begin{aligned} x_{k+1} &\leftarrow x_k - \alpha \nabla_x f(\hat{x}_k, \hat{y}_k) \\ y_{k+1} &\leftarrow y_k + \alpha \nabla_x f(\hat{x}_k, \hat{y}_k). \end{aligned} \tag{4}$$

For smooth, convex-concave functions, the above EG procedure guarantees convergence to a saddle-point at a rate of $O(1/T)$, where T is the number of iterations [15]. Moreover, to achieve this convergence, one does not need to make any assumption of a bounded domain or bounded gradients.

Now to illustrate the challenges posed by delays, let us consider solving the following problem

$$\min_x \max_y \langle x, y \rangle, \tag{5}$$

using a version of EG where all partial gradients are evaluated at iterates that are delayed by just one time-step.[3] Whereas one might have expected a slow-down in convergence due to delays, Figure 1 shows that in this specific setting, a unit delay causes EG to diverge! This demonstrates that delays can lead to non-trivial phenomena for standard min-max algorithms, thereby justifying our current study.

A rough explanation for the above phenomenon is as follows. In [8], the authors argued that EG can be studied as an approximate version of the Proximal Point (PP) algorithm, which, in turn, operates as follows:

$$\begin{aligned} x_{k+1} &\leftarrow x_k - \alpha \nabla_x f(x_{k+1}, y_{k+1}) \\ y_{k+1} &\leftarrow y_k - \alpha \nabla_y f(x_{k+1}, y_{k+1}). \end{aligned} \tag{6}$$

When the gradients on the right-hand side of the above equations are evaluated at one-step-delayed iterates, the above algorithm reduces to the GDA algorithm. Unlike EG, however, GDA can diverge for smooth, convex-concave problems like the one in Eq. (5), *even in the absence of delays*. In

[3]Formally, the delayed EG algorithm we study is outlined in Algorithm 1.

TABLE I: The table below presents a summary of our findings, outlining the conditions required for each algorithm to achieve the specified convergence rate. In the smooth convex-concave case, the convergence rate corresponds to the number of iterations needed for the duality gap to be less than ϵ. For the smooth strongly convex-strongly concave case (SC-SC), the rate corresponds to the number of iterations needed for the distance to saddle points to be less than ϵ. It is worth noting that in this table, we hide the dependence on G, L, and the strong-convexity parameter in the O notation.

Algorithm	Bounded Gradient	Bounded Domain	SC-SC	Convex-Concave	Convergence Rate
Delayed Extra-Gradient (DEG)	✓	✓	×	✓	$\mathcal{O}(\frac{\tau_{\max}}{\epsilon^2})$
Delayed Gradient Descent-Ascent (DGDA)	✓	×	×	✓	$O(\frac{\tau_{\max}}{\epsilon^2})$
Delayed Gradient Descent-Ascent (DGDA)	×	×	✓	×	$O(\tau_{\max}^3 \log(\frac{1}{\epsilon}))$

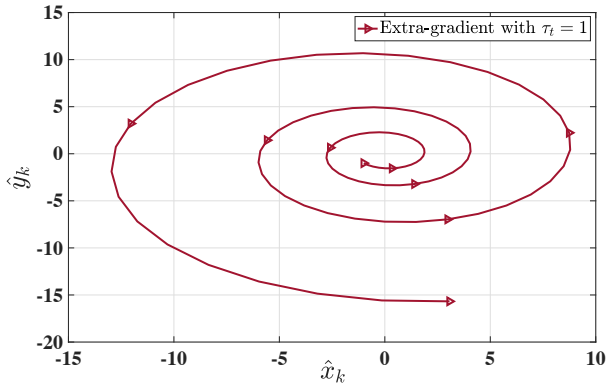

Fig. 1: The Extra-gradient algorithm fails to converge, even with just one step delay, for the optimization problem $\min_x \max_y \langle x, y\rangle$. In this plot, we used a step size of $\alpha = 0.2$. However, with the same step size and no delay, the Extra-gradient algorithm converges to the origin, which is the saddle-point for this problem.

particular, some assumption on the boundedness of domain or gradients is needed to ensure the convergence of GDA for convex-concave problems. From the above discussion, we conclude that since EG with delays tends to behave like GDA, we need to impose additional technical assumptions to ensure convergence to saddle points. As such, we will impose the following assumption of bounded gradients at various points in the paper.

Assumption 1. *There exists a constant $G > 1$ such that the following holds for all $x \in \mathcal{X}$, and all $y \in \mathcal{Y}$: $\|\nabla_x f(x,y)\| \leq G$, and $\|\nabla_y f(x,y)\| \leq G$.*[4]

We will also make the following standard assumption that the partial gradients of $f(x,y)$ are Lipschitz continuous.

Assumption 2. *There exists a constant $L > 1$ such that the following holds for all $x_1, x_2 \in \mathcal{X}$, and all $y_1, y_2 \in \mathcal{Y}$:*

$$\|\nabla_x f(x_1,y_1) - \nabla_x f(x_2,y_2)\| \leq L\left(\|x_1 - x_2\| + \|y_1 - y_2\|\right),$$
$$\|\nabla_y f(x_1,y_1) - \nabla_y f(x_2,y_2)\| \leq L\left(\|x_1 - x_2\| + \|y_1 - y_2\|\right).$$

[4]We will use $\|\cdot\|$ to represent the Euclidean norm.

III. Analysis of Delayed Extra-Gradient for Convex-Concave functions

In Section II-A, we saw that in the absence of a projection step to ensure the boundedness of iterates, the EG algorithm diverges on very simple functions, even with a one-step delay. Based on this empirical observation, in this section, we study delayed extra-gradient (DEG) - outlined in Algorithm 1 - under additional assumptions. In particular, throughout this section, we will work under Assumptions 1 and 2, i.e., we will assume that the partial gradients of $f(x,y)$ are Lipschitz continuous and uniformly bounded. It is important to note here that the divergence of DEG, as illustrated in Figure 1, occurs when we do not impose Assumption 1. Thus, this assumption will play a crucial role in our analysis.

The update rule for DEG (Algorithm 1) involves two steps. In the first step, DEG computes a midpoint $(\hat{x}_k, \hat{y}_k)$ based on partial gradients evaluated at a stale iterate $(x_{k-\tau_k}, y_{k-\tau_k})$; see Eq. (7). In the second step, DEG computes the next iterate (x_{k+1}, y_{k+1}) based on partial gradients evaluated at a stale mid-point $(\hat{x}_{k-\hat{\tau}_k}, \hat{y}_{k-\hat{\tau}_k})$; see Eq. (8). There are two important things to take note of here. First, in each of the above updates, we project onto $\mathcal{X} \times \mathcal{Y}$ to ensure the boundedness of iterates. Second, our analysis is general enough to accommodate time-varying delays; furthermore, we allow τ_k and $\hat{\tau}_k$ to also be different. That said, as mentioned before, we will work under the running assumption that all delays are bounded uniformly by $\tau_{\max}$, i.e., $\max\{\tau_k, \hat{\tau}_k\} \leq \tau_{\max}, \forall k$.

Key Insight and Outline of Analysis. The starting point of our analysis for DEG is the observation that the errors induced by delays can be interpreted as *bounded perturbations*. As we shall see in Lemma 3, the boundedness of the delay-induced errors follows as a direct consequence of Assumptions 1 and 2, and the uniform boundedness assumption on the delays. This key observation allows us to immediately make a connection to our prior work in [14], where we studied min-max optimization under adversarial perturbations. Building on this connection, we start with the following result that establishes some basic inequalities for our subsequent analysis; the proof of this result follows the same steps as that of [14, Lemma 1].

Algorithm 1 Delayed Extra-Gradient (DEG)

Require: Initial vectors $x_1 \in \mathcal{X}$, $y_1 \in \mathcal{Y}$; algorithm parameters: step-size $\alpha > 0$.

1: **for** $k = 1, \dots, T$ **do**

2:
$$\begin{aligned} \hat{x}_k &\leftarrow \Pi_{\mathcal{X}}\left(x_k - \alpha \nabla_x f(x_{k-\tau_k}, y_{k-\tau_k})\right) \\ \hat{y}_k &\leftarrow \Pi_{\mathcal{Y}}\left(y_k + \alpha \nabla_y f(x_{k-\tau_k}, y_{k-\tau_k})\right). \end{aligned} \tag{7}$$

3:
$$\begin{aligned} x_{k+1} &\leftarrow \Pi_{\mathcal{X}}\left(x_k - \alpha \nabla_x f(\hat{x}_{k-\hat{\tau}_k}, \hat{y}_{k-\hat{\tau}_k})\right) \\ y_{k+1} &\leftarrow \Pi_{\mathcal{Y}}\left(y_k + \alpha \nabla_x f(\hat{x}_{k-\hat{\tau}_k}, \hat{y}_{k-\hat{\tau}_k})\right). \end{aligned} \tag{8}$$

4: **end for**

Lemma 1. *For the DEG algorithm, the following inequalities hold for all $k \in [T], x \in \mathcal{X}$, and $y \in \mathcal{Y}$:*[5]

$$\begin{aligned} 2\alpha\langle \nabla_x f(x_{k-\tau_k}, y_{k-\tau_k}), \hat{x}_k - x\rangle &\le \|x - x_k\|^2 - \|x - \hat{x}_k\|^2 - \|\hat{x}_k - x_k\|^2 \\ -2\alpha\langle \nabla_y f(x_{k-\tau_k}, y_{k-\tau_k}), \hat{y}_k - y\rangle &\le \|y - y_k\|^2 - \|y - \hat{y}_k\|^2 - \|\hat{y}_k - y_k\|^2 \\ 2\alpha\langle \nabla_x f(\hat{x}_{k-\hat{\tau}_k}, \hat{y}_{k-\hat{\tau}_k}), x_{k+1} - x\rangle &\le \|x - x_k\|^2 - \|x - x_{k+1}\|^2 - \|x_{k+1} - x_k\|^2 \\ -2\alpha\langle \nabla_y f(\hat{x}_{k-\hat{\tau}_k}, \hat{y}_{k-\hat{\tau}_k}), y_{k+1} - y\rangle &\le \|y - y_k\|^2 - \|y - y_{k+1}\|^2 - \|y_{k+1} - y_k\|^2. \end{aligned}$$

Next, to bound the impact of delays, we introduce the following error vectors:

$$\begin{aligned} e_x(x_k, y_k) &\triangleq \nabla_x f(x_{k-\tau_k}, y_{k-\tau_k}) - \nabla_x f(x_k, y_k), \\ e_y(x_k, y_k) &\triangleq \nabla_y f(x_{k-\tau_k}, y_{k-\tau_k}) - \nabla_y f(x_k, y_k), \end{aligned}$$

and

$$\begin{aligned} e_x(\hat{x}_k, \hat{y}_k) &\triangleq \nabla_x f(\hat{x}_{k-\hat{\tau}_k}, \hat{y}_{k-\hat{\tau}_k}) - \nabla_x f(\hat{x}_k, \hat{y}_k), \\ e_y(\hat{x}_k, \hat{y}_k) &\triangleq \nabla_y f(\hat{x}_{k-\hat{\tau}_k}, \hat{y}_{k-\hat{\tau}_k}) - \nabla_y f(\hat{x}_k, \hat{y}_k). \end{aligned}$$

Let $D = \max\{D_x, D_y\}$, where D_x and D_y are the diameters of the sets $\mathcal{X}$ and $\mathcal{Y}$, respectively. Leveraging Lemma 1, our next result tracks the progress made by the mid-point sequence $\{(\hat{x}_k, \hat{y}_k)\}$ generated by DEG. The proof of this result mirrors that of [14, Lemma 2].

Lemma 2. *Suppose Assumptions 1 and 2 hold. Furthermore, suppose $\alpha \le 1/(2L)$. Then, for the DEG algorithm, the following holds for all $k \in [T], x \in \mathcal{X}$, and $y \in \mathcal{Y}$:*

$$\begin{aligned} &\alpha\langle \nabla_x f(\hat{x}_k, \hat{y}_k), \hat{x}_k - x\rangle - \alpha\langle \nabla_y f(\hat{x}_k, \hat{y}_k), \hat{y}_k - y\rangle \\ &\le \frac{1}{2}\left(\|x - x_k\|^2 - \|x - x_{k+1}\|^2 + \|y - y_k\|^2 - \|y - y_{k+1}\|^2\right) \\ &+ \alpha D\left(\|e_x(x_k, y_k)\| + \|e_x(\hat{x}_k, \hat{y}_k)\| + \|e_y(x_k, y_k)\| + \|e_y(\hat{x}_k, \hat{y}_k)\|\right). \end{aligned}$$

The above result sets things up nicely for a telescoping-sum analysis. However, the missing piece right now is to provide bounds on the delay-induced errors. We derive such bounds in the following result.

Lemma 3. *Suppose Assumptions 1 and 2 hold. For the DEG algorithm, the following error-bounds then apply $\forall k \in [T]$:*

$$\max\{\|e_x(x_k, y_k)\|, \|e_x(\hat{x}_k, \hat{y}_k)\|, \|e_y(x_k, y_k)\|, \|e_y(\hat{x}_k, \hat{y}_k)\|\} \le \Delta_T,$$

where $\Delta_T = 6\alpha G L \tau_{\max}$.

[5] Given a positive integer N, we use $[N]$ to represent the set $\{1, \dots, N\}$.

Proof. In what follows, we only show how to bound $\|e_x(x_k, y_k)\|$ and $\|e_x(\hat{x}_k, \hat{y}_k)\|$; bounds for the other two error terms can be derived in an identical manner. We start by bounding $\|e_x(x_k, y_k)\|$. From equation (8), we have

$$\begin{aligned} \|x_k - x_{k-\tau_k}\| &\le \sum_{j=k-\tau_k}^{k-1} \|x_{j+1} - x_j\| \\ &\overset{(a)}{\le} \alpha \left(\sum_{j=k-\tau_k}^{k-1} \|\nabla_x f(\hat{x}_{j-\hat{\tau}_j}, \hat{y}_{j-\hat{\tau}_j})\| \right) \\ &\overset{(b)}{\le} \alpha G \tau_{\max}, \end{aligned} \tag{9}$$

where (a) follows from the non-expansive property of the projection operator, and (b) follows from Assumption 1 and the fact that $\tau_k \le \tau_{\max}$. Using the exact same steps, one can establish the same bound for $\|y_k - y_{k-\tau_k}\|$. Thus, we have

$$\begin{aligned} \|e_x(x_k, y_k)\| &= \|\nabla_x f(x_k, y_k) - \nabla_x f(x_{k-\tau}, y_{k-\tau})\| \\ &\overset{(a)}{\le} L(\|x_k - x_{k-\tau_k}\| + \|y_k - y_{k-\tau_k}\|) \\ &\overset{(b)}{\le} 2\alpha G L \tau_{\max}, \end{aligned} \tag{10}$$

where (a) follows from smoothness, i.e., Assumption 2, and (b) follows from Eq. (9). Now to bound $e_x(\hat{x}_k, \hat{y}_k)$, observe

$$\begin{aligned} \|e_x(\hat{x}_k, \hat{y}_k)\| &= \|\nabla_x f(\hat{x}_{k-\hat{\tau}_k}, \hat{y}_{k-\hat{\tau}_k}) - \nabla_x f(\hat{x}_k, \hat{y}_k)\| \\ &\overset{(a)}{\le} L(\|\hat{x}_k - \hat{x}_{k-\hat{\tau}_k}\| + \|\hat{y}_k - \hat{y}_{k-\hat{\tau}_k}\|) \\ &\le L(\|\hat{x}_k - x_k\| + \|x_k - x_{k-\hat{\tau}_k}\| + \|x_{k-\hat{\tau}_k} - \hat{x}_{k-\hat{\tau}_k}\| \\ &+ \|\hat{y}_k - y_k\| + \|y_k - y_{k-\hat{\tau}_k}\|) + \|y_{k-\hat{\tau}_k} - \hat{y}_{k-\hat{\tau}_k}\|) \\ &\overset{(b)}{\le} 2\alpha G L \tau_{\max} + 4\alpha G L \\ &\overset{(c)}{\le} 6\alpha G L \tau_{\max}. \end{aligned}$$

In the above steps, (a) follows from Assumption 2, (b) follows from (9) and the fact that for any $j \in [T]$, $\|\hat{x}_j - x_j\| \le \alpha\|\nabla_x f(x_{j-\tau_j}, y_{j-\tau_j})\| \le \alpha G$, and (c) follows from noting that $\tau_{\max} \ge 1$. This concludes the proof. ∎

We are now in a position to prove our first main result which establishes complexity bounds for DEG for smooth convex-concave functions with bounded gradients.

Theorem 1. *Suppose Assumptions 1 and 2 hold. Moreover, suppose the number of iterations T is large enough such that $T \ge L$. Then, with*

$$\alpha = \sqrt{\frac{1}{24 G L \tau_{\max} T}},$$

the iterates generated by DEG satisfy:

$$\max_{y \in \mathcal{Y}} f(\bar{x}_T, y) - \min_{x \in \mathcal{X}} f(x, \bar{y}_T) \le 10 D^2 \sqrt{\frac{G L \tau_{\max}}{T}}, \tag{11}$$

where $\bar{x}_T = (1/T)\sum_{k \in [T]} \hat{x}_k$ and $\bar{y}_T = (1/T)\sum_{k \in [T]} \hat{y}_k$.

Proof. Let us start by noting that when $T \ge L$, the choice of step-size above satisfies $\alpha \le 1/(2L)$. Thus, we can invoke

Algorithm 2 Delayed Gradient Descent-Ascent (DGDA)

Require: Initial vector $z_1 = [x_1; y_1] \in \mathbb{R}^{m+n}$; algorithm parameters: step-size $\alpha > 0$.

1: **for** $k = 1, \ldots, T$ **do**

2:
$$z_{k+1} = z_k - \alpha \Phi(z_{k-\tau_k}). \tag{14}$$

3: **end for**

Lemma 2. From the convex-concave property of $f(x, y)$, the following inequalities hold $\forall k \in [T], x \in \mathcal{X}$, and $y \in \mathcal{Y}$:

$$\alpha \left(f(\hat{x}_k, \hat{y}_k) - f(x, \hat{y}_k) \right) \leq \alpha \langle \nabla_x f(\hat{x}_k, \hat{y}_k), \hat{x}_k - x \rangle$$
$$-\alpha \left(f(\hat{x}_k, \hat{y}_k) - f(\hat{x}_k, y) \right) \leq -\alpha \langle \nabla_y f(\hat{x}_k, \hat{y}_k), \hat{y}_k - y \rangle.$$

Summing the two inequalities above, and using Lemmas 2 and 3, we obtain:

$$\alpha \left(f(\hat{x}_k, y) - f(x, \hat{y}_k) \right) \leq \frac{1}{2} \left(\|x - x_k\|^2 - \|x - x_{k+1}\|^2 \right) + \frac{1}{2} \left(\|y - y_k\|^2 - \|y - y_{k+1}\|^2 \right) + 4\alpha D \Delta_T, \tag{12}$$

where Δ_T is as defined in Lemma 3. From the convexity of $f(x, y)$ w.r.t. x and concavity w.r.t. y, note that we have $f(\bar{x}_T, y) \leq (1/T) \sum_{k \in [T]} f(\hat{x}_k, y)$ and $f(x, \bar{y}_T) \geq (1/T) \sum_{k \in [T]} f(x, \hat{y}_k)$, respectively. Combining this with Eq. (12), we obtain

$$f(\bar{x}_T, y) - f(x, \bar{y}_T) \leq \frac{D^2}{\alpha T} + 4D\Delta_T. \tag{13}$$

The result follows by plugging into the above inequality the choice of α in the statement of the theorem, and by noting that the resulting bound holds for all $x \in \mathcal{X}$ and for all $y \in \mathcal{Y}$. This completes the proof. □

Discussion. From Theorem 1, we conclude that for smooth convex-concave functions, DEG guarantees that the primal-dual gap converges to zero at a rate $O(\sqrt{\tau_{\max}}/\sqrt{T})$. The primal-dual gap is zero if and only if $(\bar{x}_T, \bar{y}_T)$ is a saddle point of $f(x, y)$ over the set $\mathcal{X} \times \mathcal{Y}$. Thus, DEG also guarantees convergence to a saddle-point under delays. The important thing to note here is that the $O(1/T)$ convergence rate of EG gets significantly slackened in the presence of delays; whether this is an artifact of our analysis or fundamental is an open question. The $O(1/\sqrt{T})$ rate of DEG mirrors the rate of GDA in the absence of delays. In the following sections, we will further explore this connection.

IV. Analysis of Delayed Gradient Descent-Ascent for Convex-Concave functions

In this section, we will examine the convergence of a delayed version of the gradient descent ascent algorithm that we refer to as DGDA. As before, we will continue to work under Assumptions 1 and 2. However, we will set $\mathcal{X} = \mathbb{R}^m$ and $\mathcal{Y} = \mathbb{R}^n$, i.e., as a departure from the previous section, the domains of the variables x and y are no longer assumed to be bounded. As we shall soon see, this makes the analysis more challenging relative to that in Section III.

To proceed, given any $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$, we will find it convenient to define a new variable $z = [x; y]$ that resides in $\mathbb{R}^{m+n}$. Next, corresponding to any $z = [x; y]$, let us define the function $\Phi : \mathbb{R}^{m+n} \to \mathbb{R}^{m+n}$ as follows:

$$\Phi(z) = \begin{bmatrix} \nabla_x f(x, y) \\ -\nabla_y f(x, y) \end{bmatrix}, \tag{15}$$

With these notations in place, we outline the steps of DGDA in Algorithm 2; the steps are self-explanatory.

Analysis of DGDA. In our analysis, we will make use of the following result from [16], stated for our purpose.

Lemma 4. *Let $\Phi(z)$ be as defined in Eq. (15), and suppose Assumption 2 holds for all $z \in \mathbb{R}^{m+n}$. Then, the following statements are true for any $z_1, z_2 \in \mathbb{R}^{m+n}$:*

1) $\langle \Phi(z_1) - \Phi(z_2), z_1 - z_2 \rangle \geq 0$,
2) *For any saddle-point $z^* = [x^*; y^*]$ of $f(x, y)$, we have $\Phi(z^*) = 0$.*

We start with a simple result that bounds the error $e_k \triangleq \Phi(z_k) - \Phi(z_{k-\tau_k})$ induced by delays as a function of the smoothness parameter L, the uniform bound on the gradients G, and the maximum delay $\tau_{\max}$.

Lemma 5. *Suppose Assumptions 1 and 2 hold $\forall z \in \mathbb{R}^{m+n}$. Then, for any $k \in [T]$, the delay-induced error $e_k \triangleq \Phi(z_k) - \Phi(z_{k-\tau_k})$ for DGDA satisfies*

$$\|e_k\| \leq 2\alpha L G \tau_{\max}. \tag{16}$$

Proof. For any two points $z = [x; y]$ and $\hat{z} = [\hat{x}; \hat{y}]$, we have

$$\|\Phi(z) - \Phi(\hat{z})\|^2 \overset{(a)}{\leq} 2 \left(L(\|x - \hat{x}\| + \|y - \hat{y}\|) \right)^2 \leq 4L^2 \|z - \hat{z}\|^2, \tag{17}$$

where we used Assumption 2 for the first inequality. Based on the above inequality, we have

$$\begin{aligned} \|e_k\| &= \|\Phi(z_{k-\tau_k}) - \Phi(z_k)\| \\ &\leq 2L \|z_{k-\tau_k} - z_k\| \\ &\leq 2L \sum_{j=k-\tau_k}^{k-1} \|z_{j+1} - z_j\| \\ &\leq 2\alpha L \sum_{j=k-\tau_k}^{k-1} \|\Phi(z_{j-\tau_j})\| \leq 2\alpha L G \tau_{\max}, \end{aligned} \tag{18}$$

where the final step follows from Assumption 1. □

Unlike the analysis in Section III where the boundedness of the domain implied bounded iterates, we need to do more work to establish that the iterates generated by DGDA remain bounded. Leveraging Lemma 5, the following result establishes this key fact.

Lemma 6. *Suppose Assumptions 1 and 2 hold $\forall z \in \mathbb{R}^{m+n}$. Let $z^* = [x^*; y^*]$, and suppose the step-size α satisfies*

$$\alpha \leq \frac{1}{2\sqrt{LG\tau_{\max} T}}.$$

Then, for the DGDA algorithm, the following holds $\forall k \in [T]$:

$$\|z_k - z^*\|^2 \leq 10B, \text{ where } B = \max\{\|z_1 - z^*\|^2, G\}. \tag{19}$$

Proof. From Eq. (14) and the definition of e_k, we have

$$\begin{aligned}\|z_{k+1}-z\|^2=&\|z_k-\alpha\Phi(z_k)-z\|^2\\&+\alpha^2\|e_k\|^2+2\alpha\langle e_k,z_k-z-\alpha\Phi(z_k)\rangle\\=&\|z_k-z\|^2+\alpha^2\|\Phi(z_k)\|^2-2\alpha\langle\Phi(z_k),z_k-z\rangle\\&+\alpha^2\|e_k\|^2+2\alpha\langle e_k,z_k-z-\alpha\Phi(z_k)\rangle\\\leq&\|z_k-z\|^2+2\alpha^2G^2-2\alpha\langle\Phi(z_k),z_k-z\rangle\\&+4\alpha^4G^2L^2\tau_{\max}^2\underbrace{+2\alpha\langle e_k,z_k-z\rangle}_{T_1}\underbrace{-2\alpha^2\langle e_k,\Phi(z_k)\rangle}_{T_2}.\end{aligned}\tag{20}$$

We now proceed to bound T_1 and T_2. For T_2, we have:

$$\begin{aligned}T_2\overset{(a)}{\leq}&\alpha^2\|e_k\|^2+\alpha^2\|\Phi(z_k)\|^2\\\overset{(b)}{\leq}&4\alpha^4G^2L^2\tau_{\max}^2+2\alpha^2G^2,\end{aligned}$$

where (a) follows from the elementary fact that for any two scalars $c,d\in\mathbb{R}$, it holds that

$$cd\leq\frac{1}{2}c^2+\frac{1}{2}d^2.\tag{21}$$

Moreover, for (b), we used Lemma 5 and Assumption 1. For bounding T_1, observe that

$$\begin{aligned}T_1&=2\alpha\langle e_k,z_k-z\rangle\\&\leq 2\alpha\|e_k\|\|z_k-z\|\\&\overset{(a)}{\leq}4\alpha^2GL\tau_{\max}\|z_k-z\|\\&=\left(2\alpha\sqrt{GL\tau_{\max}}\right)\left(2\alpha\sqrt{GL\tau_{\max}}\|z_k-z\|\right)\\&\overset{(b)}{\leq}2\alpha^2GL\tau_{\max}+2\alpha^2GL\tau_{\max}\|z_k-z\|^2,\end{aligned}\tag{22}$$

where we again appealed to Lemma 5 for (a). For (b), we used Eq. (21). Plugging in the above bounds on T_1 and T_2 into Eq. (20), simplifying using $L,G\geq 1$, and rearranging terms, we arrive at the following inequality:

$$\begin{aligned}2\alpha\langle\Phi(z_k),z_k-z\rangle\leq&\left(1+2\alpha^2LG\tau_{\max}\right)\|z_k-z\|^2\\&-\|z_{k+1}-z\|^2+A,\end{aligned}\tag{23}$$

where $A=2\alpha^2GL\tau_{\max}(1+2G+4\alpha^2GL\tau_{\max})$. Now setting $z=z^*$ in the above inequality, and noting that $\langle\Phi(z_k),z_k-z^*\rangle\geq 0$ based on Lemma 4, we obtain the following recursive inequality that holds for all $k\in[T]$:

$$\|z_{k+1}-z^*\|^2\leq\left(1+2\alpha^2LG\tau_{\max}\right)\|z_k-z^*\|^2+A.\tag{24}$$

Defining $r_k\triangleq\|z_k-z^*\|$, $\beta\triangleq\left(1+2\alpha^2LG\tau_{\max}\right)$, and iterating the above inequality, we obtain:

$$\begin{aligned}r_k^2&\leq\beta^{k-1}r_1^2+\left(\sum_{j=0}^{k-2}\beta^j\right)A\\&\leq\beta^{k-1}r_1^2+\frac{\beta^k}{\beta-1}A\\&\leq\beta^Tr_1^2+\frac{\beta^T}{\beta-1}A.\end{aligned}\tag{25}$$

We will now bound each of the terms above by using the elementary fact that for any $c\in\mathbb{R}$, it holds that $(1+c)\leq e^c$. When the step-size α satisfies

$$\alpha\leq\frac{1}{2\sqrt{LG\tau_{\max}T}},$$

we have

$$\beta^T\leq\left(1+\frac{1}{2T}\right)^T\leq e^{0.5}\leq 2.\tag{26}$$

Furthermore, it is easy to see that

$$\frac{A}{\beta-1}\leq\left(1+2G+\frac{1}{T}\right)\leq 4G.$$

Combining the above bounds leads to the claim of the lemma. This concludes the proof. □

Based on the above result, let us introduce a set $\mathcal{H}$ as follows:

$$\mathcal{H}\triangleq\{z|\|z-z^*\|^2\leq 10B\},\tag{27}$$

where $B=\max\{\|z_1-z^*\|^2,G\}$. From Lemma 6, we note that as long as the step-size α is chosen appropriately, the iterate sequence $\{z_k\}$ generated by DGDA belongs to $\mathcal{H}$. Moreover, $z^*\in\mathcal{H}$ trivially. With these observations in place, we now prove our main convergence result for DGDA for smooth convex-concave functions with bounded gradients.

Theorem 2. *Suppose Assumptions 1 and 2 hold $\forall x\in\mathbb{R}^m$ and $\forall y\in\mathbb{R}^n$. Let the step-size be chosen to satisfy*

$$\alpha=\frac{1}{2\sqrt{LG\tau_{\max}T}}.$$

Then, the iterates generated by DGDA satisfy:

$$\max_{y:(\bar{x}_T,y)\in\mathcal{H}}f(\bar{x}_T,y)-\min_{x:(x,\bar{y}_T)\in\mathcal{H}}f(x,\bar{y}_T)\leq 44B\sqrt{\frac{GL\tau_{\max}}{T}},$$

where $\bar{x}_T=(1/T)\sum_{k\in[T]}\hat{x}_k$, $\bar{y}_T=(1/T)\sum_{k\in[T]}\hat{y}_k$, and the set $\mathcal{H}$ is as defined in Eq. (27).

Proof. Recall the following notation from Lemma 6: $r_k=\|z_k-z^*\|$ and $B=\max\{r_1^2,G\}$. Let us start by noting that the choice of step-size in the statement of the theorem complies with that used to establish Lemma 6. Thus, we can invoke Lemma 6 to conclude that for any $z\in\mathcal{H}$, the following is true:

$$\|z_k-z\|^2\leq 2r_k^2+2\|z-z^*\|^2\leq 40B,\tag{28}$$

where the last inequality follows from the definition of the set $\mathcal{H}$. Using Eq. (23) from Lemma 6, we then have for any $z\in\mathcal{H}$:

$$\begin{aligned}2\alpha\langle\Phi(z_k),z_k-z\rangle&\leq\|z_k-z\|^2-\|z_{k+1}-z\|^2+A\\&\quad+2\alpha^2LG\tau_{\max}\|z_k-z\|^2\\&\leq\|z_k-z\|^2-\|z_{k+1}-z\|^2+\bar{A},\end{aligned}$$

where $\bar{A}=A+80\alpha^2LGB\tau_{\max}$, $A=2\alpha^2GL\tau_{\max}(1+2G+4\alpha^2GL\tau_{\max})$, and we used Eq. (28). Now summing the above inequality from $k=1$ to T, we obtain

$$\sum_{k=1}^T2\alpha\langle\Phi(z_k),z_k-z\rangle\leq\|z_1-z\|^2+\bar{A}T.\tag{29}$$

Moreover, from Proposition 1 in [15], we have

$$\sum_{k=1}^{T} 2\alpha\langle\Phi(z_k), z_k - z\rangle \geq 2\alpha T(f(\bar{x}_T, y) - f(x, \bar{y}_T)). \quad (30)$$

Combining the above display with Eq. (29) then yields the following bound $\forall z = [x; y] \in \mathcal{H}$:

$$f(\bar{x}_T, y) - f(x, \bar{y}_T) \leq \frac{\|z_1 - z\|^2}{2\alpha T} + \frac{\bar{A}}{2\alpha}. \quad (31)$$

Let us simplify the bound by first noting that for α chosen as in the statement of the theorem, it holds that $\bar{A} \leq 88\alpha^2 GBL\tau_{\max}$. Moreover, since $z \in \mathcal{H}$, we have

$$\|z_1 - z\|^2 \leq 2r_1^2 + 2\|z - z^*\|^2 \leq 22B.$$

Plugging in the above bounds in Eq. (31) then gives us:

$$f(\bar{x}_T, y) - f(x, \bar{y}_T) \leq \frac{11B}{\alpha T} + 44\alpha GBL\tau_{\max}. \quad (32)$$

The result follows from simply substituting the choice of α in the statement of the theorem. □

Discussion. The main message conveyed by Theorem 2 is that for smooth convex-concave functions with bounded gradients, the convergence rates of DGDA and DEG are identical in terms of their dependence on $\tau_{\max}$ and T. This complies with the intuition developed earlier in the paper that EG under delays behaves like GDA.

V. Analysis of Delayed Gradient Descent-Ascent for Strongly Convex-Strongly Concave functions

For smooth strongly convex-strongly concave functions, it is known that GDA guarantees linear convergence to the saddle point in the absence of delays [17]. In this section, we ask: *Does DGDA (Algorithm 2) also guarantee linear convergence to the saddle point for smooth strongly convex-strongly concave functions?* Our analysis in this section will provide an answer to this question in the affirmative. Moreover, we will precisely quantify how the maximum delay $\tau_{\max}$ slackens the exponent of linear convergence relative to when there is no delay. To get started, we now provide a formal definition of strongly convex-strongly concave functions.

Assumption 3. *The function $f(x, y)$ is μ-strongly convex-μ-strongly concave (SC-SC) over $\mathcal{X} \times \mathcal{Y}$, i.e., for all $x_1, x_2 \in \mathcal{X}$ and $y_1, y_2 \in \mathcal{Y}$, the following holds:*

$$f(x_2, y_1) \geq f(x_1, y_1) + \langle\nabla_x f(x_1, y_1), x_2 - x_1\rangle + \frac{\mu}{2}\|x_2 - x_1\|^2,$$
$$f(x_1, y_2) \leq f(x_1, y_1) + \langle\nabla_y f(x_1, y_1), y_2 - y_1\rangle - \frac{\mu}{2}\|y_2 - y_1\|^2.$$

Throughout this section, we will set $\mathcal{X} = \mathbb{R}^m$ and $\mathcal{Y} = \mathbb{R}^n$, i.e., we will make no assumption of bounded domains. Furthermore, unlike prior sections, we will drop the assumption of bounded gradients, i.e., we will no longer work under Assumption 1.

Analysis of DGDA. To proceed, we start by recalling two results from [17] that will play a crucial role in our subsequent analysis; at this point, we remind the reader of the definition of $\Phi(\cdot)$ in Eq. (15).

Lemma 7 ([17]). *Suppose Assumptions 2 and 3 hold. Then, $\forall z, \hat{z} \in \mathbb{R}^{m+n}$, we have*

$$L\|z - \hat{z}\|^2 \geq \langle\Phi(z) - \Phi(\hat{z}), z - \hat{z}\rangle \geq \mu\|z - \hat{z}\|^2. \quad (33)$$

Lemma 8 ([17]). *Suppose Assumptions 2 and 3 hold. Then, $\forall z, \hat{z} \in \mathbb{R}^{m+n}$, we have*

$$\langle\Phi(z) - \Phi(\hat{z}), z - \hat{z}\rangle \geq \frac{\mu}{4L^2}\|\Phi(z) - \Phi(\hat{z})\|^2. \quad (34)$$

Recall the definitions of iterate-suboptimality and delay-induced error: $r_k = \|z_k - z^*\|$ and $e_k = \Phi(z_k) - \Phi(z_{k-\tau_k})$. As before, our starting point will be to establish a bound on $\|e_k\|$. However, to establish a linear convergence rate, we need to provide a finer analysis relative to that in Lemmas 3 and 5. In particular, unlike these results which established uniform convergence bounds on $\|e_k\|$, we will instead seek to bound $\|e_k\|$ as a function of a suitably defined iterate-suboptimality-metric. Our next result formalizes this idea.

Lemma 9. *Suppose Assumptions 2 and 3 hold $\forall z \in \mathbb{R}^{m+n}$. Then, for any $k \in [T]$, the delay-induced error $e_k = \Phi(z_k) - \Phi(z_{k-\tau_k})$ for DGDA satisfies*

$$\|e_k\| \leq 2\alpha M_k, \quad (35)$$

where $M_k = L\tau_{\max}(\frac{4L^2}{\mu} + 4L)\max_{k-2\tau_{\max} \leq t \leq k} r_t$.

Proof. For bounding e_k, observe that

$$\begin{aligned}
\|e_k\| &= \|\Phi(z_{k-\tau_k}) - \Phi(z_k)\| \\
&\overset{(a)}{\leq} 2L\|z_{k-\tau_k} - z_k\| \\
&\leq 2L\sum_{j=k-\tau_k}^{k-1}\|z_{j+1} - z_j\| \\
&\leq 2\alpha L\sum_{j=k-\tau_k}^{k-1}\|\Phi(z_{j-\tau_j})\| \\
&\leq 2\alpha L\sum_{j=k-\tau_k}^{k-1}\left(\|\Phi(z_j)\| + \|\Phi(z_{j-\tau_j}) - \Phi(z_j)\|\right) \\
&\overset{(b)}{\leq} 2\alpha L\sum_{j=k-\tau_k}^{k-1}\left(\|\Phi(z_j)\| + 2L\|z_{j-\tau_j} - z_j\|\right) \\
&\leq 2\alpha L\sum_{j=k-\tau_k}^{k-1}\left(\|\Phi(z_j)\| + 2Lr_{j-\tau_j} + 2Lr_j\right).
\end{aligned} \quad (36)$$

From Lemma 4, we know that $\Phi(z^*) = 0$. Furthermore, from Lemma 8 and the Cauchy–Schwarz inequality, we obtain

$$\|\Phi(z_k)\|\|z_k - z^*\| \geq \langle\Phi(z_k), z_k - z^*\rangle \geq \frac{\mu}{4L^2}\|\Phi(z_k)\|^2,$$

which means

$$\|\Phi(z_k)\| \leq \frac{4L^2}{\mu}\|z_k - z^*\| = \frac{4L^2}{\mu}r_k,$$

Combining the above display with Eq. (36), we obtain

$$\begin{aligned}\|e_k\| &\leq 2\alpha L \sum_{j=k-\tau_{\max}}^{k-1} \left(\frac{4L^2}{\mu} r_j + 2L r_{j-\tau_j} + 2L r_j\right) \\ &\leq 2\alpha L \tau_{\max} \left(\frac{4L^2}{\mu} + 4L\right) \max_{k-2\tau_{\max}\leq t\leq k-1} r_t \\ &\leq 2\alpha M_k.\end{aligned} \tag{37}$$

□

We will also make use of the following key result.

Lemma 10 ([18]). *Suppose we have a sequence of non-negative real numbers, V_k, satisfying the inequality*

$$V_{k+1} \leq pV_k + q \max_{k-d(k)\leq \ell\leq k} V_\ell,$$

for some non-negative constants p and q, where $k \geq 0$ and $0 \leq d(k) \leq d_{\max}$ for some positive constant $d_{\max}$. If $p+q < 1$, then we have

$$V_k \leq r^k V_0, \text{ where } r = (p+q)^{1/(1+d_{\max})}.$$

We now prove our main result for DGDA for the class of smooth strongly convex-strongly concave (SC-SC) functions.

Theorem 3. *Suppose Assumptions 2 and 3 hold $\forall z \in \mathbb{R}^{m+n}$. Let the step-size be chosen to satisfy*

$$\alpha = \frac{\mu^3}{1536 L^6 \tau_{max}^2}.$$

Then, the iterates generated by DGDA satisfy:

$$r_k \leq \left(1 - \frac{\mu^4}{3072 L^6 \tau_{\max}^2}\right)^{\frac{k-1}{6\tau_{\max}}} r_1, \tag{38}$$

where $r_k = \|z_k - z^*\|$.

Proof. From the update rule of the DGDA algorithm and Lemma 9, we have

$$\begin{aligned}&\|z_{k+1} - z^*\|^2 - (1-\alpha\mu)\|z_k - z^*\|^2 = \\ &\alpha\mu\|z_k - z^*\|^2 - 2\alpha\langle \Phi(z_{k-\tau_k}), z_k - z^*\rangle + \alpha^2\|\Phi(z_{k-\tau_k})\|^2 \\ &\leq \alpha\mu\|z_k - z^*\|^2 - 2\alpha\langle \Phi(z_k), z_k - z^*\rangle + 2\alpha^2\|\Phi(z_k)\|^2 \\ &\quad + 2\alpha\langle e_k, z_k - z^*\rangle + 2\alpha^2\|e_k\|^2 \\ &\leq \underbrace{\alpha\mu\|z_k - z^*\|^2 - 2\alpha\langle \Phi(z_k), z_k - z^*\rangle + 2\alpha^2\|\Phi(z_k)\|^2}_{f_k} \\ &\quad + \underbrace{4\alpha^2 M_k r_k + 8\alpha^4 M_k^2}_{p_k}.\end{aligned} \tag{39}$$

From Lemmas 7 and 8, we further know that

$$\langle \Phi(z_k), z_k - z^*\rangle \geq \mu\|z_k - z^*\|^2, \text{ and}$$
$$\langle \Phi(z_k), z_k - z^*\rangle \geq \frac{\mu}{4L^2}\|\Phi(z_k)\|^2.$$

When $\alpha \leq \frac{\mu}{8L^2}$ - a requirement met by the choice of step-size in the statement of the theorem - it is easy to verify that the above equations imply $f_k \leq 0$, where f_k is as in Eq. (39). We also have

$$p_k \leq 12\alpha^2 M_k^2 \leq \alpha^2 C \left(\max_{k-2\tau_{\max}\leq t\leq k} r_t^2\right),$$

where $C = 768\frac{L^6}{\mu^2}\tau_{\max}^2$, and we used $L \geq \mu$ for simplifications. From the above discussion, we conclude that

$$r_{k+1}^2 \leq (1-\alpha\mu) r_k^2 + \alpha^2 C \left(\max_{k-2\tau_{\max}\leq t\leq k} r_t^2\right).$$

From the choice of step-size in the statement of the theorem, it is easy to verify that $1 - \alpha\mu + \alpha^2 C = 1 - 0.5\alpha\mu < 1$. Thus, we can immediately apply Lemma 10 to arrive at the desired conclusion. This concludes the proof. □

Discussion. Theorem 3 reveals that for smooth SC-SC functions, DGDA guarantees *linear convergence of the iterates to the saddle-point*. The result also clearly demonstrates how the exponent of convergence gets affected by $\tau_{\max}$.

References

[1] J. Von Neumann and O. Morgenstern, *Theory of games and economic behavior*. Princeton university press, 2007.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Comm. of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[3] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski, *Robust optimization*. Princeton university press, 2009, vol. 28.

[4] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[5] G. M. Korpelevich, "The extragradient method for finding saddle points and other problems," *Matecon*, vol. 12, pp. 747–756, 1976.

[6] A. Nedić and A. Ozdaglar, "Subgradient methods for saddle-point problems," *Journal of optimization theory and applications*, vol. 142, no. 1, pp. 205–228, 2009.

[7] C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng, "Training gans with optimism," *arXiv preprint arXiv:1711.00141*, 2017.

[8] A. Mokhtari, A. Ozdaglar, and S. Pattathil, "A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 1497–1507.

[9] J. C. Duchi, S. Chaturapruek, and C. Ré, "Asynchronous stochastic convex optimization," *arXiv preprint arXiv:1508.00882*, 2015.

[10] T. T. Doan, C. L. Beck, and R. Srikant, "On the convergence rate of distributed gradient methods for finite-sum optimization under communication delays," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 1, no. 2, pp. 1–27, 2017.

[11] Y. Arjevani, O. Shamir, and N. Srebro, "A tight convergence analysis for stochastic gradient descent with delayed updates," in *Algorithmic Learning Theory*. PMLR, 2020, pp. 111–132.

[12] S. U. Stich and S. P. Karimireddy, "The error-feedback framework: Better rates for sgd with delayed gradients and compressed communication," *arXiv preprint arXiv:1909.05350*, 2019.

[13] A. Koloskova, S. U. Stich, and M. Jaggi, "Sharper convergence guarantees for asynchronous sgd for distributed and federated learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17 202–17 215, 2022.

[14] A. Adibi, A. Mitra, G. J. Pappas, and H. Hassani, "Distributed statistical min-max learning in the presence of byzantine agents," in *Proc. of the 61st IEEE Conference on Decision and Control*, 2022, pp. 4179–4184.

[15] A. Mokhtari, A. E. Ozdaglar, and S. Pattathil, "Convergence rate of $o(1/k)$ for optimistic gradient and extragradient methods in smooth convex-concave saddle point problems," *SIAM Journal on Optimization*, vol. 30, no. 4, pp. 3230–3251, 2020.

[16] A. Nemirovski, "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*, vol. 15, no. 1, pp. 229–251, 2004.

[17] A. Fallah, A. Ozdaglar, and S. Pattathil, "An optimal multistage stochastic gradient method for minimax problems," in *Proc. of the 59th IEEE Conference on Decision and Control*, 2020, pp. 3573–3579.

[18] M. Gurbuzbalaban, A. Ozdaglar, and P. A. Parrilo, "On the convergence rate of incremental aggregated gradient algorithms," *SIAM Journal on Optimization*, vol. 27, no. 2, pp. 1035–1048, 2017.