Convergence Rates of Two-Time-Scale Gradient Descent-Ascent Dynamics for Solving Nonconvex Min-Max Problems

Thinh T. Doan

Abstract

There are much recent interests in solving nonconvex min-max optimization problems due to its broad applications in many areas including machine learning, networked resource allocations, and distributed optimization. Perhaps, the most popular first-order method in solving min-max optimization is the so-called simultaneous (or single-loop) gradient descent-ascent algorithm due to its simplicity in implementation. However, theoretical guarantees on the convergence of this algorithm is very sparse since it can diverge even in a simple bilinear problem.

In this paper, our focus is to characterize the finite-time performance (or convergence rates) of the continuous-time variant of simultaneous gradient descent-ascent algorithm. In particular, we derive the rates of convergence of this method under a number of different conditions on the underlying objective function, namely, two-sided Polyak-Łojasiewicz (PL), one-sided PL, nonconvex-strongly concave, and strongly convex-nonconcave conditions. Our convergence results improve the ones in prior works under the same conditions of objective functions. The key idea in our analysis is to use the classic singular perturbation theory and coupling Lyapunov functions to address the time-scale difference and interactions between the gradient descent and ascent dynamics. Our results on the behavior of continuous-time algorithm may be used to enhance the convergence properties of its discrete-time counterpart.

1 Introduction

In this paper, we consider the following min-max optimization problems

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

where \( f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R} \) is a nonconvex function w.r.t \( x \) for a fixed \( y \) and (possibly) nonconcave w.r.t \( y \) for a fixed \( x \). The min-max problem has received much interests for years due to its broad applications in different areas including control, machine learning, and economics. In particular, many problems in these areas can be formulated as problem (1), for example, game theory [1, 2], stochastic control and reinforcement learning [3, 4], training generative adversarial networks (GANs) [5, 6], adversarial and robust machine learning [7, 8], resource allocation over networks [9], and distributed optimization [10, 11]; to name just a few.

In the existing literature, there are two types of iterative first-order methods for solving problem (1), namely, nested-loop algorithms and single-loop algorithms. Nested-loop algorithms implement multiple inner steps in each iteration to solve the maximization problem either exactly or approximately. However, this approach is not applicable to the setting when \( f(x, y) \) is nonconcave in \( y \),

*Thinh T. Doan is with the Bradley Department of Electrical and Computer Engineering, Virginia Tech, USA. Email: thinhdoan@vt.edu
since the maximization problem is NP-hard. Only finding a stationary point of the maximization problem is likely to affect the quality of solving the minimization problem.

On the other hand, single-loop algorithm simultaneously updates the iterates $x$ and $y$ by using the vanilla gradient descent and ascent steps at different time scales, respectively. As a result, this algorithm is applicable to more general settings and more practical due to its simplicity in implementation. However, single-loop algorithms may not converge in many settings, for example, they fail to converge even in a simple bilinear zero-sum game [12]. Indeed, theoretical guarantees of these methods are very sparse.

Our focus in this paper is to study the continuous-time variant of the single-loop gradient descent-ascent method for solving problem (1). Considering the continuous-time variant will help us to have a better understanding about the behavior of this method through studying the convergence of the corresponding differential equations using Lyapunov theory. Such an understanding can then be used to enhance the analysis of the discrete-time algorithms, as recently observed in the single objective optimization counterpart [13, 14, 15, 16]. Our main contributions are summarized below.

**Main Contributions.** The focus of this paper is to study the performance of the continuous-time gradient descent-ascent dynamics in solving nonconvex min-max optimization problems. In particular, we derive the rates of convergence of this method under a number of different conditions on the underlying objective function, namely, two-sided Polyak-Łojasiewicz (PL), one-sided PL, nonconvex-strongly concave, and strongly convex-nonconcave conditions. These rates are summarized in Table 1 and presented in detail in Section 3, where we show that our results improve the ones in prior works under the same conditions of objective functions. The key idea in our analysis is to use the classic singular perturbation theory and coupling Lyapunov function of the fast and slow dynamics to address the time-scale difference and interactions between the gradient descent and ascent dynamics. Proper choices of step sizes allows us to derive improved convergence properties of the two-time-scale gradient descent-ascent dynamics.

### 1.1 Related Works

**Convex-Concave Settings.** Given the broad applications of problem (1), there are a large number of works to study algorithms and their convergence in solving this problem, especially in the context of convex-concave settings. Some examples include prox-method and its variant [17, 18, 19, 20, 21], extragradient and optimistic gradient methods [22, 23, 24, 25, 26, 27], and recently Hamiltonian gradient descent methods [6, 12, 28]. Some algorithms in these settings have convergence rates matched with the lower bound complexity; see the recent work [26] for a detailed discussion.

**Nonconvex-Concave Settings.** Unlike the convex-concave settings, algorithmic development and theoretical understanding in the general nonconvex settings are very limited. Indeed, finding the global optimality of nonconvex-nonconcave problem is NP-hard, or at least as hard as solving a single nonconvex objective problem. As a result, the existing literature often aims to find a stationary point of $f$ when the max problem is concave. For example, multiple-loop algorithms have been studied in [29, 30, 31, 32, 33]. Our work in this paper is closely related to the recent literature on studying single-loop algorithm [34, 35, 36, 37, 38]. While these works study discrete-time algorithms, we consider continuous-time counterpart. We will show that for some settings, our approach improves the existing convergence results.

**Other Settings.** We also want to mention some related literature in game theory [39, 40, 41, 42, 43], two-time-scale stochastic approximation [44, 45, 46, 47, 48, 49, 50, 51, 52, 53], reinforcement
learning [54, 55, 56, 57, 58], two-time-scale optimization [59, 60], and decentralized optimization [61, 62, 63, 64, 65, 66, 67]. These works study different variants of two-time-scale methods mostly for solving a single optimization problem, and often aim to find global optimality (or fixed points) using different structure of the underlying problems (e.g., Markov structure in stochastic games and reinforcement learning or strong monotonicity in stochastic approximation). As a result, their techniques may not be applicable to the context of problem (1) considered in the current paper.

**Notation.** Given any vector $x$ we use $\|x\|$ to denote its 2-norm. We denote by $\nabla_x f(x, y)$ and $\nabla_y f(x, y)$ the partial gradients of $f$ with respect to $x$ and $y$, respectively.

## 2 Two-Time-Scale Gradient Descent-Ascent Dynamics

For solving problem (1), we are interested in studying two-time-scale gradient descent-ascent dynamics (GDAD), where we implement simultaneously the following two differential equations

$$\dot{x}(t) = \frac{d}{dt} x(t) = -\alpha \nabla_x f(x(t), y(t)), \quad \dot{y}(t) = \frac{d}{dt} y(t) = \beta \nabla_y f(x(t), y(t)), \quad (2)$$

Here, $\alpha, \beta$ are two step sizes, whose values will be specified later. In the convex-concave setting, one can choose $\alpha = \beta$. However, as observed in [68], choosing different step sizes achieves a better convergence in the context of nonconvex problem. Indeed, we will choose $\alpha \ll \beta$ since in our settings studied in the following sections, the maximization problem is often easier to solve than the minimization problem. In this case, the dynamic of $y(t)$ is implemented at a faster time scale (using larger step sizes) than $x(t)$ (using smaller step sizes). The time-scale difference is loosely defined as the ratio $\alpha/\beta \ll 1$. Thus, one has to design these two step sizes properly so that the method converges as fast as possible.

**Technical Approach.** The convergence analysis of (2) studied in this paper is mainly motivated by the classic singular perturbation theory [69]. The main idea of our approach can be explained as follows. Since $y$ is implemented at a faster time scale than $x$, one can consider $x(t) = x$ being fixed in $\dot{y}$ and separately study the stability of the system $\dot{y}$ using Lyapunov theory. Let $V_2$ be the Lyapunov function corresponding to $\dot{y}$. When $\dot{y}$ converges to an equilibrium $y$ (e.g., $\nabla_y f(x, y) = 0$), one can fix $y(t) = y$ and study the stability of $\dot{x}$. Let $V_1$ be the corresponding Lyapunov function of $\dot{x}$. We note that $V_1$ and $V_2$ both depend on $x$ and $y$, as a result, their time derivatives are coupled through the dynamics in (2). Addressing this coupling and the time-scale difference between the two dynamics is the key idea in our approach. To do that, we will consider the following Lyapunov function

$$V(x, y) = V_1(x, y) + \frac{\gamma \alpha}{\beta} V_2(x, y), \quad (3)$$

where $\alpha/\beta$ represents the time-scale difference, while the constant $\gamma$ will be properly chosen to eliminate the impact of $x$ on the convergence of $y$ and vice versa. Proper choices of these constants will also help us to derive the convergence rates of (2). Similar approach has been used in different settings of two-time-scale methods, see for example [53, 66].

We conclude this section by introducing two assumptions for our analysis studied later.
Assumption 1. The function $f(\cdot, \cdot)$ has Lipschitz continuous gradients for each variable, i.e., there exist positive constants $L_x, L_y,$ and $L_{xy}$ such that for all $x, x_2 \in \mathbb{R}^m, y_1, y_2 \in \mathbb{R}^n$ we have

$$
\|\nabla_x f(x_1, y_1) - \nabla_x f(x_2, y_2)\| \leq L_x \|x_1 - x_2\| + L_{xy} \|y_1 - y_2\|,
$$

$$
\|\nabla_y f(x_1, y_1) - \nabla_y f(x_2, y_2)\| \leq L_{xy} \|x_1 - x_2\| + L_y \|y_1 - y_2\|.
$$

(4)

Assumption 2. Given any $x$ the problem $\max_y f(x, y)$ has a nonempty solution set $\mathcal{Y}^*(x)$, i.e., there exists $y^*(x) \in \mathcal{Y}^*(x)$ such that

$$
y^*(x) = \arg\max_{y \in \mathbb{R}^n} f(x, y), \quad \text{where } f(x, y^*(x)) \text{ is finite.}
$$

Table 1: Convergence rates of GDAD for solving (1) given some accuracy $\epsilon > 0$. The abbreviations NCvx, NCave, SCvx, SCave, and PL stand for nonconvex, nonconcave, strongly convex, strongly concave, and Polyak-Lojasiewicz conditions, respectively. Condition number $\kappa$ is defined in (11), and $R$ is the size of compact set used in [33].

| Objectives | Prior Works | This Paper |
|------------|-------------|------------|
| PL & PL    | $O\left(\kappa^3 \log\left(\frac{1}{\epsilon}\right)\right)$ [36] | $O\left(\kappa^2 \log\left(\frac{1}{\epsilon}\right)\right)$ |
| NCvx & PL  | $O\left(R^2 L_{xy} \log\left(\frac{1}{\epsilon}\right)\epsilon^{-2}\right)$ [33] | $O\left(L_{xy}^2 \epsilon^{-2}\right)$ |
| NCvx & SCave| $O\left(L_{xy}^2 \epsilon^{-2}\right)$ [37] | $O\left(L_{xy}^2 \epsilon^{-2}\right)$ |
| SCvx & NCave| $O\left(L_{xy}^2 \epsilon^{-2}\right)$ [37] | $O\left(L_{xy}^2 \epsilon^{-2}\right)$ |

3 Main Results

In this section, we present the main results of this paper, where we derive the convergence rates of GDAD under different conditions on the objective function $f(x, y)$. Our results are summarized in Table 1. First, our approach improves the analysis in [36], where we show in Section 3.1 that for two-sided PL functions the convergence of GDAD only scales with $\kappa^2$ instead of $\kappa^3$ studied in [36]. Our result addresses the conjecture raised in [36], where the authors state that such an improvement may not be possible. Second, our analysis achieves a better result than the one in [33] for the case of one-sided PL function by a factor of $\log(1/\epsilon)$. We note that a nested-loop is studied in [33] while GDAD is a single-loop method. Finally, our result is the same as the one in [37] when $f(x, y)$ is either strongly concave in $y$ for fixed $x$. In Section 3.4, we will show that this observation also holds when $f(x, y)$ is either strongly convex in $x$ and nonconcave in $y$. Note that as compared to the analysis in [37], we use a simpler analysis and simpler choice of step sizes to achieve these results.

3.1 Two-Sided Polyak–Lojasiewicz Conditions

We first study the convergence rates of GDAD when $f$ satisfies a two-sided Polyak–Lojasiewicz (PL) condition, which is considered in [36] and stated here for convenience.

**Definition 1** (Two-Sided PL Conditions). A continuously differentiable function $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ is called to satisfy two-sided PL conditions if there exist two positive constants $\mu_x$ and $\mu_y$ such that $\mu_x, \mu_y \leq \min\{L_x, L_y, L_{xy}\}$ the following conditions hold for all $(x, y) \in \mathbb{R}^m \times \mathbb{R}^n$:

$$
2\mu_x [f(x, y) - \min_x f(x, y)] \leq \|\nabla_x f(x, y)\|^2,
$$

$$
2\mu_y [\max_y f(x, y) - f(x, y)] \leq \|\nabla_y f(x, y)\|^2.
$$

(5)
The two-sided PL condition, which we will assume to hold in this subsection, is a generalized variant of the popular PL condition, proposed by [\textsuperscript{70}] as a sufficient condition to guarantee that the classic gradient descent method converges exponentially to the optimal value of an unconstrained minimization problem. As shown in [\textsuperscript{71}], the PL condition also implies the quadratic growth condition, i.e., given any $x$ we have
\[
\max_{z \in \mathbb{R}^m} f(x, z) - f(x, y) \geq \frac{\mu_y}{2} \|\mathcal{P}_{\mathcal{Y}^*(x)}[y] - y\|^2, \quad \forall y \in \mathbb{R}^m,
\]
where we assume that $\mathcal{Y}^*(x)$ is a nonempty solution set of $\max_y f(x, y)$ and $\mathcal{P}_{\mathcal{Y}^*(x)}[y]$ is the projection of $y$ to this set. More discussions on PL condition can be found in [\textsuperscript{36}].

Our focus in this section is to show that GDAD converges exponentially to the global min-max solution $(x^*, y^*)$ of $f$ under the two-sided PL condition. To do that, we consider the following assumption and lemmas, which are useful for our analysis considered later. We first consider an assumption on the existence of $(x^*, y^*)$, a global min-max solution of $f$.

**Assumption 3.** There exists a global min-max solution $(x^*, y^*)$ of $f$, i.e.,
\[
x^* = \arg \min_{x \in \mathbb{R}^m} f(x, y^*) \quad \text{and} \quad y^* = \arg \max_{y \in \mathbb{R}^n} f(x^*, y).
\]

Next, we consider the following lemma about the Lipschitz continuity of the gradient of $f(x, y^*(x))$, which is a variant of the well-known Danskin lemma [\textsuperscript{72}][Proposition B.25] and studied in [\textsuperscript{33}][Lemma A.5].

**Lemma 1.** Suppose that Assumptions 1–3 hold. Then, the function $\max_y f(x, y)$ is differentiable and its gradient $\nabla_x f(x, y^*(x))$ is Lipschitz continuous with a constant $L_x + \frac{L_{xy}}{\mu_y}$.

Finally, for our analysis we consider the following two Lyapunov functions
\[
V_1(x) = \max_{y \in \mathbb{R}^n} f(x, y) - \min_{x \in \mathbb{R}^m} \max_{y \in \mathbb{R}^n} f(x, y),
\]
\[
V_2(x, y) = \max_{y \in \mathbb{R}^n} f(x, y) - f(x, y),
\]
where it is obvious to see that $V_1$ and $V_2$ are nonnegative. The time derivatives of $V_1$ and $V_2$ over the trajectories $\dot{x}$ and $\dot{y}$ are given in the following lemma, whose proof can be found in Section 4.1.

**Lemma 2.** Suppose that Assumptions 1–3 hold. Then we have
\[
\dot{V}_1(x) \leq -\frac{\alpha}{2} \|\nabla_x f(x, y^*(x))\|^2 + \frac{L_{xy}^2 \alpha}{\mu_y} V_2(x, y).
\]
\[
\dot{V}_2(x, y) \leq -\beta \|\nabla_y f(x, y)\|^2 + \frac{3\alpha}{2} \|\nabla_x f(x, y^*(x))\|^2 + \frac{5L_{xy}^2 \alpha}{\mu_y} V_2(x, y).
\]

As mentioned, the dynamics of $\dot{x}$ and $\dot{y}$ are implemented at different time scales, where this difference is often loosely defined as the ratio $\beta/\alpha > 1$. To capture such time-scale difference in our analysis, we will utilize the coupling Lyapunov function defined in (3). We denote by $\mu = \min\{\mu_x, \mu_y\}$ and the condition number
\[
\kappa = \frac{L_{xy}}{\mu} \geq 1.
\]
representing the condition number of $f(x, y)$. The convergence rate of GDAD under the two-sided PL condition is formally stated in the following theorem.
\textbf{Theorem 1.} Suppose that Assumptions 1–3 hold. Let $\gamma, \alpha, \beta$ be chosen as

$$\gamma = \frac{L_{xy}^2}{\mu_y^2}, \quad \alpha = \frac{\mu^2}{10\mu_x L_{xy}^2}, \quad \beta = \frac{\mu^2}{\mu_x \mu_y^2}. \quad (12)$$

Then we have for all $t \geq 0$

$$V(x(t), y(t)) \leq e^{-2\mu_y t} V(x(0), y(0)). \quad (13)$$

Proof. By (5) we have

$$\|\nabla_y f(x, y)\|^2 \geq 2\mu_y [\max_y f(x, y) - f(x, y)] = 2\mu_y V_2(x, y).$$

Thus, by using (9), (10), (3), and the preceding relation we have

$$\dot{V}(x(t), y(t)) = \dot{V}_1(x(t)) + \frac{\gamma \alpha}{\beta} \dot{V}_2(x(t), y(t))$$

$$\leq -\frac{\alpha}{2} \|\nabla f(x(t), y^*(x(t)))\|^2 + \frac{L_{xy}^2 \alpha}{\mu_y} V_2(x(t), y(t))$$

$$- 2\mu_y \gamma V_2(x(t), y(t)) + \frac{3\gamma \alpha^2}{2\beta} \|\nabla f(x(t), y^*(x(t)))\|^2 + \frac{5L_{xy}^2 \gamma \alpha^2}{\mu_y \beta} V_2(x(t), y(t))$$

$$= -\frac{\alpha}{4} \|\nabla f(x(t), y^*(x(t)))\|^2 - \frac{\mu_y \gamma \alpha}{2} V_2(x(t), y(t))$$

$$- \left( 1 - \frac{3\gamma \alpha}{\beta} \right) \frac{\alpha}{2} \|\nabla f(x(t), y^*(x(t)))\|^2$$

$$- \left( \frac{3\mu_y \gamma}{2} - \frac{L_{xy}^2}{\mu_y} - \frac{5L_{xy}^2 \gamma \alpha}{\mu_y \beta} \right) \alpha V_2(x(t), y(t)). \quad (14)$$

Using (12) we have

$$\frac{1}{2} - \frac{3\gamma \alpha}{\beta} = \frac{1}{2} - \frac{3L_{xy}^2}{\mu_y^2} \frac{\mu^2}{10\mu_x L_{xy}^2} \frac{\mu_x \mu_y^2}{\mu^2} = \frac{1}{5},$$

$$\frac{3\mu_y \gamma}{2} - \frac{L_{xy}^2}{\mu_y} - \frac{5L_{xy}^2 \gamma \alpha}{\mu_y \beta} = \frac{L_{xy}^2}{2\mu_y} - \frac{L_{xy}^2}{2\mu_y} = 0,$$

which when substituting into (14) and using (5) and $y^*(x) = \arg \max_y f(x, y)$ we obtain

$$\dot{V}(x(t), y(t)) \leq -\frac{\alpha}{4} \|\nabla_x f(x(t), y^*(x(t)))\|^2 - \frac{\mu_y \gamma \alpha}{2} V_2(x(t), y(t))$$

$$\leq -\frac{\mu_x \alpha}{2} \left[ \max_y f(x(t), y) - \min_x f(x, y^*(x(t))) \right] - \frac{\mu_y \gamma \alpha}{2} V_2(x, y)$$

$$\leq -\frac{\mu_x \alpha}{2} (\max_y f(x(t), y) - \max_x f(x, y)) - \frac{\mu_y \gamma \alpha}{2} V_2(x, y)$$

$$= -\frac{\mu_x \alpha}{2} V_1(x(t)) - \frac{\mu_y \beta \gamma \alpha}{2} V_2(x, y)$$

$$\leq -\frac{\mu_x \alpha}{2} (V_1(x(t)) + \frac{\gamma \alpha}{\beta} V_2(x(t), y(t))) = -\frac{\mu_x \alpha}{2} V(x(t), y(t)), \quad \text{where the last inequality is due to}$$

$$\mu_x \alpha = \frac{\mu^2}{10L_{xy}^2} \leq \mu_y \beta = \frac{\mu^2}{\mu_x \mu_y}.$$
Taking the integral on both sides of the equation above immediately gives (13), i.e.,
\[
V(x(t), y(t)) \leq e^{-\frac{\mu_y t}{2}} V(x(0), y(0)) = e^{-\frac{t}{20\sqrt{T}}} V(x(0), y(0)).
\]

\[\square\]

### 3.2 Nonconvex–Polyak-Łojasiewicz Conditions

In this subsection, we consider an extension of the result studied in the previous section, where we assume that the objective function \(f(x, \cdot)\) satisfies the Polyak-Łojasiewicz condition given any \(x\) and \(f(\cdot, y)\) is nonconvex given any \(y\).

**Assumption 4** (One-Sided PL Conditions). We assume that \(f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}\) is nonconvex in \(x\) for any fixed \(y\) and satisfies the PL condition in \(y\) for any fixed \(x\), that is, there exists a positive constants \(\mu_y\) such that the following condition hold for any \(x \in \mathbb{R}^m\):

\[2\mu_y \max_{\gamma \not= y} f(x, \gamma) - f(x, y) \leq \|\nabla_y f(x, y)\|^2. \tag{15}\]

Since \(f\) satisfies only one-sided PL condition, we are giving up the hope to find a global optimal solution of (1), as studied in Theorem 1. In stead, we will show that GDAD will return a stationary point of \(f\), as studied in [33]. Note that under Assumption 2 the result in Lemma 1 still holds since the work in [33] only assumes one-sided PL condition. In addition, since we relax the two-sided PL condition, we introduce the following two Lyapunov functions for our analysis studied later.

\[V_1(x, y) = f(x, y) - \min_{x \in \mathbb{R}^m} \min_{y \in \mathbb{R}^n} f(x, y), \tag{16}\]
\[V_2(x, y) = \max_{y} f(x, y) - f(x, y), \tag{17}\]

where it is obvious to see that \(V_1\) and \(V_2\) are nonnegative. The time derivatives of \(V_1\) and \(V_2\) over the trajectories \(\dot{x}\) and \(\dot{y}\) are given in the following lemma, whose proof is presented in Section 4.2.

**Lemma 3.** Suppose that Assumptions 1, 2, and 4 hold. Then we have

\[\dot{V}_1(x(t), y(t)) = -\alpha \|\nabla_x f(x(t), y(t))\|^2 + \beta \|\nabla_y f(x(t), y(t))\|^2. \tag{18}\]
\[\dot{V}_2(x(t), y(t)) \leq -\beta \|\nabla_y f(x(t), y(t))\|^2 + \frac{\alpha}{2} \|\nabla_x f(x(t), y(t))\|^2 + \frac{L_{xy}^2 \alpha}{\mu_y} V_2(x(t), y(t)). \tag{19}\]

Similar to the previous subsection, we utilize the following coupling Lyapunov function

\[V(x, y) = V_1(x, y) + \frac{\gamma \alpha}{\beta} V_2(x, y), \tag{20}\]

for some constant \(\gamma\), which will be defined below. The convergence rate of GDAD under the nonconvex-PL condition is formally stated in the following theorem.

**Theorem 2.** Suppose that Assumptions 1, 2, and 4 hold. Let \(\gamma, \alpha, \beta\) be chosen as

\[\gamma = \frac{32L_{xy}^2}{\mu_y^2}, \quad \alpha = \frac{1}{8L_{xy}^2}, \quad \beta = \frac{1}{\mu_y^2}. \tag{21}\]

Then we have for all \(T \geq 0\)

\[\min_{0 \leq t \leq T} \|\nabla_x f(x(t), y(t))\| \leq \frac{4L_{xy}^2 \sqrt{V_1(x(0), y(0)) + 4V_2(x(0), y(0))}}{\sqrt{T}}. \tag{22}\]
Proof. By using (18), (19), and (20) we have

\[ \dot{V}(x(t), y(t)) = \dot{V}_1(x(t)) + \frac{\gamma \alpha}{\beta} \dot{V}_2(x(t), y(t)) \]

\[ \leq -\alpha \|\nabla_x f(x(t), y(t))\|^2 + \beta \|\nabla_y f(x(t), y(t))\|^2 \]

\[ -\gamma \alpha \|\nabla_y f(x(t), y(t))\|^2 + \frac{\gamma \alpha^2}{8\beta} \|\nabla_x f(x(t), y(t))\|^2 + \frac{4L^2_{xy} \gamma \alpha^2}{\mu_y \beta} V_2(x, y) \]

\[ = -\frac{\alpha}{2} \|\nabla_x f(x(t), y(t))\|^2 + \frac{\gamma \alpha^2}{8\beta} \|\nabla_x f(x(t), y(t))\|^2 \]

\[ -\frac{\gamma \alpha}{4} \|\nabla_y f(x(t), y(t))\|^2 + \beta \|\nabla_y f(x(t), y(t))\|^2 \]

\[ \leq -\frac{\alpha}{2} \|\nabla_x f(x(t), y(t))\|^2 + \frac{\gamma \alpha^2}{8\beta} \|\nabla_x f(x(t), y(t))\|^2 \]

\[ -\frac{\gamma \alpha}{4} \|\nabla_y f(x(t), y(t))\|^2 + \beta \|\nabla_y f(x(t), y(t))\|^2 \]

\[ -\frac{\mu_y \gamma \alpha}{2} \left(1 - \frac{4L^2_{xy} \alpha}{\mu_y \beta^2}\right) V_2(x, y) - \frac{\alpha}{2} \left(1 - \frac{\gamma \alpha^2}{4\beta}\right) \|\nabla_x f(x(t), y(t))\|^2, \tag{23} \]

where in the last inequality we use (15) to have

\[ \|\nabla_y f(x, y)\|^2 \geq 2\mu_y [\max_y f(x, y) - f(x, y)] = 2\mu_y V_2(x, y). \]

Using (21) and the preceding relation we have

\[ -\frac{\gamma \alpha}{4} + \beta = -\frac{1}{\mu_y^2} + \frac{1}{\mu_y^2} = 0, \quad 1 - \frac{4L^2_{xy} \alpha}{\mu_y \beta^2} = \frac{1}{2} \quad \text{and} \quad 1 - \frac{\gamma \alpha^2}{4\beta} = 0, \]

which when substituting into (23) gives

\[ \dot{V}(x(t), y(t)) \leq -\frac{\alpha}{2} \|\nabla_x f(x(t), y(t))\|^2 - \frac{\gamma \alpha}{2} \|\nabla_y f(x(t), y(t))\|^2. \]

Taking the integral on both sides over \( t \in [0, T] \) for some \( T \geq 0 \) and rearranging we obtain

\[ \frac{\alpha}{2} \int_{t=0}^{T} \|\nabla_x f(x(t), y(t))\|^2 dt + \frac{\gamma \alpha}{2} \int_{t=0}^{T} \|\nabla_y f(x(t), y(t))\|^2 dt \leq V(x(0), y(0)), \]

which since \( \gamma \geq 1 \) and by using (21) gives

\[ \min_{0 \leq t \leq T} \left\| \begin{array}{c} \nabla_x f(x(t), y(t)) \\ \nabla_y f(x(t), y(t)) \end{array} \right\| \leq \sqrt{\frac{2V(x(0), y(0))}{\alpha T}} \]

\[ \leq \frac{4L_{xy} \sqrt{V_1(x(0), y(0))} + 4V_2(x(0), y(0))}{\sqrt{T}}, \]

which concludes our proof. \( \square \)
3.3 Nonconvex–Strongly Concave Conditions

In this subsection, we study the rate of GDAD when the function \( f(x, y) \) is nonconvex given any \( y \) and strongly concave given any \( x \). In particular, we consider the following assumption.

**Assumption 5.** The objective function \( f(\cdot, y) \) is nonconvex for any given \( y \) and \( f(x, \cdot) \) is strongly concave with constant \( \mu_y > 0 \) for any given \( x \). The latter is equivalent to

\[
f(x_1, y_1) - f(x_2, y_2) - \langle \nabla f(x_2, y_2), y_1 - y_2 \rangle \leq -\frac{\mu_y}{2} \| y_1 - y_2 \|^2, \quad \forall y_1, y_2 \in \mathbb{R}^n. \tag{24}
\]

For our analysis of in this section, we introduce the following two Lyapunov functions

\[
V_1(x, y) = f(x, y) - \min_{(x, y)} f(x, y) \tag{25}
\]

\[
V_2(x, y) = \frac{1}{2} \| \dot{y} \|^2 = \frac{1}{2} \| \beta \nabla_y f(x, y) \|^2. \tag{26}
\]

The time derivatives of \( V_1 \) and \( V_2 \) over the trajectories \( \dot{x} \) and \( \dot{y} \) are given in the following lemma, whose proof is presented in Section 4.3.

**Lemma 4.** Suppose that Assumptions 1 and 5 hold. Then we have

\[
\dot{V}_1(x(t), y(t)) \leq -\frac{1}{\alpha} \| \dot{x}(t) \|^2 + \frac{1}{\beta} \| \dot{y}(t) \|^2. \tag{27}
\]

\[
\dot{V}_2(x(t), y(t)) \leq L_{xy} \beta \| \dot{y}(t) \| \| \dot{x}(t) \| - \mu_y \beta \| \dot{y}(t) \|^2. \tag{28}
\]

We next derive the convergence rate of GDAD under Assumption 5 in the following theorem, where we show that GDAD converges sublinear to a stationary point of \( f \).

**Theorem 3.** Suppose that Assumptions 1 and 5 hold. Let \( \gamma, \alpha, \beta \) be chosen as

\[
\gamma = \mu_y L_{xy}^2, \quad \alpha = \frac{1}{L_{xy}^2}, \quad \beta = \frac{4}{\mu_y}. \tag{29}
\]

Then we have for all \( T \geq 0 \)

\[
\min_{0 \leq t \leq T} \left\| \frac{\nabla_x f(x(t), y(t))}{\nabla_y f(x(t), y(t))} \right\| \leq L_{xy} \sqrt{2 V_1(x(0), y(0))} + \frac{2 L_{xy} \| \nabla_y f(x(0), y(0)) \|}{\sqrt{\mu_y T}}. \tag{30}
\]

**Proof.** By using (27) and (28) we consider

\[
\dot{V}(x(t), y(t)) = \dot{V}_1(x(t), y(t)) + \frac{\gamma \alpha}{\beta} \dot{V}_2(x(t), y(t))
\]

\[
\leq -\frac{1}{\alpha} \| \dot{x}(t) \|^2 + \frac{1}{\beta} \| \dot{y}(t) \|^2 + L_{xy} \gamma \alpha \| \dot{y}(t) \| \| \dot{x}(t) \| - \mu_y \gamma \alpha \| \dot{y}(t) \|^2
\]

\[
= -\frac{1}{2 \alpha} \| \dot{x}(t) \|^2 - \mu_y \gamma \alpha \frac{\gamma \alpha}{4} \| \dot{y}(t) \|^2 - (\mu_y \gamma \alpha - \frac{1}{\beta}) \| \dot{y}(t) \|^2
\]

\[
- \frac{1}{2 \alpha} \| \dot{x}(t) \|^2 + L_{xy} \gamma \alpha \| \dot{y}(t) \| \| \dot{x}(t) \| - \mu_y \gamma \alpha \frac{\gamma \alpha}{2} \| \dot{y}(t) \|^2
\]

\[
= -\frac{1}{2 \alpha} \| \dot{x}(t) \|^2 - \mu_y \gamma \alpha \frac{\gamma \alpha}{4} \| \dot{y}(t) \|^2
\]

\[
- \frac{1}{2 \alpha} \| \dot{x}(t) \|^2 + L_{xy} \gamma \alpha \| \dot{y}(t) \| \| \dot{x}(t) \| - \mu_y \gamma \alpha \frac{\gamma \alpha}{2} \| \dot{y}(t) \|^2, \tag{31}
\]
Thus, the preceding relation gives (29) to have
\[ \frac{\mu_y \gamma \alpha}{4} - \frac{1}{\beta} = \frac{\mu_y^2}{2} - \frac{\mu_y^2}{4} = 0. \]

Using (29) one more time we obtain
\[ -\frac{1}{2\alpha} \| \dot{x}(t) \|^2 + L_{xy} \gamma \alpha \| \dot{y}(t) \| \| \dot{y}(t) \| - \frac{\mu_y \gamma \alpha}{2} \| \dot{y}(t) \|^2 \]
\[ = -\frac{1}{2\alpha} \| \dot{x}(t) \|^2 + \mu_y L_{xy} \| \dot{y}(t) \| \| \dot{y}(t) \| - \frac{\mu_y^2 L_{xy}^2 \alpha}{2} \| \dot{y}(t) \|^2 \leq 0, \]
which when using into (31) we obtain
\[ \dot{V}(x(t), y(t)) \leq -\frac{1}{2\alpha} \| \dot{x}(t) \|^2 - \frac{\mu_y \gamma \alpha}{4} \| \dot{y}(t) \|^2 \]
\[ = -\frac{\alpha}{2} \| \nabla_x f(x(t), y(t)) \|^2 - \frac{\mu_y \gamma \alpha \beta^2}{4} \| \nabla_y f(x(t), y(t)) \|^2 \]
\[ = -\frac{\alpha}{2} \| \nabla_x f(x(t), y(t)) \|^2 - \frac{4L_{xy}^2 \alpha}{\mu_y^2} \| \nabla_y f(x(t), y(t)) \|^2 \]
\[ \leq -\frac{\alpha}{2} \left( \| \nabla_x f(x(t), y(t)) \|^2 + \| \nabla_y f(x(t), y(t)) \|^2 \right), \]
which when taking the integral on both sides over \( t \) from 0 to \( T \) and rearrange we obtain
\[ \frac{\alpha}{2} \int_{t=0}^{T} \left( \| \nabla_x f(x(t), y(t)) \|^2 + \| \nabla_y f(x(t), y(t)) \|^2 \right) dt \leq V(x(0), y(0)). \]

Thus, the preceding relation gives (30), i.e., for all \( T > 0 \)
\[ \min_{0 \leq t \leq T} \left\| \nabla_x f(x(t), y(t)) \right\| \nabla_y f(x(t), y(t)) \right\| \leq \sqrt{2V(x(0), y(0)) \sqrt{\alpha T}} \leq \sqrt{2V_1(x(0), y(0)) \alpha T} + \sqrt{2V_2(x(0), y(0)) \alpha T} \]
\[ = \sqrt{2V_1(x(0), y(0)) \alpha T} + \sqrt{\gamma \alpha \beta \| \nabla_y f(x(0), y(0)) \|^2 \alpha T} \]
\[ = \frac{L_{xy} \sqrt{2V_1(x(0), y(0))}}{\sqrt{T}} + \frac{2L_{xy} \| \nabla_y f(x(0), y(0)) \|^2}{\sqrt{\mu_y T}}. \]

\[ \square \]

### 3.4 Strongly Convex–Nonconcave Conditions

As mentioned, the single-loop GDA method is applicable to the convex-nonconcave min-max problem, while the nested-loop GDA method is not. In this section, we complete our analysis by studying the rate of GDAD when the function \( f(x, y) \) is strongly convex given any \( y \) and nonconcave given any \( x \). In particular, we consider the following assumption.

**Assumption 6.** The objective function \( f(x, \cdot) \) is nonconcave for any given \( x \) and \( f(\cdot, y) \) is strongly convex with constant \( \mu_x > 0 \) for any given \( y \). The latter is equivalent to
\[ f(x_1, y) - f(x_2, y) - (\nabla f(x_2, y), x_1 - x_2) \geq \frac{\mu_x}{2} \| x_1 - x_2 \|^2, \quad \forall x_1, x_2 \in \mathbb{R}^m. \] (32)
For our analysis of in this section, we introduce the following two Lyapunov functions

\[ V_1(x, y) = \frac{1}{2} \| \dot{x} \|^2 = \frac{1}{2} \alpha \nabla_x f(x, y) \| \|^2 \]  

(33)

\[ V_2(x, y) = \max_{x, y} f(x, y) - f(x, y). \]  

(34)

The time derivatives of \( V_1 \) and \( V_2 \) over the trajectories \( \dot{x} \) and \( \dot{y} \) are given in the following lemma, whose proof is presented in Section 4.4.

**Lemma 5.** Suppose that Assumptions 1 and 6 hold. Then we have

\[ \dot{V}_1(x(t), y(t)) \leq -\mu_x \| \dot{x}(t) \|^2 + L_{xy} \alpha \| \dot{y}(t) \| \| \dot{x}(t) \|. \]  

(35)

\[ \dot{V}_2(x(t), y(t)) \leq \frac{1}{\alpha} \| \dot{y}(t) \|^2 - \frac{1}{\beta} \| \dot{y}(t) \|^2. \]  

(36)

Since in the strongly convex-nonconcave setting the minimization problem is easier to solve than the maximization problem, we consider the following Lyapunov function

\[ V(x, y) = V_2(x, y) + \gamma \frac{\beta}{\alpha} V_1(x, y), \]  

(37)

where \( \beta \ll \alpha \). In this case, \( \dot{x} \) is updated at a faster time scale than \( \dot{y} \). Using this Lyapunov function, we now derive the convergence rate of GDAD under Assumption 6 in the following theorem, which basically is similar to the one in Theorem 3.

**Theorem 4.** Suppose that Assumptions 1 and 6 hold. Let \( \gamma, \alpha, \beta \) be chosen as

\[ \gamma = \mu_x L_{xy}^2, \quad \alpha = \frac{4}{\mu_x^2}, \quad \beta = \frac{1}{L_{xy}^2}. \]  

(38)

Then we have for all \( T \geq 0 \)

\[ \min_{0 \leq t \leq T} \left\| \nabla_x f(x(t), y(t)) \right\| \leq \frac{L_{xy} \sqrt{2V_1(x(0), y(0))}}{\sqrt{T}} + \frac{2L_{xy} \| \nabla_x f(x(0), y(0)) \|}{\sqrt{\mu_x T}}. \]  

(39)

**Proof.** By using (35) and (36) we consider

\[ \dot{V}(x(t), y(t)) = \dot{V}_2(x(t), y(t)) + \gamma \frac{\beta}{\alpha} \dot{V}_1(x(t), y(t)) \]

\[ \leq \frac{1}{\alpha} \| \dot{x}(t) \|^2 - \frac{1}{\beta} \| \dot{y}(t) \|^2 + L_{xy} \gamma \beta \| \dot{y}(t) \| \| \dot{x}(t) \| - \mu_x \gamma \beta \| \dot{x}(t) \|^2 \]

\[ = -\frac{1}{2\beta} \| \dot{y}(t) \|^2 - \frac{\mu_x \gamma \beta}{4} || \dot{x}(t) ||^2 - \left( \frac{\mu_x \gamma \beta}{4} - \frac{1}{\alpha} \right) || \dot{x}(t) ||^2 \]

\[ - \frac{1}{2\beta} \| \dot{y}(t) \|^2 + L_{xy} \gamma \beta \| \dot{y}(t) \| \| \dot{x}(t) \| - \frac{\mu_x \gamma \beta}{2} || \dot{x}(t) ||^2 \]

\[ = -\frac{1}{2\beta} \| \dot{y}(t) \|^2 - \frac{\mu_x \gamma \beta}{4} || \dot{x}(t) ||^2 \]

\[ - \frac{1}{2\beta} \| \dot{y}(t) \|^2 + L_{xy} \gamma \beta \| \dot{y}(t) \| \| \dot{x}(t) \| - \frac{\mu_x \gamma \beta}{2} || \dot{x}(t) ||^2, \]  

(40)

where in the last equality we use (38) to have

\[ \frac{\mu_x \gamma \beta}{4} - \frac{1}{\alpha} = \frac{\mu_x^2}{4} - \frac{\mu_x^2}{4} = 0. \]
Using (38) one more time we obtain
\[- \frac{1}{2\beta} \|y(t)\|^2 + L_{xy} \gamma \beta \|x(t)\| \|\dot{x}(t)\| - \frac{\mu_x \gamma \beta}{2} \|\dot{x}(t)\|^2\]
\[= - \frac{1}{2\beta} \|y(t)\|^2 + \mu_x L_{xy} \|\dot{y}(t)\| \|\dot{x}(t)\| - \frac{\mu_x^2 L_{xy} \beta}{2} \|\dot{x}(t)\|^2 \leq 0,
\]
which when using into (40) we obtain
\[\dot{V}(x(t), y(t)) \leq - \frac{1}{2\beta} \|y(t)\|^2 - \frac{\mu_x \gamma \beta}{4} \|\dot{x}(t)\|^2\]
\[= - \frac{\beta}{2} \|\nabla_y f(x(t), y(t))\|^2 - \frac{\mu_x \gamma \beta \alpha^2}{4} \|\nabla_x f(x(t), y(t))\|^2\]
\[= - \frac{\beta}{2} \|\nabla_y f(x(t), y(t))\|^2 - \frac{4L_{xy}^2 \beta}{\mu_x^2} \|\nabla_x f(x(t), y(t))\|^2\]
\[\leq - \frac{\beta}{2} (\|\nabla_x f(x(t), y(t))\|^2 + \|\nabla_y f(x(t), y(t))\|^2),\]
which when taking the integral on both sides over \(t\) from 0 to \(T\) and rearrange we obtain
\[\frac{\beta}{2} \int_{t=0}^{T} (\|\nabla_x f(x(t), y(t))\|^2 + \|\nabla_y f(x(t), y(t))\|^2) dt \leq V(x(0), y(0)).\]

Thus, the preceding relation gives (39), i.e., for all \(T > 0\)
\[\min_{0 \leq t \leq T} \left\| \begin{array}{c} \nabla_x f(x(t), y(t)) \\ \nabla_y f(x(t), y(t)) \end{array} \right\| \leq \sqrt{2V(x(0), y(0))} \leq \sqrt{2V_1(x(0), y(0))} + \sqrt{2V_2(x(0), y(0))}\]
\[= \sqrt{2V_1(x(0), y(0))} + \sqrt{\frac{\gamma \beta \alpha \|\nabla_x f(x(0), y(0))\|^2}{\beta T}}\]
\[= \frac{L_{xy}}{\sqrt{T}} \sqrt{2V_1(x(0), y(0))} + \frac{2L_{xy}}{\sqrt{\mu_x T}} \frac{\|\nabla_x f(x(0), y(0))\|}{\sqrt{T}}.\]

\[\square\]

4 Proofs of Technical Lemmas

In this section, we present the analysis of all technical lemmas in the previous sections.

4.1 Proof of Lemma 2

Proof. For convenience, we denote by \(y^*(x) = \mathcal{P}_{\mathcal{Y}^*}(x)[y]\), where recall that \(\mathcal{Y}^*(x)\) is the solution set of \(\max_y f(x, y)\) for a given \(x\). We first show (9). The time derivative of \(V_1\) defined in (7) over the
trajectory \( \dot{x} \) in (2) is given as

\[
\dot{V}_1(x) = \frac{d}{dt} V_1(x) = \nabla_x f(x, y^*(x)) \dot{x} = -\alpha (\nabla_x f(x, y^*(x)), \nabla_x f(x, y)) \\
= -\alpha \| \nabla_x f(x, y^*(x)) \|^2 - \alpha (\nabla_x f(x, y^*(x)), \nabla_x f(x, y) - \nabla_x f(x, y^*(x))) \\
\leq -\frac{\alpha}{2} \| \nabla_x f(x, y^*(x)) \|^2 + \frac{\alpha}{2} \| \nabla_x f(x, y) - \nabla_x f(x, y^*(x)) \|^2 \\
\leq -\frac{\alpha}{2} \| \nabla_x f(x, y^*(x)) \|^2 + \frac{L_{xy}^2 \alpha}{2} \| y - y^*(x) \|^2 \\
\leq -\frac{\alpha}{2} \| \nabla_x f(x, y^*(x)) \|^2 + \frac{L_{xy}^2 \alpha}{\mu_y} (\max f(x, y) - f(x, y)),
\]

where the first and second inequalities are due to the Cauchy-Schwarz inequality and Assumption 1, respectively. In addition, the last inequality is due to (6).

Next we show (10). Indeed, using (41) we have

\[
\dot{V}_2(x, y) = \nabla_x f(x, y^*(x)) \dot{x} - \nabla_x f(x, y) \dot{x} - \nabla_y f(x, y) \dot{y} \\
\leq -\frac{\alpha}{2} \| \nabla_x f(x, y^*(x)) \|^2 + \frac{\alpha}{2} \| \nabla_x f(x, y) - \nabla_x f(x, y^*(x)) \|^2 - \beta \| \nabla_y f(x, y) \|^2 \\
+ \alpha \| \nabla_x f(x, y) \|^2 \\
\leq -\frac{\alpha}{2} \| \nabla_x f(x, y^*(x)) \|^2 - \beta \| \nabla_y f(x, y) \|^2 + \frac{\alpha}{2} \| \nabla_x f(x, y) - \nabla_x f(x, y^*(x)) \|^2 \\
+ 2\alpha \| \nabla_x f(x, y) - \nabla_x f(x, y^*(x)) \|^2 + 2\alpha \| \nabla_x f(x, y^*(x)) \|^2 \\
= -\beta \| \nabla_y f(x, y) \|^2 + \frac{3\alpha}{2} \| \nabla_x f(x, y^*(x)) \|^2 + \frac{5\alpha}{2} \| \nabla_x f(x, y) - \nabla_x f(x, y^*(x)) \|^2 \\
\leq -\beta \| \nabla_y f(x, y) \|^2 + \frac{3\alpha}{2} \| \nabla_x f(x, y^*(x)) \|^2 + \frac{5L_{xy}^2 \alpha}{\mu_y} (\max f(x, y) - f(x, y)),
\]

where the last inequality we use Assumption 1 and (6), similar to the last inequality in \( \dot{V}_1 \) above. \( \square \)

### 4.2 Proof of Lemma 3

**Proof.** For convenience, we denote by \( y^*(x) = \mathcal{P}_{\mathcal{Y}^*(x)}[y] \), where recall that \( \mathcal{Y}^*(x) \) is the solution set of \( \max_y f(x, y) \) for a given \( x \). We first show (18). The time derivative of \( V_1 \) defined in (16) over the trajectory \( \dot{x} \) in (2) is given as

\[
\dot{V}_1(x(t), y(t)) = \frac{d}{dt} V_1(x(t), y(t)) = \nabla_x f(x(t), y(t)) \dot{x} + \nabla_y f(x(t), y(t)) \dot{y} \\
= -\alpha \| \nabla_x f(x(t), y(t)) \|^2 + \beta \| \nabla_y f(x(t), y(t)) \|^2.
\]
Next we show (19). Indeed, using (2) we have
\[
\dot{V}_2(x(t), y(t)) = \nabla_x f(x(t), y^*(x(t)))\ddot{x} - \nabla_x f(x(t), y(t))\dot{x} - \nabla_y f(x(t), y(t))\dot{y} \\
= -\alpha \langle \nabla_x f(x(t), y^*(x(t))), \nabla_x f(x(t), y(t)) \rangle \ddot{x} - \beta \|\nabla_y f(x(t), y(t))\|^2 \\
+ \alpha \|\nabla_x f(x(t), y(t))\|^2 \\
= -\alpha \langle \nabla_x f(x(t), y^*(x(t))), \nabla_x f(x(t), y(t)) \rangle \ddot{x} - \beta \|\nabla_y f(x(t), y(t))\|^2 + \alpha \|\nabla_x f(x(t), y(t))\|^2 \\
- \alpha \langle \nabla_x f(x(t), y^*(x(t))) - \nabla_x f(x(t), y(t)), \nabla_x f(x(t), y(t)) \rangle \\
\leq -\beta \|\nabla_y f(x(t), y(t))\|^2 + \frac{\alpha}{8} \|\nabla_x f(x(t), y(t))\|^2 \\
+ 2\alpha \|\nabla_x f(x(t), y^*(x(t))) - \nabla_x f(x(t), y(t))\|^2 \\
\leq -\beta \|\nabla_y f(x(t), y(t))\|^2 + \frac{\alpha}{8} \|\nabla_x f(x(t), y(t))\|^2 + 2L_{xy}^2 \alpha \|y^*(x(t)) - y(t)\|^2 \\
\leq -\beta \|\nabla_y f(x(t), y(t))\|^2 + \frac{\alpha}{8} \|\nabla_x f(x(t), y(t))\|^2 + \frac{4L_{xy}^2 \alpha}{\mu_y} \dot{V}_2(x(t), y(t)),
\]
where the second inequality is due to the Cauchy-Schwarz inequality: $2xy \leq (1/\eta)x^2 + \eta y^2$ for any $\eta > 0$. In addition, the last two inequalities are due to Eqs. 1 and (6), respectively.  

\[\square\]

4.3 Proof of Lemma 4

Proof. We first show (27). Indeed, using (2)
\[
\dot{V}_1(x(t), y(t)) = \frac{d}{dt} V_1(x(t), y(t)) = \nabla_x f(x(t), y(t))\ddot{x} + \nabla_y f(x(t), y(t))\dot{y} \\
= -\frac{1}{\alpha} \|\dot{x}(t)\|^2 + \frac{1}{\beta} \|\dot{y}(t)\|^2.
\]

Next, we show (28). Using (2) and (26) we consider
\[
\dot{V}_2(x(t), y(t)) = \beta \langle \dot{y}(t), \frac{d}{dt} \nabla_y f(x, y) \rangle \\
= \beta \langle \dot{y}(t), \nabla_{yx} f(x(t), y(t))\dot{x}(t) + \nabla_{yy} f(x(t), y(t))\dot{y}(t) \rangle \\
= \beta \langle \dot{y}(t), \nabla_{yx} f(x(t), y(t))\dot{x}(t) \rangle + \beta \langle \dot{y}(t), \nabla_{yy} f(x(t), y(t))\dot{y}(t) \rangle \\
\leq L_{xy} \beta \|\dot{y}(t)\| \|\dot{x}(t)\| - \mu_y \beta \|\dot{y}(t)\|^2,
\]
where in the last inequality we use Assumptions 1 and 5 to have
\[
\|\nabla_{yx} f(x(t), y(t))\| \leq L_{xy} \quad \text{and} \quad \nabla_{yy} f(x(t), y(t)) \leq -\mu_y I, \quad \text{where} \ I \ \text{is the identity matrix}.
\]

\[\square\]

4.4 Proof of Lemma 5

Proof. We first show (35). Using (2) and (33) we consider
\[
\dot{V}_1(x(t), y(t)) = -\alpha \langle \dot{x}(t), \frac{d}{dt} \nabla_x f(x(t), y(t)) \rangle \\
= -\alpha \langle \dot{x}(t), \nabla_{xx} f(x(t), y(t))\dot{x}(t) + \nabla_{xy} f(x(t), y(t))\dot{y}(t) \rangle \\
= -\alpha \langle \dot{x}(t), \nabla_{xx} f(x(t), y(t))\dot{x}(t) \rangle - \alpha \langle \dot{x}(t), \nabla_{xy} f(x(t), y(t))\dot{y}(t) \rangle \\
\leq -\mu_x \alpha \|\dot{x}(t)\|^2 + L_{xy} \alpha \|\dot{y}(t)\| \|\dot{x}(t)\|,
\]
where in the last inequality we use Assumptions 1 and 6 to have
\[ \|\nabla_{xy} f(x(t), y(t))\| \leq L_{xy} \quad \text{and} \quad \nabla_{xx} f(x(t), y(t)) \geq \mu_x I, \] where \( I \) is the identity matrix.

Next, we show (36). Indeed, using (2)
\[
\dot{V}_2(x(t), y(t)) = \frac{d}{dt} V_2(x(t), y(t)) = -\nabla_x f(x(t), y(t)) \dot{x} - \nabla_y f(x(t), y(t)) \dot{y} \\
= \frac{1}{\alpha} \|\dot{x}(t)\|^2 - \frac{1}{\beta} \|\dot{y}(t)\|^2.
\]

\[ \square \]

5 Concluding Remarks

In this paper, we consider two-time-scale gradient descent-ascent dynamics for solving nonconvex min-max optimization problems. Our main focus is to derive the convergence rates of this method for different settings of the underlying objective functions. Our techniques are mainly motivated by the classic singular perturbation, where we show that our analysis improves the existing results under the same conditions. A natural extension from this work is to provide a better analysis for the discrete-time variant of GDAD. Another interesting future direction is to consider the stochastic setting and its accelerated counterpart.

References

[1] T. Basar and G. J. Olsder, Dynamic Noncooperative Game Theory, 2nd Edition. Society for Industrial and Applied Mathematics, 1998.

[2] L. S. Shapley, “Stochastic games,” Proceedings of the National Academy of Sciences, 1953.

[3] E. Altman, Constrained Markov decision processes. Chapman and Hall/CRC Press, 1999.

[4] J. Achiam, D. Held, A. Tamar, and P. Abbeel, “Constrained policy optimization,” in Proceedings of the 34th International Conference on Machine Learning, vol. 70, 2017, pp. 22–31.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” Commun. ACM, vol. 63, no. 11, p. 139–144, 2020.

[6] L. Mescheder, S. Nowozin, and A. Geiger, “The numerics of gans,” in Proceedings of the 31st International Conference on Neural Information Processing Systems, ser. NIPS’17, 2017, p. 1823–1833.

[7] A. Kurakin, I. J. Goodfellow, and S. Bengio, “Adversarial machine learning at scale,” in 5th International Conference on Learning Representations, ICLR, 2017.

[8] Q. Qian, S. Zhu, J. Tang, R. Jin, B. Sun, and H. Li, “Robust optimization over multiple domains,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 4739–4746, Jul. 2019.
[9] Y.-F. Liu, Y.-H. Dai, and Z.-Q. Luo, “Max-min fairness linear transceiver design for a multi-user mimo interference channel,” *IEEE Transactions on Signal Processing*, vol. 61, no. 9, pp. 2413–2423, 2013.

[10] G. Lan, S. Lee, and Y. Zhou, “Communication-efficient algorithms for decentralized and stochastic optimization,” *Mathematical Programming*, vol. 180, pp. 237–284, 2020.

[11] T.-H. Chang, M. Hong, H.-T. Wai, X. Zhang, and S. Lu, “Distributed learning in the nonconvex world: From batch data to streaming and beyond,” *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 26–38, 2020.

[12] D. Balduzzi, S. Racaniere, J. Martens, J. Foerster, K. Tuyls, and T. Graepel, “The mechanics of n-player differentiable games,” in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, 2018, pp. 354–363.

[13] W. Krichene, A. Bayen, and P. L. Bartlett, “Accelerated mirror descent in continuous and discrete time,” in *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[14] M. Raginsky and J. Bouvrie, “Continuous-time stochastic mirror descent on a network: Variance reduction, consensus, convergence,” in *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, 2012, pp. 6793–6800.

[15] W. Su, S. Boyd, and E. J. Candès, “A differential equation for modeling nesterov’s accelerated gradient method: Theory and insights,” in *Proceedings of the 27th International Conference on Neural Information Processing Systems*, ser. NIPS’14, 2014, p. 2510–2518.

[16] J. Diakonikolas and L. Orecchia, “The approximate duality gap technique: A unified theory of first-order methods,” *SIAM Journal on Optimization*, vol. 29, no. 1, pp. 660–689, 2019.

[17] 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.

[18] Y. Malitsky, “Projected reflected gradient methods for monotone variational inequalities,” *SIAM Journal on Optimization*, vol. 25, no. 1, pp. 502–520, 2015.

[19] Y. Wang and J. Li, “Improved algorithms for convex-concave minimax optimization,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 4800–4810.

[20] A. Cherukuri, B. Gharesifard, and J. Cortés, “Saddle-point dynamics: Conditions for asymptotic stability of saddle points,” *SIAM Journal on Control and Optimization*, vol. 55, no. 1, pp. 486–511, 2017.

[21] “Semi-global exponential stability of augmented primal–dual gradient dynamics for constrained convex optimization,” *Systems & Control Letters*, vol. 144, p. 104754, 2020.

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

[23] 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.
[24] R. D. C. Monteiro and B. F. Svaiter, “On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean,” *SIAM Journal on Optimization*, vol. 20, no. 6, pp. 2755–2787, 2010.

[25] N. Golowich, S. Pattathil, C. Daskalakis, and A. Ozdaglar, “Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems,” in *Proceedings of Thirty Third Conference on Learning Theory*, vol. 125, 2020, pp. 1758–1784.

[26] T. Yoon and E. K. Ryu, “Accelerated algorithms for smooth convex-concave minimax problems with $o(1/k^2)$ rate on squared gradient norm,” in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, 2021, pp. 12098–12109.

[27] C. D. Dang and G. Lan, “On the convergence properties of non-euclidean extragradient methods for variational inequalities with generalized monotone operators,” *Computational Optimization and Applications*, vol. 60, pp. 277–310, 2015.

[28] J. Abernethy, K. A. Lai, and A. Wibisono, “Last-iterate convergence rates for min-max optimization: Convergence of hamiltonian gradient descent and consensus optimization,” in *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, vol. 132, 2021, pp. 3–47.

[29] K. K. Thekumparampil, P. Jain, P. Netrapalli, and S. Oh, “Efficient algorithms for smooth minimax optimization,” in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[30] W. Kong and R. D. C. Monteiro, “An accelerated inexact proximal point method for solving nonconvex-concave min-max problems,” *SIAM Journal on Optimization*, vol. 31, no. 4, pp. 2558–2585, 2021.

[31] H. Rafique, M. Liu, Q. Lin, and T. Yang, “Weakly-convex–concave min–max optimization: provable algorithms and applications in machine learning,” *Optimization Methods and Software*, vol. 0, no. 0, pp. 1–35, 2021.

[32] T. Lin, C. Jin, and M. I. Jordan, “Near-optimal algorithms for minimax optimization,” in *Proceedings of Thirty Third Conference on Learning Theory*, vol. 125. PMLR, 09–12 Jul 2020, pp. 2738–2779.

[33] M. Nouiehed, M. Sanjabi, T. Huang, J. D. Lee, and M. Razaviyayn, “Solving a class of non-convex min-max games using iterative first order methods,” in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[34] T. Lin, C. Jin, and M. Jordan, “On gradient descent ascent for nonconvex-concave minimax problems,” in *Proceedings of the 37th International Conference on Machine Learning*, vol. 119. PMLR, 13–18 Jul 2020, pp. 6083–6093.

[35] S. Lu, I. Tsaknakis, M. Hong, and Y. Chen, “Hybrid block successive approximation for one-sided non-convex min-max problems: Algorithms and applications,” *IEEE Transactions on Signal Processing*, vol. 68, pp. 3676–3691, 2020.

[36] J. Yang, N. Kiyavash, and N. He, “Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1153–1165.
[37] Z. Xu, H.-L. Zhang, Y. Xu, and G. Lan, “A unified single-loop alternating gradient projection algorithm for nonconvex-concave and convex-nonconcave minimax problems,” ArXiv, vol. abs/2006.02032, 2020.

[38] J. Zhang, P. Xiao, R. Sun, and Z. Luo, “A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems,” in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 7377–7389.

[39] N. Loizou, H. Berard, A. Jolicoeur-Martineau, P. Vincent, S. Lacoste-Julien, and I. Mitliagkas, “Stochastic Hamiltonian gradient methods for smooth games,” in Proceedings of the 37th International Conference on Machine Learning, vol. 119, 2020, pp. 6370–6381.

[40] K. Zhang, Z. Yang, and T. Basar, “Policy optimization provably converges to nash equilibria in zero-sum linear quadratic games,” in NeurIPS, vol. 32, 2019.

[41] S. Cen, Y. Wei, and Y. Chi, “Fast policy extragradient methods for competitive games with entropy regularization,” ArXiv, vol. abs/2105.15186, 2021.

[42] J. Perolat, B. Piot, and O. Pietquin, “Actor-critic fictitious play in simultaneous move multi-stage games,” in Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, vol. 84, 2018, pp. 919–928.

[43] C. Daskalakis, D. J. Foster, and N. Golowich, “Independent policy gradient methods for competitive reinforcement learning,” in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 5527–5540.

[44] V. S. Borkar, Stochastic Approximation: A Dynamical Systems Viewpoint. Cambridge University Press, 2008.

[45] V. R. Konda and J. N. Tsitsiklis, “Convergence rate of linear two-time-scale stochastic approximation,” The Annals of Applied Probability, vol. 14, no. 2, pp. 796–819, 2004.

[46] G. Dalal, B. Szorenyi, and G. Thoppe, “A tale of two-timescale reinforcement learning with the tightest finite-time bound,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 04, pp. 3701–3708, Apr. 2020.

[47] T. T. Doan and J. Romberg, “Linear two-time-scale stochastic approximation a finite-time analysis,” in 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton), 2019, pp. 399–406.

[48] H. Gupta, R. Srikant, and L. Ying, “Finite-time performance bounds and adaptive learning rate selection for two time-scale reinforcement learning,” in Advances in Neural Information Processing Systems, 2019.

[49] T. T. Doan, “Finite-time analysis and restarting scheme for linear two-time-scale stochastic approximation,” SIAM Journal on Control and Optimization, vol. 59, no. 4, pp. 2798–2819, 2021.

[50] M. Kaledin, E. Moulines, A. Naumov, V. Tadic, and H.-T. Wai, “Finite time analysis of linear two-timescale stochastic approximation with Markovian noise,” in Proceedings of Thirty Third Conference on Learning Theory, vol. 125, 2020, pp. 2144–2203.
[51] A. Mokkadem and M. Pelletier, “Convergence rate and averaging of nonlinear two-time-scale stochastic approximation algorithms,” *The Annals of Applied Probability*, vol. 16, no. 3, pp. 1671–1702, 2006.

[52] T. T. Doan, “Finite-time convergence rates of nonlinear two-time-scale stochastic approximation under Markovian noise,” *arXiv preprint arXiv:2104.01627*, 2021.

[53] ———, “Nonlinear two-time-scale stochastic approximation: Convergence and finite-time performance,” *arXiv preprint arXiv:2011.01868*, 2020.

[54] V. S. Borkar, “An actor-critic algorithm for constrained Markov decision processes,” *Systems & Control Letters*, vol. 54, no. 3, pp. 207–213, 2005.

[55] S. Bhatnagar and K. Lakshmanan, “An online actor–critic algorithm with function approximation for constrained Markov decision processes,” *Journal of Optimization Theory and Applications*, 2012.

[56] S. Paternain, L. Chamon, M. Calvo-Fullana, and A. Ribeiro, “Constrained reinforcement learning has zero duality gap,” in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[57] D. Ding, K. Zhang, T. Basar, and M. R. Jovanovic, “Natural policy gradient primal-dual method for constrained Markov decision processes.” in *NeurIPS*, 2020.

[58] S. Zeng, T. T. Doan, and J. Romberg, “Finite-time complexity of online primal-dual natural actor-critic algorithm for constrained Markov decision processes,” *ArXiv*, vol. abs/2110.11383, 2021.

[59] M. Wang, E. X. Fang, and H. Liu, “Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions,” *Mathematical Programming*, vol. 161, no. 1, pp. 419–449, Jan 2017.

[60] S. Zeng, T. T. Doan, and J. Romberg, “A two-time-scale stochastic optimization framework with applications in control and reinforcement learning,” *ArXiv*, vol. abs/2109.14756, 2021.

[61] S. Lee and A. Nedic, “Distributed random projection algorithm for convex optimization,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 2, pp. 221–229, 2013.

[62] A. Reisizadeh, A. Mokhtari, H. Hassani, and R. Pedarsani, “An exact quantized decentralized gradient descent algorithm,” *IEEE Transactions on Signal Processing*, vol. 67, no. 19, pp. 4934–4947, 2019.

[63] T. T. Doan, S. T. Maguluri, and J. Romberg, “Convergence rates of distributed gradient methods under random quantization: A stochastic approximation approach,” *IEEE Transactions on Automatic Control*, 2020.

[64] T. T. Doan, C. L. Beck, and R. Srikant, “On the convergence rate of distributed gradient methods for finite-sum optimization under communication delays,” *Proc. ACM Meas. Anal. Comput. Syst.*, vol. 1, no. 2, 2017.

[65] A. Dutta, A. M. Boker, and T. T. Doan, “Convergence rates of distributed consensus over cluster networks: A two-time-scale approach,” Available at: [https://arxiv.org/abs/2104.07781](https://arxiv.org/abs/2104.07781), 2021.
[66] A. Dutta, N. Masrouisaadat, and T. T. Doan, “Convergence rates of decentralized gradient methods over cluster networks,” arXiv preprint arXiv:2110.06992, 2021.

[67] M. M. Vasconcelos, T. T. Doan, and U. Mitra, “Improved convergence rate for a distributed two-time-scale gradient method under random quantization,” in 2021 IEEE Conference on Decision and Control (CDC), 2021.

[68] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” in Advances in Neural Information Processing Systems, vol. 30, 2017.

[69] P. Kokotović, H. K. Khalil, and J. O’Reilly, Singular Perturbation Methods in Control: Analysis and Design. Society for Industrial and Applied Mathematics, 1999.

[70] B. Polyak, “Gradient methods for the minimisation of functionals,” Ussr Computational Mathematics and Mathematical Physics, vol. 3, pp. 864–878, 12 1963.

[71] H. Karimi, J. Nutini, and M. Schmidt, “Linear convergence of gradient and proximal-gradient methods under the polyak-losiasiewicz condition,” in Machine Learning and Knowledge Discovery in Databases, P. Frasconi, N. Landwehr, G. Manco, and J. Vreeken, Eds. Cham: Springer International Publishing, 2016, pp. 795–811.

[72] D. Bertsekas, Nonlinear Programming: 2nd Edition. Cambridge, MA: Athena Scientific, 1999.