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

Thinh T. Doan
Department of Electrical and Computer Engineering
Virginia Tech, USA

THINHDOAN@VT.EDU

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 the simplicity in its implementation. However, theoretical guarantees on the convergence of this algorithm are 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.

Keywords: Gradient Descent-Ascent Methods, Two-Time-Scale Dynamics, Singular Perturbation Theory, Min-Max Optimization.

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 Basar and Olsder (1998); Shapley (1953), stochastic control and reinforcement learning Altman (1999); Achiam et al. (2017), training generative adversarial networks (GANs) Goodfellow et al. (2020); Mescheder et al. (2017), adversarial and robust machine learning Kurakin et al. (2017); Qian et al. (2019), resource allocation over networks Liu et al. (2013), and distributed optimization Lan et al. (2020); Chang et al. (2020); 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 \( y \), 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 Balduzzi et al. (2018). 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 Krichene et al. (2015); Raginsky and Bouvrie (2012); Su et al. (2014); Diakonikolas and Orecchia (2019). 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 Nemirovski (2004); Malitsky (2015); Wang and Li (2020); Cherukuri et al. (2017), extragradient and optimistic gradient methods Korpelevich (1976); Mokhtari et al. (2020); Monteiro and Svaiter (2010); Golowich et al. (2020); Yoon and Ryu (2021); Dang and Lan (2015), and recently Hamiltonian gradient descent methods Mescheder et al. (2017); Balduzzi et al. (2018); Abernethy et al. (2021). Some algorithms in these settings have convergence rates matched with the lower bound complexity; see the recent work Yoon and Ryu (2021) 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 sta-
tionary point of \( f \) when the max problem is concave. For example, multiple-loop algorithms have
been studied in Thekumparampil et al. (2019); Kong and Monteiro (2021); Rafique et al. (2021);
Lin et al. (2020b); Nouiehed et al. (2019). Our work in this paper is closely related to the recent
literature on studying single-loop algorithm Lin et al. (2020a); Lu et al. (2020); Yang et al. (2020);
Xu et al. (2020); Zhang et al. (2020). While these works study discrete-time algorithms, we con-
sider 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 Loizou et al.
(2020); Zhang et al. (2019); Cen et al. (2021); Perolat et al. (2018); Zhang et al. (2021); Daskalakis
et al. (2020), two-time-scale stochastic approximation Borkar (2008); Konda and Tsitsiklis (2004);
Dalal et al. (2020); Doan and Romberg (2019); Gupta et al. (2019); Doan (2021b); Kaledin et al.
(2020); Mokkadem and Pelletier (2006); Doan (2021c, 2020), and reinforcement learning Bhatna-
gar and Lakshmanan (2012); Paternain et al. (2019); Ding et al. (2020); Zeng et al. (2021), two-
time-scale optimization , and decentralized optimization . 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 ap-
proximation). Thus, the techniques therein 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 dy-
namics (GDAD), where we implement simultaneously the following two differential equations

\[
\begin{align*}
\dot{x}(t) &= \frac{d}{dt} x(t) = -\alpha \nabla_x f(x(t), y(t)), \\
\dot{y}(t) &= \frac{d}{dt} y(t) = \beta \nabla_y f(x(t), y(t)),
\end{align*}
\]

(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 Heusel et al. (2017), 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 Kokotović et al. (1999), explained as follows. Since \( y \)
implemented at a faster time scale than \( x \), we 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 $y$. When $y$ 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

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

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 Chow and Kokotovic (1985); Bryik and Arcak (2008); Doan (2020); Dutta et al. (2021).

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 $x_1, x_2 \in \mathbb{R}^m$, $y_1, y_2 \in \mathbb{R}^n$ 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 NCvex, NCave, SCvex, SCave, and PL stand for nonconvex, nonconcave, strongly convex, strongly concave, and Polyak-Łojasiewicz conditions, respectively. Condition number $\kappa$ is defined in (9), and $R$ is the size of compact set used in Nouiehed et al. (2019).

<table>
<thead>
<tr>
<th>OBJECTIVES</th>
<th>PRIOR WORKS</th>
<th>THIS PAPER</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PL &amp; PL</strong></td>
<td>$O\left(\kappa^3 \log\left(\frac{1}{\epsilon}\right)\right)$ Yang et al. (2020)</td>
<td>$O\left(\kappa^2 \log\left(\frac{1}{\epsilon}\right)\right)$ - <strong>THEOREM 1</strong></td>
</tr>
<tr>
<td><strong>NCvex &amp; PL</strong></td>
<td>$O\left(R^2 L_{xy} \log\left(\frac{1}{\epsilon^2}\right)\right)$ Nouiehed et al. (2019)</td>
<td>$O\left(\kappa_{xy}^2 \epsilon^{-2}\right)$ - <strong>THEOREM 2</strong></td>
</tr>
<tr>
<td><strong>NCvex &amp; SCave</strong></td>
<td>$O\left(L_{xy}^2 \epsilon^{-2}\right)$ Xu et al. (2020)</td>
<td>$O\left(\kappa_{xy}^2 \epsilon^{-2}\right)$ - <strong>THEOREM 3</strong></td>
</tr>
<tr>
<td><strong>SCvex &amp; NCave</strong></td>
<td>$O\left(L_{xy}^2 \epsilon^{-2}\right)$ Xu et al. (2020)</td>
<td>$O\left(\kappa_{xy}^2 \epsilon^{-2}\right)$ - <strong>THEOREM 4</strong></td>
</tr>
</tbody>
</table>

### 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 $f(x, y)$. Our results are summarized in Table 1. First, our approach improves the analysis in Yang et al. (2020), 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 Yang et al. (2020). Our result addresses the conjecture raised in Yang et al. (2020), where the authors state...
that such an improvement may not be possible. Second, our analysis achieves a better result than the one in Nouiehed et al. (2019) for the case of one-sided PL function by a factor of \( \log(1/\epsilon) \). We note that a nested-loop is studied in Nouiehed et al. (2019) while GDAD is a single-loop method. Finally, our result is the same as the one in Xu et al. (2020) 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 Xu et al. (2020), we use a simpler analysis and simpler choice of step sizes to achieve these results. Due to space limitation, some proofs of technical lemmas in this paper can be found in Doan (2021a).

### 3.1. Two-Sided Polyak–Łojasiewicz Conditions

We first study the convergence rates of GDAD when \( f \) satisfies a two-sided Polyak–Łojasiewicz (PL) condition, which is considered in Yang et al. (2020) 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 \( (x, y) \in \mathbb{R}^m \times \mathbb{R}^n \):

\[
\begin{align*}
2\mu_x [f(x, y) - \min_x f(x, y)] & \leq \|\nabla_x f(x, y)\|^2, \\
2\mu_y [\max_y f(x, y) - f(x, y)] & \leq \|\nabla_y f(x, y)\|^2.
\end{align*}
\]

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 Polyak (1963) 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 Karimi et al. (2016), the PL condition also implies the quadratic growth condition, i.e., given any \( x \) we have

\[
\max_{z \in \mathbb{R}^m} f(x, z) - f(x, y) \geq \frac{\mu_y}{2} \|P_{\mathcal{Y}^*(x)}[y] - y\|^2, \quad \forall y \in \mathbb{R}^m,
\]

where we assume that \( \mathcal{Y}^*(x) \) is a nonempty solution set of \( \max_y f(x, y) \) and \( P_{\mathcal{Y}^*(x)}[y] \) is the projection of \( y \) to this set. More discussions on PL condition can be found in Karimi et al. (2016), while some examples of functions satisfying the two-sided PL condition are given in Yang et al. (2020).

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 Bertsekas (1999)[Proposition B.25] and studied in Nouiehed et al. (2019)[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 \( L_x + \frac{L_{xy}}{\mu_y} \).

5
Finally, for our analysis we consider the following two Lyapunov functions
\[ V_1(x) = \max_{y \in \mathbb{R}^m} f(x, y) - \min_{x \in \mathbb{R}^n} \max_{y \in \mathbb{R}^m} f(x, y) \quad \text{and} \quad V_2(x, y) = \max_{y} f(x, y) - f(x, y), \]
where it is obvious to see that \( V_1 \) and \( V_2 \) are nonnegative. The time derivatives of \( V_1 \) and \( V_2 \) over \( \dot{x} \) and \( \dot{y} \) are given in the following lemma, where its proof can be found in Doan (2021a).

**Lemma 2** Suppose that Assumptions 1–3 hold. Then we have
\[
\begin{align*}
\dot{V}_1(x) &\leq -\frac{\alpha}{2} \| \nabla_x f(x, y^*(x)) \|^2 + \frac{L_{xy}^2 \alpha}{\mu_y} V_2(x, y), \\
\dot{V}_2(x, y) &\leq -\beta \| \nabla_y f(x, y) \|^2 + \frac{3\alpha}{2} \| \nabla_x f(x, y^*(x)) \|^2 + \frac{5L_{xy}^2 \alpha}{\mu_y} V_2(x, y).
\end{align*}
\]  
As mentioned, the dynamics of \( \dot{x} \) and \( \dot{y} \) are implemented at different time scales, 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.

**Theorem 1** Suppose that Assumptions 1–3 hold. Let \( \gamma, \alpha, \beta \) be chosen as
\[
\gamma = -\frac{L_{xy}^2}{\mu_y^2}, \quad \alpha = \frac{\mu^2}{10\mu_x L_{xy}^2}, \quad \beta = \frac{\mu^2}{\mu_x \mu_y^2}.
\]
Then we have for all \( t \geq 0 \)
\[
V(x(t), y(t)) \leq e^{-\frac{t}{20\kappa^2}} V(x(0), y(0)).
\]

**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 Lemma 2, (3), and the preceding relation we have
\[
\begin{align*}
\dot{V}(x(t), y(t)) &= \dot{V}_1(x(t)) + \frac{\gamma \alpha}{\beta} \dot{V}_2(x(t), y(t)) \\
&\leq -\frac{\alpha}{2} \| \nabla_x f(x(t), y^*(x(t))) \|^2 + \frac{L_{xy}^2 \alpha}{\mu_y} V_2(x(t), y(t)) \\
&\quad - 2\mu_y \gamma \alpha V_2(x(t), y(t)) + \frac{3\gamma \alpha^2}{2\beta} \| \nabla_x f(x(t), y^*(x(t))) \|^2 + \frac{5L_{xy}^2 \gamma \alpha^2}{\mu_y \beta} V_2(x(t), y(t)) \\
&= -\frac{\alpha}{4} \| \nabla_x f(x(t), y^*(x(t))) \|^2 - \frac{\mu_y \gamma \alpha}{2} V_2(x(t), y(t)) \\
&\quad - \left( \frac{1}{2} - \frac{3\gamma \alpha}{\beta} \right) \frac{\alpha}{2} \| \nabla_x f(x(t), y^*(x(t))) \|^2 \\
&\quad - \left( \frac{3\mu_y \gamma}{2} - \frac{L_{xy}^2}{\mu_y} - \frac{5L_{xy}^2 \gamma \alpha}{\mu_y \beta} \right) \alpha V_2(x(t), y(t)).
\end{align*}
\]
Using (10), (5), and \( y^*(x) = \arg \max_y f(x, y) \) into (12) we obtain
\[
\dot{V}(x(t), y(t)) \leq -\frac{\alpha}{4} \| \nabla_x f(x(t), y^*(x(t))) \|^2 - \frac{\mu_y \gamma \alpha}{2} V_2(x(t), y(t))
\]
\[
\leq -\frac{\mu_x \alpha}{2} (\max_y f(x(t), y) - \min_x f(x, y)) - \frac{\mu_y \gamma \alpha}{2} V_2(x, y)
\]
\[
\leq -\frac{\mu_x \alpha}{2} (V_1(x(t)) + \frac{\gamma \alpha}{\beta} V_2(x(t), y(t))) = -\frac{\mu_x \alpha}{2} V(x(t), y(t)),
\]
where the last inequality is due to \( \mu_x \alpha = \frac{\mu_x^2}{16L_{xy}} \leq \mu_y \beta = \frac{\mu_x^2}{\mu_x \mu_y} \). Taking the integral on both sides of the equation above immediately gives (11).

3.2. Nonconvex–Polyak-Łojasiewicz Conditions

We next consider an extension of 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 \( f : \mathbb{R}^m \times \mathbb{R}^n \rightarrow \mathbb{R} \) 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 \( x \in \mathbb{R}^m \):
\[
2\mu_y [\max_y f(x, y) - f(x, y)] \leq \| \nabla_y f(x, y) \|^2.
\]

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. Instead, we will show that GDAD will return a stationary point of \( f \), as studied in Nouiehed et al. (2019). Note that under Assumption 2 the result in Lemma 1 still holds since the work in Nouiehed et al. (2019) 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.
\[
V_1(x, y) = f(x, y) - \min_{x \in \mathbb{R}^m} \min_{y \in \mathbb{R}^n} f(x, y) \quad \text{and} \quad V_2(x, y) = \max_y f(x, y) - f(x, y),
\]
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 Doan (2021a).

**Lemma 3** Suppose that Assumptions 1, 2, and 4 hold. Then we have
\[
\dot{V}_1(x(t), y(t)) = -\alpha \| \nabla_x f(x(t), y(t)) \|^2 + \beta \| \nabla_y f(x(t), y(t)) \|^2.
\]
\[
\dot{V}_2(x(t), y(t)) \leq -\beta \| \nabla_y f(x(t), y(t)) \|^2 + \frac{\alpha}{2} \| \nabla_x f(x(t), y(t)) \|^2 + \frac{L_{xy} \alpha}{\mu_y} V_2(x(t), y(t)).
\]

The convergence rate of GDAD under the nonconvex-PL condition is formally stated as follows.

**Theorem 2** Suppose that Assumptions 1, 2, and 4 hold. Let \( \gamma, \alpha, \beta \) be chosen as
\[
\gamma = \frac{32L_{xy}^2}{\mu_y^2}, \quad \alpha = \frac{1}{8L_{xy}}, \quad \beta = \frac{1}{\mu_y}.
\]

Then we have for all \( T \geq 0 \)
\[
\min_{0 \leq t \leq T} \left\| \frac{\nabla_x f(x(t), y(t))}{\nabla_y f(x(t), y(t))} \right\| \leq \frac{4L_{xy} \sqrt{V_1(x(0), y(0))} + 4V_2(x(0), y(0))}{\sqrt{T}}.
\]
Proof By using (14), (15), and (3) we have

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

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

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

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

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

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

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

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

\[ - \frac{\mu y \gamma \alpha}{2} \left(1 - \frac{4L^2_{xy} \alpha}{\mu y \beta} \right) V_2(x, y) - \alpha \left(1 - \frac{\gamma \alpha}{4\beta} \right) \| \nabla f(x(t), y(t)) \|^2, \quad (18) \]

where in the last inequality we use (13) 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 (16) and the preceding relation into (18) 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 (16) immediately gives us (17). \( \blacksquare \)

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

\[ f(x, y_1) - f(x, y_2) - \langle \nabla f(x, y_2), y_1 - y_2 \rangle \leq -\frac{\mu_y}{2} \| y_1 - y_2 \|^2, \quad \forall y_1, y_2 \in \mathbb{R}^n. \quad (19) \]
For our analysis of in this section, we introduce the following two Lyapunov functions

\[ V_1(x, y) = f(x, y) - \min_{(x,y)} f(x, y), \quad \text{and} \quad V_2(x, y) = \frac{1}{2} \|\dot{y}\|^2 = \frac{1}{2} \|\beta \nabla_y f(x, y)\|^2. \]

The time derivatives of \( V_1 \) and \( V_2 \) are given below, whose proof can be found in Doan (2021a).

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

\[ \dot{V}_1(x, t, y(t)) \leq -\frac{1}{\alpha} \|\dot{x}(t)\|^2 + \frac{1}{\beta} \|\dot{y}(t)\|^2. \]  \hspace{1cm} (20)

\[ \dot{V}_2(x, t, y(t)) \leq L_{xy} \beta \|\dot{y}(t)\| \|\dot{x}(t)\| - \mu_y \beta \|\dot{y}(t)\|^2. \]  \hspace{1cm} (21)

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

\[ \gamma = \mu_y L_{xy}^2, \quad \alpha = \frac{1}{L_{xy}^2}, \quad \beta = \frac{4}{\mu_y^2}. \]  \hspace{1cm} (22)

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

\[ \min_{0 \leq t \leq T} \left\| \frac{\nabla_x f(x, y(t))}{\nabla_y f(x, y(t))} \right\| \leq \frac{L_{xy} \sqrt{2V_1(x(0), y(0))}}{\sqrt{T}} + \frac{2L_{xy} \|\nabla_y f(x(0), y(0))\|}{\sqrt{\mu_y T}}. \]  \hspace{1cm} (23)

**Proof** By using (20) and (21) we consider

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

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

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

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

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

where in the last inequality we use (22). Using (2) we obtain

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

\[ = -\frac{\alpha}{2} \|\nabla_x f(x, y(t))\|^2 - \frac{4L_{xy}^2 \alpha}{\mu_y^2} \|\nabla_y f(x, t, y(t))\|^2 \]

\[ \leq -\frac{\alpha}{2} \left( \|\nabla_x f(x, y(t))\|^2 + \|\nabla_y f(x, t, y(t))\|^2 \right), \]
which when taking the integral on both sides over $t$ from 0 to $T$ and rearrange we obtain

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

Thus, the preceding relation gives (23), i.e., for all $T > 0$

$$\min_{0 \leq t \leq T} \left\| \nabla_x f(x(t), y(t)) \right\| \leq \frac{\sqrt{2V(x(0), y(0))}}{\sqrt{\alpha T}} + \frac{\sqrt{2V_1(x(0), y(0))}}{\sqrt{\alpha T}} = \frac{L_{xy} \sqrt{2V_1(x(0), y(0))}}{\sqrt{T}} + \frac{2L_{xy} \|\nabla_y f(x(0), y(0))\|}{\sqrt{\mu_{x} T}}.$$

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 $x$. The analysis in this section is symmetric to the one in Section 3.3, so it is omitted here for brevity. More detail can be found in Doan (2021a).

**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

$$f(x_1, y) - f(x_2, y) - \langle \nabla f(x_2, y), x_1 - x_2 \rangle \geq \frac{\mu_x}{2} \|x_1 - x_2\|^2, \quad \forall x_1, x_2 \in \mathbb{R}^m. \quad (24)$$

The convergence rate of GDAD under Assumption 6 is presented 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

$$\gamma = \mu_x L_{xy}^2, \quad \alpha = \frac{4}{\mu_x^2}, \quad \beta = \frac{1}{L_{xy}^2}.$$  \quad (25)

Then we have for all $T \geq 0$

$$\min_{0 \leq t \leq T} \left\| \nabla_x f(x(t), y(t)) \right\| \leq \frac{L_{xy} \sqrt{2V_1(x(0), y(0))}}{\sqrt{T}} + \frac{2L_{xy} \|\nabla_x f(x(0), y(0))\|}{\sqrt{\mu_x T}}.$$  \quad (26)

4. 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


