2024 European Control Conference (ECC)
June 25-28, 2024. Stockholm, Sweden

# Convergence Rates of Gradient Descent-Ascent Dynamics under Delays in Solving Nonconvex Min-Max Optimization

Duy Anh Do
*Electrical and Computer Engineering Department*
*Virginia Tech*
Blacksburg, VA, US
duyanhdo@vt.edu

Thinh T. Doan
*Electrical and Computer Engineering Department*
*Virginia Tech*
Blacksburg, VA, US
thinhdoan@vt.edu

***Abstract*—In this paper, we study the so-called two-time-scale gradient descent-ascent method for solving min-max optimization problem. Our focus is to characterize the performance of this method, in particular, its continuous-time variant, under delays in gradient computation. Delays are common issues in large-scale optimization problems, which if not properly addressed, can lead to the instability of gradient methods. Unlike the classic gradient methods where theoretical guarantees for their performance under delays are well-studied, similar results for the gradient descent-ascent algorithms are very sparse. To address this gap, we provide a new analysis to characterize the convergence rates of the two-time-scale gradient descent-ascent dynamics under delays in solving nonconvex min-max optimization under the two-sided Polyak-Łojasiewicz conditions. Our results show that these dynamics converge exponentially to the optimal solution of the problem even under the impact of delays. The key idea in our analysis is to utilize the classic singular perturbation approach to design a coupling Lyapunov function to address the interaction between the gradient descent and ascent dynamics and the effect of delays. Finally, we provide a number of numerical simulations to illustrate our theoretical results.**

***Index Terms*—Gradient descent-ascent methods, min-max optimization.**

## I. Introduction

We consider the following optimization problem:

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

where $f:\mathbb{R}^m\times\mathbb{R}^n\to\mathbb{R}$ is nonconvex with respect to $x$ and nonconcave with respect to $y$. This min-max problem is of paramount importance because of its broad applications in different areas, for instance, game theory [1], [2], stochastic control and reinforcement learning [3], [4], machine learning [5], [6], optimization [7], [8], and training generative adversarial networks [9], [10]. For solving this problem, we are interested in studying gradient-based approach, in particular, gradient descent-ascent methods [11]–[14]. These methods iteratirely update the estimates $(x,y)$ of the optimal $(x^\star,y^\star)$ of $f$ by moving along the partial gradient directions $\nabla_x f$ and $\nabla_y f$, respectively.

Our focus in this paper is to understand the convergence of these methods under delays, i.e., we only have access to the delayed values of these gradients. Such a result has not yet been fully studied in the literature of min-max optimization, unlike the non-delay counterpart.

Our motivation to study the gradient descent-ascent dynamics under delay partly comes from large-scale machine learning applications such as training deep neural networks, where the dataset size is enormous. In these scenarios, calculating the gradients can be extremely time-consuming due to the computational complexity as well as the volume of data. Additionally, in distributed optimization, communication delays are common because it takes time for the gradient information to be transmitted between different nodes or machines. Therefore, we do not always have access to the immediate gradient information, and has to rely only the values of the gradients at some earlier time step to perform an update of the model estimate.

**Main contributions.** The objective of this paper is to characterize the convergence of continuous-time gradient descent-ascent dynamics for solving problem (1) under the impact of delays. Our main result is to show that this method converges exponentially to the optimal solution of the problem when the objective function $f$ satisfies the two-sided Polyak-Łojasiewicz condition. The key idea in our analysis is to use the classic singular perturbation theory to design Lyapunov functions to study the time-scale difference and interactions between the gradient descent and ascent dynamics. Finally, we provide numerical simulations to support our theoretical results.

### A. Related works

This work was supported by the National Science Foundation under ECCS-CAREER Grant No. 2339509.

**Min-max optimization.** Because of its significance, the problem of Min-Max optimization setting has been extensively studied in general [15], [16], and in the nonconvex-concave scenario in particular [17], [18]. However, theoretical guarantees in literature are very limited. Noteworthy studies in the domain of nonconvex-nonconcave [19]–[21] often show convergence to a stationary point, while our work establishes convergence to the saddle point of the function. The paper by [14] also utilizes the PŁ condition to find the saddle point at a sublinear rate. However, our work focuses on the continuous-time dynamics, and also provides an exponential convergence.

Among the two types of first-order methods for solving such problems, the single-loop algorithm is better applicable because of its simplicity in implementation. However, it is known that single-loop algorithms may fail to converge even in simple settings [22]. Our paper delves into such intricate issues by focusing on a special case of nonconvex-nonconcave, namely the 2-sided Polyak-Łojasiewicz condition. We employ a two-time-scale approach, which incorporates both fast and slow continuous-time gradient dynamics, as well as utilizing coupling Lyapunov functions to establish convergence, even when access to gradients is not immediately available.

**Optimization under delays.** Literature that concerns optimization problems under delays is quite extensive [23], [24], however, most have been focusing on delay models in the single optimization setting. In the study by [25], when the corresponding function is smooth and convex-concave, a delayed version of the extra gradient algorithm is shown to provably converge to the saddle point at a sublinear rate. In addition, the authors show that this convergence rate is exponential when the underlying function is strongly convex and strongly concave. In this work, our focus is to study the performance of gradient descent-ascent methods when the function is nonconvex-nonconcave but satisfies the two-sided Polyak-Łojasciewicz condition.

## II. Two-Time-Scale Gradient Descent-Ascent Dynamics

To solve problem (1), we consider the two-time-scale gradient descent ascent dynamics. In this paper, our focus is to study this method in the delay regime. In particular, at any given time $t$, we only have access to the delayed value of the gradient, i.e., $\nabla f(x(t-\tau), y(t-\tau))$, where $\tau > 0$ is the constant representing the delay. In this setting, the two-time-scale gradient descent-ascent dynamics is given as:

$$\begin{aligned} \dot{x} &= \frac{d}{dt}x(t) = -\alpha \nabla_x f(x(t-\tau), y(t-\tau)), \\ \dot{y} &= \frac{d}{dt}y(t) = \beta \nabla_y f(x(t-\tau), y(t-\tau)), \end{aligned} \tag{2}$$

where $\alpha$ and $\beta$ are two different step sizes, which will be chosen properly to guarantee the convergence of these dynamics.

**Main ideas of technical analysis.** The convergence analysis of (2) studied in this paper is mainly motivated by the classic singular perturbation theory [26]. Since $y$ is updated 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

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

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 [27], [28]. For the min-max optimization problem, we will consider the following two Lyapunov functions:

$$\begin{aligned} V_1(x) &= \max_{y\in\mathbb{R}^n} f(x, y) - \min_{x\in\mathbb{R}^m} \max_{y\in\mathbb{R}^n} f(x, y), \\ V_2(x, y) &= \max_{y\in\mathbb{R}^n} f(x, y) - f(x, y). \end{aligned} \tag{4}$$

Clearly, if we can show that $V_1(x_k)$ converges to 0, then $x_k$ converges to $x^*$, and similarly, $V_2(x_k, y_k)$ converges to 0, then $y_k$ converges to the optimal $y$ value given a particular $x$.

We conclude this section with the following assumptions:

**Assumption II.1.** *The function $f(.,.)$ has Lipschitz continuous gradients for each variables, i.e., there exists positive constant $L$ such that for all $x_1, x_2 \in \mathbb{R}^m$ and $y_1, y_2 \in \mathbb{R}^n$ we have:*

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

**Assumption II.2.** *For any $x \in \mathbb{R}^m$, 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:*

$$y^*(x) = \arg\max_{y\in\mathbb{R}^n} f(x, y),$$

*where $f(x, y^*(x))$ is finite.*

**Assumption II.3.** *There exists a global min-max solution $(x^*, y^*)$ for the problem in* (1)*:*

$$x^* = \arg\min_{x\in\mathbb{R}^m} f(x, y^*) \quad \textit{and} \quad y^* = \arg\max_{y\in\mathbb{R}^n} f(x^*, y) \tag{5}$$

Next, we present the definition of 2-sided Polyak-Łojasiewicz (PŁ) condition as follows.

**Definition 1.** *A continuous differentiable function $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ is called to satisfy two-sided PŁ conditions if there exists positive constant $\mu$ such that $\mu \leq L$, and the following conditions hold for all $(x, y) \in \mathbb{R}^m \times \mathbb{R}^n$:*

$$\begin{aligned} 2\mu[f(x,y) - \min_x f(x,y)] &\leq ||\nabla_x f(x,y)||^2, \\ 2\mu[\max_y f(x,y) - f(x,y)] &\leq ||\nabla_y f(x,y)||^2. \end{aligned} \tag{6}$$

In this paper, we assume that $f$ satisfies the two-sided Polyak-Łojasiewicz (PŁ) condition, a broader form of the well-known PŁ condition introduced by [29]. This condition serves as a sufficient guarantee for the exponential convergence of the classic gradient descent method towards the optimal solution of an unconstrained optimization problem. As shown in [30], the PŁ condition also implies the quadratic growth condition, i.e., given any $x$ we have for all $y \in \mathbb{R}^m$:

$$\max_{z \in \mathbb{R}^m} f(x,z) - f(x,y) \geq \frac{\mu}{2}\|\mathcal{P}_{\mathcal{Y}^\star(x)}[y] - y\|^2, \tag{7}$$

where we assume that $\mathcal{Y}^\star(x)$ is a nonempty solution set of $\max_y f(x,y)$ and $\mathcal{P}_{\mathcal{Y}^\star(x)}[y]$ is the projection of $y$ to this set. Finally, we consider the following lemma about the Lipschitz continuity of the gradient of $f(x, y^\star(x))$, which is a variant of the Danskin lemma [31][Proposition B.25] and studied in [19][Lemma A.5].

**Lemma II.1.** *Suppose that Assumptions II.1– II.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 $L + \frac{L}{\mu}$.*

## III. Main results

We begin our technical analysis of (2) by providing a bound on the time derivatives of $V_1$ and $V_2$:

**Lemma III.1.** *Suppose that Assumptions II.1– II.3 hold and $\alpha \leq \beta$. Then we have*

$$\begin{aligned} &\dot{V}_1(x(t)) \\ &\quad\leq -\frac{\alpha}{2}\|\nabla_x f(x(t), y^\star(x(t)))\|^2 + \frac{2L^2\alpha}{\mu}V_2(x(t),y(t)) \\ &\qquad - \frac{\alpha}{2}\|\nabla_x f(x(t-\tau), y(t-\tau))\|^2 \\ &\qquad + L^2\tau\alpha\beta^2\int_{t-\tau}^t \|\nabla f(x(u-\tau), y(u-\tau))\|^2 du. \end{aligned} \tag{8}$$

$$\begin{aligned} &\dot{V}_2(x(t),y(t)) \\ &\quad\leq -\frac{\beta}{2}||\nabla_y f(x(t),y(t))||^2 + \frac{4L^2\alpha}{\mu}V_2(x(t),y(t)) \\ &\qquad - \frac{\beta}{2}||\nabla_y f(x(t-\tau), y(t-\tau))||^2 \\ &\qquad + \frac{3L^2\tau\beta^3}{2}\Big[\int_{t-\tau}^t ||\nabla f(x(u-\tau), y(u-\tau))||^2 du\Big] \\ &\qquad + \frac{\alpha}{2}||\nabla_x f(x(t), y^*(x(t)))||^2. \end{aligned} \tag{9}$$

*Proof.* For convenience, we denote by $y^\star(x) = \mathcal{P}_{\mathcal{Y}^\star(x)}[y]$, where recall that $\mathcal{Y}^\star(x)$ is the solution set of $\max_y f(x,y)$ for a given $x$. We first show (8). The time derivative of $V_1$ over the trajectory $\dot{x}$ in (2) is given as

$$\begin{aligned} &\dot{V}_1(x(t)) \\ &= \frac{d}{dt}V_1(x(t)) = \nabla_x f(x(t), y^\star(x(t)))\dot{x}(t) \\ &= -\alpha\langle \nabla_x f(x(t), y^\star(x(t))), \nabla_x f(x(t-\tau), y(t-\tau))\rangle \\ &= -\frac{\alpha}{2}\|\nabla_x f(x(t), y^\star(x(t)))\|^2 \\ &\quad - \frac{\alpha}{2}\|\nabla_x f(x(t-\tau), y(t-\tau))\|^2 \\ &\quad + \frac{\alpha}{2}\|\nabla_x f(x(t), y^\star(x(t))) - \nabla_x f(x(t-\tau), y(t-\tau))\|^2 \\ &\leq -\frac{\alpha}{2}\|\nabla_x f(x(t), y^\star(x(t)))\|^2 \\ &\quad - \frac{\alpha}{2}\|\nabla_x f(x(t-\tau), y(t-\tau))\|^2 \\ &\quad + \frac{L^2\alpha}{2}\big(\|x(t) - x(t-\tau)\|^2 + \|y^\star(x(t)) - y(t-\tau)\|^2\big) \\ &\leq -\frac{\alpha}{2}\|\nabla_x f(x(t), y^\star(x(t)))\|^2 \\ &\quad - \frac{\alpha}{2}\|\nabla_x f(x(t-\tau), y(t-\tau))\|^2 \\ &\quad + \frac{L^2\alpha}{2}\big(\|x(t) - x(t-\tau)\|^2 + 2\|y(t) - y(t-\tau)\|^2\big) \\ &\quad + L^2\alpha\|y^\star(x(t)) - y(t)\|^2., \end{aligned} \tag{10}$$

where the first inequality is due the Lipschitz continuity of $\nabla_x f$ and the last inequality is due to the Cauchy-Schwartz inequality. Taking integration on both sides of (2) over $t$ gives

$$x(t) - x(t-\tau) = -\alpha\int_{t-\tau}^t \nabla_x f(x(u-\tau), y(u-\tau))du,$$

which by using the Cauchy-Schwartz yields

$$\|x(t) - x(t-\tau)\|^2 \leq \alpha^2\tau\int_{t-\tau}^t \|\nabla_x f(x(u-\tau), y(u-\tau))\|^2 du.$$

Similarly, we obtain

$$\|y(t) - y(t-\tau)\|^2 \leq \beta^2\tau\int_{t-\tau}^t \|\nabla_y f(x(u-\tau), y(u-\tau))\|^2 du.$$

Combining the two terms above and using the fact that $||\nabla f(x,y)||^2 = ||\nabla_x f(x,y)||^2 + ||\nabla_y f(x,y)||^2$ give

$$\begin{aligned} &\frac{L^2\alpha}{2}\big(\|x(t) - x(t-\tau)\|^2 + 2\|y(t) - y(t-\tau)\|^2\big) \\ &\quad\leq L^2\tau\alpha\beta^2\int_{t-\tau}^t ||\nabla f(x(u-\tau), y(u-\tau))||^2 du, \end{aligned} \tag{11}$$

where we use $\alpha \leq \beta$. Next, using the PŁ condition of $f$ we obtain

$$\|y^\star(x(t)) - y(t)\|^2 \leq \frac{2}{\mu}\big(\max_y f(x(t),y(t)) - f(x(t),y(t))\big).$$

Substituting the preceding relations into (10) we obtain the desired inequality (8).
Next, we show (9). Using (8) we have:

$$
\begin{aligned}
&\dot{V}_2(x(t),y(t))\\
&=\nabla_x f(x(t),y^\star(x(t)))\dot{x}(t)-\nabla_x f(x(t),y(t))\dot{x}(t)\\
&\quad-\nabla_y f(x(t),y(t))\dot{y}(t),\\
&=\dot{V}_1(x(t))+\alpha\langle\nabla_x f(x(t),y(t)),\nabla_x f(x(t-\tau),y(t-\tau))\rangle\\
&\quad-\beta\langle\nabla_y f(x(t),y(t)),\nabla_y f(x(t-\tau),y(t-\tau))\rangle\\
&=\dot{V}_1(x(t))+\frac{\alpha}{2}||\nabla_x f(x(t),y(t))||^2\\
&\quad+\frac{\alpha}{2}||\nabla_x f(x(t-\tau),y(t-\tau))||^2\\
&\quad-\frac{\alpha}{2}||\nabla_x f(x(t),y(t))-\nabla_x f(x(t-\tau),y(t-\tau))||^2\\
&\quad-\frac{\beta}{2}||\nabla_y f(x(t),y(t))||^2-\frac{\beta}{2}||\nabla_y f(x(t-\tau),y(t-\tau))||^2\\
&\quad+\frac{\beta}{2}||\nabla_y f(x(t),y(t))-\nabla_y f(x(t-\tau),y(t-\tau))||^2.
\end{aligned}
\tag{12}
$$

Using the Lipschitz continuous gradient of $\nabla_x f$ we have

$$
\begin{aligned}
&\frac{\alpha}{2}||\nabla_x f(x(t),y(t))||^2\\
&\le\alpha||\nabla_x f(x(t),y^*(x(t)))||^2\\
&\qquad+\alpha||\nabla_x f(x(t),y^*(x(t)))-\nabla_x f(x(t),y(t))||^2\\
&\le\alpha||\nabla_x f(x(t),y^*(x(t)))||^2+L^2\alpha||y(t)-y^*x(t)||^2.
\end{aligned}
$$

Using the same argument as in (11), we obtain

$$
\begin{aligned}
&\frac{\beta}{2}||\nabla_y f(x(t),y(t))-\nabla_y f(x(t-\tau),y(t-\tau))||^2\\
&\le\frac{L^2\beta}{2}\big(\|x(t)-x(t-\tau)\|^2+\|y(t)-y(t-\tau)\|^2\big)\\
&\le\frac{L^2\tau\beta^3}{2}\Big[\int_{t-\tau}^{t}||\nabla f(x(u-\tau),y(u-\tau))||^2du\Big].
\end{aligned}
$$

Substituting the preceding relations into (12) gives (9), i.e.,

$$
\begin{aligned}
&\dot{V}_2(x(t),y(t))\\
&\le\dot{V}_1(x(t))+\frac{\alpha}{2}||\nabla_x f(x(t-\tau),y(t-\tau))||^2\\
&\quad-\frac{\beta}{2}||\nabla_y f(x(t),y(t))||^2-\frac{\beta}{2}||\nabla_y f(x(t-\tau),y(t-\tau))||^2\\
&\quad+\frac{L^2\tau\beta^3}{2}\Big[\int_{t-\tau}^{t}||\nabla f(x(u-\tau),y(u-\tau))||^2du\Big]\\
&\quad+\alpha||\nabla_x f(x(t),y^*(x(t)))||^2+L^2\alpha||y(t)-y^*x(t)||^2\\
&\quad-\frac{\alpha}{2}||\nabla_x f(x(t),y(t))-\nabla_x f(x(t-\tau),y(t-\tau))||^2\\
&\le-\frac{\beta}{2}||\nabla_y f(x(t),y(t))||^2-\frac{\beta}{2}||\nabla_y f(x(t-\tau),y(t-\tau))||^2\\
&\quad+\frac{3L^2\tau\beta^3}{2}\Big[\int_{t-\tau}^{t}||\nabla f(x(u-\tau),y(u-\tau))||^2du\Big]\\
&\quad+\frac{\alpha}{2}||\nabla_x f(x(t),y^*(x(t)))||^2+\frac{4L^2\alpha}{\mu}V_2(x(t),y(t))\\
&\quad-\frac{\alpha}{2}\|\nabla_x f(x(t),y(t))-\nabla_x f(x(t-\tau),y(t-\tau))\|^2
\end{aligned}
$$

where the last inequality is due to the PŁ condition. □

We next present an important result that will allow us to establish the convergence of the dynamics in (2). Our result is a continuous-time variant of Lemma 5 in [32].

**Lemma III.2.** *Let $\{V(t),W(t)\}_{t\ge0}$ be two nonnegative continuous-time sequences satisfying*

$$
\dot{V}(t)\le-\sigma V(t)-\alpha W(t-\tau)+\lambda\int_{u=t-\tau}^{t}W(u-\tau)du,
\tag{13}
$$

*where $\alpha,\tau,\sigma$, and $\lambda$ are positive constants that satisfies*

$$
\alpha-\frac{\lambda}{\sigma}e^{\sigma\tau}\ge0.
\tag{14}
$$

*Then we have*

$$
V(t)\le V(0)e^{-\sigma t}.
\tag{15}
$$

*Proof.* First, using the integral by part we have

$$
\int_{u=0}^{t}\frac{\dot{V}(u)}{e^{-\sigma u}}du=\frac{V(u)}{e^{-\sigma u}}\Big|_0^t-\int_{u=0}^{t}\frac{\sigma V(u)}{e^{-\sigma u}}du.
\tag{16}
$$

Second, using the fact that $W(t)=0$ for all $t\le0$ we consider

$$
\begin{aligned}
&-\alpha\int_{u=0}^{t}\frac{W(u-\tau)}{e^{-\sigma u}}du+\lambda\int_{u=0}^{t}\int_{s=u-\tau}^{u}\frac{W(s-\tau)}{e^{-\sigma u}}dsdu\\
&\le-\alpha\int_{u=0}^{t}\frac{W(u-\tau)}{e^{-\sigma u}}du+\lambda\int_{s=0}^{t}\int_{u=s}^{s+\tau}\frac{W(s-\tau)}{e^{-\sigma u}}duds\\
&\le-\alpha\int_{u=0}^{t}\frac{W(u-\tau)}{e^{-\sigma u}}du+\frac{\lambda}{\sigma}\int_{s=0}^{t}W(s-\tau)e^{\sigma(s+\tau)}ds\\
&\le-(\alpha-\frac{\lambda}{\sigma}e^{\sigma\tau})\int_{u=0}^{t}\frac{W(u-\tau)}{e^{-\sigma u}}du\le0,
\end{aligned}
\tag{17}
$$

where the last inequality is due to (14). Thus, by diving both sides of (13) by $e^{-\sigma t}$, taking integral both sides from $0,\dots,t$, and using Eqs. (16) and (17) we obtain (15), i.e.,

$$
V(t)e^{\sigma t}-V(0)\le0\Rightarrow V(t)\le V(0)e^{-\sigma t}.
$$

□

Finally, we present the main result of this paper in the following theorem, where we will show that the sequence $\{x(t),y(t)\}$ returned by the gradient descent-ascent dynamics in (2) converges exponentially to the optimal solution of (1).

**Theorem III.3.** *Suppose that Assumptions II.1– II.3 hold. Let $\gamma=\frac{8L^2}{\mu^2}$ and the step sizes $\alpha,\beta$ be chosen as*

$$
\alpha=\frac{\mu^7}{2^{14}L^8\tau},\ \beta=\frac{\mu^5}{2^{10}L^6\tau}.
\tag{18}
$$

*Then, we have*

$$
V(x(t),y(t))\le e^{-\mu\alpha t}V(x(0),y(0)).
\tag{19}
$$

**Remark III.1.** *As $\alpha$ is inversely proportional to the delay constant $\tau$, our result shows that the convergence rate scales linearly with the factor $1/\tau$, which is similar to the results of gradient methods under delays in solving optimization problems $\min_x f(x)$ [32].*

*Proof.* By using (8) and (9) we have

$$\begin{aligned}
&\dot{V}(x(t),y(t))\\
&=\dot{V}_1(x(t))+\frac{\gamma\alpha}{\beta}\dot{V}_2(x(t),y(t))\\
&\le \frac{-\alpha}{2}\|\nabla_x f(x(t),y^\star(x(t)))\|^2+\frac{2L^2\alpha}{\mu}V_2(x(t),y(t))\\
&\quad-\frac{\alpha}{2}\|\nabla_x f(x(t-\tau),y(t-\tau))\|^2\\
&\quad+L^2\tau\alpha\beta^2\int_{t-\tau}^{t}\|\nabla f(x(u-\tau),y(u-\tau))\|^2du\\
&\quad-\frac{\gamma\alpha}{2}||\nabla_y f(x(t),y(t))||^2\\
&\quad-\frac{\gamma\alpha}{2}||\nabla_y f(x(t-\tau),y(t-\tau))||^2\\
&\quad+\frac{3L^2\tau\alpha\beta^2}{2}\gamma\Big[\int_{t-\tau}^{t}||\nabla f(x(u-\tau),y(u-\tau))||^2du\Big]\\
&\quad-\frac{\alpha}{4}(1-\frac{\gamma\alpha}{2\beta})||\nabla_x f(x(t),y^*(x(t)))||^2\\
&\quad+\frac{4L^2\gamma\alpha^2}{\mu\beta}V_2(x(t),y(t)).
\end{aligned}$$

As a direct result of the PŁ condition in (6), we have

$$\begin{aligned}
&-\frac{\gamma\alpha}{4}||\nabla_y f(x(t),y(t))||^2\le-\frac{\gamma\mu\alpha}{2}V_2(x(t),y(t))\\
&-\frac{\alpha}{4}||\nabla_x f(x(t),y^*(x(t)))||^2\le-\frac{\mu\alpha}{2}V_1(x(t)),
\end{aligned}$$

and therefore, rearranging the terms and using $\gamma>1$ gives

$$\begin{aligned}
&\dot{V}(x(t),y(t))\\
&\le\frac{-\gamma\mu\alpha}{2}V_2(x(t),y(t))-\frac{\mu\alpha}{2}V_1(x(t))\\
&\quad-\alpha(\frac{-4L^2\gamma\alpha}{\mu\beta}-\frac{2L^2}{\mu}+\frac{\gamma\mu}{2})V_2(x(t),y(t))\\
&\quad-\frac{\alpha}{4}\Big[1-\frac{2\gamma\alpha}{\beta}\Big]||\nabla_x f(x(t),y^*(x(t)))||^2\\
&\quad+\Big[1+\frac{3\gamma}{2}\Big]L^2\tau\alpha\beta^2\Big[\int_{t-\tau}^{t}||\nabla f(x(u-\tau),y(u-\tau))||^2du\Big]\\
&\quad-\frac{\alpha}{2}||\nabla f(x(t-\tau),y(t-\tau))||^2,
\end{aligned}$$

where the inequality is due to the PŁ condition. Recall that

$$\gamma=\frac{8L^2}{\mu^2},\ \alpha=\frac{\mu^7}{2^{14}L^8\tau},\ \beta=\frac{\mu^5}{2^{10}L^6\tau},$$

which gives

$$\frac{\gamma\mu}{2}-\frac{4L^2\gamma\alpha}{\mu\beta}-\frac{2L^2}{\mu}=0\text{ and }1-\frac{2\gamma\alpha}{\beta}=\frac{1}{2}=0.$$

Thus we obtain from the equation above

$$\begin{aligned}
&\dot{V}(x(t),y(t))\\
&\le\frac{-\mu\alpha}{2}V(x(t),y(t))-\frac{\alpha}{2}||\nabla f(x(t-\tau),y(t-\tau))||^2\\
&+\Big[1+\frac{3\gamma}{2}\Big]L^2\tau\alpha\beta^2\Big[\int_{t-\tau}^{t}||\nabla f(x(u-\tau),y(u-\tau))||^2du\Big]
\end{aligned}$$

We next apply the results in Lemma (III.2) to the preceding equation. Note that with $\sigma=\frac{\mu\alpha}{2}$, $\frac{\alpha}{2}$ and $\lambda=(1+\frac{3\gamma}{2})L^2\tau\alpha\beta^2$ we have the condition in (14) is satisfied. Thus using (15) we obtain (19), i.e.,

$$\begin{aligned}
V(x(t),y(t))&\le e^{-\frac{\mu\alpha}{2}t}V(x(0),y(0))\\
&=e^{-\frac{\mu^8}{2^{15}L^8\tau}t}V(x(0),y(0)).
\end{aligned}\tag{20}$$

□

## IV. Simulations

In this section, we illustrate our theoretical results in Theorem III.3 by simulations. In particular, we will apply the dynamics in (2) to optimize the following function

$$f(x,y)=x^2+3\sin^2(x)\sin^2(y)-4y^2-10\sin^2(y),$$

which satisfies the PŁ conditions in (6). Regarding our implementation, we consider the discrete-time variant of (2) given as:

$$\begin{aligned}
x_{t+1}&=x_t-\alpha\nabla_x f(x(t-\tau),y(t-\tau)),\\
y_{t+1}&=y_t+\beta\nabla_y f(x(t-\tau),y(t-\tau)).
\end{aligned}\tag{21}$$

We will illustrate the convergence of the Lyapunov function in (4) under different values of delay constant $\tau$. In our simulation, we will choose the step sizes $\alpha=0.002$, and $\beta=0.02$.

First, we simulate the updates in (21) when $\tau=1$ to illustrate its convergence rate. Our simulation is presented in Figure 1. In this figure, we observe that $V$ decreases to zero exponentially fast, which agrees with our theoretical result in Theorem III.3.

Second, we simulate (21) for different values of delays to understand their impacts. In particular, we vary $\tau=1,2,4,6,8$ and fix the number of iterations to 100 for each simulation. The outputs of our simulations are shown in Figure 2. We again observe that the rates of $V$ decay to zero are exponential in all cases. In addition, as the values of $\tau$ increase the rates of convergence of $V$ decrease, which agree with out theoretical results.

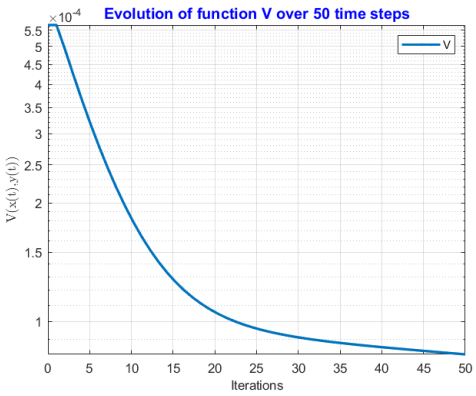


Fig. 1. Evolutions of $V$ through 100 iterations as $\tau=1$.

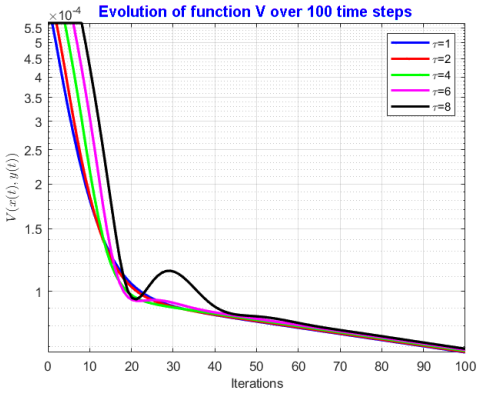


Fig. 2. Evolution of $V$ through 100 iterations over different time delay.

## REFERENCES


[1] T. Başar and G. J. Olsder, *Dynamic noncooperative game theory*. SIAM, 1998.

[2] L. S. Shapley, "Stochastic games," *Proceedings of the national academy of sciences*, vol. 39, no. 10, pp. 1095–1100, 1953.

[3] E. Altman, *Constrained Markov decision processes*. Routledge, 2021.

[4] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *International conference on machine learning*. PMLR, 2017, pp. 22–31.

[5] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," *arXiv preprint arXiv:1611.01236*, 2016.

[6] Q. Qian, S. Zhu, J. Tang, R. Jin, B. Sun, and H. Li, "Robust optimization over multiple domains," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 4739–4746.

[7] G. Lan, S. Lee, and Y. Zhou, "Communication-efficient algorithms for decentralized and stochastic optimization," *Mathematical Programming*, vol. 180, no. 1-2, pp. 237–284, 2020.

[8] 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.

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

[10] L. Mescheder, S. Nowozin, and A. Geiger, "The numerics of gans," *Advances in neural information processing systems*, vol. 30, 2017.

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

[12] 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.

[13] T. Lin, C. Jin, and M. Jordan, "On gradient descent ascent for nonconvex-concave minimax problems," in *International Conference on Machine Learning*. PMLR, 2020, pp. 6083–6093.

[14] J. Yang, N. Kiyavash, and N. He, "Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1153–1165, 2020.

[15] C. Jin, P. Netrapalli, and M. Jordan, "What is local optimality in nonconvex-nonconcave minimax optimization?" in *International conference on machine learning*. PMLR, 2020, pp. 4880–4889.

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

[17] T. Lin, C. Jin, and M. I. Jordan, "Near-optimal algorithms for minimax optimization," in *Conference on Learning Theory*. PMLR, 2020, pp. 2738–2779.

[18] 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.

[19] 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.

[20] J. Diakonikolas, C. Daskalakis, and M. I. Jordan, "Efficient methods for structured nonconvex-nonconcave min-max optimization," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 2746–2754.

[21] S. Lee and D. Kim, "Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22 588–22 600, 2021.

[22] D. Balduzzi, S. Racaniere, J. Martens, J. Foerster, K. Tuyls, and T. Graepel, "The mechanics of n-player differentiable games," in *International Conference on Machine Learning*. PMLR, 2018, pp. 354–363.

[23] H. Al-Lawati and S. C. Draper, "Gradient delay analysis in asynchronous distributed optimization," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 4207–4211.

[24] A. Agarwal and J. C. Duchi, "Distributed delayed stochastic optimization," *Advances in neural information processing systems*, vol. 24, 2011.

[25] A. Adibi, A. Mitra, and H. Hassani, "Min-max optimization under delays," *arXiv preprint arXiv:2307.06886*, 2023.

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

[27] T. T. Doan, "Nonlinear two-time-scale stochastic approximation: Convergence and finite-time performance," *arXiv preprint arXiv:2011.01868*, 2020.

[28] A. Dutta, N. Masrourisaadat, and T. T. Doan, "Convergence rates of decentralized gradient methods over cluster networks," *arXiv preprint arXiv:2110.06992*, 2021.

[29] B. T. Polyak, "Gradient methods for the minimisation of functionals," *USSR Computational Mathematics and Mathematical Physics*, vol. 3, no. 4, pp. 864–878, 1963.

[30] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*. Springer, 2016, pp. 795–811.

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

[32] H. R. Feyzmahdavian and M. Johansson, "Asynchronous iterations in optimization: New sequence results and sharper algorithmic guarantees." *J. Mach. Learn. Res.*, vol. 24, pp. 158–1, 2023.