RESEARCH ARTICLE

On the Convergence of Nonconcave-Nonconvex Max-Min Optimization Problem

Xuelin Zhang[1,*]

[1] College of Informatics, Huazhong Agricultural University, Wuhan 430070, China

Abstract

Despite extensive study of max–min problems, convergence analysis for the challenging nonconvex–nonconcave setting remains limited. This paper addresses the convergence analysis of nonconvex–nonconcave max–min problems. A novel analytical framework is developed by employing carefully constructed auxiliary functions and leveraging two-sided Polyak–Łojasiewicz (PL) and Quadratic Growth (QG) conditions to characterize the convergence behavior. Under these conditions, it is shown that the Stochastic Alternating Gradient Descent Ascent (SAGDA) algorithm achieves a convergence rate of $\mathcal{O}(1/K)$, where K denotes the number of iterations. Notably, this result matches convergence rates typically obtained in (weakly) convex–concave minimax settings while requiring significantly milder geometric assumptions. The theoretical results are further validated through empirical experiments on realistic examples.

Keywords: max-min optimization, convergence guarantee, nonconvex-nonconcave problems, two-sided polyak-lojasiewicz, quadratic growth.

1 Introduction

Consider the finite-dimensional Euclidean spaces $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, and a smooth function $\mathcal{L} : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$. Let $\Theta_1 \subset \mathbb{R}^{d_1}$ and $\Theta_2 \subset \mathbb{R}^{d_2}$ be nonempty closed convex sets. The minimax problem can be formulated as

$$\min_{\boldsymbol{\theta}_1 \in \Theta_1} \max_{\boldsymbol{\theta}_2 \in \Theta_2} \mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2). \tag{1}$$

The above minimax problem has attracted significant attention across optimization, statistics, and machine learning literatures [2, 7, 14]. The existing study on minimax problems has primarily concentrated on the convex-concave setting, where $\mathcal{L}(\boldsymbol{\theta}_1, \cdot)$ is convex for every $\boldsymbol{\theta_1} \in \mathbb{R}^{d_1}$ and $\mathcal{L}(\cdot, \boldsymbol{\theta}_2)$ is concave for every $\boldsymbol{\theta_2} \in \mathbb{R}^{d_2}$. However, in the nonconvex-nonconcave scenario, computing the minimax or saddle point is generally NP-hard.

In this work, the convergence behaviour of another max–min problem under nonconcave–nonconvex conditions is characterized under the following conditions.

$$\max_{\boldsymbol{\theta}_2 \in \Theta_2} \min_{\boldsymbol{\theta}_1 \in \Theta_1} \mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2). \tag{2}$$

The key challenges for analysing the convergence of (2) lie in the gaps between its max-min type (2) and the

Submitted: 14 October 2025
Accepted: 18 October 2025
Published: 08 December 2025

Vol. 1, **No.** 2, 2025.
10.62762/JNSPM.2025.112121

***Corresponding author:**
✉ Xuelin Zhang
zhangxuelin@webmail.hzau.edu.cn

Citation
Zhang, X. (2025). On the Convergence of Nonconcave-Nonconvex Max-Min Optimization Problem. *Journal of Numerical Simulations in Physics and Mathematics*, 1(2), 76–83.

widely investigated min-max problems (1). Under the common convex-concave hypothesis, Sion's minimax theorem [12] guarantees the equivalence of these two types of problems, where

$$\max_{\boldsymbol{\theta}_2\in\Theta_2}\min_{\boldsymbol{\theta}_1\in\Theta_1}\mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)=\min_{\boldsymbol{\theta}_1\in\Theta_1}\max_{\boldsymbol{\theta}_2\in\Theta_2}\mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2). \quad (3)$$

In contrast, this paper aims to derive fine-grained algorithmic convergence directly for a subclass of nonconcave-nonconvex max-min problems in (2) under some mild conditions to cover more practical scenarios.

2 Definitions and Assumptions

Inspired by theoretical works on min–max problems [6, 13], auxiliary functions are constructed to derive the convergence rate of the max–min problem (2), e.g.,

$$g(\boldsymbol{\theta}_2):=\min_{\boldsymbol{\theta}_1}\mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\text{ and }g^*=\max_{\boldsymbol{\theta}_2}g(\boldsymbol{\theta}_2). \quad (4)$$

Then the approximation performance of parameters $\theta=(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ at k-th iteration can be measured by

$$\begin{aligned}P_k:=&a_k+b_k\\:=&\mathbb{E}\left[\mathcal{L}\left(\boldsymbol{\theta}_1^k,\boldsymbol{\theta}_2^k\right)-g\left(\boldsymbol{\theta}_2^k\right)\right]+\mathbb{E}\left[g^*-g\left(\boldsymbol{\theta}_2^k\right)\right].\end{aligned} \quad (5)$$

Above decomposition is motivated by the definition of minimax and Nash equilibrium (saddle) points [9, 13], as $\mathcal{L}\left(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\right)-g\left(\boldsymbol{\theta}_2\right)$ and $g^*-g(\boldsymbol{\theta}_2)$ are non-negative for any $(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\in\Theta_1\times\Theta_2$, and both equal to 0 if and only if $(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ is a minimax point.

Definition 1. (*global minimax point and Nash equilibrium point* [13])

(1) $(\boldsymbol{\theta}_1^*,\boldsymbol{\theta}_2^*)$ *is a global minimax point point, if for any* $(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\in\Theta_1\times\Theta_2$,

$$\mathcal{L}\left(\boldsymbol{\theta}_1^*,\boldsymbol{\theta}_2\right)\leq\mathcal{L}\left(\boldsymbol{\theta}_1^*,\boldsymbol{\theta}_2^*\right)\leq\max_{\boldsymbol{\theta}_2'}\mathcal{L}\left(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2'\right). \quad (6)$$

(2) $(\boldsymbol{\theta}_1^*,\boldsymbol{\theta}_2^*)$ *is a Nash equilibrium point, if for any* $(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\in\Theta_1\times\Theta_2$,

$$\mathcal{L}\left(\boldsymbol{\theta}_1^*,\boldsymbol{\theta}_2\right)\leq\mathcal{L}\left(\boldsymbol{\theta}_1^*,\boldsymbol{\theta}_2^*\right)\leq\mathcal{L}\left(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2^*\right). \quad (7)$$

Definition 2. (*PL condition*) *The differentiable function* $h(\boldsymbol{\theta})$ *satisfies* μ*-PL condition if* $\forall\boldsymbol{\theta}$*, there holds*

$$\|\nabla_{\boldsymbol{\theta}}h(\boldsymbol{\theta})\|^2\geq2\mu\left(h(\boldsymbol{\theta})-\min_{\boldsymbol{\theta}}h(\boldsymbol{\theta})\right). \quad (8)$$

Definition 3. (*QG condition* [9]) *The function* $h(\boldsymbol{\theta})$ *satisfies the* Γ*-Quadratic Growth* (*QG*) *condition with constant* $\Gamma>0$ *if* $\forall\boldsymbol{\theta}$*, there holds*

$$h(\boldsymbol{\theta})-\min_{\boldsymbol{\theta}}h(\boldsymbol{\theta})\geq\frac{\Gamma}{2}\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|^2,\ \textit{where}\ \boldsymbol{\theta}^*\in\arg\min_{\boldsymbol{\theta}}h(\boldsymbol{\theta}).$$

The following assumptions have been commonly used for analyzing the convergence behavior of minimax problems [1, 5, 6, 12, 13] and projection-based algorithms [11, 17].

Assumption 1. *There exists a positive constant* $L_t>0$ *such that*

$$\begin{aligned}&\max_{\boldsymbol{\theta}\in\{\boldsymbol{\theta_1},\boldsymbol{\theta_2}\}}\left\{\left\|\nabla_{\boldsymbol{\theta}}\mathcal{L}\left(\boldsymbol{\theta_1},\boldsymbol{\theta_2}\right)-\nabla_{\boldsymbol{\theta}}\mathcal{L}\left(\boldsymbol{\theta_1}',\boldsymbol{\theta_2}'\right)\right\|_2\right\}\\\leq&L_t\left(\left\|\boldsymbol{\theta_1}-\boldsymbol{\theta_1}'\right\|_2+\left\|\boldsymbol{\theta_2}-\boldsymbol{\theta_2}'\right\|_2\right),\forall(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2),(\boldsymbol{\theta}_1',\boldsymbol{\theta}_2').\end{aligned}$$

Assumption 2. *There exist saddle points* $(\boldsymbol{\theta}_1^*,\boldsymbol{\theta}_2^*)$ *for* $\mathcal{L}$*. Naturally, we further assume that* $\min_{\boldsymbol{\theta}_1}\mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ *(or* $\max_{\boldsymbol{\theta}_2}\mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$*) has a nonempty solution set and an optimal value for any fixed* $\boldsymbol{\theta}_2$ *(or* $\boldsymbol{\theta}_1$*).*

Assumption 3. *Gradients* $\nabla_{\boldsymbol{\theta}_1}\mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2;\Xi)$ *and* $\nabla_{\boldsymbol{\theta}_2}\mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\Xi)$ *with stochastic sample* Ξ *are both unbiased estimators of* $\nabla_{\boldsymbol{\theta}_1}\mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ *and* $\nabla_{\boldsymbol{\theta}_2}\mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$*, respectively. Both of them have bounded variance* $\sigma^2>0$*.*

Assumption 4. (*Two-sided Polyak-Łojasiewicz condition* [13]) *With continuously differentiable* $\mathcal{L}$ *and for any* $\boldsymbol{\theta}_1,\boldsymbol{\theta}_2$*, there exist constants* $\mu_1,\mu_2>0$ *such that*

$$\|\nabla_{\boldsymbol{\theta}_1}\mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\|_2^2\geq2\mu_1\left[\mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)-\min_{\boldsymbol{\theta}_1}\mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\right]$$

and

$$\|\nabla_{\boldsymbol{\theta}_2}\mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\|_2^2\geq2\mu_2\left[\max_{\boldsymbol{\theta}_2}\mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)-\mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\right].$$

Assumption 4 shows that $\mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ satisfies the Polyak-Łojasiewicz (PL) condition with constant μ_1 for $\boldsymbol{\theta}_1$, and $-\mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ satisfies PL condition with constant μ_2 concerning $\boldsymbol{\theta}_2$. It does not imply convexity-concavity, and is much weaker than the strong-convexity-strong-concavity condition [10].

3 Two-sided PL and QG Properties

For simplicity, the stochastic alternating gradient descent ascent algorithm (SAGDA; see Algorithm 1) is adopted to solve the max–min problem (2). Before presenting the convergence proof of Theorem 1, several key lemmas are introduced.

Algorithm 1: SAGDA for solving max-min problem

Input: Training data $\{(x_i, y_i)\}_{i=1}^n$, step sizes at k-th iteration γ_1^k and γ_2^k, gradient estimator $\mathbf{E}_{\boldsymbol{\theta}_1}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdot)$ for $\nabla_{\boldsymbol{\theta}_1}\mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, estimator $\mathbf{E}_{\boldsymbol{\theta}_2}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdot)$ for $\nabla_{\boldsymbol{\theta}_2}\mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$.
Initialization: $\boldsymbol{\theta}_1^0$, $\boldsymbol{\theta}_2^0$.
for $k = 0$ to $K-1$ **do**
1) Randomly select i.i.d. sample pairs (Ξ_1^k, Ξ_2^k)
2) Update $\boldsymbol{\theta}_1$ by $\boldsymbol{\theta}_1^{k+1} = \boldsymbol{\theta}_1^k - \gamma_1^k \mathbf{E}_{\boldsymbol{\theta}_1}(\boldsymbol{\theta}_1^k, \boldsymbol{\theta}_2^k, \Xi_1^k)$
3) Update $\boldsymbol{\theta}_2$ by $\boldsymbol{\theta}_2^{k+1} = \boldsymbol{\theta}_2^k + \gamma_2^k \mathbf{E}_{\boldsymbol{\theta}_2}(\boldsymbol{\theta}_1^{k+1}, \boldsymbol{\theta}_2^k, \Xi_2^k)$
end for
Output: Parameter $(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) = (\boldsymbol{\theta}_1^K, \boldsymbol{\theta}_2^K)$.

In Lemma 2 and Proposition 1, the smoothness parameters of the auxiliary function

$$g(\boldsymbol{\theta_2}) := \min_{\boldsymbol{\theta_1}} \mathcal{L}(\boldsymbol{\theta_1}, \boldsymbol{\theta_2})$$

are established for the max–min setting, in contrast to the auxiliary function

$$\max_{\boldsymbol{\theta_2}} \mathcal{L}(\boldsymbol{\theta_1}, \boldsymbol{\theta_2})$$

considered in the minimax case [9].

Lemma 1 reveals the connection between PL and QG conditions.

Lemma 1. (*Corollary of Theorem 2 in* [5]). *If function $h(\boldsymbol{\theta})$ is μ-PL, then $h(\boldsymbol{\theta})$ also satisfies the Quadratic Growth condition with constant $\Gamma = 4\mu$.*

Lemma 2. *Let Assumptions 1 and 4 hold. Define $g(\boldsymbol{\theta}_2) = \min_{\boldsymbol{\theta}_1} \mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, $A(\boldsymbol{\theta}_2) = \arg\min_{\boldsymbol{\theta}_1} \mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ and assume $A(\boldsymbol{\theta}_2)$ is closed. Then for any $\boldsymbol{\theta}_2', \boldsymbol{\theta}_2''$ and $\boldsymbol{\theta}_1' \in A\left(\boldsymbol{\theta}_2'\right)$, there exists $\boldsymbol{\theta}_1'' \in A\left(\boldsymbol{\theta}_2''\right)$ such that*

$$\left\|\boldsymbol{\theta}_1' - \boldsymbol{\theta}_1''\right\| \leq \frac{L_t}{2\mu_1} \left\|\boldsymbol{\theta}_2' - \boldsymbol{\theta}_2''\right\|. \tag{9}$$

Proof. Under Assumption 1 with $\boldsymbol{\theta}_1' \in A\left(\boldsymbol{\theta}_2'\right)$, It can be derived that

$$\begin{aligned}\|\nabla_{\boldsymbol{\theta}_1}\mathcal{L}(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2') - \nabla_{\boldsymbol{\theta}_1}\mathcal{L}(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2'')\| =& \|\nabla_{\boldsymbol{\theta}_1}\mathcal{L}(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2'')\| \\ \leq& L_t\|\boldsymbol{\theta}_2' - \boldsymbol{\theta}_2''\|.\end{aligned}$$

Under μ_1-PL condition concerning $\boldsymbol{\theta}_1$ and Assumption 1, it follows that

$$\mathcal{L}(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2'') - \mathcal{L}(\boldsymbol{\theta}_1'', \boldsymbol{\theta}_2'') = \mathcal{L}(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2'') - \min_{\boldsymbol{\theta}_1} \mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2'')$$

and

$$\begin{aligned}&\mathcal{L}(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2'') - \min_{\boldsymbol{\theta}_1} \mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2'') \\ \leq& \frac{1}{2\mu_1}\|\nabla_{\boldsymbol{\theta}_1}\mathcal{L}(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2'')\|^2 \\ =& \frac{1}{2\mu_1}\|\nabla_{\boldsymbol{\theta}_1}\mathcal{L}(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2') - \nabla_{\boldsymbol{\theta}_1}\mathcal{L}(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2'')\|^2 \\ \leq& \frac{L_t^2}{2\mu_1}\|\boldsymbol{\theta}_2' - \boldsymbol{\theta}_2''\|^2.\end{aligned}$$

Lemma 1 demonstrates that there exists $\boldsymbol{\theta}_1'' = \arg\min_{\boldsymbol{\theta}_1 \in A(\boldsymbol{\theta}_2'')} \left\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_1'\right\|^2 \in A(\boldsymbol{\theta}_2'')$, such that

$$\begin{aligned}2\mu_1\|\boldsymbol{\theta}_1' - \boldsymbol{\theta}_1''\|^2 &\leq \mathcal{L}(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2'') - \mathcal{L}(\boldsymbol{\theta}_1'', \boldsymbol{\theta}_2'') \\ &\leq \frac{1}{2\mu_1}\|\nabla_{\boldsymbol{\theta}_1}\mathcal{L}(\boldsymbol{\theta}_1', \boldsymbol{\theta}_2'')\|^2,\end{aligned} \tag{10}$$

where the former inequality holds for Γ-QG condition (with parameter $\Gamma = 4\mu_1$) and the PL condition supports the latter inequality [9]. We have

$$\|\boldsymbol{\theta}_1' - \boldsymbol{\theta}_1''\| \leq \frac{L_t}{2\mu_1}\|\boldsymbol{\theta}_2' - \boldsymbol{\theta}_2''\|, \tag{11}$$

which completes the proof of Lemma 2. □

The following Proposition 1 further demonstrates the Lipschitz smoothness of $g(b) = \min_a \mathcal{L}(a, b)$.

Proposition 1. *Under Assumptions 1 and 4, there holds*

$$\nabla_{\boldsymbol{\theta}_2} g(\boldsymbol{\theta}_2) = \nabla_{\boldsymbol{\theta}_2}\mathcal{L}(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2), \tag{12}$$

where

$$\boldsymbol{\theta}_1^* \in \arg\min_{\boldsymbol{\theta}_1 \in \Theta_1} \mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2). \tag{13}$$

Moreover, g is L_g-Lipschitz smooth with $L_g = L_t + \frac{L_t^2}{2\mu_1}$.

Proof. Let $\boldsymbol{\theta}_1^* \in \arg\min_{\boldsymbol{\theta}_1 \in \Theta_1} \mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. By Lemma 2, for any scalar τ and direction $\boldsymbol{d}$, there exists

$$\boldsymbol{\theta}_1^*(\tau) \in \arg\min_{\boldsymbol{\theta}_1 \in \Theta_1} \mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 + \tau\boldsymbol{d}), \tag{14}$$

such that

$$\|\boldsymbol{\theta}_1^*(\tau) - \boldsymbol{\theta}_1^*\| \leq \frac{L_t}{2\mu_1}\tau\|\boldsymbol{d}\|. \tag{15}$$

To find the directional derivative of $g(\cdot)$, we compute the Taylor series expansion of $\mathcal{L}(\cdot)$ as follows

$$\begin{aligned}&g(\boldsymbol{\theta}_2 + \tau\boldsymbol{d}) - g(\boldsymbol{\theta}_2) \\ =& \mathcal{L}\left(\boldsymbol{\theta}_1^*(\tau), \boldsymbol{\theta}_2 + \tau\boldsymbol{d}\right) - \mathcal{L}\left(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2\right) \\ =& \nabla_{\boldsymbol{\theta}_1}\mathcal{L}(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2)(\boldsymbol{\theta}_1^*(\tau) - \boldsymbol{\theta}_1^*) + \tau\nabla_{\boldsymbol{\theta}_2}\mathcal{L}(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2)^T\boldsymbol{d} + \mathcal{O}\left(\tau^2\right) \\ =& \tau\nabla_{\boldsymbol{\theta}_2}\mathcal{L}(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2)^T\boldsymbol{d} + \mathcal{O}\left(\tau^2\right).\end{aligned}$$

Based on the directional derivative of $g(\cdot)$, for any $\boldsymbol{d}$, there holds

$$\begin{aligned} g'(\boldsymbol{\theta}_2;\boldsymbol{d}) &= \lim_{\tau\to 0^+}\frac{g(\boldsymbol{\theta}_2+\tau\boldsymbol{d})-g(\boldsymbol{\theta}_2)}{\tau} \\ &= \nabla_{\boldsymbol{\theta}_2}\mathcal{L}\left(\boldsymbol{\theta}_1^*,\boldsymbol{\theta}_2\right)^T\boldsymbol{d}. \end{aligned} \tag{16}$$

Moreover, for any $\boldsymbol{\theta}_1^* \in \arg\max_{\boldsymbol{\theta}_1\in\Theta_1}\mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = A(\boldsymbol{\theta}_2)$, we obtain

$$\nabla_{\boldsymbol{\theta}_2} g(\boldsymbol{\theta}_2) = \nabla_{\boldsymbol{\theta}_2}\mathcal{L}\left(\boldsymbol{\theta}_1^*,\boldsymbol{\theta}_2\right). \tag{17}$$

Let $A(\boldsymbol{\theta}_2) = \arg\min_{\boldsymbol{\theta}_1}\mathcal{L}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$. Denote the parameters $\boldsymbol{\theta}_1' \in A\left(\boldsymbol{\theta}_2'\right)$ and $\boldsymbol{\theta}_1'' = \arg\min_{\boldsymbol{\theta}_1\in A(\boldsymbol{\theta}_2'')}\left\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_1'\right\|^2 \in A\left(\boldsymbol{\theta}_2''\right)$. The auxiliary function g is also Lipschitz smooth since

$$\begin{aligned} &\|\nabla_{\boldsymbol{\theta}_2} g(\boldsymbol{\theta}_2') - \nabla_{\boldsymbol{\theta}_2} g(\boldsymbol{\theta}_2'')\| \\ =&\|\nabla_{\boldsymbol{\theta}_2}\mathcal{L}(\boldsymbol{\theta}_1',\boldsymbol{\theta}_2') - \nabla_{\boldsymbol{\theta}_2}\mathcal{L}(\boldsymbol{\theta}_1'',\boldsymbol{\theta}_2'')\| \\ =&\|\nabla_{\boldsymbol{\theta}_2}\mathcal{L}(\boldsymbol{\theta}_1',\boldsymbol{\theta}_2') - \nabla_{\boldsymbol{\theta}_2}\mathcal{L}(\boldsymbol{\theta}_1',\boldsymbol{\theta}_2'')\| \\ &+\|\nabla_{\boldsymbol{\theta}_2}\mathcal{L}(\boldsymbol{\theta}_1',\boldsymbol{\theta}_2'') + \nabla_{\boldsymbol{\theta}_2}\mathcal{L}(\boldsymbol{\theta}_1'',\boldsymbol{\theta}_2'')\| \\ \le& L_t(\|\boldsymbol{\theta}_2'-\boldsymbol{\theta}_2''\| + \|\boldsymbol{\theta}_1'-\boldsymbol{\theta}_1''\|) \\ \le&(L_t+\frac{L_t^2}{2\mu_1})\|\boldsymbol{\theta}_2'-\boldsymbol{\theta}_2''\|, \end{aligned} \tag{18}$$

where the last inequality is obtained by Lemma 2.

The proof for Proposition 1 is completed. □

Lemma 3. *Consider* $\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_1) = \mathbb{E}[F(\boldsymbol{\theta}_1;\Xi)]$, *where* f *is* L_t*-smooth and satisfies* μ_1*-PL condition. With step size* $\gamma_1 \le \min\{1/L_t, 1/\mu_1\}$, *the stochastic gradient descent estimator* $E_{\boldsymbol{\theta}_1}$ *at* k*-th iteration can be formulated by*

$$\boldsymbol{\theta}_1^{k+1} = \boldsymbol{\theta}_1^k - \gamma_1 E_{\boldsymbol{\theta}_1}\left(\boldsymbol{\theta}_1^k,\Xi_k\right), \tag{19}$$

where there hold $\mathbb{E}[E_{\boldsymbol{\theta}_1}(\boldsymbol{\theta}_1,\Xi) - \nabla_{\boldsymbol{\theta}_1} f(\boldsymbol{\theta}_1)] = 0$ *and* $\mathbb{E}\left[\|E_{\boldsymbol{\theta}_1}(\boldsymbol{\theta}_1,\Xi) - \nabla_{\boldsymbol{\theta}_1} f(\boldsymbol{\theta}_1)\|^2\right] \le \sigma^2$.

Then it follows that

$$\mathbb{E}\left[f(\boldsymbol{\theta}_1^{k+1}) - f^*\right] \le (1-\gamma_1\mu_1)\mathbb{E}\left[f(\boldsymbol{\theta}_1^k) - f^*\right] + \frac{L_t\gamma_1^2}{2}\sigma^2.$$

Proof. Based on the L_t smoothness of f and $L_t\gamma_1 \le 1$, it follows that

$$\begin{aligned} &f(\boldsymbol{\theta}_1^{k+1}) - f^* \\ \le& f(\boldsymbol{\theta}_1^k) - f^* + \left\langle\nabla_{\boldsymbol{\theta}_1} f(\boldsymbol{\theta}_1^k), \boldsymbol{\theta}_1^{k+1}-\boldsymbol{\theta}_1^k\right\rangle + \frac{L_t}{2}\|\boldsymbol{\theta}_1^{k+1}-\boldsymbol{\theta}_1^k\|^2 \\ =& f(\boldsymbol{\theta}_1^k) - f^* - \gamma_1\left\langle\nabla_{\boldsymbol{\theta}_1} f(\boldsymbol{\theta}_1^k), E_{\boldsymbol{\theta}_1}(\boldsymbol{\theta}_1^k,\Xi_t)\right\rangle + \frac{L_t\gamma_1^2}{2}\|E_{\boldsymbol{\theta}_1}(\boldsymbol{\theta}_1^k,\Xi_t)\|^2 \\ \le& f(\boldsymbol{\theta}_1^k) - f^* - \frac{\gamma_1}{2}\|\nabla_{\boldsymbol{\theta}_1} f(\boldsymbol{\theta}_1^k)\|^2 + \frac{L_t\gamma_1^2}{2}\|E_{\boldsymbol{\theta}_1}(\boldsymbol{\theta}_1^k,\Xi_t) - \nabla_{\boldsymbol{\theta}_1} f(\boldsymbol{\theta}_1^k)\|^2. \end{aligned}$$

Based on the PL condition with μ_1, the following holds

$$f(\boldsymbol{\theta}_1) - f^* = f(\boldsymbol{\theta}_1) - \min_{\boldsymbol{\theta}_1} f(\boldsymbol{\theta}_1) \le \frac{1}{2\mu_1}\|\nabla_{\boldsymbol{\theta}_1} f(\boldsymbol{\theta}_1)\|^2,$$

and

$$\frac{-\gamma_1}{2}\|\nabla_{\boldsymbol{\theta}_1} f(\boldsymbol{\theta}_1^k)\|^2 \le -\gamma_1\mu_1(f(\boldsymbol{\theta}_1^k) - f^*).$$

By taking expectations on both sides, it follows that

$$\mathbb{E}\left[f(\boldsymbol{\theta}_1^{k+1}) - f^*\right] \le (1-\gamma_1\mu_1)\mathbb{E}\left[f(\boldsymbol{\theta}_1^k) - f^*\right] + \frac{L_t\gamma_1^2}{2}\sigma^2.$$

□

Lemma 4. *Consider* $\max_{\boldsymbol{\theta}_2} g(\boldsymbol{\theta}_2) = \mathbb{E}[G(\boldsymbol{\theta}_2;\Xi)]$, *where* g *is* L_g*-smooth. Let Assumption 4 hold with parameters* μ_1 *and* μ_2, *and the step sizes satisfy* $\gamma_2 \le \min\{1/L_g, 1/\mu_2\}$. *Suppose the stochastic gradient ascent estimator* $E_{\boldsymbol{\theta}_2}$ *can be formulated by*

$$\boldsymbol{\theta}_2^{k+1} = \boldsymbol{\theta}_2^k + \gamma_2 E_{\boldsymbol{\theta}_2}\left(\boldsymbol{\theta}_2^k,\Xi_k\right), \tag{20}$$

where there hold $\mathbb{E}[E_{\boldsymbol{\theta}_2}(\boldsymbol{\theta}_2,\Xi) - \nabla_{\boldsymbol{\theta}_2} g(\boldsymbol{\theta}_2)] = 0$ *and* $\mathbb{E}\left[\|E_{\boldsymbol{\theta}_2}(\boldsymbol{\theta}_2,\Xi) - \nabla_{\boldsymbol{\theta}_2} g(\boldsymbol{\theta}_2)\|^2\right] \le \sigma^2$.

Then it follows that

$$\mathbb{E}\left[g^* - g(\boldsymbol{\theta}_2^{k+1})\right] \le (1-\gamma_2\mu_2)\mathbb{E}\left[g^* - f(\boldsymbol{\theta}_2^k)\right] + \frac{L_g\gamma_2^2}{2}\sigma^2.$$

Proof. Based on the L_t smoothness of function g and $-g$, it follows that

$$\begin{aligned} &g^* - g(\boldsymbol{\theta}_2^{k+1}) \\ \le& g^* - g(\boldsymbol{\theta}_2^k) + \left\langle-\nabla_{\boldsymbol{\theta}_2} g(\boldsymbol{\theta}_2^k), \boldsymbol{\theta}_2^{k+1}-\boldsymbol{\theta}_2^k\right\rangle + \frac{L_g}{2}\|\boldsymbol{\theta}_2^{k+1}-\boldsymbol{\theta}_2^k\|^2 \\ \le& g^* - g(\boldsymbol{\theta}_2^k) - \gamma_2\left\langle\nabla_{\boldsymbol{\theta}_2} g(\boldsymbol{\theta}_2^k), E_{\boldsymbol{\theta}_2}(\boldsymbol{\theta}_2^k,\Xi_t)\right\rangle + \frac{L_g\gamma_2^2}{2}\|E_{\boldsymbol{\theta}_2}(\boldsymbol{\theta}_2^k,\Xi_t)\|^2 \\ \le& g^* - g(\boldsymbol{\theta}_2^k) - \frac{L_g\gamma_2^2}{2}\|\nabla_{\boldsymbol{\theta}_2} g(\boldsymbol{\theta}_2^k)\|^2 + \frac{\gamma_2}{2}\|E_{\boldsymbol{\theta}_2}(\boldsymbol{\theta}_2^k,\Xi_t) - \nabla_{\boldsymbol{\theta}_2} g(\boldsymbol{\theta}_2^k)\|^2. \end{aligned}$$

Based on the two-sided PL condition (see Assumption 4) for maximum problem with μ_2, it follows that

$$g^* - g(\boldsymbol{\theta}_2^k) = \max_{\boldsymbol{\theta}_2} g(\boldsymbol{\theta}_2) - g(\boldsymbol{\theta}_2^k) \le \frac{1}{2\mu_2}\|\nabla_{\boldsymbol{\theta}_2} g(\boldsymbol{\theta}_2)\|^2,$$

and

$$\frac{-\gamma_2}{2}\|\nabla_{\boldsymbol{\theta}_2} g(\boldsymbol{\theta}_2^k)\|^2 \le -\gamma_2\mu_2(g^* - g(\boldsymbol{\theta}_2^k)).$$

Combining the above inequalities and taking expectations, it follows that

$$\mathbb{E}\left[g^* - g(\boldsymbol{\theta}_2^{k+1})\right] \le (1-\gamma_2\mu_2)\mathbb{E}\left[g^* - f(\boldsymbol{\theta}_2^k)\right] + \frac{L_g\gamma_2^2}{2}\sigma^2.$$

□

The following estimation follows by combining Lemmas 3 and 4.

Lemma 5. *Assume that Assumptions 4 hold with two-side PL condition parameters μ_1 and μ_2 w.r.t. $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ respectively. Define $a_k = \mathbb{E}[\mathcal{L}(\boldsymbol{\theta}_1^k, \boldsymbol{\theta}_2^k) - g(\boldsymbol{\theta}_2^k)]$ and $b_k = \mathbb{E}[g^* - g(\boldsymbol{\theta}_2^k)]$. After one iteration of Algorithm 1 with stepsizes $\gamma_1^k \le \frac{1}{L_t}$ and $\gamma_2^k \le \frac{1}{L_g}$ (L_g is specified in Proposition 1), then it follows that*

$$a_{k+1} + b_{k+1} \le \left(1 - \min\left\{\gamma_1^k \mu_1, \gamma_2^k \mu_2\right\}\right)(a_k + b_k) + \frac{L_t(\gamma_1^k)^2 + L_g(\gamma_2^k)^2}{2}\sigma^2.$$

4 Convergence Guarantee

There are some convergence results for minimax algorithms under nonconvex-nonconcave [13, 16], weakly or strongly convex-concave [6, 8] conditions recently. Compared to [6, 8, 13, 15], the following Theorem 1 concerning max-min problem (2) obtains a competitive convergence rate of $\mathcal{O}(1/K)$ under milder conditions on step size or convexity.

Theorem 1. *Let Assumptions 1-4 be true and denote $L' = \max\{L_t, L_g\}$, $\mu = \min\{\mu_1, \mu_2\}$. For the SAGDA in Algorithm 1, denote the step sizes by $\gamma_1^k = \gamma_2^k = \frac{c_0}{\Lambda+k} \le \min\{\frac{1}{L_t}, \frac{1}{L_g}\}$, where the constants satisfy $\Lambda > 0$ and $c_0 \ge 1/\mu$. Then, we have*

$$P_K \le \frac{\mathcal{V}}{\Lambda + K} = \max\left\{\Lambda P_0, \frac{L' c_0^2 \sigma^2}{c_0\mu - 1}\right\}(\Lambda + K)^{-1},$$

where the approximation measure P_K is defined in (5).

Remark 1. *The first term ΛP_0 is dependent on the initial settings, where a good initial point can be practically obtained by pretraining Algorithm 1. By setting $c_0 = 2/\mu$, the convergence rate reaches $\mathcal{O}(1/K)$.*

Proof. Here, we prove Theorem 1 by induction. When $k = 0$, the conclusion naturally holds. We assume that $P_k \le \frac{\mathcal{V}}{\Lambda+k}$ with positive constant $\Lambda > 0$. With the conclusions in Lemma 5, one can easily find that

$$P_{k+1} \le \left(1 - \min\left\{\gamma_1^k \mu_1, \gamma_2^k \mu_2\right\}\right) P_k + \frac{L_t(\gamma_1^k)^2 + L_g(\gamma_2^k)^2}{2}\sigma^2.$$

Denote $L' = \max\{L_t, L_g\}$, $\mu = \min\{\mu_1, \mu_2\}$. Let the step sizes satisfy $\gamma_1^k = \gamma_2^k = \frac{c_0}{\Lambda+k} \le \min\{\frac{1}{L_t}, \frac{1}{L_g}\}$, where the constants $c_0 \ge 1/\mu$. Then it follows that

$$\begin{aligned} P_{k+1} &\le \left(1 - \frac{c_0\mu}{\Lambda+k}\right) P_k + \left(\frac{c_0}{\Lambda+k}\right)^2 L'\sigma^2 \\ &\le \frac{\Lambda+k-c_0\mu}{\Lambda+k} \cdot \frac{\mathcal{V}}{\Lambda+k} + \left(\frac{c_0}{\Lambda+k}\right)^2 L'\sigma^2 \\ &\le \frac{\Lambda+k-1}{(\Lambda+k)^2}\mathcal{V} + \underbrace{\frac{c_0^2 L'\sigma^2}{(\Lambda+k)^2} - \frac{c_0\mu - 1}{(\Lambda+k)^2}\mathcal{V}}_{\le 0 \text{ with selected } \mathcal{V}}. \end{aligned} \tag{21}$$

Notice that with $(\Lambda + k + 1)(\Lambda + k - 1) \le (\Lambda + k)^2$, it follows that

$$\frac{\Lambda+k-1}{(\Lambda+k)^2} \le \frac{1}{\Lambda+k+1}.$$

That is

$$P_{k+1} \le \frac{\mathcal{V}}{\Lambda+k+1} \quad \text{and} \quad P_K \le \frac{\mathcal{V}}{\Lambda+K}, \tag{22}$$

where

$$\begin{aligned} \mathcal{V} &= \max\left\{\Lambda P_0, \frac{L' c_0^2 \sigma^2}{c_0\mu - 1}\right\} \\ &= \max\left\{\Lambda P_0, \frac{\max\{L_t, L_g\} c_0^2 \sigma^2}{c_0 \min\{\mu_1, \mu_2\} - 1}\right\}. \end{aligned} \tag{23}$$

Finally, the proof is completed. □

5 Example and Empirical Verification

All experiments are conducted in Python on a workstation with an Intel(R) Xeon(R) Platinum 8175M CPU and an NVIDIA RTX A6000 GPU.

5.1 Experimental Settings

To validate the theoretical findings on the convergence of nonconvex–nonconcave max–min optimization problems, the SAGDA algorithm is implemented on the GAN framework. The GAN optimization problem naturally fits the max–min formulation:

$$\max_{\boldsymbol{\theta}_2} \min_{\boldsymbol{\theta}_1} \mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2). \tag{24}$$

where θ_1 represents the generator parameters, θ_2 represents the discriminator parameters, and $\mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ is the objective function.

Dataset. The MNIST handwritten digit data[1], a standard benchmark containing 60,000 training images of 28×28 grayscale pixels representing digits 0–9, is

[1]Downloaded from http://yann.lecun.com/exdb/mnist/

used. The simplicity and widespread recognition of MNIST allows focusing on algorithmic convergence without interference from data complexity.

Model and Parameters. The Wasserstein GAN with gradient penalty (WGAN-GP[2]) [3] serves as the base model, facilitating satisfaction of the two-sided Polyak–Łojasiewicz (PL) condition assumed in the theoretical framework. The objective function is:

$$\begin{aligned} L(\theta_1, \theta_2) =& \mathbb{E}_{\tilde{x}\sim P_g}[D(\tilde{x})] - \mathbb{E}_{x\sim P_r}[D(x)] \\ &+ \lambda \mathbb{E}_{\hat{x}\sim P_{\hat{x}}}\left[\left(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1\right)^2\right], \end{aligned} \tag{25}$$

where P_g is the generator distribution, P_r is the real data distribution, $P_{\hat{x}}$ is the uniform sampling along straight lines between pairs of points from P_r and P_g, and $\lambda = 10$ is the gradient penalty coefficient.

Both the generator and discriminator employ multilayer perceptron architectures with three hidden layers (128, 256, and 512 units, respectively) and ReLU/LeakyReLU activations. According to Theorem 1, the learning rate is set as $\gamma_k^1 = \gamma_k^2 = \frac{c_0}{\Lambda+k}$ with $c_0 = 2/\mu$ and $\Lambda = 100$. A value of $\mu = 0.01$ is chosen based on preliminary experiments, resulting in an initial learning rate of 0.01. The batch size is set to 64, and training is conducted for 1000 iterations.

Evaluation Metrics. Convergence is measured using the metric $P_k = a_k + b_k$ in (5). In practice, these expectations are approximated through sampling. Additionally, the Fréchet Inception Distance (FID) [4] is used to evaluate the quality of generated samples.

5.2 Convergence Analysis

Figure 1 shows the convergence behavior of SAGDA in terms of P_k values over iterations. The observed convergence rate closely follows the theoretical $O(1/K)$ rate predicted by Theorem 1, confirming our analysis under the two-sided PL condition.

To verify the relationship between initial error and convergence, experiments were conducted using different initialization strategies. Figure 1 also compares convergence curves with random initialization versus warm-start initialization (using parameters pre-trained for 1,000 iterations). The warm-start approach demonstrates significantly faster convergence and even better convergence point (reaching obviously smaller errors), supporting our theoretical conclusion that convergence depends on the initial error P_0.

5.3 Effect of PL Constants

Table 1. Convergence Speed with Different PL Constants

μ_1	μ_2	Mean Iterations	Std Deviation
0.01	0.01	500	25
0.02	0.02	250	12
0.03	0.03	167	8
0.04	0.04	125	6
0.05	0.05	100	5
0.01	0.005	750	37
0.02	0.005	700	35
0.03	0.005	680	34
0.04	0.005	670	33
0.005	0.01	600	30
0.005	0.02	550	27
0.005	0.03	530	26
0.005	0.04	520	26
0.01	0.02	300	15
0.02	0.01	400	20
0.03	0.01	420	21
0.01	0.03	200	10
0.02	0.03	150	7
0.03	0.02	220	11
0.04	0.02	210	10
0.02	0.04	125	6
0.005	0.005	950	47
0.006	0.005	900	45
0.005	0.006	850	42
0.06	0.06	83	4
0.06	0.03	83	4
0.03	0.06	83	4

To further validate the theory, the impact of different PL constants on convergence speed was examined. Table 1 shows the number of iterations required to reach $P_k < 0.01$ for different μ_1 and μ_2 values. As predicted by our theory, larger PL constants lead to faster convergence, and the convergence rate is primarily determined by the smaller of the two constants, including $\mu = \min\{\mu_1, \mu_2\}$.

Our experimental results validate the theoretical convergence analysis of SAGDA for nonconvex-nonconcave max-min optimization problems. The observed $O(1/K)$ convergence rate matches our theoretical predictions, and the dependence on initial error P_0 confirms the importance of initialization. The relationship between PL constants and convergence speed further supports

[2]Downloaded from https://github.com/igul222/improved_wgan_training

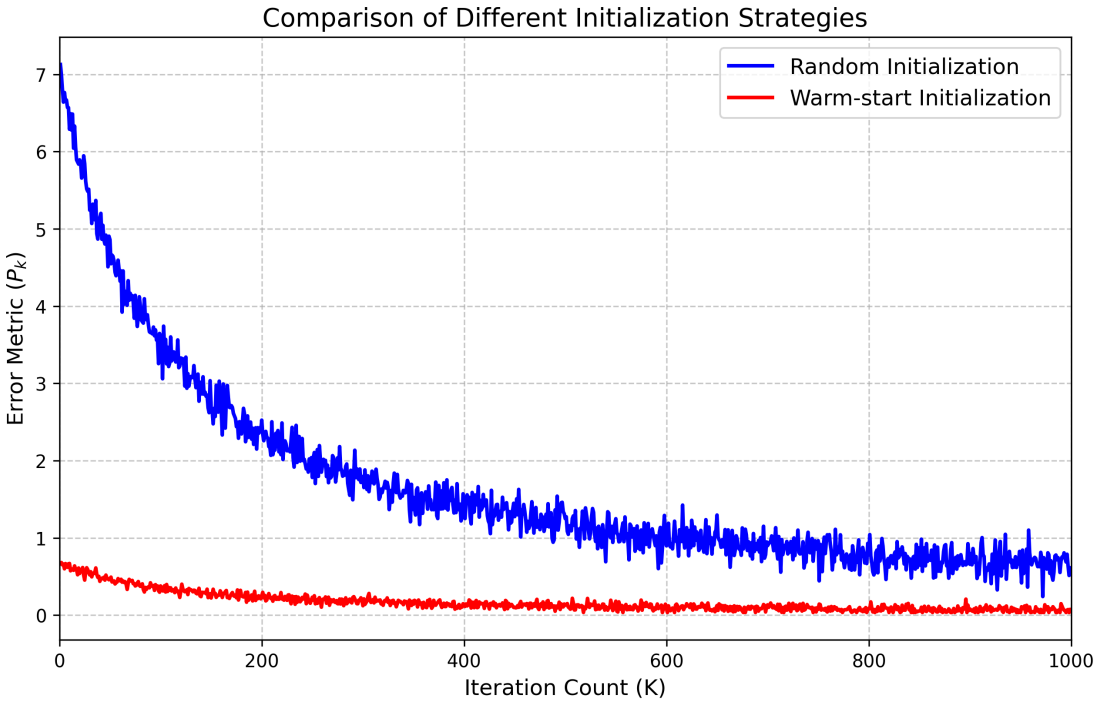

Figure 1. Convergence curves of the SAGDA algorithm, comparing random initialization versus warm-start.

our theoretical framework.

6 Discussion and Conclusion

This paper presents a novel convergence analysis for a class of nonconvex–nonconcave max–min optimization problems, a setting that remains largely underexplored due to its inherent computational hardness. By leveraging a carefully constructed auxiliary function along with two-sided Polyak–Łojasiewicz (PL) and Quadratic Growth (QG) conditions, it is established that the SAGDA algorithm achieves an $\mathcal{O}(1/K)$ convergence rate. Notably, this rate is comparable to those obtained in convex–concave settings. The result extends gradient-based optimization to a broader class of problems and is empirically validated with generative adversarial examples.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The author declares no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Donoho, D. L., & Johnstone, I. M. (1996). Neo-classical minimax problems, thresholding and adaptive function estimation. *Bernoulli*, *2*(1), 39-62. [CrossRef]

[2] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139-144. [CrossRef]

[3] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of wasserstein gans. *Advances in neural information processing systems*, *30*.

[4] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, *30*.

[5] Karimi, H., Nutini, J., & Schmidt, M. (2016, September). Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 795-811). Cham: Springer International Publishing. [CrossRef]

[6] Lin, T., Jin, C., & Jordan, M. I. (2020, July). Near-optimal algorithms for minimax optimization. In *Conference on learning theory* (pp. 2738-2779). PMLR.

[7] Lin, T., Jin, C., & Jordan, M. I. (2025). Two-timescale gradient descent ascent algorithms for nonconvex

minimax optimization. *Journal of Machine Learning Research, 26*(11), 1-45.

[8] Liu, M., Rafique, H., Lin, Q., & Yang, T. (2021). First-order convergence theory for weakly-convex-weakly-concave min-max problems. *Journal of Machine Learning Research, 22*(169), 1–34.

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

[10] Palaniappan, B., & Bach, F. (2016). Stochastic variance reduction methods for saddle-point problems. *Advances in Neural Information Processing Systems, 29*.

[11] Razaviyayn, M., Huang, T., Lu, S., Nouiehed, M., Sanjabi, M., & Hong, M. (2020). Nonconvex min-max optimization: Applications, challenges, and recent theoretical advances. *IEEE Signal Processing Magazine, 37*(5), 55-66. [CrossRef]

[12] Sion, M. (1958). On general minimax theorems. *Pacific Journal of Mathematics, 8*(1), 171–176. [CrossRef]

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

[14] Yang, S., Li, X., & Lan, G. (2025). Data-driven minimax optimization with expectation constraints. *Operations Research, 73*(3), 1345–1365. [CrossRef]

[15] Zhang, X., Chen, H., Gu, B., Gong, T., & Zheng, F. (2024). Fine-grained analysis of stability and generalization for stochastic bilevel optimization. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence* (pp. 5508–5516). [CrossRef]

[16] Jin, C., Netrapalli, P., & Jordan, M. (2020, November). What is local optimality in nonconvex-nonconcave minimax optimization?. In *International conference on machine learning* (pp. 4880-4889). PMLR.

[17] Barrett, D. G., & Dherin, B. (2020). Implicit gradient regularization. *arXiv preprint arXiv:2009.11162*.

Xuelin Zhang received the B.S. degree from China Agricultural University, Beijing, China, in 2020. He is currently pursuing a Ph.D. degree with the College of Informatics, Huazhong Agricultural University, Wuhan, China. His research interests include statistical learning theory, machine learning and applications. (Email: zhangxuelin@webmail.hzau.edu.cn)