A Single-Loop Accelerated Extra-Gradient Difference Algorithm with Improved Complexity Bounds for Constrained Minimax Optimization

Yuanyuan Liu[1], Fanhua Shang[2]*, Weixin An[1], Junhao Liu[1], Hongying Liu[3,6]*, Zhouchen Lin[4,5,6]

[1]Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education,
School of Artificial Intelligence, Xidian University, China
[2]School of Computer Science and Technology, College of Intelligence and Computing, Tianjin University
[3]Medical College, Tianjin University, China
[4]National Key Lab of General AI, School of Intelligence Science and Technology, Peking University
[5]Institute for Artificial Intelligence, Peking University
[6]Peng Cheng Laboratory
yyliu@xidian.edu.cn, fhshang@tju.edu.cn, hyliu2009@tju.edu.cn
zlin@pku.edu.cn

Abstract

In this paper, we propose a novel extra-gradient difference acceleration algorithm for solving constrained nonconvex-nonconcave (NC-NC) minimax problems. In particular, we design a new extra-gradient difference step to obtain an important quasi-cocoercivity property, which plays a key role to significantly improve the convergence rate in the constrained NC-NC setting without additional structural assumption. Then momentum acceleration is also introduced into our dual accelerating update step. Moreover, we prove that, to find an ϵ-stationary point of the function f, the proposed algorithm attains the complexity $\mathcal{O}(\epsilon^{-2})$ in the constrained NC-NC setting, while the best-known complexity bound is $\widetilde{\mathcal{O}}(\epsilon^{-4})$, where $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic factors compared to $\mathcal{O}(\cdot)$. As the special cases of the constrained NC-NC setting, the proposed algorithm can also obtain the same complexity $\mathcal{O}(\epsilon^{-2})$ for both the nonconvex-concave (NC-C) and convex-nonconcave (C-NC) cases, while the best-known complexity bounds are $\widetilde{\mathcal{O}}(\epsilon^{-2.5})$ for the NC-C case and $\widetilde{\mathcal{O}}(\epsilon^{-4})$ for the C-NC case. For fair comparison with existing algorithms, we also analyze the complexity bound to find ϵ-stationary point of the primal function ϕ for the constrained NC-C problem, which shows that our algorithm can improve the complexity bound from $\widetilde{\mathcal{O}}(\epsilon^{-3})$ to $\mathcal{O}(\epsilon^{-2})$. To the best of our knowledge, this is the first time that the proposed algorithm improves the best-known complexity bounds from $\widetilde{\mathcal{O}}(\epsilon^{-4})$ and $\widetilde{\mathcal{O}}(\epsilon^{-3})$ to $\mathcal{O}(\epsilon^{-2})$ in both the NC-NC and NC-C settings.

1 Introduction

This paper considers the following smooth minimax optimization problem,

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

[*] Corresponding authors

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

where $\mathcal{X} \subseteq \mathbb{R}^m$ and $\mathcal{Y} \subseteq \mathbb{R}^n$ are nonempty closed and convex feasible sets, and $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ is a smooth function. In recent years, this problem has drawn considerable interest from machine learning and other engineering communities such as generative adversarial networks [14], adversarial machine learning [15, 28], game theory [3], reinforcement learning [12, 41], empirical risk minimization [54, 39], and robust optimization [36, 13, 5]. While there is an extensive body of literature on minimax optimization, most prior works such as [54, 39, 42, 33, 50] focus on the convex-concave setting, where $f(x, y)$ is convex in x and concave in y. However, in many nonconvex minimax machine learning problems as in [35, 6], $f(x, y)$ is nonconvex in x and (strongly) concave in y, or $f(x, y)$ is (strongly) convex in x and nonconcave in y. For *nonconvex-strongly-concave* (NC-SC) minimax problems, a number of efficient algorithms such as [34, 25, 27] were proposed, and their complexity can be $\widetilde{\mathcal{O}}(\kappa_y^2 \epsilon^{-2})$ for achieving an ϵ-stationary point $\hat{x}$ (e.g., $\|\nabla \phi(\hat{x})\| \leq \epsilon$) of the primal function $\phi(\cdot) := \max_{y \in \mathcal{Y}} f(\cdot, y)$, where κ_y is the condition number for $f(x, \cdot)$, and $\widetilde{\mathcal{O}}$ hides logarithmic factors. More recently, [24] proposed an accelerated algorithm, which improves the gradient complexity to $\widetilde{\mathcal{O}}(\sqrt{\kappa_y} \epsilon^{-2})$, which exactly matches the lower complexity bound in [52, 23]. Therefore, we mainly consider the problem (1) in nonconvex-nonconcave (NC-NC), nonconvex-concave (NC-C) and convex-nonconcave (C-NC) settings.

1.1 Algorithms in NC-C and C-NC Settings

For *nonconvex-concave* (NC-C, but not strongly concave) and *convex-nonconcave* (C-NC) problems, there are two types of algorithms (i.e., multi-loop (including double-loop and triple-loop) and single-loop algorithms), and most of them are multi-loop algorithms such as [18, 31, 40, 24]. [31] proposed a multi-step framework that finds an ϵ-first order Nash equilibrium of $f(x, y)$ with the complexity $\widetilde{\mathcal{O}}(\epsilon^{-3.5})$. [40] designed a proximal dual implicit accelerated algorithm and proved that their algorithm finds an ϵ-stationary point of ϕ with the complexity $\widetilde{\mathcal{O}}(\epsilon^{-3})$. More recently, [24] proposed an accelerated algorithm, which achieves the complexity $\widetilde{\mathcal{O}}(\epsilon^{-2.5})$ to find an ϵ-stationary point of f. These multi-loop algorithms require at least $\mathcal{O}(\epsilon^{-2})$ outer iterations, and thus their complexities are more than $\mathcal{O}(\epsilon^{-2})$. Even though single-loop methods are more popular in practice due to their simplicity, few single-loop algorithms have been proposed for NC-C setting. The most natural approach is the gradient descent-ascent (GDA) method, which performs a gradient descent step on x and a gradient ascent step on y at each iteration. However, GDA fails to converge even for simple bilinear zero-sum games [22]. Subsequently, several improved GDA algorithms such as [8, 25, 27, 44] were proposed. For instance, [25] proved that the complexity of their two-time-scale GDA to find an ϵ-stationary point of ϕ is $\mathcal{O}(\epsilon^{-6})$ for NC-C problems. Moreover, [27] presented an efficient algorithm, which obtains an ϵ-stationary point of f with the complexity $\mathcal{O}(\epsilon^{-4})$. In fact, the complexity only counts the number of times the maximization subproblem is solved, and does not consider the complexity of solving this subproblem. [44] proposed a unified algorithm, and proved that the complexity of their algorithm to find an ϵ-stationary point of f is $\mathcal{O}(\epsilon^{-4})$ for both NC-C and C-NC problems. More recently, [51] presented a smoothed-GDA algorithm and proved that the complexity can be improved to $\mathcal{O}(\epsilon^{-2})$ for optimizing a special case of NC-C problems (i.e., minimizing the pointwise maximum of a finite collection of nonconvex functions). However, its complexity is still $\mathcal{O}(\epsilon^{-4})$ for both NC-C and C-NC problems. One natural question is: can we design a single-loop accelerated algorithm with the optimal complexity bound $\mathcal{O}(\epsilon^{-2})$ for *both NC-C and C-NC problems*?

1.2 Algorithms in NC-NC Settings

This paper mainly considers constrained NC-NC minimax problems, i.e., $f(x, y)$ is nonconvex in x and nonconcave in y in the constrained setting (called constrained NC-NC). In recent years, some works such as [7, 31, 46, 38, 20] focus on structured NC-NC problems. That is, the saddle gradient operator of such minimax problems or their objectives must satisfy one of the structured assumptions: the minty variational inequality (MVI) condition, the weak MVI condition, or negative comonotone condition and the Polyak-Łojasiewicz condition. However, structured NC-NC problems have limited practical applications because they are required to satisfy strong structural assumptions. For more practical constrained NC-NC problems or more general NC-NC problems, the convergence guarantee of the algorithms is still a challenge. In this paper, we also focus on convergence analysis

Table 1: Comparison of complexities of the minimax algorithms to find an ϵ-stationary point of $f(\cdot,\cdot)$ in the NC-C, C-NC and NC-NC settings. Note that Smoothed-GDA [51] can find an ϵ-stationary point for a special problem of (1) with $\mathcal{O}(\epsilon^{-2})$, and $\widetilde{\mathcal{O}}$ hides logarithmic factors compared to $\mathcal{O}(\cdot)$.

Optimality Criteria	References	NC-C	C-NC	NC-NC	Simplicity
Stationarity of f with smoothness and compact sets assumptions	Lu et al. [27]	$\widetilde{\mathcal{O}}(\epsilon^{-4})$	-	-	Single-Loop
	Nouiehed et al. [31]	$\widetilde{\mathcal{O}}(\epsilon^{-3.5})$	-	-	Multi-Loop
	Lin et al. [24]	$\widetilde{\mathcal{O}}(\epsilon^{-2.5})$	-	-	Multi-Loop
	Zhang et al. [51]	$\mathcal{O}(\epsilon^{-4})$	-	-	Single-Loop
	Xu et al. [44]	$\mathcal{O}(\epsilon^{-4})$	$\mathcal{O}(\epsilon^{-4})$	-	Single-Loop
	This work (Theorem 1)	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$	Single-Loop

Table 2: Comparison of complexities of existing minimax algorithms and the proposed algorithm to find an $\phi(\cdot) := \max_{y\in\mathcal{Y}} f(\cdot, y)$ in the nonconvex-concave (NC-C) setting. This table only highlights the dependence of ϵ, and compared with $\mathcal{O}(\cdot)$, $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic factors.

NC-C Settings	References	Compact set	Complexity	Simplicity
NC-C (Stationarity of ϕ)	Rafique et al. [34], Jin et al. [17]	$\mathcal{X}, \mathcal{Y}$	$\widetilde{\mathcal{O}}(\epsilon^{-6})$	Multi-Loop
	Lin et al. [25]	$\mathcal{X}, \mathcal{Y}$	$\widetilde{\mathcal{O}}(\epsilon^{-6})$	Single-Loop
	Thekumprampil et al. [40]	$\mathcal{X}, \mathcal{Y}$	$\widetilde{\mathcal{O}}(\epsilon^{-3})$	Multi-Loop
	Zhao [55], Lin et al. [24]	$\mathcal{X}, \mathcal{Y}$	$\widetilde{\mathcal{O}}(\epsilon^{-3})$	Multi-Loop
	This work (Theorem 2)	$\mathcal{Y}$	$\mathcal{O}(\epsilon^{-2})$	Single-Loop

for solving more practical constrained NC-NC problems. *Another natural question is: can we design a single-loop accelerated algorithm to further improve the bound in the constrained NC-NC setting?*

1.3 Motivations and Our Contributions

Motivations: For NC-C minimax problems, can we design a single-loop directly accelerated algorithm with the gradient complexity lower than the best-known result $\widetilde{\mathcal{O}}(\epsilon^{-2.5})$? Though Smoothed-GDA [51] can obtain the complexity $\mathcal{O}(\epsilon^{-2})$ for a special case of Problem (1), it only attains the complexity $\mathcal{O}(\epsilon^{-4})$ for NC-C minimax problems. For C-NC minimax problems, for any given x, to solve the nonconcave maximization subproblem with respect to y is NP-hard. As a result, all existing multi-loop algorithms will lose their theoretical guarantees as discussed in [44]. Can we propose a single-loop directly accelerated algorithm with the complexity lower than the best-known result $\mathcal{O}(\epsilon^{-4})$ for C-NC and NC-NC minimax problems?

Our Contributions: This paper proposes a novel single-loop extra-gradient difference acceleration algorithm to push towards optimal gradient complexities for constrained NC-NC minimax problems (1), and answer the above-mentioned problems.

We summarize the major contributions of this paper.

• We design a new single-loop accelerating algorithm for solving *constrained NC-NC* problems. In the proposed algorithm, we design a new extra-gradient difference scheme, and combine the gradient ascent and momentum acceleration steps for the dual variable update. In our algorithm, we present an important quasi-cocoercivity property. By leveraging the quasi-cocoercivity property, we can improve the complexity bound in our theoretical analysis.

• We analyze the convergence properties of the proposed algorithm for *constrained NC-NC* problems. Theorem 1 shows that to find an ϵ-stationary point of f, the proposed algorithm can obtain the gradient complexity $\mathcal{O}(\epsilon^{-2})$, which is the first time to attains the complexity bound in constrained NC-NC setting. The *constrained NC-C and C-NC* problems can be viewed as two special cases of the *constrained NC-NC* problem, and the proposed algorithm is also applicable to these two special problems. And its complexity is still $\mathcal{O}(\epsilon^{-2})$ for both NC-C and C-NC problems, which significantly improves the gradient complexity from $\mathcal{O}(\epsilon^{-4})$ of existing single-loop algorithms or $\widetilde{\mathcal{O}}(\epsilon^{-2.5})$ of existing multi-loop algorithms to $\mathcal{O}(\epsilon^{-2})$. The complexities of some recently proposed algorithms are listed in Table 1.

• In order to make a comprehensive comparison with existing algorithms, we also provide the theoretical analysis of our algorithm in terms of another convergence criteria (i.e., an ϵ-stationary point of ϕ) for constrained NC-C problems. The result shows that our algorithm improves the best-known result as in [40, 24, 55] from $\widetilde{\mathcal{O}}(\epsilon^{-3})$ to $\mathcal{O}(\epsilon^{-2})$, as shown in Table 2.

2 Preliminaries and Related Works

Notation: Throughout this paper, we use lower-case letters to denote vectors such as x, y, and calligraphic upper-case letters to denote sets such as $\mathcal{X}, \mathcal{Y}$. For a differentiable function f, $\nabla f(x)$ is the gradient of f at x. For a function $f(\cdot,\cdot)$ of two variables, $\nabla_x f(x,y)$ (or $\nabla_y f(x,y)$) is the partial gradient of f with respect to the first variable (or the second variable) at (x,y). For a vector x, $\|x\|$ denotes its ℓ_2-norm. We use $\mathcal{P}_{\mathcal{X}}$ and $\mathcal{P}_{\mathcal{Y}}$ to denote projections onto the sets $\mathcal{X}$ and $\mathcal{Y}$.

Assumption 1 (Smoothness). *$f(\cdot,\cdot)$ is continuously differentiable, and there exists a positive constant L such that*

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

holds for all $x_1, x_2 \in \mathbb{R}^m$, $y_1, y_2 \in \mathbb{R}^n$.

Definitions of the monotone operators: A operator $F(\cdot) : \mathbb{R}^n \to \mathbb{R}^n$ is monotone, if $[F(s) - F(t)]^T(s-t) \geq 0$, $\forall s, t \in \mathbb{R}^n$. If $[F(s) - F(t)]^T(s-t) \leq 0$, F is negative monotone.

A mapping $F(\cdot)$ is co-coercive if there is a positive constant α, such that

$$[F(s) - F(t)]^T(s-t) \geq \alpha\|F(s) - F(t)\|^2,\ \forall s, t \in \mathbb{R}^n.$$

Nonconvex-Concave Minimax Optimization: Due to the nonconvex nature of these minimax problems, finding the global solution is NP-hard in general. The recently proposed algorithms aim to find stationary solutions to such problems. For instance, the first-order Nash equilibrium condition (called game stationary) is used as an optimality condition in [31]. Besides game stationary, there are two main optimality criteria (i.e., an ϵ-stationary point of $f(\cdot,\cdot)$ or $\phi(\cdot) := \max_{y\in\mathcal{Y}} f(\cdot, y)$) for the convergence analysis of the algorithms such as [4, 44, 24].

For solving NC-C minimax problems, there exist a number of efficient multi-loop and single-loop algorithms such as [31, 25, 4, 44, 24]. Most of them are multi-loop algorithms, which either employ an accelerated update rule of x by adding regularization terms to its subproblem, or use multiple gradient ascent steps for the update of y to solve the subproblem exactly or inexactly. Compared with multi-loop algorithms, single-loop algorithms are easier to implement. One of the most popular single-loop algorithms is GDA. However, GDA with a constant stepsize can fail to converge even for a simple bilinear minimax problem [29].

To address this issue, only a few single-loop algorithms such as [25, 27, 44] were proposed, and most of them employed a smoothing or proximal point technique. For instance, Smoothed-GDA [51] introduces a smooth function $\varphi(x,y,z) = f(x,y) + \frac{a}{2}\|x-z\|^2$ for the update of the primal variable x, where z is an auxiliary variable and a is a constant, and its main update steps are

$$x_{t+1} = \mathcal{P}_{\mathcal{X}}(x_t - \eta_x \nabla_x \varphi(x_t, z_t, y_t)),\ y_{t+1} = \mathcal{P}_{\mathcal{Y}}(y_t + \eta_y \nabla_y f(x_{t+1}, y_t)),\ z_{t+1} = z_t + \beta(x_{t+1} - z_t),$$

where $\eta_x, \eta_y > 0$ are two stepsizes, and $0 < \beta \leq 1$. Smoothed-GDA can obtain the gradient complexity, $\mathcal{O}(\epsilon^{-4})$, for nonconvex-concave minimax problems.

Extra-Gradient Methods: There are some extra-gradient methods such as [48] for solving minimax problems. Let $\eta_0 > 0$ and u_t be the t-th iterate, the extra-gradient method [53] has the following projection-type prediction-correction step:

$$\text{Prediction} : u_{t+1/2} = \mathcal{P}_{\Omega}\left(u_t - \eta_t F(u_t)\right),\ \text{Correction} : u_{t+1} = \mathcal{P}_{\Omega}\left(u_t - \eta_t F(u_{t+1/2})\right).$$

In this paper, we call $u_{t+1/2}$ as the prediction point at the t-iteration.

3 Single-Loop Extra-Gradient Difference Acceleration Algorithm

In this section, we propose a single-loop Extra-Gradient Difference Acceleration algorithm (EGDA) for solving constrained NC-NC minimax problems.

Algorithm 1 EGDA for NC-NC minimax problems

```
Initialize: $x_0, y_0, u_0, u_{-1/2}, y_{-1}, u_{-1/2} \neq y_{-1}, \tau \geq 0.75, \eta_x, \delta, \beta$.
for $t = 0, 1, \ldots, T-1$ do
    $x_{t+1} = \mathcal{P}_{\mathcal{X}}[x_t - \eta_x \nabla_x f(x_t, y_t)]$;   $u_{t+1/2} = y_t + \beta[\nabla_y f(x_t, u_{t-1/2}) - \nabla_y f(x_t, y_{t-1})]$;
    $u_{t+1} = \mathcal{P}_{\mathcal{Y}}\left[y_t + \eta_y^t \nabla_y f(x_t, u_{t-1/2})\right]$;  $y_{t+1} = \tau y_t + (1-\tau) u_{t+1}$;
end for
Output: $(x_T, y_T)$.
```

3.1 Extra-Gradient Difference Acceleration

In recent years, many algorithms such as [27, 31, 32, 24, 51, 44] have been proposed for solving NC-C minimax problems. In essence, most of them are “indirect” acceleration algorithms, which are used to optimize the surrogate functions with a smoothing or proximal point term instead of the original function. However, this may hurt the performance of these algorithms both in theory and in practice [2, 1]. To address this issue, we propose a single-loop directly accelerating algorithm to find an ϵ-stationary points of f and ϕ with a significantly improved complexity $\mathcal{O}(\epsilon^{-2})$. The main update steps of our EGDA algorithm are designed as follows.

- Gradient descent:

$$x_{t+1} = \arg\min_{x \in \mathcal{X}} \Big\{ \langle \nabla_x f(x_t, y_t),\, x \rangle + \|x - x_t\|^2/\eta_x \Big\}. \tag{2}$$

- Extra-gradient difference prediction:

$$u_{t+1/2} = y_t + \beta \Big[\nabla_y f(x_t, u_{t-1/2}) - \nabla_y f(x_t, y_{t-1}) \Big]. \tag{3}$$

- Gradient ascent correction:

$$u_{t+1} = \arg\max_{u \in \mathcal{Y}} \Big\{ \langle \nabla_y f(x_t, u_{t-1/2}),\, u \rangle - \|u - y_t\|^2/\eta_y^t \Big\}. \tag{4}$$

- Momentum acceleration:

$$y_{t+1} = \tau y_t + (1 - \tau) u_{t+1}. \tag{5}$$

Here, $0 < \beta < \frac{1}{L}$, η_x, η_y^t are two stepsizes, $u_{t+1/2}$ and u_{t+1} are two auxiliary variables, $u_{t+1/2}$ is a prediction point, and $0 < \tau \leq 1$ is a momentum parameter. Our EGDA algorithm is formally presented in Algorithm 1. Our EGDA algorithm first performs one proximal gradient descent step on the primal variable x, and then we design a new dual-accelerating scheme in Eqs. (3), (4) and (5) for the dual variable y.

3.2 Advantages of Our Algorithm and Comparison to Related Work

In our EGDA algorithm, we first design a new prediction point $u_{t+1/2}$ in Eq. (3). Compared with existing extra-gradient-type methods such as [7, 16, 20], one of the main differences is that the proposed prediction point $u_{t+1/2}$ in (3) is updated by the gradient difference (i.e., $\nabla_y f(x_t, u_{t-1/2}) - \nabla_y f(x_t, y_{t-1})$), while the prediction point in other extra-gradient-type algorithms is updated only by using the gradient information at the correction point u_t. Then the gradient at the new prediction point $u_{t+1/2}$ is used in the gradient ascent step (4). And the dual variable y is update by the momentum acceleration step in (5).

• **Prediction Point:** The monotonicity and co-coercivity properties of gradient operators play a crucial role for convergence analysis. However, these important properties do not always hold for nonconvex problems. Some researchers have made great progress in some special nonconvex settings, such as structured nonconvex and weakly convex settings, which require a weaker condition such as weakly monotone [26], pseudo-monotone [16], and MVI [7]. However, such conditions seriously limit the application scope of Problem (1).

To address this challenge, we design a new prediction point scheme in (3), which can help us obtain a useful quasi-cocoercivity property. As a result, it does not require any monotone or structural assumption. Specifically, we find that we only require a weaker property in our theoretical analysis, that is, the weaker co-coercivity is required at some special points $\{u_{t+1/2}, y_t\}$ ($\langle \nabla_y f(x_t, u_{t+1/2}) - \nabla_y f(x_t, y_t), u_{t+1/2} - y_t \rangle \geq \beta \|\nabla_y f(x_t, u_{t+1/2}) - \nabla_y f(x_t, y_t)\|^2$ with $\beta > 0$). Thus, we develop a decoupling idea to construct the prediction point $u_{t+1/2}$. That is, we use the gradients with respect to y at $u_{t-1/2}, y_{t-1}$ instead of those at the points $u_{t+1/2}, y_t$. We can obtain a property (called the quasi-co-coercivity) in Section 4.2, which plays a key role in the theoretical analysis for the proposed algorithm.

• **Gradient Difference:** We also briefly discuss the underlying intuition of the proposed gradient difference in (3). We find that the proposed update in (3) is similar to the forms in [43] (see Eqs. (12) and (14) in [43]). [43] proposed a first-order procedure (i.e., difference of gradients) to achieve the negative curvature of a Hessian matrix. Therefore, our algorithm has a similar procedure, which contains second-order information. Moreover, we use the difference of gradients in the gradient ascent procedure for the dual update. But our EGDA algorithm only requires the Lipschitz gradient assumption for minimax problems to find first-order stationary points, while [43] requires both the Lipschitz gradient and Lipschitz Hessian assumptions for solving second-order stationary points of nonconvex minimization problems.

• **Momentum Acceleration:** We design a dual-accelerating update rule in (4) for the dual variable y in our EGDA algorithm, which is different from standard momentum acceleration schemes as in [31, 40, 24]. That is, the accelerated rules of existing algorithms are for the primal variable x, while our accelerated rule is designed for the dual variable y.

Therefore, the proposed new dual-accelerating step (including the gradient different prediction step in (3), the gradient ascent correction in (4) and momentum acceleration in (5)) is a key accelerated technique for our EGDA algorithm, and can help to improve the complexity bound from $\mathcal{O}(\epsilon^{-4})$ to $\mathcal{O}(\epsilon^{-2})$. In particular, our EGDA algorithm performs both gradient descent and ascent steps to the original function f. In contrast, many existing algorithms such as [51, 47, 44, 18] optimize their surrogate functions with smoothing terms instead of the original function. In particular, their smoothing parameters need to tune by repeatedly executing the algorithms, which may make them impractical [1]. As in our theoretical guarantees below, the proposed single-loop algorithm is able to significantly improve the best-known gradient complexity, $\mathcal{O}(\epsilon^{-4})$, of existing single-loop algorithms such as [27, 51, 44] to $\mathcal{O}(\epsilon^{-2})$.

4 Convergence Guarantees

In this section, we provide the convergence guarantees of our EGDA algorithm (i.e., Algorithm 1) for solving constrained NC-NC and NC-C problems. We first present the definitions of the two optimality criteria (i.e., an ϵ-stationary point of f or ϕ). All the proofs of the lemmas, properties and theorems below are included in the Supplementary Material.

4.1 Optimality Criteria and Key Property

Since finding a global minimum of a nonconvex optimization problem is NP-hard, finding a global saddle point (or Nash equilibrium) of a NC-NC function f is intractable [30]. As in the literature in the *NC-NC* setting, we introduce the local surrogates (i.e., the stationary point of f) and in the *NC-C setting*, we introduce the local surrogates (i.e., the stationary point of f or ϕ), whose gradient mappings are equal to zero. Below we define the following two optimality measures (i.e., an ϵ-stationary points of f or ϕ) for our theoretical analysis.

Definition 1 (An ϵ-stationary point of f [27])**.** *A point $(\overline{x}, \overline{y}) \in \mathcal{X} \times \mathcal{Y}$ is an ϵ-stationary point of $f(\cdot,\cdot)$ if $\|\pi(\overline{x}, \overline{y})\| \leq \epsilon$, where η_x and η_y are two constants, and*

$$\pi(\overline{x}, \overline{y}) := \begin{bmatrix} (1/\eta_x)(\overline{x} - \mathcal{P}_{\mathcal{X}}(\overline{x} - \eta_x \nabla_x f(\overline{x}, \overline{y}))) \\ (1/\eta_y)(\overline{y} - \mathcal{P}_{\mathcal{Y}}(\overline{y} + \eta_y \nabla_y f(\overline{x}, \overline{y}))) \end{bmatrix}. \tag{6}$$

If $\epsilon = 0$, then $(\overline{x}, \overline{y})$ is a stationary point of f.

For the *NC-C* setting, we also present another convergence criterion used in [40, 24]. Let $\phi(x) := \max_{y \in \mathcal{Y}} f(x, y)$, $\hat{x}$ is called an ϵ-stationary point of a smooth primal function $\phi : \mathbb{R}^m \to \mathbb{R}$, if $\|\nabla \phi(\hat{x})\| \leq \epsilon$. However, the function ϕ is not necessarily differentiable for minimax problems. Following [9, 40, 24], we introduce the Moreau envelope of ϕ for the optimality criterion, especially when ϕ is a weakly convex function, i.e., ϕ is L-weakly convex if the function $\phi(\cdot) + \frac{L}{2}\|\cdot\|^2$ is convex. We refer readers to [9, 24] for the comparison of these two criteria.

Definition 2 (An ϵ-stationary point of L-weakly convex function ϕ)**.** *$\hat{x}$ is an ϵ-stationary point of an L-weakly convex function $\phi:\mathbb{R}^m\to\mathbb{R}$, if $\|\nabla\phi_{1/(2L)}(\hat{x})\|\le\epsilon$, where $\phi_{1/(2L)}$ is the Moreau envelope of ϕ and is defined as: $\phi_\rho(x):=\min_z \phi(z)+(1/2\rho)\|z-x\|^2$. If $\epsilon=0$, then $\hat{x}$ is a stationary point.*

Then we give the following important property for the analysis of the proposed algorithm. By leveraging the property, we can obtain the complexity bound of the proposed algorithm.

Property 1 (**Quasi-Cocoercivity**)**.** *Let $u_{t+1/2}$ be updated in Eq. (3), then*

$$\Big\langle\nabla_y f(x_t,u_{t-1/2})-\nabla_y f(x_t,y_{t-1}),\ u_{t+1/2}-y_t\Big\rangle=\beta\|\nabla_y f(x_t,u_{t-1/2})-\nabla_y f(x_t,y_{t-1})\|^2.$$

4.2 Core Lemma

Our theoretical results mainly rely on Lemma 1 below, which plays a key role in the proofs of Theorems 1 and 2. Let $\{(x_t,y_t,u_t,u_{t-1/2})\}$ be a sequence generated by Algorithm 1, and we define the potential function

$$G_t:=f(x_t,y_t)+9L\|u_t-y_{t-1}\|^2+8L\beta^2\|\nabla_y f(x_{t-1},u_{t-3/2})-\nabla_y f(x_{t-1},y_{t-2})\|^2.$$

Next we need to prove that our defined potential function can make sufficient decrease at each iteration, i.e., $G_t-G_{t+1}>0$ as in Lemma 1 below. To prove Lemma 1, we will provide and prove the following upper bounds.

Proposition 1 (Upper bound of primal updates)**.** *Suppose Assumption 1 holds. Let $\{(x_t,y_t,u_t)\}$ be generated by Algorithm 1 with η_y^t=$\min\{\frac{\beta\|\nabla_y f(x_t,u_{t-1/2})-\nabla_y f(x_t,y_{t-1})\|^2}{4\|\nabla_y f(x_t,u_{t-1/2})\|^2},\frac{1}{28L},\eta_x\}+\delta$ and $0<\delta\le\frac{1}{28L}$ is a constant.*

$$\begin{aligned}&f(x_{t+1},u_{t+1/2})-f(x_t,y_t)\\&\le-\left(\frac{1}{\eta_x}-\frac{L}{2}\right)\|x_{t+1}-x_t\|^2+L\|u_{t+1/2}-y_t\|^2+L\|y_t-y_{t-1}\|^2+\underbrace{\big\langle\nabla_y f(x_t,y_{t-1}),u_{t+1/2}-y_t\big\rangle}_{A_1}.\end{aligned}$$

Proposition 2 (Upper bound of dual updates)**.** *Suppose Assumption 1 holds. Let $\{(x_t,y_t,u_t)\}$ be a sequence generated by Algorithm 1, then*

$$\begin{aligned}&f(x_{t+1},y_{t+1})-f(x_{t+1},u_{t+1/2})\\&\le\underbrace{\tau\langle\nabla_y f(x_t,u_{t-1/2}),y_t-u_{t+1}\rangle}_{A_2}+\underbrace{\langle\nabla_y f(x_t,u_{t-1/2}),u_{t+1}-u_{t+1/2}\rangle}_{A_3}+a_t,\end{aligned}$$

where $a_t:=2L\|x_{t+1}-x_t\|^2+6L\beta^2\|\nabla_y f(x_t,u_{t-1/2})-\nabla_y f(x_t,y_{t-1})\|^2+8L\|u_t-y_{t-1}\|^2+8L\beta^2\|\nabla_y f(x_{t-1},u_{t-3/2})-\nabla_y f(x_{t-1},y_{t-2})\|^2+3L\|u_{t+1}-y_t\|^2$.

Using the optimal condition of Problem (4) and our quasi-cocoercivity in Property 1, we can further bound $A_1+A_2+A_3$. The proof sketch of Lemma 1 is listed as follows:

$$\begin{aligned}G_{t+1}-G_t&:=\underbrace{f(x_{t+1},y_{t+1})-f(x_{t+1},u_{t+1/2})}_{\text{Proposition 1}}+\underbrace{f(x_{t+1},u_{t+1/2})-f(x_t,y_t)}_{\text{Proposition 2}}+\text{other terms}\\&=\underbrace{A_1+A_2+A_3}_{\text{Quasi-Cocoercivity in Property 1}}+\text{other terms.}\end{aligned}$$

Combining the above results and the definition of G_t, we can get the descent estimate of the potential function in Lemma 1, which is a main lemma for Theorems 1 and 2 below.

Lemma 1 (Descent estimate of G)**.** *Suppose Assumption 1 holds. Let $\{(x_t,y_t,u_t)\}$ be a sequence generated by Algorithm 1, then*

$$\begin{aligned}&G_0-G_T=\sum_{t=0}^{T-1}(G_t-G_{t+1})\\&\ge\sum_{t=0}^{T-1}\left(L\|u_{t+1}-y_t\|^2+\frac{\beta-30L\beta^2}{2}\|\nabla_y f(x_t,u_{t-1/2})-\nabla_y f(x_t,y_{t-1})\|^2+\frac{\|x_{t+1}-x_t\|^2}{2\eta_x}\right).\end{aligned}$$

4.3 Convergence Results

In this subsection, by using the optimality measures in Definitions 1 and 2, we present the following theoretical results in Theorems 1 and 2 for the constrained NC-NC setting and NC-C setting, respectively.

Assumption 2. *$\mathcal{Y}$ is a closed, convex and compact set with a diameter $D_{\mathcal{Y}} > 0$, and $\mathcal{X}$ is a closed and convex set.*

Furthermore, using Lemma 1 and the optimality measure in Definition 1, we will study the relation between $\pi(x_t, u_t)$ and the difference (i.e., $G_t - G_{t+1}$) to obtain the gradient complexity in Theorem 1 by computing the number of iterations to achieve an ϵ-stationary point of f.

Theorem 1 (Stationarity of f in constrained NC-NC settings)**.** *Suppose Assumptions 1 and 2 hold. Let the two stepsizes $\eta_x \le \frac{1}{4L}$, $\eta_y^t{=}\min\{\frac{\beta\|\nabla_y f(x_t,u_{t-1/2}) - \nabla_y f(x_t,y_{t-1})\|^2}{4\|\nabla_y f(x_t,u_{t-1/2})\|^2}, \frac{1}{28L}, \eta_x\} + \delta$, $\beta \le \frac{1}{60L}$, $\tau \ge 3/4$ and $\delta \le \frac{1}{28L}$, then the complexity of Algorithm 1 to find an ϵ-stationary point of f is bounded by*

$$\mathcal{O}\Big(\frac{G_0 - \underline{G} + 2LD_{\mathcal{Y}}^2}{\epsilon^2}\Big),$$

where $G_0 := G(x_0, y_0)$, and $\underline{G} := \min_{x\in\mathcal{X}} \phi(x)$.

Remark 1. *For constrained NC-NC problems, the gradient complexity of our EGDA algorithm to find an ϵ-stationary point of f is $\mathcal{O}(\epsilon^{-2})$. That is, our EGDA algorithm is first to obtain the gradient complexity in constrained NC-NC setting. In addition, our method can achieve the same complexity, $\mathcal{O}(\epsilon^{-2})$, as the algorithm with an additional structured assumption for NC-NC problems. However, different from existing algorithms, our algorithm is more practical. That is, it only requires that $\mathcal{Y}$ is a compact set, while existing algorithms need some stronger structured assumptions in the structured NC-NC setting. The detailed comparison is shown in Table 1. As two special cases of the constrained NC-NC problem, this theoretical result in Theorem 1 can be extended to the constrained NC-C and C-NC settings. For NC-C problems, the gradient complexity of our EGDA algorithm to find an ϵ-stationary point of f is $\mathcal{O}(\epsilon^{-2})$, while the best-known result of single-loop algorithms such as [27, 51, 44] is $\mathcal{O}(\epsilon^{-4})$, and the best-known result of multi-loop algorithms such as [24] is $\widetilde{\mathcal{O}}(\epsilon^{-2.5})$. That is, our EGDA algorithm improves the best-known gradient complexity from $\widetilde{\mathcal{O}}(\epsilon^{-2.5})$ to $\mathcal{O}(\epsilon^{-2})$. Smoothed-GDA [51] can also obtain the complexity $\mathcal{O}(\epsilon^{-2})$ for a special case of Problem (1) and $\mathcal{O}(\epsilon^{-4})$ for general NC-C minimax problems, while our algorithm attains the optimal result $\mathcal{O}(\epsilon^{-2})$ for all NC-C minimax problems. Existing algorithms such as HiBSA [27] and AGP [44] require the compactness of the domain $\mathcal{X}$ in the NC-C setting, while our EGDA algorithm does not, which significantly extends its applicability. For C-NC minimax problems, the proposed algorithm can obtain the complexity, $\mathcal{O}(\epsilon^{-2})$, while the best-known complexity as in [44] is $\mathcal{O}(\epsilon^{-4})$. In other words, the proposed algorithm can improve the best-known result from $\mathcal{O}(\epsilon^{-4})$ to $\mathcal{O}(\epsilon^{-2})$.*

By using the criterion in Definition 2 as the optimality measure, we use the definition of $\phi_{1/2L}$ and introduce the property of $\nabla\phi_{1/2L}$ as in [24]. With a similar setting for varying stepsizes as in [24], we study the relation between $\nabla\phi_{1/2L}$ and the basic descent estimation of the potential function in Lemma 1 by using the property of $\nabla\phi_{1/2L}$, and compute the number of iterations to achieve an ϵ-stationary point of ϕ. We provide the following theoretical result and its detailed proof in the Supplementary Material.

Theorem 2 (Stationarity of ϕ for constrained NC-C settings)**.** *Using the same notation as in Theorem 1, and f is concave with respect to y. Let $\{(x_t, y_t, u_t)\}$ be a sequence generated by Algorithm 1 with the stepsizes $\eta_y^t{=}\min\{\frac{\beta\|\nabla_y f(x_t,u_{t-1/2}) - \nabla_y f(x_t,y_{t-1})\|^2}{4\|\nabla_y f(x_t,u_{t-1/2})\|^2}, \frac{1}{28L}, \eta_x\} + \delta$, $\beta \le \frac{1}{60L}$, $\tau \ge 3/4$ and $\delta \le \frac{1}{28L}$. Then the gradient complexity of Algorithm 1 to find an ϵ-stationary point of ϕ is bounded by*

$$\mathcal{O}\Big(\frac{D_{\mathcal{Y}}(G_0 - \underline{G} + 2LD_{\mathcal{Y}}^2)}{\epsilon^2}\Big).$$

Remark 2. *From Theorem 2, it is clear that the gradient complexity of the proposed algorithm is $\mathcal{O}(\epsilon^{-2})$. That is, Algorithm 1 can improve the best-known gradient complexity from $\widetilde{\mathcal{O}}(\epsilon^{-3})$ as in [24] to $\mathcal{O}(\epsilon^{-2})$. Therefore, Algorithm 1 is the first algorithm, which attains the gradient complexity $\mathcal{O}(\epsilon^{-2})$ to find an ϵ-stationary point of ϕ for NC-C minimax problems.*

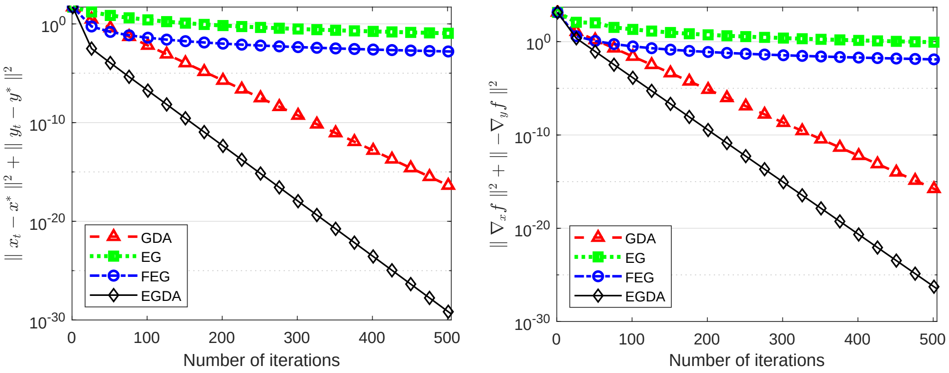

Figure 1: Comparison of all the methods for solving the NC-NC problem, $f(x, y) = x^2 + 3\sin^2 x \sin^2 y - 4y^2 - 10\sin^2 y$. Left: Convergence in terms of $\|x_t - x^*\|^2 + \|y_t - y^*\|^2$, where (x^*, y^*) is the global saddle point; Right: Convergence in terms of $\|\nabla_x f\|^2 + \|\nabla_y f\|^2$.

5 Numerical Results

We conduct many experiments to illustrate the performance of the proposed algorithm for solving NC-NC, NC-C and convex-concave problems.

NC-NC Problems. We conduct some experiments to illustrate the performance of the proposed algorithm, EGDA, for solving NC-NC problems. Moreover, we compare it against existing methods such as GDA [46], EG [10], EAG [49] and FEG [20].

We compare EGDA with GDA, EG and FEG for solving a simple NC-NC minimax problem (i.e., $f(x,y) = x^2 + 3\sin^2 x \sin^2 y - 4y^2 - 10\sin^2 y$), which satisfies the Polyak-Łojasiewicz condition, as shown in Fig. 1. All the results show that EGDA indeed converges to the global saddle point and is significantly faster than other methods including GDA and FEG in terms of both $\|x_t - x^*\|^2 + \|y_t - y^*\|^2$ and $\|\nabla_x f\|^2 + \|\nabla_y f\|^2$. Although FEG has a fast theoretical rate, $\mathcal{O}(1/t^2)$, with a negative monotone assumption, it converges much slower than EGDA in practice.

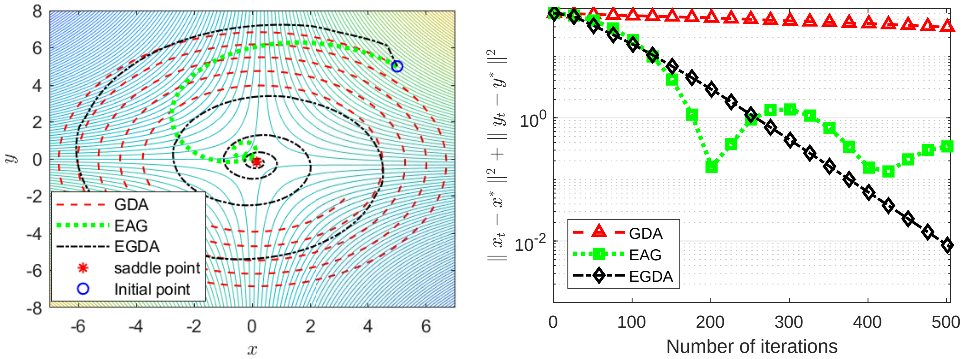

Figure 2: Comparison of all the methods for solving the convex-concave problem, $f(x, y) = \log(1 + e^x) + 3xy - \log(1 + e^y)$. Left: Trajectories of the three algorithms; Right: Convergence in terms of $\|x_t - x^*\|^2 + \|y_t - y^*\|^2$.

Convex-Concave Problems. Fig. 2 shows the convergence results of the methods including GDA, EAG [49] and EGDA for solving the convex-concave problem, $f(x, y) = \log(1 + e^x) + 3xy - \log(1 + e^y)$. It is clear that EGDA converges much faster than other methods such as EAG. We also observe empirically when the same step-size is used, even if small, GDA may not converge to stationary points, and it is proven to always have bounded iterates.

NC-C Problems. We also apply our EGDA algorithm to train robust neural networks against adversarial attacks on Fashion MNIST and MNIST, and verify our theoretical results. In [15, 28, 31],

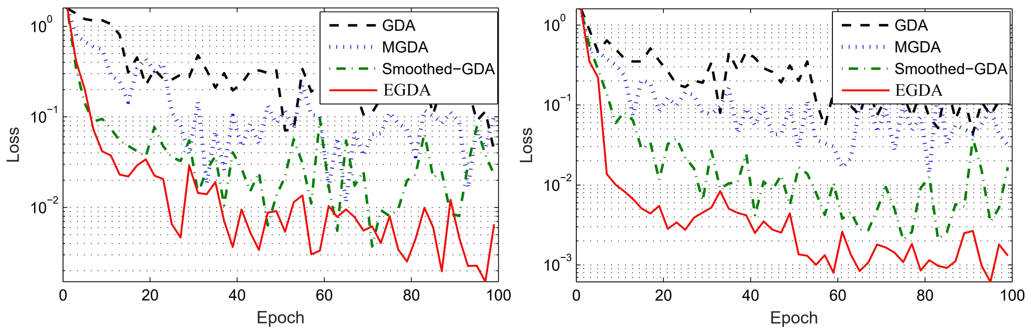

Figure 3: Convergence speed of all the algorithms on Fashion MNIST (left) and MNIST (right).

the robust training procedure can be formulated into a NC-C minimax optimization problem. It is clear that the minimax problem is nonconvex in x, but concave in y.

FGSM [15] and PGD [19] are two popular methods for generating adversarial examples. To obtain the targeted attack $\hat{x}_{ij}$, we use the same procedure as in [31, 51] as follows. The perturbation level ε is chosen from $\{0.0, 0.1, 0.2, 0.3, 0.4\}$, and the stepsize is 0.01. Note that the number of iterations is set to 40 for FGSM and 10 for PGD, respectively. We compare our EGDA method with GDA [25], MGDA [31] and Smoothed-GDA [51], and illustrate the convergence of all the algorithms on the loss function in Fig. 3. The results show that Smoothed-GDA and EGDA converge significantly faster than other algorithms. This verifies that they have a faster convergence rate, $\mathcal{O}(\epsilon^{-2})$. Moreover, EGDA converges much faster than Smoothed-GDA.

6 Conclusions and Future Work

In this paper, we proposed a new single-loop accelerated algorithm for solving various constrained NC-NC minimax problems. In the proposed algorithm, we designed a novel extra-gradient difference scheme for dual updates. Moreover, we provided the convergence guarantees for the proposed algorithm, and the theoretical results show that to find an ϵ-stationary point of f, the proposed algorithm obtains the complexity bound $\mathcal{O}(\epsilon^{-2})$ for constrained NC-NC problems. That is, this is the first time that the proposed algorithm attains the complexity bound $\mathcal{O}(\epsilon^{-2})$ in constrained NC-NC settings (including constrained NC-C and C-NC). For NC-C problems, we provided the theoretical guarantees of the proposed algorithm under the stationarity of ϕ, which show that our algorithm improves the complexity bound from $\widetilde{\mathcal{O}}(\epsilon^{-3})$ to $\mathcal{O}(\epsilon^{-2})$. Experimental results also verified the theoretical results of the proposed algorithm, which have the factors of ϵ^{-1} and ϵ^{-2} faster than existing algorithms, respectively. For further work, we will extend our directly accelerating algorithm to stochastic, non-smooth, nonconvex-nonconcave and federated learning settings as in [47, 11, 4, 55, 45, 21, 37].

Acknowledgments

We want to thank the anonymous reviewers for their valuable suggestions and comments. This work was supported by the National Key R&D Program of China (No. 2022ZD0160302), the National Natural Science Foundation of China (Nos. 62276182, 61976164, 62276004 and 61836009), the National Science Basic Research Plan in Shaanxi Province of China (No. 2022GY-061), the major key project of PCL, China (No. PCL2021A12), and Qualcomm.

References

[1] Z. Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. In *STOC*, 2017.

[2] Z. Allen-Zhu and Y. Yuan. Improved SVRG for non-strongly-convex or sum-of-non-convex objectives. In *ICML*, pages 1080–1089, 2016.

[3] J. P. Bailey, G. Gidel, and G. Piliouras. Finite regret and cycles with fixed step-size via alternating gradient descent-ascent. In *The Annual Conference on Learning Theory (COLT)*, pages 391–407, 2020.

[4] R. I. Bot and A. Bohm. Alternating proximal-gradient steps for (stochastic) nonconvex-concave minimax problems. *SIAM J. Optim.*, 2023.

[5] Y. Cai and W. Zheng. Accelerated single-call methods for constrained min-max optimization. In *International Conference on Learning Representations (ICLR)*, 2023.

[6] Z. Chen, Z. Hu, Q. Li, Z. Wang, and Y. Zhou. A cubic regularization approach for finding local minimax points in nonconvex minimax optimization. *Transactions on Machine Learning Research*, pages 1–43, 2023.

[7] C. D. Dang and G. Lan. On the convergence properties of non-euclidean extragradient methods for variational inequalities with generalized monotone operators. *Comput. Optim. Appl.*, 60(2):277–310, 2015.

[8] C. Daskalakis and I. Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In *The Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 9256–9266, 2018.

[9] D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM J. Optim.*, 29(1):207–239, 2019.

[10] J. Diakonikolas, C. Daskalakis, and M. I. Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *AISTATS*, pages 2746–2754, 2021.

[11] J. Diakonikolas, C. Daskalakis, and M. I. Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *The International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2746–2754, 2021.

[12] S. S. Du, J. Chen, L. Li, L. Xiao, and D. Zhou. Stochastic variance reduction methods for policy evaluation. In *International Conference on Machine Learning (ICML)*, pages 1049–1058, 2017.

[13] J. C. Duchi and H. Namkoong. Variance-based regularization with convex objectives. *J. Mach. Learn. Res.*, 20:1–55, 2019.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *The Annual Conference on Neural Information Processing Systems (NIPS)*, pages 2672–2680, 2014.

[15] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

[16] A. N. Iusem, A. Jofré, R. I. Oliveira, and P. Thompson. Extragradient method with variance reduction for stochastic variational inequalities. *SIAM J. Optim.*, 27(2):686–724, 2017.

[17] C. Jin, P. Netrapalli, and M. I. Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *International Conference on Machine Learning (ICML)*, pages 4880–4889, 2020.

[18] W. Kong and R. D. Monteiro. An accelerated inexact proximal point method for solving nonconvex-concave min-max problems. In *arXiv preprint arXiv:1905.13433v3*, 2021.

[19] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial machine learning at scale. In *International Conference on Learning Representations (ICLR)*, 2017.

[20] S. Lee and D. Kim. Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems. In *NeurIPS*, 2021.

[21] Y. Lei, Z. Yang, T. Yang, and Y. Ying. Stability and generalization of stochastic gradient methods for minimax problems. In *International Conference on Machine Learning (ICML)*, pages 6175–6186, 2021.

[22] A. Letcher, D. Balduzzi, S. Racaniere, J. Martens, J. N. Foerster, K. Tuyls, and T. Graepel. Differentiable game mechanics. *J. Mach. Learn. Res.*, 20:1–40, 2019.

[23] H. Li, Y. Tian, J. Zhang, and A. Jadbabaie. Complexity lower bounds for nonconvex-strongly-concave min-max optimization. In *arXiv preprint arXiv:2104.08708v1*, 2021.

[24] T. Lin, C. Jin, and M. I. Jordan. Near-optimal algorithms for minimax optimization. In *The Annual Conference on Learning Theory (COLT)*, pages 2738–2779, 2020.

[25] T. Lin, C. Jin, and M. I. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning (ICML)*, pages 6083–6093, 2020.

[26] M. Liu, H. Rafique, Q. Lin, and T. Yang. First-order convergence theory for weakly-convex-weakly-concave min-max problems. *J. Mach. Learn. Res.*, 22:169:1–169:34, 2021.

[27] S. Lu, I. Tsaknakis, M. Hong, and Y. Chen. Hybrid block successive approximation for one-sided nonconvex min-max problems: algorithms and applications. *IEEE Trans. Signal Process.*, 68:3676–3691, 2020.

[28] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

[29] A. Mokhtari, A. Ozdaglar, and S. Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *The International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1497–1507, 2020.

[30] K. G. Murty and S. N. Kabadi. Some NP-complete problems in quadratic and nonlinear programming. *Mathematical Programming*, 39(2):117–129, 1987.

[31] 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 *NeurIPS*, pages 14905–14916, 2019.

[32] D. M. Ostrovskii, A. Lowy, and M. Razaviyayn. Efficient search of first-order Nash equilibria in nonconvex-concave smooth min-max problems. *SIAM J. Optim.*, 2021.

[33] Y. Ouyang and Y. Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185:1–35, 2021.

[34] H. Rafique, M. Liu, Q. Lin, and T. Yang. Weakly-convex concave min-max optimization: Provable algorithms and applications in machine learning. *Optim. Method Softw.*, 2022.

[35] M. Razaviyayn, T. Huang, S. Lu, M. Nouiehed, M. Sanjabi, and M. Hong. Nonconvex min-max optimization: Applications, challenges, and recent theoretical advances. *IEEE Signal Process. Mag.*, 37(5):55–66, 2020.

[36] S. Shafieezadeh-Abadeh, P. M. Esfahani, and D. Kuhn. Distributionally robust logistic regression. In *The Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1576–1584, 2015.

[37] P. Sharma, R. Panda, G. Joshi, and P. K. Varshney. Federated minimax optimization: Improved convergence analyses and algorithms. In *International Conference on Machine Learning (ICML)*, pages 19683–19730, 2022.

[38] C. Song, Z. Zhou, Y. Zhou, Y. Jiang, and Y. Ma. Optimistic dual extrapolation for coherent non-monotone variational inequalities. In *NeurIPS*, 2020.

[39] C. Tan, T. Zhang, S. Ma, and J. Liu. Stochastic primal-dual method for empirical risk minimization with $o(1)$ per-iteration complexity. In *The Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 8376–8385, 2018.

[40] K. K. Thekumprampil, P. Jain, P. Netrapalli, and S. Oh. Efficient algorithms for smooth minimax optimization. In *The Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 12659–12670, 2019.

[41] H.-T. Wai, Z. Yang, Z. Wang, and M. Hong. Multi-agent reinforcement learning via double averaging primal-dual optimization. In *The Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 9672–9683, 2018.

[42] G. Xie, L. Luo, Y. Lian, and Z. Zhang. Lower complexity bounds for finite-sum convex-concave minimax optimization problems. In *International Conference on Machine Learning (ICML)*, pages 10504–10513, 2020.

[43] Y. Xu, R. Jin, and T. Yang. First-order stochastic algorithms for escaping from saddle points in almost linear time. In *NeurIPS*, pages 5535–5545, 2018.

[44] Z. Xu, H. Zhang, Y. Xu, and G. Lan. A unified single-loop alternating gradient projection algorithm for nonconvex-concave and convex-nonconcave minimax problems. *Mathematical Programming*, 2023.

[45] J. Yang, N. Kiyavash, and N. He. Global convergence and variance-reduced optimization for a class of nonconvex-nonconcave minimax problems. In *The Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[46] J. Yang, N. Kiyavash, and N. He. Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. In *NeurIPS*, 2020.

[47] J. Yang, S. Zhang, N. Kiyavash, and N. He. A catalyst framework for minimax optimization. In *The Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[48] T. Yoon and E. K. Ryu. Accelerated algorithms for smooth convex-concave minimax problems with o(1/kˆ2) rate on squared gradient norm. In M. Meila and T. Zhang, editors, *ICML*, pages 12098–12109, 2021.

[49] T. Yoon and E. K. Ryu. Accelerated algorithms for smooth convex-concave minimax problems with $o(1/k^2)$ rate on squared gradient norm. In *ICML*, pages 12098–12109, 2021.

[50] J. Zhang, M. Hong, and S. Zhang. On lower iteration complexity bounds for the saddle point problems. *Mathematical Programming*, 194:901–935, 2022.

[51] J. Zhang, P. Xiao, R. Sun, and Z. Luo. A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems. In *The Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 7377–7389, 2020.

[52] S. Zhang, J. Yang, C. Guzman, N. Kiyavash, and N. He. The complexity of nonconvex-strongly-concave minimax optimization. In *The Conference on Uncertainty in Artificial Intelligence (UAI)*, 2021.

[53] W. Zhang and D. Han. A new alternating direction method for co-coercive variational inequality problems. *Comput. Math. Appl.*, 57(7):1168–1178, 2009.

[54] Y. Zhang and L. Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. *J. Mach. Learn. Res.*, 18:1–42, 2017.

[55] R. Zhao. A primal dual smoothing framework for max-structured nonconvex optimization. *Mathematics of Operations Research*, 2023.

Appendix for "A Single-Loop Accelerated Extra-Gradient Difference Algorithm with Improved Complexity Bounds for Constrained Minimax Optimization"

Appendix A: Basic Properties

Before giving our theoretical analysis, we first present the following basic properties.

Property 2. *Given any a, b, c, then we have*

$$\langle a-b,\ a-c\rangle=\frac{1}{2}\left(\|a-b\|^2+\|a-c\|^2-\|b-c\|^2\right).$$

Property 3. *(Theorem 2.1.5 in [42]). If $f:\mathbb{R}^d\to\mathbb{R}$ is L-smooth, then for all $x, y\in\mathbb{R}^d$,*

$$\left|f(y)-f(x)-\nabla f(x)^T(y-x)\right|\leq\frac{L}{2}\|y-x\|^2.$$

Property 4. *[Lemma A.1 in [25]] Let $\phi_{1/2L}(\widehat{x}):=\min_x\{\phi(x)+L\|x-\widehat{x}\|^2\}$ and $x^*(\widehat{x})=\arg\min_x\{\phi(x)+L\|x-\widehat{x}\|^2\}$. If $f:\mathbb{R}^d\to\mathbb{R}$ is L-smooth, and $\mathcal{Y}$ is bounded, then*

- *1. $\phi_{1/2L}$ is L-smooth with $\nabla\phi_{1/2L}(\widehat{x})=2L(\widehat{x}-x^*(\widehat{x}))$.*
- *2. $\phi_{1/2L}(x_1)-\phi_{1/2L}(x_2)-(x_1-x_2)^T\nabla\phi_{1/2L}(x_2)\leq\frac{L}{2}\|x_1-x_2\|^2$ for any $x_1, x_2\in\mathbb{R}^m$.*

Appendix B: Key Property

Property 1. *(**Quasi-Cocoercivity**) Let $u_{t+1/2}$ be updated in Eq. (3), then*

$$\left\langle\nabla_y f(x_t,u_{t-1/2})-\nabla_y f(x_t,y_{t-1}),u_{t+1/2}-y_t\right\rangle=\beta\|\nabla_y f(x_t,u_{t-1/2})-\nabla_y f(x_t,y_{t-1})\|^2. \tag{7}$$

Proof. By using the update rule $u_{t+1/2}$ in Eq. (3), that is, $u_{t+1/2}=y_t+\beta\left[\nabla_y f(x_t,u_{t-1/2})-\nabla_y f(x_t,y_{t-1})\right]$, we get

$$\begin{aligned}&\left\langle\nabla_y f(x_t,u_{t-1/2})-\nabla_y f(x_t,y_{t-1}),u_{t+1/2}-y_t\right\rangle\\=&\beta\left\langle\nabla_y f(x_t,u_{t-1/2})-\nabla_y f(x_t,y_{t-1}),\nabla_y f(x_t,u_{t-1/2})-\nabla_y f(x_t,y_{t-1})\right\rangle\\=&\beta\|\nabla_y f(x_t,u_{t-1/2})-\nabla_y f(x_t,y_{t-1})\|^2.\end{aligned}$$

This completes the proof. □

Appendix C: Proof of Lemma 1

Before proving Lemma 1, we first present and prove the following two propositions for the upper bounds of the primal variable and dual variable.

Proof of Proposition 1

Proof. Using the optimality condition for the gradient descent step in Eq. (2) with respect to the primal variable $x\in\mathcal{X}$, we have

$$\left\langle\nabla_x f(x_t,y_t)+\frac{1}{\eta_x}(x_{t+1}-x_t),\ x_t-x_{t+1}\right\rangle\geq 0. \tag{8}$$

That is,

$$\langle \nabla_x f(x_t, y_t), x_{t+1} - x_t \rangle \leq -\frac{1}{\eta_x} \|x_{t+1} - x_t\|^2. \tag{9}$$

Since f is L-smooth, then

$$\begin{aligned}
f(x_{t+1}, u_{t+1/2}) \leq & f(x_t, y_t) + \langle \nabla_x f(x_t, y_t),\ x_{t+1} - x_t \rangle + \frac{L}{2} \|x_{t+1} - x_t\|^2 \\
& + \left\langle \nabla_y f(x_t, y_t),\ u_{t+1/2} - y_t \right\rangle + \frac{L}{2} \|u_{t+1/2} - y_t\|^2 \\
= & f(x_t, y_t) + \langle \nabla_x f(x_t, y_t),\ x_{t+1} - x_t \rangle + \frac{L}{2} \|x_{t+1} - x_t\|^2 \\
& + \left\langle \nabla_y f(x_t, y_{t-1}),\ u_{t+1/2} - y_t \right\rangle + \frac{L}{2} \|u_{t+1/2} - y_t\|^2 \\
& + \left\langle \nabla_y f(x_t, y_t) - \nabla_y f(x_t, y_{t-1}),\ u_{t+1/2} - y_t \right\rangle \\
\leq & f(x_t, y_t) - \left(\frac{1}{\eta_x} - \frac{L}{2} \right) \|x_{t+1} - x_t\|^2 \\
& + \left\langle \nabla_y f(x_t, y_{t-1}),\ u_{t+1/2} - y_t \right\rangle \\
& + L \|u_{t+1/2} - y_t\|^2 + \frac{1}{2L} \|\nabla_y f(x_t, y_t) - \nabla_y f(x_t, y_{t-1})\|^2 \\
\leq & f(x_t, y_t) - \left(\frac{1}{\eta_x} - \frac{L}{2} \right) \|x_{t+1} - x_t\|^2 \\
& + \left\langle \nabla_y f(x_t, y_{t-1}),\ u_{t+1/2} - y_t \right\rangle + L \|u_{t+1/2} - y_t\|^2 + L \|y_t - y_{t-1}\|^2,
\end{aligned} \tag{10}$$

where the second inequality holds due to the optimal condition in Eq. (9), and the last inequality holds due to the smoothness of f, i.e., $\|\nabla_y f(x_t, y_t) - \nabla_y f(x_t, y_{t-1})\| \leq L \|y_t - y_{t-1}\|$.

With $\eta_x < \frac{1}{L}$, and combining (9) and (10), we have

$$\begin{aligned}
& f(x_{t+1}, u_{t+1/2}) - f(x_t, y_t) \\
\leq & - \left(\frac{1}{\eta_x} - \frac{L}{2} \right) \|x_{t+1} - x_t\|^2 + L \|u_{t+1/2} - y_t\|^2 + L \|y_t - y_{t-1}\|^2 + \underbrace{\langle \nabla_y f(x_t, y_{t-1}), u_{t+1/2} - y_t \rangle}_{A_1}.
\end{aligned} \tag{11}$$

This completes the proof. □

Proof of Proposition 2

According to the update rule $y_{t+1} = \tau y_t + (1 - \tau) u_{t+1}$, we have

$$\begin{aligned}
& y_{t+1} - u_{t+1} = \tau (y_t - u_{t+1}), y_{t+1} - y_t = (1 - \tau)(u_{t+1} - y_t) \\
& u_{t+1/2} - y_t = \beta (\nabla_y f(x_t, u_{t-1/2}) - \nabla_y f(x_t, y_{t-1})).
\end{aligned} \tag{12}$$

Proof. For given x, f is L-smooth with respect to the dual variable y, and then we have

$$\begin{aligned}
& f(x_{t+1}, y_{t+1}) - f(x_{t+1}, u_{t+1/2}) \\
\leq & \ \langle \nabla_y f(x_{t+1}, u_{t+1/2}),\ y_{t+1} - u_{t+1/2} \rangle + \frac{L}{2} \|y_{t+1} - u_{t+1/2}\|^2 \\
= & \ \langle \nabla_y f(x_t, u_{t-1/2}),\ y_{t+1} - u_{t+1} \rangle + \langle \nabla_y f(x_{t+1}, u_{t+1/2}),\ u_{t+1} - u_{t+1/2} \rangle \\
& + \frac{L}{2} \|y_{t+1} - u_{t+1/2}\|^2 + \langle \nabla_y f(x_{t+1}, u_{t+1/2}) - \nabla_y f(x_t, u_{t-1/2}),\ y_{t+1} - u_{t+1} \rangle \\
= & \ \tau \langle \nabla_y f(x_t, u_{t-1/2}),\ y_t - u_{t+1} \rangle + \langle \nabla_y f(x_t, u_{t-1/2}),\ u_{t+1} - u_{t+1/2} \rangle \\
& + \langle \nabla_y f(x_{t+1}, u_{t+1/2}) - \nabla_y f(x_t, u_{t-1/2}),\ y_{t+1} - u_{t+1} \rangle + \frac{L}{2} \|y_{t+1} - u_{t+1/2}\|^2 \\
& + \langle \nabla_y f(x_{t+1}, u_{t+1/2}) - \nabla_y f(x_t, u_{t-1/2}),\ u_{t+1} - u_{t+1/2} \rangle \\
= & \ \underbrace{\tau \langle \nabla_y f(x_t, u_{t-1/2}),\ y_t - u_{t+1} \rangle}_{A_2} + \underbrace{\langle \nabla_y f(x_t, u_{t-1/2}),\ u_{t+1} - u_{t+1/2} \rangle}_{A_3} + a_t,
\end{aligned} \tag{13}$$

where the last inequality holds due to $a_t := 2L\|x_{t+1} - x_t\|^2 + 6L\beta^2\|\nabla_y f(x_t, u_{t-1/2}) - \nabla_y f(x_t, y_{t-1})\|^2 + 8L\|u_t - y_{t-1}\|^2 + 8L\beta^2\|\nabla_y f(x_{t-1}, u_{t-3/2}) - \nabla_y f(x_{t-1}, y_{t-2})\|^2 + 3L\|u_{t+1} - y_t\|^2$ and L-smoothness of f, i.e., $\|\nabla_y f(x_1, y_1) - \nabla_y f(x_2, y_2)\| \leq L\|x_1 - x_2\| + L\|y_1 - y_2\|$ and

$$
\begin{aligned}
a_t =& \langle \nabla_y f(x_{t+1}, u_{t+1/2}) - \nabla_y f(x_t, u_{t-1/2}),\ y_{t+1} - u_{t+1} \rangle + \frac{L}{2}\|y_{t+1} - u_{t+1/2}\|^2 \\
&+ \langle \nabla_y f(x_{t+1}, u_{t+1/2}) - \nabla_y f(x_t, u_{t-1/2}),\ u_{t+1} - u_{t+1/2} \rangle \\
\leq& \frac{1}{L}\|\nabla_y f(x_{t+1}, u_{t+1/2}) - \nabla_y f(x_t, u_{t-1/2})\|^2 + \frac{L}{2}\|u_{t+1} - u_{t+1/2}\|^2 \\
&+ \frac{L}{2}\|y_{t+1} - u_{t+1/2}\|^2 + \frac{L\tau^2}{2}\|u_{t+1} - y_t\|^2 \\
\leq& 2L\|x_{t+1} - x_t\|^2 + 2L\|u_{t+1/2} - u_{t-1/2}\|^2 + \frac{L}{2}\|u_{t+1} - u_{t+1/2}\|^2 + \frac{L}{2}\|y_{t+1} - u_{t+1/2}\|^2 + \frac{L}{2}\|u_{t+1} - y_t\|^2 \\
\leq& 2L\|x_{t+1} - x_t\|^2 + 4L\|u_{t+1/2} - y_t\|^2 + 8L\|y_t - y_{t-1}\|^2 + 8L\|y_{t-1} - u_{t-1/2}\|^2 + \frac{L}{2}\|u_{t+1} - y_t\|^2 \\
&+ L\|u_{t+1} - y_t\|^2 + L\|y_t - u_{t+1/2}\|^2 + L\|y_{t+1} - y_t\|^2 + L\|y_t - u_{t+1/2}\|^2 \\
\leq& 2L\|x_{t+1} - x_t\|^2 + 6L\|u_{t+1/2} - y_t\|^2 + 8L(1-\tau)^2\|u_t - y_{t-1}\|^2 + 8L\|y_{t-1} - u_{t-1/2}\|^2 \\
&+ 2L\|u_{t+1} - y_t\|^2 + L(1-\tau)^2\|u_{t+1} - y_t\|^2 \\
\leq& 2L\|x_{t+1} - x_t\|^2 + 6L\beta^2\|\nabla_y f(x_t, u_{t-1/2}) - \nabla_y f(x_t, y_{t-1})\|^2 + 8L\|u_t - y_{t-1}\|^2 \\
&+ 8L\beta^2\|\nabla_y f(x_{t-1}, u_{t-3/2}) - \nabla_y f(x_{t-1}, y_{t-2})\|^2 + 3L\|u_{t+1} - y_t\|^2.
\end{aligned} \tag{14}
$$

This completes the proof. □

Proof of Lemma 1

Proof. We first estimate the bound of $A_1 + A_3$, where A_1 is given in (11) and A_3 is given in (13)

$$
\begin{aligned}
& A_1 + A_3 \\
=& \big\langle \nabla_y f(x_t, y_{t-1}),\ u_{t+1/2} - y_t \big\rangle + \langle \nabla_y f(x_t, u_{t-1/2}),\ u_{t+1} - u_{t+1/2} \rangle \\
=& \big\langle \nabla_y f(x_t, y_{t-1}) - \nabla_y f(x_t, u_{t-1/2}), u_{t+1/2} - y_t \big\rangle + \langle \nabla_y f(x_t, u_{t-1/2}), u_{t+1} - y_t \rangle \\
=& -\beta\|\nabla_y f(x_t, u_{t-1/2}) - \nabla_y f(x_t, y_{t-1})\|^2 + \langle \nabla_y f(x_t, u_{t-1/2}), u_{t+1} - y_t - P_t \rangle \\
&+ \langle \nabla_y f(x_t, u_{t-1/2}), P_t \rangle \\
\leq& -\beta\|\nabla_y f(x_t, u_{t-1/2}) - \nabla_y f(x_t, y_{t-1})\|^2 + \frac{\delta^t}{2}\|\nabla_y f(x_t, u_{t-1/2})\|^2 \\
&+ \frac{1}{2\delta^t}\|u_{t+1} - y_t - P_t\|^2 + s_t \\
\leq& -\frac{3\beta}{4}\|\nabla_y f(x_t, u_{t-1/2}) - \nabla_y f(x_t, y_{t-1})\|^2 + \frac{1}{2\delta^t}\|u_{t+1} - y_t - P_t\|^2 + s_t,
\end{aligned} \tag{15}
$$

where the third equality holds by Quasi-Cocoercivity in Property 1 and P_t is defined below, the first inequality holds due to the Cauchy-Schwarz inequality with $\delta^t := \min\{\frac{\beta\|\nabla_y f(x_t, u_{t-1/2}) - \nabla_y f(x_t, y_{t-1})\|^2}{4\|\nabla_y f(x_t, u_{t-1/2})\|^2}, \frac{1}{28L}, \eta_x\}$ and $\eta_y^t = \delta^t + \delta$ and $s_t := \big\langle \nabla_y f(x_t, u_{t-1/2}), P_t \big\rangle$.

Furthermore, the optimality condition for the update of u_{t+1} in (4) implies that

$$
\left\langle \nabla_y f(x_t, u_{t-1/2}) - \frac{1}{\eta_y^t}(u_{t+1} - y_t), y_t - u_{t+1} \right\rangle \leq 0. \tag{16}
$$

Let $\delta^t := \min\{\frac{\beta\|\nabla_y f(x_t, u_{t-1/2}) - \nabla_y f(x_t, y_{t-1})\|^2}{4\|\nabla_y f(x_t, u_{t-1/2})\|^2}, \frac{1}{28L}, \eta_x\}$ and $\eta_y^t = \delta^t + \delta$, we also have

$$
\left(\delta^t + \delta\right) \nabla_y f(x_t, u_{t-1/2}) - (u_{t+1} - y_t) + \xi = \delta^t \nabla_y f(x_t, u_{t-1/2}) - (u_{t+1} - y_t) + P_t = 0,
$$

where $\xi \in \partial\Gamma$ is a subgradient of the indicator function Γ and $P_t := \delta\nabla_y f(x_t, u_{t-1/2}) + \xi$.

Next, we first bound A_2 in (13). Using the above two optimal conditions, we have

$$\begin{aligned}
A_2 =&\tau \left\langle \nabla_y f(x_t, u_{t-1/2}),\ y_t - u_{t+1} \right\rangle \\
=&\left\langle \nabla_y f(x_t, u_{t-1/2}),(1-\tau)(y_t-u_{t+1})+P_t\right\rangle - s_t + (2\tau-1)\left\langle \nabla_y f(x_t, u_{t-1/2}), y_t-u_{t+1}\right\rangle \\
=&\frac{1}{\delta^t}\left\langle u_{t+1} - y_t - P_t, (1-\tau)(y_t - u_{t+1}) + P_t\right\rangle - s_t + (2\tau-1)\left\langle \nabla_y f(x_t, u_{t-1/2}), y_t-u_{t+1}\right\rangle \\
\leq&\frac{(1-\tau)^2}{2\delta^t}\|u_{t+1}-y_t\|^2 - \frac{1}{2\delta^t}\|(1-\tau)(y_t-u_{t+1}) + P_t\|^2 - \frac{1}{2\delta^t}\|u_{t+1}-y_t-P_t\|^2 \\
&- s_t - \frac{(2\tau-1)}{\eta_y^t}\|u_{t+1}-y_t\|^2 \\
\leq&\frac{\beta}{4}\|\nabla_y f(x_t, u_{t-1/2}) - \nabla_y f(x_t, y_{t-1})\|^2 - \frac{(2\tau-1)}{\eta_y^t}\|u_{t+1}-y_t\|^2 - \frac{1}{2\delta^t}\|u_{t+1}-y_t-P_t\|^2 - s_t,
\end{aligned} \tag{17}$$

where the third equality holds due to the above optimality condition, s_t is defined in (15), the first inequality holds by Property 2 and the optimality condition in (16) and the last inequality holds due to $\frac{(1-\tau)^2}{2\delta^t}\|u_{t+1}-y_t\|^2 - \frac{1}{2\delta^t}\|(1-\tau)(y_t - u_{t+1}) + P_t\|^2 = \frac{(1-\tau)^2}{2\delta^t}\|P_t + \delta^t\nabla_y f(x_t, u_{t-1/2})\|^2 - \frac{1}{2\delta^t}\|\tau P_t - (1-\tau)\delta^t\nabla_y f(x_t, u_{t-1/2})\|^2 \leq \frac{(1-\tau)^2\delta_t}{2(2\tau-1)}\|\nabla_y f(x_t, u_{t-1/2})\|^2 \leq \frac{\beta}{4}\|\nabla_y f(x_t, u_{t-1/2}) - \nabla_y f(x_t, y_{t-1})\|^2$ ($\tau \geq 3/4$). We recall the results of Propositions 1 and 2 as follows:

$$\begin{aligned}
&f(x_{t+1}, u_{t+1/2}) - f(x_t, y_t) \\
\leq& -\left(\frac{1}{\eta_x} - \frac{L}{2}\right)\|x_{t+1} - x_t\|^2 + L\|u_{t+1/2} - y_t\|^2 + L\|y_t - y_{t-1}\|^2 + \underbrace{\left\langle \nabla_y f(x_t, y_{t-1}),\ u_{t+1/2} - y_t \right\rangle}_{A_1}.
\end{aligned}$$

$$f(x_{t+1}, y_{t+1}) - f(x_{t+1}, u_{t+1/2}) \leq\ A_2 + A_3 + a_t.$$

By the above inequalities and the update rules in (12), we have

$$\begin{aligned}
&f(x_{t+1}, y_{t+1}) - f(x_t, y_t) \\
&\leq A_1 + A_2 + A_3 + a_t - \left(\frac{1}{\eta_x} - \frac{L}{2}\right)\|x_{t+1} - x_t\|^2 + L\|u_{t+1/2} - y_t\|^2 + L\|y_t - y_{t-1}\|^2 \\
&\leq -\frac{2\tau-1}{\eta_y^t}\|u_{t+1} - y_t\|^2 - \frac{\beta}{2}\|\nabla_y f(x_t, u_{t-1/2}) - \nabla_y f(x_t, y_{t-1})\|^2 \\
&\quad -\left(\frac{1}{\eta_x} - \frac{L}{2}\right)\|x_{t+1} - x_t\|^2 + L\|u_{t+1/2} - y_t\|^2 + L\|y_t - y_{t-1}\|^2 + a_t.
\end{aligned} \tag{18}$$

With $a_t := 2L\|x_{t+1} - x_t\|^2 + 6L\beta^2\|\nabla_y f(x_t, u_{t-1/2}) - \nabla_y f(x_t, y_{t-1})\|^2 + 8L\|u_t - y_{t-1}\|^2 + 8L\beta^2\|\nabla_y f(x_{t-1}, u_{t-3/2}) - \nabla_y f(x_{t-1}, y_{t-2})\|^2 + 3L\|u_{t+1} - y_t\|^2$ and the update rules in (12),

$$\begin{aligned}
&f(x_{t+1}, y_{t+1}) - f(x_t, y_t) \\
\leq& -\frac{2\tau-1}{\eta_y^t}\|u_{t+1} - y_t\|^2 - \frac{\beta}{2}\|\nabla_y f(x_t, u_{t-1/2}) - \nabla_y f(x_t, y_{t-1})\|^2 \\
&-\left(\frac{1}{\eta_x} - 2L\right)\|x_{t+1} - x_t\|^2 + 7L\beta^2\|\nabla_y f(x_t, u_{t-1/2}) - \nabla_y f(x_t, y_{t-1})\|^2 + 9L\|u_t - y_{t-1}\|^2 \\
&+ 8L\beta^2\|\nabla_y f(x_{t-1}, u_{t-3/2}) - \nabla_y f(x_{t-1}, y_{t-2})\|^2 + 3L\|u_{t+1} - y_t\|^2.
\end{aligned} \tag{19}$$

Furthermore, using the definition of $G_t := f(x_t, y_t) + 9L\|u_t - y_{t-1}\|^2 + 8L\beta^2\|\nabla_y f(x_{t-1}, u_{t-3/2}) - \nabla_y f(x_{t-1}, y_{t-2})\|^2$ and the update rules in (12) with $\eta_x \leq \frac{1}{4L}$, $\eta_y^t \leq \frac{1}{14L}$ $(\delta \leq \frac{1}{28L})$ and $\tau \geq 3/4$, we have

$$\begin{aligned}
&G_t - G_{t+1} \\
\geq& \left(\frac{2\tau-1}{\eta_y^t} - 6L\right)\|u_{t+1} - y_t\|^2 + \left(\frac{\beta}{2} - 15L\beta^2\right)\|\nabla_y f(x_t, u_{t-1/2}) - \nabla_y f(x_t, y_{t-1})\|^2 + \frac{1}{2\eta_x}\|x_{t+1} - x_t\|^2 \\
\geq& L\|u_{t+1} - y_t\|^2 + \frac{\beta - 30L\beta^2}{2}\|\nabla_y f(x_t, u_{t-1/2}) - \nabla_y f(x_t, y_{t-1})\|^2 + \frac{\|x_{t+1} - x_t\|^2}{2\eta_x}.
\end{aligned} \tag{20}$$

By the above analysis, we have

$$\begin{aligned}
G_0 - G_T =& \sum_{t=0}^{T-1}(G_t - G_{t+1}) \\
\geq& \sum_{t=0}^{T-1}\left(L\|u_{t+1} - y_t\|^2 + \frac{\beta - 30L\beta^2}{2}\|\nabla_y f(x_t, u_{t-1/2}) - \nabla_y f(x_t, y_{t-1})\|^2 + \frac{\|x_{t+1} - x_t\|^2}{2\eta_x}\right)
\end{aligned} \tag{21}$$

This completes the proof. □

Appendix D: Proofs of Theorems 1 and 2

Proof of Theorem 1

Proof. Using the definition of the convergence criterion and the update rules in Algorithm 1, we define π_t as $\pi_t := \pi(x_t, y_t)$, where $\pi(x_t, y_t)$ is defined in (6) in Definition 1. Using the update rules of the primal variable x and dual variable y in Algorithm 1, non-expensive property of $\mathcal{P}_{\mathcal{X}}$ and smoothness in Assumption 1, we have

$$\begin{aligned} \|(\pi_t)_x\| &= \frac{1}{\eta_x} \|x_t - \mathcal{P}_{\mathcal{X}}\left(x_t - \eta_x \nabla_x f(x_t, y_t)\right)\| \\ &= \frac{1}{\eta_x} \|x_t - x_{t+1}\| . \end{aligned} \tag{22}$$

On the other hand, using the update rule of u_{t+1} in Algorithm 1 with $\eta_y^t = \min\{\frac{\beta\|\nabla_y f(x_t,u_{t-1/2})-\nabla_y f(x_t,y_{t-1})\|^2}{4\|\nabla_y f(x_t,u_{t-1/2})\|^2}, \frac{1}{28L}, \eta_x\} + \delta$, the non-expensive property of $\mathcal{P}_{\mathcal{Y}}$ and smoothness in Assumption 1,

$$\begin{aligned} \|(\pi_t)_y\| &= \frac{1}{\eta_y^t} \left\|y_t - \mathcal{P}_{\mathcal{Y}}\left(y_t + \eta_y^t \nabla_y f(x_t, y_t)\right)\right\| \\ &= \frac{1}{\eta_y^t} \left\|y_t - u_{t+1} + u_{t+1} - \mathcal{P}_{\mathcal{Y}}\left(y_t + \eta_y^t \nabla_y f(x_t, y_t)\right)\right\| \\ &\leq \frac{1}{\eta_y^t} \|y_t - u_{t+1}\| + \left\|\nabla_y f(x_t, u_{t-1/2}) - \nabla_y f(x_t, y_t)\right\| \\ &\leq \frac{1}{\delta} \|y_t - u_{t+1}\| + L\left\|u_{t-1/2} - y_t\right\| \\ &\leq \frac{1}{\delta} \|y_t - u_{t+1}\| + L\left\|u_{t-1/2} - y_{t-1}\right\| + L\|y_t - y_{t-1}\| , \end{aligned} \tag{23}$$

where the first inequality holds by the trigonometric inequality, the update rule in Algorithm 1 and the non-expensive property of $\mathcal{P}_{\mathcal{Y}}$ and the second inequality holds due to the smoothness of f and $1/\eta_y^t \leq 1/\delta$.

Using the above two results, then

$$\begin{aligned} \|\pi_t\|^2 &= \|(\pi_t)_x\|^2 + \|(\pi_t)_y\|^2 \\ &\leq \frac{1}{\eta_x^2} \|x_t - x_{t+1}\|^2 + \frac{2}{\delta^2} \|y_t - u_{t+1}\|^2 \\ &\quad + 4L^2 \left\|u_{t-1/2} - y_{t-1}\right\|^2 + 4L^2(1-\tau)^2 \|u_t - y_{t-1}\|^2 \\ &\leq \frac{1}{\eta_x^2} \|x_t - x_{t+1}\|^2 + \frac{2}{\delta^2} \|y_t - u_{t+1}\|^2 \\ &\quad + 4L^2\beta^2 \left\|\nabla_y f(x_t, u_{t-3/2}) - \nabla_y f(x_t, y_{t-2})\right\|^2 + 4L^2 \|u_t - y_{t-1}\|^2 . \end{aligned} \tag{24}$$

The result in (21) can be rewritten as follows:

$$\begin{aligned} &G_0 - G_T \\ &= \sum_{t=0}^{T-1} (G_t - G_{t+1}) \\ &\geq \sum_{t=0}^{T-1} \left(L\|u_{t+1} - y_t\|^2 + \frac{\beta - 30L\beta^2}{2} \|\nabla_y f(x_t, u_{t-1/2}) - \nabla_y f(x_t, y_{t-1})\|^2 + \frac{\|x_{t+1} - x_t\|^2}{2\eta_x} \right) . \end{aligned} \tag{25}$$

Similar to HiBSA [27] and APG [44], the first iteration index $T = T_\epsilon$ such that $\|\pi(x_t, y_t)\| \leq \epsilon$ is defined as: $T_\epsilon := \min\left\{t \,|\, \|\pi(x_t, y_t)\| \leq \epsilon\right\}$. Summing up the above inequality for all the iterations $t = 0, 1, \cdots, T_\epsilon - 1$ with $\beta < \frac{1}{60L}$, we obtain

$$\begin{aligned} & G_0 - G_{T_\epsilon} \\ = & \sum_{t=0}^{T_\epsilon - 1} [G_t - G_{t+1}] \\ \geq & \sum_{t=0}^{T_\epsilon - 1} \left(L\|u_{t+1} - y_t\|^2 + \frac{\beta - 30L\beta^2}{2} \|\nabla_y f(x_t, u_{t-1/2}) - \nabla_y f(x_t, y_{t-1})\|^2 + \frac{\|x_{t+1} - x_t\|^2}{2\eta_x} \right). \end{aligned} \tag{26}$$

With $\eta_x \leq \frac{1}{4L}$ and $\beta < \frac{1}{60L}$, we have

$$\sum_{t=0}^{T_\epsilon - 1} \|\pi_t\|^2 \leq \frac{160}{L\delta^2} \Big[G_0 - G_{T_\epsilon} \Big]. \tag{27}$$

Using the definition of G_t, the following result holds

$$\begin{aligned} G_0 - G_{T_\epsilon} &= G_0 - f(x_{T_\epsilon}, y_{T_\epsilon}) - 9L\|y_{T_\epsilon - 1} - u_{T_\epsilon}\|^2 - 8L\beta^2\|\nabla_x f(x_{T_\epsilon - 1}, u_{T_\epsilon - 3/2}) - \nabla_y f(x_{T_\epsilon - 1}, y_{T_\epsilon - 2})\|^2 \\ &\leq G_0 - f(x_{T_\epsilon}, y_{T_\epsilon}) \\ &= G_0 - f(x_{T_\epsilon}, u_{T_\epsilon - 1}) + f(x_{T_\epsilon}, u_{T_\epsilon - 1}) - f(x_{T_\epsilon}, y_{T_\epsilon}) \\ &\leq G_0 - f(x_{T_\epsilon}, u_{T_\epsilon - 1}) + \frac{L}{2}\|y^*_{T_\epsilon} - y_{T_\epsilon}\|^2 \\ &\leq G_0 - \underline{G} + 2LD_{\mathcal{Y}}^2. \end{aligned} \tag{28}$$

Using the two inequalities, we get

$$\sum_{t=0}^{T_\epsilon - 1} \|\pi_t\|^2 \leq \frac{160}{L\delta^2} \left(G_0 - G_{T_\epsilon} + 2LD_{\mathcal{Y}}^2 \right). \tag{29}$$

Thus,

$$\frac{1}{T_\epsilon} \sum_{t=0}^{T_\epsilon - 1} \|\pi_t\|^2 \leq \frac{160}{T_\epsilon L\delta^2} (G_0 - \underline{G} + 2LD_{\mathcal{Y}}^2),$$

which implies that $\epsilon^2 \geq 160\frac{G_0 - \underline{G} + 2LD_{\mathcal{Y}}^2}{T_\epsilon L\delta^2}$ or equivalently, $T_\epsilon \geq 160\frac{G_0 - \underline{G} + 2LD_{\mathcal{Y}}^2}{\epsilon^2 L\delta^2}$. That is, the gradient complexity of Algorithm 1 to find an ϵ-stationary point of f is

$$\mathcal{O}\left(\frac{G_0 - \underline{G} + 2LD_{\mathcal{Y}}^2}{\epsilon^2} \right).$$

This completes the proof. □

Proof of Theorem 2

Proof. Let $x^*(\widehat{x}) = \arg\min_{x \in \mathcal{X}}\{\phi(x) + L\|x - \widehat{x}\|^2\}$, $\phi(x) := \max_{y \in \mathcal{Y}} f(x, y)$ and $\phi_{1/2L}(\widehat{x}) := \min_{x \in \mathcal{X}}\{\phi(x) + L\|x - \widehat{x}\|^2\}$. Using the definitions of ϕ and $\phi_{1/2L}$, and using Property 4, we have

$$\|\nabla \phi_{1/2L}(\widehat{x})\|^2 = 4L^2 \|x^*(\widehat{x}) - \widehat{x}\|^2. \tag{30}$$

Following Property 3 and Lemma 19 in [24], the following result holds with the definition of $\phi(x)$

$$\begin{aligned} & \max_{y \in \mathcal{Y}} f(\widehat{x}, y) - \max_{y \in \mathcal{Y}} f(x^*(\widehat{x}), y) - L\|x^*(\widehat{x}) - \widehat{x}\|^2 \\ = & \phi(\widehat{x}) - \phi(x^*(\widehat{x})) - L\|x^*(\widehat{x}) - \widehat{x}\|^2 \\ \geq & \frac{L}{4}\|x^*(\widehat{x}) - \widehat{x}\|^2. \end{aligned} \tag{31}$$

Let Ω^* be the solution set of the problem $\max_{y\in\mathcal{Y}} f(\widehat{x}, y)$, and Ω^{**} be the solution set of the problem $\max_{y\in\mathcal{Y}} f(x^*(\widehat{x}), y)$. And let $y^*(\widehat{x}) \in \Omega^*$ and $y^*(x^*(\widehat{x})) \in \Omega^{**}$, respectively, we get

$$
\begin{aligned}
&\max_{y\in\mathcal{Y}} f(\widehat{x}, y) - \max_{y\in\mathcal{Y}} f(x^*(\widehat{x}), y) - L\|x^*(\widehat{x}) - \widehat{x}\|^2 \\
&= f(\widehat{x}, y^*(\widehat{x})) - f(x^*(\widehat{x}), y^*(x^*(\widehat{x})) - L\|x^*(\widehat{x}) - \widehat{x}\|^2 \\
&= f(\widehat{x}, y^*(\widehat{x})) - f(\widehat{x}, \widehat{y}) + f(\widehat{x}, \widehat{y}) - f(x^*(\widehat{x}), \widehat{y}) + f(x^*(\widehat{x}), \widehat{y}) - f(x^*(\widehat{x}), y^*(x^*(\widehat{x}))) - L\|x^*(\widehat{x}) - \widehat{x}\|^2 \\
&\le \langle \nabla_y f(\widehat{x}, \widehat{y}),\ y^*(\widehat{x}) - \widehat{y}\rangle + \langle \nabla_x f(\widehat{x}, \widehat{y}),\ \widehat{x} - x^*(\widehat{x})\rangle - \frac{L}{2}\|x^*(\widehat{x}) - \widehat{x}\|^2,
\end{aligned}
\tag{32}
$$

where the last inequality holds due to the concavity of f w.r.t. y (i.e.,$f(\widehat{x}, y^*) \le f(\widehat{x}, \widehat{y}) + \langle \nabla_y f(\widehat{x}, \widehat{y}),\ y^* - \widehat{y}\rangle$), the smoothness of f (i.e., $f(x^*(\widehat{x}), \widehat{y}) \le f(\widehat{x}, \widehat{y}) + \langle \nabla_x f(\widehat{x}, \widehat{y}),\ x^*(\widehat{x}) - \widehat{x}\rangle + \frac{L}{2}\|x^*(\widehat{x}) - \widehat{x}\|^2$) and $f(x^*(\widehat{x}), \widehat{y}) - f(x^*(\widehat{x}), y^*(x^*(\widehat{x})) \le 0$ (i.e., $y^*(x^*(\widehat{x}))$ is a solution of $\max_{y\in\mathcal{Y}} f(x^*(\widehat{x}), y)$).

With $\widehat{x} = x_t$, $\widehat{y} = u_t$, $y^*(\widehat{x}) = y^*(x_t)$, we have

$$
\begin{aligned}
&\frac{1}{16L}\|\nabla \phi_{1/2L}(\widehat{x})\|^2 = \frac{L}{4}\|x^*(\widehat{x}) - \widehat{x}\|^2 \\
&\le \max_{y\in\mathcal{Y}} f(x_t, y) - \max_{y\in\mathcal{Y}} f(x^*(x_t), y) - L\|x^*(x_t) - x_t\|^2 \\
&\le \langle \nabla_y f(x_t, u_t),\ y^*(x_t) - u_t\rangle + \langle \nabla_x f(x_{t-1}, y_{t-1}),\ x_t - x^*(x_t)\rangle - \frac{L}{2}\|x^*(x_t) - x_t\|^2 \\
&\quad + \langle \nabla_x f(x_t, u_t) - \nabla_x f(x_{t-1}, y_{t-1}),\ x_t - x^*(x_t)\rangle \\
&\le \langle \nabla_y f(x_t, u_{t-3/2}),\ y^*(x_t) - u_t\rangle + \langle \nabla_x f(x_{t-1}, y_{t-1}),\ x_t - x^*(x_t)\rangle - \frac{L}{2}\|x^*(x_t) - x_t\|^2 \\
&\quad + \langle \nabla_y f(x_t, u_t) - \nabla_y f(x_t, u_{t-3/2}),\ y^*(x_t) - u_t\rangle + \langle \nabla_x f(x_t, u_{t-1/2}) - \nabla_x f(x_{t-1}, y_{t-1}),\ x_t - x^*(x_t)\rangle \\
&\overset{\text{a}}{\le} \frac{1}{\eta_y^t}\langle u_t - y_{t-1},\ y^*(x_t) - u_t\rangle + \frac{1}{\eta_x}\langle x_t - x_{t-1},\ x^*(x_t) - x_t\rangle - \frac{L}{4}\|x^*(x_t) - x_t\|^2 \\
&\quad + 2\mathcal{D}_y\|\nabla_y f(x_t, u_t) - \nabla_y f(x_t, u_{t-3/2})\| + 4L\|x_t - x_{t-1}\|^2 + 4L\|u_{t-1/2} - y_{t-1}\|^2 \\
&\overset{\text{b}}{\le} \frac{2\mathcal{D}_y}{\delta}\|u_t - y_{t-1}\| + \frac{1}{L\eta_x^2}\|x_t - x_{t-1}\|^2 + \frac{L}{4}\|x^*(x_t) - x_t\|^2 - \frac{L}{4}\|x^*(x_t) - x_t\|^2 \\
&\quad + 2L\mathcal{D}_y(\|u_t - y_{t-1}\| + \|y_{t-1} - u_{t-1/2}\|) + 4L\|x_t - x_{t-1}\|^2 + 4L\beta^2\|\nabla_y f(x_t, y_{t-2}) - \nabla_y f(x_t, u_{t-3/2})\|^2, \\
&\le \frac{2\mathcal{D}_y(1+L)}{\delta}\|u_t - y_{t-1}\| + \frac{1}{L\eta_x^2}\|x_t - x_{t-1}\|^2 + 2\mathcal{D}_y\|\nabla_y f(x_t, y_{t-1}) - \nabla_y f(x_t, u_{t-1/2})\| \\
&\quad + 4L\|x_t - x_{t-1}\|^2 + 4L\|\nabla_y f(x_t, y_{t-2}) - \nabla_y f(x_t, u_{t-3/2})\|^2,
\end{aligned}
\tag{33}
$$

where the inequality (a) holds due to the optimality conditions in (9) and (16) and the inequality (b) holds due to the Cauchy-Schwarz inequality and the fact $\frac{1}{\eta_y^t} \le \frac{1}{\delta}$. Using Lemma 1 and by the similar derivation with Theorem 1, the complexity of Algorithm 1 to find an ϵ-stationary point of ϕ is

$$
\mathcal{O}\left(\frac{D_{\mathcal{Y}}(G_0 - \underline{G} + 2LD_{\mathcal{Y}}^2)}{\epsilon^2}\right).
$$

This completes the proof.

□