Proceedings of Machine Learning Research vol 291:1–31, 2025 38th Annual Conference on Learning Theory

Solving Convex-Concave Problems with $\tilde{\mathcal{O}}(\epsilon^{-4/7})$ Second-Order Oracle Complexity

Lesi Chen CHENLC23@MAILS.TSINGHUA.EDU.CN
IIIS, Tsinghua University
Shanghai Qizhi Institute

Chengchang Liu CCLIU22@CSE.CUHK.EDU.HK
Department of Computer Science & Engineering, The Chinese University of Hong Kong

Luo Luo LUOLUO@FUDAN.EDU.CN
School of Data Science, Fudan University
Shanghai Key Laboratory for Contemporary Applied Mathematics

Jingzhao Zhang[†] JINGZHAOZ@MAIL.TSINGHUA.EDU.CN
IIIS, Tsinghua University
Shanghai AI Lab
Shanghai Qizhi Institute

Editors: Nika Haghtalab and Ankur Moitra

Abstract

Previous algorithms can solve convex-concave minimax problems $\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} f(x,y)$ with $\mathcal{O}(\epsilon^{-2/3})$ second-order oracle calls using Newton-type methods. This result has been speculated to be optimal because the upper bound is achieved by a natural generalization of the optimal first-order method. In this work, we show an improved upper bound of $\tilde{\mathcal{O}}(\epsilon^{-4/7})$ by generalizing the optimal second-order method for convex optimization to solve the convex-concave minimax problem. We further apply a similar technique to lazy Hessian algorithms and show that our proposed algorithm can also be seen as a second-order "Catalyst" framework (Lin et al., 2018) that could accelerate any globally convergent algorithms for solving minimax problems.

Keywords: Minimax Optimization; Second-Order Methods; Acceleration

1. Introduction

We study the convex-concave minimax optimization problem over convex and compact sets $\mathcal{X}$, $\mathcal{Y}$:

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

This problem naturally arises from many applications, including finding a Nash equilibrium of a two-player zero-sum game (Von Neumann and Morgenstern, 1947; Karlin and Peres, 2017; Carmon et al., 2019, 2020b; Kornowski and Shamir, 2024), solving the Lagrangian function in constrained optimization (Ouyang and Xu, 2021; Kovalev et al., 2020; Scaman et al., 2017), and many machine learning problems such as adversarial training (Zhang et al., 2018), AUC maximization (Ying et al., 2016) and distributionally robust optimization (Ben-Tal et al., 2009; Carmon and Hausler, 2022; Curi et al., 2020; Song et al., 2022).

[†]. The corresponding author.

Problem (1) can also be viewed as variational inequality problems (Kinderlehrer and Stampacchia, 2000). Specifically, we let $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y})$ and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, then Problem (1) can be formulated by the following variational inequality (VI) problem that targets to find a solution $\boldsymbol{z}^* \in \mathcal{Z}$ such that

$$\langle \boldsymbol{F}(\boldsymbol{z}), \boldsymbol{z} - \boldsymbol{z}^* \rangle \geq 0 \;\text{ for all } \boldsymbol{z} \in \mathcal{Z}, \quad \text{where } \; \boldsymbol{F}(\boldsymbol{z}) = \begin{bmatrix} \nabla_x f(\boldsymbol{x}, \boldsymbol{y}) \\ -\nabla_y f(\boldsymbol{x}, \boldsymbol{y}) \end{bmatrix} \tag{2}$$

is monotone. The algorithms for solving monotone VIs are well-studied in the literature and can be used to solve our Problem (2). The seminal work by Korpelevich (1976) showed the extragradient (EG) method can find an ϵ-solution to Problem (1) with the $\mathcal{O}(\epsilon^{-1})$ calls of the first-order oracle $\boldsymbol{F}(\boldsymbol{z})$. Nemirovskij and Yudin (1983); Ouyang and Xu (2021) showed that any first-order algorithm for a bilinear minimax problem must make at least $\Omega(\epsilon^{-1})$ first-order queries, which proves the optimality of EG in first-order optimization. Monteiro and Svaiter (2012) proposed the Newton Proximal Extragradient (NPE) method as a second-order extension of EG, and they showed that NPE can provably find an ϵ-solution to Problem (1) with $\mathcal{O}(\epsilon^{-2/3})$ calls of the second-order oracle $(f(\boldsymbol{z}), \nabla f(\boldsymbol{z}), \nabla^2 f(\boldsymbol{z}))$. The NPE algorithm has been simplified by many subsequent works (Adil et al., 2022; Huang and Zhang, 2022; Lin et al., 2022; Bullins and Lai, 2022; Liu and Luo, 2022; Lin and Jordan, 2024), while they still require the $\mathcal{O}(\epsilon^{-2/3})$ second-order oracle calls, which are speculated to be optimal under restricted conditions (see Appendix A).

In this work, we show that it is possible to break the barrier $\Omega(\epsilon^{-2/3})$ for second-order convex-concave minimax optimization with practical algorithms. We propose the Minimax-AIPE algorithm, which generalizes the second-order methods A-NPE (Monteiro and Svaiter, 2013) for minimization problems. We prove our method requires at most $\tilde{\mathcal{O}}\big(D_x^{6/7} D_y^{6/7} (\rho/\epsilon)^{4/7}\big)$ second-order oracle calls to find an ϵ-solution for Problem (1) with the assumption of ρ-Lipschitz continuous Hessian, where D_x and D_y are the diameters of $\mathcal{X}$ and $\mathcal{Y}$, respectively. Our result significantly improves the existing upper bound of $\mathcal{O}\left(\max\{D_x^2, D_y^2\}(\rho/\epsilon)^{2/3}\right)$ achieved by the NPE method and its variants.

Our proposed Minimax-AIPE is a triple-loop algorithm: the outer loop runs $\tilde{\mathcal{O}}\big(D_x^{6/7}(\gamma/\epsilon)^{2/7}\big)$ iterations of the restarted Accelerated Inexact Proximal Exragradient (AIPE-restart, Algorithm 2) in variable $\boldsymbol{x}$; the middle loop runs $\tilde{\mathcal{O}}\big(D_y^{6/7}(\gamma/\epsilon)^{2/7}\big)$ iterations of AIPE-restart in variable $\boldsymbol{y}$; and the inner loop implements a minimax proximal step

$$\min_{\boldsymbol{x} \in \mathcal{X}} \max_{\boldsymbol{y} \in \mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y}) + \frac{\gamma}{3}\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|^3 - \frac{\gamma}{3}\|\boldsymbol{y} - \bar{\boldsymbol{y}}\|^3$$

via a linearly convergent method $\mathcal{M}$, where $(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}})$ is the proximal center and γ is a hyper-parameter to be chosen. By applying different algorithms $\mathcal{M}$ to implement the proximal step and tuning the hyper-parameter γ accordingly, we naturally obtain the acceleration of these algorithms. For instance, choosing $\mathcal{M}$ to be the NPE method (Monteiro and Svaiter, 2012) achieves the $\tilde{\mathcal{O}}\big(D_x^{6/7} D_y^{6/7} (\rho/\epsilon)^{4/7}\big)$ second-order oracle complexity as we claimed. Moreover, our method is also compatible with the lazy Hessian technique to reduce the computational complexity of second-order methods. Choosing $\mathcal{M}$ to be the recently proposed lazy version of NPE (Chen et al., 2025a) obtains a $\tilde{\mathcal{O}}\big(m + m^{5/7} D_x^{6/7} D_y^{6/7} (\rho/\epsilon)^{4/7}\big)$ upper bound of the number of iterations when reusing Hessians every m iterations. The main results of this paper are shown in Table 1.

Notations We use $\|\cdot\|$ to denote the Euclidean norm for vectors and the spectral norm for matrices. We hide logarithmic factors in the notation $\tilde{\mathcal{O}}(\,\cdot\,)$. For a set $\mathcal{S}$, we denote $\mathcal{I}_{\mathcal{S}}$ to its indicator function

Table 1: We compare the theoretical results to find an ϵ-solution for Problem (1) of representative second-order methods NPE (Monteiro and Svaiter, 2012) and LEN (Chen et al., 2025a) before and after applying our proposed Minimax-AIPE acceleration framework.

Method	Before Acc.	After Acc.	# Hessians	Reference
NPE	$\mathcal{O}(\epsilon^{-2/3})$	$\tilde{\mathcal{O}}(\epsilon^{-4/7})$	every step	Theorem 5.5
LEN	$\mathcal{O}(m + m^{2/3}\epsilon^{-2/3})$	$\tilde{\mathcal{O}}(m + m^{5/7}\epsilon^{-4/7})$	once per m steps	Theorem 5.7

and $\partial \mathcal{I}_{\mathcal{S}}$ denotes the subgradient of the function. We define $D := \max\{D_x, D_y\}$. To simplify notations, we let $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y})$ and define the operator as $\boldsymbol{F}(\boldsymbol{z}) = \begin{bmatrix} \nabla_x f(\boldsymbol{x}, \boldsymbol{y}) \\ -\nabla_y f(\boldsymbol{x}, \boldsymbol{y}) \end{bmatrix}$.

2. Related Works

We review the existing results of different methods for convex-concave minimax optimization and the related acceleration techniques used in our methods.

First-Order Methods The optimal oracle complexity of first-order algorithms to solve convex-concave minimax problems is well known. Various methods, including extragradient (Korpelevich, 1976), dual extrapolation (Nesterov, 2007), optimistic gradient descent ascent (Popov, 1980; Mokhtari et al., 2020a,b), achieve the upper bound $\mathcal{O}(\max\{D_x^2, D_y^2\}/\epsilon)$, which is known to be optimal when $D_x = D_y$ (Nemirovskij and Yudin, 1983). Inspired by the accelerated proximal point algorithm / Catalyst (Lin et al., 2018), Lin et al. (2020) proposed the Minimax-APPA algorithm which achieves the upper bound of $\tilde{\mathcal{O}}(D_x D_y/\epsilon)$, which fully matches the lower bound when $D_x \neq D_y$ (Ouyang and Xu, 2021). Subsequent works proposed enhanced algorithms that further improve the logarithmic factors (Yang et al., 2020; Kovalev and Gasnikov, 2022a; Carmon et al., 2022b) or give sharper rates under the refined (L_x, L_{xy}, L_y)-smoothness condition (Wang and Li, 2020; Jin et al., 2022).

Second-Order Methods Monteiro and Svaiter (2012) generalized the EG algorithm and proposed the Newton Proximal Extragradient (NPE) method with an $\mathcal{O}\left(\max\{D_x^2, D_y^2\}\epsilon^{-2/3}\right)$ upper bound of second-order oracle calls for monotone variational inequalities. Subsequently, researchers have proposed different generalizations of the optimal first-order methods and have demonstrated the same theoretical guarantees (Bullins and Lai, 2022; Huang and Zhang, 2022; Adil et al., 2022; Lin and Jordan, 2024; Jiang et al., 2024). Very recently, Chen et al. (2025a) proposed Lazy Extra Newton (LEN) to further reduce the computational complexity of NPE by using lazy Hessian updates (Doikov et al., 2023). By analogy with the results of first-order methods, these works conjectured that they have achieved the optimal complexity in ϵ of the second-order algorithm when $D_x = D_y$. Furthermore, some works (Adil et al., 2022; Lin and Jordan, 2024) have attempted to establish a lower bound of $\Omega(\epsilon^{-2/3})$ for this problem. However, this paper refutes the possibility of the existence of such a lower bound. In Appendix A, we discuss why the lower bounds established in Adil et al. (2022); Lin and Jordan (2024) are not applicable to ours.

Higher-Order Methods There are also works that generalized first-order and second-order methods to pth-order. Assuming a p-th order tensor step can be implemented, state-of-the-art pth-order methods achieve the iteration complexity of $\mathcal{O}(\max\{D_x^2, D_y^2\}\epsilon^{-2/(p+1)})$ (Bullins and Lai, 2022; Nesterov, 2023a; Lin and Jordan, 2024). However, the pth-order methods for minimax problems remain "conceptual" in the case $p \geq 3$ as people do not know how to implement the p-th order tensor step in general. Therefore, this work only focuses on the implementable case $p = 2$, although we expect our result to be generalizable for all p.

Acceleration Our technique for acceleration is also closely related to the acceleration in convex optimization. Monteiro and Svaiter (2013) proposed the Accelerated Hybrid Proximal Extragradient (A-HPE) framework. The second-order instance of A-HPE, called Accelerated Newton Proximal Extragradient (A-NPE) achieves the $\mathcal{O}(\epsilon^{-2/7}\log\epsilon^{-1})$ iteration complexity. Arjevani et al. (2019) showed an $\Omega(\epsilon^{-2/7})$ lower bound for second-order convex optimization. Recently, two independent works (Carmon et al., 2022a; Kovalev and Gasnikov, 2022a) proposed novel enhancements of A-NPE to remove the $\mathcal{O}(\log\epsilon^{-1})$ factor caused by line search for the extrapolation coefficient. Chen et al. (2025a) proposed a more computationally efficient version of the search-free A-NPE (Carmon et al., 2022a) by incorporating the lazy Hessian technique (Doikov et al., 2023). The A-HPE framework is a powerful tool for acceleration, which has also been used to accelerate quasi-Newton methods (Jiang and Mokhtari, 2024), tensor methods (Jiang et al., 2019; Bubeck et al., 2019b; Gasnikov et al., 2019; Carmon et al., 2022a), optimization with ball oracles (Carmon et al., 2020a, 2021, 2024; Carmon and Hausler, 2022), and parallel optimization (Bubeck et al., 2019a; Carmon et al., 2023).

3. Preliminaries

In this section, we describe the problem setup and some auxiliary definitions in this paper. We first introduce the standard convex-concave minimax problems and impose the following assumptions for Problem (1). We first assume that the function is convex-concave and the domain is bounded.

Assumption 3.1 (Convex-concavity) *We suppose that $f(\,\cdot\,\boldsymbol{y})$ is convex for any fixed $\boldsymbol{y}$ and $f(\boldsymbol{x},\,\cdot\,)$ is concave for any fixed $\boldsymbol{x}$.*

Assumption 3.2 (Bounded domain) *We suppose that both the sets $\mathcal{X}$ and $\mathcal{Y}$ are convex and compact with diameters $D_x := \sup_{\boldsymbol{x},\boldsymbol{x}'\in\mathcal{X}}\|\boldsymbol{x}-\boldsymbol{x}'\| < +\infty$ and $D_y := \sup_{\boldsymbol{y},\boldsymbol{y}'\in\mathcal{Y}}\|\boldsymbol{y}-\boldsymbol{y}'\| < +\infty$.*

We then assume that the function is Lipschitz and smooth.

Assumption 3.3 (Lipschitzness) *We suppose $f(\boldsymbol{x},\boldsymbol{y})$ is L-Lipschitz continuous for some $L > 0$:*

$$|f(\boldsymbol{x},\boldsymbol{y}) - f(\boldsymbol{x}',\boldsymbol{y}')| \leq L\|\boldsymbol{x}-\boldsymbol{y}\|, \quad \forall \boldsymbol{x},\boldsymbol{x}'\in\mathcal{X},\ \boldsymbol{y},\boldsymbol{y}'\in\mathcal{Y}.$$

Assumption 3.4 (Gradient Lipschitzness) *We suppose the gradient of $f(\boldsymbol{x},\boldsymbol{y})$ is ℓ-Lipschitz continuous for some $\ell > 0$:*

$$\|\nabla f(\boldsymbol{x},\boldsymbol{y}) - \nabla f(\boldsymbol{x}',\boldsymbol{y}')\| \leq \ell\|\boldsymbol{x}-\boldsymbol{y}\|, \quad \forall \boldsymbol{x},\boldsymbol{x}'\in\mathcal{X},\ \boldsymbol{y},\boldsymbol{y}'\in\mathcal{Y}.$$

Assumption 3.5 (Hessian Lipschitzness) *We suppose the Hessian of $f(\boldsymbol{x},\boldsymbol{y})$ is ρ-Lipschitz continuous for some $\rho > 0$:*

$$\|\nabla^2 f(\boldsymbol{x},\boldsymbol{y}) - \nabla^2 f(\boldsymbol{x}',\boldsymbol{y}')\| \leq \rho\|\boldsymbol{x}-\boldsymbol{y}\|, \quad \forall \boldsymbol{x},\boldsymbol{x}'\in\mathcal{X},\ \boldsymbol{y},\boldsymbol{y}'\in\mathcal{Y}.$$

We note that previous second-order methods generally do not impose Assumption 3.3 and 3.4. Although our additionally assumptions look more restricted than existing works, it is important to note that they are mild and can be derived from the higher-order smoothness (Assumption 3.5) in a compact set. For this problem, we want to find an approximate solution defined as follows.

Definition 3.1 *We say $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) \in \mathcal{X} \times \mathcal{Y}$ is an ϵ-solution to Problem (1) if*

$$\mathrm{Gap}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) := \max_{\boldsymbol{y} \in \mathcal{Y}} f(\hat{\boldsymbol{x}}, \boldsymbol{y}) - \min_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}, \hat{\boldsymbol{y}}) \leq \epsilon.$$

When $\epsilon = 0$, they are the exact saddle points of Problem (1). We focus on the complexity to find an ϵ-solution. By following previous works, we measure the complexity by the number of cubic regularized Newton (CRN) oracles, which is formally defined as follows.

Second-Order Oracles We introduce the second-order oracles. We then highlight several important and known lemmas for our analyses.

Definition 3.2 *A CRN oracle for Problem (1) takes the query point $\bar{\boldsymbol{z}} \in \mathcal{Z}$ and the regularization parameter $\gamma > 0$ as inputs and returns $(\boldsymbol{z}, \boldsymbol{u}) = \mathrm{CRN}(\bar{\boldsymbol{z}}, \gamma)$ satisfies:*

$$\begin{aligned} &\langle \boldsymbol{F}(\bar{\boldsymbol{z}}) + \nabla \boldsymbol{F}(\bar{\boldsymbol{z}})(\boldsymbol{z} - \bar{\boldsymbol{z}}) + \frac{\gamma}{2}\|\boldsymbol{z} - \bar{\boldsymbol{z}}\|(\boldsymbol{z} - \bar{\boldsymbol{z}}), \boldsymbol{z}' - \boldsymbol{z} \rangle \geq 0, \ \forall \boldsymbol{z}' \in \mathcal{Z}; \\ &\boldsymbol{u} = -\left(\boldsymbol{F}(\bar{\boldsymbol{z}}) + \nabla \boldsymbol{F}(\bar{\boldsymbol{z}})(\boldsymbol{z} - \bar{\boldsymbol{z}}) + \frac{\gamma}{2}\|\boldsymbol{z} - \bar{\boldsymbol{z}}\|(\boldsymbol{z} - \bar{\boldsymbol{z}})\right) \in \begin{bmatrix} \partial \mathcal{I}_{\mathcal{X}}(\boldsymbol{x}) \\ -\partial \mathcal{I}_{\mathcal{Y}}(\boldsymbol{y}) \end{bmatrix}, \end{aligned}$$

where $\boldsymbol{F}(\boldsymbol{z}) = \begin{bmatrix} \nabla_x f(\boldsymbol{x}, \boldsymbol{y}) \\ -\nabla_y f(\boldsymbol{x}, \boldsymbol{y}) \end{bmatrix}$. Particularly, for minimization problem $\min_{\boldsymbol{z} \in \mathcal{Z}} f(\boldsymbol{z})$ we have

$$\begin{aligned} \boldsymbol{z} &= \arg\min_{\boldsymbol{z}' \in \mathcal{Z}} \left\{ f(\boldsymbol{z}') + \langle \nabla f(\bar{\boldsymbol{z}}), \boldsymbol{z}' - \bar{\boldsymbol{z}} \rangle + \frac{1}{2}\langle \nabla^2 f(\bar{\boldsymbol{z}})(\boldsymbol{z}' - \bar{\boldsymbol{z}}), \boldsymbol{z}' - \bar{\boldsymbol{z}} \rangle + \frac{\gamma}{6}\|\boldsymbol{z}' - \bar{\boldsymbol{z}}\|^3 \right\}; \\ \boldsymbol{u} &= -(\nabla f(\bar{\boldsymbol{z}}) + \nabla^2 f(\bar{\boldsymbol{z}})(\boldsymbol{z} - \bar{\boldsymbol{z}}) + \frac{\gamma}{2}\|\boldsymbol{z} - \bar{\boldsymbol{z}}\|(\boldsymbol{z} - \bar{\boldsymbol{z}})) \in \partial \mathcal{I}_{\mathcal{Z}}(\boldsymbol{z}). \end{aligned}$$

It is well known that the above oracle can be implemented by transforming it into an equivalent auxiliary one-dimensional problem, which can be efficiently solved with line search procedure (Monteiro and Svaiter, 2012, Section 4).

Making Gradient Small Our method also relies on tools for making gradients small (Allen-Zhu, 2018; Foster et al., 2019; Yoon and Ryu, 2021; Chen and Luo, 2024; Lee and Kim, 2021; Cai et al., 2022). We recall the following fact for extragradient (EG) update:

$$\begin{aligned} \boldsymbol{z}_{0.5} &= \arg\min_{\boldsymbol{z} \in \mathcal{Z}} \left\{ \langle \boldsymbol{F}(\boldsymbol{z}_0), \boldsymbol{z} \rangle + \frac{1}{2\eta}\|\boldsymbol{z} - \boldsymbol{z}_0\|^2 \right\}, \\ \boldsymbol{z}_1 &= \arg\min_{\boldsymbol{z} \in \mathcal{Z}} \left\{ \langle \boldsymbol{F}(\boldsymbol{z}_{0.5}), \boldsymbol{z} \rangle + \frac{1}{2\eta}\|\boldsymbol{z} - \boldsymbol{z}_0\|^2 \right\} \end{aligned} \tag{3}$$

Lemma 3.1 (Cai et al. (2022, Lemma 12), $T = 1$) *Under Assumption 3.1 and 3.4, the extragradient update (3) with $\eta \in (0, 1/\ell)$ on Problem (1) satisfies that*

$$\|\boldsymbol{F}(\boldsymbol{z}_1) + \boldsymbol{c}_1\| \leq \frac{1 + \eta\ell + (\eta\ell)^2}{\eta\sqrt{1 - (\eta\ell)^2}}\|\boldsymbol{z}_0 - \boldsymbol{z}^*\|,$$

where $\boldsymbol{c}_1 = (\boldsymbol{z}_0 - \boldsymbol{z}_1)/\eta - \boldsymbol{F}(\boldsymbol{z}_{0.5}) \in \partial \mathcal{I}_{\mathcal{Z}}(\boldsymbol{z}_1)$.

For convenience, we also define the above EG step as an oracle.

Definition 3.3 *Given an input $\boldsymbol{z}_0 \in \mathcal{Z}$ and the operator $\boldsymbol{F}$, we let $\mathrm{EG}(\boldsymbol{z}_0, \boldsymbol{F}, \eta)$ performs the update as Eq. (3) and returns $(\boldsymbol{z}_1, \boldsymbol{c}_1)$ defined in Lemma 3.1.*

Uniform Convexity Finally, we introduce the definition of uniform convexity, which is an important property that will frequently be used in our analysis. In this paper, we only considered the third-order uniform convexity, which is formally defined as follows.

Definition 3.4 *A function $h(\boldsymbol{z}) : \mathcal{Z} \to \mathbb{R}$ is μ-uniformly convex (of order 3) for some $\mu > 0$ if*

$$h(\boldsymbol{z}) \geq h(\boldsymbol{z}') + \langle \nabla h(\boldsymbol{z}'), \boldsymbol{z} - \boldsymbol{z}' \rangle + \frac{\mu}{3} \|\boldsymbol{z} - \boldsymbol{z}'\|^3, \quad \forall \boldsymbol{z}, \boldsymbol{z}' \in \mathcal{Z}.$$

We say $h(\boldsymbol{z})$ is μ-uniformly concave if $-h(\boldsymbol{z})$ is μ-uniformly concave.

The uniform convexity of order 3 is an important class, where second-order methods such as the cubic regularized Newton method (Nesterov and Polyak, 2006) enjoy a linear convergence rate. In the following, we recall some properties of uniformly convex functions, which is useful to derive the main results of this paper.

Lemma 3.2 (Section 4.2.2, Nesterov (2018)) *Let $h(\boldsymbol{z}) : \mathcal{Z} \to \mathbb{R}$ be a μ-uniformly convex function and let $\boldsymbol{z}^* = \arg\min_{\boldsymbol{z} \in \mathcal{Z}} h(\boldsymbol{z})$. Then for any $\boldsymbol{z} \in \mathcal{Z}$, we have*

$$\frac{2}{3\sqrt{\mu}} \|\nabla h(\boldsymbol{z})\|^{\frac{3}{2}} \geq h(\boldsymbol{z}) - h(\boldsymbol{z}^*) \geq \frac{\mu}{3} \|\boldsymbol{z} - \boldsymbol{z}^*\|^3.$$

An illustrating example is the cubic function $d(\boldsymbol{z}) = (1/3)\|\boldsymbol{z}\|^3$, which has the following properties.

Lemma 3.3 (Lemma 4.2.3 and Lemma 4.2.4, Nesterov (2018)) *Let $d(\boldsymbol{z}) = (1/3)\|\boldsymbol{z}\|^3$ be the cubic function. We have that $d(\boldsymbol{z})$ is $(1/2)$-uniformly convex and has 2-Lipschitz continuous Hessians.*

4. The Subroutine: Accelerated Inexact Proximal Extragradient

Before presenting our main algorithm, we first introduce the Accelerated Inexact Proximal Extragradient (AIPE) method for minimizing a convex function $h(\boldsymbol{z}) : \mathcal{Z} \to \mathbb{R}$. It will be an important component in our main algorithm later. We present its procedure in Algorithm 1. It generalizes the recently proposed search-free A-HPE method (Carmon et al., 2022a) by allowing the following inexact proximal oracles.

Definition 4.1 (Inexact Second-Order Proximal Oracle) *An oracle is called a (δ, γ)-(second-order)-proximal oracle for function $h : \mathbb{R}^d \to \mathbb{R}$ if for every $\bar{\boldsymbol{z}} \in \mathbb{R}^d$ the points $(\boldsymbol{z}, \boldsymbol{u}) = \mathrm{iProx}_h(\bar{\boldsymbol{z}}, \gamma)$ with $\boldsymbol{z} \in \mathcal{Z}$ and $\boldsymbol{u} \in \partial \mathcal{I}_{\mathcal{Z}}(\boldsymbol{z})$ satisfy*

$$\|\nabla h(\boldsymbol{z}) + \boldsymbol{u} + \lambda(\boldsymbol{z} - \bar{\boldsymbol{z}})\| \leq \frac{\lambda}{2} \|\boldsymbol{z} - \bar{\boldsymbol{z}}\| + \delta, \quad \lambda = \gamma \|\boldsymbol{z} - \bar{\boldsymbol{z}}\|.$$

When $\delta = 0$ the above oracle reduces to the oracle used in previous works (Monteiro and Svaiter, 2013; Carmon et al., 2022a; Kovalev and Gasnikov, 2022a; Nesterov, 2021, 2023b), which can be implemented by a CRN oracle.

Algorithm 1 AIPE$(\boldsymbol{z}_0, T, \gamma, \delta)$

```
1:  $\boldsymbol{v}_0 = \boldsymbol{z}_0$, $\bar{\boldsymbol{z}}_0 = \boldsymbol{z}_0$, $A_0 = 0$
2:  $\tilde{\boldsymbol{z}}_1, \boldsymbol{u}_1 = \mathrm{iProx}_h(\bar{\boldsymbol{z}}_0, \gamma, \delta)$
3:  $\lambda_1' = \lambda_1 = \gamma \|\tilde{\boldsymbol{z}}_1 - \bar{\boldsymbol{z}}_0\|$
4:  for $t = 0, \cdots, T-1$ do
5:      Solve $a_{t+1}' > 0$ from $A_t + a_{t+1}' = 2\lambda_{t+1}' \left(a_{t+1}'\right)^2$
6:      $A_{t+1}' = A_t + a_{t+1}'$
7:      $\bar{\boldsymbol{z}}_t = \frac{A_t}{A_{t+1}'} \boldsymbol{z}_t + \frac{a_{t+1}'}{A_{t+1}'} \boldsymbol{v}_t$
8:      if $t > 0$ then
9:          $\tilde{\boldsymbol{z}}_{t+1}, \boldsymbol{u}_{t+1} = \mathrm{iProx}_h(\bar{\boldsymbol{z}}_t, \gamma, \delta)$
10:         $\lambda_{t+1} = \gamma \|\tilde{\boldsymbol{z}}_{t+1} - \bar{\boldsymbol{z}}_t\|$
11:     end if
12:     if $\lambda_{t+1} \leq \lambda_{t+1}'$ then
13:         $a_{t+1} = a_{t+1}'$, $A_{t+1} = A_{t+1}'$
14:         $\boldsymbol{z}_{t+1} = \tilde{\boldsymbol{z}}_{t+1}$
15:         $\lambda_{t+2}' = \frac{1}{2} \lambda_{t+1}'$
16:     else
17:         $\gamma_{t+1} = \frac{\lambda_{t+1}'}{\lambda_{t+1}}$
18:         $a_{t+1} = \gamma_{t+1} a_{t+1}'$, $A_{t+1} = A_t + a_{t+1}$
19:         $\boldsymbol{z}_{t+1} = \frac{(1-\gamma_{t+1})A_t}{A_{t+1}} \boldsymbol{z}_t + \frac{\gamma_{t+1} A_{t+1}'}{A_{t+1}} \tilde{\boldsymbol{z}}_{t+1}$
20:         $\lambda_{t+2}' = 2\lambda_{t+1}'$
21:     end if
22:     Get $\boldsymbol{g}_{t+1}$ such that $\|\boldsymbol{g}_{t+1} - \nabla h(\tilde{\boldsymbol{z}}_{t+1})\| \leq \delta$
23:     $\boldsymbol{v}_{t+1} = \arg\min_{\boldsymbol{v} \in \mathcal{Z}} \left\{ \langle \boldsymbol{g}_{t+1} + \boldsymbol{u}_{t+1}, \boldsymbol{v} \rangle + \frac{1}{2a_{t+1}} \|\boldsymbol{v} - \boldsymbol{v}_t\|^2 \rangle \right\}$
24: end for
25: $\boldsymbol{z}^{\mathrm{out}} = \arg\min_{0 \leq t \leq T} \left\{ \hat{h}(\boldsymbol{z}); \boldsymbol{z} \in \{\boldsymbol{z}_t, \tilde{\boldsymbol{z}}_t\} \right\}$, where $|\hat{h}(\boldsymbol{z}) - h(\boldsymbol{z})| \leq \delta$.
26: return $\boldsymbol{z}^{\mathrm{out}}$
```

Algorithm 2 AIPE-restart$(\boldsymbol{z}_0, T, \gamma, \delta, S)$

```
$\boldsymbol{z}^{(0)} = \boldsymbol{z}_0$
for $s = 0, \cdots, S-1$ do
    $\boldsymbol{z}^{(s+1)} = \mathrm{AIPE}(\boldsymbol{z}^{(s)}, T, \gamma, \delta)$
end for
return $\boldsymbol{z}^{(S)}$
```

Lemma 4.1 (Carmon et al. (2022a, Section 3.1)) *Assume $h(\boldsymbol{z}) : \mathcal{Z} \to \mathbb{R}$ has ρ-Lipschitz continuous Hessians. The CRN oracle* $\mathrm{CRN}(\,\cdot\,, 2\rho)$ *implements an* $(0, \rho)$*-proximal oracle.*

The main modification we made to the algorithm by Carmon et al. (2022a) are marked in blue in Algorithm 1, where Line 9 replaces the exact proximal oracle with an inexact one as per Definition 4.1, and Line 22 and Line 25 also allow the following inexact zeroth-order and first-order oracles.

Definition 4.2 *We call $\hat{h}(\boldsymbol{z})$ a δ-zeroth-order oracle of function $h : \mathcal{Z} \to \mathbb{R}$ if $|\hat{h}(\boldsymbol{z}) - h(\boldsymbol{z})| \leq \delta$.*

Definition 4.3 *We call $\boldsymbol{g}(\boldsymbol{z})$ a δ-first-order oracle of function $h : \mathcal{Z} \to \mathbb{R}$ if $\|\boldsymbol{g}(\boldsymbol{z}) - \nabla h(\boldsymbol{z})\| \leq \delta$.*

Since the accelerated algorithm is easily affected by errors, it causes challenges in analyzing AIPE. In our proof, we restrict the iterations in $t \leq T_\epsilon$ to avoid some ill-conditioned case after the algorithm has reached an ϵ-solution, where $T_\epsilon = \arg\min_{0 \leq t \leq T}\{h(\boldsymbol{z}); \boldsymbol{z} \in \{\boldsymbol{z}_t, \tilde{\boldsymbol{z}}_t\}\}$. Then we can follow the same steps as (Carmon et al., 2022a, Theorem 1) to show that it is sufficient to let $\delta \lesssim \epsilon / A_{T_\epsilon}$ to recover the convergence rate as A-HPE (Monteiro and Svaiter, 2013). Finally, we provide an upper bound of A_{T_ϵ} to give an explicit parameter setting of δ. The formal proof is deferred to Appendix C. We show the AIPE (Algorithm 1) can find an ϵ-solution to a convex function in $\mathcal{O}((\gamma/\epsilon)^{2/7})$ iterations. Then by incorporating the restart scheme we know Algorithm 2 has a linear convergence rate of $\mathcal{O}((\gamma/\mu)^{2/7} \log \epsilon^{-1})$ for minimizing a μ-uniformly convex function, as stated below.

Theorem 4.1 (AIPE-restart) *Assume $h(\boldsymbol{z}) : \mathcal{Z} \to \mathbb{R}$ is μ-uniformly convex. If $\delta \leq \mu\epsilon^4/(144D^2)$, then running Algorithm 1 with $T = \mathcal{O}\left((\gamma/\mu)^{2/7}\right)$ and $S = \mathcal{O}(\log(D/\epsilon))$ returns $\boldsymbol{z}^{(S)}$ such that $\|\boldsymbol{z}^{(S)} - \boldsymbol{z}^*\| \leq \epsilon$, where $\boldsymbol{z}^* = \arg\min_{\boldsymbol{z} \in \mathcal{Z}} h(\boldsymbol{z})$, $D = \sup_{\boldsymbol{z}, \boldsymbol{z}' \in \mathcal{Z}} \|\boldsymbol{z} - \boldsymbol{z}'\|$.*

Remark 1 *The inexact condition $\delta \lesssim \mu\epsilon^4/D^2$ in Theorem 4.1 can possibly be refined to $\delta \lesssim \mu\epsilon^4/d_0$ for $d_0 = \|\boldsymbol{z}_0 - \boldsymbol{z}^*\|$ by making some additional efforts such as (Bubeck et al., 2019a, Lemma 17) to show that all iterations of the algorithm lie in a bounded set $\{\boldsymbol{z} \in \mathbb{R}^d : \|\boldsymbol{z} - \boldsymbol{z}^*\| \leq \beta\|\boldsymbol{z}_0 - \boldsymbol{z}^*\|\}$ for some constant $\beta > 0$. But our inexact condition is enough to show the main result in this paper.*

As the $(0, \rho)$-proximal oracle can be implemented by the CRN oracle for a function with ρ-Lipschitz continuous Hessians, we can easily obtain the convergence result of the search-free ANPE method (Carmon et al., 2022a) under the restart scheme.

Corollary 4.1 (ANPE-restart) *Assume $h(\boldsymbol{z}) : \mathcal{Z} \to \mathbb{R}$ is μ-uniformly convex and has ρ-Lipschitz continuous Hessians. There exists a second-order algorithm that returns a point $\boldsymbol{z}$ such that $\|\boldsymbol{z} - \boldsymbol{z}^*\| \leq \epsilon$ in $\mathcal{O}\left((\rho/\mu)^{2/7} \log(D/\epsilon)\right)$ CRN oracle calls.*

5. Main Result: Achieving the $\tilde{\mathcal{O}}(\epsilon^{-4/7})$ Upper Bound

In this section, we introduce our acceleration framework and prove that it can find an ϵ-solution to a convex-concave minimax problems in $\tilde{\mathcal{O}}(\epsilon^{-4/7})$ second-order oracle calls. In Section 5.1, we first show that solving the convex-concave minimax problem can be reduced to solving a μ_x-uniformly-convex-μ_y-uniformly-concave minimax problems by adding cubic regularization. Then we show that Algorithm 3 and 4 together solve the problem fast if an inexact proximal minimax oracle (6) is available. We then describe the implementation of the oracle with existing algorithms such as NPE (Monteiro and Svaiter, 2012) and LEN (Chen et al., 2025a) in Algorithm 5.

5.1. Reducing to a Uniformly-Convex-Uniformly-Concave Problem

First of all, we apply the regularization trick to make the function uniformly-convex-uniformly-concave. The following lemma connects the approximate solution to the regularized problem and the original problem.

Lemma 5.1 *Let $\tilde{f}(\boldsymbol{x},\boldsymbol{y}) := f(\boldsymbol{x},\boldsymbol{y}) + \frac{\mu_x}{3}\|\boldsymbol{x}-\boldsymbol{x}_0\|^3 - \frac{\mu_y}{3}\|\boldsymbol{y}-\boldsymbol{y}_0\|^3$. If $(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}})$ is an $(\epsilon/3)$-solution to $\tilde{f}(\boldsymbol{x},\boldsymbol{y})$ with $\mu_x = \epsilon/(2D_x^3)$ and $\mu_y = \epsilon/(2D_y^3)$, then it is an ϵ-solution to Problem (1).*

After regularization, the objective becomes uniformly-convex-uniformly-concave. Therefore, we can reduce the problem to study the oracle complexity of second-order algorithms in finding an ϵ-solution to a function that satisfies the following assumption.

Assumption 5.1 *We suppose $f(\boldsymbol{x},\boldsymbol{y})$ is μ_x-uniformly-convex-μ_y-uniformly-concave, i.e., $f(\boldsymbol{x},\,\cdot\,)$ is μ_x-uniformly convex for any fixed $\boldsymbol{x}\in\mathbb{R}^{d_x}$ and $f(\,\cdot\,\boldsymbol{y})$ is μ_y-uniformly concave for any fixed $\boldsymbol{y}\in\mathbb{R}^{d_y}$, where $\mu_x,\mu_y>0$. We say $f(\boldsymbol{x},\boldsymbol{y})$ is convex-concave when $\mu_x=\mu_y=0$.*

Below, we present some useful lemmas which can be derived from Assumption 5.1.

Lemma 5.2 *Consider a function $f(\boldsymbol{x},\boldsymbol{y})$ that satisfies Assumption 5.1, then the primal function $\Phi(\boldsymbol{x}) = \max_{\boldsymbol{y}\in\mathcal{Y}} f(\boldsymbol{x},\boldsymbol{y})$ is μ_x-uniformly convex, and the dual function $\Psi(\boldsymbol{y}) = \min_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x},\boldsymbol{y})$ is μ_y-uniformly concave.*

Lemma 5.3 *Consider a function $f(\boldsymbol{x},\boldsymbol{y})$ that satisfies Assumption 3.4 and 5.1, let $\boldsymbol{x}^*(\boldsymbol{y}) = \arg\min_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x},\boldsymbol{y})$ and $\boldsymbol{y}^*(\boldsymbol{x}) = \arg\max_{\boldsymbol{y}\in\mathcal{Y}} f(\boldsymbol{x},\boldsymbol{y})$. Then both the mappings $\boldsymbol{x}^*(\boldsymbol{y})$ and $\boldsymbol{y}^*(\boldsymbol{x})$ are continuous. Furthermore, we have that*

$$\|\boldsymbol{y}^*(\boldsymbol{x}_1)-\boldsymbol{y}^*(\boldsymbol{x}_2)\|^2 \le \frac{\ell}{\mu_y}\|\boldsymbol{x}_1-\boldsymbol{x}_2\|, \quad \forall \boldsymbol{x}_1,\boldsymbol{x}_2\in\mathcal{X};$$
$$\|\boldsymbol{x}^*(\boldsymbol{y}_1)-\boldsymbol{x}^*(\boldsymbol{y}_2)\|^2 \le \frac{\ell}{\mu_x}\|\boldsymbol{y}_1-\boldsymbol{y}_2\|, \quad \forall \boldsymbol{y}_1,\boldsymbol{y}_2\in\mathcal{Y}.$$

Lemma 5.3 is crucial in our analysis but is not necessary for standard NPE analysis (Monteiro and Svaiter, 2012). For this reason, we additionally require the ℓ-smoothness in our algorithm, while NPE or its variants may not need. But our additional assumption is mild since the ℓ-smoothness always hold in a compact set if the function has Lipschitz continuous Hessians.

Algorithm 3 Minimax-AIPE

```
1: Run AIPE-restart (Algorithm 2) with proximal oracle given by Algorithm 4 to solve
       $$\min_{\boldsymbol{x}\in\mathcal{X}} \Phi(\boldsymbol{x})$$
   for finding $\hat{\boldsymbol{x}}$ such that $\|\hat{\boldsymbol{x}}-\boldsymbol{x}^*\| \le \zeta_1$, where $\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}\in\mathcal{X}} \Phi(\boldsymbol{x})$.
2: Run $\mathcal{M}_{\min}$ to solve
       $$\max_{\boldsymbol{y}\in\mathcal{Y}} f(\hat{\boldsymbol{x}},\,\cdot\,)$$
   for finding $\hat{\boldsymbol{y}}$ such that $\|\hat{\boldsymbol{y}}-\boldsymbol{y}^*(\hat{\boldsymbol{x}})\| \le \zeta_1$, where $\boldsymbol{y}^*(\hat{\boldsymbol{x}}) = \arg\max_{\boldsymbol{y}\in\mathcal{Y}} f(\hat{\boldsymbol{x}},\boldsymbol{y})$.
3: $(\boldsymbol{z}^{\rm out},\boldsymbol{c}^{\rm out}) \leftarrow {\rm EG}(\hat{\boldsymbol{z}},\boldsymbol{F},1/(2\ell))$
4: return $\boldsymbol{z}^{\rm out} = (\boldsymbol{x}^{\rm out},\boldsymbol{y}^{\rm out})$
```

5.2. Minimax-AIPE and its Convergence Analysis

We propose the Minimax-AIPE in Algorithm 3 for μ_x-strongly-concave-μ_y-strongly-concave minimax problems. Our algorithm is a general scheme to accelerate second-order minimax optimization. It can be applied on any linear convergent algorithms for uniformly convex minimization problems and uniformly-convex-uniformly-concave minimax problems. We denote such algorithms as $\mathcal{M}_{\rm min}$ and $\mathcal{M}_{\rm saddle}$, respectively. We assume that $\mathcal{M}_{\rm min}$ and $\mathcal{M}_{\rm saddle}$ have the following theoretical guarantee, which is generic and be satisfied by many existing algorithms.

Assumption 5.2 *Let $h(\boldsymbol{z}) : \mathcal{Z} \to \mathbb{R}$ be μ-uniformly convex and has ρ-Lipschitz continuous Hessians. We assume $\mathcal{M}_{\rm min}$ can find a point $\boldsymbol{z}$ such that $\|\boldsymbol{z} - \boldsymbol{z}^*\| \le \zeta$ in $T_{\rm min}(\rho, \mu) \log(d_0/\zeta)$ iterations, where $d_0 = \|\boldsymbol{z}_0 - \boldsymbol{z}^*\|$ and $\boldsymbol{z}^* = \arg\min_{\boldsymbol{z}\in\mathcal{Z}} h(\boldsymbol{z})$.*

Assumption 5.3 *Let $f(\boldsymbol{x}, \boldsymbol{y}) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ satisfies Assumption 3.5 and 5.1. We assume $\mathcal{M}_{\rm saddle}$ can find a point $\boldsymbol{z}$ such that $\|\boldsymbol{z} - \boldsymbol{z}^*\| \le \zeta$ in $T_{\rm saddle}(\rho, \mu) \log(d_0/\zeta)$ iterations, where $d_0 = \|\boldsymbol{z}_0 - \boldsymbol{z}^*\|$ and $\boldsymbol{z}^* = (\boldsymbol{x}^*, \boldsymbol{y}^*) = \arg\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y})$.*

The Minimax-AIPE is a triple-loop algorithm. Below, we introduce the procedures of each loop one by one. The outer loop (Algorithm 3) applies AIPE-restart (Algorithm 2) to minimize the primal objective $\Phi(\boldsymbol{x}) := \max_{\boldsymbol{y}\in\mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y})$, which requires the inexact zeroth-order oracles, first-order oracles, and second-order proximal oracles of $\Phi(\boldsymbol{x})$. As both the inexact zeroth-order and first-order oracle of $\Phi(\boldsymbol{x})$ are easily obtainable (see Theorem D.1), the non-trivial one is the proximal oracle, which will be implemented by the middle loop of Minimax-AIPE (Algorithm 4). If the middle loop can successfully return a proximal oracle, then the convergence of AIPE simply follows Theorem 4.1. Below, we show the required precision ζ_1 for Algorithm 3 to ensure an ϵ-solution to Problem (1).

Algorithm 4 $\mathrm{iProx}_\Phi(\bar{\boldsymbol{x}}, \gamma)$

```
Run AIPE-restart (Algorithm 2) with proximal oracle given by Algorithm 5 to solve
        $$\max_{\boldsymbol{y}\in\mathcal{Y}} \Psi(\boldsymbol{y}; \bar{\boldsymbol{x}})$$
    for finding $\hat{\boldsymbol{y}}$ such that $\|\hat{\boldsymbol{y}} - \boldsymbol{y}^*(\bar{\boldsymbol{x}})\| \le \zeta_2$, where $\boldsymbol{y}^*(\bar{\boldsymbol{x}}) = \arg\max_{\boldsymbol{y}\in\mathcal{Y}} \Psi(\boldsymbol{y}; \bar{\boldsymbol{x}})$.
Run $\mathcal{M}_{\rm min}$ to solve
        $$\min_{\boldsymbol{x}\in\mathcal{X}} g(\,\cdot\,, \hat{\boldsymbol{y}}; \bar{\boldsymbol{x}})$$
    for finding $\hat{\boldsymbol{x}}$ such that $\|\hat{\boldsymbol{x}} - \boldsymbol{x}^*(\hat{\boldsymbol{y}}; \bar{\boldsymbol{x}})\| \le \zeta_2$, where $\boldsymbol{x}^*(\hat{\boldsymbol{y}}; \bar{\boldsymbol{x}}) = \arg\min_{\boldsymbol{x}\in\mathcal{X}} g(\boldsymbol{x}, \hat{\boldsymbol{y}}; \bar{\boldsymbol{x}})$.
$(\boldsymbol{x}^{\rm out}, \boldsymbol{u}^{\rm out}) \leftarrow \mathrm{EG}(\hat{\boldsymbol{x}}, \nabla_x g(\boldsymbol{x}, \hat{\boldsymbol{y}}; \bar{\boldsymbol{x}}), 1/(2(\ell + 2\gamma D)))$.
return $(\boldsymbol{x}^{\rm out}, \boldsymbol{u}^{\rm out})$
```

Assumption 5.4 *Let ζ_1 be the precision in Algorithm 3. We assume that Algorithm 4 can return a (δ, γ)-proximal oracle of the primal objective $\Phi(\boldsymbol{x})$ with $\delta \le \mu_x \zeta_1^4/(144 D^2)$, where $\Phi(\boldsymbol{x}) := \max_{\boldsymbol{y}\in\mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y})$ and $D = \max\{D_x, D_y\}$.*

Theorem 5.1 (Outer-Loop Complexity) *Let $\zeta_1 \le \mu_y \epsilon^2/(147 \ell^3 D^2)$. Under Assumption 3.2, 3.4, 3.5, 5.1,5.2, and 5.4, Algorithm 3 can find an ϵ-solution to Problem (1) in $\mathcal{O}\left((\gamma/\mu_x)^{2/7} \log(D/\zeta_1)\right)$ calls of Algorithm 4, and $\mathcal{O}\left(T_{\rm min}(\rho, \mu_y) \log(D/\zeta_1)\right)$ iterations of $\mathcal{M}_{\rm min}$.*

Algorithm 5 $\mathrm{iProx}_{\Psi(\,\cdot\,;\bar{\boldsymbol{x}})}(\bar{\boldsymbol{y}}, \gamma)$

```
1: Run $\mathcal{M}_{\mathrm{saddle}}$ to solve
       $\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} h(\boldsymbol{x}, \boldsymbol{y}; \bar{\boldsymbol{x}}, \bar{\boldsymbol{y}})$
   for finding $\hat{\boldsymbol{z}}$ such that $\|\hat{\boldsymbol{z}} - \boldsymbol{z}^*(\bar{\boldsymbol{z}})\| \le \zeta_3$, where $\boldsymbol{z}^*(\bar{\boldsymbol{z}}) = \arg\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} h(\boldsymbol{x}, \boldsymbol{y}; \bar{\boldsymbol{x}}, \bar{\boldsymbol{y}})$.
2: $(\boldsymbol{z}^{\mathrm{out}}, \boldsymbol{c}^{\mathrm{out}}) \leftarrow \mathrm{EG}(\hat{\boldsymbol{z}}, \boldsymbol{F}, 1/(2(\ell + 2\gamma D))$.
3: $(\boldsymbol{x}^{\mathrm{out}}, \boldsymbol{y}^{\mathrm{out}}) = \boldsymbol{z}^{\mathrm{out}}$,  $(\boldsymbol{u}^{\mathrm{out}}, \boldsymbol{v}^{\mathrm{out}}) = \boldsymbol{c}^{\mathrm{out}}$
4: return $(\boldsymbol{y}^{\mathrm{out}}, \boldsymbol{v}^{\mathrm{out}})$
```

To introduce the middle loop of Minimax-AIPE, we denote several surrogate functions:

$$\begin{aligned} &\Phi(\boldsymbol{x}) := \max_{\boldsymbol{y}\in\mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y}), \quad g(\boldsymbol{x}, \boldsymbol{y}; \bar{\boldsymbol{x}}) := f(\boldsymbol{x}, \boldsymbol{y}) + \frac{\gamma}{3}\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|^3, \\ &\Psi(\boldsymbol{y}; \bar{\boldsymbol{x}}) := \min_{\boldsymbol{x}\in\mathcal{X}} g(\boldsymbol{x}, \boldsymbol{y}; \bar{\boldsymbol{x}}) \end{aligned} \tag{4}$$

The task of the middle loop of Minimax-AIPE (Algorithm 4) is to implement a (δ, γ)-proximal oracle for the primal objective $\Phi(\boldsymbol{x})$ such that $\delta \lesssim \mu_x \zeta_1^3 / D$ as required by Theorem 5.1. Note that

$$\mathrm{Prox}_{\Phi}(\bar{\boldsymbol{x}}, \gamma) = \arg\min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \Phi(\boldsymbol{x}) + \frac{\gamma}{3}\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|^3 \right\}$$

can be obtained by solving the equivalent subproblem

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} g(\boldsymbol{x}, \boldsymbol{y}; \bar{\boldsymbol{x}}) = \max_{\boldsymbol{y}\in\mathcal{Y}} \Psi(\boldsymbol{y}; \bar{\boldsymbol{x}}). \tag{5}$$

By Equation (5), Algorithm 4 applies AIPE-restart (Algorithm 2) to maximize $\Psi(\boldsymbol{y}; \bar{\boldsymbol{x}})$, whose proximal oracle is obtained by Algorithm 5. If Algorithm 5 can achieve the goal, then the following theorem shows that Algorithm 4 can successfully return desired oracles for $\Phi(\boldsymbol{x})$.

Assumption 5.5 *Let ζ_2 be the precision in Algorithm 4. Assume that Algorithm 5 returns a (δ, γ)-proximal oracle of the dual objective $\Psi(\boldsymbol{y}; \bar{\boldsymbol{x}})$ with $\delta \le \mu_y \zeta_2^4 / (144 D^2)$, where $\Psi(\boldsymbol{y}; \bar{\boldsymbol{x}}) := \min_{\boldsymbol{x}\in\mathcal{X}} g(\boldsymbol{x}, \boldsymbol{y}; \bar{\boldsymbol{x}})$ and $D = \max\{D_x, D_y\}$.*

Theorem 5.2 (Middle-Loop Complexity) *Let $\zeta_2 = \Omega(1/\mathrm{poly}(\rho, \ell, L, D, \gamma, \mu_x^{-1}, \mu_y^{-1}, \zeta_1^{-1}))$. Under Assumption 3.2, 3.3, 3.4, 3.5, 5.1, 5.2, and 5.5, Algorithm 4 returns a (δ, γ)-proximal oracle for $\Phi(\boldsymbol{x})$ that satisfies Assumption 5.4 ($\delta \lesssim \mu_x \zeta_1^4 / D^2$) in $\mathcal{O}\left((\gamma/\mu_y)^{2/7} \log(D/\zeta_2)\right)$ calls of Algorithm 5, and $T_{\min}(\rho + 2\gamma, \gamma/2) \log(D/\zeta_2)$ iterations of $\mathcal{M}_{\min}$. We can also obtain the δ- zeroth-order and first-order oracles for $\Phi(\boldsymbol{x})$ in $T_{\min}(\rho, \mu_y) \log(D/\zeta_2)$ iterations of $\mathcal{M}_{\min}$.*

Finally, the inner loop of Minimax-AIPE (Algorithm 5) implements a (δ, γ)-proximal oracle for the dual objective $\Psi(\boldsymbol{y}; \bar{\boldsymbol{x}})$. Note that

$$\mathrm{Prox}_{\Psi(\,\cdot\,;\bar{\boldsymbol{x}})}(\bar{\boldsymbol{y}}, \gamma) = \arg\max_{\boldsymbol{y}\in\mathcal{Y}} \left\{ \Psi(\boldsymbol{y}; \bar{\boldsymbol{x}}) - \frac{\gamma}{3}\|\boldsymbol{y} - \bar{\boldsymbol{y}}\|^3 \right\}$$

can be obtained by solving the equivalent subproblem

$$\min_{\boldsymbol{x}\in\mathcal{X}}\max_{\boldsymbol{y}\in\mathcal{Y}}\left\{h(\boldsymbol{x},\boldsymbol{y};\bar{\boldsymbol{x}},\bar{\boldsymbol{y}}):=f(\boldsymbol{x},\boldsymbol{y})+\frac{\gamma}{3}\|\boldsymbol{x}-\bar{\boldsymbol{x}}\|^3-\frac{\gamma}{3}\|\boldsymbol{y}-\bar{\boldsymbol{y}}\|^3\right\}. \tag{6}$$

It means we can apply a globally convergent algorithm $\mathcal{M}_{\text{saddle}}$ to find the saddle point of function $h(\boldsymbol{x},\boldsymbol{y};\bar{\boldsymbol{x}},\bar{\boldsymbol{y}})$ in Algorithm 5. After the above steps, we reduce the optimization of a μ_x-uniformly-convex-μ_y-uniformly-concave function (1) to the optimization of a $(\gamma/2)$-uniformly-convex-$(\gamma/2)$-uniformly-concave function (6). Since the latter has a better condition number, the new subproblem is significantly easier to optimize compared to the original problem. The following theorem states the complexity of Algorithm 5 to implement desired oracles for $\Psi(\boldsymbol{y};\bar{\boldsymbol{x}})$.

Theorem 5.3 (Inner-Loop Complexity) *Let $\zeta_3=\Omega(1/\text{poly}(\rho,\ell,L,D,\gamma,\mu_y^{-1},\zeta_2^{-1}))$. Under Assumption 3.2, 3.3, 3.4, 3.5, 5.1, and 5.3, Algorithm 5 returns a (δ,γ)-proximal oracle for $\Psi(\boldsymbol{y};\bar{\boldsymbol{x}})$ that satisfies Assumption 5.5 ($\delta\lesssim\mu_y\zeta_2^4/D^2$) in $T_{\text{saddle}}(\rho+2\gamma,\gamma/2)\log(D/\zeta_3)$ iterations of $\mathcal{M}_{\text{saddle}}$. We can also obtain the δ-zeroth-order and first-order oracles for $\Psi(\boldsymbol{y};\bar{\boldsymbol{x}})$ in $T_{\min}(\rho+2\gamma,\gamma/2)\log(D/\zeta_3)$ iterations of $\mathcal{M}_{\min}$.*

5.3. Main Result: Accelerate Existing Algorithms

From the analysis in the previous section (Theorem 5.1, 5.2 and 5.3), the total complexity of Minimax-AIPE is proportional to

$$\left(\frac{\gamma}{\mu_x}\right)^{2/7}\left(\frac{\gamma}{\mu_y}\right)^{2/7}T_{\text{saddle}}(\rho+2\gamma,\gamma/2)+\text{the cost involved in }T_{\min},$$

where $T_{\min}$ and T_{saddle} are the oracle complexity of the algorithm $\mathcal{M}_{\min}$ and $\mathcal{M}_{\text{saddle}}$ defined in Assumption 5.2 and 5.3. As minimization problems are typically easier to solve than minimax problems, in many practical scenarios the bottleneck of the complexity depends on the first term that involved in T_{saddle}, which is exactly the quantity that the best hyper-parameter γ should minimize. Below, we show the acceleration for existing methods with the optimal choice of γ.

Accelerating Newton Proximal Extragradient to Achieve the $\tilde{\mathcal{O}}(\epsilon^{-4/7})$ Complexity. If we take $\mathcal{M}_{\text{saddle}}$ = NPE-restart and $\mathcal{M}_{\min}$ = ANPE-restart, then it holds $T_{\text{saddle}}(\rho,\mu)=(\rho/\mu)^{2/3}$ by Theorem E.1 and $T_{\min}(\rho,\mu)=(\rho/\mu)^{2/7}$ by Corollary 4.1. Setting $\gamma=\rho$ and plugging in the above values of $T_{\min}$ and T_{saddle} gives an upper bound of $\tilde{\mathcal{O}}\left((\rho/\mu_x)^{2/7}(\rho/\mu_y)^{2/7}\right)$ under Assumption 5.1. Then by the reduction in Lemma 5.1, it indicates an upper bound of $\tilde{\mathcal{O}}\left((\rho D_x^3/\epsilon)^{2/7}(\rho D_y^3/\epsilon)^{2/7}\right)$ for finding an ϵ-solution, as stated in the following theorem.

Theorem 5.4 *Under Assumption 3.2, 3.3, 3.4, 3.5 and 5.1, Algorithm 3 with $\gamma=\rho$, $\mathcal{M}_{\min}$ = ANPE-restart, and $\mathcal{M}_{\text{saddle}}$ = NPE-restart finds an ϵ-solution to Problem (1) second-order oracle calls bounded by $\tilde{\mathcal{O}}\left((\rho/\mu_x)^{2/7}(\rho/\mu_y)^{2/7}\right)$.*

With the above analyses, our algorithm achieves the $\tilde{\mathcal{O}}(\epsilon^{-4/7})$ complexity for second-order convex-concave minimax optimization, and our result significantly improves the existing $\mathcal{O}(\epsilon^{-2/3})$ upper bound. The formal statements are presented as follows.

Theorem 5.5 (Main Theorem) *Under Assumption 3.1, 3.2, 3.3, 3.4, 3.5, using the reduction by Lemma 5.1, Algorithm 3, with $\gamma=\rho$ (defined in Hessian Lipchitzness), $\mathcal{M}_{\min}$ = ANPE-restart, and $\mathcal{M}_{\text{saddle}}$ = NPE-restart, can solve the ϵ-regularized minimax function and find an ϵ -solution to the problem (1) with second-order oracle calls bounded by $\tilde{\mathcal{O}}\left(D_x^{6/7}D_y^{6/7}(\rho/\epsilon)^{4/7}\right)$.*

Accelerating Lazy Extra Newton to Further Reduce the Computational Cost. We can further reduce the complexity by involving the idea of lazy Hessian updates (Doikov et al., 2023; Chen et al., 2025a). Specifically, we take $\mathcal{M}_{\text{saddle}}$ = LEN-restart and $\mathcal{M}_{\min}$ = ALEN-restart (Chen et al., 2025a). Then we know from Chen et al. (2025a) (see also Appendix E.2 for a brief review) that $T_{\text{saddle}}(\rho,\mu) = m + m^{2/3}(\rho/\mu)^{2/3}$ and $T_{\min}(\rho,\mu) = m + m^{5/7}(\rho/\mu)^{2/7}$. Setting $\gamma = \rho/\sqrt{m}$ in Minimax-AIPE yields an upper bound of $\tilde{\mathcal{O}}\left(m + m^{5/7}(\rho/\mu_x)^{2/7}(\rho/\mu_y)^{2/7}\right)$ under Assumption 5.1, which also indicates a corresponding $\tilde{\mathcal{O}}\left(m + m^{5/7}(\rho D_x^3/\epsilon)^{2/7}(\rho D_y^3/\epsilon)^{2/7}\right)$ upper bound for convex-concave minimax problems using the reduction by Lemma 5.1.

Theorem 5.6 *Under Assumption 3.2, 3.3, 3.4, 3.5 and 5.1, Algorithm 3 with $\gamma = \rho/\sqrt{m}$, $\mathcal{M}_{\min}$ = ALEN-restart, $\mathcal{M}_{\text{saddle}}$ = LEN-restart finds an ϵ-solution to Problem (1) with*

$$\tilde{\mathcal{O}}\left(m + m^{5/7}(\rho/\mu_x)^{2/7}(\rho/\mu_y)^{2/7}\right)$$

steps, and the ultimate algorithm only queries Hessians every m steps.

Then by applying similar arguments, we get the theorem below where we achieve accelerated rates than the original LEN method (Chen et al., 2025a).

Theorem 5.7 *Under Assumption 3.1, 3.2, 3.3, 3.4, 3.5, Algorithm 3 with $\gamma = \rho/\sqrt{m}$, $\mathcal{M}_{\min}$ = ALEN-restart, $\mathcal{M}_{\text{saddle}}$ = LEN-restart finds an ϵ-solution to Problem (1) with*

$$\tilde{\mathcal{O}}\left(m + m^{5/7}(D_x^3\rho/\epsilon)^{2/7}(D_y^3\rho/\epsilon)^{2/7}\right)$$

steps, and the ultimate algorithm only queries Hessians every m steps.

6. Conclusion and Future Works

In this paper, we propose the Minimax-AIPE algorithm for convex-concave minimax problems. Our theoretical result shows our proposed method can achieve the convergence rate of $\tilde{\mathcal{O}}(\epsilon^{-4/7})$, which refutes the common conjecture that the optimal rate for this setting is $\Theta(\epsilon^{-2/3})$. Our framework is also compatible with lazy Newton methods, and it yields an upper bound of $\tilde{\mathcal{O}}(m + m^{5/7}\epsilon^{-4/7})$ by reusing Hessians every m iterations. We discuss some potential future directions as follows.

Shaving Logarithmic Factors in our Upper Bound Our upper bound includes logarithmic factors is $\mathcal{O}(\epsilon^{-4/7}\log^3(\epsilon^{-1}))$. The logarithmic factors may be possibly removed using techniques to achieve optimal first-order oracle complexity in strongly-convex-strongly-concave problems (Kovalev and Gasnikov, 2022b).

More Acceleration Results Our framework can be applied to accelerate any existing linear convergent algorithms for uniformly-convex-uniformly-concave functions (Assumption 5.3). Potentially, it can be used to accelerate quasi-Newton methods (Jiang et al., 2023). But it requires additional analysis to adapt the result of (Jiang et al., 2023) from strongly-convex-strongly-concave functions to uniformly-convex-uniformly-concave cases. It is also possible to use our framework to accelerate stochastic algorithms (Chayti et al., 2024; Lin et al., 2022; Wang et al., 2019; Zhou et al., 2019; Tripuraneni et al., 2018) like in the original Catalyst method (Lin et al., 2018).

Lower Bounds It would also be important for future works to establish lower bounds for this problem. The key is to carefully design a zero-chain that couples the worse-case instances for both variables $\boldsymbol{x}$ and $\boldsymbol{y}$. Although tight lower bounds have been proved for first-order minimax optimization (Ouyang and Xu, 2021; Xie et al., 2020; Zhang et al., 2022; Li et al., 2021), this problem remains open for second-order methods.

Acknowledgment

Jingzhao Zhang is supported by National Key R&D Program of China 2024YFA1015800 and Shanghai Qi Zhi Institute Innovation Program. Luo Luo is supported by the Major Key Project of PCL under Grant PCL2024A06, the National Natural Science Foundation of China (No. 62206058), and Shanghai Basic Research Program (23JC1401000). Chengchang Liu is supported by the National Natural Science Foundation of China (624B2125).

References

Deeksha Adil, Brian Bullins, Arun Jambulapati, and Sushant Sachdeva. Optimal methods for higher-order smooth monotone variational inequalities. *arXiv preprint arXiv:2205.06167*, 2022.

Zeyuan Allen-Zhu. How to make the gradients small stochastically: Even faster convex and nonconvex sgd. In *NeurIPS*, 2018.

Yossi Arjevani, Ohad Shamir, and Ron Shiff. Oracle complexity of second-order methods for smooth convex optimization. *Mathematical Programming*, 178(1):327–360, 2019.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton university press, 2009.

Sébastien Bubeck, Qijia Jiang, Yin-Tat Lee, Yuanzhi Li, and Aaron Sidford. Complexity of highly parallel non-smooth convex optimization. In *NeurIPS*, 2019a.

Sébastien Bubeck, Qijia Jiang, Yin Tat Lee, Yuanzhi Li, and Aaron Sidford. Near-optimal method for highly smooth convex optimization. In *COLT*, 2019b.

Brian Bullins and Kevin A. Lai. Higher-order methods for convex-concave min-max optimization and monotone variational inequalities. *SIAM Journal on Optimization*, 32(3):2208–2229, 2022.

Yang Cai, Argyris Oikonomou, and Weiqiang Zheng. Finite-time last-iterate convergence for learning in multi-player games. In *NeurIPS*, 2022.

Yair Carmon and Danielle Hausler. Distributionally robust optimization via ball oracle acceleration. In *NeurIPS*, 2022.

Yair Carmon, Yujia Jin, Aaron Sidford, and Kevin Tian. Variance reduction for matrix games. In *NeurIPS*, 2019.

Yair Carmon, Arun Jambulapati, Qijia Jiang, Yujia Jin, Yin Tat Lee, Aaron Sidford, and Kevin Tian. Acceleration with a ball optimization oracle. In *NeurIPS*, 2020a.

Yair Carmon, Yujia Jin, Aaron Sidford, and Kevin Tian. Coordinate methods for matrix games. In *FOCS*, 2020b.

Yair Carmon, Arun Jambulapati, Yujia Jin, and Aaron Sidford. Thinking inside the ball: Near-optimal minimization of the maximal loss. In *COLT*, 2021.

Yair Carmon, Danielle Hausler, Arun Jambulapati, Yujia Jin, and Aaron Sidford. Optimal and adaptive monteiro-svaiter acceleration. In *NeurIPS*, 2022a.

Yair Carmon, Arun Jambulapati, Yujia Jin, and Aaron Sidford. RECAPP: Crafting a more efficient catalyst for convex optimization. In *ICML*, 2022b.

Yair Carmon, Arun Jambulapati, Yujia Jin, Yin Tat Lee, Daogao Liu, Aaron Sidford, and Kevin Tian. Resqueing parallel and private stochastic convex optimization. In *FOCS*, 2023.

Yair Carmon, Arun Jambulapati, Yujia Jin, and Aaron Sidford. A whole new ball game: A primal accelerated method for matrix games and minimizing the maximum of smooth functions. In *SODA*, 2024.

El Mahdi Chayti, Nikita Doikov, and Martin Jaggi. Unified convergence theory of stochastic and variance-reduced cubic newton methods. In *TMLR*, 2024.

Lesi Chen and Luo Luo. Near-optimal algorithms for making the gradient small in stochastic minimax optimization. *JMLR*, 25(387):1–44, 2024.

Lesi Chen, Chengchang Liu, Luo Luo, and Jingzhao Zhang. Computationally faster newton methods by lazy evaluations. *arXiv preprint arXiv:2501.17488*, 2025a.

Lesi Chen, Chengchang Liu, and Jingzhao Zhang. Second-order min-max optimization with lazy hessians. In *ICLR*, 2025b.

Sebastian Curi, Kfir Y Levy, Stefanie Jegelka, and Andreas Krause. Adaptive sampling for stochastic risk-averse learning. In *NeurIPS*, 2020.

Nikita Doikov, El Mahdi Chayti, and Martin Jaggi. Second-order optimization with lazy hessians. In *ICML*, 2023.

Dylan J. Foster, Ayush Sekhari, Ohad Shamir, Nathan Srebro, Karthik Sridharan, and Blake Woodworth. The complexity of making the gradient small in stochastic convex optimization. In *COLT*, 2019.

Alexander Gasnikov, Pavel Dvurechensky, Eduard Gorbunov, Evgeniya Vorontsova, Daniil Selikhanovych, César A Uribe, Bo Jiang, Haoyue Wang, Shuzhong Zhang, Sébastien Bubeck, et al. Near optimal methods for minimizing convex functions with lipschitz p-th derivatives. In *COLT*, 2019.

Kevin Huang and Shuzhong Zhang. An approximation-based regularized extra-gradient method for monotone variational inequalities. *arXiv preprint arXiv:2210.04440*, 2022.

Bo Jiang, Haoyue Wang, and Shuzhong Zhang. An optimal high-order tensor method for convex optimization. In *COLT*, 2019.

Ruichen Jiang and Aryan Mokhtari. Accelerated quasi-Newton proximal extragradient: Faster rate for smooth convex optimization. In *NeurIPS*, 2024.

Ruichen Jiang, Qiujiang Jin, and Aryan Mokhtari. Online learning guided curvature approximation: A quasi-newton method with global non-asymptotic superlinear convergence. In *COLT*, 2023.

Ruichen Jiang, Ali Kavis, Qiujiang Jin, Sujay Sanghavi, and Aryan Mokhtari. Adaptive and optimal second-order optimistic methods for minimax optimization. In *NeurIPS*, 2024.

Yujia Jin, Aaron Sidford, and Kevin Tian. Sharper rates for separable minimax and finite sum optimization via primal-dual extragradient methods. In *COLT*, 2022.

Anna R. Karlin and Yuval Peres. *Game theory, alive*, volume 101. American Mathematical Soc., 2017.

David Kinderlehrer and Guido Stampacchia. *An introduction to variational inequalities and their applications*. SIAM, 2000.

Guy Kornowski and Ohad Shamir. The oracle complexity of simplex-based matrix games: Linear separability and nash equilibria. *arXiv preprint arXiv:2412.06990*, 2024.

Galina M Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

Dmitry Kovalev and Alexander Gasnikov. The first optimal acceleration of high-order methods in smooth convex optimization. In *NeurIPS*, 2022a.

Dmitry Kovalev and Alexander Gasnikov. The first optimal algorithm for smooth and strongly-convex-strongly-concave minimax optimization. In *NeurIPS*, 2022b.

Dmitry Kovalev, Adil Salim, and Peter Richtárik. Optimal and practical algorithms for smooth and strongly convex decentralized optimization, 2020.

Sucheol Lee and Donghwan Kim. Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems. In *NeurIPS*, 2021.

Haochuan Li, Yi Tian, Jingzhao Zhang, and Ali Jadbabaie. Complexity lower bounds for nonconvex-strongly-concave min-max optimization. In *NeurIPS*, 2021.

Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. Catalyst acceleration for first-order convex optimization: from theory to practice. *JMLR*, 2018.

Tianyi Lin and Michael I. Jordan. Perseus: A simple high-order regularization method for variational inequalities. *Mathematical Programming*, pages 1–42, 2024.

Tianyi Lin, Chi Jin, and Michael I. Jordan. Near-optimal algorithms for minimax optimization. In *COLT*, 2020.

Tianyi Lin, Panayotis Mertikopoulos, and Michael I. Jordan. Explicit second-order min-max optimization methods with optimal convergence guarantee. *arXiv preprint arXiv:2210.12860*, 2022.

Chengchang Liu and Luo Luo. Regularized newton methods for monotone variational inequalities with h\" older continuous jacobians. *arXiv preprint arXiv:2212.07824*, 2022.

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *AISTATS*, 2020a.

Aryan Mokhtari, Asuman E. Ozdaglar, and Sarath Pattathil. Convergence rate of $\mathcal{O}(1/k)$ for optimistic gradient and extragradient methods in smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 30(4):3230–3251, 2020b.

Renato DC Monteiro and Benar Fux Svaiter. Iteration-complexity of a newton proximal extragradient method for monotone variational inequalities and inclusion problems. *SIAM Journal on Optimization*, 22(3):914–935, 2012.

Renato DC Monteiro and Benar Fux Svaiter. An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. *SIAM Journal on Optimization*, 23(2):1092–1125, 2013.

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

Yurii Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2-3):319–344, 2007.

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

Yurii Nesterov. Inexact high-order proximal-point methods with auxiliary search procedure. *SIAM Journal on Optimization*, 31(4):2807–2828, 2021.

Yurii Nesterov. High-order reduced-gradient methods for composite variational inequalities. *arXiv preprint arXiv:2311.15154*, 2023a.

Yurii Nesterov. Inexact accelerated high-order proximal-point methods. *Mathematical Programming*, pages 1–26, 2023b.

Yurii Nesterov and Boris T Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185(1):1–35, 2021.

Leonid Denisovich Popov. A modification of the arrow-hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 28:845–848, 1980.

Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *ICML*, 2017.

Chaobing Song, Cheuk Yin Lin, Stephen Wright, and Jelena Diakonikolas. Coordinate linear variance reduction for generalized linear programming. In *NeurIPS*, 2022.

Nilesh Tripuraneni, Mitchell Stern, Chi Jin, Jeffrey Regier, and Michael I. Jordan. Stochastic cubic regularization for fast nonconvex optimization. In *NeurIPS*, 2018.

John Von Neumann and Oskar Morgenstern. Theory of games and economic behavior, 2nd rev. 1947.

Yuanhao Wang and Jian Li. Improved algorithms for convex-concave minimax optimization, 2020.

Zhe Wang, Yi Zhou, Yingbin Liang, and Guanghui Lan. Stochastic variance-reduced cubic regularization for nonconvex optimization. In *AISTATS*, 2019.

Guangzeng Xie, Luo Luo, Yijiang Lian, and Zhihua Zhang. Lower complexity bounds for finite-sum convex-concave minimax optimization problems. In *ICML*, 2020.

Junchi Yang, Siqi Zhang, Negar Kiyavash, and Niao He. A catalyst framework for minimax optimization, 2020.

Yiming Ying, Longyin Wen, and Siwei Lyu. Stochastic online auc maximization. In *NeurIPS*, 2016.

TaeHo Yoon and Ernest K. Ryu. Accelerated algorithms for smooth convex-concave minimax problems with $\mathcal{O}(1/k^2)$ rate on squared gradient norm. In *ICML*, 2021.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

Junyu Zhang, Mingyi Hong, and Shuzhong Zhang. On lower iteration complexity bounds for the convex concave saddle point problems. *Mathematical Programming*, 194(1-2):901–935, 2022.

Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic variance-reduced cubic regularization methods. *JMLR*, 2019.

Appendix A. Discussions on the Existing $\Omega(\epsilon^{-2/3})$ Lower Bound

Adil et al. (2022) provided an $\Omega(\epsilon^{-2/3})$ lower bound on the second-order oracle of the primal function $\Phi(x) = \max_{\boldsymbol{y}\in\mathcal{Y}} f(\boldsymbol{x},\boldsymbol{y})$. However, practical algorithms (including our algorithm and Adil et al. (2022)'s algorithm) use information of $f(\boldsymbol{x},\boldsymbol{y})$ instead of $\Phi(\boldsymbol{x})$. It is clear that the oracle of $\Phi(\boldsymbol{x})$ is quite different from the oracle of $f(\boldsymbol{x},\boldsymbol{y})$.

Lin and Jordan (2024) adapted the hard instance of (Adil et al., 2022) to establish a lower bound with the second-order oracle of $f(\boldsymbol{x},\boldsymbol{y})$. They showed that there exists a ρ-Hessian smooth function, such that the output of any second-order method that queries no more than T second-order oracles of $f(\boldsymbol{x},\boldsymbol{y})$ must has duality gap of $\Omega(\rho D_x D_y^2/T^{3/2})$ (Lin and Jordan, 2024, Theorem 3.11). However, the diameters of sets $\mathcal{X}$ and $\mathcal{Y}$ in their construction do depend on T. Specifically, their hard instance requires taking $D_x = \Theta(T^{3/2})$ and $D_y = \Theta(T^{1/2})$. Hence, their theorem cannot directly give a lower bound $\Omega(\epsilon^{-2/3})$ when D_x and D_y are constants. Furthermore, we find that, unlike the case of first-order methods (Zhang et al., 2022), scaling down the diameters D_x and D_y in the hard instance of minimax problems is not easy, because the analysis heavily relies on their specific $\mathcal{X}$ and $\mathcal{Y}$. Therefore, the existing lower bounds do not apply to our algorithm.

Appendix B. Proofs in Section 4

B.1. Proof of Lemma 4.1

Proof We prove a general result that includes both minimax and minimization problems. From the first-order optimality condition of the CRN oracle (Definition 3.2), we know that $\boldsymbol{z} = \mathrm{CRN}(\bar{\boldsymbol{z}}, 2\rho)$ satisfies

$$\boldsymbol{u} := -\left(\boldsymbol{F}(\bar{\boldsymbol{z}}) + \nabla \boldsymbol{F}(\bar{\boldsymbol{z}})(\boldsymbol{z}-\bar{\boldsymbol{z}}) + \lambda(\boldsymbol{z}-\bar{\boldsymbol{z}})\right) \in \begin{bmatrix} \partial \mathcal{I}_{\mathcal{X}}(\boldsymbol{x}) \\ -\partial \mathcal{I}_{\mathcal{Y}}(\boldsymbol{y}) \end{bmatrix}, \tag{7}$$

where $\lambda = \rho\|\boldsymbol{z}-\bar{\boldsymbol{z}}\|$. Then using Assumption 3.5, we have that

$$\|\boldsymbol{F}(\boldsymbol{z}) + \boldsymbol{u} + \lambda(\boldsymbol{z}-\bar{\boldsymbol{z}})\| \le \|\boldsymbol{F}(\boldsymbol{z}) - \boldsymbol{F}(\bar{\boldsymbol{z}}) - \nabla \boldsymbol{F}(\bar{\boldsymbol{z}})(\boldsymbol{z}-\bar{\boldsymbol{z}})\| \le \frac{\rho}{2}\|\boldsymbol{z}-\bar{\boldsymbol{z}}\|^2, \tag{8}$$

which satisfies a $(0,\rho)$-proximal oracle according to Definition 4.1. ■

B.2. Proof of Theorem 4.1

Proof By Theorem B.1, each epoch of Algorithm 2 ensures $\|\boldsymbol{z}^{(s+1)} - \boldsymbol{z}^*\| \le \frac{1}{2}\|\boldsymbol{z}^{(s)} - \boldsymbol{z}^*\|$ if setting $T = \mathcal{O}\left((\gamma/\mu)^{2/7}\right)$. And therefore the algorithm finds a point $\boldsymbol{z}^{(S)}$ such that $\|\boldsymbol{z}^{(S)} - \boldsymbol{z}^*\| \le \epsilon$ in $S = \lceil \log_2(d_0/\epsilon)\rceil$ epochs. ■

Theorem B.1 *Under the same setting as Theorem 4.1, running Algorithm 1 finds a point $\boldsymbol{z}_t$ such that $\|\boldsymbol{z}_t - \boldsymbol{z}^*\| \le 2\epsilon$ in $T = \mathcal{O}\left(\left(\gamma d_0^3/(\mu\epsilon^3)\right)^{2/7}\right)$ iterations if $\delta \le \mu\epsilon^4/(144D^2)$, where $d_0 = \|\boldsymbol{z}_0 - \boldsymbol{z}^*\|$.*

Proof Some of our proof of error tolerance properties is motivated by the proof of (Bubeck et al., 2019a, Theorem 7). Compare to (Bubeck et al., 2019a), our analysis is more simple as we do not need to analyze the additional line search with inexact proximal oracles as (Bubeck et al., 2019a, Section E). Moreover, our analysis tackles the constrained case, but (Bubeck et al., 2019a) only considers the unconstrained case, *i.e.* $\mathcal{Z} = \mathbb{R}^d$. We define $E_t := h(\boldsymbol{z}_t) - h(\boldsymbol{z}^*)$, $R_t := \frac{1}{2}\|\boldsymbol{v}_t - \boldsymbol{z}^*\|^2$ and $N_{t+1} := \frac{1}{2}\|\tilde{\boldsymbol{z}}_{t+1} - \bar{\boldsymbol{z}}_t\|^2$. From Lemma 3.2, to find a point $\|\boldsymbol{z}_t - \boldsymbol{z}^*\| \le \epsilon$ it suffices to find $E_t \le \mu\epsilon^3/3$. Let $\mathrm{Proj}_{\mathcal{Z}}(\bar{\boldsymbol{z}}) = \arg\min_{\boldsymbol{z}\in\mathcal{Z}} \|\boldsymbol{z} - \bar{\boldsymbol{z}}\|$ be the projection operator of $\boldsymbol{z}$ onto the set $\mathcal{Z}$. The update rule of $\boldsymbol{v}_t$ gives

$$\begin{aligned} R_{t+1} &= \frac{1}{2}\left\|\mathrm{Proj}_{\mathcal{Z}}(\boldsymbol{v}_t - a_{t+1}(\boldsymbol{g}_{t+1} + \boldsymbol{u}_{t+1})) - \boldsymbol{z}^*\right\|^2 \\ &\le \frac{1}{2}\left\|(\boldsymbol{v}_t - a_{t+1}(\boldsymbol{g}_{t+1} + \boldsymbol{u}_{t+1})) - \boldsymbol{z}^*\right\|^2 \\ &= R_t + a_{t+1}\langle \boldsymbol{g}_{t+1} + \boldsymbol{u}_{t+1}, \boldsymbol{z}^* - \boldsymbol{v}_t\rangle + \frac{a_{t+1}^2}{2}\|\boldsymbol{g}_{t+1} + \boldsymbol{u}_{t+1}\|^2 \\ &\le R_t + a_{t+1}\langle \nabla h(\tilde{\boldsymbol{z}}_{t+1}) + u_{t+1}, \boldsymbol{z}^* - \boldsymbol{v}_t\rangle + a_{t+1}^2\|\nabla h(\tilde{\boldsymbol{z}}_{t+1}) + \boldsymbol{u}_{t+1}\|^2 \\ &\quad + a_{t+1}\delta\|\boldsymbol{v}_t - \boldsymbol{z}^*\| + a_{t+1}^2\delta^2. \end{aligned} \tag{9}$$

By the update rule of $\bar{\boldsymbol{z}}_t = \frac{A_t}{A'_{t+1}}\boldsymbol{z}_t + \frac{a'_{t+1}}{A'_{t+1}}\boldsymbol{v}_t$ and $A'_{t+1} = A_t + a'_{t+1}$, we have

$$a'_{t+1}\boldsymbol{v}_t = A'_{t+1}\bar{\boldsymbol{z}}_t - A_t\boldsymbol{z}_t = a'_{t+1}\tilde{\boldsymbol{z}}_{t+1} + A_t(\boldsymbol{z}_t - \tilde{\boldsymbol{z}}_{t+1}) + A'_{t+1}(\bar{\boldsymbol{z}}_t - \tilde{\boldsymbol{z}}_{t+1}).$$

Then subtracting $a'_{t+1}\boldsymbol{z}^*$ and taking inner product with $\nabla h(\tilde{\boldsymbol{z}}_{t+1}) + \boldsymbol{u}_{t+1}$ yields

$$\begin{aligned} &a'_{t+1}\langle \nabla h(\tilde{\boldsymbol{z}}_{t+1}) + \boldsymbol{u}_{t+1}, \boldsymbol{z}^* - \boldsymbol{v}_t\rangle \\ &= \langle \nabla h(\tilde{\boldsymbol{z}}_{t+1}) + \boldsymbol{u}_{t+1}, a'_{t+1}(\boldsymbol{z}^* - \tilde{\boldsymbol{z}}_{t+1}) + A_t(\boldsymbol{z}_t - \tilde{\boldsymbol{z}}_{t+1}) + A'_{t+1}(\tilde{\boldsymbol{z}}_{t+1} - \bar{\boldsymbol{z}}_t)\rangle \\ &\le a'_{t+1}(h(\boldsymbol{z}^*) - h(\tilde{\boldsymbol{z}}_{t+1})) + A_t(h(\boldsymbol{z}_t) - h(\tilde{\boldsymbol{z}}_{t+1})) + A'_{t+1}\langle \nabla h(\tilde{\boldsymbol{z}}_{t+1}) + \boldsymbol{u}_{t+1}, \tilde{\boldsymbol{z}}_{t+1} - \bar{\boldsymbol{z}}_t\rangle, \end{aligned} \tag{10}$$

where in the last step we use the convexity of h and the fact that $\boldsymbol{u}_{t+1} \in \partial\mathcal{I}_{\mathcal{Z}}(\tilde{\boldsymbol{z}}_{t+1})$. We continue to upper bound the inner product term with

$$\begin{aligned} &\langle \nabla h(\tilde{\boldsymbol{z}}_{t+1}) + \boldsymbol{u}_{t+1}, \tilde{\boldsymbol{z}}_{t+1} - \bar{\boldsymbol{z}}_t\rangle \\ &= \frac{1}{2\lambda_{t+1}}\|\nabla h(\tilde{\boldsymbol{z}}_{t+1}) + \boldsymbol{u}_{t+1} + \lambda_{t+1}(\tilde{\boldsymbol{z}}_{t+1} - \bar{\boldsymbol{z}}_t)\|^2 \\ &\quad - \frac{1}{2\lambda_{t+1}}\|\nabla h(\tilde{\boldsymbol{z}}_{t+1}) + \boldsymbol{u}_{t+1}\|^2 - \frac{\lambda_{t+1}}{2}\|\tilde{\boldsymbol{z}}_{t+1} - \bar{\boldsymbol{z}}_t\|^2. \end{aligned}$$

Substituting this upper bound into Eq. (10) yields

$$\begin{aligned} &a'_{t+1}\langle \nabla h(\tilde{\boldsymbol{z}}_{t+1}) + \boldsymbol{u}_{t+1}, \boldsymbol{z}^* - \boldsymbol{v}_t\rangle \\ &\le a'_{t+1}(h(\boldsymbol{z}^*) - h(\tilde{\boldsymbol{z}}_{t+1})) + A_t(h(\boldsymbol{z}_t) - h(\tilde{\boldsymbol{z}}_{t+1})) - \frac{A'_{t+1}}{2\lambda_{t+1}}\|\nabla h(\tilde{\boldsymbol{z}}_{t+1}) + \boldsymbol{u}_{t+1}\|^2 \\ &\quad + \frac{A'_{t+1}}{2\lambda_{t+1}}\|\nabla h(\tilde{\boldsymbol{z}}_{t+1}) + \boldsymbol{u}_{t+1} + \lambda_{t+1}(\tilde{\boldsymbol{z}}_{t+1} - \bar{\boldsymbol{z}}_t)\|^2 - \frac{A'_{t+1}\lambda_{t+1}}{2}\|\tilde{\boldsymbol{z}}_{t+1} - \bar{\boldsymbol{z}}_t\|^2. \end{aligned}$$

Furthermore, by Definition 4.1, we have that

$$\begin{aligned} &a'_{t+1}\langle \nabla h(\tilde{\boldsymbol{z}}_{t+1}) + \boldsymbol{u}_{t+1}, \boldsymbol{z}^* - \boldsymbol{v}_t\rangle \\ &\le a'_{t+1}(h(\boldsymbol{z}^*) - f(\tilde{\boldsymbol{z}}_{t+1})) + A_t(h(\boldsymbol{z}_t) - h(\tilde{\boldsymbol{z}}_{t+1})) \\ &\quad - \frac{A'_{t+1}}{2\lambda_{t+1}}\|\nabla h(\tilde{\boldsymbol{z}}_{t+1}) + \boldsymbol{u}_{t+1}\|^2 - \frac{3}{4}\lambda_{t+1}A'_{t+1}N_{t+1} + \frac{A'_{t+1}}{2\lambda_{t+1}}\delta^2. \end{aligned} \tag{11}$$

Now we separately consider the two cases $\lambda_{t+1} \le \lambda'_{t+1}$ (*Case I*) and $\lambda_{t+1} > \lambda'_{t+1}$ (*Case II*). In the first case, we have that $\boldsymbol{z}_{t+1} = \tilde{\boldsymbol{z}}_{t+1}$, $a_{t+1} = a'_{t+1}$, $A_{t+1} = A'_{t+1}$ and $A_{t+1} = 2\lambda'_{t+1}a_{t+1}^2$. Then we can directly sum up Eq. (9) and Eq. (11) to obtain that

$$\begin{aligned} \textit{Case I:} \quad & A_{t+1}E_{t+1} + D_{t+1} + \frac{3}{4}\lambda_{t+1}A_{t+1}N_{t+1} \\ & \le A_t E_t + D_t + a_{t+1}\delta\|\boldsymbol{v}_t - \boldsymbol{z}^*\| + 2a_{t+1}^2\delta^2. \end{aligned}$$

In the second case, we have that $A_{t+1} = (1-\gamma_{t+1})A_t + \gamma_{t+1}A'_{t+1}$ with $\gamma_{t+1} = \lambda'_{t+1}/\lambda_{t+1}$. By the convexity of h, we have that

$$A_{t+1}E_{t+1} \le (1-\gamma_{t+1})A_t E_t + \gamma_{t+1}A'_{t+1}(h(\tilde{\boldsymbol{z}}_{t+1}) - h(\boldsymbol{z}^*)).$$

We use Eq. (11) multiplied by γ_{t+1} to upper bounding the above inequality as

$$\begin{aligned} A_{t+1}E_{t+1} \le\ & A_t E_t + a_{t+1}\langle \nabla h(\tilde{\boldsymbol{z}}_{t+1}) + \boldsymbol{u}_{t+1}, \boldsymbol{z}^* - \boldsymbol{v}_t\rangle \\ & - \frac{3}{4}\lambda'_{t+1}A'_{t+1}N_{t+1} - \frac{\gamma_{t+1}A'_{t+1}}{2\lambda_{t+1}}\|\nabla h(\tilde{\boldsymbol{z}}_{t+1}) + \boldsymbol{u}_{t+1}\|^2 + \frac{\gamma_{t+1}A'_{t+1}}{2\lambda_{t+1}}\delta^2, \end{aligned} \tag{12}$$

where we also use facts $\gamma_{t+1}a'_{t+1} = a_{t+1}$ and $\lambda'_{t+1} = \gamma_{t+1}\lambda_{t+1}$. Note that

$$\frac{\gamma_{t+1}A'_{t+1}}{2\lambda_{t+1}} = \frac{\gamma_{t+1}^2 A'_{t+1}}{\lambda'_{t+1}} = (\gamma_{t+1}a'_{t+1}) = a_{t+1}^2.$$

Summing up Eq. (9) and Eq. (12) yields that

$$\begin{aligned} \textit{Case II:} \quad & A_{t+1}E_{t+1} + D_{t+1} + \frac{3}{4}\lambda'_{t+1}A'_{t+1}N_{t+1} \\ & \le A_t E_t + D_t + a_{t+1}\delta\|\boldsymbol{v}_t - \boldsymbol{z}^*\| + 2a_{t+1}^2\delta^2. \end{aligned}$$

The inequality for both Case I and Case II can be unified as

$$A_{t+1}E_{t+1} + R_{t+1} + \frac{3}{4}\min\{\lambda_{t+1}, \lambda'_{t+1}\}A'_{t+1}N_{t+1} \le A_t E_t + R_t + \delta_{t+1}, \tag{13}$$

where $\delta_{t+1} := a_{t+1}\delta\|\boldsymbol{v}_t - \boldsymbol{z}^*\| + 2a_{t+1}^2\delta^2$. Let T_ϵ be the first time that the algorithm achieves a point $\boldsymbol{z} \in \{\boldsymbol{z}_t, \tilde{\boldsymbol{z}}_t\}$ such that $h(\boldsymbol{z}) - h(\boldsymbol{z}^*) \le \mu\epsilon^3/3$, *i.e.* $T_\epsilon = \arg\min_{t\ge 0}\{h(\boldsymbol{z}) - h(\boldsymbol{z}^*) \le \mu\epsilon^3/3,\ \boldsymbol{z} \in \{\boldsymbol{z}_t, \tilde{\boldsymbol{z}}_t\}\}$. We let $A_{T_\epsilon}\delta \le c\|\boldsymbol{z}_0 - \boldsymbol{z}^*\|$ for some constant $c > 0$, and we will specify the value of c later. From Eq. (13) and $A_t = \sum_{j=0}^t a_t$ we know that

$$\frac{1}{2}\|\boldsymbol{v}_t - \boldsymbol{z}^*\|^2 \le \frac{1}{2}\|\boldsymbol{z}_0 - \boldsymbol{z}^*\|^2 + c\|\boldsymbol{z}_0 - \boldsymbol{z}^*\|\max_{0\le j\le t}\|\boldsymbol{v}_j - \boldsymbol{z}^*\| + 2c^2\|\boldsymbol{z}_0 - \boldsymbol{z}^*\|^2.$$

Since the above inequality holds for any $0 \le t \le T$, it must hold for $t_m \in \arg\max_{0\le j\le T}\|\boldsymbol{v}_j - \boldsymbol{z}^*\|$. Therefore, we have that

$$\|v_{t_m} - \boldsymbol{z}^*\|^2 \le (1+4c^2)\|\boldsymbol{z}_0 - \boldsymbol{z}^*\|^2 + 2c\|\boldsymbol{z}_0 - \boldsymbol{z}^*\|\|\boldsymbol{v}_{t_m} - \boldsymbol{z}^*\|.$$

Solving the above quadratic with respect to variable $\|v_{t_m} - \boldsymbol{z}^*\|$ yields

$$\|v_{t_m} - \boldsymbol{z}^*\| \le \left(2c + \sqrt{1+5c^2}\right)\|\boldsymbol{z}_0 - \boldsymbol{z}^*\| \le (1+5c)\|\boldsymbol{z}_0 - \boldsymbol{z}^*\|.$$

It implies

$$\sum_{j=0}^{t} \delta_{t+1} \leq A_{t+1}\delta\|v_{t_m} - \boldsymbol{z}^*\| + 2A_{t+1}^2\delta^2 \leq c(1+7c)\|\boldsymbol{z}_0 - \boldsymbol{z}^*\|^2 \tag{14}$$

Note that $A_0 = 0$. We telescope Eq. (13) over $t = 0, 1, \cdots, T_\epsilon - 1$ and substitute Eq. (14) to get

$$A_{T_\epsilon} E_{T_\epsilon} + R_{T_\epsilon} + \sum_{t=0}^{T_\epsilon - 1} \min\{\lambda_{t+1}, \lambda'_{t+1}\} A'_{t+1} N_{t+1} \leq R_0 + c(1+7c)\|\boldsymbol{z}_0 - \boldsymbol{z}^*\|^2.$$

We let $c \leq 1/8$ to have

$$A_{T_\epsilon} E_{T_\epsilon} + R_{T_\epsilon} + \sum_{t=0}^{T_\epsilon - 1} \min\{\lambda_{t+1}, \lambda'_{t+1}\} A'_{t+1} N_{t+1} \leq 2R_0. \tag{15}$$

Note that Eq. (15) is exactly (Carmon et al., 2022a, Eq. 9) with their R_0 now being replaced by $2R_0$. We can then follow the remaining steps as the proof of (Carmon et al., 2022a, Theorem 1) to show that the algorithm finds $E_t \leq \mu\epsilon^3/3$ in $T_\epsilon = \mathcal{O}\left(\left(\gamma d_0^3/(\mu\epsilon^3)\right)^{2/7}\right)$ iterations, where $d_0 = \|\boldsymbol{z}_0 - \boldsymbol{z}^*\|$. Therefore, we have that $\min\{\boldsymbol{z}_{T_\epsilon}, \bar{\boldsymbol{z}}_{T_\epsilon}\}\} \leq \mu\epsilon^3/3$. This further implies

$$h(\boldsymbol{z}^{\text{out}}) \leq \tilde{h}(\boldsymbol{z}^{\text{out}}) + \delta \leq \tilde{h}(\boldsymbol{z}_{T_\epsilon}) + \delta \leq h(\boldsymbol{z}_{T_\epsilon}) + 2\delta.$$

If $2\delta \leq \mu\epsilon^3/3$ and we know from $E_t \leq \mu\epsilon^3/3$ that $h(\boldsymbol{z}_{\text{out}}) - h(\boldsymbol{z}^*) \leq 2\mu\epsilon^3/3$. By Lemma 3.2, this implies that $\|\boldsymbol{z}^{\text{out}} - \boldsymbol{z}^*\| \leq 2\epsilon$. The last thing to show how to fulfill the goal of $\delta \leq d_0/(8A_{T_\epsilon})$ to ensure the above convergence rate. This can be done by giving a uniform lower bound of d_0 and a uniform upper bound of A_{T_ϵ}, as we specify below. As for d_0, we use the trivial bound $d_0 \geq \epsilon$. As for A_{T_ϵ}, we analyze its upper bound below.

From Eq. (15) we know that for all $t < T_\epsilon$ we have that $\mu\epsilon^3/3 < E_t \leq 2R_0/A_t$, which means that $A_t < 6R_0/(\mu\epsilon^3)$ for all $t < T_\epsilon$. The update of $A'_{t+1} = 2\lambda'_{t+1}\left(a'_{t+1}\right)^2$, in conjunction with $R_0 \leq D^2/2$, means that

$$a'_t = \frac{1 + \sqrt{1 + 8\lambda'_t A_{t-1}}}{4\lambda'_{t+1}} \leq \frac{1}{2\lambda'_t} + \frac{1}{2}\sqrt{\frac{2A_{t-1}}{\lambda'_t}} < \frac{1}{2\lambda'_t} + \frac{D}{2}\sqrt{\frac{6}{\mu\epsilon^3\lambda'_t}}. \tag{16}$$

Therefore, we need a lower bound of λ'_t. As an intermediate goal, we first analyze λ_t. According to Definition 4.1 we have that

$$\frac{3\lambda_t}{2}\|\tilde{\boldsymbol{z}}_t - \bar{\boldsymbol{z}}_{t-1}\| \geq \|\nabla h(\tilde{\boldsymbol{z}}_t) + \boldsymbol{u}_t\| - \delta \geq \frac{h(\tilde{\boldsymbol{z}}_t) - h(\boldsymbol{z}^*)}{\|\tilde{\boldsymbol{z}}_t - \boldsymbol{z}^*\|} - \delta.$$

By the definition of T_ϵ, we know that for all $t < T_\epsilon$, when $\delta \leq \mu\epsilon^3/(6D)$ we have

$$\lambda_t \geq \frac{2}{3\|\tilde{\boldsymbol{z}}_t - \bar{\boldsymbol{z}}_{t-1}\|}\left(\frac{h(\tilde{\boldsymbol{z}}_t) - h(\boldsymbol{z}^*)}{\|\tilde{\boldsymbol{z}}_t - \boldsymbol{z}^*\|} - \delta\right) \geq \frac{\mu\epsilon^3}{9D^2}. \tag{17}$$

Let k the the last iterate before t such that $\lambda'_k > \lambda_k$, *i.e.*, $k = \arg\max_{0\le k'\le t}\{\lambda'_{k'} > \lambda_{k'}\}$. If $t = k$, then Eq. (17) already gives a lower bound of λ'_t such that $\lambda'_t > \lambda_t \ge \mu\epsilon^3/(9D^2)$. Otherwise, if $t < k$, we know from the update rule of λ'_t that

$$\lambda'_t = 2^{t-k-1}\lambda'_{k+1} = 2^{t-k-2}\lambda'_k > \lambda_k/2 \ge \frac{\mu\epsilon^3}{18D^2}.$$

Now, if $\lambda'_{T_\epsilon-1} > \lambda_{T_\epsilon-1}$, then

$$\lambda'_{T_\epsilon} = \frac{1}{2}\lambda'_{T_\epsilon-1} > \frac{\mu\epsilon^3}{18D^2}.$$

Else, if $\lambda'_{T_\epsilon-1} \le \lambda_{T_\epsilon-1}$, then

$$\lambda'_{T_\epsilon} = 2\lambda'_{T_\epsilon-1} > \frac{\mu\epsilon^3}{9D^2}.$$

For both cases, plugging the above inequalities into Eq. (16) yields that

$$a'_{T_\epsilon} < \frac{(9+3\sqrt{3})D^2}{\mu\epsilon^3}.$$

Therefore,

$$A_{T_\epsilon} = A_{T_\epsilon-1} + a_{T_\epsilon} \le A_{T_\epsilon-1} + a'_{T_\epsilon} \le \frac{(12+3\sqrt{3})D^2}{\mu\epsilon^3}.$$

Hence, we can ensure $\delta \le d_0/(8A_{T_\epsilon})$ by setting $\delta \le \mu\epsilon^4/(144D^2)$. ■

Appendix C. Proofs in Section 5.1

Lemma C.1 *Let $f : \mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ be a μ-uniformly-convex-μ-uniformly concave function. We have*

$$\langle \boldsymbol{g}_1 - \boldsymbol{g}_2, \boldsymbol{z}_1 - \boldsymbol{z}_2\rangle \ge \frac{2\mu}{3}\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|^3, \tag{18}$$

for any $\boldsymbol{g}_1 \in \boldsymbol{A}(\boldsymbol{z}_1)$, $\boldsymbol{g}_2 \in \boldsymbol{A}(\boldsymbol{z}_2)$, where

$$\boldsymbol{A}(\boldsymbol{z}) = \begin{bmatrix}\nabla_x f(\boldsymbol{x},\boldsymbol{y})\\ -\nabla_y f(\boldsymbol{x},\boldsymbol{y})\end{bmatrix} + \begin{bmatrix}\partial\mathcal{I}_{\mathcal{X}}(\boldsymbol{x})\\ \partial\mathcal{I}_{\mathcal{Y}}(\boldsymbol{y})\end{bmatrix}.$$

Proof For any $(\boldsymbol{x}_1,\boldsymbol{y}_2), (\boldsymbol{x}_2,\boldsymbol{y}_2) \in \mathcal{X}\times\mathcal{Y}$,

$$\begin{aligned} f(\boldsymbol{x}_2,\boldsymbol{y}_1) - f(\boldsymbol{x}_1,\boldsymbol{y}_1) &\ge \langle \nabla_x f(\boldsymbol{x}_1,\boldsymbol{y}_1) + \boldsymbol{u}_1, \boldsymbol{x}_2 - \boldsymbol{x}_1\rangle + \frac{\mu}{p}\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|^p, \quad \boldsymbol{u}_1 \in \partial\mathcal{I}_{\mathcal{X}}(\boldsymbol{x}_1);\\ f(\boldsymbol{x}_1,\boldsymbol{y}_2) - f(\boldsymbol{x}_2,\boldsymbol{y}_2) &\ge \langle \nabla_x f(\boldsymbol{x}_2,\boldsymbol{y}_2) + \boldsymbol{u}_2, \boldsymbol{x}_1 - \boldsymbol{x}_2\rangle + \frac{\mu}{p}\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|^p, \quad \boldsymbol{u}_2 \in \partial\mathcal{I}_{\mathcal{X}}(\boldsymbol{x}_2). \end{aligned}$$

and

$$\begin{aligned} -f(\boldsymbol{x}_1,\boldsymbol{y}_2) + f(\boldsymbol{x}_1,\boldsymbol{y}_1) &\ge \langle -\nabla_y f(\boldsymbol{x}_1,\boldsymbol{y}_1) + \boldsymbol{v}_1, \boldsymbol{y}_2 - \boldsymbol{y}_1\rangle + \frac{\mu}{p}\|\boldsymbol{y}_2 - \boldsymbol{y}_1\|^p, \quad \boldsymbol{v}_1 \in \partial\mathcal{I}_{\mathcal{Y}}(\boldsymbol{y}_1);\\ -f(\boldsymbol{x}_2,\boldsymbol{y}_1) + f(\boldsymbol{x}_2,\boldsymbol{y}_2) &\ge \langle -\nabla_y f(\boldsymbol{x}_2,\boldsymbol{y}_2) + \boldsymbol{v}_2, \boldsymbol{y}_1 - \boldsymbol{y}_2\rangle + \frac{\mu}{p}\|\boldsymbol{y}_2 - \boldsymbol{y}_1\|^p, \quad \boldsymbol{v}_2 \in \partial\mathcal{I}_{\mathcal{Y}}(\boldsymbol{y}_2). \end{aligned}$$

Summing up the above four equations yields Eq. (18).

■

C.1. Proof of Lemma 5.1

By the definition of $\tilde{f}$, for every $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{y} \in \mathcal{Y}$, it holds that $|\tilde{f}(\boldsymbol{x}, \boldsymbol{y}) - f(\boldsymbol{x}, \boldsymbol{y})| \leq \epsilon/3$. Then

$$\max_{\boldsymbol{y} \in \mathcal{Y}} f(\hat{\boldsymbol{x}}, \boldsymbol{y}) \leq \max_{\boldsymbol{y} \in \mathcal{Y}} \tilde{f}(\hat{\boldsymbol{x}}, \boldsymbol{y}) + \epsilon/3$$
$$\min_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}, \hat{\boldsymbol{y}}) \geq \min_{\boldsymbol{x} \in \mathcal{X}} \tilde{f}(\boldsymbol{x}, \hat{\boldsymbol{y}}) - \epsilon/3.$$

If $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$ is an $(\epsilon/3)$-solution to the regularized function $\tilde{f}$, *i.e.*,

$$\max_{\boldsymbol{y} \in \mathcal{Y}} \tilde{f}(\boldsymbol{x}, \boldsymbol{y}) - \min_{\boldsymbol{x} \in \mathcal{X}} \tilde{f}(\boldsymbol{x}, \hat{\boldsymbol{y}}) \leq \epsilon/3,$$

then we can conclude that

$$\max_{\boldsymbol{y} \in \mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y}) - \min_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}, \hat{\boldsymbol{y}}) \leq \epsilon.$$

C.2. Proof of Lemma 5.2

Proof Let $y^*(\boldsymbol{x}) = \arg\max_{\boldsymbol{y} \in \mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y})$. For any $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$, we have that

$$\begin{aligned}
&\Phi(\boldsymbol{x}_1) - \Phi(\boldsymbol{x}_2) - \langle \Phi(\boldsymbol{x}_2), \boldsymbol{x}_1 - \boldsymbol{x}_2 \rangle \\
&= f(\boldsymbol{x}_1, \boldsymbol{y}^*(\boldsymbol{x}_1)) - f(\boldsymbol{x}_2, \boldsymbol{y}^*(\boldsymbol{x}_2)) - \langle \nabla_x f(\boldsymbol{x}_2, \boldsymbol{y}^*(\boldsymbol{x}_2), \boldsymbol{x}_1 - \boldsymbol{x}_2 \rangle \\
&\geq f(\boldsymbol{x}_1, \boldsymbol{y}^*(\boldsymbol{x}_2)) - f(\boldsymbol{x}_2, \boldsymbol{y}^*(\boldsymbol{x}_2)) - \langle \nabla_x f(\boldsymbol{x}_2, \boldsymbol{y}^*(\boldsymbol{x}_2), \boldsymbol{x}_1 - \boldsymbol{x}_2 \rangle \\
&\geq \frac{\mu_x}{3} \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^3,
\end{aligned}$$

where the last step uses the fact that $f(\,\cdot\,, \boldsymbol{y})$ is μ_x-uniformly convex. This is exactly the μ_x-uniform convexity by definition. Repeating the same analysis in variable $\boldsymbol{y}$ shows that $\Psi(\boldsymbol{y})$ is μ_y-uniformly concave. ■

C.3. Proof of Lemma 5.3

Proof The proof is similar to (Lin et al., 2020, Lemma B.2). Let $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$, from the optimality condition, we have that

$$\langle \nabla_y f(\boldsymbol{x}_1, \boldsymbol{y}^*(\boldsymbol{x}_1), \boldsymbol{y} - \boldsymbol{y}^*(\boldsymbol{x}_1) \rangle \leq 0, \quad \forall \boldsymbol{y} \in \mathcal{Y};$$
$$\langle \nabla_y f(\boldsymbol{x}_2, \boldsymbol{y}^*(\boldsymbol{x}_2), \boldsymbol{y}' - \boldsymbol{y}^*(\boldsymbol{x}_2) \rangle \leq 0, \quad \forall \boldsymbol{y}' \in \mathcal{Y}.$$

Plugging $\boldsymbol{y} = \boldsymbol{y}^*(\boldsymbol{x}_2)$ and $\boldsymbol{y}' = \boldsymbol{y}^*(\boldsymbol{x}_1)$ into the above inequalities and summing them up gives

$$\langle \nabla_y f(\boldsymbol{x}_1, \boldsymbol{y}^*(\boldsymbol{x}_1) - \nabla_y f(\boldsymbol{x}_2, \boldsymbol{y}^*(\boldsymbol{x}_2)), \boldsymbol{y}^*(\boldsymbol{x}_2) - \boldsymbol{y}^*(\boldsymbol{x}_1) \rangle \leq 0.$$

By the μ_y-uniform concavity in $\boldsymbol{y}$, we know from of (Nesterov, 2018, Eq. (4.2.12)) that

$$\langle \nabla_y f(\boldsymbol{x}_1, \boldsymbol{y}^*(\boldsymbol{x}_2)) - \nabla_y f(\boldsymbol{x}_1, \boldsymbol{y}^*(\boldsymbol{x}_1)), \boldsymbol{y}^*(\boldsymbol{x}_2) - \boldsymbol{y}^*(\boldsymbol{x}_1) \rangle + \mu_y \|\boldsymbol{y}^*(\boldsymbol{x}_2) - \boldsymbol{y}^*(\boldsymbol{x}_1)\|^3 \leq 0.$$

Now we can sum up the above inequality to obtain that

$$\langle \nabla_y f(\boldsymbol{x}_1, \boldsymbol{y}^*(\boldsymbol{x}_2)) - \nabla_y f(\boldsymbol{x}_2, \boldsymbol{y}^*(\boldsymbol{x}_2)), \boldsymbol{y}^*(\boldsymbol{x}_2) - \boldsymbol{y}^*(\boldsymbol{x}_1) \rangle + \mu_y \|\boldsymbol{y}^*(\boldsymbol{x}_2) - \boldsymbol{y}^*(\boldsymbol{x}_1)\|^3 \leq 0.$$

We apply the Lipschitz continuity of $\nabla_y f(\boldsymbol{x}, \boldsymbol{y})$ to get

$$\mu_y \|\boldsymbol{y}^*(\boldsymbol{x}_2) - \boldsymbol{y}^*(\boldsymbol{x}_1)\|^3 \leq \ell \|\boldsymbol{x}_1 - \boldsymbol{x}_2\| \|\boldsymbol{y}^*(\boldsymbol{x}_2) - \boldsymbol{y}^*(\boldsymbol{x}_1)\|,$$

which proves the continuity of $\boldsymbol{y}^*(\boldsymbol{x})$. Similarly, we can also show the result for $\boldsymbol{x}^*(\boldsymbol{y})$. ■

Appendix D. Proofs in Section 5.2

Below, we show that the inexact zeroth-order and first-order oracles are easily obtainable for both the primal objective $\Phi(\boldsymbol{x}) := \max_{\boldsymbol{y}\in\mathcal{Y}} f(\boldsymbol{x},\boldsymbol{y})$ and dual objective $\Psi(\boldsymbol{y};\bar{\boldsymbol{x}}) := \min_{\boldsymbol{x}\in\mathcal{X}} g(\boldsymbol{x},\boldsymbol{y};\bar{\boldsymbol{x}})$, where $g(\boldsymbol{x},\boldsymbol{y};\bar{\boldsymbol{x}}) := f(\boldsymbol{x},\boldsymbol{y}) + \frac{\gamma}{3}\|\boldsymbol{x}-\bar{\boldsymbol{x}}\|^3$.

Theorem D.1 *Under Assumption 3.3, 3.4, 3.5, 5.1 and 5.2, $\mathcal{M}_{\min}$ can*

- *find a point $\hat{\boldsymbol{y}}$ such that $|f(\boldsymbol{x},\hat{\boldsymbol{y}})-\Phi(\boldsymbol{x})|\le\delta$ in $\mathcal{O}\left(T_{\min}(\rho,\mu_y)\log(DL/\delta)\right)$ iterations.*
- *find a point $\hat{\boldsymbol{y}}$ such that $\|\nabla f(\boldsymbol{x},\hat{\boldsymbol{y}})-\nabla\Phi(\boldsymbol{x})\|\le\delta$ in $\mathcal{O}\left(T_{\min}(\rho,\mu_y)\log(D\ell/\delta)\right)$ iterations.*
- *find a point $\hat{\boldsymbol{x}}$ such that $|g(\hat{\boldsymbol{x}},\boldsymbol{y};\bar{\boldsymbol{x}})-\Psi(\boldsymbol{y};\bar{\boldsymbol{x}})|\le\delta$ in $\mathcal{O}\left(T_{\min}(\rho+2\gamma,\gamma/2)\log(DL/\delta)\right)$ iterations.*
- *finds a point $\hat{\boldsymbol{x}}$ such that $\|\nabla_y g(\hat{\boldsymbol{x}},\boldsymbol{y};\bar{\boldsymbol{x}})-\nabla\Psi(\boldsymbol{y};\bar{\boldsymbol{x}})\|\le\delta$ in $\mathcal{O}\left(T_{\min}(\rho+2\gamma,\gamma/2)\log(D\ell/\delta)\right)$ iterations.*

Proof To obtain an inexact zeroth-order oracle under Assumption 3.3, it suffices to find $\hat{\boldsymbol{y}}$ such that $\|\hat{\boldsymbol{y}}-\boldsymbol{y}^*(\boldsymbol{x})\|\le\delta/L$, requires $\mathcal{M}_{\min}$ in $\mathcal{O}\left((\rho/\mu_y)^{2/7}\log(DL/(\delta\mu_y))\right)$ iterations. Similarly, by Danskin's theorem $\nabla\Phi(\boldsymbol{x})=\nabla_x f(\boldsymbol{x},\boldsymbol{y}^*(\boldsymbol{x}))$, to obtain an inexact first-order oracle under Assumption 3.4, it suffices to find $\hat{\boldsymbol{y}}$ such that $\|\hat{\boldsymbol{y}}-\boldsymbol{y}^*(\boldsymbol{x})\|\le\delta/\ell$. This proves the first two claims. And the last two claims are similar by noting that $g(\boldsymbol{x},\boldsymbol{y};\bar{\boldsymbol{x}})$ is $(\gamma/2)$-uniformly convex in $\boldsymbol{x}$ and has $(\rho+2\gamma)$-Lipschitz continuous Hessians. ■

D.1. Proof of Theorem 5.1

Let $(\boldsymbol{x}^*,\boldsymbol{y}^*) = \arg\min_{\boldsymbol{x}\in\mathcal{X}}\max_{\boldsymbol{y}\in\mathcal{Y}} f(\boldsymbol{x},\boldsymbol{y})$. *Line 1* in Algorithm 3 finds $\|\hat{\boldsymbol{x}}-\boldsymbol{x}^*\|\le\zeta_1$ in $\mathcal{O}((D_x^3\gamma/\mu_x)^{2/7}\log(D/\zeta_1))$ calls of the subroutine Algorithm 4. *Line 2* in Algorithm 3 finds $\|\hat{\boldsymbol{y}}-\boldsymbol{y}^*(\hat{\boldsymbol{x}})\|\le\zeta_1$, where $\boldsymbol{y}^*(\boldsymbol{x})=\arg\max_{\boldsymbol{y}\in\mathcal{Y}} f(\boldsymbol{x},\boldsymbol{y})$. Then

$$\|\hat{\boldsymbol{y}}-\boldsymbol{y}^*\|\le\|\hat{\boldsymbol{y}}-\boldsymbol{y}^*(\hat{\boldsymbol{x}})\|+\frac{\ell^{1/2}}{\mu_y^{1/2}}\sqrt{\|\hat{\boldsymbol{x}}-\boldsymbol{x}^*\|},$$

where we use Lemma 5.3 and $\boldsymbol{y}^*(\boldsymbol{x}^*)=\boldsymbol{x}^*$ in the above inequality. It means

$$\|(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}})-(\boldsymbol{x}^*,\boldsymbol{y}^*)\|\le 3\sqrt{\frac{\ell\zeta_1}{\mu_y}}$$

Then from Lemma 3.1 we know *Line 3* in Algorithm 3 outputs $\boldsymbol{z}^{\text{out}}=(\boldsymbol{x}^{\text{out}},\boldsymbol{y}^{\text{out}})$ such that

$$\left\|\boldsymbol{F}(\boldsymbol{z}^{\text{out}})+\boldsymbol{c}^{\text{out}}\right\|\le 7\ell\sqrt{\frac{3\ell\zeta_1}{\mu_y}}$$

for some $(\boldsymbol{u}^{\text{out}},\boldsymbol{v}^{\text{out}})=\boldsymbol{c}^{\text{out}}\in\partial\mathcal{I}_{\mathcal{Z}}(\boldsymbol{z}^{\text{out}})$. Using the convex-concavity, we have that

$$\begin{aligned}
&f(\boldsymbol{x}^{\text{out}},\boldsymbol{y}^*(\boldsymbol{x}^{\text{out}}))-f(\boldsymbol{x}^*(\boldsymbol{y}^{\text{out}}),\boldsymbol{y}^{\text{out}})\\
&=f(\boldsymbol{x}^{\text{out}},\boldsymbol{y}^*(\boldsymbol{x}^{\text{out}}))-f(\boldsymbol{x}^{\text{out}},\boldsymbol{y}^{\text{out}})+f(\boldsymbol{x}^{\text{out}},\boldsymbol{y}^{\text{out}})-f(\boldsymbol{x}^*(\boldsymbol{y}^{\text{out}}),\boldsymbol{y}^{\text{out}})\\
&\le\langle\nabla_x f(\boldsymbol{x}^{\text{out}},\boldsymbol{y}^{\text{out}})+\boldsymbol{u}^{\text{out}},\boldsymbol{x}^*(\boldsymbol{y}^{\text{out}})-\boldsymbol{x}^{\text{out}}\rangle\\
&\quad+\langle-\nabla_y f(\boldsymbol{x}^{\text{out}},\boldsymbol{y}^{\text{out}})+\boldsymbol{v}^{\text{out}},\boldsymbol{y}^*(\boldsymbol{x}^{\text{out}})-\boldsymbol{y}^{\text{out}}\rangle.
\end{aligned}$$

Finally, the Cauchy–Schwarz inequality tells

$$
\begin{aligned}
&f(\boldsymbol{x}^{\text{out}}, \boldsymbol{y}^*(\boldsymbol{x}^{\text{out}})) - f(\boldsymbol{x}^*(\boldsymbol{y}^{\text{out}}), \boldsymbol{y}^{\text{out}}) \\
&\leq \left\| \begin{bmatrix} \nabla_x f(\boldsymbol{x}^{\text{out}}, \boldsymbol{y}^{\text{out}}) + \boldsymbol{u}^{\text{out}} \\ -\nabla_y f(\boldsymbol{x}^{\text{out}}, \boldsymbol{y}^{\text{out}}) + \boldsymbol{v}^{\text{out}} \end{bmatrix} \right\| \leq 7\ell D \sqrt{\frac{3\ell\zeta_1}{\mu_y}}.
\end{aligned}
$$

D.2. Proof of Theorem 5.2

Let $(\boldsymbol{x}^*(\bar{\boldsymbol{x}}), \boldsymbol{y}^*(\bar{\boldsymbol{x}})) = \arg\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} g(\boldsymbol{x}, \boldsymbol{y}; \bar{\boldsymbol{x}})$ and $\Psi(\boldsymbol{y}; \bar{\boldsymbol{x}}) = \min_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x}, \boldsymbol{y})$. *Line 1* in Algorithm 4 finds $\|\hat{\boldsymbol{y}} - \boldsymbol{y}^*(\bar{\boldsymbol{x}})\| \leq \zeta_2$ in $\mathcal{O}((D_y^3\gamma/\mu_y)^{2/7}\log(D/\zeta_2))$ calls of the subroutine Algorithm 5, where $\boldsymbol{y}^*(\bar{\boldsymbol{x}}) = \arg\max_{\boldsymbol{y}\in\mathcal{Y}} \Psi(\boldsymbol{y}; \bar{\boldsymbol{x}})$. *Line 2* in Algorithm 4 finds $\hat{\boldsymbol{x}}$ such that $\|\hat{\boldsymbol{x}} - \boldsymbol{x}^*(\hat{\boldsymbol{y}}; \bar{\boldsymbol{x}})\| \leq \zeta_2$. By Lemma 3.1, *Line 3* in Algorithm 4 further ensures that $\|\nabla_x g(\boldsymbol{x}^{\text{out}}, \hat{\boldsymbol{y}}; \bar{\boldsymbol{x}}) + \boldsymbol{u}^{\text{out}}\| \leq 6(\ell + 2\gamma D)\zeta_2$ for some $\boldsymbol{u}^{\text{out}} \in \partial\mathcal{I}_{\mathcal{X}}(\boldsymbol{x}^{\text{out}})$. Invoking Lemma 5.3, we have that

$$
\begin{aligned}
\|\boldsymbol{x}^*(\boldsymbol{y}_1; \bar{\boldsymbol{x}}) - \boldsymbol{x}^*(\boldsymbol{y}_2; \bar{\boldsymbol{x}})\|^2 &\leq \frac{\ell + 2\gamma D}{\gamma/2}\|\boldsymbol{y}_1 - \boldsymbol{y}_2\|, \quad \forall \boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathcal{Y}; \\
\|\boldsymbol{y}^*(\boldsymbol{x}_1; \bar{\boldsymbol{x}}) - \boldsymbol{y}^*(\boldsymbol{x}_2; \bar{\boldsymbol{x}})\|^2 &\leq \frac{\ell}{\mu_y}\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|, \quad \forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X},
\end{aligned} \tag{19}
$$

where $\boldsymbol{x}^*(\boldsymbol{y}; \bar{\boldsymbol{x}}) = \arg\min_{\boldsymbol{x}\in\mathcal{X}} g(\boldsymbol{x}, \boldsymbol{y}; \bar{\boldsymbol{x}})$ and $\boldsymbol{y}^*(\boldsymbol{x}; \bar{\boldsymbol{x}}) = \arg\max_{\boldsymbol{y}\in\mathcal{Y}} g(\boldsymbol{x}, \boldsymbol{y}; \bar{\boldsymbol{x}})$. Then

$$
\begin{aligned}
\|\boldsymbol{x}^{\text{out}} - \boldsymbol{x}^*(\bar{\boldsymbol{x}})\| &\leq \|\boldsymbol{x}^{\text{out}} - \boldsymbol{x}^*(\hat{\boldsymbol{y}}; \bar{\boldsymbol{x}})\| + \frac{(\ell + 2\gamma D)^{1/2}}{(\gamma/2)^{1/2}}\sqrt{\|\hat{\boldsymbol{y}} - \boldsymbol{y}^*(\bar{\boldsymbol{x}})\|} \\
&\leq \frac{1}{(\gamma/2)^{1/2}}\sqrt{\|\nabla_x g(\boldsymbol{x}^{\text{out}}, \hat{\boldsymbol{y}}; \bar{\boldsymbol{x}}) + \boldsymbol{u}^{\text{out}}\|} + \frac{(\ell + 2\gamma D)^{1/2}}{(\gamma/2)^{1/2}}\sqrt{\|\hat{\boldsymbol{y}} - \boldsymbol{y}^*(\bar{\boldsymbol{x}})\|},
\end{aligned} \tag{20}
$$

where we use (19) and $\boldsymbol{x}^*(\boldsymbol{y}^*(\bar{\boldsymbol{x}}); \bar{\boldsymbol{x}}) = \boldsymbol{x}^*(\bar{\boldsymbol{x}})$ in the first inequality, Lemma C.1 and the Cauchy-Schwarz inequality in the second one. Using a similar analysis, we can further show that

$$
\begin{aligned}
&\left\|\nabla\Phi(\boldsymbol{x}^{\text{out}}) + \gamma\|\boldsymbol{x}^{\text{out}} - \bar{\boldsymbol{x}}\|(\boldsymbol{x}^{\text{out}} - \bar{\boldsymbol{x}}) + \boldsymbol{u}^{\text{out}}\right\| = \|\nabla_x g(\boldsymbol{x}^{\text{out}}, \boldsymbol{y}^*(\boldsymbol{x}^{\text{out}}; \bar{\boldsymbol{x}}); \bar{\boldsymbol{x}}) + \boldsymbol{u}^{\text{out}}\| \\
&\leq \|\nabla_x g(\boldsymbol{x}^{\text{out}}, \boldsymbol{y}^*(\bar{\boldsymbol{x}}); \bar{\boldsymbol{x}}) + \boldsymbol{u}^{\text{out}}\| + \frac{(\ell + 2\rho D)\ell^{1/2}}{\mu_y^{1/2}}\sqrt{\|\boldsymbol{x}^{\text{out}} - \boldsymbol{x}^*(\bar{\boldsymbol{x}})\|} \\
&\leq \|\nabla_x g(\boldsymbol{x}^{\text{out}}, \hat{\boldsymbol{y}}; \bar{\boldsymbol{x}}) + \boldsymbol{u}^{\text{out}}\| + (\ell + 2\rho D)\|\hat{\boldsymbol{y}} - \boldsymbol{y}^*(\bar{\boldsymbol{x}})\| + \frac{(\ell + 2\rho D)\ell^{1/2}}{\mu_y^{1/2}}\sqrt{\|\boldsymbol{x}^{\text{out}} - \boldsymbol{x}^*(\bar{\boldsymbol{x}})\|},
\end{aligned}
$$

where $\boldsymbol{y}^*(\boldsymbol{x}; \bar{\boldsymbol{x}}) = \arg\max_{\boldsymbol{y}\in\mathcal{Y}} g(\boldsymbol{x}, \boldsymbol{y}; \bar{\boldsymbol{x}})$. Now we plug Eq. (20) into the above inequality to get

$$
\begin{aligned}
&\|\nabla_x g(\boldsymbol{x}^{\text{out}}, \boldsymbol{y}^*(\boldsymbol{x}^{\text{out}}; \bar{\boldsymbol{x}}); \bar{\boldsymbol{x}}) + \boldsymbol{u}^{\text{out}}\| \\
&\leq 6(\ell + 2\gamma D)\zeta_2 + \frac{(\ell + 2\rho D)\ell^{1/2}}{\mu_y^{1/2}(\gamma/2)^{1/4}}\,(6(\ell + 2\gamma D)\zeta_2)^{1/4} \\
&\quad + (\ell + 2\rho D)\zeta_2 + \frac{(\ell + 2\rho D)\ell^{1/2}(\ell + 2\gamma D)^{1/4}}{\mu_y^{1/2}(\gamma/2)^{1/4}}(\zeta_2)^{1/2},
\end{aligned} \tag{21}
$$

It means that Algorithm 4 can successfully implement a (δ, γ)-proximal oracle that satisfies Assumption 5.4 by setting $\zeta_2 = \Omega(1/\text{poly}(\rho, \ell, D, \gamma, \mu_x^{-1}, \mu_y^{-1}, \zeta_1^{-1}))$. Finally, the complexity of obtaining inexact zeroth-order and first-order oracles for $\Phi(\boldsymbol{x})$ is given in Theorem D.1.

D.3. Proof of Theorem 5.3

Let $(\boldsymbol{x}^*(\bar{\boldsymbol{x}},\bar{\boldsymbol{y}}),\boldsymbol{y}^*(\bar{\boldsymbol{x}},\bar{\boldsymbol{y}})) = \arg\min_{\boldsymbol{x}\in\mathcal{X}}\max_{\boldsymbol{y}\in\mathcal{Y}} h(\boldsymbol{x},\boldsymbol{y};\bar{\boldsymbol{x}},\bar{\boldsymbol{y}})$. By Theorem E.1, we know that *Line 1* of Algorithm 5 can find $\|(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}})-(\boldsymbol{x}^*(\bar{\boldsymbol{x}},\bar{\boldsymbol{y}}),\boldsymbol{y}^*(\bar{\boldsymbol{x}},\bar{\boldsymbol{y}}))\| \leq \zeta_3$ in $\mathcal{O}(((\rho+2\gamma)/\gamma)^{2/3}\log(D/\zeta_3))$ iterations. And by Lemma 3.1, *Line 2* of Algorithm 5 guarantees that

$$\left\|\begin{bmatrix}\nabla_x h(\boldsymbol{x}^{\rm out},\boldsymbol{y}^{\rm out};\bar{\boldsymbol{x}},\bar{\boldsymbol{y}})+\boldsymbol{u}^{\rm out}\\ -\nabla_y h(\boldsymbol{x}^{\rm out},\boldsymbol{y}^{\rm out};\bar{\boldsymbol{x}},\bar{\boldsymbol{y}})+\boldsymbol{v}^{\rm out}\end{bmatrix}\right\| \leq 6(\ell+2\gamma D)\zeta_3$$

for some $\boldsymbol{u}^{\rm out}\in\partial\mathcal{I}_{\mathcal{X}}(\boldsymbol{x})$ and $\boldsymbol{v}^{\rm out}\in\partial\mathcal{I}_{\mathcal{Y}}(\boldsymbol{y})$. Invoking Lemma 5.3, we know that

$$\|\boldsymbol{x}^*(\boldsymbol{y}_1;\bar{\boldsymbol{x}},\bar{\boldsymbol{y}})-\boldsymbol{x}^*(\boldsymbol{y}_2;\bar{\boldsymbol{x}},\bar{\boldsymbol{y}})\|^2 \leq \frac{\ell+2\gamma D}{\gamma/2}\|\boldsymbol{y}_1-\boldsymbol{y}_2\|, \quad \forall \boldsymbol{y}_1,\boldsymbol{y}_2\in\mathcal{Y},$$

where $\boldsymbol{x}^*(\boldsymbol{y};\bar{\boldsymbol{x}},\bar{\boldsymbol{y}}) = \arg\min_{\boldsymbol{x}\in\mathcal{X}} h(\boldsymbol{x},\boldsymbol{y};\bar{\boldsymbol{x}},\bar{\boldsymbol{y}})$. Then

$$\begin{aligned}
&\left\|\nabla\Psi(\boldsymbol{y}^{\rm out};\bar{\boldsymbol{x}})+\gamma\|\boldsymbol{y}^{\rm out}-\bar{\boldsymbol{y}}\|(\boldsymbol{y}^{\rm out}-\bar{\boldsymbol{y}})+\boldsymbol{v}^{\rm out}\right\| = \|\nabla_y h(\boldsymbol{x}^*(\boldsymbol{y}^{\rm out};\bar{\boldsymbol{x}},\bar{\boldsymbol{y}}),\boldsymbol{y}^{\rm out};\bar{\boldsymbol{x}},\bar{\boldsymbol{y}})+\boldsymbol{v}^{\rm out}\|\\
&\leq \|\nabla_y h(\boldsymbol{x}^*(\bar{\boldsymbol{x}},\bar{\boldsymbol{y}}),\boldsymbol{y}^{\rm out};\bar{\boldsymbol{x}},\bar{\boldsymbol{y}})+\boldsymbol{v}^{\rm out}\| + \frac{(\ell+2\gamma D)^{3/2}}{(\gamma/2)^{1/2}}\|\boldsymbol{y}^{\rm out}-\boldsymbol{y}^*(\bar{\boldsymbol{x}},\bar{\boldsymbol{y}})\|\\
&\leq \|\nabla_y h(\boldsymbol{x}^{\rm out},\boldsymbol{y}^{\rm out};\bar{\boldsymbol{x}},\bar{\boldsymbol{y}})+\boldsymbol{v}^{\rm out}\| + (\ell+2\gamma D)\|\boldsymbol{x}^{\rm out}-\boldsymbol{x}^*(\bar{\boldsymbol{x}},\bar{\boldsymbol{y}})\|\\
&\quad + \frac{(\ell+2\gamma D)^{3/2}}{(\gamma/2)^{1/2}}\sqrt{\|\boldsymbol{y}^{\rm out}-\boldsymbol{y}^*(\bar{\boldsymbol{x}},\bar{\boldsymbol{y}})\|}\\
&\leq 6(\ell+2\gamma D)\zeta_3 + \frac{\ell+2\gamma D}{(\gamma/2)^{1/2}}(6(\ell+2\gamma D)\zeta_3)^{1/2} + \frac{(\ell+2\gamma D)^{3/2}}{(\gamma/2)^{3/4}}(6(\ell+2\gamma D)\zeta_3)^{1/4},
\end{aligned} \tag{22}$$

where in the last line we use Lemma C.1 as well as the Cauchy-Schwartz inequality to upper bound the distance to saddle point with gradient norm. The above inequality means that Algorithm 5 can successfully implement a (δ,γ)-proximal oracle that satisfies Assumption 5.5 by setting $\zeta_2 = \Omega(1/\text{poly}(\rho,\ell,D,\gamma,\mu_y^{-1},\zeta_2^{-1}))$. Finally, the complexity of obtaining inexact zeroth-order and first-order oracles for $\Psi(\boldsymbol{y};\bar{\boldsymbol{x}})$ is given in Theorem D.1.

Appendix E. Proofs in Section 5.3

E.1. Guarantee of the NPE Subroutine

Monteiro and Svaiter (2012) proposed the Newton Proximal Extragradient (NPE) method for monotone variational inequalities, which can find an ϵ-solution with $\mathcal{O}(\epsilon^{-2/3})$ second-order oracle calls. As noted in (Bullins and Lai, 2022; Huang and Zhang, 2022; Lin et al., 2022; Lin and Jordan, 2024; Adil et al., 2022), the procedure of NPE can be simplified by using the cubic regularized Newton oracle. We present the simplified version in Algorithm 6. It is known (Monteiro and Svaiter, 2012; Adil et al., 2022; Bullins and Lai, 2022) that Algorithm 6 can provably find an ϵ-solution to the variational inequality problem induced by a monotone operator $\boldsymbol{F}$ in $\mathcal{O}(\epsilon^{-2/3})$ iterations. And applying the restart strategy on it (Algorithm 7) can solve a μ-strongly monotone variational inequality problem in $\mathcal{O}((\gamma/\mu)^{2/3}\log\epsilon^{-1})$ iteration complexity, as stated in the following theorem.

Theorem E.1 (NPE-restart) *Under Assumption 5.1 with $\mu_x = \mu_y = \mu$ and Assumption 3.5, running Algorithm 7 with $\gamma = 2\rho$ and $T = \mathcal{O}((\gamma/\mu)^{2/3})$ and $S = \mathcal{O}(\log(d_0/\epsilon))$ returns a point $\boldsymbol{z}^{(S)}$ such that $\|\boldsymbol{z}^{(S)}-\boldsymbol{z}^*\| \leq \epsilon$, where $\boldsymbol{z}^*$ is the unique solution to Problem (2) and $d_0 = \|\boldsymbol{z}_0-\boldsymbol{z}^*\|$.*

Algorithm 6 NPE$(\boldsymbol{z}_0, T, \gamma)$

```
for $t = 0, \cdots, T-1$ do
    $\boldsymbol{z}_{t+/2}, \boldsymbol{u}_{t+1/2} = \mathrm{CRN}(\boldsymbol{z}_t, \gamma)$
    $\eta_t = \frac{1}{2\gamma\|\boldsymbol{z}_t - \boldsymbol{z}_{t+1/2}\|}$
    $\boldsymbol{z}_{t+1} = \arg\min_{\boldsymbol{z}\in\mathcal{Z}} \left\{\langle \eta_t \boldsymbol{F}(\boldsymbol{z}_{t+1/2}), \boldsymbol{z} - \boldsymbol{z}_t\rangle + \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{z}_t\|^2\right\}$.
end for
return $z_{\mathrm{out}} = \frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t \boldsymbol{z}_{t+1/2}$
```

Algorithm 7 NPE-restart$(\boldsymbol{z}_0, T, \gamma, S)$

```
$\boldsymbol{z}^{(0)} = \boldsymbol{z}_0$
for $s = 0, \cdots, S-1$ do
    $\boldsymbol{z}^{(s+1)} = \mathrm{NPE}(\boldsymbol{z}^{(s)}, T, \gamma)$
end for
return $\boldsymbol{z}^{(S)}$
```

Proof By Theorem E.2, each epoch of Algorithm 7 ensures $\|\boldsymbol{z}^{(s+1)} - \boldsymbol{z}^*\| \leq \frac{1}{2}\|\boldsymbol{z}^{(s)} - \boldsymbol{z}^*\|$ if setting $T = \mathcal{O}\left((\gamma/\mu)^{2/3}\right)$. And therefore the algorithm finds a point $\boldsymbol{z}^{(S)}$ such that $\|\boldsymbol{z}^{(S)} - \boldsymbol{z}^*\| \leq \epsilon$ in $S = \lceil \log_2(d_0/\epsilon)\rceil$ epochs. ■

Theorem E.2 *Under the same setting as Theorem E.1, running Algorithm 6 with $\gamma = 2\rho$ outputs a point $\boldsymbol{z}_{\mathrm{out}}$ such that $\|\boldsymbol{z}_{\mathrm{out}} - \boldsymbol{z}^*\| \leq \epsilon$ in $T = \mathcal{O}\left(\left(\frac{\gamma d_0^3}{\mu\epsilon^3}\right)^{2/3}\right)$ iterations, where $\boldsymbol{z}^*$ is the unique solution to Problem (2) and $d_0 = \|\boldsymbol{z}_0 - \boldsymbol{z}^*\|$.*

Proof It is known (Monteiro and Svaiter, 2012; Adil et al., 2022; Bullins and Lai, 2022) that Algorithm 6 ensures

$$\mathrm{Reget} := \frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\langle \boldsymbol{F}(\boldsymbol{z}_{t+1/2}), \boldsymbol{z}_{t+1/2} - \boldsymbol{z}^*\rangle = \mathcal{O}\left(\frac{\gamma\|\boldsymbol{z}_0 - \boldsymbol{z}^*\|^3}{T^{3/2}}\right). \tag{23}$$

We further use Lemma C.1, the convexity of $\|\cdot\|^3$ and Jensen's inequality to derive that

$$\|\boldsymbol{z}_{\mathrm{out}} - \boldsymbol{z}^*\|^3 \leq \frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t\|\boldsymbol{z}_{t+1/2} - \boldsymbol{z}^*\|^3 \leq \frac{3}{2\mu}\mathrm{Regret},$$

which leads to the result. ■

Algorithm 8 LEN$(\boldsymbol{z}_0, T, \gamma)$

```
for $t = 0, \cdots, T-1$ do
    $\boldsymbol{z}_{t+/2}, \boldsymbol{u}_{t+1/2} = \mathrm{LazyCRN}(\boldsymbol{z}_t, \boldsymbol{z}_{\pi(t)}, \gamma)$
    $\eta_t = \frac{1}{2\gamma\|\boldsymbol{z}_t - \boldsymbol{z}_{t+1/2}\|}$
    $\boldsymbol{z}_{t+1} = \arg\min_{\boldsymbol{z}\in\mathcal{Z}} \left\{\langle \eta_t \boldsymbol{F}(\boldsymbol{z}_{t+1/2}), \boldsymbol{z} - \boldsymbol{z}_t\rangle + \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{z}_t\|^2\right\}$.
end for
return $z_{\mathrm{out}} = \frac{1}{\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t \boldsymbol{z}_{t+1/2}$
```

Algorithm 9 LEN-restart$(\boldsymbol{z}_0, T, \gamma, S)$

```
$\boldsymbol{z}^{(0)} = \boldsymbol{z}_0$
for $s = 0, \cdots, S-1$ do
    $\boldsymbol{z}^{(s+1)} = \mathrm{LEN}(\boldsymbol{z}^{(s)}, T, \gamma)$
end for
return $\boldsymbol{z}^{(S)}$
```

E.2. Guarantee of the LEN Subroutine

This section briefly reviews the results in the recent work (Chen et al., 2025b,a). Chen et al. (2025b) proposed the Lazy Extra Newton (LEN) method as the lazy version of NPE, and in the subsequent work, Chen et al. (2025a) proposed the accelerated LEN (A-LEN) as the lazy version of A-NPE. Instead of using the CRN oracle, they use the following lazy CRN oracle proposed by Doikov et al. (2023) to further reduce the computational complexity of NPE and A-NPE.

Definition E.1 *A lazy CRN oracle for Problem (1) takes the query point $\bar{\boldsymbol{z}} \in \mathcal{Z}$, the snapshot point $\boldsymbol{z}_{\mathrm{ss}}$, and the regularization parameter $\gamma > 0$ as inputs, and returns $(\boldsymbol{z}, \boldsymbol{u}) = \mathrm{CRN}(\bar{\boldsymbol{z}}, \boldsymbol{z}_{\mathrm{ss}}, \gamma)$ satisfies:*

$$\langle \boldsymbol{F}(\bar{\boldsymbol{z}}) + \nabla \boldsymbol{F}(\boldsymbol{z}_{\mathrm{ss}})(\boldsymbol{z} - \bar{\boldsymbol{z}}) + \frac{\gamma}{2}\|\boldsymbol{z} - \bar{\boldsymbol{z}}\|(\boldsymbol{z} - \bar{\boldsymbol{z}}), \boldsymbol{z}' - \boldsymbol{z}\rangle \geq 0, \ \forall \boldsymbol{z}' \in \mathcal{Z};$$
$$\boldsymbol{u} = -\left(\boldsymbol{F}(\bar{\boldsymbol{z}}) + \nabla \boldsymbol{F}(\boldsymbol{z}_{\mathrm{ss}})(\boldsymbol{z} - \bar{\boldsymbol{z}}) + \frac{\gamma}{2}\|\boldsymbol{z} - \bar{\boldsymbol{z}}\|(\boldsymbol{z} - \bar{\boldsymbol{z}})\right) \in \begin{bmatrix} \partial \mathcal{I}_{\mathcal{X}}(\boldsymbol{x}) \\ -\partial \mathcal{I}_{\mathcal{Y}}(\boldsymbol{y}) \end{bmatrix}.$$

Using their result, we know that $T_{\mathrm{saddle}}(\rho, \mu) = m + m^{2/3}(\rho/\mu)^{2/3}$ and $T_{\mathrm{min}}(\rho, \mu) = m + m^{5/7}(\rho/\mu)^{2/7}$. A small difference is that the work (Chen et al., 2025b,a) only analyzed the unconstrained case ($\mathcal{X} = \mathbb{R}^{d_x}$ and $\mathcal{Y} = \mathbb{R}^{d_y}$), but we need the results for the constraints sets $\mathcal{X}$ and $\mathcal{Y}$ here. We remark that essentially all the analysis in the work (Chen et al., 2025b,a) holds under the constrained setting. Below, we show the convergence of LEN-restart (Algorithm 9), which invokes LEN (Algorithm 8) as a subroutine. In Algorithm 8, we use the notation $\pi(t) = t - t \mod m$.

Theorem E.3 (LEN-restart) *Under Assumption 5.1 with $\mu_x = \mu_y = \mu$ and Assumption 3.5, running Algorithm 7 with $\gamma = \mathcal{O}(m\rho)$ and $T = \mathcal{O}(m + (\gamma/\mu)^{2/3})$ and $S = \mathcal{O}(\log(d_0/\epsilon))$ returns a point $\boldsymbol{z}^{(S)}$ such that $\|\boldsymbol{z}^{(S)} - \boldsymbol{z}^*\| \leq \epsilon$, where $\boldsymbol{z}^*$ is the unique solution to Problem (2) and $d_0 = \|\boldsymbol{z}_0 - \boldsymbol{z}^*\|$.*

Proof As shown in Theorem E.2, the restart scheme can transform the convergence under the convex-concave setting to the convergence under the uniformly-convex-uniformly-concave setting in a black-box manner. Therefore, we only need to show the convergence of LEN (Algorithm 8) under the convex-concave setting. The proof is essentially the same as (Chen et al., 2025b, Theorem 4.1), except here we consider the constrained case. Below, we show that all the proofs in (Chen et al., 2025b, Theorem 4.1) also hold under the constrained case.

Using the first-order optimality condition in the extragradient step, we have that

$$0 \le \langle \eta_t \boldsymbol{F}(\boldsymbol{z}_{t+1/2}) + \boldsymbol{z}_{t+1} - \boldsymbol{z}_t, \boldsymbol{z} - \boldsymbol{z}_{t+1} \rangle, \quad \forall \boldsymbol{z} \in \mathcal{Z}. \tag{24}$$

Using the first-order optimality condition in the lazy Newton step, we have that

$$0 \le \langle \eta_t \tilde{\boldsymbol{F}}(\boldsymbol{z}_{t+1/2}) + \boldsymbol{z}_{t+1/2} - \boldsymbol{z}_t, \boldsymbol{z} - \boldsymbol{z}_{t+1/2} \rangle, \quad \forall \boldsymbol{z} \in \mathcal{Z}, \tag{25}$$

where $\tilde{\boldsymbol{F}}(\boldsymbol{z}_{t+1/2}) = \boldsymbol{F}(\boldsymbol{z}_t) + \nabla \boldsymbol{F}(\boldsymbol{z}_{\pi(t)})(\boldsymbol{z}_{t+1/2} - \boldsymbol{z}_t)$. They together imply that

$$\begin{aligned}
&\eta_t \langle \boldsymbol{F}(\boldsymbol{z}_{t+1/2}), \boldsymbol{z}_{t+1/2} - \boldsymbol{z} \rangle \\
&= \eta_t \langle \boldsymbol{F}(\boldsymbol{z}_{t+1/2}), \boldsymbol{z}_{t+1} - \boldsymbol{z} \rangle + \eta_t \langle \boldsymbol{F}(\boldsymbol{z}_{t+1/2}), \boldsymbol{z}_{t+1/2} - \boldsymbol{z}_{t+1} \rangle \\
&= \eta_t \langle \boldsymbol{F}(\boldsymbol{z}_{t+1/2}), \boldsymbol{z}_{t+1} - \boldsymbol{z} \rangle + \eta_t \langle \tilde{\boldsymbol{F}}(\boldsymbol{z}_{t+1/2}), \boldsymbol{z}_{t+1/2} - \boldsymbol{z}_{t+1} \rangle \\
&\quad + \eta_t \langle \boldsymbol{F}(\boldsymbol{z}_{t+1/2}) - \tilde{\boldsymbol{F}}(\boldsymbol{z}_{t+1/2}), \boldsymbol{z}_{t+1/2} - \boldsymbol{z}_{t+1} \rangle \\
&\le \langle \boldsymbol{z}_{t+1} - \boldsymbol{z}_t, \boldsymbol{z} - \boldsymbol{z}_{t+1} \rangle + \langle \boldsymbol{z}_{t+1/2} - \boldsymbol{z}_t, \boldsymbol{z}_{t+1} - \boldsymbol{z}_{t+1/2} \rangle \\
&\quad + \eta_t \langle \boldsymbol{F}(\boldsymbol{z}_{t+1/2}) - \tilde{\boldsymbol{F}}(\boldsymbol{z}_{t+1/2}), \boldsymbol{z}_{t+1/2} - \boldsymbol{z}_{t+1} \rangle.
\end{aligned} \tag{26}$$

Next, using identity $\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \frac{1}{2}\|\boldsymbol{a}+\boldsymbol{b}\|^2 - \frac{1}{2}\|\boldsymbol{a}\|^2 + \frac{1}{2}\|\boldsymbol{b}\|^2$ and Young's inequality yields

$$\begin{aligned}
&\eta_t \langle \boldsymbol{F}(\boldsymbol{z}_{t+1/2}), \boldsymbol{z}_{t+1/2} - \boldsymbol{z} \rangle \\
&\le \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{z}_t\|^2 \cancel{-\frac{1}{2}\|\boldsymbol{z}_{t+1} - \boldsymbol{z}_t\|^2} - \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{z}_{t+1}\|^2 \\
&\quad + \cancel{\frac{1}{2}\|\boldsymbol{z}_{t+1} - \boldsymbol{z}_t\|^2} - \frac{1}{2}\|\boldsymbol{z}_{t+1/2} - \boldsymbol{z}_t\|^2 - \frac{1}{2}\|\boldsymbol{z}_{t+1} - \boldsymbol{z}_{t+1/2}\|^2 \\
&\quad + \underbrace{\eta_t^2 \|\boldsymbol{F}(\boldsymbol{z}_{t+1/2}) - \tilde{\boldsymbol{F}}(\boldsymbol{z}_{t+1/2})\|^2}_{(*)} + \frac{1}{4}\|\boldsymbol{z}_{t+1} - \boldsymbol{z}_{t+1/2}\|^2.
\end{aligned}$$

We can upper bound (*) using Young's inequality and Assumption 3.5 as

$$\begin{aligned}
&\eta_t^2 \|\boldsymbol{F}(\boldsymbol{z}_{t+1/2}) - \tilde{\boldsymbol{F}}(\boldsymbol{z}_{t+1/2})\|^2 \\
&\le 2\eta_t^2 \|\boldsymbol{F}(\boldsymbol{z}_{t+1/2}) - \boldsymbol{F}(\boldsymbol{z}_t) - \nabla \boldsymbol{F}(\boldsymbol{z}_t)(\boldsymbol{z}_{t+1/2} - \boldsymbol{z}_t)\|^2 \\
&\quad + 2\eta_t^2 \|(\nabla \boldsymbol{F}(\boldsymbol{z}_t) - \nabla \boldsymbol{F}(\boldsymbol{z}_{\pi(t)}))(\boldsymbol{z}_{t+1/2} - \boldsymbol{z}_t)\|^2 \\
&\le \frac{\eta_t^2 \rho^2}{2}\|\boldsymbol{z}_{t+1/2} - \boldsymbol{z}_t\|^4 + 2\eta_t^2 \rho^2 \|\boldsymbol{z}_t - \boldsymbol{z}_{\pi(t)}\|^2 \|\boldsymbol{z}_{t+1/2} - \boldsymbol{z}_t\|^2 \\
&= \frac{\rho^2}{8\gamma^2}\|\boldsymbol{z}_{t+1/2} - \boldsymbol{z}_t\|^2 + \frac{\rho^2}{2\gamma^2}\|\boldsymbol{z}_t - \boldsymbol{z}_{\pi(t)}\|^2.
\end{aligned}$$

Finally, it leads to

$$\begin{aligned}
&\eta_t \langle \boldsymbol{F}(\boldsymbol{z}_{t+1/2}), \boldsymbol{z}_{t+1/2} - \boldsymbol{z} \rangle \\
&\le \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{z}_t\|^2 - \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{z}_{t+1}\|^2 - \frac{1}{4}\|\boldsymbol{z}_{t+1} - \boldsymbol{z}_{t+1/2}\|^2 \\
&\quad - \left(\frac{1}{2} - \frac{\rho^2}{8\gamma^2}\right)\|\boldsymbol{z}_{t+1/2} - \boldsymbol{z}_t\|^2 + \frac{\rho^2}{2\gamma^2}\|\boldsymbol{z}_t - \boldsymbol{z}_{\pi(t)}\|^2.
\end{aligned}$$

This matches (Chen et al., 2025b, Lemma 4.1) under the unconstrained setting up to only constants. Then we can follow the proof of (Chen et al., 2025b, Theorem 4.1) to conclude Theorem E.3. ■

Similarly, the result of the (restarted) A-LEN Algorithm (Chen et al., 2025a, Theorem 5.3) also hold under the constrained setting, as stated below.

Theorem E.4 (ALEN-restart) *Assume $h(\boldsymbol{z}) : \mathcal{Z} \to \mathbb{R}$ is μ-uniformly convex and has ρ-Lipschitz continuous Hessians. There exists a second-order algorithm, specifically, the restart version of (Chen et al., 2025a, Algorithm 5.1), that reuses Hessians every m iterations, and returns a point $\boldsymbol{z}$ such that $\|\boldsymbol{z} - \boldsymbol{z}^*\| \le \epsilon$ in $\mathcal{O}\left(m + m^{5/7}(\rho/\mu)^{2/7}\log(d_0/\epsilon)\log m\right)$ lazy CRN oracle calls.*

Proof As shown in Theorem 4.1, the restart scheme transforms the convergence under the convex setting to the uniformly convex setting in a black-box manner. Hence, we only need to show the convergence of A-LEN (Chen et al., 2025a, Theorem 5.3) under the convex setting. The result for the unconstrained case can be found in (Chen et al., 2025a, Theorem 5.3). Essentially, the same result also holds under the constrained setting we consider here. It is because (Chen et al., 2025a, Algorithm 5.1) invokes the lazy CRN (Doikov et al., 2023, Algorithm 1) as a subroutine, while the proof of lazy CRN can be readily extended to the constrained case (Doikov et al., 2023, Appendix F). ■