Efficient Second-Order Methods for Min-Max Optimization Using Inexact Hessian Information

Chris Junchi Li[◇]

Department of Electrical Engineering and Computer Sciences[◇]
University of California, Berkeley

September 1, 2024

Abstract

This paper introduces a class of explicit second-order methods for convex-concave min-max optimization that achieves optimal convergence rates while addressing the computational challenges posed by inexact Hessian information. We extend traditional first-order approaches by incorporating second-order information through subsampled Newton methods, which are both computationally efficient and robust to inaccuracies in second-order data. Our proposed methods relax the strict requirements of existing algorithms, allowing for inexact subproblem solutions without compromising convergence guarantees. We demonstrate that these methods achieve an order-optimal complexity of $O(\epsilon^{-2/3})$ and significantly reduce computational overhead through the use of Schur decomposition and randomized sampling techniques. Theoretical analysis confirm that our approach outperforms current state-of-the-art algorithms, particularly in large-scale applications where exact second-order information is impractical.

Keywords: Min-Max Optimization; Second-Order Methods; Convex-Concave Functions; Inexact Hessian Information; Subsampled Newton Methods; Schur Decomposition

1 Introduction

Min-max optimization has long been a cornerstone in various fields such as game theory, economics, and machine learning, where it serves as a fundamental framework for modeling adversarial interactions and decision-making processes. The classical min-max problem involves finding a global saddle point of a function that is convex in one variable and concave in another. This setting has seen extensive research, particularly due to its applications in machine learning models like Generative Adversarial Networks (GANs) and adversarial learning. Despite significant progress in the development of optimization algorithms, the complexity and scalability of these methods remain critical challenges, especially when dealing with high-dimensional and large-scale problems.

First-order methods, such as the extragradient and optimistic gradient methods, have been the go-to algorithms for solving min-max problems due to their simplicity and relatively low computational cost. These methods are well-suited for large-scale applications where high accuracy is often unnecessary, and they have been proven to achieve optimal convergence rates in various settings. However, first-order methods can suffer from slow convergence, especially in ill-conditioned problems where the underlying objective function exhibits sharp curvature. This has motivated the exploration of second-order methods, which are known for their superior convergence properties and robustness to parameter tuning.

Second-order methods leverage curvature information to accelerate convergence, making them particularly effective in scenarios where first-order methods struggle. However, the practical use of these methods has been limited due to the computational burden associated with calculating and inverting the Hessian matrix, especially in large-scale problems. Recent advances have sought to mitigate these challenges by introducing inexact second-order methods that approximate the Hessian or use subsampling techniques to reduce computational costs. These methods offer a promising balance between accuracy and efficiency, but they often require complex implicit subproblem solving, which can be computationally expensive and challenging to implement.

In this paper, we propose a new class of explicit second-order min-max optimization algorithms that overcome the limitations of existing methods. Our approach introduces an inexact Jacobian regularization framework, which allows for the use of approximated second-order information while maintaining order-optimal convergence guarantees. By leveraging subsampled Newton methods, our algorithms achieve the optimal convergence rate of $O(\epsilon^{-2/3})$ with significantly reduced computational overhead. Furthermore, our methods are explicit, meaning that they avoid the need for implicit subproblem solving, thereby simplifying implementation and improving computational efficiency.

Mathematical Formulations. Let $\mathbb{R}^m$ and $\mathbb{R}^n$ be finite-dimensional Euclidean spaces and assume that the function $f : \mathbb{R}^m \times \mathbb{R}^n \mapsto \mathbb{R}$ has a bounded and Lipschitz-continuous Hessian. We consider the problem of finding a global saddle point of the following min-max optimization problem:

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

i.e., a tuple $(\mathbf{x}^\star, \mathbf{y}^\star) \in \mathbb{R}^m \times \mathbb{R}^n$ such that

$$f(\mathbf{x}^\star, \mathbf{y}) \leq f(\mathbf{x}^\star, \mathbf{y}^\star) \leq f(\mathbf{x}, \mathbf{y}^\star), \quad \text{for all } \mathbf{x} \in \mathbb{R}^m, \ \mathbf{y} \in \mathbb{R}^n$$

Throughout our paper, we assume that the function $f(\mathbf{x}, \mathbf{y})$ is convex in $\mathbf{x}$ for all $\mathbf{y} \in \mathbb{R}^n$ and concave in $\mathbf{y}$ for all $\mathbf{x} \in \mathbb{R}^m$. This *convex-concave* setting has been the focus of intense research in optimization, game theory, economics and computer science for several decades now [Von Neumann and Morgenstern(1953), Dantzig(1963), Blackwell and Girshick(1979), Facchinei and Pang(2007), Ben-Tal et al.(2009)], and variants of the problem have recently attracted significant interest in machine learning and data science, with applications in generative adversarial networks (GANs) [Goodfellow et al.(2014 Arjovsky et al.(2017)], adversarial learning [Sinha et al.(2018)], distributed multi-agent systems [Shamma(2008)], and many other fields; for a wide range of concrete examples, see [Facchinei and Pang(2007)] and references therein.

Motivated by these applications, several classes of optimization algorithms have been proposed and analyzed for finding a global saddle point of Eq. (1) in the convex-concave setting. An important algorithm is the extragradient (EG) method [Korpelevich(1976), Antipin(1978), Nemirovski(2004)]. The method's rate of convergence for smooth and strongly-convex-strongly-concave functions and bilinear functions (i.e., when $f(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top A\mathbf{y}$ for some square, full-rank matrix A) was shown to be linear by [Korpelevich(1976)] and [Tseng(1995)]. Subsequently, [Nemirovski(2004)] showed that the method enjoys an $O(\epsilon^{-1})$ convergence guarantee for constrained problems with a bounded domain and a convex-concave function f. In unbounded domains, [Solodov and Svaiter(1999)] generalized EG to the hybrid proximal extragradient (HPE) method which provides a framework for analyzing the iteration complexity of several existing methods, including EG and Tseng's forward-backward splitting [Tseng(2000)], while [Monteiro and Svaiter(2010)] provided an $O(\epsilon^{-1})$ guarantee for HPE in both bounded and unbounded domains. In addition to EG, there are other methods that can achieve the same convergence guarantees, such as optimistic gradient descent ascent (OGDA) [Popov(1980)] and dual extrapolation (DE) [Nesterov(2007)]; for a partial survey, see [Hsieh et al.(2019)] and the references therein. All these methods are *order-optimal first-order methods* since they match the lower bound of [Ouyang and Xu(2021)].

Focusing on convex *minimization* problems for the moment, significant effort has been devoted to developing first-order methods that are characterized by simplicity of implementation and analytic tractability [Nesterov(1983)]. Indeed, it has been recognized that first-order methods are suitable for solving large-scale machine learning problems where low-accuracy solutions may suffice [Sra et al.(2012), Lan(2020)]. However, second-order methods are known to enjoy superior convergence properties over their first-order counterparts, in both theory and practice for many other problems: the accelerated cubic regularized Newton method [Nesterov(2008)] and accelerated Newton proximal extra-gradient method [Monteiro and Svaiter(2013)] converge at a global rate of $O(\epsilon^{-1/3})$ and $\tilde{O}(\epsilon^{-2/7})$ respectively, both exceeding the best possible $O(\epsilon^{-1/2})$ bound for first-order methods [Nemirovski and Yudin(1983)]. Optimal second-order methods with the rate of $O(\epsilon^{-2/7})$ have been recently proposed by [Carmon et al.(2022)] and [Kovalev and Gasnikov(2022)], independently. In addition, first-order methods may perform poorly in ill-conditioned problems and are known to be sensitive to the parameter choices in real-world applications in which second-order methods are observed to be more robust [Pilanci and Wainwright(2017), Roosta-Khorasani and Mahoney(2019), Berahas et al.(2020)].

In the context of convex-concave min-max problems, two separate issues arise: **(i)** achieving acceleration with second-order information is less tractable analytically; and **(ii)** acquiring accurate second-order information is computationally very expensive in general. Aiming to address these issues, a line of recent work has generalized classical first-order methods to their higher-order counterparts [Monteiro and Svaiter(2012), Bullins and Lai(2022), Jiang and Mokhtari(2022)], where the best known upper iteration bound is $O(\epsilon^{-2/3}\log\log(1/\epsilon))$ [Jiang and Mokhtari(2022)] with a lower bound of $\Omega(\epsilon^{-2/3})$ [Lin and Jordan(2024)]. Closing this gap of $\log\log(1/\epsilon)$ is mainly of theoretical interest but also has important practical implications because all existing methods require a nontrivial *implicit search scheme* at each iteration, and this can be computationally expensive in practice.[1]

In a similar vein, [Huang et al.(2022)] extended the cubic regularized Newton method [Nesterov and Polyak(200 to convex-concave min-max optimization. Their method has two phases and their guarantees require an error bound condition with a parameter $0 < \theta \leq 1$ [Huang et al.(2022), Assumption 5.1]. In particular, the rate is linear under a Lipschitz-type condition ($\theta = 1$) and $O(\epsilon^{-(1-\theta)/\theta^2})$ under a Hölder-type condition ($\theta \in (0, 1)$). These conditions unfortunately exclude some important problem classes and are hard to verify in general. Another extension of cubic regularized Newton method was provided in [Nesterov(2006)] and was shown to achieve a global convergence rate of $O(\epsilon^{-1})$ without assuming any error bound condition.

Finally, it is worth mentioning that existing second-order min-max optimization algorithms require the exact second-order information; as a result, given the implicit nature of the inner loop subproblems involved, the methods' robustness to inexact information cannot be taken for granted. It is thus natural to ask: ***Can we develop explicit second-order min-max optimization algorithms that remain order-optimal even with inexact second-order information?***

Our paper offers an affirmative answer to the above question. Inspired by recent advances on variational inequalities (VIs) [Lin and Jordan(2024)], we start by presenting a second-order min-max optimization method with a global rate of $O(\epsilon^{-2/3})$. Our convergence analysis here is similar to that of [Lin and Jordan(2024)] although it is simpler in that it exploits the structure of unconstrained min-max optimization problems. More importantly, our work differs from [Lin and Jordan(2024)]

[1] By "implicit," we mean that the method's inner-loop subproblem for computing the k^{th} iterate involves the iterate being updated, leading to an implicit update rule. By contrast, "explicit" means that any inner-loop subproblem for computing the k^{th} iterate does not involve the new iterate.

in that the latter requires the exact second-order information and lacks an explicit complexity analysis for inexact subproblem solving.

Thus, the main contribution of our paper is to relax the requirements of existing work by proposing a class of second-order min-max optimization methods that require only *inexact second-order information* and *inexact subproblem solutions.* Moreover, our inexact Jacobian regularity condition allows for the use of randomized sampling for solving finite-sum min-max optimization problems. This yields considerable computational savings since the sample size increases gracefully from a very small sample set. We accordingly prove that the inexact methods achieve a global rate of $O(\epsilon^{-2/3})$ and the subsampled Newton methods achieve the same rate with high probability. Our new subroutine involves solving each subproblem via a single Schur decomposition and $O(\log\log(1/\epsilon))$ calls to a linear system solver in a quasi-upper-triangular system. As such, the total complexity bound of our method is $O((m+n)^{\omega}\epsilon^{-2/3} + (m+n)^2\epsilon^{-2/3}\log\log(1/\epsilon))$ where $\omega \approx 2.3728$ is the matrix multiplication constant.

To the best of our knowledge, our method is the first second-order min-max optimization method that does not require exact second-order information. This improves on the existing works [Lin and Jordan(2024), Adil et al.(2022)], which present solutions to convex-concave min-max optimization problems with order-optimal iteration complexity of $\Theta(\epsilon^{-2/3})$, but either require an *exact* solution of an inner explicit subproblem or lack a characterization of the complexity of solving each subproblem. Indeed, [Adil et al.(2022)] showed that solving each subproblem requires a single Schur decomposition and $O(\log(1/\epsilon))$ calls to a linear system solver in a quasi-upper-triangular system. Our work is also related to a recent literature on the line-search-based methods [Monteiro and Svaiter(2012), Bullins and Lai(2022), Jiang and Mokhtari(2022), Lin and Jordan(2023)]. Differing from these methods, our method is *explicit* and achieves an *order-optimal* iteration complexity: specifically, in terms of complexity bound guarantees, our method slightly outperforms the best known line-search-based method [Jiang and Mokhtari(2022)] that achieves the bound of $O((m+n)^{\omega}\epsilon^{-2/3}\log\log(1/\epsilon))$.

1.1 Related Work

Our work comes amid a surge of interest in optimization algorithms for a large class of emerging min-max optimization problems. For brevity, we will focus on convex-concave settings and leave other settings out of the discussion; see [Lin et al.(2020), Section 2] for a more detailed presentation.

Historically, a concrete instantiation of the convex-concave min-max optimization problem is the solution of $\min_{\mathbf{x}\in\Delta^m}\max_{\mathbf{y}\in\Delta^n}\mathbf{x}^\top A\mathbf{y}$ for $A \in \mathbb{R}^{m\times n}$ over the simplices Δ^m and Δ^n. Spurred by the von Neumann's theorem [Neumann(1928)], this problem provided the initial impetus for min-max optimization. [Sion(1958)] generalized von Neumann's result from bilinear cases to convex-concave cases and triggered a line of research on algorithms for convex-concave min-max optimization [Korpelevich(1976), Nemirovski(2004), Nesterov(2007), Nedić and Ozdaglar(2009), Mokhtari et al.(2020b)]. A notable result is that gradient descent ascent (GDA) with diminishing stepsizes can find an ϵ-global saddle point within $O(\epsilon^{-2})$ iterations if the gradients are bounded over the feasible sets [Nedić and Ozdaglar(2

Recent years have witnessed progress on the analysis of first-order min-max optimization algorithms in bilinear cases and convex-concave cases. In a bilinear case, [Daskalakis et al.(2018)] proved the convergence of OGDA to a neighborhood of a global saddle point. [Liang and Stokes(2019)] used a dynamical system approach to establish the linear convergence of OGDA for the special case when the matrix A is square and full rank. [Mokhtari et al.(2020a)] have revisited proximal point method and proposed an unified framework for achieving the sharpest convergence rates of both EG and OGDA. In the convex-concave case, [Nemirovski(2004)] demonstrated that EG finds an

ϵ-global saddle point within $O(\epsilon^{-1})$ iterations when the feasible sets are convex and compact. The same convergence guarantee was extended to unbounded feasible sets [Monteiro and Svaiter(2010), Monteiro and Svaiter(2011)] using the HPE method with different optimality criteria. [Nesterov(2007)] and [Tseng(2008)] proposed several new algorithms and refined convergence analysis with the same convergence guarantee. [Abernethy et al.(2021)] presented a Hamiltonian gradient descent method with last-iterate convergence under a "sufficiently bilinear" condition. Focusing on special min-max optimization problems with $f(\mathbf{x}, \mathbf{y}) = g(\mathbf{x}) + \mathbf{x}^\top A\mathbf{y} - h(\mathbf{y})$, [Chambolle and Pock(2011)] introduced a primal-dual hybrid gradient method that converges to a global saddle point at the rate of $O(\epsilon^{-1})$ when the functions g and h are simple and convex. [Nesterov(2005)] proposed a smoothing technique and proved that his algorithm achieves better dependence on problem parameters for convex and compact constraint sets. [He and Monteiro(2016)] and [Kolossoski and Monteiro(2017)] proved that such a result still holds true for unbounded feasible sets and non-Euclidean metrics. Moreover, [Chen et al.(2014)] developed optimal algorithms for solving min-max optimization problems with $f(\mathbf{x}, \mathbf{y}) = g(\mathbf{x}) + \mathbf{x}^\top A\mathbf{y} - h(\mathbf{y})$ even when only noisy gradients are available.

Compared to first-order methods, there has been less research on second-order methods for min-max optimization problems *with global convergence rate guarantee*. In particular, we are aware of two research thrusts [Monteiro and Svaiter(2012), Bullins and Lai(2022), Jiang and Mokhtari(2022), Huang et al.(2022), Adil et al.(2022), Lin and Jordan(2023), Lin and Jordan(2024)]. Our results contribute to this landscape by proposing the first explicit method that achieves the order-optimal iteration complexity of $O(\epsilon^{-2/3})$ and a tight complexity of $O((m+n)^\omega \epsilon^{-2/3} + (m+n)^2 \epsilon^{-2/3} \log\log(1/\epsilon))$. As far as we are aware, the complexity bound guarantees of Algorithms 2 and 3 cannot be realized by other existing second-order min-max optimization methods with exact second-order information requirements.

1.2 Notation and Organization

We use bold lower-case letters to denote vectors, as in $\mathbf{x}, \mathbf{y}, \mathbf{z}$. For a function $f(\cdot) : \mathbb{R}^n \to \mathbb{R}$, we let $\nabla f(\mathbf{z})$ denote the gradient of f at $\mathbf{z}$. For a function $f(\cdot, \cdot) : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ of two variables, $\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})$ or $\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$ to denote the partial gradient of f with respect to the first variable or the second variable at the point $(\mathbf{x}, \mathbf{y})$. We use $\nabla f(\mathbf{x}, \mathbf{y})$ to denote the gradient at $(\mathbf{x}, \mathbf{y})$ where $\nabla f(\mathbf{x}, \mathbf{y}) = (\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}), \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}))$ and $\nabla^2 f(\mathbf{x}, \mathbf{y})$ to denote the full Hessian at $(\mathbf{x}, \mathbf{y})$. We write $\|\mathbf{x}\|$ for its ℓ_2-norm. Finally, we use $O(\cdot), \Omega(\cdot)$ to hide absolute constants which do not depend on problem parameters, and $\tilde{O}(\cdot), \tilde{\Omega}(\cdot)$ to hide absolute constants and additional log factors.

The remainder of the paper is organized as follows. In Section 2, we present the setup of min-max optimization and provide the definitions and optimality criteria. In Section 3, we present an explicit yet conceptual second-order min-max optimization method and prove that it achieves an order-optimal iteration complexity of $\Theta(\epsilon^{-2/3})$. In Section 4, we propose a class of second-order min-max optimization methods with inexact second-order information and inexact subproblem solving, and we provide a crisp characterization of the complexity of solving each subproblem in Section 5. In Section 6, we propose a set of subsampled Newton methods for solving the finite-sum min-max optimization problems.

2 Preliminaries

We present the setup of min-max optimization under study, and we provide the definitions for functions as well as optimality criteria considered. In this regard, the regularity conditions that we impose for the function $f : \mathbb{R}^{m+n} \mapsto \mathbb{R}$ are as follows:

Definition 1. *A function f is ρ-Hessian Lipschitz if $\|\nabla^2 f(\mathbf{z}) - \nabla^2 f(\mathbf{z}')\| \leq \rho\|\mathbf{z} - \mathbf{z}'\|$ for all $\mathbf{z}, \mathbf{z}' \in \mathbb{R}^n$.*

Definition 2. *A differentiable function f is convex-concave if*

$$f(\mathbf{x}', \mathbf{y}) \geq f(\mathbf{x}, \mathbf{y}) + (\mathbf{x}' - \mathbf{x})^\top \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}), \text{for } \mathbf{x}', \mathbf{x} \in \mathbb{R}^m \text{ and any fixed } \mathbf{y} \in \mathbb{R}^n$$
$$f(\mathbf{x}, \mathbf{y}') \leq f(\mathbf{x}, \mathbf{y}) + (\mathbf{y}' - \mathbf{y})^\top \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}), \text{for } \mathbf{y}', \mathbf{y} \in \mathbb{R}^n \text{ and any fixed } \mathbf{x} \in \mathbb{R}^m$$

We also define the notion of global saddle points for the problem in Eq. (1).

Definition 3. *A point $\mathbf{z}^\star = (\mathbf{x}^\star, \mathbf{y}^\star) \in \mathbb{R}^m \times \mathbb{R}^n$ is a global saddle point of a function $f(\cdot, \cdot)$ if $f(\mathbf{x}^\star, \mathbf{y}) \leq f(\mathbf{x}^\star, \mathbf{y}^\star) \leq f(\mathbf{x}, \mathbf{y}^\star)$ for all $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$.*

Throughout this paper, we assume that the following conditions are satisfied.

Assumption 1. *The function $f(\mathbf{x}, \mathbf{y})$ is continuously differentiable and convex-concave, and at least one global saddle point of $f(\mathbf{x}, \mathbf{y})$ exists.*

Assumption 2. *The function $f(\mathbf{x}, \mathbf{y})$ is ρ-Hessian Lipschitz.*

The existence of a global saddle point $\mathbf{z}^\star = (\mathbf{x}^\star, \mathbf{y}^\star)$ under Assumption 1 guarantees that $f(\mathbf{x}^\star, \mathbf{y}) \leq f(\mathbf{x}^\star, \mathbf{y}^\star) \leq f(\mathbf{x}, \mathbf{y}^\star)$ for all $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$. Thus, we can adopt a restricted gap function [Nesterov(2007)] to provide a performance measure for the optimality of $\hat{\mathbf{z}} = (\hat{\mathbf{x}}, \hat{\mathbf{y}})$ in the unconstrained convex-concave setting[2].

Definition 4. *The restricted gap function is defined by*

$$\text{gap}(\hat{\mathbf{z}}, \beta) = \max_{\mathbf{y}: \|\mathbf{y} - \mathbf{y}^\star\| \leq \beta} f(\hat{\mathbf{x}}, \mathbf{y}) - \min_{\mathbf{x}: \|\mathbf{x} - \mathbf{x}^\star\| \leq \beta} f(\mathbf{x}, \hat{\mathbf{y}})$$

where β is sufficiently large such that $\|\hat{\mathbf{z}} - \mathbf{z}^\star\| \leq \beta$. Clearly, we have $\text{gap}(\hat{\mathbf{z}}, \beta) \geq 0$ since $f(\mathbf{x}^\star, \hat{\mathbf{y}}) \leq f(\hat{\mathbf{x}}, \mathbf{y}^\star)$ must hold true.

Definition 5. *A point $\hat{\mathbf{z}} = (\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is an ϵ-global saddle point of a convex-concave function $f(\cdot, \cdot)$ if $\text{gap}(\hat{\mathbf{z}}, \beta) \leq \epsilon$. If $\epsilon = 0$, it is a global saddle point.*

In our method, we denote the k^{th} iterate by $(\mathbf{x}_k, \mathbf{y}_k)$ and we define the averaged (ergodic) iterates by $(\bar{\mathbf{x}}_k, \bar{\mathbf{y}}_k)$. In particular, given a sequence of weights $\{\lambda_k\}_{k=1}^T$, we let

$$\bar{\mathbf{x}}_k = \frac{1}{\sum_{i=1}^k \lambda_i} \left(\sum_{i=1}^k \lambda_i \mathbf{x}_i \right), \quad \bar{\mathbf{y}}_k = \frac{1}{\sum_{i=1}^k \lambda_i} \left(\sum_{i=1}^k \lambda_i \mathbf{y}_i \right) \tag{2}$$

In our convergence analysis, we define the vector $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{m+n}$ and the operator $F : \mathbb{R}^{m+n} \mapsto \mathbb{R}^{m+n}$ as follows,

$$F(\mathbf{z}) = \begin{bmatrix} \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) \\ -\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) \end{bmatrix} \tag{3}$$

Accordingly, the Jacobian of F is defined as follows (note that DF is asymmetric in general),

$$DF(\mathbf{z}) = \begin{bmatrix} \nabla^2_{\mathbf{xx}} f(\mathbf{x}, \mathbf{y}) & \nabla^2_{\mathbf{xy}} f(\mathbf{x}, \mathbf{y}) \\ -\nabla^2_{\mathbf{xy}} f(\mathbf{x}, \mathbf{y}) & -\nabla^2_{\mathbf{yy}} f(\mathbf{x}, \mathbf{y}) \end{bmatrix} \in \mathbb{R}^{(m+n)\times(m+n)} \tag{4}$$

In the following lemma, we provide the properties of the operator F in Eq. (3) and its Jacobian DF in Eq. (4) under Assumption 1 and 2. We note that most of the results in the following proposition are well known [Nemirovski(2004)] so we omit their proofs.

[2] The restricted gap is also related to the classical Nikaidô-Isoda function [Nikaidô and Isoda(1955)] defined for a class of noncooperative convex games in a more general setting.

Proposition 1. *Let $F(\cdot)$ and $DF(\cdot)$ be defined as above and Assumptions 1 and 2 hold true. Then, we have*

(a) F is monotone, i.e., $(\mathbf{z} - \mathbf{z}')^\top (F(\mathbf{z}) - F(\mathbf{z}')) \geq 0$.

(b) DF is ρ-Lipschitz continuous, i.e., $\|DF(\mathbf{z}) - DF(\mathbf{z}')\| \leq \rho \|\mathbf{z} - \mathbf{z}'\|$.

(c) $F(\mathbf{z}^\star) = 0$ for any global saddle point $\mathbf{z}^\star \in \mathbb{R}^{m+n}$ of the function $f(\cdot, \cdot)$

Before proceeding to our algorithmic framework and convergence analysis, we present the following well-known result which will be used in the subsequent analysis. Given its importance, we provide the proof for completeness.

Proposition 2. *Let $(\bar{\mathbf{x}}_k, \bar{\mathbf{y}}_k)$ and $F(\cdot)$ be defined in Eq.* (2) *and* (3)*. Then, under Assumption 1, the following statement holds true,*

$$f(\bar{\mathbf{x}}_k, \mathbf{y}) - f(\mathbf{x}, \bar{\mathbf{y}}_k) \leq \frac{1}{\sum_{i=1}^k \lambda_i} \left(\sum_{i=1}^k \lambda_i (\mathbf{z}_i - \mathbf{z})^\top F(\mathbf{z}_i) \right)$$

2.1 Proof of Proposition 1

Proof of Proposition 1. Note that (a) and (c) were proven in [Nemirovski(2004)], and it suffices to prove (b). By using the definition of $DF(\cdot)$ in Eq. (4), we have

$$(DF(\mathbf{z}) - DF(\mathbf{z}'))\mathbf{h} = \begin{bmatrix} I_m & \\ & -I_n \end{bmatrix} (\nabla^2 f(\mathbf{z}) - \nabla^2 f(\mathbf{z}'))\mathbf{h} \tag{5}$$

This implies that $\|(DF(\mathbf{z}) - DF(\mathbf{z}'))\mathbf{h}\| = \|(\nabla^2 f(\mathbf{z}) - \nabla^2 f(\mathbf{z}'))\mathbf{h}\|$. Thus, we have

$$\|DF(\mathbf{z}) - DF(\mathbf{z}')\| = \sup_{\mathbf{h} \neq 0} \left\{ \frac{\|(DF(\mathbf{z}) - DF(\mathbf{z}'))\mathbf{h}\|}{\|\mathbf{h}\|} \right\} = \sup_{\mathbf{h} \neq 0} \left\{ \frac{\|(\nabla^2 f(\mathbf{z}) - \nabla^2 f(\mathbf{z}'))\mathbf{h}\|}{\|\mathbf{h}\|} \right\}$$

This equality together with Assumption 2 implies the desired result in (b). □

2.2 Proof of Proposition 2

Proof of Proposition 2. Using the definition of the operator $F(\cdot)$ in Eq. (3), we have

$$\sum_{i=1}^k \lambda_i (\mathbf{z}_i - \mathbf{z})^\top F(\mathbf{z}_i) = \sum_{i=1}^k \lambda_i ((\mathbf{x}_i - \mathbf{x})^\top \nabla_{\mathbf{x}} f(\mathbf{x}_i, \mathbf{y}_i) - (\mathbf{y}_i - \mathbf{y})^\top \nabla_{\mathbf{y}} f(\mathbf{x}_i, \mathbf{y}_i))$$

Note that Assumption 1 implies that the function $f(\mathbf{x}, \mathbf{y})$ is a convex function of $\mathbf{x}$ for any $\mathbf{y} \in \mathbb{R}^n$ and a concave function of $\mathbf{y}$ for any $\mathbf{x} \in \mathbb{R}^m$. Then, we have

$$\begin{array}{rcl} (\mathbf{x}_i - \mathbf{x})^\top \nabla_{\mathbf{x}} f(\mathbf{x}_i, \mathbf{y}_i) & \geq & f(\mathbf{x}_i, \mathbf{y}_i) - f(\mathbf{x}, \mathbf{y}_i), \\ (\mathbf{y}_i - \mathbf{y})^\top \nabla_{\mathbf{y}} f(\mathbf{x}_i, \mathbf{y}_i) & \leq & f(\mathbf{x}_i, \mathbf{y}_i) - f(\mathbf{x}_i, \mathbf{y}) \end{array}$$

Putting these pieces together with $\lambda_i > 0$ for all $1 \leq i \leq k$ yields that

$$\frac{1}{\sum_{i=1}^k \lambda_i} \left(\sum_{i=1}^k \lambda_i (\mathbf{z}_i - \mathbf{z})^\top F(\mathbf{z}_i) \right) \geq \frac{1}{\sum_{i=1}^k \lambda_i} \left(\sum_{i=1}^k \lambda_i (f(\mathbf{x}_i, \mathbf{y}) - f(\mathbf{x}, \mathbf{y}_i)) \right) \tag{6}$$

```
Algorithm 1 Newton-MinMax($\mathbf{z}_0$, $\rho$, $T$)
  Input: initial point $\mathbf{z}_0$, Lipschitz parameter $\rho$ and iteration number $T \geq 1$
  Initialization: set $\hat{\mathbf{z}}_0 = \mathbf{z}_0$
  for $k = 0, 1, 2, \ldots, T-1$ do
    STEP 1: If $\mathbf{z}_k$ is a global saddle point of the problem in Eq. (1), then stop
    STEP 2: Compute an exact solution $\Delta\mathbf{z}_k$ of the subproblem in Eq. (7)
    STEP 3: Compute $\lambda_{k+1} > 0$ such that $\frac{1}{33} \leq \lambda_{k+1}\rho\|\Delta\mathbf{z}_k\| \leq \frac{1}{13}$
    STEP 4: Compute $\mathbf{z}_{k+1} = \hat{\mathbf{z}}_k + \Delta\mathbf{z}_k$
    STEP 5: Compute $\hat{\mathbf{z}}_{k+1} = \hat{\mathbf{z}}_k - \lambda_{k+1}F(\mathbf{z}_{k+1})$
  end for
  Output: $\bar{\mathbf{z}}_T = \frac{1}{\sum_{k=1}^T \lambda_k}\left(\sum_{k=1}^T \lambda_k \mathbf{z}_k\right)$
```

Using the definition of $(\bar{\mathbf{x}}_k, \bar{\mathbf{y}}_k)$ in Eq. (2) and that f is convex-concave, we have

$$\frac{1}{\sum_{i=1}^k \lambda_i}\left(\sum_{i=1}^k \lambda_i f(\mathbf{x}_i, \mathbf{y})\right) \geq f(\bar{\mathbf{x}}_k, \mathbf{y}), \qquad \frac{1}{\sum_{i=1}^k \lambda_i}\left(\sum_{i=1}^k \lambda_i f(\mathbf{x}, \mathbf{y}_i)\right) \leq f(\mathbf{x}, \bar{\mathbf{y}}_k)$$

Plugging these two inequalities in Eq. (6), we conclude the desired inequality. □

3 Conceptual Algorithm and Convergence Analysis

As a warm-up, we describe the scheme of Newton-MinMax which is a second-order version of the method in [Lin and Jordan(2024)] for min-max optimization and which yields an optimal global rate of $\Theta(\epsilon^{-2/3})$. We emphasize that Newton-MinMax is a *conceptual* algorithmic framework in the sense that it requires exact second-order information and requires the cubic regularized subproblem to be solved exactly.

3.1 Algorithmic scheme

We summarize our second-order method, which we call Newton-MinMax($\mathbf{z}_0$, ρ, T), in Algorithm 1 where $\mathbf{z}_0 = (\mathbf{x}_0, \mathbf{y}_0) \in \mathbb{R}^m \times \mathbb{R}^n$ is an initial point, $\rho > 0$ is a Lipschitz constant for the Hessian of the function f and $T \geq 1$ is an iteration number.

Our method is a generalization of first-order extragradient method. Indeed, the k^{th} iteration consists of two important algorithmic components:

- **Gradient update:** Compute $\Delta\mathbf{z}_k \in \mathbb{R}^m \times \mathbb{R}^n$ such that it is an exact solution of the *nonlinear equation* problem given by

  $$F(\hat{\mathbf{z}}_k) + DF(\hat{\mathbf{z}}_k)\Delta\mathbf{z}_k + 6\rho\|\Delta\mathbf{z}_k\|\Delta\mathbf{z}_k = \mathbf{0} \tag{7}$$

  Compute $\mathbf{z}_{k+1} = \hat{\mathbf{z}}_k + \Delta\mathbf{z}_k$.

- **Extragradient update:** Compute $\hat{\mathbf{z}}_{k+1} = \hat{\mathbf{z}}_k - \lambda_{k+1}F(\mathbf{z}_{k+1})$

As suggested by [Lin and Jordan(2024)], we choose to update λ_k in an adaptive manner and then prove that our method can achieve an order-optimal iteration complexity of $O(\epsilon^{-2/3})$ under Assumptions 1 and 2. Intuitively, such a strategy makes sense; indeed, λ_k is the step size and would be better to increase as the iterate $\mathbf{z}_k$ approaches the set of global saddle points where the value of $\|\Delta \mathbf{z}_k\|$ measures the closeness. From a practical viewpoint, Algorithm 1 serves as an alternative to the current pipeline of line-search-based methods – which it simplifies by removing the need for an implicit binary search.

3.2 Convergence analysis

We provide our results on the iteration complexity of Algorithm 1 in the following theorem.

Theorem 1. *Suppose that Assumptions 1 and 2 hold. Then, the sequence of iterates generated by Algorithm 1 is bounded and, in addition*

$$\text{gap}(\bar{\mathbf{z}}_T, \beta) \leq \frac{2112\sqrt{3}\rho\|\mathbf{z}_0 - \mathbf{z}^\star\|^3}{T^{3/2}}$$

where $\mathbf{z}^\star = (\mathbf{x}^\star, \mathbf{y}^\star)$ is a global saddle point, $\rho > 0$ is defined as in Assumption 2, and $\beta = 7\|\mathbf{z}_0 - \mathbf{z}^\star\|$. As such, Algorithm 1 achieves an ϵ-global saddle point solution within $O(\epsilon^{-2/3})$ iterations.

Remark. Theorem 1 shows that Algorithm 1 has achieved the lower bound established in the literature for second-order VI methods [Lin and Jordan(2024)] and is thus order-optimal in this regard; in addition, it improves on the state-of-the-art bounds of [Monteiro and Svaiter(2012), Bullins and Lai(2022), Jiang and Mokhtari(2022)] by shaving off all logarithmic factors

Remark. Although Algorithm 1 is conceptual, it forms the basis for the material in the next section, where we relax the strong requirements of Algorithm 1 and propose a class of second-order min-max optimization methods that require only inexact second-order information and inexact subproblem solutions

3.3 Proof of Theorem 1

We define a Lyapunov function for Algorithm 1: $\mathcal{E}_t = \frac{1}{2}\|\hat{\mathbf{z}}_t - \mathbf{z}_0\|^2$ and use it to prove technical results that pertain to the convergence of Algorithm 1. The first lemma gives us a key descent inequality.

Lemma 1. *Suppose that Assumptions 1 and 2 hold. Then*

$$\sum_{k=1}^{t} \lambda_k (\mathbf{z}_k - \mathbf{z})^\top F(\mathbf{z}_k) \leq \mathcal{E}_0 - \mathcal{E}_t + (\mathbf{z}_0 - \hat{\mathbf{z}}_t)^\top (\mathbf{z}_0 - \mathbf{z}) - \frac{1}{24}\left(\sum_{k=1}^{t} \|\mathbf{z}_k - \hat{\mathbf{z}}_{k-1}\|^2\right)$$

This helps us prove the following lemma

Lemma 2. *Suppose that Assumptions 1 and 2 hold. Then, we have $\|\hat{\mathbf{z}}_t - \mathbf{z}_0\| \leq 2\|\mathbf{z}_0 - \mathbf{z}^\star\|$ and*

$$\sum_{k=1}^{t} \lambda_k (\mathbf{z}_k - \mathbf{z})^\top F(\mathbf{z}_k) \leq \frac{1}{2}\|\mathbf{z}_0 - \mathbf{z}\|^2, \quad \sum_{k=1}^{t} \|\mathbf{z}_k - \hat{\mathbf{z}}_{k-1}\|^2 \leq 12\|\mathbf{z}_0 - \mathbf{z}^\star\|^2$$

where $\mathbf{z} \in \mathbb{R}^{m+n}$ is any point and $\mathbf{z}^\star$ is a global saddle point

Lemma 3. *Suppose that Assumptions 1 and 2 hold. Then, for every integer $T \geq 1$, we have*

$$\sum_{k=1}^{T} \lambda_k \geq \frac{T^{\frac{3}{2}}}{66\sqrt{3}\rho\|\mathbf{z}_0 - \mathbf{z}^\star\|}$$

where $\mathbf{z}^\star$ is a global saddle point

Proof of Theorem 1. By Lemma 2, we have

$$\|\mathbf{z}_{k+1} - \hat{\mathbf{z}}_k\|^2 \leq 12\|\mathbf{z}_0 - \mathbf{z}^\star\|^2, \quad \|\hat{\mathbf{z}}_k - \mathbf{z}_0\| \leq 2\|\mathbf{z}_0 - \mathbf{z}^\star\|, \quad \text{for all } k \geq 0$$

This implies that $\|\mathbf{z}_k - \mathbf{z}_0\| \leq 6\|\mathbf{z}_0 - \mathbf{z}^\star\|$ for all $k \geq 0$. Putting these pieces together yields that both $\{\mathbf{z}_k\}_{k\geq 0}$ and $\{\hat{\mathbf{z}}_k\}_{k\geq 0}$ are bounded by a constant; indeed, we have $\|\hat{\mathbf{z}}_k - \mathbf{z}^\star\| \leq 3\|\mathbf{z}_0 - \mathbf{z}^\star\| \leq \beta$ and $\|\mathbf{z}_k - \mathbf{z}^\star\| \leq 7\|\mathbf{z}_0 - \mathbf{z}^\star\| = \beta$. For every integer $T \geq 1$, Lemma 2 also implies

$$\sum_{k=1}^{T} \lambda_k (\mathbf{z}_k - \mathbf{z})^\top F(\mathbf{z}_k) \leq \frac{1}{2}\|\mathbf{z}_0 - \mathbf{z}\|^2$$

By Proposition 2, we have

$$f(\bar{\mathbf{x}}_T, \mathbf{y}) - f(\mathbf{x}, \bar{\mathbf{y}}_T) \leq \frac{1}{\sum_{k=1}^{T} \lambda_k}\left(\sum_{k=1}^{T} \lambda_k (\mathbf{z}_k - \mathbf{z})^\top F(\mathbf{z}_k)\right)$$

Putting these pieces together yields

$$f(\bar{\mathbf{x}}_T, \mathbf{y}) - f(\mathbf{x}, \bar{\mathbf{y}}_T) \leq \frac{1}{2(\sum_{k=1}^{T} \lambda_k)}\|\mathbf{z}_0 - \mathbf{z}\|^2$$

This together with Lemma 3 yields

$$f(\bar{\mathbf{x}}_T, \mathbf{y}) - f(\mathbf{x}, \bar{\mathbf{y}}_T) \leq \frac{33\sqrt{3}\rho\|\mathbf{z}_0 - \mathbf{z}^\star\|\|\mathbf{z}_0 - \mathbf{z}\|^2}{T^{3/2}}$$

Since $\|\mathbf{z}_k - \mathbf{z}^\star\| \leq \beta$ for all $k \geq 0$, we have $\|\bar{\mathbf{z}}_T - \mathbf{z}^\star\| \leq \beta$. By the definition of the restricted gap function, we have

$$\text{gap}(\bar{\mathbf{z}}_T, \beta) \leq \frac{33\sqrt{3}\rho\|\mathbf{z}_0 - \mathbf{z}^\star\|(\|\mathbf{z}_0 - \mathbf{z}^\star\| + \beta)^2}{T^{3/2}} = \frac{2112\sqrt{3}\rho\|\mathbf{z}_0 - \mathbf{z}^\star\|^3}{T^{3/2}}$$

Therefore, we conclude from the above inequality that there exists some $T > 0$ such that the output $\hat{\mathbf{z}} = \textsf{Newton-MinMax}(\mathbf{z}_0, \rho, T)$ satisfies that $\text{gap}(\hat{\mathbf{z}}, \beta) \leq \epsilon$ and the total number of iterations is bounded by $O(\rho^{2/3}\|\mathbf{z}_0 - \mathbf{z}^\star\|^2\epsilon^{-2/3})$. □

4 Inexact Algorithm and Complexity Analysis

Building on the conceptual algorithm of the previous section, we present the Inexact-Newton-MinMax scheme, and we provide a global convergence guarantee in terms of the number of iterations required until convergence. Our inexact Jacobian regularity condition is inspired by [Xu et al.(2020)] and it allows for the use of randomized sampling for solving finite-sum min-max optimization problems. Our subroutine, which is inspired by [Adil et al.(2022)], can solve each subproblem using a single Schur decomposition and $O(\log\log(1/\epsilon))$ calls to a linear system solver in a quasi-upper-triangular system.

Algorithm 2 Inexact-Newton-MinMax($\mathbf{z}_0$, ρ, T)

```
**Input:** initial point $\mathbf{z}_0$, Lipschitz parameter $\rho$ and iteration number $T \geq 1$
**Initialization:** set $\hat{\mathbf{z}}_0 = \mathbf{z}_0$ as well as $\kappa_J > 0$, $0 < \kappa_m < \min\{1, \frac{\rho}{4}\}$ and $0 \leq \tau_0 < \frac{\rho}{4}$
**for** $k = 0, 1, 2, \ldots, T-1$ **do**
  **STEP 1:** If $\mathbf{z}_k$ is a global saddle point of the problem in Eq. (1), then **stop**
  **STEP 2:** Compute an *inexact* solution $\Delta\mathbf{z}_k$ of the subproblem
      $F(\hat{\mathbf{z}}_k) + J(\hat{\mathbf{z}}_k)\Delta\mathbf{z}_k + 6\rho\|\Delta\mathbf{z}_k\|\Delta\mathbf{z}_k = \mathbf{0}$     (8)
  such that Condition 4.1 and 4.2 hold true with a proper choice of $\tau_k \geq 0$
  **STEP 3:** Compute $\lambda_{k+1} > 0$ such that $\frac{1}{30} \leq \lambda_{k+1}\rho\|\Delta\mathbf{z}_k\| \leq \frac{1}{14}$
  **STEP 4:** Compute $\mathbf{z}_{k+1} = \hat{\mathbf{z}}_k + \Delta\mathbf{z}_k$
  **STEP 5:** Compute $\hat{\mathbf{z}}_{k+1} = \hat{\mathbf{z}}_k - \lambda_{k+1}F(\mathbf{z}_{k+1})$
**end for**
**Output:** $\bar{\mathbf{z}}_T = \frac{1}{\sum_{k=1}^T \lambda_k}\left(\sum_{k=1}^T \lambda_k\mathbf{z}_k\right)$
```

Algorithmic Scheme. We summarize our inexact second-order method, which we call Inexact-Newton-MinMax($\mathbf{z}_0$, ρ, T), in Algorithm 2 where $\mathbf{z}_0$ is an initial point, $\rho > 0$ is a Lipschitz constant for the Hessian of the function f and $T \geq 1$ is an iteration number.

Our method combines Algorithm 1 with an inexact second-order framework [Xu et al.(2020)] in the context of min-max optimization. Indeed, the key difference between Eq. (7) and Eq. (8) is that the inexact Jacobian $J(\hat{\mathbf{z}}_k)$ is used to approximate the exact Jacobian at $\hat{\mathbf{z}}_k$ and can be formed and evaluated efficiently in practice. Throughout this section, we impose the following two conditions on the inexact Jacobian construction and the inexact subproblem solving.

Condition 4.1 (Inexact Jacobian regularity)**.** *For some $\kappa_J > 0$ and $\tau_k \geq 0$, the inexact Jacobian $J(\hat{\mathbf{z}}_k)$ satisfies the following regularity conditions:*

$$\|(J(\hat{\mathbf{z}}_k) - DF(\hat{\mathbf{z}}_k))\Delta\mathbf{z}_k\| \leq \tau_k\|\Delta\mathbf{z}_k\|, \qquad \|J(\hat{\mathbf{z}}_k)\| \leq \kappa_J$$

where the iterates $\{\hat{\mathbf{z}}_k\}_{k\geq 0}$ and the updates $\{\Delta\mathbf{z}_k\}_{k\geq 0}$ are generated by Algorithm 2

Condition 4.2 (Sufficient inexact solving)**.** *Fixing $\kappa_m \in (0,1)$, we can solve the nonlinear equation problem in Eq.* (8) *inexactly to find $\Delta\mathbf{z}_k$ such that*

$$\|F(\hat{\mathbf{z}}_k) + J(\hat{\mathbf{z}}_k)\Delta\mathbf{z}_k + 6\rho\|\Delta\mathbf{z}_k\|\Delta\mathbf{z}_k\| \leq \kappa_m \cdot \min\{\|\Delta\mathbf{z}_k\|^2, \|F(\hat{\mathbf{z}}_k)\|\}$$

In addition, $\{\hat{\mathbf{z}}_k\}_{k\geq 0}$ and $\{\Delta\mathbf{z}_k\}_{k\geq 0}$ are generated by Algorithm 2

Under Conditions 4.1 and 4.2, our proposed algorithm (cf. Algorithm 2) achieves the same worst-case iteration complexity of $O(\epsilon^{-2/3})$ for computing an ϵ-global saddle point of the problem in Eq. (1) as that of the exact variant (cf. Algorithm 1).

Notably, Condition 4.1 allows for the principled use of many practical techniques to constructing the inexact Jacobian $J(\hat{\mathbf{z}}_k)$ in the context of min-max optimization. One such scheme can be described for solving the *finite-sum* min-max optimization problems in the form of

$$\min_{\mathbf{x}\in\mathbb{R}^m}\max_{\mathbf{y}\in\mathbb{R}^n} f(\mathbf{x}, \mathbf{y}) \triangleq \frac{1}{N}\sum_{i=1}^N f_i(\mathbf{x}, \mathbf{y}) \tag{9}$$

and its special instantiation

$$\min_{\mathbf{x}\in\mathbb{R}^m}\max_{\mathbf{y}\in\mathbb{R}^n}\ f(\mathbf{x},\mathbf{y}) \triangleq \frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{a}_i^\top\mathbf{x}, \mathbf{b}_i^\top\mathbf{y}) \tag{10}$$

where $N \gg 1$, each of f_i is a convex-concave function with bounded and ρ-Lipschitz Hessian, and $\{(\mathbf{a}_i, \mathbf{b}_i)\}_{i=1}^N \subseteq \mathbb{R}^m \times \mathbb{R}^n$ are a collection of data samples.

This type of problem is common in the context of optimization and machine learning [Shalev-Shwartz and Ben-D Roosta-Khorasani et al.(2014), Roosta-Khorasani and Mahoney(2019)]. Thanks to its finite-sum structure, both uniform and nonuniform subsampling schemes satisfy Condition 4.1 with high probability; indeed, we have the stronger theoretical guarantee $\|J(\hat{\mathbf{z}}_k) - DF(\hat{\mathbf{z}}_k)\| \leq \tau_k$, which implies one of two inequalities in Condition 4.1. As a consequence of Theorem 2, our subsampled Newton method achieves the order-optimal iteration complexity of $\Theta(\epsilon^{-2/3})$ for solving the finite-sum convex-concave min-max optimization problems; see Theorem 4 for the details.

Convergence Analysis. We provide our results on the iteration complexity of Algorithm 2 in the following theorem.

Theorem 2. *Suppose that Assumption 1 and 2 hold and*

$$0 \leq \tau_k \leq \min\{\tau_0, \frac{\rho(1-\kappa_m)}{4(\kappa_J + 6\rho)}\|F(\hat{\mathbf{z}}_k)\|\}, \text{ for all } k \geq 0$$

Then, the iterates generated by Algorithm 2 are bounded and, in addition,

$$\text{gap}(\bar{\mathbf{z}}_T, \beta) \leq \frac{2112\sqrt{3}\rho\|\mathbf{z}_0 - \mathbf{z}^\star\|^3}{T^{3/2}}$$

where $\mathbf{z}^\star = (\mathbf{x}^\star, \mathbf{y}^\star)$ is a global saddle point, $\rho > 0$ is defined in Assumption 2, and $\beta = 7\|\mathbf{z}_0 - \mathbf{z}^\star\|$. As such, Algorithm 2 achieves an ϵ-global saddle point solution within $O(\epsilon^{-2/3})$ iterations.

Remark. Theorem 2 shows that Algorithm 2 achieves the same iteration complexity as Algorithm 1 and is thus order-optimal regardless of inexact second-order information and inexact subproblem solving under Conditions 4.1 and 4.2.

4.1 Proof of Theorem 2

In the subsequent analysis, we use the same Lyapunov function: $\mathcal{E}_t = \frac{1}{2}\|\hat{\mathbf{z}}_t - \mathbf{z}_0\|^2$. The first lemma gives a descent inequality which is analogous to that in Lemma 1.

Lemma 4. *Suppose that Assumption 1 and 2 hold and*

$$0 < \tau_k \leq \min\{\tau_0, \frac{\rho(1-\kappa_m)}{4(\kappa_J + 6\rho)}\|F(\hat{\mathbf{z}}_k)\|\}, \text{ for all } k \geq 0$$

Then, we have

$$\sum_{k=1}^{t} \lambda_k(\mathbf{z}_k - \mathbf{z})^\top F(\mathbf{z}_k) \leq \mathcal{E}_0 - \mathcal{E}_t + (\mathbf{z}_0 - \hat{\mathbf{z}}_t)^\top(\mathbf{z}_0 - \mathbf{z}) - \frac{1}{24}\left(\sum_{k=1}^{t}\|\mathbf{z}_k - \hat{\mathbf{z}}_{k-1}\|^2\right)$$

Proof of Theorem 2. Since the descent inequalities in Lemmas 1 and 4 are the same, Lemma 2 and 3 also hold true for Algorithm 2. As such, we can apply the same argument used for proving Theorem 1 and have

$$\|\hat{\mathbf{z}}_k - \mathbf{z}^\star\| \leq 3\|\mathbf{z}_0 - \mathbf{z}^\star\| \leq \beta, \quad \|\mathbf{z}_k - \mathbf{z}^\star\| \leq 7\|\mathbf{z}_0 - \mathbf{z}^\star\| = \beta$$

and

$$\operatorname{gap}(\bar{\mathbf{z}}_T, \beta) \leq \frac{2112\sqrt{3}\rho\|\mathbf{z}_0 - \mathbf{z}^\star\|^3}{T^{3/2}}$$

Therefore, we conclude from the above inequality that there exists some $T > 0$ such that the output $\hat{\mathbf{z}} = \mathsf{Inexact\text{-}Newton\text{-}MinMax}(\mathbf{z}_0, \rho, T)$ satisfies that $\operatorname{gap}(\hat{\mathbf{z}}, \beta) \leq \epsilon$ and the total number of iterations is bounded by $O(\rho^{2/3}\|\mathbf{z}_0 - \mathbf{z}^\star\|^2\epsilon^{-2/3})$. □

5 Inexact Subproblem Solving and Complexity Analysis

We clarify how to obtain an inexact solution of the nonlinear equation problem in Eq. (8) such that Condition 4.2 holds. To that end, we present a new subroutine that solves each subproblem using a single Schur decomposition and $O(\log\log(1/\epsilon))$ calls to a linear system solver in a quasi-upper-triangular system. This gives a total complexity of $O((m+n)^\omega\epsilon^{-2/3} + (m+n)^2\epsilon^{-2/3}\log\log(1/\epsilon))$ which outperforms that of $O((m+n)^\omega\epsilon^{-2/3}\log\log(1/\epsilon))$ achieved by the best known line-search-based method.

For ease of presentation, we omit the subscript and rewrite the nonlinear equation problem in Eq. (8) as

$$F(\hat{\mathbf{z}}) + J(\hat{\mathbf{z}})\Delta\mathbf{z} + 6\rho\|\Delta\mathbf{z}\|\Delta\mathbf{z} = \mathbf{0}$$

This is then equivalent to finding a pair $(\Delta\mathbf{z}, \lambda)$ for which

$$(J(\hat{\mathbf{z}}) + \lambda I)\Delta\mathbf{z} = -F(\hat{\mathbf{z}}), \quad \lambda = 6\rho\|\Delta\mathbf{z}\| \tag{11}$$

Although $J(\hat{\mathbf{z}})$ is not symmetric, it has a Schur decomposition $J(\hat{\mathbf{z}}) = QUQ^{-1}$ where U is quasi-upper-triangular (since all entries of $J(\hat{\mathbf{z}})$ are real) in the sense that it is a block diagonal matrix with block size at most 2×2 and Q is a unitary matrix. Then, we have $J(\hat{\mathbf{z}}) + \lambda I = Q(U + \lambda I)Q^{-1}$. In the convex-concave setting, $J(\hat{\mathbf{z}})$ is positive semidefinite and thus $J(\hat{\mathbf{z}}) + \lambda I$ is positive definite.

In this regard, we define

$$\Delta\mathbf{z}(\lambda) \triangleq -(J(\hat{\mathbf{z}}) + \lambda I)^{-1}F(\hat{\mathbf{z}}) = -Q(U + \lambda I)^{-1}Q^{-1}F(\hat{\mathbf{z}}) \tag{12}$$

and obtain from Eq. (11) and the above definition that the solution of $(\Delta\mathbf{z}, \lambda)$ we are looking for depends upon the nonlinear equality $\lambda = 6\rho\|\Delta\mathbf{z}(\lambda)\|$. For convenience, we define $\psi(\lambda) = \|\Delta\mathbf{z}(\lambda)\|^2$ and examine $\psi(\lambda)$ in the following proposition.

Proposition 3. *The first-order and second-order derivatives of ψ are given by*

$$\psi'(\lambda) = -2\Delta\mathbf{z}(\lambda)^\top(J(\hat{\mathbf{z}}) + \lambda I)^{-1}\Delta\mathbf{z}(\lambda)$$

and

$$\psi''(\lambda) = 2\|(J(\hat{\mathbf{z}}) + \lambda I)^{-1}\Delta\mathbf{z}(\lambda)\|^2 + 4\Delta\mathbf{z}(\lambda)^\top(J(\hat{\mathbf{z}}) + \lambda I)^{-2}\Delta\mathbf{z}(\lambda)$$

If $F(\hat{\mathbf{z}}) \neq \mathbf{0}$, the function $\psi(\lambda)$ is strictly decreasing and convex when $\lambda > 0$

The nonlinear equality $\lambda = 6\rho\|\Delta\mathbf{z}(\lambda)\|$ is equivalent to $\lambda = 6\rho\sqrt{\psi(\lambda)}$ which can be reformulated as the following one-dimensional nonlinear equation problem:

$$\phi(\lambda) \triangleq \sqrt{\psi(\lambda)} - \frac{\lambda}{6\rho} = 0 \tag{13}$$

In the following proposition, we examine $\phi(\lambda)$ and prove several key properties.

Proposition 4. *Suppose $F(\hat{\mathbf{z}}) \neq \mathbf{0}$. Then, we have the function $\phi(\lambda)$ is strictly decreasing and convex when $\lambda > 0$. Its first-order derivative is*

$$\phi'(\lambda) = -\frac{\Delta\mathbf{z}(\lambda)^\top (J(\hat{\mathbf{z}}) + \lambda I)^{-1}\Delta\mathbf{z}(\lambda)}{\|\Delta\mathbf{z}(\lambda)\|} - \frac{1}{6\rho}$$

The required solution is the unique positive root to Eq. (13) since $\phi(\lambda)$ is decreasing. Let $\lambda^0 > 0$ be given with $\phi(\lambda^0) > 0$, we first perform one Schur decomposition of $J(\hat{\mathbf{z}})$:

$$J(\hat{\mathbf{z}}) = QUQ^{-1}$$

A typical iteration of the unit-stepsize Newton method for finding such root replaces the current iterate $\lambda^j > 0$ with the improved estimate λ^{j+1} for which

$$\lambda^{j+1} = \lambda^j - \frac{\phi(\lambda^j)}{\phi'(\lambda^j)} \tag{14}$$

The value of $\phi(\lambda^j)$ can be obtained by solving $\Delta\mathbf{z}(\lambda^j) = -Q(U + \lambda^j I)^{-1}Q^{-1}F(\hat{\mathbf{z}})$, and that of $\phi'(\lambda^j)$ can be obtained using Proposition 4 once $\Delta\mathbf{z}(\lambda^j)$ is available (solving a linear system is required). This implies that $\phi(\lambda^j)$ and $\phi'(\lambda^j)$ can be computed by two calls to a linear system solver in a quasi-upper-triangle system.

In terms of complexity, the cost of a single Schur decomposition is $O((m+n)^\omega)$ where $\omega \approx 2.3728$ is the matrix multiplication constant and the cost of one Newton iteration is $O((m+n)^2)$. In terms of convergence, the unit-stepsize Newton method by itself is not a reliable method and the iterates do not converge in general. However, the convexity of ϕ established in Proposition 4 has useful properties.

Theorem 3. *Suppose that the iterates $\{\lambda^j\}_{j\geq 0}$ are generated by a unit-stepsize Newton method with $\lambda^0 > 0$ with $\phi(\lambda^0) \geq 0$. Then, we have $\lambda^j > 0$ and $\phi(\lambda^j) \geq 0$. As a consequence, the iterates converge monotonically towards the unique solution $\lambda^\star$. The convergence rate is globally Q-linear with a factor at least $1 - \phi'(\lambda^\star)/\phi'(\lambda^0)$ and is ultimately Q-quadratic.*

Remark. Our approach is inspired by [Adil et al.(2022)], who reformulated the subproblem in their algorithm as the nonlinear equality $\lambda = 6\rho\|\Delta\mathbf{z}(\lambda)\|$ and proposed a bisection method as a subroutine. In total, this requires a single Schur decomposition and $O(\log(1/\epsilon))$ calls to a linear system solver for a quasi-upper-triangular system. Our key finding is that the Newton method is better than the bisection method for solving the nonlinear equation $\lambda = 6\rho\|\Delta\mathbf{z}(\lambda)\|$. Indeed, the simple reformulation in Eq. (13) and the unit-stepsize implementation achieve global linear and local quadratic convergence guarantees. Thus, by comparison, the proposed subroutine only requires a single Schur decomposition and $O(\log\log(1/\epsilon))$ calls to a linear system solver for a quasi-upper-triangular system

Remark. We note also that [Huang et al.(2022)] reformulated the subproblem in their algorithm as the *two-dimensional* nonlinear equation problem (see [Huang et al.(2022), Eq.(16)]) and proposed the Newton method as a subroutine. However, no theoretical guarantee was provided, even in a local sense; indeed, the proposed nonlinear equation is complex and it is not clear if it satisfies essential properties (e.g., the nonsingularity of Jacobian). Our approach is similar in spirit to [Conn et al.(2000)]; they consider the choice of $\phi \triangleq 1/\sqrt{\psi(\lambda)} - 6\rho/\lambda$ and prove that it is concave in optimization setting where the second-order information is a symmetric matrix. Their analysis is unfortunately not valid here since $J(\hat{\mathbf{z}})$ is, in general, not symmetric

5.1 Proof of Proposition 3

Proof of Proposition 3. Since $\psi(\lambda) = \|\Delta\mathbf{z}(\lambda)\|^2$, we have

$$\psi'(\lambda) = 2\Delta\mathbf{z}(\lambda)^\top \nabla_\lambda \Delta\mathbf{z}(\lambda), \quad \psi''(\lambda) = 2\|\nabla_\lambda \Delta\mathbf{z}(\lambda)\|^2 + 2\Delta\mathbf{z}(\lambda)^\top \nabla^2_{\lambda\lambda}\Delta\mathbf{z}(\lambda)$$

By differentiating the equation $(J(\hat{\mathbf{z}}) + \lambda I)\Delta\mathbf{z}(\lambda) = -F(\hat{\mathbf{z}})$, we have

$$(J(\hat{\mathbf{z}}) + \lambda I)\nabla_\lambda \Delta\mathbf{z}(\lambda) + \Delta\mathbf{z}(\lambda) = \mathbf{0}, \quad (J(\hat{\mathbf{z}}) + \lambda I)\nabla^2_{\lambda\lambda}\Delta\mathbf{z}(\lambda) + 2\nabla_\lambda \Delta\mathbf{z}(\lambda) = \mathbf{0}$$

Rearranging the first equation implies that

$$\nabla_\lambda \Delta\mathbf{z}(\lambda) = -(J(\hat{\mathbf{z}}) + \lambda I)^{-1}\Delta\mathbf{z}(\lambda)$$

Combining the second equation with the expression of $\nabla_\lambda \Delta\mathbf{z}(\lambda)$ yields

$$\nabla^2_{\lambda\lambda}\Delta\mathbf{z}(\lambda) = -2(J(\hat{\mathbf{z}}) + \lambda I)^{-2}\Delta\mathbf{z}(\lambda)$$

Putting these pieces together yields the desired expressions of $\psi'(\lambda)$ and $\psi''(\lambda)$. Since $F(\hat{\mathbf{z}}) \neq \mathbf{0}$ and $J(\hat{\mathbf{z}})$ is positive semidefinite, we have $\psi'(\lambda) < 0$ and $\psi''(\lambda) \geq 0$ when $\lambda > 0$. This completes the proof. □

5.2 Proof of Proposition 4

Proof of Proposition 4. Using Eq. (13), we have

$$\phi'(\lambda) = \frac{1}{2}\frac{\psi'(\lambda)}{\sqrt{\psi(\lambda)}} - \frac{1}{6\rho}, \quad \phi''(\lambda) = \frac{1}{4}\frac{2\psi''(\lambda)\psi(\lambda) - (\psi'(\lambda))^2}{(\psi(\lambda))^{3/2}}$$

By Proposition 3, we have

$$\phi'(\lambda) = -\frac{\Delta\mathbf{z}(\lambda)^\top (J(\hat{\mathbf{z}}) + \lambda I)^{-1}\Delta\mathbf{z}(\lambda)}{\|\Delta\mathbf{z}(\lambda)\|} - \frac{1}{6\rho} < 0, \quad \text{when } \lambda > 0$$

We also have

$$\begin{aligned} &\phi''(\lambda) \\ = \ & \frac{\|\Delta\mathbf{z}(\lambda)\|^2(\|(J(\hat{\mathbf{z}}) + \lambda I)^{-1}\Delta\mathbf{z}(\lambda)\|^2 + 2\Delta\mathbf{z}(\lambda)^\top (J(\hat{\mathbf{z}}) + \lambda I)^{-2}\Delta\mathbf{z}(\lambda)) - (\Delta\mathbf{z}(\lambda)^\top (J(\hat{\mathbf{z}}) + \lambda I)^{-1}\Delta\mathbf{z}(\lambda))^2}{\|\Delta\mathbf{z}(\lambda)\|^3} \\ \geq \ & \frac{\|\Delta\mathbf{z}(\lambda)\|^2\|(J(\hat{\mathbf{z}}) + \lambda I)^{-1}\Delta\mathbf{z}(\lambda)\|^2 - (\Delta\mathbf{z}(\lambda)^\top (J(\hat{\mathbf{z}}) + \lambda I)^{-1}\Delta\mathbf{z}(\lambda))^2}{\|\Delta\mathbf{z}(\lambda)\|^3} \\ \geq \ & 0, \quad \text{when } \lambda > 0 \end{aligned}$$

where the first inequality holds true since $J(\hat{\mathbf{z}})$ is positive semidefinite, $F(\hat{\mathbf{z}}) \neq \mathbf{0}$ and $\lambda > 0$, and the second inequality holds true because of the the Cauchy-Schwartz inequality. This completes the proof. □

5.3 Proof of Theorem 3

Proof of Theorem 3. Proposition 4 guarantees that $\phi'(\lambda^0) < 0$ when $\lambda > 0$, and it follows from Eq. (14) that $\lambda^1 > \lambda^0 > 0$. Proposition 4 also guarantees the convexity and differentiability of ϕ which together with Eq. (14) implies that

$$\phi(\lambda^1) \geq \phi(\lambda^0) + (\lambda^1 - \lambda^0)\phi'(\lambda^0) = 0$$

Repeating this argument yields that $\lambda^j > 0$ and $\phi(\lambda^j) \geq 0$. In addition, ϕ is strictly decreasing. Thus, the iterates converge monotonically towards the unique solution.

Suppose that $\lambda^\star$ is the unique solution. Then, the mean value theorem implies

$$\phi(\lambda^j) = \phi(\lambda^\star) + (\lambda^j - \lambda^\star)\phi'(\tilde{\lambda}) = (\lambda^j - \lambda^\star)\phi'(\tilde{\lambda}), \quad \text{for some } \tilde{\lambda} \in (\lambda^j, \lambda^\star)$$

Combining this with Eq. (14) yields

$$|\lambda^\star - \lambda^{j+1}| = \left|\lambda^\star - \lambda^j)\left(1 - \frac{\phi'(\tilde{\lambda})}{\phi'(\lambda^j)}\right)\right| \leq |\lambda^\star - \lambda^j| \cdot \left|1 - \frac{\phi'(\tilde{\lambda})}{\phi'(\lambda^j)}\right|$$

Proposition 4 guarantees that $\phi'(\lambda)$ is increasing and strictly less than 0. Thus, we have $\phi'(\lambda^0) \leq \phi'(\lambda^j) \leq \phi'(\tilde{\lambda}) \leq \phi'(\lambda^\star) < 0$ which implies that

$$0 \leq 1 - \frac{\phi'(\tilde{\lambda})}{\phi'(\lambda^j)} \leq 1 - \frac{\phi'(\lambda^\star)}{\phi'(\lambda^0)} < 1$$

Putting everything together implies that the convergence rate is globally Q-linear with a factor at least $1 - \phi'(\lambda^\star)/\phi'(\lambda^0)$. The asymptotic Q-quadratic convergence of the Newton iteration follows because the Jacobian of ϕ is nonsingular at the unique solution $\lambda^\star$ (i.e., $\phi'(\lambda^\star) \neq 0$). This completes the proof. □

6 Finite-Sum Min-Max Optimization

We give concrete examples to clarify the ways to construct the inexact Jacobian such that Condition 4.1 holds true. The key ingredient is random sampling which can significantly reduce the computational cost in an optimization setting [Xu et al.(2020)] and we show that such technique can be employed for solving finite-sum min-max optimization problems in Eq. (9) and (10).

We let the probability distribution of sampling $\xi \in \{1, 2, \ldots, N\}$ be defined as $\text{Prob}(\xi = i) = p_i \geq 0$ for $i = 1, 2, \ldots, N$ and $\mathcal{S} \subseteq \{1, 2, \ldots, N\}$ denote a collection of sampled indices ($|\mathcal{S}|$ is its cardinality). Then, we can construct the inexact Jacobian as follows,

$$J(\mathbf{z}) = \frac{1}{N|\mathcal{S}|}\sum_{i \in \mathcal{S}} \frac{1}{p_i} DF_i(\mathbf{z}), \quad \text{for } DF_i(\mathbf{z}) = \begin{bmatrix} \nabla^2_{\mathbf{xx}} f_i(\mathbf{x}, \mathbf{y}) & \nabla^2_{\mathbf{xy}} f_i(\mathbf{x}, \mathbf{y}) \\ -\nabla^2_{\mathbf{xy}} f_i(\mathbf{x}, \mathbf{y}) & -\nabla^2_{\mathbf{yy}} f_i(\mathbf{x}, \mathbf{y}) \end{bmatrix} \tag{15}$$

This construction is referred to as the subsampled Jacobian and can offer significant computational savings if $|\mathcal{S}| \ll N$ in big-data regime when $N \gg 1$.

In the general finite-sum setting with Eq. (9), we suppose

$$\sup_{\mathbf{z}} \|DF_i(\mathbf{z})\| \leq B_i, \quad \text{for all } i \in \{1, 2, \ldots, N\} \tag{16}$$

and let $B_{\max} = \max_{1 \leq i \leq N} B_i$. For the uniform sampling (i.e., $p_i = \frac{1}{N}$), we summarize the sample complexity results in the following lemma. The proof is omitted for brevity.

Lemma 5. *Suppose that Eq.* (16) *holds true and let* $B_{\max}$ *and* $0 < \tau, \delta < 1$ *be defined properly. A uniform sampling with or without replacement is performed to form the subsampled Jacobian; indeed,* $J(\mathbf{z})$ *is constructed using Eq.* (15) *with* $p_i = \frac{1}{n}$ *and the sample size satisfies*

$$|\mathcal{S}| \geq \Theta^U(\tau, \delta) := \frac{16B_{\max}^2}{\tau^2} \log\left(\frac{2(m+n)}{\delta}\right)$$

Then, we have

$$\text{Prob}(\|J(\mathbf{z}) - DF(\mathbf{z})\| \leq \tau) \geq 1 - \delta$$

Remark. Lemma 5 shows that the inexact Jacobian satisfies Condition 4.1 with probability $1-\delta$ under certain τ and $\kappa_J = B_{\max}$ if it is constructed using the uniform sampling and the size $|\mathcal{S}| = \Omega(\frac{B_{\max}^2}{\tau^2}\log(\frac{m+n}{\delta}))$. Indeed, the first inequality holds true with probability $1-\delta$ since $\text{Prob}(\|J(\mathbf{z}) - DF(\mathbf{z})\| \leq \tau) \geq 1-\delta$, and the second inequality holds true since $\kappa_H = B_{\max}$ (this is a deterministic statement)

In the special finite-sum setting with Eq. (10), we can construct a more "informative" distribution of sampling $\xi \in \{1, 2, \ldots, N\}$ as opposed to simplest uniform sampling. It is advantageous to bias the probability distribution towards carefully picking indices corresponding to those *relevant* f_i's in forming the Jacobian. However, constructing inexact Hessian and corresponding sample complexity guarantee from [Xu et al.(2020), Section 3.1] requires $\nabla^2 f_i$ to be rank-one, which is not valid here. To address this issue, we avail ourselves of the operator-Bernstein inequality [Gross and Nesme(2010)].

The Jacobian of F can be rewritten as $DF(\mathbf{z}) = \frac{1}{N}\sum_{i=1}^N \Lambda_i DF_i(\mathbf{a}_i^\top \mathbf{x}, \mathbf{b}_i^\top \mathbf{y})\Lambda_i^\top$ where $\Lambda_i = \begin{bmatrix} \mathbf{a}_i & \\ & \mathbf{b}_i \end{bmatrix} \in \mathbb{R}^{(m+n)\times 2}$ and $DF_i(x, y) \in \mathbb{R}^{2\times 2}$. Then, the resulting compact form is $DF(\mathbf{z}) = \Lambda^\top \Sigma \Lambda$ where

$$\Lambda^\top = \begin{bmatrix} | & \cdots & | \\ \Lambda_1 & \cdots & \Lambda_N \\ | & \cdots & | \end{bmatrix} \text{ and } \Sigma = \frac{1}{N}\begin{bmatrix} DF_1(\mathbf{a}_i^\top \mathbf{x}, \mathbf{b}_i^\top \mathbf{y}) & \cdots & \cdots \\ \vdots & \ddots & \vdots \\ \cdots & \cdots & DF_N(\mathbf{a}_i^\top \mathbf{x}, \mathbf{b}_i^\top \mathbf{y}) \end{bmatrix} \quad (17)$$

We suppose

$$\sup_{(\mathbf{x},\mathbf{y})} \|DF_i(\mathbf{a}_i^\top \mathbf{x}, \mathbf{b}_i^\top \mathbf{y})\|(\|\mathbf{a}_i\|^2 + \|\mathbf{b}_i\|^2) \leq B_i, \quad \text{for all } i \in \{1, 2, \ldots, N\} \quad (18)$$

and let $B_{\text{avg}} = \frac{1}{N}\sum_{i=1}^N B_i$. For the uniform sampling with the particular nonuniform distribution given by

$$p_i = \frac{\|DF_i(\mathbf{a}_i^\top \mathbf{x}, \mathbf{b}_i^\top \mathbf{y})\|(\|\mathbf{a}_i\|^2 + \|\mathbf{b}_i\|^2)}{\sum_{i=1}^N \|DF_i(\mathbf{a}_i^\top \mathbf{x}, \mathbf{b}_i^\top \mathbf{y})\|(\|\mathbf{a}_i\|^2 + \|\mathbf{b}_i\|^2)} \quad (19)$$

The following lemma summarizes the results on the sample complexity.

Lemma 6. *Suppose that Eq.* (18) *holds true and let* B_{avg} *and* $0 < \tau, \delta < 1$ *be defined properly. A nonuniform sampling is performed to form the subsampled Jacobian; indeed,* $J(\mathbf{z})$ *is constructed using Eq.* (15) *with* $p_i > 0$ *in Eq.* (19) *and the sample size satisfies*

$$|\mathcal{S}| \geq \Theta^N(\tau, \delta) := \frac{4B_{\text{avg}}^2}{\tau^2} \log\left(\frac{2(m+n)}{\delta}\right)$$

Then, we have

$$\text{Prob}(\|J(\mathbf{z}) - DF(\mathbf{z})\| \leq \tau) \geq 1 - \delta$$

```
Algorithm 3 Subsampled-Newton-MinMax($\mathbf{z}_0$, $\rho$, $T$, $\delta$)
  Input: initial point $\mathbf{z}_0$, Lipschitz parameter $\rho$, iteration number $T \geq 1$ and failure probability $\delta \in (0,1)$
  Initialization: set $\hat{\mathbf{z}}_0 = \mathbf{z}_0$ as well as $0 < \kappa_m < \min\{1, \frac{\rho}{4}\}$ and $0 < \tau_0 < \frac{\rho}{4}$
  for $k = 0, 1, 2, \ldots, T-1$ do
    STEP 1: If $\mathbf{z}_k$ is a global saddle point of the problem in Eq. (9) or (10), then stop
    STEP 2: Construct the inexact Jacobian $J(\hat{\mathbf{z}}_k)$ using Eq. (15) with the sample set of $|\mathcal{S}| \geq \Theta^U(\tau_k, 1 - \sqrt[T]{1-\delta})$ (uniform) or $|\mathcal{S}| \geq \Theta^N(\tau_k, 1 - \sqrt[T]{1-\delta})$ (non-uniform) given $0 < \tau_k \leq \min\{\tau_0, \frac{\rho(1-\kappa_m)}{4(B_{\max}+6\rho)}\|F(\hat{\mathbf{z}}_k)\|\}$
    STEP 3: Compute an inexact solution $\Delta\mathbf{z}_k$ of the following subproblem

      $$F(\hat{\mathbf{z}}_k) + J(\hat{\mathbf{z}}_k)\Delta\mathbf{z}_k + 6\rho\|\Delta\mathbf{z}_k\|\Delta\mathbf{z}_k = \mathbf{0}$$

    such that Condition 4.2 hold true
    STEP 4: Compute $\lambda_{k+1} > 0$ such that $\frac{1}{30} \leq \lambda_{k+1}\rho\|\Delta\mathbf{z}_k\| \leq \frac{1}{14}$
    STEP 5: Compute $\mathbf{z}_{k+1} = \hat{\mathbf{z}}_k + \Delta\mathbf{z}_k$
    STEP 6: Compute $\hat{\mathbf{z}}_{k+1} = \hat{\mathbf{z}}_k - \lambda_{k+1}F(\mathbf{z}_{k+1})$
  end for
  Output: $\bar{\mathbf{z}}_T = \frac{1}{\sum_{k=1}^T \lambda_k}\left(\sum_{k=1}^T \lambda_k \mathbf{z}_k\right)$
```

Remark. Compared to Lemma 5, the computation of sampling probability in Lemma 6 requires going through the whole dataset and the cost is $O((m+n)N)$. Nonetheless, the computational savings with smaller sample size dominates such extra cost of computing the sampling probability in optimization setting [Xu et al.(2016)]. In particular, the sample size from Lemma 6 is smaller as $B_{\text{avg}} \ll B_{\max}$ which occurs if one B_i is much larger than the others. In addition, the sample size is proportional to the log of the failure probability in Lemma 5 and 6, allowing the use of a very small failure probability without increasing the sample size significantly

Combining Algorithm 2 and these random sampling strategies gives the first class of subsampled Newton methods for solving finite-sum min-max optimization problems. We provide the scheme in Algorithm 3 and present the iteration complexity and total complexity guarantees in the high-probability sense.

Theorem 4. *Suppose that Assumption 1 and 2 hold. Then, the iterates generated by Algorithm 3 are bounded and Algorithm 3 achieves an ϵ-global saddle point solution within $O(\epsilon^{-2/3})$ iterations with the probability at least $1-\delta$. The total complexity bound of inexact methods is $O((m+n)^\omega \epsilon^{-2/3} + (m+n)^2\epsilon^{-2/3}\log\log(1/\epsilon))$ where $\omega \approx 2.3728$ is the matrix multiplication constant.*

6.1 Proof of Theorem 4

Proof of Theorem 4. Since Algorithm 3 is a combination of Algorithm 2 and the random sampling strategy in Eq. (15) with $0 < \tau_k \leq \min\{\tau_0, \frac{\rho(1-\kappa_m)}{4(B_{\max}+6\rho)}\|F(\hat{\mathbf{z}}_k)\|\}$ and $\kappa_H = B_{\max}$, we can obtain the desired results from Theorems 2 and 3 if the following statement holds true:

$$\text{Prob}(\|J(\hat{\mathbf{z}}_k) - DF(\hat{\mathbf{z}}_k)\| \leq \tau_k \text{ for all } 0 \leq k \leq T-1) \geq 1-\delta \tag{20}$$

To guarantee an overall accumulative success probability of $1-\delta$ across all T iterations, it suffices to set the per-iteration failure probability as $1 - \sqrt[T]{1-\delta}$ as we have done in Algorithm 3. Moreover, we have $1 - \sqrt[T]{1-\delta} = O(\frac{\delta}{T}) = O(\delta\epsilon^{2/3})$. Since this failure probability has only been proven to appear in the logarithmic factor for the sample size in both Lemma 5 and 6, the extra cost will not be dominating. As such, when Algorithm 3 terminates, all of the Jacobian approximations have satisfied Eq. (20). This completes the proof. □

7 Conclusion

We propose and analyze several inexact regularized Newton-type methods for finding a global saddle point of unconstrained convex-concave min-max optimization problems. Our methods are guaranteed to achieve the order-optimal iteration complexity of $O(\epsilon^{-2/3})$ and the tight complexity bound of $O((m+n)^\omega \epsilon^{-2/3} + (m+n)^2\epsilon^{-2/3}\log\log(1/\epsilon))$, where $\omega \approx 2.3728$ is the matrix multiplication constant. We demonstrate how second-order information can be leveraged to accelerate extra-gradient methods, even under inexactness, by generating iterates that remain within a bounded set and converge to an ϵ-saddle point in the sense of the restricted gap function. Specifically, we provide a simple routine for solving the subproblem at each iteration, requiring a single Schur decomposition and $O(\log\log(1/\epsilon))$ calls to a linear system solver in a quasi-upper-triangular system. This improves upon existing second-order min-max optimization methods by reducing the number of required Schur decompositions. We also introduce the first class of subsampled Newton methods for finite-sum min-max optimization problems with order-optimal iteration complexity.

Future research directions include extensions to structured nonconvex-nonconcave min-max optimization problems and customized implementations for specific applications.

References

[Abernethy et al.(2021)] J. Abernethy, K. A. Lai, and A. Wibisono. Last-iterate convergence rates for min-max optimization: Convergence of Hamiltonian gradient descent and consensus optimization. In *ALT*, pages 3–47. PMLR, 2021.

[Adil et al.(2022)] D. Adil, B. Bullins, A. Jambulapati, and S. Sachdeva. Optimal methods for higher-order smooth monotone variational inequalities. *ArXiv Preprint: 2205.06167*, 2022.

[Antipin(1978)] A. S. Antipin. Method of convex programming using a symmetric modification of Lagrange function. *Matekon*, 14(2):23–38, 1978.

[Arjovsky et al.(2017)] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *ICML*, pages 214–223. PMLR, 2017.

[Ben-Tal et al.(2009)] A. Ben-Tal, L. EL Ghaoui, and A. Nemirovski. *Robust Optimization*, volume 28. Princeton University Press, 2009.

[Berahas et al.(2020)] A. S. Berahas, R. Bollapragada, and J. Nocedal. An investigation of Newton-sketch and subsampled Newton methods. *Optimization Methods and Software*, 35(4):661–680, 2020.

[Blackwell and Girshick(1979)] D. A. Blackwell and M. A. Girshick. *Theory of Games and Statistical Decisions*. Courier Corporation, 1979.

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

[Carmon et al.(2022)] Y. Carmon, D. Hausler, A. Jambulapati, Y. Jin, and A. Sidford. Optimal and adaptive Monteiro-Svaiter acceleration. In *NeurIPS*, pages 20338–20350, 2022.

[Chambolle and Pock(2011)] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[Chen et al.(2014)] Y. Chen, G. Lan, and Y. Ouyang. Optimal primal-dual methods for a class of saddle point problems. *SIAM Journal on Optimization*, 24(4):1779–1814, 2014.

[Conn et al.(2000)] A. R. Conn, N. I. M. Gould, and P. L. Toint. *Trust Region Methods*. SIAM, 2000.

[Dantzig(1963)] G. B. Dantzig. *Linear Programming and Extensions*. Princeton University Press, 1963.

[Daskalakis et al.(2018)] C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. Training GANs with optimism. In *ICLR*, 2018. URL `https://openreview.net/forum?id=SJJySbbAZ`.

[Facchinei and Pang(2007)] F. Facchinei and J-S. Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer Science & Business Media, 2007.

[Goodfellow et al.(2014)] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[Gross and Nesme(2010)] D. Gross and V. Nesme. Note on sampling without replacing from a finite collection of matrices. *ArXiv Preprint: 1001.2738*, 2010.

[He and Monteiro(2016)] Y. He and R. D. C. Monteiro. An accelerated HPE-type algorithm for a class of composite convex-concave saddle-point problems. *SIAM Journal on Optimization*, 26(1):29–56, 2016.

[Hsieh et al.(2019)] Y-G. Hsieh, F. Iutzeler, J. Malick, and P. Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In *NeurIPS*, pages 6938–6948, 2019.

[Huang et al.(2022)] K. Huang, J. Zhang, and S. Zhang. Cubic regularized Newton method for the saddle point models: A global and local convergence analysis. *Journal of Scientific Computing*, 91(2):1–31, 2022.

[Jiang and Mokhtari(2022)] R. Jiang and A. Mokhtari. Generalized optimistic methods for convex-concave saddle point problems. *ArXiv Preprint: 2202.09674*, 2022.

[Juditsky et al.(2011)] A. Juditsky, A. Nemirovski, and C. Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

[Kolossoski and Monteiro(2017)] O. Kolossoski and R. D. C. Monteiro. An accelerated non-Euclidean hybrid proximal extragradient-type algorithm for convex-concave saddle-point problems. *Optimization Methods and Software*, 32(6):1244–1272, 2017.

[Korpelevich(1976)] G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

[Kotsalis et al.(2022)] G. Kotsalis, G. Lan, and T. Li. Simple and optimal methods for stochastic variational inequalities, I: Operator extrapolation. *SIAM Journal on Optimization*, 32(3):2041–2073, 2022.

[Kovalev and Gasnikov(2022)] D. Kovalev and A. Gasnikov. The first optimal acceleration of high-order methods in smooth convex optimization. In *NeurIPS*, pages 35339–35351, 2022.

[Lan(2020)] G. Lan. *First-Order and Stochastic Optimization Methods for Machine Learning*, volume 1. Springer, 2020.

[Liang and Stokes(2019)] T. Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *AISTATS*, pages 907–915. PMLR, 2019.

[Lin and Jordan(2023)] T. Lin and M. I. Jordan. Monotone inclusions, acceleration and closed-loop control. *Mathematics of Operations Research*, 48(4):2353–2382, 2023.

[Lin and Jordan(2024)] T. Lin and M. I. Jordan. Perseus: A simple and optimal high-order method for variational inequalities. *Mathematical Programming*, To appear, 2024.

[Lin et al.(2020)] T. Lin, C. Jin, and M. I. Jordan. Near-optimal algorithms for minimax optimization. In *COLT*, pages 2738–2779. PMLR, 2020.

[Mertikopoulos et al.(2019)] P. Mertikopoulos, B. Lecouat, H. Zenati, C-S. Foo, V. Chandrasekhar, and G. Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra(-gradient) mile. In *ICLR*, 2019. URL `https://openreview.net/forum?id=Bkg8jjC9KQ`.

[Mokhtari et al.(2020a)] A. Mokhtari, A. Ozdaglar, and S. Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *AISTATS*, pages 1497–1507. PMLR, 2020a.

[Mokhtari et al.(2020b)] A. Mokhtari, A. E. Ozdaglar, and S. Pattathil. Convergence rate of o(1/k) for optimistic gradient and extragradient methods in smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 30(4):3230–3251, 2020b.

[Monteiro and Svaiter(2010)] R. D. C. Monteiro and B. F. Svaiter. On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. *SIAM Journal on Optimization*, 20(6): 2755–2787, 2010.

[Monteiro and Svaiter(2011)] R. D. C. Monteiro and B. F. Svaiter. Complexity of variants of Tseng's modified FB splitting and Korpelevich's methods for hemivariational inequalities with applications to saddle-point and convex optimization problems. *SIAM Journal on Optimization*, 21(4):1688–1720, 2011.

[Monteiro and Svaiter(2012)] R. D. C. Monteiro and B. F. 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.

[Monteiro and Svaiter(2013)] R. D. C. Monteiro and B. F. 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.

[Nedić and Ozdaglar(2009)] A. Nedić and A. Ozdaglar. Subgradient methods for saddle-point problems. *Journal of Optimization Theory and Applications*, 142(1):205–228, 2009.

[Nemirovski(2004)] A. Nemirovski. Prox-method with rate of convergence o(1/t) for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[Nemirovski and Yudin(1983)] A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

[Nesterov(1983)] Y. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence o(1/k2). *Dokl. Akad. Nauk. SSSR*, 269(3):543–547, 1983.

[Nesterov(2005)] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

[Nesterov(2006)] Y. Nesterov. Cubic regularization of Newton's method for convex problems with constraints. Technical report, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2006.

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

[Nesterov(2008)] Y. Nesterov. Accelerating the cubic regularization of Newton's method on convex problems. *Mathematical Programming*, 112(1):159–181, 2008.

[Nesterov and Polyak(2006)] Y. Nesterov and B. T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[Neumann(1928)] J. V. Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100(1): 295–320, 1928.

[Nikaidô and Isoda(1955)] H. Nikaidô and K. Isoda. Note on non-cooperative convex game. *Pacific Journal of Mathematics*, 5(5):807–815, 1955.

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

[Pilanci and Wainwright(2017)] M. Pilanci and M. J. Wainwright. Newton sketch: A near linear-time optimization algorithm with linear-quadratic convergence. *SIAM Journal on Optimization*, 27(1):205–245, 2017.

[Popov(1980)] L. D. Popov. A modification of the Arrow-Hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980.

[Roosta-Khorasani and Mahoney(2019)] F. Roosta-Khorasani and M. W. Mahoney. Sub-sampled Newton methods. *Mathematical Programming*, 174(1):293–326, 2019.

[Roosta-Khorasani et al.(2014)] F. Roosta-Khorasani, K. Van Den Doel, and U. Ascher. Stochastic algorithms for inverse problems involving PDEs and many measurements. *SIAM Journal on Scientific Computing*, 36(5):S3–S22, 2014.

[Shalev-Shwartz and Ben-David(2014)] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms.* Cambridge University Press, 2014.

[Shamma(2008)] J. Shamma. *Cooperative Control of Distributed Multi-agent Systems.* John Wiley & Sons, 2008.

[Shen et al.(2018)] Z. Shen, A. Mokhtari, T. Zhou, P. Zhao, and H. Qian. Towards more efficient stochastic decentralized learning: Faster convergence and sparse communication. In *ICML*, pages 4624–4633. PMLR, 2018.

[Sinha et al.(2018)] A. Sinha, H. Namkoong, and J. Duchi. Certifiable distributional robustness with principled adversarial training. In *ICLR*, 2018. URL `https://openreview.net/forum?id=Hk6kPgZA-`.

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

[Solodov and Svaiter(1999)] M. V. Solodov and B. F. Svaiter. A hybrid approximate extragradient-proximal point algorithm using the enlargement of a maximal monotone operator. *Set-Valued Analysis*, 7(4): 323–345, 1999.

[Sra et al.(2012)] S. Sra, S. Nowozin, and S. J. Wright. *Optimization for Machine Learning.* MIT Press, 2012.

[Tseng(1995)] P. Tseng. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 60(1-2):237–252, 1995.

[Tseng(2000)] P. Tseng. A modified forward-backward splitting method for maximal monotone mappings. *SIAM Journal on Control and Optimization*, 38(2):431–446, 2000.

[Tseng(2008)] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 2:3, 2008.

[Von Neumann and Morgenstern(1953)] J. Von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior.* Princeton University Press, 1953.

[Xu et al.(2016)] P. Xu, J. Yang, F. Roosta-Khorasani, C. Ré, and M. W. Mahoney. Sub-sampled Newton methods with non-uniform sampling. In *NIPS*, pages 3008–3016, 2016.

[Xu et al.(2020)] P. Xu, F. Roosta, and M. W. Mahoney. Newton-type methods for non-convex optimization under inexact Hessian information. *Mathematical Programming*, 184(1):35–70, 2020.

[Yang and Ying(2022)] T. Yang and Y. Ying. AUC maximization in the era of big data and AI: A survey. *ACM Computing Surveys (CSUR)*, 2022.

[Ying et al.(2016)] Y. Ying, L. Wen, and S. Lyu. Stochastic online AUC maximization. In *NIPS*, pages 451–459, 2016.

A Proofs of Auxiliary Lemmas

A.1 Proof of Lemma 1

Proof of Lemma 1. Using the definition of the Lyapunov function, we have

$$\mathcal{E}_k - \mathcal{E}_{k-1} = \frac{1}{2}\|\hat{\mathbf{z}}_k - \mathbf{z}_0\|^2 - \frac{1}{2}\|\hat{\mathbf{z}}_{k-1} - \mathbf{z}_0\|^2 = (\hat{\mathbf{z}}_k - \hat{\mathbf{z}}_{k-1})^\top(\hat{\mathbf{z}}_k - \mathbf{z}_0) - \frac{1}{2}\|\hat{\mathbf{z}}_k - \hat{\mathbf{z}}_{k-1}\|^2 \tag{21}$$

Plugging $\hat{\mathbf{z}}_k = \hat{\mathbf{z}}_{k-1} - \lambda_k F(\mathbf{z}_k)$ into Eq. (21) yields

$$\begin{aligned}\mathcal{E}_k - \mathcal{E}_{k-1} &\leq \lambda_k(\mathbf{z}_0 - \hat{\mathbf{z}}_k)^\top F(\mathbf{z}_k) - \frac{1}{2}\|\hat{\mathbf{z}}_k - \hat{\mathbf{z}}_{k-1}\|^2 \\ &= \lambda_k(\mathbf{z}_0 - \mathbf{z})^\top F(\mathbf{z}_k) + \lambda_k(\mathbf{z} - \mathbf{z}_k)^\top F(\mathbf{z}_k) + \lambda_k(\mathbf{z}_k - \hat{\mathbf{z}}_k)^\top F(\mathbf{z}_k) - \frac{1}{2}\|\hat{\mathbf{z}}_k - \hat{\mathbf{z}}_{k-1}\|^2\end{aligned}$$

Summing up the above inequality over $k = 1, 2, \ldots, t$ yields

$$\begin{aligned}\sum_{k=1}^t \lambda_k(\mathbf{z}_k - \mathbf{z})^\top F(\mathbf{z}_k) &\leq \mathcal{E}_0 - \mathcal{E}_t \\ &\quad + \underbrace{\sum_{k=1}^t \lambda_k(\mathbf{z}_0 - \mathbf{z})^\top F(\mathbf{z}_k)}_{\mathbf{I}} + \underbrace{\sum_{k=1}^t \lambda_k(\mathbf{z}_k - \hat{\mathbf{z}}_k)^\top F(\mathbf{z}_k) - \frac{1}{2}\|\hat{\mathbf{z}}_k - \hat{\mathbf{z}}_{k-1}\|^2}_{\mathbf{II}}\end{aligned} \tag{22}$$

By using the relationship $\hat{\mathbf{z}}_k = \hat{\mathbf{z}}_{k-1} - \lambda_k F(\mathbf{z}_k)$ again, we have

$$\mathbf{I} = \sum_{k=1}^t \lambda_k(\mathbf{z}_0 - \mathbf{z})^\top F(\mathbf{z}_k) = \sum_{k=1}^t (\hat{\mathbf{z}}_{k-1} - \hat{\mathbf{z}}_k)^\top(\mathbf{z}_0 - \mathbf{z}) = (\hat{\mathbf{z}}_0 - \hat{\mathbf{z}}_t)^\top(\mathbf{z}_0 - \mathbf{z}) \tag{23}$$

In Algorithm 1, we compute $\Delta\mathbf{z}_k$ as an exact solution of the nonlinear equation problem given by

$$F(\hat{\mathbf{z}}_k) + DF(\hat{\mathbf{z}}_k)\Delta\mathbf{z}_k + 6\rho\|\Delta\mathbf{z}_k\|\Delta\mathbf{z}_k = \mathbf{0} \tag{24}$$

Using $\mathbf{z}_k = \hat{\mathbf{z}}_{k-1} + \Delta\mathbf{z}_{k-1}$ and Proposition 1, we have

$$\|F(\mathbf{z}_k) - F(\hat{\mathbf{z}}_{k-1}) - DF(\hat{\mathbf{z}}_{k-1})\Delta\mathbf{z}_{k-1}\| \leq \frac{\rho}{2}\|\Delta\mathbf{z}_{k-1}\|^2 \tag{25}$$

so it suffices to decompose $(\mathbf{z}_k - \hat{\mathbf{z}}_k)^\top F(\mathbf{z}_k)$ and bound this term using Eq. (24) and (25). Indeed, to that end, we have

$$\begin{aligned}(\mathbf{z}_k - \hat{\mathbf{z}}_k)^\top F(\mathbf{z}_k) &\leq \frac{\rho}{2}\|\Delta\mathbf{z}_{k-1}\|^2\|\mathbf{z}_k - \hat{\mathbf{z}}_k\| - 6\rho\|\Delta\mathbf{z}_{k-1}\|(\mathbf{z}_k - \hat{\mathbf{z}}_k)^\top\Delta\mathbf{z}_{k-1} \\ &\leq \frac{\rho}{2}(\|\Delta\mathbf{z}_{k-1}\|^3 + \|\Delta\mathbf{z}_{k-1}\|^2\|\|\hat{\mathbf{z}}_{k-1} - \hat{\mathbf{z}}_k\|) - 6\rho\|\Delta\mathbf{z}_{k-1}\|(\mathbf{z}_k - \hat{\mathbf{z}}_k)^\top\Delta\mathbf{z}_{k-1}\end{aligned}$$

Note that we have $(\mathbf{z}_k - \hat{\mathbf{z}}_k)^\top\Delta\mathbf{z}_{k-1} \geq \|\Delta\mathbf{z}_{k-1}\|^2 - \|\Delta\mathbf{z}_{k-1}\|\|\hat{\mathbf{z}}_{k-1} - \hat{\mathbf{z}}_k\|$. This implies

$$\|\Delta\mathbf{z}_{k-1}\|(\mathbf{z}_k - \hat{\mathbf{z}}_k)^\top\Delta\mathbf{z}_{k-1} \geq \|\Delta\mathbf{z}_{k-1}\|^3 - \|\Delta\mathbf{z}_{k-1}\|^2\|\hat{\mathbf{z}}_{k-1} - \hat{\mathbf{z}}_k\|$$

Putting these pieces together yields

$$(\mathbf{z}_k - \hat{\mathbf{z}}_k)^\top F(\mathbf{z}_k) \leq \frac{13\rho}{2}\|\Delta\mathbf{z}_{k-1}\|^2\|\|\hat{\mathbf{z}}_{k-1} - \hat{\mathbf{z}}_k\| - \frac{11\rho}{2}\|\Delta\mathbf{z}_{k-1}\|^3$$

Since $\frac{1}{33} \leq \lambda_k\rho\|\Delta\mathbf{z}_{k-1}\| \leq \frac{1}{13}$ for all $k \geq 1$, we have

$$\begin{aligned}
\mathbf{II} \ \ & \leq \ \ \sum_{k=1}^{t}\left(\frac{13\lambda_k\rho}{2}\|\Delta\mathbf{z}_{k-1}\|^2\|\|\hat{\mathbf{z}}_{k-1} - \hat{\mathbf{z}}_k\| - \frac{1}{2}\|\hat{\mathbf{z}}_{k-1} - \hat{\mathbf{z}}_k\|^2 - \frac{11\lambda_k\rho}{2}\|\Delta\mathbf{z}_{k-1}\|^3\right) \\
& \leq \ \ \sum_{k=1}^{t}\left(\frac{1}{2}\|\Delta\mathbf{z}_{k-1}\|\|\hat{\mathbf{z}}_{k-1} - \hat{\mathbf{z}}_k\| - \frac{1}{2}\|\hat{\mathbf{z}}_{k-1} - \hat{\mathbf{z}}_k\|^2 - \frac{1}{6}\|\Delta\mathbf{z}_{k-1}\|^2\right) \qquad (26) \\
& \leq \ \ \sum_{k=1}^{t}\left(\max_{\eta\geq 0}\left\{\frac{1}{2}\|\Delta\mathbf{z}_{k-1}\|\eta - \frac{1}{2}\eta^2\right\} - \frac{1}{6}\|\Delta\mathbf{z}_{k-1}\|^2\right) = -\frac{1}{24}\left(\sum_{k=1}^{t}\|\Delta\mathbf{z}_{k-1}\|^2\right)
\end{aligned}$$

Plugging Eq. (23) and (26) into Eq. (22) and using $\hat{\mathbf{z}}_0 = \mathbf{z}_0$ and $\Delta\mathbf{z}_{k-1} = \mathbf{z}_k - \hat{\mathbf{z}}_{k-1}$ yields

$$\sum_{k=1}^{t}\lambda_k(\mathbf{z}_k - \mathbf{z})^\top F(\mathbf{z}_k) \leq \mathcal{E}_0 - \mathcal{E}_t + (\mathbf{z}_0 - \hat{\mathbf{z}}_t)^\top(\mathbf{z}_0 - \mathbf{z}) - \frac{1}{24}\left(\sum_{k=1}^{t}\|\mathbf{z}_k - \hat{\mathbf{z}}_{k-1}\|^2\right)$$

This completes the proof. □

A.2 Proof of Lemma 2

Proof of Lemma 2. Since $\hat{\mathbf{z}}_0 = \mathbf{z}_0$, we have

$$\mathcal{E}_0 - \mathcal{E}_t + (\mathbf{z}_0 - \hat{\mathbf{z}}_t)^\top(\mathbf{z}_0 - \mathbf{z}) \leq -\frac{1}{2}\|\hat{\mathbf{z}}_t - \mathbf{z}_0\|^2 + \frac{1}{2}\|\hat{\mathbf{z}}_t - \mathbf{z}_0\|^2 + \frac{1}{2}\|\mathbf{z}_0 - \mathbf{z}\|^2 = \frac{1}{2}\|\mathbf{z}_0 - \mathbf{z}\|^2$$

This together with Lemma 1 yields

$$\sum_{k=1}^{t}\lambda_k(\mathbf{z}_k - \mathbf{z})^\top F(\mathbf{z}_k) \leq \frac{1}{2}\|\mathbf{z}_0 - \mathbf{z}\|^2 - \frac{1}{24}\left(\sum_{k=1}^{t}\|\mathbf{z}_k - \hat{\mathbf{z}}_{k-1}\|^2\right) \leq \frac{1}{2}\|\mathbf{z}_0 - \mathbf{z}\|^2$$

Since $\mathbf{z}^\star$ is a global saddle point, we have $(\mathbf{z}_k - \mathbf{z}^\star)^\top F(\mathbf{z}_k) \geq 0$ for all $k \geq 1$. Then, we have

$$\sum_{k=1}^{t}\|\mathbf{z}_k - \hat{\mathbf{z}}_{k-1}\|^2 \leq 12\|\mathbf{z}_0 - \mathbf{z}^\star\|^2$$

Further, Lemma 1 with $\mathbf{z} = \mathbf{z}^\star$ implies

$$\mathcal{E}_0 - \mathcal{E}_t + (\mathbf{z}_0 - \hat{\mathbf{z}}_t)^\top(\mathbf{z}_0 - \mathbf{z}^\star) \geq \sum_{k=1}^{t}\lambda_k(\mathbf{z}_k - \mathbf{z}^\star)^\top F(\mathbf{z}_k) + \frac{1}{24}\left(\sum_{k=1}^{t}\|\mathbf{z}_k - \hat{\mathbf{z}}_{k-1}\|^2\right) \geq 0$$

Using Young's inequality, we have

$$0 \leq -\frac{1}{2}\|\hat{\mathbf{z}}_t - \mathbf{z}_0\|^2 + \frac{1}{4}\|\hat{\mathbf{z}}_t - \mathbf{z}_0\|^2 + \|\mathbf{z}_0 - \mathbf{z}^\star\|^2 = -\frac{1}{4}\|\hat{\mathbf{z}}_t - \mathbf{z}_0\|^2 + \|\mathbf{z}_0 - \mathbf{z}^\star\|^2$$

This completes the proof. □

A.3 Proof of Lemma 3

Proof of Lemma 3. Without loss of generality, we assume that $\mathbf{z}_0 \neq \mathbf{z}^\star$. Then, Lemma 2 implies

$$\sum_{k=1}^t (\lambda_k)^{-2}(\frac{1}{33\rho})^2 \leq \sum_{k=1}^t (\lambda_k)^{-2}(\lambda_k \|\mathbf{z}_k - \hat{\mathbf{z}}_{k-1}\|)^2 = \sum_{k=1}^t \|\mathbf{z}_k - \hat{\mathbf{z}}_{k-1}\|^2 \leq 12\|\mathbf{z}_0 - \mathbf{z}^\star\|^2$$

By the Hölder inequality, we have

$$\sum_{k=1}^t 1 = \sum_{k=1}^t \left((\lambda_k)^{-2}\right)^{\frac{1}{3}} (\lambda_k)^{\frac{2}{3}} \leq \left(\sum_{k=1}^t (\lambda_k)^{-2}\right)^{\frac{1}{3}} \left(\sum_{k=1}^t \lambda_k\right)^{\frac{2}{3}}$$

Putting these pieces together yields

$$t \leq \left(66\sqrt{3}\rho\|\mathbf{z}_0 - \mathbf{z}^\star\|\right)^{\frac{2}{3}} \left(\sum_{k=1}^t \lambda_k\right)^{\frac{2}{3}}$$

Letting $t = T$ and rearranging yields the desired result. □

A.4 Proof of Lemma 6

Proof of Lemma 6. Fixing $\mathbf{z} \in \mathbb{R}^{m+n}$, we obtain from Eq. (15) and (17) that $J(\mathbf{z}) = \frac{1}{|\mathcal{S}|}\sum_{j=1}^{|\mathcal{S}|} J_j$ where each random matrix J_j is random and satisfies that $\text{Prob}(J_j = \frac{1}{p_i}\Lambda_i \Sigma_{ii} \Lambda_i^\top) = p_i$ with $p_i > 0$ in Eq. (19). For simplicity, we define

$$X_j = J_j - DF(\mathbf{z}) = J_j - \Lambda^\top \Sigma \Lambda, \qquad X = \sum_{j=1}^{|\mathcal{S}|} X_j = |\mathcal{S}|(J(\mathbf{z}) - \Lambda^\top \Sigma \Lambda)$$

It is easy to verify that $\mathbb{E}[X_j] = 0$ and

$$\|\mathbb{E}[X_j^2]\| \leq \left(\frac{1}{N}\sum_{i=1}^N \|DF_i(\mathbf{a}_i^\top \mathbf{x}, \mathbf{b}_i^\top \mathbf{y})\|(\|\mathbf{a}_i\|^2 + \|\mathbf{b}_i\|^2)\right)^2 \overset{\text{Eq. (18)}}{\leq} B_{\text{avg}}^2$$

Applying the operator-Bernstein inequality yields

$$\text{Prob}(\|J(\mathbf{z}) - DF(\mathbf{z})\| \geq \tau) = \text{Prob}(\|X\| \geq \tau|\mathcal{S}|) \leq 2(m+n)\exp\left(\frac{\tau^2|\mathcal{S}|}{4B_{\text{avg}}^2}\right) \leq \delta$$

This completes the proof. □

A.5 Proof of Lemma 4

Proof of Lemma 4. By using the same argument as used in Lemma 1, we have

$$\sum_{k=1}^t \lambda_k(\mathbf{z}_k - \mathbf{z})^\top F(\mathbf{z}_k) \leq \mathcal{E}_0 - \mathcal{E}_t + (\hat{\mathbf{z}}_0 - \hat{\mathbf{z}}_t)^\top(\mathbf{z}_0 - \mathbf{z}) + \sum_{k=1}^t \lambda_k(\mathbf{z}_k - \hat{\mathbf{z}}_k)^\top F(\mathbf{z}_k) - \frac{1}{2}\|\hat{\mathbf{z}}_k - \hat{\mathbf{z}}_{k-1}\|^2 \quad (27)$$

In what follows, we bound $\sum_{k=1}^{t} \lambda_k (\mathbf{z}_k - \hat{\mathbf{z}}_k)^\top F(\mathbf{z}_k) - \frac{1}{2}\|\hat{\mathbf{z}}_k - \hat{\mathbf{z}}_{k-1}\|^2$. In Algorithm 2, we compute $\Delta\mathbf{z}_k$ such that it is an *inexact* solution of the nonlinear equation problem given by Eq. (8) under Conditions 4.1 and 4.2. Note that we have

$$\|F(\hat{\mathbf{z}}_k) + J(\hat{\mathbf{z}}_k)\Delta\mathbf{z}_k + 6\rho\|\Delta\mathbf{z}_k\|\Delta\mathbf{z}_k\| \leq \kappa_m \cdot \min\{\|\Delta\mathbf{z}_k\|^2, \|F(\hat{\mathbf{z}}_k)\|\} \tag{28}$$

Using $\mathbf{z}_k = \hat{\mathbf{z}}_{k-1} + \Delta\mathbf{z}_{k-1}$ and Proposition 1, we have

$$\|F(\mathbf{z}_k) - F(\hat{\mathbf{z}}_{k-1}) - DF(\hat{\mathbf{z}}_{k-1})\Delta\mathbf{z}_{k-1}\| \leq \frac{\rho}{2}\|\Delta\mathbf{z}_{k-1}\|^2 \tag{29}$$

It suffices to decompose $(\mathbf{z}_k - \hat{\mathbf{z}}_k)^\top F(\mathbf{z}_k)$ and bound this term using Condition 4.1, Eq. (28) and (29). Indeed, we have

$$\begin{aligned}(\mathbf{z}_k - \hat{\mathbf{z}}_k)^\top F(\mathbf{z}_k) &\leq \|\mathbf{z}_k - \hat{\mathbf{z}}_k\|\|F(\mathbf{z}_k) - F(\hat{\mathbf{z}}_{k-1}) - DF(\hat{\mathbf{z}}_{k-1})\Delta\mathbf{z}_{k-1}\| \\ &\quad + \|\mathbf{z}_k - \hat{\mathbf{z}}_k\|\|F(\hat{\mathbf{z}}_{k-1}) + J(\hat{\mathbf{z}}_{k-1})\Delta\mathbf{z}_{k-1} + 6\rho\|\Delta\mathbf{z}_{k-1}\|\Delta\mathbf{z}_{k-1}\| \\ &\quad + \|\mathbf{z}_k - \hat{\mathbf{z}}_k\|\|(DF(\hat{\mathbf{z}}_{k-1}) - J(\hat{\mathbf{z}}_{k-1}))\Delta\mathbf{z}_{k-1}\| - 6\rho\|\Delta\mathbf{z}_{k-1}\|(\mathbf{z}_k - \hat{\mathbf{z}}_k)^\top \Delta\mathbf{z}_{k-1}\end{aligned}$$

The first and second terms are bounded using Eq. (28) and (29). For the third term, we derive from Condition 4.1 that $\|(DF(\hat{\mathbf{z}}_{k-1}) - J(\hat{\mathbf{z}}_{k-1}))\Delta\mathbf{z}_{k-1}\| \leq \tau_{k-1}\|\Delta\mathbf{z}_{k-1}\|$. The fourth term is then bounded using the same argument from the proof of Lemma 1. Putting these pieces together yields that

$$\begin{aligned}(\mathbf{z}_k - \hat{\mathbf{z}}_k)^\top F(\mathbf{z}_k) \quad \leq \quad & \|\mathbf{z}_k - \hat{\mathbf{z}}_k\|\left(\left(\frac{\rho}{2} + \kappa_m\right)\|\Delta\mathbf{z}_{k-1}\|^2 + \tau_{k-1}\|\Delta\mathbf{z}_{k-1}\|\right) \\ & -6\rho\|\Delta\mathbf{z}_{k-1}\|^3 + 6\rho\|\Delta\mathbf{z}_{k-1}\|^2\|\hat{\mathbf{z}}_{k-1} - \hat{\mathbf{z}}_k\|\end{aligned} \tag{30}$$

We claim that

$$\left(\frac{\rho}{2} + \kappa_m\right)\|\Delta\mathbf{z}_{k-1}\|^2 + \tau_{k-1}\|\Delta\mathbf{z}_{k-1}\| \leq \rho\|\Delta\mathbf{z}_{k-1}\|^2 \tag{31}$$

Indeed, for the case of $\|\Delta\mathbf{z}_{k-1}\| \geq 1$, we have

$$\left(\frac{\rho}{2} + \kappa_m\right)\|\Delta\mathbf{z}_{k-1}\|^2 + \tau_{k-1}\|\Delta\mathbf{z}_{k-1}\| \leq \left(\frac{\rho}{2} + \kappa_m + \tau_{k-1}\right)\|\Delta\mathbf{z}_{k-1}\|^2$$

which together with the fact that $0 < \kappa_m < \min\{1, \frac{\rho}{4}\}$ and $\tau_{k-1} \leq \tau_0 < \frac{\rho}{4}$ can yield Eq. (31). Otherwise, we have $\|\Delta\mathbf{z}_{k-1}\| < 1$ and obtain from Conditions 4.1 and 4.2 that

$$\begin{aligned}\kappa_m\|F(\hat{\mathbf{z}}_{k-1})\| \quad & \geq \quad \|F(\hat{\mathbf{z}}_{k-1}) + J(\hat{\mathbf{z}}_{k-1})\Delta\mathbf{z}_{k-1} + 6\rho\|\Delta\mathbf{z}_{k-1}\|\Delta\mathbf{z}_{k-1}\| \\ & \geq \quad \|F(\hat{\mathbf{z}}_{k-1})\| - \kappa_J\|\Delta\mathbf{z}_{k-1}\| - 6\rho\|\Delta\mathbf{z}_{k-1}\|^2 \\ & \geq \quad \|F(\hat{\mathbf{z}}_{k-1})\| - (\kappa_J + 6\rho)\|\Delta\mathbf{z}_{k-1}\|\end{aligned}$$

Rearranging the above inequality and using $0 \leq \tau_{k-1} \leq \frac{\rho(1-\kappa_m)}{4(\kappa_J+6\rho)}\|F(\hat{\mathbf{z}}_{k-1})\|$ yields

$$\|\Delta\mathbf{z}_{k-1}\| \geq \frac{1-\kappa_m}{\kappa_J + 6\rho}\|F(\hat{\mathbf{z}}_{k-1})\| \geq \frac{4\tau_{k-1}}{\rho}$$

Since $0 < \kappa_m < \min\{1, \frac{\rho}{4}\}$ again, we get Eq. (31) as follows,

$$\left(\frac{\rho}{2} + \kappa_m\right)\|\Delta\mathbf{z}_{k-1}\|^2 + \tau_{k-1}\|\Delta\mathbf{z}_{k-1}\| \leq \left(\frac{\rho}{2} + \kappa_m + \frac{\tau_{k-1}}{\|\Delta\mathbf{z}_{k-1}\|}\right)\|\Delta\mathbf{z}_{k-1}\|^2 \leq \rho\|\Delta\mathbf{z}_{k-1}\|^2$$

Plugging Eq. (31) into Eq. (30) and using $\|\mathbf{z}_k - \hat{\mathbf{z}}_k\| \leq \|\Delta\mathbf{z}_{k-1}\| + \|\hat{\mathbf{z}}_{k-1} - \hat{\mathbf{z}}_k\|$ yields

$$(\mathbf{z}_k - \hat{\mathbf{z}}_k)^\top F(\mathbf{z}_k) \leq 7\rho\|\Delta\mathbf{z}_{k-1}\|^2\|\|\hat{\mathbf{z}}_{k-1} - \hat{\mathbf{z}}_k\| - 5\rho\|\Delta\mathbf{z}_{k-1}\|^3$$

Since $\frac{1}{30} \leq \lambda_k\rho\|\Delta\mathbf{z}_{k-1}\| \leq \frac{1}{14}$ for all $k \geq 1$, we have

$$\begin{aligned}
&\sum_{k=1}^t \lambda_k(\mathbf{z}_k - \hat{\mathbf{z}}_k)^\top F(\mathbf{z}_k) - \frac{1}{2}\|\hat{\mathbf{z}}_k - \hat{\mathbf{z}}_{k-1}\|^2 \\
&\quad \leq \quad \sum_{k=1}^t \left(7\lambda_k\rho\|\Delta\mathbf{z}_{k-1}\|^2\|\|\hat{\mathbf{z}}_{k-1} - \hat{\mathbf{z}}_k\| - \frac{1}{2}\|\hat{\mathbf{z}}_{k-1} - \hat{\mathbf{z}}_k\|^2 - 5\lambda_k\rho\|\Delta\mathbf{z}_{k-1}\|^3\right) \\
&\quad \leq \quad \sum_{k=1}^t \left(\frac{1}{2}\|\Delta\mathbf{z}_{k-1}\|\|\hat{\mathbf{z}}_{k-1} - \hat{\mathbf{z}}_k\| - \frac{1}{2}\|\hat{\mathbf{z}}_{k-1} - \hat{\mathbf{z}}_k\|^2 - \frac{1}{6}\|\Delta\mathbf{z}_{k-1}\|^2\right) \\
&\quad \leq \quad \sum_{k=1}^t \left(\max_{\eta \geq 0}\left\{\frac{1}{2}\|\Delta\mathbf{z}_{k-1}\|\eta - \frac{1}{2}\eta^2\right\} - \frac{1}{6}\|\Delta\mathbf{z}_{k-1}\|^2\right) = -\frac{1}{24}\left(\sum_{k=1}^t \|\Delta\mathbf{z}_{k-1}\|^2\right)
\end{aligned}$$

Therefore, we conclude from Eq. (27), $\hat{\mathbf{z}}_0 = \mathbf{z}_0$ and $\Delta\mathbf{z}_{k-1} = \mathbf{z}_k - \hat{\mathbf{z}}_{k-1}$ that

$$\sum_{k=1}^t \lambda_k(\mathbf{z}_k - \mathbf{z})^\top F(\mathbf{z}_k) \leq \mathcal{E}_0 - \mathcal{E}_t + (\mathbf{z}_0 - \hat{\mathbf{z}}_t)^\top(\mathbf{z}_0 - \mathbf{z}) - \frac{1}{24}\left(\sum_{k=1}^t \|\mathbf{z}_k - \hat{\mathbf{z}}_{k-1}\|^2\right)$$

This completes the proof. □