An Explicit Second-Order Min-Max Optimization Method with Optimal Convergence Guarantees

Tianyi Lin[◇] Panayotis Mertikopoulos[⋆] Michael I. Jordan[◇,†]

[◇]Department of Electrical Engineering and Computer Sciences
[†]Department of Statistics
University of California, Berkeley
[⋆]Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble, France

November 18, 2022

Abstract

We propose and analyze exact and inexact regularized Newton-type methods for finding a global saddle point of a *convex-concave* unconstrained min-max optimization problem. Compared to their first-order counterparts, investigations of second-order methods for min-max optimization are relatively limited, as obtaining global rates of convergence with second-order information is much more involved. In this paper, we highlight how second-order information can be used to speed up the dynamics of dual extrapolation methods despite inexactness. Specifically, we show that the proposed algorithms generate iterates that remain within a bounded set and the averaged iterates converge to an ϵ-saddle point within $O(\epsilon^{-2/3})$ iterations in terms of a gap function. Our algorithms match the theoretically established lower bound in this context and our analysis provides a simple and intuitive convergence analysis for second-order methods without requiring any compactness assumptions. Finally, we present a series of numerical experiments on synthetic and real data that demonstrate the efficiency of the proposed algorithms.

1 Introduction

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:

$$\min_{\mathbf{x}\in\mathbb{R}^m} \max_{\mathbf{y}\in\mathbb{R}^n} \ f(\mathbf{x}, \mathbf{y}), \tag{1.1}$$

i.e., a pair of points $(\mathbf{x}^\star, \mathbf{y}^\star) \in \mathbb{R}^m \times \mathbb{R}^n$ 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 so-called *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 range of concrete examples, see Facchinei and Pang [2007] and references therein.

Owing to the above, several classes of optimization algorithms have been proposed and analyzed for finding a global saddle point of Eq. (1.1) in the convex-concave case. An important representative algorithm is the extragradient (EG) method [Korpelevich, 1976, Antipin, 1978]. 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 compact domain, even if f is not strongly convex-concave or bilinear. For a more general unconstrained setting, Solodov and Svaiter [1999] considered a variant of EG (namely hybrid proximal extragradient (HPE) method) and proved that it achieved the same convergence rate.

In addition to the extra-gradient algorithm itself, there are several variants that achieve similar convergence rate guarantees with a lighter updating structure, such as optimistic gradient descent ascent (OGDA) [Popov, 1980, Mokhtari et al., 2020, Kotsalis et al., 2022], forward-backward splitting [Tseng, 2000] and dual extrapolation [Nesterov, 2007]; for a partial survey, see Hsieh et al. [2019] and references therein. All these methods are *order-optimal first-order methods* as they match the lower bound of Ouyang and Xu [2021].

Transitioning for the moment to convex minimization problems, second-order methods are known to enjoy superior convergence properties over their first-order counterparts, in both theory and practice. In particular, the celebrated accelerated cubic regularization of Newton's method converges at a rate of $O(\epsilon^{-1/3})$ [Nesterov, 2008] which outperforms the $O(\epsilon^{-1/2})$ lower bound for first-order methods [Nemirovski and Yudin, 1983]. Moreover, first-order methods may perform poorly in ill-conditioned problems and can be sensitive to the parameter choices in real application problems while the second-order methods are often shown to be more robust in the same context [Pilanci and Wainwright, 2017, Roosta-Khorasani and Mahoney, 2019, Berahas et al., 2020].

However, when moving back to the context of convex-concave min-max problems, two separate series of issues arise: (a) achieving acceleration with second-order information is more involved than in the min-max setting; and (b) acquiring exact second-order information is more expensive in general. Aiming to address these issues, a line of recent work has generalized classical methods from first- to second-order, including EG and OGDA [Monteiro and Svaiter, 2012, Lin and Jordan, 2021, Bullins and Lai, 2022, Jiang and Mokhtari, 2022]. These methods achieve a global rate of $O(\epsilon^{-2/3}\log(1/\epsilon))$,[1] but they require solving a nontrivial *implicit* binary search problem at each iteration, and this can be computationally expensive from a practical viewpoint.[2] In a similar vein, Huang et al. [2022] extended the cubic regularization approach of Newton's method [Nesterov and Polyak, 2006] to min-max optimization but their convergence analysis for convex-concave problems requires an error-bound condition. It is also worth mentioning that all these existing second-order min-max optimization algorithms require *exact* second-order information, and given the implicit nature of the inner loop subproblems involved,

[1] The lower bound of $\Omega(\epsilon^{-2/3})$ for second-order min-max optimization algorithms has been recently established in the VI literature [Lin and Jordan, 2022, Adil et al., 2022]. Although such bound is only valid under a linear span assumption, it can be viewed as a benchmark for the algorithms that we consider in this paper.

[2] By "implicit", we mean here that the method's inner loop subproblem for computing the t-th iterate involves the iterate being updated, so it results in an implicit update rule. By contrast, the term "explicit" means that any inner loop subproblem for computing the t-th iterate does not involve said iterate.

the methods' robustness to inexact information cannot be taken for granted. In view of all this, it is natural to ask:

Can we develop explicit second-order min-max optimization algorithms that remain order-optimal even with inexact second-order information?

This paper provides an affirmative answer to the above question. By leveraging the recent progress in the variational inequality (VI) literature [Lin and Jordan, 2022], we begin by proposing a conceptual second-order min-max optimization algorithm with a global convergence guarantee of $O(\epsilon^{-2/3})$ in the convex-concave case. The proposed algorithm does not contain any binary search procedure but still requires exact second-order information and an exact solution of an inner explicit subproblem. To relax these requirements, we subsequently propose a class of second-order min-max optimization algorithms that only equire *inexact second-order information* and *inexact subproblem solutions*. The approximation condition is inspired by Xu et al. [2020, Condition 1] and allows for direct construction of such oracles through randomized sampling in the case of finite-sum problems [Drineas and Mahoney, 2018]. As far as we are aware, this is the first class of sub-sampled Newton methods for solving finite-sum min-max optimization problems, gaining considerable computational savings since the sample size increases gracefully from a very small sample set. In terms of theoretical guarantees, we prove that these inexact algorithms achieve the convergence guarantee of $O(\epsilon^{-2/3})$ and the sub-sampled Newton methods achieve the same rate in high probability.

Overall, our paper stands at the interface of two synergistic research thrusts in the literature: one is to generalize second-order methods for convex minimization problems [Huang et al., 2022] and the other is to generalize first-order methods for min-max optimization problems [Monteiro and Svaiter, 2012, Lin and Jordan, 2021, Bullins and Lai, 2022, Jiang and Mokhtari, 2022]. To the best of our knowledge, there are no *explicit* second-order methods in the literature that (i) achieve an order-optimal convergence guarantee of $\Theta(\epsilon^{-2/3})$ in the convex-concave case; and (ii) do *not* require exact second-order information in their implementation. Experimental results on both real and synthetic data demonstrate the efficiency of the proposed algorithms.

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 point $(\mathbf{x}, \mathbf{y})$. We use $\nabla f(\mathbf{x}, \mathbf{y})$ to denote the full 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 $(\mathbf{x}, \mathbf{y})$. For a vector $\mathbf{x}$, 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 any problem parameter, and $\tilde{O}(\cdot), \tilde{\Omega}(\cdot)$ to hide absolute constants and log factors.

The remainder of the paper is organized as follows. In Section 2, we present the setup of smooth min-max optimization and provide the definitions for functions and optimality criteria. In Section 3, we propose an exact second-order min-max optimization algorithm without any binary search procedure and prove that it achieves a global convergence rate of $\Omega(\epsilon^{-2/3})$ in the convex-concave case. In Section 4, we propose a broad class of second-order min-max optimization algorithms under inexact second-order information and inexact subproblem solving and prove the same convergence guarantee. We also provide the subsampled Newton method for solving the finite-sum min-max optimization problems. In Section 5, we conduct the experiments on synthetic and real data to demonstrate the efficiency of our algorithms.

2 Preliminaries

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

Definition 2.1 *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 2.2 *A differentiable function f is convex-concave if*

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

We also define the notion of global saddle points for the min-max problem in Eq. (1.1).

Definition 2.3 *A point $(\mathbf{x}^\star, \mathbf{y}^\star) \in \mathbb{R}^m \times \mathbb{R}^n$ is a global saddle point of a convex-concave function $f(\cdot, \cdot)$ if we have $f(\mathbf{x}^\star, \mathbf{y}) \leq f(\mathbf{x}^\star, \mathbf{y}^\star) \leq f(\mathbf{x}, \mathbf{y}^\star)$ for all $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$.*

Finally, throughout this paper, we will assume that the following conditions are satisfied.

Assumption 2.4 *The function $f(\mathbf{x}, \mathbf{y})$ is continuously differentiable. Furthermore, 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 $\mathbf{x} \in \mathbb{R}^m$. There also exists at least one global saddle point of $f(\mathbf{x}, \mathbf{y})$ for the min-max problem in Eq. (1.1).*

Assumption 2.5 *The function $f(\mathbf{x}, \mathbf{y}) : \mathbb{R}^m \times \mathbb{R}^n \mapsto \mathbb{R}$ is ρ-Hessian Lipschitz. Formally, we have*

$$\|\nabla^2 f(\mathbf{x}, \mathbf{y}) - \nabla^2 f(\mathbf{x}', \mathbf{y}')\| \leq \rho \left\| \begin{bmatrix} \mathbf{x} - \mathbf{x}' \\ \mathbf{y} - \mathbf{y}' \end{bmatrix} \right\|, \quad \text{for all } (\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}') \in \mathbb{R}^m \times \mathbb{R}^n.$$

Under Assumption 2.4, we have $f(\mathbf{x}^\star, \mathbf{y}) \leq f(\mathbf{x}^\star, \mathbf{y}^\star) \leq f(\mathbf{x}, \mathbf{y}^\star)$ for all $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$. This gives rise to the Nikaido-Isoda gap function $\text{GAP}(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = f(\hat{\mathbf{x}}, \mathbf{y}^\star) - f(\mathbf{x}^\star, \hat{\mathbf{y}}) \geq 0$ which provides a performance measure for the closeness of $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ to a global saddle point. Formally, we have

Definition 2.6 *A point $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathbb{R}^m \times \mathbb{R}^n$ is an ϵ-global saddle point of a convex-concave function $f(\cdot, \cdot)$ if $\text{GAP}(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \leq \epsilon$. If $\epsilon = 0$, we have that $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is a global saddle point of the function $f(\cdot, \cdot)$.*

In the subsequent sections of this paper, we propose a new regularized Newton method for solving the min-max optimization problem in Eq. (1.1) and prove an optimal global convergence rate in the sense that our upper bound on the required iteration number to return an ϵ-optimal solution matching the known lower bound of $\Omega(\epsilon^{-2/3})$ [Lin and Jordan, 2022, Adil et al., 2022].

In our algorithm, we denote the k^{th} iterate by $(\mathbf{x}_k, \mathbf{y}_k)$ and define the averaged (ergodic) iterates by $(\bar{\mathbf{x}}_k, \bar{\mathbf{y}}_k)$. More specifically, we have given a sequence of weights λ_n, $n = 1, \ldots, T$, we let

$$\bar{\mathbf{x}}_k = \frac{1}{\sum_{i=1}^k \lambda_i} \left(\sum_{i=1}^k \lambda_i \mathbf{x}_i \right), \quad \bar{\mathbf{y}}_k = \frac{1}{\sum_{i=1}^k \lambda_i} \left(\sum_{i=1}^k \lambda_i \mathbf{y}_i \right). \tag{2.1}$$

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,

$$F(\mathbf{z}) = \begin{bmatrix} \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) \\ -\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) \end{bmatrix}. \tag{2.2}$$

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

$$DF(\mathbf{z}) = \begin{bmatrix} \nabla^2_{\mathbf{xx}} f(\mathbf{x}, \mathbf{y}) & \nabla^2_{\mathbf{xy}} f(\mathbf{x}, \mathbf{y}) \\ -\nabla^2_{\mathbf{xy}} f(\mathbf{x}, \mathbf{y}) & -\nabla^2_{\mathbf{yy}} f(\mathbf{x}, \mathbf{y}) \end{bmatrix} \in \mathbb{R}^{(m+n)\times(m+n)}. \tag{2.3}$$

In the following lemma, we summarize the properties of the operator F in Eq. (2.2) and its Jacobian in Eq. (2.3) under Assumption 2.4 and 2.5. We note that most of the results in the following lemma are well known [Nemirovski, 2004, Mokhtari et al., 2020] so we omit their proofs.

Lemma 2.7 *Let $F(\cdot)$ and $DF(\cdot)$ be defined in Eq. (2.2) and (2.3). Under Assumption 2.4 and 2.5, the following statements hold true,*

(a) F is monotone, i.e., for any $\mathbf{z}, \mathbf{z}' \in \mathbb{R}^{m+n}$, we have $(\mathbf{z} - \mathbf{z}')^\top (F(\mathbf{z}) - F(\mathbf{z}')) \geq 0$.

(b) DF is ρ-Lipschitz continuous, i.e., for any $\mathbf{z}', \mathbf{z} \in \mathbb{R}^{m+n}$, we have $\|DF(\mathbf{z}) - DF(\mathbf{z}')\| \leq \rho\|\mathbf{z} - \mathbf{z}'\|$.

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

Proof. Note that (a) and (c) have been proven in earlier work [Nemirovski, 2004, Mokhtari et al., 2020], and it suffices to prove (b). By using the definition of $DF(\cdot)$ in Eq. (2.3), we have

$$(DF(\mathbf{z}) - DF(\mathbf{z}'))\mathbf{h} = \begin{bmatrix} I_m & \\ & -I_n \end{bmatrix} (\nabla^2 f(\mathbf{z}) - \nabla^2 f(\mathbf{z}'))\mathbf{h}. \tag{2.4}$$

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

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

This equality together with Assumption 2.5 implies the desired result in (b). $\square$

Before proceeding to our algorithm and 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 2.8 *Let $(\bar{\mathbf{x}}_k, \bar{\mathbf{y}}_k)$ and $F(\cdot)$ be defined in Eq. (2.1) and (2.2). Then, under Assumption 2.4, the following statement holds true,*

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

Proof. Using the definition of the operator $F(\cdot)$ in Eq. (2.2), we have

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

Note that Assumption 2.4 guarantees 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

$$(\mathbf{x}_i - \mathbf{x}^\star)^\top \nabla_{\mathbf{x}} f(\mathbf{x}_i, \mathbf{y}_i) \geq f(\mathbf{x}_i, \mathbf{y}_i) - f(\mathbf{x}^\star, \mathbf{y}_i), \qquad (\mathbf{y}_i - \mathbf{y}^\star)^\top \nabla_{\mathbf{y}} f(\mathbf{x}_i, \mathbf{y}_i) \leq f(\mathbf{x}_i, \mathbf{y}_i) - f(\mathbf{x}_i, \mathbf{y}^\star).$$

Algorithm 1 Newton-MinMax$((\mathbf{x}_0, \mathbf{y}_0),\ \rho,\ T)$

Input: initial point $(\mathbf{x}_0, \mathbf{y}_0)$, Lipschitz parameter ρ and iteration number $T \geq 1$.
Initialization: set $\hat{\mathbf{x}}_0 = \mathbf{x}_0$ and $\hat{\mathbf{y}}_0 = \mathbf{y}_0$.
for $k = 0, 1, 2, \ldots, T-1$ **do**
STEP 1: If $(\mathbf{x}_k, \mathbf{y}_k)$ is a global saddle point of the min-max optimization problem, then **stop**.
STEP 2: Compute an *exact solution* $(\Delta\mathbf{x}_k, \Delta\mathbf{y}_k) \in \mathbb{R}^m \times \mathbb{R}^n$ of the cubic regularized min-max optimization (we let $\hat{\mathbf{z}}_k = (\hat{\mathbf{x}}_k, \hat{\mathbf{y}}_k)$ for brevity):

$$\min_{\Delta\mathbf{x}} \max_{\Delta\mathbf{y}} \left\{ \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\mathbf{y} \end{bmatrix}^\top \begin{bmatrix} \nabla_{\mathbf{x}} f(\hat{\mathbf{z}}_k) \\ \nabla_{\mathbf{y}} f(\hat{\mathbf{z}}_k) \end{bmatrix} + \frac{1}{2} \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\mathbf{y} \end{bmatrix}^\top \begin{bmatrix} \nabla^2_{\mathbf{xx}} f(\hat{\mathbf{z}}_k) & \nabla^2_{\mathbf{xy}} f(\hat{\mathbf{z}}_k) \\ \nabla^2_{\mathbf{xy}} f(\hat{\mathbf{z}}_k) & \nabla^2_{\mathbf{yy}} f(\hat{\mathbf{z}}_k) \end{bmatrix} \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\mathbf{y} \end{bmatrix} + 2\rho\|\Delta\mathbf{x}\|^3 - 2\rho\|\Delta\mathbf{y}\|^3 \right\}. \tag{3.1}$$

STEP 3: Compute $\lambda_{k+1} > 0$ such that $\frac{1}{15} \leq \lambda_{k+1}\rho\sqrt{\|\Delta\mathbf{x}_k\|^2 + \|\Delta\mathbf{y}_k\|^2} \leq \frac{1}{13}$.
STEP 4a: Compute $\mathbf{x}_{k+1} = \hat{\mathbf{x}}_k + \Delta\mathbf{x}_k$.
STEP 4b: Compute $\mathbf{y}_{k+1} = \hat{\mathbf{y}}_k + \Delta\mathbf{y}_k$.
STEP 5a: Compute $\hat{\mathbf{x}}_{k+1} = \hat{\mathbf{x}}_k - \lambda_{k+1}\nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})$.
STEP 5b: Compute $\hat{\mathbf{y}}_{k+1} = \hat{\mathbf{y}}_k + \lambda_{k+1}\nabla_{\mathbf{y}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})$.
end for
Output: $(\bar{\mathbf{x}}_T, \bar{\mathbf{y}}_T) = \frac{1}{\sum_{k=1}^T \lambda_k}\left(\sum_{k=1}^T \lambda_k \mathbf{x}_k, \sum_{k=1}^T \lambda_k \mathbf{y}_k\right)$.

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

$$\frac{1}{\sum_{i=1}^k \lambda_i}\left(\sum_{i=1}^k \lambda_i (\mathbf{z}_i - \mathbf{z}^\star)^\top F(\mathbf{z}_i)\right) \geq \frac{1}{\sum_{i=1}^k \lambda_i}\left(\sum_{i=1}^k \lambda_i (f(\mathbf{x}_i, \mathbf{y}^\star) - f(\mathbf{x}^\star, \mathbf{y}_i))\right). \tag{2.5}$$

Using the definition of $(\bar{\mathbf{x}}_k, \bar{\mathbf{y}}_k)$ in Eq. (2.1) and that f is convex-concave, we have

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

Plugging the above two inequalities in Eq. (2.5), we conclude the desired inequality. □

3 Conceptual Algorithm and Convergence Analysis

In this section, we present the scheme of Newton-MinMax and establish a global convergence rate guarantee. Moreover, we provide intuition into why Newton-MinMax yields an optimal rate of global convergence by leveraging the second-order information. It is worth mentioning that Newton-MinMax is a conceptual algorithmic framework in which the *exact* second-order information is required and the cubic regularized subproblem needs to be solved *exactly* at each iteration.

3.1 Algorithmic scheme

We summarize our second-order method, which we call Newton-MinMax$((\mathbf{x}_0, \mathbf{y}_0),\ \rho,\ T)$, in Algorithm 1 where $(\mathbf{x}_0, \mathbf{y}_0) \in \mathbb{R}^m \times \mathbb{R}^n$ 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 the classical first-order dual extrapolation method [Nesterov, 2007] in the context of min-max optimization. More specifically, with $(\mathbf{x}_0, \mathbf{y}_0) \in \mathbb{R}^m \times \mathbb{R}^n$ and $(\hat{\mathbf{x}}_0, \hat{\mathbf{y}}_0) = (\mathbf{x}_0, \mathbf{y}_0)$, the k^{th} iteration of dual extrapolation for solving Eq. (1.1) is given by

- Compute a pair $(\Delta \mathbf{x}_k, \Delta \mathbf{y}_k) \in \mathbb{R}^m \times \mathbb{R}^n$ such that it is an *exact* solution of the following *quadratic regularized min-max optimization* problem: (we let $\hat{\mathbf{z}}_k = (\hat{\mathbf{x}}_k, \hat{\mathbf{y}}_k)$ for brevity)

$$\min_{\Delta \mathbf{x}} \max_{\Delta \mathbf{y}} \left\{ \begin{bmatrix} \Delta \mathbf{x} \\ \Delta \mathbf{y} \end{bmatrix}^\top \begin{bmatrix} \nabla_{\mathbf{x}} f(\hat{\mathbf{z}}_k) \\ \nabla_{\mathbf{y}} f(\hat{\mathbf{z}}_k) \end{bmatrix} + \ell \|\Delta \mathbf{x}\|^2 - \ell \|\Delta \mathbf{y}\|^2 \right\}. \tag{DE}$$

- Compute $\mathbf{x}_{k+1} = \hat{\mathbf{x}}_k + \Delta \mathbf{x}_k$ and $\mathbf{y}_{k+1} = \hat{\mathbf{y}}_k + \Delta \mathbf{y}_k$.
- Compute $\hat{\mathbf{x}}_{k+1} = \hat{\mathbf{x}}_k - \lambda \nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})$ and $\hat{\mathbf{y}}_{k+1} = \hat{\mathbf{y}}_k + \lambda \nabla_{\mathbf{y}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})$.

Intuitively, the above scheme can be viewed as an instance of EG in the dual space and the rule that transforms $(\hat{\mathbf{x}}_k, \hat{\mathbf{y}}_k)$ into $(\hat{\mathbf{x}}_{k+1}, \hat{\mathbf{y}}_{k+1})$ is well known as the dual extrapolation step. If we set $\lambda = \Theta(\ell)$ where $\ell > 0$ is the Lipschitz constant for the gradient of the function f, Nesterov [2007, Theorem 2] guarantees that the averaged iterates converge to a global saddle point at a rate of $O(\epsilon^{-1})$. By exploiting exact second-order information, Nesterov [2006] generalized his approach by replacing the subproblem in Eq. (DE) with the cubic regularized subproblem that is similar to the one in Eq. (3.1) from Algorithm 1. However, the resulting scheme with constant $\lambda = \Theta(\rho)$ is only guaranteed to achieve the global convergence rate of $O(\epsilon^{-1})$ which is the same as its first-order counterpart [Nesterov, 2006, Theorem 4].

Inspired by Lin and Jordan [2022], we propose an adaptive strategy for updating λ_k in Algorithm 1 and prove that our algorithm achieve an improved global rate of $O(\epsilon^{-2/3})$ under Assumptions 2.4 and 2.5. Intuitively, such a strategy will work well; in particular, λ_k is the step size in the dual space and needs to increase as the iterate $(\mathbf{x}_k, \mathbf{y}_k)$ approaches the set of global saddle points (note that the value of $\sqrt{\|\Delta \mathbf{x}_k\|^2 + \|\Delta \mathbf{y}_k\|^2}$ quantifies the quality of $(\mathbf{x}_k, \mathbf{y}_k)$ in terms of the gap function). From a practical viewpoint, Algorithm 1 can be valuable since it simplifies the existing schemes for second-order min-max optimization by removing any line search procedure. We will relax the requirement of exact second-order information and exact subproblem solving in Section 4 and present numerical results in Section 5.

3.2 Convergence analysis

We provide our main results on the convergence rate for Algorithm 1 in terms of the number of calls of the subproblem solvers. The following theorem gives us the global convergence rate of Algorithm 1 for convex-concave min-max optimization problems.

Theorem 3.1 *Suppose that Assumptions 2.4 and 2.5 hold. Then the iterates generated by Algorithm 1 are bounded,*

and its output state enjoys the guarantee

$$\mathrm{GAP}(\bar{\mathbf{x}}_T, \bar{\mathbf{y}}_T) \leq \frac{15\sqrt{3}\rho \|\mathbf{z}_0 - \mathbf{z}^\star\|^3}{T^{\frac{3}{2}}} \tag{3.2}$$

i.e., the algorithm achieves an ϵ-optimal state within $O(\epsilon^{-2/3})$ iterations.

Remark 3.2 *Theorem 3.1 demonstrates that Algorithm 1 achieves the lower bound established in the literature on variational inequalities for second-order methods [Lin and Jordan, 2022, Adil et al., 2022] 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. It is also worth mentioning that Theorem 3.1 is not a consequence of Lin and Jordan [2022, Theorem 3.1] since the bounded feasible sets are necessary for their convergence analysis.*

We define a Lyapunov function for the iterates generated by Algorithm 1 as follows:

$$\mathcal{E}_t = \frac{1}{2}\left(\|\hat{\mathbf{x}}_t - \mathbf{x}_0\|^2 + \|\hat{\mathbf{y}}_t - \mathbf{y}_0\|^2\right), \tag{3.3}$$

This function will be used to prove technical results that pertain to the dynamics of Algorithm 1. Recall that $\mathbf{z} = [\mathbf{x}; \mathbf{y}]$, $\hat{\mathbf{z}} = [\hat{\mathbf{x}}; \hat{\mathbf{y}}]$ and F is defined in Eq. (2.2). The first lemma gives us a key descent inequality.

Lemma 3.3 *Suppose that Assumption 2.4 and 2.5 hold true. Then, we have*

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

where $\mathbf{z}^\star = (\mathbf{x}^\star, \mathbf{y}^\star)$ is a global saddle point (cf. Assumption 2.4).

Proof. Using the definition of the Lyapunov function in Eq. (3.3) and $\hat{\mathbf{z}} = [\hat{\mathbf{x}}; \hat{\mathbf{y}}]$, we have

$$\mathcal{E}_k - \mathcal{E}_{k-1} = \frac{1}{2}\|\hat{\mathbf{z}}_k - \mathbf{z}_0\|^2 - \frac{1}{2}\|\hat{\mathbf{z}}_{k-1} - \mathbf{z}_0\|^2 = (\hat{\mathbf{z}}_k - \hat{\mathbf{z}}_{k-1})^\top(\hat{\mathbf{z}}_k - \mathbf{z}_0) - \frac{1}{2}\|\hat{\mathbf{z}}_k - \hat{\mathbf{z}}_{k-1}\|^2. \tag{3.4}$$

Note that **Step 5** shows that $\hat{\mathbf{z}}_k = \hat{\mathbf{z}}_{k-1} - \lambda_k F(\mathbf{z}_k)$. Plugging it into Eq. (3.4) yields

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

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

$$\sum_{k=1}^{t} \lambda_k(\mathbf{z}_k - \mathbf{z}^\star)^\top F(\mathbf{z}_k) \le \mathcal{E}_0 - \mathcal{E}_t + \underbrace{\sum_{k=1}^{t} \lambda_k(\mathbf{z}_0 - \mathbf{z}^\star)^\top F(\mathbf{z}_k)}_{\mathbf{I}} + \underbrace{\sum_{k=1}^{t} \lambda_k(\mathbf{z}_k - \hat{\mathbf{z}}_k)^\top F(\mathbf{z}_k) - \frac{1}{2}\|\hat{\mathbf{z}}_k - \hat{\mathbf{z}}_{k-1}\|^2}_{\mathbf{II}}. \tag{3.5}$$

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

$$\mathbf{I} = \sum_{k=1}^{t} \lambda_k(\mathbf{z}_0 - \mathbf{z}^\star)^\top F(\mathbf{z}_k) = \sum_{k=1}^{t}(\hat{\mathbf{z}}_{k-1} - \hat{\mathbf{z}}_k)^\top(\mathbf{z}_0 - \mathbf{z}^\star) = (\hat{\mathbf{z}}_0 - \hat{\mathbf{z}}_t)^\top(\mathbf{z}_0 - \mathbf{z}^\star). \tag{3.6}$$

In **Step 2** of Algorithm 1, we compute a pair $(\Delta\mathbf{x}_k, \Delta\mathbf{y}_k) \in \mathbb{R}^m \times \mathbb{R}^n$ such that it is an *exact* solution of the cubic regularized min-max optimization problem. Since this is an unconstrained and convex-concave min-max optimization problem, we can write down its optimality condition as follows,

$$\begin{bmatrix}\nabla_{\mathbf{x}} f(\hat{\mathbf{z}}_k) \\ \nabla_{\mathbf{y}} f(\hat{\mathbf{z}}_k)\end{bmatrix} + \begin{bmatrix}\nabla^2_{\mathbf{xx}} f(\hat{\mathbf{z}}_k) & \nabla^2_{\mathbf{xy}} f(\hat{\mathbf{z}}_k) \\ \nabla^2_{\mathbf{xy}} f(\hat{\mathbf{z}}_k) & \nabla^2_{\mathbf{yy}} f(\hat{\mathbf{z}}_k)\end{bmatrix}\begin{bmatrix}\Delta\mathbf{x}_k \\ \Delta\mathbf{y}_k\end{bmatrix} + \begin{bmatrix}6\rho\|\Delta\mathbf{x}_k\|\Delta\mathbf{x}_k \\ -6\rho\|\Delta\mathbf{y}_k\|\Delta\mathbf{y}_k\end{bmatrix} = \mathbf{0}.$$

Equivalently, we have

$$F(\hat{\mathbf{z}}_k) + DF(\hat{\mathbf{z}}_k)\Delta\mathbf{z}_k + 6\rho\begin{bmatrix}\|\Delta\mathbf{x}_k\|\Delta\mathbf{x}_k \\ \|\Delta\mathbf{y}_k\|\Delta\mathbf{y}_k\end{bmatrix} = \mathbf{0}. \tag{3.7}$$

Note that **Step 4** in the compact form is equivalent to $\mathbf{z}_k = \hat{\mathbf{z}}_{k-1} + \Delta\mathbf{z}_{k-1}$. By using Lemma 2.7, we have

$$\|F(\mathbf{z}_k) - F(\hat{\mathbf{z}}_{k-1}) - DF(\hat{\mathbf{z}}_{k-1})\Delta\mathbf{z}_{k-1}\| \leq \tfrac{\rho}{2}\|\Delta\mathbf{z}_{k-1}\|^2. \tag{3.8}$$

It suffices to decompose $(\mathbf{z}_k - \hat{\mathbf{z}}_k)^\top F(\mathbf{z}_k)$ and bound it using Eq. (3.7) and (3.8). Indeed, we have

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

Note that we have

$$\begin{aligned}
(\mathbf{x}_k - \hat{\mathbf{x}}_k)^\top\Delta\mathbf{x}_{k-1} &= \|\Delta\mathbf{x}_{k-1}\|^2 + (\Delta\mathbf{x}_{k-1})^\top(\hat{\mathbf{x}}_{k-1} - \hat{\mathbf{x}}_k) \geq \|\Delta\mathbf{x}_{k-1}\|^2 - \|\Delta\mathbf{x}_{k-1}\|\|\hat{\mathbf{x}}_{k-1} - \hat{\mathbf{x}}_k\|, \\
(\mathbf{y}_k - \hat{\mathbf{y}}_k)^\top\Delta\mathbf{y}_{k-1} &= \|\Delta\mathbf{y}_{k-1}\|^2 + (\Delta\mathbf{y}_{k-1})^\top(\hat{\mathbf{y}}_{k-1} - \hat{\mathbf{y}}_k) \geq \|\Delta\mathbf{y}_{k-1}\|^2 - \|\Delta\mathbf{y}_{k-1}\|\|\hat{\mathbf{y}}_{k-1} - \hat{\mathbf{y}}_k\|.
\end{aligned}$$

This implies

$$\begin{aligned}
&\|\Delta\mathbf{x}_{k-1}\|(\mathbf{x}_k - \hat{\mathbf{x}}_k)^\top\Delta\mathbf{x}_{k-1} + \|\Delta\mathbf{y}_{k-1}\|(\mathbf{y}_k - \hat{\mathbf{y}}_k)^\top\Delta\mathbf{y}_{k-1} \\
&\quad\geq \|\Delta\mathbf{x}_{k-1}\|^3 - \|\Delta\mathbf{x}_{k-1}\|^2\|\hat{\mathbf{x}}_{k-1} - \hat{\mathbf{x}}_k\| + \|\Delta\mathbf{y}_{k-1}\|^3 - \|\mathbf{y}_k - \hat{\mathbf{y}}_{k-1}\|^2\|\hat{\mathbf{y}}_{k-1} - \hat{\mathbf{y}}_k\| \\
&\quad\geq (\|\Delta\mathbf{x}_{k-1}\|^3 + \|\Delta\mathbf{y}_{k-1}\|^3) - \|\Delta\mathbf{z}_{k-1}\|^2\|\hat{\mathbf{z}}_{k-1} - \hat{\mathbf{z}}_k\| \\
&\quad\geq \tfrac{1}{2}\|\Delta\mathbf{z}_{k-1}\|^3 - \|\Delta\mathbf{z}_{k-1}\|^2\|\hat{\mathbf{z}}_{k-1} - \hat{\mathbf{z}}_k\|.
\end{aligned}$$

Putting these pieces together yields

$$(\mathbf{z}_k - \hat{\mathbf{z}}_k)^\top F(\mathbf{z}_k) \leq \tfrac{13\rho}{2}\|\Delta\mathbf{z}_{k-1}\|^2\|\|\hat{\mathbf{z}}_{k-1} - \hat{\mathbf{z}}_k\| - \tfrac{5\rho}{2}\|\Delta\mathbf{z}_{k-1}\|^3.$$

Since $\Delta\mathbf{z}_{k-1} = [\Delta\mathbf{x}_{k-1}; \Delta\mathbf{y}_{k-1}]$, we have **Step 3** of Algorithm 1 implies that $\frac{1}{15} \leq \lambda_k\rho\|\Delta\mathbf{z}_{k-1}\| \leq \frac{1}{13}$ for all $k \geq 1$. Thus, we have

$$\begin{aligned}
\mathbf{II} &\leq \sum_{k=1}^t \left(\tfrac{13\lambda_k\rho}{2}\|\Delta\mathbf{z}_{k-1}\|^2\|\|\hat{\mathbf{z}}_{k-1} - \hat{\mathbf{z}}_k\| - \tfrac{1}{2}\|\hat{\mathbf{z}}_{k-1} - \hat{\mathbf{z}}_k\|^2 - \tfrac{5\lambda_k\rho}{2}\|\Delta\mathbf{z}_{k-1}\|^3\right) \\
&\leq \sum_{k=1}^t \left(\tfrac{1}{2}\|\Delta\mathbf{z}_{k-1}\|\|\hat{\mathbf{z}}_{k-1} - \hat{\mathbf{z}}_k\| - \tfrac{1}{2}\|\hat{\mathbf{z}}_{k-1} - \hat{\mathbf{z}}_k\|^2 - \tfrac{1}{6}\|\Delta\mathbf{z}_{k-1}\|^2\right) \\
&\leq \sum_{k=1}^t \left(\max_{\eta \geq 0}\left\{\tfrac{1}{2}\|\Delta\mathbf{z}_{k-1}\|\eta - \tfrac{1}{2}\eta^2\right\} - \tfrac{1}{6}\|\Delta\mathbf{z}_{k-1}\|^2\right) \\
&= -\tfrac{1}{24}\left(\sum_{k=1}^t \|\Delta\mathbf{z}_{k-1}\|^2\right).
\end{aligned} \tag{3.9}$$

Plugging Eq. (3.6) and Eq. (3.9) into Eq. (3.5) and using $\hat{\mathbf{z}}_0 = \mathbf{z}_0$ and $\Delta\mathbf{z}_{k-1} = \mathbf{z}_k - \hat{\mathbf{z}}_{k-1}$ yields

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

This completes the proof. □

Lemma 3.4 *Suppose that Assumption 2.4 and 2.5 hold true. Then, we have*

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

where $\mathbf{z}^\star = (\mathbf{x}^\star, \mathbf{y}^\star)$ *is a global saddle point (cf. Assumption 2.4).*

Proof. Using the notation $\hat{\mathbf{z}} = [\hat{\mathbf{x}}; \hat{\mathbf{y}}]$, we have $\mathcal{E}_t = \frac{1}{2}\|\hat{\mathbf{z}}_t - \mathbf{z}_0\|^2$. Since $\hat{\mathbf{z}}_0 = \mathbf{z}_0$, we have

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

Since $\mathbf{z}^\star \in \mathbb{R}^{m+n}$ 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, this together with Lemma 3.3, yields

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

Further, Lemma 3.3 implies

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

Using Young's inequality, we have

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

This completes the proof. □

We provide a technical lemma establishing a lower bound for $\sum_{k=1}^t \lambda_k$.

Lemma 3.5 *Suppose that Assumption 2.4 and 2.5 hold true. Then, for every integer* $T \geq 1$*, we have*

$$\sum_{k=1}^{T} \lambda_k \geq \frac{T^{\frac{3}{2}}}{30\sqrt{3}\rho\|\mathbf{z}_0 - \mathbf{z}^\star\|},$$

where $\mathbf{z}^\star = (\mathbf{x}^\star, \mathbf{y}^\star)$ *is a global saddle point of* $f(\mathbf{x}, \mathbf{y})$ *in Assumption 2.4.*

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

$$\sum_{k=1}^{t} (\lambda_k)^{-2} (\tfrac{1}{15\rho})^2 \leq \sum_{k=1}^{t} (\lambda_k)^{-2} (\lambda_k \|\mathbf{z}_k - \hat{\mathbf{z}}_{k-1}\|)^2 = \sum_{k=1}^{t} \|\mathbf{z}_k - \hat{\mathbf{z}}_{k-1}\|^2 \overset{\text{Lemma 3.4}}{\leq} 12\|\mathbf{z}_0 - \mathbf{z}^\star\|^2.$$

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

$$t \leq \left(30\sqrt{3}\rho\|\mathbf{z}_0 - \mathbf{z}^\star\|\right)^{\frac{2}{3}} \left(\sum_{k=1}^{t} \lambda_k\right)^{\frac{2}{3}}.$$

Letting $t = T$ and rearranging yields

$$\sum_{k=1}^{T} \lambda_k \geq \frac{T^{\frac{3}{2}}}{30\sqrt{3}\rho\|\mathbf{z}_0 - \mathbf{z}^\star\|}.$$

This completes the proof. □

Proof of Theorem 3.1. By Lemma 3.4, 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 yields that the iterates $\{\mathbf{z}_k\}_{k\geq 0}$ and $\{\hat{\mathbf{z}}_k\}_{k\geq 0}$ generated by Algorithm 1 are bounded by an universal constant. For every integer $T \geq 1$, Lemma 3.4 also implies

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

By Proposition 2.8, we have

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

Putting these pieces together yields

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

This together with Lemma 3.5 yields

$$f(\bar{\mathbf{x}}_T, \mathbf{y}^\star) - f(\mathbf{x}^\star, \bar{\mathbf{y}}_T) \leq \frac{15\sqrt{3}\rho\|\mathbf{z}_0 - \mathbf{z}^\star\|^3}{T^{\frac{3}{2}}}.$$

By the definition of a gap function in Section 2, we have

$$\textsc{gap}(\bar{\mathbf{x}}_T, \bar{\mathbf{y}}_T) \leq \frac{15\sqrt{3}\rho\|\mathbf{z}_0 - \mathbf{z}^\star\|^3}{T^{\frac{3}{2}}}. \tag{3.10}$$

Therefore, we conclude from Eq. (3.10) that there exists some $T > 0$ such that the output $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = \mathsf{Newton\text{-}MinMax}((\mathbf{x}_0, \mathbf{y}_0), \rho, T)$ satisfies that $\textsc{gap}(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \leq \epsilon$ and the total number of calls of the subproblem solvers is bounded by

$$O\left(\left(\frac{\rho(\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 + \|\mathbf{y}_0 - \mathbf{y}^\star\|^2)^{\frac{3}{2}}}{\epsilon}\right)^{\frac{2}{3}}\right).$$

This completes the proof.

Algorithm 2 Inexact-Newton-MinMax$((\mathbf{x}_0, \mathbf{y}_0), \rho, T)$

Input: initial point $(\mathbf{x}_0, \mathbf{y}_0)$, Lipschitz parameter ρ and iteration number $T \geq 1$.
Initialization: set $\hat{\mathbf{x}}_0 = \mathbf{x}_0$ and $\hat{\mathbf{y}}_0 = \mathbf{y}_0$ as well as parameters $\kappa_H > 0$, $0 < \kappa_m < \min\{1, \frac{\rho}{4}\}$ and $0 < \tau_0 < \frac{\rho}{4}$.
for $k = 0, 1, 2, \ldots, T-1$ **do**
  STEP 1: If $(\mathbf{x}_k, \mathbf{y}_k)$ is a global saddle point of the min-max optimization problem, then **stop**.
  STEP 2: Compute an *approximate* solution $(\Delta\mathbf{x}_k, \Delta\mathbf{y}_k) \in \mathbb{R}^m \times \mathbb{R}^n$ of the cubic regularized min-max problem (we let $\hat{\mathbf{z}}_k = (\hat{\mathbf{x}}_k, \hat{\mathbf{y}}_k)$ for brevity):

$$\min_{\Delta\mathbf{x}} \max_{\Delta\mathbf{y}} \left\{ m_k(\Delta\mathbf{x}, \Delta\mathbf{y}) = \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\mathbf{y} \end{bmatrix}^\top \begin{bmatrix} \nabla_{\mathbf{x}} f(\hat{\mathbf{z}}_k) \\ \nabla_{\mathbf{y}} f(\hat{\mathbf{z}}_k) \end{bmatrix} + \frac{1}{2} \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\mathbf{y} \end{bmatrix}^\top H(\hat{\mathbf{z}}_k) \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\mathbf{y} \end{bmatrix} + 2\rho\|\Delta\mathbf{x}\|^3 - 2\rho\|\Delta\mathbf{y}\|^3 \right\}, \quad (4.1)$$

  given that Condition 4.1 and 4.2 hold true with a proper choice of $\tau_k > 0$.
  STEP 3: Compute $\lambda_{k+1} > 0$ such that $\frac{1}{15} \leq \lambda_{k+1}\rho\sqrt{\|\Delta\mathbf{x}_k\|^2 + \|\Delta\mathbf{y}_k\|^2} \leq \frac{1}{14}$.
  STEP 4a: Compute $\mathbf{x}_{k+1} = \hat{\mathbf{x}}_k + \Delta\mathbf{x}_k$.
  STEP 4b: Compute $\mathbf{y}_{k+1} = \hat{\mathbf{y}}_k + \Delta\mathbf{y}_k$.
  STEP 5a: Compute $\hat{\mathbf{x}}_{k+1} = \hat{\mathbf{x}}_k - \lambda_{k+1}\nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})$.
  STEP 5b: Compute $\hat{\mathbf{y}}_{k+1} = \hat{\mathbf{y}}_k + \lambda_{k+1}\nabla_{\mathbf{y}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})$.
end for
Output: $(\bar{\mathbf{x}}_T, \bar{\mathbf{y}}_T) = \frac{1}{\sum_{k=1}^T \lambda_k}\left(\sum_{k=1}^T \lambda_k \mathbf{x}_k, \sum_{k=1}^T \lambda_k \mathbf{y}_k\right)$.

4 Inexact Algorithm and Convergence Analysis

In this section, we present the scheme of Inexact-Newton-MinMax and prove the same global convergence guarantee. Different from the conceptual framework of Newton-MinMax in Algorithm 1, our algorithm is compatible with both *inexact second-order information* and *inexact subproblem solving*. Our inexactness conditions are inspired by Xu et al. [2020, Condition 1 and 4] and allows for direct construction through randomized sampling. As the consequence of our results, the sub-sampled Newton method is proposed for solving finite-sum min-max optimization with a global convergence rate guarantee.

4.1 Algorithmic scheme

We summarize our inexact second-order method, which we call Inexact-Newton-MinMax$((\mathbf{x}_0, \mathbf{y}_0), \rho, T)$, in Algorithm 2 where $(\mathbf{x}_0, \mathbf{y}_0) \in \mathbb{R}^m \times \mathbb{R}^n$ 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 combination of Algorithm 1 and the inexact second-order framework for nonconvex optimization [Xu et al., 2020] in the context of min-max optimization. More specifically, the difference between the subproblems in Eq. (3.1) and Eq. (4.1) is that the inexact Hessian $H(\hat{\mathbf{z}}_k) \in \mathbb{R}^{(m+n)\times(m+n)}$ is used to approximate the exact Hessian at $\hat{\mathbf{z}}_k$ and could be formed and evaluated *efficiently* in practice. Inspired by Xu et al. [2020, Condition 1] and Chen et al. [2022, Condition 3.1], we impose the conditions on the inexact Hessian and the inexact subproblem solving. Formally, we have

Condition 4.1 (Inexact Hessian Regularity) *For some $\kappa_H > 0$ and $\tau_k > 0$, the inexact Hessian $H(\hat{\mathbf{z}}_k)$ satisfies the following regularity conditions:*

$$\|(H(\hat{\mathbf{z}}_k) - \nabla^2 f(\hat{\mathbf{z}}_k))\Delta\mathbf{z}_k\| \leq \tau_k\|\Delta\mathbf{z}_k\|, \qquad \|H(\hat{\mathbf{z}}_k)\| \leq \kappa_H,$$

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) *For some $\kappa_m \in (0,1)$, we are able to solve the min-max subproblem in Eq. (4.1) approximately to find $\Delta \mathbf{z}_k = (\Delta \mathbf{x}_k, \Delta \mathbf{y}_k)$ such that*

$$\|\nabla m_k(\Delta \mathbf{z}_k)\| \leq \kappa_m \cdot \min\{\|\Delta \mathbf{z}_k\|^2, \|\nabla f(\hat{\mathbf{z}}_k)\|\},$$

where $m_k(\cdot)$ is defined in Eq. (4.1). In addition, $\{\hat{\mathbf{z}}_k\}_{k\geq 0}$ and $\{\Delta \mathbf{z}_k\}_{k \geq 0}$ are generated by Algorithm 2.

Under Condition 4.1 and 4.2, we prove that our proposed algorithm (cf. Algorithm 2) achieve the same *worst-case* iteration complexity to obtain an ϵ-global saddle point of the min-max optimization problem in Eq. (1.1) as that of the exact variant (cf. Algorithm 1); see Theorem 4.1 for the details.

For the merits of Condition 4.1 and its advantages over other approximation conditions in nonconvex optimization, we refer to Xu et al. [2020, Section 1.3.2]. Despite the convex-concave structure, f is nonconvex in $(\mathbf{x}, \mathbf{y})$. Thus, Condition 4.1 is suitable for min-max optimization and allows for theoretically principled use of many practical techniques to constructing $H(\hat{\mathbf{z}}_k)$.

One such scheme can be described explicitly in the context of large-scale *finite-sum* min-max optimization problems of the form

$$\min_{\mathbf{x} \in \mathbb{R}^m} \max_{\mathbf{y} \in \mathbb{R}^n} \ f(\mathbf{x}, \mathbf{y}) \triangleq \tfrac{1}{N} \sum_{i=1}^N f_i(\mathbf{x}, \mathbf{y}), \tag{4.2}$$

and its special instantiation

$$\min_{\mathbf{x} \in \mathbb{R}^m} \max_{\mathbf{y} \in \mathbb{R}^n} \ f(\mathbf{x}, \mathbf{y}) \triangleq \tfrac{1}{N} \sum_{i=1}^N f_i(\mathbf{a}_i^\top \mathbf{x}, \mathbf{b}_i^\top \mathbf{y}), \tag{4.3}$$

where $N \gg 1$, each of f_i is a convex-concave function with bounded and Lipschitz-continuous Hessian, and $\{(\mathbf{a}_i, \mathbf{b}_i)\}_{i=1}^N \subseteq \mathbb{R}^m \times \mathbb{R}^n$ are a few given data samples. This type of problems are common in machine learning and scientific computing [Shalev-Shwartz and Ben-David, 2014, Roosta-Khorasani et al., 2014, Roosta-Khorasani and Mahoney, 2019]. Then, we present both *uniform* and *nonuniform* sub-sampling schemes to guarantee Condition 4.1 with high probability. In fact, the theoretical results are a little bit stronger, namely

$$\|H(\hat{\mathbf{z}}_k) - \nabla^2 f(\hat{\mathbf{z}}_k)\| \leq \tau_k, \tag{4.4}$$

which will imply one of two key inequalities in Condition 4.1. As a consequence of Theorem 4.1, we give the first sub-sampled Newton method for solving finite-sum min-max optimization problems and prove the global convergence rate of $O(\epsilon^{-2/3})$ in the convex-concave case; see Theorem 4.8 for the details.

It remains to clarify how to approximately solve the cubic regularized min-max optimization problem in Eq. (4.1) such that Condition 4.2 holds true. We have conducted numerical experiments (see Section 5) and found that the inexact subproblem solving can be efficiently done using the semismooth Newton (SSN) method [Qi and Sun, 1993]. In terms of theoretical guarantee, Solodov and Svaiter [1998] established the global convergence by exploiting the monotone structure and derived a local superlinear convergence under the conditions that the generalized Jacobian is semismooth and nonsingular at an optimal solution. These results were later extended by Zhou and Toh [2005] to the setting where the generalized Jacobian is not necessarily nonsingular. In terms of practical efficiency, the SSN method has received a considerable amount of attention due to its success in solving a wide range of structured application problems [Ulbrich, 2011, Zhao et al., 2010, Milzarek and Ulbrich, 2014, Li et al., 2018a,b, Milzarek et al., 2019].

In what follows, we clarify how to approximately solve Eq. (4.1) using the SSN method. In particular, we rewrite this problem equivalently in the following form of

$$\min_{(\Delta \mathbf{x}, u) \in \mathcal{X}} \max_{(\Delta \mathbf{y}, v) \in \mathcal{Y}} \begin{bmatrix} \Delta \mathbf{x} \\ \Delta \mathbf{y} \end{bmatrix}^\top \begin{bmatrix} \nabla_{\mathbf{x}} f(\hat{\mathbf{z}}_k) \\ \nabla_{\mathbf{y}} f(\hat{\mathbf{z}}_k) \end{bmatrix} + \frac{1}{2} \begin{bmatrix} \Delta \mathbf{x} \\ \Delta \mathbf{y} \end{bmatrix}^\top \begin{bmatrix} H_{\mathbf{xx}}(\hat{\mathbf{z}}_k) & H_{\mathbf{xy}}(\hat{\mathbf{z}}_k) \\ H_{\mathbf{xy}}(\hat{\mathbf{z}}_k) & H_{\mathbf{yy}}(\hat{\mathbf{z}}_k) \end{bmatrix} \begin{bmatrix} \Delta \mathbf{x} \\ \Delta \mathbf{y} \end{bmatrix} + 2\rho u^3 - 2\rho v^3,$$

where the constraint sets are $\mathcal{X} = \{(\Delta \mathbf{x}, u) : \|\Delta \mathbf{x}\| \leq u\}$ and $\mathcal{Y} = \{(\Delta \mathbf{y}, v) : \|\Delta \mathbf{y}\| \leq v\}$ and are known as second-order cones [Lobo et al., 1998, Alizadeh and Goldfarb, 2003]. For simplicity, we let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $\Delta \mathbf{z} = (\Delta \mathbf{x}, u, \Delta \mathbf{y}, v)$. Then, it suffices to solve the nonsmooth equation in the form of

$$E(\Delta \mathbf{z}) \triangleq \Delta \mathbf{z} - P_{\mathcal{Z}} \left(\Delta \mathbf{z} - \begin{bmatrix} \nabla_{\mathbf{x}} f(\hat{\mathbf{z}}_k) + H_{\mathbf{xx}}(\hat{\mathbf{z}}_k) \Delta \mathbf{x} + H_{\mathbf{xy}}(\hat{\mathbf{z}}_k) \Delta \mathbf{y} \\ 6\rho u^2 \\ -\nabla_{\mathbf{y}} f(\hat{\mathbf{z}}_k) - H_{\mathbf{xy}}(\hat{\mathbf{z}}_k) \Delta \mathbf{x} - H_{\mathbf{yy}}(\hat{\mathbf{z}}_k) \Delta \mathbf{y} \\ 6\rho v^2 \end{bmatrix} \right) = 0. \tag{4.5}$$

We will use the SSN method to solve Eq. (4.5). Indeed, it is clear that the operator E is monotone and locally Lipschitz continuous over $\mathcal{Z}$ and we can apply the SSN method to find a zero of $E(\cdot)$. Then, it suffices to compute an element of the generalized Jacobian of $E(\Delta \mathbf{z})$. To that end, we define

$$J(\Delta \mathbf{z}) = I_{m+n+2} - J_{P_{\mathcal{Z}}}(\mathbf{w}) \cdot \left(I_{m+n+2} - \begin{bmatrix} H_{\mathbf{xx}}(\hat{\mathbf{z}}_k) & 0 & H_{\mathbf{xy}}(\hat{\mathbf{z}}_k) & 0 \\ 0 & 12\rho u & 0 & 0 \\ -H_{\mathbf{xy}}(\hat{\mathbf{z}}_k) & 0 & -H_{\mathbf{yy}}(\hat{\mathbf{z}}_k) & 0 \\ 0 & 0 & 0 & 12\rho v \end{bmatrix} \right), \tag{4.6}$$

where $\mathbf{w} \in \mathbb{R}^{m+n+2}$ is defined by

$$\mathbf{w} = \Delta \mathbf{z} - \begin{bmatrix} \nabla_{\mathbf{x}} f(\hat{\mathbf{z}}_k) + H_{\mathbf{xx}}(\hat{\mathbf{z}}_k) \Delta \mathbf{x} + H_{\mathbf{xy}}(\hat{\mathbf{z}}_k) \Delta \mathbf{y} \\ 6\rho u^2 \\ -\nabla_{\mathbf{y}} f(\hat{\mathbf{z}}_k) - H_{\mathbf{xy}}(\hat{\mathbf{z}}_k) \Delta \mathbf{x} - H_{\mathbf{yy}}(\hat{\mathbf{z}}_k) \Delta \mathbf{y} \\ 6\rho v^2 \end{bmatrix} = \begin{bmatrix} \mathbf{w}_{\mathbf{x}} \\ \mathbf{w}_u \\ \mathbf{w}_{\mathbf{y}} \\ \mathbf{w}_v \end{bmatrix}.$$

If $J_{P_{\mathcal{Z}}}$ is an element of the generalized Jacobian of $P_{\mathcal{Z}}$, we have $J(\Delta \mathbf{z})$ is an element of the generalized Jacobian of $E(\Delta \mathbf{z})$. Even though it is not easy to compute $J_{P_{\mathcal{Z}}}$ for a general convex set $\mathcal{Z}$, the structure of being the product of second-order cones implies that

$$J_{P_{\mathcal{Z}}}(\mathbf{w}) = \begin{bmatrix} J_{P_{\mathcal{X}}}(\mathbf{w}_{\mathbf{x}}, \mathbf{w}_u) & 0 \\ 0 & J_{P_{\mathcal{Y}}}(\mathbf{w}_{\mathbf{y}}, \mathbf{w}_v) \end{bmatrix} \in \mathbb{R}^{(m+n+2) \times (m+n+2)},$$

where $J_{P_{\mathcal{X}}}(\mathbf{w}_{\mathbf{x}}, \mathbf{w}_u)$ and $J_{P_{\mathcal{Y}}}(\mathbf{w}_{\mathbf{y}}, \mathbf{w}_v)$ are (see Kanzow et al. [2009, Lemma 2.5])

$$J_{P_{\mathcal{X}}}(\mathbf{w}_{\mathbf{x}}, \mathbf{w}_u) = \begin{cases} 0, & \mathbf{w}_u \in (-\infty, -\|\mathbf{w}_{\mathbf{x}}\|), \\ I_{m+1}, & \mathbf{w}_u \in (\|\mathbf{w}_{\mathbf{x}}\|, +\infty), \\ D, & \text{otherwise}, \end{cases} \quad \text{for } D = \frac{1}{2} \begin{bmatrix} \left(\frac{\mathbf{w}_u}{\|\mathbf{w}_{\mathbf{x}}\|} + 1 \right) I_m + \frac{\mathbf{w}_u \mathbf{w}_{\mathbf{x}} \mathbf{w}_{\mathbf{x}}^\top}{\|\mathbf{w}_{\mathbf{x}}\|^3} & \frac{\mathbf{w}_{\mathbf{x}}}{\|\mathbf{w}_{\mathbf{x}}\|} \\ \frac{\mathbf{w}_{\mathbf{x}}^\top}{\|\mathbf{w}_{\mathbf{x}}\|} & 1 \end{bmatrix},$$

and

$$J_{P_{\mathcal{Y}}}(\mathbf{w}_{\mathbf{y}}, \mathbf{w}_v) = \begin{cases} 0, & \mathbf{w}_v \in (-\infty, -\|\mathbf{w}_{\mathbf{y}}\|), \\ I_{n+1}, & \mathbf{w}_v \in (\|\mathbf{w}_{\mathbf{y}}\|, +\infty), \\ D, & \text{otherwise}, \end{cases} \quad \text{for } D = \frac{1}{2} \begin{bmatrix} \left(\frac{\mathbf{w}_v}{\|\mathbf{w}_{\mathbf{y}}\|} + 1 \right) I_n + \frac{\mathbf{w}_v \mathbf{w}_{\mathbf{y}} \mathbf{w}_{\mathbf{y}}^\top}{\|\mathbf{w}_{\mathbf{y}}\|^3} & \frac{\mathbf{w}_{\mathbf{y}}}{\|\mathbf{w}_{\mathbf{y}}\|} \\ \frac{\mathbf{w}_{\mathbf{y}}^\top}{\|\mathbf{w}_{\mathbf{y}}\|} & 1 \end{bmatrix}.$$

As a practical matter, we consider a simple combination of an adaptive regularized SSN method proposed by Xiao et al. [2018] and a generalized minimum residual (GMRES) method [Saad and Schultz, 1986]. Given the current iterate $\Delta \mathbf{z} \in \mathbb{R}^{m+n+2}$, we first compute the direction $\mathbf{d} \in \mathbb{R}^{m+n+2}$ by solving $(J(\Delta \mathbf{z}) + \eta I_{m+n+2})\mathbf{d} = -E(\Delta \mathbf{z})$ inexactly using the GMRES method. Then, we follow the strategy in Xiao et al. [2018] and decide if this $\mathbf{d}$ is accepted. Roughly speaking, if there is a sufficient decrease from $\|E(\Delta \mathbf{z})\|$ to $\|E(\Delta \mathbf{z} + \mathbf{d})\|$, we accept $\mathbf{d}$ and set the next iterate as $\Delta \mathbf{z} + \mathbf{d}$. Otherwise, a safeguard step is taken. To evaluate $E(\Delta \mathbf{z})$ at each iteration, we derive the closed-form expression of $P_{\mathcal{Z}}(\mathbf{w})$ as follows,

$$P_{\mathcal{Z}}(\mathbf{w}) = \begin{bmatrix} P_{\mathcal{X}}(\mathbf{w}_{\mathbf{x}}, \mathbf{w}_u) \\ P_{\mathcal{Y}}(\mathbf{w}_{\mathbf{y}}, \mathbf{w}_v) \end{bmatrix} \in \mathbb{R}^{m+n+2},$$

where $P_{\mathcal{X}}(\mathbf{w}_{\mathbf{x}}, \mathbf{w}_u)$ and $P_{\mathcal{Y}}(\mathbf{w}_{\mathbf{y}}, \mathbf{w}_v)$ are (see Kanzow et al. [2009, Lemma 2.2])

$$P_{\mathcal{X}}(\mathbf{w}_{\mathbf{x}}, \mathbf{w}_u) = \begin{cases} 0, & \mathbf{w}_u \in (-\infty, -\|\mathbf{w}_{\mathbf{x}}\|), \\ \begin{bmatrix} \mathbf{w}_{\mathbf{x}} \\ \mathbf{w}_u \end{bmatrix}, & \mathbf{w}_u \in (\|\mathbf{w}_{\mathbf{x}}\|, +\infty), \\ \frac{1}{2}\left(1 + \frac{\mathbf{w}_u}{\|\mathbf{w}_{\mathbf{x}}\|}\right) \begin{bmatrix} \mathbf{w}_{\mathbf{x}} \\ \|\mathbf{w}_{\mathbf{x}}\| \end{bmatrix}, & \text{otherwise,} \end{cases}$$

and

$$P_{\mathcal{Y}}(\mathbf{w}_{\mathbf{y}}, \mathbf{w}_v) = \begin{cases} 0, & \mathbf{w}_v \in (-\infty, -\|\mathbf{w}_{\mathbf{y}}\|), \\ \begin{bmatrix} \mathbf{w}_{\mathbf{y}} \\ \mathbf{w}_v \end{bmatrix}, & \mathbf{w}_v \in (\|\mathbf{w}_{\mathbf{y}}\|, +\infty), \\ \frac{1}{2}\left(1 + \frac{\mathbf{w}_v}{\|\mathbf{w}_{\mathbf{y}}\|}\right) \begin{bmatrix} \mathbf{w}_{\mathbf{y}} \\ \|\mathbf{w}_{\mathbf{y}}\| \end{bmatrix}, & \text{otherwise.} \end{cases}$$

Remarkably, Kanzow et al. [2009, Lemma 2.3] guarantees that $P_{\mathcal{Z}}$ is strongly semismooth, which has been used to establish a strong local convergence guarantee for the adaptive regularized SSN method. For more details on the algorithm and its convergence property, we refer the reader to Xiao et al. [2018].

4.2 Convergence analysis

We provide our main results on the convergence rate for Algorithm 2 in terms of the number of calls of the subproblem solvers. The following theorem gives us the global convergence rate of Algorithm 2 for convex-concave min-max optimization problems.

Theorem 4.1 *Suppose that Assumption 2.4 and 2.5 hold true, $0 < \tau_k \leq \min\{\tau_0, \frac{\rho(1-\kappa_m)}{4(\kappa_H + 6\rho)}\|\nabla f(\hat{\mathbf{z}}_k)\|\}$ for all $k \geq 0$ and $\epsilon \in (0, 1)$, the iterates generated by Algorithm 2 are bounded by an universal constant. Moreover, there exists some $T > 0$ such that the output $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) =$ Inexact-Newton-MinMax$((\mathbf{x}_0, \mathbf{y}_0), \rho, T)$ satisfies that $\text{GAP}(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \leq \epsilon$ and the total number of calls of the subproblem solvers is bounded by*

$$O\left(\left(\frac{\rho(\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 + \|\mathbf{y}_0 - \mathbf{y}^\star\|^2)^{\frac{3}{2}}}{\epsilon}\right)^{\frac{2}{3}}\right),$$

where $(\mathbf{x}^\star, \mathbf{y}^\star) \in \mathbb{R}^m \times \mathbb{R}^n$ is a global saddle point (cf. Assumption 2.4) and $\rho > 0$ is a Lipschitz constant for the Hessian of the function f (cf. Assumption 2.5).

Remark 4.2 *Theorem 4.1 demonstrates that Algorithm 2 achieves the same global convergence rate as Algorithm 1. To be more precise, Algorithm 2 can achieve the optimal convergence guarantee regardless of inexact Hessian and inexact subproblem solving under Condition 4.1 and 4.2. Our subsequent analysis is based on a combination of techniques for proving Theorem 3.1 and Xu et al. [2020, Lemma 14].*

We use a Lyapunov function defined in Eq. (3.3) but for the iterates generated by Algorithm 2. Recall that $\mathbf{z} = [\mathbf{x}; \mathbf{y}]$, $\hat{\mathbf{z}} = [\hat{\mathbf{x}}; \hat{\mathbf{y}}]$ and F is defined in Eq. (2.2). The first lemma is analogous to Lemma 3.3.

Lemma 4.3 *Suppose that Assumption 2.4 and 2.5 hold true and* $0 < \tau_k \leq \min\{\tau_0, \frac{\rho(1-\kappa_m)}{4(\kappa_H+6\rho)}\|\nabla f(\hat{\mathbf{z}}_k)\|\}$. *Then, we have*

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

where $\mathbf{z}^\star = (\mathbf{x}^\star, \mathbf{y}^\star)$ *is a global saddle point (cf. Assumption 2.4).*

Proof. By using the same argument as used in Lemma 3.3, we have

$$\sum_{k=1}^t \lambda_k(\mathbf{z}_k - \mathbf{z}^\star)^\top F(\mathbf{z}_k) \leq \mathcal{E}_0 - \mathcal{E}_t + (\hat{\mathbf{z}}_0 - \hat{\mathbf{z}}_t)^\top(\mathbf{z}_0 - \mathbf{z}^\star) + \sum_{k=1}^t \lambda_k(\mathbf{z}_k - \hat{\mathbf{z}}_k)^\top F(\mathbf{z}_k) - \tfrac{1}{2}\|\hat{\mathbf{z}}_k - \hat{\mathbf{z}}_{k-1}\|^2. \quad (4.7)$$

It remains to bound the term $\sum_{k=1}^t \lambda_k(\mathbf{z}_k - \hat{\mathbf{z}}_k)^\top F(\mathbf{z}_k) - \frac{1}{2}\|\hat{\mathbf{z}}_k - \hat{\mathbf{z}}_{k-1}\|^2$. Indeed, in **Step 2** of Algorithm 2, we compute a pair $(\Delta\mathbf{x}_k, \Delta\mathbf{y}_k) \in \mathbb{R}^m \times \mathbb{R}^n$ such that it is an *inexact* solution of the min-max optimization subproblem in Eq. (4.1) under Condition 4.1 and 4.2. Note that Condition 4.2 can be written as follows,

$$\|\nabla m_k(\Delta\mathbf{z}_k)\| = \left\|\nabla f(\hat{\mathbf{z}}_k) + H(\hat{\mathbf{z}}_k)\Delta\mathbf{z}_k + \begin{bmatrix} 6\rho\|\Delta\mathbf{x}_k\|\Delta\mathbf{x}_k \\ -6\rho\|\Delta\mathbf{y}_k\|\Delta\mathbf{y}_k \end{bmatrix}\right\| \leq \kappa_m \cdot \min\{\|\Delta\mathbf{z}_k\|^2, \|\nabla f(\hat{\mathbf{z}}_k)\|\}.$$

Define $J = \begin{bmatrix} I_m & \\ & -I_n \end{bmatrix}$, we have

$$\left\|F(\hat{\mathbf{z}}_k) + JH(\hat{\mathbf{z}}_k)\Delta\mathbf{z}_k + 6\rho\begin{bmatrix} \|\Delta\mathbf{x}_k\|\Delta\mathbf{x}_k \\ \|\Delta\mathbf{y}_k\|\Delta\mathbf{y}_k \end{bmatrix}\right\| \leq \kappa_m \cdot \min\{\|\Delta\mathbf{z}_k\|^2, \|\nabla f(\hat{\mathbf{z}}_k)\|\}. \quad (4.8)$$

Combining **Step 4** of Algorithm 2 and Lemma 2.7, we have

$$\|F(\mathbf{z}_k) - F(\hat{\mathbf{z}}_{k-1}) - DF(\hat{\mathbf{z}}_{k-1})\Delta\mathbf{z}_{k-1}\| \leq \tfrac{\rho}{2}\|\Delta\mathbf{z}_{k-1}\|^2. \quad (4.9)$$

It suffices to decompose $(\mathbf{z}_k - \hat{\mathbf{z}}_k)^\top F(\mathbf{z}_k)$ and bound it using Condition 4.1, Eq. (4.8) and (4.9). Indeed, we have

$$\begin{aligned}
(\mathbf{z}_k - \hat{\mathbf{z}}_k)^\top F(\mathbf{z}_k) \leq{} & \|\mathbf{z}_k - \hat{\mathbf{z}}_k\|\|F(\mathbf{z}_k) - F(\hat{\mathbf{z}}_{k-1}) - DF(\hat{\mathbf{z}}_{k-1})\Delta\mathbf{z}_{k-1}\| \\
& + \|\mathbf{z}_k - \hat{\mathbf{z}}_k\|\left\|F(\hat{\mathbf{z}}_{k-1}) + JH(\hat{\mathbf{z}}_{k-1})\Delta\mathbf{z}_{k-1} + 6\rho\begin{bmatrix} \|\Delta\mathbf{x}_{k-1}\|\Delta\mathbf{x}_{k-1} \\ \|\Delta\mathbf{y}_{k-1}\|\Delta\mathbf{y}_{k-1} \end{bmatrix}\right\| \\
& - 6\rho(\|\Delta\mathbf{x}_{k-1}\|(\mathbf{x}_k - \hat{\mathbf{x}}_k)^\top\Delta\mathbf{x}_{k-1} + \|\Delta\mathbf{y}_{k-1}\|(\mathbf{y}_k - \hat{\mathbf{y}}_k)^\top\Delta\mathbf{y}_{k-1}) \\
& + \|\mathbf{z}_k - \hat{\mathbf{z}}_k\|\|(DF(\hat{\mathbf{z}}_{k-1}) - JH(\hat{\mathbf{z}}_{k-1}))\Delta\mathbf{z}_{k-1}\|.
\end{aligned}$$

The first and second terms can be bounded using Eq. (4.8) and (4.9). The third term can be bounded using the same argument from the proof of Lemma 3.3. For the fourth term, we have

$$\begin{aligned}\|(DF(\hat{\mathbf{z}}_{k-1}) - JH(\hat{\mathbf{z}}_{k-1}))\Delta\mathbf{z}_{k-1}\| &= \|J(\nabla^2 f(\hat{\mathbf{z}}_{k-1}) - H(\hat{\mathbf{z}}_{k-1}))\Delta\mathbf{z}_{k-1}\| \\ = \quad \|(\nabla^2 f(\hat{\mathbf{z}}_{k-1}) - H(\hat{\mathbf{z}}_{k-1}))\Delta\mathbf{z}_{k-1}\| &\overset{\text{Condition 4.1}}{\leq} \tau_{k-1}\|\Delta\mathbf{z}_{k-1}\|.\end{aligned}$$

Putting these pieces together yields that

$$\begin{aligned}(\mathbf{z}_k - \hat{\mathbf{z}}_k)^\top F(\mathbf{z}_k) \quad \leq \quad & \|\mathbf{z}_k - \hat{\mathbf{z}}_k\| \left(\left(\tfrac{\rho}{2} + \kappa_m\right)\|\Delta\mathbf{z}_{k-1}\|^2 + \tau_{k-1}\|\Delta\mathbf{z}_{k-1}\|\right) - 3\rho\|\Delta\mathbf{z}_{k-1}\|^3 \\ & +6\rho\|\Delta\mathbf{z}_{k-1}\|^2\|\hat{\mathbf{z}}_{k-1} - \hat{\mathbf{z}}_k\|.\end{aligned} \tag{4.10}$$

We claim that

$$\left(\tfrac{\rho}{2} + \kappa_m\right)\|\Delta\mathbf{z}_{k-1}\|^2 + \tau_{k-1}\|\Delta\mathbf{z}_{k-1}\| \leq \rho\|\Delta\mathbf{z}_{k-1}\|^2. \tag{4.11}$$

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

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

The above inequality together with $0 < \kappa_m < \min\{1, \frac{\rho}{4}\}$ and $\tau_{k-1} \leq \tau_0 < \frac{\rho}{4}$ yields Eq. (4.11). Otherwise, we have $\|\Delta\mathbf{z}_{k-1}\| < 1$ and obtain from Condition 4.1 and 4.2 that

$$\begin{aligned}\kappa_m\|\nabla f(\hat{\mathbf{z}}_{k-1})\| \quad \geq \quad & \left\|\nabla f(\hat{\mathbf{z}}_{k-1}) + H(\hat{\mathbf{z}}_{k-1})\Delta\mathbf{z}_{k-1} + \begin{bmatrix} 6\rho\|\Delta\mathbf{x}_{k-1}\|\Delta\mathbf{x}_{k-1} \\ -6\rho\|\Delta\mathbf{y}_{k-1}\|\Delta\mathbf{y}_{k-1}\end{bmatrix}\right\| \\ \geq \quad & \|\nabla f(\hat{\mathbf{z}}_{k-1})\| - \kappa_H\|\Delta\mathbf{z}_{k-1}\| - 6\rho\|\Delta\mathbf{z}_{k-1}\|^2 \\ \geq \quad & \|\nabla f(\hat{\mathbf{z}}_{k-1})\| - (\kappa_H + 6\rho)\|\Delta\mathbf{z}_{k-1}\|.\end{aligned}$$

Rearranging the above inequality and using $0 < \tau_k \leq \frac{\rho(1-\kappa_m)}{4(\kappa_H+6\rho)}\|\nabla f(\hat{\mathbf{z}}_k)\|$ yields

$$\|\Delta\mathbf{z}_{k-1}\| \geq \tfrac{1-\kappa_m}{\kappa_H+6\rho}\|\nabla f(\hat{\mathbf{z}}_{k-1})\| \geq \tfrac{4\tau_{k-1}}{\rho}.$$

Using $0 < \kappa_m < \min\{1, \frac{\rho}{4}\}$ again, we have

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

Plugging Eq. (4.11) into Eq. (4.10) 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

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

Since $\Delta\mathbf{z}_{k-1} = [\Delta\mathbf{x}_{k-1}; \Delta\mathbf{y}_{k-1}]$, we have **Step 3** of Algorithm 2 implies that $\frac{1}{15} \leq \lambda_k\rho\|\Delta\mathbf{z}_{k-1}\| \leq \frac{1}{14}$

for all $k \geq 1$. Thus, we have

$$\begin{aligned}
&\sum_{k=1}^t \lambda_k (\mathbf{z}_k - \hat{\mathbf{z}}_k)^\top F(\mathbf{z}_k) - \tfrac{1}{2}\|\hat{\mathbf{z}}_k - \hat{\mathbf{z}}_{k-1}\|^2 \\
&\quad \leq \sum_{k=1}^t \left(7\lambda_k \rho \|\Delta \mathbf{z}_{k-1}\|^2 \|\|\hat{\mathbf{z}}_{k-1} - \hat{\mathbf{z}}_k\| - \tfrac{1}{2}\|\hat{\mathbf{z}}_{k-1} - \hat{\mathbf{z}}_k\|^2 - 2\lambda_k \rho \|\Delta \mathbf{z}_{k-1}\|^3\right) \\
&\quad \leq \sum_{k=1}^t \left(\tfrac{1}{2}\|\Delta \mathbf{z}_{k-1}\|\|\hat{\mathbf{z}}_{k-1} - \hat{\mathbf{z}}_k\| - \tfrac{1}{2}\|\hat{\mathbf{z}}_{k-1} - \hat{\mathbf{z}}_k\|^2 - \tfrac{2}{15}\|\Delta \mathbf{z}_{k-1}\|^2\right) \\
&\quad \leq \sum_{k=1}^t \left(\max_{\eta \geq 0} \left\{\tfrac{1}{2}\|\Delta \mathbf{z}_{k-1}\|\eta - \tfrac{1}{2}\eta^2\right\} - \tfrac{2}{15}\|\Delta \mathbf{z}_{k-1}\|^2\right) \\
&\quad = -\tfrac{1}{40}\left(\sum_{k=1}^t \|\Delta \mathbf{z}_{k-1}\|^2\right).
\end{aligned}$$

Therefore, we conclude from Eq. (4.7), $\hat{\mathbf{z}}_0 = \mathbf{z}_0$ and $\Delta \mathbf{z}_{k-1} = \mathbf{z}_k - \hat{\mathbf{z}}_{k-1}$ that

$$\sum_{k=1}^t \lambda_k (\mathbf{z}_k - \mathbf{z}^\star)^\top F(\mathbf{z}_k) \leq \mathcal{E}_0 - \mathcal{E}_t + (\mathbf{z}_0 - \hat{\mathbf{z}}_t)^\top (\mathbf{z}_0 - \mathbf{z}^\star) - \tfrac{1}{40}\left(\sum_{k=1}^t \|\mathbf{z}_k - \hat{\mathbf{z}}_{k-1}\|^2\right).$$

This completes the proof. □

Proof of Theorem 4.1. Applying the same argument for proving Lemma 3.4 with Lemma 4.3 instead of Lemma 3.3, we have

$$\|\mathbf{z}_{k+1} - \hat{\mathbf{z}}_k\|^2 \leq 20\|\mathbf{z}_0 - \mathbf{z}^\star\|^2, \quad \|\hat{\mathbf{z}}_k - \mathbf{z}_0\| \leq 2\|\mathbf{z}_0 - \mathbf{z}^\star\|, \quad \text{for all } k \geq 0,$$

and

$$\sum_{k=1}^T \lambda_k (\mathbf{z}_k - \mathbf{z}^\star)^\top F(\mathbf{z}_k) \leq \tfrac{1}{2}\|\mathbf{z}_0 - \mathbf{z}^\star\|^2, \quad \text{for every integer } T \geq 1.$$

This above inequalities imply that $\|\mathbf{z}_k - \mathbf{z}_0\| \leq 7\|\mathbf{z}_0 - \mathbf{z}^\star\|$ for all $k \geq 0$. Putting these pieces yields that the iterates $\{\mathbf{z}_k\}_{k \geq 0}$ and $\{\hat{\mathbf{z}}_k\}_{k \geq 0}$ generated by Algorithm 2 are bounded by a universal constant. By Proposition 2.8, we have

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

Putting these pieces together yields

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

By using the same argument as used in Lemma 3.5 with $\|\mathbf{z}_{k+1} - \hat{\mathbf{z}}_k\|^2 \leq 20\|\mathbf{z}_0 - \mathbf{z}^\star\|^2$, we have $\sum_{k=1}^T \lambda_k \geq \frac{T^{\frac{3}{2}}}{30\sqrt{5}\rho\|\mathbf{z}_0 - \mathbf{z}^\star\|}$. Applying the same argument used for proving Theorem 3.1, we have

$$\text{GAP}(\bar{\mathbf{x}}_T, \bar{\mathbf{y}}_T) \leq \frac{15\sqrt{5}\rho\|\mathbf{z}_0 - \mathbf{z}^\star\|^3}{T^{\frac{3}{2}}}.$$

Therefore, there exists some $T > 0$ such that the output $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = \mathsf{Inexact\text{-}Newton\text{-}MinMax}((\mathbf{x}_0, \mathbf{y}_0), \rho, T)$ satisfies that $\text{GAP}(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \leq \epsilon$ and the total number of calls of the subproblem solvers is bounded by

$$O\left(\left(\frac{\rho(\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 + \|\mathbf{y}_0 - \mathbf{y}^\star\|^2)^{\frac{3}{2}}}{\epsilon}\right)^{\frac{2}{3}}\right).$$

This completes the proof.

4.3 Finite-sum min-max optimization

We give concrete examples to clarify the ways to construct the inexact Hessian such that Condition 4.1 holds true. The key ingredient here 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 the finite-sum min-max optimization problems in the form of Eq. (4.2) and (4.3).

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), we can construct the inexact Hessian as follows,

$$H(\mathbf{z}) = \tfrac{1}{N|\mathcal{S}|} \sum_{i \in \mathcal{S}} \tfrac{1}{p_i} \nabla^2 f_i(\mathbf{z}). \tag{4.12}$$

This construction is referred to as the sub-sampled Hessian and offers significant computational savings if $|\mathcal{S}| \ll N$ in big-data regime when $N \gg 1$.

In the general finite-sum setting with Eq. (4.2), we suppose

$$\sup_{(\mathbf{x},\mathbf{y}) \in \mathbb{R}^m \times \mathbb{R}^n} \|\nabla^2 f_i(\mathbf{z})\| \leq B_i, \quad \text{for all } i \in \{1, 2, \ldots, N\}, \tag{4.13}$$

and let $B_{\max} = \max_{1 \leq i \leq N} B_i$. Then, we consider sampling with the uniform distribution over $\{1, 2, \ldots, N\}$, i.e., $p_i = \frac{1}{N}$ and summarize the sample complexity results in the following lemma. The proof is omitted for brevity and we refer to Xu et al. [2020, Lemma 16] for the details.

Lemma 4.4 *Suppose that Eq. (4.13) holds true and let $B_{\max}$ be defined accordingly and $0 < \tau, \delta < 1$. A uniform sampling with or without replacement is performed to form the sub-sampled Hessian; indeed, $H(\mathbf{z})$ is constructed using Eq. (4.12) with $p_i = \frac{1}{n}$ and the sample size satisfies*

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

Then, we have

$$\text{Prob}(\|H(\mathbf{z}) - \nabla^2 f(\mathbf{z})\| \leq \tau) \geq 1 - \delta.$$

Remark 4.5 *Lemma 4.4 demonstrates that the inexact Hessian satisfies Condition 4.1 with probability $1-\delta$ under certain τ and κ_H if it is constructed using the uniform sampling and $|\mathcal{S}| = \Omega(\frac{B_{\max}^2}{\tau^2} \log(\frac{m+n}{\delta}))$. Indeed, the first inequality holds true with probability $1 - \delta$ since $\text{Prob}(\|H(\mathbf{z}) - \nabla^2 f(\mathbf{z})\| \leq \tau) \geq 1 - \delta$, and the second inequality is satisfied with $\kappa_H = B_{\max}$ (this is a deterministic statement).*

In the special finite-sum setting with Eq. (4.3), we are able to construct a more "informative" distribution of sampling $\xi \in \{1, 2, \ldots, N\}$ as opposed to simple uniform sampling. In particular, it is advantageous to bias the probability distribution towards picking indices corresponding to those *relevant* f_i's carefully in forming the Hessian. However, the construction of inexact Hessian and corresponding sample complexity guarantee from Xu et al. [2020, Section 3.1] in nonconvex optimization requires that $\nabla^2 f_i$ has rank one, which is not valid in our case. To generalize Xu et al. [2020, Lemma 17], we avail ourselves directly of the operator-Bernstein inequality [Gross and Nesme, 2010, Theorem 1].

It is clear that the Hessian of f in this case can be written as $\nabla^2 f(\mathbf{z}) = \frac{1}{N}\sum_{i=1}^N D_i f_i''(\mathbf{a}_i^\top \mathbf{x}, \mathbf{b}_i^\top \mathbf{y}) D_i^\top$ where $D_i = \begin{bmatrix} \mathbf{a}_i & \\ & \mathbf{b}_i \end{bmatrix} \in \mathbb{R}^{(m+n)\times 2}$. In the compact form, we have $\nabla^2 f(\mathbf{z}) = D^\top \Sigma D$ where

$$D^\top = \begin{bmatrix} | & \cdots & | \\ D_1 & \cdots & D_N \\ | & \cdots & | \end{bmatrix} \text{ and } \Sigma = \frac{1}{N}\begin{bmatrix} f_1''(\mathbf{a}_i^\top \mathbf{x}, \mathbf{b}_i^\top \mathbf{y}) & \cdots & \cdots \\ \vdots & \ddots & \vdots \\ \cdots & \cdots & f_N''(\mathbf{a}_i^\top \mathbf{x}, \mathbf{b}_i^\top \mathbf{y}) \end{bmatrix}. \tag{4.14}$$

We suppose

$$\sup_{(\mathbf{x},\mathbf{y})\in\mathbb{R}^m\times\mathbb{R}^n} \|f_i''(\mathbf{a}_i^\top \mathbf{x}, \mathbf{b}_i^\top \mathbf{y})\|(\|\mathbf{a}_i\|^2 + \|\mathbf{b}_i\|^2) \leq B_i, \quad \text{for all } i \in \{1, 2, \ldots, N\}, \tag{4.15}$$

and let $B_{\text{avg}} = \frac{1}{N}\sum_{i=1}^N B_i$. Then, we consider sampling with the nonuniform distribution over $\{1, 2, \ldots, N\}$ as follows,

$$p_i = \frac{\|f_i''(\mathbf{a}_i^\top \mathbf{x}, \mathbf{b}_i^\top \mathbf{y})\|(\|\mathbf{a}_i\|^2 + \|\mathbf{b}_i\|^2)}{\sum_{j=1}^N \|f_j''(\mathbf{a}_j^\top \mathbf{x}, \mathbf{b}_j^\top \mathbf{y})\|(\|\mathbf{a}_j\|^2 + \|\mathbf{b}_j\|^2)}. \tag{4.16}$$

The following lemma summarizes the results on the sample complexity.

Lemma 4.6 *Suppose that Eq. (4.15) holds true and let B_{avg} be defined accordingly and $0 < \tau, \delta < 1$. A nonuniform sampling is performed to form the sub-sampled Hessian; indeed, $H(\mathbf{z})$ is constructed using Eq. (4.12) with $p_i > 0$ in Eq. (4.16) 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

$$\text{Prob}(\|H(\mathbf{z}) - \nabla^2 f(\mathbf{z})\| \leq \tau) \geq 1 - \delta.$$

Proof. For any fixed $\mathbf{z} \in \mathbb{R}^{m+n}$, we obtain from Eq. (4.12) and (4.14) that $H(\mathbf{z}) = \frac{1}{|\mathcal{S}|}\sum_{j=1}^{|\mathcal{S}|} H_j$ where each random matrix $H_j \in \mathbb{R}^{(m+n)\times(m+n)}$ is random and satisfies that $\text{Prob}(H_j = \frac{1}{p_i} D_i \Sigma_{ii} D_i^\top) = p_i$ with $p_i > 0$ in Eq. (4.16). For simplicity, we define

$$X_j = H_j - \nabla^2 f(\mathbf{z}) = H_j - D^\top \Sigma D, \qquad X = \sum_{j=1}^{|\mathcal{S}|} X_j = |\mathcal{S}|\left(H(\mathbf{z}) - D^\top \Sigma D\right).$$

Applying a similar argument as used for proving Xu et al. [2020, Lemma 17], we have $\mathbb{E}[X_j] = 0$ and

$$\begin{aligned} \|\mathbb{E}[X_j^2]\| \leq \|\mathbb{E}[H_j^2]\| &= \left\|\sum_{i=1}^N p_i \left(\frac{1}{p_i} D_i \Sigma_{ii} D_i^\top\right)^2\right\| = \left\|\sum_{i=1}^N \frac{1}{p_i} D_i \Sigma_{ii} D_i^\top D_i \Sigma_{ii} D_i^\top\right\| \\ &\leq \sum_{i=1}^N \frac{1}{N^2 p_i}\|f_i''(\mathbf{a}_i^\top \mathbf{x}, \mathbf{b}_i^\top \mathbf{y})\|^2(\|\mathbf{a}_i\|^2 + \|\mathbf{b}_i\|^2)^2 = \left(\frac{1}{N}\sum_{i=1}^N \|f_i''(\mathbf{a}_i^\top \mathbf{x}, \mathbf{b}_i^\top \mathbf{y})\|(\|\mathbf{a}_i\|^2 + \|\mathbf{b}_i\|^2)\right)^2. \end{aligned}$$

Algorithm 3 Subsampled-Newton-MinMax($(\mathbf{x}_0, \mathbf{y}_0)$, ρ, T, δ)

Input: initial point $(\mathbf{x}_0, \mathbf{y}_0)$, Lipschitz parameter ρ, iteration number $T \geq 1$ and failure probability $\delta \in (0, 1)$.
Initialization: set $\hat{\mathbf{x}}_0 = \mathbf{x}_0$ and $\hat{\mathbf{y}}_0 = \mathbf{y}_0$ as well as parameters $0 < \kappa_m < \min\{1, \frac{\rho}{4}\}$ and $0 < \tau_0 < \frac{\rho}{4}$.
for $k = 0, 1, 2, \ldots, T-1$ **do**
STEP 1: If $(\mathbf{x}_k, \mathbf{y}_k)$ is a global saddle point of the min-max optimization problem, then **stop**.
STEP 2: Construct the inexact Hessian $H(\hat{\mathbf{z}}_k)$ using Eq. (4.12) and the sample set $|\mathcal{S}| \geq \Theta^U(\tau_k, 1 - \sqrt[T]{1-\delta})$ (uniform) or $|\mathcal{S}| \geq \Theta^N(\tau_k, 1 - \sqrt[T]{1-\delta})$ (non-uniform) given $0 < \tau_k \leq \min\{\tau_0, \frac{\rho(1-\kappa_m)}{4(B_{\max}+6\rho)}\|\nabla f(\hat{\mathbf{z}}_k)\|\}$.
STEP 3: Compute a pair $(\Delta\mathbf{x}_k, \Delta\mathbf{y}_k) \in \mathbb{R}^m \times \mathbb{R}^n$ such that it is an *inexact* solution of the following cubic regularized min-max optimization problem: (we let $\hat{\mathbf{z}}_k = (\hat{\mathbf{x}}_k, \hat{\mathbf{y}}_k)$ for brevity)

$$\min_{\Delta\mathbf{x}} \max_{\Delta\mathbf{y}} \left\{ m_k(\Delta\mathbf{x}, \Delta\mathbf{y}) = \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\mathbf{y} \end{bmatrix}^\top \begin{bmatrix} \nabla_{\mathbf{x}} f(\hat{\mathbf{z}}_k) \\ \nabla_{\mathbf{y}} f(\hat{\mathbf{z}}_k) \end{bmatrix} + \frac{1}{2} \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\mathbf{y} \end{bmatrix}^\top H(\hat{\mathbf{z}}_k) \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\mathbf{y} \end{bmatrix} + 2\rho\|\Delta\mathbf{x}\|^3 - 2\rho\|\Delta\mathbf{y}\|^3 \right\},$$

given that Condition 4.2 hold true.
STEP 3: Compute $\lambda_{k+1} > 0$ such that $\frac{1}{15} \leq \lambda_{k+1}\rho\sqrt{\|\Delta\mathbf{x}_k\|^2 + \|\Delta\mathbf{y}_k\|^2} \leq \frac{1}{14}$.
STEP 4a: Compute $\mathbf{x}_{k+1} = \hat{\mathbf{x}}_k + \Delta\mathbf{x}_k$.
STEP 4b: Compute $\mathbf{y}_{k+1} = \hat{\mathbf{y}}_k + \Delta\mathbf{y}_k$.
STEP 5a: Compute $\hat{\mathbf{x}}_{k+1} = \hat{\mathbf{x}}_k - \lambda_{k+1}\nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})$.
STEP 5b: Compute $\hat{\mathbf{y}}_{k+1} = \hat{\mathbf{y}}_k + \lambda_{k+1}\nabla_{\mathbf{y}} f(\mathbf{x}_{k+1}, \mathbf{y}_{k+1})$.
end for
Output: $(\bar{\mathbf{x}}_T, \bar{\mathbf{y}}_T) = \frac{1}{\sum_{k=1}^T \lambda_k}\left(\sum_{k=1}^T \lambda_k\mathbf{x}_k, \sum_{k=1}^T \lambda_k\mathbf{y}_k\right)$.

This together with Eq. (4.15) implies that $\|\mathbb{E}[X_j^2]\| \leq B_{\text{avg}}^2$. Putting these pieces with the aforementioned operator-Bernstein inequality yields

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

This completes the proof. □

Remark 4.7 *Compared to Lemma 4.4, the computation of sampling probability in Lemma 4.6 requires going through all data points and the total cost amounts to one full gradient evaluation. However, the computational savings stems from smaller sample size could dominate the extra cost of computing the sampling probability in practice [Xu et al., 2016]. In particular, the sample size from Lemma 4.6 could be smaller as $B_{\text{avg}} \leq B_{\max}$ which occurs if some B_i's are much larger than the others. In addition, the sample size is proportional to the log of the failure probability in Lemma 4.4 and 4.6, allowing the use of a very small failure probability without increasing the sample size significantly.*

Combining Algorithm 2 and these random sampling schemes gives the first class of sub-sampled Newton methods for solving the finite-sum min-max optimization problems in the form of Eq. (4.2) and (4.3). We summarize the detailed scheme in Algorithm 3 and prove the global convergence rate guarantee in the high-probability sense. Formally, we have

Theorem 4.8 *Suppose that Assumption 2.4 and 2.5 hold true and let $\epsilon \in (0,1)$, the iterates generated by Algorithm 3 are bounded by an universal constant. Moreover, there exists some $T > 0$ such that the output $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ = Subsampled-Newton-MinMax($(\mathbf{x}_0, \mathbf{y}_0), \rho, T, \delta$) satisfies that $\text{GAP}(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \leq \epsilon$ with probability $1-\delta$ and the total number of calls of the subproblem solvers is bounded by $O(\epsilon^{-\frac{2}{3}})$.*

Proof of Theorem 4.8. Since Algorithm 3 is a straightforward combination of Algorithm 2 and the random sampling schemes in Eq. (4.12) with $\kappa_H = B_{\max}$ and $0 < \tau_k \le \min\{\tau_0, \frac{\rho(1-\kappa_m)}{4(B_{\max}+6\rho)}\|\nabla f(\hat{\mathbf{z}}_k)\|\}$, we conclude the desired results from Theorem 4.1 if the following statement holds true:

$$\text{Prob}\left(\|H(\hat{\mathbf{z}}_k) - \nabla^2 f(\hat{\mathbf{z}}_k)\| \le \tau_k \text{ for all } 0 \le k \le T-1\right) \ge 1 - \delta. \tag{4.17}$$

To guarantee an overall accumulative success probability of $1-\delta$ across all the T iterations, it suffices to set the per-iteration failure probability as $1 - \sqrt[T]{1-\delta}$ as we have done in **Step 2** of Algorithm 3. In addition, 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 4.4 and 4.6, the extra cost will not be dominating. Thus, when Algorithm 3 terminates, all Hessian approximations have satisfied Eq. (4.17). This completes the proof.

5 Numerical Experiment

In this section, we study the numerical performance of our algorithms for min-max problems with both synthetic and real datasets. The baseline approaches include extragradient (EG) method, optimistic gradient descent ascent (OGDA) method, stochastic EG and OGDA methods and second-order variants of EG and OGDA methods.[3] All of these algorithms were implemented using MATLAB R2021b on a MacBook Pro with an Intel Core i9 2.4GHz and 16GB memory.

5.1 Cubic regularized bilinear min-max problem

Following the setup of Jiang and Mokhtari [2022], we consider the problem in the following form:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \max_{\mathbf{y}\in\mathbb{R}^n} \; f(\mathbf{x}, \mathbf{y}) = \tfrac{\rho}{6}\|\mathbf{x}\|^3 + \mathbf{y}^\top(A\mathbf{x} - \mathbf{b}). \tag{5.1}$$

where $\rho > 0$, the entries of $\mathbf{b} \in \mathbb{R}^n$ are generated independently from $[-1, 1]$ and $A \in \mathbb{R}^{n\times n}$ is given by

$$A = \begin{bmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1 \\ & & & & 1 \end{bmatrix}.$$

This min-max optimization problem is obviously convex-concave and the function f is ρ-Hessian Lipschitz. This problem admits a global saddle point given by $\mathbf{x}^\star = A^{-1}\mathbf{b}$ and $\mathbf{y}^\star = -\frac{\rho}{2}\|\mathbf{x}^\star\|(A^\top)^{-1}\mathbf{x}^\star$. Thus, we use the gap function $\text{GAP}(\mathbf{x}, \mathbf{y})$ defined in Section 2 as the evaluation metric. In our experiment, the parameters are chosen as $\rho = \frac{1}{20n}$ and $n \in \{50, 100, 200\}$. Since the exact Hessian of f is available and the subproblem can be computed up to high accuracy (using nonlinear equation solvers in MATLAB), we apply Algorithm 1. For the baseline approaches, we include the EG and OGDA methods and their second-order variants [Bullins and Lai, 2022, Jiang and Mokhtari, 2022] which both require the exact Hessian at each iteration. We implement second-order EG using the pseudocode of Bullins and Lai [2022, Algorithm 5.2 and 5.3] with fine-tuning parameters and second-order OGDA using the code from the author of Jiang and Mokhtari [2022] with $(\alpha, \beta) = (0.5, 0.8)$. For a fair comparison, we solve all the subproblems in second-order methods using nonlinear system solvers in MATLAB (i.e., fsolve).

[3] We exclude the algorithms in Huang et al. [2022] since their theoretical guarantee is proved under additional conditions.

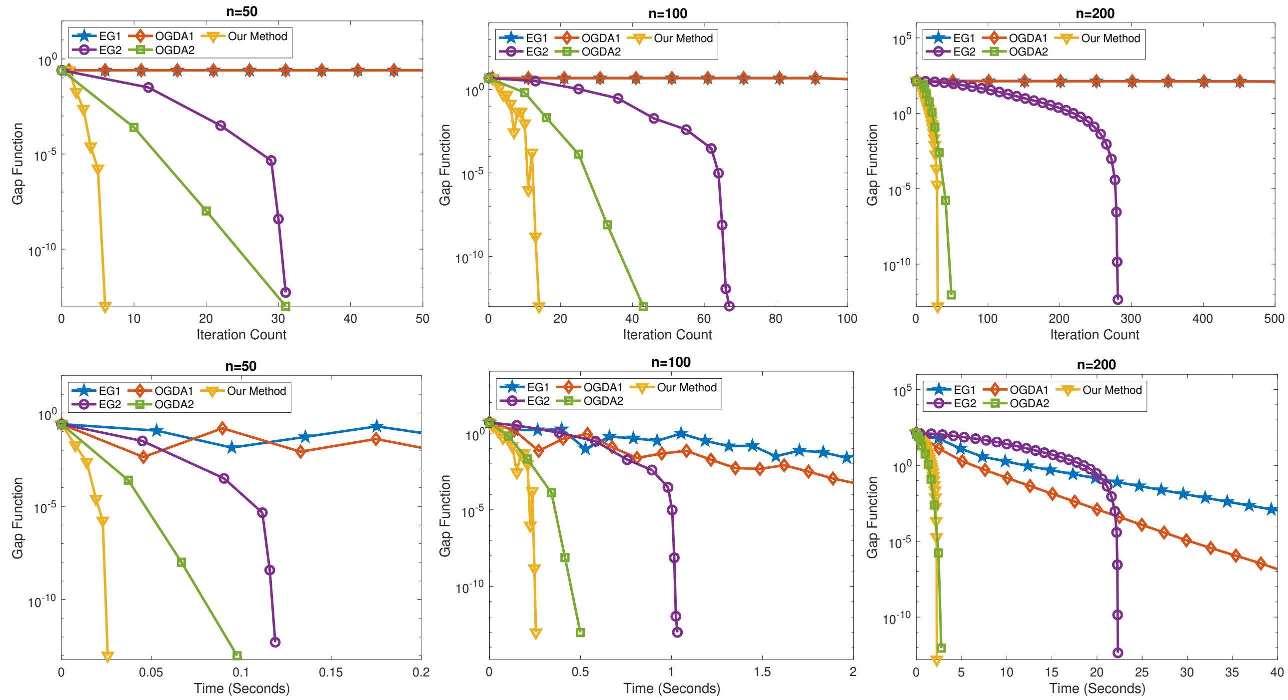

Figure 1: Performance of all the algorithms for $n \in \{50, 100, 200\}$ when $\rho = \frac{1}{20n}$ is set. The numerical results are presented in terms of iteration count (Top) and computational time (Bottom).

In Figure 1, we compare our algorithm with four other baseline algorithms in terms of gap function. Our results evidence that the second-order method can be superior to the first-order method in terms of solution accuracy: the first-order method barely makes any progress when the second-order method converges successfully. Moreover, our algorithm outperforms the second-order methods [Bullins and Lai, 2022, Jiang and Mokhtari, 2022] in terms of both iteration numbers and computational time thanks to its simple scheme without line search. However, this does not eliminate the potential advantages of using line search in min-max optimization. In fact, we find that the hybrid line search scheme suggested by Jiang and Mokhtari [2022] is quite powerful in practice and their algorithms with aggressive choices of (α, β) can sometimes outperform our algorithm. However, such choices make their algorithm unstable. As such, we set $(\alpha, \beta) = (0.5, 0.8)$ which is a bit conservative yet more robust. Finally, we believe that it is promising to investigate the line search scheme further in Jiang and Mokhtari [2022] and see if modifications can speed up second-order min-max optimization in a universal manner.

5.2 AUC maximization problem

The problem of maximizing area under the receiver operating characteristic curve [Hanley and McNeil, 1982] is a learning paradigm that learns a classifier for imbalanced data. It has a long history in machine learning [Graepel and Obermayer, 2000] and has motivated many studies ranging from formulations to algorithms and theory [Yang and Ying, 2022]. The goal is to find a classifier $\theta \in \mathbb{R}^n$ that maximizes the AUC score on a set of samples $\{(\mathbf{a}_i, b_i)\}_{i=1}^N$, where $\mathbf{a}_i \in \mathbb{R}^n$ and $b_i \in \{-1, +1\}$.

Table 1: Statistics of datasets for AUC maximization.

Name	Description	N	n	Scaled Interval
a9a	UCI adult	48842	123	$[0, 1]$
covetype	forest covetype	581012	54	$[0, 1]$
w8a	-	64700	300	$[0, 1]$

We consider the min-max formulation for AUC maximization [Ying et al., 2016, Shen et al., 2018]:

$$\min_{\mathbf{x}=(\theta,u,v)\in\mathbb{R}^{n+2}} \max_{y\in\mathbb{R}} \ \tfrac{1-\hat{p}}{N}\left\{\sum_{i=1}^N (\theta^\top \mathbf{a}_i - u)^2 \mathbb{I}_{[b_i=1]}\right\} + \tfrac{\hat{p}}{N}\left\{\sum_{i=1}^N (\theta^\top \mathbf{a}_i - v)^2 \mathbb{I}_{[b_i=-1]}\right\} \tag{5.2}$$
$$+\tfrac{2(1+y)}{N}\left\{\sum_{i=1}^N \theta^\top \mathbf{a}_i(\hat{p}\mathbb{I}_{[b_i=-1]} - (1-\hat{p})\mathbb{I}_{[b_i=1]})\right\} + \tfrac{\rho}{6}\|\mathbf{x}\|^3 - \hat{p}(1-\hat{p})y^2,$$

where $\lambda > 0$ is a scalar, $\mathbb{I}_{[\cdot]}$ is an indicator function and $\hat{p} = \frac{\#\{i:b_i=1\}}{N}$ be the proportion of samples with positive label. It is clear that the min-max optimization problem in Eq. (5.2) is convex-concave and has the finite-sum structure in the form of Eq. (4.2) with the function $f_i(\mathbf{x}, y)$ given by

$$f_i(\mathbf{x}, y) = (1-\hat{p})(\theta^\top \mathbf{a}_i - u)^2 \mathbb{I}_{[b_i=1]} + \hat{p}(\theta^\top \mathbf{a}_i - v)^2 \mathbb{I}_{[b_i=-1]}$$
$$+2(1+y)\theta^\top \mathbf{a}_i(\hat{p}\mathbb{I}_{[b_i=-1]} - (1-\hat{p})\mathbb{I}_{[b_i=1]}) + \tfrac{\rho}{6}\|\mathbf{x}\|^3 - \hat{p}(1-\hat{p})y^2.$$

This min-max optimization problem is in the cubic form and admits a global saddle point. Thus, we also use the gap function $\textsc{gap}(\mathbf{x}, \mathbf{y})$ defined in Section 2 as the evaluation metric. In our experiment, the parameter is chosen as $\rho = \frac{1}{N}$ and we use 3 LIBSVM datasets[4] for AUC maximization (see Table 1). Since the problem in Eq. (5.2) has a finite-sum structure, we apply Algorithm 3 with uniform sampling. For the baseline approaches, we include stochastic EG and OGDA methods [Juditsky et al., 2011, Hsieh et al., 2019, Mertikopoulos et al., 2019, Kotsalis et al., 2022]. The stepsizes for SEG and SOGDA are in the form of $\frac{c}{\sqrt{k+1}}$ where $c > 0$ is tuned using grid search and k is the iteration count. For our algorithm, we choose $\delta = 0.01$, $\kappa_m = 0.1$ and $|\mathcal{S}_k| = \frac{160\log(100(d+3))}{\min\{\|\nabla f(\hat{\mathbf{z}}_k)\|^2, \|\nabla f(\mathbf{z}_k)\|^2\}}$[5]. For the subproblem solution, we apply the semi-smooth Newton method as described in Section 4.1.

In Figure 2, we compare our algorithm with two other baseline algorithms in terms of gap function. Our results show that the sub-sampled second-order method can be superior to the mini-batch stochastic first-order method in terms of solution accuracy: stochastic first-order method barely makes any progress when our sub-sampled second-order method converges successfully. Moreover, it is worth remarking that our algorithm exhibits the (super)-linear convergence as the iterate approaches the optimal solution. This intriguing property has been rigorously justified for the sub-sampled Newton method in the context of convex optimization [Roosta-Khorasani and Mahoney, 2019]. Can we extend their results to convex-concave min-max optimization and prove the local (super)-linear convergence for Algorithm 3 under certain regularity condition? This is an interesting open question.

[4] https://www.csie.ntu.edu.tw/∼cjlin/libsvm/

[5] We observe that $\|\nabla f(\hat{\mathbf{z}}_k)\|$ fluctuates dramatically in the first few iterations when the sample size is small. Only using $\|\nabla f(\hat{\mathbf{z}}_k)\|$ to set the sample size in our algorithm makes it stuck at low-accurate solutions. In contrast, $\|\nabla f(\mathbf{z}_k)\|$ decreases steadily regardless of relatively small sample size. As such, we use this strategy to set the sample size.

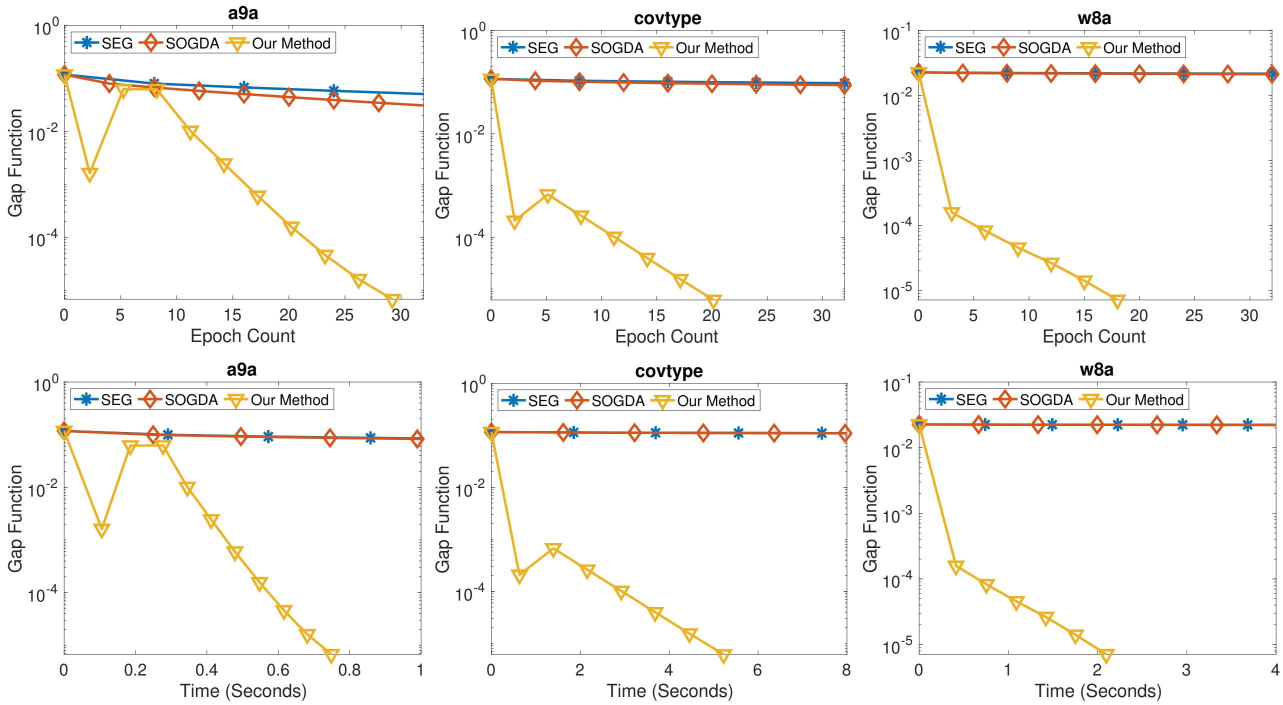

Figure 2: Performance of all the algorithms with 3 LIBSVM datasets when $\rho = \frac{1}{N}$ is set. The numerical results are presented in terms of epoch count (Top) and computational time (Bottom).

6 Conclusions

In this paper, we propose and analyze exact and inexact regularized Newton-type methods for finding a global saddle point of unconstrained and convex-concave min-max optimization problems with bounded and Lipschitz-continuous Hessian. In terms of theoretical guarantee, our methods are proved to achieve an optimal convergence guarantee of $O(\epsilon^{-2/3})$. Moreover, we show that our framework and convergence analysis on inexact algorithms lead to the first sub-sampled Newton method for solving the finite-sum min-max optimization problems with global convergence guarantee. Future research directions include the extension of our results to more general nonconvex-nonconcave min-max optimization problems and the customized implementation of our algorithms in real application problems.

Acknowledgments

This work was supported in part by the Mathematical Data Science program of the Office of Naval Research under grant number N00014-18-1-2764 and by the Vannevar Bush Faculty Fellowship program under grant number N00014-21-1-2941.

References

D. Adil, B. Bullins, A. Jambulapati, and S. Sachdeva. Line search-free methods for higher-order smooth monotone variational inequalities. *ArXiv Preprint: 2205.06167*, 2022. (Cited on pages 2, 4, and 7.)

F. Alizadeh and D. Goldfarb. Second-order cone programming. *Mathematical Programming*, 95(1):3–51, 2003. (Cited on page 14.)

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

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *ICML*, pages 214–223. PMLR, 2017. (Cited on page 2.)

A. Ben-Tal, L. EL Ghaoui, and A. Nemirovski. *Robust Optimization*, volume 28. Princeton University Press, 2009. (Cited on page 1.)

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. (Cited on page 2.)

D. A. Blackwell and M. A. Girshick. *Theory of Games and Statistical Decisions.* Courier Corporation, 1979. (Cited on page 1.)

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. (Cited on pages 2, 3, 7, 22, and 23.)

X. Chen, B. Jiang, T. Lin, and S. Zhang. Accelerating adaptive cubic regularization of Newton's method via random sampling. *Journal of Machine Learning Research*, 23(90):1–38, 2022. (Cited on page 12.)

G. B. Dantzig. *Linear Programming and Extensions.* Princeton University Press, 1963. (Cited on page 1.)

P. Drineas and M. W. Mahoney. Lectures on randomized numerical linear algebra. *The Mathematics of Data*, 25(1), 2018. (Cited on page 3.)

F. Facchinei and J-S. Pang. *Finite-dimensional Variational Inequalities and Complementarity Problems.* Springer Science & Business Media, 2007. (Cited on pages 1 and 2.)

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NeurIPS*, pages 2672–2680, 2014. (Cited on page 1.)

T. Graepel and K. Obermayer. Large margin rank boundaries for ordinal regression. In *Advances in Large Margin Classifiers*, pages 115–132. The MIT Press, 2000. (Cited on page 23.)

D. Gross and V. Nesme. Note on sampling without replacing from a finite collection of matrices. *ArXiv Preprint: 1001.2738*, 2010. (Cited on page 20.)

J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982. (Cited on page 23.)

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. (Cited on pages 2 and 24.)

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. (Cited on pages 2, 3, and 22.)

R. Jiang and A. Mokhtari. Generalized optimistic methods for convex-concave saddle point problems. *ArXiv Preprint: 2202.09674*, 2022. (Cited on pages 2, 3, 7, 22, and 23.)

A. Juditsky, A. Nemirovski, and C. Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011. (Cited on page 24.)

C. Kanzow, I. Ferenczi, and M. Fukushima. On the local convergence of semismooth Newton methods for linear and nonlinear second-order cone programs without strict complementarity. *SIAM Journal on Optimization*, 20(1):297–320, 2009. (Cited on pages 14 and 15.)

G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976. (Cited on page 2.)

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. (Cited on pages 2 and 24.)

X. Li, D. Sun, and K-C. Toh. A highly efficient semismooth Newton augmented Lagrangian method for solving Lasso problems. *SIAM Journal on Optimization*, 28(1):433–458, 2018a. (Cited on page 13.)

Y. Li, Z. Wen, C. Yang, and Y-X. Yuan. A semismooth Newton method for semidefinite programs and its applications in electronic structure calculations. *SIAM Journal on Scientific Computing*, 40(6): A4131–A4157, 2018b. (Cited on page 13.)

T. Lin and M. I. Jordan. On monotone inclusions, acceleration and closed-loop control. *ArXiv Preprint: 2111.08093*, 2021. (Cited on pages 2 and 3.)

T. Lin and M. I. Jordan. Perseus: A simple high-order regularization method for variational inequalities. *ArXiv Preprint: 2205.03202*, 2022. (Cited on pages 2, 3, 4, 7, and 8.)

M. S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret. Applications of second-order cone programming. *Linear Algebra and Its Applications*, 284(1-3):193–228, 1998. (Cited on page 14.)

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. (Cited on page 24.)

A. Milzarek and M. Ulbrich. A semismooth Newton method with multidimensional filter globalization for l_1-optimization. *SIAM Journal on Optimization*, 24(1):298–333, 2014. (Cited on page 13.)

A. Milzarek, X. Xiao, S. Cen, Z. Wen, and M. Ulbrich. A stochastic semismooth Newton method for nonsmooth nonconvex optimization. *SIAM Journal on Optimization*, 29(4):2916–2948, 2019. (Cited on page 13.)

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, 2020. (Cited on pages 2 and 5.)

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. (Cited on pages 2, 3, and 7.)

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. (Cited on pages 2 and 5.)

A. S. Nemirovski and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983. (Cited on page 2.)

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. (Cited on page 7.)

Y. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2):319–344, 2007. (Cited on pages 2, 6, and 7.)

Y. Nesterov. Accelerating the cubic regularization of newton's method on convex problems. *Mathematical Programming*, 112(1):159–181, 2008. (Cited on page 2.)

Y. Nesterov and B. T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006. (Cited on page 2.)

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. (Cited on page 2.)

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. (Cited on page 2.)

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. (Cited on page 2.)

L. Qi and J. Sun. A nonsmooth version of Newton's method. *Mathematical Programming*, 58(1):353–367, 1993. (Cited on page 13.)

F. Roosta-Khorasani and M. W. Mahoney. Sub-sampled Newton methods. *Mathematical Programming*, 174(1):293–326, 2019. (Cited on pages 2, 13, and 24.)

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. (Cited on page 13.)

Y. Saad and M. H. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 7(3):856–869, 1986. (Cited on page 15.)

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014. (Cited on page 13.)

J. Shamma. *Cooperative Control of Distributed Multi-agent Systems*. John Wiley & Sons, 2008. (Cited on page 2.)

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. (Cited on page 24.)

A. Sinha, H. Namkoong, and J. Duchi. Certifiable distributional robustness with principled adversarial training. In *ICLR*, 2018. (Cited on page 2.)

M. V. Solodov and B. F. Svaiter. A globally convergent inexact Newton method for systems of monotone equations. In *Reformulation: Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods*, pages 355–369. Springer, 1998. (Cited on page 13.)

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. (Cited on page 2.)

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. (Cited on page 2.)

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

M. Ulbrich. *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*. SIAM, 2011. (Cited on page 13.)

J. Von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1953. (Cited on page 1.)

X. Xiao, Y. Li, Z. Wen, and L. Zhang. A regularized semi-smooth Newton method with projection steps for composite convex programs. *Journal of Scientific Computing*, 76(1):364–389, 2018. (Cited on page 15.)

P. Xu, J. Yang, F. Roosta-Khorasani, C. Ré, and M. W. Mahoney. Sub-sampled Newton methods with non-uniform sampling. In *NeurIPS*, pages 3008–3016, 2016. (Cited on page 21.)

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. (Cited on pages 3, 12, 13, 16, 19, and 20.)

T. Yang and Y. Ying. AUC maximization in the era of big data and AI: A survey. *ACM Computing Surveys (CSUR)*, 2022. (Cited on page 23.)

Y. Ying, L. Wen, and S. Lyu. Stochastic online AUC maximization. In *NIPS*, pages 451–459, 2016. (Cited on page 24.)

X-Y. Zhao, D. Sun, and K-C. Toh. A Newton-CG augmented Lagrangian method for semidefinite programming. *SIAM Journal on Optimization*, 20(4):1737–1765, 2010. (Cited on page 13.)

G. Zhou and K-C. Toh. Superlinear convergence of a Newton-type algorithm for monotone equations. *Journal of Optimization Theory and Applications*, 125(1):205–221, 2005. (Cited on page 13.)