The Complexity of Nonconvex-Strongly-Concave Minimax Optimization

Siqi Zhang[1] **Junchi Yang**[1] **Cristóbal Guzmán**[2] **Negar Kiyavash**[3] **Niao He**[4]

[1]University of Illinois at Urbana-Champaign (UIUC)
[2]University of Twente & Pontificia Universidad Católica de Chile
[3]École polytechnique fédérale de Lausanne (EPFL)
[4]ETH Zürich

Abstract

This paper studies the complexity for finding approximate stationary points of *nonconvex-strongly-concave (NC-SC)* smooth minimax problems, in both *general* and *averaged smooth finite-sum* settings. We establish nontrivial lower complexity bounds of $\Omega(\sqrt{\kappa}\Delta L\epsilon^{-2})$ and $\Omega(n+\sqrt{n\kappa}\Delta L\epsilon^{-2})$ for the two settings, respectively, where κ is the condition number, L is the smoothness constant, and Δ is the initial gap. Our result reveals substantial gaps between these limits and best-known upper bounds in the literature. To close these gaps, we introduce a *generic acceleration scheme* that deploys existing gradient-based methods to solve a sequence of crafted strongly-convex-strongly-concave subproblems. In the general setting, the complexity of our proposed algorithm nearly matches the lower bound; in particular, it removes an additional poly-logarithmic dependence on accuracy present in previous works. In the averaged smooth finite-sum setting, our proposed algorithm improves over previous algorithms by providing a nearly-tight dependence on the condition number.

1 INTRODUCTION

In this paper, we consider general minimax problems of the form $(n, d_1, d_2 \in \mathbb{N}^+)$:

$$\min_{x\in\mathbb{R}^{d_1}} \max_{y\in\mathbb{R}^{d_2}} f(x,y), \tag{1}$$

as well as their finite-sum counterpart:

$$\min_{x\in\mathbb{R}^{d_1}} \max_{y\in\mathbb{R}^{d_2}} f(x,y) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_i(x,y), \tag{2}$$

where f, f_i are continuously differentiable and f is L-Lipschitz smooth jointly in x and y. We focus on the setting when f is μ-strongly concave in y and perhaps nonconvex in x, i.e., f is *nonconvex-strongly-concave (NC-SC)*. Such problems arise ubiquitously in machine learning, e.g., GANs with regularization [Lei et al., 2020], Wasserstein robust models [Sinha et al., 2018], robust learning over multiple domains [Qian et al., 2019], and off-policy reinforcement learning [Dai et al., 2017, 2018, Huang and Jiang, 2020]. Since the problem is nonconvex in general, a natural goal is to find an approximate stationary point $\bar{x}$, such that $\|\nabla\Phi(\bar{x})\| \le \epsilon$, for a given accuracy ϵ, where $\Phi(x) \triangleq \max_y f(x,y)$ is the primal function. This goal is meaningful for the aforementioned applications, e.g., in adversarial models the primal function quantifies the worst-case loss for the learner, with respect to adversary's actions.

There exists a number of algorithms for solving NC-SC problems in the general setting, including GDmax [Nouiehed et al., 2019], GDA [Lin et al., 2020a], alternating GDA [Yang et al., 2020a, Boţ and Böhm, 2020, Xu et al., 2020], Minimax-PPA [Lin et al., 2020b]. Specifically, GDA and its alternating variant both achieve the complexity of $O(\kappa^2\Delta L\epsilon^{-2})$ [Lin et al., 2020a, Yang et al., 2020a], where $\kappa \triangleq \frac{L}{\mu}$ is the condition number and $\Delta \triangleq \Phi(x_0) - \inf_x \Phi(x)$ is the initial function gap. Recently, [Lin et al., 2020b] provided the best-known complexity of $O\big(\sqrt{\kappa}\Delta L\epsilon^{-2}\cdot\log^2(\frac{\kappa L}{\epsilon})\big)$ achieved by Minimax-PPA, which improves the dependence on the condition number but suffers from an extra poly-logarithmic factor in $\frac{1}{\epsilon}$.

In the finite-sum setting, several algorithms have been proposed recently, e.g., PGSMD [Rafique et al., 2018], SGDmax [Jin et al., 2020], Stochastic GDA [Lin et al., 2020a], SREDAs [Luo et al., 2020]. In particular, [Lin et al., 2020a] proved that Stochastic GDA attains the complexity of $O(\kappa^3\epsilon^{-4})$. [Luo et al., 2020] recently showed the state-of-the-art result achieved by SREDA: when $n \ge \kappa^2$, the complexity is $\tilde{O}\big(n\log\frac{\kappa}{\epsilon} + \sqrt{n}\kappa^2\Delta L\epsilon^{-2}\big)$, which is sharper than the batch Minimax-PPA algorithm; when $n \le \kappa^2$, the complexity is $O\big(\big(n\kappa+\kappa^2\big)\Delta L\epsilon^{-2}\big)$, which is sharper than Stochastic GDA.

Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021), PMLR 161:482–492.

Table 1: Upper and lower complexity bounds for finding an approximate stationary point. Here $\tilde{O}(\cdot)$ hides poly-logarithmic factor in L, μ and κ. L: Lipschitz smoothness parameter; μ: strong concavity parameter, κ: condition number $\frac{L}{\mu}$; Δ: initial gap of the primal function.

Setting	Our Lower Bound	Our Upper Bound	Previous Upper Bound
NC-SC, general	$\Omega\big(\sqrt{\kappa}\Delta L\epsilon^{-2}\big)$ Theorem 3.1	$\tilde{O}\big(\sqrt{\kappa}\Delta L\epsilon^{-2}\big)$ Section 4.2	$O(\kappa^2\Delta L\epsilon^{-2})$ [Lin et al., 2020a] $\tilde{O}\big(\sqrt{\kappa}\Delta L\epsilon^{-2}\log^2\frac{1}{\epsilon}\big)$ [Lin et al., 2020b]
NC-SC, FS, AS[1]	$\Omega\big(n+\sqrt{n\kappa}\Delta L\epsilon^{-2}\big)$ Theorem 3.2	$\tilde{O}\Big(\Big(n+n^{\frac{3}{4}}\sqrt{\kappa}\Big)\Delta L\epsilon^{-2}\Big)$ Section 4.2	$\begin{cases}\tilde{O}(n+\sqrt{n}\kappa^2\Delta L\epsilon^{-2}) & n\geq\kappa^2\\ O\big((n\kappa+\kappa^2)\Delta L\epsilon^{-2}\big) & n\leq\kappa^2\end{cases}$ [Luo et al., 2020]

[1] FS: finite-sum, AS: averaged smooth; see Section 2 for definitions.

Despite this active line of research, whether these state-of-the-art complexity bounds can be further improved remains elusive. As a special case by restricting the domain of y to a singleton, lower bounds for nonconvex smooth minimization, e.g., [Carmon et al., 2019a,b, Fang et al., 2018, Zhou and Gu, 2019, Arjevani et al., 2019], hardly capture the dependence on the condition number κ, which plays a crucial role in the complexity for general NC-SC smooth minimax problems. In many of the aforementioned machine learning applications, the condition number is inversely proportional the regularization parameter, and can be very large in practice. For example, in statistical learning, where n represents the sample size, the optimal regularization parameter (i.e. with optimal empirical/generalization trade-off) leads to $\kappa = \Omega(\sqrt{n})$ [Shalev-Shwartz and Ben-David, 2014].

This motivates the following fundamental questions: *What is the complexity limit for NC-SC problems in the general and finite-sum settings? Can we design new algorithms to meet the performance limits and attain optimal dependence on the condition number?*

1.1 CONTRIBUTIONS

Our contributions, summarized in Table 1, are as follows:

- We establish nontrivial lower complexity bounds for finding an approximate stationary point of nonconvex-strongly-concave (NC-SC) minimax problems. In the general setting, we prove an $\Omega\big(\sqrt{\kappa}\Delta L\epsilon^{-2}\big)$ lower complexity bound which applies to arbitrary deterministic linear-span algorithms interacting with the classical first-order oracle. In the finite-sum setting, we prove an $\Omega\big(n+\sqrt{n\kappa}\Delta L\epsilon^{-2}\big)$ lower complexity bound (when $\kappa = \Omega(n)$)[1] for the class of *averaged smooth functions* and arbitrary linear-span algorithms interacting with a (randomized) incremental first-order oracle (precise definitions in Sections 2 and 3).

  Our lower bounds build upon two main ideas: first, we start from an NC-SC function whose primal function mimics the lower bound construction in smooth nonconvex minimization [Carmon et al., 2019a]. Crucially, the smoothness parameter of this primal function is boosted by an $\Omega(\kappa)$ factor, which strengthens the lower bound. Second, the function has an alternating zero-chain structure, as utilized in lower bounds for convex-concave settings [Ouyang and Xu, 2019]. The combination of these features leads to a hard instance for our problem.

- To bridge the gap between the lower bounds and existing upper bounds in both settings, we introduce a generic Catalyst acceleration framework for NC-SC minimax problems, inspired by [Lin et al., 2018a, Yang et al., 2020b], which applies existing gradient-based methods to solving a sequence of crafted strongly-convex-strongly-concave (SC-SC) minimax subproblems. When combined with the extragradient (EG) method [Tseng, 1995], the resulting algorithm achieves an $\tilde{O}\big(\sqrt{\kappa}\Delta L\epsilon^{-2}\big)$ complexity in terms of gradient evaluations, which tightly matches the lower bound in the general setting (up to logarithmic terms in constants) and shaves off the extra poly-logarithmic term in $\frac{1}{\epsilon}$ required by the state-of-the-art [Lin et al., 2020b]. When combined with stochastic variance-reduced method, the resulting algorithm achieves an overall $\tilde{O}\big((n+n^{3/4}\sqrt{\kappa})\Delta L\epsilon^{-2}\big)$ complexity for averaged smooth finite-sum problems, which has nearly-tight dependence on the condition number and improves on the best-known upper bound when $n \leq \kappa^4$.

1.2 RELATED WORK

Lower bounds for minimax problems. Information-based complexity (IBC) theory [Traub et al., 1988], which derives the minimal number of oracle calls to attain an approximate solution with a desired accuracy, is often used in lower bound analysis of optimization algorithms. Unlike the case of minimization, e.g., [Nemirovski and Yudin, 1983, Carmon et al., 2019a,b, Arjevani et al., 2019], lower bounds for minimax optimization are far less understood; only a few

[1] A concurrent work by Han et al. [2021] provided a similar lower bound result for finite-sum NC-SC problems under probabilistic arguments based on geometric random variables.

recent works provided lower bounds for finding an approximate saddle point of (strongly)-convex-(strongly)-concave minimax problems [Ouyang and Xu, 2019, Zhang et al., 2019, Ibrahim et al., 2020, Xie et al., 2020, Yoon and Ryu, 2021]. Instead, this paper considers lower bounds for NC-SC problems of finding an stationary point, which requires different techniques for constructing zero-chain properties. Note that there exists another line of research on the purely stochastic setting, e.g., [Rafique et al., 2018, Luo et al., 2020]; constructing lower bounds in that setting is out of the scope of this paper.

Complexity of making gradient small. In nonconvex optimization, most lower complexity bounds we mentioned before measure the convergence in gradient norm, and corresponding upper bounds are able to match with them [Carmon et al., 2019a]. Although for the convex optimization objective value is commonly considered as convergence criterion, Nesterov [2012], Allen-Zhu [2018], Foster et al. [2019], Carmon et al. [2019b] provide algorithms with gradient norm convergence, which is easier to check and arguably more informative than the traditional optimality gap. Recently, Diakonikolas [2020], Diakonikolas and Wang [2021], Yoon and Ryu [2021] attain convergence results in terms of gradient norm for convex-concave minimax problems. In particular, Yoon and Ryu [2021] propose an algorithm combining Halpern iteration and extragradient to attain a complexity of $O(L/\epsilon)$, also with a matching lower bound.

Nonconvex minimax optimization. In NC-SC setting, as we mentioned, there has been several substantial works. Among them, Lin et al. [2020b] achieved the best dependency on condition number by combining proximal point algorithm with accelerated gradient descent. Luo et al. [2020] introduced a variance reduction algorithm, SREDA. In addition, nonconvex-concave minimax optimization, i.e., the function f is only concave in y, is extensively explored, e.g., [Zhang et al., 2020, Thekumparampil et al., 2019, Nouiehed et al., 2019, Yang et al., 2020b]. Recently, [Daskalakis et al., 2020] showed that for general smooth nonconvex-nonconcave objectives, finding locally optimal solutions is intractable. Therefore, another line of research is devoted to solving under additional structural properties, e.g., [Yang et al., 2020c,a, Diakonikolas et al., 2020, Lin et al., 2018b].

Catalyst acceleration. The catalyst framework was initially studied in [Lin et al., 2015] for convex minimization and extended to nonconvex minimization in [Paquette et al., 2018] to obtain accelerated algorithms. A similar idea to accelerate SVRG appeared in [Frostig et al., 2015]. These work are rooted on the proximal point algorithm (PPA) [Rockafellar, 1976] and inexact accelerated PPA [Güler, 1992]. Recently, [Yang et al., 2020b] generalized the idea and obtained state-of-the-art results for solving strongly-convex-concave and nonconvex-concave minimax problems. In contrast, this paper introduces a new catalyst acceleration scheme in the nonconvex-strongly-concave setting, which relies on different parameter settings and stopping criterion.

2 PRELIMINARIES

Notations Throughout the paper, we use $\mathbf{dom}\, F$ as the domain of a function F, $\nabla F = (\nabla_x F, \nabla_y F)$ as the full gradient, $\|\cdot\|$ as the ℓ_2-norm. We use 0 to represent zero vectors or scalars, e_i to represent unit vector with the i-th element being 1. For nonnegative functions $f(x)$ and $g(x)$, we say $f = O(g)$ if $f(x) \le cg(x)$ for some $c > 0$, and further write $f = \tilde{O}(g)$ to omit poly-logarithmic terms on constants L, μ and κ, while $f = \Omega(g)$ if $f(x) \ge cg(x)$ (see more in Appendix).

We introduce definitions and assumptions used throughout.

Definition 2.1 (Primal and Dual Functions) *For a function $f(x,y)$, we define $\Phi(x) \triangleq \max_y f(x,y)$ as the primal function, and $\Psi(y) \triangleq \min_x f(x,y)$ as the dual function. We also define the primal-dual gap at a point $(\bar{x},\bar{y})$ as $\mathrm{gap}_f(\bar{x},\bar{y}) \triangleq \max_{y\in\mathbb{R}^{d_2}} f(\bar{x},y) - \min_{x\in\mathbb{R}^{d_1}} f(x,\bar{y})$.*

Definition 2.2 (Lipschitz Smoothness) *We say a function $f(x,y)$ is L-Lipschitz smooth (L-S) jointly in x and y if it is differentiable and for any $(x_1,y_1),(x_2,y_2) \in \mathbb{R}^{d_1}\times\mathbb{R}^{d_2}$, $\|\nabla_x f(x_1,y_1) - \nabla_x f(x_2,y_2)\| \le L(\|x_1-x_2\|+\|y_1-y_2\|)$ and $\|\nabla_y f(x_1,y_1) - \nabla_y f(x_2,y_2)\| \le L(\|x_1-x_2\|+\|y_1-y_2\|)$, for some $L > 0$.*

Definition 2.3 (Average / Individual Smoothness) *We say $f(x,y) = \frac{1}{n}\sum_{i=1}^n f_i(x,y)$ or $\{f_i\}_{i=1}^n$ is L-averaged smooth (L-AS) if each f_i is differentiable, and for any $(x_1,y_1),(x_2,y_2) \in \mathbb{R}^{d_1}\times\mathbb{R}^{d_2}$, we have $\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x_1,y_1) - \nabla f_i(x_2,y_2)\|^2 \le L^2\left(\|x_1-x_2\|^2 + \|y_1-y_2\|^2\right)$. We say f or $\{f_i\}_{i=1}^n$ is L-individually smooth (L-IS) if each f_i is L-S.*

Average smoothness is a weaker condition than the common Lipschitz smoothness assumption of each component in finite-sum / stochastic minimization [Fang et al., 2018, Zhou and Gu, 2019]. Similarly in minimax problems, the following proposition summarizes the relationship among these different notions of smoothness.

Proposition 2.1 *Let $f(x,y) = \frac{1}{n}\sum_{i=1}^n f_i(x,y)$. Then we have: (a) If the function f is L-IS or L-AS, then it is L-S. (b) If f is L-IS, then it is $(2L)$-AS. (c) If f is L-AS, then $f(x,y)+\frac{\tau_x}{2}\|x-\tilde{x}\|^2-\frac{\tau_y}{2}\|y-\tilde{y}\|^2$ is $\sqrt{2}(L+\max\{\tau_x,\tau_y\})$-AS for any $\tilde{x}$ and $\tilde{y}$.*

Definition 2.4 (Strong Convexity) *A differentiable function $g : \mathbb{R}^{d_1} \to \mathbb{R}$ is convex if $g(x_2) \ge g(x_1) +$*

$\langle \nabla g(x_1), x_2 - x_1 \rangle$ for any $x_1, x_2 \in \mathbb{R}^{d_1}$. Given $\mu \geq 0$, we say f is μ-strongly convex if $g(x) - \frac{\mu}{2}\|x\|^2$ is convex, and it is μ-strongly concave if $-g$ is μ-strongly convex.

Assumption 2.1 (Main Settings) *We assume that $f(x, y)$ in* (1) *is a nonconvex-strongly-concave (NC-SC) function such that f is L-S, and $f(x, \cdot)$ is μ-strongly concave for any fixed $x \in \mathbb{R}^{d_1}$; for the finite-sum case, we further assume that $\{f_i\}_{i=1}^n$ is L-AS. We assume that the initial primal suboptimality is bounded: $\Phi(x_0) - \inf_x \Phi(x) \leq \Delta$.*

Under Assumption 2.1, the primal function $\Phi(\cdot)$ is differentiable and $2\kappa L$-Lipschitz smooth [Lin et al., 2020b, Lemma 23] where $\kappa \triangleq \frac{L}{\mu}$. Throughout this paper, we use the stationarity of $\Phi(\cdot)$ as the convergence criterion.

Definition 2.5 (Convergence Criterion) *For a differentiable function Φ, a point $\bar{x} \in \mathbf{dom}\,\Phi$ is called an ϵ-stationary point of Φ if $\|\nabla \Phi(\bar{x})\| \leq \epsilon$.*

Another commonly used criterion is the stationarity of f, i.e., $\|\nabla_x f(\bar{x}, \bar{y})\| \leq \epsilon, \|\nabla_y f(\bar{x}, \bar{y})\| \leq \epsilon$. This is a weaker convergence criterion. We refer readers to [Lin et al., 2020a, Section 4.3] for the comparison of these two criteria.

3 LOWER BOUNDS FOR NC-SC MINIMAX PROBLEMS

In this section, we establish lower complexity bounds (LB) for finding approximate stationary points of NC-SC minimax problems, in both general and finite-sum settings. We first present the basic components of the oracle complexity framework [Nemirovski and Yudin, 1983] and then proceed to the details for each case. For simplicity, in this section only, we denote x_d as the d-th coordinate of x and x^t as the variable x in the t-th iteration.

3.1 FRAMEWORK AND SETUP

We study the lower bound of finding primal stationary point under the well-known oracle complexity framework [Nemirovski and Yudin, 1983], here we first present the basics of the framework.

Function class We consider the *nonconvex-strongly-concave (NC-SC)* function class, as defined in Assumption 2.1, with parameters $L, \mu, \Delta > 0$, denoted by $\mathcal{F}_{\mathrm{NCSC}}^{L,\mu,\Delta}$.

Oracle class We consider different oracles for the general and finite-sum settings. Define $z \triangleq (x, y)$.

- For the general setting, we consider the *first-order oracle (FO)*, denoted as $\mathbb{O}_{\mathrm{FO}}(f, \cdot)$, that for each query on point z, it returns the gradient $\mathbb{O}_{\mathrm{FO}}(f, z) \triangleq (\nabla_x f(x, y), \nabla_y f(x, y))$.
- For the finite-sum setting, *incremental first-order oracle (IFO)* is often used in lower bound analysis [Agarwal and Bottou, 2015]. This oracle for a function $f(x, y) = \frac{1}{n}\sum_{i=1}^n f_i(x, y)$, is such that for each query on point z and given $i \in [n]$, it returns the gradient of the i-th component, i.e., $\mathbb{O}_{\mathrm{IFO}}(f, z, i) \triangleq (\nabla_x f_i(x, y), \nabla_y f_i(x, y)),$. Here, we consider *averaged smooth IFO* and *individually smooth IFO*, denoted as $\mathbb{O}_{\mathrm{IFO}}^{L,\mathrm{AS}}(f)$ and $\mathbb{O}_{\mathrm{IFO}}^{L,\mathrm{IS}}(f)$, where $\{f_i\}_{i=1}^n$ is L-AS or L-IS, respectively.

Algorithm class In this work, we consider the class of *linear-span algorithms* interacting with oracle $\mathbb{O}$, denoted as $\mathcal{A}(\mathbb{O})$. These algorithms satisfy the following property: if we let $(z^t)_t$ be the sequence of queries by the algorithm, where $z^t = (x^t, y^t)$; then for all t, we have

$$z^{t+1} \in \mathrm{Span}\big\{z^0, \cdots, z^t; \mathbb{O}\big(f, z^0\big), \cdots, \mathbb{O}\big(f, z^t\big)\big\}. \quad (3)$$

For the finite-sum case, the above protocol fits with many existing deterministic and randomized linear-span algorithms. We distinguish the general and finite-sum setting by specifying the used oracle, which is $\mathbb{O}_{\mathrm{FO}}$ or $\mathbb{O}_{\mathrm{IFO}}$, respectively. Most existing first-order algorithms, including simultaneous and alternating update algorithms, can be formulated as linear-span algorithms. It is worth pointing out that typically the linear span assumption is used without loss of generality, since there is a standard reduction from deterministic linear-span algorithms to arbitrary oracle based deterministic algorithms [Nemirovsky, 1991, 1992, Ouyang and Xu, 2019]. We defer this extension for future work.

Complexity measures The efficiency of algorithms is quantified by the *oracle complexity* [Nemirovski and Yudin, 1983] of finding an ϵ-stationary point of the primal function: for an algorithm $\mathtt{A} \in \mathcal{A}(\mathbb{O})$ interacting with a FO oracle $\mathbb{O}$, an instance $f \in \mathcal{F}$, we define

$$T_\epsilon(f, \mathtt{A}) \triangleq \inf\big\{T \in \mathbb{N} \big| \|\nabla \Phi\big(x^T\big)\| \leq \epsilon\big\} \quad (4)$$

as the minimum number of oracle calls $\mathtt{A}$ makes to reach stationarity convergence. For the general case, we define the *worst-case complexity*

$$\mathrm{Compl}_\epsilon(\mathcal{F}, \mathcal{A}, \mathbb{O}) \triangleq \sup_{f \in \mathcal{F}} \inf_{\mathtt{A} \in \mathcal{A}(\mathbb{O})} T_\epsilon(f, \mathtt{A}). \quad (5)$$

For finite-sum cases, we lower bound the randomized complexity by the *distributional complexity* [Braun et al., 2017]:

$$\mathrm{Compl}_\epsilon(\mathcal{F}, \mathcal{A}, \mathbb{O}) \triangleq \sup_{f \in \mathcal{F}} \inf_{\mathtt{A} \in \mathcal{A}(\mathbb{O})} \mathbb{E}\, T_\epsilon(f, \mathtt{A}). \quad (6)$$

Following the motivation of analysis discussed in Section 1.1, we will use the zero-chain argument for the analysis. First we define the notion of (first-order) zero-chain [Carmon et al., 2019b] and activation as follows.

Definition 3.1 (Zero Chain, Activation) *A function $f : \mathbb{R}^d \to \mathbb{R}$ is a first-order zero-chain if for any $x \in \mathbb{R}^d$,*

$$\operatorname{supp}\{x\} \subseteq \{1, \cdots, i-1\} \Rightarrow \operatorname{supp}\{\nabla f(x)\} \subseteq \{1, \cdots, i\},$$

where $\operatorname{supp}\{x\} \triangleq \{i \in [d] \mid x_i \neq 0\}$ *and* $[d] = \{1, \cdots, d\}$. *For an algorithm initialized at $0 \in \mathbb{R}^d$, with iterates $\{x^t\}_t$, we say coordinate i is activated at x^t, if $x_i^t \neq 0$ and $x_i^s = 0$, for any $s < t$.*

3.2 GENERAL NC-SC PROBLEMS

First we consider the *general NC-SC (Gen-NC-SC)* minimax optimization problems. Following the above framework, we choose function class $\mathcal{F}_{\text{NCSC}}^{L,\mu,\Delta}$, oracle $\mathbb{O}_{\text{FO}}$, linear-span algorithms $\mathcal{A}$, and we analyze the complexity defined in (5).

Hard instance construction Inspired by the hard instances constructed in [Ouyang and Xu, 2019, Carmon et al., 2019b], we introduce the following function $F_d : \mathbb{R}^{d+1} \times \mathbb{R}^{d+2} \to \mathbb{R}$ ($d \in \mathbb{N}^+$) and

$$\begin{aligned} F_d(x, y; \lambda, \alpha) \triangleq{}& \lambda_1 \langle B_d x, y\rangle - \lambda_2 \|y\|^2 - \frac{\lambda_1^2 \sqrt{\alpha}}{2\lambda_2} \langle e_1, x\rangle \\ &+ \frac{\lambda_1^2 \alpha}{2\lambda_2} \sum_{i=1}^{d} \Gamma(x_i) - \frac{\lambda_1^2 \alpha}{4\lambda_2} x_{d+1}^2 + \frac{\lambda_1^2 \sqrt{\alpha}}{4\lambda_2}, \end{aligned} \tag{7}$$

where $\lambda = (\lambda_1, \lambda_2) \in \mathbb{R}^2$ is the parameter vector, $e_1 \in \mathbb{R}^{d+1}$ is the unit vector with the only non-zero element in the first dimension, $\Gamma : \mathbb{R} \to \mathbb{R}$ and $B_d \in \mathbb{R}^{(d+2)\times(d+1)}$ are

$$B_d = \begin{bmatrix} & & & & 1 \\ & & & 1 & -1 \\ & & \cdot^{\cdot^{\cdot}} & \cdot^{\cdot^{\cdot}} & \\ & 1 & -1 & & \\ 1 & -1 & & & \\ \sqrt[4]{\alpha} & & & & \end{bmatrix}, \Gamma(x) = 120 \int_1^x \frac{t^2(t-1)}{1+t^2} dt. \tag{8}$$

Matrix B_d essentially triggers the activation of variables at each iteration, and function Γ introduces nonconvexity in x to the instance. By the first-order optimality condition of $F_d(x, \cdot; \lambda, \alpha)$, we can compute its primal function, Φ_d:

$$\begin{aligned} &\Phi_d(x; \lambda, \alpha) \triangleq \max_{y \in \mathbb{R}^{d+1}} F_d(x, y; \lambda, \alpha) \\ &= \frac{\lambda_1^2}{2\lambda_2} \left(\frac{1}{2} x^\top A_d x - \sqrt{\alpha} x_1 + \frac{\sqrt{\alpha}}{2} + \alpha \sum_{i=1}^{d} \Gamma(x_i) + \frac{1-\alpha}{2} x_{d+1}^2 \right), \end{aligned} \tag{9}$$

where $A_d \in \mathbb{R}^{(d+1)\times(d+1)}$ is

$$\begin{aligned} A_d &= \left(B_d^\top B_d - e_{d+1} e_{d+1}^\top\right) \\ &= \begin{bmatrix} 1+\sqrt{\alpha} & -1 & & & & \\ -1 & 2 & -1 & & & \\ & & & & & \\ & -1 & 2 & \ddots & & \\ & & \ddots & \ddots & -1 & \\ & & & & & \\ & & & \ddots & 2 & -1 \\ & & & & -1 & 1 \end{bmatrix}. \end{aligned} \tag{10}$$

The resulting primal function resembles the worst-case functions used in lower bound analysis of minimization problems [Nesterov, 2018, Carmon et al., 2019b].

Zero-Chain Construction First we summarize key properties of the instance and its zero-chain mechanism. We further denote $\hat{e}_i \in \mathbb{R}^{d+2}$ as the unit vector for the variable y and define ($k \geq 1$)

$$\begin{aligned} &\mathcal{X}_k \triangleq \operatorname{Span}\{e_1, e_2, \cdots, e_k\},\ \mathcal{X}_0 \triangleq \{0\}, \\ &\mathcal{Y}_k \triangleq \operatorname{Span}\{\hat{e}_{d+2}, \hat{e}_{d+1}, \cdots, \hat{e}_{d-k+2}\},\ \mathcal{Y}_0 \triangleq \{0\}, \end{aligned} \tag{11}$$

then we have the following properties for F_d.

Lemma 3.1 (Properties of F_d) *For any $d \in \mathbb{N}^+$ and $\alpha \in [0, 1]$, $F_d(x, y; \lambda, \alpha)$ in (7) satisfies:*

(i) The function $F_d(x, \cdot; \lambda, \alpha)$ is L_F-Lipschitz smooth where $L_F = \max\left\{\frac{200\lambda_1^2\alpha}{\lambda_2}, 2\lambda_1, 2\lambda_2\right\}$.

(ii) For each fixed $x \in \mathbb{R}^{d+1}$, $F_d(x, \cdot; \lambda, \alpha)$ is μ_F-strongly concave where $\mu_F = 2\lambda_2$.

(iii) The following properties hold:

a) $x = y = 0 \Rightarrow \nabla_x F_d \in \mathcal{X}_1,\ \nabla_y F_d = 0$.

b) $x \in \mathcal{X}_k,\ y \in \mathcal{Y}_k \Rightarrow \nabla_x F_d \in \mathcal{X}_{k+1},\ \nabla_y F_d \in \mathcal{Y}_k$.

c) $x \in \mathcal{X}_{k+1}, y \in \mathcal{Y}_k \Rightarrow \nabla_x F_d \in \mathcal{X}_{k+1}, \nabla_y F_d \in \mathcal{Y}_{k+1}$.

(iv) For $L \geq \mu > 0$, if $\lambda = \lambda^* = (\lambda_1^*, \lambda_2^*) = (\frac{L}{2}, \frac{\mu}{2})$ and $\alpha \leq \frac{\mu}{100L}$, F_d is L-Lipschitz smooth, and for any fixed $x \in \mathbb{R}^{d+1}$, $F_d(x, \cdot; \lambda, \alpha)$ is μ-strongly concave.

The proof of Lemma 3.1 is deferred to Appendix. The first two properties show that function F_d is Lipschitz smooth and NC-SC; the third property above suggests that, starting from $(x, y) = (0, 0)$, the activation process follows an "alternating zero-chain" form [Ouyang and Xu, 2019]. That is, for a linear-span algorithm, when $x \in \mathcal{X}_k,\ y \in \mathcal{Y}_k$, the next iterate will at most activate the $(k+1)$-th coordinate of x while keeping y fixed; similarly when $x \in \mathcal{X}_{k+1},\ y \in \mathcal{Y}_k$, the next iterate will at most activate the $(d-k+1)$-th element of y. We need the following properties of Φ_d for the lower bound argument.

Lemma 3.2 (Properties of Φ_d) *For any $\alpha \in [0, 1]$ and $x \in \mathbb{R}^{d+1}$, if $x_d = x_{d+1} = 0$, we have:*

(i) $\|\nabla \Phi_d(x; \lambda, \alpha)\| \geq \frac{\lambda_1^2}{8\lambda_2} \alpha^{3/4}$.

(ii) $\Phi_d(0; \lambda, \alpha) - \inf_{x \in \mathbb{R}^{d+1}} \Phi_d(x; \lambda, \alpha) \leq \frac{\lambda_1^2}{2\lambda_2} \left(\frac{\sqrt{\alpha}}{2} + 10\alpha d\right)$.

We defer the proof of Lemma 3.2 to Appendix. This lemma indicates that, starting from $(x, y) = (0, 0)$ with appropriate parameter settings, the primal function Φ_d will not approximate stationarity until the last two coordinates are activated. Now we are ready to present our final lower bound result for the general NC-SC case.

Theorem 3.1 (LB for Gen-NC-SC) *For the linear-span first-order algorithm class $\mathcal{A}$, parameters $L, \mu, \Delta > 0$, and accuracy ϵ satisfying $\epsilon^2 \le \min\left(\frac{\Delta L}{6400}, \frac{\Delta L\sqrt{\kappa}}{38400}\right)$, we have*

$$\mathrm{Compl}_\epsilon\left(\mathcal{F}_{\mathrm{NCSC}}^{L,\mu,\Delta}, \mathcal{A}, \mathbb{O}_{\mathrm{FO}}\right) = \Omega\left(\sqrt{\kappa}\Delta L\epsilon^{-2}\right). \quad (12)$$

The hard instance in the proof is established based on F_d in (7). We choose the scaled function $f(x,y) = \eta^2 F_d(\frac{x}{\eta}, \frac{y}{\eta}; \lambda^*, \alpha)$ as the final hard instance, which preserves the smoothness and strong convexity , while appropriate setting of η will help to fulfill the requirements on the initial gap and large gradient norm (before thorough activation) of the primal function. The detailed statement and proof of Theorem 3.1 are presented in Appendix.

Remark 3.1 (Tightness) *The best-known upper bounds for general NC-SC problems are $O(\Delta L\kappa^2\epsilon^{-2})$ [Lin et al., 2020a, Boţ and Böhm, 2020] and $\tilde{O}\left(\Delta\sqrt{\kappa}L\epsilon^{-2}\log^2\frac{1}{\epsilon}\right)$ [Lin et al., 2020b]. Therefore, our result exhibits significant gaps in terms of the dependence on ϵ and κ. In order to mitigate these gaps, we propose faster algorithms in Section 4. On the other hand, compared to the $\Omega(\Delta L\epsilon^{-2})$ lower bound for nonconvex smooth minimization [Carmon et al., 2019a], our result reveals an explicit dependence on κ.*

3.3 FINITE-SUM NC-SC PROBLEMS

The second case we consider is *finite-sum NC-SC (FS-NC-SC)* minimax problems, for the function class $\mathcal{F}_{\mathrm{NCSC}}^{L,\mu,\Delta}$, the linear-span algorithm class $\mathcal{A}$ and the averaged smooth IFO class $\mathbb{O}_{\mathrm{IFO}}^{L,\mathrm{AS}}$. The complexity is defined in (6).

Hard instance construction To derive the finite-sum hard instance, we modify F_d in (7) with orthogonal matrices defined as follows.

Definition 3.2 (Orthogonal Matrices) *For positive integers $a, b, n \in \mathbb{N}^+$, we define a matrix sequence $\{\mathbf{U}^{(i)}\}_{i=1}^n \in \mathbf{Orth}(a,b,n)$ if for each $i, j \in \{1, \cdots, n\}$ and $i \ne j$, $\mathbf{U}^{(i)}, \mathbf{U}^{(j)} \in \mathbb{R}^{a\times b}$ and $\mathbf{U}^{(i)}(\mathbf{U}^{(i)})^\top = \mathbf{I} \in \mathbb{R}^{a\times a}$ and $\mathbf{U}^{(i)}(\mathbf{U}^{(j)})^\top = \mathbf{0} \in \mathbb{R}^{a\times a}$.*

Here the intuition for the finite-sum hard instance is combining n independent copies of the hard instance in the general case (7), then appropriate orthogonal matrices will convert the n independent variables with dimension d into one variable with dimension $n \times d$, which results in the desired hard instance. To preserve the zero chain property, for $\{\mathbf{U}^{(i)}\}_{i=1}^n \in \mathbf{Orth}(d+1, n(d+1), n)$, $\{\mathbf{V}^{(i)}\}_{i=1}^n \in \mathbf{Orth}(d+2, n(d+2), n)$, $\forall n, d \in \mathbb{N}^+$ and $x \in \mathbb{R}^{n(d+1)}$, $y \in \mathbb{R}^{n(d+2)}$, we set $\mathbf{U}^{(i)}$ and $\mathbf{V}^{(i)}$ by concatenating n matrices:

$$\begin{aligned}\mathbf{U}^{(i)} &= \left[\mathbf{0}_{d+1} \ \cdots \ \mathbf{0}_{d+1} \ \mathbf{I}_{d+1} \ \mathbf{0}_{d+1} \ \cdots \ \mathbf{0}_{d+1}\right],\\ \mathbf{V}^{(i)} &= \left[\mathbf{0}_{d+2} \ \cdots \ \mathbf{0}_{d+2} \ \mathbf{I}_{d+2} \ \mathbf{0}_{d+2} \ \cdots \ \mathbf{0}_{d+2}\right],\end{aligned} \quad (13)$$

where $\mathbf{0}_d, \mathbf{I}_d \in \mathbb{R}^{d\times d}$ are the zero and identity matrices respectively, while the i-th matrix above is the identity matrix. Hence, $\mathbf{U}^{(i)}x$ will be the $(id-d+1)$-th to the (id)-th elements of x, similar property also holds for $\mathbf{V}^{(i)}y$.

The hard instance construction here follows the idea of that in the deterministic hard instance (7), the basic motivation is that its primal function will be a finite-sum form of the primal function Φ_d defined in the deterministic case (9). We choose the following functions $H_d : \mathbb{R}^{d+1} \times \mathbb{R}^{d+2} \to \mathbb{R}$, $\Gamma_d^n : \mathbb{R}^{n(d+1)} \to \mathbb{R}$, $\Gamma_d^n(x) \triangleq \sum_{i=1}^n \sum_{j=i(d+1)-d}^{i(d+1)-1} \Gamma(x_j)$,

$$\begin{aligned}H_d(x,y;\lambda,\alpha) \triangleq \lambda_1\langle B_d x, y\rangle - \lambda_2\|y\|^2 - \frac{\lambda_1^2\sqrt{\alpha}}{2\lambda_2}\langle e_1, x\rangle\\ - \frac{\lambda_1^2\alpha}{4\lambda_2}x_{d+1}^2 + \frac{\lambda_1^2\sqrt{\alpha}}{4\lambda_2},\end{aligned} \quad (14)$$

then $\bar{f}_i, \bar{f} : \mathbb{R}^{n(d+1)} \times \mathbb{R}^{n(d+2)} \to \mathbb{R}$, $\{\mathbf{U}^{(i)}\}_{i=1}^n \in \mathbf{Orth}(d+1, n(d+1), n)$, $\{\mathbf{V}^{(i)}\}_{i=1}^n \in \mathbf{Orth}(d+2, n(d+2), n)$, $\bar{f}(x,y) \triangleq \frac{1}{n}\sum_{i=1}^n \bar{f}_i(x,y)$ and

$$\begin{aligned}&\bar{f}_i(x,y) \triangleq H_d\left(\mathbf{U}^{(i)}x, \mathbf{V}^{(i)}y; \lambda, \alpha\right) + \frac{\lambda_1^2\alpha}{2n\lambda_2}\Gamma_d^n(x),\\ &= \frac{1}{n}\sum_{i=1}^n\left[H_d\left(\mathbf{U}^{(i)}x, \mathbf{V}^{(i)}y; \lambda, \alpha\right) + \frac{\lambda_1^2\alpha}{2n\lambda_2}\Gamma_d^n(x)\right],\end{aligned} \quad (15)$$

note that by denoting $\Gamma_d(x) \triangleq \sum_{i=1}^d \Gamma(x_i)$, it is easy to find that

$$\begin{aligned}\Gamma_d^n(x) &= \sum_{i=1}^n \sum_{j=i(d+1)-d}^{i(d+1)-1} \Gamma(x_j) = \sum_{i=1}^n \Gamma_d\left(\mathbf{U}^{(i)}x\right)\\ &= \sum_{i=1}^n\sum_{j=1}^d \Gamma\left(\left(\mathbf{U}^{(i)}x\right)_j\right).\end{aligned} \quad (16)$$

Define $u^{(i)} \triangleq \mathbf{U}^{(i)}x$, we summarize the properties of the above functions in the following lemma.

Lemma 3.3 (Properties of $\bar{f}$) *For the above functions $\{\bar{f}_i\}_i$ and $\bar{f}$ in (15), they satisfy that:*

(i) $\{\bar{f}_i\}_i$ is L_F-average smooth where $L_F = \sqrt{\frac{1}{n}\max\left\{16\lambda_1^2 + 8\lambda_2^2, \frac{C_\gamma^2\lambda_1^4\alpha^2}{n\lambda_2^2} + \frac{\lambda_1^4\alpha^2}{\lambda_2^2} + 8\lambda_1^2\right\}}$.

(ii) $\bar{f}$ is μ_F-strongly concave on y where $\mu_F = \frac{2\lambda_2}{n}$.

(iii) For $n \in \mathbb{N}^+$, $L \ge 2n\mu > 0$, if we set $\lambda = \lambda^* = (\lambda_1^*, \lambda_2^*) = \left(\sqrt{\frac{n}{40}}L, \frac{n\mu}{2}\right)$, $\alpha = \frac{n\mu}{50L} \in [0,1]$, then $\{\bar{f}_i\}_i$ is L-AS and $\bar{f}$ is μ-strongly concave on y.

(iv) With Φ_d is defined in (9), let $\bar{\Phi}(x) \triangleq \max_y \bar{f}(x,y)$, we have $\bar{\Phi}(x) = \frac{1}{n}\sum_{i=1}^n \bar{\Phi}_i(x)$, $\bar{\Phi}_i(x) \triangleq \Phi_d(\mathbf{U}^{(i)}x)$.

We defer the proof of Lemma 3.3 to Appendix. From

Lemma 3.2, we have

$$\bar{\Phi}(0) - \inf_{x\in\mathbb{R}^{n(d+1)}} \bar{\Phi}(x) \le \frac{1}{n}\sum_{i=1}^{n} \sup_{x\in\mathbb{R}^{d+1}} \left(\bar{\Phi}(0) - \bar{\Phi}_i(x)\right) \le \frac{\lambda_1^2}{2\lambda_2}\left(\frac{\sqrt{\alpha}}{2} + 10\alpha d\right). \tag{17}$$

Define the index set $\mathcal{I}$ as all the indices $i \in [n]$ such that $u_d^{(i)} = u_{d+1}^{(i)} = 0,\ \forall i \in \mathcal{I}$. Suppose that $|\mathcal{I}| > \frac{n}{2}$, by orthogonality and Lemma 3.2 we have

$$\begin{aligned}\|\nabla\bar{\Phi}(x)\|^2 &= \left\|\frac{1}{n}\sum_{i=1}^{n}\nabla\bar{\Phi}_i(x)\right\|^2 = \frac{1}{n^2}\sum_{i=1}^{n}\left\|\nabla\Phi_d\left(u^{(i)}\right)\right\|^2 \\ &\ge \frac{1}{n^2}\sum_{i\in\mathcal{I}}\left\|\nabla\Phi_d\left(u^{(i)}\right)\right\|^2 \ge \frac{1}{n^2}\frac{n}{2}\left(\frac{\lambda_1^2}{8\lambda_2}\alpha^{\frac{3}{4}}\right)^2 = \frac{\lambda_1^4}{128n\lambda_2^2}\alpha^{\frac{3}{2}}.\end{aligned} \tag{18}$$

Now we arrive at our final theorem for the averaged smooth FS-NC-SC case as follows.

Theorem 3.2 (LB for AS FS-NC-SC) *For the linear-span algorithm class $\mathcal{A}$, parameters $L, \mu, \Delta > 0$ and component size $n \in \mathbb{N}^+$, if $L \ge 2n\mu > 0$, the accuracy ϵ satisfies that $\epsilon^2 \le \min\left(\frac{\sqrt{\alpha}L^2\Delta}{76800n\mu}, \frac{\alpha L^2\Delta}{1280n\mu}, \frac{L^2\Delta}{\mu}\right)$ where $\alpha = \frac{n\mu}{50L} \in [0,1]$, then we have*

$$\mathrm{Compl}_\epsilon\left(\mathcal{F}_{\mathrm{NCSC}}^{L,\mu,\Delta}, \mathcal{A}, \mathbb{O}_{\mathrm{IFO}}^{L,\mathrm{AS}}\right) = \Omega\big(n + \sqrt{n\kappa}\Delta L\epsilon^{-2}\big). \tag{19}$$

The theorem above indicates that for any $\mathtt{A} \in \mathcal{A}$, we can construct a function $f(x,y) = \frac{1}{n}\sum_{i=1}^{n} f_i(x,y)$, such that $f \in \mathcal{F}_{\mathrm{NCSC}}^{L,\mu,\Delta}$ and $\{f_i\}_i$ is L-AS, and $\mathtt{A}$ requires at least $\Omega\big(n + \sqrt{n\kappa}\Delta L\epsilon^{-2}\big)$ IFO calls to attain an approximate primal stationary point. The hard instance construction is based on $\bar{f}$ and $\bar{f}_i$ above (15), combined with a scaling trick similar to the one in the general case. Also we remark that lower bound holds for small enough ϵ, while the requirement on ϵ is comparable to those in existing literature, e.g. [Zhou and Gu, 2019, Han et al., 2021]. The detailed statement and proof of the theorem are deferred to Appendix.

Remark 3.2 (Tightness) *The state-of-the-art upper bound for NC-SC finite-sum problems is $\tilde{O}(n + \sqrt{n}\kappa^2\Delta L\epsilon^{-2})$ when $n \ge \kappa^2$ and $O\big((n\kappa + \kappa^2)\Delta L\epsilon^{-2}\big)$ when $n \le \kappa^2$ [Luo et al., 2020]. Note that there is still a large gap between upper and lower bounds on the dependence in terms of κ and n, which motivates the design of faster algorithms, we address this in Section 4. Note that a weaker result on the lower bound of nonconvex finite-sum averaged smooth minimization is $\Omega(\sqrt{n}\Delta L\epsilon^{-2})$ [Zhou and Gu, 2019]; here, our result presents explicitly the dependence on κ.*

4 FASTER ALGORITHMS FOR NC-SC MINIMAX PROBLEMS

In this section, we introduce a generic Catalyst acceleration scheme that turns existing optimizers for (finite-sum) SC-SC minimax problems into efficient, near-optimal algorithms for (finite-sum) NC-SC minimax optimization. Rooted in the inexact accelerated proximal point algorithm, the idea of Catalyst acceleration was introduced in Lin et al. [2015] for convex minimization and later extended to nonconvex minimization in Paquette et al. [2018] and nonconvex-concave minimax optimization in Yang et al. [2020b]. In stark contrast, we focus on NC-SC minimax problems.

The backbone of our Catalyst framework is to repeatedly solve regularized subproblems of the form:

$$\min_{x\in\mathbb{R}^{d_1}}\max_{y\in\mathbb{R}^{d_2}} f(x,y) + L\|x - \tilde{x}_t\|^2 - \frac{\tau}{2}\|y - \tilde{y}_t\|^2,$$

where $\tilde{x}_t$ and $\tilde{y}_t$ are carefully chosen prox-centers, and the parameter $\tau \ge 0$ is selected such that the condition numbers for x-component and y-component of these subproblems are well-balanced. Since f is L-Lipschitz smooth and μ-strongly concave in y, the above auxiliary problem is L-strongly convex in x and $(\mu + \tau)$-strongly concave in y. Therefore, it can be solved by a wide family of off-the-shelf first-order algorithms with linear convergence rate.

Our Catalyst framework, presented in Algorithm 1, consists of three crucial components: an inexact proximal point step for primal update, an inexact accelerated proximal point step for dual update, and a linear-convergent algorithm for solving the subproblems.

Inexact proximal point step in the primal. For the x-update in the outer loop, $\{x_0^t\}_{t=1}^T$, can be viewed as applying an inexact proximal point method to the primal function $\Phi(x)$, requiring to solve a sequence of auxiliary problems:

$$\min_{x\in\mathbb{R}^{d_1}}\max_{y\in\mathbb{R}^{d_2}} \left[\hat{f}_t(x,y) \triangleq f(x,y) + L\|x - x_0^t\|^2\right]. \tag{$\star$}$$

Inexact proximal point methods have been explored in minimax optimization in several work, e.g. [Lin et al., 2020b, Rafique et al., 2018]. Our scheme is distinct from these work in two aspects: (i) we introduce a new subroutine to approximately solve the auxiliary problems ($\star$) with near-optimal complexity, and (ii) the inexactness is measured by an adaptive stopping criterion using the gradient norms:

$$\|\nabla\hat{f}_t(x_0^{t+1}, y_0^{t+1})\|^2 \le \alpha_t\|\nabla\hat{f}_t(x_0^t, y_0^t)\|^2, \tag{20}$$

where $\{\alpha_t\}_t$ is carefully chosen. Using the adaptive stopping criterion significantly reduces the complexity of solving the auxiliary problems. We will show that the number of steps required is only logarithmic in L, μ without any dependence on target accuracy ϵ. Although the auxiliary problem is (L,μ)-SC-SC and can be solved with linear convergence by algorithms such as extragradient, OGDA, etc., these algorithms are not optimal in terms of the dependency on the condition number when $L > \mu$ [Zhang et al., 2019].

Algorithm 1 Catalyst for NC-SC Minimax Problems

```
Input: objective $f$, initial point $(x_0, y_0)$, smoothness constant $L$, strong-concavity const. $\mu$, and param. $\tau > 0$.

Let $(x_0^0, y_0^0) = (x_0, y_0)$ and $q = \frac{\mu}{\mu+\tau}$.
for all $t = 0, 1, ..., T$ do
    Let $z_1 = y_0^t$ and $k = 1$.
    Let $\hat{f}_t(x, y) \triangleq f(x, y) + L\|x - x_0^t\|^2$.
    repeat
        Find inexact solution $(x_k^t, y_k^t)$ to the problem below by algorithm $\mathcal{M}$ with initial point $(x_{k-1}^t, y_{k-1}^t)$:
            $\min_{x\in\mathbb{R}^{d_1}} \max_{y\in\mathbb{R}^{d_2}} [\tilde{f}_{t,k}(x, y) \triangleq f(x, y) + L\|x - x_0^t\|^2 - \frac{\tau}{2}\|y - z_k\|^2]$   $(\star\star)$
        such that $\|\nabla \tilde{f}_{t,k}(x_k^t, y_k^t)\|^2 \le \epsilon_k^t$.
        Let $z_{k+1} = y_k^t + \frac{\sqrt{q}-q}{\sqrt{q}+q}(y_k^t - y_{k-1}^t), k = k + 1$.
    until $\|\nabla \hat{f}_t(x_k^t, y_k^t)\|^2 \le \alpha_t \|\nabla \hat{f}_t(x_0^t, y_0^t)\|^2$
    Set $(x_0^{t+1}, y_0^{t+1}) = (x_k^t, y_k^t)$.
end for
Output: $\hat{x}_T$, which is uniformly sampled from $x_0^1, ..., x_0^T$.
```

Inexact accelerated proximal point step in the dual. To solve the auxiliary problem with optimal complexity, we introduce an inexact accelerated proximal point scheme. The key idea is to add an extra regularization in y to the objective such that the strong-convexity and strong-concavity are well-balanced. Therefore, we propose to iteratively solve the subproblems:

$$\min_{x\in\mathbb{R}^{d_1}} \max_{y\in\mathbb{R}^{d_2}} \left[\tilde{f}_{t,k}(x, y) \triangleq \hat{f}_t(x, y) - \frac{\tau}{2}\|y - z_k\|^2\right], \quad (\star\star)$$

where $\{z_k\}_k$ is updated analogously to Nesterov's accelerated method [Nesterov, 2005] and $\tau \ge 0$ is the regularization parameter. For example, by setting $\tau = L - \mu$, the subproblems become (L, L)-SC-SC and can be approximately solved by extragradient method with optimal complexity, to be discussed in more details in next section. Finally, when solving these subproblems, we use the following stopping criterion $\|\nabla \tilde{f}_{t,k}(x, y)\|^2 \le \epsilon_k^t$ with time-varying accuracy ϵ_k^t that decays exponentially with k.

Linearly-convergent algorithms for SC-SC subproblems. Let $\mathcal{M}$ be any algorithm that solves the subproblem $(\star\star)$ (denoting (x^*, y^*) as the optimal solution) at a linear convergence rate such that after N iterations:

$$\|x_N - x^*\|^2 + \|y_N - y^*\|^2 \le \left(1 - \frac{1}{\Lambda_{\mu,L}^{\mathcal{M}}(\tau)}\right)^N [\|x_0 - x^*\|^2 + \|y_0 - y^*\|^2], \tag{21}$$

if $\mathcal{M}$ is a deterministic algorithm; or taking expectation to the left-hand side above if $\mathcal{M}$ is randomized. The choices for $\mathcal{M}$ include, but are not limited to, EG [Tseng, 1995], optimistic gradient descent ascent (OGDA) [Gidel et al., 2018], SVRG [Balamurugan and Bach, 2016] , SPD1-VR [Tan et al., 2018], Point-SAGA [Luo et al., 2019]. For example, in the case of EG, $\Lambda_{\mu,L}^{\mathcal{M}}(\tau) = \frac{L+\max\{2L,\tau\}}{4\min\{L,\mu+\tau\}}$ [Tseng, 1995].

4.1 CONVERGENCE ANALYSIS

In this section, we analyze the complexity of each of the three components we discussed. Let T denote the outer-loop complexity, K the inner-loop complexity, and N the number of iterations for $\mathcal{M}$ (expected number if $\mathcal{M}$ is randomized) to solve subproblem $(\star\star)$. The total complexity of Algorithm 1 is computed by multiplying K, T and N. Later, we will provide a guideline for choosing parameter τ to achieve the best complexity, given an algorithm $\mathcal{M}$.

Theorem 4.1 (Outer loop) *Suppose function f is NC-SC with strong convexity parameter μ and L-Lipschitz smooth. If we choose $\alpha_t = \frac{\mu^5}{504L^5}$ for $t > 0$ and $\alpha_0 = \frac{\mu^5}{576\max\{1,L^7\}}$, the output $\hat{x}_T$ from Algorithm 1 satisfies*

$$\mathbb{E}\|\nabla\Phi(\hat{x}_T)\|^2 \le \frac{268L}{5T}\Delta + \frac{28L}{5T}D_y^0, \tag{22}$$

where $\Delta = \Phi(x_0) - \inf_x \Phi(x)$, $D_y^0 = \|y_0 - y^*(x_0)\|^2$ and $y^*(x_0) = \arg\max_{y\in\mathbb{R}^{d_2}} f(x_0, y)$.

This theorem implies that the algorithm finds an ϵ stationary point of Φ after inexactly solving $(\star)$ for $T = O\left(L(\Delta + D_y^0)\epsilon^{-2}\right)$ times. The dependency on D_y^0 can be eliminated if we select the initialization y_0 close enough to $y^*(x_0)$, which only requires an additional logarithmic cost by maximizing a strongly concave function.

Theorem 4.2 (Inner loop) *Under the same assumptions in Theorem 4.1, if we choose $\epsilon_k^t = \frac{\sqrt{2}\mu}{2}(1-\rho)^k \operatorname{gap}_{\hat{f}_t}(x_0^t, y_0^t)$ with $\rho < \sqrt{q} = \sqrt{\frac{\mu}{\mu+\tau}}$, we have*

$$\|\nabla\hat{f}_t(x_k^t, y_k^t)\|^2 \le \left[\frac{5508L^2}{\mu^2(\sqrt{q}-\rho)^2} + \frac{18\sqrt{2}L^2}{\mu}\right](1-\rho)^k\|\nabla\hat{f}_t(x_0^t, y_0^t)\|^2.$$

Particularly, setting $\rho = 0.9\sqrt{q}$, Theorem 4.2 implies after inexactly solving $(\star\star)$ for $K = \tilde{O}\left(\sqrt{(\tau+\mu)/\mu}\log\frac{1}{\alpha_t}\right)$ times, the stopping criterion (20) is satisfied. This complexity decreases with τ. However, we should not choose τ too small, because the smaller τ is, the harder it is for $\mathcal{M}$ to solve $(\star\star)$. The following theorem captures the complexity for algorithm $\mathcal{M}$ to solve the subproblem.

Theorem 4.3 (Complexity of solving subproblems $(\star\star)$) *Under the same assumptions in Theorem 4.1 and the choice* *of ϵ_k^t in Theorem 4.2, the number of iterations (expected number of iterations if $\mathcal{M}$ is stochastic) for $\mathcal{M}$ to solve ($\star\star$) such that $\|\nabla \tilde{f}_{t,k}(x,y)\|^2 \leq \epsilon_k^t$ is*

$$N = O\left(\Lambda_{\mu,L}^{\mathcal{M}}(\tau)\log\left(\frac{\max\{1,L,\tau\}}{\min\{1,\mu\}}\right)\right).$$

The above result implies that the subproblems can be solved within constant iterations that only depends on L, μ, τ and $\Lambda_{\mu,L}^{\mathcal{M}}$. This largely benefits from the use of warm-starting and stopping criterion with time-varying accuracy. In contrast, other inexact proximal point algorithms in minimax optimization, such as [Yang et al., 2020b, Lin et al., 2020b], fix the target accuracy, thus their complexity of solving the subproblems usually has an extra logarithmic factor in $1/\epsilon$.

The overall complexity of the algorithm follows immediately after combining the above three theorems:

Corollary 4.1 *Under the same assumptions in Theorem 4.1 and setting in Theorem 4.2, the total number (expected number if $\mathcal{M}$ is randomized) of gradient evaluations for Algorithm 1 to find an ϵ-stationary point of Φ, is*

$$\tilde{O}\left(\frac{\Lambda_{\mu,L}^{\mathcal{M}}(\tau)L(\Delta + D_y^0)}{\epsilon^2}\sqrt{\frac{\mu+\tau}{\mu}}\right). \tag{23}$$

In order to minimize the total complexity, we should choose the regularization parameter τ that minimizes $\Lambda_{\mu,L}^{\mathcal{M}}(\tau)\sqrt{\mu+\tau}$.

4.2 SPECIFIC ALGORITHMS AND COMPLEXITIES

In this subsection, we discuss specific choices for $\mathcal{M}$ and the corresponding optimal choices of τ, as well as the resulting total complexities for solving NC-SC problems.

Catalyst-EG/OGDA algorithm. When solving NC-SC minimax problems in the general setting, we set $\mathcal{M}$ to be either extra-gradient method (EG) or optimistic gradient descent ascent (OGDA). Hence, we have $\Lambda_{\mu,L}^{\mathcal{M}}(\tau) = \frac{L+\max\{2L,\tau\}}{4\min\{L,\mu+\tau\}}$ [Gidel et al., 2018, Azizian et al., 2020]. Minimizing $\Lambda_{\mu,L}^{\mathcal{M}}(\tau)\sqrt{\mu+\tau}$ yields that the optimal choice for τ is $L-\mu$. This leads to a total complexity of $\tilde{O}\left(\sqrt{\kappa}L(\Delta + D_y^0)\epsilon^{-2}\right)$.

$$\tilde{O}\left(\sqrt{\kappa}L(\Delta + D_y^0)\epsilon^{-2}\right). \tag{24}$$

Remark 4.1 *The above complexity matches the lower bound in Theorem 3.1, up to a logarithmic factor in L and κ. It improves over Minimax-PPA [Lin et al., 2020b] by $\log^2(1/\epsilon)$, GDA [Lin et al., 2020a] by $\kappa^{\frac{3}{2}}$ and therefore achieves the best of two worlds in terms of dependency on κ and ϵ. In addition, our Catalyst-EG/OGDA algorithm does not require the bounded domain assumption on y, unlike [Lin et al., 2020b].*

Catalyst-SVRG/SAGA algorithm. When solving NC-SC minimax problems in the averaged smooth finite-sum setting, we set $\mathcal{M}$ to be either SVRG or SAGA. Hence, we have $\Lambda_{\mu,L}^{\mathcal{M}}(\tau) \propto n + \left(\frac{L+\sqrt{2}\max\{2L,\tau\}}{\min\{L,\mu+\tau\}}\right)^2$ [Balamurugan and Bach, 2016][2] [3]. Minimizing $\Lambda_{\mu,L}^{\mathcal{M}}(\tau)\sqrt{\mu+\tau}$, the best choice for τ is (proportional to) $\max\left\{\frac{L}{\sqrt{n}} - \mu, 0\right\}$, which leads to the total complexity of

$$\tilde{O}\left(\left(n + n^{\frac{3}{4}}\sqrt{\kappa}\right)L(\Delta + D_y^0)\epsilon^{-2}\right). \tag{25}$$

Remark 4.2 *According to the lower bound established in Theorem 3.2, the dependency on κ in the above upper bound is nearly tight, up to logarithmic factors. Recall that SREDA [Luo et al., 2020] achieves the complexity of $\tilde{O}\left(\kappa^2\sqrt{n}\epsilon^{-2} + n + (n+\kappa)\log(\kappa)\right)$ for $n \geq \kappa^2$ and $O\left(\left(\kappa^2 + \kappa n\right)\epsilon^{-2}\right)$ for $n \leq \kappa^2$. Hence, our Catalyst-SVRG/SAGA algorithm attains better complexity in the regime $n \leq \kappa^4$. Particularly, in the critical regime $\kappa = \Omega(\sqrt{n})$ arising in statistical learning [Shalev-Shwartz and Ben-David, 2014], our algorithm performs strictly better.*

5 CONCLUSION

In this work, we take an initial step towards understanding the fundamental limits of minimax optimization in the nonconvex-strongly-concave setting for both general and finite-sum cases, and bridge the gaps between lower and upper bounds. It remains interesting to investigate whether the dependence on n can be further tightened in the complexity for finite-sum NC-SC minimax optimization.

Author Contributions

Siqi Zhang and Junchi Yang contributed equally to this paper.

Acknowledgements

This work is supported by NSF CMMI #1761699, NSF CCF #1934986. CG's research is partially supported by INRIA through the INRIA Associate Teams, and FONDECYT 1210362 project. NK's research is supported by NSF CCF #1704970.

[2] Although Balamurugan and Bach [2016] assumes individual smoothness, its analysis can be extended to average smoothness.

[3] SVRG in [Balamurugan and Bach, 2016] requires computing the proximal operator of an (μ,μ)-SC-SC function. For any (μ,μ)-SC-SC function in the form of $\sum_i f_i(x,y)$, we can rewrite it as $\sum_i[f_i(x,y) - \frac{\mu}{2n}||x||^2 + \frac{\mu}{2n}||y||^2] + \frac{\mu}{2}(||x||^2 - ||y||^2)$, where the first term is convex-concave, and the second term is (μ,μ)-SC-SC and admits a simple proximal operator.

References

A. Agarwal and L. Bottou. A lower bound for the optimization of finite sums. In *ICML*, pages 78–86, 2015.

Z. Allen-Zhu. How to make the gradients small stochastically: even faster convex and nonconvex sgd. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 1165–1175, 2018.

Y. Arjevani, Y. Carmon, J. C. Duchi, D. J. Foster, N. Srebro, and B. Woodworth. Lower bounds for non-convex stochastic optimization. *arXiv:1912.02365*, 2019.

W. Azizian, I. Mitliagkas, S. Lacoste-Julien, and G. Gidel. A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games. In *AISTATS*, pages 2863–2873, 2020.

P. Balamurugan and F. Bach. Stochastic variance reduction methods for saddle-point problems. In *NIPS*, pages 1416–1424, 2016.

R. I. Boţ and A. Böhm. Alternating proximal-gradient steps for (stochastic) nonconvex-concave minimax problems. *arXiv preprint arXiv:2007.13605*, 2020.

G. Braun, C. Guzmán, and S. Pokutta. Lower bounds on the oracle complexity of nonsmooth convex optimization via information theory. *IEEE Transactions on Information Theory*, 63(7):4709–4724, 2017.

Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, pages 1–50, 2019a.

Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Lower bounds for finding stationary points ii: first-order methods. *Mathematical Programming*, pages 1–41, 2019b.

B. Dai, N. He, Y. Pan, B. Boots, and L. Song. Learning from conditional distributions via dual embeddings. In *AISTATS*, pages 1458–1467, 2017.

B. Dai, A. Shaw, L. Li, L. Xiao, N. He, Z. Liu, J. Chen, and L. Song. SBEED: Convergent reinforcement learning with nonlinear function approximation. In *ICML*, 2018.

C. Daskalakis, S. Skoulakis, and M. Zampetakis. The complexity of constrained min-max optimization. *arXiv preprint arXiv:2009.09623*, 2020.

J. Diakonikolas. Halpern iteration for near-optimal and parameter-free monotone inclusion and strong solutions to variational inequalities. In *Conference on Learning Theory*, pages 1428–1451. PMLR, 2020.

J. Diakonikolas and P. Wang. Potential function-based framework for making the gradients small in convex and min-max optimization. *arXiv preprint arXiv:2101.12101*, 2021.

J. Diakonikolas, C. Daskalakis, and M. I. Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. *arXiv:2011.00364*, 2020.

C. Fang, C. J. Li, Z. Lin, and T. Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *NIPS*, pages 689–699, 2018.

D. J. Foster, A. Sekhari, O. Shamir, N. Srebro, K. Sridharan, and B. Woodworth. The complexity of making the gradient small in stochastic convex optimization. In *Conference on Learning Theory*. PMLR, 2019.

R. Frostig, R. Ge, S. Kakade, and A. Sidford. Unregularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization. In *ICML*. PMLR, 2015.

G. Gidel, H. Berard, G. Vignoud, P. Vincent, and S. Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *ICLR*, 2018.

O. Güler. New proximal point algorithms for convex minimization. *SIAM Journal on Optimization*, 2(4):649–664, 1992.

Y. Han, G. Xie, and Z. Zhang. Lower complexity bounds of finite-sum optimization problems: The results and construction. *arXiv preprint arXiv:2103.08280*, 2021.

J. Huang and N. Jiang. On the convergence rate of density-ratio based off-policy policy gradient methods. In *NeurIPS Offline Reinforcement Learning Workshop*, 2020.

A. Ibrahim, W. Azizian, G. Gidel, and I. Mitliagkas. Linear lower bounds and conditioning of differentiable games. In *ICML*, pages 4583–4593. PMLR, 2020.

C. Jin, P. Netrapalli, and M. Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *ICML*, 2020.

Q. Lei, J. Lee, A. Dimakis, and C. Daskalakis. Sgd learns one-layer networks in wgans. In *ICML*, 2020.

H. Lin, J. Mairal, and Z. Harchaoui. A universal catalyst for first-order optimization. *NIPS*, 2015.

H. Lin, J. Mairal, and Z. Harchaoui. Catalyst acceleration for first-order convex optimization: from theory to practice. *JMLR*, 18(1):7854–7907, 2018a.

Q. Lin, M. Liu, H. Rafique, and T. Yang. Solving weakly-convex-weakly-concave saddle-point problems as weakly-monotone variational inequality. *arXiv preprint arXiv:1810.10207*, 5, 2018b.

T. Lin, C. Jin, and M. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *ICML*, 2020a.

T. Lin, C. Jin, and M. I. Jordan. Near-optimal algorithms for minimax optimization. In *COLT*, 2020b.

L. Luo, C. Chen, Y. Li, G. Xie, and Z. Zhang. A stochastic proximal point algorithm for saddle-point problems. *arXiv preprint arXiv:1909.06946*, 2019.

L. Luo, H. Ye, Z. Huang, and T. Zhang. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. *NeurIPS*, 33, 2020.

A. S. Nemirovski and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

A. S. Nemirovsky. On optimality of krylov's information when solving linear operator equations. *J. Complex.*, 7 (2):121–130, 1991.

A. S. Nemirovsky. Information-based complexity of linear operator equations. *J. Complexity*, 8(2):153–175, 1992.

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

Y. Nesterov. How to make the gradients small. *Optima. Mathematical Optimization Society Newsletter*, pages 10–11, 2012.

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

M. Nouiehed, M. Sanjabi, T. Huang, J. D. Lee, and M. Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. *NeurIPS*, 2019.

Y. Ouyang and Y. Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, pages 1–35, 2019.

C. Paquette, H. Lin, D. Drusvyatskiy, J. Mairal, and Z. Harchaoui. Catalyst for gradient-based nonconvex optimization. In *AISTATS*, pages 613–622. PMLR, 2018.

Q. Qian, S. Zhu, J. Tang, R. Jin, B. Sun, and H. Li. Robust optimization over multiple domains. In *AAAI*, volume 33, pages 4739–4746, 2019.

H. Rafique, M. Liu, Q. Lin, and T. Yang. Non-convex min-max optimization: Provable algorithms and applications in machine learning. *arXiv preprint arXiv:1810.02060*, 2018.

R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM J. Control and Optimization*, 1976.

S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

A. Sinha, H. Namkoong, and J. Duchi. Certifying some distributional robustness with principled adversarial training. In *ICLR*, 2018.

C. Tan, T. Zhang, S. Ma, and J. Liu. Stochastic primal-dual method for empirical risk minimization with $o(1)$ per-iteration complexity. In *NIPS*, pages 8366–8375, 2018.

K. K. Thekumparampil, P. Jain, P. Netrapalli, and S. Oh. Efficient algorithms for smooth minimax optimization. *NeurIPS*, 32:12680–12691, 2019.

J. F. Traub, G. W. Wasilkowski, and H. Woźniakowski. *Information-based complexity*. Boston, MA: Academic Press, Inc., 1988. ISBN 0-12-697545-0.

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.

G. Xie, L. Luo, Y. Lian, and Z. Zhang. Lower complexity bounds for finite-sum convex-concave minimax optimization problems. In *ICML*, 2020.

Z. Xu, H. Zhang, Y. Xu, and G. Lan. A unified single-loop alternating gradient projection algorithm for nonconvex-concave and convex-nonconcave minimax problems. *arXiv preprint arXiv:2006.02032*, 2020.

J. Yang, N. Kiyavash, and N. He. Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. *NeurIPS*, 33, 2020a.

J. Yang, S. Zhang, N. Kiyavash, and N. He. A catalyst framework for minimax optimization. In *NeurIPS*, 2020b.

Y. Yang, N. Kiyavash, L. Song, and N. He. The devil is in the detail: A framework for macroscopic prediction via microscopic models. *NeurIPS*, 33, 2020c.

T. Yoon and E. K. Ryu. Accelerated algorithms for smooth convex-concave minimax problems with $\mathcal{O}(1/k^2)$ rate on squared gradient norm. *arXiv:2102.07922*, 2021.

J. Zhang, M. Hong, and S. Zhang. On lower iteration complexity bounds for the saddle point problems. *arXiv preprint arXiv:1912.07481*, 2019.

J. Zhang, P. Xiao, R. Sun, and Z.-Q. Luo. A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems. *arXiv preprint arXiv:2010.15768*, 2020.

D. Zhou and Q. Gu. Lower bounds for smooth nonconvex finite-sum optimization. In *International Conference on Machine Learning*, pages 7574–7583. PMLR, 2019.