Lower Complexity Bounds for Finite-Sum Convex-Concave Minimax Optimization Problems

Guangzeng Xie [1] Luo Luo [2] Yijiang Lian [3] Zhihua Zhang [4 5]

Abstract

This paper studies the lower bound complexity for minimax optimization problem whose objective function is the average of n individual smooth convex-concave functions. We consider the algorithm which has access to gradient and proximal oracle for each individual component. For the strongly-convex-strongly-concave case, we prove such an algorithm can not reach an ε-saddle point in fewer than $\Omega\left((n+\kappa)\log(1/\varepsilon)\right)$ iterations, where κ is the condition number of the objective function. This lower bound matches the upper bound of the existing proximal incremental first-order oracle algorithm in some specific case. We develop a novel construction to show the above result, which partitions the tridiagonal matrix of classical examples into n groups. This construction is friendly to the analysis of incremental gradient and proximal oracle and we also extend the analysis to general convex-concave cases.

1. Introduction

We consider the following minimax optimization problem

$$\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x},\mathbf{y}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x},\mathbf{y}), \tag{1}$$

where each individual component $f_i(\mathbf{x},\mathbf{y})$ is L-smooth, convex in $\mathbf{x}$ and concave in $\mathbf{y}$; the feasible sets $\mathcal{X}$ and $\mathcal{Y}$ are close and convex such that $\mathcal{X}\subseteq\mathbb{R}^{d_x}$ and $\mathcal{Y}\subseteq\mathbb{R}^{d_y}$. This formulation contains several popular machine learning applications such as matrix games (Carmon et al., 2019; Ibrahim et al., 2019), regularized empirical risk minimization (Zhang & Xiao, 2017; Tan et al., 2018), AUC maximization (Joachims, 2005; Ying et al., 2016; Shen et al., 2018), robust optimization (Ben-Tal et al., 2009; Yan et al., 2019) and reinforcement learning (Du et al., 2017).

A popular approach for solving minimax problems is the first order algorithm which iterates with gradient and proximal point operation (Korpelevich, 1977; Chen & Rockafellar, 1997; Chambolle & Pock, 2011; 2016; Mokhtari et al., 2019a;b; Thekumparampil et al., 2019). Zhang et al. (2019); Ibrahim et al. (2019) presented tight lower bounds for solving strongly-convex-strongly-concave minimax problems by first order algorithms. Ouyang & Xu (2018) studied a more general case that the objective function is possibly not strongly-convex or strongly-concave. However, these analyses (Ouyang & Xu, 2018; Zhang et al., 2019; Ibrahim et al., 2019) do not consider the specific finite-sum structure as in Problem (1). They only consider the deterministic first order algorithms which are based on the full gradient and exact proximal point iteration.

In big data regime, the number of components n in Problem (1) could be very large and we would like to devise stochastic optimization algorithms that avoid accessing the full gradient frequently. For example, Palaniappan & Bach (2016) used stochastic variance reduced gradient algorithms to solve (1). Similar to convex optimization, one can accelerate it by catalyst (Lin et al., 2018; Palaniappan & Bach, 2016) and proximal point (Defazio, 2016; Luo et al., 2019) techniques. Although stochastic optimization algorithms are widely used for solving minimax problems, the study of their lower bounds complexity is still open. All of existing lower bound analysis for stochastic optimization are focused on convex or nonconvex minimization problems (Agarwal & Bottou, 2015; Woodworth & Srebro, 2016; Carmon et al., 2017; Lan & Zhou, 2017; Fang et al., 2018; Arjevani et al., 2019).

This paper focuses on stochastic first order methods for solving Problem (1), which access to the Proximal Incremental

[1] Academy for Advanced Interdisciplinary Studies, Peking University [2] Department of Mathematics, Hong Kong University of Science and Technology [3] Baidu Inc., Beijing [4] School of Mathematical Sciences, Peking University [5] Huawei Noah's Ark Lab, Beijing. Correspondence to: Guangzeng Xie <smsxgz@pku.edu.cn>, Luo Luo <luoluo@ust.hk>, Zhihua Zhang <zhzhang@math.pku.edu.cn>, Yijiang Lian <lianyijiang@baidu.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

First-order Oracle (PIFO), that is,

$$h_{f_i}(\mathbf{x}, \mathbf{y}, \gamma) \triangleq \left[f_i(\mathbf{x}, \mathbf{y}), \nabla f_i(\mathbf{x}, \mathbf{y}), \mathrm{prox}_{f_i}^{\gamma}(\mathbf{x}, \mathbf{y}), \mathcal{P}_{\mathcal{X}}(\mathbf{x}), \mathcal{P}_{\mathcal{Y}}(\mathbf{y})\right], \tag{2}$$

where $i \in \{1, \dots, n\}$, $\gamma > 0$, the proximal operator is defined as

$$\mathrm{prox}_{f_i}^{\gamma}(\mathbf{x}, \mathbf{y}) \triangleq \arg\min_{\mathbf{u} \in \mathbb{R}^{d_x}, \mathbf{v} \in \mathbb{R}^{d_y}} \max \left\{ f_i(\mathbf{u}, \mathbf{v}) + \frac{1}{2\gamma} \|\mathbf{x} - \mathbf{u}\|_2^2 - \frac{1}{2\gamma} \|\mathbf{y} - \mathbf{v}\|_2^2 \right\},$$

and the projection operator is defined as

$$\mathcal{P}_{\mathcal{X}}(\mathbf{x}) = \arg\min_{\mathbf{u} \in \mathcal{X}} \|\mathbf{u} - \mathbf{x}\|_2, \mathcal{P}_{\mathcal{Y}}(\mathbf{y}) = \arg\min_{\mathbf{v} \in \mathcal{Y}} \|\mathbf{v} - \mathbf{y}\|_2.$$

We also define the Incremental First-order Oracle (IFO)

$$g_{f_i}(\mathbf{x}, \mathbf{y}, \gamma) \triangleq \left[f_i(\mathbf{x}, \mathbf{y}), \nabla f_i(\mathbf{x}, \mathbf{y}), \mathcal{P}_{\mathcal{X}}(\mathbf{x}), \mathcal{P}_{\mathcal{Y}}(\mathbf{y})\right].$$

PIFO provides more information than IFO and it would be potentially more powerful than IFO in first order optimization algorithms. Our goal is to find an ε-saddle point whose Euclidean squared distance to the exact solution of Problem (1) is not larger than ε or ε-suboptimal solution such that the primal dual gap is not larger than ε.

In this paper we show that the PIFO algorithm requires at least $\Omega((n + L/\mu)\log(1/\varepsilon))$ complexity to find an ε-saddle point of Problem (1) when each f_i is L-smooth and convex-concave; f is μ-strongly-convex-μ-strongly-concave. This result matches the upper bound of the existing PIFO algorithm (Zhang & Xiao, 2017; Lan & Zhou, 2017) for some specific bilinear problems. We also consider more general cases. When f is μ-strongly-concave but possibly non-strongly-concave, we establish a PIFO lower bound complexity $\Omega(n + L/\sqrt{\mu\varepsilon})$. If there is neither strongly-convexity nor strongly-concavity assumption, we prove that the PIFO lower bound will be $\Omega(n + L/\varepsilon)$.

The above results are mainly due to a novel lower bound analysis framework proposed in this paper, which is quite different from previous work. Our construction decomposes Nesterov's classical tridiagonal matrix into n groups and it facilitates the analysis for both the IFO and PIFO algorithms. In contrast, previous work is based on an aggregation method (Lan & Zhou, 2017; Zhou & Gu, 2019) or a very complicated adversarial construction (Woodworth & Srebro, 2016). Their results do not cover the minimax problems.

The remainder of the paper is organized as follows. In Section 2, we present preliminaries. In Section 3, we introduce the basic idea of our analysis framework. In Section 4, we provide the specific construction for the lower bound analysis. We compare our method to related work in Section 5 and conclude this work in Section 6.

2. Preliminaries

We first introduce the preliminaries used in this paper.

Definition 1. *For a differentiable function $\varphi(\mathbf{x}, \mathbf{y})$ from $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$ and $L > 0$, φ is said to be L-smooth if its gradient is L-Lipschitz continuous; that is, for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ and $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$, we have*

$$\|\nabla\varphi(\mathbf{x}_1, \mathbf{y}_1) - \nabla\varphi(\mathbf{x}_2, \mathbf{y}_2)\|_2 \leq L \left\| \begin{matrix} \mathbf{x}_1 - \mathbf{x}_2 \\ \mathbf{y}_1 - \mathbf{y}_2 \end{matrix} \right\|_2.$$

Definition 2. *For a differentiable function $\varphi(\mathbf{x}, \mathbf{y})$ from $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$, φ is said to be convex-concave, if φ is convex in $\mathbf{x}$ and concave in $\mathbf{y}$; that is, for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$, $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$ we have*

$$\varphi(\mathbf{x}_2, \mathbf{y}) \geq \varphi(\mathbf{x}_1, \mathbf{y}) + \nabla_{\mathbf{x}}\varphi(\mathbf{x}_1, \mathbf{y})^\top(\mathbf{x}_2 - \mathbf{x}_1),$$
$$\varphi(\mathbf{x}, \mathbf{y}_2) \leq \varphi(\mathbf{x}, \mathbf{y}_1) + \nabla_{\mathbf{y}}\varphi(\mathbf{x}, \mathbf{y}_1)^\top(\mathbf{y}_2 - \mathbf{y}_1).$$

Definition 3. *For constants $\mu_x, \mu_y \geq 0$, φ is said to be (μ_x, μ_y)-convex-concave, if the function*

$$\hat{\varphi}(\mathbf{x}, \mathbf{y}) = \varphi(\mathbf{x}, \mathbf{y}) - \frac{\mu_x}{2}\|\mathbf{x}\|_2^2 + \frac{\mu_y}{2}\|\mathbf{y}\|_2^2$$

is convex-concave.

Remark 1. *In Definition 3, we allow both μ_x and μ_y could be 0. In other words, we say that $\varphi(\mathbf{x}, \mathbf{y})$ is $(0, 0)$-convex-concave means the function is general convex-concave and $(0, \mu)$-convex-concave means it is μ-strongly-concave in $\mathbf{y}$ but possibly non-strongly-convex in $\mathbf{x}$.*

Definition 4. *We call a minimax optimization problem $\min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} \varphi(\mathbf{x}, \mathbf{y})$ satisfying strong duality condition if*

$$\min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} \varphi(\mathbf{x}, \mathbf{y}) = \max_{\mathbf{y} \in \mathcal{Y}} \min_{\mathbf{x} \in \mathcal{X}} \varphi(\mathbf{x}, \mathbf{y}).$$

The goal of a stochastic optimization algorithm for solving the minimax problem is finding an ε-suboptimal solution or ε-saddle point which are defined as follows.

Definition 5. *Suppose the strong duality of Problem (1) holds. We call $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathcal{X} \times \mathcal{Y}$ an ε-suboptimal solution to Problem (1), if*

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

Definition 6. *Suppose Problem (1) has an exact optimal solution $(\mathbf{x}^*, \mathbf{y}^*) \in \mathcal{X} \times \mathcal{Y}$ such that*

$$f(\mathbf{x}^*, \mathbf{y}) \leq f(\mathbf{x}^*, \mathbf{y}^*) \leq f(\mathbf{x}, \mathbf{y}^*)$$

for all $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$. We call $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathcal{X} \times \mathcal{Y}$ an ε-saddle point of Problem (1), if $\|\hat{\mathbf{x}} - \mathbf{x}^*\|_2^2 + \|\hat{\mathbf{y}} - \mathbf{y}^*\|_2^2 \leq \varepsilon$.

We define PIFO algorithms as follows.

Definition 7. *Consider a stochastic optimization algorithm $\mathcal{A}$ to solve Problem (1). Denote $(\mathbf{x}_t, \mathbf{y}_t)$ to be the point obtained by $\mathcal{A}$ at time-step t. The algorithm is said to be a PIFO algorithm if for any $t > 0$, we have*

$$\begin{aligned}
\tilde{\mathbf{x}}_t \in \operatorname{span}\big\{&\mathbf{x}_0, \cdots, \mathbf{x}_{t-1}, \mathbf{u}_1, \cdots, \mathbf{u}_t, \\
&\nabla_{\mathbf{x}} f_{i_1}(\mathbf{x}_0, \mathbf{y}_0), \cdots, \nabla_{\mathbf{x}} f_{i_t}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\big\}, \\
\tilde{\mathbf{y}}_t \in \operatorname{span}\big\{&\mathbf{y}_0, \cdots, \mathbf{y}_{t-1}, \mathbf{v}_1, \cdots, \mathbf{v}_t, \\
&\nabla_{\mathbf{y}} f_{i_1}(\mathbf{x}_0, \mathbf{y}_0), \cdots, \nabla_{\mathbf{y}} f_{i_t}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\big\}, \\
\mathbf{x}_t = \mathcal{P}_{\mathcal{X}}(&\tilde{\mathbf{x}}_t), \text{ and } \mathbf{y}_t = \mathcal{P}_{\mathcal{Y}}(\tilde{\mathbf{y}}_t),
\end{aligned}$$

where $(\mathbf{u}_t, \mathbf{v}_t) = \operatorname{prox}_{f_{i_t}}^{\gamma_t}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})$ and i_t is a random variable supported on $[n]$ by taking $\mathbb{P}(i_t = j) = p_j$ for each $t \geq 1$ and $1 \leq j \leq n$ along with $\sum_{j=1}^n p_j = 1$.

Without loss of generality, we assume that the PIFO algorithm $\mathcal{A}$ starts from $(\mathbf{x}_0, \mathbf{y}_0) = (\mathbf{0}_{d_x}, \mathbf{0}_{d_y})$ and $p_1 \leq p_2 \leq \cdots \leq p_n$ to simplify our analysis. Otherwise, we can take $\{\tilde{f}_i(\mathbf{x}, \mathbf{y}) = f_i(\mathbf{x}+\mathbf{x}_0, \mathbf{y}+\mathbf{y}_0)\}_{i=1}^n$ into consideration. On the other hand, suppose that $p_{s_1} \leq p_{s_2} \leq \cdots \leq p_{s_n}$ where $\{s_i\}_{i=1}^n$ is a permutation of $[n]$. We can define $\{\hat{f}_i\}_{i=1}^n$ such that $\hat{f}_{s_i} = f_i$ and consider $\mathcal{A}$ to take the component $\hat{f}_{s_i}$ by probability p_{s_i}.

3. A General Analysis Framework

In this section we introduce our construction and show that it enjoys some elegant properties when we use PIFO algorithms to solve it.

3.1. Construction

We first introduce the following class of matrices:

$$\mathbf{B}(m, \omega) \triangleq \begin{bmatrix} & & & -1 & 1 \\ & & -1 & 1 & \\ & \iddots & \iddots & & \\ -1 & 1 & & & \\ \omega & & & & \end{bmatrix} \in \mathbb{R}^{m \times m}.$$

Denote the l-th row of the matrix $\mathbf{B}(m, \omega)$ by $\mathbf{b}_l(m, \omega)^\top$.

Then we define

$$\mathbf{A}(m, \omega) \triangleq \begin{bmatrix} \omega^2+1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{bmatrix}.$$

It is easy to check the fact that

$$\mathbf{A}(m, \omega) = \mathbf{B}(m, \omega)^\top \mathbf{B}(m, \omega). \tag{3}$$

The matrix $\mathbf{A}(m, \omega)$ is widely-used in the analysis of lower bounds for first order optimization algorithms (Nesterov, 2013; Agarwal & Bottou, 2015; Lan & Zhou, 2017; Carmon et al., 2017; Zhou & Gu, 2019; Ouyang & Xu, 2018; Zhang et al., 2019).

We partition the rows of $\mathbf{B}(m, \omega)$ by index sets $\mathcal{L}_1, \ldots, \mathcal{L}_n$, where $\mathcal{L}_i = \big\{l : 1 \leq l \leq m, l \equiv i-1 \pmod n\big\}$. Then we construct the following class of functions by this partition:

$$r(\mathbf{x}, \mathbf{y}; \boldsymbol{\lambda}, m, \omega) \triangleq \frac{1}{n}\sum_{i=1}^n r_i(\mathbf{x}, \mathbf{y}; \boldsymbol{\lambda}, m, \omega), \tag{4}$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ and

$$\begin{aligned}
&r_i(\mathbf{x}, \mathbf{y}; \boldsymbol{\lambda}, m, \omega) \\
&= \begin{cases} \lambda_1 \sum_{l \in \mathcal{L}_1} \mathbf{y}^\top \mathbf{e}_l \mathbf{b}_l(m, \omega)^\top \mathbf{x} - \lambda_4 \langle \mathbf{e}_m, \mathbf{x} \rangle \\ \quad +\lambda_2 \|\mathbf{x}\|_2^2 - \lambda_3 \|\mathbf{y}\|_2^2, \quad \text{for } i = 1, \\ \lambda_1 \sum_{l \in \mathcal{L}_i} \mathbf{y}^\top \mathbf{e}_l \mathbf{b}_l(m, \omega)^\top \mathbf{x} \\ \quad +\lambda_2 \|\mathbf{x}\|_2^2 - \lambda_3 \|\mathbf{y}\|_2^2, \quad \text{for } i = 2, 3, \cdots, n. \end{cases}
\end{aligned}$$

The lower bound analysis of the PIFO algorithm for the minimax problem in this paper is based on the function $r(\mathbf{x}, \mathbf{y}; \boldsymbol{\lambda}, m, \omega)$ and its finite-sum formulation (4).

We show the smoothness, convexity, and concavity of the component function r_i in Lemma 1.

Lemma 1. *For any $\lambda_2 \geq 0, \lambda_3 \geq 0, \omega < \sqrt{2}$, we have that the r_i is $2\sqrt{\lambda_1^2 + 2\max\{\lambda_2, \lambda_3\}^2}$-smooth and $(2\lambda_2, 2\lambda_3)$ convex-concave.*

Consider the following minimax optimization problem

$$\min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} r(\mathbf{x}, \mathbf{y}; \boldsymbol{\lambda}, m, \omega), \tag{5}$$

where r is defined as Eq. (4) and

$$\mathcal{X} = \begin{cases} \mathbb{R}^m, & \text{if } \lambda_2 > 0, \\ \{\mathbf{x} \in \mathbb{R}^m : \|\mathbf{x}\|_2 \leq R_x\}, & \text{if } \lambda_2 = 0, \end{cases}$$

$$\mathcal{Y} = \begin{cases} \mathbb{R}^m, & \text{if } \lambda_3 > 0, \\ \{\mathbf{y} \in \mathbb{R}^m : \|\mathbf{y}\|_2 \leq R_y\}, & \text{if } \lambda_3 = 0, \end{cases}$$

where $R_x > 0$ and $R_y > 0$.

Note that the strong duality of the problem (5) holds.

Lemma 2. *For any $\lambda_2 \geq 0, \lambda_3 \geq 0, R_x > 0, R_y > 0$, we always have*

$$\min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} r(\mathbf{x}, \mathbf{y}; \boldsymbol{\lambda}, m, \omega) = \max_{\mathbf{y} \in \mathcal{Y}} \min_{\mathbf{x} \in \mathcal{X}} r(\mathbf{x}, \mathbf{y}; \boldsymbol{\lambda}, m, \omega)$$

3.2. Properties of the PIFO Algorithm

Now consider using the PIFO algorithm to solve the problem (5).

We define subspaces $\mathcal{F}_t = \mathrm{span}\{\mathbf{e}_m, \mathbf{e}_{m-1}, \cdots, \mathbf{e}_{m-t+1}\}$ for convex variable $\mathbf{x}$ and $\mathcal{G}_t = \mathrm{span}\{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_t\}$ for concave variable $\mathbf{y}$, where $t \in \{1, 2, \dots, m\}$. Additionally, we let $\mathcal{F}_0 = \mathcal{G}_0 = \{\mathbf{0}\}$. The following technical lemma plays a crucial role in our proofs.

Lemma 3. *Suppose that $n \geq 2$ and each function r_i satisfies $\lambda_1 \neq 0, \lambda_2, \lambda_3 \geq 0$. Denote $\mathrm{prox}_{r_i}^{\gamma}(\mathbf{x}, \mathbf{y})$ by $(\mathbf{u}_i, \mathbf{v}_i)$. Then we have the following results (we omit the parameters of r_i to simplify the presentation):*

1. *If $\mathbf{x} \in \mathcal{F}_k$, then we have $\mathcal{P}_{\mathcal{X}}(\mathbf{x}) \in \mathcal{F}_k$; and if $\mathbf{y} \in \mathcal{G}_k$, then we have $\mathcal{P}_{\mathcal{Y}}(\mathbf{y}) \in \mathcal{G}_k$.*
2. *If $\mathbf{x} \in \mathcal{F}_k, \mathbf{y} \in \mathcal{G}_k$ and $0 \leq k < m$, we have that*
$$\nabla_{\mathbf{x}} r_i(\mathbf{x}, \mathbf{y}), \mathbf{u}_i \in \begin{cases} \mathcal{F}_{k+1}, & \text{if } k \equiv i - 1 \pmod{n}, \\ \mathcal{F}_k, & \text{otherwise,} \end{cases}$$
and $\nabla_{\mathbf{y}} r_i(\mathbf{x}, \mathbf{y}), \mathbf{v}_i \in \mathcal{G}_k$.
3. *If $\mathbf{x} \in \mathcal{F}_{k+1}, \mathbf{y} \in \mathcal{G}_k$ and $0 \leq k < m$, we have that $\nabla_{\mathbf{x}} r_i(\mathbf{x}, \mathbf{y}), \mathbf{u}_i \in \mathcal{F}_{k+1}$ and*
$$\nabla_{\mathbf{y}} r_i(\mathbf{x}, \mathbf{y}), \mathbf{v}_i \in \begin{cases} \mathcal{G}_{k+1}, & \text{if } k \equiv i \pmod{n}, \\ \mathcal{G}_k, & \text{otherwise.} \end{cases}$$

Proof. The results about projection operator are trivial. Next, we can give the closed form expression of the gradient and proximal operation of r_i as follows

$$\begin{aligned}
\nabla_{\mathbf{x}} r_i(\mathbf{x}, \mathbf{y}) &= 2\lambda_2 \mathbf{x} + \lambda_1 \sum_{l \in \mathcal{L}_i} (\mathbf{e}_l^\top \mathbf{y}) \mathbf{b}_l + c_i \mathbf{e}_m, \\
\nabla_{\mathbf{y}} r_i(\mathbf{x}, \mathbf{y}) &= -2\lambda_3 \mathbf{y} + \lambda_1 \sum_{l \in \mathcal{L}_i} (\mathbf{b}_l^\top \mathbf{x}) \mathbf{e}_l, \\
\mathbf{u}_i &= \frac{1}{1 + 2\gamma\lambda_2} \left(\mathbf{x} - \gamma\lambda_1 \sum_{l \in \mathcal{L}_i} (\mathbf{e}_l^\top \mathbf{y}) \mathbf{b}_l - \gamma c_i \mathbf{e}_m \right), \\
\mathbf{v}_i &= \frac{1}{1 + 2\gamma\lambda_3} \left(\mathbf{y} + \gamma\lambda_1 \sum_{l \in \mathcal{L}_i} (\mathbf{b}_l^\top \mathbf{x}) \mathbf{e}_l \right),
\end{aligned}$$

where $c_1 = -1$ and $c_i = 0$ for $i = 2, \dots, n$.

If $\mathbf{x} = \mathbf{y} = \mathbf{0}$, then we have $\nabla_{\mathbf{y}} r_i(\mathbf{x}, \mathbf{y}) = \mathbf{v}_i = \mathbf{0}$ and $\nabla_{\mathbf{x}} r_i(\mathbf{x}, \mathbf{y}) = \mathbf{u}_i = \mathbf{0}$ for $i \geq 2$. Only when $i = 1$, we have $\nabla_{\mathbf{x}} r_1(\mathbf{x}, \mathbf{y}), \mathbf{u}_1 \in \mathcal{F}_1$.

Observe that $\mathbf{b}_l^\top \mathbf{x} = 0$ for $\mathbf{x} \in \mathcal{F}_k, l > k$ and $\mathbf{b}_l \in \mathcal{F}_{l+1}$ for $1 \leq l < m$. Then, we have

- if $\mathbf{y} \in \mathcal{G}_k$, $k \geq 1$, then $y_l \mathbf{b}_l \in \mathcal{F}_k$ for $l \neq k$ and $y_k \mathbf{b}_k \in \mathcal{F}_{k+1}$;
- if $\mathbf{x} \in \mathcal{F}_k$, $k \geq 1$, then $(\mathbf{b}_l^\top \mathbf{x}) \mathbf{e}_l \in \mathcal{G}_{k-1}$ and $(\mathbf{b}_k^\top \mathbf{x}) \mathbf{e}_k \in \mathcal{G}_k$.

Consequently, we can derive the result of the lemma:

- If $\mathbf{x} \in \mathcal{F}_k, \mathbf{y} \in \mathcal{G}_k$, $k \geq 1$, then
  - $\nabla_{\mathbf{y}} r_i(\mathbf{x}, \mathbf{y}), \mathbf{v}_i \in \mathcal{G}_k$,
  - $\nabla_{\mathbf{x}} r_i(\mathbf{x}, \mathbf{y}), \mathbf{u}_i \in \mathcal{F}_k$ for $k \notin \mathcal{L}_i$;
  - $\nabla_{\mathbf{x}} r_i(\mathbf{x}, \mathbf{y}), \mathbf{u}_i \in \mathcal{F}_{k+1}$ for $k \in \mathcal{L}_i$.
- If $\mathbf{x} \in \mathcal{F}_{k+1}, \mathbf{y} \in \mathcal{G}_k$, $k \geq 1$, then
  - $\nabla_{\mathbf{x}} r_i(\mathbf{x}, \mathbf{y}), \mathbf{u}_i \in \mathcal{F}_{k+1}$,
  - $\nabla_{\mathbf{y}} r_i(\mathbf{x}, \mathbf{y}), \mathbf{v}_i \in \mathcal{G}_k$ for $k + 1 \notin \mathcal{L}_i$,
  - $\nabla_{\mathbf{y}} r_i(\mathbf{x}, \mathbf{y}), \mathbf{v}_i \in \mathcal{G}_{k+1}$ for $k + 1 \in \mathcal{L}_i$.

□

Suppose the time-step t_0 of a PIFO algorithm $\mathcal{A}$ satisfies $\mathbf{x}_{t_0} \in \mathcal{F}_k$ and $\mathbf{y}_{t_0} \in \mathcal{G}_k$. Then Lemma 3 implies that $\mathbf{x}_t \in \mathcal{F}_k$ and $\mathbf{y}_t \in \mathcal{G}_k$ $(t > t_0)$ will hold until the algorithm $\mathcal{A}$ draws the component f_i such that $k \in \mathcal{L}_i$. After that, $\mathbf{x}_t \in \mathcal{F}_{k+1}$ and $\mathbf{y}_t \in \mathcal{G}_k$ will hold until $\mathcal{A}$ draws the component f_j such that $k + 1 \in \mathcal{L}_j$.

We can describe the process of using PIFO algorithm $\mathcal{A}$ to solve Problem (5) by the following lemma.

Lemma 4. *Let $T_0 = 0$ and*

$$T_k = \min\{t : t > T_{k-1}, i_t \equiv \lfloor k/2 \rfloor + 1 (\mathrm{mod}\ n)\} \quad (6)$$

for any $k \geq 1$. Then we have $\mathbf{x}_t \in \mathcal{F}_{k-1}$ for $t < T_{2k-1}$ and $\mathbf{y}_t \in \mathcal{G}_{k-1}$ for $t < T_{2k}$. Moreover, we can write T_k as the sum of k independent random variables $\{Y_l\}_{l=1}^k$, i.e., $T_k = \sum_{l=1}^k Y_l$, where Y_l follows a geometric distribution with success probability $q_l = p_{l'}$ such that

$$l' \equiv \lfloor l/2 \rfloor + 1 \pmod{n} \ \text{ and } \ 1 \leq l' \leq n.$$

The basic idea of the lower bound analysis is that we guarantee the PIFO algorithm to extend the spaces of $\mathrm{span}\{\mathbf{x}_0, \dots, \mathbf{x}_t\}$ and $\mathrm{span}\{\mathbf{y}_0, \dots, \mathbf{y}_t\}$ slowly as t is increasing. Lemma 4 shows $\mathrm{span}\{\mathbf{x}_0, \dots, \mathbf{x}_{T_{2k}}\} \subseteq \mathcal{F}_k$ and $\mathrm{span}\{\mathbf{y}_0, \dots, \mathbf{x}_{T_{2k+1}}\} \subseteq \mathcal{G}_k$. Then we can regard quantity T_k as the one that reflects how $\mathrm{span}\{\mathbf{x}_0, \dots, \mathbf{x}_t\}$ and $\mathrm{span}\{\mathbf{y}_0, \dots, \mathbf{y}_t\}$ vary. Because T_k can be written as the sum of geometrically distributed random variables, we introduce the following lemma for further analysis.

Lemma 5. *Let $\{Y_i\}_{1 \leq i \leq N}$ be independent random variables, and Y_i follows a geometric distribution with success probability p_i. Then*

$$\mathbb{P}\left(\sum_{i=1}^N Y_i > \frac{N^2}{4(\sum_{i=1}^N p_i)} \right) \geq 1 - \frac{16}{9N}.$$

Based on Lemmas 4 and 5, we can estimate how many PIFO calls that $\mathcal{A}$ needs to obtain an output which is close to the solution of Problem (5) sufficiently.

Lemma 6. *We consider the minimax Problem (5) and any criterion $H(\mathbf{x}, \mathbf{y})$ of measuring how $\mathbf{x}, \mathbf{y}$ close to solution to the problem. Suppose that $M \geq 1$, $N = nM/2$ and M satisfies $\min_{\mathbf{x} \in \mathcal{X} \cap \mathcal{F}_M} \min_{\mathbf{y} \in \mathcal{Y} \cap \mathcal{G}_M} H(\mathbf{x}, \mathbf{y}) \geq 9\varepsilon$, then we have $\min_{t \leq N} \mathbb{E}\left(H(\mathbf{x}_t, \mathbf{y}_t)\right) \geq \varepsilon$.*

Proof. For any $t \le N$, we have

$$\begin{aligned}
&\min_{t\le N} \mathbb{E}\left(H(\mathbf{x}_t, \mathbf{y}_t)\right) \\
&\ge \min_{t\le N} \mathbb{E}\left(H(\mathbf{x}_t, \mathbf{y}_t) \mid N < T_{2M+1}\right) \mathbb{P}\left(N < T_{2M+1}\right) \\
&\ge \mathbb{E}\left(\min_{\mathbf{x}\in\mathcal{X}\cap\mathcal{F}_M} \min_{\mathbf{y}\in\mathcal{Y}\cap\mathcal{G}_M} H(\mathbf{x}, \mathbf{y})\right) \mathbb{P}\left(N < T_{2M+1}\right) \\
&\ge 9\varepsilon \mathbb{P}\left(T_{2M+1} > N\right),
\end{aligned}$$

where T_k is defined in Eq. (6), and the second inequality follows from $\mathbf{x}_t \in \mathcal{F}_M$ and $\mathbf{y}_t \in \mathcal{G}_M$ for $t < T_{2M+1}$ by Lemma 4.

Then, according to Lemma 4, we have $T_{2M+1} = \sum_{l=1}^{2M+1} Y_l$. Here $\{Y_l\}_{l=1}^{2M+1}$ are independent random variables where Y_l follows a geometric distribution with success probability $q_l = p_{l'}$ such that $l' \equiv \lfloor l/2 \rfloor + 1 (\bmod n)$ and $1 \le l' \le n$.

Suppose $M = s_1 n + s_2$ and $0 \le s_2 < n$. Recalling that $p_1 \le p_2 \le \cdots \le p_n$, we have

$$\begin{aligned}
\sum_{l=1}^{2M+1} q_l =& 2s_1 + 2\sum_{l=1}^{s_2+1} p_l - p_1 \le 2s_1 + 2\sum_{l=1}^{s_2+1} p_l \\
\le& 2s_1 + 2\cdot\frac{s_2+1}{n} = \frac{2M+2}{n}.
\end{aligned}$$

Hence, we can use Lemma 5 to obtain

$$\begin{aligned}
\mathbb{P}\left(\sum_{l=1}^{2M+1} Y_l > \frac{nM}{2}\right) \ge& \mathbb{P}\left(\sum_{l=1}^{2M+1} Y_l > \frac{(2M+1)^2 n}{4(2M+2)}\right) \\
\ge& 1 - \frac{16}{9(M+1)} \ge \frac{1}{9},
\end{aligned}$$

where the first inequality follows from $(2M+1)^2 > 4M(M+1)$. Therefore, we achieve the desired result

$$\min_{t\le N} \mathbb{E}\left(H(\mathbf{x}_t, \mathbf{y}_t)\right) \ge 9\varepsilon \mathbb{P}\left(T_{2M+1} > N\right) \ge \varepsilon.$$

□

4. Main Results

In this section we show the specific construction for the lower bound analysis of minimax problems in different kinds of assumptions. We start with strongly-convex-strongly-concave setting, then consider more general cases.

4.1. Strongly-Convex-Strongly-Concave Case

For the lower bound analysis of the strongly-convex-strongly-concave minimax problem, we define the following class of component functions.

Definition 8. *For fixed L, μ and n such that $L/\mu \ge \sqrt{2}, \mu > 0, n \ge 2$, let*

$$\alpha = \sqrt{\frac{L^2 - 2\mu^2}{n^2\mu^2} + 1} \text{ and } \boldsymbol{\lambda}_{SC} = \left(\sqrt{\frac{L^2-2\mu^2}{4}}, \frac{\mu}{2}, \frac{\mu}{2}, 1\right).$$

Define functions $f_{SC,i} : \mathbb{R}^{2m} \to \mathbb{R}$ for $i = 1, \dots, n$

$$f_{SC,i}(\mathbf{x}, \mathbf{y}) = r_i\left(\mathbf{x}, \mathbf{y}; \boldsymbol{\lambda}_{SC}, m, \sqrt{\frac{2}{\alpha+1}}\right),$$

and the minimax problem

$$\min_{\mathbf{x}\in\mathbb{R}^m} \max_{\mathbf{y}\in\mathbb{R}^m} F_{SC}(\mathbf{x}, \mathbf{y}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_{SC,i}(\mathbf{x}, \mathbf{y}). \tag{7}$$

The following lemma shows that F_{SC} is (μ, μ)-convex-concave and we can present the closed form of the optimal solution for Problem (7).

Lemma 7. *Consider minimax problem (7) in Definition 8. Then we have following properties.*

1. *Each component function $f_{SC,i}$ is L-smooth and (μ, μ)-convex-concave.*
2. *The saddle point of Problem (7) is*

$$\begin{cases}
\mathbf{x}^* = \frac{2n\mu(\alpha+1)}{L^2-2\mu^2}(q^m, q^{m-1}, \cdots, q)^\top, \\
\mathbf{y}^* = \frac{2}{\sqrt{L^2-2\mu^2}}\left(q, q^2, \cdots, q^{m-1}, \sqrt{\frac{\alpha+1}{2}} q^m\right)^\top,
\end{cases}$$

where $q = \frac{\alpha-1}{\alpha+1}$.

Proof. The first statement of this lemma can be directly obtained by Lemma 1. The remainder of the proof is focus on the solution of Problem (7).

We can rewrite the function F_{SC} as follows

$$\begin{aligned}
F_{\text{SC}}(\mathbf{x}, \mathbf{y}) =& \frac{\mu}{2}\left(\|\mathbf{x}\|_2^2 - \|\mathbf{y}\|_2^2\right) - \frac{1}{n}\langle \mathbf{e}_m, \mathbf{x}\rangle \\
&+ \sqrt{\frac{L^2-2\mu^2}{4n^2}}\left\langle \mathbf{B}(m, \omega)\mathbf{x}, \mathbf{y}\right\rangle,
\end{aligned}$$

where $\omega = \sqrt{\frac{2}{\alpha+1}}$.

Letting the gradient of $F_{SC}(\mathbf{x}, \mathbf{y})$ be zero, we obtain

$$\begin{cases}
\mu\mathbf{x} + \sqrt{\frac{L^2-2\mu^2}{4n^2}}\mathbf{B}(m,\omega)^\top\mathbf{y} - \frac{1}{n}\mathbf{e}_m = \mathbf{0}, \\
-\mu\mathbf{y} + \sqrt{\frac{L^2-2\mu^2}{4n^2}}\mathbf{B}(m,\omega)\mathbf{x} = \mathbf{0},
\end{cases}$$

which implies

$$\mathbf{y} = \sqrt{\frac{L^2-2\mu^2}{4n^2\mu^2}}\mathbf{B}(m,\omega)\mathbf{x}, \tag{8}$$

$$\left(\mu\mathbf{I}+\frac{L^2-2\mu^2}{4n^2\mu}\mathbf{B}(m,\omega)^\top\mathbf{B}(m,\omega)\right)\mathbf{x}=\frac{1}{n}\mathbf{e}_m. \quad (9)$$

The equation (9) is equivalent to

$$\begin{bmatrix}\omega^2+1+\beta & -1 & & & \\ -1 & 2+\beta & -1 & & \\ & \ddots & \ddots & & \\ & & -1 & 2+\beta & -1 \\ & & & -1 & 1+\beta\end{bmatrix}\mathbf{x}=\begin{bmatrix}0\\0\\ \vdots \\ 0\\ \frac{\beta}{n\mu}\end{bmatrix}, \quad (10)$$

where $\beta=\frac{4n^2\mu^2}{L^2-2\mu^2}$.

Let $q=\frac{\alpha-1}{\alpha+1}$ which is a root of the equation

$$z^2-\left(2+\frac{4n^2\mu^2}{L^2-2\mu^2}\right)z+1=0.$$

Then, we can check that the solution of (10) equation is

$$\mathbf{x}^*=\frac{2n\mu(\alpha+1)}{L^2-2\mu^2}(q^m,q^{m-1},\cdots,q)^\top.$$

Substituting above result into (8), we have

$$\mathbf{y}^*=\frac{2}{\sqrt{L^2-2\mu^2}}(q,q^2,\cdots,q^{m-1},\sqrt{\frac{\alpha+1}{2}}q^m)^\top.$$

□

We now can prove the lower bound complexity for finding $\mathcal{O}(\varepsilon)$-saddle point of Problem (7) by PIFO algorithms.

Theorem 1. *Consider minimax problem (7) and $\varepsilon>0$ such that*

$$\frac{L}{\mu}\geq\sqrt{n^2+2},\ \ \varepsilon\leq\frac{1}{2}\left(\frac{\sqrt{2}-1}{\sqrt{2}+1}\right)^2,\ \ \textit{and}$$
$$m=\left\lfloor\frac{1}{2}\left(\sqrt{\frac{L^2-2\mu^2}{n^2\mu^2}}+1\right)\log\left(\frac{1}{18\varepsilon}\right)\right\rfloor.$$

In order to find $(\hat{\mathbf{x}},\hat{\mathbf{y}})$ such that

$$\mathbb{E}\big[\|\hat{\mathbf{x}}-\mathbf{x}^*\|_2^2+\|\hat{\mathbf{y}}-\mathbf{y}^*\|_2^2\big]<\varepsilon\big[\|\mathbf{x}_0-\mathbf{x}^*\|_2^2+\|\mathbf{y}_0-\mathbf{y}^*\|_2^2\big],$$

the PIFO algorithm $\mathcal{A}$ needs at least

$$\Omega\left(\left(n+\frac{L}{\mu}\right)\log\left(\frac{1}{18\varepsilon}\right)\right)$$

PIFO queries.

Proof. We use the same definition of q as Lemma 7. For $L/\mu\geq\sqrt{n^2+2}$, we have $\alpha=\sqrt{\frac{L^2-2\mu^2}{n^2\mu^2}+1}\geq\sqrt{2}$ and $q=\frac{\alpha-1}{\alpha+1}\geq\frac{\sqrt{2}-1}{\sqrt{2}+1}$. The assumption on ε means $\varepsilon\leq\frac{1}{2}q^2$. Note that the function $h(\beta)=\frac{1}{\log\left(\frac{\beta+1}{\beta-1}\right)}-\frac{\beta}{2}$ is increasing when $\beta>1$ and $\lim_{\beta\to+\infty}h(\beta)=0$. Thus there holds

$$\frac{\alpha}{2}+h(\sqrt{2})\leq-\frac{1}{\log q}\leq\frac{\alpha}{2}.$$

Let $M=\left\lfloor\frac{\log(18\varepsilon)}{2\log q}\right\rfloor$, $\xi=\frac{2n\mu(\alpha+1)}{L^2-2\mu^2}$ and $\eta=\frac{2}{\sqrt{L^2-2\mu^2}}$, then we have $M\geq 1$ and

$$m=\left\lfloor\frac{\alpha}{2}\log\left(\frac{1}{18\varepsilon}\right)\right\rfloor\geq\left\lfloor\frac{\log(18\varepsilon)}{\log q}\right\rfloor\geq 2M.$$

Consequently, we can achieve

$$\begin{aligned}&\frac{\min_{\mathbf{x}\in\mathcal{F}_M}\|\mathbf{x}-\mathbf{x}^*\|_2^2+\min_{\mathbf{y}\in\mathcal{G}_M}\|\mathbf{y}-\mathbf{y}^*\|_2^2}{\|\mathbf{x}_0-\mathbf{x}^*\|_2^2+\|\mathbf{y}_0-\mathbf{y}^*\|_2^2}\\ =&\frac{(\xi^2+\eta^2)\cdot\frac{q^{2(M+1)}-q^{2(m+1)}}{1-q^2}+\eta^2\cdot\frac{\alpha-1}{2}q^{2m}}{(\xi^2+\eta^2)\cdot\frac{q^2-q^{2(m+1)}}{1-q^2}+\eta^2\cdot\frac{\alpha-1}{2}q^{2m}}\\ \geq&\frac{q^{2M}-q^{2m}}{1-q^{2m}}\geq\frac{q^{2M}}{2}\geq 9\varepsilon,\end{aligned}$$

where the first inequality is according to $\frac{a+c}{b+c}\geq\frac{a}{b}$ for $b\geq a$ and $c\geq 0$, and the second inequality is due to $M\leq m/2$.

Hence, following from Lemma 6 along with $H(\mathbf{x},\mathbf{y})=\frac{\|\mathbf{x}-\mathbf{x}^*\|_2^2+\|\mathbf{y}-\mathbf{y}^*\|_2^2}{\|\mathbf{x}_0-\mathbf{x}^*\|_2^2+\|\mathbf{y}_0-\mathbf{y}^*\|_2^2}$, $M=\left\lfloor\frac{\log(18\varepsilon)}{2\log q}\right\rfloor$ and $N=nM/2$, we know that

$$\min_{t\leq N}\mathbb{E}\left(\frac{\|\mathbf{x}_t-\mathbf{x}^*\|_2^2+\|\mathbf{y}_t-\mathbf{y}^*\|_2^2}{\|\mathbf{x}_0-\mathbf{x}^*\|_2^2+\|\mathbf{y}_0-\mathbf{y}^*\|_2^2}\right)\geq\varepsilon.$$

Therefore, in order to find $(\hat{\mathbf{x}},\hat{\mathbf{y}})$ such that

$$\mathbb{E}\big[\|\hat{\mathbf{x}}-\mathbf{x}^*\|_2^2+\|\hat{\mathbf{y}}-\mathbf{y}^*\|_2^2\big]<\varepsilon\big[\|\mathbf{x}_0-\mathbf{x}^*\|_2^2+\|\mathbf{y}_0-\mathbf{y}^*\|_2^2\big],$$

the PIFO algorithm $\mathcal{A}$ needs at least N PIFO queries.

At last, we can estimate N by

$$\begin{aligned}-\frac{1}{\log(q)}&=\frac{1}{\log\left(\frac{\alpha+1}{\alpha-1}\right)}\geq\frac{\alpha}{2}+h(\sqrt{2})\\ &=\frac{1}{2}\sqrt{\frac{L^2-2\mu^2}{n^2\mu^2}+1}+h(\sqrt{2})\\ &\geq\frac{\sqrt{2}}{4}\left(\sqrt{\frac{L^2-2\mu^2}{n^2\mu^2}}+1\right)+h(\sqrt{2})\\ &\geq\frac{\sqrt{2L^2/\mu^2-4}}{4n}+\frac{\sqrt{2}}{4}+h(\sqrt{2}),\end{aligned}$$

and

$$N=Mn/2=\frac{n}{2}\left\lfloor\frac{\log(18\varepsilon)}{2\log q}\right\rfloor$$

$$\geq \frac{n}{8}\left(-\frac{1}{\log(q)}\right)\log\left(\frac{1}{18\varepsilon}\right)$$
$$\geq \frac{n}{8}\left(\frac{\sqrt{2L^2/\mu^2-4}}{4n}+\frac{\sqrt{2}}{4}+h(\sqrt{2})\right)\log\left(\frac{1}{18\varepsilon}\right)$$
$$=\Omega\left(\left(n+\frac{L}{\mu}\right)\log\left(\frac{1}{18\varepsilon}\right)\right),$$

where we use the fact $2\lfloor\beta\rfloor\geq\beta$ for $\beta\geq 1$. □

Zhang & Xiao (2017) considered a specific bilinear case of Problem (1) with $\mathcal{X}=\mathbb{R}^d$, $\mathcal{Y}=\mathbb{R}^n$ and each individual component function has the form of

$$f_i(\mathbf{x},\mathbf{y})=h(\mathbf{x})+y_i\langle\mathbf{a}_i,\mathbf{x}\rangle-J_i(y_i),$$

where h is μ_x-strongly-convex and J_i is μ_y-strongly-convex. They proposed stochastic primal-dual coordinate (SPDC) method which can find $\mathcal{O}(\varepsilon)$-saddle point with at most $\mathcal{O}\left(\left(n+\sqrt{\frac{nL^2}{\mu_x\mu_y}}\right)\log(1/\varepsilon)\right)$ PIFO queries. Note that f is $(\mu_x,\mu_y/n)$-convex-concave and if we set $\mu_x=\mu_y/n=\mu$, the complexity will be $\mathcal{O}\left(\left(n+\frac{L}{\mu}\right)\log(1/\varepsilon)\right)$, which implies that our lower bound is tight for this problem.

In general strongly-convex-strongly-concave case, the best known upper bound complexity for IFO/PIFO algorithms is $\mathcal{O}\left(\left(n+\frac{\sqrt{n}L}{\mu}\right)\log(1/\varepsilon)\right)$ (Palaniappan & Bach, 2016; Luo et al., 2019), which still exist a $\sqrt{n}$ gap to our lower bound.

4.2. Convex-Strongly-Concave Case

We now consider the finite-sum minimax problem whose each individual component is strongly-concave but possibly non-strongly-convex. Our analysis is based on the following functions.

Definition 9. *For fixed L,μ,n and R_x such that $L/\mu\geq\sqrt{2},\mu>0,R_x>0,n\geq 2$, let*

$$\gamma=\frac{R_x(L^2-2\mu^2)}{4n\mu(m+1)^{3/2}}\text{ and }\boldsymbol{\lambda}_{SCC}=\left(\sqrt{\frac{L^2-2\mu^2}{4}},0,\frac{\mu}{2},\gamma\right).$$

Define functions $f_{SCC,i}:\mathbb{R}^{2m}\to\mathbb{R}$ for $i=1,\dots,n$ as

$$f_{SCC,i}(\mathbf{x},\mathbf{y})=r_i\left(\mathbf{x},\mathbf{y};\boldsymbol{\lambda}_{SCC},m,1\right)$$

and the minimax problem

$$\min_{\mathbf{x}\in\mathcal{X}'}\max_{\mathbf{y}\in\mathbb{R}^m}F_{SCC}(\mathbf{x},\mathbf{y})\triangleq\frac{1}{n}\sum_{i=1}^n f_{SCC,i}(\mathbf{x},\mathbf{y}),\qquad(11)$$

where $\mathcal{X}'=\{\mathbf{x}:\|\mathbf{x}\|_2\leq R_x\}$.

It is easily checked each component function $f_{\mathrm{SCC},i}$ is L-smooth and $(0,\mu)$-convex-concave by Lemma 1.

The following lemma helps us to establish the lower bound with respect to the primal dual gap.

Lemma 8. *Let $\phi(\mathbf{x})\triangleq\max_{\mathbf{y}}F_{SCC}(\mathbf{x},\mathbf{y})$ and $\psi(\mathbf{y})\triangleq\min_{\mathbf{x}\in\mathcal{X}'}F_{SCC}(\mathbf{x},\mathbf{y})$. Then, for $k=\left\lfloor\frac{m+1}{2}\right\rfloor$, we have*

$$\min_{\mathbf{x}\in\mathcal{X}'\cap\mathcal{F}_k}\phi(\mathbf{x})-\max_{\mathbf{y}\in\mathcal{G}_k}\psi(\mathbf{y})\geq\frac{(L^2-2\mu^2)R_x^2}{16n^2\mu(k+1)^2}.$$

Proof. We prove the result as follows

$$\min_{\mathbf{x}\in\mathcal{X}'\cap\mathcal{F}_k}\phi(\mathbf{x})-\max_{\mathbf{y}\in\mathcal{G}_k}\psi(\mathbf{y})\geq-\frac{2\mu k\gamma^2}{L^2-2\mu^2}+\frac{R_x\gamma}{n\sqrt{k+1}}$$
$$=\frac{(L^2-2\mu^2)R_x^2}{8n^2\mu}\frac{2(m+1)^{3/2}-k\sqrt{k+1}}{(m+1)^3\sqrt{k+1}}$$
$$\geq\frac{(L^2-2\mu^2)R_x^2}{8n^2\mu}\frac{4\sqrt{2}-1}{8(k+1)^2}>\frac{(L^2-2\mu^2)R_x^2}{16n^2\mu(k+1)^2},$$

where the equality is due to $\gamma=\frac{R_x(L^2-2\mu^2)}{4n\mu(m+1)^{3/2}}$, the first inequality is based on Lemma 17 in Appendix D, and the second inequality is according to $m+1<2\left\lfloor\frac{m+1}{2}+1\right\rfloor=2(k+1)$, $h(\beta)=\frac{2\beta^{3/2}-\beta_0^{3/2}}{\beta^3}$ is a decreasing function when $\beta>\beta_0$. □

Finally, we obtain the PIFO lower bound complexity for finite-sum $(0,\mu)$-convex-concave minimax problem.

Theorem 2. *Suppose that*

$$\varepsilon\leq\frac{(L^2-2\mu^2)R_x^2}{576n^2\mu},\text{ and }m=\left\lfloor\frac{R_x}{6n}\sqrt{\frac{L^2-2\mu^2}{\mu\varepsilon}}\right\rfloor-3.$$

In order to find $(\hat{\mathbf{x}},\hat{\mathbf{y}})$ such that $\mathbb{E}\left(\phi(\hat{\mathbf{x}})-\psi(\hat{\mathbf{y}})\right)<\varepsilon$, the PIFO algorithm $\mathcal{A}$ needs at least $\Omega\left(n+\frac{R_xL}{\sqrt{\mu\varepsilon}}\right)$ queries.

Proof. Note that $M\triangleq\left\lfloor\frac{m+1}{2}\right\rfloor=\left\lfloor\frac{R_x}{12n}\sqrt{\frac{L^2-2\mu^2}{\mu\varepsilon}}\right\rfloor-1\geq 1$. Following Lemma 8, we have

$$\min_{\mathbf{x}\in\mathcal{X}'\cap\mathcal{F}_M}\phi(\mathbf{x})-\max_{\mathbf{y}\in\mathcal{G}_M}\psi(\mathbf{y})\geq\frac{(L^2-2\mu^2)R_x^2}{16n^2\mu(M+1)^2}\geq 9\varepsilon,$$

where the last inequality is due to $M+1\leq\frac{R_x}{12n}\sqrt{\frac{L^2-2\mu^2}{\mu\varepsilon}}$. Hence, following from Lemma 6 with $H(\mathbf{x},\mathbf{y})=\phi(\mathbf{x})-\psi(\mathbf{y})$, for $N=nM/2$, we know that

$$\min_{t\leq N}\mathbb{E}\left(\phi(\hat{\mathbf{x}})-\psi(\hat{\mathbf{y}})\right)\geq\varepsilon.$$

Therefore, in order to find suboptimal solution $(\hat{\mathbf{x}},\hat{\mathbf{y}})$ such that $\mathbb{E}\left(\phi(\hat{\mathbf{x}})-\psi(\hat{\mathbf{y}})\right)<\varepsilon$, algorithm $\mathcal{A}$ needs at least N PIFO queries, where

$$N=\frac{n}{2}\left(\left\lfloor\frac{R_x}{12n}\sqrt{\frac{L^2-2\mu^2}{\mu\varepsilon}}\right\rfloor-1\right)=\Omega\left(n+\frac{R_xL}{\sqrt{\mu\varepsilon}}\right).$$

□

We can also provide the lower bound $\Omega(n)$ if $\varepsilon < LR_x^2/4$ (see Lemma 21 in Appendix F) and an improved result in convex-strongly-concave case which is formally presented in Corollary 1.

Corollary 1. *For any PIFO algorithm $\mathcal{A}$ and any $L, \mu, R_x, n, \varepsilon$ such that $L/\mu \geq \sqrt{2}, R_x > 0, n \geq 2$ and $\varepsilon \leq \min\{\frac{LR_x^2}{4}, \frac{(L^2-2\mu^2)R_x^2}{576n^2\mu}\}$, there exist a dimension $m = \mathcal{O}\left(1 + \frac{R_xL}{n\sqrt{\mu\varepsilon}}\right)$ and n L-smooth and $(0,\mu)$-convex-concave functions $\{f_i : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}\}_{i=1}^n$. In order to find ε-suboptimal solution to the problem $\min_{\|\mathbf{x}\|_2 \leq R_x} \max_{\mathbf{y}} \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x}, \mathbf{y})$, algorithm $\mathcal{A}$ needs at least $\Omega\left(n + R_xL/\sqrt{\mu\varepsilon}\right)$ queries to h_f.*

4.3. General Convex-Concave Case

The analysis for general convex-concave case is similar to the one of Section 4.2. We consider the following functions.

Definition 10. *For fixed L, R_x, R_y and n such that $L, R_x, R_y > 0, n \geq 2$, let $\boldsymbol{\lambda}_C = \left(\frac{L}{2}, 0, 0, \frac{LR_y}{2\sqrt{m}}\right)$. Define functions $f_{C,i} : \mathbb{R}^{2m} \to \mathbb{R}$ for $i = 1, \dots, n$ as*

$$f_{C,i}(\mathbf{x}, \mathbf{y}) = r_i\left(\mathbf{x}, \mathbf{y}; \boldsymbol{\lambda}_C, m, 1\right)$$

and the minimax problem

$$\min_{\mathbf{x}\in\mathcal{X}'} \max_{\mathbf{y}\in\mathcal{Y}'} F_C(\mathbf{x}, \mathbf{y}) \triangleq \frac{1}{n}\sum_{i=1}^n f_{C,i}(\mathbf{x}, \mathbf{y}), \qquad (12)$$

where $\mathcal{X}' = \{\mathbf{x} : \|\mathbf{x}\|_2 \leq R_x\}$ and $\mathcal{Y}' = \{\mathbf{y} : \|\mathbf{y}\|_2 \leq R_y\}$.

We can prove each component function $f_{\mathrm{C},i}$ is L-smooth and convex-concave by Lemma 1.

The following lemma helps us to establish the lower bound with respect to the primal dual gap.

Lemma 9. *Let $\phi_C(\mathbf{x}) \triangleq \max_{\mathbf{y}\in\mathcal{Y}'} F_C(\mathbf{x}, \mathbf{y})$ and $\psi_C \triangleq \min_{\mathbf{x}\in\mathcal{X}'} F_C(\mathbf{x}, \mathbf{y})$. Then for $1 \leq k = \lfloor (m-1)/2 \rfloor$, we have*

$$\min_{\mathbf{x}\in\mathcal{X}'\cap\mathcal{F}_k} \phi_C(\mathbf{x}) - \max_{\mathbf{y}\in\mathcal{Y}'\cap\mathcal{G}_k} \psi_C \geq \frac{LR_xR_y}{2\sqrt{2}n(k+1)}.$$

Proof. By closed-form expression of $\min_{\mathbf{x}\in\mathcal{X}'\cap\mathcal{F}_k} \phi_{\mathrm{C}}(\mathbf{x})$ and $\max_{\mathbf{y}\in\mathcal{Y}'\cap\mathcal{G}_k} \psi_{\mathrm{C}}(\mathbf{y})$ from Lemma 19 in Appendix D, we know that

$$\begin{aligned}&\min_{\mathbf{x}\in\mathcal{X}'\cap\mathcal{F}_k} \phi_{\mathrm{C}}(\mathbf{x}) - \max_{\mathbf{y}\in\mathcal{Y}'\cap\mathcal{G}_k} \psi_{\mathrm{C}}(\mathbf{y}) \\ =& \frac{LR_xR_y}{2n\sqrt{m(k+1)}} \geq \frac{LR_xR_y}{2\sqrt{2}n(k+1)}.\end{aligned}$$

□

Then, we obatin a PIFO lower bound complexity for general finite-sum convex-concave minimax problem.

Theorem 3. *Suppose that*

$$\varepsilon \leq \frac{LR_xR_y}{36\sqrt{2}n}, \text{ and } m = \left\lfloor \frac{LR_xR_y}{9\sqrt{2}n\varepsilon} \right\rfloor - 1.$$

In order to find $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ such that $\mathbb{E}\left(\phi_C(\hat{\mathbf{x}}) - \psi_C(\hat{\mathbf{y}})\right) < \varepsilon$, the PIFO algorithm $\mathcal{A}$ needs at least $\Omega\left(n + \frac{R_xL}{\sqrt{\mu\varepsilon}}\right)$ queries.

Proof. Let $M \triangleq \lfloor (m-1)/2 \rfloor = \left\lfloor \frac{LR_xR_y}{18\sqrt{2}n\varepsilon} \right\rfloor - 1 \geq 1$. Following Lemma 9, we have

$$\begin{aligned}&\min_{\mathbf{x}\in\mathcal{X}'\cap\mathcal{F}_M} \phi_{\mathrm{C}}(\mathbf{x}) - \max_{\mathbf{y}\in\mathcal{Y}'\cap\mathcal{G}_M} \psi_{\mathrm{C}}(\mathbf{y}) \\ \geq& \frac{LR_xR_y}{2\sqrt{2}n\lfloor (m+1)/2 \rfloor} \geq \frac{LR_xR_y}{\sqrt{2}n(m+1)} \geq 9\varepsilon.\end{aligned}$$

Hence, following from Lemma 6 with $H(\mathbf{x}, \mathbf{y}) = \phi_{\mathrm{C}}(\mathbf{x}) - \psi_{\mathrm{C}}(\mathbf{y})$, for $N = nM/2$, we know that

$$\min_{t\leq N} \mathbb{E}\left(\phi_{\mathrm{C}}(\mathbf{x}_t) - \psi_{\mathrm{C}}(\mathbf{y}_t)\right) \geq \varepsilon.$$

Therefore, in order to find an approximate solution $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ such that $\mathbb{E}\left(\phi_{\mathrm{C}}(\hat{\mathbf{x}}) - \psi_{\mathrm{C}}(\hat{\mathbf{y}})\right) < \varepsilon$, the algorithm $\mathcal{A}$ needs at least N PIFO queries, where

$$N = \frac{n}{2}\left(\left\lfloor \frac{LR_xR_y}{18\sqrt{2}n\varepsilon} \right\rfloor - 1\right) = \Omega\left(n + \frac{LR_xR_y}{\varepsilon}\right).$$

□

Note that Theorem 3 requires the condition $\varepsilon \leq \mathcal{O}(L/n)$ to obtain the desired lower bound. In fact, this assumption can be relaxed into $\varepsilon \leq \mathcal{O}(L)$ and we show the more general result formally in Corollary 2.

Corollary 2. *For any PIFO algorithm $\mathcal{A}$ and any $L, R_x, R_y, n, \varepsilon$ such that $L, R_x, R_y > 0, \varepsilon \leq LR_xR_y/4$ and $n \geq 2$, there exist a dimension $m = \mathcal{O}\left(1 + \frac{LR_xR_y}{n\varepsilon}\right)$ and n L-smooth and convex-concave functions $\{f_i : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}\}_{i=1}^n$. In order to find ε-suboptimal solution to the problem $\min_{\|\mathbf{x}\|_2\leq R_x} \max_{\|\mathbf{y}\|_2\leq R_y} \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x}, \mathbf{y})$, $\mathcal{A}$ needs at least $\Omega\left(n + LR_xR_y/\varepsilon\right)$ queries to h_f.*

5. Comparison with Related Work

For deterministic convex optimization, Nesterov (2013) introduced a type of quadratic functions based on matrix $\mathbf{A}(m, \omega)$ to analyze the lower bound of gradient based algorithms. Lan & Zhou (2017) considered the first order stochastic algorithm for finite-sum convex optimization. They constructed a block diagonal matrix by aggregating several ones in the form of $\mathbf{A}(m, \omega)$ to obtain a tight lower bound. Zhou & Gu (2019) extended the results to more general cases, including sum-of-nonconvex problem and nonconvex optimization. Woodworth & Srebro (2016) designed a type of adversary constructions to analyze finite-sum convex optimization which is also valid for stochastic proximal point iteration.

Ouyang & Xu (2018) first studied the lower bound complexity of first order algorithms for the convex-concave minimax problem. They constructed a class of bilinear functions based on the formulation (3). Recently, Zhang et al. (2019) established a lower bound for strongly-convex-strongly-concave objective functions. However, both of them (Ouyang & Xu, 2018; Zhang et al., 2019) do not cover the stochastic optimization algorithms, which are very popular in machine learning applications.

Our proposed lower bounds analysis framework is the first one which considers the finite-sum minimax problem for PIFO algorithms. Our construction is based on the decomposition of matrix $\mathbf{B}(m, \omega)$ as formulation (4) in Section 3.1. This strategy is quite different from previous art and it provides a very concise analysis for the query of proximal incremental first-order oracle.

6. Conclusion

In this paper, we have studied lower bounds of PIFO algorithms for finite-sum convex-concave minimax optimization problems. We have proposed a novel construction framework, which is very useful to the analysis of stochastic proximal point algorithms. With this framework, we have demonstrated the lower bounds of PIFO algorithms in strongly-convex-strongly-concave case, convex-strongly-concave case and general convex-concave case.

There are still some open problems. Although SPDC matches our lower bound in a specific minimax problem, the upper bound in the general strongly-convex-strongly-concave case remains a $\sqrt{n}$ gap. Furthermore, to the best of our knowledge, there is no stochastic optimization algorithm that could match our lower bounds for convex-strongly-concave and general convex-concave cases. It would be interesting to devise more efficient algorithms for these settings or improve our lower bounds further. It is also possible to use our framework to address the lower bounds of minimax problems without the convex-concave assumption.

Acknowledgments

Xie and Zhang have been supported by the National Key Research and Development Project of China (No. 2018AAA0101004), Beijing Natural Science Foundation (Z190001), and Beijing Academy of Artificial Intelligence (BAAI).

References

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

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

Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. *Robust optimization*, volume 28. Princeton University Press, 2009.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. Lower bounds for finding stationary points I. *arXiv preprint:1710.11606*, 2017.

Carmon, Y., Jin, Y., Sidford, A., and Tian, K. Variance reduction for matrix games. In *Advances in Neural Information Processing Systems*, pp. 11381–11392, 2019.

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

Chambolle, A. and Pock, T. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, 159(1-2):253–287, 2016.

Chen, G. H. and Rockafellar, R. T. Convergence rates in forward–backward splitting. *SIAM Journal on Optimization*, 7(2):421–444, 1997.

Defazio, A. A simple practical accelerated method for finite sums. In *NIPS*, 2016.

Du, S. S., Chen, J., Li, L., Xiao, L., and Zhou, D. Stochastic variance reduction methods for policy evaluation. In *ICML*, 2017.

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

Ibrahim, A., Azizian, W., Gidel, G., and Mitliagkas, I. Linear lower bounds and conditioning of differentiable games. *arXiv preprint arXiv:1906.07300*, 2019.

Joachims, T. A support vector method for multivariate performance measures. In *ICML*, 2005.

Korpelevich, G. Extragradient method for finding saddle points and other problems. *Matekon*, 13(4):35–49, 1977.

Lan, G. and Zhou, Y. An optimal randomized incremental gradient method. *Mathematical programming*, pp. 1–49, 2017.

Lin, H., Mairal, J., and Harchaoui, Z. Catalyst acceleration for first-order convex optimization: from theory to practice. *Journal of Machine Learning Research*, 18(212): 1–54, 2018.

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

Mokhtari, A., Ozdaglar, A., and Pattathil, S. Proximal point approximations achieving a convergence rate of $O(1/k)$ for smooth convex-concave saddle point problems: Optimistic gradient and extra-gradient methods. *arXiv preprint:1906.01115*, 2019a.

Mokhtari, A., Ozdaglar, A., and Pattathil, S. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. *arXiv preprint:1901.08511*, 2019b.

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Ouyang, Y. and Xu, Y. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *arXiv preprint:1808.02901*, 2018.

Palaniappan, B. and Bach, F. Stochastic variance reduction methods for saddle-point problems. In *NIPS*, 2016.

Shen, Z., Mokhtari, A., Zhou, T., Zhao, P., and Qian, H. Towards more efficient stochastic decentralized learning: Faster convergence and sparse communication. In *ICML*, 2018.

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

Thekumparampil, K. K., Jain, P., Netrapalli, P., and Oh, S. Efficient algorithms for smooth minimax optimization. In *NIPS*, 2019.

Woodworth, B. and Srebro, N. Tight complexity bounds for optimizing composite objectives. In *NIPS*, 2016.

Yan, Y., Xu, Y., Lin, Q., Zhang, L., and Yang, T. Stochastic primal-dual algorithms with faster convergence than $O(1/\sqrt{T})$ for problems without bilinear structure. *arXiv preprint arXiv:1904.10112*, 2019.

Ying, Y., Wen, L., and Lyu, S. Stochastic online AUC maximization. In *NIPS*, 2016.

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

Zhang, Y. and Xiao, L. Stochastic primal-dual coordinate method for regularized empirical risk minimization. *The Journal of Machine Learning Research*, 18(1):2939–2980, 2017.

Zhou, D. and Gu, Q. Lower bounds for smooth nonconvex finite-sum optimization. In *ICML*, 2019.