Journal of Machine Learning Research 25 (2024) 1-86 Submitted 3/21; Revised 1/24; Published 1/24

Lower Complexity Bounds of Finite-Sum Optimization Problems: The Results and Construction

Yuze Han HANYUZE97@PKU.EDU.CN
School of Mathematical Sciences
Peking University
Beijing, China

Guangzeng Xie SMSXGZ@PKU.EDU.CN
Academy for Advanced Interdisciplinary Studies
Peking University
Beijing, China

Zhihua Zhang ZHZHANG@MATH.PKU.EDU.CN
School of Mathematical Sciences
Peking University
Beijing, China

Editor: Francis Bach

Abstract

In this paper we study the lower complexity bounds for finite-sum optimization problems, where the objective is the average of n individual component functions. We consider a so-called proximal incremental first-order oracle (PIFO) algorithm, which employs the individual component function's gradient and proximal information provided by PIFO to update the variable. To incorporate loopless methods, we also allow the PIFO algorithm to obtain the full gradient infrequently. We develop a novel approach to constructing the hard instances, which partitions the tridiagonal matrix of classical examples into n groups. This construction is friendly to the analysis of PIFO algorithms. Based on this construction, we establish the lower complexity bounds for finite-sum minimax optimization problems when the objective is convex-concave or nonconvex-strongly-concave and the class of component functions is L-average smooth. Most of these bounds are nearly matched by existing upper bounds up to log factors. We also derive similar lower bounds for finite-sum minimization problems as previous work under both smoothness and average smoothness assumptions. Our lower bounds imply that proximal oracles for smooth functions are not much more powerful than gradient oracles.

Keywords: minimax optimization, lower bound, proximal incremental first-order oracle, finite-sum optimization

1. Introduction

We consider the following 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 the feasible sets $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ and $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$ are closed and convex. This formulation contains several popular machine learning applications such as matrix games (Carmon et al., 2019, 2020b; Ibrahim et al., 2020), regularized empirical risk minimization (Zhang and 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; Dai et al., 2018).

A popular approach for solving minimax problems is the first-order algorithm which iterates with gradient and proximal point operation (Chambolle and Pock, 2011, 2016; Mokhtari et al., 2020b,a; Thekumparampil et al., 2019; Luo et al., 2019). Along this line, Zhang et al. (2022) and Ibrahim et al. (2020) presented tight lower bounds for solving strongly-convex-strongly-concave minimax problems by first-order algorithms. Ouyang and Xu (2021) studied a more general case that the objective function is only convex-concave. However, these analyses (Ouyang and Xu, 2021; Zhang et al., 2022; Ibrahim et al., 2020) do not consider the specific finite-sum structure as in Problem (1). They only considered the deterministic first-order algorithms which are based on the full gradient and exact proximal point iteration.

In big data regimes, the number of components n in Problem (1) could be very large and we would like to devise randomized optimization algorithms that avoid accessing the full gradient frequently. For example, Palaniappan and Bach (2016) used stochastic variance reduced gradient (SVRG) algorithms to solve Problem (1). Similar to convex optimization, one can accelerate it by catalyst (Lin et al., 2018; Yang et al., 2020) and proximal point techniques (Defazio, 2016; Luo et al., 2019). Note that SVRG is a double-loop algorithm, where the full gradient is calculated periodically with a constant interval. There are also some loopless algorithms where a coin flip decides whether to calculate the full gradient at each iteration (Alacaoglu and Malitsky, 2022; Luo et al., 2021). Although randomized algorithms are widely used for solving minimax problems, the study of their lower bounds is still open. All of the existing lower bound analysis focuses on convex or nonconvex minimization problems (Agarwal and Bottou, 2015; Woodworth and Srebro, 2016; Arjevani and Shamir, 2016; Lan and Zhou, 2018; Hannah et al., 2018; Fang et al., 2018).

This paper considers randomized PIFO algorithms for solving Problem (1), which are formally defined in Definition 11. These algorithms have access to the *proximal incremental first-order oracle* (PIFO)

$$h_{f_i}^{\mathrm{PIFO}}(\mathbf{x}, \mathbf{y}, \gamma) \triangleq \left[f_i(\mathbf{x}, \mathbf{y}), \nabla_{\mathbf{x}} f_i(\mathbf{x}, \mathbf{y}), -\nabla_{\mathbf{y}} f_i(\mathbf{x}, \mathbf{y}), \mathrm{prox}_{f_i}^{\gamma}(\mathbf{x}, \mathbf{y})\right], \tag{2}$$

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

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

Compared with *incremental first-order oracle* (IFO), which is defined as $h_{f_i}^{\mathrm{IFO}}(\mathbf{x}, \mathbf{y}) \triangleq [f_i(\mathbf{x}, \mathbf{y}), \nabla_{\mathbf{x}} f_i(\mathbf{x}, \mathbf{y}), -\nabla_{\mathbf{y}} f_i(\mathbf{x}, \mathbf{y})]$, PIFO additionally provides the proximal oracle of the component function. To incorporate loopless methods, we also allow PIFO algorithms to access the full gradient infrequently with the interval obeying geometric distributions.

We consider the general setting where $f(\mathbf{x}, \mathbf{y})$ is L-smooth and (μ_x, μ_y)-convex-concave, i.e., the function $f(\cdot, \mathbf{y}) - \frac{\mu_x}{2}\|\cdot\|_2^2$ is convex for any $\mathbf{y} \in \mathcal{Y}$ and the function $-f(\mathbf{x}, \cdot) - \frac{\mu_y}{2}\|\cdot\|_2^2$

is convex for any $\mathbf{x} \in \mathcal{X}$. When $\mu_x, \mu_y \geq 0$, our goal is to find an ε-suboptimal solution $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ to Problem (1) such that the primal-dual gap is less than ε, i.e.,

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

On the other hand, when $\mu_x < 0, \mu_y > 0$, $f(\mathbf{x}, \mathbf{y})$ is called a nonconvex-strongly-concave function, which has been widely studied in Rafique et al. (2022); Lin et al. (2020); Ostrovskii et al. (2021); Luo et al. (2020). In this case, our goal is instead to find an ε-stationary point $\hat{\mathbf{x}}$ of $\phi_f(\mathbf{x}) \triangleq \max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y})$, which is defined as

$$\|\nabla \phi_f(\hat{\mathbf{x}})\|_2 < \varepsilon.$$

It is worth noting that by setting the feasible set of $\mathbf{y}$ as a singleton, the minimax problem becomes a minimization problem. Then we can omit the dependence of f on $\mathbf{y}$ and rewrite the function as $f(\mathbf{x})$ with some abuse of notation. When $f(\mathbf{x})$ is convex, our goal is to find an ε-suboptimal solution $\hat{\mathbf{x}}$ such that $f(\hat{\mathbf{x}}) - \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) < \varepsilon$, while when $f(\mathbf{x})$ is nonconvex, our goal is to find an ε-stationary point $\hat{\mathbf{x}}$ such that $\|\nabla f(\hat{\mathbf{x}})\|_2 < \varepsilon$.

1.1 Contributions

Our contributions are summarized as follows.

1. We propose a novel construction framework to analyze lower complexity bounds for finite-sum optimization problems. Different from previous work, we decompose the classical tridiagonal matrix in Nesterov (2013) into n groups and each component function is defined in terms of only one group. Such a construction facilitates the analysis for both IFO and PIFO algorithms (see Definition 11).

2. We establish the lower complexity bounds for finite-sum minimax problems when f is convex-concave or nonconvex-strongly-concave and $\{f_i\}_{i=1}^n$ is L-average smooth (see Definition 2). When f is convex-concave, our lower bounds nearly match existing upper bounds up to log factors. The results are summarized in Table 1.[1]

3. For finite-sum minimization problems, we derive similar lower bounds as Woodworth and Srebro (2016); Hannah et al. (2018); Zhou and Gu (2019) when each f_i is L-smooth or $\{f_i\}_{i=1}^n$ is L-average smooth. The results are summarized in Tables 3 and 4 in Section 6. Compared to previous work, our framework provides more intuition about the optimizing process and requires fewer dimensions to construct the hard instances.

4. For most cases, our lower bounds are nearly matched by IFO algorithms. This implies that the proximal oracles for smooth functions are not much more powerful than gradient oracles, which is consistent with the observation in Woodworth and Srebro (2016).

1. The work of Zhang et al. (2021) appeared on arXiv during the review process of our work.

Cases	Upper or Lower bounds	References
$\mu_x > 0, \mu_y > 0$	$\tilde{\mathcal{O}}\left(\sqrt{n\left(\sqrt{n}+\kappa_x\right)\left(\sqrt{n}+\kappa_y\right)}\log(1/\varepsilon)\right)$	Luo et al. (2021)
	$\Omega\left(\sqrt{n\left(\sqrt{n}+\kappa_x\right)\left(\sqrt{n}+\kappa_y\right)}\log(1/\varepsilon)\right)$	Theorem 22
$\mu_x = 0, \mu_y > 0$	$\tilde{\mathcal{O}}\Big(\big(n + R_x n^{3/4}\sqrt{\frac{L}{\varepsilon}} + R_x\sqrt{\frac{nL\kappa_y}{\varepsilon}} + n^{3/4}\sqrt{\kappa_y}\big)\log\left(\frac{1}{\varepsilon}\right)\Big)$	Luo et al. (2021)
	$\Omega\Big(n + R_x n^{3/4}\sqrt{\frac{L}{\varepsilon}} + R_x\sqrt{\frac{nL\kappa_y}{\varepsilon}} + n^{3/4}\sqrt{\kappa_y}\log\left(\frac{1}{\varepsilon}\right)\Big)$	Theorem 23
$\mu_x = 0, \mu_y = 0$	$\tilde{\mathcal{O}}\Big(\big(n + \frac{\sqrt{n}LR_xR_y}{\varepsilon} + (R_x+R_y)n^{3/4}\sqrt{\frac{L}{\varepsilon}}\big)\log\left(\frac{1}{\varepsilon}\right)\Big)$	Luo et al. (2021)
	$\Omega\Big(n + \frac{\sqrt{n}LR_xR_y}{\varepsilon} + (R_x+R_y)n^{3/4}\sqrt{\frac{L}{\varepsilon}}\Big)$	Theorem 24
$\mu_x < 0, \mu_y > 0$, $\kappa_y = \Omega(n)$	$\tilde{\mathcal{O}}\left((n + n^{3/4}\sqrt{\kappa_y})\Delta L\varepsilon^{-2}\right)$	Zhang et al. (2021)
	$\Omega\left(n + \sqrt{n\kappa_y}\Delta L\varepsilon^{-2}\right)$	Theorem 25; Zhang et al. (2021)

Table 1: Upper and lower bounds under the assumption that $\{f_i\}_{i=1}^n$ is L-average smooth and f is (μ_x, μ_y)-convex-concave. When $\mu_x \geq 0$ and $\mu_y \geq 0$, the goal is to find an ε-suboptimal solution with $\mathrm{diam}(\mathcal{X}) \leq 2R_x, \mathrm{diam}(\mathcal{Y}) \leq 2R_y$. And when $\mu_x < 0$, the goal is to find an ε-stationary point of the function $\phi_f(\mathbf{x}) \triangleq \max_{\mathbf{y}\in\mathcal{Y}} f(\cdot, \mathbf{y})$ with $\Delta = \phi_f(\mathbf{x}_0) - \min_{\mathbf{x}} \phi_f(\mathbf{x})$ and $\mathcal{X} = \mathbb{R}^{d_x}, \mathcal{Y} = \mathbb{R}^{d_y}$. The condition numbers are defined as $\kappa_x = L/\mu_x$ and $\kappa_y = L/\mu_y$ when $\mu_x, \mu_y > 0$.

1.2 Related Work

To contextualize our results, we provide some background related to our topic.

Lower bounds for finite-sum minimization problems. There has been extensive study on this topic. Agarwal and Bottou (2015) established the lower bound $\Omega(n + \sqrt{n(\kappa-1)}\log(1/\varepsilon))$ when each component is L-smooth and their average is μ-strongly convex by a resisting oracle construction, where $\kappa = L/\mu$ is the condition number. However, their lower bound only applies to deterministic algorithms. Lan and Zhou (2018) obtained the lower bound $\Omega((n + \sqrt{n\kappa})\log(1/\varepsilon))$ for randomized incremental gradient methods, but their bound does not apply to multi-loop methods such as SVRG (Johnson and Zhang, 2013) and SARAH (Nguyen et al., 2017). Woodworth and Srebro (2016) provided the lower bound $\Omega(n + \sqrt{n\kappa}\log(1/\varepsilon))$ for any randomized algorithms using gradient and proximal oracles. Moreover, when the objective is only convex, their lower bound is $\Omega(n + \sqrt{nL/\varepsilon})$. Arjevani and Shamir (2016) established a similar lower bound for the strongly convex case and their bound also applies to stochastic coordinate-descent methods. Hannah et al. (2018) improved this bound to $\Omega\big(\frac{n\log(1/\varepsilon)}{(1+\log(n/\kappa))_+}\big)$ when $\kappa = \mathcal{O}(n)$. Zhou and Gu (2019) proved lower bounds $\Omega(n + n^{3/4}\sqrt{\kappa}\log(1/\varepsilon))$ and $\Omega(n + n^{3/4}\sqrt{L/\varepsilon})$ for the strongly convex and convex case respectively under the weaker condition that the class of component function is L-average smooth.

When the objective is nonconvex, Fang et al. (2018) proved the lower bound $\Omega(L\sqrt{n}/\varepsilon^2)$ for $\varepsilon = \mathcal{O}(\sqrt{L}/n^{1/4})$ under the average smooth condition. Li et al. (2021) improved the bound to $\Omega(n + L\sqrt{n}/\varepsilon^2)$ for an arbitrary ε. Under a more refined condition that objective is μ-weakly convex (see Definition 4), Zhou and Gu (2019) established the lower bound to $\Omega(1/\varepsilon^2 \min\{n^{3/4}\sqrt{L\mu}, \sqrt{n}L\})$ when ε is sufficiently small. They also provided the lower bound $\Omega(1/\varepsilon^2 \min\{\sqrt{nL\mu}, L\})$ when each component is L-smooth.

Upper bounds for finite-sum minimax problems. For Problem (1), if $\mu_x, \mu_y \geq 0$ and each f_i is L-smooth, the best known upper bound is $\mathcal{O}\left((n + \sqrt{n}(\kappa_x+\kappa_y)) \log(1/\varepsilon)\right)$ (Carmon et al., 2019; Luo et al., 2019). Furthermore, if each f_i has L-cocoercive gradient, which is a stronger assumption than smoothness, Chavdarova et al. (2019) provided an upper bound $\mathcal{O}\left((n + \kappa_x + \kappa_y) \log(1/\varepsilon)\right)$. If $\{f_i\}_{i=1}^n$ is L-average smooth, Accelerated SVRG (Palaniappan and Bach, 2016) attained the upper bound $\tilde{\mathcal{O}}\left((n + \sqrt{n}(\kappa_x + \kappa_y)) \log(1/\varepsilon)\right)$ and Alacaoglu and Malitsky (2022) obtained the bound $\mathcal{O}\left((n + \sqrt{n}(\kappa_x + \kappa_y)) \log(1/\varepsilon)\right)$. Then Luo et al. (2021) improved this bound to $\tilde{\mathcal{O}}(\sqrt{n(\sqrt{n} + \kappa_x)(\sqrt{n} + \kappa_y)} \log(1/\varepsilon))$ by catalyst acceleration. The same technique was also employed to derive lower bounds for the convex-strongly-concave case where $\mu_x = 0, \mu_y > 0$ (Yang et al., 2020; Luo et al., 2021).

For the convex-concave case ($\mu_x = \mu_y = 0$), Carmon et al. (2019) established the upper bound $\mathcal{O}(n + \sqrt{n}L/\varepsilon)$ under the smoothness assumption, while Alacaoglu and Malitsky (2022) developed the same upper bound under the average smoothness assumption. Luo et al. (2021) still used the catalyst acceleration and derived a similar bound.

In terms of the nonconvex-strongly-concave case ($\mu_x < 0, \mu_y > 0$), Luo et al. (2020) proposed an upper bound $\tilde{\mathcal{O}}\left(n + \min\{\sqrt{n}\kappa_y^2, \kappa_y^2 + n\kappa_y\}\varepsilon^{-2}\right)$, while Zhang et al. (2021) developed an upper bound $\tilde{\mathcal{O}}\left((n + n^{3/4}\sqrt{\kappa_y})L\varepsilon^{-2}\right)$. The latter is better when $n = \mathcal{O}(\kappa^4)$. We emphasize that both results are under the average smoothness assumption.

Loopless methods. Variance-reduced methods designed for finite-sum minimization problems such as SVRG (Johnson and Zhang, 2013), Katyusha (Allen-Zhu, 2018a) and SARAH (Nguyen et al., 2017) have a double-loop design where the full gradient needs to be calculated periodically. Recently, many researchers have aimed to study their loopless variants or devise new loopless methods such that whether to access the full gradient depends on a coin toss with a small head probability. Equivalently speaking, the inner loop size obeys the geometric distribution with a small success probability. Such a design facilitates theoretical analysis without deteriorating the convergence rates. For example, loopless SVRG (L-SVRG) was first proposed in Hofmann et al. (2015) and then further analyzed in Kovalev et al. (2020); Qian et al. (2021) together with loopless Katyusha (L-Katyusha). Loopless SARAH (L2S) was developed in Li et al. (2020). Other loopless methods include but are not limited to KatyushaX (Allen-Zhu, 2018b), PAGE (Li et al., 2021) and ANITA (Li, 2021). For finite-sum minimax problems, there are also many loopless methods (Loizou et al., 2020; Alacaoglu and Malitsky, 2022; Beznosikov et al., 2023).

The proximal oracle. The proximal oracle provides more information than the gradient oracle and has been used in algorithm design (Shalev-Shwartz and Zhang, 2013; Defazio, 2016; Lan and Zhou, 2018; Luo et al., 2019). Compared with catalyst acceleration, employing proximal oracles would neither increase the number of loops nor induce additional parameter tuning. When each component function enjoys a simple form (Zhang and Xiao, 2017; Du et al., 2017; Lan and Zhou, 2018; Carmon et al., 2019), the proximal operator can be computed efficiently. In terms of the power of proximal oracles, Woodworth and Srebro (2016) have shown that for smooth functions, the gradient oracle is sufficient for the optimal rate. As a comparison, for nonsmooth functions, having access to proximal oracles does reduce the complexity and Woodworth and Srebro (2016) presented optimal methods that improve over those only using gradient oracles.

1.3 Organization

The remainder of this paper is organized as follows. In Section 2, we introduce some necessary notation and definitions and give a concentration inequality for geometric distributions. In Section 3, we present and discuss the definition of PIFO algorithms. In Section 4, we define the optimization complexity and construct the hard instances for Problem (1). In Sections 5 and 6, we provide and analyze our lower bounds for finite-sum minimax and minimization problems respectively. Finally, in Section 7, we summarize our results and propose some future research directions.

2. Preliminaries

In this section, we present some necessary notation and definitions used in our paper and then give a concentration inequality about geometric distributions.

Notation. We denote the set $\{1, 2, \dots, n\}$ by $[n]$. $a_+ \triangleq \max\{a, 0\}$ represent the positive part of a real number. The projection operator is defined as $\mathcal{P}_{\mathcal{X}}(\mathbf{x}) \triangleq \arg\min_{\mathbf{x}' \in \mathcal{X}} \|\mathbf{x}' - \mathbf{x}\|_2$ where $\mathcal{X}$ is a convex set and $\|\cdot\|_2$ is the Euclidean norm. We use $\mathbf{0}$ for all-zero vectors and $\mathbf{e}_i$ for the unit vector with the i-th element equal to 1 and others equal to 0. Their dimensions will be specified by an additional subscript, if necessary, and otherwise are clear from the context. We use $\mathrm{Geo}(p)$ to denote the geometric distribution with success probability p, i.e., $Y \sim \mathrm{Geo}(p)$ implies $\mathbb{P}[Y = k] = (1-p)^{k-1}p$ for $0 < p \leq 1,\ k \in \{1, 2, \dots\}$. Finally, we use the notation $\mathcal{O}(\cdot), \Omega(\cdot), \Theta(\cdot)$ to hide absolute constants that do not depend on any problem parameter, and notation $\tilde{\mathcal{O}}(\cdot)$ to hide absolute constants and log factors.

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

$$\|\nabla \varphi(\mathbf{x}_1) - \nabla \varphi(\mathbf{x}_2)\|_2 \leq L \|\mathbf{x}_1 - \mathbf{x}_2\|_2 .$$

Definition 2 *For a class of differentiable functions $\{\varphi_i(\mathbf{x}) : \mathcal{X} \to \mathbb{R}\}_{i=1}^n$ and $L > 0$, $\{\varphi_i\}_{i=1}^n$ is said to be L-average smooth if for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$, we have*

$$\frac{1}{n} \sum_{i=1}^{n} \|\nabla \varphi_i(\mathbf{x}_1) - \nabla \varphi_i(\mathbf{x}_2)\|_2^2 \leq L^2 \|\mathbf{x}_1 - \mathbf{x}_2\|_2^2 .$$

The assumption of average smoothness is widely used in many finite-sum optimizations (Zhou and Gu, 2019; Fang et al., 2018; Zhou and Gu, 2019; Alacaoglu and Malitsky, 2022).

Now we discuss the relationship between smoothness and average smoothness. For a class of differentiable functions $\{\varphi_i(\mathbf{x}) : \mathcal{X} \to \mathbb{R}\}_{i=1}^n$ and their average $\bar{\varphi}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n \varphi_i(\mathbf{x})$, we have the following result

$$\varphi_i \text{ is } L\text{-smooth}, \forall i \Longrightarrow \{\varphi_i\}_{i=1}^n \text{ is } L\text{-average smooth} \Longrightarrow \bar{\varphi} \text{ is } L\text{-smooth}.$$

Moreover, suppose that φ_i is L_i-smooth, $\bar{\varphi}$ is L-smooth and $\{\varphi_i\}_{i=1}^n$ is L'-average smooth, we have $L \leq L' \leq \sqrt{\frac{1}{n}\sum_{i=1}^n L_i^2}$ and $L \leq \frac{1}{n}\sum_{i=1}^n L_1$.

However, L and L' can be much smaller than L_i. For example, if $\varphi_i(\mathbf{x}) = \frac{1}{2}\left(\langle \mathbf{e}_i, \mathbf{x}\rangle\right)^2$, then we have $L_i = 1$, $L = 1/n$ and $L' = 1/\sqrt{n}$. As a result, it is more restrictive to say that each φ_i is L-smooth than to say that $\{\varphi_i\}_{i=1}^n$ is L-average smooth.

Definition 3 *For a differentiable function $\varphi(\mathbf{x})$ from $\mathcal{X}$ to $\mathbb{R}$, φ is said to be convex if for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$, we have*

$$\varphi(\mathbf{x}_2) \geq \varphi(\mathbf{x}_1) + \langle \nabla\varphi(\mathbf{x}_1), \mathbf{x}_2 - \mathbf{x}_1 \rangle .$$

Definition 4 *For a constant μ, if the function $\hat{\varphi}(\mathbf{x}) = \varphi(\mathbf{x}) - \frac{\mu}{2}\|\mathbf{x}\|_2^2$ is convex, then φ is said to be μ-strongly convex if $\mu > 0$ and φ is said to be μ-weakly convex if $\mu < 0$.*

One can check that if φ is L-smooth, then it is $(-L)$-weakly-convex.

Definition 5 *For a differentiable function $\varphi(\mathbf{x})$ from $\mathcal{X}$ to $\mathbb{R}$, we call $\hat{\mathbf{x}}$ an ε-stationary point of φ if*

$$\|\nabla\varphi(\hat{\mathbf{x}})\|_2 < \varepsilon.$$

Definition 6 *For a differentiable function $f(\mathbf{x}, \mathbf{y})$ from $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$, f is said to be convex-concave, if the function $f(\cdot, \mathbf{y})$ is convex for any $\mathbf{y} \in \mathcal{Y}$ and the function $-f(\mathbf{x}, \cdot)$ is convex for any $\mathbf{x} \in \mathcal{X}$. Furthermore, f is said to be (μ_x, μ_y)-convex-concave, if the function $f(\mathbf{x}, \mathbf{y}) - \frac{\mu_x}{2}\|\mathbf{x}\|_2^2 + \frac{\mu_y}{2}\|y\|_2^2$ is convex-concave.*

Definition 7 *We call a minimax optimization problem $\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y})$ satisfying the strong duality condition if*

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

By Sion's minimax theorem, if $\varphi(\mathbf{x}, \mathbf{y})$ is convex-concave and either $\mathcal{X}$ or $\mathcal{Y}$ is a compact set, then the strong duality condition holds.

Definition 8 *We call $(\mathbf{x}^*, \mathbf{y}^*) \in \mathcal{X} \times \mathcal{Y}$ the saddle point of $f(\mathbf{x}, \mathbf{y})$ if*

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

for all $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$.

Definition 9 *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*

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

2.1 A Concentration Inequality about Geometric Distributions

In this subsection, we introduce a concentration inequality about geometric distributions.

Lemma 10 *Let $\{Y_i\}_{i=1}^m$ be independent random variables, and Y_i follows a geometric distribution with success probability p_i. Then for $m \geq 2$, we have*

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

Lemma 10 implies that at least with a constant probability, the sum of geometric random variables is larger than a constant number, which depends on the number of variables and their success probabilities. Then we can obtain a lower bound of $\mathbb{E}\sum_{i=1}^{m} Y_i$, which is helpful to the construction in Section 4. The proof is deferred to Appendix A.

3. PIFO Algorithms

In this section, we present our definition of PIFO algorithms. We first discuss previous definitions in Section 3.1 and our formal definition is given in Section 3.2.

3.1 Discussion on Previous Definitions

In this subsection, we discuss the definitions of oracles and algorithms in previous work on the minimization problem $\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})$. With some abuse of notation, we do not distinguish the oracles for minimization problems from those for minimax problems.

IFO and PIFO. The *incremental first-order oracle* (IFO) is defined as $h_{f_i}^{\mathrm{IFO}}(\mathbf{x}) \triangleq [f_i(\mathbf{x}), \nabla f_i(\mathbf{x})]$, which takes as input a point $\mathbf{x} \in \mathcal{X}$ and a component function f_i and returns the function value and the gradient of f_i at $\mathbf{x}$. Many lower bounds for minimization optimization are based on this oracle, e.g., Agarwal and Bottou (2015); Lan and Zhou (2018); Zhou and Gu (2019). They all consider *linear-span randomized first-order algorithms*.[2] For these algorithms, the current point lies in the linear span of previous points and gradients returned by earlier IFO calls.

Woodworth and Srebro (2016) consider *proximal incremental first-order oracle* (PIFO) which is stronger than IFO and is defined as $h_{f_i}^{\mathrm{PIFO}}(\mathbf{x}, \gamma) \triangleq [f_i(\mathbf{x}), \nabla f_i(\mathbf{x}), \mathrm{prox}_{f_i}^{\gamma}(\mathbf{x})]$ with the proximal operator $\mathrm{prox}_{f_i}^{\gamma}(\mathbf{x}) \triangleq \arg\min_{\mathbf{u}}\left\{f_i(\mathbf{u}) + \frac{1}{2\gamma}\|\mathbf{x}-\mathbf{u}\|_2^2\right\}$. When f_i is convex, any $\gamma > 0$ is feasible. Different from IFO, PIFO provides global information about the function. To see this, letting $\gamma \to \infty$ yields the exact minimizer of f_i. Based on PIFO, Woodworth and Srebro (2016) consider the class of *any* randomized algorithms, a more general class than *linear-span randomized first-order algorithms*. We also emphasize that when f_i is nonconvex, γ should be sufficiently small such that $f_i(\mathbf{u}) + \frac{1}{2\gamma}\|\mathbf{x}-\mathbf{u}\|_2^2$ is a convex function of $\mathbf{u}$. Otherwise, it can be pretty hard to calculate $\mathrm{prox}_{f_i}^{\gamma}(\mathbf{x})$. Specially, if f is $(-\mu)$-weakly convex, we need to ensure $0 < \gamma < 1/\mu$.

2. The formal definition is given in Definition 3.3 in Zhou and Gu (2019). Although the results of Agarwal and Bottou (2015) do not rely on the linear span assumption, this assumption can be made without loss of generality, as shown in Appendix A of their work.

Sampling of the component function. Note that both IFO and PIFO depend on a specific component f_i. Different methods use different ways to choose the index i. Some of them, e.g., SAGA (Defazio et al., 2014), RPDG (Lan and Zhou, 2018), pick i randomly according to some distribution over $[n]$ and the full gradient is calculated only at the initial point. However, much more methods need to calculate the full gradient periodically, either with a deterministic or random interval. For multi-loop methods, e.g., SVRG (Johnson and Zhang, 2013), Katyusha (Allen-Zhu, 2018a) and SPIDER (Fang et al., 2018), the interval is predetermined, while for loopless methods, e.g., KatyushaX (Allen-Zhu, 2018b), L2S (Li et al., 2020), L-SVRG (Kovalev et al., 2020), the interval is a geometric random variable.

The lower bound of Lan and Zhou (2018) requires that the index i_t at iteration t is sampled from a predetermined distribution over $[n]$. Thus their bound does not apply to methods such as SVRG and L-SVRG. Woodworth and Srebro (2016); Zhou and Gu (2019) do not specify the way to choose i_t. As a result, their class of algorithms does include those multi-loop or loopless methods.

Arjevani and Shamir (2016) and Hannah et al. (2018) consider p-CLI algorithms equipped with the generalized first-order oracle, where the current point and the gradient can be left-multiplied by preconditioning matrices. They do not specify the way to choose i_t, either. Thus their lower bounds apply to all the methods mentioned above. Moreover, their framework can also be equipped with the steepest coordinate descent oracle to incorporate methods such as SDCA (Shalev-Shwartz, 2016).

3.2 Our Definition

In this subsection, we come back to the minimax problem (1) and formally introduce the definition of PIFO algorithms.

Recall that the PIFO has been defined in (2). For convenience, we also define the *First-order Oracle* (FO) as $h_f^{\mathrm{FO}}(\mathbf{x},\mathbf{y}) \triangleq [f(\mathbf{x},\mathbf{y}), \nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{y}), -\nabla_{\mathbf{y}} f(\mathbf{x},\mathbf{y})]$, which returns the full gradient information. Since the feasible set of Problem (1) is not necessarily the whole space, the algorithm should also have access to the projection operators $\mathcal{P}_{\mathcal{X}}$ and $\mathcal{P}_{\mathcal{Y}}$. Then we can define the PIFO algorithms we focus on in our paper.

Definition 11 *Consider a randomized PIFO algorithm $\mathcal{A}$ to solve Problem (1). Denote the point obtained by $\mathcal{A}$ after step t by $(\mathbf{x}_t,\mathbf{y}_t)$, which is generated by the following procedure.*

1. *Initialize the set $\mathcal{H}$ as $\{(\mathbf{x}_0,\mathbf{y}_0)\}$, the distribution $\mathcal{D}$ over $[n]$, a positive number $q \le c_0/n$ and set $t=1$.*
2. *Sample $i_t \sim \mathcal{D}$ and query the oracle $h_{f_{i_t}}^{\mathrm{PIFO}}$ at the current point $(\mathbf{x}_{t-1},\mathbf{y}_{t-1})$ and also at a previous point in $\{(\mathbf{x}_l,\mathbf{y}_l)\}_{0\le l<t-1}$, if necessary.*
3. *Sample a Bernoulli random variable a_t with expectation equal to q. If $a_t=1$, query the oracle h_f^{FO} at point $(\mathbf{x}_{t-1},\mathbf{y}_{t-1})$ and add $(\mathbf{x}_{t-1},\mathbf{y}_{t-1})$ to $\mathcal{H}$.*
4. *Obtain $(\tilde{\mathbf{x}}_t,\tilde{\mathbf{y}}_t)$ following the linear-span protocol*

$$
\begin{aligned}
(\tilde{\mathbf{x}}_t,\tilde{\mathbf{y}}_t) \in \operatorname{span}\big\{&(\mathbf{x}_0,\mathbf{y}_0),\dots,(\mathbf{x}_{t-1},\mathbf{y}_{t-1}), \operatorname{prox}_{f_{i_j}}^{\gamma_j}(\mathbf{x}_l,\mathbf{y}_l) \text{ for } l<j\le t,\\
&(\nabla_{\mathbf{x}} f_{i_j}(\mathbf{x}_l,\mathbf{y}_l),\mathbf{0}_{d_y}), (\mathbf{0}_{d_x}, -\nabla_{\mathbf{y}} f_{i_j}(\mathbf{x}_l,\mathbf{y}_l)) \text{ for } l<j\le t,\\
&(\nabla_{\mathbf{x}} f(\mathbf{u},\mathbf{v}),\mathbf{0}_{d_y}), (\mathbf{0}_{d_x}, -\nabla_{\mathbf{y}} f(\mathbf{u},\mathbf{v})) \text{ for } (\mathbf{u},\mathbf{v})\in\mathcal{H}\big\}.
\end{aligned}
$$

5. *Projection step:* $\mathbf{x}_t = \mathcal{P}_{\mathcal{X}}(\tilde{\mathbf{x}}_t), \mathbf{y}_t = \mathcal{P}_{\mathcal{Y}}(\tilde{\mathbf{y}}_t)$.

6. *Output* $(\mathbf{x}_t, \mathbf{y}_t)$, *or increase* t *by* 1 *and go back to step 2.*

Let $\mathscr{A}$ *be the class of all such PIFO algorithms. A PIFO algorithm becomes an IFO algorithm if it queries the IFO at step 2.*

Remark 12 *We remark on some details in our definition of PIFO algorithms.*

(i) *The random vector sequence* $\{(i_t, a_t)\}_{t\geq 1}$ *are mutually independent and each* i_t *is also independent of* a_t.

(ii) *Note that in part 2, we allow PIFO queries simultaneously at the current point and a previous point. Such a simultaneous query is commonly employed in variance-reduced methods (Johnson and Zhang, 2013; Fang et al., 2018; Zhou et al., 2020; Luo et al., 2020).*

(iii) *In part 4, we allow the algorithm to reuse previously obtained information, e.g.,* $(\nabla_{\mathbf{x}} f_{i_j}(\mathbf{x}_l, \mathbf{y}_l), \mathbf{0}_{d_y})$, $(\mathbf{0}_{d_x}, -\nabla_{\mathbf{y}} f_{i_j}(\mathbf{x}_l, \mathbf{y}_l))$ *and* $\operatorname{prox}_{f_{i_j}}^{\gamma_j}(\mathbf{x}_l, \mathbf{y}_l)$ *for* $l < j < t$. *The set* $\mathcal{H}$ *collects the points where the full gradient is calculated.*

(iv) *When* f_i *is not convex-concave,* γ *should be chosen such that* $f_i(\mathbf{u}, \mathbf{v}) + \frac{1}{2\gamma}\|\mathbf{x}-\mathbf{u}\|_2^2 - \frac{1}{2\gamma}\|\mathbf{y}-\mathbf{v}\|_2^2$ *is convex-concave w.r.t. to* $(\mathbf{u}, \mathbf{v})$.

(v) *Without loss of generality, we assume that the PIFO algorithm* $\mathcal{A}$ *starts from* $(\mathbf{x}_0, \mathbf{y}_0) = (\mathbf{0}_{d_x}, \mathbf{0}_{d_y})$ *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.*

(vi) *Let* $p_i = \mathbb{P}_{Z\sim\mathcal{D}}[Z = i]$ *for* $i \in [n]$. *The distribution* $\mathcal{D}$ *can be the uniform distribution or based on the smoothness of the component functions, e.g.,* $p_i \propto L_i$ *(Xiao and Zhang, 2014) or* $p_i \propto L_i^2$ *(Allen-Zhu, 2018b) for* $i \in [n]$, *where* L_i *is the smoothness parameter of* f_i. *We can assume that* $p_1 \leq p_2 \leq \cdots \leq p_n$ *by rearranging the component functions* $\{f_i\}_{i=1}^n$. *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 consider* $\{\hat{f}_i\}_{i=1}^n$ *and categorical distribution* $\mathcal{D}'$ *with* $\hat{f}_i \triangleq f_{s_i}$ *and* $\mathbb{P}_{Z\sim\mathcal{D}'}[Z = i] = p_{s_i}$.

Recall that by setting $\mathcal{Y}$ as a singleton, we can obtain the definition of IFO and PIFO algorithms for finite-sum minimization problems.

We emphasize that only the proximal operator of the individual component function f_i is allowed. The algorithm is not accessible to the proximal operator of the averaged function f. In practice, each f_i usually depends on a single sample and enjoys a simple form (Zhang and Xiao, 2017; Du et al., 2017; Lan and Zhou, 2018; Carmon et al., 2019). Then $\operatorname{prox}_{f_i}^{\gamma}(\mathbf{x}, \mathbf{y})$ is easy to calculate. For example, if $f_i(\mathbf{x}, \mathbf{y}) = \frac{1}{2}(\mathbf{a}_i^\top \mathbf{x})^2$, then $\operatorname{prox}_{f_i}^{\gamma}(\mathbf{x}, \mathbf{y}) = (\mathbf{I} + \gamma\,\mathbf{a}_i\mathbf{a}_i^\top)^{-1}\mathbf{x}$; if $f_i(\mathbf{x}, \mathbf{y}) = \mathbf{y}^\top \mathbf{b}_i \mathbf{a}_i^\top \mathbf{x}$, then $\operatorname{prox}_{f_i}^{\gamma}(\mathbf{x}, \mathbf{y}) = \begin{bmatrix} \mathbf{I} & \gamma\mathbf{a}_i\mathbf{b}_i^\top \\ -\gamma\mathbf{b}_i\mathbf{a}_i^\top & \mathbf{I} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}$. For these two examples, $\operatorname{prox}_{f_i}^{\gamma}(\mathbf{x}, \mathbf{y})$ can be efficiently computed by the Sherman–Morrison formula. However, computing $\operatorname{prox}_{f}^{\gamma}(\mathbf{x}, \mathbf{y})$ could be as hard as solving the original problem (1). To see this, letting $\gamma \to \infty$ yields the exact solution of problem (1).

Methods for minimization problems. One-loop methods such as SAGA (Defazio et al., 2014) and PointSAGA (Defazio, 2016) belong to PIFO algorithms. Some methods such as SVRG (Johnson and Zhang, 2013) and Katyusha (Allen-Zhu, 2018a) have two loops and the full gradient needs to be calculated at each iteration of the outer loop. Although these two-loop methods do not satisfy our definition, their loopless variants do. These loopless variants only have one loop and whether to compute the full gradient depends on a coin toss with a small head probability, i.e., q in Definition 11. Kovalev et al. (2020) have shown that loopless SVRG (L-SVRG) and loopless Katyusha (L-Katyusha) enjoy the same theoretical properties as the original methods. With a constant q, these loopless methods can also be viewed as two-loop methods with a random inner-loop size that obeys the geometric distribution with success probability q. Other loopless methods that satisfy our definition include KatyushaX (Allen-Zhu, 2018b), L2S (Li et al., 2020), PAGE (Li et al., 2021), ANITA (Li, 2021) and so on. For these methods, the order of q is usually $\Theta(1/n)$. And it suffices to set $c_0 = 2$.

Now we consider catalyst-accelerated methods. Seemingly these methods do not satisfy our definition, since they have two loops and the full gradient is calculated at each iteration of the outer loop. Nevertheless, we can slightly change them without affecting the convergence rate. Firstly, we can replace the algorithm used to solve the inner-loop subproblem, e.g., SVRG, with its loopless variant. Secondly, note that the complexity of each iteration of the outer loop is of the order $\Omega(n)$ (all the components need to be sampled at least once in the inner loop). At each iteration of the outer loop, we do not update the current point until the FO is called. In expectation, we need $\Theta(1/q)$ more steps. Thus, if we choose $q = \Theta(1/n)$, then $\Theta(1/q) = \Theta(n)$, implying that such a change makes no difference to the order of the complexity.

Methods for minimax problems. One can check SAGA (Palaniappan and Bach, 2016) and PointSAGA (Defazio, 2016) are PIFO algorithms. Although SVRG (Palaniappan and Bach, 2016) does not satisfy our definition, we believe a loopless variant can share the same convergence properties. Existing loopless methods that belong to PIFO algorithms include L-SVRHG (Loizou et al., 2020) and L-SVRE[3] (Alacaoglu and Malitsky, 2022). Moreover, similar to the analysis above, the catalyst-accelerated methods in Luo et al. (2021); Zhang et al. (2021) also satisfy our definition. For these methods, the order of q is still $\Theta(1/n)$ and we can set $c_0 = 2$.

Finally, we emphasize that all the methods analyzed above except PointSAGA (Defazio et al., 2014; Defazio, 2016) are also IFO algorithms. From the results in Table 1 and the analysis in Section 6, we find that IFO algorithms are powerful enough for smooth functions.

4. Framework of Construction

In this section, we introduce the framework of our construction to establish the lower bound for Problem (1). In Section 4.1, we give the definition of the optimization complexity. In Section 4.2, we construct the hard instances used to prove the lower bound and present some fundamental lemmas. Now we first highlight the key idea of our construction.

3. The method was renamed by Luo et al. (2021) and we adopt the new name.

Key idea. To construct the hard instance, we partition the tridiagonal matrix in Nesterov (2013) into n groups and each component function is defined in terms of only one group. Then the hard instance satisfies a variant of *zero-chain* property: starting from the origin, only when a specific component is drawn, can we increase the nonzero elements of the current point by at most 2. Meanwhile, the number of PIFO calls required to draw this component obeys the geometric distribution. Consequently, once we prove that we cannot obtain any ε-suboptimal solution or ε-stationary point unless we span all the dimensions, the lower bound on the complexity can be derived by using the concentration inequality of geometric distributions, as outlined in Lemma 10. As a comparison, previous span-based construction (Lan and Zhou, 2018; Zhou and Gu, 2019) partitions the variable. In their construction, the number of nonzero elements of the current point can increase no matter which component is drawn. A more detailed analysis is deferred to Section 6.1.

4.1 Optimization Complexity

Before presenting the definition of the optimization complexity, we first introduce the function class we consider. Define the *primal function* as $\phi_f(\mathbf{x}) = \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y})$ and the *dual function* as $\psi_f(\mathbf{y}) = \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}, \mathbf{y})$.

Function class. We develop lower bounds for PIFO algorithms that find a suboptimal solution or near stationary point of Problem (1) in the following sets.

$$\mathcal{F}_{\mathrm{CC}}(R_x, R_y, L, \mu_x, \mu_y) = \Bigg\{ f(\mathbf{x}, \mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}, \mathbf{y}) \;\Bigg|\; f\colon \mathcal{X}\times\mathcal{Y} \to \mathbb{R},\ \mathrm{diam}(\mathcal{X}) \le 2R_x, ,$$
$$\mathrm{diam}(\mathcal{Y}) \le 2R_y,\ \{f_i\}_{i=1}^n \text{ is } L\text{-average smooth}, f \text{ is } (\mu_x, \mu_y)\text{-convex-concave} \Bigg\}.$$
$$\mathcal{F}_{\mathrm{NCC}}(\Delta, L, \mu_x, \mu_y) = \Bigg\{ f(\mathbf{x}, \mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}, \mathbf{y}) \;\Bigg|\; f\colon \mathcal{X}\times\mathcal{Y} \to \mathbb{R},\ \phi(\mathbf{0}) - \inf_{\mathbf{x}\in\mathcal{X}} \phi(\mathbf{x}) \le \Delta,$$
$$\{f_i\}_{i=1}^n \text{ is } L\text{-average smooth}, f \text{ is } (-\mu_x, \mu_y)\text{-convex-concave} \Bigg\}.$$

We remark that for the second class, μ_x measures how nonconvex the function is. A natural upper bound of μ_x is L. Moreover, we do not specify the dimensions of the feasible set. That is to say, the two classes include functions defined on $\mathcal{X}\times\mathcal{Y} \subseteq \mathbb{R}^{d_x}\times\mathbb{R}^{d_y}$ with any positive integers d_x and d_y.

Optimization complexity. Then we formally define the optimization complexity.

Definition 13 *For a function f, a PIFO algorithm $\mathcal{A}$ and a tolerance $\varepsilon > 0$, the number of queries to PIFO needed by $\mathcal{A}$ to find an ε-suboptimal solution to Problem (1) or an ε-stationary point of $\phi_f(\mathbf{x})$ is defined as*

$$T(\mathcal{A}, f, \varepsilon) = \begin{cases} \inf\left\{T\in\mathbb{N} \mid \mathbb{E}\phi_f(\mathbf{x}_{\mathcal{A},T}) - \mathbb{E}\psi_f(\mathbf{y}_{\mathcal{A},T}) < \varepsilon\right\}, & \text{if } f \in \mathcal{F}_{\mathrm{CC}}(R_x, R_y, L, \mu_x, \mu_y), \\ \inf\left\{T\in\mathbb{N} \mid \mathbb{E}\left\|\nabla\phi_f(\mathbf{x}_{\mathcal{A},T})\right\|_2 < \varepsilon\right\}, & \text{if } f \in \mathcal{F}_{\mathrm{NCC}}(\Delta, L, \mu_x, \mu_y), \end{cases}$$

where $(\mathbf{x}_{\mathcal{A},T}, \mathbf{y}_{\mathcal{A},T})$ *is the point obtained by the algorithm* $\mathcal{A}$ *at time-step* $T-1$*. The optimization complexity with respect to the two function classes is defined as*[4]

$$\mathfrak{m}^{\mathrm{CC}}(\varepsilon, R_x, R_y, L, \mu_x, \mu_y) \triangleq \inf_{\mathcal{A}\in\mathscr{A}} \sup_{f\in\mathcal{F}_{\mathrm{CC}}(R_x,R_y,L,\mu_x,\mu_y)} T(\mathcal{A}, f, \varepsilon).$$
$$\mathfrak{m}^{\mathrm{NCC}}(\varepsilon, \Delta, L, \mu_x, \mu_y) \triangleq \inf_{\mathcal{A}\in\mathscr{A}} \sup_{f\in\mathcal{F}_{\mathrm{NCC}}(\Delta,L,\mu_x,\mu_y)} T(\mathcal{A}, f, \varepsilon).$$

When f is convex-concave, the functions we consider have a bounded feasible set and L-average smooth components. By Sion's minimax theorem, the strong duality condition holds. Then the primal-dual gap is a natural measurement of the optimality because this gap equals zero at the saddle point. In particular, if f is strongly-convex-strongly-concave, the saddle point $(\mathbf{x}^\star, \mathbf{y}^\star)$ is unique. Then the squared distance $\|\mathbf{x}-\mathbf{x}^*\|_2^2 + \|\mathbf{y}-\mathbf{y}^\star\|_2^2$ is also a widely-used measurement of optimality (Luo et al., 2021). In this case, ϕ_f is μ_x-strongly convex and $2L^2/\mu_x$-smooth with minimizer $\mathbf{x}^*$ by Lin et al. (2020, Lemma 23). Similarly, $-\psi_f$ is μ_y-strongly convex and $2L^2/\mu_y$-smooth with minimizer $\mathbf{y}^*$. If $(\mathbf{x}^*, \mathbf{y}^*)$ is an interior point of the feasible set, we have that $\frac{\mu_x}{2}\|\mathbf{x}-\mathbf{x}^*\|_2^2 + \frac{\mu_y}{2}\|\mathbf{y}-\mathbf{y}^*\|_2^2 \le \phi_f(\mathbf{x}) - \psi_f(\mathbf{y}) \le \frac{L^2}{\mu_x}\|\mathbf{x}-\mathbf{x}^*\|_2^2 + \frac{L^2}{\mu_y}\|\mathbf{y}-\mathbf{y}^*\|_2^2$ holds.[5] Then the squared distance and the primal-dual gap are equivalent up to constant factors. As for the nonconvex-strongly-concave case, we aim to find the stationary point of the primal function and use the norm of the gradient of the primal function as the measurement.

Note that we use the number of PIFO calls to measure the complexity. We claim that the infrequent FO calls do not influence the order of this complexity. At each step, the FO is called with probability $q = \mathcal{O}\left(\frac{1}{n}\right)$. Since the computational cost of each FO call is no larger than that of n PIFO calls, the total cost of PO calls is no larger than the order of the number of PIFO calls in expectation. Thus our definition of complexity is reasonable, due to that we usually ignore the influence of constants.

4.2 The Hard Instances

In this subsection, we construct the (unscaled) hard instances used to prove the lower bound. The constructions for the convex-concave case and the nonconvex-strongly-concave case are slightly different and presented in Sections 4.2.1 and 4.2.2 respectively. However, they are both based on the following class of matrices

$$\mathbf{B}(m,\omega,\zeta) = \begin{bmatrix} \omega & & & & \\ 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1 \\ & & & & \zeta \end{bmatrix} \in \mathbb{R}^{(m+1)\times m}, \tag{3}$$

which is also used in the proof of lower bounds in deterministic minimax optimization (Ouyang and Xu, 2021; Zhang et al., 2022).

4. Our definition follows from Carmon et al. (2020a).

5. When f is strongly-convex-strongly-concave, we can take the feasible set as $\{(\mathbf{x},\mathbf{y})\colon \phi_f(\mathbf{x}) - \psi_y(\mathbf{y}) \le \phi_f(\mathbf{x}_0) - \psi_f(\mathbf{y}_0)\}$ with $(\mathbf{x}_0, \mathbf{y}_0)$ the initial point. Then $(\mathbf{x}^*, \mathbf{y}^*)$ is naturally an interior point.

For convenience, we denote the l-th row of the matrix $\mathbf{B}(m,\omega,\zeta)$ by $\mathbf{b}_{l-1}(m,\omega,\zeta)^\top$. To construct a hard instance for the finite-sum optimization problem, we partition the row vectors of $\mathbf{B}(m,\omega,\zeta)$ according to the index sets $\mathcal{L}_i = \{l : 0 \le l \le m, l \equiv i-1 \pmod n\}$. The i-th component is constructed in terms of $\{\mathbf{b}_l(m,\omega,\zeta) : l \in \mathcal{L}_i\}$. This way of partition is different from those used in Lan and Zhou (2018) and Zhou and Gu (2019) (a detailed comparison is deferred to Section 6.1). We find that the $\mathbf{b}_l(m,\omega,\zeta)$ have at most two nonzero elements and the vectors whose indices lie in the same index sets are mutually orthogonal, as long as $n \ge 2$.

4.2.1 Convex-Concave Case

The hard instance for the convex-concave case is constructed as

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} r^{\mathrm{CC}}(\mathbf{x},\mathbf{y};m,\zeta,\mathbf{c}^{\mathrm{CC}}) \triangleq \frac{1}{n}\sum_{i=1}^n r_i^{\mathrm{CC}}(\mathbf{x},\mathbf{y};m,\zeta,\mathbf{c}^{\mathrm{CC}}), \tag{4}$$

where $\mathbf{c}^{\mathrm{CC}} = (c_1^{\mathrm{CC}}, c_2^{\mathrm{CC}})$, $\mathcal{X} = \{\mathbf{x}\in\mathbb{R}^m : \|\mathbf{x}\|_2 \le R_x\}, \mathcal{Y} = \{\mathbf{y}\in\mathbb{R}^m : \|\mathbf{y}\|_2 \le R_y\}$ and

$$\begin{aligned} & r_i^{\mathrm{CC}}(\mathbf{x},\mathbf{y};m,\zeta,\mathbf{c}^{\mathrm{CC}}) \\ &= \begin{cases} n\sum\limits_{l\in\mathcal{L}_i} \mathbf{y}^\top\mathbf{e}_l\mathbf{b}_l(m,0,\zeta)^\top\mathbf{x} + \frac{c_1^{\mathrm{CC}}}{2}\|\mathbf{x}\|_2^2 - \frac{c_2^{\mathrm{CC}}}{2}\|\mathbf{y}\|_2^2 - n\langle\mathbf{e}_1,\mathbf{x}\rangle, & \text{for } i=1, \\ n\sum\limits_{l\in\mathcal{L}_i} \mathbf{y}^\top\mathbf{e}_l\mathbf{b}_l(m,0,\zeta)^\top\mathbf{x} + \frac{c_1^{\mathrm{CC}}}{2}\|\mathbf{x}\|_2^2 - \frac{c_2^{\mathrm{CC}}}{2}\|\mathbf{y}\|_2^2, & \text{for } i=2,3,\dots,n. \end{cases} \end{aligned}$$

Note that $\mathbf{b}_0(m,0,\zeta) = \mathbf{0}$, which implies that this hard instance is based on the last m rows of $\mathbf{B}(m,\omega,\zeta)$. Then we can determine the smoothness and strong convexity coefficients of r_i^{CC} as follows.

Proposition 14 *For $c_1^{\mathrm{CC}}, c_2^{\mathrm{CC}} \ge 0$ and $0 \le \zeta \le \sqrt{2}$, we have that r_i^{CC} is L-smooth and $(c_1^{\mathrm{CC}}, c_2^{\mathrm{CC}})$-convex-concave, and $\{r_i^{\mathrm{CC}}\}_{i=1}^n$ is L'-average smooth, where*

$$L = \sqrt{4n^2 + 2\max\{c_1^{\mathrm{CC}}, c_2^{\mathrm{CC}}\}^2} \quad \text{and} \quad L' = \sqrt{8n + 2\max\{c_1^{\mathrm{CC}}, c_2^{\mathrm{CC}}\}^2}.$$

We find if $\max\{c_1^{\mathrm{CC}}, c_2^{\mathrm{CC}}\} = \mathcal{O}(\sqrt{n})$, then $L/L' = \Theta(\sqrt{n})$.

Define the subspaces $\{\mathcal{F}_k\}_{k=0}^m$ as

$$\mathcal{F}_k = \begin{cases} \mathrm{span}\{\mathbf{e}_1,\mathbf{e}_2,\dots,\mathbf{e}_k\}, & \text{for } 1\le k\le m, \\ \{\mathbf{0}\}, & \text{for } k=0. \end{cases} \tag{5}$$

Now we show that the hard instance satisfies a variant of the *zero-chain* property (Carmon et al., 2020a).

Lemma 15 *Suppose that $n \ge 2$ and $\mathcal{F}_{-1} = \mathcal{F}_0$. Then for $(\mathbf{x},\mathbf{y}) \in \mathcal{F}_k \times \mathcal{F}_{k-1}$ and $0 \le k < m$, we have that*

$$\begin{pmatrix} \nabla_{\mathbf{x}} r_i^{\mathrm{CC}}(\mathbf{x},\mathbf{y}) \\ -\nabla_{\mathbf{y}} r_i^{\mathrm{CC}}(\mathbf{x},\mathbf{y}) \end{pmatrix}, \mathrm{prox}_{r_i^{\mathrm{CC}}}^{\gamma}(\mathbf{x},\mathbf{y}) \in \begin{cases} \mathcal{F}_{k+1}\times\mathcal{F}_k, & \text{if } i \equiv k+1 \pmod n, \\ \mathcal{F}_k \times \mathcal{F}_{k-1}, & \text{otherwise,} \end{cases}$$

where we omit the parameters of r_i^{CC} to simplify the presentation.

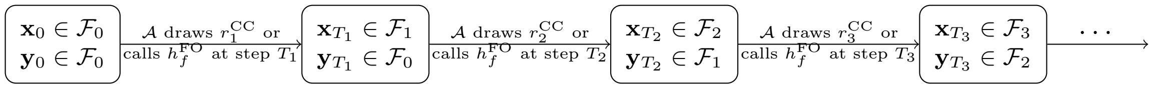

Figure 1: An illustration of the process of solving Problem (4) with a PIFO algorithm $\mathcal{A}$.

If the current point is $(\mathbf{x}, \mathbf{y})$, the information provided by the PIFO call at $(\mathbf{x}, \mathbf{y})$ will not increase the nonzero elements of $(\mathbf{x}, \mathbf{y})$ unless a specific component function is drawn. Moreover, if such a specific component is drawn, the increase is at most 2. This variant of *zero-chain* property is different from the conventional *zero-chain* property in finite-sum minimization problems (Lan and Zhou, 2018; Zhou and Gu, 2019), where regardless of which component is drawn, the nonzero elements of the current point can increase. Such a difference comes from different ways of partitioning and ensures that our construction requires a lower dimension (see the analysis in Section 6.2). The proofs of Proposition 14 and Lemma 15 are given in Appendix C.1.

When we apply a PIFO algorithm $\mathcal{A}$ to solve Problem (4), Lemma 15 implies that $\mathbf{x}_t = \mathbf{y}_t = \mathbf{0}$ will hold until algorithm $\mathcal{A}$ draws the component f_1 or calls the FO. Then, for any $t < T_1 = \min_t\{t : i_t = 1 \text{ or } a_t = 1\}$, we have $\mathbf{x}_t, \mathbf{y}_t \in \mathcal{F}_0$ while $\mathbf{x}_{T_1} \in \mathcal{F}_1$ and $\mathbf{y}_{T_1} \in \mathcal{F}_0$. The value of T_1 can be regarded as the smallest integer such that $\mathbf{x}_{T_1} \in \mathcal{F}_1 \setminus \mathcal{F}_0$ could hold. Similarly, for $T_1 \le t < T_2 = \min_t\{t > T_1 : i_t = 2 \text{ or } a_t = 1\}$, it holds that $\mathbf{x}_t \in \mathcal{F}_1$ and $\mathbf{y}_t \in \mathcal{F}_0$ while we can ensure that $\mathbf{x}_{T_2} \in \mathcal{F}_2$ and $\mathbf{y}_{T_2} \in \mathcal{F}_1$. Figure 1 illustrates this optimization process.

We can define T_k to be the smallest integer such that $\mathbf{x}_{T_k} \in \mathcal{F}_k \setminus \mathcal{F}_{k-1}$ and $\mathbf{y}_{T_k} \in \mathcal{F}_{k-1} \setminus \mathcal{F}_{k-2}$ could hold. The following corollary demonstrates that we can connect T_k to geometrically distributed random variables.

Corollary 16 *Assume we employ a PIFO algorithm $\mathcal{A}$ to solve Problem (4). Let*

$$T_0 = 0, \quad \textit{and} \;\; T_k = \min_t\{t : t > T_{k-1}, i_t \equiv k \pmod{n} \;\textit{or}\; a_t = 1\} \;\; \textit{for } k \ge 1. \tag{6}$$

Then we have

$$(\mathbf{x}_t, \mathbf{y}_t) \in \mathcal{F}_{k-1} \times \mathcal{F}_{k-2}, \qquad \textit{for } t < T_k, k \ge 1.$$

Moreover, the random variables $\{Y_k\}_{k \ge 1}$ such that $Y_k \triangleq T_k - T_{k-1}$ are mutually independent and Y_k follows a geometric distribution with success probability $p_{k'} + q - p_{k'}q$ where $k' \in [n]$ satisfies $k' \equiv k \pmod{n}$.

The basic idea of our analysis is that we guarantee that the ε-suboptimal solution of Problem (4) does not lie in $\mathcal{F}_k \times \mathcal{F}_k$ for $k < m$ and assure that the PIFO algorithm extends the space $\mathrm{span}\{(\mathbf{x}_0, \mathbf{y}_0), (\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_t, \mathbf{y}_t)\}$ slowly with t increasing. By Corollary 16, we know that $\mathrm{span}\{(\mathbf{x}_0, \mathbf{y}_0), (\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_{T_k-1}, \mathbf{y}_{T_k-1})\} \subseteq \mathcal{F}_{k-1} \times \mathcal{F}_{k-1}$. Hence, T_k is the quantity that measures how $\mathrm{span}\{(\mathbf{x}_0, \mathbf{y}_0), (\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_t, \mathbf{y}_t)\}$ expands. Note that T_k can be written as the sum of geometrically distributed random variables. Recalling Lemma 10, we can obtain how many PIFO calls we need.

Lemma 17 *If M satisfies $1 \leq M < m$,*

$$\min_{\substack{\mathbf{x}\in\mathcal{X}\cap\mathcal{F}_M \\ \mathbf{y}\in\mathcal{Y}\cap\mathcal{F}_M}} \left(\max_{\mathbf{v}\in\mathcal{Y}} r^{\mathrm{CC}}(\mathbf{x},\mathbf{v}) - \min_{\mathbf{u}\in\mathcal{X}} r^{\mathrm{CC}}(\mathbf{u},\mathbf{y}) \right) \geq 9\varepsilon \tag{7}$$

and $N = \frac{n(M+1)}{4(1+c_0)}$, then we have

$$\min_{t\leq N} \mathbb{E} \left(\max_{\mathbf{v}\in\mathcal{Y}} r^{\mathrm{CC}}(\mathbf{x}_t,\mathbf{v}) - \min_{\mathbf{u}\in\mathcal{X}} r^{\mathrm{CC}}(\mathbf{u},\mathbf{y}_t) \right) \geq \varepsilon.$$

Note that rescaling will not influence the *zero-chain* property. Thus Lemma 17 still holds for any rescaled version of r^{CC}. It remains to specify the parameters carefully, obtain a condition of the form (7) and then estimate the order of N. These steps depend on the specific problem and are deferred to Sections 5.2 to 5.4 and Appendices D.1 to D.3.

The proofs of Corollary 16 and Lemma 17 are given in Appendix C.2.

4.2.2 Nonconvex-Strongly-Concave Case

For the nonconvex-strongly-concave case, the hard instance is constructed as

$$\min_{\mathbf{x}\in\mathbb{R}^m} \max_{\mathbf{y}\in\mathbb{R}^m} r^{\mathrm{NCC}}(\mathbf{x},\mathbf{y};m,\omega,\mathbf{c}^{\mathrm{NCC}}) \triangleq \frac{1}{n}\sum_{i=1}^{n} r_i^{\mathrm{NCC}}(\mathbf{x},\mathbf{y};m,\omega,\mathbf{c}^{\mathrm{NCC}}) \tag{8}$$

where $\mathbf{c}^{\mathrm{NCC}} = (c_1^{\mathrm{NCC}}, c_2^{\mathrm{NCC}}, c_3^{\mathrm{NCC}})$ and

$$r_i^{\mathrm{NCC}}(\mathbf{x},\mathbf{y};m,\omega,\mathbf{c}^{\mathrm{NCC}})$$
$$= \begin{cases} n \sum_{l\in\mathcal{L}_i} \mathbf{y}^\top \mathbf{e}_{l+1}\mathbf{b}_l(m,\omega,0)^\top\mathbf{x} - \frac{c_1^{\mathrm{NCC}}}{2}\|\mathbf{y}\|_2^2 + c_2^{\mathrm{NCC}}\sum_{i=1}^{m-1}\Gamma(c_3^{\mathrm{NCC}}x_i) - n\langle\mathbf{e}_1,\mathbf{y}\rangle, & \text{for } i=1, \\ n \sum_{l\in\mathcal{L}_i} \mathbf{y}^\top \mathbf{e}_{l+1}\mathbf{b}_l(m,\omega,0)^\top\mathbf{x} - \frac{c_1^{\mathrm{NCC}}}{2}\|\mathbf{y}\|_2^2 + c_2^{\mathrm{NCC}}\sum_{i=1}^{m-1}\Gamma(c_3^{\mathrm{NCC}}x_i), & \text{for } i=2,3,\dots,n. \end{cases}$$

The nonconvex function $\Gamma : \mathbb{R}\to\mathbb{R}$ is

$$\Gamma(x) \triangleq 120\int_1^x \frac{t^2(t-1)}{1+t^2}dt,$$

which was introduced by Carmon et al. (2021). Since $\mathbf{b}_m(m,\omega,0) = \mathbf{0}_m$, the vector $\mathbf{e}_{m+1}$ will not appear in the definition of r^{NCC}. Thus r^{NCC} is well-defined and only depends on the first m rows of $\mathbf{B}(m,\omega,\zeta)$. We can determine the smoothness and strong convexity coefficients of r_i^{NCC} as follows.

Proposition 18 *For $c_1^{\mathrm{NCC}} \geq 0$, $c_2^{\mathrm{NCC}}, c_3^{\mathrm{NCC}} > 0$ and $0 \leq \omega \leq \sqrt{2}$, we have that r_i^{NCC} is L-smooth and $\left(-45(\sqrt{3}-1)c_2^{\mathrm{NCC}}(c_3^{\mathrm{NCC}})^2, c_1^{\mathrm{NCC}}\right)$-convex-concave, and $\{r_i^{\mathrm{NCC}}\}_{i=1}^n$ is L'-average smooth, where*

$$L = \sqrt{4n^2+2(c_1^{\mathrm{NCC}})^2} + 180c_2^{\mathrm{NCC}}(c_3^{\mathrm{NCC}})^2 \text{ and } L' = 2\sqrt{4n+(c_1^{\mathrm{NCC}})^2+16200(c_2^{\mathrm{NCC}})^2(c_3^{\mathrm{NCC}})^4}.$$

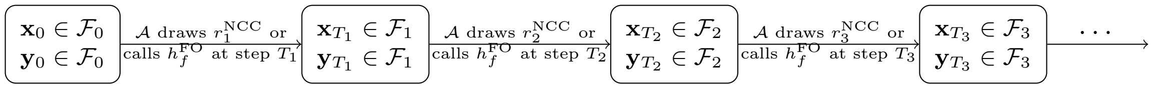

Figure 2: An illustration of the process of solving Problem (8) with a PIFO algorithm $\mathcal{A}$.

We find if $\max\{c_1^{\mathrm{NCC}}, c_2^{\mathrm{NCC}}(c_3^{\mathrm{NCC}})^2\} = \mathcal{O}(\sqrt{n})$, then $L/L' = \Theta(\sqrt{n})$.

The next lemma shows that the r_i^{NCC} share the similar *zero-chain* property as Lemma 15.

Lemma 19 *Suppose that $n \geq 2$, $c_2^{\mathrm{NCC}}, c_3^{\mathrm{NCC}} > 0$ and $\gamma < \frac{\sqrt{2}+1}{60 c_2^{\mathrm{NCC}}(c_3^{\mathrm{NCC}})^2}$. If $(\mathbf{x}, \mathbf{y}) \in \mathcal{F}_k \times \mathcal{F}_k$ and $0 \leq k < m-1$, we have that*

$$\begin{pmatrix} \nabla_x r_i^{\mathrm{NCC}}(\mathbf{x}, \mathbf{y}) \\ -\nabla_y r_i^{\mathrm{NCC}}(\mathbf{x}, \mathbf{y}) \end{pmatrix}, \operatorname{prox}_{r_i^{\mathrm{NCC}}}^{\gamma}(\mathbf{x}, \mathbf{y}) \in \begin{cases} \mathcal{F}_{k+1} \times \mathcal{F}_{k+1}, & \text{if } i \equiv k+1 \pmod{n}, \\ \mathcal{F}_k \times \mathcal{F}_k, & \text{otherwise}, \end{cases}$$

where we omit the parameters of r_i^{NCC} to simplify the presentation.

The proofs of Proposition 18 and Lemma 19 are given in Appendix C.3.

It is worth emphasizing that the assumption on γ naturally holds. Recall that the choice of γ should satisfy that $r_i(\mathbf{u}, \mathbf{y}) + \frac{1}{2\gamma}\|\mathbf{x} - \mathbf{u}\|_2^2 - \frac{1}{2\gamma}\|\mathbf{y} - \mathbf{v}\|_2^2$ is convex-concave in $(\mathbf{u}, \mathbf{v})$. Proposition 18 implies that we must have $\gamma \leq \frac{1}{45(\sqrt{3}-1)c_2^{\mathrm{NCC}}(c_3^{\mathrm{NCC}})^2} \leq \frac{\sqrt{2}+1}{60 c_2^{\mathrm{NCC}}(c_3^{\mathrm{NCC}})^2}$.

When we apply a PIFO algorithm to solve Problem (8), the optimization process is similar to the process related to Problem (4). We demonstrate the optimization process in Figure 2 and present a formal statement in Corollary 20.

Corollary 20 *Assume we employ a PIFO algorithm $\mathcal{A}$ to solve Problem (8). Let*

$$T_0 = 0, \quad \text{and} \quad T_k = \min_t\{t : t > T_{k-1}, i_t \equiv k \pmod{n} \text{ or } a_t = 1\} \quad \text{for } k \geq 1.$$

Then we have

$$(\mathbf{x}_t, \mathbf{y}_t) \in \mathcal{F}_{k-1} \times \mathcal{F}_{k-1}, \qquad \text{for } t < T_k, k \geq 1.$$

Moreover, the random variables $\{Y_k\}_{k\geq 1}$ such that $Y_k \triangleq T_k - T_{k-1}$ are mutual independent and Y_k follows a geometric distribution with success probability $p_{k'} + q - p_{k'}q$ where $k' \in [n]$ satisfies $k' \equiv k \pmod{n}$.

The proof of Corollary 20 is similar to that of Corollary 16. Furthermore, the prime-dual gap in Lemma 17 can be replaced with the gradient norm of the primal function in the nonconvex-strongly-concave case.

Lemma 21 *Let $\phi_{r^{\mathrm{NCC}}}(\mathbf{x}) \triangleq \max_{\mathbf{y}\in\mathbb{R}^m} r^{\mathrm{NCC}}(\mathbf{x}, \mathbf{y})$. If M satisfies $1 \leq M < m$ and*

$$\min_{\mathbf{x}\in\mathcal{F}_M} \|\nabla \phi_{r^{\mathrm{NCC}}}(\mathbf{x})\|_2 \geq 9\varepsilon \tag{9}$$

and $N = \frac{n(M+1)}{4(1+c_0)}$, then we have

$$\min_{t\leq N} \mathbb{E}\|\nabla \phi_{r^{\mathrm{NCC}}}(\mathbf{x}_t)\|_2 \geq \varepsilon.$$

Lemma 21 also holds for any rescaled version of r^{NCC}. It remains to specify the parameters carefully, obtain a condition of the form (9) and then estimate the order of N. The details are deferred to Section 5.5 and Appendix D.4.

5. Lower Complexity Bounds for the Minimax Problems

In this section, we focus on the minimax problem (1), which is restated as follows.

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

We assume that the function class $\{f_i(\mathbf{x},\mathbf{y})\}_{i=1}^n$ is L-average smooth, and the feasible sets $\mathcal{X}$ and $\mathcal{Y}$ are closed and convex. In addition, $f(\mathbf{x},\mathbf{y})$ is convex in $\mathbf{x}$ and concave in $\mathbf{y}$ or $f(\mathbf{x},\mathbf{y})$ is nonconvex in $\mathbf{x}$ and strongly-concave in $\mathbf{y}$. The lower bound results are shown in Section 5.1. The detailed constructions for different cases are shown in Sections 5.2 to 5.5. Finally, in Section 5.6, we consider the more constrained case where each f_i is L-smooth and briefly introduce the results.

5.1 Main Results

Recall that the comparison of the upper and lower bounds is already shown in Table 1. In this subsection, we present the formal statements of our lower bounds and give some interpretation. We emphasize that the methods in Luo et al. (2021); Zhang et al. (2021) are IFO algorithms from the analysis in Section 3.2, which implies PIFO oracles are not much more powerful than IFO oracles.

We start with the case where the objective function f is μ_x-strongly-convex in $\mathbf{x}$ and μ_y-strongly-concave in $\mathbf{y}$. Define the condition numbers $\kappa_x \triangleq L/\mu_x$ and $\kappa_y \triangleq L/\mu_y$. Without loss of generality, we assume $\mu_x \le \mu_y$. According to the relationship between κ_x, κ_y and n, we can classify the problem into three cases: (a) f is extremely ill-conditioned w.r.t. both $\mathbf{x}$ and $\mathbf{y}$, i.e., $\kappa_x, \kappa_y = \Omega(\sqrt{n})$; (b) f is only extremely ill-conditioned w.r.t. $\mathbf{x}$, i.e., $\kappa_x = \Omega(\sqrt{n}), \kappa_y = \mathcal{O}(\sqrt{n})$; (c) f is relatively well-conditioned w.r.t. both $\mathbf{x}$ and $\mathbf{y}$, i.e., $\kappa_x, \kappa_y = \mathcal{O}(\sqrt{n})$. For the three cases, we can prove different lower bounds as follows.

Theorem 22 *Let $n \ge 4$ be a positive integer and $L, \mu_x, \mu_y, R_x, R_y, \varepsilon$ be positive parameters. Assume additionally that $\kappa_x \ge \kappa_y \ge 2$ and $\varepsilon \le \min\left\{\frac{n\mu_x R_x^2}{800\kappa_x\kappa_y}, \frac{\mu_x R_x^2}{720}, \frac{\mu_y R_y^2}{800}\right\}$. Then we have*

$$\mathfrak{m}^{\mathrm{CC}}(\varepsilon, R_x, R_y, L, \mu_x, \mu_y) = \begin{cases} \Omega\left(\left(n+\sqrt{\kappa_x\kappa_y n}\right)\log(1/\varepsilon)\right), & \text{for } \kappa_x, \kappa_y = \Omega(\sqrt{n}), \\ \Omega\left(\left(n+n^{3/4}\sqrt{\kappa_x}\right)\log(1/\varepsilon)\right), & \text{for } \kappa_x = \Omega(\sqrt{n}), \kappa_y = \mathcal{O}(\sqrt{n}), \\ \Omega(n), & \text{for } \kappa_x, \kappa_y = \mathcal{O}(\sqrt{n}). \end{cases}$$

We mainly focus on the first two cases where at least one condition number is of the order $\Omega(\sqrt{n})$. Then the lower bound can be summarized as $\Omega(\sqrt{n(\sqrt{n}+\kappa_x)(\sqrt{n}+\kappa_y)}\log(1/\varepsilon))$, as shown in Table 1.

Some works focus on the balanced case $\kappa_x = \kappa_y$. For example, the upper bound of Accelerated SVRG/SAGA (Palaniappan and Bach, 2016) is $\mathcal{O}\left(\left(n + \frac{\sqrt{n}L}{\min\{\mu_x,\mu_y\}}\right)\log(1/\varepsilon)\right)$. L-SVRE (Alacaoglu and Malitsky, 2022) also achieves the same upper bound.[6] At least for the balanced case, their upper bounds nearly match our lower bound. However, for the

6. The setting in Section 4.3 of Alacaoglu and Malitsky (2022) is slightly different from ours here. However, the proof of their result can be adapted to the strongly-convex-strongly-concave case.

unbalanced case, there still exists a gap. Luo et al. (2021) focus on the unbalanced case. They employ the catalyst technique to accelerate L-SVRE and propose the method AL-SVRE, which achieves the upper bound $\tilde{\mathcal{O}}(\sqrt{n(\sqrt{n}+\kappa_x)(\sqrt{n}+\kappa_y)}\log(1/\varepsilon))$. This bound nearly matches our lower bound for the unbalanced case up to log factors.

Then we consider the lower bound when the objective function is not strongly convex in $\mathbf{x}$, i.e., $\mu_x = 0$. In this case, only the condition number w.r.t. $\mathbf{y}$ is well-defined. According to the relationship between κ_y and $\sqrt{n}$, we can also classify the problem into two cases: (a) f is extremely ill-conditioned w.r.t. $\mathbf{y}$, i.e., $\kappa_y = \Omega(\sqrt{n})$; (b) f is relatively well-conditioned w.r.t. $\mathbf{y}$, i.e., $\kappa_y = \mathcal{O}(\sqrt{n})$. We can prove the lower bounds as follows.

Theorem 23 *Let $n \geq 4$ be a positive integer and $L, \mu_y, R_x, R_y, \varepsilon$ be positive parameters. Assume additionally that $\kappa_y \geq 2$ and $\varepsilon \leq \min\left\{\frac{LR_x^2}{4}, \frac{\mu_y R_y^2}{36}\right\}$. Then we have*

$$\mathfrak{m}^{\mathrm{CC}}(\varepsilon, R_x, R_y, L, 0, \mu_y) = \begin{cases} \Omega\Big(n + R_x n^{3/4}\sqrt{\frac{L}{\varepsilon}} + R_x\sqrt{\frac{nL\kappa_y}{\varepsilon}} + n^{3/4}\sqrt{\kappa_y}\log(\frac{1}{\varepsilon})\Big), & \text{for } \kappa_y = \Omega(\sqrt{n}), \\ \Omega\Big(n + R_x n^{3/4}\sqrt{\frac{L}{\varepsilon}} + R_x\sqrt{\frac{nL\kappa_y}{\varepsilon}}\Big), & \text{for } \kappa_y = \mathcal{O}(\sqrt{n}). \end{cases}$$

For both cases, the leading term w.r.t. ε is of the order $\Omega(\sqrt{1/\varepsilon})$ and the only difference between the two bounds is the term $\Omega(n^{3/4}\sqrt{\kappa_y}\log(\frac{1}{\varepsilon}))$, which is usually much smaller than the $\Omega(\sqrt{1/\varepsilon})$ term, especially when ε is small. The upper bound of AL-SVRE (Luo et al., 2021) for this case is $\mathcal{O}\Big(\big(n + R_x n^{3/4}\sqrt{\frac{L}{\varepsilon}} + R_x\sqrt{\frac{nL\kappa_y}{\varepsilon}} + n^{3/4}\sqrt{\kappa_y}\big)\log(\frac{1}{\varepsilon})\Big)$, which nearly matches our lower bound up to log factors.

For the general convex-concave case where $\mu_x = \mu_y = 0$, we have the following lower bound.

Theorem 24 *Let $n \geq 2$ be a positive integer and L, R_x, R_y, ε be positive parameters. Assume additionally that $\varepsilon \leq \frac{L}{4}\min\{R_x^2, R_y^2\}$. Then we have*

$$\mathfrak{m}^{\mathrm{CC}}(\varepsilon, R_x, R_y, L, 0, 0) = \Omega\left(n + \frac{\sqrt{n}LR_xR_y}{\varepsilon} + (R_x + R_y)n^{3/4}\sqrt{\frac{L}{\varepsilon}}\right).$$

The leading term w.r.t. ε is of the order $\Omega(1/\varepsilon)$. If $\varepsilon = \mathcal{O}\left(\frac{LR_x^2R_y^2}{\sqrt{n}(R_x+R_y)^2}\right)$, our lower bound is $\Omega\left(n + \frac{\sqrt{n}LR_xR_y}{\varepsilon}\right)$, which matches the upper bound $\mathcal{O}\left(n + \frac{\sqrt{n}L(R_x^2+R_y^2)}{\varepsilon}\right)$ in Alacaoglu and Malitsky (2022) in terms of n, L and ε. The upper bound of AL-SVRE (Luo et al., 2021) for this case is $\mathcal{O}\Big(\big(n + \frac{\sqrt{n}LR_xR_y}{\varepsilon} + (R_x + R_y)n^{3/4}\sqrt{\frac{L}{\varepsilon}}\big)\log(\frac{1}{\varepsilon})\Big)$, which nearly matches our lower bound up to log factors.

Finally, we give the lower bound when the objective function is nonconvex in $\mathbf{x}$ but strongly concave in $\mathbf{y}$.

Theorem 25 *Let $n \geq 2$ be a positive integer and $L, \mu_x, \mu_y, R_x, R_y, \varepsilon$ be positive parameters. Assume additionally that $\varepsilon^2 \leq \frac{\Delta L^2\alpha}{435456 n\mu_y}$, where $\alpha = \min\left\{1, \frac{128(\sqrt{3}+1)n\mu_x\mu_y}{45L^2}, \frac{32n\mu_y}{135L}\right\}$. Then we have*

$$\mathfrak{m}^{\mathrm{NCC}}(\varepsilon, \Delta, L, \mu_x, \mu_y) = \Omega\left(n + \frac{\Delta L^2\sqrt{\alpha}}{\mu_y\varepsilon^2}\right).$$

For $\kappa_y = L/\mu_y \geq 32n/135$, we have

$$\Omega\left(n + \frac{\Delta L^2\sqrt{\alpha}}{\mu_y \varepsilon^2}\right) = \Omega\left(n + \frac{\Delta L\sqrt{n}}{\varepsilon^2}\min\left\{\sqrt{\kappa_y}, \sqrt{\frac{\mu_x}{\mu_y}}\right\}\right).$$

We mainly focus on the ill-conditioned setting $\kappa_y = \Omega(n)$, where the lower bound has a more concise expression. Recall that μ_x measures the nonconvexity of the function. When L is fixed, we must have $\mu_x \leq L$. If we are uninterested in the dependence of the lower bound on μ_x, then we can consider the largest function class $\mathcal{F}_{\mathrm{NCC}}(\Delta, L, L, \mu_y)$. which corresponds to the complexity $\mathfrak{m}^{\mathrm{NCC}}(\varepsilon, \Delta, L, L, \mu_y) = \Omega\left(n + \frac{\Delta L\sqrt{n\kappa_y}}{\varepsilon^2}\right)$, as shown in Table 1.[7]

As for the upper bound, Luo et al. (2020) propose the method SREDA and establish the upper bound $\mathcal{O}\left(n\log(\kappa_y/\varepsilon) + L\kappa_y^2\sqrt{n}\varepsilon^{-2}\right)$ for $n \geq \kappa_y^2$ and $\mathcal{O}\left((\kappa_y^2 + \kappa_y n)L\varepsilon^{-2}\right)$ for $n < \kappa_y^2$. Zhang et al. (2021) propose Catalyst-SVRG/SAGA and obtain the upper bound $\tilde{\mathcal{O}}\left((n + n^{3/4}\sqrt{\kappa_y})\Delta L\varepsilon^{-2}\right)$. When $n \leq \kappa^4$, the upper bound of Zhang et al. (2021) is better; otherwise, the upper bound of Luo et al. (2020) is better. Since we focus on the ill-conditioned setting, the upper and lower bounds nearly match in terms of κ_y. And there is still a $n^{1/4}$ gap in terms of n.

5.2 Construction for the Strongly-Convex-Strongly-Concave Case

In this subsection, we give the exact forms of the hard instance when the objective function is strongly convex in $\mathbf{x}$ and strongly concave in $\mathbf{y}$. We still assume $\mu_x \leq \mu_y$. Then we have $\kappa_y \leq \kappa_x$. This means that the max part has a smaller condition number and is easier to solve. According to the magnitude of κ_x and κ_y, the construction can be divided into three cases.

Case 1: $\kappa_x, \kappa_y = \Omega(\sqrt{n})$. When both condition numbers are no smaller than $\Theta(\sqrt{n})$, the analysis depends on the following construction.

Definition 26 *For fixed $L, \mu_x, \mu_y, R_x, R_y$ and n such that $\mu_x \leq \mu_y$ and $\kappa_x \geq \kappa_y \geq 2$, we define $f_{\mathrm{SCSC},i} : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ as follows*

$$f_{\mathrm{SCSC},i}(\mathbf{x}, \mathbf{y}) = \lambda\, r_i^{\mathrm{CC}}\left(\mathbf{x}/\beta, \mathbf{y}/\beta; m, \sqrt{\frac{2}{\alpha+1}}, \mathbf{c}^{\mathrm{SCSC}}\right), \text{ for } 1 \leq i \leq n,$$

where

$$\alpha = \sqrt{\frac{(\kappa_y - 2/\kappa_y)\,\kappa_x}{2n} + 1}, \quad \mathbf{c}^{\mathrm{SCSC}} = \left(\frac{2\kappa_y}{\kappa_x}\sqrt{\frac{2n}{\kappa_y^2 - 2}}, 2\sqrt{\frac{2n}{\kappa_y^2 - 2}}\right),$$

$$\beta = \min\left\{2R_x\sqrt{\frac{2\alpha n}{\kappa_x^2(1 - 2/\kappa_y^2)}}, \frac{4R_x}{\alpha+1}\sqrt{\frac{\alpha n}{\kappa_x^2(1 - 2/\kappa_y^2)}}, \frac{\sqrt{2\alpha}R_y}{\alpha - 1}\right\} \text{ and } \lambda = \frac{\beta^2}{2}\sqrt{\frac{L^2 - 2\mu_y^2}{2n}}.$$

Consider the minimax problem

$$\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}} f_{\mathrm{SCSC}}(\mathbf{x}, \mathbf{y}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_{\mathrm{SCSC},i}(\mathbf{x}, \mathbf{y}), \tag{10}$$

7. A concurrent work by Zhang et al. (2021) obtains a similar lower bound.

where $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^m : \|\mathbf{x}\|_2 \le R_x\}$ *and* $\mathcal{Y} = \{\mathbf{y} \in \mathbb{R}^m : \|\mathbf{y}\|_2 \le R_y\}$. *Define* $\phi_{\mathrm{SCSC}}(\mathbf{x}) = \max_{\mathbf{y}\in\mathcal{Y}} f_{\mathrm{SCSC}}(\mathbf{x},\mathbf{y})$ *and* $\psi_{\mathrm{SCSC}}(\mathbf{y}) = \min_{\mathbf{x}\in\mathcal{X}} f_{\mathrm{SCSC}}(\mathbf{x},\mathbf{y})$.

One can check that f_{SCSC} belongs to $\mathcal{F}_{\mathrm{CC}}(R_x, R_y, L, \mu_x, \mu_y)$ and satisfies a condition of the form (7) (please see Proposition 47 in Appendix D.1). Then we can establish the lower bound of the complexity for finding ε-suboptimal point of Problem (10) by PIFO algorithms.

Theorem 27 *Consider the minimax problem (10) and* $\varepsilon > 0$. *Let* $\alpha = \sqrt{\frac{(\kappa_y - 2/\kappa_y)\kappa_x}{2n} + 1}$. *Suppose that*

$$n \ge 2,\ \kappa_x \ge \kappa_y \ge \sqrt{2n+2},\ \ \varepsilon \le \frac{1}{800}\min\left\{\frac{n\mu_x R_x^2}{\kappa_x\kappa_y}, \mu_y R_y^2\right\},$$
$$\text{and } m = \left\lfloor \frac{\alpha}{4}\log\left(\frac{\max\left\{\mu_x R_x^2, \mu_y R_y^2\right\}}{9\varepsilon}\right)\right\rfloor + 1.$$

In order to find $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathcal{X}\times\mathcal{Y}$ *such that* $\mathbb{E}\phi_{\mathrm{SCSC}}(\hat{\mathbf{x}}) - \mathbb{E}\psi_{\mathrm{SCSC}}(\hat{\mathbf{y}}) < \varepsilon$, *any PIFO algorithm* $\mathcal{A}$ *needs at least* N *queries to PIFO, where* $N = \Omega\left(\left(n + \sqrt{n\kappa_x\kappa_y}\right)\log\left(\frac{1}{\varepsilon}\right)\right)$.

The proof of Theorem 27 is deferred to Appendix D.1.

Case 2: $\kappa_x = \Omega(\sqrt{n})$**,** $\kappa_y = \mathcal{O}(\sqrt{n})$. When only κ_y is no smaller than $\Theta(\sqrt{n})$, the lower bound is characterized by the following theorem.

Theorem 28 *For any* $L, \mu_x, \mu_y, n, R_x, R_y, \varepsilon$ *such that* $n \ge 4$,

$$n \ge 4,\ \kappa_x \ge \sqrt{2n+2} \ge \kappa_y \ge 2,\ \ \varepsilon \le \frac{1}{720}\mu_x R_x^2,\ \ \tilde{L} = \sqrt{n(L^2 - \mu_x^2)/2 - \mu_x^2},$$
$$\text{and } m = \left\lfloor \frac{1}{4}\left(\sqrt{\frac{2(\tilde{L}/\mu_x - 1)}{n}} + 1\right)\log\left(\frac{\mu_x R_x^2}{9\varepsilon}\right)\right\rfloor + 1,$$

there exist n *functions* $\{f_i : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}\}_{i=1}^n$ *such that the average* $f = \frac{1}{n}\sum_{i=1}^n f_i \in \mathcal{F}_{\mathrm{CC}}(R_x, R_y, L, \mu_x, \mu_y)$. *Let* $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^m : \|\mathbf{x}\|_2 \le R_x\}$ *and* $\mathcal{Y} = \{\mathbf{y} \in \mathbb{R}^m : \|\mathbf{y}\|_2 \le R_y\}$. *In order to find* $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathcal{X}\times\mathcal{Y}$ *such that* $\mathbb{E}\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - \mathbb{E}\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}, \hat{\mathbf{y}}) < \varepsilon$, *any PIFO algorithm* $\mathcal{A}$ *needs at least* N *queries to PIFO, where* $N = \Omega\left(\left(n + n^{3/4}\sqrt{\kappa_x}\right)\log\left(\frac{1}{\varepsilon}\right)\right)$.

We find that κ_y does not appear in the lower bound. In fact, since κ_y is relatively small, the max part is easier to solve than the min part and the min part becomes the main obstacle. To construct the hard instance, it suffices to consider the separable function of the form $f(\mathbf{x}, \mathbf{y}) = f_x(\mathbf{x}) - f_y(\mathbf{y})$ where f_x is the hard instance used for finite-sum minimization problems and $f_y(\mathbf{y}) = \frac{\mu_y}{2}\|\mathbf{y}\|_2^2$. For the details, see Appendix D.1.

Case 3: $\kappa_x, \kappa_y = \mathcal{O}(\sqrt{n})$. When both the condition numbers are relatively small, the lower bound is $\Omega(n)$, which means that the number of component functions becomes the main obstacle.

Lemma 29 *For any $L, \mu_x, \mu_y, n, R_x, R_y, \varepsilon$ such that $n \geq 2$, $L \geq \mu_x$, $L \geq \mu_y$ and $\varepsilon \leq \frac{1}{4}LR_x^2$, there exist n functions $\{f_i : \mathbb{R}\times\mathbb{R} \to \mathbb{R}\}_{i=1}^n$ such that $f = \frac{1}{n}\sum_{i=1}^n f_i \in \mathcal{F}_{\mathrm{CC}}(R_x, R_y, L, \mu_x, \mu_y)$. Let $\mathcal{X} = \{x \in \mathbb{R} : |x| \leq R_x\}$ and $\mathcal{Y} = \{y \in \mathbb{R} : |y| \leq R_y\}$. In order to find $(\hat{x}, \hat{y}) \in \mathcal{X} \times \mathcal{Y}$ such that $\mathbb{E}\max_{y\in\mathcal{Y}} f(\hat{x}, y) - \mathbb{E}\min_{x\in\mathcal{X}} f(x, \hat{y}) < \varepsilon$, any PIFO algorithm $\mathcal{A}$ needs at least $N = \Omega(n)$ queries to PIFO.*

This bound is trivial in some sense, since we usually need to compute the full gradient at least once, whose complexity is of the order $\Omega(n)$. The proof is also deferred to Appendix D.1. Combining Theorems 27, 28 and Lemma 29, we can obtain Theorem 22.

5.3 Construction for the Convex-Strongly-Concave Case

In this subsection, we construct the hard instance when f is convex in $\mathbf{x}$ and strongly concave in $\mathbf{y}$. The condition number κ_y is still well-defined. Our analysis is based on the following functions.

Definition 30 *For fixed L, μ_y, n, R_x, R_y such that $\kappa_y \geq 2$, we define $f_{\mathrm{CSC},i} : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ as follows*

$$f_{\mathrm{CSC},i}(\mathbf{x}, \mathbf{y}) = \lambda\, r_i^{\mathrm{CC}}\left(\mathbf{x}/\beta, \mathbf{y}/\beta; m, 1, \mathbf{c}^{\mathrm{CSC}}\right),$$

where

$$\mathbf{c}^{\mathrm{CSC}} = \left(0, 2\sqrt{\frac{2n}{\kappa_y^2 - 2}}\right),\ \beta = \min\left\{\frac{R_x\sqrt{\frac{\kappa_y^2-2}{2n}}}{2(m+1)^{3/2}}, \frac{R_y}{\sqrt{m}}\right\} \text{ and } \lambda = \frac{\beta^2}{2}\sqrt{\frac{L^2 - 2\mu_y^2}{2n}}.$$

Consider the minimax problem

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} f_{\mathrm{CSC}}(\mathbf{x}, \mathbf{y}) \triangleq \frac{1}{n}\sum_{i=1}^n f_{\mathrm{CSC},i}(\mathbf{x}, \mathbf{y}), \tag{11}$$

where $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^m : \|\mathbf{x}\|_2 \leq R_x\}$ and $\mathcal{Y} = \{\mathbf{y} \in \mathbb{R}^m : \|\mathbf{y}\|_2 \leq R_y\}$. Define $\phi_{\mathrm{CSC}}(\mathbf{x}) = \max_{\mathbf{y}\in\mathcal{Y}} f_{\mathrm{CSC}}(\mathbf{x}, \mathbf{y})$ and $\psi_{\mathrm{CSC}}(\mathbf{y}) = \min_{\mathbf{x}\in\mathcal{X}} f_{\mathrm{CSC}}(\mathbf{x}, \mathbf{y})$.

One can check that f_{CSC} belongs to $\mathcal{F}_{\mathrm{CC}}(R_x, R_y, L, 0, \mu_y)$ and satisfies a condition of the form (7) (please see Proposition 48 in Appendix D.2). Then we can establish the lower bound of the complexity for finding ε-suboptimal point of Problem (11) by PIFO algorithms.

Theorem 31 *Consider the minimax problem (11) and $\varepsilon > 0$. Suppose that*

$$n \geq 2,\ \kappa_y \geq 2,\ \varepsilon \leq \min\left\{\frac{L^2R_x^2}{5184\, n\mu_y}, \frac{\mu_y R_y^2}{36}\right\} \text{ and } m = \left\lfloor \frac{R_x}{6}\sqrt{\frac{L^2 - 2\mu_y^2}{2n\mu_y\varepsilon}} \right\rfloor - 2.$$

In order to find $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathcal{X}\times\mathcal{Y}$ such that $\mathbb{E}\phi_{\mathrm{CSC}}(\hat{\mathbf{x}}) - \mathbb{E}\psi_{\mathrm{CSC}}(\hat{\mathbf{y}}) < \varepsilon$, any PIFO algorithm $\mathcal{A}$ needs at least N queries to PIFO, where $N = \Omega\left(n + R_x\sqrt{nL\kappa_y/\varepsilon}\right)$.

When κ_y is small, the second term of N is also small. When $\kappa_y = \mathcal{O}(\sqrt{n})$, we can provide a tighter lower bound as follows.

Theorem 32 *For any $L, \mu_y, n, R_x, R_y, \varepsilon$ such that $n \geq 2$, $L \geq \mu_y$, $\varepsilon \leq \frac{\sqrt{2}R_x^2 L}{768\sqrt{n}}$ and $m = \left\lfloor \frac{\sqrt[4]{18}}{12} R_x n^{-1/4}\sqrt{\frac{L}{\varepsilon}} \right\rfloor - 1$, there exist n functions $\{f_i : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}\}_{i=1}^n$ such that $f = \frac{1}{n}\sum_{i=1}^n f_i \in \mathcal{F}_{\rm CC}(R_x, R_y, L, 0, \mu_y)$. Let $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^m : \|\mathbf{x}\|_2 \leq R_x\}$ and $\mathcal{Y} = \{\mathbf{y} \in \mathbb{R}^m : \|\mathbf{y}\|_2 \leq R_y\}$. In order to find $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathcal{X}\times\mathcal{Y}$ such that $\mathbb{E}\max_{\mathbf{y}\in\mathcal{Y}} f(\hat{\mathbf{x}}, \mathbf{y}) - \mathbb{E}\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}, \hat{\mathbf{y}}) < \varepsilon$, any PIFO algorithm $\mathcal{A}$ needs at least $N = \Omega\left(n + R_x n^{3/4}\sqrt{L/\varepsilon}\right)$ queries to PIFO.*

The construction of Theorem 32 is similar to that of Theorem 28. We still consider the separable function $f(\mathbf{x}, \mathbf{y}) = f_x(\mathbf{x}) - f_y(\mathbf{y})$ where f_x is the hard instance used for finite-sum minimization problems and $f_y(\mathbf{y}) = \frac{\mu_y}{2}\|\mathbf{y}\|_2^2$. The proofs of Theorems 31 and 32 are deferred to Appendix D.2.

Now we give the proof of Theorem 23.

Proof [Proof of Theorem 23] By Lemma 29, we have the lower bound $\Omega(n)$ if $\varepsilon \leq LR_x^2/4$. Note that if $\varepsilon \geq \frac{L^2R_x^2}{5184 n\mu_y}$, $\Omega(n) = \Omega\left(n + R_x\sqrt{\frac{nL\kappa_y}{\varepsilon}}\right)$. And if $\varepsilon \geq \frac{\sqrt{2}R_x^2L}{768\sqrt{n}}$, $\Omega(n) = \Omega\left(n + R_x n^{3/4}\sqrt{\frac{L}{\varepsilon}}\right)$. Then for $\varepsilon \leq \min\left\{\frac{LR_x^2}{4}, \frac{\mu_y R_y^2}{36}\right\}$, we have $\mathfrak{m}^{\rm CC}(\varepsilon, R_x, R_y, L, 0, \mu_y) = \Omega\left(n + R_x\sqrt{\frac{nL}{\varepsilon}} + \frac{R_xL}{\sqrt{\mu_y\varepsilon}}\right)$. It remains to add the term $\Omega(n^{3/4}\sqrt{\kappa_y}\log(\frac{1}{\varepsilon}))$ for $\kappa_y = \Omega(\sqrt{n})$.

Now we construct $\{H_{{\rm CSC},i}\}_{i=1}^n, H_{\rm CSC} : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ as follows.

$$
\begin{aligned}
H_{{\rm CSC},i}(\mathbf{x}, \mathbf{y}) &= \frac{L}{2}\|\mathbf{x}\|_2^2 - g_{{\rm SC},i}(\mathbf{y}), \\
H_{\rm CSC}(\mathbf{x}, \mathbf{y}) &= \frac{1}{n}\sum_{i=1}^n H_{{\rm CSC},i}(\mathbf{x}, \mathbf{y}) = \frac{L}{2}\|\mathbf{x}\|_2^2 - g_{\rm SC}(\mathbf{y}),
\end{aligned}
$$

where $g_{\rm SC}(\mathbf{y})$ is μ_y-convex and $\{g_{{\rm SC},i}(\mathbf{y})\}_{i=1}^n$ is L-average smooth. It is easy to check $H_{\rm CSC} \in \mathcal{F}_{\rm CC}(R_x, R_y, L, 0, \mu_y)$,

$$
\min_{\mathbf{x}\in\mathcal{X}} H_{\rm CSC}(\mathbf{x}, \mathbf{y}) = -g_{\rm SC}(\mathbf{y}) \quad \text{and} \quad \max_{\mathbf{y}\in\mathcal{Y}} H_{\rm CSC}(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x}\|_2^2 - \min_{\mathbf{y}\in\mathcal{Y}} g_{\rm SC}(\mathbf{y}).
$$

It follows that for any $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathcal{X} \times \mathcal{Y}$, we have

$$
\max_{\mathbf{y}\in\mathcal{Y}} H_{\rm CSC}(\hat{\mathbf{x}}, \mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} H_{\rm CSC}(\mathbf{x}, \hat{\mathbf{y}}) \geq g_{\rm SC}(\hat{\mathbf{y}}) - \min_{\mathbf{y}\in\mathcal{Y}} g_{\rm SC}(\mathbf{y}).
$$

By the result of Theorem 66, for $\varepsilon \leq \frac{LR_y^2}{4}$ and $\kappa_y = \Omega(\sqrt{n})$, we have $\mathfrak{m}_\varepsilon^{\rm CC}(R_x, R_y, L, 0, \mu_y) = n^{3/4}\sqrt{\kappa_y}\log\left(\frac{1}{\varepsilon}\right)$. This completes the proof. ∎

5.4 Construction for the Convex-Concave Case

For the general convex-concave case, the hard instance is constructed as follows.

Definition 33 *For fixed L, n, R_x, R_y such that $n \geq 2$, we define $f_{{\rm CC},i} : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ as follows*

$$
f_{{\rm CC},i}(\mathbf{x}, \mathbf{y}) = \lambda\, r_i^{\rm CC}\left(\mathbf{x}/\beta, \mathbf{y}/\beta; m, 1, \mathbf{0}\right).
$$

where $\lambda = \frac{LR_y^2}{m\sqrt{8n}}$ *and* $\beta = \frac{R_y}{\sqrt{m}}$. *Consider the minimax problem*

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} f_{\mathrm{CC}}(\mathbf{x}, \mathbf{y}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_{\mathrm{CC},i}(\mathbf{x}, \mathbf{y}), \tag{12}$$

where $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^m : \|\mathbf{x}\|_2 \leq R_x\}$ *and* $\mathcal{Y} = \{\mathbf{y} \in \mathbb{R}^m : \|\mathbf{y}\|_2 \leq R_y\}$. *Define* $\phi_{\mathrm{CC}}(\mathbf{x}) = \max_{\mathbf{y}\in\mathcal{Y}} f_{\mathrm{CC}}(\mathbf{x}, \mathbf{y})$ *and* $\psi_{\mathrm{CC}}(\mathbf{y}) = \min_{\mathbf{x}\in\mathcal{X}} f_{\mathrm{CC}}(\mathbf{x}, \mathbf{y})$.

One can check that f_{CC} belongs to $\mathcal{F}_{\mathrm{CC}}(R_x, R_y, L, 0, 0)$ and satisfies a condition of the form (7) (please see Proposition 49 in Appendix D.3). Then, we can obtain a PIFO lower bound complexity for the general finite-sum convex-concave minimax problem.

Theorem 34 *Consider minimax problem (12) and* $\varepsilon > 0$. *Suppose that*

$$n \geq 2,\ \varepsilon \leq \frac{LR_xR_y}{72\sqrt{n}},\ \ \textit{and}\ \ m = \left\lfloor \frac{LR_xR_y}{18\varepsilon\sqrt{n}} \right\rfloor - 1.$$

In order to find $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathcal{X}\times\mathcal{Y}$ *such that* $\mathbb{E}\phi_{\mathrm{CC}}(\hat{\mathbf{x}}) - \mathbb{E}\psi_{\mathrm{CC}}(\hat{\mathbf{y}}) < \varepsilon$, *any PIFO algorithm* $\mathcal{A}$ *needs at least* $N = \Omega\left(n + \sqrt{n}LR_xR_y/\varepsilon\right)$ *queries to PIFO.*

Note that Theorem 31 requires the condition $\varepsilon \leq \mathcal{O}(L/\sqrt{n})$ to obtain the desired lower bound. For large ε, we can apply the following lemma.

Lemma 35 *For any positive* $L, n, R_x, R_y, \varepsilon$ *such that* $n \geq 2$ *and* $\varepsilon \leq \frac{1}{4}LR_xR_y$ *there exist* n *functions* $\{f_i : \mathbb{R}\times\mathbb{R} \to \mathbb{R}\}_{i=1}^n$ *such that* $f = \frac{1}{n}\sum_{i=1}^n f_i \in \mathcal{F}_{\mathrm{CC}}(R_x, R_y, L, 0, 0)$. *Let* $\mathcal{X} = \{x \in \mathbb{R} : |x| \leq R_x\}$ *and* $\mathcal{Y} = \{y \in \mathbb{R} : |y| \leq R_y\}$. *In order to find* $(\hat{x}, \hat{y}) \in \mathcal{X}\times\mathcal{Y}$ *such that* $\mathbb{E}\max_{y\in\mathcal{Y}} f(\hat{x}, y) - \mathbb{E}\min_{x\in\mathcal{X}} f(x, \hat{y}) < \varepsilon$, *any PIFO algorithm* $\mathcal{A}$ *needs at least* $N = \Omega(n)$ *queries to PIFO.*

This Lemma is similar to Lemma 29. The proofs of Theorem 34 and Lemma 35 are deferred to Appendix D.3.

Now we can give the proof of Theorem 24.

Proof [Proof of Theorem 24] Note that for $\varepsilon \geq \frac{LR_xR_y}{72\sqrt{n}}$, we have $\Omega\left(n + \frac{\sqrt{n}LR_xR_y}{\varepsilon}\right) = \Omega(n)$. Combining Theorem 34 and Lemma 32, we obtain the lower bound $\Omega\left(n + \frac{\sqrt{n}LR_xR_y}{\varepsilon}\right)$ for $\varepsilon \leq LR_xR_y/4$. On the other hand, G_{CSC} defined in the proof of Theorem 32 and H_{SCSC} defined in the proof of Lemma 29 are also convex-concave and $\varepsilon \geq \frac{\sqrt{2}R_x^2L}{768\sqrt{n}}$ implies $\Omega\left(n + R_x n^{3/4}\sqrt{\frac{L}{\varepsilon}}\right) = \Omega(n)$. Thus, we have the lower bound $\Omega\left(n + R_x n^{3/4}\sqrt{\frac{L}{\varepsilon}}\right)$ for $\varepsilon \leq LR_x^2/4$. It is also worth noting that if $f(\mathbf{x}, \mathbf{y})$ is convex in $\mathbf{x}$ and concave in $\mathbf{y}$, then $-f(\mathbf{x}, \mathbf{y})$ is convex in $\mathbf{y}$ and concave in $\mathbf{x}$. This implies the symmetry of $\mathbf{x}$ and $\mathbf{y}$. Thus, we can also obtain the lower bound $\Omega\left(n + R_y n^{3/4}\sqrt{\frac{L}{\varepsilon}}\right)$ for $\varepsilon \leq LR_y^2/4$. In summary, for $\varepsilon \leq \frac{LR_xR_y}{4}$, the lower bound is $\Omega\left(n + \frac{LR_xR_y}{\varepsilon} + (R_x + R_y)n^{3/4}\sqrt{\frac{L}{\varepsilon}}\right)$. ■

5.5 Construction for the Nonconvex-Strongly-Concave Case

In this subsection, we consider the finite-sum minimax problem where the objective function is strongly concave in $\mathbf{y}$ but nonconvex in $\mathbf{x}$. The analysis is based on the following construction.

Definition 36 *For fixed $L, \mu_x, \mu_y, \Delta, n$, we define $f_{\mathrm{NCSC},i} : \mathbb{R}^{m+1} \times \mathbb{R}^{m+1} \to \mathbb{R}$ as follows*

$$f_{\mathrm{NCSC},i}(\mathbf{x}, \mathbf{y}) = \lambda\, r_i^{\mathrm{NCC}} \left(\mathbf{x}/\beta, \mathbf{y}/\beta; m+1, \sqrt[4]{\alpha}, \mathbf{c}^{\mathrm{NCSC}}\right), \text{ for } 1 \le i \le n,$$

where

$$\alpha = \min\left\{1, \frac{32n\mu_y}{135L}, \frac{128(\sqrt{3}+1)n\mu_x\mu_y}{45L^2}\right\}, \; \mathbf{c}^{\mathrm{NCSC}} = \left(\frac{16\sqrt{n}\mu_y}{L}, \frac{\sqrt{\alpha}L}{16\sqrt{n}\mu_y}, \sqrt[4]{\alpha}\right),$$

$$\lambda = \frac{5308416 n^{3/2}\mu_y^2\varepsilon^2}{L^3\alpha}, \; \beta = 4\sqrt{\lambda\sqrt{n}/L} \;\text{ and }\; m = \left\lfloor \frac{\Delta L^2\sqrt{\alpha}}{3483648 n\varepsilon^2\mu_y} \right\rfloor.$$

Define $\phi_{\mathrm{NCSC}}(\mathbf{x}) = \max_{\mathbf{y}\in\mathbb{R}^{m+1}} f_{\mathrm{NCSC}}(\mathbf{x}, \mathbf{y})$. Consider the minimax problem

$$\min_{\mathbf{x}\in\mathbb{R}^{m+1}} \max_{\mathbf{y}\in\mathbb{R}^{m+1}} f_{\mathrm{NCSC}}(\mathbf{x}, \mathbf{y}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_{\mathrm{NCSC},i}(\mathbf{x}, \mathbf{y}). \tag{13}$$

One can check that f_{NCSC} belongs to $\mathcal{F}_{\mathrm{NCC}}(\Delta, L, \mu_x, \mu_y)$ and satisfies a condition of the form (9) (please see Proposition 50 in Appendix D.4). With Proposition 50, we can give the proof of Theorem 55.

Proof [Proof of Theorem 25] Combining Lemma 21 and the third property of Proposition 50, for $N = \frac{nm}{4(1+c_0)}$, we have $\min_{t\le N} \mathbb{E}\left\|\nabla\phi_{\mathrm{NCSC}}(\mathbf{x}_t)\right\|_2 \ge \varepsilon$. Thus, in order to find $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ such that $\mathbb{E}\left\|\nabla\phi_{\mathrm{NCSC}}(\hat{\mathbf{x}})\right\|_2 < \varepsilon$, any PIFO algorithm $\mathcal{A}$ needs at least N PIFO queries, where $N = \frac{nm}{4(1+c_0)} = \Omega\left(\frac{\Delta L^2\sqrt{\alpha}}{\varepsilon^2\mu_y}\right)$. Since $\varepsilon^2 \le \frac{\Delta L^2\alpha}{6767296n\mu_y}$ and $\alpha \le 1$, we have $\Omega\left(\frac{\Delta L^2\sqrt{\alpha}}{\varepsilon^2\mu_y}\right) = \Omega\left(n + \frac{\Delta L^2\sqrt{\alpha}}{\varepsilon^2\mu_y}\right)$. ∎

5.6 Smooth Cases

In this subsection, we focus on the more constrained function classes where each component f_i is L-smooth. The results are summarized in Table 2. We defer the definitions of the function class and optimization complexity and the formal statements of our lower bounds to Appendix D.5.

In Table 2, we only present the upper bounds of some methods designed for smooth f_i's.[8] Methods designed for the average-smooth functions also apply here and thus the upper bounds in Table 1 are still valid. However, there exists a gap in all cases.

Compared to the lower bounds in Table 1, the lower bounds in Table 2 have the same dependence on $L, \kappa_x, \kappa_y, \varepsilon$, but with a weaker dependence on n. Specially. if we replace L,

8. Although the method in Carmon et al. (2019) has two loops and does not satisfy our definition, we list it here for a better comparison.

Cases	Upper or Lower Bounds	References
$\mu_x > 0, \mu_y > 0$	$\tilde{\mathcal{O}}\left(\left(n + \frac{\sqrt{n}L}{\min\{\mu_x, \mu_y\}}\right)\log(1/\varepsilon)\right)$	Carmon et al. (2019); Luo et al. (2019)
	$\Omega\left(\sqrt{(n+\kappa_x)(n+\kappa_y)}\log(1/\varepsilon)\right)$	Theorem 52
$\mu_x = 0, \mu_y > 0$	$\Omega\left(n + R_x\sqrt{\frac{nL}{\varepsilon}} + R_x\sqrt{\frac{L\kappa_y}{\varepsilon}} + \sqrt{n\kappa_y}\log\left(\frac{1}{\varepsilon}\right)\right)$	Theorem 53
$\mu_x = 0, \mu_y = 0$	$\tilde{\mathcal{O}}\left(n + \frac{\sqrt{n}L(R_x^2+R_y^2)}{\varepsilon}\right)$	Carmon et al. (2019)
	$\Omega\left(n + \frac{LR_xR_y}{\varepsilon} + (R_x+R_y)\sqrt{\frac{nL}{\varepsilon}}\right)$	Theorem 54
$\mu_x < 0, \mu_y > 0$ $\kappa_y = \Omega(\sqrt{n})$	$\Omega\left(n + \frac{\Delta L\sqrt{\kappa_y}}{\varepsilon^2}\right)$	Theorem 55

Table 2: Upper and lower bounds under the assumption that f_i is L-smooth and f is (μ_x, μ_y)-convex-concave. The condition numbers are defined as $\kappa_x = L/\mu_x$ and $\kappa_y = L/\mu_y$ when $\mu_x, \mu_y > 0$. The definitions of R_x, R_y and Δ are given in Table 1.

κ_x and κ_y in Table 2 by $\sqrt{n}L$, $\sqrt{n}\kappa_x$ and $\sqrt{n}\kappa_y$ respectively, we can obtain the lower bounds in Table 1.[9] This is due to the way of partitioning the matrix $\mathbf{B}(m, \omega, \zeta)$ in Section 4.2. Intuitively, we partition the Hessian matrix of the coupling term between $\mathbf{x}$ and $\mathbf{y}$ and each component only gets a low-rank part. Propositions 14 and 18 have shown the $\sqrt{n}$ gap between the smoothness and average smoothness parameters as long as the non-coupling term is not too large.

Convex-concave cases. We speculate that when f is convex-concave, the lower bounds in Table 2 are the best ones within our framework because the corresponding lower bounds under the average smoothness assumption have been nearly matched by existing upper bounds. To further improve the lower bounds, one may have to resort to new constructions.

As for the upper bounds, we notice that most work only uses the average smoothness condition. We guess that the smoothness property of each component function needs to be better employed because the upper and lower bounds for convex minimization problems under the two smoothness conditions nearly match (see Tables 3 and 4),

Nonconvex-strongly-concave case. When f is nonconvex-strongly-concave, there exists a gap between the upper and lower bounds under both smoothness and average smoothness assumptions. Since the nonconvexity poses more difficulty to the problem, it remains an open problem whether the upper bounds, the lower bounds, or both can be further tightened.

9. For the nonconvex-strongly-concave case, we just need to replace L by $\sqrt{n}L$.

6. Lower Complexity Bounds for the Minimization Problems

In this section, we focus on the minimization problem

$$\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}), \tag{14}$$

where each individual component $f_i(\mathbf{x})$ is L-smooth or the function class $\{f_i(\mathbf{x})\}_{i=1}^n$ is L-average smooth, the feasible set $\mathcal{X}$ is closed and convex such that $\mathcal{X} \subseteq \mathbb{R}^d$. We show that we can obtain similar lower bounds as those in Woodworth and Srebro (2016); Hannah et al. (2018); Zhou and Gu (2019).

Recall that Problem (1) becomes Problem (14) if we set $\mathcal{Y}$ as a singleton. Then the definitions of function classes and optimization complexity follow directly from their counterparts in Sections 4.1. The details are deferred to Appendix E.1.

In Section 6.1, we construct the hard instances for Problem (14). In Section 6.2, we summarize our results and compare them with previous work.

6.1 The Hard Instances

In this subsection, we present the construction of hard instances for Problem (14) and compare our construction with some related work.

The construction is also based on the class of matrices $\mathbf{B}(m,\omega,\zeta)$ define in Equation (3). We still use $\mathbf{b}_{l-1}(m,\omega,\zeta)^\top$ to denote the l-th row of $\mathbf{B}(m,\omega,\zeta)$ and defined the index sets $\mathcal{L}_1,\dots,\mathcal{L}_n$ as $\mathcal{L}_i = \left\{l : 0 \le l \le m, l \equiv i-1 \pmod n\right\}$. Then the hard instance is constructed as

$$\min_{\mathbf{x}\in\mathcal{X}} r(\mathbf{x}; m,\omega,\zeta,\mathbf{c}) \triangleq \frac{1}{n}\sum_{i=1}^{n} r_i(\mathbf{x}; m,\omega,\zeta,\mathbf{c}), \tag{15}$$

where $\mathbf{c} = (c_1, c_2, c_3)$, $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^m : \|\mathbf{x}\|_2 \le R_x\}$ or $\mathbb{R}^m$, and

$$\begin{aligned} &r_i(\mathbf{x}; m,\omega,\zeta,\mathbf{c}) \\ &= \begin{cases} \frac{n}{2}\sum\limits_{l\in\mathcal{L}_i}\left\|\mathbf{b}_l(m,\omega,\zeta)^\top\mathbf{x}\right\|_2^2 + \frac{c_1}{2}\|\mathbf{x}\|_2^2 + c_2\sum\limits_{i=1}^{m-1}\Gamma(x_i) - c_3 n\langle \mathbf{e}_1, \mathbf{x}\rangle, & \text{for } i=1, \\ \frac{n}{2}\sum\limits_{l\in\mathcal{L}_i}\left\|\mathbf{b}_l(m,\omega,\zeta)^\top\mathbf{x}\right\|_2^2 + \frac{c_1}{2}\|\mathbf{x}\|_2^2 + c_2\sum\limits_{i=1}^{m-1}\Gamma(x_i), & \text{for } i=2,3,\dots,n. \end{cases} \end{aligned}$$

The nonconvex function $\Gamma : \mathbb{R} \to \mathbb{R}$ is $\Gamma(x) \triangleq 120\int_1^x \frac{t^2(t-1)}{1+t^2}dt$. We can determine the smoothness and strong convexity parameters of r_i similar to Propositions 14 and 18. The details are deferred to Proposition 59 in Appendix E.2.

One can check that $r(\mathbf{x}; m,\omega,\zeta,\mathbf{c}) = \frac{1}{2}\mathbf{x}^\top\mathbf{A}(m,\omega,\zeta)\,\mathbf{x} + \frac{c_1}{2}\|\mathbf{x}\|_2^2 + c_2\sum_{i=1}^{m-1}\Gamma(\mathbf{x}_i) - c_3\langle \mathbf{e}_1, \mathbf{x}\rangle$, where

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

The matrix $\mathbf{A}(m, \omega, \zeta)$ is widely used in the analysis of lower bounds for convex optimization (Nesterov, 2013; Agarwal and Bottou, 2015; Lan and Zhou, 2018; Carmon et al., 2020a; Zhou and Gu, 2019).

Now we compare our construction with Lan and Zhou (2018) and Zhou and Gu (2019). In our construction, we partition the row vectors of $\mathbf{B}(m, \omega, \zeta)$ into n parts and each component function is defined in terms of only one part. All the component functions share the same $\mathbf{x}$. However, in Lan and Zhou (2018), different component functions have the same form except that they are based on different subvectors of the high-dimensional $\mathbf{x}$. Intuitively speaking, we partition the Hessian matrix while Lan and Zhou (2018) partition the variable. The construction of Zhou and Gu (2019) is more complex than Lan and Zhou (2018) but the basic idea is the same.

Recall the subspaces $\{\mathcal{F}_k\}_{k=0}^m$ defined in (5). The next lemma shows that the hard instance also satisfies a variant of the *zero-chain* property.

Lemma 37 *Suppose that $n \geq 2$, $c_1 \geq 0$ and $\mathbf{x} \in \mathcal{F}_k$, $0 \leq k < m$. If (i) (convex case) $c_2 = 0$ and $\omega = 0$, or (ii) (nonconvex case) $c_1 = 0$, $c_2 > 0$, $\zeta = 0$ and $\gamma < \frac{\sqrt{2}+1}{60c_2}$, we have*

$$\nabla r_i(\mathbf{x}), \operatorname{prox}_{r_i}^{\gamma}(\mathbf{x}) \in \begin{cases} \mathcal{F}_{k+1}, & \text{if } i \equiv k+1 (\operatorname{mod} n), \\ \mathcal{F}_k, & \text{otherwise.} \end{cases}$$

We omit the parameters of r_i to simplify the presentation.

The proof of Lemma 37 are given in Appendix E.7.

We emphasize that the assumption on γ naturally holds. Recall that the choice of γ should satisfy that $r_i(\mathbf{u}) + \frac{1}{2\gamma}\|\mathbf{x} - \mathbf{u}\|_2^2$ is a convex function of $\mathbf{u}$ for a fixed $\mathbf{x}$. Proposition 59 implies that we must have $\gamma \leq \frac{1}{45(\sqrt{3}-1)c_2} \leq \frac{\sqrt{2}+1}{60c_2}$.

In short, if $\mathbf{x} \in \mathcal{F}_k$, then there exists only one $i \in \{1, \dots, n\}$ such that $h_{f_i}^{\mathrm{PIFO}}$ could provide additional information in $\mathcal{F}_{k+1}$. This property is the main difference between the constructions in Lan and Zhou (2018); Zhou and Gu (2019) and ours. In Lan and Zhou (2018); Zhou and Gu (2019), no matter which component is drawn, the number of the nonzero elements of the current point can increase. Such a difference results from the different ways of partitioning. As a consequence, their hard instances need to be constructed in a space with a higher dimension than ours. Moreover, our construction also works for PIFO oracles while the constructions of Lan and Zhou (2018) and Zhou and Gu (2019) only apply to IFO oracles.

With Lemma 37, we can obtain how many PIFO calls we need as what we did in Section 4.2. The details are deferred to Appendix E.2.

6.2 Results

In this subsection, we present our lower bounds in Tables 3 and 4, and compare them with previous upper and lower bounds. It is worth emphasizing that we are not trying to list all the upper bounds, just to provide a few algorithms that could match our lower bounds. The formal statements of our lower bounds are deferred to Appendix E.3.

Cases	Upper or Lower Bounds	References
$\mu > 0$	$\mathcal{O}\left((n+\sqrt{\kappa n})\log(1/\varepsilon)\right)$	Defazio (2016); Li (2021)
	$\mathcal{O}\left(n + \frac{n\log(1/\varepsilon)}{1+(\log(n/\kappa))_+}\right), \quad \kappa = \mathcal{O}(n)$	Hannah et al. (2018)
	$\begin{cases} \Omega\left((n+\sqrt{\kappa n})\log(1/\varepsilon)\right), & \kappa = \Omega(n), \\ \Omega\left(n + \dfrac{n\log(1/\varepsilon)}{1+(\log(n/\kappa))_+}\right), & \kappa = \mathcal{O}(n). \end{cases}$	Hannah et al. (2018); Theorem 63
$\mu = 0$	$\tilde{\mathcal{O}}\left(n + R\sqrt{nL/\varepsilon}\right)$	Li (2021)
	$\Omega\left(n+R\sqrt{nL/\varepsilon}\right)$	Woodworth and Srebro (2016); Theorem 64
$\mu < 0$	$\tilde{\mathcal{O}}\left(n + \frac{\Delta}{\varepsilon^2}\min\{\sqrt{n}L, n\lvert\mu\rvert + \sqrt{n\lvert\mu\rvert L}\}\right)$	Lan and Yang (2019); Li et al. (2020)
	$\Omega\left(n + \frac{\Delta}{\varepsilon^2}\min\{L, \sqrt{n\lvert\mu\rvert L}\}\right)$	Zhou and Gu (2019); Theorem 65

Table 3: The upper and lower bounds under the assumption that f_i is L-smooth and f is μ-strongly convex, convex or μ-weakly convex. $\kappa = L/\mu$ for $\mu > 0$. The definitions of R, Δ and optimization complexity are given in Appendix E.1.

Smooth cases. Table 3 illustrates the upper and lower bounds when each f_i is L-smooth.[10] For the strongly convex and convex cases, the upper bounds and lower bounds nearly match up to log factors, while for the nonconvex case, there is still a $\sqrt{n}$ gap. Specially, when $\kappa = \Omega(n)$, the lower bound is $\Omega(n + \Delta\sqrt{n\lvert\mu\rvert L}/\varepsilon^2)$ and has been achieved by Lan and Zhou (2018) up to log factors. When $\kappa = \mathcal{O}(n)$, the lower bound is $\Omega(n+\Delta/\varepsilon^2)$, while the upper bound by Li et al. (2020) is $\mathcal{O}(n+\sqrt{n}\Delta/\varepsilon^2)$. From the analysis in Section 3.2, the algorithms in Defazio (2016); Hannah et al. (2018); Li (2021); Lan and Yang (2019); Li et al. (2020) all belong to PIFO algorithms. In fact, except the one in Defazio (2016), others are also IFO algorithms.

As for the lower bounds, Hannah et al. (2018) consider the class of p-CLI oblivious algorithms introduced by Arjevani and Shamir (2016). For these algorithms, we can left-multiply the gradient by a preconditioning matrix. Thus, the linear-span assumption can be violated. However, they do not take proximal operators into account. Woodworth and Srebro (2016) prove the lower bounds for arbitrary randomized algorithms with access to PIFO oracles. Although smaller than that in Woodworth and Srebro (2016), our class of algorithms is large enough to include many near-optimal algorithms. Moreover, our construction is simpler than Woodworth and Srebro (2016). As a result, such a construction can not only provide more intuition about the optimization process, but also requires fewer dimensions to construct the hard instances. Specially, for the convex case, our construction only requires the dimension to be $\mathcal{O}\left(1 + R\sqrt{L/(n\varepsilon)}\right)$ (see Appendix E.5), which is much smaller than $\mathcal{O}\left(\frac{L^2R^4}{\varepsilon^2}\log\left(\frac{nLR^2}{\varepsilon}\right)\right)$ in Woodworth and Srebro (2016).

10. The lower bound of Hannah et al. (2018) for $\kappa = \Omega(n)$ uses the lower bound in Woodworth and Srebro (2016).

Cases	Upper or Lower Bounds	References
$\mu > 0$, $\kappa = \Omega(\sqrt{n})$	$\mathcal{O}\left(\left(n + n^{3/4}\sqrt{\kappa}\right)\log(1/\varepsilon)\right)$	Allen-Zhu (2018b)
	$\Omega\left(\left(n + n^{3/4}\sqrt{\kappa}\right)\log(1/\varepsilon)\right)$	Zhou and Gu (2019); Theorem 66
$\mu = 0$	$\mathcal{O}\left(n\log(1/\varepsilon) + Rn^{3/4}\sqrt{L/\varepsilon}\right)$	Allen-Zhu (2018b)
	$\Omega\left(n + Rn^{3/4}\sqrt{L/\varepsilon}\right)$	Zhou and Gu (2019); Theorem 67
$\mu < 0$	$\tilde{\mathcal{O}}\left(n + \frac{\Delta}{\varepsilon^2}\min\{\sqrt{n}L, n^{3/4}\sqrt{\lvert\mu\rvert L}\}\right)$	Allen-Zhu (2017); Li et al. (2021)
	$\Omega\left(n + \frac{\Delta}{\varepsilon^2}\min\{\sqrt{n}L, n^{3/4}\sqrt{\lvert\mu\rvert L}\}\right)$	Zhou and Gu (2019); Theorem 68

Table 4: The upper and lower bounds with the assumption that $\{f_i\}_{i=1}^n$ is L-average smooth and f is μ-strongly convex, convex or μ-weakly convex. $\kappa = L/\mu$ for $\mu > 0$. The definitions of R, Δ and optimization complexity are given in Appendix E.1.

Zhou and Gu (2019) only consider the class of IFO algorithms, which is only a subset of PIFO algorithms. Moreover, our construction still requires fewer dimensions. For the non-convex case, our construction only requires the dimension to be $\mathcal{O}(1+\frac{\Delta}{\varepsilon^2}\min\{L/n, \sqrt{\mu L/n}\})$ (see Appendix E.6), which is much smaller than $\mathcal{O}\left(\frac{\Delta}{\varepsilon^2}\min\{L, \sqrt{n\mu L}\}\right)$ in Zhou and Gu (2019).

Average smooth cases. For the average smooth cases, the upper and lower bounds nearly match up to log factors for all three cases. Specially, for the nonconvex case, when $\kappa = \Omega(\sqrt{n})$, the lower bound is $\Omega(n + \Delta n^{3/4}\sqrt{\lvert\mu\rvert L}/\varepsilon^2)$ and has been achieved by repeatedSVRG in Agarwal et al. (2017); Carmon et al. (2018); Allen-Zhu (2017) up to log factors.[11] When $\kappa = \mathcal{O}(\sqrt{n})$, the lower bound is $\Omega(n + \Delta L\sqrt{n}/\varepsilon^2)$ and has been achieved by Li et al. (2021). One can check that the algorithms in Allen-Zhu (2018b); Li et al. (2021) are both IFO algorithms. The method repeatedSVRG in Allen-Zhu (2017) can also be modified into IFO algorithms.[12] As for the lower bounds, our results have the same orders as those in Zhou and Gu (2019) and can apply to PIFO algorithms. And our constructions also require fewer dimensions than Zhou and Gu (2019). The details are deferred to Appendix E.3.

IFO and PIFO algorithms. From the above analysis, we find that PIFO oracles are no more powerful than IFO oracles in terms of the complexity for smooth functions. The PIFO lower bounds have been nearly matched by many IFO algorithms. This is consistent with the observation in Woodworth and Srebro (2016). From the results in Table 1, this phenomenon also appears in finite-sum minimax problems under the average smoothness assumption. As a comparison, Woodworth and Srebro (2016) shows that for Lipschitz but nonsmooth functions, having access to proximal oracles does reduce the complexity.

11. This method was implicitly proposed in Agarwal et al. (2017); Carmon et al. (2018) and formally named as repeatedSVRG by Allen-Zhu (2017).
12. Similar to the analysis for catalyst-accelerated methods in Section 3.2.

7. Concluding Remarks

In this paper, focusing on finite-sum minimax and minimization optimization problems, we have given a new definition of PIFO algorithms, which have access to proximal and gradient oracles for each component function and can obtain the full gradient infrequently. This class of PIFO algorithms is large enough to include many near-optimal methods. We have developed a novel approach to constructing the hard instance. Instead of partitioning the variable (Lan and Zhou, 2018; Zhou and Gu, 2019), we partition the classical tridiagonal matrix in Nesterov (2013) into n groups. Such a construction is friendly to the analysis of both IFO and PIFO algorithms, providing some intuition of the optimization process and requiring fewer dimensions than those in Woodworth and Srebro (2016); Zhou and Gu (2019).

Based on our approach, we have established the lower bounds for finite-sum minimax problems when f is convex-concave or nonconvex-strongly-concave and $\{f_i\}_i^n$ is L-average smooth. Most of the lower bounds are nearly matched by existing upper bounds up to log factors. For minimization problems, we have derived similar lower bounds as in Woodworth and Srebro (2016); Hannah et al. (2018); Zhou and Gu (2019). The comparison of upper and lower bounds shows that for smooth functions, the proximal oracles are not much more powerful than gradient oracles.

Finally, we propose several future research directions.

- When f is nonconvex-strongly-concave or each f_i is L-smooth, there still exists some gap between the upper and lower bounds. It remains open to design faster algorithms or tighten the lower bound to close the gap.
- It would be interesting to apply our construction framework to prove the lower bounds for nonconvex-concave cases.
- The definition of PIFO algorithms can be further extended to include more methods. For example, the distribution $\mathcal{D}$ over $[n]$ and the expectation q of the Bernoulli random variable need not be stationary over time. Sampling without replacement and methods that break the linear-span protocol are also worth considering.

Acknowledgments

This work has been supported by the National Natural Science Foundation of China (No. 12350001 and No. 12271011).

Appendix A. Results of the Sum of Geometric Distributions

In this section, we present the approach to proving Lemma 10. We can view the probability $\mathbb{P}\left[\sum_{i=1}^m Y_i > j\right]$ as a function of m variables $p_1, p_2, \ldots, p_m$:

$$f_{m,j}(p_1, p_2, \ldots, p_m) \triangleq \mathbb{P}\left[\sum_{i=1}^m Y_i > j\right]. \tag{16}$$

We first provide the following useful result about the function $f_{m,j}$.

Lemma 38 *For $m \geq 2$ and $j \geq 1$, we have that*

$$f_{m,j}(p_1, p_2, \ldots, p_m) \geq f_{m,j}\left(\frac{\sum_{i=1}^m p_i}{m}, \ldots, \frac{\sum_{i=1}^m p_i}{m}\right).$$

This lemma implies that with the sum of the p_i unchanged, the uniform case (all the p_i are equal) is the least heavy-tailed. Since we aim to give a lower bound, it suffices to only focus on the uniform case. The proof of Lemma 38 is given in Appendix A.1. With Lemma 38 in hand, we give the proof of Lemma 10.

Proof [Proof of Lemma 10]. Let $p = \frac{\sum_{i=1}^m p_i}{m}$ and $\{Z_i \sim \text{Geo}\,(p)\}_{i=1}^m$ be independent geometric random variables. Then we have

$$\mathbb{P}\left[\sum_{i=1}^m Y_i > \frac{m^2}{4(\sum_{i=1}^m p_i)}\right] > \mathbb{P}\left[\sum_{i=1}^m Z_i > \frac{m}{4p}\right].$$

Denote $\sum_{i=1}^m Z_i$ by τ. It is easily checked that $\mathbb{E}[\tau] = \frac{m}{p}$ and $\text{Var}(\tau) = \frac{m(1-p)}{p^2}$. Hence, we have

$$\begin{aligned}\mathbb{P}\left[\tau > \frac{1}{4}\mathbb{E}\tau\right] &= \mathbb{P}\left[\tau - \mathbb{E}\tau > -\frac{3}{4}\mathbb{E}\tau\right] \\ &= 1 - \mathbb{P}\left[\tau - \mathbb{E}\tau \leq -\frac{3}{4}\mathbb{E}\tau\right] \geq 1 - \mathbb{P}\left[|\tau - \mathbb{E}\tau| \geq \frac{3}{4}\mathbb{E}\tau\right] \\ &\geq 1 - \frac{16\text{Var}(\tau)}{9(\mathbb{E}\tau)^2} = 1 - \frac{16m(1-p)}{9m^2} \geq 1 - \frac{16}{9m} \geq \frac{1}{9},\end{aligned}$$

which completes the proof. ■

A.1 Proof of Lemma A.1

Before giving the proof of Lemma 38, we first present some results about $f_{2,j}$. which is defined in Equation (16).

Lemma 39 *The following properties hold for the function $f_{2,j}$.*

1. *For $j \geq 1$, $p_1, p_2 \in (0, 1]$, it holds that*

$$f_{2,j}(p_1, p_2) = \begin{cases} jp_1(1-p_1)^{j-1} + (1-p_1)^j, & \text{if } p_1 = p_2, \\ \frac{p_2(1-p_1)^j - p_1(1-p_2)^j}{p_2 - p_1}, & \text{otherwise.} \end{cases}$$

2. *For $j \geq 2, p_1 \neq p_2$, we have*

$$f_{2,j}(p_1, p_2) > f_{2,j}\left(\frac{p_1 + p_2}{2}, \frac{p_1 + p_2}{2}\right).$$

Proof *1.* Let $Y_1 \sim \text{Geo}(p_1), Y_2 \sim \text{Geo}(p_2)$ be two independent random variables. Then

$$\mathbb{P}[Y_1 + Y_2 > j] = \sum_{l=1}^{j} \mathbb{P}[Y_1 = l]\,\mathbb{P}[Y_2 > j - l] + \mathbb{P}[Y_1 > j]$$

$$= \sum_{l=1}^{j} (1-p_1)^{l-1} p_1 (1-p_2)^{j-l} + (1-p_1)^j$$

$$= p_1(1-p_2)^{j-1} \sum_{l=1}^{j} \left(\frac{1-p_1}{1-p_2}\right)^{l-1} + (1-p_1)^j.$$

If $p_1 = p_2$, Then $\mathbb{P}[Y_1 + Y_2 > j] = jp_1(1-p_1)^{j-1} + (1-p_1)^j$; if $p_1 < p_2$, we have

$$\mathbb{P}[Y_1 + Y_2 > j] = p_1 \frac{(1-p_1)^j - (1-p_2)^j}{p_2 - p_1} + (1-p_1)^j = \frac{p_2(1-p_1)^j - p_1(1-p_2)^j}{p_2 - p_1}.$$

2. Now we suppose that $p_1 + p_2 = c$ and $p_1 < p_2$. Consider

$$h(p_1) \triangleq f_{2,j}(p_1, c - p_1) = \frac{(c-p_1)(1-p_1)^j - p_1(1+p_1-c)^j}{c - 2p_1},$$

where $p_1 \in (0, c/2)$. It is clear that

$$h(c/2) \triangleq \lim_{p_1 \to c/2} h(p_1) = f_{2,j}\left(c/2,\, c/2\right).$$

If $h'(p_1) < 0$ for $p_1 \in (0, c/2)$, then there holds $h(p_1) > h(c/2)$, i.e.,

$$f_{2,j}(p_1, p_2) > f_{2,j}\left(\frac{p_1+p_2}{2}, \frac{p_1+p_2}{2}\right).$$

Note that

$$\begin{aligned} h'(p_1) = &\frac{-(1-p_1)^j - j(c-p_1)(1-p_1)^{j-1} - (1+p_1-c)^j - jp_1(1+p_1-c)^{j-1}}{c-2p_1} \\ &+ 2\frac{(c-p_1)(1-p_1)^j - p_1(1+p_1-c)^j}{(c-2p_1)^2} \\ = &\frac{[c(1-p_1) - j(c-p_1)(c-2p_1)](1-p_1)^{j-1} - [c(1+p_1-c) + jp_1(c-2p_1)](1+p_1-c)^{j-1}}{(c-2p_1)^2}. \end{aligned}$$

Hence $h'(p_1) < 0$ is equivalent to

$$\frac{c(1-p_1) - j(c-p_1)(c-2p_1)}{c(1+p_1-c) + jp_1(c-2p_1)} < \left(\frac{1+p_1-c}{1-p_1}\right)^{j-1}. \tag{17}$$

Observe that

$$\frac{c(1-p_1) - j(c-p_1)(c-2p_1)}{c(1+p_1-c) + jp_1(c-2p_1)} = 1 - \frac{(j-1)c(c-2p_1)}{c(1+p_1-c) + jp_1(c-2p_1)} = 1 - \frac{j-1}{\frac{1+p_1-c}{c-2p_1} + j\frac{p_1}{c}}.$$

Letting $x = \frac{1+p_1-c}{c-2p_1}$, inequality (17) can be written as $1 - \frac{j-1}{x + jp_1/c} < \left(\frac{x}{x+1}\right)^{j-1}$. Note that

$$(x+1)^j - j/2(x+1)^{j-1} = x^j + \sum_{l=0}^{j-1} \left[\binom{j}{l} - \frac{j}{2}\binom{j-1}{l}\right] x^l$$

$$
\begin{aligned}
&= x^j + \sum_{l=0}^{j-1} \left[\left(\frac{j}{j-l} - \frac{j}{2} \right) \binom{j-1}{l} \right] x^l \\
&\le x^j + j/2 x^{j-1} = x^{j-1}(x + j/2).
\end{aligned}
$$

That is $(x+1)^{j-1}(x+j/2) - (j-1)(x+1)^{j-1} \le x^{j-1}(x+j/2)$. Consequently, we have

$$
\left(\frac{x}{x+1} \right)^{j-1} \ge 1 - \frac{j-1}{x+j/2} > 1 - \frac{j-1}{x + jp_1/c},
$$

which is the result we desired. ■

Now we can give the proof of Lemma 38.

Proof [Proof of Lemma 38] We first prove the continuity of the function $f_{m,j}$. Actually, we can prove that

$$
|f_{m,j}(p_1, p_2, \ldots, p_m) - f_{m,j}(p_1', p_2, \ldots, p_m)| \le j|p_1 - p_1'|. \tag{18}
$$

Recall that $f_{m,j}(p_1, p_2, \ldots, p_m) \triangleq \mathbb{P}\left[\sum_{i=1}^m Y_i > j\right]$, where $\{Y_i \sim \mathrm{Geo}(p_i)\}_{i=1}^m$ are independent geometric random variables. Let $Y_1' \sim \mathrm{Geo}(p_1')$ be independent of the Y_i, then by mean value theorem for $1 \le l \le j-1$, there holds

$$
\begin{aligned}
\left| \mathbb{P}\left[Y_1 > l\right] - \mathbb{P}\left[Y_1' > l\right] \right| &= \left| (1-p_1)^l - (1-p_1')^l \right| \\
&= \left| l(1-\xi)^{l-1} \right| \left| p_1 - p_1' \right| \\
&\le l \left| p_1 - p_1' \right| \le j \left| p_1 - p_1' \right|,
\end{aligned}
$$

where ξ lies on the interval $[p_1, p_1']$. Consequently, with $Z \triangleq \sum_{i=2}^m Y_i$, we conclude that

$$
\begin{aligned}
&\left| f_{m,j}(p_1, p_2, \ldots, p_m) - f_{m,j}(p_1', p_2, \ldots, p_m) \right| \\
&= \left| \mathbb{P}\left[Y_1 + Z > j\right] - \mathbb{P}\left[Y_1' + Z > j\right] \right| \\
&= \left| \sum_{l=1}^{j-1} \mathbb{P}\left[Z = l\right] \mathbb{P}\left[Y_1 > j-l\right] + \mathbb{P}\left[Z > j-1\right] - \sum_{l=1}^{j-1} \mathbb{P}\left[Z = l\right] \mathbb{P}\left[Y_1' > j-l\right] + \mathbb{P}\left[Z > j-1\right] \right| \\
&\le \sum_{l=1}^{j-1} \mathbb{P}\left[Z = l\right] \left| \mathbb{P}\left[Y_1 > j-l\right] - \mathbb{P}\left[Y_1' > j-l\right] \right| \\
&\le j|p_1 - p_1'| \sum_{l=1}^{j-1} \mathbb{P}\left[Z = l\right] \\
&= j|p_1 - p_1'| \mathbb{P}\left[1 \le Z \le j-1\right] \le j|p_1 - p_1'|,
\end{aligned}
$$

where we have used $\mathbb{P}\left[Y_1 > 0\right] = 1$ in the second equality.

Following from Equation (18) and the symmetry of the function $f_{m,j}$, we know that

$$
|f_{m,j}(p_1, p_2, \ldots, p_m) - f_{m,j}(p_1', p_2', \ldots, p_m')| \le j \sum_{i=1}^m |p_i - p_i'|,
$$

which implies that $f_{m,j}$ is a continuous function.

Furthermore, following the way we obtain the Equation (18) and the fact that

$$|(1-p_1)^l - 1| \leq lp_1, \quad l = 1, 2, \cdots, j-1,$$

we have $|f_{m,j}(p_1, p_2, \ldots, p_m) - 1| \leq jp_1$. Moreover, by symmetry of the function $f_{m,j}$, it holds that

$$1 - f_{m,j}(p_1, p_2, \ldots, p_m) \leq j \min\{p_1, p_2, \ldots, p_m\}. \tag{19}$$

For $1 \leq j \leq m-1$, we have $f_{m,j}(p_1, p_2, \ldots, p_m) \equiv 1$ and the desired result is apparent. Then Lemma 39 implies the desired result holds for $m = 2$.

For $m \geq 3$, $j \geq m$ and $c \in (0, m)$, our goal is to find the minimal value of $f_{m,j}(p_1, p_2, \ldots, p_m)$ with the domain

$$\mathcal{B} = \left\{ (p_1, p_2, \ldots, p_m) \middle| \sum_{i=1}^m p_m = c,\ p_i \in (0, 1] \text{ for } i \in [m] \right\}.$$

For $j \geq m$, note that

$$\begin{aligned} f_{m,j}(c/m, c/m, \ldots, c/m) &= \mathbb{P}\left[\sum_{i=1}^m Z_i > j\right] \leq \mathbb{P}\left[\sum_{i=1}^m Z_i > m\right] \\ &= 1 - \mathbb{P}\left[\sum_{i=1}^m Z_i \leq m\right] = 1 - \mathbb{P}\left[Z_1 = 1, Z_2 = 1, \cdots, Z_m = 1\right] \\ &= 1 - \left(\frac{c}{m}\right)^m < 1, \end{aligned}$$

where $\{Z_i \sim \mathrm{Geo}(c/m)\}_{i=1}^m$ are independent random variables, and we have used that $\mathbb{P}[Z_i \geq 1] = 1$ for $i \in [m]$.

By Equation (19), if there is an index i satisfies $p_i < \delta \triangleq \frac{1 - f_{m,j}(c/m, c/m, \ldots, c/m)}{j} > 0$, then we have

$$f_{m,j}(p_1, p_2, \ldots, p_m) \geq 1 - jp_i > f_{m,j}(c/m, c/m, \ldots, c/m).$$

Therefore, we just need to find the minimal value of $f_{m,j}(p_1, p_2, \ldots, p_m)$ with the domain

$$\mathcal{B}' = \left\{ (p_1, p_2, \ldots, p_m) \middle| \sum_{i=1}^m p_m = c,\ p_i \in [\delta, 1] \text{ for } i \in [m] \right\},$$

which is a compact set. Hence, by continuity of $f_{m,j}$, we know that there exists $(q_1, q_2, \ldots, q_m) \in \mathcal{B}'$ such that

$$\min_{(p_1, p_2, \ldots, p_m) \in \mathcal{B}'} f_{m,j}(p_1, p_2, \ldots, p_m) = f_{m,j}(q_1, q_2, \ldots, q_m).$$

Suppose that there are indexes $k, l \in [m]$ such that $q_k < q_l$. By symmetry of the function $f_{m,j}$, we assume that $q_1 < q_2$.

Let $\{X_1', X_2'\} \cup \{X_i\}_{i=1}^m$ be independent geometric random variables and $X_1', X_2' \sim \mathrm{Geo}\left(\frac{q_1+q_2}{2}\right)$, $X_i \sim \mathrm{Geo}(q_i)$ for $i \in [m]$. Denoting $Z' = \sum_{i=3}^m X_i$, we have

$$\begin{aligned}
&f_{m,j}(q_1, q_2, \dots, q_m) \\
&= \mathbb{P}\left[X_1 + X_2 + Z' > j\right] \\
&= \sum_{l=1}^{j-1} \mathbb{P}\left[Z' = l\right] \mathbb{P}\left[X_1 + X_2 > j - l\right] + \mathbb{P}\left[Z' > j-1\right] \\
&\geq \sum_{l=1}^{j-1} \mathbb{P}\left[Z' = l\right] \mathbb{P}\left[X_1' + X_2' > j - l\right] + \mathbb{P}\left[Z' > j-1\right] \\
&= \mathbb{P}\left[X_1' + X_2' + Z' > j\right] \\
&= f_{m,j}\left(\frac{q_1+q_2}{2}, \frac{q_1+q_2}{2}, \dots, q_m\right),
\end{aligned}$$

where the inequality is according to Lemma 39.

However, for $l = m-2$, by Lemma 39, it holds that $\mathbb{P}\left[Z' = m-2\right] = 1 - \prod_{i=2}^m q_i > 0$ and $\mathbb{P}\left[X_1 + X_2 > j - m + 2\right] > \mathbb{P}\left[X_1' + X_2' > j - m + 2\right]$, which implies that

$$f_{m,j}(q_1, q_2, \dots, q_m) > f_{m,j}\left(\frac{q_1+q_2}{2}, \frac{q_1+q_2}{2}, \dots, q_m\right).$$

Note that $\frac{q_1+q_2}{2} + \frac{q_1+q_2}{2} + \sum_{i=2}^m q_i = c$ and $\frac{q_1+q_2}{2} \in [\delta, 1]$. Hence we have

$$\left(\frac{q_1+q_2}{2}, \frac{q_1+q_2}{2}, \dots, q_m\right) \in \mathcal{B}',$$

which contradicts the fact that $(q_1, q_2, \dots, q_m)$ is the optimal point in $\mathcal{B}'$.

Therefore, we can conclude that

$$f_{m,j}(p_1, p_2, \dots, p_m) \geq f_{m,j}\left(\frac{\sum_{i=1}^m p_i}{m}, \frac{\sum_{i=1}^m p_i}{m}, \dots, \frac{\sum_{i=1}^m p_i}{m}\right).$$

This completes the proof. ∎

Appendix B. Technical Lemmas

In this section, we present some technical lemmas.

Lemma 40 *Suppose $f(\mathbf{x}, \mathbf{y})$ is (μ_x, μ_y)-convex-concave and L-smooth, then the function $\hat{f}(\mathbf{x}, \mathbf{y}) = \lambda f(\mathbf{x}/\beta, \mathbf{y}/\beta)$ is $\left(\frac{\lambda \mu_x}{\beta^2}, \frac{\lambda \mu_y}{\beta^2}\right)$-convex-concave and $\frac{\lambda L}{\beta^2}$-smooth. If $\{f_i(\mathbf{x}, \mathbf{y})\}_{i=1}^n$ is L'-average smooth, then the function class $\{\hat{f}_i(\mathbf{x}, \mathbf{y}) \triangleq \lambda f_i(\mathbf{x}/\beta, \mathbf{y}/\beta)\}_{i=1}^n$ is $\frac{\lambda L'}{\beta^2}$-average smooth.*

Lemma 41 *Suppose that* $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \le R_x\}$*, then we have*

$$\mathcal{P}_{\mathcal{X}}(\mathbf{x}) = \begin{cases} \mathbf{x}, & \text{if } \mathbf{x} \in \mathcal{X}, \\ \frac{R_X}{\|\mathbf{x}\|_2}\mathbf{x}, & \text{otherwise.} \end{cases}$$

Remark 42 *By Lemma 41, vectors* $\mathcal{P}_{\mathcal{X}}(\mathbf{x})$ *and* $\mathbf{x}$ *are always collinear.*

Proposition 43 (Carmon et al. 2021, Lemmas 2, 3 and 4) *Let* $G_{NC} : \mathbb{R}^{m+1} \to \mathbb{R}$ *be*

$$G_{NC}(\mathbf{x}; \omega, m+1) = \frac{1}{2}\|\mathbf{B}(m+1, \omega, 0)\mathbf{x}\|_2^2 - \omega^2 \langle \mathbf{e}_1, \mathbf{x} \rangle + \omega^4 \sum_{i=1}^{m} \Gamma(x_i).$$

For any $0 < \omega \le 1$*, it holds that*

1. $\Gamma(x)$ *is* 180*-smooth and* $[-45(\sqrt{3}-1)]$*-weakly convex.*
2. $G_{NC}(\mathbf{0}_{m+1}; \omega, m+1) - \min_{\mathbf{x}\in\mathbb{R}^{m+1}} G_{NC}(\mathbf{x}; \omega, m+1) \le \omega^2/2 + 10\omega^4 m.$
3. *For any* $\mathbf{x} \in \mathbb{R}^{m+1}$ *such that* $x_m = x_{m+1} = 0$*,* $G_{NC}(\mathbf{x}; \omega, m)$ *is* $(4 + 180\omega^4)$*-smooth and* $[-45(\sqrt{3}-1)\omega^4]$*-weakly convex and*

$$\|\nabla G_{NC}(\mathbf{x}; \omega, m)\|_2 \ge \omega^3/4.$$

Lemma 44 *Suppose that* $0 < \lambda_2 < (2 + 2\sqrt{2})\lambda_1$*, then* $z = 0$ *is the only real solution to the equation*

$$\lambda_1 z + \lambda_2 \frac{z^2(z-1)}{1+z^2} = 0. \tag{20}$$

Proof Since $0 < \lambda_2 < (2 + 2\sqrt{2})\lambda_1$, we have

$$\lambda_2^2 - 4\lambda_1(\lambda_1 + \lambda_2) < 0,$$

and consequently, for any z, $(\lambda_1 + \lambda_2)z^2 - \lambda_2 z + \lambda_1 > 0$.

On the other hand, we can rewrite Equation (20) as

$$z\big((\lambda_1 + \lambda_2)z^2 - \lambda_2 z + \lambda_1\big) = 0.$$

Clearly, $z = 0$ is the only real solution to Equation (20). ■

Lemma 45 *Suppose that* $0 < \lambda_2 < (2 + 2\sqrt{2})\lambda_1$ *and* $\lambda_3 > 0$*, then* $z_1 = z_2 = 0$ *is the only real solution to the equation*

$$\begin{cases} \lambda_1 z_1 + \lambda_3(z_1 - z_2) + \lambda_2 \frac{z_1^2(z_1-1)}{1+z_1^2} = 0. \\ \lambda_1 z_2 + \lambda_3(z_2 - z_1) + \lambda_2 \frac{z_2^2(z_2-1)}{1+z_2^2} = 0. \end{cases} \tag{21}$$

Proof If $z_1 = 0$, then $z_2 = 0$. So let assume that $z_1 z_2 \neq 0$. Rewrite the first equation of Equations (21) as

$$\frac{\lambda_1 + \lambda_3}{\lambda_3} + \frac{\lambda_2}{\lambda_3}\frac{z_1(z_1 - 1)}{1 + z_1^2} = \frac{z_2}{z_1}$$

Note that

$$\frac{1 - \sqrt{2}}{2} \leq \frac{z(z-1)}{1+z^2}.$$

Thus, we have

$$\frac{\lambda_1 + \lambda_3}{\lambda_3} + \frac{\lambda_2}{\lambda_3}\frac{1 - \sqrt{2}}{2} \leq \frac{z_2}{z_1}.$$

Similarly, it also holds

$$\frac{\lambda_1 + \lambda_3}{\lambda_3} + \frac{\lambda_2}{\lambda_3}\frac{1 - \sqrt{2}}{2} \leq \frac{z_1}{z_2}.$$

By $0 < \lambda_2 < (2 + 2\sqrt{2})\lambda_1$, we know that $\lambda_1 + \frac{1-\sqrt{2}}{2}\lambda_2 > 0$. Thus

$$\frac{\lambda_1 + \lambda_3}{\lambda_3} + \frac{\lambda_2}{\lambda_3}\frac{1 - \sqrt{2}}{2} > 1.$$

Since $z_1/z_2 > 1$ and $z_2/z_1 > 1$ can not hold at the same time, so we get a contradiction. ■

Lemma 46 *Define the function*

$$J_{k,\beta}(y_1, y_2, \ldots, y_k) \triangleq y_k^2 + \sum_{i=2}^{k}(y_i - y_{i-1})^2 + (y_1 - \beta)^2. \tag{22}$$

Then we have $\min J_{k,\beta}(y_1, \ldots, y_k) = \frac{\beta^2}{k+1}$.

Proof Letting the gradient of $J_{k,\beta}$ equal to zero, we get

$$2y_k - y_{k-1} = 0, \quad 2y_1 - y_2 - \beta = 0, \text{ and } y_{i+1} - 2y_i + y_{i-1} = 0, \text{ for } i = 2, 3, \ldots, k-1.$$

That is,

$$y_i = \frac{k - i + 1}{k + 1}\beta \text{ for } i = 1, 2, \ldots, k. \tag{23}$$

Thus by substituting Equation (23) into the expression of $J_{k,\beta}(y_1, y_2, \ldots, y_k)$, we achieve the desired result. ■

Appendix C. Proofs for Section 4

In this section, we present some omitted proofs in Section 4.

C.1 Proofs of Proposition 14 and Lemma 15

Let $\widetilde{\mathbf{B}}(m,\zeta)$ denote the last m rows of $\mathbf{B}(m,0,\zeta)$ and $\tilde{\mathbf{b}}_l(m,\zeta) = \mathbf{b}_l(m,0,\zeta)$ for $0 \le l \le m$. Note that $\tilde{\mathbf{b}}_0(m,\zeta) = \mathbf{0}$. For simplicity, we omit the parameters of $\widetilde{\mathbf{B}}$, $\tilde{\mathbf{b}}_l$ and $\tilde{r}_i$. Then we have $\widetilde{\mathbf{B}} = (\tilde{\mathbf{b}}_1, \tilde{\mathbf{b}}_2, \dots, \tilde{\mathbf{b}}_m)^\top$.

Recall that

$$\mathcal{L}_i = \{l : 0 \le l \le m, l \equiv i-1 (\bmod n)\},\ i = 1,2,\dots,n.$$

For $1 \le i \le n$, let $\widetilde{\mathbf{B}}_i$ be the submatrix of $\widetilde{\mathbf{B}}$ whose rows are $\left\{\tilde{\mathbf{b}}_l^\top\right\}_{l\in\mathcal{L}_i}$. Note that $\widetilde{\mathbf{B}} = \sum_{l=1}^m \mathbf{e}_l\tilde{\mathbf{b}}_l^\top$ and $\widetilde{\mathbf{B}}_i = \sum_{l\in\mathcal{L}_i} \mathbf{e}_l\tilde{\mathbf{b}}_l^\top$. Then $\tilde{r}_i$ can be written as

$$\tilde{r}_i(\mathbf{x},\mathbf{y}) = n\left\langle \mathbf{y}, \widetilde{\mathbf{B}}_i\mathbf{x}\right\rangle + \frac{\tilde{c}_1}{2}\|\mathbf{x}\|_2^2 - \frac{\tilde{c}_2}{2}\|\mathbf{y}\|_2^2 - n\langle \mathbf{e}_1, \mathbf{x}\rangle \mathbb{1}_{\{i=1\}}.$$

Proof [Proof of Proposition 14] Firstly, it is clear that $\tilde{r}_i$ is $(\tilde{c}_1,\tilde{c}_2)$-convex-concave.

Next, note that for $l_1, l_2 \in \mathcal{L}_i$ and $l_1 \ne l_2$, we have $|l_1 - l_2| \ge n \ge 2$, thus $\tilde{\mathbf{b}}_{l_1}^\top\tilde{\mathbf{b}}_{l_2} = 0$. Since $\zeta \le 2$, $\tilde{\mathbf{b}}_l^\top\tilde{\mathbf{b}}_l \le 2$, it follows that

$$\left\|\sum_{l\in\mathcal{L}_i}\tilde{\mathbf{b}}_l\mathbf{e}_l^\top\mathbf{y}\right\|_2^2 = \sum_{l\in\mathcal{L}_i}\mathbf{y}^\top\mathbf{e}_l\tilde{\mathbf{b}}_l^\top\tilde{\mathbf{b}}_l\mathbf{e}_l^\top\mathbf{y} \le 2\sum_{l\in\mathcal{L}_i}\left(\mathbf{e}_l^\top\mathbf{y}\right)^2 \le 2\|\mathbf{y}\|_2^2,$$
$$\left\|\sum_{l\in\mathcal{L}_i}\mathbf{e}_l\tilde{\mathbf{b}}_l^\top\mathbf{x}\right\|_2^2 = \sum_{l\in\mathcal{L}_i}\left(\tilde{\mathbf{b}}_l^\top\mathbf{x}\right)^2 \le \sum_{l\in\mathcal{L}_i\setminus\{m\}} 2\left(x_l^2 + x_{l+1}^2\right) + \zeta^2 x_m^2\mathbb{1}_{\{m\in\mathcal{L}_i\}} \le 2\|\mathbf{x}\|_2^2.$$

Note that

$$\begin{aligned}\nabla_{\mathbf{x}}\tilde{r}_i(\mathbf{x},\mathbf{y}) &= n\widetilde{\mathbf{B}}_i^\top\mathbf{y} + \tilde{c}_1\mathbf{x} - n\mathbf{e}_1\mathbb{1}_{\{i=1\}},\\ \nabla_{\mathbf{y}}\tilde{r}_i(\mathbf{x},\mathbf{y}) &= n\widetilde{\mathbf{B}}_i\mathbf{x} - \tilde{c}_2\mathbf{y}.\end{aligned}$$

With $\mathbf{u} = \mathbf{x}_1 - \mathbf{x}_2$ and $\mathbf{v} = \mathbf{y}_1 - \mathbf{y}_2$, we have

$$\begin{aligned}&\|\nabla\tilde{r}_i(\mathbf{x}_1,\mathbf{y}_1) - \nabla\tilde{r}_i(\mathbf{x}_2,\mathbf{y}_2)\|_2^2\\ =&\|\nabla_{\mathbf{x}}\tilde{r}_i(\mathbf{x}_1,\mathbf{y}_1) - \nabla_{\mathbf{x}}\tilde{r}_i(\mathbf{x}_2,\mathbf{y}_2)\|_2^2 + \|\nabla_{\mathbf{y}}\tilde{r}_i(\mathbf{x}_1,\mathbf{y}_1) - \nabla_{\mathbf{y}}\tilde{r}_i(\mathbf{x}_1,\mathbf{y}_2)\|_2^2\\ =&\left\|\tilde{c}_1\mathbf{u} + n\sum_{l\in\mathcal{L}_i}\tilde{\mathbf{b}}_l\mathbf{e}_l^\top\mathbf{v}\right\|_2^2 + \left\|\tilde{c}_2\mathbf{v} - n\sum_{l\in\mathcal{L}_i}\mathbf{e}_l\tilde{\mathbf{b}}_l^\top\mathbf{u}\right\|_2^2\\ \le& 2\left(\tilde{c}_1^2\|\mathbf{u}\|_2^2 + \tilde{c}_2^2\|\mathbf{v}\|_2^2\right) + 2n^2\left\|\sum_{l\in\mathcal{L}_i}\tilde{\mathbf{b}}_l\mathbf{e}_l^\top\mathbf{v}\right\|_2^2 + 2n^2\left\|\sum_{l\in\mathcal{L}_i}\mathbf{e}_l\tilde{\mathbf{b}}_l^\top\mathbf{u}\right\|_2^2\\ \le& 2\left(\tilde{c}_1^2\|\mathbf{u}\|_2^2 + \tilde{c}_2^2\|\mathbf{v}\|_2^2\right) + 4n^2\sum_{l\in\mathcal{L}_i}\left(\mathbf{e}_l^\top\mathbf{v}\right)^2 + 2n^2\sum_{l\in\mathcal{L}_i}\left(\tilde{\mathbf{b}}_l^\top\mathbf{u}\right)^2\end{aligned}$$

$$\leq \left(2\max\{\tilde{c}_1, \tilde{c}_2\}^2 + 4n^2\right)\left(\|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2\right),$$

where the first inequality follows from $(a+b)^2 \leq 2(a^2+b^2)$. In addition,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \|\nabla \tilde{r}_i(\mathbf{x}_1, \mathbf{y}_1) - \nabla \tilde{r}_i(\mathbf{x}_2, \mathbf{y}_2)\|_2^2 \\
&\leq 2\left(\tilde{c}_1^2 \|\mathbf{u}\|_2^2 + \tilde{c}_2^2 \|\mathbf{v}\|_2^2\right) + 4n\sum_{l=1}^{m}\left(\mathbf{e}_l^\top \mathbf{v}\right)^2 + 2n\sum_{l=1}^{m}\left(\tilde{\mathbf{b}}_l^\top \mathbf{u}\right)^2 \\
&\leq 2\left(\tilde{c}_1^2 \|\mathbf{u}\|_2^2 + \tilde{c}_2^2 \|\mathbf{v}\|_2^2\right) + 4n\|\mathbf{v}\|_2^2 + 8n\|\mathbf{u}\|_2^2 \\
&\leq \left(2\max\{\tilde{c}_1, \tilde{c}_2\}^2 + 8n\right)\left(\|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2\right).
\end{aligned}$$

Thus, $\tilde{r}_i$ is $\sqrt{4n^2 + 2\max\{\tilde{c}_1, \tilde{c}_2\}^2}$-smooth, and $\{\tilde{r}_i\}_{i=1}^n$ is $\sqrt{8n + 2\max\{\tilde{c}_1, \tilde{c}_2\}^2}$-average smooth. ■

Proof [Proof of Lemma 15] Note that

$$\mathbf{e}_l \tilde{\mathbf{b}}_l^\top \mathbf{x} = \begin{cases} (x_l - x_{l+1})\mathbf{e}_l, & 1 \leq l < m, \\ \zeta x_m \mathbf{e}_m, & l = m, \end{cases} \text{ and } \tilde{\mathbf{b}}_l \mathbf{e}_l^\top \mathbf{y} = \begin{cases} y_l(\mathbf{e}_l - \mathbf{e}_{l+1}), & 1 \leq l < m, \\ \zeta y_m \mathbf{e}_m, & l = m. \end{cases}$$

For $\mathbf{x}, \mathbf{y} \in \mathcal{F}_k$ with $1 \leq k < m$, we have

$$\mathbf{e}_l \tilde{\mathbf{b}}_l^\top \mathbf{x} \in \begin{cases} \mathcal{F}_k, & l = k, \\ \mathcal{F}_{k-1}, & l \neq k. \end{cases} \text{ and } \tilde{\mathbf{b}}_l \mathbf{e}_l^\top \mathbf{y} \in \begin{cases} \mathcal{F}_{k+1}, & l = k, \\ \mathcal{F}_k, & l \neq k. \end{cases} \tag{24}$$

Recall that

$$\begin{aligned}
\nabla_{\mathbf{x}} \tilde{r}_i(\mathbf{x}, \mathbf{y}) &= n\sum_{l \in \mathcal{L}_i} \tilde{\mathbf{b}}_l \mathbf{e}_l^\top \mathbf{y} + \tilde{c}_1 \mathbf{x} - n\mathbf{e}_1 \mathbb{1}_{\{i=1\}}, \\
\nabla_{\mathbf{y}} \tilde{r}_i(\mathbf{x}, \mathbf{y}) &= n\sum_{l \in \mathcal{L}_i} \mathbf{e}_l \tilde{\mathbf{b}}_l^\top \mathbf{x} - \tilde{c}_2 \mathbf{y}.
\end{aligned}$$

By Inclusions (24), we have the following results.

1. Suppose that $\mathbf{x}, \mathbf{y} \in \mathcal{F}_0$. It holds that $\nabla_{\mathbf{x}} \tilde{r}_1(\mathbf{x}, \mathbf{y}) = n\mathbf{e}_1 \in \mathcal{F}_1$, $\nabla_{\mathbf{x}} \tilde{r}_j(\mathbf{x}, \mathbf{y}) = \mathbf{0}$ for $j \geq 2$ and $\nabla_{\mathbf{y}} \tilde{r}_j(\mathbf{x}, \mathbf{y}) = \mathbf{0}$ for any j.

2. Suppose that $\mathbf{x} \in \mathcal{F}_1$ and $\mathbf{y} \in \mathcal{F}_0$ and $1 \in \mathcal{L}_i$. It holds that $\nabla_{\mathbf{x}} \tilde{r}_j(\mathbf{x}, \mathbf{y}) = \tilde{c}_1 \mathbf{x} + n\mathbf{e}_1 \mathbb{1}_{\{i=1\}} \in \mathcal{F}_1$ for any j, $\nabla_{\mathbf{y}} \tilde{r}_i(\mathbf{x}, \mathbf{y}) \in \mathcal{F}_1$ and $\nabla_{\mathbf{y}} \tilde{r}_j(\mathbf{x}, \mathbf{y}) = \mathbf{0}$ for $j \neq i$.

3. Suppose that $\mathbf{x} \in \mathcal{F}_{k+1}$, $\mathbf{y} \in \mathcal{F}_k$, $1 \leq k < m$ and $k+1 \in \mathcal{L}_i$. It holds that $\nabla_{\mathbf{x}} \tilde{r}_j(\mathbf{x}, \mathbf{y}) \in \mathcal{F}_{k+1}$ for any j, $\nabla_{\mathbf{y}} \tilde{r}_i(\mathbf{x}, \mathbf{y}) \in \mathcal{F}_{k+1}$ and $\nabla_{\mathbf{y}} \tilde{r}_j(\mathbf{x}, \mathbf{y}) \in \mathcal{F}_k$ for $j \neq i$.

Now we turn to consider $(\mathbf{u}_i, \mathbf{v}_i) = \mathrm{prox}_{\tilde{r}_i}^{\gamma}(\mathbf{x}, \mathbf{y})$. We have

$$\nabla_{\mathbf{x}} \tilde{r}_i(\mathbf{u}_i, \mathbf{v}_i) + \frac{1}{\gamma}(\mathbf{u}_i - \mathbf{x}) = \mathbf{0},$$

$$\nabla_{\mathbf{y}} \tilde{r}_i(\mathbf{u}_i, \mathbf{v}_i) - \frac{1}{\gamma}(\mathbf{v}_i - \mathbf{y}) = \mathbf{0},$$

that is

$$\begin{bmatrix} \left(\tilde{c}_1 + \frac{1}{\gamma}\right)\mathbf{I}_m & n\widetilde{\mathbf{B}}_i^\top \\ -n\widetilde{\mathbf{B}}_i & \left(\tilde{c}_2 + \frac{1}{\gamma}\right)\mathbf{I}_m \end{bmatrix} \begin{bmatrix} \mathbf{u}_i \\ \mathbf{v}_i \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{x}}_i \\ \tilde{\mathbf{y}} \end{bmatrix},$$

where $\tilde{\mathbf{x}}_i = \mathbf{x}/\gamma + n\mathbf{e}_1 \mathbb{1}_{\{i=1\}}$ and $\tilde{\mathbf{y}} = \mathbf{y}/\gamma$. Recall that for $l_1, l_2 \in \mathcal{L}_i$ and $l_1 \neq l_2$, $\tilde{\mathbf{b}}_{l_1}^\top \tilde{\mathbf{b}}_{l_2} = 0$. It follows that

$$\widetilde{\mathbf{B}}_i \widetilde{\mathbf{B}}_i^\top = \left(\sum_{l\in\mathcal{L}_i} \mathbf{e}_l \tilde{\mathbf{b}}_l^\top\right)\left(\sum_{l\in\mathcal{L}_i} \tilde{\mathbf{b}}_l \mathbf{e}_l^\top\right) = \sum_{l\in\mathcal{L}_i} \mathbf{e}_l \tilde{\mathbf{b}}_l^\top \tilde{\mathbf{b}}_l \mathbf{e}_l^\top,$$

which is a diagonal matrix. Assuming that

$$\mathbf{D}_i \triangleq \left(\tilde{c}_2 + \frac{1}{\gamma}\right)\mathbf{I}_m + \frac{n^2}{\tilde{c}_1 + 1/\gamma}\widetilde{\mathbf{B}}_i \widetilde{\mathbf{B}}_i^\top = \operatorname{diag}(d_{i,1}, d_{i,2}, \dots, d_{j,m}),$$

we have

$$\begin{aligned} \begin{bmatrix} \mathbf{u}_i \\ \mathbf{v}_i \end{bmatrix} &= \begin{bmatrix} \left(\tilde{c}_1 + \frac{1}{\gamma}\right)\mathbf{I}_m & n\widetilde{\mathbf{B}}_i^\top \\ -n\widetilde{\mathbf{B}}_i & \left(\tilde{c}_2 + \frac{1}{\gamma}\right)\mathbf{I}_m \end{bmatrix}^{-1} \begin{bmatrix} \tilde{\mathbf{x}}_i \\ \tilde{\mathbf{y}} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{\tilde{c}_1+1/\gamma}\mathbf{I}_m - \frac{n^2}{(\tilde{c}_1+1/\gamma)^2}\widetilde{\mathbf{B}}_i^\top \mathbf{D}_i^{-1}\widetilde{\mathbf{B}}_i & -\frac{n}{\tilde{c}_1+1/\gamma}\widetilde{\mathbf{B}}_i^\top \mathbf{D}_i^{-1} \\ \frac{n}{\tilde{c}_1+1/\gamma}\mathbf{D}_i^{-1}\widetilde{\mathbf{B}}_i & \mathbf{D}_i^{-1} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{x}}_i \\ \tilde{\mathbf{y}} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{\tilde{c}_1+1/\gamma}\tilde{\mathbf{x}}_i - \frac{n^2}{(\tilde{c}_1+1/\gamma)^2}\sum_{l\in\mathcal{L}_i} d_{i.l}^{-1}\tilde{\mathbf{b}}_l\tilde{\mathbf{b}}_l^\top \tilde{\mathbf{x}}_i - \frac{n}{\tilde{c}_1+1/\gamma}\sum_{l\in\mathcal{L}_i}\tilde{\mathbf{b}}_l\mathbf{e}_l^\top \mathbf{D}_i^{-1}\tilde{\mathbf{y}} \\ \frac{n}{\tilde{c}_1+1/\gamma}\sum_{l\in\mathcal{L}_i} d_{i.l}^{-1}\mathbf{e}_l\tilde{\mathbf{b}}_l^\top\tilde{\mathbf{x}}_i + \mathbf{D}_i^{-1}\tilde{\mathbf{y}} \end{bmatrix}. \end{aligned} \tag{25}$$

Note that for $1 \le k \le m$, $\mathbf{y} \in \mathcal{F}_k$ implies $\mathbf{D}_i^{-1}\tilde{\mathbf{y}} \in \mathcal{F}_k$ and $\mathbf{x} \in \mathcal{F}_k$ implies $\tilde{\mathbf{x}}_i \in \mathcal{F}_k$. And recall that

$$\tilde{\mathbf{b}}_l \tilde{\mathbf{b}}_l^\top \mathbf{x} = \begin{cases} (x_l - x_{l+1})(\mathbf{e}_l - \mathbf{e}_{l+1}), & l < m, \\ \zeta^2 x_m \mathbf{e}_m, & l = m. \end{cases}$$

Then for $\mathbf{x} \in \mathcal{F}_k$ with $1 \le k < m$, we have

$$\tilde{\mathbf{b}}_l \tilde{\mathbf{b}}_l^\top \mathbf{x} \in \begin{cases} \mathcal{F}_{k+1}, & l = k, \\ \mathcal{F}_k, & l \neq k. \end{cases} \tag{26}$$

By Inclusions (24), (26) and Equation (25), we have the following results.

1. Suppose that $\mathbf{x}, \mathbf{y} \in \mathcal{F}_0$. It holds that $\tilde{\mathbf{x}}_1 \in \mathcal{F}_1$ and $\tilde{\mathbf{x}}_j = \mathbf{0}$ for $j \ge 2$, which implies $\mathbf{u}_1 \in \mathcal{F}_1$ and $\mathbf{u}_j = \mathbf{0}$ for $j \ge 2$. Moreover, $\mathbf{v}_j = \mathbf{0}$ for any j.

2. Suppose that $\mathbf{x} \in \mathcal{F}_1$, $\mathbf{y} \in \mathcal{F}_0$ and $1 \in \mathcal{L}_i$. It holds that $\mathbf{u}_i \in \mathcal{F}_2$, $\mathbf{v}_i \in \mathcal{F}_1$ and $\mathbf{u}_j \in \mathcal{F}_1$, $\mathbf{v}_j \in \mathcal{F}_0$ for $j \neq i$.

3. Suppose that $\mathbf{x} \in \mathcal{F}_{k+1}$, $\mathbf{y} \in \mathcal{F}_k$, $1 \le k < m-1$ and $k+1 \in \mathcal{L}_i$. It holds that $\mathbf{u}_i \in \mathcal{F}_{k+2}$, $\mathbf{v}_i \in \mathcal{F}_{k+1}$ and $\mathbf{u}_j \in \mathcal{F}_{k+1}$, $\mathbf{v}_j \in \mathcal{F}_k$ for $j \neq i$.

This completes the proof. ■

C.2 Proofs of Corollary 16 and Lemma 17

Proof [Proof of Corollary 16] First, we note that by Lemma 41, the projection operations $\mathcal{P}_{\mathcal{X}}(\mathbf{x})$ and $\mathcal{P}_{\mathcal{Y}}(\mathbf{y})$ do not affect the nonzero elements of the vectors $\mathbf{x}$ and $\mathbf{y}$.

Then we prove the first claim by induction on k. Clearly, it holds that $(\mathbf{x}_0, \mathbf{y}_0) = (\mathbf{0}, \mathbf{0}) \in \mathcal{F}_0 \times \mathcal{F}_{-1}$. Suppose that $(\mathbf{x}_t, \mathbf{y}_t) \in \mathcal{F}_{k(t)-1} \times \mathcal{F}_{k(t)-2}$ for any $t \le t_0$ where $k(t)$ is the positive integer such that $T_{k(t)-1} \le t < T_{k(t)}$. By Lemma 15, for $t < T_{k(t_0)-1}$, $\nabla r_i^{\mathrm{CC}}(\mathbf{x}_t, \mathbf{y}_t), \operatorname{prox}^{\gamma}_{r_i^{\mathrm{CC}}}(\mathbf{x}_t, \mathbf{y}_t) \in \mathcal{F}_{k(t_0)-1} \times \mathcal{F}_{k(t_0)-2}$ for any i; for $T_{k(t_0)-1} \le t \le t_0$, $a_t = 0$ and $\nabla r_{i_j}^{\mathrm{CC}}(\mathbf{x}_t, \mathbf{y}_t), \operatorname{prox}^{\gamma}_{r_{i_j}^{\mathrm{CC}}}(\mathbf{x}_t, \mathbf{y}_t) \in \mathcal{F}_{k(t_0)-1} \times \mathcal{F}_{k(t_0)-2}$ for any $t < j \le t_0$. It remains to check $\nabla r_{i_{t_0+1}}^{\mathrm{CC}}(\mathbf{x}_t, \mathbf{y}_t), \operatorname{prox}^{\gamma}_{r_{i_{t_0+1}}^{\mathrm{CC}}}(\mathbf{x}_t, \mathbf{y}_t)$ for $T_{k(t_0)-1} \le t \le t_0$ and the value of a_{t_0+1}. By Lemma 15,

$$\nabla r_{i_{t_0+1}}^{\mathrm{CC}}(\mathbf{x}_t, \mathbf{y}_t),\ \operatorname{prox}^{\gamma}_{r_{i_{t_0+1}}^{\mathrm{CC}}}(\mathbf{x}_t, \mathbf{y}_t) \in \begin{cases} \mathcal{F}_{k(t_0)} \times \mathcal{F}_{k(t_0)-1}, & \text{if } i_{t_0+1} \equiv k(t_0) \pmod{n}, \\ \mathcal{F}_{k(t_0)-1} \times \mathcal{F}_{k(t_0)-2}, & \text{otherwise.} \end{cases}$$

Thus, if $i_{t_0+1} \equiv k(t_0) \pmod{n}$ or $a_{t_0+1} = 1$, we have $T_{k(t_0)} = t_0 + 1 < T_{k(t_0)+1}$. Thus, $k(t_0+1) = k(t_0) + 1$ and $(\mathbf{x}_{t_0+1}, \mathbf{y}_{t_0+1}) \in \mathcal{F}_{k(t_0+1)-1} \times \mathcal{F}_{k(t_0+1)-2}$. Otherwise, we still have $T_{k(t_0)-1} \le t_0 + 1 < T_{k(t_0)}$. Thus, $k(t_0+1) = k(t_0)$ and $(\mathbf{x}_{t_0+1}, \mathbf{y}_{t_0+1}) \in \mathcal{F}_{k(t_0+1)-1} \times \mathcal{F}_{k(t_0+1)-2}$.

Consequently, we have $(\mathbf{x}_t, \mathbf{y}_t) \in \mathcal{F}_{k(t)-1} \times \mathcal{F}_{k(t)-2}$ for any t. Since $k(t)$ is monotone increasing, we have $(\mathbf{x}_t, \mathbf{y}_t) \in \mathcal{F}_{k-1} \times \mathcal{F}_{k-2}$ for any $t > T_k$ and $k \ge 1$.

Next, note that

$$\begin{aligned} &\mathbb{P}\left[T_k - T_{k-1} = s\right] \\ =&\mathbb{P}\left[i_{T_{k-1}+1} \neq k', \dots, i_{T_{k-1}+s-1} \neq k', a_{T_{k-1}+1} = 0, \dots, a_{T_{k-1}+s-1} = 0, i_{T_{k-1}+s} = k' \text{ or } a_{T_{k-1}+s} = 1\right] \\ =&(1 - p_{k'})^{s-1}(1-q)^{s-1}(p_{k'} + q - p_{k'}q), \end{aligned}$$

where $k' \equiv k (\text{mod } n), 1 \le k' \le n$ and the last equality is due to the independence of $\{(i_t, a_t)\}_{t \ge 1}$. So $Y_k = T_k - T_{k-1}$ is a geometric random variable with success probability $p_{k'} + q - p_{k'}q$. The independence of $\{Y_k\}_{k \ge 1}$ is just according to the independence of $\{(i_t, a_t)\}_{t \ge 1}$. ■

Proof [Proof of Lemma 17] For $t \le N$, we have

$$\begin{aligned} &\mathbb{E}\left(\max_{\mathbf{v} \in \mathcal{Y}} r^{\mathrm{CC}}(\mathbf{x}_t, \mathbf{v}) - \min_{\mathbf{u} \in \mathcal{X}} r^{\mathrm{CC}}(\mathbf{u}, \mathbf{y}_t)\right) \\ \ge\ &\mathbb{E}\left(\max_{\mathbf{v} \in \mathcal{Y}} r^{\mathrm{CC}}(\mathbf{x}_t, \mathbf{v}) - \min_{\mathbf{u} \in \mathcal{X}} r^{\mathrm{CC}}(\mathbf{u}, \mathbf{y}_t) \middle| N < T_{M+1}\right) \mathbb{P}\left[N < T_{M+1}\right] \end{aligned}$$

$$\geq 9\varepsilon \mathbb{P}\left[N < T_{M+1}\right],$$

where T_{M+1} is defined in (6), and the second inequality follows from Corollary 16 (if $N < T_{M+1}$, then $\mathbf{x}_t \in \mathcal{F}_M$ and $\mathbf{y}_t \in \mathcal{F}_{M-1} \subset \mathcal{F}_M$ for $t \leq N$).

By Corollary 16, T_{M+1} can be written as $T_{M+1} = \sum_{l=1}^{M+1} Y_l$, where $\{Y_l\}_{1\leq l \leq M+1}$ are independent random variables, and Y_l follows a geometric distribution with success probability $q_l \triangleq p_{l'} + q - p_{l'}q$ where $l' \equiv l(\text{mod } n)$, $1 \leq l' \leq n$. Moreover, recalling that $p_1 \leq p_2 \leq \cdots \leq p_n$, we have $\sum_{l=1}^{M+1} q_l \leq (M+1)\left(\frac{1}{n} + q\right) \leq (M+1)(1+c_0)/n$. Therefore, by Lemma 10, we have

$$\mathbb{P}\left[T_{M+1} > N\right] = \mathbb{P}\left[\sum_{l=1}^{M+1} Y_l > \frac{(M+1)n}{4(1+c_0)}\right] \geq \frac{1}{9},$$

which implies our desired result. ■

C.3 Proofs of Proposiiton 18 and Lemma 19

Let $\widehat{\mathbf{B}}(m,\omega)$ denote the first m rows of $\mathbf{B}(m,\omega,0)$ by and $\hat{\mathbf{b}}_l(m,\omega) = \mathbf{b}_l(m,\omega,0)$ for $0 \leq l \leq m$. Note that $\hat{\mathbf{b}}_m(m,\omega) = \mathbf{0}$. For simplicity, we omit the parameters of $\widehat{\mathbf{B}}$, $\hat{\mathbf{b}}_l$ and $\hat{r}_i$. Then we have $\widehat{\mathbf{B}} = (\hat{\mathbf{b}}_0, \hat{\mathbf{b}}_1, \ldots, \hat{\mathbf{b}}_{m-1})^\top$.

Let $G(\mathbf{x}) \triangleq \sum\limits_{i=1}^{m-1} \Gamma(x_i)$. Recall that

$$\mathcal{L}_i = \{l : 0 \leq l \leq m, l \equiv i-1(\text{mod} n)\},\ i = 1,2,\ldots,n.$$

For $1 \leq i \leq n$, let $\widehat{\mathbf{B}}_i$ be the submatrix whose rows are $\left\{\hat{\mathbf{b}}_l^\top\right\}_{l\in\mathcal{L}_i}$. Note that $\widehat{\mathbf{B}} = \sum_{l=0}^{m-1} \mathbf{e}_{l+1}\hat{\mathbf{b}}_l^\top$ and $\widehat{\mathbf{B}}_i = \sum_{l\in\mathcal{L}_i} \mathbf{e}_{l+1}\hat{\mathbf{b}}_l^\top$. Then $\hat{r}_i$ can be written as

$$\hat{r}_i(\mathbf{x},\mathbf{y}) = n\left\langle \mathbf{y}, \widehat{\mathbf{B}}_i\mathbf{x}\right\rangle - \frac{\hat{c}_1}{2}\|\mathbf{y}\|_2^2 + \hat{c}_2 G(\hat{c}_3\mathbf{x}) - n\left\langle \mathbf{e}_1, \mathbf{y}\right\rangle \mathbb{1}_{\{i=1\}}.$$

Proof [Proof of Proposition 18] Denote $s_i(\mathbf{x},\mathbf{y}) = \hat{r}_i(\mathbf{x},\mathbf{y}) - \hat{c}_2 G(\hat{c}_3\mathbf{x})$. Similar to the proof of Proposition 14, we can establish that for any $\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}_1, \mathbf{y}_2$,

$$\|\nabla s_i(\mathbf{x}_1,\mathbf{y}_1) - \nabla s_i(\mathbf{x}_2,\mathbf{y}_2)\|_2^2 \leq \left(4n^2 + 2\hat{c}_1^2\right)\left(\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2 + \|\mathbf{y}_1 - \mathbf{y}_2\|_2^2\right),$$

and

$$\frac{1}{n}\sum_{i=1}^n \|\nabla s_i(\mathbf{x}_1,\mathbf{y}_1) - \nabla s_i(\mathbf{x}_2,\mathbf{y}_2)\|_2^2 \leq \left(8n + 2\hat{c}_1^2\right)\left(\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2 + \|\mathbf{y}_1 - \mathbf{y}_2\|_2^2\right).$$

By Proposition 43 and the inequality $(a+b)^2 \leq 2(a^2+b^2)$, we have $\hat{r}_i$ is $\left(-45(\sqrt{3}-1)\hat{c}_2\hat{c}_3^2, \hat{c}_1\right)$-convex-concave,

$$\|\nabla \hat{r}_i(\mathbf{x}_1,\mathbf{y}_1) - \nabla \hat{r}_i(\mathbf{x}_2,\mathbf{y}_2)\|_2 \leq \left(\sqrt{4n^2 + 2\hat{c}_1^2} + 180\hat{c}_2\hat{c}_3^2\right)\sqrt{\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2 + \|\mathbf{y}_1 - \mathbf{y}_2\|_2^2},$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla\hat{r}_i(\mathbf{x}_1,\mathbf{y}_1)-\nabla\hat{r}_i(\mathbf{x}_2,\mathbf{y}_2)\|_2^2 \le \left(16n+4\hat{c}_1^2+64800\hat{c}_2\hat{c}_3^2\right)\left(\|\mathbf{x}_1-\mathbf{x}_2\|_2^2+\|\mathbf{y}_1-\mathbf{y}_2\|_2^2\right).$$

■

Now we prove the Lemma 19.

Proof [Proof of Lemma 19] Note that

$$\mathbf{e}_{l+1}\hat{\mathbf{b}}_l^\top\mathbf{x}=\begin{cases}\omega x_1\mathbf{e}_1, & l=0,\\ (x_l-x_{l+1})\mathbf{e}_{l+1}, & 1\le l<m.\end{cases} \text{ and } \hat{\mathbf{b}}_l\mathbf{e}_{l+1}^\top\mathbf{y}=\begin{cases}\omega y_1\mathbf{e}_1, & l=0,\\ y_{l+1}(\mathbf{e}_l-\mathbf{e}_{l+1}), & 1\le l<m.\end{cases}$$

For $\mathbf{x},\mathbf{y}\in\mathcal{F}_k$ with $1\le k<m$, we have

$$\mathbf{e}_{l+1}\hat{\mathbf{b}}_l^\top\mathbf{x}\in\begin{cases}\mathcal{F}_{k+1}, & l=k,\\ \mathcal{F}_k, & l\ne k.\end{cases} \text{ and } \hat{\mathbf{b}}_l\mathbf{e}_{l+1}^\top\mathbf{y}\in\begin{cases}\mathcal{F}_k, & l=k-1,\\ \mathcal{F}_{k-1}, & l\ne k-1.\end{cases} \tag{27}$$

Recall that

$$\begin{aligned}\nabla_{\mathbf{x}}\hat{r}_i(\mathbf{x},\mathbf{y}) &= n\sum_{l\in\mathcal{L}_i}\hat{\mathbf{b}}_l\mathbf{e}_{l+1}^\top\mathbf{y}+\hat{c}_2\hat{c}_3\nabla G(\hat{c}_3\mathbf{x}),\\ \nabla_{\mathbf{y}}\hat{r}_i(\mathbf{x},\mathbf{y}) &= n\sum_{l\in\mathcal{L}_i}\mathbf{e}_{l+1}\hat{\mathbf{b}}_l^\top\mathbf{x}-\hat{c}_1\mathbf{y}+n\mathbf{e}_1\mathbb{1}_{\{i=1\}}.\end{aligned}$$

By Inclusions (27), we have the following results.

1. Suppose that $\mathbf{x},\mathbf{y}\in\mathcal{F}_0$. It holds that $\nabla_{\mathbf{x}}\hat{r}_j(\mathbf{x},\mathbf{y})=\mathbf{0}$ for any j, $\nabla_{\mathbf{y}}\hat{r}_1(\mathbf{x},\mathbf{y})=n\mathbf{e}_1\in\mathcal{F}_1$ and $\nabla_{\mathbf{y}}\hat{r}_j(\mathbf{x},\mathbf{y})=\mathbf{0}$ for $j\ge 2$.
2. Suppose that $\mathbf{x},\mathbf{y}\in\mathcal{F}_k$, $1\le k<m$ and $k\in\mathcal{L}_i$. It holds that $\nabla_{\mathbf{x}}\hat{r}_j(\mathbf{x},\mathbf{y})\in\mathcal{F}_k$ for any j, $\nabla_{\mathbf{y}}\hat{r}_i(\mathbf{x},\mathbf{y})\in\mathcal{F}_{k+1}$ and $\nabla_{\mathbf{y}}\hat{r}_j(\mathbf{x},\mathbf{y})\in\mathcal{F}_k$ for $j\ne i$.

Now we turn to consider $(\mathbf{u}_i,\mathbf{v}_i)=\operatorname{prox}_{\hat{r}_i}^{\gamma}(\mathbf{x},\mathbf{y})$. We have

$$\begin{aligned}\nabla_{\mathbf{x}}\hat{r}_i(\mathbf{u}_i,\mathbf{v}_i)+\frac{1}{\gamma}(\mathbf{u}_i-\mathbf{x})&=\mathbf{0},\\ \nabla_{\mathbf{y}}\hat{r}_i(\mathbf{u}_i,\mathbf{v}_i)-\frac{1}{\gamma}(\mathbf{v}_i-\mathbf{y})&=\mathbf{0},\end{aligned}$$

that is

$$\begin{bmatrix}\frac{1}{\gamma}\mathbf{I}_m & n\widehat{\mathbf{B}}_i^\top\\ -n\widehat{\mathbf{B}}_i & \left(\hat{c}_1+\frac{1}{\gamma}\right)\mathbf{I}_m\end{bmatrix}\begin{bmatrix}\mathbf{u}_i\\ \mathbf{v}_i\end{bmatrix}=\begin{bmatrix}\hat{\mathbf{x}}-\hat{\mathbf{u}}_i\\ \hat{\mathbf{y}}_i\end{bmatrix},$$

where $\hat{\mathbf{x}}=\mathbf{x}/\gamma$, $\hat{\mathbf{y}}_i=\mathbf{y}/\gamma+n\mathbf{e}_1\mathbb{1}_{\{i=1\}}$ and $\hat{\mathbf{u}}_i=\hat{c}_2\hat{c}_3\nabla G(\hat{c}_3\mathbf{u}_i)$. Recall that for $l_1,l_2\in\mathcal{L}_i$ and $l_1\ne l_2$, $\hat{\mathbf{b}}_{l_1}^\top\hat{\mathbf{b}}_{l_2}=0$. It follows that

$$\widehat{\mathbf{B}}_i\widehat{\mathbf{B}}_i^\top=\left(\sum_{l\in\mathcal{L}_i}\mathbf{e}_{l+1}\hat{\mathbf{b}}_l^\top\right)\left(\sum_{l\in\mathcal{L}_i}\hat{\mathbf{b}}_l\mathbf{e}_{l+1}^\top\right)=\sum_{l\in\mathcal{L}_i}\mathbf{e}_{l+1}\hat{\mathbf{b}}_l^\top\hat{\mathbf{b}}_l\mathbf{e}_{l+1}^\top,$$

which is a diagonal matrix. Denote

$$\mathbf{D}_i \triangleq \left(\hat{c}_1 + \frac{1}{\gamma}\right)\mathbf{I}_m + \gamma n^2 \widehat{\mathbf{B}}_i \widehat{\mathbf{B}}_i^\top = \mathrm{diag}\left(d_{i,1}, d_{i,2}, \ldots, d_{i,m}\right).$$

For $0 < l < m$, $l \in \mathcal{L}_i$ implies $d_{i,l+1} = \hat{c}_1 + \frac{1}{\gamma} + 2\gamma n^2$. Then we have

$$\begin{aligned}
\begin{bmatrix}\mathbf{u}_i \\ \mathbf{v}_i\end{bmatrix} &= \begin{bmatrix}\frac{1}{\gamma}\mathbf{I}_m & n\widehat{\mathbf{B}}_i^\top \\ -n\widehat{\mathbf{B}}_i & \left(\hat{c}_1 + \frac{1}{\gamma}\right)\mathbf{I}_m\end{bmatrix}^{-1} \begin{bmatrix}\hat{\mathbf{x}} - \hat{\mathbf{u}}_i \\ \hat{\mathbf{y}}_i\end{bmatrix} \\
&= \begin{bmatrix}\gamma \mathbf{I}_m - \gamma^2 n^2 \widehat{\mathbf{B}}_i^\top \mathbf{D}_i^{-1}\widehat{\mathbf{B}}_i & -\gamma n \widehat{\mathbf{B}}_i^\top \mathbf{D}_i^{-1} \\ \gamma n \mathbf{D}_i^{-1}\widehat{\mathbf{B}}_i & \mathbf{D}_i^{-1}\end{bmatrix} \begin{bmatrix}\hat{\mathbf{x}} - \hat{\mathbf{u}}_i \\ \hat{\mathbf{y}}_i\end{bmatrix} \\
&= \begin{bmatrix}\gamma(\hat{\mathbf{x}} - \hat{\mathbf{u}}_i) - \gamma^2 n^2 \sum_{l\in\mathcal{L}_i} d_{i.l+1}^{-1}\hat{\mathbf{b}}_l\hat{\mathbf{b}}_l^\top(\hat{\mathbf{x}} - \hat{\mathbf{u}}_i) - \gamma\sum_{l\in\mathcal{L}_i}\hat{\mathbf{b}}_l\mathbf{e}_{l+1}^\top \mathbf{D}_i^{-1}\hat{\mathbf{y}}_i \\ \gamma\sum_{l\in\mathcal{L}_i} d_{i.l+1}^{-1}\mathbf{e}_{l+1}\hat{\mathbf{b}}_l^\top(\hat{\mathbf{x}} - \hat{\mathbf{u}}_i) + \mathbf{D}_i^{-1}\hat{\mathbf{y}}_i\end{bmatrix},
\end{aligned}$$

that is

$$\mathbf{u}_i + \gamma\hat{\mathbf{u}}_i - \gamma^2 n^2 \sum_{l\in\mathcal{L}_i} d_{i.l+1}^{-1}\hat{\mathbf{b}}_l\hat{\mathbf{b}}_l^\top\hat{\mathbf{u}}_i = \gamma\hat{\mathbf{x}} - \gamma^2 n^2 \sum_{l\in\mathcal{L}_i} d_{i.l+1}^{-1}\hat{\mathbf{b}}_l\hat{\mathbf{b}}_l^\top\hat{\mathbf{x}} - \gamma\sum_{l\in\mathcal{L}_i}\hat{\mathbf{b}}_l\mathbf{e}_{l+1}^\top\mathbf{D}_i^{-1}\hat{\mathbf{y}}_i. \quad (28)$$

$$\mathbf{v}_i = \gamma\sum_{l\in\mathcal{L}_i} d_{i.l+1}^{-1}\mathbf{e}_{l+1}\hat{\mathbf{b}}_l^\top(\hat{\mathbf{x}} - \hat{\mathbf{u}}_i) + \mathbf{D}_i^{-1}\hat{\mathbf{y}}_i. \quad (29)$$

We first focus on Equations (28). Recall that $\hat{\mathbf{u}}_i = \hat{c}_2\hat{c}_3\nabla G(\hat{c}_3\mathbf{u}_i)$ and

$$\hat{\mathbf{b}}_l\hat{\mathbf{b}}_l^\top x = \begin{cases}\omega^2 x_1\mathbf{e}_1, & l = 0, \\ (x_l - x_{l+1})(\mathbf{e}_l - \mathbf{e}_{l+1}), & 0 < l < m.\end{cases}$$

For simplicity, let $\mathbf{u}_i = (u_1, u_2, \ldots, u_m)^\top$ and $\hat{\mathbf{u}}_i = (\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_m)^\top$, and denote the right hand side of Equations (28) by $\mathbf{w}$. Recalling the definition of $G(\mathbf{x})$, we have $\hat{u}_l = 120\hat{c}_2\hat{c}_3\frac{\hat{c}_3^2 u_l^2(\hat{c}_3 u_l - 1)}{1 + \hat{c}_3^2 u_l^2}$ for $l < m$ and $\hat{u}_m = 0$. We can establish the following claims.

1. If $0 < l < m-1$ and $l \in \mathcal{L}_i$, we have

$$\begin{aligned}
u_l + \left(\gamma - \gamma^2 n^2 d_{i,l+1}^{-1}\right)\hat{u}_l + \gamma^2 n^2 d_{i,l+1}^{-1}\hat{u}_{l+1} &= w_l, \\
u_{l+1} + \gamma^2 n^2 d_{i,l+1}^{-1}\hat{u}_l + \left(\gamma - \gamma^2 n^2 d_{i,l+1}^{-1}\right)\hat{u}_{l+1} &= w_{l+1}.
\end{aligned} \quad (30)$$

   Setting $w_l = w_{l+1} = 0$ yields

$$\begin{aligned}
\left(1 - 2\gamma n^2 d_{i,l+1}^{-1}\right)u_l + \gamma n^2 d_{i,l+1}^{-1}(u_l - u_{l+1}) + \left(\gamma - 2\gamma^2 n^2 d_{i,l+1}^{-1}\right)\hat{u}_l &= 0, \\
\left(1 - 2\gamma n^2 d_{i,l+1}^{-1}\right)u_{l+1} + \gamma n^2 d_{i,l+1}^{-1}(u_{l+1} - u_l) + \left(\gamma - 2\gamma^2 n^2 d_{i,l+1}^{-1}\right)\hat{u}_{l+1} &= 0.
\end{aligned}$$

   Recalling that $d_{i,l+1} = \hat{c}_1 + 1/\gamma + 2\gamma n^2$, we find $1 - 2\gamma n^2 d_{i,l+1}^{-1} > 0$. Since $\gamma < \frac{\sqrt{2}+1}{60\hat{c}_2\hat{c}_3^2}$, we can apply Lemma 45 with $z_1 = \hat{c}_3 u_l$ and $z_2 = \hat{c}_3 u_{l+1}$ and conclude that $u_l = u_{l+1} = 0$.

2. If $m-1 \in \mathcal{L}_i$, we have

$$\begin{aligned} u_{m-1} + \left(\gamma - \gamma^2 n^2 d_{i,m}^{-1}\right) \hat{u}_{m-1} &= w_{m-1}, \\ u_m + \gamma^2 n^2 d_{i,m}^{-1} \hat{u}_{m-1} &= w_m. \end{aligned} \tag{31}$$

Setting $w_{m-1} = w_m = 0$ yields

$$\begin{aligned} u_{m-1} + \left(\gamma - \gamma^2 n^2 d_{i,m}^{-1}\right) \hat{u}_{m-1} &= 0, \\ \gamma n^2 d_{i,m}^{-1} u_{m-1} - \left(1 - \gamma n^2 d_{i,m}^{-1}\right) u_m &= 0. \end{aligned}$$

Recalling that $d_{i,l+1} = \hat{c}_1 + 1/\gamma + 2\gamma n^2$ and $\gamma < \frac{\sqrt{2}+1}{60\hat{c}_2\hat{c}_3^2}$, we have $0 < \gamma - \gamma^2 n^2 d_{i,m}^{-1} < \gamma < \frac{\sqrt{2}+1}{60\hat{c}_2\hat{c}_3^2}$. Applying Lemma 44 with $z = \hat{c}_3 u_{m-1}$, we conclude that $u_{m-1} = 0$. It follows that $u_m = 0$.

3. If $0 < l < m$ and $l, l-1 \notin \mathcal{L}_i$, we have

$$u_l + \gamma \hat{u}_l = w_l. \tag{32}$$

Setting $w_l = 0$ and applying Lemma 44 with $z = \hat{c}_3 u_l$, we conclude that $u_l = 0$.

Note that for $1 \leq k \leq m$, $\mathbf{x} \in \mathcal{F}_k$ implies $\hat{\mathbf{x}} \in \mathcal{F}_k$ and $\mathbf{y} \in \mathcal{F}_k$ implies $\mathbf{D}_i^{-1}\hat{\mathbf{y}}_i \in \mathcal{F}_k$. And for $\mathbf{x} \in \mathcal{F}_k$ with $1 \leq k < m$, we have

$$\hat{\mathbf{b}}_l \hat{\mathbf{b}}_l^\top x \in \begin{cases} \mathcal{F}_{k+1}, & l = k, \\ \mathcal{F}_k, & l \neq k. \end{cases} \tag{33}$$

Then we can provide the following analysis.

1. Suppose that $\mathbf{x}, \mathbf{y} \in \mathcal{F}_0$. Note that $0 \in \mathcal{L}_1$.

   For $j = 1$, we have $\hat{\mathbf{x}} = \mathbf{0}$ and $\hat{\mathbf{y}}_1 \in \mathcal{F}_1$. Since $0 \in \mathcal{L}_1$, Inclusion (27) implies $\mathbf{w} \in \mathcal{F}_1$. Then we consider the solution to Equations (28). Since $n \geq 2$, we have $1 \notin \mathcal{L}_1$. If $2 \in \mathcal{L}_1$, we can consider the solution to Equations (30) or (31) and conclude that $u_2 = 0$. If $2 \notin \mathcal{L}_1$, we can consider the solution to Equation (32) and conclude that $u_2 = 0$. Similarly, we obtain $u_l = 0$ for $l \geq 2$, which implies $\mathbf{u}_1 \in \mathcal{F}_1$. Since $1 \notin \mathcal{L}_1$, by Inclusion (27) and Equations (29), we have $\mathbf{v}_1 \in \mathcal{F}_1$.

   For $j \neq 1$, we have $\hat{\mathbf{x}} = \hat{\mathbf{y}}_j = \mathbf{0}$. It follows that $\mathbf{w} = \mathbf{0}$. Note that $0 \notin \mathcal{L}_j$. If $1 \in \mathcal{L}_j$, we can consider the solution to Equations (30) or (31) and conclude that $u_1 = 0$. If $1 \notin \mathcal{L}_j$, we can consider the solution to Equation (32) and conclude that $u_1 = 0$. Similarly, we obtain $u_l = 0$ for all l, which implies $\mathbf{u}_j = \mathbf{0}$. By Equations (29), we have $\mathbf{v}_j = \mathbf{0}$.

2. Suppose that $\mathbf{x}, \mathbf{y} \in \mathcal{F}_k$, $1 \leq k < m$ and $k \in \mathcal{L}_i$.

   For $j = i$, we have $\hat{\mathbf{x}}, \hat{\mathbf{y}}_i \in \mathcal{F}_k$. If $k = m-1$, clearly $\mathbf{u}_i, \mathbf{v}_i \in \mathcal{F}_m$. Now we assume $k < m-1$. Inclusions (33) and (27) imply $\mathbf{w} \in \mathcal{F}_{k+1}$. Then we consider the solution to Equations (28). Since $n \geq 2$, we have $k+1 \notin \mathcal{L}_i$. If $k+2 \in \mathcal{L}_i$, we can consider the solution to Equations (30) or (31) and conclude that $u_{k+2} = 0$. If $k+2 \notin \mathcal{L}_i$, we can consider the solution to Equation (32) and conclude that $u_{k+2} = 0$. Similarly, we obtain $u_l = 0$ for $l \geq k+2$, which implies $\mathbf{u}_i \in \mathcal{F}_{k+1}$. Since $k+1 \notin \mathcal{L}_i$, by Inclusion (27) and Equations (29), we have $\mathbf{v}_i \in \mathcal{F}_{k+1}$.

For $j \neq i$, we also have $\hat{\mathbf{x}}, \hat{\mathbf{y}}_i \in \mathcal{F}_k$. Since $k \notin \mathcal{L}_j$, by Inclusions (27) and (33), we have $\mathbf{w} \in \mathcal{F}_k$. If $k+1 \in \mathcal{L}_j$, we can consider the solution to Equations (30) or (31) and conclude that $u_{k+1} = 0$. If $k+1 \notin \mathcal{L}_j$, we can consider the solution to Equation (32) and conclude that $u_{k+1} = 0$. Similarly, we obtain $u_l = 0$ for $l \geq k+1$, which implies $\mathbf{u}_j \in \mathcal{F}_k$. Since $k \notin \mathcal{L}_j$, by Inclusion (27) and Equations (29), we have $\mathbf{v}_j \in \mathcal{F}_k$.

This completes the proof. ■

Appendix D. Proofs for Section 5

In this section, we present the omitted proofs in Section 5.

D.1 Proofs for the Strongly-Convex-Strongly-Concave Case

With f_{SCSC} and $\{f_{\mathrm{SCSC},i}\}_{i=1}^n$ defined in Definition 26, we have the following proposition.

Proposition 47 *For any $n \geq 2$, $m \geq 2$, $f_{\mathrm{SCSC},i}$ and f_{SCSC} in Definition 26 satisfy:*

1. *$\{f_{\mathrm{SCSC},i}\}_{i=1}^n$ is L-average smooth and each $f_{\mathrm{SCSC},i}$ is (μ_x, μ_y)-convex-concave. Thus, f_{SCSC} is (μ_x, μ_y)-convex-concave.*

2. *The saddle point of Problem (10) is*

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

   where $q = \frac{\alpha-1}{\alpha+1}$. Moreover, $\|\mathbf{x}^*\|_2 \leq R_x$, $\|\mathbf{y}^*\|_2 \leq R_y$.

3. *For $1 \leq k \leq m-1$, we have*

$$\min_{\mathbf{x} \in \mathcal{X} \cap \mathcal{F}_k} \phi_{\mathrm{SCSC}}(\mathbf{x}) - \max_{\mathbf{y} \in \mathcal{Y} \cap \mathcal{F}_k} \psi_{\mathrm{SCSC}}(\mathbf{y}) \geq \frac{\beta^2\left(L^2 - 2\mu_y^2\right)}{8n(\alpha+1)\mu_x} q^{2k}.$$

Proof

1. Just recall Proposition 14 and Lemma 40.

2. It is easy to check $f_{\mathrm{SCSC}}(\mathbf{x}, \mathbf{y}) = \xi\left\langle \mathbf{y}, \widetilde{\mathbf{B}}(m, \zeta)\mathbf{x}\right\rangle + \frac{\mu_x}{2}\|\mathbf{x}\|_2^2 - \frac{\mu_y}{2}\|\mathbf{y}\|_2^2 - \beta\xi\langle \mathbf{e}_1, \mathbf{x}\rangle$, where $\zeta = \sqrt{\frac{2}{\alpha+1}}$ and $\xi = \lambda/\beta^2 = \frac{1}{2}\sqrt{\frac{L^2-2\mu_y^2}{2n}}$. Letting the gradient of $f_{\mathrm{SCSC}}(\mathbf{x}, \mathbf{y})$ be zero, we obtain

$$\mathbf{y} = \frac{\xi}{\mu_y}\widetilde{\mathbf{B}}(m, \zeta)\mathbf{x}, \quad \left(\mu_x \mathbf{I} + \frac{\xi^2}{\mu_y}\widetilde{\mathbf{B}}(m, \zeta)^\top \widetilde{\mathbf{B}}(m, \zeta)\right)\mathbf{x} = \beta\xi\mathbf{e}_1. \tag{34}$$

Note that

$$\frac{\mu_x\mu_y}{\xi^2} = \frac{8n\mu_x\mu_y}{L^2-2\mu_y^2} = \frac{8n\mu_x}{(\kappa_y^2-2)\mu_y} = \frac{8n}{(\kappa_y - 2/\kappa_y)\,\kappa_x} = \frac{4}{\alpha^2-1}.$$

One can check q is a root of the equation $z^2-\left(2+\frac{\mu_x\mu_y}{\xi^2}\right)z+1=0$. By some calculation, the solution of (34) equation is

$$\mathbf{x}^* = \frac{\beta\mu_y}{(1-q)\xi}(q, q^2, \ldots, q^m)^\top, \quad \mathbf{y}^* = \beta\left(q, q^2, \ldots, q^{m-1}, \frac{q^m}{\zeta}\right)^\top.$$

Moreover, from the definition of β, we have

$$\|\mathbf{x}^*\|_2^2 = \frac{\beta^2\mu_y^2\left(q^2-q^{2m+2}\right)}{(1-q)^2\left(1-q^2\right)\xi^2} \le \frac{\beta^2\mu_y^2 q^2}{(1-q)^2\left(1-q^2\right)\xi^2} \le \frac{\beta^2\kappa_x^2(1-2/\kappa_y^2)}{8n\alpha} \le R_x^2,$$

and

$$\|\mathbf{y}^*\|_2^2 = \beta^2\left(\frac{q^2-q^{2m}}{1-q^2}+\frac{q^{2m}}{\zeta^2}\right) = \beta^2\frac{q^2+q^{2m+1}}{1-q^2} \le \beta^2\frac{2q^2}{1-q^2} = \beta^2\frac{(\alpha-1)^2}{4\alpha} \le R_y^2.$$

3. Define $\tilde{\phi}_{\mathrm{SCSC}}(\mathbf{x}) = \max_{\mathbf{y}\in\mathbb{R}^m} f_{\mathrm{SCSC}}(\mathbf{x},\mathbf{y})$ and $\tilde{\psi}_{\mathrm{SCSC}}(\mathbf{y}) = \min_{\mathbf{x}\in\mathbb{R}^m} f_{\mathrm{SCSC}}(\mathbf{x},\mathbf{y})$. We first show that

$$\min_{\mathbf{x}\in\mathcal{F}_k}\tilde{\phi}_{\mathrm{SCSC}}(\mathbf{x}) - \max_{\mathbf{y}\in\mathcal{F}_k}\tilde{\psi}_{\mathrm{SCSC}}(\mathbf{y}) \ge \frac{\beta^2\xi^2}{(\alpha+1)\mu_x}q^{2k},$$

where $\xi = \frac{1}{2}\sqrt{\frac{L^2-2\mu_y^2}{2n}}$. Recall that $f_{\mathrm{SCSC}}(\mathbf{x},\mathbf{y}) = \xi\left\langle \mathbf{y}, \widetilde{\mathbf{B}}(m,\zeta)\,\mathbf{x}\right\rangle + \frac{\mu_x}{2}\|\mathbf{x}\|_2^2 - \frac{\mu_y}{2}\|\mathbf{y}\|_2^2 - \beta\xi\langle\mathbf{e}_1,\mathbf{x}\rangle$, where $\zeta = \sqrt{\frac{2}{\alpha+1}}$. Then we can rewrite $f_{\mathrm{SCSC}}(\mathbf{x},\mathbf{y})$ as

$$f_{\mathrm{SCSC}}(\mathbf{x},\mathbf{y}) = -\frac{\mu_y}{2}\left\|\mathbf{y}-\frac{\xi}{\mu_y}\widetilde{\mathbf{B}}(m,\zeta)\mathbf{x}\right\|_2^2 + \frac{\xi^2}{2\mu_y}\left\|\widetilde{\mathbf{B}}(m,\zeta)\mathbf{x}\right\|_2^2 + \frac{\mu_x}{2}\|\mathbf{x}\|_2^2 - \beta\xi\langle\mathbf{e}_1,\mathbf{x}\rangle. \tag{35}$$

Thus $\tilde{\phi}_{\mathrm{SCSC}}(\mathbf{x}) = \frac{\xi^2}{2\mu_y}\left\|\widetilde{\mathbf{B}}(m,\zeta)\mathbf{x}\right\|_2^2 + \frac{\mu_x}{2}\|\mathbf{x}\|_2^2 - \beta\xi\langle\mathbf{e}_1,\mathbf{x}\rangle$. For $\mathbf{x}\in\mathcal{F}_k$, let $\tilde{\mathbf{x}}$ be the first k coordinates of $\mathbf{x}$. Then we can rewrite $\tilde{\phi}_{\mathrm{SCSC}}$ as $\tilde{\phi}_k(\tilde{\mathbf{x}}) \triangleq \tilde{\phi}_{\mathrm{SCSC}}(\mathbf{x}) = \frac{\xi^2}{2\mu_y}\left\|\widetilde{\mathbf{B}}(k,1)\tilde{\mathbf{x}}\right\|_2^2 + \frac{\mu_x}{2}\|\tilde{\mathbf{x}}\|_2^2 - \beta\xi\langle\mathbf{e}_1,\tilde{\mathbf{x}}\rangle$, where $\hat{\mathbf{e}}_1$ is the first k coordinates of $\mathbf{e}_1$. Letting $\nabla\tilde{\phi}_k(\tilde{\mathbf{x}}) = \mathbf{0}_k$, we obtain

$$\frac{\xi^2}{\mu_y}\widetilde{\mathbf{B}}(k,1)^\top\widetilde{\mathbf{B}}(k,1)\tilde{\mathbf{x}} + \mu_x\tilde{\mathbf{x}} = \beta\xi\hat{\mathbf{e}}_1. \tag{36}$$

Recall that $\frac{\mu_x\mu_y}{\xi^2} = \frac{4}{\alpha^2-1}$ and $q = \frac{\alpha-1}{\alpha+1}$. One can check q and $1/q$ are two roots of the equation $z^2-\left(2+\frac{\mu_x\mu_y}{\xi^2}\right)z+1=0$. By some calculations, the solution to Equations (36) is

$$\tilde{\mathbf{x}}^* = \frac{\beta\mu_y(\alpha+1)q^{k+1}}{2\xi\left(1+q^{2k+1}\right)}\left(q^{-k}-q^k, q^{-k+1}-q^{k-1}, \ldots, q^{-1}-q\right)^\top,$$

and the value of $\min_{\mathbf{x}\in\mathcal{F}_k} \tilde{\phi}_{\text{SCSC}}(\mathbf{x})$ is $\min_{\mathbf{x}\in\mathcal{F}_k} \tilde{\phi}_{\text{SCSC}}(\mathbf{x}) = -\frac{\beta^2\mu_y(\alpha+1)}{4}\frac{q-q^{2k+1}}{1+q^{2k+1}}$.

On the other hand, observe that

$$f_{\text{SCSC}}(\mathbf{x},\mathbf{y}) = \frac{\mu_x}{2}\left\|\mathbf{x}+\frac{\xi}{\mu_x}\widetilde{\mathbf{B}}(m,\zeta)^\top\mathbf{y}-\frac{\beta\xi}{\mu_x}\mathbf{e}_1\right\|_2^2 - \frac{\xi^2}{2\mu_x}\left\|\widetilde{\mathbf{B}}(m,\zeta)^\top\mathbf{y}-\beta\mathbf{e}_1\right\|_2^2 - \frac{\mu_y}{2}\|\mathbf{y}\|_2^2. \tag{37}$$

It follows that $\tilde{\psi}_{\text{SCSC}}(\mathbf{y}) = -\frac{\xi^2}{2\mu_x}\left\|\widetilde{\mathbf{B}}(m,\zeta)^\top\mathbf{y}-\beta\mathbf{e}_1\right\|_2^2 - \frac{\mu_y}{2}\|\mathbf{y}\|_2^2$. For $\mathbf{y}\in\mathcal{F}_k$, let $\tilde{\mathbf{y}}$ be the first k coordinated of $\mathbf{y}$. Then we can rewrite $\tilde{\psi}_{\text{SCSC}}$ as $\tilde{\psi}_k(\tilde{\mathbf{y}}) \triangleq \tilde{\psi}_{\text{SCSC}}(\mathbf{y}) = -\frac{\xi^2}{2\mu_x}\left\|\widetilde{\mathbf{B}}(k,1)^\top\tilde{\mathbf{y}}-\beta\hat{\mathbf{e}}_1\right\|_2^2 - -\frac{\xi^2}{2\mu_x}\langle\hat{\mathbf{e}}_k,\tilde{\mathbf{y}}\rangle^2 - \frac{\mu_y}{2}\|\tilde{\mathbf{y}}\|_2^2$, where $\hat{\mathbf{e}}_1,\hat{\mathbf{e}}_k$ are the first k ordinates of $\mathbf{e}_1$ and $\mathbf{e}_k$ respectively. Letting $\nabla\tilde{\psi}_k(\tilde{\mathbf{y}}) = \mathbf{0}_k$, we obtain

$$\frac{\xi^2}{\mu_x}\left(\widetilde{\mathbf{B}}(k,1)\widetilde{\mathbf{B}}(k,1)^\top + \hat{\mathbf{e}}_k\hat{\mathbf{e}}_k^\top\right)\tilde{\mathbf{y}} + \mu_y\tilde{\mathbf{y}} = \frac{\beta\xi^2}{\mu_x}\widetilde{\mathbf{B}}(k,1)\hat{\mathbf{e}}_1.$$

Then, we can check that the solution to the above equations is

$$\tilde{\mathbf{y}}^* = \frac{\beta q^{k+1}}{1-q^{2k+2}}(q^{-k}-q^k, q^{-k+1}-q^{k-1},\ldots,q^{-1}-q)^\top,$$

and the optimal value of $\tilde{\psi}_{\text{SCSC}}(\mathbf{y})$ is $\min_{\mathbf{y}\in\mathcal{F}_k} \tilde{\psi}_{\text{SCSC}}(\mathbf{y}) = -\frac{\beta^2\xi^2}{\mu_x(\alpha+1)}\frac{1+q^{2k+1}}{1-q^{2k+2}}$. It follows that

$$\begin{aligned}
&\min_{\mathbf{x}\in\mathcal{F}_k}\tilde{\phi}_{\text{SCSC}}(\mathbf{x}) - \max_{\mathbf{y}\in\mathcal{F}_k}\tilde{\psi}_{\text{SCSC}}(\mathbf{y}) \\
&= -\frac{\beta^2\mu_y(\alpha+1)}{4}\frac{q-q^{2k+1}}{1+q^{2k+1}} + \frac{\beta^2\xi^2}{\mu_x(\alpha+1)}\frac{1+q^{2k+1}}{1-q^{2k+2}} \\
&= -\frac{\beta^2\xi^2}{\mu_x(\alpha+1)}\frac{\mu_x\mu_y(\alpha+1)^2q}{4\xi^2}\frac{1-q^{2k}}{1+q^{2k+1}} + \frac{\beta^2\xi^2}{\mu_x(\alpha+1)}\frac{1+q^{2k+1}}{1-q^{2k+2}} \\
&= \frac{\beta^2\xi^2}{\mu_x(\alpha+1)}\left(\frac{1+q^{2k+1}}{1-q^{2k+2}} - \frac{1-q^{2k}}{1+q^{2k+1}}\right) \\
&= \frac{\beta^2\xi^2}{\mu_x(\alpha+1)}\frac{2q^{2k+1}+q^{2k}+q^{2k+2}}{(1-q^{2k+2})(1+q^{2k+1})} \geq \frac{\beta^2\xi^2}{\mu_x(\alpha+1)}q^{2k}.
\end{aligned}$$

Clearly, we have $\min_{\mathbf{x}\in\mathcal{X}\cap\mathcal{F}_k}\phi_{\text{SCSC}}(\mathbf{x}) - \max_{\mathbf{y}\in\mathcal{Y}\cap\mathcal{F}_k}\psi_{\text{SCSC}}(\mathbf{y}) \geq \min_{\mathbf{x}\in\mathcal{F}_k}\phi_{\text{SCSC}}(\mathbf{x}) - \max_{\mathbf{y}\in\mathcal{F}_k}\psi_{\text{SCSC}}(\mathbf{y})$. It remains to show that $\min_{\mathbf{x}\in\mathcal{F}_k}\phi_{\text{SCSC}}(\mathbf{x}) = \min_{\mathbf{x}\in\mathcal{F}_k}\tilde{\phi}_{\text{SCSC}}(\mathbf{x})$ and $\max_{\mathbf{y}\in\mathcal{F}_k}\psi_{\text{SCSC}}(\mathbf{y}) = \max_{\mathbf{y}\in\mathcal{F}_k}\tilde{\psi}_{\text{SCSC}}(\mathbf{y})$. Recall the expressions (35) and (37). It suffices to prove $\|\hat{\mathbf{x}}\|_2 \leq R_x$ and $\|\hat{\mathbf{y}}\|_2 \leq R_y$ where

$$\hat{\mathbf{x}} = -\frac{\xi}{\mu_x}\widetilde{\mathbf{B}}(m,\zeta)\top\begin{bmatrix}\tilde{\mathbf{y}}^*\\ \mathbf{0}_{m-k}\end{bmatrix} + \frac{\beta\xi}{\mu_x}\mathbf{e}_1, \quad \hat{\mathbf{y}} = \frac{\xi}{\mu_y}\widetilde{\mathbf{B}}(m,\zeta)\begin{bmatrix}\tilde{\mathbf{x}}^*\\ \mathbf{0}_{m-k}\end{bmatrix}.$$

By some calculation, we have

$$\|\hat{\mathbf{x}}\|_2^2 = \frac{\beta^2\xi^2(1-q)^2}{\mu_x^2(1-q^{2k+2})^2}\left(\frac{1-q^{4k+2}}{1-q^2} + 2(k+1)q^{2k+1}\right),$$

$$\|\hat{\mathbf{y}}\|_2^2 = \frac{\beta^2}{(1+q^{2k+1})^2}\left(\frac{q^2 - q^{4k+2}}{1-q^2} + 2kq^{2k+1}\right).$$

Note that $\max_{x>0} xq^x = \log\frac{1}{q}\, e^{-\left(\log\frac{1}{q}\right)^2}$ and $\log r - r^2 \le -r$ for any $r > 0$. It follows that $\max_{x>0} xq^x \le e^{-\log\frac{1}{q}} = q$. Then we have

$$\|\hat{\mathbf{x}}\|_2^2 \le \frac{\beta^2\xi^2(1-q)^2}{\mu_x^2(1-q)^2}\left(\frac{1}{1-q^2}+1\right) \le \frac{2\beta^2\xi^2}{\mu_x^2(1-q^2)} = \frac{\beta^2\left(L^2-2\mu_x^2\right)(\alpha+1)^2}{16n\mu_y^2\alpha} \le R_x^2,$$

$$\|\hat{\mathbf{y}}\|_2^2 \le \beta^2\left(\frac{q^2}{1-q^2}+q^2\right) \le \frac{2\beta^2 q^2}{1-q^2} \le \frac{\beta^2(\alpha-1)^2}{2\alpha} \le R_y^2.$$

This completes the proof. ■

Proof [Proof of Theorem 27] Let $q = \frac{\alpha-1}{\alpha+1}$. For $\kappa_x \ge \kappa_y \ge \sqrt{2n+2}$, we have $\alpha = \sqrt{\frac{(\kappa_y - 2/\kappa_y)\kappa_x}{2n}+1} \ge \sqrt{2}$, $q = \frac{\alpha-1}{\alpha+1} \ge \frac{\sqrt{2}-1}{\sqrt{2}+1}$ and $\kappa_y - 2/\kappa_y \ge \kappa_y/2$.

Let $M = \left\lfloor \frac{\log\left(9(\alpha+1)\mu_x\varepsilon/\beta^2\xi^2\right)}{2\log q} \right\rfloor$ where $\xi = \frac{1}{2}\sqrt{\frac{L^2-2\mu_y^2}{2n}}$. Then we have

$$\min_{\mathbf{x}\in\mathcal{X}\cap\mathcal{F}_M} \phi_{\text{SCSC}}(\mathbf{x}) - \max_{\mathbf{y}\in\mathcal{Y}\cap\mathcal{F}_M} \psi_{\text{SCSC}}(\mathbf{y}) \ge \frac{\beta^2\xi^2}{(\alpha+1)\mu_x} q^{2M} \ge 9\varepsilon.$$

where the first inequality follows from the third property of Proposition 47.

First, we need to ensure $1 \le M < m$. Note that $M \ge 1$ is equivalent to $\varepsilon \le \frac{q^2\beta^2\xi^2}{9(\alpha+1)\mu_x}$. Recall that

$$\beta = \min\left\{2R_x\sqrt{\frac{2\alpha n}{\kappa_x^2(1-2/\kappa_y^2)}},\ \frac{4R_x}{\alpha+1}\sqrt{\frac{\alpha n}{\kappa_x^2(1-2/\kappa_y^2)}},\ \frac{\sqrt{2\alpha}R_y}{\alpha-1}\right\}.$$

When $\beta = 2R_x\sqrt{\frac{2\alpha n}{\kappa_x^2(1-2/\kappa_y^2)}}$, noticing that $\frac{\alpha(\alpha-1)^2}{(\alpha+1)^3}$ is increasing for $\alpha > 1$, we have

$$\frac{q^2\beta^2\xi^2}{9(\alpha+1)\mu_x} = \frac{\alpha(\alpha-1)^2}{9(\alpha+1)^3}\mu_x R_x^2 \ge \frac{\sqrt{2}\left(\sqrt{2}-1\right)^5}{9}\mu_x R_x^2,$$

When $\beta = \frac{4R_x}{\alpha+1}\sqrt{\frac{\alpha n}{\kappa_x^2(1-2/\kappa_y^2)}}$, noticing that $\alpha^2 - 1 = \frac{(\kappa_y-2/\kappa_y)\kappa_x}{2n} \le \frac{\kappa_x\kappa_y}{2n}$ and $\frac{\alpha(\alpha-1)^3}{(\alpha+1)^4}$ is increasing for $\alpha > 1$, we have

$$\frac{q^2\beta^2\xi^2}{9(\alpha+1)\mu_x} = \frac{2\alpha(\alpha-1)^3}{9(\alpha+1)^5(\alpha-1)}\mu_x R_x^2 \ge \frac{4\sqrt{2}\left(\sqrt{2}-1\right)^7}{9}\frac{n\mu_x R_x^2}{\kappa_x\kappa_y}.$$

When $\beta = \frac{\sqrt{2\alpha}R_y}{\alpha-1}$, noticing that $\frac{\mu_x\mu_y}{\xi^2} = \frac{4}{\alpha^2-1}$ and $\frac{\alpha(\alpha-1)}{(\alpha+1)^2}$ is increasing for $\alpha > 1$, we have

$$\frac{q^2\beta^2\xi^2}{9(\alpha+1)\mu_x} = \frac{\alpha(\alpha-1)}{18(\alpha+1)^2}\mu_y R_y^2 \ge \frac{\sqrt{2}\left(\sqrt{2}-1\right)^3}{18}\mu_y R_y^2.$$

Thus, $\varepsilon \le \frac{1}{800} \min\left\{\frac{n\mu_x R_x^2}{\kappa_x\kappa_y}, \mu_y R_y^2\right\}$ is a sufficient condition for $M \ge 1$. Similarly, we can obtain

$$\frac{\beta^2\xi^2}{9(\alpha+1)\mu_x} \ge \frac{1}{25} \min\left\{\frac{n\mu_x R_x^2}{\kappa_x\kappa_y}, \mu_y R_y^2\right\}. \tag{38}$$

On the other hand, since $\frac{\alpha}{\alpha-1} \le \frac{\sqrt{2}}{\sqrt{2}-1}$, we have

$$\frac{\beta^2\xi^2}{9(\alpha+1)\mu_x} \le \min\left\{\frac{\alpha\mu_x R_x^2}{9(\alpha+1)}, \frac{2\alpha(\alpha-1)\mu_x R_x^2}{9(\alpha+1)^3(\alpha-1)}, \frac{\alpha\mu_y R_y^2}{18(\alpha-1)}\right\} \le \frac{2}{9}\min\{\mu_x R_x^2, \mu_y R_y^2\}. \tag{39}$$

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 $\lim_{\beta\to+\infty} h(\beta) = 0$. With $q = \frac{\alpha-1}{\alpha+1}$,there holds $h(\sqrt{2}) \le -\frac{1}{\log q} - \frac{\alpha}{2} \le 0$, which implies $\frac{\alpha}{2} \ge -\frac{1}{\log q} \ge \frac{\alpha}{2} + h(\sqrt{2})$. Then by (39) we have

$$m = \left\lfloor \frac{\alpha}{4}\log\left(\frac{2\min\left\{\mu_x R_x^2, \mu_y R_y^2\right\}}{9\varepsilon}\right)\right\rfloor + 1 \ge \left\lfloor -\frac{\log\left(\frac{\beta^2\xi^2}{9(\alpha+1)\mu_x\varepsilon}\right)}{2\log q}\right\rfloor + 1 > M.$$

Thus, we have verified $1 \le M < m$. Moreover, $-\frac{1}{\log q} \ge \frac{\alpha}{2} + h(\sqrt{2})$ implies

$$-\frac{1}{\log(q)} \ge \frac{1}{2}\sqrt{\frac{(\kappa_y - 2/\kappa_y)\,\kappa_x}{2n} + 1} + h\left(\sqrt{2}\right) \ge \frac{\sqrt{2}}{4}\left(\sqrt{\frac{\kappa_x\kappa_y}{4n}} + 1\right) + h\left(\sqrt{2}\right),$$

where the last inequality is due to $\kappa_y - 2/\kappa_y \ge \kappa_y/2$ and $\sqrt{2(a+b)} \ge \sqrt{a} + \sqrt{b}$ for $a, b > 0$.

By Lemma 17, for $M \ge 1$ and $N = \frac{(M+1)n}{4(1+c_0)}$, we have that $\min_{t\le N}\mathbb{E}\phi_{\text{SCSC}}(\mathbf{x}_t) - \min_{t\le N}\mathbb{E}\psi_{\text{SCSC}}(\mathbf{y}_t) \ge \varepsilon$ holds. Therefore, in order to find $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathcal{X}\times\mathcal{Y}$ such that $\mathbb{E}\phi_{\text{SCSC}}(\hat{\mathbf{x}}) - \mathbb{E}\psi_{\text{SCSC}}(\hat{\mathbf{y}}) \ge \varepsilon$, $\mathcal{A}$ needs at least N PIFO queries, where

$$\begin{aligned}
N = \frac{(M+1)n}{4(1+c_0)} &\ge \frac{n}{4(1+c_0)}\left(-\frac{1}{\log(q)}\right)\log\left(\frac{\beta^2\xi^2}{9(\alpha+1)\mu_x\varepsilon}\right) \\
&\ge \frac{n}{4(1+c_0)}\left(\sqrt{\frac{\kappa_x\kappa_y}{32n}} + \frac{\sqrt{2}}{4} + h\left(\sqrt{2}\right)\right)\log\left(\frac{\min\left\{n\mu_x R_x^2/(\kappa_x\kappa_y), \mu_y R_y^2\right\}}{25\varepsilon}\right) \\
&= \Omega\left(\left(n + \sqrt{n\kappa_x\kappa_y}\right)\log\left(\frac{1}{\varepsilon}\right)\right),
\end{aligned}$$

where the second inequality is by (39). This completes the proof. ■

Proof [Proof of Theorem 28] Let $\alpha = \sqrt{\frac{2(\kappa_x-1)}{n} + 1}$. Consider the functions $\{f_{\text{SC},i}\}_{i=1}^n$ and f_{SC} defined in Definition 69 with μ and R replaced by μ_x and R_x. We construct $\{G_{\text{SCSC},i}\}_{i=1}^n, G_{\text{SCSC}} : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ as follows

$$G_{\text{SCSC},i}(\mathbf{x},\mathbf{y}) = f_{\text{SC},i}(\mathbf{x}) - \frac{\mu_y}{2}\|\mathbf{y}\|_2^2,$$

$$G_{\mathrm{SCSC}}(\mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} G_{\mathrm{SCSC},i}(\mathbf{x}, \mathbf{y}) = f_{\mathrm{SC}}(\mathbf{x}) - \frac{\mu_y}{2} \|\mathbf{y}\|_2^2 .$$

By Proposition 59 and Lemma 40, we can check that each component function $G_{\mathrm{SCSC},i}$ is L-smooth and (μ_x, μ_y)-convex-concave. Then G_{SCSC} is (μ_x, μ_y)-convex-concave. Moreover, we have

$$\max_{\mathbf{y} \in \mathcal{Y}} G_{\mathrm{SCSC}}(\mathbf{x}, \mathbf{y}) = f_{\mathrm{SC}}(\mathbf{x}) \quad \text{and} \quad \min_{\mathbf{x} \in \mathcal{X}} G_{\mathrm{SCSC}}(\mathbf{x}, \mathbf{y}) = \min_{\mathbf{x} \in \mathcal{X}} f_{\mathrm{SC}}(\mathbf{y}) - \frac{\mu_y}{2} \|\mathbf{y}\|_2^2 .$$

It follows that for any $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathcal{X} \times \mathcal{Y}$, we have

$$\max_{\mathbf{y} \in \mathcal{Y}} G_{\mathrm{SCSC}}(\hat{\mathbf{x}}, \mathbf{y}) - \min_{\mathbf{x} \in \mathcal{X}} G_{\mathrm{SCSC}}(\mathbf{x}, \hat{\mathbf{y}}) \geq f_{\mathrm{SC}}(\hat{\mathbf{x}}) - \min_{\mathbf{x} \in \mathcal{X}} f_{\mathrm{SC}}(\mathbf{x}).$$

Note that $\kappa_x \geq \sqrt{2n+2} = \Omega(\sqrt{n})$. By Theorem 74, for $\tilde{L} = \sqrt{\frac{n(L^2 - \mu_x^2)}{2} - \mu_x^2}$,

$$\varepsilon \leq \frac{\mu_x R_x^2}{18} \left(\frac{\alpha - 1}{\alpha + 1} \right)^2 \text{ and } m = \left\lfloor \frac{1}{4} \left(\sqrt{\frac{2(\tilde{L}/\mu_x - 1)}{n} + 1} \right) \log \left(\frac{\mu_x R_x^2}{9\varepsilon} \right) \right\rfloor + 1,$$

in order to find $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathcal{X} \times \mathcal{Y}$ such that $\mathbb{E} \left(\max_{\mathbf{y} \in \mathcal{Y}} G_{\mathrm{SCSC}}(\hat{\mathbf{x}}, \mathbf{y}) - \min_{\mathbf{x} \in \mathcal{X}} G_{\mathrm{SCSC}}(\mathbf{x}, \hat{\mathbf{y}}) \right) < \varepsilon$, PIFO algorithm $\mathcal{A}$ needs at least $N = \Omega \left(\left(n + n^{3/4} \sqrt{\kappa_x} \right) \log \left(\frac{1}{\varepsilon} \right) \right)$ queries.

Moreover, $\kappa_x \geq n/2 + 1$ implies $\alpha \geq \sqrt{2}$. Then we have $\left(\frac{\alpha - 1}{\alpha + 1} \right)^2 \geq \left(\frac{\sqrt{2} - 1}{\sqrt{2} + 1} \right)^2 \geq \frac{1}{40}$. This completes the proof. ■

Proof [Proof of Lemma 29] Consider the functions $\{H_{\mathrm{SCSC},i} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}\}_{i=1}^n$ where

$$H_{\mathrm{SCSC},i}(x, y) = \begin{cases} \frac{L}{2} \left(x^2 - y^2 \right) - nLR_x x, & \text{for } i = 1, \\ \frac{L}{2} \left(x^2 - y^2 \right), & \text{otherwise,} \end{cases}$$

and $H_{\mathrm{SCSC}}(x, y) = \frac{1}{n} \sum_{i=1}^n H_{\mathrm{SCSC},i}(x, y) = \frac{L}{2}(x^2 - y^2) - LR_x x$. It is easy to check that $\{H_{\mathrm{SCSC},i}\}_{i=1}^n$ L-average smooth and (μ_x, μ_y)-convex-concave for any $0 \leq \mu_x, \mu_y \leq L$. Moreover, we have

$$\max_{|y| \leq R_y} H_{\mathrm{SCSC}}(x, y) = \frac{L}{2} x^2 - LR_x x \quad \text{and} \quad \min_{|x| \leq R_x} H_{\mathrm{SCSC}}(x, y) = -\frac{LR_x^2}{2} - \frac{L}{2} y^2.$$

Note that for $i \geq 2$, it holds that

$$\nabla_x H_{\mathrm{SCSC},i}(x, y) = Lx \quad \text{and} \quad \mathrm{prox}_{H_{\mathrm{SCSC},i}}^{\gamma}(x, y) = \left(\frac{x}{L\gamma + 1}, \frac{y}{L\gamma + 1} \right).$$

This implies $x_t = x_0 = 0$ will hold till the PIFO algorithm $\mathcal{A}$ draws $H_{\mathrm{SCSC},1}$. Denote $T = \min\{t : i_t = 1\}$. Then, the random variable T follows a geometric distribution with success probability p_1, and satisfies $\mathbb{P}\left[T \geq n/2\right] = (1 - p_1)^{\lfloor (n-1)/2 \rfloor} \geq (1 - 1/n)^{(n-1)/2} \geq 1/2$, where the last inequality is according to that $h(\beta) = (\frac{\beta}{\beta+1})^{\beta/2}$ is a decreasing function and $\lim_{\beta\to\infty} h(\beta) = 1/\sqrt{e} \geq 1/2$. Consequently, for $N = n/2$ and $t < N$, we know that

$$\begin{aligned}
&\mathbb{E}\left(\max_{|y|\leq R_y} H_{\text{SCSC}}(x_t, y) - \min_{|x|\leq R_x} H_{\text{SCSC}}(x, y_t)\right) \\
\geq\, &\mathbb{E}\left(\max_{|y|\leq R_y} H_{\text{SCSC}}(x_t, y) - \min_{|x|\leq R_x} H_{\text{SCSC}}(x, y_t)\middle| t < T\right)\mathbb{P}\left[T > t\right] \\
=\, &\mathbb{E}\left(\max_{|y|\leq R_y} H_{\text{SCSC}}(0, y) - \min_{|x|\leq R_x} H_{\text{SCSC}}(0, y_t)\middle| t < T\right)\mathbb{P}\left[T > t\right] \\
\geq\, &\frac{LR_x^2}{2}\mathbb{P}\left[T \geq N\right] \geq LR_x^2/4 \geq \varepsilon.
\end{aligned}$$

Thus, to find $(\hat{x}, \hat{y}) \in \mathcal{X} \times \mathcal{Y}$ such that $\mathbb{E}\max_{|y|\leq R_y} H_{\text{SCSC}}(\hat{x}, y) - \mathbb{E}\min_{|x|\leq R_x} H_{\text{SCSC}}(x, \hat{y}) < \varepsilon$, PIFO algorithm $\mathcal{A}$ needs at least $N = \Omega(n)$ queries. ■

D.2 Proofs for the Convex-Strongly-Concave Case

With f_{CSC} and $\{f_{\text{CSC},i}\}_{i=1}^n$ defined in Definition 30, we have the following proposition.

Proposition 48 *For any $n \geq 2$, $m \geq 2$, $f_{\text{CSC},i}$ and f_{CSC} in Definition 30 satisfy:*

1. *$\{f_{\text{CSC},i}\}_{i=1}^n$ is L-smooth and each $f_{\text{CSC},i}$ is $(0, \mu_y)$-convex-concave. Thus, f_{CSC} is $(0, \mu_y)$-convex-concave.*
2. *For $1 \leq k \leq m-1$, we have*

$$\min_{\mathbf{x}\in\mathcal{X}\cap\mathcal{F}_k} \phi_{\text{CSC}}(\mathbf{x}) - \max_{\mathbf{y}\in\mathcal{Y}\cap\mathcal{F}_k} \psi_{\text{CSC}}(\mathbf{y}) \geq -\frac{k\mu_y\beta^2}{2} + \frac{R_x\beta}{2}\sqrt{\frac{L^2 - 2\mu_y^2}{2n(k+1)}},$$

where $\beta = \min\left\{\frac{R_x\sqrt{(L^2/\mu_y^2 - 2)/(2n)}}{2(m+1)^{3/2}}, \frac{R_y}{\sqrt{m}}\right\}$.

Proof

1. Just recall Proposition 14 and Lemma 40.
2. It it easy to check $f_{\text{CSC}}(\mathbf{x}, \mathbf{y}) = \xi\left\langle \mathbf{y}, \widetilde{\mathbf{B}}(m,1)\,\mathbf{x}\right\rangle - \frac{\mu_y}{2}\|\mathbf{y}\|_2^2 - \beta\xi\langle \mathbf{e}_1, \mathbf{x}\rangle$, where $\xi = \lambda/\beta^2 = \frac{1}{2}\sqrt{\frac{L^2 - 2\mu_y^2}{2n}}$. Define $\tilde{\phi}_{\text{CSC}}(\mathbf{x}) = \max_{\mathbf{y}\in\mathbb{R}^m} f_{\text{CSC}}(\mathbf{x}, \mathbf{y})$. We first show that

$$\min_{\mathbf{x}\in\mathcal{X}\cap\mathcal{F}_k} \tilde{\phi}_{\text{CSC}}(\mathbf{x}) - \max_{\mathbf{y}\in\mathcal{Y}\cap\mathcal{F}_k} \psi_{\text{CSC}}(\mathbf{y}) \geq -\frac{k\mu_y\beta^2}{2} + \frac{R_x\xi\beta}{\sqrt{k+1}}.$$

On one hand, we have

$$\begin{aligned}
\tilde{\phi}_{\text{CSC}}(\mathbf{x}) &= \max_{\mathbf{y}\in\mathbb{R}^m}\left(\xi\left\langle \mathbf{y}, \widetilde{\mathbf{B}}(m,1)\,\mathbf{x}\right\rangle - \frac{\mu_y}{2}\|\mathbf{y}\|_2^2 - \beta\xi\langle \mathbf{e}_1, \mathbf{x}\rangle\right) \\
&= \max_{\mathbf{y}\in\mathbb{R}^m}\left(-\frac{\mu_y}{2}\left\|\mathbf{y} - \frac{\xi}{\mu_y}\widetilde{\mathbf{B}}(m,1)\mathbf{x}\right\|_2^2 + \frac{\xi^2}{2\mu_y}\left\|\widetilde{\mathbf{B}}(m,1)\mathbf{x}\right\|_2^2 - \beta\xi\langle \mathbf{e}_1, \mathbf{x}\rangle\right) \quad (40) \\
&= \frac{\xi^2}{2\mu_y}\left\|\widetilde{\mathbf{B}}(m,1)\mathbf{x}\right\|_2^2 - \beta\xi\langle \mathbf{e}_1, \mathbf{x}\rangle.
\end{aligned}$$

For $\mathbf{x} \in \mathcal{F}_k$, let $\tilde{\mathbf{x}}$ be the first k coordinates of $\mathbf{x}$. We can rewrite $\tilde{\phi}_{\mathrm{CSC}}(\mathbf{x})$ as $\tilde{\phi}_k(\tilde{\mathbf{x}}) \triangleq \tilde{\phi}_{\mathrm{CSC}}(\mathbf{x}) = \frac{\xi^2}{2\mu_y}\left\|\widetilde{\mathbf{B}}(k,1)\tilde{\mathbf{x}}\right\|_2^2 - \beta\xi\left\langle \hat{\mathbf{e}}_1, \tilde{\mathbf{x}}\right\rangle$, where $\hat{\mathbf{e}}_1$ is the first k coordinates of $\mathbf{e}_1$. Letting $\nabla\tilde{\phi}_k(\tilde{\mathbf{x}}) = \mathbf{0}_k$, we get $\widetilde{\mathbf{B}}(k,1)^\top\widetilde{\mathbf{B}}(k,1)\tilde{\mathbf{x}} = \frac{\beta\mu_y}{\xi}\hat{\mathbf{e}}_1$. The solution is $\tilde{\mathbf{x}}^* = \frac{\beta\mu_y}{\xi}(k, k-1, \ldots, 1)^\top$. Noting that $\|\tilde{\mathbf{x}}^*\|_2^2 = \frac{\beta^2\mu_y^2}{\xi^2}\frac{k(k+1)(2k+1)}{6} \le \frac{8n\beta^2}{L^2/\mu_y^2-2}(m+1)^3 \le R_x^2$, we obtain $\min_{\mathbf{x}\in\mathcal{X}\cap\mathcal{F}_k}\tilde{\phi}_{\mathrm{CSC}}(\mathbf{x}) = -\frac{k\mu_y\beta^2}{2}$.

On the other hand,

$$\begin{aligned}\min_{\mathbf{x}\in\mathcal{X}}\left\langle \mathbf{x}, \widetilde{\mathbf{B}}(m,1)^\top\mathbf{y} - \beta\mathbf{e}_1\right\rangle &\ge \min_{\|\mathbf{x}\|_2\le R_x} -\|\mathbf{x}\|_2\left\|\widetilde{\mathbf{B}}(m,1)^\top\mathbf{y} - \beta\mathbf{e}_1\right\|_2 \\ &\ge -R_x\left\|\widetilde{\mathbf{B}}(m,1)^\top\mathbf{y} - \beta\mathbf{e}_1\right\|_2,\end{aligned} \tag{41}$$

where the equality will hold when either $\mathbf{x} = -\frac{R_x}{\widetilde{\mathbf{B}}(m,1)^\top\mathbf{y}-\beta\mathbf{e}_1}\left(\widetilde{\mathbf{B}}(m,1)^\top\mathbf{y} - \beta\mathbf{e}_1\right)$ or $\widetilde{\mathbf{B}}(m,1)^\top\mathbf{y} - \beta\mathbf{e}_1 = \mathbf{0}$. It follows that

$$\begin{aligned}\psi_{\mathrm{CSC}}(\mathbf{y}) &= \min_{\mathbf{x}\in\mathcal{X}}\left(\xi\left\langle \mathbf{x}, \widetilde{\mathbf{B}}\left(m,1\right)^\top\mathbf{y} - \beta\mathbf{e}_1\right\rangle - \frac{\mu_y}{2}\|\mathbf{y}\|_2^2\right) \\ &= -R_x\xi\left\|\widetilde{\mathbf{B}}\left(m,1\right)^\top\mathbf{y} - \beta\mathbf{e}_1\right\|_2 - \frac{\mu_y}{2}\|\mathbf{y}\|_2^2.\end{aligned} \tag{42}$$

We can upper bound $\max_{\mathbf{y}\in\mathcal{Y}\cap\mathcal{F}_k}\psi_{\mathrm{CSC}}(\mathbf{y})$ as

$$\begin{aligned}\max_{\mathbf{y}\in\mathcal{Y}\cap\mathcal{F}_k}\psi_{\mathrm{CSC}}(\mathbf{y}) &= \max_{\mathbf{y}\in\mathcal{Y}\cap\mathcal{F}_k}\left(-R_x\xi\left\|\widetilde{\mathbf{B}}(m,1)^\top\mathbf{y} - \beta\mathbf{e}_1\right\|_2 - \frac{\mu_y}{2}\|\mathbf{y}\|_2^2\right) \\ &\le \max_{\mathbf{y}\in\mathcal{F}_k}\left(-R_x\xi\left\|\widetilde{\mathbf{B}}(m,1)^\top\mathbf{y} - \beta\mathbf{e}_1\right\|_2\right) \\ &= -R_x\xi\sqrt{J_{k,\beta}(y_1,y_2,\ldots,y_k)} \le -\frac{R_x\xi\beta}{\sqrt{k+1}},\end{aligned}$$

where $J_{k,\beta}$ is defined in (22) and the last inequality follows from Lemma 46.

It remains to prove $\min_{\mathbf{x}\in\mathcal{X}\cap\mathcal{F}_k}\phi_{\mathrm{CSC}}(\mathbf{x}) = \min_{\mathbf{x}\in\mathcal{X}\cap\mathcal{F}_k}\tilde{\phi}_{\mathrm{CSC}}(\mathbf{x})$. Recall the expression (40). It suffices to show that $\|\hat{\mathbf{y}}\|_2 \le R_y$ where $\hat{\mathbf{y}} = \frac{\xi}{\mu_y}\widetilde{\mathbf{B}}(m,1)\begin{bmatrix}\tilde{\mathbf{x}}^*\\ \mathbf{0}_{m-k}\end{bmatrix}$. Since $\beta \le \frac{R_y}{\sqrt{m}}$, one can check $\|\hat{\mathbf{y}}\|_2 \le R_y$ does hold.

This completes the proof. ■

Proof [Proof of Theorem 31] Since $L/\mu_y \ge 2$, we have $L^2 - 2\mu_y^2 \ge L^2/2$. Then $\varepsilon \le \frac{L^2R_x^2}{5184\,n\mu_y} \le \frac{(L^2-2\mu_y^2)R_x^2}{2592n\mu_y}$, which implies that $m \ge 4$ and $\frac{R_x}{6}\sqrt{\frac{L^2-2\mu_y^2}{2n\mu_y\varepsilon}} - 2 \ge \frac{R_x}{12}\sqrt{\frac{L^2-2\mu_y^2}{2n\mu_y\varepsilon}} + 1$. It follows that $m \ge \frac{R_x}{12}\sqrt{\frac{L^2-2\mu_y^2}{2n\mu_y\varepsilon}}$. Then with $\varepsilon \le \frac{\mu_yR_y^2}{36}$, we have

$$\frac{R_x\sqrt{\frac{L^2/\mu_y^2-2}{2n}}}{2(m+1)^{3/2}} < \frac{R_x\sqrt{\frac{L^2/\mu_y^2-2}{2n}}}{2m^{3/2}} \le 6\sqrt{\frac{\varepsilon}{\mu_ym}} \le \frac{R_y}{\sqrt{m}},$$

which implies that $\beta = \min\left\{\frac{R_x\sqrt{(L^2/\mu_y^2-2)/(2n)}}{2(m+1)^{3/2}}, \frac{R_y}{\sqrt{m}}\right\} = \frac{R_x\sqrt{(L^2/\mu_y^2-2)/(2n)}}{2(m+1)^{3/2}}$. Following Proposition 48, for $1 \le k \le m-1$, we have

$$\begin{aligned}\min_{\mathbf{x}\in\mathcal{X}\cap\mathcal{F}_k} \phi_{\text{CSC}}(\mathbf{x}) - \max_{\mathbf{y}\in\mathcal{Y}\cap\mathcal{F}_k} \psi_{\text{CSC}}(\mathbf{y}) &\ge -\frac{k\mu_y\beta^2}{2} + \frac{R_x\beta}{2}\sqrt{\frac{L^2-2\mu_y^2}{2n(k+1)}} \\ &= \frac{(L^2-2\mu_y^2)R_x^2}{16n\mu_y}\frac{2(m+1)^{3/2}-k\sqrt{k+1}}{(m+1)^3\sqrt{k+1}}.\end{aligned}$$

Define $M \triangleq \left\lfloor \frac{m}{2} \right\rfloor$. Then we have $M = \left\lfloor \frac{R_x}{12}\sqrt{\frac{L^2-2\mu^2}{2n\mu\varepsilon}} \right\rfloor - 1 \ge 2$ and $M < m$.

Since $2(M+1) = 2\left\lfloor \frac{m}{2} \right\rfloor + 2 \ge m+1$ and $h(\beta) = \frac{2\beta^{3/2}-\beta_0^{3/2}}{\beta^3}$ is a decreasing function when $\beta > \beta_0$, for $k = M$ we have

$$\min_{\mathbf{x}\in\mathcal{X}\cap\mathcal{F}_M} \phi_{\text{CSC}}(\mathbf{x}) - \max_{\mathbf{y}\in\mathcal{Y}\cap\mathcal{F}_M} \psi_{\text{CSC}}(\mathbf{y}) \ge \frac{(L^2-2\mu_y^2)R_x^2}{16n\mu_y}\frac{4\sqrt{2}-1}{8(M+1)^2} > \frac{(L^2-2\mu_y^2)R_x^2}{32n\mu_y(M+1)^2} \ge 9\varepsilon,$$

where the last inequality is due to $M+1 \le \frac{R_x}{12}\sqrt{\frac{L^2-2\mu_y^2}{2n\mu_y\varepsilon}}$.

By Lemma 17, for $N = \frac{n(M+1)}{4(1+c_0)}$, we have $\min_{t\le N}\mathbb{E}\left(\phi_{\text{CSC}}(\mathbf{x}_t) - \psi_{\text{CSC}}(\mathbf{y}_t)\right) \ge \varepsilon$. Therefore, in order to find suboptimal solution $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathcal{X}\times\mathcal{Y}$ such that $\mathbb{E}\left(\phi_{\text{CSC}}(\hat{\mathbf{x}}) - \psi_{\text{CSC}}(\hat{\mathbf{y}})\right) < \varepsilon$, algorithm $\mathcal{A}$ needs at least N PIFO queries, where

$$N = \frac{n}{4(1+c_0)}\left(\left\lfloor \frac{R_x}{12}\sqrt{\frac{L^2-2\mu_y^2}{2n\mu_y\varepsilon}} \right\rfloor\right) = \Omega\left(n + R_xL\sqrt{\frac{n}{\mu_y\varepsilon}}\right).$$

This completes the proof. ■

Proof [Proof of Theorem 32] Consider the functions $\{f_{\text{C},i}\}_{i=1}^n$ and f_{C} defined in Definition 79 with R replaced by R_x. We construct $\{G_{\text{CSC},i}\}_{i=1}^n, G_{\text{CSC}} : \mathbb{R}^m\times\mathbb{R}^m \to \mathbb{R}$ as follows

$$\begin{aligned}G_{\text{CSC},i}(\mathbf{x},\mathbf{y}) &= f_{\text{C},i}(\mathbf{x}) - \frac{\mu_y}{2}\|\mathbf{y}\|_2^2, \\ G_{\text{CSC}}(\mathbf{x},\mathbf{y}) &= \frac{1}{n}\sum_{i=1}^n G_{\text{CSC},i}(\mathbf{x},\mathbf{y}) = f_{\text{C}}(\mathbf{x}) - \frac{\mu_y}{2}\|\mathbf{y}\|_2^2.\end{aligned}$$

By Proposition 59 and Lemma 40, we can check that each component function $G_{\text{CSC},i}$ is L-smooth and $(0,\mu_y)$-convex-concave. Then G_{CSC} is $(0,\mu_y)$-convex-concave. Moreover, we have

$$\max_{\mathbf{y}\in\mathcal{Y}} G_{\text{CSC}}(\mathbf{x},\mathbf{y}) = f_{\text{C}}(\mathbf{x}) \quad \text{and} \quad \min_{\mathbf{x}\in\mathcal{X}} G_{\text{CSC}}(\mathbf{x},\mathbf{y}) = \min_{\mathbf{x}\in\mathcal{X}} f_{\text{C}}(\mathbf{x}) - \frac{\mu_y}{2}\|\mathbf{y}\|_2^2.$$

It follows that for any $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathcal{X}\times\mathcal{Y}$, we have

$$\max_{\mathbf{y}\in\mathcal{Y}} G_{\text{CSC}}(\hat{\mathbf{x}},\mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} G_{\text{CSC}}(\mathbf{x},\hat{\mathbf{y}}) \ge f_{\text{C}}(\hat{\mathbf{x}}) - \min_{\mathbf{x}\in\mathcal{X}} f_{\text{C}}(\mathbf{x}).$$

By Theorem 80, for

$$\varepsilon \leq \frac{\sqrt{2}R_x^2 L}{768\sqrt{n}} \text{ and } m = \left\lfloor \frac{\sqrt[4]{18}}{12} R_x n^{-1/4} \sqrt{\frac{L}{\varepsilon}} \right\rfloor - 1,$$

in order to find $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathcal{X} \times \mathcal{Y}$ such that $\mathbb{E}\left(\max_{\mathbf{y}\in\mathcal{Y}} G_{\mathrm{CSC}}(\hat{\mathbf{x}}, \mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} G_{\mathrm{CSC}}(\mathbf{x}, \hat{\mathbf{y}})\right) < \varepsilon$, PIFO algorithm $\mathcal{A}$ needs at least $N = \Omega\left(n + R_x n^{3/4}\sqrt{\frac{L}{\varepsilon}}\right)$ queries. ■

D.3 Proofs for the Convex-Concave Case

With f_{CC} and $\{f_{\mathrm{CC},i}\}_{i=1}^n$ defined in Definition 33, we have the following proposition.

Proposition 49 *For any $n \geq 2$, $m \geq 3$, $f_{\mathrm{CC},i}$ and f_{CC} in Definition 33 satisfy:*

1. *$\{f_{\mathrm{CC},i}\}_{i=1}^n$ is L-average smooth and each $f_{\mathrm{CC},i}$ convex-concave. Thus, f_{CC} is convex-concave.*
2. *For $1 \leq k \leq m-1$, we have*

$$\min_{\mathbf{x}\in\mathcal{X}\cap\mathcal{F}_k} \phi_{\mathrm{CC}}(\mathbf{x}) - \max_{\mathbf{y}\in\mathcal{Y}\cap\mathcal{F}_k} \psi_{\mathrm{CC}}(\mathbf{y}) \geq \frac{LR_xR_y}{\sqrt{8nm(k+1)}}.$$

Proof

1. Just recall Proposition 14 and Lemma 40.
2. It is easy to check $f_{\mathrm{CC}}(\mathbf{x}, \mathbf{y}) = \frac{L}{\sqrt{8n}}\left\langle \mathbf{y}, \widetilde{\mathbf{B}}(m,1)\mathbf{x}\right\rangle - \frac{LR_y}{\sqrt{8nm}}\langle \mathbf{e}_1, \mathbf{x}\rangle$. By similar analysis from Equation (41) to Equation (42) of the proof of Proposition 48, we can conclude that

$$\phi_{\mathrm{CC}}(\mathbf{x}) = \frac{LR_y}{\sqrt{8n}}\left\|\widetilde{\mathbf{B}}(m,1)\mathbf{x}\right\|_2 - \frac{LR_y}{\sqrt{8nm}}\langle \mathbf{e}_1, \mathbf{x}\rangle,$$
$$\psi_{\mathrm{CC}}(\mathbf{y}) = -\frac{LR_x}{\sqrt{8n}}\left\|\widetilde{\mathbf{B}}(m,1)^\top \mathbf{y} - \frac{R_y}{\sqrt{m}}\mathbf{e}_1\right\|_2.$$

Note that $\phi_{\mathrm{CC}}(\mathbf{x}) = \max_{\mathbf{y}\in\mathcal{Y}} f_{\mathrm{CC}}(\mathbf{x}, \mathbf{y}) \geq \max_{\mathbf{y}\in\mathcal{Y}} \min_{\mathbf{x}\in\mathcal{X}} f_{\mathrm{CC}}(\mathbf{x}, \mathbf{y}) = \max_{\mathbf{y}\in\mathcal{Y}} \psi(\mathbf{y}) \geq \psi(\mathbf{y}^*) = 0$, where $\mathbf{y}^* = \frac{R_y}{\sqrt{m}}\mathbf{1}_m \in \mathcal{Y}$. Therefore, we have $\min_{\mathbf{x}\in\mathcal{X}\cap\mathcal{F}_k} \phi_{\mathrm{CC}}(\mathbf{x}) = \phi_{\mathrm{CC}}(\mathbf{0}) = 0$. On the other hand, following Lemma 46, we can obtain

$$\max_{\mathbf{y}\in\mathcal{Y}\cap\mathcal{F}_k} \psi_{\mathrm{CC}}(\mathbf{y}) = \max_{\mathbf{y}\in\mathcal{Y}\cap\mathcal{F}_k} -\frac{LR_x}{\sqrt{8n}}\left\|\widetilde{\mathbf{B}}(m,1)^\top \mathbf{y} - \frac{R_y}{\sqrt{m}}\mathbf{e}_1\right\|_2 = -\frac{LR_x}{\sqrt{8n}}\frac{R_y}{\sqrt{m(k+1)}},$$

where the optimal point is $\tilde{\mathbf{y}}^* = \frac{R_y}{(k+1)\sqrt{m}}(k, k-1, \ldots, 1, 0, \ldots, 0)^\top$, which satisfies $\|\tilde{\mathbf{y}}^*\|_2 = \frac{R_y}{(k+1)\sqrt{m}}\sqrt{\frac{k(k+1)(2k+1)}{6}} \leq R_y$. Finally, note that $k+1 \geq m/2$. Thus we obtain

$$\min_{\mathbf{x}\in\mathcal{X}\cap\mathcal{F}_k} \phi_{\mathrm{CC}}(\mathbf{x}) - \max_{\mathbf{y}\in\mathcal{Y}\cap\mathcal{F}_k} \psi_{\mathrm{CC}}(\mathbf{y}) \geq \frac{LR_xR_y}{\sqrt{8nm(k+1)}}.$$

This completes the proof. ∎

Proof [Proof of Theorem 34] The assumption on ε implies $m \geq 3$. Let $M \triangleq \lfloor (m-1)/2 \rfloor = \left\lfloor \frac{LR_xR_y}{36\varepsilon\sqrt{n}} \right\rfloor - 1$. Then we have $M \geq 1$ and $m/2 \leq M+1 \leq (m+1)/2$. By Proposition 49, we have

$$\min_{\mathbf{x}\in\mathcal{X}\cap\mathcal{F}_M} \phi_{\mathrm{CC}}(\mathbf{x}) - \max_{\mathbf{y}\in\mathcal{Y}\cap\mathcal{F}_M} \psi_{\mathrm{CC}}(\mathbf{y}) \geq \frac{LR_xR_y}{\sqrt{8nm(M+1)}} \geq \frac{LR_xR_y}{4(M+1)\sqrt{n}} \geq \frac{LR_xR_y}{2(m+1)\sqrt{n}} \geq 9\varepsilon.$$

Hence, by Lemma 17, for $N = \frac{n(M+1)}{4(1+c_0)}$, we know that $\min_{t\leq N} \mathbb{E}\left(\phi_{\mathrm{CC}}(\mathbf{x}_t) - \psi_{\mathrm{CC}}(\mathbf{y}_t)\right) \geq \varepsilon$. Thus, to find an approximate solution $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \mathcal{X} \times \mathcal{Y}$ such that $\mathbb{E}\left(\phi_{\mathrm{CC}}(\hat{\mathbf{x}}) - \psi_{\mathrm{CC}}(\hat{\mathbf{y}})\right) < \varepsilon$, the PIFO algorithm $\mathcal{A}$ needs at least N queries, where $N = \frac{n}{4(1+c_0)}\left(\left\lfloor \frac{LR_xR_y}{36\varepsilon\sqrt{n}} \right\rfloor\right) = \Omega\left(n + \frac{\sqrt{n}LR_xR_y}{\varepsilon}\right)$. ∎

Proof [Proof of Lemma 35] Consider the functions $\{H_{\mathrm{CC},i} : \mathbb{R}\times\mathbb{R} \to \mathbb{R}\}_{i=1}^n$ where

$$H_{\mathrm{CC},i}(x,y) = \begin{cases} Lxy - nLR_xy, & \text{for } i = 1, \\ Lxy, & \text{otherwise,} \end{cases}$$

and $H_{\mathrm{CC}}(x,y) = \frac{1}{n}\sum_{i=1}^n H_{\mathrm{CC},i}(x,y) = Lxy - LR_xy$. Consider the minimax problem $\min_{|x|\leq R_x} \max_{|y|\leq R_y} H_{\mathrm{CC}}(x,y)$. It is easy to check that $\{H_{\mathrm{CC},i}\}_{i=1}^n$ is L-smooth and each $H_{\mathrm{CC},i}$ is convex-concave. Moreover, we have

$$\max_{|y|\leq R_y} H_{\mathrm{CC}}(x,y) = LR_y|x - R_x|, \quad \text{and} \quad \min_{|x|\leq R_x} H_{\mathrm{CC}}(x,y) = -LR_x(|y| + y) \leq 0,$$

and it holds that $\min_{|x|\leq R_x} \max_{|y|\leq R_y} H_{\mathrm{CC}}(x,y) = \max_{|y|\leq R_y} \min_{|x|\leq R_x} H_{\mathrm{CC}}(x,y) = 0$.

Note that for $i \geq 2$, we have

$$\nabla_x H_{\mathrm{CC},i}(x,y) = Ly,\ \nabla_y H_{\mathrm{CC},i}(x,y) = Lx, \text{ and } \operatorname{prox}^{\gamma}_{H_{\mathrm{CC},i}}(x,y) = \left(\frac{L\gamma x + y}{L^2\gamma^2 + 1}, \frac{x - L\gamma y}{L^2\gamma^2+1}\right),$$

which implies $x_t = y_t = x_0 = y_0 = 0$ will hold till the PIFO algorithm $\mathcal{A}$ draws $H_{\mathrm{CC},1}$.

Let $T = \min\{t : i_t = 1\}$. Then, the random variable T follows a geometric distribution with success probability p_1, and satisfies $\mathbb{P}\left[T \geq n/2\right] = (1-p_1)^{\lfloor (n-1)/2 \rfloor} \geq (1-1/n)^{(n-1)/2} \geq 1/2$, where the last inequality is according to that $h(\beta) = (\frac{\beta}{\beta+1})^{\beta/2}$ is a decreasing function and $\lim_{\beta\to\infty} h(\beta) = 1/\sqrt{e} \geq 1/2$. For $N = n/2$ and $t < N$, we know that

$$\begin{aligned}
&\mathbb{E}\left(\max_{|y|\leq R_y} H_{\mathrm{CC}}(x_t, y) - \min_{|x|\leq R_x} H_{\mathrm{CC}}(x, y_t)\right) \\
&\geq \mathbb{E}\left(\max_{|y|\leq R_y} H_{\mathrm{CC}}(x_t, y) - \min_{|x|\leq R_x} H_{\mathrm{CC}}(x, y_t) \;\Big|\; t < T\right)\mathbb{P}\left[T > t\right] \\
&= \mathbb{E}\left(\max_{|y|\leq R_y} H_{\mathrm{CC}}(0, y) - \min_{|x|\leq R_x} H_{\mathrm{CC}}(x, 0) \;\Big|\; t < T\right)\mathbb{P}\left[T > t\right] \\
&= \frac{LR_xR_y}{2}\mathbb{P}\left[T \geq N\right] \geq LR_xR_y/4 \geq \varepsilon.
\end{aligned}$$

Thus, to find $(\hat{x}, \hat{y}) \in \mathcal{X} \times \mathcal{Y}$ such that $\mathbb{E} \max_{|y| \leq R_y} H_{\text{CC}}(\hat{x}, y) - \mathbb{E} \min_{|x| \leq R_x} H_{\text{CC}}(x, \hat{y}) < \varepsilon$, algorithm $\mathcal{A}$ needs at least $N = \Omega(n)$ PIFO queries. ■

D.4 Proofs for the Nonconvex-Strongly-Concave Case

With f_{NCSC}, ϕ_{NCSC} and $\{f_{\text{NCSC},i}\}_{i=1}^n$ defined in Definition 36, we have the following proposition.

Proposition 50 *For any $n \geq 2$, $L/\mu_y \geq 4$ and $\varepsilon^2 \leq \frac{\Delta L^2 \alpha}{6967296 n \mu_y}$, the following properties hold:*

1. *$\{f_{\text{NCSC},i}\}_{i=1}^n$ is L-average smooth and each $f_{\text{NCSC},i}$ is $(-\mu_x, \mu_y)$-convex-concave.*
2. *$\phi_{\text{NCSC}}(\mathbf{0}) - \min_{\mathbf{x} \in \mathbb{R}^{m+1}} \phi_{\text{NCSC}}(\mathbf{x}^*) \leq \Delta$.*
3. *$m \geq 2$ and for $M = m - 1$, $\min_{\mathbf{x} \in \mathcal{F}_M} \|\nabla \phi_{\text{NCSC}}(\mathbf{x})\|_2 \geq 9\varepsilon$.*

Proof [Proof of Proposition 50]

1. By Proposition 18 and Lemma 40, we have that $f_{\text{NCSC},i}$ is $(-\mu_1, \mu_2)$-convex-concave and $\{f_{\text{NCSC},i}\}_{i=1}^n$ l-average smooth where

$$\mu_1 = \frac{45(\sqrt{3}-1)L^2\alpha}{256 n \mu_y} \leq \mu_x, \qquad \mu_2 = \mu_y,$$
$$l = \frac{L}{8\sqrt{n}} \sqrt{4n + \frac{256 n \mu_y^2}{L^2} + 16200 \frac{\alpha^2 L^2}{256 n \mu_y^2}} \leq \frac{L}{8\sqrt{n}} \left(2\sqrt{n} + 16 \frac{\sqrt{n}\mu_y}{L} + \frac{45\sqrt{2}\alpha L}{8\sqrt{n}\mu_y} \right) \leq L.$$

   Thus each component $f_{\text{NCSC, i}}$ is $(-\mu_x, \mu_y)$-convex-concave and $\{f_{\text{NCSC},i}\}_{i=1}^n$ is L-smooth.

2. We first give a closed form expression of ϕ_{NCSC}. For simplicity, we omit the parameters of $\widehat{\mathbf{B}}$. It is easy to check

$$f_{\text{NCSC}}(\mathbf{x}, \mathbf{y}) = \frac{L}{16\sqrt{n}} \left\langle \mathbf{y}, \widehat{\mathbf{B}}\mathbf{x} \right\rangle - \frac{\mu_y}{2} \|\mathbf{y}\|_2^2 + \frac{\sqrt{\alpha}\lambda L}{16\sqrt{n}\mu_y} \sum_{i=1}^m \Gamma\left(\frac{1}{4}\sqrt{\frac{\sqrt{\alpha}L}{\lambda\sqrt{n}}} x_i \right) - \frac{1}{4}\sqrt{\frac{\lambda L}{\sqrt{n}}} \langle \mathbf{e}_1, \mathbf{y} \rangle.$$

   Then we can rewrite $f_{\text{NCSC}}(\mathbf{x}, \mathbf{y})$ as

$$f_{\text{NCSC}}(\mathbf{x}, \mathbf{y}) = -\frac{\mu_y}{2} \left\| \mathbf{y} - \frac{1}{\mu_y}\left(\frac{L}{16\sqrt{n}} \widehat{\mathbf{B}}\mathbf{x} - \frac{1}{4}\sqrt{\frac{\lambda L}{\sqrt{n}}} \mathbf{e}_1 \right) \right\|_2^2 + \frac{1}{2\mu_y} \left\| \frac{L}{16\sqrt{n}} \widehat{\mathbf{B}}\mathbf{x} - \frac{1}{4}\sqrt{\frac{\lambda L}{\sqrt{n}}} \mathbf{e}_1 \right\|_2^2 + \frac{\sqrt{\alpha}\lambda L}{16\sqrt{n}\mu_y} \sum_{i=1}^m \Gamma\left(\frac{1}{4}\sqrt{\frac{\sqrt{\alpha}L}{\lambda\sqrt{n}}} x_i \right).$$

It follows that

$$\begin{aligned}\phi_{\mathrm{NCSC}}(\mathbf{x}) &= \frac{1}{2\mu_y}\left\|\frac{L}{16\sqrt{n}}\widehat{\mathbf{B}}\mathbf{x} - \frac{1}{4}\sqrt{\frac{\lambda L}{\sqrt{n}}}\mathbf{e}_1\right\|_2^2 + \frac{\sqrt{\alpha}\lambda L}{16\sqrt{n}\mu_y}\sum_{i=1}^{m}\Gamma\left(\frac{1}{4}\sqrt{\frac{\sqrt{\alpha}L}{\lambda\sqrt{n}}}x_i\right) \\ &= \frac{L^2}{512n\mu_y}\left\|\widehat{\mathbf{B}}\mathbf{x}\right\|_2^2 - \frac{L}{64\mu_y}\sqrt{\frac{\sqrt{\alpha}\lambda L}{n^{3/2}}}\langle\mathbf{x},\mathbf{e}_1\rangle + \frac{\sqrt{\alpha}\lambda L}{16\sqrt{n}\mu_y}\sum_{i=1}^{m}\Gamma\left(\frac{1}{4}\sqrt{\frac{\sqrt{\alpha}L}{\lambda\sqrt{n}}}x_i\right) \\ &\quad + \frac{\lambda L}{32\sqrt{n}\mu_y}.\end{aligned}$$

Letting $\tilde{\mathbf{x}} = \frac{1}{4}\sqrt{\frac{\sqrt{\alpha}L}{\lambda\sqrt{n}}}\mathbf{x}$, we have

$$\tilde{\phi}_{\mathrm{NCSC}}(\tilde{\mathbf{x}}) \triangleq \phi_{\mathrm{NCSC}}(\mathbf{x}) = \frac{\lambda L}{16\mu_y\sqrt{\alpha n}}\left(\frac{1}{2}\left\|\widehat{\mathbf{B}}\tilde{\mathbf{x}}\right\|_2^2 - \sqrt{\alpha}\langle\tilde{\mathbf{x}},\mathbf{e}_1\rangle + \alpha\sum_{i=1}^{m}\Gamma(\tilde{x}_i)\right) + \frac{\lambda L}{32\sqrt{n}\mu_y}.$$

By Proposition 43,

$$\begin{aligned}&\phi_{\mathrm{NCSC}}(\mathbf{0}) - \min_{\mathbf{x}\in\mathbb{R}^{m+1}}\phi_{\mathrm{NCSC}}(\mathbf{x}) = \tilde{\phi}_{\mathrm{NCSC}}(\mathbf{0}) - \min_{\tilde{\mathbf{x}}\in\mathbb{R}^{m+1}}\tilde{\phi}_{\mathrm{NCSC}}(\tilde{\mathbf{x}}) \\ &= \frac{\lambda L}{16\mu_y\sqrt{\alpha n}}\left(\frac{\sqrt{\alpha}}{2} + 10\alpha m\right) \le \frac{165888n\mu_y\varepsilon^2}{L^2\alpha} + \frac{3311760n\mu_y\varepsilon^2 m}{L^2\sqrt{\alpha}} \\ &\le \frac{165888}{3483648}\Delta + \frac{3317760}{3483648}\Delta \le \Delta.\end{aligned}$$

3. Since $\alpha \le 1$, we have $\frac{\Delta L^2\sqrt{\alpha}}{3483648n\varepsilon^2\mu_y} \ge \frac{\Delta L^2\alpha}{3483648n\varepsilon^2\mu_y} \ge 2$ and consequently $m \ge 2$. By Proposition 43,

$$\begin{aligned}\min_{\mathbf{x}\in\mathcal{F}_M}\|\nabla\phi_{\mathrm{NCSC}}(\mathbf{x})\|_2 &= \frac{1}{4}\sqrt{\frac{\sqrt{\alpha}L}{\lambda\sqrt{n}}}\min_{\hat{\mathbf{x}}\in\mathcal{F}_M}\left\|\nabla\tilde{\phi}_{\mathrm{NCSC}}(\tilde{\mathbf{x}})\right\|_2 \\ &\ge \frac{1}{4}\sqrt{\frac{\sqrt{\alpha}L}{\lambda\sqrt{n}}}\frac{\lambda L}{16\mu_y\sqrt{\alpha n}}\frac{\alpha^{3/4}}{4} \ge 9\varepsilon.\end{aligned}$$

This completes the proof. ■

D.5 Results for the Smooth Cases

In this subsection, we give the formal statements of the lower bounds in Table 2.

Function class We develop lower bounds for PIFO algorithms that find a suboptimal solution or near stationary point of Problem 1 in the following sets

$$\mathcal{F}_{\mathrm{CC}}^{*}(R_x, R_y, L, \mu_x, \mu_y) = \left\{f(\mathbf{x},\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n}f_i(\mathbf{x},\mathbf{y}) \;\middle|\; f\colon \mathcal{X}\times\mathcal{Y}\to\mathbb{R},\ \mathrm{diam}(\mathcal{X}) \le 2R_x,\right.$$

$$\text{diam}(\mathcal{Y}) \le 2R_y,\ f_i \text{ is } L\text{-smooth}, f \text{ is } (\mu_x, \mu_y)\text{-convex-concave} \Bigg\},$$

$$\mathcal{F}^*_{\text{NCC}}(\Delta, L, \mu_x, \mu_y) = \Bigg\{ f(\mathbf{x}, \mathbf{y}) = \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x}, \mathbf{y}) \;\Big|\; f\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R},\ \phi(\mathbf{0}) - \inf_{\mathbf{x}\in\mathcal{X}} \phi(\mathbf{x}) \le \Delta,$$

$$f_i \text{ is } L\text{-smooth}, f \text{ is } (-\mu_x, \mu_y)\text{-convex-concave} \Bigg\}.$$

We can also add a condition that each component function is (μ_x, μ_y)-convex-concave to the definition. This induces a more restrictive function class but will not affect our construction. Such a definition better matches the assumptions of some upper bounds, e.g., Luo et al. (2019).

Optimization complexity We formally define the optimization complexity as follows.

Definition 51 *The optimization complexity with respect to function classes $\mathcal{F}^*_{\text{CC}}(\Delta, R, L, \mu)$ and $\mathcal{F}^*_{\text{NCC}}(\Delta, R, L, \mu)$ is defined as*

$$\mathfrak{m}^{\text{CC}}_*(\varepsilon, R_x, R_y, L, \mu_x, \mu_y) \triangleq \inf_{\mathcal{A}\in\mathscr{A}} \sup_{f\in\mathcal{F}_{\text{CC}}(R_x,R_y,L,\mu_x,\mu_y)} T(\mathcal{A}, f, \varepsilon),$$
$$\mathfrak{m}^{\text{NCC}}_*(\varepsilon, \Delta, L, \mu_x, \mu_y) \triangleq \inf_{\mathcal{A}\in\mathscr{A}} \sup_{f\in\mathcal{F}_{\text{NCC}}(\Delta,L,\mu_x,\mu_y)} T(\mathcal{A}, f, \varepsilon),$$

where $T(\mathcal{A}, f, \varepsilon)$ is defined in Definition 13 with $\mathcal{F}_{\text{CC}}$ and $\mathcal{F}_{\text{NCC}}$ replaced by $\mathcal{F}^*_{\text{CC}}$ and $\mathcal{F}^*_{\text{NCC}}$.

The lower bounds are listed as follows. Let $\kappa_x = L/\mu_x$ and $\kappa_y = L/\mu_y$ denote the condition number if they are well-defined.

Theorem 52 *Let $n \ge 4$ be a positive integer and $L, \mu_x, \mu_y, R_x, R_y, \varepsilon$ be positive parameters. Assume additionally that $\kappa_x \ge \kappa_y \ge 2$ and $\varepsilon = \mathcal{O}\left(\min\left\{\frac{n^2\mu_x R_x^2}{\kappa_x\kappa_y}, \mu_y R_y^2, \mu_y R_y^2\right\}\right)$. Then we have*

$$\mathfrak{m}^{\text{CC}}_*(\varepsilon, R_x, R_y, L, \mu_x, \mu_y) = \begin{cases} \Omega\left(\left(n+\sqrt{\kappa_x\kappa_y}\right)\log\left(1/\varepsilon\right)\right), & \textit{for } \kappa_x, \kappa_y = \Omega(n), \\ \Omega\left(\left(n+\sqrt{\kappa_x n}\right)\log\left(1/\varepsilon\right)\right), & \textit{for } \kappa_x = \Omega(n), \kappa_y = \mathcal{O}(n), \\ \Omega\left(n\right), & \textit{for } \kappa_x, \kappa_y = \mathcal{O}(n). \end{cases}$$

The best known upper bound complexity in this case for IFO/PIFO algorithms is $\mathcal{O}\left(\left(n + \frac{\sqrt{n}L}{\min\{\mu_x,\mu_y\}}\right)\log(1/\varepsilon)\right)$ (Luo et al., 2019). There still exists a $\sqrt{n}$ gap to our lower bound.

Theorem 53 *Let $n \ge 2$ be a positive integer and $L, \mu_y, R_x, R_y, \varepsilon$ be positive parameters. Assume additionally that $\kappa_y \ge 2$ and $\varepsilon = \mathcal{O}\left(\min\left\{LR_x^2, \mu_y R_y^2\right\}\right)$. Then we have*

$$\mathfrak{m}^{\text{CC}}_*(\varepsilon, R_x, R_y, L, 0, \mu_y) = \begin{cases} \Omega\left(n + R_x\sqrt{\frac{nL}{\varepsilon}} + R_x\sqrt{\frac{L\kappa_y}{\varepsilon}} + \sqrt{n\kappa_y}\log\left(\frac{1}{\varepsilon}\right)\right), & \textit{for } \kappa_y = \Omega(n), \\ \Omega\left(n + R_x\sqrt{\frac{nL}{\varepsilon}} + R_x\sqrt{\frac{L\kappa_y}{\varepsilon}}\right), & \textit{for } \kappa_y = \mathcal{O}(n). \end{cases}$$

Theorem 54 *Let $n \geq 2$ be a positive integer and L, R_x, R_y, ε be positive parameters. Assume additionally that $\varepsilon \leq \frac{LR_xR_y}{4}$. Then we have*

$$\mathfrak{m}_*^{\mathrm{CC}}(\varepsilon, R_x, R_y, L, 0, 0) = \Omega\left(n + \frac{LR_xR_y}{\varepsilon} + (R_x + R_y)\sqrt{\frac{nL}{\varepsilon}}\right).$$

Theorem 55 *Let $n \geq 2$ be a positive integer and $L, \mu_x, \mu_y, \Delta, \varepsilon$ be positive parameters. Assume additionally that $\varepsilon^2 \leq \frac{\Delta L^2 \alpha}{27216 n^2 \mu_y}$, where $\alpha = \min\left\{1, \frac{8(\sqrt{3}+1)n^2\mu_x\mu_y}{45L^2}, \frac{n^2\mu_y}{90L}\right\}$. Then we have*

$$\mathfrak{m}_*^{\mathrm{NCC}}(\varepsilon, \Delta, L, \mu_x, \mu_y) = \Omega\left(n + \frac{\Delta L^2 \sqrt{\alpha}}{n\mu_y\varepsilon^2}\right).$$

For $\kappa_y \geq n^2/90$, we have

$$\Omega\left(n + \frac{\Delta L^2 \sqrt{\alpha}}{n\mu_y\varepsilon^2}\right) = \Omega\left(n + \frac{\Delta L}{\varepsilon^2}\min\left\{\sqrt{\kappa_y}, \sqrt{\frac{\mu_x}{\mu_y}}\right\}\right).$$

With $\mu_x = L$, we obtain the result in Table 2.

The proofs of these theorems are similar to those of Theorems 22 to 25. We just list some key lemmas here and omit the lengthy proofs.

When f is convex-concave, without loss of generality, we still assume $\mu_x \leq \mu_y$. The hard instances for Theorems 52 to 54 can be directly derived previous constructions. Specially, for the hard instances constructed in Definitions 26, 30 and 33. it suffices to replace L by $\tilde{L} \triangleq \sqrt{\frac{2(L^2-2\mu_y^2)}{n} + 2\mu_y^2}$. One can check that by Proposition 59 and Lemma 40, each component function is L-smooth and (μ_x, μ_y)-convex-concave after this replacement. Moreover, we have $\tilde{L} \geq \sqrt{\frac{2}{n}}L$ for $n \geq 2$ and $\tilde{L} \leq \sqrt{\frac{4}{n}}L$ as long as $n \geq 4$ and $L^2/\mu_y^2 \geq n-2 \geq 2$. For the hard instances constructed in the proofs of Theorems 28 and 32, there are also corresponding lower bounds in terms of the smoothness parameter. And the hard instances constructed in the proofs of Lemmas 29 and 35 also have L-smooth and (μ_x, μ_y)-convex-concave component functions. As a result, the lower bounds in terms of the average smooth parameter can be transformed into those in terms of the smooth parameter.

When f is nonconvex in $\mathbf{x}$ and strongly concave in $\mathbf{y}$, the hard instance is constructed as follows.

Definition 56 *For fixed $L, \mu_x, \mu_y, \Delta, n$, we define $f_{\mathrm{NCSC},i} : \mathbb{R}^{m+1} \times \mathbb{R}^{m+1} \to \mathbb{R}$ as follows*

$$f_{\mathrm{NCSC},i}^*(\mathbf{x}, \mathbf{y}) = \lambda r_i^{\mathrm{NCC}}\left(\mathbf{x}/\beta, \mathbf{y}/\beta; m+1, \sqrt[4]{\alpha}, \mathbf{c}_*^{\mathrm{NCSC}}\right), \text{ for } 1 \leq i \leq n,$$

where

$$\alpha = \min\left\{1, \frac{n^2\mu_y}{90L}, \frac{8(\sqrt{3}+1)n^2\mu_x\mu_y}{45L^2}\right\}, \ \mathbf{c}_*^{\mathrm{NCSC}} = \left(\frac{4n\mu_y}{L}, \frac{\sqrt{\alpha}L}{4n\mu_y}, \sqrt[4]{\alpha}\right),$$
$$\lambda = \frac{82944 n^3 \mu_y^2 \varepsilon^2}{L^3 \alpha}, \ \beta = 2\sqrt{\lambda n/L} \ \text{ and } \ m = \left\lfloor \frac{\Delta L^2 \sqrt{\alpha}}{217728 n^2 \varepsilon^2 \mu_y} \right\rfloor.$$

Consider the minimax problem

$$\min_{\mathbf{x}\in\mathbb{R}^{m+1}} \max_{\mathbf{y}\in\mathbb{R}^{m+1}} f^*_{\mathrm{NCSC}}(\mathbf{x},\mathbf{y}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f^*_{\mathrm{NCSC},i}(\mathbf{x},\mathbf{y}). \tag{43}$$

Define $\phi^*_{\mathrm{NCSC}}(\mathbf{x}) = \max_{\mathbf{y}\in\mathbb{R}^{m+1}} f^*_{\mathrm{NCSC}}(\mathbf{x},\mathbf{y})$.

Then we have the following proposition, whose proof is similar to that of Proposition 50 and is omitted.

Proposition 57 *For any* $n \geq 2$, $L/\mu_y \geq 4$ *and* $\varepsilon^2 \leq \frac{\Delta L^2 \alpha}{435456 n^2 \mu_y}$, *the following properties hold:*

1. $f^*_{\mathrm{NCSC},i}$ *is* L*-smooth and* $(-\mu_x, \mu_y)$*-convex-concave.*
2. $\phi^*_{\mathrm{NCSC}}(\mathbf{0}_{m+1}) - \min_{\mathbf{x}\in\mathbb{R}^{m+1}} \phi^*_{\mathrm{NCSC}}(\mathbf{x}^*) \leq \Delta$.
3. $m \geq 2$ *and for* $M = m-1$, $\min_{\mathbf{x}\in\mathcal{F}_M} \|\nabla \phi^*_{\mathrm{NCSC}}(\mathbf{x})\|_2 \geq 9\varepsilon$.

The proof of Theorem 55 is similar to that of Theorem 25 and is omitted.

Appendix E. Details for Section 6

In this section, we provide the details and omitted proofs in Section 6.

E.1 The Setup

Function class We develop lower bounds for PIFO algorithms that find the suboptimal solution or near stationary point of Problem (14) in the following four sets.

$$\mathcal{F}^*_{\mathrm{C}}(R, L, \mu) = \left\{ f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}) \;\middle|\; f: \mathcal{X} \to \mathbb{R},\ \mathrm{diam}(\mathcal{X}) \leq 2R, \right.$$
$$\left. f_i \text{ is } L\text{-smooth}, f \text{ is } \mu\text{-strongly convex} \right\},$$

$$\mathcal{F}^*_{\mathrm{NC}}(\Delta, L, \mu) = \left\{ f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}) \;\middle|\; f: \mathcal{X} \to \mathbb{R},\ f(\mathbf{0}) - \inf_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) \leq \Delta, \right.$$
$$\left. f_i \text{ is } L\text{-smooth}, f \text{ is } (-\mu)\text{-weakly convex} \right\},$$

$$\mathcal{F}_{\mathrm{C}}(R, L, \mu) = \left\{ f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}) \;\middle|\; f: \mathcal{X} \to \mathbb{R},\ \mathrm{diam}(\mathcal{X}) \leq 2R, \right.$$
$$\left. \{f_i\}_{i=1}^{n} \text{ is } L\text{-average smooth}, f \text{ is } \mu\text{-strongly convex} \right\}.$$

$$\mathcal{F}_{\mathrm{NC}}(\Delta, L, \mu) = \left\{ f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}) \;\middle|\; f: \mathcal{X} \to \mathbb{R},\ f(\mathbf{0}) - \inf_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) \leq \Delta, \right.$$

$$\{f_i\}_{i=1}^n \text{ is } L\text{-average smooth}, f \text{ is } (-\mu)\text{-weakly convex} \Big\}.$$

For the definitions of $\mathcal{F}_{\mathrm{C}}^*$ and $\mathcal{F}_{\mathrm{NC}}^*$, we can also add a condition that each component function is μ-strongly convex or $(-\mu)$-weakly convex to the definitions respectively. This induces a more restrictive function class but will not affect our construction. In fact, the component function of the hard instances constructed in Woodworth and Srebro (2016), Hannah et al. (2018) and Zhou and Gu (2019) is also μ-strongly convex or $(-\mu)$-weakly convex.

Optimization complexity We formally define the optimization complexity as follows.

Definition 58 *For a function f, a PIFO algorithm $\mathcal{A}$ and a tolerance $\varepsilon > 0$, the number of queries needed by $\mathcal{A}$ to find ε-suboptimal solution to the Problem (14) or the ε-stationary point of $f(\mathbf{x})$ is defined as*

$$T(\mathcal{A}, f, \varepsilon) = \begin{cases} \inf\left\{T \in \mathbb{N} \,|\, \mathbb{E} f(\mathbf{x}_{\mathcal{A},T}) - \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) < \varepsilon\right\}, & \text{if } f \in \mathcal{F}_{\mathrm{C}}^*(R, L, \mu) \cup \mathcal{F}_{\mathrm{C}}(R, L, \mu), \\ \inf\left\{T \in \mathbb{N} \,|\, \mathbb{E}\left\|\nabla f(\mathbf{x}_{\mathcal{A},T})\right\|_2 < \varepsilon\right\}, & \text{if } f \in \mathcal{F}_{\mathrm{NC}}^*(\Delta, L, \mu) \cup \mathcal{F}_{\mathrm{NC}}(\Delta, L, \mu) \end{cases}$$

where $\mathbf{x}_{\mathcal{A},T}$ is the point obtained by the algorithm $\mathcal{A}$ at time-step T.

Furthermore, the optimization complexity with respect to these function classes are defined as

$$\begin{aligned}
\mathfrak{m}_*^{\mathrm{C}}(\varepsilon, R, L, \mu) &\triangleq \inf_{\mathcal{A}\in\mathscr{A}} \sup_{f\in\mathcal{F}_{\mathrm{C}}^*(R,L,\mu)} T(\mathcal{A}, f, \varepsilon), \\
\mathfrak{m}^{\mathrm{C}}(\varepsilon, R, L, \mu) &\triangleq \inf_{\mathcal{A}\in\mathscr{A}} \sup_{f\in\mathcal{F}_{\mathrm{C}}(R,L,\mu)} T(\mathcal{A}, f, \varepsilon). \\
\mathfrak{m}_*^{\mathrm{NC}}(\varepsilon, \Delta, L, \mu) &\triangleq \inf_{\mathcal{A}\in\mathscr{A}} \sup_{f\in\mathcal{F}_{\mathrm{NC}}^*(\Delta,L,\mu)} T(\mathcal{A}, f, \varepsilon), \\
\mathfrak{m}^{\mathrm{NC}}(\varepsilon, \Delta, L, \mu) &\triangleq \inf_{\mathcal{A}\in\mathscr{A}} \sup_{f\in\mathcal{F}_{\mathrm{NC}}(\Delta,L,\mu)} T(\mathcal{A}, f, \varepsilon).
\end{aligned}$$

E.2 More Properties of the Hard Instances

In this subsection, we present more properties of the hard instance $\{r_i\}_{i=1}^n$ constructed in Section 6.1. First, We can determine the smoothness and strong convexity parameters of r_i as follows.

Proposition 59 *Suppose that $0 \le \omega, \zeta \le \sqrt{2}$ and $c_1 \ge 0$.*

1. ***Convex case.*** *For $c_2 = 0$, we have that r_i is $(2n+c_1)$-smooth and c_1-strongly-convex, and $\{r_i\}_{i=1}^n$ is L'-average smooth where*

$$L' = \sqrt{\frac{4}{n}\left[(n+c_1)^2 + n^2\right] + c_1^2}.$$

2. ***Non-convex case.*** *For $c_1 = 0$, we have that r_i is $(2n+180c_2)$-smooth and $[-45(\sqrt{3}-1)c_2]$-weakly-convex, and $\{r_i\}_{i=1}^n$ is $4\sqrt{n+4050c_2^2}$-average smooth.*

The proof of Proposition 59 is given in Appendix E.7.

Recall the subspaces $\{\mathcal{F}_k\}_{k=0}^m$ which are defined as

$$\mathcal{F}_k = \begin{cases} \text{span}\{\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k\}, & \text{for } 1 \leq k \leq m, \\ \{\mathbf{0}\}, & \text{for } k = 0. \end{cases}$$

When we apply a PIFO algorithm $\mathcal{A}$ to solve the Problem (15), Lemma 37 implies that $\mathbf{x}_t = \mathbf{0}$ will hold until algorithm $\mathcal{A}$ draws the component r_1 or calls the FO. Then for any $t < T_1 = \min_t\{t : i_t = 1 \text{ or } a_t = 1\}$, we have $\mathbf{x}_t \in \mathcal{F}_0$ while $\mathbf{x}_{T_1} \in \mathcal{F}_1$ holds. The value of T_1 can be regarded as the smallest integer such that $\mathbf{x}_{T_1} \in \mathcal{F}_1 \setminus \mathcal{F}_0$ could hold. Similarly, for $T_1 \leq t < T_2 = \min_t\{t > T_1 : i_t = 2 \text{ or } a_t = 1\}$, there holds $\mathbf{x}_t \in \mathcal{F}_1$ while we can ensure that $\mathbf{x}_{T_2} \in \mathcal{F}_2$.

We can define T_k to be the smallest integer such that $\mathbf{x}_{T_k} \in \mathcal{F}_k \setminus \mathcal{F}_{k-1}$ could hold. We give the formal definition of T_k recursively and connect it to geometrically distributed random variables in the following corollary.

Corollary 60 *Assume we employ a PIFO algorithm $\mathcal{A}$ to solve the Problem (15). Let*

$$T_0 = 0, \quad \textit{and} \;\; T_k = \min_t\{t : t > T_{k-1}, i_t \equiv k \pmod{n} \;\textit{or}\; a_t = 1\} \;\; \textit{for } k \geq 1. \tag{44}$$

Then we have

$$\mathbf{x}_t \in \mathcal{F}_{k-1}, \qquad \textit{for } t < T_k, k \geq 1.$$

Moreover, the random variables $\{Y_k\}_{k\geq 1}$ such that $Y_k \triangleq T_k - T_{k-1}$ are mutual independent and Y_k follows a geometric distribution with success probability $p_{k'} + q - p_{k'}q$ where $k' \in [n]$ satisfies $k' \equiv k \pmod{n}$.

The proof of Corollary 60 is similar to that of Corollary 16. The basic idea of our analysis is that we guarantee that the minimizer of r does not lie in $\mathcal{F}_k$ for $k < m$ and assure that the PIFO algorithm extends the space of $\text{span}\{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_t\}$ slowly with t increasing. We know that $\text{span}\{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{T_k}\} \subseteq \mathcal{F}_{k-1}$ by Corollary 60. Hence, T_k is just the quantity that measures how $\text{span}\{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_t\}$ expands. Note that T_k can be written as the sum of geometrically distributed random variables. Recalling Lemma 10, we can obtain how many PIFO calls we need.

Lemma 61 *Let $H_r(\mathbf{x})$ be a criterion of measuring how $\mathbf{x}$ is close to solution to Problem (15). If M satisfies $1 \leq M < m$, $\min_{\mathbf{x}\in\mathcal{X}\cap\mathcal{F}_M} H_r(\mathbf{x}) \geq 9\varepsilon$ and $N = \frac{n(M+1)}{4(1+c_0)}$, then we have*

$$\min_{t\leq N} \mathbb{E} H_r(\mathbf{x}_t) \geq \varepsilon$$

Remark 62 *If $r(\mathbf{x})$ is convex in $\mathbf{x}$, we set $H_r(\mathbf{x}) = r(\mathbf{x}) - \min_{\mathbf{x}\in\mathcal{X}} r(\mathbf{x})$. If $r(\mathbf{x})$ is non-convex, we set $H_r(\mathbf{x}) = \|\nabla r(\mathbf{x})\|_2$.*

The proof of Lemma 61 is similar to that of Lemma 17.

E.3 Main Results

In this subsection, we present the formal statements of the lower bounds in Tables 3 and 4.

E.3.1 Smooth Cases

We first focus on the smooth cases, i.e., the results in Table 3. When $\mu \neq 0$, the conditon number is denoted by $\kappa = L/\mu$.

When f is strongly-convex, we have the following result.

Theorem 63 *Let $n \geq 2$ be a positive integer and L, μ, R, ε be positive parameters. Assume additionally that $\kappa = L/\mu \geq 2$ and $\varepsilon \leq LR^2/4$. Then we have*

$$\mathfrak{m}_*^{\mathrm{C}}(\varepsilon, R, L, \mu) = \begin{cases} \Omega\left((n+\sqrt{\kappa n})\log(1/\varepsilon)\right), & \textit{for } \kappa = \Omega(n), \\ \Omega\left(n + \left(\frac{n}{1+(\log(n/\kappa))_+}\right)\log(1/\varepsilon)\right), & \textit{for } \kappa = \mathcal{O}(n). \end{cases}$$

From the analysis in Appendix E.4, our construction only requires the dimension to be $\mathcal{O}(1 + \sqrt{\kappa/n}\log(1/\varepsilon))$, which is much smaller than $\tilde{\mathcal{O}}(\kappa n/\varepsilon)$ in Woodworth and Srebro (2016).

Next, we give the lower bound when the objective function is not strongly convex.

Theorem 64 *Let $n \geq 2$ be a positive integer and L, R, ε be positive parameters. Assume additionally that $\varepsilon \leq LR^2/4$. Then we have*

$$\mathfrak{m}_*^{\mathrm{C}}(\varepsilon, R, L, 0) = \Omega\left(n+R\sqrt{nL/\varepsilon}\right)$$

From Appendix E.5, our construction requires the dimension to be $\mathcal{O}(1+R\sqrt{L/(n\varepsilon)})$, which is much smaller than $\tilde{\mathcal{O}}(L^2R^4/\varepsilon^2)$ in Woodworth and Srebro (2016).

Finally, we give the lower bound when the objective function is non-convex.

Theorem 65 *Let $n \geq 2$ be a positive integer and $L, \mu, \Delta, \varepsilon$ be positive parameters. Assume additionally that $\varepsilon^2 \leq \frac{\Delta L \alpha}{81648n}$, where $\alpha = \min\left\{1, \frac{(\sqrt{3}+1)n\mu}{30L}, \frac{n}{180}\right\}$. Then we have*

$$\mathfrak{m}_*^{\mathrm{NC}}(\varepsilon, \Delta, L, \mu) = \Omega\left(n + \frac{\Delta L\sqrt{\alpha}}{\varepsilon^2}\right)$$

For $n > 180$, we have

$$\Omega\left(n + \frac{\Delta L\sqrt{\alpha}}{\varepsilon^2}\right) = \Omega\left(n + \frac{\Delta}{\varepsilon^2}\min\{L, \sqrt{n\mu L}\}\right).$$

From the analysis in Appendix E.6, our construction only requires the dimension to be $\mathcal{O}\left(1 + \frac{\Delta}{\varepsilon^2}\min\{L/n, \sqrt{\mu L/n}\}\right)$, which is much smaller than $\mathcal{O}\left(\frac{\Delta}{\varepsilon^2}\min\{L, \sqrt{n\mu L}\}\right)$ in Zhou and Gu (2019).

E.3.2 Average Smooth Case

Then we give the results for the average smooth cases, i.e., the results in Table 4. When $\mu \neq 0$, the condition number is still denoted by $\kappa = L/\mu$.

When f is strongly convex, we have the following result.

Theorem 66 *Let $n \geq 4$ be a positive integer and L, μ, R, ε be positive parameters. Assume additionally that $\kappa = L/\mu \geq 2$ and $\varepsilon \leq LR^2/4$. Then we have*

$$\mathfrak{m}^{\mathrm{C}}(\varepsilon, R, L, \mu) = \begin{cases} \Omega\left(\left(n + n^{3/4}\sqrt{\kappa}\right)\log\left(1/\varepsilon\right)\right), & \text{for } \kappa = \Omega(\sqrt{n}), \\ \Omega\left(n + \left(\frac{n}{1 + (\log(\sqrt{n}/\kappa))_+}\right)\log\left(1/\varepsilon\right)\right), & \text{for } \kappa = \mathcal{O}(\sqrt{n}). \end{cases}$$

From the analysis in Appendix E.4, our construction only requires the dimension to be $\mathcal{O}(1 + \sqrt{\kappa}/\sqrt[4]{n}\log(1/\varepsilon))$, which is much smaller than $\mathcal{O}(n + n^{3/4}\sqrt{\kappa}\log(1/\varepsilon))$ in Zhou and Gu (2019).

The next theorem gives the lower bound when f is only convex.

Theorem 67 *Let $n \geq 2$ be a positive integer and L, R, ε be positive parameters. Assume additionally that $\varepsilon \leq LR^2/4$. Then we have*

$$\mathfrak{m}^{\mathrm{C}}(\varepsilon, R, L, 0) = \Omega\left(n + Rn^{3/4}\sqrt{L/\varepsilon}\right)$$

From Appendix E.5, our construction requires the dimension to be $\mathcal{O}(1 + R\sqrt{L/(\sqrt{n}\varepsilon)})$, which is much smaller than $\mathcal{O}(n + n^{3/4}\sqrt{L/\varepsilon})$ in Zhou and Gu (2019).

Finally, we give the lower bound when the objective function is non-convex.

Theorem 68 *Let $n \geq 2$ be a positive integer and $L, \mu, \Delta, \varepsilon$ be positive parameters. Assume additionally that $\varepsilon^2 \leq \frac{\Delta L \alpha}{435456\sqrt{n}}$, where $\alpha = \min\left\{1, \frac{8(\sqrt{3}+1)\sqrt{n}\mu}{45L}, \sqrt{\frac{n}{270}}\right\}$. Then we have*

$$\mathfrak{m}_{\varepsilon}^{\mathrm{NC}}(\varepsilon, \Delta, L, \mu) = \Omega\left(n + \frac{\Delta L\sqrt{n\alpha}}{\varepsilon^2}\right)$$

For $n > 270$, we have

$$\Omega\left(n + \frac{\Delta L\sqrt{n\alpha}}{\varepsilon^2}\right) = \Omega\left(n + \frac{\Delta}{\varepsilon^2}\min\left\{\sqrt{n}L, n^{3/4}\sqrt{\mu L}\right\}\right).$$

From the analysis in Appendix E.6, our construction only requires the dimension to be $\mathcal{O}\left(1 + \frac{\Delta}{\varepsilon^2}\min\{L/\sqrt{n}, \sqrt{\mu L/\sqrt{n}}\}\right)$, which is much smaller than $\mathcal{O}\left(\frac{\Delta}{\varepsilon^2}\min\{\sqrt{n}L, n^{3/4}\sqrt{\mu L}\}\right)$ in Zhou and Gu (2019).

E.4 Construction for the Strongly-Convex Case

The analysis of lower bound complexity for the strongly convex case depends on the following construction.

Definition 69 *For fixed L, μ, R, n such that $L/\mu \geq 2$, let $\alpha = \sqrt{\frac{2(L/\mu-1)}{n} + 1}$. We define $f_{\mathrm{SC},i} : \mathbb{R}^m \to \mathbb{R}$ as follows*

$$f_{\mathrm{SC},i}(\mathbf{x}) = \lambda r_i \left(\mathbf{x}/\beta; m, 0, \sqrt{\frac{2}{\alpha+1}}, \mathbf{c}^{\mathrm{SC}} \right), \text{ for } 1 \leq i \leq n,$$

where $\mathbf{c}^{\mathrm{SC}} = \left(\frac{2n}{L/\mu-1}, 0, 1\right)$, $\lambda = \frac{2\mu R^2 \alpha n}{L/\mu-1}$ and $\beta = \frac{2R\sqrt{\alpha}n}{L/\mu-1}$. Consider the minimization problem

$$\min_{\mathbf{x}\in\mathcal{X}} f_{\mathrm{SC}}(\mathbf{x}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_{\mathrm{SC},i}(\mathbf{x}). \tag{45}$$

where $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^m : \|\mathbf{x}\|_2 \leq R\}$.

With this definition, we have the following proposition.

Proposition 70 *For any $n \geq 2$, $m \geq 2$, $f_{\mathrm{SC},i}$ and f_{SC} in Definition 69 satisfy:*

1. *$f_{\mathrm{SC},i}$ is L-smooth and μ-strongly-convex. Thus, f_{SC} is μ-strongly-convex.*
2. *The minimizer of the function f_{SC} is*

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^m} f_{\mathrm{SC}}(\mathbf{x}) = \frac{2R\sqrt{\alpha}}{\alpha-1}(q^1, q^2, \ldots, q^m)^\top,$$

   where $\alpha = \sqrt{\frac{2(L/\mu-1)}{n}+1}$ and $q = \frac{\alpha-1}{\alpha+1}$. Moreover, $f_{\mathrm{SC}}(\mathbf{x}^*) = -\frac{\mu R^2\alpha}{\alpha+1}$ and $\|\mathbf{x}^*\|_2 \leq R$.
3. *For $1 \leq k \leq m-1$, we have*

$$\min_{\mathbf{x}\in\mathcal{X}\cap\mathcal{F}_k} f_{\mathrm{SC}}(\mathbf{x}) - \min_{\mathbf{x}\in\mathcal{X}} f_{\mathrm{SC}}(\mathbf{x}) \geq \frac{\mu R^2\alpha}{\alpha+1} q^{2k}.$$

Proof

1. Just recall Proposition 59 and Lemma 40.
2. It is easy to check $f_{\mathrm{SC}}(\mathbf{x}) = \frac{\xi}{2}\|\mathbf{B}(m,0,\zeta)\mathbf{x}\|_2^2 + \frac{\mu}{2}\|\mathbf{x}\|_2^2 - \xi\beta\langle \mathbf{e}_1, \mathbf{x}\rangle$, where $\xi = \lambda/\beta^2 = (L-\mu)/(2n)$ and $\zeta = \sqrt{\frac{2}{\alpha+1}}$. Letting $\nabla f_{\mathrm{SC}}(\mathbf{x}) = \mathbf{0}$, we have

$$(\xi\mathbf{A}(m,0,\zeta) + \mu\mathbf{I})\mathbf{x} = \xi\beta\mathbf{e}_1. \tag{46}$$

   Since $\frac{\mu}{\xi} = \frac{2n\mu}{L-\mu} = \frac{4}{\alpha^2-1}$, $q = \frac{\alpha-1}{\alpha+1}$ is a root of the equation $z^2 - \left(2 + \frac{2n\mu}{L-\mu}\right)z + 1 = 0$. Note that $\zeta^2 + 1 + \frac{2n\mu}{L-\mu} = \frac{1}{q}$, one can check that the solution to Equations (46) is $\mathbf{x}^* = \frac{\beta(\alpha+1)}{2}(q^1, q^2, \ldots, q^m)^\top = \frac{2R\sqrt{\alpha}}{\alpha-1}(q^1, q^2, \ldots, q^m)^\top$, and $f_{\mathrm{SC}}(\mathbf{x}^*) = -\frac{\xi\beta^2(\alpha+1)q}{4} = -\frac{\lambda(\alpha-1)}{4} = -\frac{\mu R^2\alpha}{\alpha+1}$. Moreover, we have

$$\|\mathbf{x}^*\|_2^2 = \frac{4R^2\alpha}{(\alpha-1)^2}\cdot\frac{q^2-q^{2m+2}}{1-q^2} \leq \frac{4R^2\alpha}{(\alpha-1)^2}\cdot\frac{q^2}{1-q^2} = R^2.$$

3. If $\mathbf{x} \in \mathcal{F}_k$, $1 \le k < m$, then $x_{k+1} = x_{k+2} = \cdots = x_m = 0$.

   Let $\mathbf{y}$ be the first k coordinates of $\mathbf{x}$ and $\mathbf{A}_k$ be first k rows and columns of $\mathbf{A}(m, 0, \zeta)$. Then we can rewrite $f_{\mathrm{SC}}(\mathbf{x})$ as $f_k(\mathbf{y}) \triangleq f_{\mathrm{SC}}(\mathbf{x}) = \frac{\xi}{2}\mathbf{y}^\top \mathbf{A}_k \mathbf{y} - \xi\beta \langle \hat{\mathbf{e}}_1, \mathbf{y}\rangle$, where $\hat{\mathbf{e}}_1$ is the first k coordinates of $\mathbf{e}_1$. Let $\nabla f_k(\mathbf{y}) = \mathbf{0}_k$. By some calculation, the solution is

$$\frac{(\alpha-1)\beta q^k}{2(1+q^{2k+1})}\left(q^{-k} - q^k, q^{-k+1} - q^{k-1}, \ldots, q^{-1} - q^1\right)^\top.$$

   Thus, $\min_{\mathbf{x}\in\mathcal{F}_k} f_{\mathrm{SC}}(\mathbf{x}) = \min_{\mathbf{y}\in\mathbb{R}^k} f_k(\mathbf{y}) = -\frac{\lambda(\alpha-1)}{4}\cdot\frac{1-q^{2k}}{1+q^{2k+1}}$, and

$$\begin{aligned}
&\min_{\mathbf{x}\in\mathcal{X}\cap\mathcal{F}_k} f_{\mathrm{SC}}(\mathbf{x}) - \min_{\mathbf{x}\in\mathcal{X}} f_{\mathrm{SC}}(\mathbf{x}) \\
\ge &\min_{\mathbf{x}\in\mathcal{F}_k} f_{\mathrm{SC}}(\mathbf{x}) - f_{\mathrm{SC}}(\mathbf{x}^*) = \frac{\lambda(\alpha-1)}{4}\left(1 - \frac{1-q^{2k}}{1+q^{2k+1}}\right) \\
= &\frac{\lambda(\alpha-1)}{4} q^{2k}\frac{1+q}{1+q^{2k+1}} \ge \frac{\mu R^2\alpha}{\alpha+1} q^{2k}
\end{aligned}$$

This completes the proof. ■

With this hard instance, we have the following result.

Theorem 71 *Consider the minimization problem (45) and $\varepsilon > 0$. Suppose that*

$$n \ge 2,\ \ \varepsilon \le \frac{\mu R^2}{18}\left(\frac{\alpha-1}{\alpha+1}\right)^2 \ \textit{and}\ m = \left\lfloor \frac{1}{4}\left(\sqrt{2\frac{L/\mu-1}{n}} + 1\right)\log\left(\frac{\mu R^2}{9\varepsilon}\right)\right\rfloor + 1,$$

where $\alpha = \sqrt{\frac{2(L/\mu-1)}{n} + 1}$. In order to find $\hat{\mathbf{x}} \in \mathcal{X}$ such that $\mathbb{E} f_{\mathrm{SC}}(\hat{\mathbf{x}}) - \min_{\mathbf{x}\in\mathcal{X}} f_{\mathrm{SC}}(\mathbf{x}) < \varepsilon$, any PIFO algorithm $\mathcal{A}$ needs at least N queries to PIFO, where

$$N = \begin{cases} \Omega\left(\left(n + \sqrt{\frac{nL}{\mu}}\right)\log\left(\frac{1}{\varepsilon}\right)\right), & \textit{for } \frac{L}{\mu} \ge \frac{n}{2} + 1, \\ \Omega\left(n + \left(\frac{n}{1 + (\log(n\mu/L))_+}\right)\log\left(\frac{1}{\varepsilon}\right)\right), & \textit{for } 2 \le \frac{L}{\mu} < \frac{n}{2} + 1. \end{cases}$$

Proof Let $\Delta = \frac{\mu R^2\alpha}{\alpha+1}$. Since $\alpha > 1$, we have $\frac{\mu R^2}{2} < \Delta < \mu R^2$. Let $M = \left\lfloor \frac{\log(9\varepsilon/\Delta)}{2\log q}\right\rfloor$, then we have $\min_{\mathbf{x}\in\mathcal{X}\cap\mathcal{F}_M} f_{\mathrm{SC}}(\mathbf{x}) - \min_{\mathbf{x}\in\mathcal{X}} f_{\mathrm{SC}}(\mathbf{x}) \ge \Delta q^{2M} \ge 9\varepsilon$, where the first inequality is according to the third property of Proposition 70.

By Lemma 61, if $1 \le M < m$ and $N = \frac{(M+1)n}{4(1+c_0)}$, we have that $\min_{t\le N} \mathbb{E} f_{\mathrm{SC}}(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} f_{\mathrm{SC}}(\mathbf{x}) \ge \varepsilon$ holds. Thus, in order to find $\hat{\mathbf{x}} \in \mathcal{X}$ such that $\mathbb{E} f_{\mathrm{SC}}(\hat{\mathbf{x}}) - \min_{\mathbf{x}\in\mathcal{X}} f_{\mathrm{SC}}(\mathbf{x}) < \varepsilon$, $\mathcal{A}$ needs at least N queries.

We estimate $-\log(q)$ and N in two cases.

1. If $L/\mu \ge n/2 + 1$, then $\alpha = \sqrt{2\frac{L/\mu-1}{n} + 1} \ge \sqrt{2}$. Observe that function $h(\beta) = \frac{1}{\log\left(\frac{\beta+1}{\beta-1}\right)} - \frac{\beta}{2}$ is increasing when $\beta > 1$. Thus, we have

$$-\frac{1}{\log(q)} = \frac{1}{\log\left(\frac{\alpha+1}{\alpha-1}\right)} \ge \frac{\alpha}{2} + h(\sqrt{2}) = \frac{1}{2}\sqrt{2\frac{L/\mu-1}{n} + 1} + h(\sqrt{2})$$

$$\geq \frac{\sqrt{2}}{4}\left(\sqrt{2\frac{L/\mu-1}{n}}+1\right)+h(\sqrt{2})$$

and

$$N=\frac{(M+1)n}{4(1+c_0)}=\frac{n}{4(1+c_0)}\left(\left\lfloor\frac{\log(9\varepsilon/\Delta)}{2\log q}\right\rfloor+1\right)\geq\frac{n}{8(1+c_0)}\left(-\frac{1}{\log(q)}\right)\log\left(\frac{\Delta}{9\varepsilon}\right)$$
$$\geq\frac{n}{8(1+c_0)}\left(\frac{1}{2}\sqrt{\frac{L/\mu-1}{n}}+\frac{\sqrt{2}}{4}+h(\sqrt{2})\right)\log\left(\frac{\mu R^2}{18\varepsilon}\right)=\Omega\left(\left(n+\sqrt{\frac{nL}{\mu}}\right)\log\left(\frac{1}{\varepsilon}\right)\right).$$

2. If $2\leq L/\mu<n/2+1$, then we have

$$-\log(q)=\log\left(\frac{\alpha+1}{\alpha-1}\right)=\log\left(1+\frac{2(\alpha-1)}{\alpha^2-1}\right)=\log\left(1+\frac{\sqrt{2\frac{L/\mu-1}{n}+1}-1}{\frac{L/\mu-1}{n}}\right)$$
$$\leq\log\left(1+\frac{(\sqrt{2}-1)n}{L/\mu-1}\right)\leq\log\left(\frac{(\sqrt{2}-1/2)n}{L/\mu-1}\right)\leq\log\left(\frac{(2\sqrt{2}-1)n}{L/\mu}\right),\quad(47)$$

where the first inequality and second inequality follow from $L/\mu-1<n/2$ and the last inequality is according to $\frac{1}{x-1}\leq\frac{2}{x}$ for $x\geq 2$.

Note that $n\geq 2$, thus $\frac{n}{n-1}\leq 2\leq\frac{n}{L/\mu-1}$, and hence $n\geq L/\mu$, i.e. $\log(n\mu/L)\geq 0$.

Therefore,

$$N=\frac{(M+1)n}{4(1+c_0)}\geq\frac{n}{8(1+c_0)}\left(-\frac{1}{\log(q)}\right)\log\left(\frac{\mu R^2}{18\varepsilon}\right)=\Omega\left(\left(\frac{n}{1+\log(n\mu/L)}\right)\log\left(\frac{1}{\varepsilon}\right)\right).$$

Recalling that we assume that $q^2\geq\frac{18\varepsilon}{\mu R^2}>\frac{9\varepsilon}{\Delta}$, thus we have

$$N\geq\frac{n}{8(1+c_0)}\left(-\frac{1}{\log(q)}\right)\log\left(\frac{\Delta}{9\varepsilon}\right)\geq\frac{n}{8(1+c_0)}\left(-\frac{1}{\log(q)}\right)(-2\log(q))=\frac{n}{4(1+c_0)}.$$

Therefore, $N=\Omega\left(n+\left(\frac{n}{1+\log(n\mu/L)}\right)\log\left(\frac{1}{\varepsilon}\right)\right)$.

At last, we must ensure that $1\leq M<m$, that is $1\leq\frac{\log(9\varepsilon/\Delta)}{2\log q}<m$. Note that $\lim_{\beta\to+\infty}h(\beta)=0$, so $-1/\log(q)\leq\alpha/2$. Thus the above conditions are satisfied when

$$m=\left\lfloor\frac{\log(\mu R^2/(9\varepsilon))}{2(-\log q)}\right\rfloor+1\leq\frac{1}{4}\left(\sqrt{2\frac{L/\mu-1}{n}+1}\right)\log\left(\frac{\mu R^2}{9\varepsilon}\right)+1=\mathcal{O}\left(\sqrt{\frac{L}{n\mu}}\log\left(\frac{1}{\varepsilon}\right)\right),$$

and $\frac{\varepsilon}{\Delta}\leq\frac{1}{9}\left(\frac{\alpha-1}{\alpha+1}\right)^2$. ■

For the average smooth case, the hard instance can be directly derived from Definition 69.

Definition 72 *For fixed L, μ, R, n such that $L/\mu \geq 2$, consider $\{f_{\mathrm{SC},i}\}_{i=1}^n$, f_{SC} and Problem (45) defined in Definition 69 with L replaced by $\tilde{L} \triangleq \sqrt{\frac{n(L^2-\mu^2)}{2} - \mu^2}$.*

The following proposition ensures the hard instance is L-average smooth and gives the relationship between the smoothness parameter and the average smoothness parameter.

Proposition 73 *Consider $\{f_{\mathrm{SC},i}\}_{i=1}^n$ and f_{SC} defined in Definition 72. For $n \geq 4$ and $\kappa = \frac{L}{\mu} \geq 2$, we have that*

1. *$f_{\mathrm{SC}}(\mathbf{x})$ is μ-strongly-convex and $\{f_{\mathrm{SC},i}\}_{i=1}^n$ is L-average smooth.*
2. *$\frac{\sqrt{n}}{2} L \leq \tilde{L} \leq \sqrt{\frac{n}{2}} L$ and $\tilde{\kappa} = \frac{\tilde{L}}{\mu} \geq 2$.*

Proof

1. It is easy to check that $f_{\mathrm{SC}}(\mathbf{x})$ is μ-strongly-convex. By Proposition 59 and Lemma 40, $\{f_{\mathrm{SC},i}\}_{i=1}^n$ is $\hat{L}$-average smooth, where
$$\hat{L} = \frac{\tilde{L}-\mu}{2n}\sqrt{\frac{4}{n}\left[\left(\frac{n\tilde{L}/\mu + n}{\tilde{L}/\mu - 1}\right)^2 + n^2\right] + \left(\frac{2n}{\tilde{L}/\mu - 1}\right)^2} = \sqrt{\frac{2(\tilde{L}^2+\mu^2)}{n} + \mu^2} = L.$$

2. Clearly, $\tilde{L} = \sqrt{\frac{n(L^2-\mu^2)}{2} - \mu^2} \leq \sqrt{\frac{n}{2}} L$.
Furthermore, according to $\kappa \geq 2$ and $n \geq 4$, we have
$$\tilde{L}^2 - \frac{n}{4}L^2 = \frac{n}{4}L^2 - \frac{n}{2}\mu^2 - \mu^2 = \mu^2\left(\frac{n}{4}\kappa^2 - \frac{n}{2} - 1\right) \geq \mu^2\left(\frac{n}{2} - 1\right) \geq 0.$$
and $\tilde{\kappa} = \frac{\tilde{L}}{\mu} \geq \frac{\sqrt{n}L}{2\mu} \geq \kappa \geq 2$.

This completes the proof. ■

Recalling Theorem 71, we have the following result.

Theorem 74 *Consider the minimization problem (45) and $\varepsilon > 0$. Suppose that $\kappa = L/\mu \geq 2$, $n \geq 4$, $\varepsilon \leq \frac{\mu R^2}{18}\left(\frac{\alpha-1}{\alpha+1}\right)^2$ and $m = \left\lfloor \frac{1}{4}\left(\sqrt{2\frac{\tilde{L}/\mu - 1}{n}} + 1\right)\log\left(\frac{\mu R^2}{9\varepsilon}\right)\right\rfloor + 1$ where $\alpha = \sqrt{\frac{2(\tilde{L}/\mu-1)}{n} + 1}$, and $\tilde{L} = \sqrt{\frac{n(L^2-\mu^2)}{2} - \mu^2}$, In order to find $\hat{\mathbf{x}} \in \mathcal{X}$ such that $\mathbb{E} f_{\mathrm{SC}}(\hat{\mathbf{x}}) - \min_{\mathbf{x}\in\mathcal{X}} f_{\mathrm{SC}}(\mathbf{x}) < \varepsilon$, any PIFO algorithm $\mathcal{A}$ needs at least N queries to PIFO, where*
$$N = \begin{cases} \Omega\left(\left(n + n^{3/4}\sqrt{\kappa}\right)\log(1/\varepsilon)\right), & \text{for } \kappa = \Omega(\sqrt{n}), \\ \Omega\left(n + \left(\frac{n}{1 + (\log(\sqrt{n}/\kappa))_+}\right)\log(1/\varepsilon)\right), & \text{for } \kappa = \mathcal{O}(\sqrt{n}). \end{cases}$$

For larger ε, we can apply the following Lemma.

Lemma 75 *For any $L, \mu, n, R, \varepsilon$ such that $n \geq 2$ and $\varepsilon \leq LR^2/4$, there exist n functions $\{f_i : \mathbb{R} \to \mathbb{R}\}_{i=1}^n$ such that $f_i(x)$ is L-smooth, $\{f_i\}_{i=1}^n$ is L-average smooth and $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$ is μ-strongly-convex. In order to find $|\hat{x}| \leq R$ such that $\mathbb{E} f(\hat{x}) - \min_{|x|\leq R} f(x) < \varepsilon$, PIFO algorithm $\mathcal{A}$ needs at least $N = \Omega(n)$ queries.*

Proof Consider the following functions $\{G_{\mathrm{SC},i}\}_{1\le i\le n}$, $G_{\mathrm{SC}}:\mathbb{R}\to\mathbb{R}$, where

$$G_{\mathrm{SC},i}(x)=\frac{L}{2}x^2-nLRx,\quad \text{for } i=1,$$
$$G_{\mathrm{SC},i}(x)=\frac{L}{2}x^2,\qquad\qquad \text{for } i=2,3,\ldots,n,$$

and $G_{\mathrm{SC}}(x)=\frac{1}{n}\sum_{i=1}^n G_{\mathrm{SC},i}(x)=\frac{L}{2}x^2-LRx$. Note that $\{G_{\mathrm{SC},i}\}_{i=1}^n$ is L smooth and μ-strongly-convex for any $\mu\le L$. Observe that $x^*=\arg\min_{x\in\mathbb{R}}G_{\mathrm{SC}}(x)=R$, $G_{\mathrm{SC}}(0)-G_{\mathrm{SC}}(x^*)=\frac{LR^2}{2}$ and $|x^*|=R$. Thus $x^*=\arg\min_{|x|\le R}G_{\mathrm{SC}}(x)$.

For $i>1$, we have $\frac{dG_{\mathrm{SC},i}(x)}{dx}|_{x=0}=0$ and $\mathrm{prox}^{\gamma}_{G_{\mathrm{SC},i}}(0)=0$. Thus $x_t=0$ will hold till our first-order method $\mathcal{A}$ draws the component $G_{\mathrm{SC},1}$. That is, for $t<T=\arg\min\{t: i_t=1\}$, we have $x_t=0$.

Hence, for $t\le\frac{1}{2p_1}$, we have

$$\mathbb{E}G_{\mathrm{SC}}(x_t)-G_{\mathrm{SC}}(x^*)\ge\mathbb{E}\left[G_{\mathrm{SC}}(x_t)-G_{\mathrm{SC}}(x^*)\Big|\frac{1}{2p_1}<T\right]\mathbb{P}\left[\frac{1}{2p_1}<T\right]=\frac{LR^2}{2}\mathbb{P}\left[\frac{1}{2p_1}<T\right].$$

Note that T follows a geometric distribution with success probability $p_1\le 1/n$, and

$$\mathbb{P}\left[T>\frac{1}{2p_1}\right]=\mathbb{P}\left[T>\left\lfloor\frac{1}{2p_1}\right\rfloor\right]=(1-p_1)^{\left\lfloor\frac{1}{2p_1}\right\rfloor}\ge(1-p_1)^{\frac{1}{2p_1}}\ge(1-1/n)^{n/2}\ge\frac{1}{2},$$

where the second inequality follows from $h(z)=\frac{\log(1-z)}{2z}$ is a decreasing function.

Thus, for $t\le\frac{1}{2p_1}$, we have $\mathbb{E}G_{\mathrm{SC}}(x_t)-G_{\mathrm{SC}}(x^*)\ge\frac{LR^2}{4}\ge\varepsilon$. Thus, in order to find $|\hat{x}|\le R$ such that $\mathbb{E}G_{\mathrm{SC}}(\hat{x})-G_{\mathrm{SC}}(x^*)<\varepsilon$, $\mathcal{A}$ needs at least $\frac{1}{2p_1}\ge n/2=\Omega(n)$ queries. ∎

Proof [Proof of Theorem 63] It remains to explain that the lower bound in Lemma 75 is the same as the lower bound in Theorem 71 for $\varepsilon>\frac{\mu R^2}{18}\left(\frac{\alpha-1}{\alpha+1}\right)^2$. Suppose that $\frac{\varepsilon}{\mu R^2}>\frac{1}{18}\left(\frac{\alpha-1}{\alpha+1}\right)^2$, $\alpha=\sqrt{2\frac{\kappa-1}{n}+1}$ and $\kappa=\frac{L}{\mu}$.

1. If $\kappa\ge n/2+1$, then we have $\alpha\ge\sqrt{2}$ and

$$\begin{aligned}\left(n+\sqrt{\kappa n}\right)\log\left(\frac{\mu R^2}{18\varepsilon}\right)&\le 2\left(n+\sqrt{\kappa n}\right)\log\left(\frac{\alpha+1}{\alpha-1}\right)\\ &\le\frac{4\left(n+\sqrt{\kappa n}\right)}{\alpha-1}=\mathcal{O}(n)+\frac{4\sqrt{\kappa n}}{(1-\sqrt{2}/2)\alpha}\\ &\le\mathcal{O}(n)+\frac{4}{\sqrt{2}-1}\frac{\sqrt{\kappa n}}{\sqrt{\kappa/n}}=\mathcal{O}(n),\end{aligned}$$

where the second inequality follows from $\log(1+x)\le x$ and the last inequality is according to $\alpha\ge\sqrt{2\kappa/n}$. Then we have $\Omega(n)=\Omega\left(\left(n+\sqrt{\kappa n}\right)\log\left(\frac{1}{\varepsilon}\right)\right)$.

2. If $2 \le L/\mu < n/2+1$, then we have

$$\left(\frac{n}{1+(\log(n\mu/L))_+}\right)\log\left(\frac{\mu R^2}{18\varepsilon}\right) \le \left(\frac{n}{1+(\log(n\mu/L))_+}\right)\left(2\log\left(\frac{\alpha+1}{\alpha-1}\right)\right)$$
$$\le \left(\frac{n}{1+(\log(n\mu/L))_+}\right)\left(2\log\left(\frac{(2\sqrt{2}-1)n}{L/\mu}\right)\right) = \mathcal{O}(n),$$

where the second inequality is by applying (47). As a consequence, we have $\Omega(n) = \Omega\left(\left(\frac{n}{1+(\log(n\mu/L))_+}\right)\log\left(\frac{1}{\varepsilon}\right)+n\right)$.

This completes the proof. ∎

The proof of Theorem 66 is similar to that of Theorem 63.

E.5 Construction for the Convex Case

The analysis of lower bound complexity for the convex case depends on the following construction.

Definition 76 *For fixed L, R, n, we define $f_{\mathrm{C},i} : \mathbb{R}^m \to \mathbb{R}$ as follows*

$$f_{\mathrm{C},i}(\mathbf{x}) = \lambda r_i\left(\mathbf{x}/\beta; m, 0, 1, \mathbf{c}^{\mathrm{C}}\right), \text{ for } 1 \le i \le n,$$

where $\mathbf{c}^{\mathrm{C}} = (0,0,1)$, $\lambda = \frac{3LR^2}{2n(m+1)^3}$ and $\beta = \frac{\sqrt{3}R}{(m+1)^{3/2}}$. Consider the minimization problem

$$\min_{\mathbf{x}\in\mathcal{X}} f_{\mathrm{C}}(\mathbf{x}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_{\mathrm{C},i}(\mathbf{x}). \tag{48}$$

where $\mathcal{X} = \{\mathbf{x}\in\mathbb{R}^m : \|\mathbf{x}\|_2 \le R\}$.

Then we have the following proposition.

Proposition 77 *For any $n \ge 2$, $m \ge 2$, the following properties hold:*

1. *$f_{\mathrm{C},i}$ is L-smooth and convex. Thus, f_{C} is convex.*
2. *The minimizer of the function f_{C} is*

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^m} f_{\mathrm{C}}(\mathbf{x}) = \frac{2\xi}{L}\left(m, m-1, \dots, 1\right)^\top,$$

where $\xi = \frac{\sqrt{3}}{2}\frac{RL}{(m+1)^{3/2}}$. Moreover, $f_{\mathrm{C}}(\mathbf{x}^*) = -\frac{m\xi^2}{nL}$ and $\|\mathbf{x}^*\|_2 \le R$.

3. *For $1 \le k \le m$, we have*

$$\min_{\mathbf{x}\in\mathcal{X}\cap\mathcal{F}_k} f_{\mathrm{C}}(\mathbf{x}) - \min_{\mathbf{x}\in\mathcal{X}} f_{\mathrm{C}}(\mathbf{x}) = \frac{\xi^2}{nL}(m-k).$$

Proof

1. Just recall Proposition 59 and Lemma 40.

2. It is easy to check $f_{\mathrm{C}}(\mathbf{x}) = \frac{L}{4n}\|\mathbf{B}(m,1)\mathbf{x}\|_2^2 - \frac{\xi}{n}\langle \mathbf{e}_1, \mathbf{x}\rangle$, where $\xi = \frac{\sqrt{3}}{2}\frac{BL}{(m+1)^{3/2}n}$. Let $\nabla f_{\mathrm{C}}(\mathbf{x}) = \mathbf{0}$, that is $\frac{L}{2n}\mathbf{A}(m,0,1)\mathbf{x} = \frac{\xi}{n}\mathbf{e}_1$. One can check that the solution is $\mathbf{x}^* = \frac{2\xi}{L}(m, m-1, \dots, 1)^\top$, and $f_{\mathrm{C}}(\mathbf{x}^*) = -\frac{m\xi^2}{nL}$. Moreover, we have
$$\|\mathbf{x}^*\|_2^2 = \frac{4\xi^2}{L^2}\frac{m(m+1)(2m+1)}{6} \le \frac{4\xi^2}{3L^2}(m+1)^3 = R^2.$$

3. The second property implies $\min_{\mathbf{x}\in\mathcal{X}} f_{\mathrm{C}}(\mathbf{x}) = -\frac{m\xi^2}{nL}$. Following some similar calculations to the above proof, we have $\arg\min_{\mathbf{x}\in\mathcal{X}\cap\mathcal{F}_k} f_{\mathrm{C}}(\mathbf{x}) = \frac{2\xi}{L}(k, k-1, \dots, 1, 0, \dots, 0)^\top$, and $\min_{\mathbf{x}\in\mathcal{X}\cap\mathcal{F}_k} f_{\mathrm{C}}(\mathbf{x}) = -\frac{k\xi^2}{nL}$. Thus $\min_{\mathbf{x}\in\mathcal{X}\cap\mathcal{F}_k} f_{\mathrm{C}}(\mathbf{x}) - \min_{\mathbf{x}\in\mathcal{X}} f_{\mathrm{C}}(\mathbf{x}) = \frac{\xi^2}{nL}(m-k)$.

This completes the proof. ■

Next, we show the lower bound for functions $f_{\mathrm{C},i}$ defined above.

Theorem 78 *Consider the minimization problem (48) and $\varepsilon > 0$. Suppose that*
$$n \ge 2,\ \ \varepsilon \le \frac{R^2 L}{384n} \ \ \text{and} \ \ m = \left\lfloor \sqrt{\frac{R^2L}{24n\varepsilon}} \right\rfloor - 1.$$
In order to find $\hat{\mathbf{x}} \in \mathcal{X}$ such that $\mathbb{E} f_{\mathrm{C}}(\hat{\mathbf{x}}) - \min_{\mathbf{x}\in\mathcal{X}} f_{\mathrm{C}}(\mathbf{x}) < \varepsilon$, any PIFO algorithm $\mathcal{A}$ needs at least N queries to PIFO, where $N = \Omega\left(n + R\sqrt{nL/\varepsilon}\right)$.

Proof Since $\varepsilon \le \frac{R^2L}{384n}$, we have $m \ge 3$. Let $\xi = \frac{\sqrt{3}}{2}\frac{RL}{(m+1)^{3/2}}$. For $M = \left\lfloor \frac{m-1}{2} \right\rfloor \ge 1$, we have $m - M \ge (m+1)/2$, and
$$\min_{\mathbf{x}\in\mathcal{X}\cap\mathcal{F}_M} f_{\mathrm{C}}(\mathbf{x}) - \min_{\mathbf{x}\in\mathcal{X}} f_{\mathrm{C}}(\mathbf{x}) = \frac{\xi^2}{nL}(m-M) = \frac{3R^2L}{4n}\frac{m-M}{(m+1)^3} \ge \frac{3R^2L}{8n}\frac{1}{(m+1)^2} \ge 9\varepsilon,$$
where the first equation is according to the 3rd property in Proposition 77 and the last inequality follows from $m + 1 \le R\sqrt{L/(24n\varepsilon)}$.

Similar to the proof of Theorem 71, we have $\min_{t\le N} \mathbb{E} f_{\mathrm{C}}(\mathbf{x}_t) - \min_{\mathbf{x}\in\mathcal{X}} f_{\mathrm{C}}(\mathbf{x}) \ge \varepsilon$ by Lemma 61. In other words, in order to find $\hat{\mathbf{x}} \in \mathcal{X}$ such that $\mathbb{E} f_{\mathrm{C}}(\hat{\mathbf{x}}) - \min_{\mathbf{x}\in\mathcal{X}} f_{\mathrm{C}}(\mathbf{x}) < \varepsilon$, $\mathcal{A}$ needs at least N queries.

At last, observe that
$$N = \frac{(M+1)n}{4(1+c_0)} = \frac{n}{4(1+c_0)}\left\lfloor \frac{m+1}{2} \right\rfloor \ge \frac{n(m-1)}{8} \ge \frac{n}{8}\left(\sqrt{\frac{R^2L}{24n\varepsilon}} - 2\right) = \Omega\left(n + R\sqrt{\frac{nL}{\varepsilon}}\right),$$
where we have recalled $\varepsilon \le \frac{B^2L}{384n}$ in last equation. ■

The hard instance for the average smooth case can be derived from Definition 76.

Definition 79 *For fixed L, R, n, consider $\{f_{\mathrm{C},i}\}_{i=1}^n$ and f_{C} defined in Definition 76 with L replaced by $\sqrt{\frac{n}{2}}L$.*

It follows from Proposition 59 and Lemma 40 that f_{C} is convex and $\{f_{\mathrm{C},i}\}_{i=1}^n$ is L-average smooth. By Theorem 78, we have the following conclusion.

Theorem 80 *Consider the minimization problem (48) and $\varepsilon > 0$. Suppose that*

$$n \geq 2,\ \varepsilon \leq \frac{\sqrt{2}}{768}\frac{R^2L}{\sqrt{n}} \quad \text{and} \quad m = \left\lfloor \frac{\sqrt[4]{18}}{12}Rn^{-1/4}\sqrt{\frac{L}{\varepsilon}} \right\rfloor - 1.$$

In order to find $\hat{\mathbf{x}} \in \mathcal{X}$ such that $\mathbb{E}f_{\mathrm{C}}(\hat{\mathbf{x}}) - \min_{\mathbf{x}\in\mathcal{X}} f_{\mathrm{C}}(\mathbf{x}) < \varepsilon$, any PIFO algorithm $\mathcal{A}$ needs at least N queries to PIFO, where $N = \Omega\left(n + Rn^{3/4}\sqrt{\frac{L}{\varepsilon}}\right)$.

Proof [Proof of Theorem 64] To derive Theorem 64, it remains to consider the case $\varepsilon > \frac{\sqrt{2}R^2L}{768\sqrt{n}}$. By Lemma 75, there exist n functions $\{f_i : \mathbb{R} \to \mathbb{R}\}$ such that $f_i(x)$ is L-smooth and $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$ is convex. In order to find $|\hat{x}| \leq R$ such that $\mathbb{E}f(\hat{x}) - \min_{|x|\leq R} f(x) < \varepsilon$, PIFO algorithm $\mathcal{A}$ needs at least $N = \Omega(n)$ queries. Since $\varepsilon > \frac{\sqrt{2}R^2L}{768\sqrt{n}}$, $\Omega(n) = \Omega\left(n + R\sqrt{\frac{nL}{\varepsilon}}\right)$. This completes the proof. ■

The proof of Theorem 67 is similar to that of Theorem 64.

E.6 Construction for the Nonconvex Case

The analysis of lower bound complexity for the nonconvex case depends on the following construction.

Definition 81 *For fixed L, μ, Δ, n, we define $f_{\mathrm{NC},i} : \mathbb{R}^{m+1} \to \mathbb{R}$ as follows*

$$f_{\mathrm{NC},i}(\mathbf{x}) = \lambda r_i\left(\mathbf{x}/\beta; m+1, \sqrt[4]{\alpha}, 0, \mathbf{c}^{\mathrm{NC}}\right), \ \text{for } 1 \leq i \leq n,$$

where

$$\alpha = \min\left\{1, \frac{(\sqrt{3}+1)n\mu}{30L}, \frac{n}{180}\right\},\ \mathbf{c}^{\mathrm{NC}} = \left(0, \alpha, \sqrt{\alpha}\right),$$
$$m = \left\lfloor \frac{\Delta L\sqrt{\alpha}}{40824n\varepsilon^2} \right\rfloor,\ \lambda = \frac{3888n\varepsilon^2}{L\alpha^{3/2}} \quad \text{and } \beta = \sqrt{3\lambda n/L}.$$

Consider the minimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^{m+1}} f_{\mathrm{NC}}(\mathbf{x}) \triangleq \frac{1}{n}\sum_{i=1}^n f_{\mathrm{NC},i}(\mathbf{x}). \tag{49}$$

Then we have the following proposition.

Proposition 82 *For any $n \geq 2$ and $\varepsilon^2 \leq \frac{\Delta L\alpha}{81648n}$, the following properties hold:*

1. *$f_{\mathrm{NC},i}$ is L-smooth and $(-\mu)$-weakly-convex. Thus, f_{NC} is $(-\mu)$-weakly-convex.*

2. *$f_{\mathrm{NC}}(\mathbf{0}) - \min_{\mathbf{x}\in\mathbb{R}^{m+1}} f_{\mathrm{NC}}(\mathbf{x}) \leq \Delta$.*

3. $m \geq 2$ and for $M = m-1$, $\min_{\mathbf{x}\in\mathcal{F}_M} \|\nabla f_{\mathrm{NC}}(\mathbf{x})\|_2 \geq 9\varepsilon$.

Proof

1. By Proposition 59 and Lemma 40, $f_{\mathrm{NC},i}$ is $(-l_1)$-weakly convex and l_2-smooth where
$$l_1 = \frac{45(\sqrt{3}-1)\alpha\lambda}{\beta^2} = \frac{45(\sqrt{3}-1)L}{3n}\alpha \leq \frac{45(\sqrt{3}-1)L}{3n}\frac{(\sqrt{3}+1)n\mu}{30L} = \mu,$$
$$l_2 = \frac{(2n+180\alpha)\lambda}{\beta^2} = \frac{L}{3n}(2n+180\alpha) \leq L.$$
Thus each f_i is L-smooth and $(-\mu)$-weakly convex.

2. By Proposition 43, we know that
$$f_{\mathrm{NC}}(\mathbf{0}) - \min_{\mathbf{x}\in\mathbb{R}^{m+1}} f_{\mathrm{NC}}(\mathbf{x}) \leq \lambda(\sqrt{\alpha}/2 + 10\alpha m) = \frac{1944n\varepsilon^2}{L\alpha} + \frac{38880n\varepsilon^2}{L\sqrt{\alpha}}m$$
$$\leq \frac{1944}{40824}\Delta + \frac{38880}{40824}\Delta = \Delta.$$

3. Since $\alpha \leq 1$, we have $\frac{\Delta L^2\sqrt{\alpha}}{40824n\varepsilon^2} \geq \frac{\Delta L^2 \alpha}{40824n\varepsilon^2}$ and consequently $m \geq 2$. By Proposition 43, we know that
$$\min_{\mathbf{x}\in\mathcal{F}_M} \|\nabla f_{\mathrm{NC}}(\mathbf{x})\|_2 \geq \frac{\alpha^{3/4}\lambda}{4\beta} = \frac{\alpha^{3/4}\lambda}{4\sqrt{3\lambda n/L}} = \sqrt{\frac{\lambda L}{3n}}\frac{\alpha^{3/4}}{4} = 9\varepsilon.$$

This completes the proof. ∎

Next we prove Theorem 65.

Proof [Proof of Theorem 65] By Lemma 61 and the third property of Proposition 82, in order to find $\hat{\mathbf{x}} \in \mathbb{R}^{m+1}$ such that $\mathbb{E}\|\nabla f_{\mathrm{NC}}(\hat{\mathbf{x}})\|_2 < \varepsilon$, PIFO algorithm $\mathcal{A}$ needs at least N queries, where $N = \frac{nm}{4(1+c_0)} = \Omega\left(\frac{\Delta L\sqrt{\alpha}}{\varepsilon^2}\right)$. Since $\varepsilon^2 \leq \frac{\Delta L\alpha}{81648n}$ and $\alpha \leq 1$, we have $\Omega\left(\frac{\Delta L\sqrt{\alpha}}{\varepsilon^2}\right) = \Omega\left(n + \frac{\Delta L\sqrt{\alpha}}{\varepsilon^2}\right)$. ∎

The analysis of lower bound complexity for the non-convex case under the average smooth assumption depends on the following construction.

Definition 83 *For fixed L, μ, Δ, n, we define $\bar{f}_{\mathrm{NC},i} : \mathbb{R}^{m+1} \to \mathbb{R}$ as follows*
$$\bar{f}_{\mathrm{NC},i}(\mathbf{x}) = \lambda r_i\left(\mathbf{x}/\beta; m+1, \sqrt[4]{\alpha}, 0, \bar{\mathbf{c}}^{\mathrm{NC}}\right), \text{ for } 1 \leq i \leq n,$$
where
$$\alpha = \min\left\{1, \frac{8(\sqrt{3}+1)\sqrt{n}\mu}{45L}, \sqrt{\frac{n}{270}}\right\},\ \bar{\mathbf{c}}^{\mathrm{NC}} = \left(0, \alpha, \sqrt{\alpha}\right),$$
$$m = \left\lfloor \frac{\Delta L\sqrt{\alpha}}{217728\sqrt{n}\varepsilon^2} \right\rfloor,\ \lambda = \frac{20736\sqrt{n}\varepsilon^2}{L\alpha^{3/2}} \text{ and } \beta = 4\sqrt{\lambda\sqrt{n}/L}.$$
Consider the minimization problem
$$\min_{\mathbf{x}\in\mathbb{R}^{m+1}} \bar{f}_{\mathrm{NC}}(\mathbf{x}) \triangleq \frac{1}{n}\sum_{i=1}^{n} \bar{f}_{\mathrm{NC},i}(\mathbf{x}). \tag{50}$$

Then we have the following proposition.

Proposition 84 *For any $n \geq 2$ and $\varepsilon^2 \leq \frac{\Delta L \alpha}{435456\sqrt{n}}$, the following properties hold:*

1. *$\bar{f}_{\mathrm{NC},i}$ is $(-\mu)$-weakly-convex and $\{\bar{f}_{\mathrm{NC},i}\}_{i=1}^n$ is L-average smooth. Thus, f_{NC} is $(-\mu)$-weakly-convex.*

2. *$f_{\mathrm{NC}}(\mathbf{0}) - \min_{\mathbf{x}\in\mathbb{R}^{m+1}} f_{\mathrm{NC}}(\mathbf{x}) \leq \Delta$.*

3. *$m \geq 2$ and for $M = m-1$, $\min_{\mathbf{x}\in\mathcal{F}_M} \|\nabla f_{\mathrm{NC}}(\mathbf{x})\|_2 \geq 9\varepsilon$.*

Proof

1. By Proposition 59 and Lemma 40, $\bar{f}_{\mathrm{NC},i}$ is $(-l_1)$-weakly convex and $\{\bar{f}_{\mathrm{NC},i}\}_{i=1}^n$ is l_2-average smooth where

$$l_1 = \frac{45(\sqrt{3}-1)\alpha\lambda}{\beta^2} = \frac{45(\sqrt{3}-1)L'}{16\sqrt{n}}\alpha \leq \frac{45(\sqrt{3}-1)L'}{16\sqrt{n}}\frac{8(\sqrt{3}+1)\sqrt{n}\mu}{45L'} = \mu,$$
$$l_2 = 4\sqrt{n+4050\alpha^2}\frac{\lambda}{\beta^2} = \frac{L'}{4\sqrt{n}}\sqrt{n+4050\alpha^2} \leq L'.$$

2. By Proposition 43, we know that

$$f_{\mathrm{NC}}(\mathbf{0}) - \min_{\mathbf{x}\in\mathbb{R}^{m+1}} f_{\mathrm{NC}}(\mathbf{x}) \leq \lambda(\sqrt{\alpha}/2 + 10\alpha m) = \frac{10368\sqrt{n}\varepsilon^2}{L'\alpha} + \frac{207360\sqrt{n}\varepsilon^2}{L'\sqrt{\alpha}}m$$
$$\leq \frac{10368}{217728}\Delta + \frac{207360}{217728}\Delta = \Delta.$$

3. Since $\alpha \leq 1$, we have $\frac{\Delta L'\sqrt{\alpha}}{217728\sqrt{n}\varepsilon^2} \geq \frac{\Delta L'\alpha}{217728\sqrt{n}\varepsilon^2}$ and consequently $m \geq 2$. By Proposition 43, we know that

$$\min_{\mathbf{x}\in\mathcal{F}_M} \|\nabla f_{\mathrm{NC}}(\mathbf{x})\|_2 \geq \frac{\alpha^{3/4}\lambda}{4\beta} = \frac{\alpha^{3/4}\lambda}{4\sqrt{16\lambda\sqrt{n}/L'}} = \frac{\sqrt{\lambda L'}}{\sqrt[4]{n}}\frac{\alpha^{3/4}}{16} = 9\varepsilon.$$

This completes the proof. ■

Next we prove Theorem 68.

Proof [Proof of Theorem 68] By Lemma 61 and the third property of Proposition 84, in order to find $\hat{\mathbf{x}} \in \mathbb{R}^{m+1}$ such that $\mathbb{E}\|\nabla f_{\mathrm{NC}}(\hat{\mathbf{x}})\|_2 < \varepsilon$, PIFO algorithm $\mathcal{A}$ needs at least N queries , where $N = \frac{nm}{4(1+c_0)} = \Omega\left(\frac{\Delta L\sqrt{n\alpha}}{\varepsilon^2}\right)$. Since $\varepsilon^2 \leq \frac{\Delta L\alpha}{435456\sqrt{n}}$ and $\alpha \leq 1$, we have $\Omega\left(\frac{\Delta L\sqrt{n\alpha}}{\varepsilon^2}\right) = \Omega\left(n + \frac{\Delta L\sqrt{n\alpha}}{\varepsilon^2}\right)$. ■

E.7 Proofs of Proposition 59 and Lemma 37

We use $\|\mathbf{A}\|$ to denote the spectral radius of $\mathbf{A}$. Recall that $\mathbf{b}_{l-1}^\top$ is the l-th row of $\mathbf{B}$, $G(\mathbf{x}) = \sum_{i=1}^{m-1} \Gamma(x_i)$ and

$$\mathcal{L}_i = \{l : 0 \le l \le m, l \equiv i - 1(\text{mod} n)\}, i = 1, 2, \ldots, n.$$

For simplicity, we omit the parameters of $\mathbf{B}$, $\mathbf{b}_l$ and r_i.

For $1 \le i \le n$, let $\mathbf{B}_i$ be the submatrix whose rows are $\left\{\mathbf{b}_l^\top\right\}_{l \in \mathcal{L}_i}$. Then r_i can be written as

$$r_i(\mathbf{x}) = \frac{n}{2} \|\mathbf{B}_i \mathbf{x}\|_2^2 + \frac{c_1}{2} \|\mathbf{x}\|_2^2 + c_2 G(\mathbf{x}) - c_3 n \langle \mathbf{e}_1, \mathbf{x} \rangle \mathbb{1}_{\{i=1\}}.$$

Proof [Proof of Proposition 59]

1. For the convex case,

$$r_i(\mathbf{x}) = \frac{n}{2} \|\mathbf{B}_i \mathbf{x}\|_2^2 + \frac{c_1}{2} \|\mathbf{x}\|_2^2 - c_3 n \langle \mathbf{e}_1, \mathbf{x} \rangle \mathbb{1}_{\{i=1\}}.$$

Obviously, r_i is c_1-strongly convex. Note that

$$\begin{aligned}\left\langle \mathbf{u}, \mathbf{B}_i^\top \mathbf{B}_i \mathbf{u} \right\rangle &= \|\mathbf{B}_i \mathbf{u}\|_2^2 = \sum_{l \in \mathcal{L}_i} (\mathbf{b}_l^\top \mathbf{u})^2 \\ &= \sum_{l \in \mathcal{L}_i \setminus \{0, m\}} (u_l - u_{l+1})^2 + \omega^2 u_1^2 \mathbb{1}_{\{0 \in \mathcal{L}_i\}} + \zeta^2 u_m^2 \mathbb{1}_{\{m \in \mathcal{L}_i\}} \le 2 \|\mathbf{u}\|_2^2,\end{aligned}$$

where the last inequality is according to $(x+y)^2 \le 2(x^2 + y^2)$, and $|l_1 - l_2| \ge n \ge 2$ for $l_1, l_2 \in \mathcal{L}_i$. Hence, $\left\|\mathbf{B}_i^\top \mathbf{B}_i\right\| \le 2$, and

$$\left\|\nabla^2 r_i(\mathbf{x})\right\| = \left\|n \mathbf{B}_i^\top \mathbf{B}_i + c_1 \mathbf{I}\right\| \le 2n + c_1.$$

Next, observe that

$$\|\nabla r_i(\mathbf{x}_1) - \nabla r_i(\mathbf{x}_2)\|_2^2 = \left\|(n \mathbf{B}_i^\top \mathbf{B}_i + c_1 \mathbf{I})(\mathbf{x}_1 - \mathbf{x}_2)\right\|_2^2$$

Let $\mathbf{u} = \mathbf{x}_1 - \mathbf{x}_2$. Note that

$$\mathbf{b}_l \mathbf{b}_l^\top \mathbf{u} = \begin{cases} (u_l - u_{l+1})(\mathbf{e}_l - \mathbf{e}_{l+1}), & 0 < l < m, \\ \omega^2 u_1 \mathbf{e}_1, & l = 0, \\ \zeta^2 u_m \mathbf{e}_m, & l = m. \end{cases}$$

Thus,

$$\left\|(n \mathbf{B}_i^\top \mathbf{B}_i + c_1 \mathbf{I}) \mathbf{u}\right\|_2^2$$

$$
\begin{aligned}
&= \left\| n \sum_{l \in \mathcal{L}_i \setminus \{0,m\}} (u_l - u_{l+1})(\mathbf{e}_l - \mathbf{e}_{l+1}) + n\omega^2 u_1^2 \mathbb{1}_{\{0 \in \mathcal{L}_i\}} + n\zeta^2 u_m^2 \mathbb{1}_{\{m \in \mathcal{L}_i\}} + c_1 \mathbf{u} \right\|_2^2 \\
&= \sum_{l \in \mathcal{L}_i \setminus \{0,m\}} \left[(n(u_l - u_{l+1}) + c_1 u_l)^2 + (-n(u_l - u_{l+1}) + c_1 u_{l+1})^2 \right] \\
&\quad + (n\omega^2 + c_1)^2 u_1^2 \mathbb{1}_{\{0 \in \mathcal{L}_i\}} + (n\zeta^2 + c_1)^2 u_m^2 \mathbb{1}_{\{m \in \mathcal{L}_i\}} + \sum_{\substack{l-1, l \notin \mathcal{L}_i \\ l \neq 0, m}} c_1^2 u_l^2 \\
&\leq 2\left[(n + c_1)^2 + n^2 \right] \left[\sum_{l \in \mathcal{L}_i \setminus \{0,m\}} (u_l^2 + u_{l+1}^2) + u_1^2 \mathbb{1}_{\{0 \in \mathcal{L}_i\}} + u_m^2 \mathbb{1}_{\{m \in \mathcal{L}_i\}} \right] + c_1^2 \left\| \mathbf{u} \right\|_2^2,
\end{aligned}
$$

where we have used $(2n + c_1)^2 \leq 2\left[(n + c_1)^2 + n^2\right]$.

Therefore, we have

$$
\begin{aligned}
&\frac{1}{n} \sum_{i=1}^{n} \left\| \nabla r_i(\mathbf{x}_1) - \nabla r_i(\mathbf{x}_2) \right\|_2^2 \\
\leq& \frac{1}{n} \sum_{l=0}^{m} 4\left[(n + c_1)^2 + n^2 \right] u_l^2 + c_1^2 \left\| \mathbf{u} \right\|_2^2 \\
\leq& \frac{4}{n} \left[\left[(n + c_1)^2 + n^2 \right] \right] \left\| \mathbf{u} \right\|_2^2 + c_1^2 \left\| \mathbf{u} \right\|_2^2,
\end{aligned}
$$

In summary, we get that $\{r_i\}_{i=1}^n$ is L'-average smooth, where

$$
L' = \sqrt{\frac{4}{n} \left[(n + c_1)^2 + n^2 \right] + c_1^2}.
$$

2. The results of the non-convex case follow from the above proof, Proposition 43 and the inequality $(a + b)^2 \leq 2(a^2 + b^2)$.

This completes the proof. ■

Proof [Proof of Lemma 37]

1. For the convex case,

$$
r_j(\mathbf{x}) = \frac{n}{2} \left\| \mathbf{B}_j \mathbf{x} \right\|_2^2 + \frac{c_1}{2} \left\| \mathbf{x} \right\|_2^2 - c_3 n \left\langle \mathbf{e}_1, \mathbf{x} \right\rangle \mathbb{1}_{\{j=1\}}.
$$

Recall that

$$
\mathbf{b}_l \mathbf{b}_l^\top \mathbf{x} = \begin{cases} (x_l - x_{l+1})(\mathbf{e}_l - \mathbf{e}_{l+1}), & 0 < l < m, \\ \omega^2 x_1 \mathbf{e}_1, & l = 0, \\ \zeta^2 x_m \mathbf{e}_m, & l = m. \end{cases}
$$

For $\mathbf{x} \in \mathcal{F}_0$, we have $\mathbf{x} = \mathbf{0}$, and

$$
\nabla r_1(\mathbf{x}) = c_3 n \mathbf{e}_1 \in \mathcal{F}_1,
$$

$$\nabla r_j(\mathbf{x}) = \mathbf{0} \ (j \geq 2).$$

For $\mathbf{x} \in \mathcal{F}_k$ $(1 \leq k < m)$, we have

$$\mathbf{b}_l \mathbf{b}_l^\top \mathbf{x} \in \begin{cases} \mathcal{F}_k, & l \neq k, \\ \mathcal{F}_{k+1}, & l = k. \end{cases}$$

Moreover, we suppose $k \in \mathcal{L}_i$. Since

$$\begin{aligned} \nabla r_j(\mathbf{x}) &= n\mathbf{B}_j^\top \mathbf{B}_j \mathbf{x} + c_1 \mathbf{x} - c_3 n \mathbf{e}_1 \mathbb{1}_{\{j=1\}} \\ &= n \sum_{l \in \mathcal{L}_j} \mathbf{b}_l \mathbf{b}_l^\top \mathbf{x} + c_1 \mathbf{x} - c_3 n \mathbf{e}_1 \mathbb{1}_{\{j=1\}}, \end{aligned}$$

it follows that $\nabla r_i(\mathbf{x}) \in \mathcal{F}_{k+1}$ and $\nabla r_j(\mathbf{x}) \in \mathcal{F}_k$ $(j \neq i)$.

Now, we turn to consider $\mathbf{u} = \mathrm{prox}_{r_j}^{\gamma}(\mathbf{x})$. We have

$$\left(n\mathbf{B}_j^\top \mathbf{B}_j + \left(c_1 + \frac{1}{\gamma} \right) \mathbf{I} \right) \mathbf{u} = c_3 n \mathbf{e}_1 \mathbb{1}_{\{j=1\}} + \frac{1}{\gamma} \mathbf{x},$$

i.e.,

$$\mathbf{u} = d_1 (\mathbf{I} + d_2 \mathbf{B}_j^\top \mathbf{B}_j)^{-1} \mathbf{y},$$

where $d_1 = \frac{1}{c_1 + 1/\gamma}$, $d_2 = \frac{n}{c_1 + 1/\gamma}$, and $\mathbf{y} = c_3 n \mathbf{e}_1 \mathbb{1}_{\{j=1\}} + \frac{1}{\gamma}\mathbf{x}$.
Note that

$$(\mathbf{I} + d_2 \mathbf{B}_j^\top \mathbf{B}_j)^{-1} = \mathbf{I} - \mathbf{B}_j^\top \left(\frac{1}{d_2} \mathbf{I} + \mathbf{B}_j \mathbf{B}_j^\top \right)^{-1} \mathbf{B}_j.$$

If $k = 0$ and $j > 1$, we have $\mathbf{y} = \mathbf{0}$ and $\mathbf{u} = \mathbf{0}$.
If $k = 0$ and $j = 1$, we have $\mathbf{y} = c_3 n \mathbf{e}_1$. Since $\omega = 0$, $\mathbf{B}_1 \mathbf{e}_1 = \mathbf{0}$, so $\mathbf{u} = c_1 \mathbf{y} \in \mathcal{F}_1$.

For $k \geq 1$, we know that $\mathbf{y} \in \mathcal{F}_k$. And observe that if $|l - l'| \geq 2$, then $\mathbf{b}_l^\top \mathbf{b}_{l'} = 0$, and consequently $\mathbf{B}_j \mathbf{B}_j^\top$ is a diagonal matrix, so we can assume that $\frac{1}{d_2}\mathbf{I} + \mathbf{B}_j \mathbf{B}_j^\top = \mathrm{diag}(\beta_{j,1}, \ldots, \beta_{j,|\mathcal{L}_j|})$. Therefore,

$$\mathbf{u} = d_1 \mathbf{y} - d_1 \sum_{s=1}^{|\mathcal{L}_j|} \beta_{j,s} \mathbf{b}_{l_{j,s}} \mathbf{b}_{l_{j,s}}^\top \mathbf{y},$$

where we assume that $\mathcal{L}_j = \{l_{j,1}, \ldots, l_{j,|\mathcal{L}_j|}\}$.
Thus, we have $\mathrm{prox}_{r_i}^{\gamma}(\mathbf{x}) \in \mathcal{F}_{k+1}$ for $k \in \mathcal{L}_i$ and $\mathrm{prox}_{r_j}^{\gamma}(\mathbf{x}) \in \mathcal{F}_k$ $(j \neq i)$.

2. For the non-convex case,

$$r_j(\mathbf{x}) = \frac{n}{2} \|\mathbf{B}_j \mathbf{x}\|_2^2 + c_2 G(\mathbf{x}) - c_3 n \langle \mathbf{e}_1, \mathbf{x} \rangle \mathbb{1}_{\{j=1\}}.$$

Let $\Gamma'(x)$ be the derivative of $\Gamma(x)$. First note that $\Gamma'(0) = 0$, so if $\mathbf{x} \in \mathcal{F}_k$, then

$$\nabla G(\mathbf{x}) = \left(\Gamma'(x_1), \Gamma'(x_2), \ldots, \Gamma'(x_{m-1}), 0\right)^\top \in \mathcal{F}_k.$$

For $\mathbf{x} \in \mathcal{F}_0$, we have $\mathbf{x} = \mathbf{0}$, and

$$\begin{aligned} \nabla r_1(\mathbf{x}) &= c_3 n \mathbf{e}_1 \in \mathcal{F}_1, \\ \nabla r_j(\mathbf{x}) &= \mathbf{0} \ (j \geq 2). \end{aligned}$$

For $\mathbf{x} \in \mathcal{F}_k$ $(1 \leq k < m)$, recall that

$$\mathbf{b}_l \mathbf{b}_l^\top \mathbf{x} = \begin{cases} (x_l - x_{l+1})(\mathbf{e}_l - \mathbf{e}_{l+1}), & 0 < l < m, \\ \omega^2 x_1 \mathbf{e}_1, & l = 0, \\ \zeta^2 x_m \mathbf{e}_m, & l = m. \end{cases}$$

Suppose $k \in \mathcal{L}_i$. Since

$$\begin{aligned} \nabla r_j(\mathbf{x}) &= n \mathbf{B}_j^\top \mathbf{B}_j \mathbf{x} + c_2 \nabla G(\mathbf{x}) - c_3 n \mathbf{e}_1 \mathbb{1}_{\{j=1\}} \\ &= n \sum_{l \in \mathcal{L}_j} \mathbf{b}_l \mathbf{b}_l^\top \mathbf{x} + c_2 \nabla G(\mathbf{x}) - c_3 n \mathbf{e}_1 \mathbb{1}_{\{j=1\}}, \end{aligned}$$

it follows that $\nabla r_i(\mathbf{x}) \in \mathcal{F}_{k+1}$ and $\nabla r_j(\mathbf{x}) \in \mathcal{F}_k$ $(j \neq i)$.

Now, we turn to consider $\mathbf{u} = \operatorname{prox}_{r_j}^{\gamma}(\mathbf{x})$. We have

$$\nabla r_j(\mathbf{u}) + \frac{1}{\gamma}(\mathbf{u} - \mathbf{x}) = \mathbf{0},$$

that is

$$\left(n \sum_{l \in \mathcal{L}_j} \mathbf{b}_l \mathbf{b}_l^\top + \frac{1}{\gamma} \mathbf{I} \right) \mathbf{u} + c_2 \nabla G(\mathbf{u}) = \mathbf{y},$$

where $\mathbf{y} = c_3 n \mathbf{e}_1 \mathbb{1}_{\{j=1\}} + \frac{1}{\gamma}\mathbf{x}$. Since $\gamma < \frac{\sqrt{2}+1}{60 c_2}$, we have the following claims.

(a) If $0 < l < m-1$ and $l \in \mathcal{L}_j$, we have

$$\begin{aligned} n(u_l - u_{l+1}) + \frac{1}{\gamma} u_l + 120 c_2 \frac{u_l^2 (u_l - 1)}{1 + u_l^2} &= y_l \\ n(u_{l+1} - u_l) + \frac{1}{\gamma} u_{l+1} + 120 c_2 \frac{u_{l+1}^2 (u_{l+1} - 1)}{1 + u_{l+1}^2} &= y_{l+1}. \end{aligned} \tag{51}$$

By Lemma 45, $y_l = y_{l+1} = 0$ implies $u_l = u_{l+1} = 0$.

(b) If $m-1 \in \mathcal{L}_j$, we have

$$\begin{aligned} n(u_{m-1}-u_m) + \frac{1}{\gamma}u_{m-1} + 120c_2\frac{u_{m-1}^2(u_{m-1}-1)}{1+u_{m-1}^2} &= y_{m-1} \\ n(u_m - u_{m-1}) + \frac{1}{\gamma}u_m &= y_m. \end{aligned} \tag{52}$$

If $y_{m-1} = y_m = 0$, we obtain

$$\begin{aligned} \frac{1+2\gamma n}{\gamma(1+\gamma n)}u_{m-1} + 120c_2\frac{u_{m-1}^2(u_{m-1}-1)}{1+u_{m-1}^2} &= 0 \\ \left(n+\frac{1}{\gamma}\right)u_m - \frac{1}{\gamma}u_{m-1} &= 0. \end{aligned}$$

By Lemma 44, $u_{m-1} = u_m = 0$.

(c) If $m \in \mathcal{L}_j$, we have

$$n\zeta^2 u_m + \frac{1}{\gamma}u_m = y_m. \tag{53}$$

$y_m = 0$ implies $u_m = 0$.

(d) If $l > 0$ and $l-1, l \notin \mathcal{L}_j$, we have

$$\frac{1}{\gamma}u_l + 120c_2\frac{u_l^2(u_l-1)}{1+u_l^2}\mathbb{1}_{\{l<m\}} = y_l. \tag{54}$$

By Lemma 44, $y_l = 0$ implies $u_l = 0$.

For $\mathbf{x} \in \mathcal{F}_0$ and $j=1$, we have $\mathbf{x} = \mathbf{0}$ and $\mathbf{y} = n\omega^2\mathbf{e}_1$. Since $n \geq 2$, we have $1 \notin \mathcal{L}_1$. If $2 \in \mathcal{L}_1$, we can consider the solution to Equations (51), (52) or (53) and conclude that $u_2 = 0$. If $2 \notin \mathcal{L}_1$, we can consider the solution to Equation (54) and conclude that $u_2 = 0$. Similarly, we can obtain $u_l = 0$ for $l \geq 2$, which implies $\mathbf{u} \in \mathcal{F}_1$.

For $\mathbf{x} \in \mathcal{F}_0$ and $j > 1$, we have $\mathbf{y} = \mathbf{0}$ and $0 \notin \mathcal{L}_j$. If $1 \in \mathcal{L}_j$, we can consider the solution to Equations (51) or (52) and conclude that $u_1 = 0$. If $1 \notin \mathcal{L}_j$, we can consider the solution to Equation (54) and conclude that $u_1 = 0$. Similarly, we can obtain $u_l = 0$ for all l, which implies $\mathbf{u} = \mathbf{0} \in \mathcal{F}_0$.

For $k \geq 1$, we know that $\mathbf{y} \in \mathcal{F}_k$. Suppose $k \in \mathcal{L}_i$.

If $j = i$, we have $k+1 \notin \mathcal{L}_i$. If $k = m-1$, clearly we have $\mathbf{u} \in \mathcal{F}_{k+1}$. Now we suppose $k < m-1$. If $k+2 \in \mathcal{L}_i$, we can consider the solution to Equations (51), (52) or (53) and conclude that $u_{k+2} = 0$. If $k+2 \notin \mathcal{L}_1$, we can consider the solution to Equation (54) and conclude that $u_{k+2} = 0$. Similarly, we can obtain $u_l = 0$ for $l \geq k+2$, which implies $\mathbf{u} \in \mathcal{F}_{k+1}$.

If $j \neq i$, we have $k \notin \mathcal{L}_j$. If $k+1 \in \mathcal{L}_j$, we can consider the solution to Equations (51), (52) or (53) and conclude that $u_{k+1} = 0$. If $k+1 \notin \mathcal{L}_j$, we can consider the solution to Equation (54) and conclude that $u_{k+1} = 0$. Similarly, we can obtain $u_l = 0$ for $l \geq k+1$, which implies $\mathbf{u} \in \mathcal{F}_k$.

This completes the proof. ■

References

Alekh Agarwal and Leon Bottou. A lower bound for the optimization of finite sums. In *International Conference on Machine Learning*, pages 78–86. PMLR, 2015.

Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding approximate local minima faster than gradient descent. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1195–1199, 2017.

Ahmet Alacaoglu and Yura Malitsky. Stochastic variance reduction for variational inequality methods. In *Conference on Learning Theory*, pages 778–816. PMLR, 2022.

Zeyuan Allen-Zhu. Natasha: Faster non-convex stochastic optimization via strongly non-convex parameter. In *International Conference on Machine Learning*, pages 89–97. PMLR, 2017.

Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *Journal of Machine Learning Research*, 18(221):1–51, 2018a.

Zeyuan Allen-Zhu. Katyusha X: Practical momentum method for stochastic sum-of-nonconvex optimization. In *International Conference on Machine Learning*, pages 179–185. PMLR, 2018b.

Yossi Arjevani and Ohad Shamir. Dimension-free iteration complexity of finite sum optimization problems. *Advances in Neural Information Processing Systems*, 29, 2016.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*, volume 28. Princeton University Press, 2009.

Aleksandr Beznosikov, Eduard Gorbunov, Hugo Berard, and Nicolas Loizou. Stochastic gradient descent-ascent: Unified theory and new efficient methods. In *International Conference on Artificial Intelligence and Statistics*, pages 172–235. PMLR, 2023.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1751–1772, 2018.

Yair Carmon, Yujia Jin, Aaron Sidford, and Kevin Tian. Variance reduction for matrix games. *Advances in Neural Information Processing Systems*, 32, 2019.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points I. *Mathematical Programming*, 184(1-2):71–120, 2020a.

Yair Carmon, Yujia Jin, Aaron Sidford, and Kevin Tian. Coordinate methods for matrix games. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 283–293. IEEE, 2020b.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points II: first-order methods. *Mathematical Programming*, 185(1-2):315–355, 2021.

Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40: 120–145, 2011.

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

Tatjana Chavdarova, Gauthier Gidel, François Fleuret, and Simon Lacoste-Julien. Reducing noise in GAN training with variance reduced extragradient. *Advances in Neural Information Processing Systems*, 32, 2019.

Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. SBEED: Convergent reinforcement learning with nonlinear function approximation. In *International Conference on Machine Learning*, pages 1125–1134. PMLR, 2018.

Aaron Defazio. A simple practical accelerated method for finite sums. *Advances in Neural Information Processing Systems*, 29, 2016.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in Neural Information Processing Systems*, 27, 2014.

Simon S Du, Jianshu Chen, Lihong Li, Lin Xiao, and Dengyong Zhou. Stochastic variance reduction methods for policy evaluation. In *International Conference on Machine Learning*, pages 1049–1058. PMLR, 2017.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. SPIDER: Near-optimal nonconvex optimization via stochastic path-integrated differential estimator. *Advances in Neural Information Processing Systems*, 31, 2018.

Robert Hannah, Yanli Liu, Daniel O'Connor, and Wotao Yin. Breaking the span assumption yields fast finite-sum minimization. *Advances in Neural Information Processing Systems*, 31, 2018.

Thomas Hofmann, Aurelien Lucchi, Simon Lacoste-Julien, and Brian McWilliams. Variance reduced stochastic gradient descent with neighbors. *Advances in Neural Information Processing Systems*, 28, 2015.

Adam Ibrahim, Waıss Azizian, Gauthier Gidel, and Ioannis Mitliagkas. Linear lower bounds and conditioning of differentiable games. In *International Conference on Machine Learning*, pages 4583–4593. PMLR, 2020.

Thorsten Joachims. A support vector method for multivariate performance measures. In *International Conference on Machine Learning*, pages 377–384, 2005.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in Neural Information Processing Systems*, 26, 2013.

Dmitry Kovalev, Samuel Horváth, and Peter Richtárik. Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop. In *Algorithmic Learning Theory*, pages 451–467. PMLR, 2020.

Guanghui Lan and Yu Yang. Accelerated stochastic algorithms for nonconvex finite-sum and multiblock optimization. *SIAM Journal on Optimization*, 29(4):2753–2784, 2019.

Guanghui Lan and Yi Zhou. An optimal randomized incremental gradient method. *Mathematical programming*, 171:167–215, 2018.

Bingcong Li, Meng Ma, and Georgios B Giannakis. On the convergence of SARAH and beyond. In *International Conference on Artificial Intelligence and Statistics*, pages 223–233. PMLR, 2020.

Zhize Li. ANITA: An optimal loopless accelerated variance-reduced gradient method. *arXiv preprint arXiv:2103.11333*, 2021.

Zhize Li, Hongyan Bao, Xiangliang Zhang, and Peter Richtárik. PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International Conference on Machine Learning*, pages 6286–6295. PMLR, 2021.

Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. Catalyst acceleration for first-order convex optimization: from theory to practice. *Journal of Machine Learning Research*, 18 (1):7854–7907, 2018.

Tianyi Lin, Chi Jin, and Michael I Jordan. Near-optimal algorithms for minimax optimization. In *Conference on Learning Theory*, pages 2738–2779. PMLR, 2020.

Nicolas Loizou, Hugo Berard, Alexia Jolicoeur-Martineau, Pascal Vincent, Simon Lacoste-Julien, and Ioannis Mitliagkas. Stochastic hamiltonian gradient methods for smooth games. In *International Conference on Machine Learning*, pages 6370–6381. PMLR, 2020.

Luo Luo, Cheng Chen, Yujun Li, Guangzeng Xie, and Zhihua Zhang. A stochastic proximal point algorithm for saddle-point problems. *arXiv preprint arXiv:1909.06946*, 2019.

Luo Luo, Haishan Ye, Zhichao Huang, and Tong Zhang. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. *Advances in Neural Information Processing Systems*, 33, 2020.

Luo Luo, Guangzeng Xie, Tong Zhang, and Zhihua Zhang. Near optimal stochastic algorithms for finite-sum unbalanced convex-concave minimax optimization. *arXiv preprint arXiv:2106.01761*, 2021.

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, pages 1497–1507. PMLR, 2020a.

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

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2013.

Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *International conference on machine learning*, pages 2613–2621. PMLR, 2017.

Dmitrii M Ostrovskii, Andrew Lowy, and Meisam Razaviyayn. Efficient search of first-order nash equilibria in nonconvex-concave smooth min-max problems. *SIAM Journal on Optimization*, 31(4):2508–2538, 2021.

Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185(1-2): 1–35, 2021.

Balamurugan Palaniappan and Francis Bach. Stochastic variance reduction methods for saddle-point problems. *Advances in Neural Information Processing Systems*, 29, 2016.

Xun Qian, Zheng Qu, and Peter Richtárik. L-SVRG and L-Katyusha with arbitrary sampling. *Journal of Machine Learning Research*, 22(1):4991–5039, 2021.

Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Weakly-convex–concave min–max optimization: provable algorithms and applications in machine learning. *Optimization Methods and Software*, 37(3):1087–1121, 2022.

Shai Shalev-Shwartz. Sdca without duality, regularization, and individual convexity. In *International Conference on Machine Learning*, pages 747–754. PMLR, 2016.

Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(1):567–599, 2013.

Zebang Shen, Aryan Mokhtari, Tengfei Zhou, Peilin Zhao, and Hui Qian. Towards more efficient stochastic decentralized learning: Faster convergence and sparse communication. In *International Conference on Machine Learning*, pages 4624–4633. PMLR, 2018.

Conghui Tan, Tong Zhang, Shiqian Ma, and Ji Liu. Stochastic primal-dual method for empirical risk minimization with $O(1)$ per-iteration complexity. *Advances in Neural Information Processing Systems*, 31, 2018.

Kiran K Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Efficient algorithms for smooth minimax optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

Blake E Woodworth and Nati Srebro. Tight complexity bounds for optimizing composite objectives. *Advances in Neural Information Processing Systems*, 29, 2016.

Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

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

Junchi Yang, Siqi Zhang, Negar Kiyavash, and Niao He. A catalyst framework for minimax optimization. *Advances in Neural Information Processing Systems*, 33, 2020.

Yiming Ying, Longyin Wen, and Siwei Lyu. Stochastic online auc maximization. *Advances in Neural Information Processing Systems*, 29, 2016.

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

Siqi Zhang, Junchi Yang, Cristóbal Guzmán, Negar Kiyavash, and Niao He. The complexity of nonconvex-strongly-concave minimax optimization. In *Uncertainty in Artificial Intelligence*, pages 482–492. PMLR, 2021.

Yuchen Zhang and Lin Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. *Journal of Machine Learning Research*, 18(84):1–42, 2017.

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

Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic nested variance reduction for nonconvex optimization. *Journal of Machine Learning Research*, 21(1):4130–4192, 2020.