A NOVEL ANALYSIS FRAMEWORK OF LOWER COMPLEXITY BOUNDS FOR FINITE-SUM OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper studies the lower bound complexity for the optimization problem whose objective function is the average of n individual smooth convex functions. We consider the algorithm which gets access to gradient and proximal oracle for each individual component. For the strongly-convex case, we prove such an algorithm can not reach an ε-suboptimal point in fewer than $\Omega((n+\sqrt{\kappa n})\log(1/\varepsilon))$ iterations, where κ is the condition number of the objective function. This lower bound is tighter than previous results and perfectly matches the upper bound of the existing proximal incremental first-order oracle algorithm Point-SAGA. We develop a novel construction to show the above result, which partitions the tridiagonal matrix of classical examples into n groups to make the problem difficult enough to stochastic algorithms. This construction is friendly to the analysis of proximal oracle and also could be used in general convex and average smooth cases naturally.

1 INTRODUCTION

We consider the minimization of the following optimization problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x}), \tag{1}$$

where the $f_i(\boldsymbol{x})$ are L-smooth and μ-strongly convex. Accordingly, the condition number is defined as $\kappa = L/\mu$, which is typically larger than n in real-world applications. Many machine learning models can be formulated as the above problem such as ridge linear regression, ridge logistic regression, smoothed support vector machines, graphical models, etc. This paper focuses on the first order methods for solving Problem (1), which access to the Proximal Incremental First-order Oracle (PIFO) for each individual component, that is,

$$h_f(\boldsymbol{x}, i, \gamma) \triangleq \left[f_i(\boldsymbol{x}), \nabla f_i(\boldsymbol{x}), \operatorname{prox}_{f_i}^{\gamma}(\boldsymbol{x})\right], \tag{2}$$

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

$$\operatorname{prox}_{f_i}^{\gamma}(\boldsymbol{x}) = \arg\min_{\boldsymbol{u}} \left\{ f_i(\boldsymbol{x}) + \frac{1}{2\gamma}\|\boldsymbol{x}-\boldsymbol{u}\|_2^2 \right\}.$$

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

$$g_f(\boldsymbol{x}, i, \gamma) \triangleq \left[f_i(\boldsymbol{x}), \nabla f_i(\boldsymbol{x})\right].$$

PIFO provides more information than IFO and it would be potentially more powerful than IFO in first order optimization algorithms. Our goal is to find an ε-suboptimal solution $\hat{\boldsymbol{x}}$ such that

$$f(\hat{\boldsymbol{x}}) - \min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) \le \varepsilon$$

by using PIFO or IFO.

There are several first-order stochastic algorithms to solve Problem (1). The key idea to leverage the structure of f is variance reduction which is effective for ill-conditioned problems. For example, SVRG (Zhang et al., 2013; Johnson and Zhang, 2013; Xiao and Zhang, 2014) can find an ε-suboptimal solution in $\mathcal{O}((n+\kappa)\log(1/\varepsilon))$ IFO calls, while the complexity of the classical Nesterov's acceleration (Nesterov, 1983) is $\mathcal{O}(n\sqrt{\kappa}\log(1/\varepsilon))$. Similar results[1] also hold for SAG (Schmidt et al., 2017) and SAGA (Defazio et al., 2014). In fact, there exists an accelerated stochastic gradient method with $\sqrt{\kappa}$ dependency. Defazio (2016) introduced a simple and practical accelerated method called Point SAGA, which reduces the iteration complexity to $\mathcal{O}((n+\sqrt{\kappa n})\log(1/\varepsilon))$. The advantage of Point SAGA is in that it has only one parameter to be tuned, but the iteration depends on PIFO rather than IFO. Allen-Zhu (2017) proposed the Katyusha momentum to accelerate variance reduction algorithms, which achieves the same iteration complexity as Point-SAGA but only requires IFO calls.

The lower bound complexities of IFO algorithms for convex optimization have been well studied (Agarwal and Bottou, 2015; Arjevani and Shamir, 2015; Woodworth and Srebro, 2016; Carmon et al., 2017; Lan and Zhou, 2017; Zhou and Gu, 2019). Specifically, Lan and Zhou (2017) showed that at least $\Omega((n+\sqrt{\kappa n})\log(1/\varepsilon))$ IFO calls[2] are needed to obtain an ε-suboptimal solution for some complicated objective functions. This lower bound is optimal because it matches the upper bound complexity of Katyusha (Allen-Zhu, 2017).

It would be interesting whether we can establish a more efficient PIFO algorithm than IFO one. Woodworth and Srebro (2016) provided a lower bound $\Omega(n+\sqrt{\kappa n}\log(1/\varepsilon))$ for PIFO algorithms, while the known upper bound of the PIFO algorithm Point SAGA [3] is $\mathcal{O}((n+\sqrt{\kappa n})\log(1/\varepsilon))$. The difference of dependency on n implies that the existing theory of PIFO algorithm is not perfect. This gap can not be ignored because the number of components n is typically very large in many machine learning problems. A natural question is can we design a PIFO algorithm whose upper bound complexity matches Woodworth and Srebro's lower bound, or can we improve the lower bound complexity of PIFO to match the upper bound of Point SAGA.

In this paper, we prove the lower bound complexity of PIFO algorithm is $\Omega((n+\sqrt{\kappa n})\log(1/\varepsilon))$ for smooth and strongly-convex f_i, which means the existing Point-SAGA (Defazio, 2016) has achieved optimal complexity and PIFO can not lead to a tighter upper bound than IFO. We provide a novel construction, showing the above result by decomposing the classical tridiagonal matrix (Nesterov, 2013) into n groups. This technique is quite different from the previous lower bound complexity analysis (Agarwal and Bottou, 2015; Woodworth and Srebro, 2016; Lan and Zhou, 2017; Zhou and Gu, 2019). Moreover, it is very friendly to the analysis of proximal operation and easy to follow. We also use this technique to study general convex and average smooth cases (Allen-Zhu, 2018; Zhou and Gu, 2019), obtaining the similar lower bounds to the previous work (Woodworth and Srebro, 2016; Zhou and Gu, 2019).

2 Our Analysis Framework

In this paper, we consider the Proximal Incremental First-order Oracle (PIFO) algorithm for smooth convex finite-sum optimization. All proofs in this section can be found in Appendices B and C for a detailed version. We analyze the lower bounds of the algorithms when the objective functions are respectively strongly convex, general convex, smooth and average smooth (Zhou and Gu, 2019).

Definition 2.1. *For any differentiable function $f:\mathbb{R}^m\to\mathbb{R}$,*

- *f is convex, if for any $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^m$ it satisfies $f(\boldsymbol{y})\geq f(\boldsymbol{x})+\langle\nabla f(\boldsymbol{x}),\boldsymbol{y}-\boldsymbol{x}\rangle$.*
- *f is μ-strongly convex, if for any $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^m$ it satisfies*
  $$f(\boldsymbol{y})\geq f(\boldsymbol{x})+\langle\nabla f(\boldsymbol{x}),\boldsymbol{y}-\boldsymbol{x}\rangle+\frac{\mu}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2.$$
- *f is L-smooth, if for any $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^m$ it satisfies $\|\nabla f(\boldsymbol{x})-\nabla f(\boldsymbol{y})\|_2\leq L\|\boldsymbol{x}-\boldsymbol{y}\|_2$.*

[1]SVRG, SAG and SAGA only need to introduce the proximal operation for composite objective, that is, $f_i(\boldsymbol{x})=g_i(\boldsymbol{x})+h(\boldsymbol{x})$, where h may be non-smooth. Their iterations only depend on IFO when all the $f_i(x)$ are smooth. Hence, we regard these algorithms only require IFO calls in this paper.

[2]Lan and Zhou's construction requires f to be μ-strongly convex and every f_i to be convex, while this paper studies the lower bound with stronger condition that is every f_i is μ-strongly convex. For the same lower bound complexity, the result with stronger assumptions on the objective functions is stronger.

	Upper Bounds	Previous Lower Bounds	Our Lower Bounds
f_i is L-smooth and μ-strongly convex	$\mathcal{O}\left((n+\sqrt{\kappa n})\log(\frac{1}{\varepsilon})\right)$ (Allen-Zhu, 2017), IFO (Defazio, 2016), PIFO	$\Omega\left(n+\sqrt{\kappa n}\log(\frac{1}{\varepsilon})\right)$ (Woodworth and Srebro, 2016) PIFO	$\Omega\left((n+\sqrt{\kappa n})\log(\frac{1}{\varepsilon})\right)$ [Theorem 3.1] PIFO
f_i is L-smooth and convex	$\mathcal{O}\left(n\log(\frac{1}{\varepsilon})+\sqrt{\frac{nL}{\varepsilon}}\right)$ (Allen-Zhu, 2017) IFO	$\Omega\left(n+\sqrt{\frac{nL}{\varepsilon}}\right)$ (Woodworth and Srebro, 2016) PIFO	$\Omega\left(n+\sqrt{\frac{nL}{\varepsilon}}\right)$ [Theorem 3.3] PIFO
$\{f_i\}_{i=1}^n$ is L-average smooth and f is μ-strongly convex	$\mathcal{O}\left(\left(n+n^{3/4}\sqrt{\kappa}\right)\log\left(\frac{1}{\varepsilon}\right)\right)$ (Allen-Zhu, 2018) IFO	$\Omega\left(n+n^{3/4}\sqrt{\kappa}\log\left(\frac{1}{\varepsilon}\right)\right)$ (Zhou and Gu, 2019) IFO	$\Omega\left(\left(n+n^{3/4}\sqrt{\kappa}\right)\log\left(\frac{1}{\varepsilon}\right)\right)$ [Theorem 3.5] PIFO
$\{f_i\}_{i=1}^n$ is L-average smooth and f is convex	$\mathcal{O}\left(n+n^{3/4}\sqrt{\frac{L}{\varepsilon}}\right)$ (Allen-Zhu, 2018) IFO	$\Omega\left(n+n^{3/4}\sqrt{\frac{L}{\varepsilon}}\right)$ (Zhou and Gu, 2019) IFO	$\Omega\left(n+n^{3/4}\sqrt{\frac{L}{\varepsilon}}\right)$ [Theorem 3.7] PIFO

Table 1: We compare our PIFO lower bounds with existing results of IFO or PIFO algorithms, where $\kappa = L/\mu$. Note that the call of PIFO could obtain more information than IFO. Hence, any PIFO lower bound also can be regarded as an IFO lower bound, not vice versa.

Definition 2.2. *We say differentiable functions $\{f_i\}_{i=1}^n,\ f_i : \mathbb{R}^m \to \mathbb{R}$, to be L-average smooth if for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^m$, they satisfy*

$$\frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{y})\right\|_2^2 \le L^2 \left\|\boldsymbol{x}-\boldsymbol{y}\right\|_2^2. \tag{3}$$

Remark 2.3. *We point out that*

1. *if each f_i is L-smooth, then $\{f_i\}_{i=1}^n$ are L-average smooth.*
2. *if $\{f_i\}_{i=1}^n$ are L-average smooth, then $f(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^n f_i(\boldsymbol{x})$ is L-smooth.*

We present the formal definition for PIFO algorithm.

Definition 2.4. *Consider a stochastic optimization algorithm $\mathcal{A}$ to solve Problem (1). Let $\boldsymbol{x}_t$ be the point obtained at time-step t and the algorithm starts with $\boldsymbol{x}_0$. The algorithm $\mathcal{A}$ is said to be a PIFO algorithm if for any $t \ge 0$, we have*

$$\boldsymbol{x}_t \in \operatorname{span}\left\{\boldsymbol{x}_0, \dots, \boldsymbol{x}_{t-1}, \nabla f_{i_1}(\boldsymbol{x}_0), \cdots, \nabla f_{i_t}(\boldsymbol{x}_{t-1}), \operatorname{prox}_{f_{i_1}}^{\gamma_1}(\boldsymbol{x}_0), \cdots, \operatorname{prox}_{f_{i_t}}^{\gamma_t}(\boldsymbol{x}_{t-1})\right\}, \tag{4}$$

where i_t is a random variable supported on $[n]$ and takes $\mathbb{P}(i_t = j) = p_j$ for each $t \ge 0$ and $1 \le j \le n$ where $\sum_{j=1}^n p_j = 1$.

Without loss of generality, we assume $\boldsymbol{x}_0 = \boldsymbol{0}$ and $p_1 \le p_2 \le \cdots \le p_n$ to simplify our analysis. Otherwise, we can take $\{\hat{f}_i(\boldsymbol{x}) = f_i(\boldsymbol{x}+\boldsymbol{x}_0)\}_{i=1}^n$ into consideration. On the other hand, suppose that $p_{s_1} \le p_{s_2} \le \cdots \le p_{s_n}$ where $\{s_i\}_{i=1}^n$ is a permutation of $[n]$. Define $\{\tilde{f}_i\}_{i=1}^n$ such that $\tilde{f}_{s_i} = f_i$, then $\mathcal{A}$ takes component $\tilde{f}_{s_i}$ in probability p_{s_i}, i.e., $\mathcal{A}$ takes f_i in probability p_{s_i}.

To demonstrate the construction of adversarial functions, we first introduce the following class of matrices:

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

Then we define

$$\boldsymbol{A}(m,\omega) \triangleq \boldsymbol{B}(m,\omega)^\top \boldsymbol{B}(m,\omega) = \begin{bmatrix} \omega^2+1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{bmatrix}. \tag{5}$$

The matrix $\boldsymbol{A}(m,\omega)$ is widely-used in the analysis of lower bounds for convex optimization (Nesterov, 2013; Agarwal and Bottou, 2015; Lan and Zhou, 2017; Carmon et al., 2017; Zhou and Gu, 2019). We now present a decomposition of $\boldsymbol{A}(m,\omega)$ based on Eq. (5).

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

$$\mathcal{L}_i = \left\{l : 1 \le l \le m, l \equiv i-1 (\bmod n)\right\}, \quad i = 1,2,\cdots,n.$$

Our construction is based on the following class of functions

$$r(\boldsymbol{x};\lambda_0,\lambda_1,\lambda_2,m,\omega) \triangleq \frac{1}{n}\sum_{i=1}^n r_i(\boldsymbol{x};\lambda_0,\lambda_1,\lambda_2,m,\omega),$$

where

$$r_i(\boldsymbol{x};\lambda_0,\lambda_1,\lambda_2,m,\omega) = \begin{cases} \lambda_1 \sum\limits_{l\in\mathcal{L}_1} \left\|\boldsymbol{b}_l(m,\omega)^\top \boldsymbol{x}\right\|_2^2 + \lambda_2 \|\boldsymbol{x}\|_2^2 - \lambda_0\langle \boldsymbol{e}_m, \boldsymbol{x}\rangle, & \text{for } i=1, \\ \lambda_1 \sum\limits_{l\in\mathcal{L}_i} \left\|\boldsymbol{b}_l(m,\omega)^\top \boldsymbol{x}\right\|_2^2 + \lambda_2 \|\boldsymbol{x}\|_2^2, & \text{for } i=2,3,\cdots,n. \end{cases} \tag{6}$$

We can determine the smooth and strongly-convex coefficients of r_i as follows.

Proposition 2.5. *For any $\lambda_1 > 0, \lambda_2 \ge 0, \omega < \sqrt{2}$, we have that the r_i are $(4\lambda_1 + 2\lambda_2)$-smooth and λ_2-strongly convex, and $\{r_i\}_{i=1}^n$ is L'-average smooth where*

$$L' = 2\sqrt{\frac{4}{n}\left[(\lambda_1+\lambda_2)^2 + \lambda_1^2\right] + \lambda_2^2}.$$

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

$$\mathcal{F}_k = \begin{cases} \operatorname{span}\{\boldsymbol{e}_m, \boldsymbol{e}_{m-1}, \cdots, \boldsymbol{e}_{m-k+1}\}, & \text{for } 1 \le k \le m, \\ \{\boldsymbol{0}\}, & \text{for } k = 0. \end{cases}$$

The following technical lemma plays a crucial role in our proof.

Lemma 2.6. *For any $\lambda_0 \ne 0, \lambda_1 > 0, \lambda_2 \ge 0$ and $\boldsymbol{x} \in \mathcal{F}_k$, $0 \le k < m$, we have that*

$$\nabla r_i(\boldsymbol{x};\lambda_0,\lambda_1,\lambda_2,m,\omega) \;\; \textit{and} \;\; \operatorname{prox}_{r_i}^{\gamma}(\boldsymbol{x}) \in \begin{cases} \mathcal{F}_{k+1}, & \textit{if } k \equiv i-1 (\bmod n), \\ \mathcal{F}_k, & \textit{otherwise}. \end{cases}$$

In short, if $\boldsymbol{x} \in \mathcal{F}_k$ and let $f_i(\boldsymbol{x}) \triangleq r_i(\boldsymbol{x};\lambda_0,\lambda_1,\lambda_2,\omega)$, then there exists only one $i \in \{1,\dots,n\}$ such that $h_f(\boldsymbol{x}, i, \gamma)$ could (and only could) provide additional information in $\mathcal{F}_{k+1}$. The "only one" property is important to the lower bound analysis for first order stochastic optimization algorithms (Lan and Zhou, 2017; Zhou and Gu, 2019), but these prior constructions only work for IFO rather than PIFO.

Lemma 2.6 implies that $\boldsymbol{x}_t = \boldsymbol{0}$ will host until algorithm $\mathcal{A}$ draws the component f_1. Then, for any $t < T_1 = \min_t\{t : i_t = 1\}$, we have $\boldsymbol{x}_t \in \mathcal{F}_0$ and $\boldsymbol{x}_{T_1} \in \mathcal{F}_1$. The value of T_1 can be regarded as the smallest integer such that $\boldsymbol{x}_{T_1}$ could host. Similarly, we can define T_k to be the smallest integer such that $\boldsymbol{x}_{T_k} \in \mathcal{F}_k$ could host. We give the formal definition of T_k recursively and connect it to geometrically distributed random variables in the following corollary.

Corollary 2.7. *Let*

$$T_0 = 0, \;\; \textit{and} \;\; T_k = \min_t\{t : t > T_{k-1}, i_t \equiv k \; (\bmod n)\} \;\; \textit{for } k \ge 1. \tag{7}$$

Then for any $k \ge 1$ and $t < T_k$, we have $\boldsymbol{x}_t \in \mathcal{F}_{k-1}$. Moreover, T_k can be written as sum of k independent random variables $\{Y_l\}_{1\le l\le k}$, i.e., $T_k = \sum_{l=1}^k Y_l$, where Y_l follows a geometric distribution with success probability $q_l = p_{l'}$ where $l' \equiv l \; (\bmod n), 1 \le l' \le n$.

The basic idea of our analysis is that we guarantee the minimizer of r lies in $\mathcal{F}_m$ and assure the PIFO algorithm extend the space of $\mathrm{span}\{\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_t\}$ slowly with t increasing. We know that $\mathrm{span}\{\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{T_k}\} \subseteq \mathcal{F}_k$ by Corollary 2.7. Hence, T_k is just the quantity that reflects how $\mathrm{span}\{\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_t\}$ verifies. Because T_k can be written as the sum of geometrically distributed random variables, we needs to introduce some properties of such random variables which derive the lower bounds of our construction.

Lemma 2.8. *Let $\{Y_i\}_{1 \leq i \leq N}$ be independent random variables, and Y_i follows a geometric distribution with success probability p_i. Then*

$$\mathbb{P}\left(\sum_{i=1}^{N} Y_i > \frac{N^2}{4(\sum_{i=1}^{N} p_i)}\right) \geq 1 - \frac{16}{9N}. \tag{8}$$

From Lemma 2.8, the following result implies how many PIFO calls we need.

Lemma 2.9. *If $M \geq 1$ satisfies $\min_{\boldsymbol{x} \in \mathcal{F}_M} f(\boldsymbol{x}) - \min_{\boldsymbol{x} \in \mathbb{R}^m} f(\boldsymbol{x}) \geq 9\varepsilon$ and $N = n(M+1)/4$, then we have*

$$\min_{t \leq N} \mathbb{E} f(\boldsymbol{x}_t) - \min_{\boldsymbol{x} \in \mathbb{R}^m} f(\boldsymbol{x}) \geq \varepsilon.$$

Proof. Denote $\min_{\boldsymbol{x} \in \mathbb{R}^m} f(\boldsymbol{x})$ by f^*. For $t \leq N$, we have

$$\begin{aligned}
\mathbb{E} f(\boldsymbol{x}_t) - f^* &\geq \mathbb{E}[f(\boldsymbol{x}_t) - f^* | N < T_{M+1}]\mathbb{P}\left(N < T_{M+1}\right) \\
&\geq \mathbb{E}[\min_{\boldsymbol{x} \in \mathcal{F}_M} f(\boldsymbol{x}) - f^* | N < T_{M+1}]\mathbb{P}\left(N < T_{M+1}\right) \\
&\geq 9\varepsilon \mathbb{P}\left(T_{M+1} > N\right),
\end{aligned}$$

where T_{M+1} is defined in (7), and the second inequality follows from Corollary 2.7 (if $N < T_{M+1}$, then $\boldsymbol{x}_t \in \mathcal{F}_M$ for $t \leq N$).

By Corollary 2.7, 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 $q_l = p_{l'}$ ($l' \equiv l (\mathrm{mod}\ n)$, $1 \leq l' \leq n$). Moreover, recalling that $p_1 \leq p_2 \leq \cdots \leq p_n$, we have $\sum_{l=1}^{M+1} q_l \leq \frac{M+1}{n}$.

Therefore, by Lemma 2.8, we have

$$\mathbb{P}\left(T_{M+1} > N\right) = \mathbb{P}\left(\sum_{l=1}^{M+1} Y_l > \frac{(M+1)n}{4}\right) \geq 1 - \frac{16}{9(M+1)} \geq \frac{1}{9},$$

Hence, we can conclude that $\mathbb{E} f(\boldsymbol{x}_N) - f^* \geq 9\varepsilon \mathbb{P}\left(T_{M+1} > N\right) \geq \varepsilon$. □

Remark In fact, a more strong conclusion hosts:

$$\mathbb{E}\left[\min_{t \leq N} f(\boldsymbol{x}_t)\right] - \min_{\boldsymbol{x} \in \mathbb{R}^m} f(\boldsymbol{x}) \geq \varepsilon.$$

3 Main Results

We present the our lower bound results for PIFO algorithms and summarize all of results in Table 1 and 2 . We first start with smooth and strongly convex setting, then consider the general convex and average smooth cases.

Theorem 3.1. *For any PIFO algorithm $\mathcal{A}$ and any $L, \mu, n, \Delta, \varepsilon$ such that $\kappa = L/\mu \geq n/2 + 1$, and $\varepsilon/\Delta \leq 0.00327$, there exist a dimension $d = \mathcal{O}\left(\sqrt{\kappa/n} \log\left(\Delta/\varepsilon\right)\right)$ and n L-smooth and μ-strongly convex functions $\{f_i : \mathbb{R}^d \to \mathbb{R}\}_{i=1}^n$ such that $f(\boldsymbol{x}_0) - f(\boldsymbol{x}^*) = \Delta$. In order to find $\hat{\boldsymbol{x}} \in \mathbb{R}^d$ such that $\mathbb{E} f(\hat{\boldsymbol{x}}) - f(\boldsymbol{x}^*) < \varepsilon$, $\mathcal{A}$ needs at least $\Omega\left((n + \sqrt{\kappa n}) \log\left(\Delta/\varepsilon\right)\right)$ queries to h_f.*

Remark 3.2. *In fact, the upper bound of the existing PIFO algorithm Point SAGA (Defazio, 2016) [3] is $\mathcal{O}\left((n + \sqrt{\kappa n}) \log\left(1/\varepsilon\right)\right)$. Hence, the lower bound in Theorem 3.1 is tight, while Woodworth and*

[3]Defazio (2016) proves Point SAGA requires $\mathcal{O}\left((n + \sqrt{\kappa n}) \log\left(1/\varepsilon\right)\right)$ PIFO calls to find $\hat{\boldsymbol{x}}$ such that $\mathbb{E}\|\hat{\boldsymbol{x}} - \boldsymbol{x}^*\|_2^2 < \varepsilon$, where $\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}} f(\boldsymbol{x})$, which is not identical to the condition $\mathbb{E} f(\hat{\boldsymbol{x}}) - f(\boldsymbol{x}^*) < \varepsilon\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2$ in Theorem 3.1. However, it is unnecessary to worry about it because we also establish a PIFO lower bound $\Omega\left((n + \sqrt{\kappa n}) \log\left(1/\varepsilon\right)\right)$ for $\mathbb{E}\|\hat{\boldsymbol{x}} - \boldsymbol{x}^*\|_2^2 < \varepsilon\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2$ in Theorem 4.3.

Srebro (2016) only provided lower bound $\Omega\left(n+\sqrt{\kappa n}\log\left(1/\varepsilon\right)\right)$ which is not optimal to n dependency. The theorem also shows that the PIFO algorithm can not be more powerful than the IFO algorithm in the worst case, because the upper bound of the IFO algorithm (Allen-Zhu, 2017) is also $\mathcal{O}\left(\left(n+\sqrt{\kappa n}\right)\log\left(1/\varepsilon\right)\right)$.

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

Theorem 3.3. *For any PIFO algorithm $\mathcal{A}$ and any L, n, B, ε such that $\varepsilon \leq LB^2/4$, there exist a dimension $d = \mathcal{O}\left(1 + B\sqrt{L/(n\varepsilon)}\right)$ and n L-smooth and convex functions $\{f_i : \mathbb{R}^d \to \mathbb{R}\}_{i=1}^n$ such that $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2 \leq B$. In order to find $\hat{\boldsymbol{x}} \in \mathbb{R}^d$ such that $\mathbb{E}f(\hat{\boldsymbol{x}}) - f(\boldsymbol{x}^*) < \varepsilon$, $\mathcal{A}$ needs at least $\Omega\left(n+B\sqrt{nL/\varepsilon}\right)$ queries to h_f.*

Remark 3.4. *The lower bound in Theorem 3.3 is the same as the one of Woodworth and Srebro's result. However, our construction only requires the dimension be $\mathcal{O}\left(1 + B\sqrt{L/(n\varepsilon)}\right)$, which is much smaller than $\mathcal{O}\left(\frac{L^2B^4}{\varepsilon^2}\log\left(\frac{nLB^2}{\varepsilon}\right)\right)$ in (Woodworth and Srebro, 2016).*

Then we extend our results to the weaker assumption: that is, the objective function F is L-average smooth (Zhou and Gu, 2019). We start with the case that F is strongly convex.

Theorem 3.5. *For any PIFO algorithm $\mathcal{A}$ and any $L, \mu, n, \Delta, \varepsilon$ such that $\kappa = L/\mu \geq \sqrt{3/n}\left(\frac{n}{2}+1\right)$, and $\varepsilon/\Delta \leq 0.00327$, there exist a dimension $d = \mathcal{O}\left(n^{-1/4}\sqrt{\kappa}\log\left(\Delta/\varepsilon\right)\right)$ and n functions $\{f_i : \mathbb{R}^d \to \mathbb{R}\}_{i=1}^n$ where the $\{f_i\}_{i=1}^n$ are L-average smooth and f is μ-strongly convex, such that $f(\boldsymbol{x}_0) - f(\boldsymbol{x}^*) = \Delta$. In order to find $\hat{\boldsymbol{x}} \in \mathbb{R}^d$ such that $\mathbb{E}f(\hat{\boldsymbol{x}}) - f(\boldsymbol{x}^*) < \varepsilon$, $\mathcal{A}$ needs at least $\Omega\left(\left(n+n^{3/4}\sqrt{\kappa}\right)\log\left(\Delta/\varepsilon\right)\right)$ queries to h_f.*

Remark 3.6. *Compared with Zhou and Gu's lower bound $\Omega\left(n + n^{3/4}\sqrt{\kappa}\log\left(\Delta/\varepsilon\right)\right)$ for IFO algorithms, Theorem 3.5 shows tighter dependency on n and supports PIFO algorithms additionally.*

We also give the lower bound for general convex case under the L-average smooth condition.

Theorem 3.7. *For any PIFO algorithm $\mathcal{A}$ and any L, n, B, ε such that $\varepsilon \leq LB^2/4$, there exist a dimension $d = \mathcal{O}\left(1 + Bn^{-1/4}\sqrt{L/\varepsilon}\right)$ and n functions $\{f_i : \mathbb{R}^d \to \mathbb{R}\}_{i=1}^n$ which the $\{f_i\}_{i=1}^n$ are L-average smooth and f is convex, such that $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2 \leq B$. In order to find $\hat{\boldsymbol{x}} \in \mathbb{R}^d$ such that $\mathbb{E}f(\hat{\boldsymbol{x}}) - f(\boldsymbol{x}^*) < \varepsilon$, $\mathcal{A}$ needs at least $\Omega\left(n + Bn^{3/4}\sqrt{L/\varepsilon}\right)$ queries to h_f.*

Remark 3.8. *The lower bound in Theorem 3.7 is comparable to the one of Zhou and Gu's result, but our construction only requires the dimension be $\mathcal{O}\left(1 + Bn^{-1/4}\sqrt{L/\varepsilon}\right)$, which is much smaller than $\mathcal{O}\left(n + Bn^{3/4}\sqrt{L/\varepsilon}\right)$ in (Zhou and Gu, 2019).*

4 Constructions in Proof of Main Theorems

We demonstrate the detailed constructions for PIFO lower bounds in this section. All the omitted proof in this section can be found in Appendix for a detailed version.

4.1 Strongly Convex Case

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

Definition 4.1. *For fixed L, μ, Δ, n, let $\alpha = \sqrt{\frac{2(L/\mu-1)}{n}+1}$. We define $f_{SC,i} : \mathbb{R}^m \to \mathbb{R}$ as follows*

$$f_{SC,i}(\boldsymbol{x}) = r_i\left(\boldsymbol{x}; \sqrt{\frac{2(L-\mu)n\Delta}{\alpha-1}}, \frac{L-\mu}{4}, \frac{\mu}{2}, m, \sqrt{\frac{2}{\alpha+1}}\right), \text{ for } 1 \leq i \leq n, \tag{9}$$

and

$$F_{SC}(\boldsymbol{x}) \triangleq \frac{1}{n}\sum_{i=1}^n f_{SC,i}(\boldsymbol{x}) = \frac{L-\mu}{4n}\left\|\boldsymbol{B}\left(m, \sqrt{\frac{2}{\alpha+1}}\right)\boldsymbol{x}\right\|_2^2 + \frac{\mu}{2}\|\boldsymbol{x}\|_2^2 - \sqrt{\frac{2(L-\mu)\Delta}{n(\alpha-1)}}\langle \boldsymbol{e}_m, \boldsymbol{x}\rangle.$$

Note that the $f_{\text{SC},i}$ are L-smooth and μ-strongly convex, and $F_{\text{SC}}(\boldsymbol{x}_0) - F_{\text{SC}}(\boldsymbol{x}^*) = \Delta$ (see Proposition D.1 in Appendix for more details). Next we show that the functions $\{f_{\text{SC},i}\}_{i=1}^n$ are "hard enough" for any PIFO algorithm $\mathcal{A}$, and deduce the conclusion of Theorem 3.1.

Theorem 4.2. *Suppose that*

$$\frac{L}{\mu} \geq \frac{n}{2} + 1,\ \varepsilon \leq \frac{\Delta}{9}\left(\frac{\sqrt{2}-1}{\sqrt{2}+1}\right)^2,\ \textit{and}\ m = \frac{1}{4}\left(\sqrt{2\frac{L/\mu - 1}{n} + 1}\right)\log\left(\frac{\Delta}{9\varepsilon}\right) + 1.$$

In order to find $\hat{\boldsymbol{x}} \in \mathbb{R}^m$ such that $\mathbb{E}F_{SC}(\hat{\boldsymbol{x}}) - F_{SC}(\boldsymbol{x}^*) < \varepsilon$, PIFO algorithm $\mathcal{A}$ needs at least $\Omega\left(\left(n + \sqrt{\frac{nL}{\mu}}\right)\log\left(\frac{\Delta}{\varepsilon}\right)\right)$ queries to $h_{F_{SC}}$.

Defazio (2016) showed that the PIFO algorithm Point SAGA has the convergence result $\mathbb{E}\left\|\boldsymbol{x}_t - \boldsymbol{x}^*\right\|_2^2 \leq (q')^t \left\|\boldsymbol{x}_0 - \boldsymbol{x}^*\right\|_2$, where q' satisfies $-1/\log(q') = \mathcal{O}\left(n + \sqrt{nL/\mu}\right)$. To match this form of upper bound, we point out that a similar result holds for $\{f_{\text{SC},i}\}_{i=1}^n$.

Theorem 4.3. *Suppose that*

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

In order to find $\hat{\boldsymbol{x}} \in \mathbb{R}^m$ such that $\mathbb{E}\left\|\hat{\boldsymbol{x}} - \boldsymbol{x}^*\right\|_2^2 < \varepsilon \left\|\boldsymbol{x}_0 - \boldsymbol{x}^*\right\|_2^2$, PIFO algorithm $\mathcal{A}$ needs at least $\Omega\left(\left(n + \sqrt{\frac{nL}{\mu}}\right)\log\left(\frac{1}{\varepsilon}\right)\right)$ queries to $h_{F_{SC}}$.

4.2 Convex Case

The analysis of lower bound complexity for non strongly-convex cases depends on the following construction.

Definition 4.4. *For fixed L, B, n, we define $f_{C,i} : \mathbb{R}^m \to \mathbb{R}$ as follows*

$$f_{C,i}(\boldsymbol{x}) = r_i\left(\boldsymbol{x}; \frac{\sqrt{3}}{2}\frac{BL}{(m+1)^{3/2}}, \frac{L}{4}, 0, m, 1\right) \tag{10}$$

and

$$F_C(\boldsymbol{x}) \triangleq \frac{1}{n}\sum_{i=1}^n f_{C,i}(\boldsymbol{x}) = \frac{L}{4n}\left\|\boldsymbol{B}(m,1)\boldsymbol{x}\right\|_2^2 - \frac{\sqrt{3}}{2}\frac{BL}{(m+1)^{3/2}n}\langle \boldsymbol{e}_m, \boldsymbol{x}\rangle.$$

Note that the $f_{\text{C},i}$ are L-smooth and convex, and $\left\|\boldsymbol{x}_0 - \boldsymbol{x}^*\right\|_2 \leq B$ (see Proposition F.1 in Appendix for more details). Next we show the lower bound for functions $f_{\text{C},i}$ defined above.

Theorem 4.5. *Suppose that*

$$\varepsilon \leq \frac{B^2 L}{384n}\ \textit{and}\ m = \left\lfloor\sqrt{\frac{B^2 L}{24n\varepsilon}}\right\rfloor - 1.$$

In order to find $\hat{\boldsymbol{x}} \in \mathbb{R}^m$ such that $\mathbb{E}F_C(\hat{\boldsymbol{x}}) - F_C(\boldsymbol{x}^*) < \varepsilon$, $\mathcal{A}$ needs at least $\Omega\left(n + B\sqrt{\frac{nL}{\varepsilon}}\right)$ queries to h_{F_C}.

To derive Theorem 3.3, we also need the following lemma in the case $\varepsilon > \frac{B^2 L}{384n}$.

Lemma 4.6. *For any PIFO algorithm $\mathcal{A}$ and any L, n, B, ε such that $\varepsilon \leq LB^2/4$, there exist n L-smooth and convex functions $\{f_i : \mathbb{R} \to \mathbb{R}\}_{i=1}^n$ such that $|x_0 - x^*| \leq B$. In order to find $\hat{x} \in \mathbb{R}$ such that $\mathbb{E}F(\hat{x}) - F(x^*) < \varepsilon$, $\mathcal{A}$ needs at least $\Omega(n)$ queries to h_F.*

It is worth noting that if $\varepsilon > \frac{B^2 L}{384n}$, then $\Omega(n) = \Omega\left(n + B\sqrt{\frac{nL}{\varepsilon}}\right)$. Thus combining Theorem 4.5 and Lemma 4.6, we obtain Theorem 3.3.

4.3 Average Smooth Case

Zhou and Gu (2019) established lower bounds of IFO complexity under the average smooth assumption. Here we demonstrate that our technique can also develop lower bounds of PIFO algorithm under this assumption.

4.3.1 F is Strongly Convex

For fixed $L', \mu, \Delta, n, \varepsilon$, we set $L = \sqrt{\frac{n(L'^2-\mu^2)}{2} - \mu^2}$, and consider $\{f_{\mathrm{SC},i}\}_{i=1}^n$ and F_{SC} defined in Definition 4.1.

Proposition 4.7. *For $n \geq 2$, we have that*

1. *$F_{SC}(\boldsymbol{x})$ is μ-strongly convex and $\{f_{SC,i}\}_{i=1}^n$ is L'-average smooth.*
2. *If $\frac{L'}{\mu} \geq \sqrt{\frac{3}{n}(\frac{n}{2}+1)}$, then we have $\sqrt{\frac{n}{3}}L' \leq L \leq \sqrt{\frac{n}{2}}L'$ and $L/\mu \geq n/2+1$.*

Theorem 4.8. *Suppose that*

$$\frac{L'}{\mu} \geq \sqrt{\frac{3}{n}}\left(\frac{n}{2}+1\right), \; \varepsilon \leq \frac{\Delta}{9}\left(\frac{\sqrt{2}-1}{\sqrt{2}+1}\right)^2, \; \text{and } m = \frac{1}{4}\left(\sqrt{\sqrt{\frac{2}{n}}\frac{L'}{\mu}+1}\right)\log\left(\frac{\Delta}{9\varepsilon}\right)+1.$$

In order to find $\hat{\boldsymbol{x}} \in \mathbb{R}^m$ such that $\mathbb{E}F_{SC}(\hat{\boldsymbol{x}}) - F_{SC}(\boldsymbol{x}^*) < \varepsilon$, PIFO algorithm $\mathcal{A}$ needs at least $\Omega\left(\left(n + n^{3/4}\sqrt{\frac{L'}{\mu}}\right)\log\left(\frac{\Delta}{\varepsilon}\right)\right)$ queries to $h_{F_{SC}}$.

4.3.2 F is Convex

For fixed L', B, n, ε, we set $L = \sqrt{\frac{n}{2}}L'$, and consider $\{f_{\mathrm{C},i}\}_{i=1}^n$ and F_{C} defined in Definition 4.4. It follows from Proposition 2.5 that $\{f_{\mathrm{C},i}\}_{i=1}^n$ is L'-average smooth.

Theorem 4.9. *Suppose that*

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

In order to find $\hat{\boldsymbol{x}} \in \mathbb{R}^m$ such that $\mathbb{E}F_C(\hat{\boldsymbol{x}}) - F_C(\boldsymbol{x}^*) < \varepsilon$, $\mathcal{A}$ needs at least $\Omega\left(n + Bn^{3/4}\sqrt{\frac{L'}{\varepsilon}}\right)$ queries to h_{F_C}.

Similar to Lemma 4.6, we also need the following lemma for the case $\varepsilon > \frac{\sqrt{2}}{768}\frac{B^2L'}{\sqrt{n}}$.

Lemma 4.10. *For any PIFO algorithm $\mathcal{A}$ and any L, n, B, ε such that $\varepsilon \leq LB^2/4$, there exist n functions $\{f_i : \mathbb{R} \to \mathbb{R}\}_{i=1}^n$ which is L-average smooth, such that $F(x)$ is convex and $\|x_0 - x^*\|_2 \leq B$. In order to find $\hat{x} \in \mathbb{R}$ such that $\mathbb{E}F(\hat{x}) - F(x^*) < \varepsilon$, $\mathcal{A}$ needs at least $\Omega(n)$ queries to h_F.*

Similarly, note that if $\varepsilon > \frac{\sqrt{2}}{768}\frac{B^2L'}{\sqrt{n}}$, then $\Omega(n) = \Omega\left(n + Bn^{3/4}\sqrt{\frac{L'}{\varepsilon}}\right)$. In summary, we obtain Theorem 3.7.

5 Conclusion and Future Work

In this paper we have studied lower bound of PIFO algorithm for smooth convex finite-sum optimization. We have given a tight lower bound of PIFO algorithms in the strongly convex case. We have proposed a novel construction framework that is very useful to the analysis of proximal algorithms. Based on this framework, We have also extended our result to non-strongly convex, average smooth and non-convex problems (see Appendix I). It would be interesting to prove tight lower bounds in more general setting, such as F is of (σ, L)-smoothness while each f_i is (l, L)-smoothness.

References

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

Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *Journal of Machine Learning Research*, 18(1):8194–8244, 2017.

Zeyuan Allen-Zhu. Katyusha X: Practical momentum method for stochastic sum-of-nonconvex optimization. In *ICML*, 2018.

Yossi Arjevani and Ohad Shamir. Communication complexity of distributed convex learning and optimization. In *NIPS*, 2015.

Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points I. *arXiv preprint arXiv:1710.11606*, 2017.

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

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *NIPS*, 2014.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, 2013.

Guanghui Lan and Yi Zhou. An optimal randomized incremental gradient method. *Mathematical programming*, pages 1–49, 2017.

Yurii Nesterov. A method for solving the convex programming problem with convergence rate o(1/kˆ2). In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547, 1983.

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

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

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

Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

Lijun Zhang, Mehrdad Mahdavi, and Rong Jin. Linear convergence with condition number independent access of full gradients. In *NIPS*, 2013.

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

A Comparison of required number of dimensions

	Previous Lower Bounds	Our Lower Bounds
f_i is L-smooth and μ-strongly convex	$\#\text{PIFO} = \Omega\left(n + \sqrt{\kappa n}\log(\frac{1}{\varepsilon})\right)$ $d = \mathcal{O}\left(\frac{\kappa n}{\varepsilon}\log^5\left(\frac{1}{\varepsilon}\right)\right)$ (Woodworth and Srebro, 2016)	$\#\text{PIFO} = \Omega\left((n + \sqrt{\kappa n})\log(\frac{1}{\varepsilon})\right)$ $d = \mathcal{O}\left(\sqrt{\frac{\kappa}{n}}\log\left(\frac{1}{\varepsilon}\right)\right)$ [Theorem 3.1]
f_i is L-smooth and convex	$\#\text{PIFO} = \Omega\left(n + \sqrt{\frac{nL}{\varepsilon}}\right)$ $d = \mathcal{O}\left(\frac{L^2}{\varepsilon^2}\log\left(\frac{1}{\varepsilon}\right)\right)$ (Woodworth and Srebro, 2016)	$\#\text{PIFO} = \Omega\left(n + \sqrt{\frac{nL}{\varepsilon}}\right)$ $d = \mathcal{O}\left(1 + \sqrt{\frac{L}{n\varepsilon}}\right)$ [Theorem 3.3]
$\{f_i\}_{i=1}^n$ is L-average smooth and f is μ-strongly convex	$\#\text{IFO} = \Omega\left(n + n^{3/4}\sqrt{\kappa}\log\left(\frac{1}{\varepsilon}\right)\right)$ $d = \mathcal{O}\left(n + n^{3/4}\sqrt{\kappa}\log\left(\frac{1}{\varepsilon}\right)\right)$ (Zhou and Gu, 2019)	$\#\text{PIFO} = \Omega\left((n + n^{3/4}\sqrt{\kappa})\log\left(\frac{1}{\varepsilon}\right)\right)$ $d = \mathcal{O}\left(n^{-1/4}\sqrt{\kappa}\log\left(\frac{1}{\varepsilon}\right)\right)$ [Theorem 3.5]
$\{f_i\}_{i=1}^n$ is L-average smooth and f is convex	$\#\text{IFO} = \Omega\left(n + n^{3/4}\sqrt{\frac{L}{\varepsilon}}\right)$ $d = \mathcal{O}\left(n + n^{3/4}\sqrt{\frac{L}{\varepsilon}}\right)$ (Zhou and Gu, 2019)	$\#\text{PIFO} = \Omega\left(n + n^{3/4}\sqrt{\frac{L}{\varepsilon}}\right)$ $d = \mathcal{O}\left(1 + n^{-1/4}\sqrt{\frac{L}{\varepsilon}}\right)$ [Theorem 3.7]

Table 2: We compare our PIFO lower bounds with previous results, including the number of PIFO or IFO calls to obtain ε-suboptimal point and the required number of dimensions in corresponding construction.

B Detailed Proof for Section 2

In this section, we use $\|\boldsymbol{A}\|$ to denote the spectral radius of $\boldsymbol{A}$.

For simplicity, let

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

$\boldsymbol{b}_l^\top$ is the l-th row of $\boldsymbol{B}$, and $f_i(\boldsymbol{x}) = r_i(\boldsymbol{x}; \lambda_0, \lambda_1, \lambda_2, m, \omega)$.

Recall that

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

For $1 \leq i \leq n$, let $\boldsymbol{B}_i$ be a submatrix which is formed from rows $\mathcal{L}_i$ of $\boldsymbol{B}$, that is

$$\boldsymbol{B}_i = \boldsymbol{B}[\mathcal{L}_i;]$$

Then f_i can be wriiten as

$$f_i(\boldsymbol{x}) = \lambda_1 \|\boldsymbol{B}_i \boldsymbol{x}\|_2^2 + \lambda_2 \|\boldsymbol{x}\|_2^2 - \eta_i \langle \boldsymbol{e}_m, \boldsymbol{x} \rangle,$$

where $\eta_1 = \lambda_0, \eta_i = 0, i \geq 2$.

Proof of Proposition 2.5. Note that

$$\begin{aligned}\langle \boldsymbol{u}, \boldsymbol{B}_i^\top \boldsymbol{B}_i \boldsymbol{u}\rangle &= \|\boldsymbol{B}_i\boldsymbol{u}\|_2^2\\ &= \sum_{l\in\mathcal{L}_i}(\boldsymbol{b}_l^\top\boldsymbol{u})^2\\ &= \begin{cases}\sum_{l\in\mathcal{L}_i\setminus\{m\}}(u_{m-l}-u_{m-l+1})^2+\omega^2u_m^2 & (\text{if } m\in\mathcal{L}_i)\\ \sum_{l\in\mathcal{L}_i}(u_{m-l}-u_{m-l+1})^2\end{cases}\\ &\leq 2\|\boldsymbol{u}\|_2^2,\end{aligned}$$

where the last inequality is according to $(x+y)^2 \leq 2(x^2+y^2)$, and $|l_1-l_2|\geq n\geq 2$ for $l_1, l_2 \in \mathcal{L}_i$. Hence, $\left\|\boldsymbol{B}_i^\top\boldsymbol{B}_i\right\|\leq 2$, and

$$\left\|\nabla^2 f_i(\boldsymbol{x})\right\| = \left\|2\lambda_1\boldsymbol{B}_i^\top\boldsymbol{B}_i+2\lambda_2\boldsymbol{I}\right\|\leq 4\lambda_1+2\lambda_2.$$

Next, observe that

$$\|\nabla f_i(\boldsymbol{x})-\nabla f_i(\boldsymbol{y})\|_2^2 = \left\|(2\lambda_1\boldsymbol{B}_i^\top\boldsymbol{B}_i+2\lambda_2\boldsymbol{I})(\boldsymbol{x}-\boldsymbol{y})\right\|_2^2$$

Let $\boldsymbol{u}=\boldsymbol{x}-\boldsymbol{y}$.
Note that

$$\boldsymbol{b}_l\boldsymbol{b}_l^\top\boldsymbol{u} = \begin{cases}(u_{m-l}-u_{m-l+1})(\boldsymbol{e}_{m-l}-\boldsymbol{e}_{m-l+1}), & l<m,\\ \omega^2u_1\boldsymbol{e}_1, & l=m.\end{cases}$$

Thus, if $m\notin\mathcal{L}_i$, then

$$\begin{aligned}&\left\|(2\lambda_1\boldsymbol{B}_i^\top\boldsymbol{B}_i+2\lambda_2\boldsymbol{I})\boldsymbol{u}\right\|_2^2\\ =&\left\|2\lambda_1\sum_{l\in\mathcal{L}_i}(u_{m-l}-u_{m-l+1})(\boldsymbol{e}_{m-l}-\boldsymbol{e}_{m-l+1})+2\lambda_2\boldsymbol{u}\right\|_2^2\\ =&\sum_{m-l\in\mathcal{L}_i}\left[(2\lambda_1(u_l-u_{l+1})+2\lambda_2u_l)^2+(-2\lambda_1(u_l-u_{l+1})+2\lambda_2u_{l+1})^2\right]+\sum_{\substack{m-l\notin\mathcal{L}_i\\ m-l+1\notin\mathcal{L}_i}}(2\lambda_2u_l)^2\\ \leq&\sum_{m-l\in\mathcal{L}_i}8\left[(\lambda_1+\lambda_2)^2+\lambda_1^2\right](u_l^2+u_{l+1}^2)+4\lambda_2^2\|\boldsymbol{u}\|_2^2.\end{aligned}$$

Similarly, if $m\in\mathcal{L}_i$, then

$$\begin{aligned}&\left\|(2\lambda_1\boldsymbol{B}_i^\top\boldsymbol{B}_i+2\lambda_2\boldsymbol{I})\boldsymbol{u}\right\|_2^2\\ \leq&\sum_{\substack{m-l\in\mathcal{L}_i\\ l\neq 0}}8\left[(\lambda_1+\lambda_2)^2+\lambda_1^2\right](u_l^2+u_{l+1}^2)+4(\lambda_1\omega^2+\lambda_2)^2u_1^2+4\lambda_2^2\|\boldsymbol{u}\|_2^2.\end{aligned}$$

Therefore, we have

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n\|\nabla f_i(\boldsymbol{x})-\nabla f_i(\boldsymbol{y})\|_2^2\\ \leq&\frac{1}{n}\left[\sum_{l=1}^{m-1}8\left[(\lambda_1+\lambda_2)^2+\lambda_1^2\right](u_l^2+u_{l+1}^2)+4(2\lambda_1+\lambda_2)^2u_1^2\right]+4\lambda_2^2\|\boldsymbol{u}\|_2^2\\ \leq&\frac{16}{n}\left[(\lambda_1+\lambda_2)^2+\lambda_1^2\right]\|\boldsymbol{u}\|_2^2+4\lambda_2^2\|\boldsymbol{u}\|_2^2,\end{aligned}$$

where we have used $(2\lambda_1+\lambda_2)^2\leq 2\left[(\lambda_1+\lambda_2)^2+\lambda_1^2\right]$.

In summary, we get that $\{f_i\}_{1\le i\le n}$ is L'-average smooth, where

$$L' = 2\sqrt{\frac{4}{n}\left[(\lambda_1+\lambda_2)^2+\lambda_1^2\right]+\lambda_2^2}.$$

□

Proof of Lemma 2.6. For $\boldsymbol{x}\in\mathcal{F}_k$ $(k\ge 1)$, we have

$$\begin{aligned}\boldsymbol{b}_l^\top \boldsymbol{x} &= 0 \text{ for } l>k,\\ \boldsymbol{b}_l &\in \mathcal{F}_k \text{ for } l<k,\\ \boldsymbol{b}_k &\in \mathcal{F}_{k+1}.\end{aligned}$$

Consequently, for $l\ne k$, $\boldsymbol{b}_l\boldsymbol{b}_l^\top\boldsymbol{x} = (\boldsymbol{b}_l^\top\boldsymbol{x})\boldsymbol{b}_l\in\mathcal{F}_k$, and $\boldsymbol{b}_k\boldsymbol{b}_k^\top\boldsymbol{x}\in\mathcal{F}_{k+1}$.

For $k=0$, we have $\boldsymbol{x}=\boldsymbol{0}$, and

$$\begin{aligned}\nabla f_1(\boldsymbol{x}) &= \lambda_0\boldsymbol{e}_m\in\mathcal{F}_1,\\ \nabla f_j(\boldsymbol{x}) &= \boldsymbol{0}\ (j\ge 2).\end{aligned}$$

Moreover, we suppose $k\ge 1$, $k\in\mathcal{L}_i$. Since

$$\begin{aligned}\nabla f_j(\boldsymbol{x}) &= 2\lambda_1\boldsymbol{B}_j^\top\boldsymbol{B}_j\boldsymbol{x}+2\lambda_2\boldsymbol{x}-\eta_j\boldsymbol{e}_m\\ &= 2\lambda_1\sum_{l\in\mathcal{L}_j}\boldsymbol{b}_l^\top\boldsymbol{b}_l\boldsymbol{x}+2\lambda_2\boldsymbol{x}-\eta_j\boldsymbol{e}_m.\end{aligned}$$

Hence, $\nabla f_i(\boldsymbol{x})\in\mathcal{F}_{k+1}$ and $\nabla f_j(\boldsymbol{x})\in\mathcal{F}_k$ $(j\ne i)$.

Now, we turn to consider $\boldsymbol{u}=\mathrm{prox}_{f_j}^{\gamma}(\boldsymbol{x})$. We have

$$\left(2\lambda_1\boldsymbol{B}_j^\top\boldsymbol{B}_j+\left(2\lambda_2+\frac{1}{\gamma}\right)\boldsymbol{I}\right)\boldsymbol{u}=\eta_j\boldsymbol{e}_m+\frac{1}{\gamma}\boldsymbol{x},$$

i.e.,

$$\boldsymbol{u}=c_1(\boldsymbol{I}+c_2\boldsymbol{B}_j^\top\boldsymbol{B}_j)^{-1}\boldsymbol{y},$$

where $c_1=\frac{1}{2\lambda_2+1/\gamma}$, $c_2=\frac{2\lambda_1}{2\lambda_2+1/\gamma}$, and $\boldsymbol{y}=\eta_j\boldsymbol{e}_m+\frac{1}{\gamma}\boldsymbol{x}$.

Note that

$$(\boldsymbol{I}+c_2\boldsymbol{B}_j^\top\boldsymbol{B}_j)^{-1}=\boldsymbol{I}-\boldsymbol{B}_j^\top\left(\frac{1}{c_2}\boldsymbol{I}+\boldsymbol{B}_j\boldsymbol{B}_j^\top\right)^{-1}\boldsymbol{B}_j.$$

If $k=0$ and $j>1$, we have $\boldsymbol{y}=\boldsymbol{0}$ and $\boldsymbol{u}=\boldsymbol{0}$.
If $k=0$ and $j=1$, we have $\boldsymbol{y}=\lambda_0\boldsymbol{e}_m$. On this case, $\boldsymbol{B}_1\boldsymbol{e}_m=\boldsymbol{0}$, so $\boldsymbol{u}=c_1\boldsymbol{y}\in\mathcal{F}_1$.

For $k\ge 1$, we know that $\boldsymbol{y}\in\mathcal{F}_k$. And observe that if $|l-l'|\ge 2$, then $\boldsymbol{b}_l^\top\boldsymbol{b}_{l'}=0$, and consequently $\boldsymbol{B}_j\boldsymbol{B}_j^\top$ is a diagonal matrix, so we can assume that $\frac{1}{c_2}\boldsymbol{I}+\boldsymbol{B}_j\boldsymbol{B}_j^\top=\mathrm{diag}(\beta_{j,1},\cdots,\beta_{j,|\mathcal{L}_j|})$. Therefore,

$$\boldsymbol{u}=c_1\boldsymbol{y}-c_1\sum_{s=1}^{|\mathcal{L}_j|}\beta_{j,s}\boldsymbol{b}_{l_{j,s}}\boldsymbol{b}_{l_{j,s}}^\top\boldsymbol{y},$$

where we assume that $\mathcal{L}_j=\{l_{j,1},\cdots,l_{j,|\mathcal{L}_j|}\}$.

Thus, we have $\mathrm{prox}_{f_i}^{\gamma}(\boldsymbol{x})\in\mathcal{F}_{k+1}$ for $k\in\mathcal{L}_i$ and $\mathrm{prox}_{f_j}^{\gamma}(\boldsymbol{x})\in\mathcal{F}_k$ $(j\ne i)$.

□

Proof of Corollary 2.7. Denote

$$\text{span}\{\nabla f_{i_1}(\boldsymbol{x}_0), \cdots, \nabla f_{i_t}(\boldsymbol{x}_{t-1}), \text{prox}_{f_{i_1}}^{\gamma_1}(\boldsymbol{x}_0), \cdots, \text{prox}_{f_{i_t}}^{\gamma_t}(\boldsymbol{x}_{t-1})\}$$

by $\mathcal{M}_t$. We know that $\boldsymbol{x}_t \in \mathcal{M}_t$.

Suppose that $\mathcal{M}_T \subseteq \mathcal{F}_{k-1}$ for some T and let $T' = \arg\min t : t > T, i_t \equiv k(\text{mod } n)$.

By Lemma 2.6, for $T < t < T'$, we can use a simple induction to obtain that

$$\text{span}\{\nabla f_{i_t}(\boldsymbol{x}_{t-1}), \text{prox}_{f_{i_t}}^{\gamma_t}(\boldsymbol{x}_{t-1})\} \subseteq \mathcal{F}_{k-1}$$

and $\mathcal{M}_t \subseteq \mathcal{F}_{k-1}$.

Moreover, since $i_{T'} \equiv k(\text{mod } n)$, we have

$$\text{span}\{\nabla f_{i_{T'}}(\boldsymbol{x}_{T'-1}), \text{prox}_{f_{i_{T'}}}^{\gamma_{T'}}(\boldsymbol{x}_{T'-1})\} \subseteq \mathcal{F}_k$$

and $\mathcal{M}_{T'} \subseteq \mathcal{F}_k$.

Following from above statement, it is easily to check that for $t < T_k$, we have $\boldsymbol{x}_t \in \mathcal{M}_t \subseteq \mathcal{F}_{k-1}$.

Next, note that

$$\begin{aligned}
&\mathbb{P}\left(T_k - T_{k-1} = s\right) \\
&= \mathbb{P}\left(i_{T_{k-1}+1} \not\equiv k(\text{mod } n), \cdots, i_{T_{k-1}+s-1} \not\equiv k(\text{mod } n), i_{T_{k-1}+s} \equiv k(\text{mod } n)\right) \\
&= \mathbb{P}\left(i_{T_{k-1}+1} \neq k', \cdots, i_{T_{k-1}+s-1} \neq k', i_{T_{k-1}+s} = k'\right) \\
&= (1 - p_{k'})^{s-1} p_{k'},
\end{aligned}$$

where $k' \equiv k(\text{mod } n), 1 \leq k' \leq n$. So $T_k - T_{k-1}$ is a geometric random variable with success probability $p_{k'}$.

On the other hand, $T_k - T_{k-1}$ is just dependent on $i_{T_{k-1}+1}, \cdots, i_{T_k}$, thus for $l \neq k$, $T_l - T_{l-1}$ is independent with $T_k - T_{k-1}$.

Therefore,

$$T_k = \sum_{l=1}^{k} (T_l - T_{l-1}) = \sum_{i=1}^{k} Y_l,$$

where Y_l follows a geometric distribution with success probability $q_l = p_{l'}$ where $l' \equiv l(\text{mod } n), 1 \leq l' \leq n$.

□

Proof of Remark 2.3. If each f_i is L-smooth, then for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^m$ we have

$$\|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{y})\|_2^2 \leq L^2 \|\boldsymbol{x} - \boldsymbol{y}\|_2^2,$$

and consequently,

$$\frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{y})\|_2^2 \leq L^2 \|\boldsymbol{x} - \boldsymbol{y}\|_2^2. \tag{11}$$

If $\{f_i\}_{i=1}^n$ is L-average smooth, then for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^m$ we have

$$\begin{aligned}
\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2 &= \frac{1}{n^2} \left\| \sum_{i=1}^{n} (\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{y})) \right\|_2^2 \\
&\leq \frac{1}{n^2} \left(\sum_{i=1}^{n} \|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{y})\|_2 \right)^2 \\
&\leq \frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{y})\|_2^2 \\
&\leq L^2 \|\boldsymbol{x} - \boldsymbol{y}\|_2^2.
\end{aligned}$$

□

C Results about Sum of Geometric Distributed Random Variables

Lemma C.1. *Let $X_1 \sim \mathrm{Geo}(p_1), X_2 \sim \mathrm{Geo}(p_2)$ be independent random variables. For any positive integer j, if $p_1 \neq p_2$, then*

$$\mathbb{P}\left(X_1 + X_2 > j\right) = \frac{p_2(1-p_1)^j - p_1(1-p_2)^j}{p_2 - p_1}, \tag{12}$$

and if $p_1 = p_2$, then

$$\mathbb{P}\left(X_1 + X_2 > j\right) = jp_1(1-p_1)^{j-1} + (1-p_1)^j. \tag{13}$$

Proof.

$$\begin{aligned}
\mathbb{P}\left(X_1 + X_2 > j\right) &= \sum_{l=1}^{j} \mathbb{P}\left(X_1 = l\right)\mathbb{P}\left(X_2 > j-l\right) + \mathbb{P}\left(X_1 > j\right) \\
&= \sum_{l=1}^{j} (1-p_1)^{l-1} p_1 (1-p_2)^{j-l} + (1-p_1)^j \\
&= p_1(1-p_2)^{j-1} \sum_{l=1}^{j} \left(\frac{1-p_1}{1-p_2}\right)^{l-1} + (1-p_1)^j
\end{aligned}$$

Thus if $p_1 = p_2$, $\mathbb{P}\left(X_1 + X_2 > j\right) = jp_1(1-p_1)^{j-1} + (1-p_1)^j$.

For $p_1 \neq p_2$,

$$\begin{aligned}
\mathbb{P}\left(X_1 + X_2 > j\right) &= p_1 \frac{(1-p_1)^j - (1-p_2)^j}{p_2 - p_1} + (1-p_1)^j \\
&= \frac{p_2(1-p_1)^j - p_1(1-p_2)^j}{p_2 - p_1}.
\end{aligned}$$

□

Lemma C.2. *For $x \geq 0$ and $j \geq 2$,*

$$1 - \frac{j-1}{x + j/2} \leq \left(\frac{x}{x+1}\right)^{j-1}. \tag{14}$$

Proof. We just need to show that

$$(x+1)^{j-1}(x+j/2) - (j-1)(x+1)^{j-1} \leq x^{j-1}(x+j/2),$$

that is

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

Note that for $l \leq j-2$,

$$\binom{j}{l} - \frac{j}{2}\binom{j-1}{l} = \left(1 - \frac{j-l}{2}\right)\binom{j}{l} \leq 0,$$

thus inequality (14) hosts for $x \geq 0$ and $j \geq 2$. □

Lemma C.3. *Let $X_1 \sim \mathrm{Geo}(p_1), X_2 \sim \mathrm{Geo}(p_2), Y_1, Y_2 \sim \mathrm{Geo}\left(\frac{p_1+p_2}{2}\right)$ be independent random variables with $0 < p_1 \leq p_2 \leq 1$. Then for any positive integer j, we have*

$$\mathbb{P}\left(X_1 + X_2 > j\right) \geq \mathbb{P}\left(Y_1 + Y_2 > j\right).$$

Proof. If $j=1$, then $\mathbb{P}\left(X_1+X_2>j\right)=1=\mathbb{P}\left(Y_1+Y_2>j\right)$.
If $p_1=p_2=1$, then $\mathbb{P}\left(X_1+X_2>j\right)=0=\mathbb{P}\left(Y_1+Y_2>j\right)$ for $j\geq 2$.

Let $j\geq 2$, and $c\triangleq p_1+p_2<2$ be a given constant.

We prove that $f(p_1)\triangleq\mathbb{P}\left(X_1+X_2>j\right)$ is a decreasing function.

Employing equation (12), for $p_1<c/2$, we have

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

and

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

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

$$\frac{c(1-p_1)-j(c-p_1)(c-2p_1)}{c(1+p_1-c)+jp_1(c-2p_1)}<\left(\frac{1+p_1-c}{1-p_1}\right)^{j-1}. \tag{15}$$

Note that

$$\begin{aligned}&\frac{c(1-p_1)-j(c-p_1)(c-2p_1)}{c(1+p_1-c)+jp_1(c-2p_1)}\\&=1-\frac{(j-1)c(c-2p_1)}{c(1+p_1-c)+jp_1(c-2p_1)}\\&=1-\frac{j-1}{\frac{1+p_1-c}{c-2p_1}+j\frac{p_1}{c}}\end{aligned}$$

Denote $x=\frac{1+p_1-c}{c-2p_1}$. If $c\leq 1$, then $p_1>0$ and $x>\frac{1-c}{c}\geq 0$. And if $c>1$, then $p_1\geq c-1$ and $x\geq\frac{1+c-1-c}{2-c}=0$.
Rewrite inequality (15) as

$$1-\frac{j-1}{x+jp_1/c}<\left(\frac{x}{x+1}\right)^{j-1}.$$

Recall inequality (14), we have

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

Consequently, $f'(p_1)<0$ hosts for $p_1<c/2$ and $j\geq 2$.
With the fact that $\lim_{p_1\to c/2}f(p_1)=f(c/2)$ according to equation (13), we have

$$\mathbb{P}\left(X_1+X_2>j\right)\geq\mathbb{P}\left(Y_1+Y_2>j\right).$$

for any positive integer j and $0<p_1\leq p_2\leq 1$. □

Corollary C.4. *Let $X_1\sim\mathrm{Geo}(p_1), X_2\sim\mathrm{Geo}(p_2), Y_1, Y_2\sim\mathrm{Geo}\left(\frac{p_1+p_2}{2}\right)$ be independent random variables with $0<p_1\leq p_2\leq 1$. Suppose Z is a random variable that takes nonnegative integer values, and Z is independent with X_1, X_2, Y_1, Y_2. Then for any positive integer j, we have*

$$\mathbb{P}\left(Z+X_1+X_2>j\right)\geq\mathbb{P}\left(Z+Y_1+Y_2>j\right).$$

Proof. With applying Lemma C.3, we have

$$\begin{aligned}\mathbb{P}\left(Z + X_1 + X_2 > j\right) &= \sum_{l=0}^{j-1} \mathbb{P}\left(Z = l\right) \mathbb{P}\left(X_1 + X_2 > l - j\right) + \mathbb{P}\left(Z > j - 1\right) \\ &\geq \sum_{l=0}^{j-1} \mathbb{P}\left(Z = l\right) \mathbb{P}\left(Y_1 + Y_2 > l - j\right) + \mathbb{P}\left(Z > j - 1\right) \\ &= \mathbb{P}\left(Z + Y_1 + Y_2 > j\right).\end{aligned}$$

□

Corollary C.5. *Let $\{X_i\}_{1\leq i\leq m}$ be independent variables, and X_i follow a geometric distribution with success probability p_i. For any positive integer j, we have*

$$\mathbb{P}\left(\sum_{i=1}^{m} X_i \geq j\right) \geq \mathbb{P}\left(\sum_{i=1}^{m} Y_i \geq j\right),$$

where $\{Y_i\}_{1\leq i\leq m}$ are i.i.d. random variables, $Y_i \sim \mathrm{Geo}(\sum_{i=1}^{m} p_i/m)$, and Y_i is independent with $X_{i'}(1 \leq i' \leq m)$.

Proof. Let

$$f(p_1, p_2, \cdots, p_m) \triangleq \mathbb{P}\left(\sum_{i=1}^{m} X_i \geq j\right).$$

Our goal is to minimize $f(p_1, p_2, \cdots, p_m)$ such that $\sum_{i=1}^{m} p_i = S < 1$.

By Corollary C.4, we know that

$$f(p_1, p_2, \cdots, p_i, \cdots, p_j, \cdots, p_m) \geq f(p_1, p_2, \cdots, \frac{p_i + p_j}{2}, \cdots, \frac{p_i + p_j}{2}, \cdots, p_m).$$

This fact implies that $(p_1, p_2, \cdots, p_m)$ such that $p_1 = p_2 = \cdots = p_m = S/m$ is a minimizer of the function f.

□

Lemma C.6. *Let $\{X_i\}_{1\leq i\leq m}$ be i.i.d. random variables, and X_i follows a geometric distribution with success probability p. We have*

$$\mathbb{P}\left(\sum_{i=1}^{m} X_i > \frac{m}{4p}\right) \geq 1 - \frac{16}{9m} \tag{16}$$

Proof. Denote $\sum_{i=1}^{m} X_i$ by τ. We know that

$$\mathbb{E}\tau = \frac{m}{p}, \;\; \mathrm{Var}(\tau) = \frac{m(1-p)}{p^2}.$$

Hence, we have

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

□

Corollary C.7. *Let $\{X_i\}_{1 \leq i \leq m}$ be independent random variables, and X_i follows a geometric distribution with success probability p_i. Then*

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

D PROOF OF THEOREM 4.2

Proposition D.1. *For any $n \geq 2$, $m \geq 2$, $f_{SC,i}$ and F_{SC} in Definition 4.1 satisfy:*

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

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x} \in \mathbb{R}^m} F_{SC}(\boldsymbol{x}) = \sqrt{\frac{2\Delta n(\alpha+1)^2}{(L-\mu)(\alpha-1)}}(q^m, q^{m-1}, \cdots, q)^\top,$$

   where $q = \frac{\alpha-1}{\alpha+1}$. Moreover, $F_{SC}(\boldsymbol{x}^*) = -\Delta$.

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

$$\min_{\boldsymbol{x} \in \mathcal{F}_k} F_{SC}(\boldsymbol{x}) - F_{SC}(\boldsymbol{x}^*) \geq \Delta q^{2k}. \tag{17}$$

Proof.

1. Just recall Proposition 2.5.
2. Denote $\xi = \sqrt{\frac{2\Delta n(\alpha+1)^2}{(L-\mu)(\alpha-1)}}$.

   Let $\nabla F_{\mathrm{SC}}(\boldsymbol{x}) = 0$, that is

$$\left(\frac{L-\mu}{2n}\boldsymbol{A}\left(\sqrt{\frac{2}{\alpha+1}}\right) + \mu \boldsymbol{I}\right)\boldsymbol{x} = \frac{L-\mu}{n(\alpha+1)}\xi \boldsymbol{e}_m,$$

   or

$$\begin{bmatrix} \omega^2 + 1 + \frac{2n\mu}{L-\mu} & -1 & & & \\ -1 & 2 + \frac{2n\mu}{L-\mu} & -1 & & \\ & \ddots & \ddots & & \\ & & -1 & 2 + \frac{2n\mu}{L-\mu} & -1 \\ & & & -1 & 1 + \frac{2n\mu}{L-\mu} \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \frac{2\xi}{\alpha+1} \end{bmatrix} \tag{18}$$

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

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

   and

$$\omega^2 + 1 + \frac{2n\mu}{L-\mu} = \frac{1}{q},$$
$$\frac{2}{\alpha+1} = 1 - q = -q^2 + (1 + \frac{2n\mu}{L-\mu})q.$$

   Hence, it is easily to check that the solution to Equation (18) is

$$\boldsymbol{x}^* = \xi(q^m, q^{m-1}, \cdots, q)^\top,$$

   and

$$F_{\mathrm{SC}}(\boldsymbol{x}^*) = -\frac{L-\mu}{2n(\alpha+1)}\xi^2 q = -\Delta.$$

3. If $\boldsymbol{x} \in \mathcal{F}_k$, $1 \leq k < m$, then $x_1 = x_2 = \cdots = x_{m-k} = 0$.

   Let $\boldsymbol{y} = \boldsymbol{x}_{m-k+1:m} \in \mathbb{R}^k$ and $\boldsymbol{A}_k$ be last k rows and columns of the matrix in Equation (19). Then we can rewrite $F(\boldsymbol{x})$ as
$$F_k(\boldsymbol{y}) \triangleq F_{\text{SC}}(\boldsymbol{x}) = \frac{L-\mu}{4n}\boldsymbol{y}^\top \boldsymbol{A}_k \boldsymbol{y} - \frac{L-\mu}{n(\alpha+1)}\xi\langle \boldsymbol{e}_m, \boldsymbol{y}\rangle.$$
   Let $\nabla F_k(\boldsymbol{y}) = 0$, that is
$$\begin{bmatrix} 2+\frac{2n\mu}{L-\mu} & -1 & & & \\ -1 & 2+\frac{2n\mu}{L-\mu} & -1 & & \\ & \ddots & \ddots & & \\ & & -1 & 2+\frac{2n\mu}{L-\mu} & -1 \\ & & & -1 & 1+\frac{2n\mu}{L-\mu} \end{bmatrix} \boldsymbol{y} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \frac{2\xi}{\alpha+1} \end{bmatrix}. \tag{19}$$
   By some calculation, the solution to above equation is
$$\frac{\xi q^{k+1}}{1+q^{2k+1}}\left(q^{-1}-q, q^{-2}-q^2, \cdots, q^{-k}-q^k\right)^\top.$$
   Thus
$$\min_{\boldsymbol{x}\in\mathcal{F}_k} F_{\text{SC}}(\boldsymbol{x}) = \min_{\boldsymbol{y}\in\mathbb{R}^k} F_k(\boldsymbol{y}) = -\frac{L-\mu}{2n(\alpha+1)}\xi^2 q \frac{1-q^{2k}}{1+q^{2k+1}} = \Delta\frac{1-q^{2k}}{1+q^{2k+1}},$$
   and
$$\begin{aligned}\min_{\boldsymbol{x}\in\mathcal{F}_k} F_{\text{SC}}(\boldsymbol{x}) - F_{\text{SC}}(\boldsymbol{x}^*) &= \Delta\left(1 - \frac{1-q^{2k}}{1+q^{2k+1}}\right) \\ &= \Delta q^{2k}\frac{1+q}{1+q^{2k+1}} \\ &\geq \Delta q^{2k}.\end{aligned}$$

□

Proof of Theorem 4.2. Let $M = \left\lfloor \frac{\log(9\varepsilon/\Delta)}{2\log q}\right\rfloor$, then we have
$$\arg\min_{\boldsymbol{x}\in\mathcal{F}_M} F_{\text{SC}}(\boldsymbol{x}) - F_{\text{SC}}(\boldsymbol{x}^*) \geq \Delta q^{2M} \geq 9\varepsilon,$$
where the first inequality is according to the third property of Proposition D.1.

Following from Lemma 2.9, for $M \geq 1$ and $N = (M+1)n/4$, we have
$$\min_{t\leq N} \mathbb{E}F_{\text{SC}}(\boldsymbol{x}_t) - F_{\text{SC}}(\boldsymbol{x}^*) \geq \varepsilon.$$
Therefore, in order to find $\hat{\boldsymbol{x}} \in \mathbb{R}^m$ such that $\mathbb{E}F_{\text{SC}}(\hat{\boldsymbol{x}}) - F_{\text{SC}}(\boldsymbol{x}^*) < \varepsilon$, $\mathcal{A}$ needs at least N queries to $h_{F_{\text{SC}}}$.

Next, observe that function $h(\beta) = \frac{1}{\log\left(\frac{\beta+1}{\beta-1}\right)} - \frac{\beta}{2}$ is increasing when $\beta > 1$ and $L/\mu \geq n/2+1$, $\alpha = \sqrt{2\frac{L/\mu-1}{n}+1} \geq \sqrt{2}$. Thus, we have
$$\begin{aligned}-\frac{1}{\log(q)} = \frac{1}{\log\left(\frac{\alpha+1}{\alpha-1}\right)} &\geq \frac{\alpha}{2} + h(\sqrt{2}) \\ &= \frac{1}{2}\sqrt{2\frac{L/\mu-1}{n}+1} + h(\sqrt{2}) \\ &\geq \frac{\sqrt{2}}{4}\left(\sqrt{2\frac{L/\mu-1}{n}}+1\right) + h(\sqrt{2}) \\ &\geq \frac{1}{2}\sqrt{\frac{L/\mu-1}{n}} + \frac{\sqrt{2}}{4} + h(\sqrt{2}),\end{aligned}$$

and

$$
\begin{aligned}
N = (M+1)n/4 &= \frac{n}{4}\left(\left\lfloor \frac{\log(9\varepsilon/\Delta)}{2\log q}\right\rfloor + 1\right) \\
&\geq \frac{n}{8}\left(-\frac{1}{\log(q)}\right)\log\left(\frac{\Delta}{9\varepsilon}\right) \\
&\geq \frac{n}{8}\left(\frac{1}{2}\sqrt{\frac{L/\mu-1}{n}} + \frac{\sqrt{2}}{4} + h(\sqrt{2})\right)\log\left(\frac{\Delta}{9\varepsilon}\right) \\
&= \Omega\left(\left(n + \sqrt{\frac{nL}{\mu}}\right)\log\left(\frac{\Delta}{9\varepsilon}\right)\right)
\end{aligned}
$$

At last, we must to ensure that $1 \leq M < m$, that is

$$
1 \leq \frac{\log(9\varepsilon/\Delta)}{2\log q} < m. \tag{20}
$$

Note that $\lim_{\beta\to+\infty} h(\beta) = 0$, so $-1/\log(q) \leq \alpha/2$. Thus the above conditions are satisfied when

$$
m = \frac{\log(\Delta/(9\varepsilon))}{2(-\log q)} + 1 \leq \frac{1}{4}\left(\sqrt{2\frac{L/\mu-1}{n}} + 1\right)\log\left(\frac{\Delta}{9\varepsilon}\right) + 1 = \mathcal{O}\left(\sqrt{\frac{L}{n\mu}}\log\left(\frac{\Delta}{\varepsilon}\right)\right),
$$

and

$$
\frac{\varepsilon}{\Delta} \leq \frac{1}{9}\left(\frac{\sqrt{2}-1}{\sqrt{2}+1}\right)^2 \leq \frac{1}{9}\left(\frac{\alpha-1}{\alpha+1}\right)^2.
$$

□

E Proof of Theorem 4.3

Proof of Theorem 4.3. Denote $\xi = \sqrt{\frac{2\Delta n(\alpha+1)^2}{(L-\mu)(\alpha-1)}}$, and $M = \left\lfloor \frac{\log(18\varepsilon)}{2\log q}\right\rfloor$.
For $1 \leq M \leq m/2$, $N = n(M+1)/4$ and $t \leq N$, we have

$$
\begin{aligned}
\mathbb{E}\,\|\boldsymbol{x}_t - \boldsymbol{x}^*\|_2^2 &\geq \mathbb{E}\left[\|\boldsymbol{x}_t - \boldsymbol{x}^*\|_2^2 \,\middle|\, N < T_{M+1}\right]\mathbb{P}\left(N < T_{M+1}\right) \\
&\geq \mathbb{E}\left[\min_{\boldsymbol{x}\in\mathcal{F}_M}\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 \,\middle|\, N < T_{M+1}\right]\mathbb{P}\left(N < T_{M+1}\right) \\
&\geq \frac{1}{9}\min_{\boldsymbol{x}\in\mathcal{F}_M}\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2.
\end{aligned}
$$

where T_{M+1} is defined in (7), the second inequality follows from Corollary 2.7 (if $N < T_{M+1}$, then $\boldsymbol{x}_t \in \mathcal{F}_M$ for $t \leq N$), and the last inequality is established because of our Corollary 2.7 (More detailed explanation refer to our proof of Lemma 2.9).

By Proposition D.1, we know that $\boldsymbol{x}^* = \xi(q^m, q^{m-1}, \cdots, q)^\top$, and

$$
\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2 = \|\boldsymbol{x}^*\|_2^2 = \xi^2\frac{q^2 - q^{2(m+1)}}{1-q^2}.
$$

Note that if $\boldsymbol{x} \in \mathcal{F}_M$, then $x_1 = x_2 = \cdots = x_{m-M} = 0$, thus

$$
\min_{\boldsymbol{x}\in\mathcal{F}_M}\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 = \xi^2\sum_{l=m-M}^{m} q^{2(m-l+1)} = \xi^2\frac{q^{2(M+1)} - q^{2(m+1)}}{1-q^2}.
$$

Thus, for $t \leq N$ and $M \leq m/2$, we have

$$\begin{aligned}\frac{\mathbb{E}\left\|\boldsymbol{x}_t-\boldsymbol{x}^*\right\|_2^2}{\left\|\boldsymbol{x}_t-\boldsymbol{x}^*\right\|_2^2} &\geq \frac{1}{9}\frac{q^{2M}-q^{2m}}{1-q^{2m}} \\ &\geq \frac{1}{18}q^{2M}=\frac{1}{18}q^{2\left\lfloor\frac{\log(18\varepsilon)}{2\log q}\right\rfloor}\geq\varepsilon,\end{aligned}$$

where the second inequality is due to

$$\begin{aligned}\frac{q^{2M}-q^{2m}}{1-q^{2m}}-\frac{q^{2M}}{2} &= \frac{q^{2M}-2q^{2m}+q^{2(m+M)}}{2(1-q^{2m})} \\ &= \frac{q^{2M}}{2(1-q^{2m})}(1-2q^{2(m-M)}+q^{2m}) \\ &\geq \frac{q^{2M}}{2(1-q^{2m})}(1-2q^{m}+q^{2m})\geq 0.\end{aligned}$$

Therefore, in order to find $\hat{\boldsymbol{x}}\in\mathbb{R}^m$ such that $\frac{\mathbb{E}\|\hat{\boldsymbol{x}}-\boldsymbol{x}^*\|_2^2}{\|\boldsymbol{x}_0-\boldsymbol{x}^*\|_2^2}<\varepsilon$, $\mathcal{A}$ needs at least N queries to $h_{F_{\text{SC}}}$.

As we have showed in proof of Theorem 4.2, for $L/\mu\geq n/2+1$, we have

$$\frac{1}{2}\sqrt{2\frac{L/\mu-1}{n}+1}\geq-\frac{1}{\log(q)}\geq c_1\left(\sqrt{\frac{L/\mu-1}{n}}+1\right),$$

and

$$\begin{aligned}N=\frac{n}{4}(M+1) &\geq \frac{n}{4}\frac{\log(18\varepsilon)}{2\log q} \\ &\geq \frac{c_1}{8}\left(n+\sqrt{n(L/\mu-1)}\right)\log\left(\frac{1}{18\varepsilon}\right) \\ &= \Omega\left(\left(n+\sqrt{\frac{nL}{\mu}}\right)\log\left(\frac{1}{\varepsilon}\right)\right).\end{aligned}$$

At last, we have to ensure that $1\leq M\leq m/2$, that is

$$1\leq\frac{\log(18\varepsilon)}{2\log q}<m/2.$$

The above conditions are satisfied when

$$m=\frac{\log(1/(18\varepsilon))}{-\log q}+1\leq\frac{1}{2}\left(\sqrt{2\frac{L/\mu-1}{n}+1}\right)\log\left(\frac{1}{18\varepsilon}\right)+1=\mathcal{O}\left(\sqrt{\frac{L}{n\mu}}\log\left(\frac{1}{\varepsilon}\right)\right),$$

and

$$\varepsilon\leq\frac{1}{18}q^2.$$

Observe that when $L/\mu\leq n/2+1$, we have $\alpha\geq\sqrt{2}$ and $q=\frac{\alpha-1}{\alpha+1}\geq\frac{\sqrt{2}-1}{\sqrt{2}+1}$. Hence, we just need $\varepsilon\leq\frac{1}{18}\left(\frac{\sqrt{2}-1}{\sqrt{2}+1}\right)^2\approx 0.00164$.

□

F Proof of Theorem 4.5

Proposition F.1. *For any $n\geq 2$, $m\geq 2$, following properties hold:*

1. *$f_{C,i}$ is L-smooth and convex.*
2. *The minimizer of the function F_C is*

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}\in\mathbb{R}^m} F_C(\boldsymbol{x}) = \frac{2\xi}{L}\left(1, 2, \cdots, m\right)^\top,$$

   where $\xi = \frac{\sqrt{3}}{2}\frac{BL}{(m+1)^{3/2}}$. Moreover, $F_C(\boldsymbol{x}^*) = -\frac{m\xi^2}{nL}$ and $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2 \le B^2$.

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

$$\min_{\boldsymbol{x}\in\mathcal{F}_k} F_C(\boldsymbol{x}) - F_C(\boldsymbol{x}^*) = \frac{\xi^2}{nL}(m-k). \tag{21}$$

Proof.

1. Just recall Proposition 2.5.
2. Denote $\xi = \frac{\sqrt{3}}{2}\frac{BL}{(m+1)^{3/2}n}$. Let $\nabla F_{\mathrm{C}}(\boldsymbol{x}) = 0$, that is

$$\frac{L}{2n}\boldsymbol{A}(1)\boldsymbol{x} = \frac{\xi}{n}\boldsymbol{e}_m,$$

   or

$$\begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{bmatrix}\boldsymbol{x} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \frac{2\xi}{L} \end{bmatrix}. \tag{22}$$

   Hence, it is easily to check that the solution to Equation (22) is

$$\boldsymbol{x}^* = \frac{2\xi}{L}(1, 2, \cdots, m)^\top,$$

   and

$$F_{\mathrm{C}}(\boldsymbol{x}^*) = -\frac{m\xi^2}{nL}.$$

   Moreover, we have

$$\begin{aligned}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2 &= \frac{4\xi^2}{L^2}\frac{m(m+1)(2m+1)}{6} \\ &\le \frac{4\xi^2}{3L^2}(m+1)^3 = B^2.\end{aligned}$$

3. By similar calculation to above proof, we have

$$\arg\min_{\boldsymbol{x}\in\mathcal{F}_k} F_{\mathrm{C}}(\boldsymbol{x}) = \frac{2\xi}{L}(1, 2, \cdots, k)^\top,$$

   and

$$\min_{\boldsymbol{x}\in\mathcal{F}_k} F_{\mathrm{C}}(\boldsymbol{x}) = -\frac{k\xi^2}{nL}.$$

   Thus

$$\min_{\boldsymbol{x}\in\mathcal{F}_k} F_{\mathrm{C}}(\boldsymbol{x}) - F_{\mathrm{C}}(\boldsymbol{x}^*) = \frac{\xi^2}{nL}(m-k).$$

□

Proof of Theorem 4.5. Since $\varepsilon \leq \frac{B^2 L}{384n}$, we have $m \geq 3$. Let $\xi = \frac{\sqrt{3}}{2} \frac{BL}{(m+1)^{3/2}}$.

For $M = \left\lfloor \frac{m-1}{2} \right\rfloor \geq 1$, we have $m - M \geq (m+1)/2$, and

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathcal{F}_M} F_{\mathrm{C}}(\boldsymbol{x}) - F_{\mathrm{C}}(\boldsymbol{x}^*) &= \frac{\xi^2}{nL}(m - M) = \frac{3B^2 L}{4n} \frac{m - M}{(m+1)^3} \\ &\geq \frac{3B^2 L}{8n} \frac{1}{(m+1)^2} \geq 9\varepsilon, \end{aligned}$$

where the first equation is according to the 3rd property in Proposition F.1 and the last inequality follows from $m + 1 \leq B\sqrt{L/(24n\varepsilon)}$.

Similar to the proof of Theorem 4.2, by Lemma 2.9, we have

$$\min_{t \leq N} \mathbb{E} F_{\mathrm{C}}(\boldsymbol{x}_t) - F_{\mathrm{C}}(\boldsymbol{x}^*) \geq \varepsilon.$$

In other words, in order to find $\hat{\boldsymbol{x}} \in \mathbb{R}^m$ such that $\mathbb{E} F_{\mathrm{C}}(\hat{\boldsymbol{x}}) - F_{\mathrm{C}}(\boldsymbol{x}^*) < \varepsilon$, $\mathcal{A}$ needs at least N queries to h_F.

At last, observe that

$$\begin{aligned} N = (M+1)n/4 &= \frac{n}{4} \left\lfloor \frac{m+1}{2} \right\rfloor \\ &\geq \frac{n(m-1)}{8} \\ &\geq \frac{n}{8} \left(\sqrt{\frac{B^2 L}{24n\varepsilon}} - 2 \right) \\ &= \Omega\left(n + B\sqrt{\frac{nL}{\varepsilon}} \right), \end{aligned}$$

where we have recalled $\varepsilon \leq \frac{B^2 L}{384n}$ in last equation.

□

G Proof of Lemma 4.6 and Lemma 4.10

Proof of Lemma 4.6. Consider the following functions $\{g_i\}_{1 \leq i \leq n}$, $g_i : \mathbb{R} \to \mathbb{R}$, where

$$\begin{aligned} g_1(x) &= \frac{L}{2}x^2 - nLBx, \\ g_i(x) &= \frac{L}{2}x^2, \\ G(x) &= \frac{1}{n}\sum_{i=1}^{n} g_i(x) = \frac{L}{2}x^2 - LBx. \end{aligned}$$

First observe that

$$\begin{aligned} x^* &= \arg\min_{x \in \mathbb{R}} G(x) = B, \\ G(0) - G(x^*) &= \frac{LB^2}{2}, \end{aligned}$$

and $|x_0 - x^*| = B$.

For $i > 1$, we have $\frac{dg_i(x)}{dx}|_{x=0} = 0$ and $\mathrm{prox}_{g_i}^{\gamma}(0) = 0$. Thus $x_t = 0$ will host till our first-order method $\mathcal{A}$ draws the component f_1. That is, for $t < T = \arg\min\{t : i_t = 1\}$, we have $x_t = 0$.

Hence, for $t \leq \frac{1}{2p_1}$, we have

$$\begin{aligned}\mathbb{E}G(x_t) - F(x^*) &\geq \mathbb{E}\left[G(x_t) - G(x^*)\Big|\frac{1}{2p_1} < T\right]\mathbb{P}\left(\frac{1}{2p_1} < T\right)\\ &= \frac{LB^2}{2}\mathbb{P}\left(\frac{1}{2p_1} < T\right).\end{aligned}$$

Note that T follows a geometric distribution with success probability $p_1 \leq 1/n$, and

$$\begin{aligned}\mathbb{P}\left(T > \frac{1}{2p_1}\right) = \mathbb{P}\left(T > \left\lfloor\frac{1}{2p_1}\right\rfloor\right) &= (1-p_1)^{\left\lfloor\frac{1}{2p_1}\right\rfloor}\\ &\geq (1-p_1)^{\frac{1}{2p_1}} \geq (1-1/n)^{n/2} \geq \frac{1}{2},\end{aligned}$$

where the second inequality follows from $h(z) = \frac{\log(1-z)}{2z}$ is a decreasing function.

Thus, for $t \leq \frac{1}{2p_1}$, we have

$$\mathbb{E}G(x_t) - F(x^*) \geq \frac{LB^2}{4} \geq \varepsilon$$

Thus, in order to find $\hat{x} \in \mathbb{R}$ such that $\mathbb{E}F(\hat{x}) - F(x^*) < \varepsilon$, $\mathcal{A}$ needs at least $\frac{1}{2p_1} \geq n/2 = \Omega(n)$ queries to h_G.

□

Proof of Lemma 4.10. Note that $\{g_i\}_{i=1}^n$ defined in above proof is also L-average smooth, so Lemma 4.10 hosts for the same reason. □

H Detailed Proof for Section 4.3

Proof of Proposition 4.7.

1. It is easily to check that $F_{\text{SC}}(\boldsymbol{x})$ is μ-strongly convex. Following from Proposition 2.5, then $\{f_{\text{SC},i}\}_{i=1}^n$ is $\hat{L}$-average smooth, where

$$\begin{aligned}\hat{L} &= \sqrt{\frac{16}{n}\left[\left(\frac{L+\mu}{4}\right)^2 + \left(\frac{L-\mu}{4}\right)^2\right] + \mu^2}\\ &= \sqrt{\frac{2(L^2+\mu^2)}{n} + \mu^2} = L'.\end{aligned}$$

2. Clearly, $L = \sqrt{\frac{n(L'^2-\mu^2)}{2} - \mu^2} \leq \sqrt{\frac{n}{2}}L'$.
   Furthermore, according to $\frac{L'}{\mu} \geq \sqrt{\frac{3}{n}}(\frac{n}{2}+1)$, we have

$$\begin{aligned}L^2 - \frac{n}{3}L'^2 &= \frac{n}{2}(L'^2-\mu^2) - \mu^2 - \frac{n}{3}L'^2\\ &= \frac{1}{2}\left(\frac{n}{2}+1\right)^2\mu^2 - \frac{n+2}{2}\mu^2\\ &= \left(\frac{n^2}{8} - \frac{1}{2}\right)\mu^2 \geq 0,\end{aligned}$$

   and, $L/\mu \geq \sqrt{\frac{n}{3}}L'/\mu \geq n/2+1$.

□

Proof of Theorem 4.8. By 2nd property of Proposition 4.7, we know that $L/\mu \geq n/2 + 1$. Moreover,

$$\begin{aligned} m &= \frac{1}{4}\left(\sqrt{\sqrt{\frac{2}{n}}\frac{L'}{\mu} + 1}\right)\log\left(\frac{\Delta}{9\varepsilon}\right) + 1 \\ &\geq \frac{1}{4}\left(\sqrt{2\frac{L/\mu - 1}{n} + 1}\right)\log\left(\frac{\Delta}{9\varepsilon}\right) + 1, \end{aligned}$$

Then, by Theorem 4.2 [4], in order to find $\hat{\boldsymbol{x}} \in \mathbb{R}^m$ such that $\mathbb{E}F_{\text{SC}}(\hat{\boldsymbol{x}}) - F_{\text{SC}}(\boldsymbol{x}^*) < \varepsilon$, $\mathcal{A}$ needs at least N queries to $h_{F_{\text{SC}}}$, where

$$\begin{aligned} N &= \Omega\left(\left(n + \sqrt{\frac{nL}{\mu}}\right)\log\left(\frac{\Delta}{\varepsilon}\right)\right) \\ &= \Omega\left(\left(n + \sqrt{\frac{n\sqrt{n/3}L'}{\mu}}\right)\log\left(\frac{\Delta}{\varepsilon}\right)\right) \\ &= \Omega\left(\left(n + n^{3/4}\sqrt{\frac{L'}{\mu}}\right)\log\left(\frac{\Delta}{\varepsilon}\right)\right). \end{aligned}$$

□

Proof of Theorem 4.9. Note that

$$\begin{aligned} \varepsilon &\leq \frac{\sqrt{2}}{768}\frac{B^2 L'}{\sqrt{n}} = \frac{B^2 L}{384n}, \\ m &= \left\lfloor \frac{\sqrt[4]{18}}{12} B n^{-1/4}\sqrt{\frac{L'}{\varepsilon}} \right\rfloor - 1 = \left\lfloor \sqrt{\frac{B^2 L}{24n\varepsilon}} \right\rfloor - 1. \end{aligned}$$

By Theorem 4.5, in order to find $\hat{\boldsymbol{x}} \in \mathbb{R}^m$ such that $\mathbb{E}F_{\text{C}}(\hat{\boldsymbol{x}}) - F_{\text{C}}(\boldsymbol{x}^*) < \varepsilon$, $\mathcal{A}$ needs at least N queries to $h_{F_{\text{C}}}$, where

$$\begin{aligned} N &= \Omega\left(n + B\sqrt{\frac{nL}{\varepsilon}}\right) \\ &= \Omega\left(n + B\sqrt{\frac{n\sqrt{n/2}L'}{\varepsilon}}\right) \\ &= \Omega\left(n + Bn^{3/4}\sqrt{\frac{L'}{\varepsilon}}\right). \end{aligned}$$

□

I NON-CONVEX CASE

In non-convex case, our goal is to find an ε-approximate stationary point $\hat{\boldsymbol{x}}$ of our objective function f, which satisfies

$$\|\nabla f(\hat{\boldsymbol{x}})\|_2 \leq \varepsilon. \tag{23}$$

[4]By the proof of Theorem 4.2, a larger dimension m does not affect the conclusion of the theorem.

I.1 PRELIMINARIES

We first introduce a general concept about smoothness.

Definition I.1. *For any differentiable function $f : \mathbb{R}^{m+1} \to \mathbb{R}$, we say f is (l, L)-smooth, if for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^m$ we have*

$$\frac{l}{2} \left\| \boldsymbol{x} - \boldsymbol{y} \right\|_2^2 \leq f(\boldsymbol{x}) - f(\boldsymbol{y}) - \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \leq \frac{L}{2} \left\| \boldsymbol{x} - \boldsymbol{y} \right\|_2^2,$$

where $L > 0, l \in \mathbb{R}$.

Especially, if f is L-smooth, then it can be checked that f is $(-L, L)$-smooth.

If f is $(-\sigma, L)$-smooth, in order to make the operator $\operatorname{prox}_f^\gamma$ valid, we set $\frac{1}{\gamma} > \sigma$ to ensure the function

$$\hat{f}(\boldsymbol{u}) \triangleq f(\boldsymbol{u}) + \frac{1}{2\gamma} \left\| \boldsymbol{x} - \boldsymbol{u} \right\|_2^2$$

is a convex function.

Next, we introduce a class of function which is original proposed in (Carmon et al., 2017). Let $G_{\text{NC}} : \mathbb{R}^{m+1} \to \mathbb{R}$ be

$$G_{\text{NC}}(\boldsymbol{x}; \alpha, m) = \frac{1}{2} \left\| \boldsymbol{B}(m+1, \sqrt[4]{\alpha}) \boldsymbol{x} \right\|_2^2 - \sqrt{\alpha} \langle \boldsymbol{e}_1, \boldsymbol{x} \rangle + \alpha \sum_{i=1}^{m} \Gamma(x_i),$$

where the non-convex function $\Gamma : \mathbb{R} \to \mathbb{R}$ is

$$\Gamma(x) \triangleq 120 \int_1^x \frac{t^2(t-1)}{1+t^2} dt. \tag{24}$$

We need following properties about $G_{\text{NC}}(\boldsymbol{x}; \alpha, m)$.

Proposition I.2 (Lemmas 3,4, Carmon et al. (2017))**.** *For any $0 < \alpha \leq 1$, it holds that*

1. *$\Gamma(x)$ is $(-45(\sqrt{3}-1), 180)$-smooth and $G_{NC}(\boldsymbol{x}; \alpha, m)$ is $(-45(\sqrt{3}-1)\alpha, 4 + 180\alpha)$-smooth.*
2. *$G_{NC}(\boldsymbol{0}; \alpha, m) - \min_{\boldsymbol{x} \in \mathbb{R}^{m+1}} G_{NC}(\boldsymbol{x}; \alpha, m) \leq \sqrt{\alpha}/2 + 10\alpha m$.*
3. *For $\boldsymbol{x}$ which satisfies that $x_m = x_{m+1} = 0$, we have*

$$\left\| \nabla G_{NC}(\boldsymbol{x}; \alpha, m) \right\|_2 \geq \alpha^{3/4}/4.$$

I.2 OUR RESULT

Theorem I.3. *For any PIFO algorithm $\mathcal{A}$ and any $L, \sigma, n, \Delta, \varepsilon$ such that $\varepsilon^2 \leq \frac{\Delta L \alpha}{81648n}$, there exist a dimension $d = \left\lfloor \frac{\Delta L \sqrt{\alpha}}{40824 n \varepsilon^2} \right\rfloor + 1$ and n $(-\sigma, L)$-smooth nonconvex functions $\{f_i : \mathbb{R}^d \to \mathbb{R}\}_{i=1}^n$ such that $f(\boldsymbol{x}_0) - f(\boldsymbol{x}^*) \leq \Delta$. In order to find $\hat{\boldsymbol{x}} \in \mathbb{R}^d$ such that $\mathbb{E} \left\| \nabla f(\hat{\boldsymbol{x}}) \right\|_2 < \varepsilon$, $\mathcal{A}$ needs at least $\Omega\left(\frac{\Delta L \sqrt{\alpha}}{\varepsilon^2}\right)$ queries to h_f, where we set $\alpha = \min\left\{1, \frac{(\sqrt{3}+1)n\sigma}{30L}, \frac{n}{180}\right\}$.*

Remark I.4. *For $n > 180$, wehave*

$$\Omega\left(\frac{\Delta L \sqrt{\alpha}}{\varepsilon^2}\right) = \Omega\left(\frac{\Delta}{\varepsilon^2} \min\left\{L, \sqrt{\frac{\sqrt{3}+1}{30}} \sqrt{n\sigma L}, \frac{\sqrt{n}L}{\sqrt{180}}\right\}\right) = \Omega\left(\frac{\Delta}{\varepsilon^2} \min\{L, \sqrt{n\sigma L}\}\right).$$

Thus, our result is comparable to the one of Zhou and Gu's result (their result only related to IFO algorithms, so our result is more strong), but our construction only requires the dimension be $\mathcal{O}\left(1 + \frac{\Delta}{\varepsilon^2} \min\{L/n, \sqrt{\sigma L/n}\}\right)$, which is much smaller than $\mathcal{O}\left(\frac{\Delta}{\varepsilon^2} \min\{L, \sqrt{n\sigma L}\}\right)$ in (Zhou and Gu, 2019).

I.3 CONSTRUCTIONS

Consider

$$F(\boldsymbol{x};\alpha,m,\lambda,\beta)=\lambda G_{\text{NC}}(\boldsymbol{x}/\beta;\alpha,m). \tag{25}$$

Similar to our construction we introduced in Section 2, we denote the l-th row of the matrix $\boldsymbol{B}(m+1,\sqrt[4]{\alpha})$ by $\boldsymbol{b}_l$ and

$$\mathcal{L}_i=\{l:1\le l\le m, m+1-l\equiv i(\mathrm{mod}\ n)\}, i=1,2,\cdots,n. \tag{26}$$

Let $\mathcal{G}_k=\mathrm{span}\{\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_k\}$, $1\le k\le m$, $\mathcal{G}_0=\{\boldsymbol{0}\}$ and compose $F(\boldsymbol{x};\alpha,m,\lambda,\beta)$ to

$$\begin{cases} f_1(\boldsymbol{x};\alpha,m,\lambda,\beta)=\frac{\lambda n}{2\beta^2}\sum\limits_{l\in\mathcal{L}_i}\left\|\boldsymbol{b}_l^\top\boldsymbol{x}\right\|_2^2-\frac{\lambda n\sqrt{\alpha}}{\beta}\langle\boldsymbol{e}_1,\boldsymbol{x}\rangle+\lambda\alpha\sum\limits_{i=1}^m\Gamma(x_i/\beta),\\ f_i(\boldsymbol{x};\alpha,m,\lambda,\beta)=\frac{\lambda n}{2\beta^2}\sum\limits_{l\in\mathcal{L}_i}\left\|\boldsymbol{b}_l^\top\boldsymbol{x}\right\|_2^2+\lambda\alpha\sum\limits_{i=1}^m\Gamma(x_i/\beta),\ \text{for } i\ge 2.\end{cases} \tag{27}$$

Clearly, $F(\boldsymbol{x};\alpha,m,\lambda,\beta)=\frac{1}{n}\sum_{i=1}^n f_i(\boldsymbol{x};\alpha,m,\lambda,\beta)$. Moreover, by Proposition I.2, we have following properties about $F(\boldsymbol{x};\alpha,m,\lambda,\beta)$ and $\{f_i(\boldsymbol{x};\alpha,m,\lambda,\beta)\}_{i=1}^n$.

Proposition I.5. *For any $0<\alpha\le 1$, it holds that*

1. *$f_i(\boldsymbol{x};\alpha,m,\lambda,\beta)$ is $\left(\frac{-45(\sqrt{3}-1)\alpha\lambda}{\beta^2},\frac{(2n+180\alpha)\lambda}{\beta^2}\right)$-smooth.*
2. *$F(\boldsymbol{0};\alpha,m,\lambda,\beta)-\min_{\boldsymbol{x}\in\mathbb{R}^{m+1}}F(\boldsymbol{x};\alpha,m,\lambda,\beta)\le\lambda(\sqrt{\alpha}/2+10\alpha m)$.*
3. *For $\boldsymbol{x}$ which satisfies that $x_m=x_{m+1}=0$, we have*
$$\|\nabla F(\boldsymbol{x};\alpha,m,\lambda,\beta)\|_2\ge\frac{\alpha^{3/4}\lambda}{4\beta}.$$

Similar to Lemma 2.6, similar conclusion hosts for $\{f_i(\boldsymbol{x};\alpha,m,\lambda,\beta)\}_{i=1}^n$.

Lemma I.6. *For $\boldsymbol{x}\in\mathcal{F}_k$, $0\le k<m$ and $\gamma<\frac{\sqrt{2}+1}{60}\frac{\beta^2}{\lambda\alpha}$, we have*

$$\nabla f_i(\boldsymbol{x};\alpha,m,\lambda,\beta),\mathrm{prox}_{f_i}^{\gamma}(\boldsymbol{x})\in\begin{cases}\mathcal{G}_{k+1},\ \textit{if } k\equiv i-1(\mathrm{mod}\ n),\\ \mathcal{G}_k,\ \textit{otherwise}.\end{cases}$$

Proof. Let $G(\boldsymbol{x})\triangleq\sum\limits_{i=1}^m\Gamma(x_i)$ and $\Gamma'(x)$ be the derivative of $\Gamma(x)$.

First note that $\Gamma'(0)=0$, so if $\boldsymbol{x}\in\mathcal{G}_k$, then

$$\nabla G(\boldsymbol{x})=\big(\Gamma'(x_1),\Gamma'(x_2),\cdots,\Gamma'(x_m)\big)^\top\in\mathcal{G}_k.$$

Moreover, for $\boldsymbol{x}\in\mathcal{F}_G$ $(k\ge 1)$, we have

$$\begin{aligned}\boldsymbol{b}_l^\top\boldsymbol{x}&=0\text{ for }l<m-k,\\ \boldsymbol{b}_l&\in\mathcal{G}_k\text{ for }l>m-k,\\ \boldsymbol{b}_{m-k}&\in\mathcal{G}_{k+1}.\end{aligned}$$

Consequently, for $l\ne m-k$, $\boldsymbol{b}_l\boldsymbol{b}_l^\top\boldsymbol{x}=(\boldsymbol{b}_l^\top\boldsymbol{x})\boldsymbol{b}_l\in\mathcal{G}_k$, and $\boldsymbol{b}_{m-k}\boldsymbol{b}_{m-k}^\top\boldsymbol{x}\in\mathcal{G}_{k+1}$.

For $k=0$, we have $\boldsymbol{x}=\boldsymbol{0}$, and

$$\begin{aligned}\nabla f_1(\boldsymbol{x})&=\lambda n\sqrt{\alpha}/\beta\ \boldsymbol{e}_1\in\mathcal{G}_1,\\ \nabla f_j(\boldsymbol{x})&=\boldsymbol{0}\ (j\ge 2).\end{aligned}$$

For $k\ge 1$, we suppose that $m-k\in\mathcal{L}_i$. Since

$$\nabla f_j(\boldsymbol{x})=\frac{\lambda n}{\beta^2}\sum_{l\in\mathcal{L}_j}\boldsymbol{b}_l^\top\boldsymbol{b}_l\boldsymbol{x}+\frac{\lambda\alpha}{\beta}\nabla G(\boldsymbol{x}/\beta)-\eta_j\boldsymbol{e}_1,$$

where $\eta_1 = \lambda n\sqrt{\alpha}/\beta$, $\eta_j = 0$ for $j \geq 2$.
Hence, $\nabla f_i(\boldsymbol{x}) \in \mathcal{F}_{k+1}$ and $\nabla f_j(\boldsymbol{x}) \in \mathcal{F}_k$ $(j \neq i)$.

Now, we turn to consider $\boldsymbol{v} = \mathrm{prox}_{f_j}^{\gamma}(\boldsymbol{x})$.

We have

$$\nabla f_j(\boldsymbol{v}) + \frac{1}{\gamma}(\boldsymbol{v} - \boldsymbol{x}) = \boldsymbol{0},$$

that is

$$\left(\frac{\lambda n}{\beta^2}\sum_{l\in\mathcal{L}_j}\boldsymbol{b}_l^\top\boldsymbol{b}_l + \frac{1}{\gamma}\boldsymbol{I}\right)\boldsymbol{v} + \frac{\lambda\alpha}{\beta}\nabla G(\boldsymbol{v}/\beta) = \eta_j\boldsymbol{e}_1 + \frac{1}{\gamma}\boldsymbol{x}. \tag{28}$$

Denote

$$\boldsymbol{A} = \frac{\lambda n}{\beta}\sum_{l\in\mathcal{L}_j}\boldsymbol{b}_l^\top\boldsymbol{b}_l + \frac{\beta}{\gamma}\boldsymbol{I},\ \boldsymbol{u} = \frac{1}{\beta}\boldsymbol{v},\ \boldsymbol{y} = \eta_j\boldsymbol{e}_1 + \frac{1}{\gamma}\boldsymbol{x},$$

then we have

$$\boldsymbol{A}\boldsymbol{u} + \frac{\lambda\alpha}{\beta}\nabla G(\boldsymbol{u}) = \boldsymbol{y}. \tag{29}$$

Next, if s satisfies

$$\begin{cases} s > \max\{1, k\} & \text{for } j = 1, \\ s > k & \text{for } j > 1, \end{cases} \tag{30}$$

then we know that the s-th element of $\boldsymbol{y}$ is 0.

If s satisfies (30) and $m - s \in \mathcal{L}_j$, then the s-th and $(s+1)$-th elements of $\boldsymbol{A}\boldsymbol{u}$ is $((\xi + \beta/\gamma)u_s - \xi u_{s+1})$ and $(-\xi u_s + (\xi + \beta/\gamma)u_{s+1})$ respectively where $\xi = \lambda n/\beta$. So by Equation (29), we have

$$\begin{cases} \frac{\beta}{\gamma}u_s + \xi(u_s - u_{s+1}) + \frac{120\lambda\alpha}{\beta}\frac{u_s^2(u_s-1)}{1+u_s^2} = 0. \\ \frac{\beta}{\gamma}u_{s+1} + \xi(u_{s+1} - u_s) + \frac{120\lambda\alpha}{\beta}\frac{u_{s+1}^2(u_{s+1}-1)}{1+u_{s+1}^2} = 0. \end{cases}$$

Following from Lemma I.9, for $\frac{120\lambda\alpha}{\beta} < \frac{(2+2\sqrt{2})\beta}{\gamma}$, we have $u_s = u_{s+1} = 0$.
That is

1. if $m - s \in \mathcal{L}_j$ and s satisfies (30), then $u_s = 0$.
2. if $m - s + 1 \in \mathcal{L}_j$ and $s - 1$ satisfies (30), then $u_s = 0$.

For s which satisfies (30), if $m - s \notin \mathcal{L}_j$ and $m - s + 1 \notin \mathcal{L}_j$, then the s-th element of $\boldsymbol{A}\boldsymbol{u}$ is $(\beta/\gamma\, u_s)$. Similarly, by Equation (29), we have

$$\frac{\beta}{\gamma}u_s + \frac{120\lambda\alpha}{\beta}\frac{u_s^2(u_s-1)}{1+u_s^2} = 0.$$

Following from Lemma I.8, for $\frac{120\lambda\alpha}{\beta} < \frac{(2+2\sqrt{2})\beta}{\gamma}$, we have $u_s = 0$.

Therefore, we can conclude that

1. if $s - 1$ satisfies (30), then $u_s = 0$.

2. if s satisfies (30) and $m-s+1 \notin \mathcal{L}_j$, then $u_s = 0$.

Moreover, we have that

1. if $k=0$ and $j=1$, then $m-1, m-2 \notin \mathcal{L}_j$, so $u_2 = 0$.
2. if $k=0$ and $j>1$, then for $s=1$, we have $m-s+1 \notin \mathcal{L}_j$, so $u_1 = 0$.
3. if $k=0$, then for $s>2$, we have $s-1>1$ satisfies (30), so $u_s=0$.
4. if $k>0$, then for $s>k+1$, we have $s-1>k$ satisfies (30), so $u_s=0$.
5. if $k>0$ and $m-k \notin \mathcal{L}_j$, then for $s=k+1$, we have $m-s+1 \notin \mathcal{L}_j$, so $u_{k+1}=0$.

In short,

1. if $k=0$ and $j>1$, then $\boldsymbol{u} \in \mathcal{G}_0$.
2. if $k=0$ and $j=1$, then $\boldsymbol{u} \in \mathcal{G}_1$.
3. if $k>1$ and $m-k \notin \mathcal{L}_j$, then $\boldsymbol{u} \in \mathcal{G}_k$.
4. if $k>1$ and $m-k \in \mathcal{L}_j$, then $\boldsymbol{u} \in \mathcal{G}_{k+1}$.

□

Remark I.7. *In order to make the operator* $\mathrm{prox}_{f_i}^{\gamma}$ *valid,* γ *need to satisfy*

$$\gamma < \frac{\sqrt{3}+1}{90}\frac{\beta^2}{\lambda\alpha} < \frac{\sqrt{2}+1}{60}\frac{\beta^2}{\lambda\alpha}.$$

So for any valid PIFO call, the condition about γ *in Lemma I.6 must be satisfied.*

Lemma I.8. *Suppose that* $0 < \lambda_2 < (2+2\sqrt{2})\lambda_1$*, then* $z=0$ *is the only real solution to the equation*

$$\lambda_1 z + \lambda_2 \frac{z^2(z-1)}{1+z^2} = 0. \tag{31}$$

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 (31) 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 (31).

□

Lemma I.9. *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*

$$\begin{cases} \lambda_1 z_1 + \lambda_3(z_1 - z_2) + \lambda_2 \frac{z_1^2(z_1-1)}{1+z_1^2} = 0. \\ \lambda_1 z_2 + \lambda_3(z_2 - z_1) + \lambda_2 \frac{z_2^2(z_2-1)}{1+z_2^2} = 0. \end{cases} \tag{32}$$

Proof. If $z_1 = 0$, then $z_2 = 0$. So let assume that $z_1 z_2 \neq 0$. Rewrite the first equation of Equation (32) 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. ☐

Following from Lemma I.6, we know following Lemma which is similar to Lemma 2.9.

Lemma I.10. *If $M \geq 1$ satisfies $\min_{\boldsymbol{x}\in\mathcal{G}_M} \|\nabla F(\boldsymbol{x})\|_2 \geq 9\varepsilon$ and $N = n(M+1)/4$, then we have*

$$\min_{t\leq N} \mathbb{E}\|\nabla F(\boldsymbol{x}_t)\|_2 \geq \varepsilon.$$

Theorem I.11. *Set*

$$\alpha = \min\left\{1, \frac{(\sqrt{3}+1)n\sigma}{30L}, \frac{n}{180}\right\},$$
$$\lambda = \frac{3888n\varepsilon^2}{L\alpha^{3/2}},$$
$$\beta = \sqrt{3\lambda n/L},$$
$$m = \left\lfloor \frac{\Delta L\sqrt{\alpha}}{40824n\varepsilon^2} \right\rfloor$$

Suppose that $\varepsilon^2 \leq \frac{\Delta L\alpha}{81648n}$. In order to find $\hat{\boldsymbol{x}} \in \mathbb{R}^{m+1}$ such that $\mathbb{E}\|\nabla F(\hat{\boldsymbol{x}})\|_2 < \varepsilon$, PIFO algorithm $\mathcal{A}$ needs at least $\Omega\left(\frac{\Delta L\sqrt{\alpha}}{\varepsilon^2}\right)$ queries to h_F.

Proof. First, note that f_i is $(-l_1, l_2)$-smooth, where

$$l_1 = \frac{45(\sqrt{3}-1)\alpha\lambda}{\beta^2} = \frac{45(\sqrt{3}-1)L}{3n}\alpha \leq \frac{45(\sqrt{3}-1)L}{3n}\frac{(\sqrt{3}+1)n\sigma}{30L} = \sigma,$$
$$l_2 = \frac{(2n+180\alpha)\lambda}{\beta^2} = \frac{L}{3n}(2n+180\alpha) \leq L.$$

Thus each f_i is $(-\sigma, L)$-smooth.

Next, observe that

$$F(\boldsymbol{x}_0) - F(\boldsymbol{x}^*) \leq \lambda(\sqrt{\alpha}/2 + 10\alpha m) = \frac{1944n\varepsilon^2}{L\alpha} + \frac{38880n\varepsilon^2}{L\sqrt{\alpha}}m$$
$$\leq \frac{1944}{40824}\Delta + \frac{38880}{40824}\Delta = \Delta.$$

For $M = m-1$, we know that

$$\min_{\boldsymbol{x}\in\mathcal{G}_M} \|\nabla F(\boldsymbol{x})\|_2 \geq \frac{\alpha^{3/4}\lambda}{4\beta} = \frac{\alpha^{3/4}\lambda}{4\sqrt{3\lambda n/L}} = \sqrt{\frac{\lambda L}{3n}}\frac{\alpha^{3/4}}{4} = 9\varepsilon.$$

With recalling Lemma I.10, in order to find $\hat{\boldsymbol{x}} \in \mathbb{R}^{m+1}$ such that $\mathbb{E}\left\|\nabla F(\hat{\boldsymbol{x}})\right\|_2 < \varepsilon$, PIFO algorithm $\mathcal{A}$ needs at least N queries to h_F, where

$$N = n(M+1)/4 = nm/4 = \Omega\left(\frac{\Delta L\sqrt{\alpha}}{\varepsilon^2}\right).$$

At last, we need to ensure that $m \geq 2$. By $\varepsilon^2 \leq \frac{\Delta L\alpha}{81648n}$, we have

$$\frac{\Delta L\sqrt{\alpha}}{40824n\varepsilon^2} \geq \frac{\Delta L\alpha}{40824n\varepsilon^2} \geq 2,$$

and consequently $m \geq 2$. $\square$