Lower Bounds for Smooth Nonconvex Finite-Sum Optimization

Dongruo Zhou [1] **Quanquan Gu** [1]

Abstract

Smooth finite-sum optimization has been widely studied in both convex and nonconvex settings. However, existing lower bounds for finite-sum optimization are mostly limited to the setting where each component function is (strongly) convex, while the lower bounds for nonconvex finite-sum optimization remain largely unsolved. In this paper, we study the lower bounds for smooth nonconvex finite-sum optimization, where the objective function is the average of n nonconvex component functions. We prove tight lower bounds for the complexity of finding ϵ-suboptimal point and ϵ-approximate stationary point in different settings, for a wide regime of the smallest eigenvalue of the Hessian of the objective function (or each component function). Given our lower bounds, we can show that existing algorithms including KatyushaX (Allen-Zhu, 2018), Natasha (Allen-Zhu, 2017b) and StagewiseKatyusha (Chen & Yang, 2018) have achieved optimal Incremental First-order Oracle (IFO) complexity (i.e., number of IFO calls) up to logarithm factors for nonconvex finite-sum optimization. We also point out potential ways to further improve these complexity results, in terms of making stronger assumptions or by a different convergence analysis.

1. Introduction

We consider minimizing the following unconstrained finite-sum optimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}), \tag{1.1}$$

where each $f_i(\mathbf{x}) : \mathbb{R}^m \to \mathbb{R}$ is smooth and *nonconvex* function. We are interested in the algorithmic performance of *first-order algorithms* for solving (1.1), which have accesses to the Incremental First-order Oracle (IFO) (Agarwal & Bottou, 2015) defined as follows:

Given $\mathbf{x}$ and $i \in [n]$, an IFO returns $[f_i(\mathbf{x}), \nabla f_i(\mathbf{x})]$.

In this paper, we consider the very general setting where $F(\mathbf{x})$ is of (l, L)-smoothness (Allen-Zhu, 2017b), i.e., there exist some constant $l \in \mathbb{R}$ and $L > 0$, such that for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$\begin{aligned}\frac{l}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 \le F(\mathbf{x}) - F(\mathbf{y}) - \langle \nabla F(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle \\ \le \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|_2^2,\end{aligned} \tag{1.2}$$

where $l \in \mathbb{R}$ [1] is the lower smoothness parameter, and $L > 0$ is the upper smoothness parameter. Note that conventional L-smoothness definition is a special case of (1.2), where $l = -L$. (1.2) is quite general, because with different choice of l, (1.1) and (1.2) together can cover various kinds of smooth finite-sum optimization problems. For example, when $l \ge 0$, $F(\mathbf{x})$ is convex function, and $F(\mathbf{x})$ is σ-strongly convex if $l = \sigma > 0$. Such a sum-of-nonconvex optimization problem (convex functions that are average of nonconvex ones) was originally identified in Shalev-Shwartz (2015), and appears in various machine learning problems such as the shift-and-inverse procedure for principal component analsyis (PCA) (Garber et al., 2016; Allen-Zhu & Li, 2016). In specific, in order to calculate the leading eigenvector of $\mathbf{A} = 1/n\sum_{i=1}^n \mathbf{a}_i\mathbf{a}_i^\top$, the shift-and-inverse procedure minimizes $F(\mathbf{x}) = 1/n\sum_{i=1}^n f_i(\mathbf{x})$, where $f_i(\mathbf{x}) = 1/2 \cdot \mathbf{x}^\top(\mu\mathbf{I} - \mathbf{a}_i\mathbf{a}_i^\top)\mathbf{x} + \mathbf{b}^\top\mathbf{x}$, $\mu > 0$ such that $\mu\mathbf{I} \succ \mathbf{A}$, and $\mathbf{b}$ is a vector. It can be seen that $F(\mathbf{x})$ is convex because its Hessian $\mu\mathbf{I} - \mathbf{A}$ is positive definite, but some $f_i(\mathbf{x})$'s can be nonconvex since their Hessian $\mu\mathbf{I} - \mathbf{a}_i\mathbf{a}_i^\top$ can be negative definite. With $l \ge 0$, our goal is to find an ϵ-suboptimal solution $\widehat{\mathbf{x}}$ (Woodworth & Srebro, 2016) to (1.1), which satisfies

$$\mathbb{E}F(\widehat{\mathbf{x}}) - \inf_{\mathbf{x}} F(\mathbf{x}) \le \epsilon. \tag{1.3}$$

On the other hand, when $l = -\sigma < 0$, $F(\mathbf{x})$ is nonconvex,

[1]Department of Computer Science, University of California, Los Angeles. Correspondence to: Quanquan Gu <qgu@cs.ucla.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

[1]We allow l to be nonnegative, which covers the definitions of convex and strongly convex functions.

and it is called σ-almost convex (Carmon et al., 2018) [2]. It is known that finding an ϵ-suboptimal solution in such nonconvex setting is NP-hard (Murty & Kabadi, 1987). Thus, our goal is instead to find an ϵ-approximate stationary point $\widehat{\mathbf{x}}$ of $F(\mathbf{x})$ for general nonconvex case, which is defined as follows

$$\mathbb{E}\|\nabla F(\widehat{\mathbf{x}})\|_2 \leq \epsilon. \tag{1.4}$$

There is a vast literature on finding either (1.3) or (1.4) for (1.1), such as SDCA without Duality (Shalev-Shwartz, 2016), Natasha (Allen-Zhu, 2017b), KatyushaX (Allen-Zhu, 2018), RapGrad (Lan & Yang, 2018), StagewiseKatyusha (Chen & Yang, 2018), RepeatSVRG (Agarwal et al., 2017; Carmon et al., 2018), to mention a few. In specific, this line of work can be divided into two categories based on the smoothness assumption over $\{f_i(\mathbf{x})\}_{i=1}^n$. The first category of work (Shalev-Shwartz, 2016; Allen-Zhu, 2017b; 2018; Agarwal et al., 2017; Carmon et al., 2018) makes the assumption that each individual component function $f_i(\mathbf{x})$ is L-smooth and $F(\mathbf{x})$ is (l, L) smooth. Under such an assumption, when $F(\mathbf{x})$ is convex or σ-strongly convex, SDCA without Duality and KatyushaX can find the ϵ-suboptimal solution within $O(n + n^{3/4}\sqrt{L/\epsilon})$ or $O(n + n^{3/4}\sqrt{L/\sigma}\log(1/\epsilon))$ IFO calls respectively. When $F(\mathbf{x})$ is σ-almost convex, Natasha and RepeatSVRG can find the ϵ-approximate stationary point with $O((n^{3/4}\sqrt{\sigma L} \wedge \sqrt{n}L)/\epsilon^2)$ IFO calls.

The second category of work Allen-Zhu (2017b; 2018); Lan & Yang (2018); Chen & Yang (2018) assumes that each $f_i(\mathbf{x})$ is $(-\sigma, L)$-smooth [3]. With such an assumption, RapGrad and StagewiseKatyusha find ϵ-approximate stationary point with $O((n\sigma + \sqrt{n\sigma L})/\epsilon^2)$ IFO calls.

Given the above IFO complexity results, a natural research question is:

Are these upper bounds of IFO complexity already optimal?

We answer this question in an affirmative way by proving lower bounds on the IFO complexity for a wide regime of l, using carefully constructed functions. More specifically, our contributions are summarized as follows:

1. For the case that $F(\mathbf{x})$ is convex or σ-strongly convex (a.k.a., sum-of-nonconvex optimization), we show that without the L-smoothness assumption on each component function $f_i(\mathbf{x})$, the lower bound of IFO complexity for any linear-span first-order randomized algorithms (See Definition 3.3) to find ϵ-suboptimal solution is $\Omega(n + n^{3/4}\sqrt{L/\sigma}\log(1/\epsilon))$ when F is σ-strongly convex, and $\Omega(n + n^{3/4}\sqrt{L/\epsilon})$ when $F(\mathbf{x})$ is convex, where L is the average smoothness parameter on $\{f_i(\mathbf{x})\}_{i=1}^n$ (See Definition 3.2). That is in contrast to the lower bounds $\Omega(n + n^{1/2}\sqrt{L/\sigma}\log(1/\epsilon))$ and $\Omega(n + n^{1/2}\sqrt{L/\epsilon})$ proved by Woodworth & Srebro (2016) when each component function $f_i(\mathbf{x})$ is L-smooth.

2. For the case that $F(\mathbf{x})$ is σ-almost convex, we show that the lower bound of IFO complexity for any linear-span first-order randomized algorithms to find ϵ-approximate stationary point is $\Omega(1/\epsilon^2(n^{3/4}\sqrt{L\sigma} \wedge \sqrt{n}L))$ when $\{f_i(\mathbf{x})\}_{i=1}^n$ is L-average smooth, and $\Omega(1/\epsilon^2(\sqrt{nL\sigma} \wedge L))$ when each $f_i(\mathbf{x})$ is $(-\sigma, L)$-smooth. To our best knowledge, this is the first lower bound result which precisely characterizes the dependency on the lower smoothness parameter for finding approximate stationary point.

3. We show that many existing algorithms including SDCA without Duality (Shalev-Shwartz, 2016), Natasha (Allen-Zhu, 2017b), KatyushaX (Allen-Zhu, 2018), RapGrad (Lan & Yang, 2018), StagewiseKatyusha (Chen & Yang, 2018) and RepeatSVRG (Agarwal et al., 2017; Carmon et al., 2018) have indeed achieved optimal IFO complexity for a large regime of the lower smoothness parameter, with slight modification of their original convergence analyses.

Notation We use $a(x) = O(b(x))$ if $a(x) \leq Cb(x)$, where C is a universal constant. We use $\widetilde{O}(\cdot)$ to hide polynomial logarithm terms. For any vector $\mathbf{v} \in \mathbb{R}^m$, we use $\mathbf{v}_i$ to denote the i-th coordinate of $\mathbf{v}$, and $\|\mathbf{v}\|_2$ to denote its 2-norm. For any vector sequence $\{\mathbf{v}^{(i)}\}_{i=1}^n$, we use $\mathbf{v}^{(i)}$ to denote the i-th vector. We say a matrix sequence $\{\mathbf{U}^{(i)}\}_{i=1}^n \in \mathcal{O}(a, b, n)$ where for each i, $\mathbf{U}^{(i)} \in \mathbb{R}^{a\times b}$, if $\mathbf{U}^{(i)}(\mathbf{U}^{(i)})^\top = \mathbf{I}$ and $\mathbf{U}^{(i)}(\mathbf{U}^{(j)})^\top = \mathbf{0}$ for any $1 \leq i \neq j \leq n$. For any sets $A, B \subseteq \mathbb{R}^d$, we define the distance between them as $\text{dist}(A, B) = \inf_{\mathbf{a}\in A, \mathbf{b}\in B}\|\mathbf{a} - \mathbf{b}\|_2$. For any $A \subseteq \mathbb{R}^d$, we denote by $\text{Lin}\{A\}$ the linear space spanned by $a \in A$. In the rest of this paper, we use $F(\mathbf{x})$, $f_i(\mathbf{x})$ and F, f_i interchangeably when there is no confusion.

2. Additional Related Work

In this section, we review additional related work that is not discussed in the introduction section.

Existing lower bounds for nonconvex optimization: To the best of our knowledge, the only existing lower bounds for nonconvex optimization are proved in Carmon et al. (2017a;b); Fang et al. (2018). Carmon et al. (2017a;b) proved the lower bounds for both deterministic and randomized algorithms on nonconvex optimization with high-order smoothness assumption. However, they did not consider the finite-sum structure which will bring additional dependency

[2] It is also known as σ-weakly convex (Chen & Yang, 2018) or σ-bounded nonconvex Allen-Zhu (2017b).)

[3] In fact, Allen-Zhu (2017b; 2018) fall into both categories.

on the lower-smoothness parameter l and the number of component functions n. Fang et al. (2018) proved a lower bound for nonconvex finite-sum optimization under conventional smoothness assumption, i.e., $l = -L$. Our work extends this line of research, and proves matching lower bounds for nonconvex finite-sum optimization under the refined (l, L)-smooth assumption.

Existing upper bounds for first-order convex optimization: There existing a bunch of work focusing on establishing upper complexity bounds to find ϵ-suboptimal solution for convex finite-sum optimization problems. It is well known that by treating $F(\mathbf{x})$ as a whole part, gradient descent can achieve $O(nL/\epsilon)$ IFO complexity for convex functions and $O(nL/\sigma \log(1/\epsilon))$ for σ-strongly convex functions, and accelerated gradient descent (AGD) (Nesterov, 1983) can achieve $O(n\sqrt{L/\epsilon})$ IFO complexity for convex functions and $O(n\sqrt{L/\sigma}\log(1/\epsilon))$ for σ-strongly convex functions. Both IFO complexities achieved by AGD are optimal when $n = 1$ (Nesterov, 1983). By using variance reduction technique (Roux et al., 2012; Johnson & Zhang, 2013; Xiao & Zhang, 2014; Defazio et al., 2014; Mairal, 2015; Bietti & Mairal, 2017), the IFO complexity can be improved to be $O((n + L/\sigma)\log(1/\epsilon))$ for strongly convex functions. By combining variance reduction and Nesterov's acceleration techniques (Nesterov, 1983), the IFO complexity can be further reduced to $O(n\log(1/\epsilon) + \sqrt{nL/\epsilon})$ for convex functions, and $O((n + \sqrt{nL/\sigma})\log(1/\epsilon))$ for σ-strongly convex functions (Allen-Zhu, 2017a), which matches the lower bounds up to a logarithm factor.

Existing lower bounds for first-order convex optimization: For deterministic optimization algorithms, it has been proved that one needs $\Omega(\sqrt{L/\epsilon})$ IFO calls for convex functions, and $\Omega(\sqrt{L/\sigma}\log(1/\epsilon))$ IFO calls for σ-strongly convex functions to find an ϵ-suboptimal solution. There is a line of work (Woodworth & Srebro, 2016; Lan & Zhou; Agarwal & Bottou, 2015; Arjevani & Shamir, 2016) establishing the lower bounds for first-order algorithms to find ϵ-suboptimal solution to the convex finite-sum optimization. More specifically, Agarwal & Bottou (2015) proved a lower bound $\Omega(n + \sqrt{nL/\sigma}\log(1/\epsilon))$ for strongly convex finite-sum optimization problems, which is valid for deterministic algorithms. Arjevani & Shamir (2016) provided a dimension-free lower bound $\Omega(n + \sqrt{nL/\sigma}\log(1/\epsilon))$ for first-order algorithms with the assumption that any new iterate generated by the algorithm lies in the linear span of gradients and iterates up to the current iteration. Lan & Zhou proved a lower bound $\Omega(n + \sqrt{L/\sigma}\log(1/\epsilon))$ for a class of randomized first-order algorithms where each component function will be selected by fixed probabilities. Woodworth & Srebro (2016) proved a set of lower bounds including $\Omega(n + \sqrt{L/\epsilon})$ for convex functions and $\Omega(n + \sqrt{L/\sigma}\log(1/\epsilon))$ for σ-strongly convex functions. Besides, Woodworth & Srebro (2016)'s results do not need the assumption that the new iterate lies in the span of all the iterates up to the iteration, which is a more general result.

For more details on the upper bound and lower bound results, please refer to Tables 1 and 2.

3. Preliminaries

We first present the formal definitions of (l, L)-smoothness and average smoothness, which will be used throughout the proof.

Definition 3.1. For any differentiable function $f : \mathbb{R}^m \to \mathbb{R}$, we say f is (l, L)-smooth for some $l \in \mathbb{R}$ and $L \in \mathbb{R}^+$ if for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$, it holds that

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

We denote such a function class by $\mathcal{S}^{(l,L)}$. In particular, we say f is L-smooth if $f \in \mathcal{S}^{(-L,L)}$.

Note that if f is twice differentiable, then $f \in \mathcal{S}^{(l,L)}$ if and only if $l\mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}$ for any $\mathbf{x} \in \mathbb{R}^m$.

Definition 3.2. For any differentiable functions $\{f_i\}_{i=1}^n : \mathbb{R}^m \to \mathbb{R}$, we say $\{f_i\}_{i=1}^n$ is L-average smooth for some $L > 0$ if $\mathbb{E}_i\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|_2^2 \leq L^2\|\mathbf{x} - \mathbf{y}\|_2^2$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$, where $\mathbb{E}_i X(i) = 1/n \cdot \sum_{i=1}^n X(i)$ for any random variable $X(i)$. We denote such a function class by $\mathcal{V}^{(L)}$.

It is worth noting that if $\{f_i\}$ satisfy that for each i, $f_i \in \mathcal{S}^{(-L,L)}$, then $\{f_i\} \in \mathcal{V}^{(L)}$.

In this work, we focus on the *linear-span randomized first-order algorithm*, which is defined as follows:

Definition 3.3. Given an initial point $\mathbf{x}^{(0)}$, a *linear-span randomized first-order algorithm* $\mathcal{A}$ is defined as a measurable mapping from functions $\{f_i\}_{i=1}^n$ to an infinite sequence of point and index pairs $\{(\mathbf{x}^{(t)}, i_t)\}_{t=0}^\infty$ with random variable $i_t \in [n]$, which satisfies

$$\mathbf{x}^{(t+1)} \in \text{Lin}\{\mathbf{x}^{(0)}, \ldots, \mathbf{x}^{(t)}, \nabla f_{i_0}(\mathbf{x}^{(0)}), \ldots, \nabla f_{i_t}(\mathbf{x}^{(t)})\}.$$

It can be easily checked that most first-order primal finite-sum optimization algorithms, such as SAG (Roux et al., 2012), SVRG (Johnson & Zhang, 2013), SAGA (Defazio et al., 2014) and Katyusha (Allen-Zhu, 2017a), KatyushaX (Allen-Zhu, 2018), are linear-span randomized first-order algorithms.

In this work, we prove the lower bounds by constructing adversarial functions which are "hard enough" for any linear-span randomized first-order algorithms. To demonstrate

Table 1. IFO Complexity comparison with the assumption that $\{f_i\}_{i=1}^n$ is average L-smooth and F is (l, L)-smooth. Here $\Delta = F(\mathbf{x}^{(0)}) - \inf_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x})$ and $B = \mathrm{dist}(\mathbf{x}^{(0)}, \mathcal{X}^*)$, where $\mathcal{X}^* = \mathrm{argmin}_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x})$. When $l = \sigma$ or 0, the goal is to find an ϵ-suboptimal solution; and when $l = -\sigma < 0$, the goal is to find an ϵ-approximate stationary point.

$\sigma > 0$	(σ, L)	$(0, L)$	$(-\sigma, L)$
Upper Bounds	$O\Big(\big(n + n^{3/4}\sqrt{\frac{L}{\sigma}}\big)\log\frac{\Delta}{\epsilon}\Big)$ (Allen-Zhu, 2018)	$O\big(n + n^{3/4}B\sqrt{\frac{L}{\epsilon}}\big)$ (Allen-Zhu, 2018)	$\widetilde{O}\big(\frac{\Delta}{\epsilon^2}(n^{3/4}\sqrt{\sigma L} \wedge \sqrt{n}L)\big)$ (Allen-Zhu, 2017b; Fang et al., 2018)
Lower Bounds	$\Omega\big(n + n^{3/4}\sqrt{\frac{L}{\sigma}}\log\frac{\Delta}{\epsilon}\big)$ (Theorem 4.1)	$\Omega\big(n + n^{3/4}B\sqrt{\frac{L}{\epsilon}}\big)$ (Theorem 4.2)	$\Omega\big(\frac{\Delta}{\epsilon^2}(n^{3/4}\sqrt{\sigma L} \wedge \sqrt{n}L)\big)$ (Theorem 4.5)

Table 2. IFO Complexity comparison with the assumption that each f_i is (l, L)-smooth. Here $\Delta = F(\mathbf{x}^{(0)}) - \inf_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x})$ and $B = \mathrm{dist}(\mathbf{x}^{(0)}, \mathcal{X}^*)$ where $\mathcal{X}^* = \mathrm{argmin}_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x})$. When $l = \sigma$ or 0, the goal is to find an ϵ-suboptimal solution; and when $l = -\sigma < 0$, the goal is to find an ϵ-approximate stationary point.

$\sigma > 0$	(σ, L)	$(0, L)$	$(-\sigma, L)$
Upper Bounds	$O\Big(\big(n + \sqrt{\frac{nL}{\sigma}}\big)\log\frac{\Delta}{\epsilon}\Big)$ (Allen-Zhu, 2017a)	$O\big(n + B\sqrt{\frac{nL}{\epsilon}}\big)$ (Allen-Zhu, 2017a)	$\widetilde{O}\big(\frac{\Delta}{\epsilon^2}(n\sigma + \sqrt{n\sigma L}) \wedge \sqrt{n}L)\big)$ (Lan & Yang, 2018; Fang et al., 2018)
Lower Bounds	$\Omega\big(n + \sqrt{\frac{nL}{\sigma}}\log\frac{\Delta}{\epsilon}\big)$ (Woodworth & Srebro, 2016)	$\Omega\big(n + B\sqrt{\frac{nL}{\epsilon}}\big)$ (Woodworth & Srebro, 2016)	$\Omega\big(\frac{\Delta}{\epsilon^2}(\sqrt{n\sigma L} \wedge L)\big)$ (Theorem 4.7)

the construction of adversarial functions, we first introduce the following quadratic function class, which comes from Nesterov (2013).

Definition 3.4. Let $Q(\mathbf{x}; \xi, m, \zeta) : \mathbb{R}^m \to \mathbb{R}$ be:

$$Q(\mathbf{x}; \xi, m, \zeta) := \frac{\xi}{2}(\mathbf{x}_1 - 1)^2 + \frac{1}{2}\sum_{t=1}^{m-1}(\mathbf{x}_{t+1} - \mathbf{x}_t)^2 + \frac{\zeta}{2}(\mathbf{x}_m)^2.$$

In our construction, we need the following two important properties of $Q(\mathbf{x}; \xi, m, \zeta)$.

Proposition 3.5. For any $0 \le \xi, \zeta \le 1$ and $m \ge 1$, the following properties hold:

1. $Q(\mathbf{x}; \xi, m, \zeta) \in \mathcal{S}^{(0,4)}$.
2. Suppose that $\mathbf{U} \in \mathbb{R}^{m\times d}$ satisfying $\mathbf{U}\mathbf{U}^\top = \mathbf{I}$. Suppose that $\mathbf{U} = [\mathbf{u}^{(1)}, ...\mathbf{u}^{(m)}]^\top$. Then for any $\bar{\mathbf{x}}$ satisfying $\mathbf{U}\bar{\mathbf{x}} \in \mathrm{Lin}\{\mathbf{u}^{(1)}, ..., \mathbf{u}^{(t)}\}$, and any differentiable function $\mu : \mathbb{R} \to \mathbb{R}$, we have $\nabla[Q(\mathbf{U}\bar{\mathbf{x}}; \xi, m, \zeta) + \sum_{i=1}^m \mu(\bar{\mathbf{x}}^\top\mathbf{u}^{(i)})] \in \mathrm{Lin}\{\mathbf{u}^{(1)}, ..., \mathbf{u}^{(t+1)}\}$.

In short, the first property of $Q(\mathbf{x}; \xi, m, \zeta)$ says that Q is a convex function with 4-smoothness, and the second property says that for any orthogonal matrix $\mathbf{U}$, the composite function $Q(\mathbf{U}\mathbf{x}; \xi, m, \zeta) + \sum_{i=1}^m \mu(\mathbf{x}^\top\mathbf{u}^{(i)})$ enjoys the so-called *zero-chain* property (Carmon et al., 2017a): if the current point is $\bar{\mathbf{x}}$, then the information brought by an IFO call at the current point can at most increase the dimension of linear space which $\bar{\mathbf{x}}$ belongs to by 1, which is very important for the proof of lower bounds.

Based on Definition 3.4, one can define the following three function classes: $f_{\mathcal{N}\mathrm{sc}}$, $f_{\mathcal{N}\mathrm{c}}$ from Nesterov (2013) and $f_{\mathcal{C}}$ from Carmon et al. (2017b). We first introduce a class of strongly convex functions $f_{\mathcal{N}\mathrm{sc}}$, which is originally defined in Nesterov (2013).

Definition 3.6. (Nesterov, 2013) Let $f_{\mathcal{N}\mathrm{sc}}(\mathbf{x}; \alpha, m) : \mathbb{R}^m \to \mathbb{R}$ be

$$f_{\mathcal{N}\mathrm{sc}}(\mathbf{x}; \alpha, m) := \frac{1-\alpha}{4}Q\bigg(\mathbf{x}; 1, m, \frac{2\sqrt{\alpha}}{\sqrt{\alpha}+1}\bigg) + \frac{\alpha}{2}\|\mathbf{x}\|_2^2. \tag{3.1}$$

For $f_{\mathcal{N}\mathrm{sc}}(\mathbf{x}; \alpha, m)$, we have the following properties.

Proposition 3.7 (Chapter 2.1.4, Nesterov (2013))**.** For any $0 \le \alpha \le 1$, let $q := (1 - \sqrt{\alpha})/(1 + \sqrt{\alpha})$, it holds that

1. $f_{\mathcal{N}\mathrm{sc}}(\mathbf{x}; \alpha, m) \in \mathcal{S}^{(\alpha,1)}$.
2. $f_{\mathcal{N}\mathrm{sc}}(0; \alpha, m) - \inf_{\mathbf{x}\in\mathbb{R}^m} f_{\mathcal{N}\mathrm{sc}}(\mathbf{x}; \alpha, m) \le q^2(1 - q^2)$.

3. For any $\mathbf{x}$ satisfying $\mathbf{x}_m = 0$, we have

$$f_{\mathcal{N}\mathrm{sc}}(\mathbf{x};\alpha,m) - \inf_{\mathbf{x}} f_{\mathcal{N}\mathrm{sc}}(\mathbf{x};\alpha,m) \geq \frac{\alpha}{2}q^{2m+2}.$$

Next we introduce a class of general convex functions $f_{\mathcal{N}\mathrm{c}}(\mathbf{x};m)$, which is also defined in Nesterov (2013).

Definition 3.8. (Nesterov, 2013) Let $f_{\mathcal{N}\mathrm{c}}(\mathbf{x};m) : \mathbb{R}^{2m-1} \to 1$ be

$$f_{\mathcal{N}\mathrm{c}}(\mathbf{x};m) := \frac{1}{4}Q(\mathbf{x};1,2m-1,1). \tag{3.2}$$

We have the following properties about $f_{\mathcal{N}\mathrm{c}}(\mathbf{x};m)$.

Proposition 3.9 (Chapter 2.1.2, Nesterov (2013))**.** We have

1. $f_{\mathcal{N}\mathrm{c}}(\mathbf{x};m) \in \mathcal{S}^{(0,1)}$.
2. Let $\mathcal{X}^* = \mathrm{argmin}_{\mathbf{x}\in\mathbb{R}^d} f_{\mathcal{N}\mathrm{c}}(\mathbf{x};m)$ be the optimal solution set, we have $\mathrm{dist}^2(0,\mathcal{X}^*) \leq 2m/3$.
3. For any $\mathbf{x}$ which satisfies that $|\mathbf{x}_m| = ... = |\mathbf{x}_{2m-1}| = 0$, we have $f_{\mathcal{N}\mathrm{c}}(\mathbf{x};m) - \inf_{\mathbf{x}\in\mathbb{R}^{2m-1}} f_{\mathcal{N}\mathrm{c}}(\mathbf{x};m) \geq 1/(16m)$.

The above two function classes $f_{\mathcal{N}\mathrm{sc}}$ and $f_{\mathcal{N}\mathrm{c}}$ will be used to prove the lower bounds for convex optimization. Finally we introduce $f_{\mathcal{C}}$, which is original proposed in Carmon et al. (2017b), and we will use it to prove the lower bounds for nonconvex optimization.

Definition 3.10. Let $f_{\mathcal{C}}(\mathbf{x};\alpha,m) : \mathbb{R}^{m+1} \to \mathbb{R}$ be

$$f_{\mathcal{C}}(\mathbf{x};\alpha,m) := Q(\mathbf{x};\sqrt{\alpha},m+1,0) + \alpha\Gamma(\mathbf{x}),$$

where $\Gamma(\mathbf{x}) : \mathbb{R}^{m+1} \to \mathbb{R}$ is defined as

$$\Gamma(\mathbf{x}) := \sum_{i=1}^{m} 120 \int_1^{\mathbf{x}_i} \frac{t^2(t-1)}{1+t^2}dt.$$

We have the following properties about $f_{\mathcal{C}}$.

Proposition 3.11 (Lemmas 2, 3, 4, Carmon et al. (2017b))**.** Let $c_\gamma = 360$. Then for any $0 \leq \alpha \leq 1$, it holds that

1. $\Gamma(\mathbf{x}) \in \mathcal{S}^{(-c_\gamma,c_\gamma)}$ and $f_{\mathcal{C}}(\mathbf{x};\alpha,m) \in \mathcal{S}^{(-\alpha c_\gamma,4+\alpha c_\gamma)}$.
2. $f_{\mathcal{C}}(0;\alpha,m) - \inf_{\mathbf{x}\in\mathbb{R}^{m+1}} f_{\mathcal{C}}(\mathbf{x};\alpha,m) \leq \sqrt{\alpha}/2 + 10\alpha m$.
3. For $\mathbf{x}$ which satisfies that $\mathbf{x}_m = \mathbf{x}_{m+1} = 0$, we have $\|\nabla f_{\mathcal{C}}(\mathbf{x};\alpha,m)\|_2 \geq \alpha^{3/4}/4$.

4. Main Results

In this section we present our lower bound results. We start with the sum-of-nonconvex (but convex) optimization setting, then move on to the general nonconvex finite-sum optimization setting.

4.1. F is Convex – Suboptimal Solution

We first show the result when F is σ-strongly convex and $\{f_i\}_{i=1}^n \in \mathcal{V}^{(L)}$, and our goal is to find an ϵ-suboptimal solution.

Theorem 4.1. For any linear-span randomized first-order algorithm $\mathcal{A}$ and any $L,\sigma,n,\Delta,\epsilon$ such that $\epsilon \leq 8\Delta n^{7/4}\sigma^{3/2}L^{-3/2}$, there exist a dimension $d = O(n + n^{3/4}\sqrt{L/\sigma}\log(1/\epsilon))$ and functions $\{f_i\}_{i=1}^n : \mathbb{R}^d \to \mathbb{R}$ which satisfy that $\{f_i\}_{i=1}^n \in \mathcal{V}^{(L)}$, $F \in \mathcal{S}^{(\sigma,L)}$ and $F(\mathbf{x}^{(0)}) - \inf_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) \leq \Delta$. In order to find $\widehat{\mathbf{x}} \in \mathbb{R}^d$ such that $\mathbb{E}F(\widehat{\mathbf{x}}) - \inf_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) \leq \epsilon$, $\mathcal{A}$ needs at least

$$\Omega\left(n + n^{3/4}\sqrt{\frac{L}{\sigma}}\log\left(\frac{n\sigma\Delta}{L\epsilon}\right)\right) \tag{4.1}$$

IFO calls.

Next we show the result when F is convex and $\{f_i\}_{i=1}^n \in \mathcal{V}^{(L)}$.

Theorem 4.2. For any linear-span randomized first-order algorithm $\mathcal{A}$ and any L,n,B,ϵ such that $\epsilon \leq LB^2/4$ there exist a dimension $d = O(n + n^{3/4}\sqrt{L/\epsilon})$ and functions $\{f_i\}_{i=1}^n : \mathbb{R}^d \to \mathbb{R}$ which satisfy that $\{f_i\}_{i=1}^n \in \mathcal{V}^{(L)}$, $F \in \mathcal{S}^{(0,L)}$, and $\mathrm{dist}(\mathbf{x}^{(0)},\mathcal{X}^*) \leq B$ where $\mathcal{X}^* = \mathrm{argmin}_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x})$. In order to find $\widehat{\mathbf{x}} \in \mathbb{R}^d$ such that $\mathbb{E}F(\widehat{\mathbf{x}}) - \inf_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) \leq \epsilon$, $\mathcal{A}$ needs at least

$$\Omega\left(n + n^{3/4}B\sqrt{\frac{L}{\epsilon}}\right) \tag{4.2}$$

IFO calls.

Remark 4.3. Our lower bounds (4.1) and (4.2) are tight, because they have been achieved by SDCA without Duality (Shalev-Shwartz, 2016) for $l = \sigma$ and KatyushaX (Allen-Zhu, 2018) for $l = \sigma$ and $l = 0$ up to a logarithm factor.

Remark 4.4. It is interesting to compare (4.1) and (4.2) with the corresponding lower bounds for convex finite-sum optimization in Woodworth & Srebro (2016), which proves $\widetilde{\Omega}(n + \sqrt{nL/\sigma})$ lower bound for strongly convex functions and $\Omega(n + \sqrt{nL/\epsilon})$ for convex functions, where each f_i is L-smooth. The dependence on n is $n^{3/4}$ in our lower bounds when $\epsilon \ll 1$, as opposed to $n^{1/2}$ in Woodworth & Srebro (2016). This gap has been observed firstly by Shalev-Shwartz (2016) from the view of upper bounds, and was conjectured to be caused by nonconvexity of each component function f_i. Our lower bound results suggest that this gap is due to the L-smooth on each component function f_i and the L-average smooth on $\{f_i\}_{i=1}^n$, and such a gap cannot be removed.

4.2. F is Nonconvex – Approximate Stationary Point

Next we show the lower bounds when F is σ-almost convex. For this case our goal is to find an ϵ-approximate stationary

point. We first present the lower result when $\{f_i\}_{i=1}^n \in \mathcal{V}^{(L)}$.

Theorem 4.5. For any linear-span randomized first-order algorithm $\mathcal{A}$ and any L, n, Δ, ϵ with $\epsilon^2 \leq (\Delta\sigma \wedge L\Delta n^{-1/2})/10^5$, there exist a dimension $d = O(\Delta/\epsilon^2 \cdot (n^{3/4}\sqrt{\sigma L} \wedge \sqrt{n}L))$ and functions $\{f_i\}_{i=1}^n : \mathbb{R}^d \to \mathbb{R}$ which satisfy that $\{f_i\}_{i=1}^n \in \mathcal{V}^{(L)}$, $F \in \mathcal{S}^{(-\sigma,L)}$ and $F(\mathbf{x}^{(0)}) - \inf_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) \leq \Delta$. In order to find $\widehat{\mathbf{x}} \in \mathbb{R}^d$ such that $\mathbb{E}\|\nabla F(\widehat{\mathbf{x}})\|_2 \leq \epsilon$, $\mathcal{A}$ needs at least

$$\Omega\bigg(\frac{\Delta}{\epsilon^2}\big[n^{3/4}\sqrt{\sigma L} \wedge \sqrt{n}L\big]\bigg) \tag{4.3}$$

IFO calls.

Remark 4.6. Our lower bound (4.3) is tight for the following reasons. (4.3) becomes $\Omega(\Delta/\epsilon^2 \cdot n^{3/4}\sqrt{\sigma L})$ when $\sigma = O(L/\sqrt{n})$, and such IFO complexity has been achieved by RepeatSVRG up to a logarithm factor (Carmon et al., 2018; Agarwal et al., 2017). For the case $\sigma = \Omega(L/\sqrt{n})$, (4.3) becomes $\Omega(\Delta/\epsilon^2 \cdot \sqrt{n}L)$, and such IFO complexity has been achieved by SPIDER (Fang et al., 2018) and SNVRG (Zhou et al., 2018) up to a logarithm factor.

Next we show lower bounds under a slightly stronger assumption that each $f_i \in \mathcal{S}^{(-\sigma,L)}$. Our result shows that with such a stronger assumption, the optimal dependency on n will be smaller.

Theorem 4.7. For any linear-span randomized first-order algorithm $\mathcal{A}$ and any L, n, Δ, ϵ which satisfies that $\epsilon^2 \leq (\Delta L n^{-1} \wedge \Delta\sigma)/10^3$, there exist a dimension $d = O(\Delta/\epsilon^2 \cdot (\sqrt{n\sigma L} \wedge L))$ and functions $\{f_i\}_{i=1}^n : \mathbb{R}^d \to \mathbb{R}$ which satisfy that each $f_i \in \mathcal{S}^{(-\sigma,L)}$ and $F(\mathbf{x}^{(0)}) - \inf_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) \leq \Delta$. In order to find $\widehat{\mathbf{x}} \in \mathbb{R}^d$ such that $\mathbb{E}\|\nabla F(\widehat{\mathbf{x}})\|_2 \leq \epsilon$, $\mathcal{A}$ needs at least

$$\Omega\bigg(\frac{\Delta}{\epsilon^2}\big[\sqrt{n\sigma L} \wedge L\big]\bigg) \tag{4.4}$$

IFO calls.

Remark 4.8. Our lower bound (4.4) is tight for the case $\sigma = O(L/n)$, where (4.4) becomes $\Omega(\Delta/\epsilon^2 \cdot \sqrt{n\sigma l})$. Such IOF complexity has been achieved by Natasha (Allen-Zhu, 2017b), RapGrad (Lan & Yang, 2018) and StagewiseKatyusha (Chen & Yang, 2018) up to a logarithm factor. Nevertheless, for the case $\sigma = \Omega(L/n)$, (4.4) becomes $\Omega(\Delta/\epsilon^2 \cdot L)$, which does not match the best-known upper bound $O(\Delta/\epsilon^2 \cdot \sqrt{n}L)$ (Fang et al., 2018) by a factor of $\sqrt{n}$ on the dependency of n. We leave it as a future work to close this gap.

4.3. Discussion on the Average Smoothness Assumption

Careful readers may have already found that in our Theorems 4.1, 4.2 and 4.5, we only assume that $\{f_i\}_{i=1}^n \in \mathcal{V}^{(L)}$. In other words, the above lower bound results (except Theorem 4.7) hold for $\{f_i\}_{i=1}^n$ that is average smooth. Nevertheless, most of the upper bound results achieved by existing finite-sum optimization algorithms (i.e., SDCA without Duality (Shalev-Shwartz, 2016), Natasha (Allen-Zhu, 2017b), KatyushaX (Allen-Zhu, 2018), RapGrad (Lan & Yang, 2018), StagewiseKatyusha (Chen & Yang, 2018) and RepeatSVRG (Agarwal et al., 2017; Carmon et al., 2018)) are proved under the assumption that $f_i \in \mathcal{S}^{(-L,L)}$ for each $i \in [n]$, which is stronger than assuming $\{f_i\}_{i=1}^n \in \mathcal{V}^{(L)}$, which only appears in Zhou et al. (2018) and Fang et al. (2018). Therefore, it is important to verify that these upper bounds results still hold under the weaker assumption that $\{f_i\}_{i=1}^n$ that is average smooth.

To verify this, we need to rethink about the role that the assumption $f_i \in \mathcal{S}^{(-L,L)}$ for each $i \in [n]$ plays in the convergence analyses for those algorithms. In detail, in the convergence analyses of those nonconvex finite-sum optimization algorithms including SDCA without Duality (Shalev-Shwartz, 2016), Natasha (Allen-Zhu, 2017b), KatyushaX (Allen-Zhu, 2018), one needs the assumption that $f_i \in \mathcal{S}^{(-L,L)}$ for each $i \in [n]$ in the following two scenarios: First, it is used to show that $F \in \mathcal{S}^{(-L,L)}$, which can be derived as follows: for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$\begin{aligned}\|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\|_2^2 &\leq \mathbb{E}_i\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|_2^2 \\ &\leq L^2\|\mathbf{x} - \mathbf{y}\|_2^2. \end{aligned} \tag{4.5}$$

Second, it is used to upper bound the variance of the semi-stochastic gradient at each iteration, which is an unbiased estimator of the true gradient. More specifically, let $\mathbf{v}$ be

$$\mathbf{v} = \nabla f_i(\mathbf{x}) - \nabla f_i(\widehat{\mathbf{x}}) + \nabla F(\widehat{\mathbf{x}}),$$

where $\widehat{\mathbf{x}}$ is the global minimum of F when F is convex or any snapshot of $\mathbf{x}$ when F is nonconvex. Then we have

$$\begin{aligned}&\mathbb{E}_i\|\mathbf{v} - \nabla F(\mathbf{x})\|_2^2 \\ &= \mathbb{E}_i\|\nabla f_i(\mathbf{x}) - \nabla f_i(\widehat{\mathbf{x}}) + \nabla F(\widehat{\mathbf{x}}) - \nabla F(\mathbf{x})\|_2^2 \\ &\leq 2\Big[\mathbb{E}_i\|\nabla f_i(\mathbf{x}) - \nabla f_i(\widehat{\mathbf{x}})\|_2^2 + \|\nabla F(\widehat{\mathbf{x}}) - \nabla F(\mathbf{x})\|_2^2\Big] \\ &\leq 2L^2\|\mathbf{x} - \widehat{\mathbf{x}}\|_2^2. \end{aligned} \tag{4.6}$$

We can see that in both scenarios, the weaker assumption $\{f_i\}_{i=1}^n \in \mathcal{V}^{(L)}$ is sufficient to make (4.5) and (4.6) hold. Thus, we make the following informal statement, which may be regarded as a slight improvement/modification in terms of assumptions over existing algorithms for nonconvex finite-sum optimization problems.

Proposition 4.9. For existing nonconvex finite-sum optimization algorithms including SDCA without Duality (Shalev-Shwartz, 2016), Natasha (Allen-Zhu, 2017b), KatyushaX (Allen-Zhu, 2018), RapGrad (Lan & Yang,

2018), StagewiseKatyusha (Chen & Yang, 2018) and RepeatSVRG (Agarwal et al., 2017; Carmon et al., 2018), we can replace the smoothness assumption that $f_i \in \mathcal{S}^{(-L,L)}$ with $\{f_i\}_{i=1}^n \in \mathcal{V}^{(L)}$, without affecting their IFO complexities.

5. Proof of Main Theorems

In this section, we provide the detailed proofs for the lower bounds presented in Section 4. Due to space limit, we only provide the proofs for Theorems 4.1 and 4.5, and defer the proofs for the other theorems in the supplementary material.

5.1. Technical Lemmas

Our proofs are based on the following three technical lemmas, whose proofs can be found in the supplementary material.

The first lemma provides the upper bound for the average smoothness parameter of finite-sum functions, when each component function is lower and upper smooth.

Lemma 5.1. For any $g : \mathbb{R}^m \to \mathbb{R}$ and $g \in \mathcal{S}^{(\xi,\zeta)}$ where $0 \le |\xi| \le \zeta$, suppose that $\{\mathbf{U}^{(i)}\}_{i=1}^n \in \mathcal{O}(m, mn, n)$. Then for $\bar{g}_i : \mathbb{R}^{mn} \to \mathbb{R}$ where $\bar{g}_i(\mathbf{x}) := \sqrt{n} g(\mathbf{U}^{(i)}\mathbf{x})$, we have that $\{\bar{g}_i\}_{i=1}^n \in \mathcal{V}^{(\zeta)}$. For $\bar{G}(\mathbf{x}) = \sum_{i=1}^n \bar{g}_i(\mathbf{U}^{(i)}\mathbf{x})/n$, we also have $\bar{G} \in \mathcal{S}^{(\xi/\sqrt{n},\zeta)}$.

In the proof we need to do scale transformation to the given functions. The following lemma describes how problem dependent quantities change with respect to scale transformation.

Lemma 5.2. Let $\{\bar{g}_i\}_{i=1}^n, \bar{g}_i : \mathbb{R}^d \to \mathbb{R}$ be functions satisfying $\{\bar{g}_i\}_{i=1}^n \in \mathcal{V}^{(L')}$, $\bar{g}_i \in \mathcal{S}^{(\xi',\zeta')}$. We further define $\bar{G} = \sum_{i=1}^n \bar{g}_i/n$, and $\mathcal{Z}^* = \operatorname{argmin}_{\mathbf{z}\in\mathbb{R}^d} \bar{G}(\mathbf{z})$. Suppose that $\bar{G}(0) - \inf_{\mathbf{x}\in\mathbb{R}^d} \bar{G}(\mathbf{x}) \le \Delta'$ and $\operatorname{dist}(0, \mathcal{Z}^*) \le B'$. For any $\lambda, \beta > 0$, we define $\{g_i\}_{i=1}^n$ satisfying $g_i(\mathbf{x}) = \lambda \bar{g}_i(\mathbf{x}/\beta)$ and $G = \sum_{i=1}^n g_i/n$. Let $(\mathcal{Z}')^* = \operatorname{argmin}_{\mathbf{z}\in\mathbb{R}^d} G(\mathbf{z})$. Then we have that $\{g_i\}_{i=1}^n \in \mathcal{V}^{(\lambda/\beta^2\cdot L')}$, $g_i \in \mathcal{S}^{(\lambda/\beta^2\cdot\xi',\lambda/\beta^2\cdot\zeta')}$, $G(0) - \inf_{\mathbf{x}\in\mathbb{R}^d} G(\mathbf{x}) \le \lambda\Delta'$ and $\operatorname{dist}(0, (\mathcal{Z}')^*) \le \beta B'$.

We also need the following lemma to guarantee an $\Omega(n)$ lower bound for finding an ϵ-suboptimal solution when F is convex.

Lemma 5.3. For any linear-span randomized first-order algorithm $\mathcal{A}$ and any $L, \sigma, n, \Delta, \epsilon$ with $\epsilon < \Delta/4$, there exist functions $\{f_i\}_{i=1}^n : \mathbb{R}^n \to \mathbb{R}$ and $F = \sum_{i=1}^n f_i/n$ which satisfy that $\{f_i\}_{i=1}^n \in \mathcal{V}^{(L)}$, $F \in \mathcal{S}^{(\sigma,L)}$ and $F(\mathbf{x}^{(0)}) - \inf_{\mathbf{x}\in\mathbb{R}^n} F(\mathbf{x}) \le \Delta$. In order to find $\widehat{\mathbf{x}} \in \mathbb{R}^n$ such that $\mathbb{E}F(\widehat{\mathbf{x}}) - \inf_{\mathbf{x}\in\mathbb{R}^n} F(\mathbf{x}) \le \epsilon$, $\mathcal{A}$ needs at least $\Omega(n)$ IFO calls.

We now begin our proof. Without loss of generality, we assume that $\mathbf{x}^{(0)} = \mathbf{0}$, otherwise we can replace function $f(\mathbf{x})$ with $\widehat{f}(\mathbf{x}) = f(\mathbf{x} - \mathbf{x}^{(0)})$.

5.2. Proofs for: F is Convex

Proof of Theorem 4.1. Let $\{\mathbf{U}^{(i)}\}_{i=1}^n \in \mathcal{O}(T, Tn, n)$. We choose $\bar{f}_i(\mathbf{x}) : \mathbb{R}^{Tn} \to \mathbb{R}$ as follows:

$$\bar{f}_i(\mathbf{x}) := \sqrt{n} f_{\mathcal{N}\text{sc}}(\mathbf{U}^{(i)}\mathbf{x}; \alpha, T),$$
$$\bar{F}(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^n \bar{f}_i(\mathbf{x}).$$

First, we claim that $\{\bar{f}_i(\mathbf{x})\}_{i=1}^n \in \mathcal{V}^{(1)}$ and $\bar{F} \in \mathcal{S}^{(\alpha/\sqrt{n},1)}$ due to Lemma 5.1 where $f_{\mathcal{N}\text{sc}} \in \mathcal{S}^{(\alpha,1)}$, $\alpha \le 1$. Next, we claim $\bar{F}(0) - \inf_{\mathbf{x}} \bar{F}(\mathbf{x}) \le 1/\sqrt{n}\sum_{i=1}^n [f_{\mathcal{N}\text{sc}}(0; \alpha, T) - \inf_{\mathbf{x}} f_{\mathcal{N}\text{sc}}(\mathbf{U}^{(i)}\mathbf{x}; \alpha, T)] \le q^2/\sqrt{n}(1-q^2)$, because

$$\begin{aligned}
&\bar{F}(0) - \inf_{\mathbf{x}} \bar{F}(\mathbf{x}) \\
&= \frac{\sqrt{n}}{n}\sum_{i=1}^n f_{\mathcal{N}\text{sc}}(0; \alpha, T) - \inf_{\mathbf{x}} \frac{\sqrt{n}}{n}\sum_{i=1}^n f_{\mathcal{N}\text{sc}}(\mathbf{U}^{(i)}\mathbf{x}; \alpha, T) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n [f_{\mathcal{N}\text{sc}}(0; \alpha, T) - \inf_{\mathbf{x}} f_{\mathcal{N}\text{sc}}(\mathbf{x}; \alpha, T)] \\
&\le \frac{q^2}{\sqrt{n}(1-q^2)},
\end{aligned}$$

where the second equality holds due to the fact that $\inf_{\mathbf{x}} \sum_{i=1}^n f_{\mathcal{N}\text{sc}}(\mathbf{U}^{(i)}\mathbf{x}; \alpha, T) = \sum_{i=1}^n \inf_{\mathbf{x}} f_{\mathcal{N}\text{sc}}(\mathbf{x}; \alpha, T)$. Finally, let $\mathbf{y}^{(i)} = \mathbf{U}^{(i)}\mathbf{x}$. If there exists $\mathcal{I} \subset [n], |\mathcal{I}| > n/2$ and for each $i \in \mathcal{I}$, $\mathbf{y}_T^{(i)} = 0$. Then, by Proposition 3.7, for each $i \in \mathcal{I}$, we have $f_{\mathcal{N}\text{sc}}(\mathbf{y}^{(i)}, \alpha, T) - \inf_{\mathbf{z}} f_{\mathcal{N}\text{sc}}(\mathbf{z}, \alpha, T) \ge \alpha q^{2T+2}/2$, which implies

$$\begin{aligned}
&\bar{F}(\mathbf{x}) - \inf_{\mathbf{z}} \bar{F}(\mathbf{z}) \\
&\ge \frac{1}{\sqrt{n}}\sum_{i\in\mathcal{I}} [f_{\mathcal{N}\text{sc}}(\mathbf{y}^{(i)}, \alpha, T) - \inf_{\mathbf{z}} f_{\mathcal{N}\text{sc}}(\mathbf{z}, \alpha, T)] \\
&\ge \alpha\sqrt{n} q^{2T+2}/2. \qquad (5.1)
\end{aligned}$$

With the above properties, we can choose $f_i(\mathbf{x}) = \lambda \bar{f}_i(\mathbf{x}/\beta)$ in the following proof. We first consider any fixed index sequence $\{i_t\}$. In the sequel, we consider two cases: (1) $\sqrt{n}\sigma/L \le 1/4$; and (2) $\sqrt{n}\sigma/L > 1/4$.

Case (1): $\sqrt{n}\sigma/L \le 1/4$, we set $\alpha, \lambda, \beta, T$ as follows

$$\begin{aligned}
\alpha &= \frac{\sqrt{n}\sigma}{L} \\
\lambda &= \frac{4\sqrt{n\alpha}\Delta}{(1-\sqrt{\alpha})^2} \\
\beta &= \sqrt{\lambda/L} \\
T &= \sqrt{\frac{L}{\sqrt{n}\sigma}} \cdot \log\left[\left(\frac{\sigma}{L}\right)^{3/2} \frac{8n^{7/4}\Delta}{\epsilon}\right].
\end{aligned}$$

Then by Lemma 5.2, we have that $\{f_i\}_{i=1}^n \in \mathcal{V}^{(L)}$, $F \in \mathcal{S}^{(\sigma,L)}$, $F(\mathbf{0}) - \inf_{\mathbf{z}} F(\mathbf{z}) \leq \Delta$ due to $\sqrt{n}\sigma/L \leq 1/4$. By Proposition 3.5, we know that for any algorithm output $\mathbf{x}^{(t)}$ where t is less than

$$\frac{nT}{2} = n^{3/4}\sqrt{\frac{L}{\sigma}}\log\left[\left(\frac{\sigma}{L}\right)^{3/2}\frac{8n^{7/4}\Delta}{\epsilon}\right], \tag{5.2}$$

there exists $\mathcal{I} \subset [n], |\mathcal{I}| > n - nT/(2T) = n/2$ and for each $i \in \mathcal{I}$, $\mathbf{y}_T^{(i)} = 0$, where $\mathbf{y}^{(i)} = \mathbf{U}^{(i)}\mathbf{x}^{(t)}$. Thus, $\mathbf{x}^{(t)}$ satisfies

$$F(\mathbf{x}^{(t)}) - \inf_{\mathbf{z}} F(\mathbf{z}) \geq \lambda\alpha\sqrt{n}q^{2T+2}/2 \geq \epsilon,$$

where the first inequality holds due to (5.1). We now have proved that for any fixed index sequence $\{i_t\}$, the output $\mathbf{x}^{(T)}$ of a deterministic linear-span algorithm is not an ϵ-suboptimal solution, i.e. $F(\mathbf{x}^{(T)}) - F^* > \epsilon$. Then, applying Yao's minimax theorem[4], we have that for any randomized index sequence $\{i_t\}$, we have the lower bound (5.2).

Case (2): $\sqrt{n}\sigma/L > 1/4$, by Lemma 5.3 we know that there exists an $\Omega(n)$ lower bound.

By combining Cases (1) and (2), we have the lower bound (4.1). □

5.3. Proofs for: F is Nonconvex

Proof of Theorem 4.5. Let $\{\mathbf{U}^{(i)}\}_{i=1}^n \in \mathcal{O}(T+1, (T+1)n, n)$. We choose $\bar{f}_i(\mathbf{x}) : \mathbb{R}^{Tn} \to \mathbb{R}$ as follows:

$$\begin{aligned}\bar{f}_i(\mathbf{x}) &:= \sqrt{n} f_{\mathcal{C}}(\mathbf{U}^{(i)}\mathbf{x}; \alpha, T),\\ \bar{F}(\mathbf{x}) &:= \frac{1}{n}\sum_{i=1}^n \bar{f}_i(\mathbf{x}).\end{aligned} \tag{5.3}$$

We have the following properties. First, we claim that $\{\bar{f}_i(\mathbf{x})\}_{i=1}^n \in \mathcal{V}^{(4+\alpha c_\gamma)}$ and $\bar{F}(\mathbf{x}) \in \mathcal{S}^{(-\alpha c_\gamma/\sqrt{n}, 4+\alpha c_\gamma)}$ by Lemma 5.1 where $f_{\mathcal{C}} \in \mathcal{S}^{(-\alpha c_\gamma, 4+\alpha c_\gamma)}$ and $\alpha c_\gamma < 4 + \alpha c_\gamma$. Next, we have

$$\begin{aligned}&\bar{F}(0) - \inf_{\mathbf{x}} \bar{F}(\mathbf{x})\\ &\leq 1/\sqrt{n}\sum_{i=1}^n [f_{\mathcal{N}\text{sc}}(0; \alpha, T) - \inf_{\mathbf{x}} f_{\mathcal{N}\text{sc}}(\mathbf{U}^{(i)}\mathbf{x}; \alpha, T)]\\ &\leq \sqrt{n}(\sqrt{\alpha} + 10\alpha T).\end{aligned} \tag{5.4}$$

Finally, let $\mathbf{y}^{(i)} = \mathbf{U}^{(i)}\mathbf{x}$. If there exists $\mathcal{I}, |\mathcal{I}| > n/2$ and for each $i \in \mathcal{I}$, $\mathbf{y}_T^{(i)} = \mathbf{y}_{T+1}^{(i)} = 0$, then by Proposition 3.11, we have

$$\begin{aligned}\|\nabla \bar{F}(\mathbf{x})\|_2^2 &\geq \frac{1}{n}\sum_{i\in\mathcal{I}} \|(\mathbf{U}^{(i)})^\top \nabla[f_{\mathcal{C}}(\mathbf{U}^{(i)}\mathbf{x}; \alpha, T)]\|_2^2\\ &\geq \frac{1}{n}\frac{n}{2}(\alpha^{3/4}/4)^2\\ &= \alpha^{3/2}/32.\end{aligned} \tag{5.5}$$

With above properties, we choose $f_i(\mathbf{x}) = \lambda\bar{f}_i(\mathbf{x}/\beta)$ in the following proof. We first consider any fixed index sequence $\{i_t\}$. We set $\alpha, \lambda, \beta, T$ as

$$\begin{aligned}\alpha &= \min\left\{\frac{5\sigma\sqrt{n}}{c_\gamma L}, \frac{1}{c_\gamma}\right\}\\ \lambda &= \frac{5\epsilon^2}{L\alpha^{3/2}}\\ \beta &= \sqrt{5\lambda/L}\\ T &= \frac{L\Delta}{55\sqrt{n}\epsilon^2}\sqrt{\min\left\{\frac{5\sigma\sqrt{n}}{c_\gamma L}, \frac{1}{c_\gamma}\right\}},\end{aligned}$$

Then by Lemma 5.2, we have that $\{f_i\}_{i=1}^n \in \mathcal{V}^{(L)}$, $F \in \mathcal{S}^{(\sigma,L)}$, $F(0) - \inf_{\mathbf{z}} F(\mathbf{z}) \leq \Delta$ with the assumption that $\epsilon^2 \leq L\alpha\Delta/(55\sqrt{n})$. By Proposition 3.5, we know that for any algorithm output $\mathbf{x}^{(t)}$ where t is less than

$$\frac{nT}{2} = \frac{L\sqrt{n}\Delta}{110\epsilon^2}\sqrt{\min\left\{\frac{5\sigma\sqrt{n}}{c_\gamma L}, \frac{1}{c_\gamma}\right\}}, \tag{5.6}$$

there exists $\mathcal{I} \subset [n], |\mathcal{I}| > n - nT/(2T) = n/2$ and for each i, $\mathbf{y}_T^{(i)} = \mathbf{y}_{T+1}^{(i)} = 0$ where $\mathbf{y}^{(i)} = \mathbf{U}^{(i)}\mathbf{x}^{(t)}$. Thus, by (5.5), $\mathbf{x}^{(t)}$ satisfies

$$\|\nabla F(\mathbf{x}^{(t)})\|_2 \geq \lambda/\beta \cdot \sqrt{\alpha^{3/2}/32} \geq \epsilon.$$

Then, applying Yao's minimax theorem, we have that for any randomized index sequence $\{i_t\}$, we have the lower bound (5.6), which implies (4.3). □

6. Conclusions and Future Work

In this paper we proved the lower bounds of IFO complexity for linear-span randomized first-order algorithms to find ϵ-suboptimal points or ϵ-approximate stationary points for smooth nonconvex finite-sum optimization, where the objective function is the average of n nonconvex functions. While our lower bound results are proved for linear-span randomized first-order algorithms, they can be extended to more general randomized algorithms without linear-span assumption (Woodworth & Srebro, 2016; Carmon et al., 2017a; Fang et al., 2018). We leave it as a future work. On the other hand, we would like to consider more general setting, such as F is of (σ, L)-smoothness while each f_i is (l, L)-smoothness. We are also interested in proving lower bound results for high-order finite-sum optimization problems (Arjevani et al.; Agarwal & Hazan, 2017).

[4] Yao's minimax principle (Yao, 1977) states that the expected cost of a randomized algorithm on the worst case input, is lower bounded by the expected cost of the deterministic algorithm against a worst-case probability distribution on the inputs.

Acknowledgement

We would like to thank the anonymous reviewers for their helpful comments. We would also like to thank the anonymous Area Chair and Ohad Shamir for pointing out that the $n^{3/4}$ factor in the lower bound of sum-of-nonconvex optimization is due to the average smoothness rather than the nonconvexity of each component function. This research was sponsored in part by the National Science Foundation IIS-1906169. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

Agarwal, A. and Bottou, L. A lower bound for the optimization of finite sums. In *International Conference on Machine Learning*, pp. 78–86, 2015.

Agarwal, N. and Hazan, E. Lower bounds for higher-order convex optimization. *arXiv preprint arXiv:1710.10329*, 2017.

Agarwal, N., Allen-Zhu, Z., Bullins, B., Hazan, E., and Ma, T. Finding approximate local minima faster than gradient descent. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1195–1199. ACM, 2017.

Allen-Zhu, Z. Katyusha: The first direct acceleration of stochastic gradient methods. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1200–1205. ACM, 2017a.

Allen-Zhu, Z. Natasha: Faster non-convex stochastic optimization via strongly non-convex parameter. *arXiv preprint arXiv:1702.00763*, 2017b.

Allen-Zhu, Z. Katyusha x: Practical momentum method for stochastic sum-of-nonconvex optimization. *arXiv preprint arXiv:1802.03866*, 2018.

Allen-Zhu, Z. and Li, Y. Lazysvd: Even faster svd decomposition yet without agonizing pain. In *Advances in Neural Information Processing Systems*, pp. 974–982, 2016.

Arjevani, Y. and Shamir, O. Dimension-free iteration complexity of finite sum optimization problems. In *Advances in Neural Information Processing Systems*, pp. 3540–3548, 2016.

Arjevani, Y., Shamir, O., and Shiff, R. Oracle complexity of second-order methods for smooth convex optimization. *Mathematical Programming*, pp. 1–34.

Bietti, A. and Mairal, J. Stochastic optimization with variance reduction for infinite datasets with finite sum structure. In *Advances in Neural Information Processing Systems*, pp. 1622–1632, 2017.

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

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. Lower bounds for finding stationary points ii: First-order methods. *arXiv preprint arXiv:1711.00841*, 2017b.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1751–1772, 2018.

Chen, Z. and Yang, T. A variance reduction method for non-convex optimization with improved convergence under large condition number. *arXiv preprint arXiv:1809.06754*, 2018.

Defazio, A., Bach, F., and Lacoste-Julien, S. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pp. 1646–1654, 2014.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, pp. 687–697, 2018.

Garber, D., Hazan, E., Jin, C., Kakade, S. M., Musco, C., Netrapalli, P., and Sidford, A. Faster eigenvector computation via shift-and-invert preconditioning. In *ICML*, pp. 2626–2634, 2016.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pp. 315–323, 2013.

Lan, G. and Yang, Y. Accelerated stochastic algorithms for nonconvex finite-sum and multi-block optimization. *arXiv preprint arXiv:1805.05411*, 2018.

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

Mairal, J. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015.

Murty, K. G. and Kabadi, S. N. Some np-complete problems in quadratic and nonlinear programming. *Mathematical programming*, 39(2):117–129, 1987.

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

Nesterov, Y. E. A method for solving the convex programming problem with convergence rate $o(1/ksp2)$. *Dokl.akad.nauk Sssr*, (3):543–547, 1983.

Roux, N. L., Schmidt, M., and Bach, F. R. A stochastic gradient method with an exponential convergence _rate for finite training sets. In *Advances in Neural Information Processing Systems*, pp. 2663–2671, 2012.

Shalev-Shwartz, S. Sdca without duality. *arXiv preprint arXiv:1502.06177*, 2015.

Shalev-Shwartz, S. Sdca without duality, regularization, and individual convexity. In *International Conference on Machine Learning*, pp. 747–754, 2016.

Woodworth, B. E. and Srebro, N. Tight complexity bounds for optimizing composite objectives. In *Advances in neural information processing systems*, pp. 3639–3647, 2016.

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

Yao, A. C.-C. Probabilistic computations: Toward a unified measure of complexity. In *Foundations of Computer Science, 1977., 18th Annual Symposium on*, pp. 222–227. IEEE, 1977.

Zhou, D., Xu, P., and Gu, Q. Stochastic nested variance reduced gradient descent for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pp. 3925–3936, 2018.