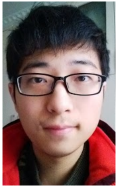
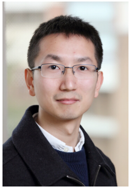
# Lower Bounds for Smooth Nonconvex Finite-Sum Optimization





Dongruo Zhou Quanquan Gu

Computer Science Department
University of California, Los Angeles

# Problem Setup

- **Nonconvex finite-sum optimization:**

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

- $F(\mathbf{x})$ is of $(l, L)$-smoothness, $l \in \mathbb{R}$ and $L > 0$,

$$\frac{l}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 \le F(\mathbf{x}) - F(\mathbf{y}) - \langle \nabla F(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle \le \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|_2^2, \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

# Problem Setup

- **Nonconvex finite-sum optimization:**

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

▷ $F(\mathbf{x})$ is of $(l, L)$-smoothness, $l \in \mathbb{R}$ and $L > 0$,

$$\frac{l}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 \le F(\mathbf{x}) - F(\mathbf{y}) - \langle \nabla F(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle \le \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|_2^2, \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

- **Optimization goals:**

▷ For $l \ge 0$, the goal is to find an $\epsilon$-suboptimal solution $\widehat{\mathbf{x}}$,

$$F(\widehat{\mathbf{x}}) - \inf_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) \le \epsilon.$$

▷ For $l < 0$, the goal is to find an $\epsilon$-stationary point $\widehat{\mathbf{x}}$,

$$\|\nabla F(\widehat{\mathbf{x}})\|_2 \le \epsilon.$$

# Definitions

- **Optimization oracle:** *Incremental First-order Oracle (IFO)*
  - Given $\mathbf{x} \in \mathbb{R}^d$ and $i \in [n]$, an IFO returns $[f_i(\mathbf{x}), \nabla f_i(\mathbf{x})]$.

# Definitions

- **Optimization oracle:** *Incremental First-order Oracle (IFO)*
- ▷ Given $\mathbf{x} \in \mathbb{R}^d$ and $i \in [n]$, an IFO returns $[f_i(\mathbf{x}), \nabla f_i(\mathbf{x})]$.
- **Algorithm class:** *Linear-span first-order randomized algorithms*
- ▷ Given an initial point $\mathbf{x}^{(0)}$.
- ▷ $\mathcal{A} : \{f_i\}_{i=1}^n \to \{\mathbf{x}_t, i_t\}_{t=0}^\infty$ 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 index $i_t \in [n]$, which satisfies

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

# Definitions

- **Optimization oracle:** *Incremental First-order Oracle (IFO)*
  - ▷ Given $\mathbf{x} \in \mathbb{R}^d$ and $i \in [n]$, an IFO returns $[f_i(\mathbf{x}), \nabla f_i(\mathbf{x})]$.
- **Algorithm class:** *Linear-span first-order randomized algorithms*
  - ▷ Given an initial point $\mathbf{x}^{(0)}$.
  - ▷ $\mathcal{A} : \{f_i\}_{i=1}^n \to \{\mathbf{x}_t, i_t\}_{t=0}^\infty$ 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 index $i_t \in [n]$, which satisfies

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

- **Questions:**
  - ▷ *Are existing algorithms (KatyushaX, RapGrad, ...) already optimal?*
  - ▷ *What is the lower bound of IFO complexity for any linear-span first-order randomized algorithm to find $\epsilon$-suboptimal solution or stationary point?*

# Smoothness Assumption

- **Smoothness:** 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 $f \in \mathcal{S}^{(l,L)}$.

- **Average smoothness:** 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 for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$

$$\mathbb{E}_i\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|_2^2 \leq L^2\|\mathbf{x} - \mathbf{y}\|_2^2,$$

  where $\mathbb{E}_i X(i) = 1/n \cdot \sum_{i=1}^n X(i)$ for any random variable $X(i)$. We denote $\{f_i\}_{i=1}^n \in \mathcal{V}^{(L)}$.

- Let $\Delta = F(\mathbf{x}^{(0)}) - \inf_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x})$, $B = \min_{\mathbf{x}\in\mathcal{X}^*} \|\mathbf{x} - \mathbf{x}^{(0)}\|_2$, where $\mathcal{X}^* = \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x})$.
- $F \in \mathcal{S}^{(\sigma,L)}$ or $F \in \mathcal{S}^{(0,L)}$, $\sigma > 0$, find an $\epsilon$-suboptimal solution.
- The lower bounds are tight.

| $\epsilon$-suboptimal solution | $(\sigma, L), \{f_i\} \in \mathcal{V}^{(L)}$ | $(0, L), \{f_i\} \in \mathcal{V}^{(L)}$ |
|---|---|---|
| Upper Bounds | $O\left(\left(n + n^{3/4}\sqrt{\frac{L}{\sigma}}\right)\log\frac{\Delta}{\epsilon}\right)$ (Allen-Zhu, 2018) | $O\left(n + n^{3/4}B\sqrt{\frac{L}{\epsilon}}\right)$ (Allen-Zhu, 2018) |
| Lower Bounds | $\Omega\left(n + n^{3/4}\sqrt{\frac{L}{\sigma}}\log\frac{\Delta}{\epsilon}\right)$ (This work) | $\Omega\left(n + n^{3/4}B\sqrt{\frac{L}{\epsilon}}\right)$ (This work) |

# Lower Bound Results – Nonconvex Case

- $F \in \mathcal{S}^{(-\sigma,L)}$, find an $\epsilon$- stationary point.
- The lower bounds are tight in most regime of parameters.

| $\epsilon$-stationary point | $(-\sigma, L), \{f_i\} \in \mathcal{V}^{(L)}$ | $(-\sigma, L), f_i \in \mathcal{S}^{(-\sigma,L)}$ |
|---|---|---|
| Upper Bounds | $\widetilde{O}\left(\frac{\Delta}{\epsilon^2}(n^{3/4}\sqrt{\sigma L} \wedge \sqrt{n}L)\right)$ (Allen-Zhu, 2017b) (Zhou et al., 2018) | $\widetilde{O}\left(\frac{\Delta}{\epsilon^2}(n\sigma + \sqrt{n\sigma L}) \wedge \sqrt{n}L\right)$ (Lan and Yang, 2018) (Zhou et al, 2018) |
| Lower Bounds | $\Omega\left(\frac{\Delta}{\epsilon^2}(n^{3/4}\sqrt{\sigma L} \wedge \sqrt{n}L)\right)$ (This work) | $\Omega\left(\frac{\Delta}{\epsilon^2}(\sqrt{n\sigma L} \wedge L)\right)$ (This work) |

# Overview of Proof Technique

- We want to construct the following adversarial functions to show the lower bounds of IFO complexity.

# Overview of Proof Technique

- We want to construct the following adversarial functions to show the lower bounds of IFO complexity.
- **Quadratic function class:** For any $0 \leq \xi, \zeta \leq 1$, we define $Q(\mathbf{x}; \xi, m, \zeta) : \mathbb{R}^m \to \mathbb{R}$ as

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

# Overview of Proof Technique

- We want to construct the following adversarial functions to show the lower bounds of IFO complexity.
- **Quadratic function class:** For any $0 \leq \xi, \zeta \leq 1$, we define $Q(\mathbf{x}; \xi, m, \zeta) : \mathbb{R}^m \to \mathbb{R}$ as

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

  - $Q(\mathbf{x}; \xi, m, \zeta) \in \mathcal{S}^{(0,4)}$.
  - 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 $\nabla[Q(\mathbf{U}\bar{\mathbf{x}}; \xi, m, \zeta) + \sum_{i=1}^m \mu(\bar{\mathbf{x}}^\top \mathbf{u}^{(i)})] \in \mathrm{Lin}\{\mathbf{u}^{(1)}, ..., \mathbf{u}^{(t+1)}\}$.

# Lower Bound Function Class

- **Strongly convex case:** $f_{N\mathsf{sc}}$ (Nesterov, 2014)
- For $0 \le \alpha \le 1$, we define $f_{N\mathsf{sc}}(\mathbf{x};\alpha,m):\mathbb{R}^m \to \mathbb{R}$ as

$$f_{N\mathsf{sc}}(\mathbf{x};\alpha,m) := \frac{1-\alpha}{4}Q\left(\mathbf{x};1,m,\frac{2\sqrt{\alpha}}{\sqrt{\alpha}+1}\right) + \frac{\alpha}{2}\|\mathbf{x}\|_2^2.$$

# Lower Bound Function Class

- **Strongly convex case:** $f_{N\mathsf{sc}}$ (Nesterov, 2014)

▷ For $0 \leq \alpha \leq 1$, we define $f_{N\mathsf{sc}}(\mathbf{x}; \alpha, m) : \mathbb{R}^m \to \mathbb{R}$ as

$$f_{N\mathsf{sc}}(\mathbf{x}; \alpha, m) := \frac{1-\alpha}{4} Q\left(\mathbf{x}; 1, m, \frac{2\sqrt{\alpha}}{\sqrt{\alpha}+1}\right) + \frac{\alpha}{2}\|\mathbf{x}\|_2^2.$$

- **Convex case:** $f_{N\mathsf{c}}$ (Nesterov, 2014)

▷ We define $f_{N\mathsf{c}}(\mathbf{x}; m) : \mathbb{R}^{2m-1} \to 1$ as

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

# Lower Bound Function Class

- **Strongly convex case:** $f_{N\text{sc}}$ (Nesterov, 2014)
  - ▷ For $0 \leq \alpha \leq 1$, we define $f_{N\text{sc}}(\mathbf{x}; \alpha, m) : \mathbb{R}^m \to \mathbb{R}$ as

$$f_{N\text{sc}}(\mathbf{x}; \alpha, m) := \frac{1-\alpha}{4} Q\left(\mathbf{x}; 1, m, \frac{2\sqrt{\alpha}}{\sqrt{\alpha}+1}\right) + \frac{\alpha}{2}\|\mathbf{x}\|_2^2.$$

- **Convex case:** $f_{N\text{c}}$ (Nesterov, 2014)
  - ▷ We define $f_{N\text{c}}(\mathbf{x}; m) : \mathbb{R}^{2m-1} \to 1$ as

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

- **Nonconvex case:** $f_{\mathcal{C}}$ (Carmon et al., 2017b)
  - ▷ For $0 \leq \alpha \leq 1$, we define $f_{\mathcal{C}}(\mathbf{x}; \alpha, m) : \mathbb{R}^{m+1} \to \mathbb{R}$ as

$$f_{\mathcal{C}}(\mathbf{x}; \alpha, m) : = Q(\mathbf{x}; \sqrt{\alpha}, m+1, 0) + \alpha\Gamma(\mathbf{x}), \ \Gamma(\mathbf{x}) := \sum_{i=1}^{m} 120 \int_1^{\mathbf{x}_i} \frac{t^2(t-1)}{1+t^2} dt.$$

# Thank you!

Poster session:
**Tue Jun 11th 06:30 – 09:00 PM**
PM @ Pacific Ballroom 94