Complexity of Highly Parallel Non-Smooth Convex Optimization

Sébastien Bubeck
Qijia Jiang
Yin Tat Lee
Yuanzhi Li
Aaron Sidford

arXiv:1906.10655
Spotlight NeurIPS 2019
Non-smooth Convex Optimization

Goal

• Compute "\( \epsilon \)-optimal point": \( f(x) \leq OPT + \epsilon \)

Assumptions

• \( f \) is convex
• \( f \) is 1-Lipschitz, \( \|f(y) - f(x)\| \leq \|y - x\| \)
• \( OPT = f(x_*) \) where \( \|x_*\| \leq 1 \)

Query: \( x \in \mathbb{R}^d \)

First Order Oracle

\( f(x), \nabla f(x) \)

\[
\min_{x \in \mathbb{R}^d} f(x)
\]
Algorithms

(Sub)-Gradient Descent
- $x_{k+1} = x_k - \eta \nabla f(x_k)$
- Output average $\bar{x}_k = \frac{1}{k} \sum x_k$
- $O(1/\epsilon^2)$ queries suffice

Cutting Plane Methods
- Center of gravity / high dimensional binary search.
- $O(d \log(1/\epsilon))$ queries suffice

Unimprovable when
$\epsilon = \omega(1/\sqrt{d})$

Unimprovable when
$\epsilon = O(1/\sqrt{d})$

Parallelizable?

Oracle: first order
Assumptions: 1-Lipschitz, $\|x_*\|_2 \leq 1$

Optimal?
Yes!
Parallel Non-smooth Convex Optimization

**Assumptions**
- $f$ is convex and 1-Lipschitz
- $OPT = f(x_*)$ for $||x_*|| \leq 1$

**Goal**
- Compute $\epsilon$-optimal point

**Parallel Complexity**
- **Depth**: # queries to parallel oracle
- **Work**: # gradients computed / functions evaluated

**This Talk**
- Focus on “highly parallel setting”
- $k = \text{poly}(d)$, work $= \text{poly}(d)$

---

Minimize $\min_{x \in \mathbb{R}^d} f(x)$

Query: $x \in \mathbb{R}^d$ to First Order Oracle
- $f(x), \nabla f(x)$

Parallel First Order Oracle
- Query: $x_1, \ldots, x_k \in \mathbb{R}^d$
- $f(x_1), \nabla f(x_1), \ldots, f(x_k), \nabla f(x_k)$

[Nemirovski 1994]
State-of-the-Art

(Sub)-Gradient Descent
Depth $O(1/\varepsilon^2)$

[DBW12]
Depth $O(d^{1/4}/\varepsilon)$

Cutting Plane Methods
Depth $O(d \log(1/\varepsilon))$

Best known when:
$\varepsilon \geq d^{-1/4}$
depth $\leq \sqrt{d}$

This is the regime when is optimal.

$\varepsilon \in [d^{-3/4}, d^{-1/4}]$
depth $\in [\sqrt{d}, d]$

$\varepsilon < d^{-3/4}$
depth $> d$

Why?
Accelerated stochastic method

Goal: $\varepsilon$-optimal point for convex $f$
Oracle: highly parallel first order
Assumptions: 1-Lipschitz, $\|x_*\|_2 \leq 1$

Sequential: best when $\varepsilon < 1/\sqrt{d}$
Parallel: best known when $\varepsilon < 1/d$

Our Results
High order accelerated stochastic method

Depth $\widetilde{O}(d^{1/3}/\varepsilon^{2/3})$

Best known when
$\varepsilon \in [d^{-1}, d^{-1/4}]$, depth $\in [\sqrt{d}, d]$

No randomized algorithm improves when $\varepsilon = \tilde{\omega}(d^{-1/3})$, depth $= \tilde{O}(d^{1/3})$

No randomized algorithm improves when $\varepsilon = \tilde{\omega}(d^{-1/4})$, depth $= \tilde{O}(\sqrt{d})$