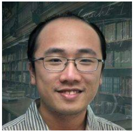
A Gradient Sampling Method with Complexity Guarantees for Lipschitz Functions in Low and High Dimensions

Damek Davis [1], Dmitriy Drusvyatskiy [2], Yin Tat Lee [2], Swati Padmanabhan[2], Guanghao Ye[3]

[1]Cornell University; [2]University of Washington, Seattle;
[3]Massachussetts Institute of Technology
Authors ordered alphabetically

NeurIPS 2022 (Oral)





Guiding Research Question

Given an optimization problem with black-box oracle access, can we obtain improved complexity guarantees for approximately solving it?

Guiding Research Question

Given an optimization problem with black-box oracle access, can we obtain improved complexity guarantees for approximately solving it?

Talk outline:

1. A faster algorithm for a general nonconvex nonsmooth problem
2. Improved rates of the above result for a special case

The Subgradient Method: Background

black-box optimization

with first-order oracle

Gradient-based methods are ubiquitous in optimization

A typical template is the subgradient method:

$$x_{t+1} = x_t - \sum_{i \le t} \alpha_{i,t} \cdot v_i, \text{ for } v_i \in \partial f(x_i),$$

where the set $\partial f(x)$ is the *Clarke subdifferential*:

$$\partial f(x) = \operatorname{conv}\{\lim_{i\to\infty} \nabla f(x_i) : x_i \to x, x_i \in \operatorname{dom}(f)\}.$$

The Subgradient Method: Background

black-box optimization

with first-order oracle

Gradient-based methods are ubiquitous in optimization

A typical template is the subgradient method:

$$\boxed{x_{t+1} = x_t - \textstyle\sum_{i \le t} \alpha_{i,t} \cdot v_i, \text{ for } v_i \in \partial f(x_i),}$$

where the set $\partial f(x)$ is the *Clarke subdifferential*:

$$\boxed{\partial f(x) = \operatorname{conv}\{\lim_{i\to\infty} \nabla f(x_i) : x_i \to x, x_i \in \operatorname{dom}(f)\}.}$$

gradient for smooth f

The Subgradient Method: Background

black-box optimization

with first-order oracle

Gradient-based methods are ubiquitous in optimization

A typical template is the subgradient method:

$$\boxed{x_{t+1} = x_t - \textstyle\sum_{i \leq t} \alpha_{i,t} \cdot v_i, \text{ for } v_i \in \partial f(x_i),}$$

where the set $\partial f(x)$ is the *Clarke subdifferential*:

$$\boxed{\partial f(x) = \operatorname{conv}\left\{\lim_{i \to \infty} \nabla f(x_i) : x_i \to x, x_i \in \operatorname{dom}(f)\right\}.}$$

subdifferential for nonsmooth f

The Subgradient Method: Convergence Guarantees

The subgradient method:

oracle access

$$\boxed{x_{t+1} = x_t - \sum_{i \le t} \alpha_{i,t} \cdot v_i, \text{ for } v_i \in \partial f(x_i).}$$

The Subgradient Method: Convergence Guarantees

The subgradient method:

oracle access

$$\boxed{x_{t+1} = x_t - \textstyle\sum_{i \le t} \alpha_{i,t} \cdot v_i, \text{ for } v_i \in \partial f(x_i).}$$

✓ Nonasymptotic guarantees for convex problems

global function error bound

The Subgradient Method: Convergence Guarantees

The subgradient method:

oracle access

$$\boxed{x_{t+1} = x_t - \textstyle\sum_{i\le t} \alpha_{i,t} \cdot v_i, \text{ for } v_i \in \partial f(x_i).}$$

- ✓ Nonasymptotic guarantees for convex problems
- ✓ Nonasymptotic guarantees for smooth nonconvex problems

gradient norm bound

(stationary point)

The Subgradient Method: Convergence Guarantees

The subgradient method:

oracle access

$$\boxed{x_{t+1} = x_t - \sum_{i\leq t} \alpha_{i,t} \cdot v_i, \text{ for } v_i \in \partial f(x_i).}$$

- ✓ Nonasymptotic guarantees for convex problems
- ✓ Nonasymptotic guarantees for smooth nonconvex problems
- ✓ Asymptotic guarantees for nonsmooth nonconvex problems:
  - Benaim, Hofbauer, Sorin (2005)
  - Kiwiel (2007)
  - Majewski, Miasojedow, Moulines (2018)
  - Davis & Drusvyatskiy (2019)
  - Bolte & Pauwels (2019)

stationary point

(specified later)

The Subgradient Method: Convergence Guarantees

The subgradient method:

oracle access

$$\boxed{x_{t+1} = x_t - \textstyle\sum_{i\leq t} \alpha_{i,t} \cdot v_i, \text{ for } v_i \in \partial f(x_i).}$$

- ✓ Nonasymptotic guarantees for convex problems
- ✓ Nonasymptotic guarantees for smooth nonconvex problems
- ✓ Asymptotic guarantees for nonsmooth nonconvex problems
- ▶ Nonasymptotic guarantees for nonsmooth nonconvex problems?
  - ▶ Breakthrough by Zhang, Lin, Jegelka, Sra, Jadbabaie (2020):

The Subgradient Method: Convergence Guarantees

The subgradient method:

oracle access

$$\boxed{x_{t+1} = x_t - \sum_{i \le t} \alpha_{i,t} \cdot v_i, \text{ for } v_i \in \partial f(x_i).}$$

- ✓ Nonasymptotic guarantees for convex problems
- ✓ Nonasymptotic guarantees for smooth nonconvex problems
- ✓ Asymptotic guarantees for nonsmooth nonconvex problems
- ▶ Nonasymptotic guarantees for nonsmooth nonconvex problems?
  - ▶ Breakthrough by Zhang, Lin, Jegelka, Sra, Jadbabaie (2020): However, their algorithm uses an unusually strong oracle

The Subgradient Method: Convergence Guarantees

The subgradient method:

oracle access

$$\boxed{x_{t+1} = x_t - \textstyle\sum_{i\leq t} \alpha_{i,t} \cdot v_i, \text{ for } v_i \in \partial f(x_i).}$$

- ✓ Nonasymptotic guarantees for convex problems
- ✓ Nonasymptotic guarantees for smooth nonconvex problems
- ✓ Asymptotic guarantees for nonsmooth nonconvex problems
- ▶ Nonasymptotic guarantees for nonsmooth nonconvex problems?
  - ▶ Breakthrough by Zhang, Lin, Jegelka, Sra, Jadbabaie (2020): However, their algorithm uses an unusually strong oracle

No nonasymptotic guarantees for nonsmooth nonconvex problems!

The Subgradient Method: Convergence Guarantees

The subgradient method:

oracle access

$$\boxed{x_{t+1} = x_t - \sum_{i \leq t} \alpha_{i,t} \cdot v_i, \text{ for } v_i \in \partial f(x_i).}$$

- ✓ Nonasymptotic guarantees for convex problems
- ✓ Nonasymptotic guarantees for smooth nonconvex problems
- ✓ Asymptotic guarantees for nonsmooth nonconvex problems
- ▶ Nonasymptotic guarantees for nonsmooth nonconvex problems?
  - ▶ Breakthrough by Zhang, Lin, Jegelka, Sra, Jadbabaie (2020): However, their algorithm uses an unusually strong oracle

No nonasymptotic guarantees for nonsmooth nonconvex problems!

deep learning

A Meaningful Notion of Convergence

Problem Class:

Nonsmooth Nonconvex

A Meaningful Notion of Convergence

Problem Class:

Nonsmooth Nonconvex

- Cannot bound global function error

A Meaningful Notion of Convergence

Problem Class:

Nonsmooth Nonconvex

- Cannot bound global function error
- Cannot attain ϵ-stationarity (Zhang et al (2020))

A Meaningful Notion of Convergence

Problem Class:

Nonsmooth Nonconvex

- Cannot bound global function error
- Cannot attain ϵ-stationarity (Zhang et al (2020))
- Cannot attain near-ϵ-stationarity (Kornowski & Shamir (2022))

A Meaningful Notion of Convergence

Problem Class:

Nonsmooth Nonconvex

- Cannot bound global function error
- Cannot attain ϵ-stationarity (Zhang et al (2020))
- Cannot attain near-ϵ-stationarity (Kornowski & Shamir (2022))
- Smoothing doesn't work (Kornowski & Shamir (2022))

A Meaningful Notion of Convergence

Problem Class:

Nonsmooth Nonconvex

- Cannot bound global function error
- Cannot attain ϵ-stationarity (Zhang et al (2020))
- Cannot attain near-ϵ-stationarity (Kornowski & Shamir (2022))
- Smoothing doesn't work (Kornowski & Shamir (2022))

Alternate notion: A bound on the convex combination of nearby gradients!

A Meaningful Notion of Convergence

Problem Class:

Nonsmooth Nonconvex

- Cannot bound global function error
- Cannot attain ϵ-stationarity (Zhang et al (2020))
- Cannot attain near-ϵ-stationarity (Kornowski & Shamir (2022))
- Smoothing doesn't work (Kornowski & Shamir (2022))

Alternate notion: A bound on the convex combination of nearby gradients!

Definition (Goldstein (1977))

A point x is (δ, ϵ)**-stationary** for a Lipschitz function f if

$$\min_{g \in \partial_\delta f(x)} \|g\| \leq \epsilon.$$

A Meaningful Notion of Convergence

Problem Class:

Nonsmooth Nonconvex

- Cannot bound global function error
- Cannot attain ϵ-stationarity (Zhang et al (2020))
- Cannot attain near-ϵ-stationarity (Kornowski & Shamir (2022))
- Smoothing doesn't work (Kornowski & Shamir (2022))

Alternate notion: A bound on the convex combination of nearby gradients!

Definition (Goldstein (1977))

A point x is (δ, ϵ)**-stationary** for a Lipschitz function f if

$$\min_{g \in \partial_\delta f(x)} \|g\| \leq \epsilon.$$

$\partial_\delta f(x) := \mathrm{conv}(\cup_{y \in \mathbb{B}_\delta(x)} \partial f(y))$

"Goldstein subdifferential"

Our Main Result: Informal Statement

Goal: Find a (δ, ϵ)-stationary point for a given Lipschitz function

Our Main Result: Informal Statement

Goal: Find a (δ, ϵ)-stationary point for a given Lipschitz function

Theorem 1: (informal; Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022)

Given an L-Lipschitz function with first-order oracle access to it.

Our Main Result: Informal Statement

Goal: Find a (δ, ϵ)-stationary point for a given Lipschitz function

Theorem 1: (informal; Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022)

Given an L-Lipschitz function with first-order oracle access to it. We provide a randomized algorithm, which, with high probability, in $\mathrm{poly}(L, \epsilon, \delta)$ iterations, converges to a (δ, ϵ)-stationary point.

Our Main Result: Informal Statement

Goal: Find a (δ, ϵ)-stationary point for a given Lipschitz function

Theorem 1: (informal; Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022)

Given an L-Lipschitz function with first-order oracle access to it. We provide a randomized algorithm, which, with high probability, in $\mathrm{poly}(L, \epsilon, \delta)$ iterations, converges to a (δ, ϵ)-stationary point.

- First such guarantee using a standard oracle!

Towards an Overview of Our Algorithm & Analysis

A General Algorithmic Framework

Goal: Given an L-Lipschitz function f and accuracy parameters ϵ and δ, find a point x such that $\min_{g \in \partial_\delta f(x)} \|g\| \leq \epsilon$.

A General Algorithmic Framework

Goal: Given an L-Lipschitz function f and accuracy parameters ϵ and δ, find a point x such that $\min_{g \in \partial_\delta f(x)} \|g\| \leq \epsilon$.

Goldstein's Conceptual Descent Algorithm (Goldstein (1977)):
Let $g_t^\star \in \arg\min_{g \in \partial_\delta f(x_t)} \|g\|$ and $x_{t+1} = x_t - \delta \frac{g_t^\star}{\|g_t^\star\|}$. Then,

$$f(x_{t+1}) \leq f(x_t) - \delta \|g_t^\star\|.$$

A General Algorithmic Framework

Goal: Given an L-Lipschitz function f and accuracy parameters ϵ and δ, find a point x such that $\min_{g \in \partial_\delta f(x)} \|g\| \leq \epsilon$.

Goldstein descent step

Goldstein's Conceptual Descent Algorithm (Goldstein (1977)):
Let $g_t^\star \in \arg\min_{g \in \partial_\delta f(x_t)} \|g\|$ and $x_{t+1} = x_t - \delta \frac{g_t^\star}{\|g_t^\star\|}$. Then,

$$f(x_{t+1}) \leq f(x_t) - \delta \|g_t^\star\|.$$

A General Algorithmic Framework

Goal: Given an L-Lipschitz function f and accuracy parameters ϵ and δ, find a point x such that $\min_{g \in \partial_\delta f(x)} \|g\| \leq \epsilon$.

Goldstein's Conceptual Descent Algorithm (Goldstein (1977)):
Let $g_t^\star \in \arg\min_{g \in \partial_\delta f(x_t)} \|g\|$ and $x_{t+1} = x_t - \delta \frac{g_t^\star}{\|g_t^\star\|}$. Then,

$$f(x_{t+1}) \leq f(x_t) - \delta \|g_t^\star\|.$$

- A Goldstein descent step decreases function value by at least $\delta\epsilon$

A General Algorithmic Framework

Goal: Given an L-Lipschitz function f and accuracy parameters ϵ and δ, find a point x such that $\min_{g \in \partial_\delta f(x)} \|g\| \leq \epsilon$.

Goldstein's Conceptual Descent Algorithm (Goldstein (1977)):
Let $g_t^\star \in \arg\min_{g \in \partial_\delta f(x_t)} \|g\|$ and $x_{t+1} = x_t - \delta \frac{g_t^\star}{\|g_t^\star\|}$. Then,

$$f(x_{t+1}) \leq f(x_t) - \delta \|g_t^\star\|$$

- A Goldstein descent step decreases function value by at least $\delta\epsilon$

A General Algorithmic Framework

Goal: Given an L-Lipschitz function f and accuracy parameters ϵ and δ, find a point x such that $\min_{g \in \partial_\delta f(x)} \|g\| \leq \epsilon$.

Goldstein's Conceptual Descent Algorithm (Goldstein (1977)):
Let $g_t^\star \in \arg\min_{g \in \partial_\delta f(x_t)} \|g\|$ and $x_{t+1} = x_t - \delta \frac{g_t^\star}{\|g_t^\star\|}$. Then,

$$f(x_{t+1}) \leq f(x_t) - \delta \|g_t^\star\|.$$

- A Goldstein descent step decreases function value by at least $\delta\epsilon$
- Assuming the initial function error to be Δ...

A General Algorithmic Framework

Goal: Given an L-Lipschitz function f and accuracy parameters ϵ and δ, find a point x such that $\min_{g \in \partial_\delta f(x)} \|g\| \leq \epsilon$.

Goldstein's Conceptual Descent Algorithm (Goldstein (1977)):
Let $g_t^\star \in \arg\min_{g \in \partial_\delta f(x_t)} \|g\|$ and $x_{t+1} = x_t - \delta \frac{g_t^\star}{\|g_t^\star\|}$. Then,

$$f(x_{t+1}) \leq f(x_t) - \delta \|g_t^\star\|.$$

- A Goldstein descent step decreases function value by at least $\delta\epsilon$
- Assuming the initial function error to be Δ...
- ... guarantees a (δ, ϵ)-stationary point in $O\left(\frac{\Delta}{\delta\epsilon}\right)$ iterations.

A General Algorithmic Framework

Goal: Given an L-Lipschitz function f and accuracy parameters ϵ and δ, find a point x such that $\min_{g \in \partial_\delta f(x)} \|g\| \le \epsilon$.

Goldstein's Conceptual Descent Algorithm (Goldstein (1977)):
Let $g_t^\star \in \arg\min_{g \in \partial_\delta f(x_t)} \|g\|$ and $x_{t+1} = x_t - \delta \frac{g_t^\star}{\|g_t^\star\|}$. Then,

$$f(x_{t+1}) \le f(x_t) - \delta \|g_t^\star\|.$$

- A Goldstein descent step decreases function value by at least $\delta\epsilon$
- Assuming the initial function error to be Δ...
- ... guarantees a (δ, ϵ)-stationary point in $O\left(\frac{\Delta}{\delta\epsilon}\right)$ iterations.

requires $\arg\min_{g \in \partial_\delta f(x)} \|g\|$

A General Algorithmic Framework

Goal: Given an L-Lipschitz function f and accuracy parameters ϵ and δ, find a point x such that $\min_{g \in \partial_\delta f(x)} \|g\| \leq \epsilon$.

Goldstein's Conceptual Descent Algorithm (Goldstein (1977)):
Let $g_t^\star \in \arg\min_{g \in \partial_\delta f(x_t)} \|g\|$ and $x_{t+1} = x_t - \delta \frac{g_t^\star}{\|g_t^\star\|}$. Then,

$$f(x_{t+1}) \leq f(x_t) - \delta \|g_t^\star\|.$$

- A Goldstein descent step decreases function value by at least $\delta\epsilon$
- Assuming the initial function error to be Δ...
- ... guarantees a (δ, ϵ)-stationary point in $O\left(\frac{\Delta}{\delta\epsilon}\right)$ iterations.

Central Technical Question:
How to compute $\arg\min_{g \in \partial_\delta f(x)} \|g\|$ using a first-order oracle?

Towards a Min-Norm Element: A Sketch

Suppose a candidate $g \in \partial_\delta f(x)$ satisfies

$\boxed{\|g\| \geq \epsilon}$

$$f\left(x - \delta \cdot \frac{g}{\|g\|}\right) \geq f(x) - \frac{\delta}{2} \cdot \|g\|.$$

Towards a Min-Norm Element: A Sketch

$$f\left(x - \delta \frac{g}{\|g\|}\right) \le f(x) - \delta\|g\|$$

Goldstein descent

Suppose a candidate $g \in \partial_\delta f(x)$ satisfies

$\|g\| \ge \epsilon$

$$f\left(x - \delta \cdot \frac{g}{\|g\|}\right) \ge f(x) - \frac{\delta}{2} \cdot \|g\|.$$

not satisfying

Goldstein's descent

Towards a Min-Norm Element: A Sketch

$$f\left(x - \delta \frac{g}{\|g\|}\right) \leq f(x) - \delta\|g\|$$

Goldstein descent

Suppose a candidate $g \in \partial_\delta f(x)$ satisfies

$$f\left(x - \delta \cdot \frac{g}{\|g\|}\right) \geq f(x) - \frac{\delta}{2} \cdot \|g\|.$$

Want to construct $g' \in \partial_\delta f(x)$ that is a minimal norm element of $\partial_\delta f(x)$

Towards a Min-Norm Element: A Sketch

$$f\left(x - \delta \frac{g}{\|g\|}\right) \le f(x) - \delta\|g\|$$

Goldstein descent

Suppose a candidate $g \in \partial_\delta f(x)$ satisfies

$$f\left(x - \delta \cdot \frac{g}{\|g\|}\right) \ge f(x) - \frac{\delta}{2} \cdot \|g\|.$$

Want to construct $g' \in \partial_\delta f(x)$ that is a minimal norm element of $\partial_\delta f(x)$

Task reduces to finding some $u \in \partial_\delta f(x)$ satisfying $\langle u, g\rangle \le \frac{1}{2}\|g\|^2$.

Towards a Min-Norm Element: A Sketch

$$f\left(x - \delta \frac{g}{\|g\|}\right) \leq f(x) - \delta\|g\|$$

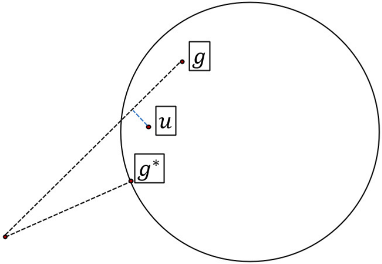
Goldstein descent

Suppose a candidate $g \in \partial_\delta f(x)$ satisfies

$$f\left(x - \delta \cdot \frac{g}{\|g\|}\right) \geq f(x) - \frac{\delta}{2} \cdot \|g\|.$$

Want to construct $g' \in \partial_\delta f(x)$ that is a minimal norm element of $\partial_\delta f(x)$

Task reduces to finding some $u \in \partial_\delta f(x)$ satisfying $\langle u, g \rangle \leq \frac{1}{2}\|g\|^2$.



A Solution under a Strong Assumption

Given a vector $g \in \partial_\delta f(x)$ not satisfying the descent condition, construct a vector $u \in \partial_\delta f(x)$ satisfying $\langle u, g \rangle \leq \frac{1}{2}\|g\|^2$.

A Solution under a Strong Assumption

Given a vector $g \in \partial_\delta f(x)$ not satisfying the descent condition, construct a vector $u \in \partial_\delta f(x)$ satisfying $\langle u, g\rangle \leq \frac{1}{2}\|g\|^2$.

"Inner Product Oracle"

A Solution under a Strong Assumption

Given a vector $g \in \partial_\delta f(x)$ not satisfying the descent condition, construct a vector $u \in \partial_\delta f(x)$ satisfying $\langle u, g\rangle \leq \frac{1}{2}\|g\|^2$.

Suppose f were differentiable along $\left[x, x - \delta \cdot \frac{g}{\|g\|}\right]$.

strong assumption!

A Solution under a Strong Assumption

Given a vector $g \in \partial_\delta f(x)$ not satisfying the descent condition, construct a vector $u \in \partial_\delta f(x)$ satisfying $\langle u, g\rangle \leq \frac{1}{2}\|g\|^2$.

Suppose f were differentiable along $\left[x, x - \delta \cdot \frac{g}{\|g\|}\right]$. Then, we have

A Solution under a Strong Assumption

Given a vector $g \in \partial_\delta f(x)$ not satisfying the descent condition, construct a vector $u \in \partial_\delta f(x)$ satisfying $\langle u, g\rangle \leq \frac{1}{2}\|g\|^2$.

Suppose f were differentiable along $\left[x, x - \delta \cdot \frac{g}{\|g\|}\right]$. Then, we have

$$\frac{1}{2}\|g\| \geq \frac{f(x) - f\left(x - \delta \frac{g}{\|g\|}\right)}{\delta}$$

since Goldstein descent

not satisfied

A Solution under a Strong Assumption

Given a vector $g \in \partial_\delta f(x)$ not satisfying the descent condition, construct a vector $u \in \partial_\delta f(x)$ satisfying $\langle u, g\rangle \leq \frac{1}{2}\|g\|^2$.

Suppose f were differentiable along $\left[x, x - \delta \cdot \frac{g}{\|g\|}\right]$. Then, we have

$$\frac{1}{2}\|g\| \geq \frac{f(x) - f\left(x - \delta \frac{g}{\|g\|}\right)}{\delta} = \frac{1}{\delta}\int_{\tau=0}^{\delta} \left\langle \nabla f\left(x - \tau\frac{g}{\|g\|}\right), \frac{g}{\|g\|}\right\rangle d\tau.$$

by above assumption

A Solution under a Strong Assumption

Given a vector $g \in \partial_\delta f(x)$ not satisfying the descent condition, construct a vector $u \in \partial_\delta f(x)$ satisfying $\langle u, g\rangle \leq \frac{1}{2}\|g\|^2$.

Suppose f were differentiable along $\left[x, x - \delta \cdot \frac{g}{\|g\|}\right]$. Then, we have

$$\frac{1}{2}\|g\| \geq \frac{f(x) - f\left(x - \delta \frac{g}{\|g\|}\right)}{\delta} = \frac{1}{\delta}\int_{\tau=0}^{\delta} \left\langle \nabla f\left(x - \tau \frac{g}{\|g\|}\right), \frac{g}{\|g\|}\right\rangle d\tau.$$

Thus, a point $y \overset{u.a.r.}{\sim} \left[x, x - \delta \frac{g}{\|g\|}\right]$ satisfies $\mathbb{E}\langle \nabla f(y), g\rangle \leq \frac{1}{2}\|g\|_2^2$.

A Solution under a Strong Assumption

Given a vector $g \in \partial_\delta f(x)$ not satisfying the descent condition, construct a vector $u \in \partial_\delta f(x)$ satisfying $\langle u, g\rangle \le \frac{1}{2}\|g\|^2$.

Suppose f were differentiable along $\left[x, x - \delta \cdot \frac{g}{\|g\|}\right]$. Then, we have

$$\frac{1}{2}\|g\| \ge \frac{f(x) - f\left(x - \delta\frac{g}{\|g\|}\right)}{\delta} = \frac{1}{\delta}\int_{\tau=0}^{\delta}\left\langle \nabla f\left(x - \tau\frac{g}{\|g\|}\right), \frac{g}{\|g\|}\right\rangle d\tau.$$

Thus, a point $y \overset{u.a.r.}{\sim} \left[x, x - \delta\frac{g}{\|g\|}\right]$ satisfies $\mathbb{E}\langle \nabla f(y), g\rangle \le \frac{1}{2}\|g\|_2^2$.

Using randomization, we get this result without the above assumption!

The Idea for Our Algorithm

- We start with the algorithm of Zhang et al (2020)...
  - ... interpreting it in the Goldstein descent framework
- and use randomization to replace Zhang et al (2020)'s strong oracle ("ZO") with a standard first-order oracle

First, Zhang et al (2020)'s Algorithm

First, Zhang et al (2020)'s Algorithm

1. **for** T iterations **do**:
   - Compute $g = \textsc{MinNorm}(x_t, \delta, \epsilon)$
   - Update $x_{t+1} = x_t - \delta \frac{g}{\|g\|}$
2. Return x_T

First, Zhang et al (2020)'s Algorithm

1. **for** T iterations **do**:
   - Compute $g = \textsc{MinNorm}(x_t, \delta, \epsilon)$
   - Update $x_{t+1} = x_t - \delta \frac{g}{\|g\|}$ `Goldstein descent step`
2. Return x_T

First, Zhang et al (2020)'s Algorithm

1. **for** T iterations **do**:
   - Compute $g = \text{MINNORM}(x_t, \delta, \epsilon)$
   - Update $x_{t+1} = x_t - \delta \frac{g}{\|g\|}$
2. Return x_T

Zhang et al (2020)'s MINNORM(x, δ, ϵ)

1. **while** $\|g_k\| \geq \epsilon$ and $\frac{\delta}{4}\|g_k\| \geq f(x) - f\left(x - \delta \frac{g_k}{\|g_k\|}\right)$, **do**
   - Choose $y_k \overset{u.a.r.}{\sim} \left[x, x - \delta \frac{g_k}{\|g_k\|}\right]$
   - Let $u_k = \text{ZO}(y_k, g_k)$
   - Update $g_{k+1} = \arg\min_{z \in [g_k, u_k]} \|z\|$, and update $k = k + 1$
2. Return g_k

First, Zhang et al (2020)'s Algorithm

1. **for** T iterations **do**:
   - Compute $g = \textsc{MinNorm}(x_t, \delta, \epsilon)$
   - Update $x_{t+1} = x_t - \delta \frac{g}{\|g\|}$
2. Return x_T

Zhang et al (2020)'s MinNorm(x, δ, ϵ)

1. **while** $\|g_k\| \geq \epsilon$ and $\frac{\delta}{4}\|g_k\| \geq f(x) - f\left(x - \delta \frac{g_k}{\|g_k\|}\right)$, **do**
   - Choose $y_k \overset{u.a.r.}{\sim} \left[x, x - \delta \frac{g_k}{\|g_k\|}\right]$
   - Let $u_k = \text{ZO}(y_k, g_k)$
   - Update $g_{k+1} = \arg\min_{z \in [g_k, u_k]} \|z\|$, and update $k = k + 1$
2. Return g_k

Next, Our Algorithm

1. **for** T iterations **do**:
   - Compute $g = \textsc{MinNorm}(x_t, \delta, \epsilon)$
   - Update $x_{t+1} = x_t - \delta \frac{g}{\|g\|}$
2. Return x_T

Our MinNorm(x, δ, ϵ)

1. **while** $\|g_k\| \geq \epsilon$ and $\frac{\delta}{4}\|g_k\| \geq f(x) - f\left(x - \delta \frac{g_k}{\|g_k\|}\right)$, **do**
   - Choose $y_k \overset{u.a.r.}{\sim} \left[x, x - \delta \frac{\xi_k}{\|\xi_k\|}\right]$ where $\xi_k \overset{u.a.r.}{\sim} B_r(g_k)$
   - Let $u_k = \nabla f(y_k)$
   - Update $g_{k+1} = \arg\min_{z \in [g_k, u_k]} \|z\|$, and update $k = k + 1$
2. Return g_k

The Issue with Zhang et al (2020)'s Oracle

Zhang et al (2020)'s algorithm requires the following oracle access:

The Issue with Zhang et al (2020)'s Oracle

Zhang et al (2020)'s algorithm requires the following oracle access: given $x, g \in \mathbb{R}^d$, solve the auxiliary convex feasibility problem

$$\boxed{\text{find } \; u \in \partial f(x) \; \text{ subject to } \; \langle u, g \rangle = f'(x, g).}$$

The Issue with Zhang et al (2020)'s Oracle

Zhang et al (2020)'s algorithm requires the following oracle access: given $x, g \in \mathbb{R}^d$, solve the auxiliary convex feasibility problem

$$\boxed{\text{find } \; u \in \partial f(x) \; \text{ subject to } \; \langle u, g \rangle = f'(x, g).}$$

- The set $\partial f(x)$ could be extremely complicated

The Issue with Zhang et al (2020)'s Oracle

Zhang et al (2020)'s algorithm requires the following oracle access: given $x, g \in \mathbb{R}^d$, solve the auxiliary convex feasibility problem

$$\boxed{\text{find } \; u \in \partial f(x) \; \text{ subject to } \; \langle u, g \rangle = f'(x, g).}$$

- The set $\partial f(x)$ could be extremely complicated
- The chain rule fails for subdifferentials

Analysis of Our Algorithm

Guarantee of Our MinNorm Subroutine

Our MinNorm(x, δ, ϵ)

1. **while** $\|g_k\| \geq \epsilon$ **and** $\frac{\delta}{4}\|g_k\| \geq f(x) - f\left(x - \delta\frac{g_k}{\|g_k\|}\right)$, **do**
   - Choose $y_k \overset{u.a.r.}{\sim} \left[x, x - \delta\frac{\xi_k}{\|\xi_k\|}\right]$ where $\xi_k \overset{u.a.r.}{\sim} B_r(g_k)$
   - Let $u_k = \nabla f(y_k)$
   - Update $g_{k+1} = \arg\min_{z \in [g_k, u_k]} \|z\|$, and update $k = k + 1$
2. Return g_k

Theorem 2: (informal; Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022**)**

Let $\{g_\ell\}$ be generated by $\texttt{MinNorm}(x, \delta, \epsilon)$, and let τ be its termination time. Then, for a fixed $k \geq 0$, we have $\mathbb{E}[\|g_k\|^2 \mathbf{1}_{\tau > k}] \leq \frac{L^2}{1+k}$.

Guarantee of Our MinNorm Subroutine

Theorem 3: (informal; Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022**)**

Let $\{g_\ell\}$ be generated by $\texttt{MinNorm}(x, \delta, \epsilon)$, and let τ be its termination time. Then, for a fixed $k \geq 0$, we have $\mathbb{E}[\|g_k\|^2 \mathbf{1}_{\tau > k}] \leq \frac{L^2}{1+k}$.

Guarantee of Our `MinNorm` Subroutine

Theorem 3: (informal; Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022**)**

Let $\{g_\ell\}$ be generated by $\texttt{MinNorm}(x, \delta, \epsilon)$, and let τ be its termination time. Then, for a fixed $k \geq 0$, we have $\mathbb{E}[\|g_k\|^2 \mathbf{1}_{\tau > k}] \leq \frac{L^2}{1+k}$.

Proof. Let $\widehat{u} := u/\|u\|$; Then, almost surely, conditioned on g_k, we have:

Guarantee of Our `MinNorm` Subroutine

Theorem 3: (informal; Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022)

Let $\{g_\ell\}$ be generated by $\texttt{MinNorm}(x, \delta, \epsilon)$, and let τ be its termination time. Then, for a fixed $k \geq 0$, we have $\mathbb{E}[\|g_k\|^2 \mathbf{1}_{\tau > k}] \leq \frac{L^2}{1+k}$.

Proof. Let $\widehat{u} := u/\|u\|$; Then, almost surely, conditioned on g_k, we have:

$$\frac{1}{2}\|g_k\| \geq \frac{1}{\delta}[f(x) - f(x - \delta\widehat{g}_k)]$$

since Goldstein descent

not satisfied

Guarantee of Our MinNorm Subroutine

Theorem 3: (informal; Davis, Drusvyatskiy, Lee, Padmanabhan, Ye; 2022)

Let $\{g_\ell\}$ be generated by $\texttt{MinNorm}(x, \delta, \epsilon)$, and let τ be its termination time. Then, for a fixed $k \geq 0$, we have $\mathbb{E}[\|g_k\|^2 \mathbf{1}_{\tau > k}] \leq \frac{L^2}{1+k}$.

Proof. Let $\widehat{u} := u/\|u\|$; Then, almost surely, conditioned on g_k, we have:

$$\frac{1}{2}\|g_k\| \geq \frac{1}{\delta}[f(x) - f(x - \delta \widehat{g}_k)] \geq \frac{1}{\delta}[f(x) - f(x - \delta \widehat{\xi}_k)] - L\|\widehat{g}_k - \widehat{\xi}_k\|$$

L-Lipschitzness

Guarantee of Our MinNorm Subroutine

Theorem 3: (informal; Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022**)**

Let $\{g_\ell\}$ be generated by $\texttt{MinNorm}(x, \delta, \epsilon)$, and let τ be its termination time. Then, for a fixed $k \geq 0$, we have $\mathbb{E}[\|g_k\|^2 \mathbf{1}_{\tau > k}] \leq \frac{L^2}{1+k}$.

Proof. Let $\widehat{u} := u/\|u\|$; Then, almost surely, conditioned on g_k, we have:

$$\frac{1}{2}\|g_k\| \geq \frac{1}{\delta}[f(x) - f(x - \delta\widehat{g}_k)] \geq \frac{1}{\delta}[f(x) - f(x - \delta\widehat{\xi}_k)] - L\|\widehat{g}_k - \widehat{\xi}_k\|$$

$$= \frac{1}{\delta}\int_{s=0}^{\delta} \langle \nabla f(x - s\widehat{\xi}_k), \widehat{\xi}_k \rangle ds - L\|\widehat{g}_k - \widehat{\xi}_k\|$$

by randomization and

fundamental thm. of calc.

Guarantee of Our `MinNorm` Subroutine

Theorem 3: (informal; Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022**)**

Let $\{g_\ell\}$ be generated by $\texttt{MinNorm}(x, \delta, \epsilon)$, and let τ be its termination time. Then, for a fixed $k \geq 0$, we have $\mathbb{E}[\|g_k\|^2 \mathbf{1}_{\tau > k}] \leq \frac{L^2}{1+k}$.

Proof. Let $\widehat{u} := u/\|u\|$; Then, almost surely, conditioned on g_k, we have:

$$\frac{1}{2}\|g_k\| \geq \frac{1}{\delta}[f(x) - f(x - \delta\widehat{g}_k)] \geq \frac{1}{\delta}[f(x) - f(x - \delta\widehat{\xi}_k)] - L\|\widehat{g}_k - \widehat{\xi}_k\|$$

$$= \frac{1}{\delta}\int_{s=0}^{\delta} \langle \nabla f(x - s\widehat{\xi}_k), \widehat{\xi}_k \rangle ds - L\|\widehat{g}_k - \widehat{\xi}_k\|$$

$$\geq \frac{1}{\delta}\int_{s=0}^{\delta} \langle \nabla f(x - s\widehat{\xi}_k), \widehat{g}_k \rangle ds - 2L\|\widehat{g}_k - \widehat{\xi}_k\|$$

L-Lipschitzness

Guarantee of Our `MinNorm` Subroutine

Theorem 3: (informal; Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022**)**

Let $\{g_\ell\}$ be generated by $\texttt{MinNorm}(x, \delta, \epsilon)$, and let τ be its termination time. Then, for a fixed $k \geq 0$, we have $\mathbb{E}[\|g_k\|^2 \mathbf{1}_{\tau > k}] \leq \frac{L^2}{1+k}$.

Proof. Let $\widehat{u} := u/\|u\|$; Then, almost surely, conditioned on g_k, we have:

$$
\begin{aligned}
\frac{1}{2}\|g_k\| &\geq \frac{1}{\delta}[f(x) - f(x - \delta\widehat{g}_k)] \geq \frac{1}{\delta}[f(x) - f(x - \delta\widehat{\xi}_k)] - L\|\widehat{g}_k - \widehat{\xi}_k\| \\
&= \frac{1}{\delta}\int_{s=0}^{\delta} \langle \nabla f(x - s\widehat{\xi}_k), \widehat{\xi}_k \rangle ds - L\|\widehat{g}_k - \widehat{\xi}_k\| \\
&\geq \frac{1}{\delta}\int_{s=0}^{\delta} \langle \nabla f(x - s\widehat{\xi}_k), \widehat{g}_k \rangle ds - 2L\|\widehat{g}_k - \widehat{\xi}_k\| \\
&= \mathbb{E}_k \langle \nabla f(y_k), \widehat{g}_k \rangle - 2L\|\widehat{g}_k - \widehat{\xi}_k\|.
\end{aligned}
$$

definition of y_k

Guarantee of Our MinNorm Subroutine

Theorem 3: (informal; Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022)

Let $\{g_\ell\}$ be generated by $\texttt{MinNorm}(x, \delta, \epsilon)$, and let τ be its termination time. Then, for a fixed $k \geq 0$, we have $\mathbb{E}[\|g_k\|^2 \mathbf{1}_{\tau > k}] \leq \frac{L^2}{1+k}$.

Proof. Let $\widehat{u} := u/\|u\|$; Then, almost surely, conditioned on g_k, we have:

$$\begin{aligned}\frac{1}{2}\|g_k\| &\geq \frac{1}{\delta}[f(x) - f(x - \delta\widehat{g}_k)] \geq \frac{1}{\delta}[f(x) - f(x - \delta\widehat{\xi}_k)] - L\|\widehat{g}_k - \widehat{\xi}_k\| \\ &= \frac{1}{\delta}\int_{s=0}^{\delta} \langle \nabla f(x - s\widehat{\xi}_k), \widehat{\xi}_k \rangle ds - L\|\widehat{g}_k - \widehat{\xi}_k\| \\ &\geq \frac{1}{\delta}\int_{s=0}^{\delta} \langle \nabla f(x - s\widehat{\xi}_k), \widehat{g}_k \rangle ds - 2L\|\widehat{g}_k - \widehat{\xi}_k\| \\ &= \mathbb{E}_k \langle \nabla f(y_k), \widehat{g}_k \rangle - 2L\|\widehat{g}_k - \widehat{\xi}_k\|.\end{aligned}$$

Guarantee of Our MinNorm Subroutine

Theorem 3: (informal; Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022)

Let $\{g_\ell\}$ be generated by $\texttt{MinNorm}(x, \delta, \epsilon)$, and let τ be its termination time. Then, for a fixed $k \geq 0$, we have $\mathbb{E}[\|g_k\|^2 \mathbf{1}_{\tau > k}] \leq \frac{L^2}{1+k}$.

Proof. Let $\widehat{u} := u/\|u\|$; Then, almost surely, conditioned on g_k, we have:

$$\begin{aligned}
\frac{1}{2}\|g_k\| &\geq \frac{1}{\delta}[f(x) - f(x - \delta\widehat{g}_k)] \geq \frac{1}{\delta}[f(x) - f(x - \delta\widehat{\xi}_k)] - L\|\widehat{g}_k - \widehat{\xi}_k\| \\
&= \frac{1}{\delta}\int_{s=0}^{\delta} \langle \nabla f(x - s\widehat{\xi}_k), \widehat{\xi}_k \rangle ds - L\|\widehat{g}_k - \widehat{\xi}_k\| \\
&\geq \frac{1}{\delta}\int_{s=0}^{\delta} \langle \nabla f(x - s\widehat{\xi}_k), \widehat{g}_k \rangle ds - 2L\|\widehat{g}_k - \widehat{\xi}_k\| \\
&= \mathbb{E}_k \langle \nabla f(y_k), \widehat{g}_k \rangle - 2L\|\widehat{g}_k - \widehat{\xi}_k\|.
\end{aligned}$$

This matches the requirement for $u \in \partial_\delta f(x)$ with $\langle u, g \rangle \leq \frac{1}{2}\|g\|^2$. ■

Guarantee of Our `MinNorm` Subroutine

Theorem 3: (informal; Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022)

Let $\{g_\ell\}$ be generated by $\texttt{MinNorm}(x, \delta, \epsilon)$, and let τ be its termination time. Then, for a fixed $k \geq 0$, we have $\mathbb{E}[\|g_k\|^2 \mathbf{1}_{\tau > k}] \leq \frac{L^2}{1+k}$.

Proof. Let $\widehat{u} := u/\|u\|$; Then, almost surely, conditioned on g_k, we have:

$$
\begin{aligned}
\frac{1}{2}\|g_k\| \geq \frac{1}{\delta}[f(x) - f(x - \delta\widehat{g}_k)] \geq & \frac{1}{\delta}[f(x) - f(x - \delta\widehat{\xi}_k)] - L\|\widehat{g}_k - \widehat{\xi}_k\| \\
= & \frac{1}{\delta}\int_{s=0}^{\delta} \langle \nabla f(x - s\widehat{\xi}_k), \widehat{\xi}_k \rangle ds - L\|\widehat{g}_k - \widehat{\xi}_k\| \\
\geq & \frac{1}{\delta}\int_{s=0}^{\delta} \langle \nabla f(x - s\widehat{\xi}_k), \widehat{g}_k \rangle ds - 2L\|\widehat{g}_k - \widehat{\xi}_k\| \\
= & \mathbb{E}_k \langle \nabla f(y_k), \widehat{g}_k \rangle - 2L\|\widehat{g}_k - \widehat{\xi}_k\|.
\end{aligned}
$$

Inner Product Oracle

This matches the requirement for $u \in \partial_\delta f(x)$ with $\langle u, g \rangle \leq \frac{1}{2}\|g\|^2$. ∎

Our Main Result: Formal Statement

Theorem 4: (Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022)

Given an L-Lipschitz function f, fix an initial point $x_0 \in \mathbb{R}^d$, and define $f(x_0) - \inf_x f(x)$. Then, with probability $1 - \gamma$, our algorithm returns x_T satisfying $\min_{g \in \partial_\delta f(x_T)} \|g\| \leq \epsilon$ in at most

$\lceil \frac{4\Delta}{\delta\epsilon} \rceil \cdot \lceil \frac{64L^2}{\epsilon^2} \rceil \cdot \lceil 2 \log \left(\frac{4\Delta}{\gamma\delta\epsilon} \right) \rceil$ function-value and gradient evaluations.

Our Main Result: Formal Statement

Theorem 4: (Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022)

Given an L-Lipschitz function f, fix an initial point $x_0 \in \mathbb{R}^d$, and define $f(x_0) - \inf_x f(x)$. Then, with probability $1-\gamma$, our algorithm returns x_T satisfying $\min_{g \in \partial_\delta f(x_T)} \|g\| \leq \epsilon$ in at most

$\lceil \frac{4\Delta}{\delta\epsilon} \rceil \cdot \lceil \frac{64L^2}{\epsilon^2} \rceil \cdot \lceil 2 \log\left(\frac{4\Delta}{\gamma\delta\epsilon}\right) \rceil$ function-value and gradient evaluations.

Goldstein descent iterations

Our Main Result: Formal Statement

Theorem 4: (Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022)

Given an L-Lipschitz function f, fix an initial point $x_0 \in \mathbb{R}^d$, and define $f(x_0) - \inf_x f(x)$. Then, with probability $1-\gamma$, our algorithm returns x_T satisfying $\min_{g \in \partial_\delta f(x_T)} \|g\| \leq \epsilon$ in at most

$\lceil \frac{4\Delta}{\delta\epsilon} \rceil \cdot \lceil \frac{64L^2}{\epsilon^2} \rceil \cdot \lceil 2 \log \left(\frac{4\Delta}{\gamma\delta\epsilon} \right) \rceil$ function-value and gradient evaluations.

Goldstein descent iterations

`MinNorm` iterations

Our Second Question in this Thread

Problem Overview

Recall that $g \in \partial_\delta f(x)$ satisfies the descent condition at x if

$$f\left(x - \delta\frac{g}{\|g\|}\right) \leq f(x) - \frac{\delta\epsilon}{3}.$$

Problem Overview

Recall that $g \in \partial_\delta f(x)$ satisfies the descent condition at x if

$$f\left(x - \delta \frac{g}{\|g\|}\right) \leq f(x) - \frac{\delta\epsilon}{3}.$$

If not, the Inner Product Oracle outputs $u \in \partial_\delta f(x)$ such that

$$\langle u, g \rangle \leq \frac{\epsilon}{3}\|g\|.$$

Problem Overview

Recall that $g \in \partial_\delta f(x)$ satisfies the descent condition at x if

$$f\left(x - \delta \frac{g}{\|g\|}\right) \leq f(x) - \frac{\delta\epsilon}{3}.$$

If not, the Inner Product Oracle outputs $u \in \partial_\delta f(x)$ such that

$$\langle u, g \rangle \leq \frac{\epsilon}{3}\|g\|.$$

This vector u is combined with g to generate a vector that either corresponds to the desired stationarity or is a descent direction

Problem Overview

Recall that $g \in \partial_\delta f(x)$ satisfies the descent condition at x if

$$f\left(x - \delta \frac{g}{\|g\|}\right) \leq f(x) - \frac{\delta \epsilon}{3}.$$

If not, the Inner Product Oracle outputs $u \in \partial_\delta f(x)$ such that

$$\langle u, g \rangle \leq \frac{\epsilon}{3}\|g\|.$$

This vector u is combined with g to generate a vector that either corresponds to the desired stationarity or is a descent direction

Are there settings in which we can use the vector u more efficiently?

Our Main Idea

Recall that given $g \in \partial_\delta f(x)$ not satisfying the descent condition, we can output $u \in \partial_\delta f(x)$ such that $\langle u, g \rangle \leq \frac{\epsilon}{2} \|g\|$.

Inner Product Oracle

Our Main Idea

Recall that given $g \in \partial_\delta f(x)$ not satisfying the descent condition, we can output $u \in \partial_\delta f(x)$ such that $\langle u, g\rangle \leq \frac{\epsilon}{2}\|g\|$.

Inner Product Oracle

Our Key Insight.

The above oracle is essentially the gradient oracle of the `MinNorm` element problem.

Our Main Idea

Recall that given $g \in \partial_\delta f(x)$ not satisfying the descent condition, we can output $u \in \partial_\delta f(x)$ such that $\langle u, g \rangle \leq \frac{\epsilon}{2}\|g\|$.

Inner Product Oracle

Our Key Insight.
The above oracle is essentially the gradient oracle of the `MinNorm` element problem. We can therefore use it in a cutting-plane method.

Using the Inner Product Oracle

Notation Denote $Q := \partial_\delta f(x)$; and $\widehat{x} := x/\|x\|$ for some vector x

Using the Inner Product Oracle

Notation Denote $Q := \partial_\delta f(x)$; and $\widehat{x} := x/\|x\|$ for some vector x

Lemma 1: (Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022)

Let $g \in Q$ be a vector not satisfying the descent condition, and let $u \in Q$ be the output of the inner product oracle. Let $g_Q^\star \in \min_{g\in Q} \|g\| \geq \epsilon/2$. Then, $\widehat{g_Q^\star} \in \{w \in \mathbb{R}^d : \langle u, \widehat{g} - w\rangle \leq 0\}$.

Using the Inner Product Oracle

Notation Denote $Q := \partial_\delta f(x)$; and $\widehat{x} := x/\|x\|$ for some vector x

Lemma 1: (Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022)

Let $g \in Q$ be a vector not satisfying the descent condition, and let $u \in Q$ be the output of the inner product oracle. Let $g_Q^\star \in \min_{g \in Q} \|g\| \geq \epsilon/2$. Then, $\widehat{g_Q^\star} \in \{w \in \mathbb{R}^d : \langle u, \widehat{g} - w \rangle \leq 0\}$.

Proof Combining the above definitions and a technical lemma gives:

Using the Inner Product Oracle

Notation Denote $Q := \partial_\delta f(x)$; and $\widehat{x} := x/\|x\|$ for some vector x

Lemma 1: (Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022)

Let $g \in Q$ be a vector not satisfying the descent condition, and let $u \in Q$ be the output of the inner product oracle. Let $g_Q^\star \in \min_{g \in Q} \|g\| \geq \epsilon/2$. Then, $\widehat{g_Q^\star} \in \{w \in \mathbb{R}^d : \langle u, \widehat{g} - w\rangle \leq 0\}$.

Proof Combining the above definitions and a technical lemma gives:

The inner product oracle guarantees: $\boxed{\langle u, \widehat{g}\rangle \leq \frac{\epsilon}{2}}$

Using the Inner Product Oracle

Notation Denote $Q := \partial_\delta f(x)$; and $\widehat{x} := x/\|x\|$ for some vector x

Lemma 1: (Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022)

Let $g \in Q$ be a vector not satisfying the descent condition, and let $u \in Q$ be the output of the inner product oracle. Let $g_Q^\star \in \min_{g \in Q} \|g\| \geq \epsilon/2$. Then, $\widehat{g_Q^\star} \in \{w \in \mathbb{R}^d : \langle u, \widehat{g} - w \rangle \leq 0\}$.

Proof Combining the above definitions and a technical lemma gives:

The inner product oracle guarantees: $\boxed{\langle u, \widehat{g} \rangle \leq \frac{\epsilon}{2}}$

The technical lemma (extra slide) shows: $\boxed{\langle u, \widehat{g_Q^\star} \rangle \geq \|g_Q^\star\|}$

Using the Inner Product Oracle

Notation Denote $Q := \partial_\delta f(x)$; and $\widehat{x} := x/\|x\|$ for some vector x

Lemma 1: (Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022)

Let $g \in Q$ be a vector not satisfying the descent condition, and let $u \in Q$ be the output of the inner product oracle. Let $g_Q^\star \in \min_{g \in Q} \|g\| \geq \epsilon/2$. Then, $\widehat{g_Q^\star} \in \{w \in \mathbb{R}^d : \langle u, \widehat{g} - w\rangle \leq 0\}$.

Proof Combining the above definitions and a technical lemma gives:

The inner product oracle guarantees: $\boxed{\langle u, \widehat{g}\rangle \leq \frac{\epsilon}{2}}$

The technical lemma (extra slide) shows: $\boxed{\langle u, \widehat{g_Q^\star}\rangle \geq \|g_Q^\star\|}$

Combining these two inequalities yields: $\boxed{\langle u, \widehat{g} - \widehat{g_Q^\star}\rangle \leq \frac{\epsilon}{2} - \|g_Q^\star\| \leq 0}$

Our Second Result: Complete Statement

Theorem 5: (Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022)

Given an L-Lipschitz function f. Fix an initial point $x_0 \in \mathbb{R}^d$, and define $f(x_0) - \inf_x f(x)$. Then, with probability $1-\gamma$, our algorithm returns x_T satisfying $\min_{g \in \partial_\delta f(x_T)} \|g\| \leq \epsilon$ in at most

$\lceil \frac{4\Delta}{\delta\epsilon} \rceil \cdot \lceil 8d \log\left(\frac{8L}{\epsilon}\right) \rceil \cdot \lceil \frac{36L}{\epsilon} \rceil$ function-value and gradient evaluations.

Our Second Result: Complete Statement

Theorem 5: (Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022)

Given an L-Lipschitz function f. Fix an initial point $x_0 \in \mathbb{R}^d$, and define $f(x_0) - \inf_x f(x)$. Then, with probability $1-\gamma$, our algorithm returns x_T satisfying $\min_{g \in \partial_\delta f(x_T)} \|g\| \leq \epsilon$ in at most

$$\underbrace{\left\lceil \frac{4\Delta}{\delta\epsilon} \right\rceil}_{\text{Goldstein descent iterations}} \cdot \left\lceil 8d \log\left(\frac{8L}{\epsilon}\right) \right\rceil \cdot \left\lceil \frac{36L}{\epsilon} \right\rceil$$

function-value and gradient evaluations.

Our Second Result: Complete Statement

Theorem 5: (Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022)

Given an L-Lipschitz function f. Fix an initial point $x_0 \in \mathbb{R}^d$, and define $f(x_0) - \inf_x f(x)$. Then, with probability $1-\gamma$, our algorithm returns x_T satisfying $\min_{g \in \partial_\delta f(x_T)} \|g\| \leq \epsilon$ in at most

$$\underbrace{\left\lceil \frac{4\Delta}{\delta\epsilon} \right\rceil}_{\text{Goldstein descent iterations}} \cdot \underbrace{\left\lceil 8d \log\left(\frac{8L}{\epsilon}\right) \right\rceil}_{\text{cutting-plane iterations}} \cdot \left\lceil \frac{36L}{\epsilon} \right\rceil$$ function-value and gradient evaluations.

Our Second Result: Complete Statement

Theorem 5: (Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022)

Given an L-Lipschitz function f. Fix an initial point $x_0 \in \mathbb{R}^d$, and define $f(x_0) - \inf_x f(x)$. Then, with probability $1 - \gamma$, our algorithm returns x_T satisfying $\min_{g \in \partial_\delta f(x_T)} \|g\| \leq \epsilon$ in at most

$$\lceil \frac{4\Delta}{\delta\epsilon} \rceil \cdot \lceil 8d \log\left(\frac{8L}{\epsilon}\right) \rceil \cdot \lceil \frac{36L}{\epsilon} \rceil$$ function-value and gradient evaluations.

Goldstein descent iterations

cutting-plane iterations

Inner Product Oracle iterations

A Technical Lemma

Notation. Let $\phi_Q(v) := \min_{g \in Q} \langle g, v \rangle$; let $\widehat{x} := x/\|x\|$.

A Technical Lemma

Notation. Let $\phi_Q(v) := \min_{g \in Q} \langle g, v \rangle$; let $\widehat{x} := x / \|x\|$.

Lemma 2: (informal; Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022**)**

Let $g_Q^\star \in \arg\min_Q \|g\|$. Then, $\widehat{g_Q^\star}$ satisfies two properties:

A Technical Lemma

Notation. Let $\phi_Q(v) := \min_{g \in Q} \langle g, v \rangle$; let $\widehat{x} := x/\|x\|$.

Lemma 2: (informal; Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022)

Let $g_Q^\star \in \arg\min_Q \|g\|$. Then, $\widehat{g_Q^\star}$ satisfies two properties:

- $\langle \widehat{g_Q^\star}, g \rangle \geq \|g_Q^\star\|$ for all $g \in Q$,
- $\widehat{g_Q^\star} = \arg\max_{\|v\| \leq 1} \phi_Q(v)$.

A Technical Lemma

Notation. Let $\phi_Q(v) := \min_{g \in Q} \langle g, v \rangle$; let $\widehat{x} := x / \|x\|$.

Lemma 2: (informal; Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022)

Let $g_Q^\star \in \arg\min_Q \|g\|$. Then, $\widehat{g_Q^\star}$ satisfies two properties:

- $\langle \widehat{g_Q^\star}, g \rangle \geq \|g_Q^\star\|$ for all $g \in Q$,
- $\widehat{g_Q^\star} = \arg\max_{\|v\| \leq 1} \phi_Q(v)$.

Proof. The first inequality holds by definition of $g_Q^\star$. We drop Q for notational simplicity in the rest of the proof.

A Technical Lemma

Notation. Let $\phi_Q(v) := \min_{g \in Q} \langle g, v \rangle$; let $\widehat{x} := x/\|x\|$.

Lemma 2: (informal; Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022**)**

Let $g_Q^\star \in \arg\min_Q \|g\|$. Then, $\widehat{g_Q^\star}$ satisfies two properties:

- $\langle \widehat{g_Q^\star}, g \rangle \geq \|g_Q^\star\|$ for all $g \in Q$,
- $\widehat{g_Q^\star} = \arg\max_{\|v\| \leq 1} \phi_Q(v)$.

Proof. The first inequality holds by definition of $g_Q^\star$. We drop Q for notational simplicity in the rest of the proof.

$$\phi(\widehat{g^\star}) = \|g^\star\| \quad .$$

first inequality

& definition of ϕ_Q

A Technical Lemma

Notation. Let $\phi_Q(v) := \min_{g\in Q}\langle g, v\rangle$; let $\widehat{x} := x/\|x\|$.

Lemma 2: (informal; Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022**)**

Let $g_Q^\star \in \arg\min_Q \|g\|$. Then, $\widehat{g_Q^\star}$ satisfies two properties:

- $\langle \widehat{g_Q^\star}, g\rangle \geq \|g_Q^\star\|$ for all $g \in Q$,
- $\widehat{g_Q^\star} = \arg\max_{\|v\|\leq 1} \phi_Q(v)$.

Proof. The first inequality holds by definition of $g_Q^\star$. We drop Q for notational simplicity in the rest of the proof.

$$\phi(\widehat{g^\star}) = \|g^\star\| = \min_Q \|g\| \qquad .$$

definition of ϕ_Q

A Technical Lemma

Notation. Let $\phi_Q(v) := \min_{g \in Q} \langle g, v \rangle$; let $\widehat{x} := x/\|x\|$.

Lemma 2: (informal; Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022**)**

Let $g_Q^\star \in \arg\min_Q \|g\|$. Then, $\widehat{g_Q^\star}$ satisfies two properties:

- $\langle \widehat{g_Q^\star}, g \rangle \geq \|g_Q^\star\|$ for all $g \in Q$,
- $\widehat{g_Q^\star} = \arg\max_{\|v\| \leq 1} \phi_Q(v)$.

Proof. The first inequality holds by definition of $g_Q^\star$. We drop Q for notational simplicity in the rest of the proof.

$$\phi(\widehat{g^\star}) = \|g^\star\| = \min_Q \|g\| = \min_Q \max_{\|v\| \leq 1} \langle g, v \rangle \quad .$$

dual representation

A Technical Lemma

Notation. Let $\phi_Q(v) := \min_{g\in Q}\langle g, v\rangle$; let $\widehat{x} := x/\|x\|$.

Lemma 2: (informal; Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022**)**

Let $g_Q^\star \in \arg\min_Q \|g\|$. Then, $\widehat{g_Q^\star}$ satisfies two properties:

- $\langle \widehat{g_Q^\star}, g\rangle \geq \|g_Q^\star\|$ for all $g \in Q$,
- $\widehat{g_Q^\star} = \arg\max_{\|v\|\leq 1} \phi_Q(v)$.

Proof. The first inequality holds by definition of $g_Q^\star$. We drop Q for notational simplicity in the rest of the proof.

$$\phi(\widehat{g^\star}) = \|g^\star\| = \min_{Q} \|g\| = \min_{Q} \max_{\|v\|\leq 1} \langle g, v\rangle = \max_{\|v\|\leq 1} \min_{Q} \langle g, v\rangle \quad .$$

Sion's minmax theorem

A Technical Lemma

Notation. Let $\phi_Q(v) := \min_{g \in Q} \langle g, v \rangle$; let $\widehat{x} := x/\|x\|$.

Lemma 2: (informal; Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022**)**

Let $g_Q^\star \in \arg\min_Q \|g\|$. Then, $\widehat{g_Q^\star}$ satisfies two properties:

- $\langle \widehat{g_Q^\star}, g \rangle \geq \|g_Q^\star\|$ for all $g \in Q$,
- $\widehat{g_Q^\star} = \arg\max_{\|v\| \leq 1} \phi_Q(v)$.

Proof. The first inequality holds by definition of $g_Q^\star$. We drop Q for notational simplicity in the rest of the proof.

$$\phi(\widehat{g^\star}) = \|g^\star\| = \min_Q \|g\| = \min_Q \max_{\|v\| \leq 1} \langle g, v \rangle = \max_{\|v\| \leq 1} \min_Q \langle g, v \rangle = \max_{\|v\| \leq 1} \phi(v).$$

Definition of ϕ

A Technical Lemma

Notation. Let $\phi_Q(v) := \min_{g\in Q}\langle g, v\rangle$; let $\widehat{x} := x/\|x\|$.

Lemma 2: (informal; Davis, Drusvyatskiy, Lee, **Padmanabhan**, Ye; 2022**)**

Let $g_Q^\star \in \arg\min_Q \|g\|$. Then, $\widehat{g_Q^\star}$ satisfies two properties:

- $\langle \widehat{g_Q^\star}, g\rangle \geq \|g_Q^\star\|$ for all $g \in Q$,
- $\widehat{g_Q^\star} = \arg\max_{\|v\|\leq 1} \phi_Q(v)$.

Proof. The first inequality holds by definition of $g_Q^\star$. We drop Q for notational simplicity in the rest of the proof.

$$\phi(\widehat{g^\star}) = \|g^\star\| = \min_Q \|g\| = \min_Q \max_{\|v\|\leq 1} \langle g, v\rangle = \max_{\|v\|\leq 1} \min_Q \langle g, v\rangle = \max_{\|v\|\leq 1} \phi(v).$$

■

Takeaways & Future Directions

1. A faster algorithm for nonsmooth nonconvex optimization
2. Improved (optimal) rates in low dimensions
3. Key ideas: randomization; cutting-plane methods

Takeaways & Future Directions

1. A faster algorithm for nonsmooth nonconvex optimization
2. Improved (optimal) rates in low dimensions
3. Key ideas: randomization; cutting-plane methods
4. Future Direction. More practical notions of convergence?

Thank You!