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; $^3$Massachusetts 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

Gradient-based methods are ubiquitous in optimization

A typical template is the subgradient method:

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

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

\[ \partial f(x) = \text{conv} \{ \lim_{i \to \infty} \nabla f(x_i) : x_i \to x, x_i \in \text{dom}(f) \}. \]
The Subgradient Method: Background

Gradient-based methods are ubiquitous in optimization

A typical template is the subgradient method:

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

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

\[ \partial f(x) = \text{conv} \{ \lim_{i \to \infty} \nabla f(x_i) : x_i \to x, x_i \in \text{dom}(f) \}. \]
The Subgradient Method: Background

Gradient-based methods are ubiquitous in optimization.

A typical template is the subgradient method:

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

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

\[ \partial f(x) = \text{conv} \{ \lim_{i \to \infty} \nabla f(x_i) : x_i \to x, x_i \in \text{dom}(f) \}. \]
The Subgradient Method: Convergence Guarantees

The subgradient method:

\[ x_{t+1} = x_t - \sum_{i \leq t} \alpha_{i,t} \cdot v_i, \text{ for } v_i \in \partial f(x_i). \]
The Subgradient Method: Convergence Guarantees

The subgradient method:

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

✓ Nonasymptotic guarantees for **convex** problems

oracle access

global function error bound
The Subgradient Method: Convergence Guarantees

The subgradient method:

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

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

oracle access

gradient norm bound
(stationary point)
The Subgradient Method: Convergence Guarantees

The subgradient method:

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

✓ 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)
The Subgradient Method: Convergence Guarantees

The subgradient method:

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

✓ 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):
The Subgradient Method: Convergence Guarantees

The subgradient method:

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

- 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.
The Subgradient Method: Convergence Guarantees

The subgradient method:

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

- 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!
The Subgradient Method: Convergence Guarantees

The subgradient method:

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

✓ 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!
A Meaningful Notion of Convergence

Problem Class: Nonsmooth Nonconvex

Definition (Goldstein)

A point $x$ is $(\delta, \epsilon)$-stationary for a Lipschitz function $f$ if

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

$\partial \delta f(x) := \text{conv}(\bigcup_{y \in B_\delta(x)} \partial f(y))$  

"Goldstein subdifferential"
A Meaningful Notion of Convergence

- Cannot bound global function error

Problem Class:
Nonsmooth Nonconvex
A Meaningful Notion of Convergence

- Cannot bound global function error
- Cannot attain $\epsilon$-stationarity (Zhang et al (2020))

Problem Class:
Nonsmooth Nonconvex
A Meaningful Notion of Convergence

- Cannot bound global function error
- Cannot attain $\epsilon$-stationarity (Zhang et al (2020))
- Cannot attain near-$\epsilon$-stationarity (Kornowski & Shamir (2022))
A Meaningful Notion of Convergence

- Cannot bound global function error
- Cannot attain $\varepsilon$-stationarity (Zhang et al (2020))
- Cannot attain near-$\varepsilon$-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 $\epsilon$-stationarity (Zhang et al (2020))
- Cannot attain near-$\epsilon$-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 $\epsilon$-stationarity (Zhang et al. (2020))
- Cannot attain near-$\epsilon$-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 $(\delta, \epsilon)$-stationary for a Lipschitz function $f$ if

\[
\min_{g \in \partial_\delta f(x)} \|g\| \leq \epsilon.
\]
A Meaningful Notion of Convergence

- Cannot bound global function error
- Cannot attain $\epsilon$-stationarity (Zhang et al (2020))
- Cannot attain near-$\epsilon$-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 $(\delta, \epsilon)$-stationary for a Lipschitz function $f$ if

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

$$\partial_\delta f(x) := \text{conv}(\bigcup_{y \in B_\delta(x)} \partial f(y))$$

“Goldstein subdifferential”
Our Main Result: Informal Statement

Goal: Find a \((\delta, \epsilon)\)-stationary point for a given Lipschitz function
Our Main Result: Informal Statement

**Goal**: Find a \((\delta, \epsilon)\)-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 $(\delta, \epsilon)$-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 \(\text{poly}(L, \epsilon, \delta)\) iterations, converges to a $(\delta, \epsilon)$-stationary point.
Our Main Result: Informal Statement

**Goal:** Find a \((\delta, \epsilon)\)-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 \(\text{poly}(L, \epsilon, \delta)\) iterations, converges to a \((\delta, \epsilon)\)-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 $\epsilon$ and $\delta$, find a point $x$ such that $\min_{g \in \partial_\delta f(x)} \|g\| \leq \epsilon$. 

Goldstein’s Conceptual Descent Algorithm:

Let $g^{\star}_t \in \arg\min_{g \in \partial_\delta f(x_t)} \|g\|$ and $x_{t+1} = x_t - \delta g^{\star}_t \|g^{\star}_t\$. Then, $f(x_{t+1}) \leq f(x_t) - \delta \|g^{\star}_t\|$. 

Goldstein descent step ▶

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

Assuming the initial function error to be $\Delta \ldots$ guarantees a $(\delta, \epsilon)$-stationary point in $O(\Delta \delta \epsilon)$ iterations.

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

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

**Goldstein’s Conceptual Descent Algorithm** (Goldstein (1977)):

Let $g_t^* \in \arg \min_{g \in \partial \delta f(x_t)} \| g \|$ and $x_{t+1} = x_t - \delta \frac{g_t^*}{\| g_t^* \|}$. Then,

$$f(x_{t+1}) \leq f(x_t) - \delta \| g_t^* \|.$$
A General Algorithmic Framework

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

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

$$f(x_{t+1}) \leq f(x_t) - \delta \|g_t^*\|.$$
A General Algorithmic Framework

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

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

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

- 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 $\epsilon$ and $\delta$, find a point $x$ such that $\min_{g \in \partial_{\delta} f(x)} \|g\| \leq \epsilon$.

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

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

▶ 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 $\epsilon$ and $\delta$, find a point $x$ such that $\min_{g \in \partial \delta f(x)} \|g\| \leq \epsilon$.

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

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

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

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

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

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

- A Goldstein descent step decreases function value by at least $\delta \epsilon$
- Assuming the initial function error to be $\Delta$...
- ... guarantees a $(\delta, \epsilon)$-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 $\epsilon$ and $\delta$, find a point $x$ such that $\min_{g \in \partial f(x)} \|g\| \leq \epsilon$.

**Goldstein’s Conceptual Descent Algorithm** (*Goldstein (1977)*):

Let $g^*_t \in \arg \min_{g \in \partial f(x_t)} \|g\|$ and $x_{t+1} = x_t - \delta \frac{g^*_t}{\|g^*_t\|}$. Then,

$$f(x_{t+1}) \leq f(x_t) - \delta \|g^*_t\|.$$ 

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

requires $\arg \min_{g \in \partial f(x)} \|g\|$
A General Algorithmic Framework

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

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

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

- A Goldstein descent step decreases function value by at least $\delta \epsilon$.
- Assuming the initial function error to be $\Delta$... 
- ...guarantees a $(\delta, \epsilon)$-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
\[
\|g\| \geq \epsilon
\]
\[
f \left( x - \delta \cdot \frac{g}{\|g\|} \right) \geq f(x) - \frac{\delta}{2} \cdot \|g\|.
\]
Towards a Min-Norm Element: A Sketch

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

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

Goldstein’s descent

Not satisfying

Goldstein’s descent
Towards a Min-Norm Element: A Sketch

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

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

Goldstein descent

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

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 \( \langle u, g \rangle \leq \frac{1}{2} \|g\|^2 \).
Towards a Min-Norm Element: A Sketch

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 $\langle u, g \rangle \leq \frac{1}{2} \|g\|^2$. 

"Inner Product Oracle"

Suppose $f$ were differentiable along $[x, x - \delta \cdot g \|g\|]$. Then, we have $\frac{1}{2} \|g\| \geq f(x) - f(x - \delta g \|g\|) \delta = \frac{1}{\delta} \int_{\tau = 0}^{\delta} \langle \nabla f(x - \tau g \|g\|), g \|g\| \rangle d\tau$.

since Goldstein descent not satisfied

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

Using randomization, we get this result without the 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 $\langle u, g \rangle \leq \frac{1}{2} \|g\|^2$.

"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 $[x, x - \delta \cdot \frac{g}{\|g\|}]$. strongly 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 $[x, x - \delta \cdot \frac{g}{\|g\|}]$. 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 \( [x, x - \delta \cdot \frac{g}{\|g\|}] \). 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 \([x, x - \delta \cdot \frac{g}{\|g\|}]\). Then, we have

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

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

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 \( \mathbb{E} \left\langle \nabla f(y), g \right\rangle \leq \frac{1}{2} \| g \|^2 \).

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

Compute $g = M/i.sc/n.scN/o.sc/r.sc/m.sc(x_t, \delta, \epsilon)$

Update $x_{t+1} = x_t - \delta g \|g\|$ (Goldstein descent step)

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 $x_T$
First, Zhang et al (2020)’s Algorithm

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

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 $y_k \sim \text{u.a.r.} \left[ x, x - \delta \frac{g_k}{\|g_k\|} \right]$
   - 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 $g = \text{MINNORM}(x_t, \delta, \epsilon)$
   ▶ Update $x_{t+1} = x_t - \delta \frac{g}{\|g\|}$
2. Return $x_T$

Zhang et al (2020)’s $\text{MINNORM}(x, \delta, \epsilon)$

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 $y_k \sim_{u.a.r.} \left[x, x - \delta \frac{g_k}{\|g_k\|}\right]$
   ▶ 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 $g = \text{MINNORM}(x_t, \delta, \epsilon)$
   - Update $x_{t+1} = x_t - \delta \frac{g}{\|g\|}$
2. Return $x_T$

**Our MINNORM($x, \delta, \epsilon$)**

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 $g_{k+1} = \arg \min_{z \in \left[g_k, u_k\right]} \|z\|$, 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:

\[
\text{given } x, g \in \mathbb{R}^d, \text{ solve the auxiliary convex feasibility problem: find } u \in \partial f(x) \text{ subject to } \langle u, g \rangle = f'(x, g).
\]

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

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

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

- The set $\partial f(x)$ could be extremely complicated
Zhang et al. (2020)’s algorithm requires the following oracle access: given $x, g \in \mathbb{R}^d$, solve the auxiliary convex feasibility problem

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

- 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 \((x, \delta, \epsilon)\)

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 \(y_k \sim u.a.r. \left[x, x - \delta \frac{\xi_k}{||\xi_k||}\right]\), where \(\xi_k \sim B_r(g_k)\)
   - 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 MinNorm \((x, \delta, \epsilon)\), and let \(\tau\) be its termination time. Then, for a fixed \(k \geq 0\), we have \(\mathbb{E}[||g_k||^2 1_{\tau > k}] \leq \frac{L^2}{1 + k}\).
Guarantee of Our MinNorm Subroutine

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

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

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

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

**Proof.** Let \( \hat{u} := u/\|u\| \); 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 \( \text{MinNorm}(x, \delta, \epsilon) \), and let \( \tau \) be its termination time. Then, for a fixed \( k \geq 0 \), we have

\[
\mathbb{E}[\|g_k\|^2 1_{\tau > k}] \leq \frac{L^2}{1+k}.
\]

### Proof

Let \( \hat{u} := u/\|u\| \); Then, almost surely, conditioned on \( g_k \), we have:

\[
\frac{1}{2} \|g_k\| \geq \frac{1}{\delta} \left[f(x) - f(x - \delta \hat{g}_k)\right]
\]

since Goldstein descent not satisfied

---

L - Lipschitzness 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 \( \text{MinNorm}(x, \delta, \epsilon) \), and let \( \tau \) be its termination time. Then, for a fixed \( k \geq 0 \), we have \( \mathbb{E}[\|g_k\|^2 1_{\tau > k}] \leq \frac{L^2}{1+k} \).

**Proof.** Let \( \hat{u} := u/\|u\| \); Then, almost surely, conditioned on \( g_k \), we have:

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

*\( L \)-Lipschitzness
Guarantee of Our MinNorm Subroutine

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

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

**Proof.** Let \( \hat{u} := u/\|u\| \); Then, almost surely, conditioned on \( g_k \), we have:

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

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

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

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

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

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

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

\( L \)-Lipschitzness
Guarantee of Our MinNorm Subroutine

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

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

**Proof.** Let \( \hat{u} := u/\|u\| \); Then, almost surely, conditioned on \( g_k \), we have:

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

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

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

\[
= \mathbb{E}_k \langle \nabla f(y_k), \hat{g}_k \rangle - 2L \|\hat{g}_k - \hat{\xi}_k\|.
\]

**definition of \( y_k \)**
Guarantee of Our MinNorm Subroutine

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

Let \( \{g_\ell\} \) be generated by \( \text{MinNorm}(x, \delta, \epsilon) \), and let \( \tau \) 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 \( \hat{u} := u/\|u\| \); Then, almost surely, conditioned on \( g_k \), we have:

\[
\frac{1}{2} \|g_k\| \geq \frac{1}{\delta} [f(x) - f(x - \delta \hat{g}_k)] \geq \frac{1}{\delta} [f(x) - f(x - \delta \hat{\xi}_k)] - L \|\hat{g}_k - \hat{\xi}_k\| \\
= \frac{1}{\delta} \int_{s=0}^{\delta} \langle \nabla f(x - s \hat{\xi}_k), \hat{\xi}_k \rangle ds - L \|\hat{g}_k - \hat{\xi}_k\| \\
\geq \frac{1}{\delta} \int_{s=0}^{\delta} \langle \nabla f(x - s \hat{\xi}_k), \hat{g}_k \rangle ds - 2L \|\hat{g}_k - \hat{\xi}_k\| \\
= \mathbb{E}_k \langle \nabla f(y_k), \hat{g}_k \rangle - 2L \|\hat{g}_k - \hat{\xi}_k\|. 
\]
Guarantee of Our MinNorm Subroutine

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

Let \( \{g_\ell\} \) be generated by MinNorm\((x, \delta, \epsilon)\), and let \( \tau \) 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 \( \hat{u} := u/\|u\| \); Then, almost surely, conditioned on \( g_k \), we have:

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

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

\[
= \mathbb{E}_k \langle \nabla f(y_k), \hat{g}_k \rangle - 2L \|\hat{g}_k - \hat{\xi}_k\|.
\]

This matches the requirement for \( u \in \partial_\delta f(x) \) with \( \langle u, g \rangle \leq \frac{1}{2} \|g\|^2 \). \( \blacksquare \)
Guarantee of Our MinNorm Subroutine

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

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

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

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

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

$$\left\lceil \frac{4\Delta}{\delta\epsilon} \right\rceil \cdot \left\lceil \frac{64L^2}{\epsilon^2} \right\rceil \cdot \left\lceil 2 \log \left( \frac{4\Delta}{\gamma\delta\epsilon} \right) \right\rceil$$

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 $\min_{g \in \partial f(x_T)} \|g\| \leq \epsilon$ in at most

\[
\left\lceil \frac{4\Delta}{\delta \epsilon} \right\rceil \cdot \left\lceil \frac{64L^2}{\epsilon^2} \right\rceil \cdot \left\lceil 2 \log \left( \frac{4\Delta}{\gamma \delta \epsilon} \right) \right\rceil
\]

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 $\min_{g \in \partial f(x_T)} \|g\| \leq \epsilon$ in at most

$$\left\lceil \frac{4\Delta}{\delta \epsilon} \right\rceil \cdot \left\lceil \frac{64L^2}{\epsilon^2} \right\rceil \cdot \left\lceil 2 \log \left( \frac{4\Delta}{\gamma \delta \epsilon} \right) \right\rceil$$

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 \( g \in \partial_{\delta} f(x) \) \textbf{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 \textbf{the desired stationarity} or is \textbf{a descent direction}
Problem Overview

Recall that $g \in \partial_\delta f(x)$ \textbf{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 \textbf{the desired stationarity} or is \textbf{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
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 $\langle u, g \rangle \leq \frac{\epsilon}{2} \|g\|$.

**Our Key Insight.**
The above oracle is essentially the gradient oracle of the MinNorm element problem.
Our Main Idea

Recall that given $g \in \partial f(x)$ not satisfying the descent condition, we can output $u \in \partial f(x)$ 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 $\hat{x} := x / \|x\|$ for some vector $x$
Using the Inner Product Oracle

**Notation** Denote $Q := \partial_\delta f(x)$; and $\hat{x} := x/\|x\|$ 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 $g^*_Q \in \min_{g \in Q} \|g\| \geq \epsilon/2$. Then, $\hat{g}^*_Q \in \{ w \in \mathbb{R}^d : \langle u, \hat{g} - w \rangle \leq 0 \}$. 
Using the Inner Product Oracle

**Notation** Denote $Q := \partial_\delta f(x)$; and $\hat{x} := x/\|x\|$ 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 $g^*_Q \in \min_{g \in Q} \|g\| \geq \epsilon/2$. Then, $\hat{g}^*_Q \in \{w \in \mathbb{R}^d : \langle u, \hat{g} - w \rangle \leq 0\}$.

**Proof** Combining the above definitions and a technical lemma gives:
Using the Inner Product Oracle

**Notation** Denote $Q := \partial \delta f(x)$; and $\hat{x} := x/\|x\|$ 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 $g^*_Q \in \min_{g \in Q} \|g\| \geq \epsilon/2$. Then, $\hat{g}^*_Q \in \{w \in \mathbb{R}^d : \langle u, \hat{g} - w \rangle \leq 0\}$.

**Proof** Combining the above definitions and a technical lemma gives:

The inner product oracle guarantees: $\langle u, \hat{g} \rangle \leq \frac{\epsilon}{2}$
Using the Inner Product Oracle

**Notation** Denote $Q := \partial_\delta f(x)$; and $\hat{x} := x/\|x\|$ 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 $g^*_Q \in \min_{g \in Q} \|g\| \geq \epsilon/2$. Then, $g^*_Q \in \{w \in \mathbb{R}^d : \langle u, \hat{g} - w \rangle \leq 0\}$.

**Proof** Combining the above definitions and a technical lemma gives:

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

The technical lemma (extra slide) shows: $\langle u, \hat{g}^*_Q \rangle \geq \|g^*_Q\|$
Using the Inner Product Oracle

**Notation** Denote $Q := \partial_\delta f(x)$; and $\hat{x} := x/\|x\|$ 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 $g^*_Q \in \min_{g \in Q} \|g\| \geq \epsilon/2$. Then, $\hat{g}^*_Q \in \{w \in \mathbb{R}^d : \langle u, \hat{g} - w \rangle \leq 0\}$.

**Proof** Combining the above definitions and a technical lemma gives:

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

The technical lemma (extra slide) shows: $\langle u, \hat{g}^*_Q \rangle \geq \|g^*_Q\|$

Combining these two inequalities yields: $\langle u, \hat{g} - \hat{g}^*_Q \rangle \leq \frac{\epsilon}{2} - \|g^*_Q\| \leq 0$
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 $\min_{g \in \partial f(x_T)} \|g\| \leq \epsilon$ in at most

$$\left\lceil \frac{4\Delta}{\delta \epsilon} \right\rceil \cdot \left\lceil 8d \log \left( \frac{8L}{\epsilon} \right) \right\rceil \cdot \left\lceil \frac{36L}{\epsilon} \right\rceil$$

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

$$\left\lceil \frac{4\Delta}{\delta \epsilon} \right\rceil \cdot \left\lceil 8d \log \left( \frac{8L}{\epsilon} \right) \right\rceil \cdot \left\lceil \frac{36L}{\epsilon} \right\rceil$$

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 $\min_{g \in \partial f(x_T)} \|g\| \leq \epsilon$ in at most

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

Goldstein descent iterations

Cutting-plane iterations
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

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

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 $\hat{x} := x/\|x\|$. 
A Technical Lemma

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

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

Let $g_Q^* \in \arg \min_Q \|g\|$. Then, $g_Q^*$ satisfies two properties:
A Technical Lemma

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

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

Let \( g^*_Q \in \arg \min_Q \|g\| \). Then, \( \hat{g}^*_Q \) satisfies two properties:

- \( \langle \hat{g}^*_Q, g \rangle \geq \|g^*_Q\| \) for all \( g \in Q \),
- \( \hat{g}^*_Q = \arg \max_{\|v\| \leq 1} \phi_Q(v) \).
A Technical Lemma

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

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

Let \( g^*_Q \in \arg\min_Q \|g\| \). Then, \( \hat{g}^*_Q \) satisfies two properties:

- \( \langle \hat{g}^*_Q, g \rangle \geq \|g^*_Q\| \) for all \( g \in Q \),
- \( \hat{g}^*_Q = \arg\max_{\|v\| \leq 1} \phi_Q(v) \).

**Proof.** The first inequality holds by definition of \( g^*_Q \). 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 $\hat{x} := x/\|x\|$. 

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

Let $g_Q^* \in \arg \min_Q \|g\|$. Then, $\hat{g}_Q^*$ satisfies two properties:

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

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

\[ \phi(\hat{g}^*) = \|g^*\| \]

first inequality

& definition of $\phi_Q$
A Technical Lemma

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

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

Let \( g_Q^* \in \arg \min_Q \|g\| \). Then, \( \hat{g}_Q^* \) satisfies two properties:

1. \( \langle \hat{g}_Q^*, g \rangle \geq \|g_Q^*\| \) for all \( g \in Q \),
2. \( \hat{g}_Q^* = \arg \max_{\|v\| \leq 1} \phi_Q(v) \).

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

\[
\phi(\hat{g}^*) = \|g^*\| = \min_Q \|g\|
\]
**A Technical Lemma**

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

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

Let \( g^*_Q \in \arg \min_Q \|g\| \). Then, \( \hat{g}^*_Q \) satisfies two properties:

- \( \langle \hat{g}^*_Q, g \rangle \geq \|g^*_Q\| \) for all \( g \in Q \),
- \( \hat{g}^*_Q = \arg \max_{\|v\| \leq 1} \phi_Q(v) \).

**Proof.** The first inequality holds by definition of \( g^*_Q \). We drop \( Q \) for notational simplicity in the rest of the proof.

\[
\phi(\hat{g}^*) = \|\hat{g}^*\| = \min_Q \|g\| = \min_Q \max_{\|v\| \leq 1} \langle g, v \rangle
\]

\( \text{dual representation} \)
A Technical Lemma

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

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

Let $g_Q^* \in \text{arg min}_Q \|g\|$. Then, $g_Q^*$ satisfies two properties:

- $\langle g_Q^*, g \rangle \geq \|g_Q^*\|$ for all $g \in Q$,
- $\hat{g}_Q^* = \text{arg max}_{\|v\| \leq 1} \phi_Q(v)$.

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

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

Sion’s minmax theorem
A Technical Lemma

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

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

Let $g_Q^* \in \arg \min_Q \|g\|$. Then, $\hat{g}_Q^*$ satisfies two properties:

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

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

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

Definition of $\phi$
**A Technical Lemma**

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

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

Let \( g_Q^* \in \arg \min_Q \|g\| \). Then, \( \hat{g}_Q^* \) satisfies two properties:

1. \( \langle \hat{g}_Q^*, g \rangle \geq \|g_Q^*\| \) for all \( g \in Q \),
2. \( \hat{g}_Q^* = \arg \max_{\|v\| \leq 1} \phi_Q(v) \).

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

\[
\phi(\hat{g}^*) = \|g^*\| = \min_Q \|g\| = \min_Q \max_{\|v\| \leq 1} \langle g, v \rangle = \max_{\|v\| \leq 1} \min_Q \langle g, v \rangle = \max_{\|v\| \leq 1} \phi(v).
\]
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!