Homeworks due today
Project questions?
Nonconvexity...
Nonconvex: hardness of global optima

Does there exist a subset of \( \{a_1, \ldots, a_n\} \) that sums to \( s \)?
Nonconvex: hardness of global optima

Does there exist a subset of \( \{a_1, \ldots, a_n\} \) that sums to \( s \)?

\textsc{SubsetSum}, well-known to be NP-Hard
Nonconvex: hardness of global optima

Does there exist a subset of \( \{a_1, \ldots, a_n\} \) that sums to \( s \)?

**SUBSETSUM**, well-known to be NP-Hard

**SUBSETSUM via nonconvex opt**

\[
\begin{align*}
\min_z & \quad (\sum_{i=1}^{n} z_i a_i - s)^2 + \sum_i z_i (1 - z_i) \\
\text{s.t.} & \quad 0 \leq z_i \leq 1, \ i = 1, \ldots, n.
\end{align*}
\]
Nonconvex: hardness of global optima

Does there exist a subset of \( \{a_1, \ldots, a_n\} \) that sums to \( s \)?

**SUBSETSUM**, well-known to be NP-Hard

**SUBSETSUM via nonconvex opt**

\[
\begin{align*}
\min_z & \quad (\sum_{i=1}^{n} za_i - s)^2 + \sum_i z_i (1 - z_i) \\
\text{s.t.} & \quad 0 \leq z_i \leq 1, \ i = 1, \ldots, n.
\end{align*}
\]

Is the **global min** of above problem equal to 0?
Nonconvex: hardness of global optima

Does there exist a subset of \( \{a_1, \ldots, a_n\} \) that sums to \( s \)?

**SUBSETSUM**, well-known to be NP-Hard

**SUBSETSUM via nonconvex opt**

\[
\min_z \left( \sum_{i=1}^{n} z_i a_i - s \right)^2 + \sum_i z_i (1 - z_i)
\]

s.t. \( 0 \leq z_i \leq 1, \ i = 1, \ldots, n \).

Is the **global min** of above problem equal to 0?

**Concrete proof of intractability**

To be pedantic, need to care for model of computing used.
Nonconvex: what about local minima?

Let \( f(x) = (1 - 1/s)^{\max_i |x_i| - \min_i |x_i| + |a^T x|} \) where \( a \in \mathbb{Z}^n_+ \), \( s = \sum_i a_i \geq 1 \).

Clearly \( f(0) = 0 \), but!

NP-Hard to decide if there's an \( x \) s.t. \( f(x) < 0 \)?

Assume \( y \in \{\pm 1\}^n \) satisfies \( a^T y = 0 \). Then, \( f(y) = -1/s \).

Let \( \max_i |x_i| = 1 \) and \( \delta = |a^T x| \).

If \( f(x) < 0 \), then \( |x_i| > 1 - 1/s + \delta \) for \( 1 \leq i \leq n \).

If \( y_i = sgn \ x_i \); then \( y_i x_i > 1 - 1/s + \delta \) and \( |y_i - x_i| = 1 - y_i x_i < 1 - \delta \); so \( |a^T y| \leq |a^T x| + |a^T (y - x)| \leq \delta + \max_i |y_i - x_i| < (1 - s) \delta + 1 \leq 1 \).

Since \( a \in \mathbb{Z}^n_+ \), this is possible iff \( a^T y = 0 \) (latter is like subset-sum).
Nonconvex: what about local minima?

Let \( f(x) = (1 - \frac{1}{s}) \max_i |x_i| - \min_i |x_i| + |a^T x| \)
where \( a \in \mathbb{Z}^n_+, s = \sum_i a_i \geq 1. \)

(Ref: Example due to Y. Nesterov.)
Nonconvex: what about local minima?

Let \( f(x) = \left(1 - \frac{1}{s}\right) \max_i |x_i| - \min_i |x_i| + |a^T x| \)

where \( a \in \mathbb{Z}_+^n \), \( s = \sum_i a_i \geq 1 \).

(Ref: Example due to Y. Nesterov.)

Clearly \( f(0) = 0 \), but!
Let \( f(x) = (1 - \frac{1}{s}) \max_i |x_i| - \min_i |x_i| + |a^T x| \)
where \( a \in \mathbb{Z}^n_+, s = \sum_i a_i \geq 1. \)

(Ref: Example due to Y. Nesterov.)

Clearly \( f(0) = 0, \text{ but!} \)

NP-Hard to decide if there's an \( x \) s.t. \( f(x) < 0? \)
Nonconvex: what about local minima?

Let \( f(x) = (1 - \frac{1}{s}) \max_i |x_i| - \min_i |x_i| + |a^T x| \)
where \( a \in \mathbb{Z}_+^n, s = \sum_i a_i \geq 1. \)

(Ref: Example due to Y. Nesterov.)

Clearly \( f(0) = 0, \) but!

NP-Hard to decide if there’s an \( x \) s.t. \( f(x) < 0? \)

- Assume \( y \in \{\pm 1\}^n \) satisfies \( a^T y = 0. \) Then, \( f(y) = -1/s. \)
Nonconvex: what about local minima?

Let \( f(x) = (1 - \frac{1}{s}) \max_i |x_i| - \min_i |x_i| + |a^T x| \)

where \( a \in \mathbb{Z}_+^n, s = \sum_i a_i \geq 1. \)

(Ref: Example due to Y. Nesterov.)

Clearly \( f(0) = 0, \) but!

NP-Hard to decide if there’s an \( x \) s.t. \( f(x) < 0? \)

- Assume \( y \in \{ \pm 1 \}^n \) satisfies \( a^T y = 0. \) Then, \( f(y) = -1/s. \)
- Let \( \max_i |x_i| = 1 \) and \( \delta = |a^T x| \)
Nonconvex: what about local minima?

Let \( f(x) = (1 - \frac{1}{s}) \max_i |x_i| - \min_i |x_i| + |a^T x| \)

where \( a \in \mathbb{Z}_+^n, s = \sum_i a_i \geq 1. \)

(Ref: Example due to Y. Nesterov.)

Clearly \( f(0) = 0, \) but!

NP-Hard to decide if there’s an \( x \) s.t. \( f(x) < 0? \)

- Assume \( y \in \{ \pm 1 \}^n \) satisfies \( a^T y = 0. \) Then, \( f(y) = -1/s. \)
- Let \( \max_i |x_i| = 1 \) and \( \delta = |a^T x| \)
- If \( f(x) < 0, \) then \( |x_i| > 1 - \frac{1}{s} + \delta \) for \( 1 \leq i \leq n \)
Nonconvex: what about local minima?

Let \( f(x) = (1 - \frac{1}{s}) \max_i |x_i| - \min_i |x_i| + |a^T x| \)

where \( a \in \mathbb{Z}^n_+, s = \sum_i a_i \geq 1. \)

(Ref: Example due to Y. Nesterov.)

Clearly \( f(0) = 0 \), but!

NP-Hard to decide if there’s an \( x \) s.t. \( f(x) < 0? \)

- Assume \( y \in \{\pm 1\}^n \) satisfies \( a^T y = 0 \). Then, \( f(y) = -1/s. \)
- Let \( \max_i |x_i| = 1 \) and \( \delta = |a^T x| \)
- If \( f(x) < 0 \), then \( |x_i| > 1 - \frac{1}{s} + \delta \) for \( 1 \leq i \leq n \)
- If \( y_i = \text{sgn} \; x_i \); then \( y_i x_i > 1 - \frac{1}{s} + \delta \) and \( |y_i - x_i| = 1 - y_i x_i < \frac{1}{s} - \delta \); so

\[
|a^T y| \leq |a^T x| + |a^T (y - x)| \leq \delta + s \max_i |y_i - x_i|
\]

\[
< (1 - s)\delta + 1 \leq 1.
\]
Nonconvex: what about local minima?

Let \( f(x) = (1 - \frac{1}{s}) \max_i |x_i| - \min_i |x_i| + |a^T x| \)
where \( a \in \mathbb{Z}_+^n, s = \sum_i a_i \geq 1. \)

(Ref: Example due to Y. Nesterov.)

Clearly \( f(0) = 0, \) but!

NP-Hard to decide if there’s an \( x \) s.t. \( f(x) < 0? \)

- Assume \( y \in \{\pm 1\}^n \) satisfies \( a^T y = 0. \) Then, \( f(y) = -1/s. \)
- Let \( \max_i |x_i| = 1 \) and \( \delta = |a^T x| \)
- If \( f(x) < 0, \) then \( |x_i| > 1 - \frac{1}{s} + \delta \) for \( 1 \leq i \leq n \)
- If \( y_i = \text{sgn} \ x_i; \) then \( y_i x_i > 1 - \frac{1}{s} + \delta \) and \( |y_i - x_i| = 1 - y_i x_i < \frac{1}{s} - \delta; \) so
  \[
  |a^T y| \leq |a^T x| + |a^T (y - x)| \leq \delta + s \max_i |y_i - x_i| \\
  < (1 - s)\delta + 1 \leq 1.
  \]
- Since \( a \in \mathbb{Z}_+^n, \) this is possible iff \( a^T y = 0 \) (latter is like subset-sum)
Convex but hard
Hardness due to a fundamental failure

Consider the following subset of real symmetric matrices:

\[ CP_n := \{ A \in \mathbb{S}^{n \times n} \mid x^T A x \geq 0 \text{ for all } x \geq 0 \} \]
Consider the following subset of real symmetric matrices:

\[
CP_n := \{ A \in S^{n \times n} \mid x^T A x \geq 0 \text{ for all } x \geq 0 \}
\]

**Exercise:** Verify that \( CP_n \) is a convex cone.

**Challenge.** Given matrix \( A \), decide if \( A \in CP_n \)?

---

Suvrit Sra (suvrit@mit.edu) 6.881 Optimization for Machine Learning (3/02/21; Lecture 5)
Hardness due to a fundamental failure

Consider the following subset of real symmetric matrices:

\[ CP_n := \{ A \in S^{n \times n} \mid x^T A x \geq 0 \text{ for all } x \geq 0 \} \]

**Exercise:** Verify that \( CP_n \) is a convex cone.

**Challenge.** Given matrix \( A \), decide if \( A \in CP_n \)?

\[
\begin{align*}
\min_x & \quad x^T A x \\
\text{s.t.} & \quad x \geq 0 \\
\text{Is there an } x \text{ s.t. } x^T A x < 0? \\
\text{Is } x = 0 \text{ a local min?}
\end{align*}
\]

Amounts to checking if \( A \) is copositive, known to be co-NPC (which implies that checking copositivity is NP-Hard).

Explore: the topic "testing copositivity".

Hardness due to a fundamental failure

Consider the following subset of real symmetric matrices:

\[ CP_n := \{ A \in S^{n \times n} \mid x^T A x \geq 0 \text{ for all } x \geq 0 \} \]

**Exercise:** Verify that \( CP_n \) is a convex cone.

**Challenge.** Given matrix \( A \), decide if \( A \in CP_n \)?

\[
\min_x \quad x^T A x \quad \text{s.t. } x \geq 0
\]

Is there an \( x \) s.t. \( x^T A x < 0 \)?

Is \( x = 0 \) a local min?

Amounts to checking if \( A \) is *copositive*, known to be co-NPC (which implies that checking copositivity is NP-Hard).
Consider the following subset of real symmetric matrices:

$$CP_n := \{ A \in \mathbb{S}^{n \times n} \mid x^T A x \geq 0 \text{ for all } x \geq 0 \}$$

**Exercise:** Verify that $CP_n$ is a convex cone.

**Challenge.** Given matrix $A$, decide if $A \in CP_n$?

$$\min_x \ x^T A x \ \text{s.t.} \ x \geq 0$$

Is there an $x$ s.t. $x^T A x < 0$?

Is $x = 0$ a local min?

Amounts to checking if $A$ is **copositive**, known to be co-NPC (which implies that checking copositivity is NP-Hard).

**Explore:** the topic “testing copositivity”.

Exercise: Verify that the following matrix is copositive

\[ A := \begin{bmatrix} 1 & -1 & 1 & 1 & -1 \\ -1 & 1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & 1 & -1 \\ -1 & 1 & 1 & -1 & 1 \end{bmatrix}. \]
Exercise: Verify that the following matrix is copositive

\[
A := \begin{bmatrix}
1 & -1 & 1 & 1 & -1 \\
-1 & 1 & -1 & 1 & 1 \\
1 & -1 & 1 & -1 & 1 \\
1 & 1 & -1 & 1 & -1 \\
-1 & 1 & 1 & -1 & 1
\end{bmatrix}.
\]

Exercise: Non-negative matrix factorization (NMF) seeks to solve

\[
\min_{B, C \geq 0} \|A - BC\|_F^2,
\]

for a given \( A \geq 0 \) (elementwise). Restricting \( C = B^T \), rewrite NMF as a “copositive programming” problem.
Maximizing convex functions

Theorem. Let \( f \) be a convex function and let \( C = \text{conv} S \), where \( S \) is an arbitrary set of points. Then,

\[
\sup \{ f(x) | x \in C \} = \sup \{ f(x) | x \in S \},
\]

where the first sup is attained only when the second one is.

Theorem. Let \( f \) be convex; \( C \) be a closed convex set in \( \text{dom} f \). Suppose \( C \) contains no lines. Then, if the sup of \( f \) relative to \( C \) is attained at all, it is attained at some extreme point of \( C \).

Example: LP optimum at a vertex (vertices extreme points for polyhedra)

Maximizing convex functions

**Theorem.** Let $f$ be a convex function and let $C = \text{conv } S$, where $S$ is an arbitrary set of points. Then,

$$\sup \{f(x) \mid x \in C\} = \sup \{f(x) \mid x \in S\},$$

where the first sup is attained only when the second one is.
Maximizing convex functions

**Theorem.** Let $f$ be a convex function and let $C = \text{conv } S$, where $S$ is an *arbitrary* set of points. Then,

$$\sup \{ f(x) \mid x \in C \} = \sup \{ f(x) \mid x \in S \},$$

where the first sup is attained only when the second one is.

**Theorem.** Let $f$ be convex; $C$ be a closed convex set in $\text{dom } f$. Suppose $C$ contains no lines. Then, if the sup of $f$ relative to $C$ is attained at all, it is attained at some *extreme point* of $C$.

**Example:** LP optimum at a vertex (vertices extreme points for polyhedra)

**Ref.** See Section 32 of R. T. Rockafellar, Convex Analysis.
How hard is global opt?
Complexity of global optimization

How much computation required to ensure $f(x) - f^* \leq \epsilon$?

How to measure complexity?
How much computation required to ensure
\[ f(x) - f^* \leq \epsilon? \]

How to measure complexity?

**Oracle** based complexity: count number of calls to an “oracle”
How much computation required to ensure $f(x) - f^* \leq \epsilon$?

How to measure complexity?

**Oracle** based complexity: count number of calls to an “oracle”

- **Zeroth order** oracle: inputs a point $x$, outputs $f(x)$
- **First-order** oracle: inputs a point $x$, outputs $f(x), \nabla f(x)$

Higher order oracles can also be considered; also, later, we’ll consider other oracles (stochastic, inexact, etc.)
Complexity of global optimization

How much computation required to ensure

\[ f(x) - f^* \leq \epsilon \]
Complexity of global optimization

How much computation required to ensure $f(x) - f^* \leq \epsilon$?

Problem: $f^* = \min_x \{f(x) \mid x \in [0, 1]^n\}$
Complexity of global optimization

How much computation required to ensure 
\[ f(x) - f^* \leq \epsilon? \]

**Problem:** \[ f^* = \min_{x} \{ f(x) \mid x \in [0, 1]^n \} \]

**Problem class:** \( f \) is **L-Lipschitz** on \([0, 1]^n\)
\[
|f(x) - f(y)| \leq L\|x - y\|_\infty \quad \text{for constant } L \text{ and } x, y \in [0, 1]^n.
\]
Complexity of global optimization

How much computation required to ensure

\[ f(x) - f^* \leq \epsilon? \]

Problem: \( f^* = \min_x \{ f(x) \mid x \in [0, 1]^n \} \)

Problem class: \( f \) is \( L \)-Lipschitz on \([0, 1]^n\)

\[ |f(x) - f(y)| \leq L \|x - y\|_{\infty} \]

for constant \( L \) and \( x, y \in [0, 1]^n \).

Algorithm: Brute force search.

- Pick integer \( p \geq 1 \) and place a uniform grid (width \( 1/2p \))
  over \([0, 1]^n\) centered around \( p^n \) points
Complexity of global optimization

How much computation required to ensure
\[ f(x) - f^* \leq \epsilon? \]

**Problem:** \[ f^* = \min_x \{ f(x) \mid x \in [0, 1]^n \} \]

**Problem class:** \( f \) is \( L \)-Lipschitz on \( [0, 1]^n \)
\[ |f(x) - f(y)| \leq L\|x - y\|_\infty \] for constant \( L \) and \( x, y \in [0, 1]^n \).

**Algorithm:** Brute force search.

- Pick integer \( p \geq 1 \) and place a uniform grid (width \( 1/2p \)) over \( [0, 1]^n \) centered around \( p^n \) points
- We can ensure \( f(\bar{x}) - f^* \leq L/2p \) in \( O(p^n) \) calls of oracle \( f(x) \)
Complexity of global optimization

How much computation required to ensure
\[ f(x) - f^* \leq \epsilon? \]

**Problem:** \( f^* = \min_x \{ f(x) \mid x \in [0, 1]^n \} \)

**Problem class:** \( f \) is \( L \)-Lipschitz on \([0, 1]^n\)

\[ |f(x) - f(y)| \leq L \|x - y\|_{\infty} \] for constant \( L \) and \( x, y \in [0, 1]^n \).

**Algorithm:** Brute force search.

- Pick integer \( p \geq 1 \) and place a uniform grid (width \( 1/2p \)) over \([0, 1]^n\) centered around \( p^n \) points
- We can ensure \( f(\bar{x}) - f^* \leq L/2p \) in \( O(p^n) \) calls of oracle \( f(x) \)
- (this translates into \( O((L/2\epsilon)^n) \) for \( p \geq L/2\epsilon \))
Complexity of global optimization

How much computation required to ensure $f(x) - f^* \leq \epsilon$?

Problem: $f^* = \min_x \{ f(x) \mid x \in [0, 1]^n \}$

Problem class: $f$ is $L$-Lipschitz on $[0, 1]^n$

$|f(x) - f(y)| \leq L\|x - y\|_\infty$ for constant $L$ and $x, y \in [0, 1]^n$.

Algorithm: Brute force search.

- Pick integer $p \geq 1$ and place a uniform grid (width $1/2p$) over $[0, 1]^n$ centered around $p^n$ points
- We can ensure $f(\bar{x}) - f^* \leq L/2p$ in $O(p^n)$ calls of oracle $f(x)$
- (this translates into $O((L/2\epsilon)^n)$ for $p \geq L/2\epsilon$)

The brute force method is worst-case optimal!
Constructing the lower bound

Idea: Create “resisting” oracles.

\[ p = \lfloor L_2 \varepsilon \rfloor. \]

Suppose, we have a method that needs \( N < p \) oracle calls to solve problems to accuracy \( \varepsilon \) in problem class.

◦ Resisting oracle

Return \( f(x) = 0 \) at any test point \( x \)

(\( \text{s.t.} \) \( f(\bar{x}) = 0 \))

But \( N < p \), so there's a box with no test points.

Thus, put \( x^* \) inside this box of width \( \varepsilon / L \) and set \( f(x) = \min\{0, L \|x - x^*\| - \varepsilon\} \)
Constructing the lower bound

**Idea:** Create “resisting” oracles.

Let \( p = \lfloor \frac{L}{2\epsilon} \rfloor \). Suppose, we have a method that needs \( N < p^n \) oracle calls to solve problems to accuracy \( \epsilon \) in problem class.
Constructing the lower bound

Idea: Create “resisting” oracles.
Let \( p = \left\lfloor \frac{L}{2\epsilon} \right\rfloor \). Suppose, we have a method that needs \( N < p^n \) oracle calls to solve problems to accuracy \( \epsilon \) in problem class.

Resisting oracle

Return \( f(x) = 0 \) at any test point \( x \)

(so method can only find \( \bar{x} \in [0, 1]^n \) s.t. \( f(\bar{x}) = 0 \))

But \( N < p^n \), so there’s a box with no test points.
Constructing the lower bound

Idea: Create “resisting” oracles.
Let \( p = \left\lfloor \frac{L}{2\epsilon} \right\rfloor \). Suppose, we have a method that needs \( N < p^n \) oracle calls to solve problems to accuracy \( \epsilon \) in problem class.

Resisting oracle

<table>
<thead>
<tr>
<th>Return ( f(x) = 0 ) at any test point ( x )</th>
</tr>
</thead>
</table>

(so method can only find \( \bar{x} \in [0, 1]^n \) s.t. \( f(\bar{x}) = 0 \))

But \( N < p^n \), so there’s a box with no test points.

Thus, put \( x^* \) inside this box of width \( \epsilon/L \) and set

\[
f(x) = \min \{ 0, L\|x - x^*\| - \epsilon \}
\]
Lower bound for global optimization

\[ f(x) = \min \{0, L\|x - x^*\| - \epsilon\} \]

This function is $L$-Lipschitz, its accuracy is $\epsilon$.

Thus, without at least $p^n$ points, accuracy cannot be better than $\epsilon$. 
Lower bound for global optimization

\[ f(x) = \min \{ 0, L\|x - x^*\| - \epsilon \} \]

This function is \( L \)-Lipschitz, its accuracy is \( \epsilon \).

Thus, without at least \( p^n \) points, accuracy cannot be better than \( \epsilon \).

In general, brute force (exponential time) method the best. Moreover, vastly worse than “just” \( 2^n \)!

**Exercise:** Provide similar lower bounds for \( C^1 \) functions.

**Ref.** Section 1.1 of *Yu. Nesterov, “Lectures on Convex Optimization”*
Stationarity

(More modest goal)
More modest goal: stationarity

First-order necessary condition

Assuming \( f \in C^1, \nabla f(x) = 0 \) necessary

Weak requirement: \( \|\nabla f(x)\| \leq \epsilon \)
More modest goal: stationarity

First-order necessary condition

Assuming $f \in C^1$, $\nabla f(x) = 0$ necessary

**Weak requirement:** $||\nabla f(x)|| \leq \epsilon$

Consider $f(x) = x^3$ on the set $[-1, 1]$. Global opt is at $-1$, while $f'(x) = 3x^2 = 0$ as $x = 0$. 
More modest goal: stationarity

First-order necessary condition

Assuming \( f \in C^1 \), \( \nabla f(x) = 0 \) necessary

**Weak requirement:** \( \|\nabla f(x)\| \leq \epsilon \)

Consider \( f(x) = x^3 \) on the set \([-1, 1]\). Global opt is at \(-1\), while \( f'(x) = 3x^2 = 0 \) as \( x = 0 \).

Second-order necessary conditions

Assume \( f \in C^2 \). Then, \( \nabla f(x) = 0 \) and \( \nabla^2 f(x) \succeq 0 \)
More modest goal: stationarity

First-order necessary condition

Assuming $f \in C^1$, $\nabla f(x) = 0$ necessary

**Weak requirement:** $\|\nabla f(x)\| \leq \epsilon$

Consider $f(x) = x^3$ on the set $[-1, 1]$. Global opt is at $-1$, while $f'(x) = 3x^2 = 0$ as $x = 0$.

Second-order necessary conditions

Assume $f \in C^2$. Then, $\nabla f(x) = 0$ and $\nabla^2 f(x) \succeq 0$

Second-order sufficient conditions (local opt)

Assume $f \in C^2$. Then, $\nabla f(x) = 0$ and $\nabla^2 f(x) \succ 0$
Second-order necessary conditions

Assume $f \in C^2$. Then, $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \succeq 0$.

Taylor expand $f(x^* + td)$, where $d$ is arbitrary and $t > 0$:

$$f(x^* + td) = f(x^*) + t\nabla f(x^*)^T d + \frac{t^2}{2} d^T \nabla^2 f(x^*) d + o(t^2).$$
Assume $f \in C^2$. Then, $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \succeq 0$.

Taylor expand $f(x^* + td)$, where $d$ is arbitrary and $t > 0$:

$$f(x^* + td) = f(x^*) + t\nabla f(x^*)^T d + \frac{t^2}{2} d^T \nabla^2 f(x^*) d + o(t^2).$$

Since $x^*$ is a local min, $\nabla f(x^*) = 0$ holds. Thus,

$$\frac{f(x^* + td) - f(x^*)}{t^2} = \frac{1}{2} d^T \nabla^2 f(x^*) d + \frac{o(t^2)}{t^2}.$$
Second-order necessary conditions

Assume $f \in C^2$. Then, $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \succeq 0$

Taylor expand $f(x^* + td)$, where $d$ is arbitrary and $t > 0$:

$$f(x^* + td) = f(x^*) + t\nabla f(x^*)^T d + \frac{t^2}{2} d^T \nabla^2 f(x^*) d + o(t^2).$$

Since $x^*$ is a local min, $\nabla f(x^*) = 0$ holds. Thus,

$$\frac{f(x^* + td) - f(x^*)}{t^2} = \frac{1}{2} d^T \nabla^2 f(x^*) d + \frac{o(t^2)}{t^2}$$

Since $x^*$ is local min, for small enough $t$ lhs above is $\geq 0$. Thus,

$$0 \leq \lim_{t \downarrow 0} \frac{1}{2} d^T \nabla^2 f(x^*) d + \frac{o(t^2)}{t^2} \implies d^T \nabla^2 f(x^*) d \geq 0 \iff \nabla^2 f(x^*) \succeq 0.$$
Sufficient condition

Assume $f \in C^2$, $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \succ 0$.

Exercise: Prove that $x^*$ is a local minimum. (*Hint:* Analyze $f(x^* + y) - f(x^*)$ via Taylor series, use $\nabla^2 f(x^*) \succeq \delta I$ for some $\delta > 0$.)
**Sufficient condition**

Assume \( f \in C^2, \nabla f(x^*) = 0 \) and \( \nabla^2 f(x^*) \succ 0 \).

**Exercise:** Prove that \( x^* \) is a local minimum. (*Hint:* Analyze \( f(x^* + y) - f(x^*) \) via Taylor series, use \( \nabla^2 f(x^*) \succeq \delta I \) for some \( \delta > 0 \)).

**Remark:** It can still happen that \( \nabla^2 f(x^*) \not \succ 0 \) but \( x^* \) is a local min (e.g., consider \( f(x) = x^4 + 2 \) at \( x = 0 \)). Such critical points are called *degenerate*; functions without degenerate critical points called “Morse functions” (*Explore!*).
Sufficient condition

Assume \( f \in C^2, \nabla f(x^*) = 0 \) and \( \nabla^2 f(x^*) \succ 0 \).

**Exercise:** Prove that \( x^* \) is a local minimum. (*Hint:* Analyze \( f(x^* + y) - f(x^*) \) via Taylor series, use \( \nabla^2 f(x^*) \succeq \delta I \) for some \( \delta > 0 \).)

**Remark:** It can still happen that \( \nabla^2 f(x^*) \not\succ 0 \) but \( x^* \) is a local min (e.g., consider \( f(x) = x^4 + 2 \) at \( x = 0 \)). Such critical points are called *degenerate*; functions without degenerate critical points called “Morse functions” (*Explore!*).

---

**Useful convergence criterion:** \((\epsilon, \delta)\)-stationarity

\[
\|\nabla f(x)\|_2 \leq \epsilon \text{ and } \nabla^2 f(x) \succeq -\sqrt{\delta} I
\]
Nonsmooth & Nonconvex

(Introduction)
First-order conditions

- For convex, $0 \in \partial f$ necessary and sufficient for global opt.
First-order conditions

- For convex, $0 \in \partial f$ necessary and sufficient for global opt.
- For nonconvex, we hope for only (first-order) stationarity.
First-order conditions

► For convex, $0 \in \partial f$ necessary and sufficient for global opt.
► For nonconvex, we hope for only (first-order) stationarity.

How should we define $\partial f$?
How to generalize $\partial f$?

- If $f$ is nonsmooth, nonconvex, $\partial f$ defined via
  
  $\partial f(x) := \{g \mid f(y) \geq f(x) + \langle g, y - x \rangle \ \forall \ y\}$ not helpful!

- It is a global notion; we seek a local one.

- Regularity assumption: locally Lipschitz functions

For convex functions, $\partial f$ intimately related to directional derivative $f'(x; d) := \lim_{t \downarrow 0} \frac{f(x + td) - f(x)}{t}$.

A key property of $f'(x; d)$ and $\partial f$: $f'(x; d) = \max \{\langle g, d \rangle \mid g \in \partial f(x)\}$.

Thus, generalize $\partial f$ via directional derivatives.
How to generalize $\partial f$?

- If $f$ is nonsmooth, nonconvex, $\partial f$ defined via $\partial f(x) := \{g \mid f(y) \geq f(x) + \langle g, y - x \rangle \ \forall \ y\}$ is not helpful!
- It is a global notion; we seek a local one.
- Regularity assumption: locally Lipschitz functions
How to generalize $\partial f$?

- If $f$ is nonsmooth, nonconvex, $\partial f$ defined via
  $$\partial f(x) := \{g \mid f(y) \geq f(x) + \langle g, y - x \rangle \ \forall \ y\}$$ not helpful!
- It is a global notion; we seek a local one.
- Regularity assumption: locally Lipschitz functions

For convex functions, $\partial f$ intimately related to *directional derivative*

$$f'(x; d) := \lim_{t \downarrow 0} \frac{f(x + td) - f(x)}{t}.$$
How to generalize $\partial f$?

- If $f$ is nonsmooth, nonconvex, $\partial f$ defined via
  
  \[
  \partial f(x) := \{g \mid f(y) \geq f(x) + \langle g, y - x \rangle \ \forall \ y \} \text{ not helpful!}
  \]

- It is a global notion; we seek a local one.

- Regularity assumption: locally Lipschitz functions

For convex functions, $\partial f$ intimately related to *directional derivative*

\[
f'(x; d) := \lim_{t \downarrow 0} \frac{f(x + td) - f(x)}{t}.
\]

**A key property of $f'(x; d)$ and $\partial f$**

\[
f'(x; d) = \max \{ \langle g, d \rangle \mid g \in \partial f(x) \}\]
How to generalize $\partial f$?

► If $f$ is nonsmooth, nonconvex, $\partial f$ defined via
$\partial f(x) := \{g \mid f(y) \geq f(x) + \langle g, y - x \rangle \ \forall \ y\}$ not helpful!
► It is a global notion; we seek a local one.
► Regularity assumption: locally Lipschitz functions

For convex functions, $\partial f$ intimately related to directional derivative

$$f'(x; d) := \lim_{t \downarrow 0} \frac{f(x + td) - f(x)}{t}.$$ 

**A key property of $f'(x; d)$ and $\partial f$**

$$f'(x; d) = \max \{ \langle g, d \rangle \mid g \in \partial f(x) \}$$

Thus, generalize $\partial f$ via directional derivatives.
Clarke directional derivative

\[ f^\circ(x; d) := \limsup_{y \to x, t \downarrow 0} \frac{f(y + td) - f(y)}{t} \]
Clarke directional derivative

\[ f^\circ(x; d) := \limsup_{y \to x, t \downarrow 0} \frac{f(y + td) - f(y)}{t} \]

**Prop.** \( f^\circ(x; \cdot) \) is positively homogeneous and subadditive.

**Proof sketch:** homogeneity is clear; we prove subadditivity.

\[
\begin{align*}
f^\circ(x; u + v) &= \limsup_{y \to x} \frac{f(y + t(u + v)) - f(y)}{t} \\
&\leq \limsup_{y \to x} \frac{f(y + tu + tv) - f(y + tv)}{t} + \limsup_{y \to x} \frac{f(y + tv) - f(y)}{t} \\
&= f^\circ(x; u) + f^\circ(x; v).
\end{align*}
\]

(first limsup is \( f^\circ(x; u) \) since \( y + tv \) essentially dummy var converging to \( x \))

Exercises

**Exercise:** Let \( f(x) = x^2 \sin(1/x) \). This function is Lipschitz near 0. Show that \( f^\circ(0; v) = |v| \).

**Exercise:** What should \( \partial_o f(0) \) be? (Answer: \([-1, 1]\); why?)

**Exercise:** What is \( f^\circ(0; v) \) for \( f = -|x| \)? (Verify it is \( |v| \).)
Clarke subdifferential

\[ \partial f(x) := \{ g \in X \mid \langle g, d \rangle \leq f^\circ(x; d) \text{ for all } d \in X \}. \]

**Exercise:** Prove that \( \partial f(x) \) is a convex, compact set.
Clarke subdifferential

Clarke subdifferential

\[ \partial_\circ f(x) := \{ g \in X \mid \langle g, d \rangle \leq f^\circ(x; d) \text{ for all } d \in X \} . \]

Exercise: Prove that \( \partial_\circ f(x) \) is a convex, compact set.

**Theorem.**

**A.** When \( f \) is \( C^1 \), \( \partial_\circ f(x) = \{ \nabla f(x) \} \).

**B.** If \( f \) is convex, then \( \partial_\circ f(x) = \partial f(x) \).
Clarke subdifferential

Theorem. A. When $f$ is $C^1$, $\partial \circ f(x) = \{\nabla f(x)\}$.

B. If $f$ is convex, then $\partial \circ f(x) = \partial f(x)$.

Prop. Let $f \in C^0_L$. $f^\circ(x; d) = \max \{\langle g, d \rangle | g \in \partial \circ f(x)\}$

Proof: Assume $\exists v$ s.t. $f^\circ(x; v)$ exceeds the given max. Then, there exists (why?) a linear functional $\zeta$ majorized by $f^\circ(x; v)$ agreeing with it at $v$. It follows that $\zeta \in \partial \circ f(x)$, leading to a contradiction.

(we used definition of $\partial \circ f$ along with sublinearity of $f^\circ(x; \cdot)$)

Exercise: Prove that for a locally Lipschitz function, $f'(x; d)$ is the support function of the (convex) set $\partial \circ f(x)$.

Exercise: Prove that $\partial \circ f(x)$ is a convex, compact set.
Theorem. Necessary condition for optimality: $0 \in \partial f(x)$
**Theorem.** Necessary condition for optimality: $0 \in \partial f(x)$

**Proof:** Since $\partial(-f) = -\partial f$, suffices to consider when $x$ is a local minimum. When $x$ is a local min, as before, starting from

$$ \frac{f(y + td) - f(y)}{t} $$

evident that $f^\circ(x; d) \geq 0$. Thus, $\zeta = 0$ belongs to $\partial f(x)$ because of the “max-rule” which implies that

$$ \zeta \in \partial f(x) \iff f^\circ(x; d) \geq \langle \zeta, d \rangle \quad \forall d \in X. $$
Theorem. Necessary condition for optimality: $0 \in \partial \circ f(x)$

Proof: Since $\partial(-f) = -\partial f$, suffices to consider when $x$ is a local minimum. When $x$ is a local min, as before, starting from

$$\frac{f(y + td) - f(y)}{t}$$

evident that $f\circ(x; d) \geq 0$. Thus, $\zeta = 0$ belongs to $\partial \circ f(x)$ because of the “max-rule” which implies that

$$\zeta \in \partial \circ f(x) \quad \text{iff} \quad f\circ(x; d) \geq \langle \zeta, d \rangle \quad \forall \, d \in X.$$  

Could use $\text{dist}(0, \partial \circ f(x)) \leq \epsilon$ as stationarity criterion
Theorem. Let $f \in C^1$ and $g$ convex. Then, $\partial(f + g) = \nabla f + \partial g$
Theorem. Let \( f \in C^1 \) and \( g \) convex. Then, \( \partial(f + g) = \nabla f + \partial g \)

Theorem. If \( f \) and \( g \) are LL around a point \( x \in X \), then \( \partial(f + g)(x) \subset \partial f(x) + \partial g(x) \)
Clarke subdifferential – key properties

**Theorem.** Let \( f \in C^1 \) and \( g \) convex. Then, \( \partial (f + g) = \nabla f + \partial g \)

**Theorem.** If \( f \) and \( g \) are LL around a point \( x \in X \), then
\[
\partial (f + g)(x) \subset \partial f(x) + \partial g(x)
\]

Recalling Rademacher’s theorem, we can “simplify” \( \partial f \)

**Theorem.** An LL function is a.e. differentiable
Clarke subdifferential – key properties

**Theorem.** Let $f \in C^1$ and $g$ convex. Then, $\partial_\circ (f + g) = \nabla f + \partial g$

**Theorem.** If $f$ and $g$ are LL around a point $x \in X$, then $\partial_\circ (f + g)(x) \subset \partial f(x) + \partial g(x)$

Recalling Rademacher’s theorem, we can “simplify” $\partial_\circ f$

**Theorem.** An LL function is a.e. differentiable

**Theorem.** Let $f$ be LL around $x \in X$ and let $S \subset X$ have measure zero. Then, $\partial_\circ f(x) = \text{conv} \{ \lim_r \nabla f(x^r) \mid x^r \to x, x^r \notin S \}$

**Corollary.** Approximate $\partial_\circ f(x)$ using “gradient sampling”