1 Introduction

Many popular ML models involve nondifferentiable objective functions, e.g., Lasso introduced as a special case of weighted least squares models. We generalize the concept of gradient for differentiable functions to the so-called subgradient for nondifferentiable convex functions.

2 Subgradients and Subdifferentials

Definition 1. A function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is called proper if

1. $\exists x \in \mathbb{R}^n$, such that $f(x) < \infty$;
2. $f(x) > -\infty$, $\forall x \in \mathbb{R}^n$.

Definition 2. Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ be a proper convex function and let $x \in \text{dom } f$. A vector $g \in \mathbb{R}^n$ such that

$$f(y) \geq f(x) + \langle g, y - x \rangle, \forall y \in \mathbb{R}^n$$

is called a subgradient of $f$ at $x$.

![Figure 1: A subgradient.](image)

Question 1. In Definition 2, shall we ask $y \in \text{dom } f$?

Example 1. Consider function $f(x) = |x|, x \in \mathbb{R}$. Find the subgradient of $f$ at 0.

Solution: Let $g \in \partial f(0)$. Then

$$f(y) = |y| \geq f(0) + g(y - 0) = gy.$$  

Clearly, the above inequality holds for all $y \in \mathbb{R}$ if and only if $g \in [-1, 1]$. Thus, we have

$$\partial f(0) = [-1, 1],$$
which is not unique.

**Remark 1 (A geometric interpretation of subdifferential).** Inspired by Fig. 1, we can link the subgradient of \( f \) to its epigraph. Indeed, for any \((y, t) \in \text{epi } f\), we have
\[
t \geq f(y) \geq f(x) + \langle g, y - x \rangle,
\]
which can be rewritten as
\[
\begin{pmatrix} g \\ -1 \end{pmatrix} \cdot \begin{pmatrix} y \\ t \end{pmatrix} - \begin{pmatrix} x \\ f(x) \end{pmatrix} \leq 0.
\]
(2)
The inequality (2) is the variational inequality characterizing the projection of a point lying on the ray with base \((x, f(x))\) and direction \((g, -1)\) onto the set \(\text{epi } f\).

Furthermore, Fig. 1 implies that the vector \((g, -1)\) determines a hyperplane supporting \(\text{epi } f\) at the point \((x, f(x))\).

**Definition 3.** The set of all subgradients of \( f \) at \( x \) is called the subdifferential of \( f \) at \( x \) and is denoted by \( \partial f(x) \).

## 3 Subdifferential Calculus

**Theorem 1.** Let \( f : \mathbb{R}^n \to \mathbb{R} \) be convex and \( x \in \text{int}(\text{dom } f) \). Then, \( f \) is locally Lipschitz continuous at \( x \), that is, \( \exists \epsilon > 0 \) and \( M \geq 0 \) such that
\[
|f(y) - f(x)| \leq M \|y - x\|, \forall \{y : \|y - x\| \leq \epsilon\}.
\]

**Theorem 2.** [1] Let \( f : \mathbb{R}^n \to \mathbb{R} \) be convex and let \( x \in \text{int}(\text{dom } f) \). Then
1. the subdifferential \( \partial f(x) \) is a nonempty, bounded, closed, and convex set;
2. for any \( v \in \mathbb{R}^n \), we have
\[
f'(x; v) = \lim_{t \downarrow 0} \frac{f(x + tv) - f(x)}{t} = \max_{g \in \partial f(x)} \langle v, g \rangle,
\]
where \( f'(x; v) \) is the directional derivative of \( f \) at \( x \) along the direction \( v \);
3. if \( f \) is differentiable at \( x \), then \( \partial f(x) = \{\nabla f(x)\} \).

**Proof.**

1. We first show that \( \partial f(x) \) is nonempty.

As the point \((x, f(x))\) is a boundary point of \(\text{epi } f\), the supporting hyperplane theorem implies that we can separate \((x, f(x))\) and \(\text{epi } f\) by a hyperplane. That is, there exists a \((d, \alpha) \in \mathbb{R}^{n+1}\) and \((d, \alpha) \neq 0\) such that
\[
\langle (d, \alpha), (y, t) \rangle \leq \langle (d, \alpha), (x, f(x)) \rangle, \forall (y, t) \in \text{epi } f,
\]
which can be rewritten as
\[
\langle d, y \rangle + \alpha t \leq \langle d, x \rangle + \alpha f(x), \forall (y, t) \in \text{epi } f.
\]
(3)
As the inequality (3) holds for all \((y,t) \in \text{epi } f\), we conclude \(\alpha \leq 0\). We further claim that \(\alpha \neq 0\). Suppose not, that is, \(\alpha = 0\) (and thus \(d \neq 0\)), the inequality (3) becomes
\[
\langle d, y - x \rangle \leq 0, \forall (y,t) \in \text{epi } f.
\]
As \(x \in \text{int } (\text{dom } f)\), there exists a small number \(\epsilon > 0\) such that \(x + \epsilon d \in \text{dom } f\). Replacing \(y\) in (4) by \(x + \epsilon d\) leads to a contradiction. Thus, we must have \(\alpha < 0\). Then, by replacing \(t\) by \(f(y)\) in (3) and dividing both sides by \(\alpha\), we have
\[
f(y) \geq f(x) + \langle -d/\alpha, y - x \rangle, \forall y,
\]
which implies that \(-d/\alpha \in \partial f(x)\). Therefore, the set \(\partial f(x)\) is nonempty.

We next show the boundedness of \(\partial f(x)\). Due to Theorem 1, we can find an \(\epsilon > 0\) and \(M \geq 0\) such that \(\forall \|y - x\| \leq \epsilon\), we have
\[
|f(y) - f(x)| \leq M\|y - x\|.
\]
For any \(g \in \partial f(x)\) and \(g \neq 0\), we choose
\[
x' = x + \epsilon g/\|g\|,
\]
which leads to
\[
\epsilon \|g\| = \langle g, x' - x \rangle \leq f(x') - f(x) \leq M\|x' - x\| = M\epsilon.
\]
Thus, \(\partial f(x)\) is bounded.

The closedness and convexity of \(\partial f(x)\) can be seen from its definition that, it is the intersection of a set of closed half-spaces.

2. We omit the proof here.

3. For any \(v \in \mathbb{R}^n\) and \(g \in \partial f(x)\), we have
\[
\langle \nabla f(x), v \rangle = f'(x; v) \geq \langle g, v \rangle.
\]
Changing the sign of \(v\), we conclude that
\[
\langle \nabla f(x), v \rangle = \langle g, v \rangle.
\]
By letting \(v = e_k, k = 1, \ldots, n\), we have \(g = \nabla f(x)\).  

\(\Box\)

**Lemma 1.** [2] Suppose that \(f : \mathbb{R}^n \to \mathbb{R}\) is a convex function. For \(\alpha > 0\), let \(h(x) = \alpha f(x)\). Then, \(h\) is convex, and \(\partial h(x) = \alpha \partial f(x)\) for every \(x\).

**Proof.** We show the result directly from the definition. Indeed, \(g \in \partial f(x)\) if and only if for all \(y\)
\[
h(y) = \alpha f(y) \geq \alpha [f(x) + \langle g, y - x \rangle] = h(x) + \langle \alpha g, y - x \rangle,
\]
which implies that \(\alpha g \in \partial h(x)\).  

\(\Box\)
Lemma 2. [2] Suppose that \( f : \mathbb{R}^m \to \mathbb{R} \) is a convex function, \( A \in \mathbb{R}^{m \times n} \), and \( b \in \mathbb{R}^m \). Let \( h(x) = f(Ax + b) \). Then, for any \( x \), we have
\[
\partial h(x) = A^T \partial f(Ax + b).
\]
Proof. We show the result directly from the definition. Indeed, we have \( g \in \partial f(Ax + b) \) if and only if
\[
h(y) = f(Ay + b) \geq f(Ax + b) + \langle g, Ay - Ax \rangle = h(x) + \langle A^T g, y - x \rangle,
\]
which implies that \( A^T g \in \partial h(x) \).

Theorem 3 (Moreau-Rockafellar Theorem). [2] Assume that \( f = f_1 + f_2 \), where \( f_i : \mathbb{R}^n \to \mathbb{R} \), \( i = 1, 2 \), are convex proper functions. If there exists a point \( x_0 \in \text{dom} \ f \) such that \( f_1 \) is continuous at \( x_0 \), then
\[
\partial f(x) = \partial f_1(x) + \partial f_2(x), \forall x \in \text{dom} \ f.
\]

Definition 4. A convex function is called closed if its epigraph is a closed set.

Lemma 3. [1] Let functions \( f_i(x), i = 1, \ldots, m \), be closed and convex. Then function \( f(x) = \max_{1 \leq i \leq m} f_i(x) \) is also closed and convex. For any \( x \in \text{int} (\text{dom} \ f) = \bigcap_{i=1}^m \text{int} (\text{dom} \ f_i) \), we have
\[
\partial f(x) = \text{conv} \{ \partial f_i(x) : i \in I(x) \},
\]
where \( I(x) = \{ i : f_i(x) = f(x) \} \).

Lemma 4. Let \( \Delta \) be a set and \( f(x) = \sup \{ \phi(y, x) : y \in \Delta \} \). Suppose that for any fixed \( y \in \Delta \), the function \( \phi(y, x) \) is closed and convex in \( x \). Then, \( f(x) \) is closed and convex. For and \( x \) from
\[
\text{dom} \ f = \{ x \in \mathbb{R}^n : \exists \gamma \text{ such that } \phi(y, x) \leq \gamma, \forall y \in \Delta \},
\]
we have
\[
\partial f(x) \supseteq \text{conv} \{ \partial \phi_{x}(y, x) : y \in I(x) \},
\]
where \( I(x) = \{ y : \phi(y, x) = f(x) \} \). When \( \Delta \) is compact and \( \phi(y, x') \) is continuous (upper semicontinuous) in \( y \) for all \( x' \) in a neighborhood of \( x \), we get an equality above.

Example 2. Consider function \( f(x) = |x|, x \in \mathbb{R} \). Find \( \partial f(x) \).

Solution: We find \( \partial f(x) \) by two different approaches.

1. We have derived that \( \partial f(0) = [-1, 1] \). Moreover, by noting that \( f(x) \) is differentiable for \( x \neq 0 \), we have
\[
\partial f(x) = \begin{cases} 
1, & \text{if } x > 0, \\
[-1, 1], & \text{if } x = 0, \\
-1, & \text{if } x < 0.
\end{cases}
\]

2. Let \( f_1(x) = x \) and \( f_2(x) = -x \). Clearly, we have \( \partial f_1(x) = \{ \nabla f_1(x) \} = \{ 1 \} \), and similarly \( \partial f_2(x) = \{ -1 \} \).
Moreover, it is easy to see that $f(x) = \max\{f_1(x), f_2(x)\}$, and thus
\[
\partial f(x) = \text{conv} \{\partial f_i(x) : f_i(x) = f(x)\}
\]
is given by
\[
\begin{cases}
1, & \text{if } x > 0, \\
[-1, 1], & \text{if } x = 0, \\
-1, & \text{if } x < 0.
\end{cases}
\]

Example 3. Let $f(x) = \|x\|_1$, where $x \in \mathbb{R}^n$. Find $\partial f(x)$.

Solution: We compute $\partial f(x)$ by two different approaches.

1. By Lemma 2 and Theorem 3, we have
\[
f(x) = \|x\|_1 = \sum_{i=1}^n |x_i| = \sum_{i=1}^n |e_i^T x|
\]
\[
\Rightarrow \partial f(x) = \partial \left( \sum_{i=1}^n |e_i^T x| \right) = \sum_{i=1}^n \partial |e_i^T x| = \sum_{i=1}^n e_i \partial |x_i|
\]
is given by
\[
\begin{cases}
v \in \mathbb{R}^n : v_i = \begin{cases}
1, & \text{if } x_i > 0, \\
[-1, 1], & \text{if } x_i = 0, \\
-1, & \text{if } x_i < 0.
\end{cases}
\end{cases}
\]

2. By Lemma 3, we have
\[
f(x) = \|x\|_1 = \sum_{i=1}^n |x_i| = \max \{ \langle s, x \rangle : s \in \mathbb{R}^n, |s_i| = 1, \forall i \}
\]
\[
\Rightarrow \partial f(x) = \text{conv} \{ s : s \in \mathbb{R}^n, |s_i| = 1, \forall i, \langle s, x \rangle = \|x\|_1 \}
\]
is given by
\[
\begin{cases}
v \in \mathbb{R}^n : v_i = \begin{cases}
1, & \text{if } x_i > 0, \\
[-1, 1], & \text{if } x_i = 0, \\
-1, & \text{if } x_i < 0.
\end{cases}
\end{cases}
\]

Example 4. Let $f(x) = \|x\|_\infty$, where $x \in \mathbb{R}^n$. Find $\partial f(x)$.

Solution: We compute $\partial f(x)$ by two different approaches.

1. Let $f_i(x) = |x_i|, i = 1, 2, \ldots, n$, where $x \in \mathbb{R}^n$. Then we have
\[
f(x) = \max_{1 \leq i \leq n} f_i(x).
\]
\[
\text{It’s easy to see that } f_i \text{ is closed and convex for } i = 1, 2, \ldots, n.
\]
We first find \( \partial f_i(x) \), \( i = 1, 2, \ldots, n \). Since \( f_i(x) = \max\{x_i, -x_i\} \), we have

\[
\partial f_i(x) = \begin{cases} 
  e_i, & x_i > 0, \\
  \text{conv} \{-e_i, e_i\}, & x_i = 0, \\
  -e_i, & x_i < 0.
\end{cases}
\]

Note that \( \text{conv} \{-e_i, e_i\} \) is the line segment connecting \(-e_i\) and \( e_i \).

By Lemma 3, we have

\[
\partial f(0) = \text{conv} \{ \text{conv} \{-e_i, e_i\} : i = 1, 2, \ldots, n \}
\]

\[
= \{ x \in \mathbb{R}^n : \sum_{i=1}^{n} |x_i| \leq 1 \}.
\]

Besides, for all \( x = (x_1, \ldots, x_n) \in \mathbb{R}^n \setminus \{0\} \), suppose that

\[
\Delta_x = \{ i : |x_i| = \|x\|_\infty \} = \{i_{\alpha}\}_{\alpha=1}^{m} \cup \{j_{\beta}\}_{\beta=1}^{k},
\]

where

\[
x_{i_{\alpha}} > 0, \quad \alpha = 1, \ldots, m,
\]

\[
x_{j_{\beta}} < 0, \quad \beta = 1, \ldots, k.
\]

Hence we have

\[
\partial f(x) = \text{conv} \{ \partial f_i(x) : i \in \Delta_x \}
\]

\[
= \text{conv} \{ e_{i_1}, \ldots, e_{i_m}, -e_{j_1}, \ldots, -e_{j_k} \}
\]

\[
= \left\{ y \in \mathbb{R}^n : \sum_{i=1}^{n} \varepsilon_i y_i = 1, \varepsilon_i y_i \geq 0, y_i = 0 \text{ if } \varepsilon_i = 0 \right\},
\]

where \( \varepsilon_i \) is defined as

\[
\varepsilon_i = \begin{cases} 
  1, & x_i = \|x\|, \\
  0, & |x_i| < \|x\|, \\
  -1, & x_i = -\|x\|.
\end{cases}
\]

Therefore, we have

\[
\partial f(x) = \begin{cases} 
  \{ y \in \mathbb{R}^n : \|y\|_1 \leq 1 \}, & x = 0, \\
  \{ y \in \mathbb{R}^n : \sum_{i=1}^{n} \varepsilon_i y_i = 1, \varepsilon_i y_i \geq 0, y_i = 0 \text{ if } \varepsilon_i = 0 \}, & x \neq 0.
\end{cases}
\]

2. By Lemma 4, we have

\[
f(x) = \|x\|_\infty = \max_{1 \leq i \leq n} |x_i| = \sup_{1 \leq i \leq n} \langle x, y \rangle = \|y\|_1 \leq 1 \}
\]

\[
\Rightarrow \partial f(x) = \text{conv} \{ y \in \mathbb{R}^n : \langle x, y \rangle = \|x\|_\infty, \|y\|_1 \leq 1 \}.
\]
It’s easy to see that \( \{ y \in \mathbb{R}^n : \langle x, y \rangle = 1, \| y \|_1 \leq 1 \} \) is convex. Hence we have

\[
\partial f(x) = \{ y \in \mathbb{R}^n : \langle x, y \rangle = \| x \|_\infty, \| y \|_1 \leq 1 \}, \quad \forall x \in \mathbb{R}^n.
\]

**Question 2.** We got two forms of \( \partial f(x) \) by two approaches. Are they the same?

**Example 5.** Let \( f : \mathbb{S}^n \to \mathbb{R} \) be defined by \( f(X) = \lambda_{\max}(X) \). Find \( \partial f(X) \) [3].

**Solution:**

By eigen-decomposition, a symmetric matrix can be written as

\[
X = U\sigma U^\top,
\]

where \( U^\top U = I \) and \( \sigma = \text{diag}(\lambda_1, \ldots, \lambda_n) \) with \( \lambda_1 \geq \cdots \geq \lambda_n \). Let \( U = (u_1, \ldots, u_n) \), i.e., \( u_i \) is the \( i \)th eigenvector corresponding to \( \lambda_i \). We then write \( f(X) \) as the maximum of a set of linear functions over \( X \):

\[
f(X) = \max \{ \langle s, Xs \rangle : \| s \| = 1 \}
= \max \{ \langle ss^\top, X \rangle : \| s \| = 1 \}
\]

Assume that \( \lambda_{\max} = \lambda_1 = \cdots = \lambda_r \), where \( 1 \leq r \leq n \). We can see that \( u_i \in \arg\max_{\| s \|=1} \langle ss^\top, X \rangle \), \( i = 1, \ldots, r \). Let \( U^r = (u_1, \ldots, u_r) \). Then,

\[
S^* := \arg\max_{\| s \|=1} \langle ss^\top, X \rangle = \{ v : v \in \text{span } U^r, \| v \| = 1 \}
= \{ v : v = U^r Q, Q \in \mathbb{R}^{r \times r}, Q^\top Q = I \}

\Rightarrow \partial f(X) = \text{conv} \left\{ vv^\top : v \in S^* \right\} = \left\{ U^r G(U^r)^\top : G \succeq 0, \text{trace } G = 1 \right\}.
\]
References

