EE3001 Machine Learning **Fall 2023**

Lecture 07. Subdifferential

Lecturer: Jie Wang Date: Oct 26, 2023

Last Update: November 2, 2023

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 $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ ($\bar{\mathbb{R}} := \mathbb{R} \cup \{\infty, -\infty\}$) is called *proper* if

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

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

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle,\ \forall\, \mathbf{y} \in \mathbb{R}^n \tag{1}$$

is called a *subgradient* of f at $\mathbf{x}$.

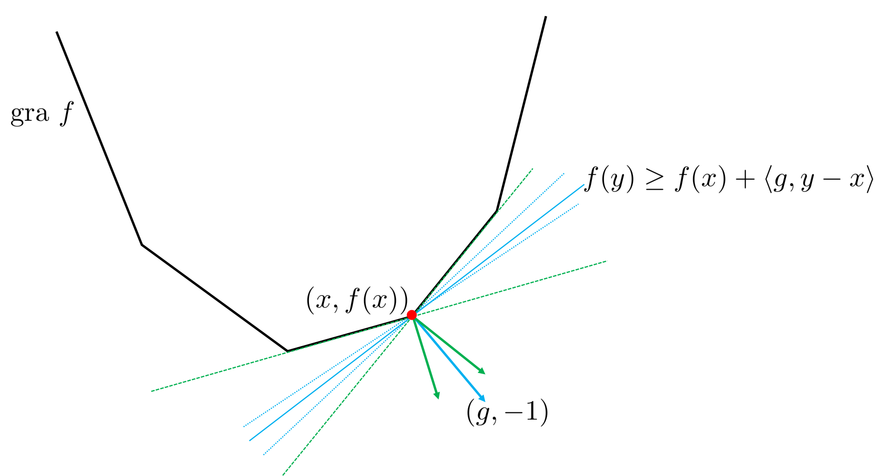

Figure 1: A subgradient.

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

***Remark* 1.** In view of Definition 2, the subgradient is defined for **convex** functions.

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* 2 (A geometric interpretation of subdifferential).** Inspired by Fig. 1, we can link the subgradient of f to its epigraph. Indeed, for any $(\mathbf{y}, t) \in \mathbf{epi}\ f$, we have

$$t \geq f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle,$$

which can be rewritten as

$$\left\langle \begin{pmatrix} \mathbf{g} \\ -1 \end{pmatrix}, \begin{pmatrix} \mathbf{y} \\ t \end{pmatrix} - \begin{pmatrix} \mathbf{x} \\ f(\mathbf{x}) \end{pmatrix} \right\rangle \leq 0. \tag{2}$$

The inequality (2) is the **variational inequality** characterizing the projection of a point lying on the ray with base $(\mathbf{x}, f(\mathbf{x}))$ and direction $(\mathbf{g}, -1)$ onto the set $\mathbf{epi}\ f$.

Furthermore, Fig. 1 implies that the vector $(\mathbf{g}, -1) \in \mathbb{R}^{n+1}$ determines a hyperplane supporting $\mathbf{epi}\ f$ at the point $(\mathbf{x}, f(\mathbf{x}))$. Can you find the expression of this hyperplane?

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

Theorem 1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and $\mathbf{x} \in \mathbf{int}\,(\mathbf{dom}\ f)$. Then, f is locally Lipschitz continuous at $\mathbf{x}$, that is, $\exists\ \epsilon > 0$ and $M \geq 0$ such that*

$$|f(\mathbf{y}) - f(\mathbf{x})| \leq M\|\mathbf{y} - \mathbf{x}\|, \forall \{\mathbf{y} : \|\mathbf{y} - \mathbf{x}\| \leq \epsilon\}.$$

***Remark* 3.** The value of the parameter M in Theorem 1 may depend on $\mathbf{x}$.

Theorem 2. [1] *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and let $\mathbf{x} \in \mathbf{int}\,(\mathbf{dom}\ f)$. Then*

1. *the subdifferential $\partial f(\mathbf{x})$ is a nonempty, bounded, closed, and convex set;*
2. *for any $\mathbf{v} \in \mathbb{R}^n$, we have*

$$f'(\mathbf{x}; \mathbf{v}) = \lim_{t \downarrow 0} \frac{f(\mathbf{x} + t\mathbf{v}) - f(\mathbf{x})}{t} = \max_{\mathbf{g} \in \partial f(\mathbf{x})} \langle \mathbf{v}, \mathbf{g} \rangle,$$

   where $f'(\mathbf{x}; \mathbf{v})$ is the directional derivative of f at $\mathbf{x}$ along the direction $\mathbf{v}$;

3. *if f is differentiable at $\mathbf{x}$, then $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$.*

Proof.

1. We first show that $\partial f(\mathbf{x})$ is **nonempty**. The working horse is the **supporting hyperplane theorem**.

As the point $(\mathbf{x}, f(\mathbf{x}))$ is a boundary point of $\textbf{epi}\ f$, the supporting hyperplane theorem implies that we can separate $(\mathbf{x}, f(\mathbf{x}))$ and $\textbf{epi}\ f$ by a hyperplane. That is, there exists a $(\mathbf{d}, \alpha) \in \mathbb{R}^{n+1}$ and $(\mathbf{d}, \alpha) \neq 0$ such that

$$\langle(\mathbf{d}, \alpha), (\mathbf{y}, t)\rangle \leq \langle(\mathbf{d}, \alpha), (\mathbf{x}, f(\mathbf{x}))\rangle, \forall (\mathbf{y}, t) \in \textbf{epi}\ f,$$

which can be rewritten as

$$\langle\mathbf{d}, \mathbf{y}\rangle + \alpha t \leq \langle\mathbf{d}, \mathbf{x}\rangle + \alpha f(\mathbf{x}), \forall (\mathbf{y}, t) \in \textbf{epi}\ f. \tag{3}$$

As the inequality (3) holds for all $(\mathbf{y}, t) \in \textbf{epi}\ f$, we conclude $\alpha \leq 0$. We further claim that $\alpha \neq 0$. Suppose not, that is, $\alpha = 0$ (and thus $\mathbf{d} \neq 0$), the inequality (3) becomes

$$\langle\mathbf{d}, \mathbf{y} - \mathbf{x}\rangle \leq 0, \forall (\mathbf{y}, t) \in \textbf{epi}\ f. \tag{4}$$

As $\mathbf{x} \in \textbf{int}\,(\textbf{dom}\ f)$, there exists a small number $\epsilon > 0$ such that $\mathbf{x} + \epsilon\mathbf{d} \in \textbf{dom}\ f$. Replacing $\mathbf{y}$ in (4) by $\mathbf{x} + \epsilon\mathbf{d}$ leads to a contradiction. Thus, we must have $\alpha < 0$. Then, by replacing t by $f(\mathbf{y})$ in (3) and dividing both sides by α, we have

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle -\mathbf{d}/\alpha, \mathbf{y} - \mathbf{x}\rangle, \forall \mathbf{y},$$

which implies that $-\mathbf{d}/\alpha \in \partial f(\mathbf{x})$. Therefore, the set $\partial f(\mathbf{x})$ is nonempty.

We next show the **boundedness** of $\partial f(\mathbf{x})$. As $\mathbf{x} \in \textbf{int}\,(\textbf{dom}\ f)$, we can find a a small number $\epsilon_1 > 0$ such that $\{\mathbf{y} : \|\mathbf{y} - \mathbf{x}\| < \epsilon_1\} \subseteq \textbf{dom}\ f$. Moreover, by Theorem 1, we can find an $\epsilon_2 > 0$ and $M \geq 0$ such that $\forall \|\mathbf{y} - \mathbf{x}\| \leq \epsilon_2$, we have

$$|f(\mathbf{y}) - f(\mathbf{x})| \leq M\|\mathbf{y} - \mathbf{x}\|.$$

Let $\epsilon = \min\{\epsilon_1, \epsilon_2\}$. For any $\mathbf{g} \in \partial f(\mathbf{x})$ and $\mathbf{g} \neq 0$, we choose

$$\mathbf{x}' = \mathbf{x} + \epsilon\mathbf{g}/\|\mathbf{g}\|,$$

which leads to

$$\epsilon\|\mathbf{g}\| = \langle\mathbf{g}, \mathbf{x}' - \mathbf{x}\rangle \leq f(\mathbf{x}') - f(\mathbf{x}) \leq M\|\mathbf{x}' - \mathbf{x}\| = M\epsilon.$$

Thus, $\partial f(\mathbf{x})$ is bounded.

The **closedness** and **convexity** of $\partial f(\mathbf{x})$ 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 $\mathbf{v} \in \mathbb{R}^n$ and $\mathbf{g} \in \partial f(\mathbf{x})$, we have

$$\langle\nabla f(\mathbf{x}), \mathbf{v}\rangle = f'(\mathbf{x}; \mathbf{v}) \geq \langle\mathbf{g}, \mathbf{v}\rangle.$$

Changing the sign of $\mathbf{v}$, we conclude that

$$\langle\nabla f(\mathbf{x}), \mathbf{v}\rangle = \langle\mathbf{g}, \mathbf{v}\rangle.$$

By letting $\mathbf{v} = \mathbf{e}_k$, $k = 1, \ldots, n$, we have $\mathbf{g} = \nabla f(\mathbf{x})$.

□

Question 2. Consider Theorem 2. The condition that $\mathbf{x} \in \textbf{int}\,(\textbf{dom}\ f)$ is fundamentally important in deriving the conclusions.

1. If $\mathbf{x} \in \textbf{dom}\ f$ but it is not an interior point of $\textbf{dom}\ f$, is it possible that $\partial f(\mathbf{x}) = \emptyset$?
2. If $x \in \textbf{relint}\,(\textbf{dom}\ f)$, is it possible that $\partial f(\mathbf{x})$ is unbounded?

3 Subdifferential Calculus

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

Proof. We show the result directly from the definition. Indeed, $\mathbf{g} \in \partial f(\mathbf{x})$ if and only if for all $\mathbf{y}$

$$h(\mathbf{y}) = \alpha f(\mathbf{y}) \geq \alpha[f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle] = h(\mathbf{x}) + \langle \alpha \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle,$$

which implies that $\alpha \mathbf{g} \in \partial h(\mathbf{x})$. □

Lemma 2. [2] *Suppose that* $f : \mathbb{R}^m \to \bar{\mathbb{R}}$ *is a convex function,* $A \in \mathbb{R}^{m \times n}$, *and* $\mathbf{b} \in \mathbb{R}^m$. *Let* $h(\mathbf{x}) = f(A\mathbf{x} + \mathbf{b})$. *Then, for any* $\mathbf{x}$, *we have*

$$\partial h(\mathbf{x}) = A^\top \partial f(A\mathbf{x} + \mathbf{b}).$$

Proof. We show the result directly from the definition. Indeed, we have $\mathbf{g} \in \partial f(A\mathbf{x} + \mathbf{b})$ if and only if

$$h(\mathbf{y}) = f(A\mathbf{y} + \mathbf{b}) \geq f(A\mathbf{x} + \mathbf{b}) + \langle \mathbf{g}, A\mathbf{y} - A\mathbf{x} \rangle = h(\mathbf{x}) + \langle A^\top \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle,$$

which implies that $A^\top \mathbf{g} \in \partial h(\mathbf{x})$. □

Theorem 3 (Moreau-Rockafellar Theorem). [2] *Assume that* $f = f_1 + f_2$, *where* $f_i : \mathbb{R}^n \to \bar{\mathbb{R}}$, $i = 1, 2$, *are convex proper functions. If there exists a point* $\mathbf{x}_0 \in \textbf{dom}\ f$ *such that* f_1 *is continuous at* $\mathbf{x}_0$, *then*

$$\partial f(\mathbf{x}) = \partial f_1(\mathbf{x}) + \partial f_2(\mathbf{x}), \forall\, \mathbf{x} \in \textbf{dom}\ f.$$

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

Lemma 3. [1] *Let functions* $f_i(\mathbf{x})$, $i = 1, \ldots, m$, *be closed and convex. Then function*

$$f(\mathbf{x}) = \max_{1 \leq i \leq m} f_i(\mathbf{x})$$

is also closed and convex. For any $\mathbf{x} \in \textbf{int}\,(\textbf{dom}\ f) = \cap_{i=1}^m \textbf{int}\,(\textbf{dom}\ f_i)$, *we have*

$$\partial f(\mathbf{x}) = \textbf{conv}\,\{\partial f_i(\mathbf{x}) : i \in \mathcal{I}(\mathbf{x})\},$$

where $\mathcal{I}(\mathbf{x}) = \{i : f_i(\mathbf{x}) = f(\mathbf{x})\}$.

Lemma 4. *Let* Δ *be a set and*

$$f(\mathbf{x}) = \sup\{\phi(\mathbf{y}, \mathbf{x}) : \mathbf{y} \in \Delta\}.$$

Suppose that for any fixed $\mathbf{y} \in \Delta$, *the function* $\phi(\mathbf{y}, \mathbf{x})$ *is closed and convex in* $\mathbf{x}$. *Then,* $f(\mathbf{x})$ *is closed and convex. For and* $\mathbf{x}$ *from*

$$\textbf{dom}\ f = \{\mathbf{x} \in \mathbb{R}^n : \exists\, \gamma \text{ such that } \phi(\mathbf{y}, \mathbf{x}) \leq \gamma, \forall\, \mathbf{y} \in \Delta\},$$

we have

$$\partial f(\mathbf{x}) \supseteq \textbf{conv}\,\{\partial \phi_{\mathbf{x}}(\mathbf{y}, \mathbf{x}) : \mathbf{y} \in \mathcal{I}(\mathbf{x})\},$$

where $\mathcal{I}(\mathbf{x}) = \{\mathbf{y} : \phi(\mathbf{y}, \mathbf{x}) = f(\mathbf{x})\}$. *When* Δ *is compact and* $\phi(\mathbf{y}, \mathbf{x}')$ *is continuous (upper semi-continuous) in* $\mathbf{y}$ *for all* $\mathbf{x}'$ *in a neighborhood of* $\mathbf{x}$, *we get an equality above.*

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

Solution: Clearly, $f(x)$ is a convex function. 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

$$\begin{aligned} \partial f(x) &= \mathbf{conv}\, \{\partial f_i(x) : f_i(x) = f(x)\} \\ &= \begin{cases} 1, & \text{if } x > 0, \\ [-1, 1], & \text{if } x = 0, \\ -1, & \text{if } x < 0. \end{cases} \end{aligned}$$

■

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

Solution: It is easy to see that $f(\mathbf{x})$ is a convex function. We compute $\partial f(\mathbf{x})$ by two different approaches.

1. By Lemma 2 and Theorem 3, we have

$$\begin{aligned} f(\mathbf{x}) = \|\mathbf{x}\|_1 &= \sum_{i=1}^n |x_i| = \sum_{i=1}^n |\mathbf{e}_i^\top \mathbf{x}| \\ \Rightarrow \partial f(\mathbf{x}) &= \partial \left(\sum_{i=1}^n |\mathbf{e}_i^\top \mathbf{x}| \right) = \sum_{i=1}^n \partial |\mathbf{e}_i^\top \mathbf{x}| = \sum_{i=1}^n \mathbf{e}_i \partial |x_i| \\ &= \left\{ \mathbf{v} \in \mathbb{R}^n : v_i = \begin{cases} 1, & \text{if } x_i > 0, \\ [-1, 1], & \text{if } x_i = 0, \\ -1, & \text{if } x_i < 0. \end{cases} \right\} \end{aligned}$$

2. We first write $f(\mathbf{x})$ as the supreme of a set of linear functions, that is,

$$f(\mathbf{x}) = \|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i| = \max\{\langle \mathbf{s}, \mathbf{x} \rangle : \mathbf{s} \in \mathbb{R}^n, |s_i| = 1, \forall i\}.$$

   Let $f_{\mathbf{s}}(\mathbf{x}) = \langle \mathbf{s}, \mathbf{x} \rangle$ and $\Delta = \{\mathbf{s} \in \mathbb{R}^n : |s_i| = 1, i = 1, \dots, n\}$. Then,

$$f(\mathbf{x}) = \|\mathbf{x}\|_1 = \max\{f_{\mathbf{s}}(\mathbf{x}) : \mathbf{s} \in \Delta\}.$$

Clearly, the function $f_{\mathbf{s}}(\mathbf{x})$ is continuously differentiable and $\nabla f_{\mathbf{s}}(\mathbf{x}) = \mathbf{s}$. Then, by Lemma 3, we have

$$\begin{aligned}\partial f(\mathbf{x}) &= \mathbf{conv}\,\{\mathbf{s} : \mathbf{s} \in \Delta, f_{\mathbf{s}}(\mathbf{x}) = \langle \mathbf{s}, \mathbf{x}\rangle = \|\mathbf{x}\|_1\} \\ &= \left\{ \mathbf{v} \in \mathbb{R}^n : v_i = \begin{cases} 1, & \text{if } x_i > 0, \\ [-1,1], & \text{if } x_i = 0, \\ -1, & \text{if } x_i < 0. \end{cases} \right\}\end{aligned}$$

■

Example 4. Let $f : \mathbb{R}^n \to \mathbb{R}$ be defined by $f(\mathbf{x}) = \max\{x_i, i = 1, \ldots, n\}$, where x_i is the i^{th} component of $\mathbf{x}$.

Solution: To see that $f(\mathbf{x})$ is convex, it suffices to note that

$$f(\mathbf{x}) = \max_{i=1,\ldots,n} \langle \mathbf{e}_i, \mathbf{x}\rangle.$$

Let $f_i(\mathbf{x}) = \langle \mathbf{e}_i, \mathbf{x}\rangle$ and $\mathcal{I} = \{1, 2, \ldots, n\}$. Clearly, $\nabla f_i(\mathbf{x}) = \mathbf{e}_i$. Thus, by Lemma 3, we have

$$\partial f(\mathbf{x}) = \mathbf{conv}\,\{\mathbf{e}_i : i \in \Delta, f_i(\mathbf{x}) = \langle \mathbf{e}_i, \mathbf{x}\rangle = f(\mathbf{x})\} = \{\mathbf{v} : \mathbf{v} \in \mathbb{R}^n_+, \|\mathbf{v}\|_1 = 1, \langle \mathbf{v}, \mathbf{x}\rangle = f(\mathbf{x})\}.$$

■

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

Solution: From the last lecture, we have shown that $f(X)$ is a convex function. By eigen-decomposition, a symmetric matrix can be written as

$$X = U \Lambda U^\top,$$

where $U^\top U = I$ and $\Lambda = \mathbf{diag}\,(\lambda_1, \ldots, \lambda_n)$ with $\lambda_1 \geq \cdots \geq \lambda_n$. Let $U = (\mathbf{u}_1, \ldots, \mathbf{u}_n)$, i.e., $\mathbf{u}_i$ is the eigenvector corresponding to λ_i. We then write $f(X)$ as the maximum of a set of linear functions over X:

$$f(X) = \max\{\langle \mathbf{s}, X\mathbf{s}\rangle : \|\mathbf{s}\| = 1\} = \max\{\langle \mathbf{s}\mathbf{s}^\top, X\rangle : \|\mathbf{s}\| = 1\},$$

where

$$\langle X, Y\rangle = \mathrm{tr}(X^\top Y) = \sum_{i,j} x_{i,j} y_{i,j}$$

denotes the inner product of two matrices X and Y. Let $f_{\mathbf{s}}(X) = \{\langle \mathbf{s}\mathbf{s}^\top, X\rangle$ and $\Delta = \{\mathbf{s} : \|\mathbf{s}\| = 1\}$. Clearly, the function $f_{\mathbf{s}}(\mathbf{x})$ is continuously differentiable and $\nabla f_{\mathbf{s}}(\mathbf{x}) = \mathbf{s}\mathbf{s}^\top$. Then,

$$\partial f(X) = \mathbf{conv}\,\{\mathbf{s}\mathbf{s}^\top : \mathbf{s} \in \Delta, f_{\mathbf{s}}(X) = \langle \mathbf{s}\mathbf{s}^\top, X\rangle = f(X)\}.$$

Next, let us find out which $\mathbf{s}$ from Δ makes $f_{\mathbf{s}}(X) = f(X)$ holds. Assume that $\lambda_{\max} = \lambda_1 =$

$\cdots = \lambda_r$, where $1 \leq r \leq n$. We can see that

$$\mathbf{u}_i \in \underset{\|\mathbf{s}\|=1}{\mathbf{argmax}} \langle \mathbf{s}\mathbf{s}^\top, X \rangle,\ i = 1, \ldots, r.$$

Let $U^r = (\mathbf{u}_1, \ldots, \mathbf{u}_r)$. Then,

$$\Delta^* := \underset{\mathbf{s}\in\Delta}{\mathbf{argmax}} \langle \mathbf{s}\mathbf{s}^\top, X \rangle = \{\mathbf{v} : \mathbf{v} \in \mathbf{span}\ U^r, \|\mathbf{v}\| = 1\} = \{\mathbf{v} : \mathbf{v} = U^r\mathbf{q}, \mathbf{q} \in \mathbb{R}^r, \|\mathbf{q}\| = 1\}.$$

By Lemma 4, we have

$$\begin{aligned}
\partial f(X) =& \mathbf{conv} \left\{\mathbf{v}\mathbf{v}^\top : \mathbf{v} \in \Delta^*\right\} \\
=& \mathbf{conv} \{U^r\mathbf{q}\mathbf{q}^\top (U^r)^\top : \mathbf{q} \in \mathbb{R}^r, \|\mathbf{q}\| = 1\} \\
=& \{U^r G (U^r)^\top : G \succeq 0, \mathrm{tr}(G) = 1\}.
\end{aligned}$$

■

References

[1] Y. Nesterov. *Introductory Lectures on Convex Optimization*. Kluwer Academic Publishers, 2004.

[2] A. Ruszczyński. *Nonlinear Optimization*. Princeton University Press, 2006.