Chapter 3

Subgradients

3.1 Definitions and First Examples

Definition 3.1 (subgradient). *Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper function and let $\mathbf{x} \in \mathrm{dom}(f)$. A vector $\mathbf{g} \in \mathbb{E}^*$ is called* **a subgradient** *of f at $\mathbf{x}$ if*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle \text{ for all } \mathbf{y} \in \mathbb{E}. \tag{3.1}$$

Recall (see Section 1.11) that we use in this book the convention that the elements of $\mathbb{E}^*$ are exactly the elements of $\mathbb{E}$, whereas the asterisk just marks the fact that the endowed norm on $\mathbb{E}^*$ is the dual norm $\|\cdot\|_*$ rather than the endowed norm $\|\cdot\|$ on $\mathbb{E}$.

The inequality (3.1) is also called *the subgradient inequality*. It actually says that each subgradient is associated with an underestimate affine function, which is tangent to the surface of the function at $\mathbf{x}$. Since the subgradient inequality (3.1) is trivial for $\mathbf{y} \notin \mathrm{dom}(f)$, it is frequently restricted to points in $\mathrm{dom}(f)$ and is thus written as

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle \text{ for all } \mathbf{y} \in \mathrm{dom}(f).$$

Given a point $\mathbf{x} \in \mathrm{dom}(f)$, there might be more than one subgradient of f at $\mathbf{x}$, and the set of all subgradients is called the *subdifferential*.

Definition 3.2 (subdifferential). *The set of all subgradients of f at $\mathbf{x}$ is called* **the subdifferential** *of f at $\mathbf{x}$ and is denoted by $\partial f(\mathbf{x})$:*

$$\partial f(\mathbf{x}) \equiv \{\mathbf{g} \in \mathbb{E}^* : f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle \text{ for all } \mathbf{y} \in \mathbb{E}\}.$$

When $\mathbf{x} \notin \mathrm{dom}(f)$, we define $\partial f(\mathbf{x}) = \emptyset$. Actually, for proper functions, this is a direct consequence of the definition of the subdifferential set since the subgradient inequality (3.1) does not hold for $\mathbf{x} \notin \mathrm{dom} f$ and $\mathbf{y} \in \mathrm{dom} f$.

Example 3.3 (subdifferential of norms at 0). Let $f : \mathbb{E} \to \mathbb{R}$ be given by $f(\mathbf{x}) = \|\mathbf{x}\|$, where $\|\cdot\|$ is the endowed norm on $\mathbb{E}$. We will show that the subdifferential of f at $\mathbf{x} = \mathbf{0}$ is the dual norm unit ball:

35

$$\partial f(\mathbf{0}) = B_{\|\cdot\|_*}[\mathbf{0}, 1] = \{\mathbf{g} \in \mathbb{E}^* : \|\mathbf{g}\|_* \leq 1\}. \tag{3.2}$$

To show (3.2), note that $\mathbf{g} \in \partial f(\mathbf{0})$ if and only if

$$f(\mathbf{y}) \geq f(\mathbf{0}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{0} \rangle \text{ for all } \mathbf{y} \in \mathbb{E},$$

which is the same as

$$\|\mathbf{y}\| \geq \langle \mathbf{g}, \mathbf{y} \rangle \text{ for all } \mathbf{y} \in \mathbb{E}. \tag{3.3}$$

We will prove that the latter holds true if and only if $\|\mathbf{g}\|_* \leq 1$. Indeed, if $\|\mathbf{g}\|_* \leq 1$, then by the generalized Cauchy–Schwarz inequality (Lemma 1.4),

$$\langle \mathbf{g}, \mathbf{y} \rangle \leq \|\mathbf{g}\|_* \|\mathbf{y}\| \leq \|\mathbf{y}\| \text{ for any } \mathbf{y} \in \mathbb{E},$$

implying (3.3). In the reverse direction, assume that (3.3) holds. Taking the maximum of both sides of (3.3) over all $\mathbf{y}$ satisfying $\|\mathbf{y}\| \leq 1$, we get

$$\|\mathbf{g}\|_* = \max_{\mathbf{y}:\|\mathbf{y}\|\leq 1} \langle \mathbf{g}, \mathbf{y} \rangle \leq \max_{\mathbf{y}:\|\mathbf{y}\|\leq 1} \|\mathbf{y}\| = 1.$$

We have thus established the equivalence between (3.3) and the inequality $\|\mathbf{g}\|_* \leq 1$, which is the same as the result (3.2). ∎

Example 3.4 (subdifferential of the l_1-norm at 0). Let $f : \mathbb{R}^n \to \mathbb{R}$ be given by $f(\mathbf{x}) = \|\mathbf{x}\|_1$. Then, since this is a special case of Example 3.3 with $\|\cdot\| = \|\cdot\|_1$, and since the l_∞-norm is the dual of the l_1-norm, it follows that

$$\partial f(0) = B_{\|\cdot\|_\infty}[\mathbf{0}, 1] = [-1, 1]^n.$$

In particular, when $n = 1$, then $f(x) = |x|$, and we have

$$\partial f(0) = [-1, 1].$$

The linear underestimators that correspond to -0.8, -0.3, and $0.7 \in \partial f(0)$, meaning $-0.8x$, $-0.3x$, and $0.7x$, are described in Figure 3.1. ∎

For the next example, we need the definition of the *normal cone*. Given a set $S \subseteq \mathbb{E}$ and a point $\mathbf{x} \in S$, the normal cone of S at $\mathbf{x}$ is defined as

$$N_S(\mathbf{x}) = \{\mathbf{y} \in \mathbb{E}^* : \langle \mathbf{y}, \mathbf{z} - \mathbf{x} \rangle \leq 0 \text{ for any } \mathbf{z} \in S\}.$$

The normal cone, in addition to being a cone, is closed and convex as an intersection of half-spaces. When $\mathbf{x} \notin S$, we define $N_S(\mathbf{x}) = \emptyset$.
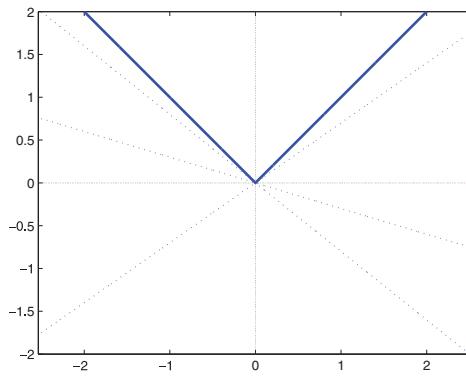
Example 3.5 (subdifferential of indicator functions). Suppose that $S \subseteq \mathbb{E}$ is nonempty and consider the indicator function δ_S. Then for any $\mathbf{x} \in S$, we have that $\mathbf{y} \in \partial \delta_S(\mathbf{x})$ if and only if

$$\delta_S(\mathbf{z}) \geq \delta_S(\mathbf{x}) + \langle \mathbf{y}, \mathbf{z} - \mathbf{x} \rangle \text{ for all } \mathbf{z} \in S,$$

which is the same as

$$\langle \mathbf{y}, \mathbf{z} - \mathbf{x} \rangle \leq 0 \text{ for all } \mathbf{z} \in S.$$

Therefore, we have that

Figure 3.1. *The linear underestimators of $|x|$ corresponding to -0.8, -0.3, $0.7 \in \partial f(0)$; see Example* 3.4.

$$\partial \delta_S(\mathbf{x}) = N_S(\mathbf{x}) \text{ for all } \mathbf{x} \in S. \tag{3.4}$$

For $\mathbf{x} \notin S$, $\partial \delta_S(\mathbf{x}) = N_S(\mathbf{x}) = \emptyset$ by convention, so we obtain that (3.4) holds also for $\mathbf{x} \notin S$. ∎

Example 3.6 (subdifferential of the indicator function of the unit ball). As a special case of Example 3.5, let

$$S = B[\mathbf{0}, 1] = \{\mathbf{x} \in \mathbb{E} : \|\mathbf{x}\| \leq 1\}.$$

Then $\partial \delta_S(\mathbf{x}) = N_S(\mathbf{x})$, where $N_S(\mathbf{x})$ is given by

$$N_S(\mathbf{x}) = \{\mathbf{y} \in \mathbb{E}^* : \langle \mathbf{y}, \mathbf{z} - \mathbf{x} \rangle \leq 0 \text{ for all } \mathbf{z} \in S\}.$$

We will find a more explicit representation for $N_S(\mathbf{x})$. If $\mathbf{x} \notin S$, then $N_S(\mathbf{x}) = \emptyset$. Suppose that $\|\mathbf{x}\| \leq 1$. A vector $\mathbf{y} \in \mathbb{E}^*$ satisfies $\mathbf{y} \in N_S(\mathbf{x})$ if and only if

$$\langle \mathbf{y}, \mathbf{z} - \mathbf{x} \rangle \leq 0 \text{ for any } \mathbf{z} \text{ satisfying } \|\mathbf{z}\| \leq 1,$$

which is the same as the inequality,

$$\max_{\mathbf{z}:\|\mathbf{z}\| \leq 1} \langle \mathbf{y}, \mathbf{z} \rangle \leq \langle \mathbf{y}, \mathbf{x} \rangle.$$

Using the definition of the dual norm, we obtain that the latter can be rewritten as

$$\|\mathbf{y}\|_* \leq \langle \mathbf{y}, \mathbf{x} \rangle.$$

Therefore,

$$\partial \delta_{B[\mathbf{0},1]}(\mathbf{x}) = N_{B[\mathbf{0},1]}(\mathbf{x}) = \begin{cases} \left\{ \mathbf{y} \in \mathbb{E}^* : \|\mathbf{y}\|_* \leq \langle \mathbf{y}, \mathbf{x} \rangle \right\}, & \|\mathbf{x}\| \leq 1, \\ \\ \emptyset, & \|\mathbf{x}\| > 1. \end{cases}$$

∎

Example 3.7 (subgradient of the dual function). Consider the minimization problem

$$\min\{f(\mathbf{x}) : \mathbf{g}(\mathbf{x}) \leq \mathbf{0}, \mathbf{x} \in X\}, \tag{3.5}$$

where $\emptyset \neq X \subseteq \mathbb{E}$, $f : \mathbb{E} \to \mathbb{R}$ and $\mathbf{g} : \mathbb{E} \to \mathbb{R}^m$ is a vector-valued function. The Lagrangian dual objective function is given by

$$q(\boldsymbol{\lambda}) = \min_{\mathbf{x} \in X} \left\{ L(\mathbf{x}; \boldsymbol{\lambda}) \equiv f(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x}) \right\}.$$

The dual problem consists of maximizing q on its effective domain, which is given by

$$\mathrm{dom}(-q) = \{\boldsymbol{\lambda} \in \mathbb{R}_+^m : q(\boldsymbol{\lambda}) > -\infty\}.$$

No matter whether the primal problem (3.5) is convex or not, the dual problem

$$\max_{\boldsymbol{\lambda} \in \mathbb{R}^m} \{q(\boldsymbol{\lambda}) : \boldsymbol{\lambda} \in \mathrm{dom}(-q)\}$$

is always convex, meaning that q is a concave function and $\mathrm{dom}(-q)$ is a convex set. Let $\boldsymbol{\lambda}_0 \in \mathrm{dom}(-q)$ and assume that the minimum in the minimization problem defining $q(\boldsymbol{\lambda}_0)$,

$$q(\boldsymbol{\lambda}_0) = \min_{\mathbf{x} \in X} \left\{ f(\mathbf{x}) + \boldsymbol{\lambda}_0^T \mathbf{g}(\mathbf{x}) \right\},$$

is attained at $\mathbf{x}_0 \in X$, that is,

$$L(\mathbf{x}_0; \boldsymbol{\lambda}_0) = f(\mathbf{x}_0) + \boldsymbol{\lambda}_0^T \mathbf{g}(\mathbf{x}_0) = q(\boldsymbol{\lambda}_0).$$

We seek to find a subgradient of the convex function $-q$ at $\boldsymbol{\lambda}_0$. For that, note that for any $\boldsymbol{\lambda} \in \mathrm{dom}(-q)$,

$$\begin{aligned} q(\boldsymbol{\lambda}) &= \min_{\mathbf{x} \in X} \left\{ f(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x}) \right\} \\ &\leq f(\mathbf{x}_0) + \boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x}_0) \\ &= f(\mathbf{x}_0) + \boldsymbol{\lambda}_0^T \mathbf{g}(\mathbf{x}_0) + (\boldsymbol{\lambda} - \boldsymbol{\lambda}_0)^T \mathbf{g}(\mathbf{x}_0) \\ &= q(\boldsymbol{\lambda}_0) + \mathbf{g}(\mathbf{x}_0)^T (\boldsymbol{\lambda} - \boldsymbol{\lambda}_0). \end{aligned}$$

Thus,

$$-q(\boldsymbol{\lambda}) \geq -q(\boldsymbol{\lambda}_0) + (-\mathbf{g}(\mathbf{x}_0))^T (\boldsymbol{\lambda} - \boldsymbol{\lambda}_0) \text{ for any } \boldsymbol{\lambda} \in \mathrm{dom}(-q),$$

concluding that

$$-\mathbf{g}(\mathbf{x}_0) \in \partial(-q)(\boldsymbol{\lambda}_0).$$

∎

Example 3.8 (subgradient of the maximum eigenvalue function). Consider the function $f : \mathbb{S}^n \to \mathbb{R}$ given by $f(\mathbf{X}) = \lambda_{\max}(\mathbf{X})$ (recall that $\mathbb{S}^n$ is the set of all $n \times n$ symmetric matrices). Let $\mathbf{X} \in \mathbb{S}^n$ and let $\mathbf{v}$ be a normalized eigenvector of $\mathbf{X}$ ($\|\mathbf{v}\|_2 = 1$) associated with the maximum eigenvalue of $\mathbf{X}$. We will establish the relation

$$\mathbf{v}\mathbf{v}^T \in \partial f(\mathbf{X}). \tag{3.6}$$

To show this, note that for any $\mathbf{Y} \in \mathbb{S}^n$,

$$\begin{aligned}
\lambda_{\max}(\mathbf{Y}) &= \max_{\mathbf{u}}\{\mathbf{u}^T\mathbf{Y}\mathbf{u} : \|\mathbf{u}\|_2 = 1\} \\
&\geq \mathbf{v}^T\mathbf{Y}\mathbf{v} \\
&= \mathbf{v}^T\mathbf{X}\mathbf{v} + \mathbf{v}^T(\mathbf{Y} - \mathbf{X})\mathbf{v} \\
&= \lambda_{\max}(\mathbf{X})\|\mathbf{v}\|_2^2 + \mathrm{Tr}(\mathbf{v}^T(\mathbf{Y} - \mathbf{X})\mathbf{v}) \\
&= \lambda_{\max}(\mathbf{X}) + \mathrm{Tr}(\mathbf{v}\mathbf{v}^T(\mathbf{Y} - \mathbf{X})) \\
&= \lambda_{\max}(\mathbf{X}) + \langle \mathbf{v}\mathbf{v}^T, \mathbf{Y} - \mathbf{X}\rangle,
\end{aligned}$$

establishing (3.6). ∎

There is an intrinsic difference between the results in Examples 3.7 and 3.8 and the results in Examples 3.3, 3.4, 3.5, and 3.6. Only one subgradient is computed in Examples 3.7 and 3.8; such results are referred to as *weak results*. On the other hand, in Examples 3.3, 3.4, 3.5, and 3.6 the entire subdifferential set is characterized—such results are called *strong results*.

3.2 Properties of the Subdifferential Set

Note that the subdifferential sets computed in the previous section are all closed and convex. This is not a coincidence. Subdifferential sets are *always* closed and convex.

Theorem 3.9 (closedness and convexity of the subdifferential set). *Let* $f : \mathbb{E} \to (-\infty, \infty]$ *be a proper function. Then the set* $\partial f(\mathbf{x})$ *is closed and convex for any* $\mathbf{x} \in \mathbb{E}$.

Proof. For any $\mathbf{x} \in \mathbb{E}$, the subdifferential set can be represented as

$$\partial f(\mathbf{x}) = \bigcap_{\mathbf{y} \in \mathbb{E}} H_{\mathbf{y}},$$

where $H_{\mathbf{y}} = \{\mathbf{g} \in \mathbb{E}^* : f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x}\rangle\}$. Since the sets $H_{\mathbf{y}}$ are half-spaces and, in particular, closed and convex, it follows that $\partial f(\mathbf{x})$ is closed and convex. □

The subdifferential set $\partial f(\mathbf{x})$ may be empty. When it is nonempty at a given $\mathbf{x} \in \mathbb{E}$, the function f is called *subdifferentiable* at $\mathbf{x}$.

Definition 3.10 (subdifferentiability). *A proper function* $f : \mathbb{E} \to (-\infty, \infty]$ *is called* **subdifferentiable** *at* $\mathbf{x} \in \mathrm{dom}(f)$ *if* $\partial f(\mathbf{x}) \neq \emptyset$.

The collection of points of subdifferentiability is denoted by $\mathrm{dom}(\partial f)$:

$$\mathrm{dom}(\partial f) = \{\mathbf{x} \in \mathbb{E} : \partial f(\mathbf{x}) \neq \emptyset\}.$$

We will now show that if a function is subdifferentiable at any point in its domain, which is assumed to be convex, then it is necessarily convex.

Lemma 3.11 (nonemptiness of subdifferential sets $\Rightarrow$ convexity). *Let f : $\mathbb{E} \to (-\infty, \infty]$ be a proper function and assume that $\mathrm{dom}(f)$ is convex. Suppose that for any $\mathbf{x} \in \mathrm{dom}(f)$, the set $\partial f(\mathbf{x})$ is nonempty. Then f is convex.*

Proof. Let $\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$ and $\alpha \in [0,1]$. Define $\mathbf{z}_\alpha = (1-\alpha)\mathbf{x} + \alpha\mathbf{y}$. By the convexity of $\mathrm{dom}(f)$, $\mathbf{z}_\alpha \in \mathrm{dom}(f)$, and hence there exists $\mathbf{g} \in \partial f(\mathbf{z}_\alpha)$, which in particular implies the following two inequalities:

$$f(\mathbf{y}) \geq f(\mathbf{z}_\alpha) + \langle \mathbf{g}, \mathbf{y} - \mathbf{z}_\alpha \rangle = f(\mathbf{z}_\alpha) + (1-\alpha)\langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle,$$
$$f(\mathbf{x}) \geq f(\mathbf{z}_\alpha) + \langle \mathbf{g}, \mathbf{x} - \mathbf{z}_\alpha \rangle = f(\mathbf{z}_\alpha) - \alpha\langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle.$$

Multiplying the first inequality by α, the second by $1-\alpha$, and summing them yields

$$f((1-\alpha)\mathbf{x} + \alpha\mathbf{y}) = f(\mathbf{z}_\alpha) \leq (1-\alpha)f(\mathbf{x}) + \alpha f(\mathbf{y}).$$

Since the latter holds for any $\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$ with $\mathrm{dom}(f)$ being convex, it follows that the function f is convex. $\quad\square$

We have thus shown that if a function is subdifferentiable at any point in its (assumed to be) convex domain, then it is convex. However, this does not mean that the reverse direction is correct. The next example describes a convex function, which is not subdifferentiable at one of the points in its domain.

Example 3.12. Consider the convex function $f : \mathbb{R} \to (-\infty, \infty]$ defined by

$$f(x) = \begin{cases} -\sqrt{x}, & x \geq 0, \\ \infty, & \text{else.} \end{cases}$$

The function is plotted in Figure 3.2. It is not subdifferentiable at $x = 0$. To show this, suppose by contradiction that there exists $g \in \mathbb{R}$ such that $g \in \partial f(0)$. Then

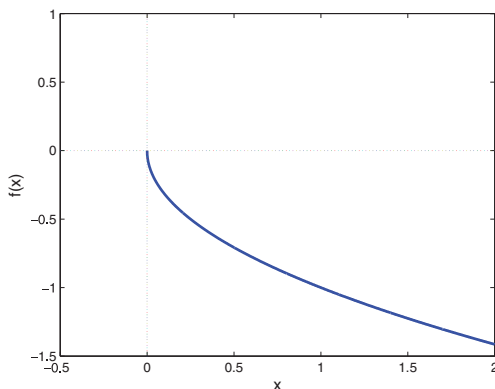$$f(y) \geq f(0) + g(y - 0) \text{ for any } y \geq 0,$$

which is the same as

$$-\sqrt{y} \geq gy \text{ for any } y \geq 0. \tag{3.7}$$

The above is impossible since substituting $y = 1$, we obtain that $g \leq -1$ (and in particular $g < 0$), while substituting $y = \frac{1}{2g^2}$ in (3.7) yields the inequality

$$-\sqrt{1/(2g^2)} \geq 1/(2g),$$

which is equivalent to the impossible inequality (utilizing the fact that $g < 0$)

$$\frac{1}{2g^2} \leq \frac{1}{4g^2}. \quad\blacksquare$$

Figure 3.2. *The function $f(x) = -\sqrt{x}$ with $\mathrm{dom}(f) = \mathbb{R}_+$. The function is not subdifferentiable at $x = 0$.*

Although, as demonstrated in Example 3.12, convex functions are not necessarily subdifferentiable at any point in their domain, they must be subdifferentiable at any point in the interior of their domain. This is stated in Theorem 3.14 below, which also shows the boundedness of the subdifferential set in this setting. The proof of the theorem strongly relies on the supporting hyperplane theorem stated explicitly below.

Theorem 3.13 (supporting hyperplane theorem [29, Proposition 2.4.1]). *Let $\emptyset \neq C \subseteq \mathbb{E}$ be a convex set, and let $\mathbf{y} \notin \mathrm{int}(C)$. Then there exists $\mathbf{0} \neq \mathbf{p} \in \mathbb{E}^*$ such that*
$$\langle \mathbf{p}, \mathbf{x} \rangle \leq \langle \mathbf{p}, \mathbf{y} \rangle \text{ for any } \mathbf{x} \in C.$$

Theorem 3.14 (nonemptiness and boundedness of the subdifferential set at interior points of the domain). *Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper convex function, and assume that $\tilde{\mathbf{x}} \in \mathrm{int}(\mathrm{dom}(f))$. Then $\partial f(\tilde{\mathbf{x}})$ is nonempty and bounded.*

Proof. Recall that the inner product in the product space $\mathbb{E} \times \mathbb{R}$ is defined as (see Section 1.9)
$$\langle (\mathbf{y}_1, \beta_1), (\mathbf{y}_2, \beta_2) \rangle \equiv \langle \mathbf{y}_1, \mathbf{y}_2 \rangle + \beta_1 \beta_2, \quad (\mathbf{y}_1, \beta_1), (\mathbf{y}_2, \beta_2) \in \mathbb{E} \times \mathbb{R}.$$

Since $(\tilde{\mathbf{x}}, f(\tilde{\mathbf{x}}))$ is on the boundary of $\mathrm{epi}(f) \subseteq \mathbb{E} \times \mathbb{R}$, it follows by the supporting hyperplane theorem (Theorem 3.13) that there exists a separating hyperplane between $(\tilde{\mathbf{x}}, f(\tilde{\mathbf{x}}))$ and $\mathrm{epi}(f)$, meaning that there exists a nonzero vector $(\mathbf{p}, -\alpha) \in \mathbb{E}^* \times \mathbb{R}$ for which
$$\langle \mathbf{p}, \tilde{\mathbf{x}} \rangle - \alpha f(\tilde{\mathbf{x}}) \geq \langle \mathbf{p}, \mathbf{x} \rangle - \alpha t \text{ for any } (\mathbf{x}, t) \in \mathrm{epi}(f). \tag{3.8}$$

Note that $\alpha \geq 0$ since $(\tilde{\mathbf{x}}, f(\tilde{\mathbf{x}}) + 1) \in \mathrm{epi}(f)$, and hence plugging $\mathbf{x} = \tilde{\mathbf{x}}$ and $t = f(\tilde{\mathbf{x}}) + 1$ into (3.8) yields
$$\langle \mathbf{p}, \tilde{\mathbf{x}} \rangle - \alpha f(\tilde{\mathbf{x}}) \geq \langle \mathbf{p}, \tilde{\mathbf{x}} \rangle - \alpha(f(\tilde{\mathbf{x}}) + 1),$$

implying that $\alpha \geq 0$. Since $\tilde{\mathbf{x}} \in \text{int}(\text{dom}(f))$, it follows by the local Lipschitz continuity property of convex functions (Theorem 2.21) that there exist $\varepsilon > 0$ and $L > 0$ such that $B_{\|\cdot\|}[\tilde{\mathbf{x}}, \varepsilon] \subseteq \text{dom}(f)$ and

$$|f(\mathbf{x}) - f(\tilde{\mathbf{x}})| \leq L\|\mathbf{x} - \tilde{\mathbf{x}}\| \text{ for any } \mathbf{x} \in B_{\|\cdot\|}[\tilde{\mathbf{x}}, \varepsilon]. \tag{3.9}$$

Since $B_{\|\cdot\|}[\tilde{\mathbf{x}}, \varepsilon] \subseteq \text{dom}(f)$, it follows that $(\mathbf{x}, f(\mathbf{x})) \in \text{epi}(f)$ for any $\mathbf{x} \in B_{\|\cdot\|}[\tilde{\mathbf{x}}, \varepsilon]$. Therefore, plugging $t = f(\mathbf{x})$ into (3.8), yields that

$$\langle \mathbf{p}, \mathbf{x} - \tilde{\mathbf{x}} \rangle \leq \alpha(f(\mathbf{x}) - f(\tilde{\mathbf{x}})) \text{ for any } \mathbf{x} \in B_{\|\cdot\|}[\tilde{\mathbf{x}}, \varepsilon]. \tag{3.10}$$

Combining (3.9) and (3.10), we obtain that for any $\mathbf{x} \in B_{\|\cdot\|}[\tilde{\mathbf{x}}, \varepsilon]$,

$$\langle \mathbf{p}, \mathbf{x} - \tilde{\mathbf{x}} \rangle \leq \alpha(f(\mathbf{x}) - f(\tilde{\mathbf{x}})) \leq \alpha L\|\mathbf{x} - \tilde{\mathbf{x}}\|. \tag{3.11}$$

Take $\mathbf{p}^\dagger \in \mathbb{E}$ satisfying $\langle \mathbf{p}, \mathbf{p}^\dagger \rangle = \|\mathbf{p}\|_*$ and $\|\mathbf{p}^\dagger\| = 1$. Since $\tilde{\mathbf{x}} + \varepsilon\mathbf{p}^\dagger \in B_{\|\cdot\|}[\tilde{\mathbf{x}}, \varepsilon]$, we can plug $\mathbf{x} = \tilde{\mathbf{x}} + \varepsilon\mathbf{p}^\dagger$ into (3.11) and obtain that

$$\varepsilon\|\mathbf{p}\|_* = \varepsilon\langle \mathbf{p}, \mathbf{p}^\dagger \rangle \leq \alpha L\varepsilon\|\mathbf{p}^\dagger\| = \alpha L\varepsilon.$$

Therefore, $\alpha > 0$, since otherwise we would have $\alpha = 0$ and $\mathbf{p} = \mathbf{0}$, which is impossible by the fact that the vector $(\mathbf{p}, \alpha)$ is not the zeros vector. Taking $t = f(\mathbf{x})$ in (3.8) and dividing by α yields

$$f(\mathbf{x}) \geq f(\tilde{\mathbf{x}}) + \langle \mathbf{g}, \mathbf{x} - \tilde{\mathbf{x}} \rangle \text{ for all } \mathbf{x} \in \text{dom}(f), \tag{3.12}$$

where $\mathbf{g} = \mathbf{p}/\alpha$. Thus, $\mathbf{g} \in \partial f(\tilde{\mathbf{x}})$, establishing the nonemptiness of $\partial f(\tilde{\mathbf{x}})$. To show the boundedness of $\partial f(\tilde{\mathbf{x}})$, let $\mathbf{g} \in \partial f(\tilde{\mathbf{x}})$, meaning that (3.12) holds. Take $\mathbf{g}^\dagger \in \mathbb{E}$ for which $\|\mathbf{g}\|_* = \langle \mathbf{g}, \mathbf{g}^\dagger \rangle$ and $\|\mathbf{g}^\dagger\| = 1$. Then plugging $\mathbf{x} = \tilde{\mathbf{x}} + \varepsilon\mathbf{g}^\dagger$ in (3.12) yields

$$\varepsilon\|\mathbf{g}\|_* = \varepsilon\langle \mathbf{g}, \mathbf{g}^\dagger \rangle = \langle \mathbf{g}, \mathbf{x} - \tilde{\mathbf{x}} \rangle \leq f(\mathbf{x}) - f(\tilde{\mathbf{x}}) \overset{(3.9)}{\leq} L\|\mathbf{x} - \tilde{\mathbf{x}}\| = L\varepsilon,$$

showing that $\partial f(\tilde{\mathbf{x}}) \subseteq B_{\|\cdot\|_*}[\mathbf{0}, L]$, and hence establishing the boundedness of $\partial f(\tilde{\mathbf{x}})$. $\quad\square$

The result of Theorem 3.14 can be stated as the following inclusion relation:

$$\text{int}(\text{dom}(f)) \subseteq \text{dom}(\partial f).$$

A direct consequence of Theorem 3.14 is that real-valued convex functions (namely, convex functions f with $\text{dom}(f) = \mathbb{E}$) are subdifferentiable at any point.

Corollary 3.15 (subdifferentiability of real-valued convex functions). *Let $f : \mathbb{E} \to \mathbb{R}$ be a convex function. Then f is subdifferentiable over $\mathbb{E}$.*

We can extend the boundedness result of Theorem 3.14 and show that subgradients of points in a given compact set contained in the interior of the domain are always bounded.

Theorem 3.16 (boundedness of subgradients over compact sets). *Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper convex function, and assume that $X \subseteq \text{int}(\text{dom}(f))$ is nonempty and compact. Then $Y = \bigcup_{\mathbf{x} \in X} \partial f(\mathbf{x})$ is nonempty and bounded.*

Proof. The set Y is nonempty, since by Theorem 3.14 $\partial f(\mathbf{x}) \neq \emptyset$ for any $\mathbf{x} \in X$. To prove the boundedness, assume by contradiction that there exists a sequence $\{\mathbf{x}_k\}_{k\geq 1} \subseteq X$ and $\mathbf{g}_k \in \partial f(\mathbf{x}_k)$ such that $\|\mathbf{g}_k\|_* \to \infty$ as $k \to \infty$. For any k, let $\mathbf{g}_k^\dagger$ be a vector satisfying $\langle \mathbf{g}_k, \mathbf{g}_k^\dagger \rangle = \|\mathbf{g}_k\|_*$ and $\|\mathbf{g}_k^\dagger\| = 1$. Since X is compact, $(\text{int}(\text{dom}(f)))^c$ (the complement of $\text{int}(\text{dom}(f))$) is closed, and $X \cap (\text{int}(\text{dom}(f)))^c = \emptyset$, it follows that the distance between the two sets is nonempty, meaning in particular that there exists an $\varepsilon > 0$ for which[10]

$$\|\mathbf{x} - \mathbf{y}\| \geq \varepsilon \text{ for any } \mathbf{x} \in X, \mathbf{y} \notin \text{int}(\text{dom}(f)). \tag{3.13}$$

The relation $\mathbf{g}_k \in \partial f(\mathbf{x}_k)$ implies in particular that

$$f\left(\mathbf{x}_k + \frac{\varepsilon}{2}\mathbf{g}_k^\dagger\right) - f(\mathbf{x}_k) \geq \frac{\varepsilon}{2}\langle \mathbf{g}_k, \mathbf{g}_k^\dagger \rangle = \frac{\varepsilon}{2}\|\mathbf{g}_k\|_*, \tag{3.14}$$

where we used the fact that by (3.13), $\mathbf{x}_k + \frac{\varepsilon}{2}\mathbf{g}_k^\dagger \in \text{int}(\text{dom}(f))$. We will show that the left-hand side of (3.14) is bounded. Suppose by contradiction that it is not bounded. Then there exist subsequences $\{\mathbf{x}_k\}_{k\in T}$, $\{\mathbf{g}_k^\dagger\}_{k\in T}$ (T being the set of indices of the subsequences) for which

$$f\left(\mathbf{x}_k + \frac{\varepsilon}{2}\mathbf{g}_k^\dagger\right) - f(\mathbf{x}_k) \to \infty \text{ as } k \xrightarrow{T} \infty. \tag{3.15}$$

Since both $\{\mathbf{x}_k\}_{k\in T}$ and $\{\mathbf{g}_k^\dagger\}_{k\in T}$ are bounded, it follows that there exist convergent subsequences $\{\mathbf{x}_k\}_{k\in S}$, $\{\mathbf{g}_k^\dagger\}_{k\in S}$ ($S \subseteq T$) whose limits will be denoted by $\bar{\mathbf{x}}$ and $\bar{\mathbf{g}}$. Consequently, $\mathbf{x}_k + \frac{\varepsilon}{2}\mathbf{g}_k^\dagger \to \bar{\mathbf{x}} + \frac{\varepsilon}{2}\bar{\mathbf{g}}$ as $k \xrightarrow{S} \infty$. Since $\mathbf{x}_k, \mathbf{x}_k + \frac{\varepsilon}{2}\mathbf{g}_k^\dagger, \bar{\mathbf{x}} + \frac{\varepsilon}{2}\bar{\mathbf{g}}$ are all[11] in $\text{int}(\text{dom}(f))$, it follows by the continuity of f over $\text{int}(\text{dom}(f))$ (Theorem 2.21) that

$$f\left(\mathbf{x}_k + \frac{\varepsilon}{2}\mathbf{g}_k^\dagger\right) - f(\mathbf{x}_k) \to f\left(\bar{\mathbf{x}} + \frac{\varepsilon}{2}\bar{\mathbf{g}}^\dagger\right) - f(\bar{\mathbf{x}}) \text{ as } k \xrightarrow{S} \infty,$$

which is a contradiction of (3.15). We can thus conclude that the left-hand side of (3.14) is bounded and hence that the right-hand side of (3.14) is also bounded, in contradiction to our assumption that $\|\mathbf{g}_k\|_*$ goes to ∞ as $k \to \infty$. $\quad\square$

Subdifferentiability can be guaranteed for points that are not necessarily in the interior of the domain but are in the interior of the domain w.r.t. its affine hull. This is the notion of *relative interior* that we now recall:

$$\text{ri}(S) = \{\mathbf{x} \in \text{aff}(S) : B[\mathbf{x}, \varepsilon] \cap \text{aff}(S) \subseteq S \text{ for some } \varepsilon > 0\}.$$

One key property of the relative interior is that it is nonempty for convex sets.

Theorem 3.17 (nonemptiness of the relative interior [108, Theorem 6.2]).
Let $C \subseteq \mathbb{E}$ be a nonempty convex set. Then $\text{ri}(C)$ is nonempty.

---

[10]The proof of (3.13) is simple. Suppose by contradiction that there exist sequences $\{\mathbf{x}_k\}_{k\geq 1} \subseteq X$ and $\{\mathbf{y}_k\}_{k\geq 1} \subseteq (\text{int}(\text{dom}(f)))^c$ satisfying $\|\mathbf{x}_k - \mathbf{y}_k\| \to 0$ as $k \to \infty$. Since $\{\mathbf{x}_k\}_{k\geq 1}$ is bounded, there exists $M > 0$ for which $\|\mathbf{x}_k\| \leq M$ for all k. Therefore, $\|\mathbf{y}_k\| \leq \|\mathbf{x}_k - \mathbf{y}_k\| + \|\mathbf{x}_k\| \leq \|\mathbf{x}_k - \mathbf{y}_k\| + M$, and we can conclude by the boundedness of $\{\|\mathbf{x}_k - \mathbf{y}_k\|\}_{k\geq 1}$ that $\{\mathbf{y}_k\}_{k\geq 1}$ is bounded. By the Bolzano–Weierstrass theorem, there exist convergent subsequences $\mathbf{x}_{k_j} \to \bar{\mathbf{x}}, \mathbf{y}_{k_j} \to \bar{\mathbf{y}}$, and by the closedness of X and $(\text{int}(\text{dom}(f)))^c$, we have that $\bar{\mathbf{x}} \in X, \bar{\mathbf{y}} \in (\text{int}(\text{dom}(f)))^c$. The limit $\|\mathbf{x}_{k_j} - \mathbf{y}_{k_j}\| \to 0$ as $j \to \infty$ now brings us to the impossible equality $\bar{\mathbf{x}} = \bar{\mathbf{y}}$.

[11]The fact that $\bar{\mathbf{x}} + \frac{\varepsilon}{2}\bar{\mathbf{g}} \in \text{int}(\text{dom}(f))$ follows by (3.13) and the relations $\bar{\mathbf{x}} \in X$ and $\|\bar{\mathbf{g}}\| = 1$.

A well-known result is that a proper convex function is always subdifferentiable at relative interior points of its domain. We state this result without a proof.

Theorem 3.18 (nonemptiness of the subdifferential set at relative interior points [108, Theorem 23.4]). *Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper convex function, and let $\tilde{\mathbf{x}} \in \mathrm{ri}(\mathrm{dom}(f))$. Then $\partial f(\tilde{\mathbf{x}})$ is nonempty.*

The result stated in Theorem 3.18 can be written as the inclusion

$$\mathrm{ri}(\mathrm{dom}(f)) \subseteq \mathrm{dom}(\partial f).$$

Since the relative interior of $\mathrm{dom}(f)$ is always nonempty (Theorem 3.17), we can conclude that there always exists a point in the domain for which the subdifferential set is nonempty.

Corollary 3.19. *Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper convex function. Then there exists $\mathbf{x} \in \mathrm{dom}(f)$ for which $\partial f(\mathbf{x})$ is nonempty.*

One instance in which the subdifferential set $\partial f(\mathbf{x})$ is guaranteed to be *unbounded* is when the dimension of the domain of the function is strictly smaller than the dimension of the underlying space $\mathbb{E}$.

Theorem 3.20 (unboundedness of the subdifferential set when $\dim(\mathrm{dom}(f)) < \dim(\mathbb{E})$). *Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper convex function. Suppose that $\dim(\mathrm{dom}(f)) < \dim(\mathbb{E})$ and let $\mathbf{x} \in \mathrm{dom}(f)$. If $\partial f(\mathbf{x}) \neq \emptyset$, then $\partial f(\mathbf{x})$ is unbounded.*

Proof. Let $\boldsymbol{\eta}$ be an arbitrary vector in $\partial f(\mathbf{x})$. The set[12] $\mathbb{V} \equiv \mathrm{aff}(\mathrm{dom}(f)) - \{\mathbf{x}\}$ is a vector space. The dimension condition translates to $\dim(\mathbb{V}) < \dim(\mathbb{E})$, which in particular implies that there exists a nonzero vector $\mathbf{v} \in \mathbb{E}$ such that $\langle \mathbf{v}, \mathbf{w} \rangle = 0$ for any $\mathbf{w} \in \mathbb{V}$. Take any $\beta \in \mathbb{R}$. For any $\mathbf{y} \in \mathrm{dom}(f)$,

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \boldsymbol{\eta}, \mathbf{y} - \mathbf{x} \rangle = f(\mathbf{x}) + \langle \boldsymbol{\eta} + \beta \mathbf{v}, \mathbf{y} - \mathbf{x} \rangle,$$

where the equality is due to the fact that $\mathbf{y} - \mathbf{x} \in \mathbb{V}$. We thus obtained that $\boldsymbol{\eta} + \beta \mathbf{v} \in \partial f(\mathbf{x})$ for any β, implying the unboundedness of $\partial f(\mathbf{x})$. $\square$

3.3   Directional Derivatives

3.3.1   Definition and Basic Properties

Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper function and let $\mathbf{x} \in \mathrm{int}(\mathrm{dom}(f))$. The *directional derivative* of f at $\mathbf{x}$ in a given direction $\mathbf{d} \in \mathbb{E}$, if it exists, is defined by

$$f'(\mathbf{x}; \mathbf{d}) \equiv \lim_{\alpha \to 0^+} \frac{f(\mathbf{x} + \alpha \mathbf{d}) - f(\mathbf{x})}{\alpha}.$$

A well-known result states that convex functions have directional derivatives in all directions at points in the interior of their domains.

---

[12]Here the notation "$-$" stands for the Minkowski difference.

Theorem 3.21 ([108, Theorem 23.1][13]**).** *Let* $f : \mathbb{E} \to (-\infty, \infty]$ *be a proper convex function and let* $\mathbf{x} \in \text{int}(\text{dom}(f))$. *Then for any* $\mathbf{d} \in \mathbb{E}$, *the directional derivative* $f'(\mathbf{x}; \mathbf{d})$ *exists.*

It is important to establish some basic properties of the function $\mathbf{d} \mapsto f'(\mathbf{x}; \mathbf{d})$. The next theorem shows that it is convex and homogeneous of degree 1.

Lemma 3.22 (convexity and homogeneity of $\mathbf{d} \mapsto f'(\mathbf{x}; \mathbf{d})$**).** *Let* $f : \mathbb{E} \to (-\infty, \infty]$ *be a proper convex function and let* $\mathbf{x} \in \text{int}(\text{dom}(f))$. *Then*

(a) *the function* $\mathbf{d} \mapsto f'(\mathbf{x}; \mathbf{d})$ *is convex;*

(b) *for any* $\lambda \geq 0$ *and* $\mathbf{d} \in \mathbb{E}$, *it holds that* $f'(\mathbf{x}; \lambda \mathbf{d}) = \lambda f'(\mathbf{x}; \mathbf{d})$.

Proof. (a) To show that the function $g(\mathbf{d}) \equiv f'(\mathbf{x}; \mathbf{d})$ is convex, take $\mathbf{d}_1, \mathbf{d}_2 \in \mathbb{E}$ and $\lambda \in [0, 1]$. Then

$$
\begin{aligned}
f'(\mathbf{x}; &\lambda \mathbf{d}_1 + (1 - \lambda)\mathbf{d}_2) \\
&= \lim_{\alpha \to 0^+} \frac{f(\mathbf{x} + \alpha[\lambda \mathbf{d}_1 + (1 - \lambda)\mathbf{d}_2]) - f(\mathbf{x})}{\alpha} \\
&= \lim_{\alpha \to 0^+} \frac{f(\lambda(\mathbf{x} + \alpha \mathbf{d}_1) + (1 - \lambda)(\mathbf{x} + \alpha \mathbf{d}_2)) - f(\mathbf{x})}{\alpha} \\
&\leq \lim_{\alpha \to 0^+} \frac{\lambda f(\mathbf{x} + \alpha \mathbf{d}_1) + (1 - \lambda)f(\mathbf{x} + \alpha \mathbf{d}_2) - f(\mathbf{x})}{\alpha} \\
&= \lambda \lim_{\alpha \to 0^+} \frac{f(\mathbf{x} + \alpha \mathbf{d}_1) - f(\mathbf{x})}{\alpha} + (1 - \lambda) \lim_{\alpha \to 0^+} \frac{f(\mathbf{x} + \alpha \mathbf{d}_2) - f(\mathbf{x})}{\alpha} \\
&= \lambda f'(\mathbf{x}; \mathbf{d}_1) + (1 - \lambda)f'(\mathbf{x}; \mathbf{d}_2),
\end{aligned}
$$

where the inequality follows from Jensen's inequality for convex functions.
(b) If $\lambda = 0$, the claim is trivial. Take $\lambda > 0$. Then

$$
f'(\mathbf{x}; \lambda \mathbf{d}) = \lim_{\alpha \to 0^+} \frac{f(\mathbf{x} + \alpha \lambda \mathbf{d}) - f(\mathbf{x})}{\alpha} = \lambda \lim_{\alpha \to 0^+} \frac{f(\mathbf{x} + \alpha \lambda \mathbf{d}) - f(\mathbf{x})}{\alpha \lambda} = \lambda f'(\mathbf{x}; \mathbf{d}).
$$

$\square$

The next result highlights a connection between function values and directional derivatives under a convexity assumption.

Lemma 3.23. *Let* $f : \mathbb{E} \to (-\infty, \infty]$ *be a proper convex function, and let* $\mathbf{x} \in \text{int}(\text{dom}(f))$. *Then*

$$
f(\mathbf{y}) \geq f(\mathbf{x}) + f'(\mathbf{x}; \mathbf{y} - \mathbf{x}) \text{ for all } \mathbf{y} \in \text{dom}(f).
$$

---

[13]See also [10, Theorem 7.37].

Proof. By the definition of the directional derivative,

$$
\begin{aligned}
f'(\mathbf{x}; \mathbf{y} - \mathbf{x}) &= \lim_{\alpha \to 0^+} \frac{f(\mathbf{x} + \alpha(\mathbf{y} - \mathbf{x})) - f(\mathbf{x})}{\alpha} \\
&= \lim_{\alpha \to 0^+} \frac{f((1 - \alpha)\mathbf{x} + \alpha\mathbf{y}) - f(\mathbf{x})}{\alpha} \\
&\leq \lim_{\alpha \to 0^+} \frac{-\alpha f(\mathbf{x}) + \alpha f(\mathbf{y})}{\alpha} \\
&= f(\mathbf{y}) - f(\mathbf{x}),
\end{aligned}
$$

where the inequality follows by Jensen's inequality. $\Box$

A useful "calculus" rule for directional derivatives shows how to compute the directional derivative of maximum of a finite collection of functions without any convexity assumptions.

Theorem 3.24 (directional derivative of maximum of functions). *Suppose that $f(\mathbf{x}) = \max\{f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x})\}$, where $f_1, f_2, \ldots, f_m : \mathbb{E} \to (-\infty, \infty]$ are proper functions. Let $\mathbf{x} \in \bigcap_{i=1}^m \operatorname{int}(\operatorname{dom}(f_i))$ and $\mathbf{d} \in \mathbb{E}$. Assume that $f_i'(\mathbf{x}; \mathbf{d})$ exist for any $i \in \{1, 2, \ldots, m\}$. Then*

$$
f'(\mathbf{x}; \mathbf{d}) = \max_{i \in I(\mathbf{x})} f_i'(\mathbf{x}; \mathbf{d}),
$$

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

Proof. For any $i \in \{1, 2, \ldots, m\}$,

$$
\lim_{t \to 0^+} f_i(\mathbf{x} + t\mathbf{d}) = \lim_{t \to 0^+} \left[t \frac{f_i(\mathbf{x} + t\mathbf{d}) - f_i(\mathbf{x})}{t} + f_i(\mathbf{x}) \right] = 0 \cdot f_i'(\mathbf{x}; \mathbf{d}) + f_i(\mathbf{x}) = f_i(\mathbf{x}).
$$

$$\tag{3.16}$$

By the definition of $I(\mathbf{x})$, $f_i(\mathbf{x}) > f_j(\mathbf{x})$ for any $i \in I(\mathbf{x}), j \notin I(\mathbf{x})$. Utilizing (3.16), it follows that there exists an $\varepsilon > 0$ such that $f_i(\mathbf{x} + t\mathbf{d}) > f_j(\mathbf{x} + t\mathbf{d})$ for any $i \in I(\mathbf{x}), j \notin I(\mathbf{x})$ and $t \in (0, \varepsilon]$. Therefore, for any $t \in (0, \varepsilon]$,

$$
f(\mathbf{x} + t\mathbf{d}) = \max_{i=1,2,\ldots,m} f_i(\mathbf{x} + t\mathbf{d}) = \max_{i \in I(\mathbf{x})} f_i(\mathbf{x} + t\mathbf{d}).
$$

Consequently, for any $t \in (0, \varepsilon]$,

$$
\frac{f(\mathbf{x} + t\mathbf{d}) - f(\mathbf{x})}{t} = \frac{\max_{i \in I(\mathbf{x})} f_i(\mathbf{x} + t\mathbf{d}) - f(\mathbf{x})}{t} = \max_{i \in I(\mathbf{x})} \frac{f_i(\mathbf{x} + t\mathbf{d}) - f_i(\mathbf{x})}{t},
$$

where the last equality follows from the fact that $f(\mathbf{x}) = f_i(\mathbf{x})$ for any $i \in I(\mathbf{x})$. Finally, taking $t \to 0^+$, we obtain that

$$
\begin{aligned}
f'(\mathbf{x}; \mathbf{d}) &= \lim_{t \to 0^+} \frac{f(\mathbf{x} + t\mathbf{d}) - f(\mathbf{x})}{t} \\
&= \lim_{t \to 0^+} \max_{i \in I(\mathbf{x})} \frac{f_i(\mathbf{x} + t\mathbf{d}) - f_i(\mathbf{x})}{t} \\
&= \max_{i \in I(\mathbf{x})} \lim_{t \to 0^+} \frac{f_i(\mathbf{x} + t\mathbf{d}) - f_i(\mathbf{x})}{t} \\
&= \max_{i \in I(\mathbf{x})} f_i'(\mathbf{x}; \mathbf{d}).   \quad \Box
\end{aligned}
$$

Note that an assumption of Theorem 3.24 is that the directional derivatives $f_i'(\mathbf{x}; \mathbf{d})$ exist. This assumption is automatically satisfied when the functions $f_1, f_2, \ldots, f_m$ are convex. We can thus write the next corollary that replaces the condition on the existence of the directional derivatives by a convexity assumption.

Corollary 3.25 (directional derivative of maximum of functions—convex case). *Suppose that $f(\mathbf{x}) = \max\{f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x})\}$, where $f_1, f_2, \ldots, f_m : \mathbb{E} \to (-\infty, \infty]$ are proper convex functions. Let $\mathbf{x} \in \bigcap_{i=1}^m \mathrm{int}(\mathrm{dom}(f_i))$ and $\mathbf{d} \in \mathbb{E}$. Then*

$$f'(\mathbf{x}; \mathbf{d}) = \max_{i \in I(\mathbf{x})} f_i'(\mathbf{x}; \mathbf{d}),$$

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

3.3.2 The Max Formula

We will now prove an extremely important and useful result, known as the *max formula*, that connects subgradients and directional derivatives.

Theorem 3.26 (max formula). *Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper convex function. Then for any $\mathbf{x} \in \mathrm{int}(\mathrm{dom}(f))$ and $\mathbf{d} \in \mathbb{E}$,*

$$f'(\mathbf{x}; \mathbf{d}) = \max\left\{\langle \mathbf{g}, \mathbf{d} \rangle : \mathbf{g} \in \partial f(\mathbf{x})\right\}. \tag{3.17}$$

Proof. Let $\mathbf{x} \in \mathrm{int}(\mathrm{dom}(f))$ and $\mathbf{d} \in \mathbb{E}$. By the subgradient inequality, we have that for any $\mathbf{g} \in \partial f(\mathbf{x})$,

$$f'(\mathbf{x}; \mathbf{d}) = \lim_{\alpha \to 0^+} \frac{1}{\alpha}(f(\mathbf{x} + \alpha \mathbf{d}) - f(\mathbf{x})) \geq \lim_{\alpha \to 0^+} \langle \mathbf{g}, \mathbf{d} \rangle = \langle \mathbf{g}, \mathbf{d} \rangle \tag{3.18}$$

and, consequently,

$$f'(\mathbf{x}; \mathbf{d}) \geq \max\{\langle \mathbf{g}, \mathbf{d} \rangle : \mathbf{g} \in \partial f(\mathbf{x})\}. \tag{3.19}$$

All that is left is to show the reverse direction of the above inequality. For that, define the function $h(\mathbf{w}) \equiv f'(\mathbf{x}; \mathbf{w})$. Then by Lemma 3.22(a), h is a real-valued convex function and is thus subdifferentiable over $\mathbb{E}$ (Corollary 3.15). Let $\tilde{\mathbf{g}} \in \partial h(\mathbf{d})$. Then for any $\mathbf{v} \in \mathbb{E}$ and $\alpha \geq 0$, using the homogeneity of h (Lemma 3.22(b)),

$$\alpha f'(\mathbf{x}; \mathbf{v}) = f'(\mathbf{x}; \alpha \mathbf{v}) = h(\alpha \mathbf{v}) \geq h(\mathbf{d}) + \langle \tilde{\mathbf{g}}, \alpha \mathbf{v} - \mathbf{d} \rangle = f'(\mathbf{x}; \mathbf{d}) + \langle \tilde{\mathbf{g}}, \alpha \mathbf{v} - \mathbf{d} \rangle.$$

Therefore,

$$\alpha(f'(\mathbf{x}; \mathbf{v}) - \langle \tilde{\mathbf{g}}, \mathbf{v} \rangle) \geq f'(\mathbf{x}; \mathbf{d}) - \langle \tilde{\mathbf{g}}, \mathbf{d} \rangle. \tag{3.20}$$

Since the above inequality holds for any $\alpha \geq 0$, it follows that the coefficient of α in the left-hand side expression is nonnegative (otherwise, inequality (3.20) would be violated for large enough α), meaning that

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

Thus, by Lemma 3.23, for any $\mathbf{y} \in \mathrm{dom}(f)$,

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

showing that $\tilde{\mathbf{g}} \in \partial f(\mathbf{x})$. Taking $\alpha = 0$ in (3.20), we have that $f'(\mathbf{x}; \mathbf{d}) \leq \langle \tilde{\mathbf{g}}, \mathbf{d} \rangle$, so that

$$f'(\mathbf{x}; \mathbf{d}) \leq \langle \tilde{\mathbf{g}}, \mathbf{d} \rangle \leq \max\{\langle \mathbf{g}, \mathbf{d} \rangle : \mathbf{g} \in \partial f(\mathbf{x})\},$$

establishing the desired result. $\square$

Remark 3.27. *The max formula* (3.17) *can also be rewritten using the support function notation as follows:*

$$f'(\mathbf{x}; \mathbf{d}) = \sigma_{\partial f(\mathbf{x})}(\mathbf{d}).$$

3.3.3   Differentiability

Definition 3.28 (differentiability). *Let $f : \mathbb{E} \to (-\infty, \infty]$ and $\mathbf{x} \in \mathrm{int}(\mathrm{dom}f)$. The function f is said to be **differentiable** at $\mathbf{x}$ if there exists $\mathbf{g} \in \mathbb{E}^*$ such that*

$$\lim_{\mathbf{h} \to \mathbf{0}} \frac{f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) - \langle \mathbf{g}, \mathbf{h} \rangle}{\|\mathbf{h}\|} = 0. \tag{3.21}$$

The unique[14] vector $\mathbf{g}$ satisfying (3.21) *is called the **gradient** of f at $\mathbf{x}$ and is denoted by $\nabla f(\mathbf{x})$.*

The above is actually a definition of Fréchet differentiability, which is the one used in this book.

If f is differentiable at $\mathbf{x} \in \mathrm{int}(\mathrm{dom}f)$, then the directional derivative has a simple formula.

Theorem 3.29 (directional derivatives at points of differentiability). *Let $f : \mathbb{E} \to (-\infty, \infty]$ be proper, and suppose that f is differentiable at $\mathbf{x} \in \mathrm{int}(\mathrm{dom}f)$. Then for any $\mathbf{d} \in \mathbb{E}$*

$$f'(\mathbf{x}; \mathbf{d}) = \langle \nabla f(\mathbf{x}), \mathbf{d} \rangle. \tag{3.22}$$

Proof. The formula is obviously correct for $\mathbf{d} = \mathbf{0}$. Suppose that $\mathbf{d} \neq \mathbf{0}$. The differentiability of f implies that

$$\begin{aligned}
0 &= \lim_{\alpha \to 0^+} \frac{f(\mathbf{x} + \alpha \mathbf{d}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \alpha \mathbf{d} \rangle}{\|\alpha \mathbf{d}\|} \\
&= \lim_{\alpha \to 0^+} \left[\frac{f(\mathbf{x} + \alpha \mathbf{d}) - f(\mathbf{x})}{\alpha \|\mathbf{d}\|} - \frac{\langle \nabla f(\mathbf{x}), \mathbf{d} \rangle}{\|\mathbf{d}\|} \right].
\end{aligned}$$

Therefore,

$$\begin{aligned}
f'(\mathbf{x}; \mathbf{d}) &= \lim_{\alpha \to 0^+} \frac{f(\mathbf{x} + \alpha \mathbf{d}) - f(\mathbf{x})}{\alpha} \\
&= \lim_{\alpha \to 0^+} \left\{ \|\mathbf{d}\| \left[\frac{f(\mathbf{x} + \alpha \mathbf{d}) - f(\mathbf{x})}{\alpha \|\mathbf{d}\|} - \frac{\langle \nabla f(\mathbf{x}), \mathbf{d} \rangle}{\|\mathbf{d}\|} \right] + \langle \nabla f(\mathbf{x}), \mathbf{d} \rangle \right\} \\
&= \langle \nabla f(\mathbf{x}), \mathbf{d} \rangle.   \square
\end{aligned}$$

---

[14]The uniqueness can be shown by the following argument. Suppose that (3.21) is satisfied by both $\mathbf{g} = \mathbf{g}_1$ and $\mathbf{g} = \mathbf{g}_2$. Then by subtracting the two limits, we obtain that $\lim_{\mathbf{h} \to \mathbf{0}} \langle \mathbf{g}_1 - \mathbf{g}_2, \mathbf{h} \rangle / \|\mathbf{h}\| = 0$, which immediately shows that $\mathbf{g}_1 = \mathbf{g}_2$.

Example 3.30 (directional derivative of maximum of differentiable functions). Consider the function $f(\mathbf{x}) = \max_{i=1,2,\ldots,m} f_i(\mathbf{x})$, where $f_i : \mathbb{E} \to (-\infty, \infty]$ are proper functions. Assume that $f_1, f_2, \ldots, f_m$ are differentiable at a given point $\mathbf{x} \in \cap_{i=1}^m \operatorname{int}(\operatorname{dom}(f_i))$. Then by Theorem 3.29, for any $\mathbf{d} \in \mathbb{E}$, $f_i'(\mathbf{x}; \mathbf{d}) = \langle \nabla f_i(\mathbf{x}), \mathbf{d} \rangle$. Therefore, invoking Theorem 3.24,

$$f'(\mathbf{x}; \mathbf{d}) = \max_{i \in I(\mathbf{x})} f_i'(\mathbf{x}; \mathbf{d}) = \max_{i \in I(\mathbf{x})} \langle \nabla f_i(\mathbf{x}), \mathbf{d} \rangle,$$

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

Example 3.31 (gradient of $\frac{1}{2} d_C^2(\cdot)$). [15] Suppose that $\mathbb{E}$ is a Euclidean space, and let $C \subseteq \mathbb{E}$ be nonempty closed and convex set. Consider the function $\varphi_C : \mathbb{E} \to \mathbb{R}$ given by $\varphi_C(\mathbf{x}) \equiv \frac{1}{2} d_C^2(\mathbf{x}) = \frac{1}{2} \|\mathbf{x} - P_C(\mathbf{x})\|^2$, where P_C is the so-called *orthogonal projection* mapping defined by

$$P_C(\mathbf{x}) \equiv \operatorname{argmin}_{\mathbf{y} \in C} \|\mathbf{y} - \mathbf{x}\|.$$

It is well known that P_C is well defined (exists and unique) when the underlying set C is nonempty, closed, and convex. [16] We will show that for any $\mathbf{x} \in \mathbb{E}$,

$$\nabla \varphi_C(\mathbf{x}) = \mathbf{x} - P_C(\mathbf{x}). \tag{3.23}$$

For that, fix $\mathbf{x} \in \mathbb{E}$ and define the function $g_{\mathbf{x}}$ by

$$g_{\mathbf{x}}(\mathbf{d}) \equiv \varphi_C(\mathbf{x} + \mathbf{d}) - \varphi_C(\mathbf{x}) - \langle \mathbf{d}, \mathbf{z}_{\mathbf{x}} \rangle,$$

where $\mathbf{z}_{\mathbf{x}} = \mathbf{x} - P_C(\mathbf{x})$. By the definition of the gradient, to show (3.23), it is enough to establish that

$$\frac{g_{\mathbf{x}}(\mathbf{d})}{\|\mathbf{d}\|} \to 0 \text{ as } \mathbf{d} \to 0. \tag{3.24}$$

To prove (3.24), note that by the definition of the orthogonal projection, we have for any $\mathbf{d} \in \mathbb{E}$

$$\|\mathbf{x} + \mathbf{d} - P_C(\mathbf{x} + \mathbf{d})\|^2 \leq \|\mathbf{x} + \mathbf{d} - P_C(\mathbf{x})\|^2,$$

which implies that for any $\mathbf{d} \in \mathbb{E}$,

$$\begin{aligned}
g_{\mathbf{x}}(\mathbf{d}) &= \frac{1}{2} \|\mathbf{x} + \mathbf{d} - P_C(\mathbf{x} + \mathbf{d})\|^2 - \frac{1}{2} \|\mathbf{x} - P_C(\mathbf{x})\|^2 - \langle \mathbf{d}, \mathbf{z}_{\mathbf{x}} \rangle \\
&\leq \frac{1}{2} \|\mathbf{x} + \mathbf{d} - P_C(\mathbf{x})\|^2 - \frac{1}{2} \|\mathbf{x} - P_C(\mathbf{x})\|^2 - \langle \mathbf{d}, \mathbf{z}_{\mathbf{x}} \rangle \\
&= \frac{1}{2} \|\mathbf{x} - P_C(\mathbf{x})\|^2 + \langle \mathbf{d}, \mathbf{x} - P_C(\mathbf{x}) \rangle + \frac{1}{2} \|\mathbf{d}\|^2 - \frac{1}{2} \|\mathbf{x} - P_C(\mathbf{x})\|^2 - \langle \mathbf{d}, \mathbf{z}_{\mathbf{x}} \rangle \\
&= \frac{1}{2} \|\mathbf{d}\|^2. \tag{3.25}
\end{aligned}$$

In particular, we also have

$$g_{\mathbf{x}}(-\mathbf{d}) \leq \frac{1}{2} \|\mathbf{d}\|^2. \tag{3.26}$$

---

[15] The proof in Example 3.31 follows Beck and Teboulle [20, proof of Theorem 4.1].

[16] See, for example, [10, Theorem 8.8]. In addition, see Section 6.4.

Since φ_C is convex, so is $g_{\mathbf{x}}$. Therefore, by Jensen's inequality, and noting that $g_{\mathbf{x}}(\mathbf{0}) = 0$,

$$0 = g_{\mathbf{x}}(\mathbf{0}) = g_{\mathbf{x}}\left(\frac{\mathbf{d} + (-\mathbf{d})}{2}\right) \leq \frac{1}{2}(g_{\mathbf{x}}(\mathbf{d}) + g_{\mathbf{x}}(-\mathbf{d})). \tag{3.27}$$

Combining (3.26) and (3.27), we get

$$g_{\mathbf{x}}(\mathbf{d}) \geq -g_{\mathbf{x}}(-\mathbf{d}) \geq -\frac{1}{2}\|\mathbf{d}\|^2. \tag{3.28}$$

Finally, by (3.25) and (3.28), it follows that $|g_{\mathbf{x}}(\mathbf{d})| \leq \frac{1}{2}\|\mathbf{d}\|^2$, from which the limit (3.24) follows and hence also the desired result (3.23). ∎

Remark 3.32 (what is the gradient?). *We will now illustrate the fact that the gradient depends on the choice of the inner product in the underlying space. Let $\mathbb{E} = \mathbb{R}^n$ be endowed with the dot product. By Theorem 3.29 we know that when f is differentiable at $\mathbf{x}$, then*

$$(\nabla f(\mathbf{x}))_i = \langle \nabla f(\mathbf{x}), \mathbf{e}_i \rangle = f'(\mathbf{x}; \mathbf{e}_i);$$

that is, in this case, the ith component of $\nabla f(\mathbf{x})$ is equal to $\frac{\partial f}{\partial x_i}(\mathbf{x}) = f'(\mathbf{x}; \mathbf{e}_i)$—the ith partial derivative of f at $\mathbf{x}$—so that $\nabla f(\mathbf{x}) = D_f(\mathbf{x})$, where

$$D_f(\mathbf{x}) \equiv \begin{pmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}) \\ \frac{\partial f}{\partial x_2}(\mathbf{x}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\mathbf{x}) \end{pmatrix}. \tag{3.29}$$

Note that the definition of the directional derivative does not depend on the choice of the inner product in the underlying space, so we can arbitrarily choose the inner product in the formula (3.22) as the dot product and obtain (recalling that in this case $\nabla f(\mathbf{x}) = D_f(\mathbf{x})$)

$$f'(\mathbf{x}; \mathbf{d}) = D_f(\mathbf{x})^T \mathbf{d} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\mathbf{x}) d_i. \tag{3.30}$$

Formula (3.30) holds for any choice of inner product in the space. However, $\nabla f(\mathbf{x})$ is not necessarily equal to $D_f(\mathbf{x})$ when the endowed inner product is not the dot product. For example, suppose that the inner product is given by

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{H} \mathbf{y}, \tag{3.31}$$

where $\mathbf{H}$ is a given $n \times n$ positive definite matrix. In this case,

$$\begin{aligned}
(\nabla f(\mathbf{x}))_i = \nabla f(\mathbf{x})^T \mathbf{e}_i &= \nabla f(\mathbf{x})^T \mathbf{H}\left(\mathbf{H}^{-1}\mathbf{e}_i\right) \\
&= \langle \nabla f(\mathbf{x}), \mathbf{H}^{-1}\mathbf{e}_i \rangle && [by\ (3.31)] \\
&= f'(\mathbf{x}; \mathbf{H}^{-1}\mathbf{e}_i) && [by\ (3.22)] \\
&= D_f(\mathbf{x})^T \mathbf{H}^{-1}\mathbf{e}_i. && [by\ (3.30)]
\end{aligned}$$

Hence, we obtain that with respect to the inner product (3.31), *the gradient is actually a "scaled"/"weighted" gradient:*

$$\nabla f(\mathbf{x}) = \mathbf{H}^{-1} D_f(\mathbf{x}).$$

Now consider the space $\mathbb{E} = \mathbb{R}^{m \times n}$ *of all* $m \times n$ *real-valued matrices with the dot product as the endowed inner product:*

$$\langle \mathbf{x}, \mathbf{y} \rangle = \text{Tr}(\mathbf{x}^T \mathbf{y}) \text{ for any } \mathbf{x}, \mathbf{y} \in \mathbb{R}^{m \times n}.$$

Given a proper function $f : \mathbb{R}^{m \times n} \to (-\infty, \infty]$ *and* $\mathbf{x} \in \text{int}(\text{dom}(f))$, *the gradient, if it exists, is given by* $\nabla f(\mathbf{x}) = D_f(\mathbf{x})$, *where* $D_f(\mathbf{x})$ *is the* $m \times n$ *matrix*

$$D_f(\mathbf{x}) = \left(\frac{\partial f}{\partial x_{ij}}(\mathbf{x}) \right)_{i,j}.$$

If the inner product is replaced by

$$\langle \mathbf{x}, \mathbf{y} \rangle = \text{Tr}(\mathbf{x}^T \mathbf{H} \mathbf{y}),$$

where $\mathbf{H}$ *is a given* $m \times m$ *positive definite matrix, then a similar argument to the one given previously shows that*

$$\nabla f(\mathbf{x}) = \mathbf{H}^{-1} D_f(\mathbf{x}).$$

We will now show that when a convex function is differentiable at a point in the interior of its domain, then the subdifferential set is the singleton (i.e., a set containing a single vector) consisting of the gradient at the point. The reverse is also correct in the sense that if the subdifferential set is a singleton $\{\mathbf{g}\}$, then the function is differentiable at the given point with $\mathbf{g}$ being its gradient.

Theorem 3.33 (the subdifferential at points of differentiability). *Let* $f : \mathbb{E} \to (-\infty, \infty]$ *be a proper convex function, and let* $\mathbf{x} \in \text{int}(\text{dom}(f))$. *If* f *is differentiable at* $\mathbf{x}$, *then* $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$. *Conversely, if* f *has a unique subgradient at* $\mathbf{x}$, *then it is differentiable at* $\mathbf{x}$ *and* $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$.

Proof. Let $\mathbf{x} \in \text{int}(\text{dom}(f))$ and assume that f is differentiable at $\mathbf{x}$. Then by Theorem 3.29 it follows that for any $\mathbf{d} \in \mathbb{E}$,

$$f'(\mathbf{x}; \mathbf{d}) = \langle \nabla f(\mathbf{x}), \mathbf{d} \rangle. \tag{3.32}$$

Let $\mathbf{g} \in \partial f(\mathbf{x})$. We will show that $\mathbf{g} = \nabla f(\mathbf{x})$. Combining (3.32) with the max formula (Theorem 3.26) we have

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

so that

$$\langle \mathbf{g} - \nabla f(\mathbf{x}), \mathbf{d} \rangle \leq 0.$$

Taking the maximum over all $\mathbf{d}$ satisfying $\|\mathbf{d}\| \leq 1$, we obtain that $\|\mathbf{g} - \nabla f(\mathbf{x})\|_* \leq 0$ and consequently that $\nabla f(\mathbf{x}) = \mathbf{g}$. We have thus shown that the only possible

subgradient in $\partial f(\mathbf{x})$ is $\nabla f(\mathbf{x})$. Combining this with the fact that the subdifferential set is nonempty (Theorem 3.14) yields the desired result $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$.

For the reverse direction, suppose that f has a unique subgradient $\mathbf{g}$ at $\mathbf{x} \in \operatorname{int}(\operatorname{dom}(f))$. Consider the convex function

$$h(\mathbf{u}) \equiv f(\mathbf{x} + \mathbf{u}) - f(\mathbf{x}) - \langle \mathbf{g}, \mathbf{u} \rangle.$$

We will show that

$$\lim_{\mathbf{u} \to \mathbf{0}} \frac{h(\mathbf{u})}{\|\mathbf{u}\|} = 0.$$

This will establish (by definition) that $\mathbf{g} = \nabla f(\mathbf{x})$. Obviously, $\mathbf{0}$ is the unique subgradient of h at $\mathbf{0}$ and $\mathbf{0} \in \operatorname{int}(\operatorname{dom}(h))$, and hence by the max formula (Theorem 3.26), for any $\mathbf{d} \in \mathbb{E}$,

$$h'(\mathbf{0}; \mathbf{d}) = \sigma_{\partial h(\mathbf{0})}(\mathbf{d}) = 0.$$

We can thus conclude that for any $\mathbf{d} \in \mathbb{E}$,

$$0 = h'(\mathbf{0}; \mathbf{d}) = \lim_{\alpha \to 0^+} \frac{h(\alpha \mathbf{d}) - h(\mathbf{0})}{\alpha} = \lim_{\alpha \to 0^+} \frac{h(\alpha \mathbf{d})}{\alpha}. \tag{3.33}$$

Let $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$ be an orthonormal basis of $\mathbb{E}$. Since $\mathbf{0} \in \operatorname{int}(\operatorname{dom}(h))$, there exists $\varepsilon \in (0, 1)$ such that $\varepsilon \mathbf{v}_i, -\varepsilon \mathbf{v}_i \in \operatorname{dom}(h)$ for any $i = 1, 2, \ldots, k$. Therefore, since $\operatorname{dom}(h)$ is convex, the set

$$D = \operatorname{conv}\left(\{\pm \varepsilon \mathbf{v}_i\}_{i=1}^k\right)$$

satisfies $D \subseteq \operatorname{dom}(h)$. Let $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$ be the Euclidean norm corresponding to the endowed inner product on $\mathbb{E}$. Note that $B_{\|\cdot\|}[\mathbf{0}, \gamma] \subseteq D$, where $\gamma = \frac{\varepsilon}{k}$. Indeed, let $\mathbf{w} \in B_{\|\cdot\|}[\mathbf{0}, \gamma]$. Then since $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$ is an orthonormal basis of $\mathbb{E}$, we have

$$\mathbf{w} = \sum_{i=1}^k \langle \mathbf{w}, \mathbf{v}_i \rangle \mathbf{v}_i$$

as well as

$$\|\mathbf{w}\|^2 = \sum_{i=1}^k \langle \mathbf{w}, \mathbf{v}_i \rangle^2. \tag{3.34}$$

Since $\|\mathbf{w}\|^2 \leq \gamma^2$, it follows by (3.34) that $|\langle \mathbf{w}, \mathbf{v}_i \rangle| \leq \gamma$, and hence

$$\mathbf{w} = \sum_{i=1}^k \langle \mathbf{w}, \mathbf{v}_i \rangle \mathbf{v}_i = \sum_{i=1}^k \frac{|\langle \mathbf{w}, \mathbf{v}_i \rangle|}{\varepsilon} [\operatorname{sgn}(\langle \mathbf{w}, \mathbf{v}_i \rangle) \varepsilon \mathbf{v}_i] + \left(1 - \sum_{i=1}^k \frac{|\langle \mathbf{w}, \mathbf{v}_i \rangle|}{\varepsilon}\right) \cdot \mathbf{0} \in D,$$

where the membership in D follows by the fact that $\mathbf{0}, \pm \varepsilon \mathbf{v}_i \in D$ and $\sum_{i=1}^k \frac{|\langle \mathbf{w}, \mathbf{v}_i \rangle|}{\varepsilon} \leq \frac{k\gamma}{\varepsilon} = 1$. We have therefore established the inclusion $B_{\|\cdot\|}[\mathbf{0}, \gamma] \subseteq D$. Denote the $2k$ vectors $\{\pm \varepsilon \mathbf{v}_i\}_{i=1}^k$ by $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{2k}$. Take $\mathbf{0} \neq \mathbf{u} \in B_{\|\cdot\|}[\mathbf{0}, \gamma^2]$. We have that $\gamma \frac{\mathbf{u}}{\|\mathbf{u}\|} \in B_{\|\cdot\|}[\mathbf{0}, \gamma] \subseteq D$, and hence there exists $\boldsymbol{\lambda} \in \Delta_{2k}$ such that

$$\gamma \frac{\mathbf{u}}{\|\mathbf{u}\|} = \sum_{i=1}^{2k} \lambda_i \mathbf{z}_i.$$

Therefore,

$$
\begin{aligned}
\frac{h(\mathbf{u})}{\|\mathbf{u}\|} &= \frac{h\left(\frac{\|\mathbf{u}\|}{\gamma}\gamma\frac{\mathbf{u}}{\|\mathbf{u}\|}\right)}{\|\mathbf{u}\|} = \frac{h\left(\sum_{i=1}^{2k}\lambda_i\frac{\|\mathbf{u}\|}{\gamma}\mathbf{z}_i\right)}{\|\mathbf{u}\|} \\
&\le \sum_{i=1}^{2k}\lambda_i\frac{h\left(\|\mathbf{u}\|\frac{\mathbf{z}_i}{\gamma}\right)}{\|\mathbf{u}\|} \\
&\le \max_{i=1,2,\ldots,2k}\left\{\frac{h\left(\|\mathbf{u}\|\frac{\mathbf{z}_i}{\gamma}\right)}{\|\mathbf{u}\|}\right\},
\end{aligned}
\tag{3.35}
$$

where the first inequality follows by the convexity of h and by the fact that $\|\mathbf{u}\|\frac{\mathbf{z}_i}{\gamma}\in B_{\|\cdot\|}[\mathbf{0},\gamma]\subseteq D\subseteq\operatorname{dom}(h)$. By (3.33),

$$
\lim_{\mathbf{u}\to\mathbf{0}}\frac{h\left(\|\mathbf{u}\|\frac{\mathbf{z}_i}{\gamma}\right)}{\|\mathbf{u}\|} = \lim_{\|\mathbf{u}\|\to\mathbf{0}}\frac{h\left(\|\mathbf{u}\|\frac{\mathbf{z}_i}{\gamma}\right)}{\|\mathbf{u}\|} = \lim_{\alpha\to0^+}\frac{h\left(\alpha\frac{\mathbf{z}_i}{\gamma}\right)}{\alpha} = 0,
$$

which, combined with (3.35), implies that $\frac{h(\mathbf{u})}{\|\mathbf{u}\|}\to0$ as $\mathbf{u}\to\mathbf{0}$, proving the desired result. $\quad\square$

Example 3.34 (subdifferential of the l_2-norm). Let $f:\mathbb{R}^n\to\mathbb{R}$ be given by $f(\mathbf{x})=\|\mathbf{x}\|_2$. Then the subdifferential set of f at $\mathbf{x}=\mathbf{0}$ was already computed in Example 3.3. When $\mathbf{x}\neq\mathbf{0}$, the function is differentiable with gradient $\frac{\mathbf{x}}{\|\mathbf{x}\|_2}$. Thus, using Theorem 3.33, we can summarize and write the subdifferential set as

$$
\partial f(\mathbf{x}) = \begin{cases} \left\{\frac{\mathbf{x}}{\|\mathbf{x}\|_2}\right\}, & \mathbf{x}\neq\mathbf{0}, \\[2mm] B_{\|\cdot\|_2}[\mathbf{0},1], & \mathbf{x}=\mathbf{0}. \end{cases}
$$

In particular, when considering the case $n=1$, we obtain that for the one-dimensional function $g(x)=|x|$, we have

$$
\partial g(x) = \begin{cases} \{\operatorname{sgn}(x)\}, & x\neq0, \\[2mm] [-1,1], & x=0. \end{cases} \qquad\blacksquare
$$

3.4 Computing Subgradients

This section establishes several useful calculus rules for subgradients and subdifferentials. Some of the results are "weak results" (rules for computing some of the subgradients in the subdifferential set), and some are "strong" (full characterization of the subdifferential set).

3.4.1 Multiplication by a Positive Scalar

Theorem 3.35. *Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper function and let $\alpha > 0$. Then for any $\mathbf{x} \in \mathrm{dom}(f)$*

$$\partial(\alpha f)(\mathbf{x}) = \alpha \partial f(\mathbf{x}).$$

Proof. We have that $\mathbf{g} \in \partial f(\mathbf{x})$ if and only if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle \text{ for any } \mathbf{y} \in \mathrm{dom}(f).$$

Multiplying the inequality by $\alpha > 0$, we can conclude that the above inequality holds if and only if

$$\alpha f(\mathbf{y}) \geq \alpha f(\mathbf{x}) + \langle \alpha \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle \text{ for any } \mathbf{y} \in \mathrm{dom}(\alpha f), \tag{3.36}$$

where we used the obvious fact that $\mathrm{dom}(\alpha f) = \mathrm{dom}(f)$. The statement (3.36) is equivalent to the relation $\alpha \mathbf{g} \in \partial(\alpha f)(\mathbf{x})$. $\quad\square$

3.4.2 Summation

The following result contains both weak and strong results on the subdifferential set of a sum of functions. The weak result is also "weak" in the sense that its proof only requires the definition of the subgradient. The strong result utilizes the max formula.

Theorem 3.36. *Let $f_1, f_2 : \mathbb{E} \to (-\infty, \infty]$ be proper convex functions, and let $\mathbf{x} \in \mathrm{dom}(f_1) \cap \mathrm{dom}(f_2)$.*

(a) *The following inclusion holds:*

$$\partial f_1(\mathbf{x}) + \partial f_2(\mathbf{x}) \subseteq \partial(f_1 + f_2)(\mathbf{x}).$$

(b) *If $\mathbf{x} \in \mathrm{int}(\mathrm{dom}(f_1)) \cap \mathrm{int}(\mathrm{dom}(f_2))$, then*

$$\partial(f_1 + f_2)(\mathbf{x}) = \partial f_1(\mathbf{x}) + \partial f_2(\mathbf{x}).$$

Proof. (a) Let $\mathbf{g} \in \partial f_1(\mathbf{x}) + \partial f_2(\mathbf{x})$. Then there exist $\mathbf{g}_1 \in \partial f_1(\mathbf{x})$ and $\mathbf{g}_2 \in \partial f_2(\mathbf{x})$ such that $\mathbf{g} = \mathbf{g}_1 + \mathbf{g}_2$. By the definition of $\mathbf{g}_1$ and $\mathbf{g}_2$, it follows that for any $\mathbf{y} \in \mathrm{dom}(f_1) \cap \mathrm{dom}(f_2)$,

$$f_1(\mathbf{y}) \geq f_1(\mathbf{x}) + \langle \mathbf{g}_1, \mathbf{y} - \mathbf{x} \rangle,$$
$$f_2(\mathbf{y}) \geq f_2(\mathbf{x}) + \langle \mathbf{g}_2, \mathbf{y} - \mathbf{x} \rangle.$$

Summing the two inequalities, we obtain that for any $\mathbf{y} \in \mathrm{dom}(f_1) \cap \mathrm{dom}(f_2)$,

$$f_1(\mathbf{y}) + f_2(\mathbf{y}) \geq f_1(\mathbf{x}) + f_2(\mathbf{x}) + \langle \mathbf{g}_1 + \mathbf{g}_2, \mathbf{y} - \mathbf{x} \rangle,$$

that is, $\mathbf{g} = \mathbf{g}_1 + \mathbf{g}_2 \in \partial(f_1 + f_2)(\mathbf{x})$.

(b) Let $\mathbf{d} \in \mathbb{E}$ and define $f \equiv f_1 + f_2$. Then since $\mathbf{x} \in \mathrm{int}(\mathrm{dom}(f)) = \mathrm{int}(\mathrm{dom}(f_1)) \cap \mathrm{int}(\mathrm{dom}(f_2))$, it follows by the max formula (Theorem 3.26) that

$$\sigma_{\partial f(\mathbf{x})}(\mathbf{d}) = \max\{\langle \mathbf{g}, \mathbf{d}\rangle : \mathbf{g} \in \partial f(\mathbf{x})\} = f'(\mathbf{x}; \mathbf{d}).$$

Using the additivity of the directional derivative and the max formula (again), we also obtain

$$\begin{aligned}
\sigma_{\partial f(\mathbf{x})}(\mathbf{d}) &= f'(\mathbf{x}; \mathbf{d}) \\
&= f_1'(\mathbf{x}; \mathbf{d}) + f_2'(\mathbf{x}; \mathbf{d}) \\
&= \max\{\langle \mathbf{g}_1, \mathbf{d}\rangle : \mathbf{g}_1 \in \partial f_1(\mathbf{x})\} + \max\{\langle \mathbf{g}_2, \mathbf{d}\rangle : \mathbf{g}_2 \in \partial f_2(\mathbf{x})\} \\
&= \max\{\langle \mathbf{g}_1 + \mathbf{g}_2, \mathbf{d}\rangle : \mathbf{g}_1 \in \partial f_1(\mathbf{x}), \mathbf{g}_2 \in \partial f_2(\mathbf{x})\}\} \\
&= \sigma_{\partial f_1(\mathbf{x})+\partial f_2(\mathbf{x})}(\mathbf{d}).
\end{aligned}$$

By Theorems 3.9 and 3.14, $\partial f(\mathbf{x})$, $\partial f_1(\mathbf{x})$, and $\partial f_2(\mathbf{x})$ are nonempty compact convex sets, which also implies (simple exercise) that $\partial f_1(\mathbf{x}) + \partial f_2(\mathbf{x})$ is nonempty compact and convex. Finally, invoking Lemma 2.34, it follows that $\partial f(\mathbf{x}) = \partial f_1(\mathbf{x}) + \partial f_2(\mathbf{x})$. $\quad\square$

Remark 3.37. *Note that the proof of part* (a) *of Theorem* 3.36 *does not require a convexity assumption on f_1 and f_2.*

A simple induction argument can be used to generalize the last result to an arbitrary number of functions.

Corollary 3.38. *Let $f_1, f_2, \ldots, f_m : \mathbb{E} \to (-\infty, \infty]$ be proper convex functions, and let $\mathbf{x} \in \cap_{i=1}^m \mathrm{dom}(f_i)$.*

(a) **(weak sum rule of subdifferential calculus)** *The following inclusion holds:*

$$\sum_{i=1}^m \partial f_i(\mathbf{x}) \subseteq \partial\left(\sum_{i=1}^m f_i\right)(\mathbf{x}).$$

(b) *If $\mathbf{x} \in \cap_{i=1}^m \mathrm{int}(\mathrm{dom}(f_i))$, then*

$$\partial\left(\sum_{i=1}^m f_i\right)(\mathbf{x}) = \sum_{i=1}^m \partial f_i(\mathbf{x}). \tag{3.37}$$

Another direct consequence is that if $f_1, f_2, \ldots, f_m$ are real-valued, meaning that their domain is the entire space $\mathbb{E}$, then the sum formula (3.37) holds.

Corollary 3.39. *Let $f_1, f_2, \ldots, f_m : \mathbb{E} \to \mathbb{R}$ be real-valued convex functions. Then for any $\mathbf{x} \in \mathbb{E}$*

$$\partial\left(\sum_{i=1}^m f_i\right)(\mathbf{x}) = \sum_{i=1}^m \partial f_i(\mathbf{x}).$$

A result with a less restrictive assumption than the one in Corollary 3.38(b) states that if the intersection $\cap_{i=1}^m \mathrm{ri}(\mathrm{dom}(f_i))$ is nonempty, then the sum formula is correct at *any* point. We state this result without a proof.

Theorem 3.40 (sum rule of subdifferential calculus [108, Theorem 23.8]). *Let $f_1, f_2, \ldots, f_m : \mathbb{E} \to (-\infty, \infty]$ be proper convex functions, and assume that $\bigcap_{i=1}^m \mathrm{ri}(\mathrm{dom}(f_i)) \neq \emptyset$. Then for any $\mathbf{x} \in \mathbb{E}$*

$$\partial \left(\sum_{i=1}^m f_i \right) (\mathbf{x}) = \sum_{i=1}^m \partial f_i(\mathbf{x}).$$

Example 3.41 (subdifferential set of the l_1-norm function—strong result). Consider the function $f : \mathbb{R}^n \to \mathbb{R}$ given by $f(\mathbf{x}) = \|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$. Then $f = \sum_{i=1}^n f_i$, where $f_i(\mathbf{x}) \equiv |x_i|$. We have (see also Example 3.34)

$$\partial f_i(\mathbf{x}) = \begin{cases} \{\mathrm{sgn}(x_i)\mathbf{e}_i\}, & x_i \neq 0, \\ [-\mathbf{e}_i, \mathbf{e}_i], & x_i = 0. \end{cases}$$

Thus, by Corollary 3.39,

$$\partial f(\mathbf{x}) = \sum_{i=1}^n \partial f_i(\mathbf{x}) = \sum_{i \in I_{\neq}(\mathbf{x})} \mathrm{sgn}(x_i)\mathbf{e}_i + \sum_{i \in I_0(\mathbf{x})} [-\mathbf{e}_i, \mathbf{e}_i],$$

where

$$I_{\neq}(\mathbf{x}) = \{i : x_i \neq 0\}, \ I_0(\mathbf{x}) = \{i : x_i = 0\},$$

and hence

$$\partial f(\mathbf{x}) = \{\mathbf{z} \in \mathbb{R}^n : z_i = \mathrm{sgn}(x_i), i \in I_{\neq}(\mathbf{x}), |z_j| \leq 1, j \in I_0(\mathbf{x})\}.$$

∎

Example 3.42 (a subgradient of the l_1-norm function—weak result). Using the formula for the subdifferential set described in Example 3.41, we can readily conclude that

$$\mathrm{sgn}(\mathbf{x}) \in \partial f(\mathbf{x}).$$

∎

3.4.3 Affine Transformation

The following theorem states strong and weak results on the subdifferential set of a composition of a convex function with an affine transformation.

Theorem 3.43. *Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper convex function and $\mathcal{A} : \mathbb{V} \to \mathbb{E}$ be a linear transformation. Let $h(\mathbf{x}) = f(\mathcal{A}(\mathbf{x}) + \mathbf{b})$ with $\mathbf{b} \in \mathbb{E}$. Assume that h is proper, meaning that*

$$\mathrm{dom}(h) = \{\mathbf{x} \in \mathbb{V} : \mathcal{A}(\mathbf{x}) + \mathbf{b} \in \mathrm{dom}(f)\} \neq \emptyset.$$

(a) **(weak affine transformation rule of subdifferential calculus)** *For any* $\mathbf{x} \in \text{dom}(h)$,

$$\mathcal{A}^T(\partial f(\mathcal{A}(\mathbf{x}) + \mathbf{b})) \subseteq \partial h(\mathbf{x}).$$

(b) **(affine transformation rule of subdifferential calculus)** *If* $\mathbf{x} \in \text{int}(\text{dom}(h))$ *and* $\mathcal{A}(\mathbf{x}) + \mathbf{b} \in \text{int}(\text{dom}(f))$, *then*

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

Proof. (a) Let $\mathbf{x} \in \text{dom}(h)$ and assume that $\mathbf{g} \in \mathcal{A}^T(\partial f(\mathcal{A}(\mathbf{x}) + \mathbf{b}))$. Then there exists $\mathbf{d} \in \mathbb{E}^*$ for which $\mathbf{g} = \mathcal{A}^T(\mathbf{d})$, where

$$\mathbf{d} \in \partial f(\mathcal{A}(\mathbf{x}) + \mathbf{b}). \tag{3.38}$$

For any $\mathbf{y} \in \text{dom}(h)$, we have $\mathcal{A}(\mathbf{y}) + \mathbf{b} \in \text{dom}(f)$, and hence, by (3.38),

$$f(\mathcal{A}(\mathbf{y}) + \mathbf{b}) \geq f(\mathcal{A}(\mathbf{x}) + \mathbf{b}) + \langle \mathbf{d}, \mathcal{A}(\mathbf{y}) + \mathbf{b} - \mathcal{A}(\mathbf{x}) - \mathbf{b} \rangle,$$

and therefore

$$h(\mathbf{y}) \geq h(\mathbf{x}) + \langle \mathcal{A}^T(\mathbf{d}), \mathbf{y} - \mathbf{x} \rangle \text{ for all } \mathbf{y} \in \text{dom}(h).$$

Hence, $\mathbf{g} = \mathcal{A}^T(\mathbf{d}) \in \partial h(\mathbf{x})$, proving that $\mathcal{A}^T(\partial f(\mathcal{A}(\mathbf{x}) + \mathbf{b})) \subseteq \partial h(\mathbf{x})$.

(b) Since $\mathbf{x} \in \text{int}(\text{dom}(h))$, then for any $\mathbf{d} \in \mathbb{V}$, by the max formula (Theorem 3.26),

$$h'(\mathbf{x}; \mathbf{d}) = \sigma_{\partial h(\mathbf{x})}(\mathbf{d}). \tag{3.39}$$

In addition, by the definition of the directional derivative, we have

$$\begin{aligned}
h'(\mathbf{x}; \mathbf{d}) &= \lim_{\alpha \to 0^+} \frac{h(\mathbf{x} + \alpha \mathbf{d}) - h(\mathbf{x})}{\alpha} \\
&= \lim_{\alpha \to 0^+} \frac{f(\mathcal{A}(\mathbf{x}) + \mathbf{b} + \alpha \mathcal{A}(\mathbf{d})) - f(\mathcal{A}(\mathbf{x}) + \mathbf{b})}{\alpha} \\
&= f'(\mathcal{A}(\mathbf{x}) + \mathbf{b}; \mathcal{A}(\mathbf{d})),
\end{aligned}$$

which, combined with (3.39), yields

$$\sigma_{\partial h(\mathbf{x})}(\mathbf{d}) = f'(\mathcal{A}(\mathbf{x}) + \mathbf{b}; \mathcal{A}(\mathbf{d})).$$

Therefore, using the max formula again and the assumption that $\mathcal{A}(\mathbf{x}) + \mathbf{b} \in \text{int}(\text{dom}(f))$, we obtain that

$$\begin{aligned}
\sigma_{\partial h(\mathbf{x})}(\mathbf{d}) &= f'(\mathcal{A}(\mathbf{x}) + \mathbf{b}; \mathcal{A}(\mathbf{d})) \\
&= \max_{\mathbf{g}} \{ \langle \mathbf{g}, \mathcal{A}(\mathbf{d}) \rangle : \mathbf{g} \in \partial f(\mathcal{A}(\mathbf{x}) + \mathbf{b}) \} \\
&= \max_{\mathbf{g}} \{ \langle \mathcal{A}^T(\mathbf{g}), \mathbf{d} \rangle : \mathbf{g} \in \partial f(\mathcal{A}(\mathbf{x}) + \mathbf{b}) \} \\
&= \max_{\tilde{\mathbf{g}}} \{ \langle \tilde{\mathbf{g}}, \mathbf{d} \rangle : \tilde{\mathbf{g}} \in \mathcal{A}^T(\partial f(\mathcal{A}(\mathbf{x}) + \mathbf{b})) \} \\
&= \sigma_{\mathcal{A}^T(\partial f(\mathcal{A}(\mathbf{x}) + \mathbf{b}))}(\mathbf{d}).
\end{aligned}$$

Since $\mathbf{x} \in \text{int}(\text{dom}(h))$, it follows by Theorems 3.9 and 3.14 that $\partial h(\mathbf{x})$ is nonempty compact and convex. Similarly, since $\mathcal{A}(\mathbf{x}) + \mathbf{b} \in \text{int}(\text{dom}(f))$, the set $\partial f(\mathcal{A}(\mathbf{x}) + \mathbf{b})$ is nonempty, compact, and convex, which implies that $\mathcal{A}^T(\partial f(\mathcal{A}(\mathbf{x}) + \mathbf{b}))$ is also nonempty, compact, and convex. Finally, invoking Lemma 2.34, we obtain that $\partial h(\mathbf{x}) = \mathcal{A}^T(\partial f(\mathcal{A}(\mathbf{x}) + \mathbf{b}))$. $\quad\square$

Example 3.44 (subdifferential of $\|\mathbf{Ax} + \mathbf{b}\|_1$). Let $f : \mathbb{R}^n \to \mathbb{R}$ be the function given by $f(\mathbf{x}) = \|\mathbf{Ax} + \mathbf{b}\|_1$, where $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{b} \in \mathbb{R}^m$. Then $f(\mathbf{x}) = g(\mathbf{Ax} + \mathbf{b})$ with $g : \mathbb{R}^m \to \mathbb{R}$ given by $g(\mathbf{y}) = \|\mathbf{y}\|_1$. By the affine transformation rule of subdifferential calculus (Theorem 3.43(b)), we have that

$$\partial f(\mathbf{x}) = \mathbf{A}^T \partial g(\mathbf{Ax} + \mathbf{b}). \tag{3.40}$$

Denote the ith row of $\mathbf{A}$ by $\mathbf{a}_i^T$ and define

$$I_{\neq}(\mathbf{x}) = \{i : \mathbf{a}_i^T \mathbf{x} + b_i \neq 0\},$$
$$I_0(\mathbf{x}) = \{i : \mathbf{a}_i^T \mathbf{x} + b_i = 0\}.$$

In this terminology, by Example 3.41,

$$\partial g(\mathbf{Ax} + \mathbf{b}) = \sum_{i \in I_{\neq}(\mathbf{x})} \text{sgn}(\mathbf{a}_i^T \mathbf{x} + b_i)\mathbf{e}_i + \sum_{i \in I_0(\mathbf{x})} [-\mathbf{e}_i, \mathbf{e}_i].$$

Thus, by (3.40),

$$\partial f(\mathbf{x}) = \mathbf{A}^T \partial g(\mathbf{Ax} + \mathbf{b})$$
$$= \sum_{i \in I_{\neq}(\mathbf{x})} \text{sgn}(\mathbf{a}_i^T \mathbf{x} + b_i)\mathbf{A}^T\mathbf{e}_i + \sum_{i \in I_0(\mathbf{x})} [-\mathbf{A}^T\mathbf{e}_i, \mathbf{A}^T\mathbf{e}_i].$$

Using the relation $\mathbf{A}^T\mathbf{e}_i = \mathbf{a}_i$, we finally conclude that

$$\partial f(\mathbf{x}) = \sum_{i \in I_{\neq}(\mathbf{x})} \text{sgn}(\mathbf{a}_i^T \mathbf{x} + b_i)\mathbf{a}_i + \sum_{i \in I_0(\mathbf{x})} [-\mathbf{a}_i, \mathbf{a}_i].$$

The above is a strong result characterizing the entire subdifferential set. A weak result indicating one possible subgradient is

$$\mathbf{A}^T\text{sgn}(\mathbf{Ax} + \mathbf{b}) \in \partial f(\mathbf{x}). \quad \blacksquare$$

Example 3.45 (subdifferential of $\|\mathbf{Ax} + \mathbf{b}\|_2$). Let $f : \mathbb{R}^n \to \mathbb{R}$ be the function $f(\mathbf{x}) = \|\mathbf{Ax} + \mathbf{b}\|_2$, where $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{b} \in \mathbb{R}^m$. Then $f(\mathbf{x}) = g(\mathbf{Ax} + \mathbf{b})$ with $g : \mathbb{R}^m \to \mathbb{R}$ given by $g(\mathbf{y}) = \|\mathbf{y}\|_2$. By Example 3.34,

$$\partial g(\mathbf{Ax} + \mathbf{b}) = \begin{cases} \frac{\mathbf{Ax}+\mathbf{b}}{\|\mathbf{Ax}+\mathbf{b}\|_2}, & \mathbf{Ax} + \mathbf{b} \neq \mathbf{0}, \\ B_{\|\cdot\|_2}[\mathbf{0}, 1], & \mathbf{Ax} + \mathbf{b} = \mathbf{0}. \end{cases}$$

Thus, by the affine transformation rule of subdifferential calculus (Theorem 3.43(b)),

$$\partial f(\mathbf{x}) = \mathbf{A}^T \partial g(\mathbf{A}\mathbf{x} + \mathbf{b}) = \begin{cases} \frac{\mathbf{A}^T(\mathbf{A}\mathbf{x}+\mathbf{b})}{\|\mathbf{A}\mathbf{x}+\mathbf{b}\|_2}, & \mathbf{A}\mathbf{x} + \mathbf{b} \neq \mathbf{0}, \\ \mathbf{A}^T B_{\|\cdot\|_2}[\mathbf{0}, 1], & \mathbf{A}\mathbf{x} + \mathbf{b} = \mathbf{0}. \end{cases}$$

Note that at a vector $\mathbf{x}$ satisfying $\mathbf{A}\mathbf{x} + \mathbf{b} = \mathbf{0}$, the subdifferential set can be explicitly written as

$$\partial f(\mathbf{x}) = \mathbf{A}^T B_{\|\cdot\|_2}[\mathbf{0}, 1] = \left\{ \mathbf{A}^T \mathbf{y} : \|\mathbf{y}\|_2 \leq 1 \right\}.$$

If a weak result is required, then since $\mathbf{0} \in B_{\|\cdot\|_2}[\mathbf{0}, 1]$, we can write $\mathbf{0} = \mathbf{A}^T \mathbf{0} \in \partial f(\mathbf{x})$ for any $\mathbf{x}$ satisfying $\mathbf{A}\mathbf{x} + \mathbf{b} = \mathbf{0}$. ∎

3.4.4 Composition

The derivative of a composition of differentiable functions can be computed by using the well-known chain rule. We recall here the classical result on the derivative of the composition of two one-dimensional functions. The result is a small variation of the result from [112, Theorem 5.5].

Theorem 3.46. *Suppose that f is continuous on $[a, b]$ $(a < b)$ and that $f'_+(a)$ exists. Let g be a function defined on an open interval I which contains the range of f, and assume that g is differentiable at $f(a)$. Then the function*

$$h(t) = g(f(t)) \quad (a \leq t \leq b)$$

is right differentiable at $t = a$ and

$$h'_+(a) = g'(f(a))f'_+(a).$$

Proof.

$$\begin{aligned} h'_+(a) &= \lim_{t \to a^+} \frac{g(f(t)) - g(f(a))}{t - a} \\ &= \lim_{t \to a^+} \frac{g(f(t)) - g(f(a))}{f(t) - f(a)} \cdot \frac{f(t) - f(a)}{t - a} = g'(f(a))f'_+(a). \quad \square \end{aligned}$$

We will now show how the one-dimensional chain rule can be used with the help of the max formula (Theorem 3.26) to show a multidimensional version of the chain rule.

Theorem 3.47 (chain rule of subdifferential calculus). Let $f : \mathbb{E} \to \mathbb{R}$ be a convex function and $g : \mathbb{R} \to \mathbb{R}$ be a nondecreasing convex function. Let $\mathbf{x} \in \mathbb{E}$, and suppose that g is differentiable at the point $f(\mathbf{x})$. Let $h = g \circ f$. Then

$$\partial h(\mathbf{x}) = g'(f(\mathbf{x}))\partial f(\mathbf{x}).$$

Proof. For any $\mathbf{d} \in \mathbb{E}$, define the following one-dimensional functions:

$$\begin{aligned} f_{\mathbf{x},\mathbf{d}}(t) &= f(\mathbf{x} + t\mathbf{d}), \quad t \in \mathbb{R}, \\ h_{\mathbf{x},\mathbf{d}}(t) &= h(\mathbf{x} + t\mathbf{d}), \quad t \in \mathbb{R}. \end{aligned}$$

We have the following simple relation between $f_{\mathbf{x},\mathbf{d}}$ and $h_{\mathbf{x},\mathbf{d}}$:

$$h_{\mathbf{x},\mathbf{d}}(t) = h(\mathbf{x} + t\mathbf{d}) = g(f(\mathbf{x} + t\mathbf{d})) = g(f_{\mathbf{x},\mathbf{d}}(t)), \quad t \in \mathbb{R}. \tag{3.41}$$

The function f is convex by the premise of the theorem, and h is convex since it is a composition of a nondecreasing convex function with a convex function. Therefore, the directional derivatives of f and h exist in every direction (Theorem 3.21), and we have by the definition of the directional derivative that

$$(f_{\mathbf{x},\mathbf{d}})'_+(0) = f'(\mathbf{x}; \mathbf{d}), \tag{3.42}$$
$$(h_{\mathbf{x},\mathbf{d}})'_+(0) = h'(\mathbf{x}; \mathbf{d}). \tag{3.43}$$

Since $h_{\mathbf{x},\mathbf{d}} = g \circ f_{\mathbf{x},\mathbf{d}}$ (by (3.41)), $f_{\mathbf{x},\mathbf{d}}$ is right differentiable at 0, and g is differentiable at $f_{\mathbf{x},\mathbf{d}}(0) = f(\mathbf{x})$, it follows by the chain rule for one-dimensional functions (Theorem 3.46) that

$$(h_{\mathbf{x},\mathbf{d}})'_+(0) = g'(f(\mathbf{x}))(f_{\mathbf{x};\mathbf{d}})'_+(0).$$

Plugging (3.42) and (3.43) into the latter equality, we obtain

$$h'(\mathbf{x}; \mathbf{d}) = g'(f(\mathbf{x}))f'(\mathbf{x}; \mathbf{d}).$$

By the max formula (Theorem 3.26), since f and h are convex and $\mathbf{x} \in \text{int}(\text{dom}(f)) = \text{int}(\text{dom}(h)) = \mathbb{E}$,

$$h'(\mathbf{x}; \mathbf{d}) = \sigma_{\partial h(\mathbf{x})}(\mathbf{d}), \ f'(\mathbf{x}; \mathbf{d}) = \sigma_{\partial f(\mathbf{x})}(\mathbf{d}),$$

and hence

$$\sigma_{\partial h(\mathbf{x})}(\mathbf{d}) = h'(\mathbf{x}; \mathbf{d}) = g'(f(\mathbf{x}))f'(\mathbf{x}; \mathbf{d}) = g'(f(\mathbf{x}))\sigma_{\partial f(\mathbf{x})}(\mathbf{d}) = \sigma_{g'(f(\mathbf{x}))\partial f(\mathbf{x})}(\mathbf{d}),$$

where the last equality is due to Lemma 2.24(c) and the fact that $g'(f(\mathbf{x})) \geq 0$. Finally, by Theorems 3.9 and 3.14 the sets $\partial h(\mathbf{x}), \partial f(\mathbf{x})$ are nonempty, closed, and convex, and thus by Lemma 2.34

$$\partial h(\mathbf{x}) = g'(f(\mathbf{x}))\partial f(\mathbf{x}). \quad \square$$

Example 3.48 (subdifferential of $\|\cdot\|_1^2$). Consider the function $h : \mathbb{R}^n \to \mathbb{R}$ given by $h(\mathbf{x}) = \|\mathbf{x}\|_1^2$, which can be written as the composition $h = g \circ f$, where $f(\mathbf{x}) = \|\mathbf{x}\|_1$ and $g(t) = [t]_+^2 = \max\{t, 0\}^2$. Both f and g are real-valued convex functions, and g is nondecreasing and differentiable over $\mathbb{R}$ with derivative $g'(t) = 2[t]_+$. Therefore, by the chain rule of subdifferential calculus (Theorem 3.47), for any $\mathbf{x} \in \mathbb{R}^n$,

$$\partial h(\mathbf{x}) = g'(f(\mathbf{x}))\partial f(\mathbf{x}) = 2\left[\|\mathbf{x}\|_1\right]_+ \partial f(\mathbf{x}) = 2\|\mathbf{x}\|_1 \partial f(\mathbf{x}).$$

Using the general form of $\partial f(\mathbf{x})$ as derived in Example 3.41, we can write $\partial h(\mathbf{x})$ explicitly as follows:

$$\partial h(\mathbf{x}) = 2\|\mathbf{x}\|_1 \left\{ \mathbf{z} \in \mathbb{R}^n : z_i = \mathrm{sgn}(x_i), i \in I_{\neq}(\mathbf{x}), |z_j| \leq 1, j \in I_0(\mathbf{x}) \right\},$$
where $I_{\neq}(\mathbf{x}) = \{i : x_i \neq 0\}, I_0(\mathbf{x}) = \{i : x_i = 0\}$.

Plugging $\mathbf{x} = \mathbf{0}$ into the above formula, we obtain that

$$\partial h(\mathbf{0}) = \{\mathbf{0}\}.$$

Since h has a unique subgradient at $\mathbf{x} = \mathbf{0}$, it follows by Theorem 3.33 that h is differentiable at $\mathbf{x} = \mathbf{0}$ and $\nabla h(\mathbf{0}) = \mathbf{0}$. Note that the function is obviously not differentiable over $\mathbb{R}^n$. For example, when $n = 2$, the nondifferentiability points are $\{(x_1, 0)^T : x_1 \neq 0\} \cup \{(0, x_2)^T : x_2 \neq 0\}$, as illustrated in Figure 3.3. ∎
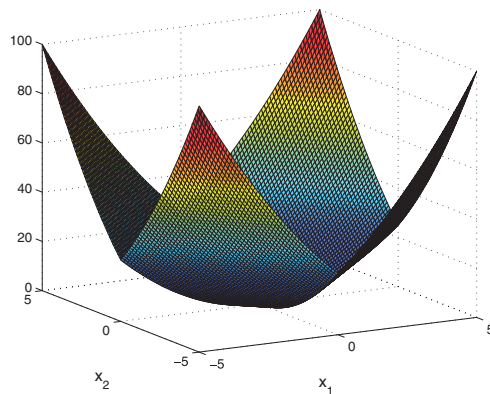
Figure 3.3. *Surface plot of the function $f(x_1, x_2) = (|x_1| + |x_2|)^2$.*

Example 3.49 (subdifferential of $d_C(\cdot)$). Suppose that $\mathbb{E}$ is a Euclidean space, and let $C \subseteq \mathbb{E}$ be a nonempty closed and convex set (see Example 2.20). The distance function d_C is convex. We will show that

$$\partial d_C(\mathbf{x}) = \begin{cases} \left\{ \frac{\mathbf{x} - P_C(\mathbf{x})}{d_C(\mathbf{x})} \right\}, & \mathbf{x} \notin C, \\ N_C(\mathbf{x}) \cap B[\mathbf{0}, 1], & \mathbf{x} \in C. \end{cases}$$

By Example 3.31, we know that the function $\varphi_C(\mathbf{x}) = \frac{1}{2} d_C^2(\mathbf{x})$ is differentiable and

$$\partial \varphi_C(\mathbf{x}) = \{\mathbf{x} - P_C(\mathbf{x})\} \tag{3.44}$$

for any $\mathbf{x} \in \mathbb{E}$. Note that $\varphi_C = g \circ d_C$, where $g(t) = \frac{1}{2}[t]_+^2$ is a nonincreasing real-valued convex differentiable function. Then by the chain rule of subdifferential calculus (Theorem 3.47),

$$\partial \varphi_C(\mathbf{x}) = g'(d_C(\mathbf{x})) \partial d_C(\mathbf{x}) = [d_C(\mathbf{x})]_+ \partial d_C(\mathbf{x}) = d_C(\mathbf{x}) \partial d_C(\mathbf{x}). \tag{3.45}$$

If $\mathbf{x} \notin C$, then $d_C(\mathbf{x}) \neq 0$, and thus by (3.44) and (3.45),

$$\partial d_C(\mathbf{x}) = \left\{ \frac{\mathbf{x} - P_C(\mathbf{x})}{d_C(\mathbf{x})} \right\} \text{ for any } \mathbf{x} \notin C.$$

Since $\partial d_C(\mathbf{x})$ is a singleton for any $\mathbf{x} \notin C$, it follows in particular, by Theorem 3.33, that d_C is differentiable at points outside C.

Now assume that $\mathbf{x} \in C$. We will show that

$$\partial d_C(\mathbf{x}) = N_C(\mathbf{x}) \cap B[\mathbf{0}, 1].$$

Indeed, if $\mathbf{d} \in \partial d_C(\mathbf{x})$, then

$$d_C(\mathbf{y}) \geq \langle \mathbf{d}, \mathbf{y} - \mathbf{x} \rangle \text{ for any } \mathbf{y} \in \mathbb{E}. \tag{3.46}$$

This means in particular that for any $\mathbf{y} \in C$

$$\langle \mathbf{d}, \mathbf{y} - \mathbf{x} \rangle \leq 0,$$

that is, $\mathbf{d} \in N_C(\mathbf{x})$. In addition, taking $\mathbf{y} = \mathbf{x} + \mathbf{d}$ in (3.46), we get

$$\|\mathbf{d}\|^2 = \langle \mathbf{d}, \mathbf{x} + \mathbf{d} - \mathbf{x} \rangle \leq d_C(\mathbf{x} + \mathbf{d}) \leq \|\mathbf{x} + \mathbf{d} - \mathbf{x}\| = \|\mathbf{d}\|,$$

which readily implies that $\|\mathbf{d}\| \leq 1$. We conclude that $\partial d_C(\mathbf{x}) \subseteq N_C(\mathbf{x}) \cap B[\mathbf{0}, 1]$. To show the reverse direction, take $\mathbf{d} \in N_C(\mathbf{x}) \cap B[\mathbf{0}, 1]$. Then for any $\mathbf{y} \in \mathbb{E}$,

$$\langle \mathbf{d}, \mathbf{y} - \mathbf{x} \rangle = \langle \mathbf{d}, \mathbf{y} - P_C(\mathbf{y}) \rangle + \langle \mathbf{d}, P_C(\mathbf{y}) - \mathbf{x} \rangle. \tag{3.47}$$

Since $\mathbf{d} \in N_C(\mathbf{x})$ and $P_C(\mathbf{y}) \in C$, it follows by the definition of the normal cone that $\langle \mathbf{d}, P_C(\mathbf{y}) - \mathbf{x} \rangle \leq 0$, which, combined with (3.47), the Cauchy–Schwarz inequality, and the assertion that $\|\mathbf{d}\| \leq 1$, implies that for any $\mathbf{y} \in \mathbb{E}$

$$\langle \mathbf{d}, \mathbf{y} - \mathbf{x} \rangle \leq \langle \mathbf{d}, \mathbf{y} - P_C(\mathbf{y}) \rangle \leq \|\mathbf{d}\| \cdot \|\mathbf{y} - P_C(\mathbf{y})\| \leq \|\mathbf{y} - P_C(\mathbf{y})\| = d_C(\mathbf{y}),$$

and hence $\mathbf{d} \in \partial d_C(\mathbf{x})$. ∎

3.4.5 Maximization

The following result shows how to compute the subdifferential set of a maximum of a finite collection of convex functions.

Theorem 3.50 (max rule of subdifferential calculus). *Let $f_1, f_2, \ldots, f_m : \mathbb{E} \to (-\infty, \infty]$ be proper convex functions, and define*

$$f(\mathbf{x}) = \max\{f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x})\}.$$

Let $\mathbf{x} \in \bigcap_{i=1}^m \text{int}(\text{dom}(f_i))$. Then

$$\partial f(\mathbf{x}) = \text{conv}\left(\cup_{i \in I(\mathbf{x})} \partial f_i(\mathbf{x}) \right),$$

where $I(\mathbf{x}) = \{i \in \{1, 2, \ldots, m\} : f_i(\mathbf{x}) = f(\mathbf{x})\}$.

Proof. First note that f, as a maximum of convex functions, is convex (see Theorem 2.16(c)) and that by Corollary 3.25 for any $\mathbf{d} \in \mathbb{E}$,

$$f'(\mathbf{x}; \mathbf{d}) = \max_{i \in I(\mathbf{x})} f_i'(\mathbf{x}; \mathbf{d}).$$

For the sake of simplicity of notation, we will assume that $I(\mathbf{x}) = \{1, 2, \ldots, k\}$ for some $k \in \{1, 2, \ldots, m\}$. Now, using the max formula (Theorem 3.26), we obtain

$$f'(\mathbf{x}; \mathbf{d}) = \max_{i=1,2,\ldots,k} f_i'(\mathbf{x}; \mathbf{d}) = \max_{i=1,2,\ldots,k} \max_{\mathbf{g}_i \in \partial f_i(\mathbf{x})} \langle \mathbf{g}_i, \mathbf{d} \rangle. \tag{3.48}$$

Using the fact that for any $a_1, a_2, \ldots, a_k \in \mathbb{R}$ the identity

$$\max\{a_1, a_2, \ldots, a_k\} = \max_{\boldsymbol{\lambda} \in \Delta_k} \sum_{i=1}^{k} \lambda_i a_i$$

holds, we can continue (3.48) and write

$$f'(\mathbf{x}; \mathbf{d}) = \max_{\boldsymbol{\lambda} \in \Delta_k} \left\{ \sum_{i=1}^{k} \lambda_i \max\{\langle \mathbf{g}_i, \mathbf{d} \rangle : \mathbf{g}_i \in \partial f_i(\mathbf{x})\} \right\}$$

$$= \max \left\{ \left\langle \sum_{i=1}^{k} \lambda_i \mathbf{g}_i, \mathbf{d} \right\rangle : \mathbf{g}_i \in \partial f_i(\mathbf{x}), \boldsymbol{\lambda} \in \Delta_k \right\}$$

$$= \max \left\{ \langle \mathbf{g}, \mathbf{d} \rangle : \mathbf{g} \in \mathrm{conv}\left(\cup_{i=1}^{k} \partial f_i(\mathbf{x}) \right) \right\}$$

$$= \sigma_A(\mathbf{d}),$$

where $A = \mathrm{conv}\left(\cup_{i=1}^{k} \partial f_i(\mathbf{x}) \right)$. By the max formula (Theorem 3.26), since $\mathbf{x} \in \mathrm{int}(\mathrm{dom}(f))$,

$$f'(\mathbf{x}; \mathbf{d}) = \sigma_{\partial f(\mathbf{x})}(\mathbf{d}),$$

and hence

$$\sigma_A(\mathbf{d}) = \sigma_{\partial f(\mathbf{x})}(\mathbf{d}) \text{ for any } \mathbf{d} \in \mathbb{E}. \tag{3.49}$$

The set $\partial f(\mathbf{x})$ is closed and convex by Theorem 3.9, and since $\mathbf{x} \in \mathrm{int}(\mathrm{dom}(f))$, it is also nonempty and bounded by Theorem 3.14. Similarly, $\partial f_i(\mathbf{x}), i = 1, 2, \ldots, k$, are nonempty and compact sets, and hence also is $\cup_{i=1}^{k} \partial f_i(\mathbf{x})$. We can conclude that the set $A = \mathrm{conv}(\cup_{i=1}^{k} \partial f_i(\mathbf{x}))$, as a convex hull of a nonempty compact set, is also nonempty and compact.[17] In addition, by the definition of the convex hull, A is convex.

To conclude, both A and $\partial f(\mathbf{x})$ are nonempty closed and convex, and thus (3.49) implies by Lemma 2.34 that

$$\partial f(\mathbf{x}) = A,$$

which is the desired result. $\quad\square$

Example 3.51 (subdifferential of the max function). Let $f : \mathbb{R}^n \to \mathbb{R}$ be given by $f(\mathbf{x}) = \max\{x_1, x_2, \ldots, x_n\}$. Obviously, $f(\mathbf{x}) = \max\{f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_n(\mathbf{x})\}$, where $f_i(\mathbf{x}) = x_i$, and hence $\partial f_i(\mathbf{x}) = \{\mathbf{e}_i\}$ for any $i = 1, 2, \ldots, n$. Denote

$$I(\mathbf{x}) = \{i : f(\mathbf{x}) = x_i\}.$$

---

[17] This follows by [10, Proposition 6.31].

Then by the max rule of subdifferential calculus (Theorem 3.50),

$$\partial f(\mathbf{x}) = \text{conv}(\cup_{i \in I(\mathbf{x})} \partial f_i(\mathbf{x})) = \text{conv}(\cup_{i \in I(\mathbf{x})} \{\mathbf{e}_i\}),$$

and hence

$$\partial f(\mathbf{x}) = \left\{ \sum_{i \in I(\mathbf{x})} \lambda_i \mathbf{e}_i : \sum_{i \in I(\mathbf{x})} \lambda_i = 1, \lambda_j \geq 0, j \in I(\mathbf{x}) \right\}.$$

In particular,

$$\partial f(\alpha \mathbf{e}) = \Delta_n \text{ for any } \alpha \in \mathbb{R}. \quad \blacksquare$$

Example 3.52 (subdifferential of the l_∞-norm). Let $f : \mathbb{R}^n \to \mathbb{R}$ be given by $f(\mathbf{x}) = \|\mathbf{x}\|_\infty$. There are two options. If $\mathbf{x} = \mathbf{0}$, then by Example 3.3 $\partial f(\mathbf{0})$ is the dual-norm unit ball, and in this case,

$$\partial f(\mathbf{0}) = B_{\|\cdot\|_1}[\mathbf{0}, 1] = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_1 \leq 1\}.$$

Suppose that $\mathbf{x} \neq \mathbf{0}$. Note that $f(\mathbf{x}) = \max\{f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_n(\mathbf{x})\}$ with $f_i(\mathbf{x}) = |x_i|$ and set

$$I(\mathbf{x}) = \{i : |x_i| = \|\mathbf{x}\|_\infty\}.$$

For any $i \in I(\mathbf{x})$ we have $x_i \neq 0$, and hence for any such i, $\partial f_i(\mathbf{x}) = \{\text{sgn}(x_i)\mathbf{e}_i\}$. Thus, by the max rule of subdifferential calculus (Theorem 3.50),

$$\begin{aligned} \partial f(\mathbf{x}) &= \text{conv}\left(\cup_{i \in I(\mathbf{x})} \partial f_i(\mathbf{x})\right) \\ &= \text{conv}\left(\cup_{i \in I(\mathbf{x})} \{\text{sgn}(x_i)\mathbf{e}_i\}\right) \\ &= \left\{ \sum_{i \in I(\mathbf{x})} \lambda_i \text{sgn}(x_i)\mathbf{e}_i : \sum_{i \in I(\mathbf{x})} \lambda_i = 1, \lambda_j \geq 0, j \in I(\mathbf{x}) \right\}. \end{aligned}$$

To conclude,

$$\partial f(\mathbf{x}) = \begin{cases} B_{\|\cdot\|_1}[\mathbf{0}, 1], & \mathbf{x} = \mathbf{0}, \\ \left\{ \displaystyle\sum_{i \in I(\mathbf{x})} \lambda_i \text{sgn}(x_i)\mathbf{e}_i : \sum_{i \in I(\mathbf{x})} \lambda_i = 1, \lambda_j \geq 0, j \in I(\mathbf{x}) \right\}, & \mathbf{x} \neq \mathbf{0}. \end{cases}$$

$\blacksquare$

Example 3.53 (subdifferential of piecewise linear functions). Consider the piecewise linear function $f : \mathbb{R}^n \to \mathbb{R}$ given by

$$f(\mathbf{x}) = \max_{i=1,2,\ldots,m} \{\mathbf{a}_i^T \mathbf{x} + b_i\},$$

where $\mathbf{a}_i \in \mathbb{R}^n, b_i \in \mathbb{R}, i = 1, 2, \ldots, m$. The function f can be written as $f(\mathbf{x}) = \max\{f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x})\}$, where $f_i(\mathbf{x}) = \mathbf{a}_i^T \mathbf{x} + b_i, i = 1, 2, \ldots, m$. Obviously, $\partial f_i(\mathbf{x}) = \{\mathbf{a}_i\}$. Thus, by the max rule of subdifferential calculus (Theorem 3.50),

$$\partial f(\mathbf{x}) = \left\{ \sum_{i \in I(\mathbf{x})} \lambda_i \mathbf{a}_i : \sum_{i \in I(\mathbf{x})} \lambda_i = 1, \lambda_j \geq 0, j \in I(\mathbf{x}) \right\},$$

where $I(\mathbf{x}) = \{i : f(\mathbf{x}) = \mathbf{a}_i^T \mathbf{x} + b_i\}$.

∎

Example 3.54 (subdifferential of $\|\mathbf{A}\mathbf{x} + \mathbf{b}\|_\infty$). Consider the function $f : \mathbb{R}^n \to \mathbb{R}$ given by $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} + \mathbf{b}\|_\infty$, where $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$. Then $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x} + \mathbf{b})$, where $g : \mathbb{R}^m \to \mathbb{R}$ is given by $g(\mathbf{y}) = \|\mathbf{y}\|_\infty$. By Example 3.52, we have, for any $\mathbf{y} \in \mathbb{R}^m$,

$$\partial g(\mathbf{y}) = \begin{cases} B_{\|\cdot\|_1}[\mathbf{0}, 1], & \mathbf{y} = \mathbf{0}, \\ \left\{ \sum_{i \in I(\mathbf{y})} \lambda_i \mathrm{sgn}(y_i) \mathbf{e}_i : \sum_{i \in I(\mathbf{y})} \lambda_i = 1, \lambda_j \geq 0, j \in I(\mathbf{y}) \right\}, & \mathbf{y} \neq \mathbf{0}, \end{cases}$$

where

$$I(\mathbf{y}) = \{i \in \{1, 2, \ldots, m\} : |y_i| = \|\mathbf{y}\|_\infty\}.$$

We can thus use the affine transformation rule of subdifferential calculus (Theorem 3.43(b)) to conclude that $\partial f(\mathbf{x}) = \mathbf{A}^T \partial g(\mathbf{A}\mathbf{x} + \mathbf{b})$ is given by

$$\partial f(\mathbf{x}) = \begin{cases} \mathbf{A}^T B_{\|\cdot\|_1}[\mathbf{0}, 1], & \mathbf{A}\mathbf{x} + \mathbf{b} = \mathbf{0}, \\ \left\{ \sum_{i \in I_{\mathbf{x}}} \lambda_i \mathrm{sgn}(\mathbf{a}_i^T \mathbf{x} + b_i) \mathbf{a}_i : \sum_{i \in I_{\mathbf{x}}} \lambda_i = 1, \lambda_j \geq 0, j \in I_{\mathbf{x}} \right\}, & \mathbf{A}\mathbf{x} + \mathbf{b} \neq \mathbf{0}, \end{cases}$$

where $\mathbf{a}_1^T, \mathbf{a}_2^T, \ldots, \mathbf{a}_m^T$ are the rows of $\mathbf{A}$ and $I_{\mathbf{x}} = I(\mathbf{A}\mathbf{x} + \mathbf{b})$. ∎

When the index set is arbitrary (for example, infinite), it is still possible to prove a weak subdifferential calculus rule.

Theorem 3.55 (weak maximum rule of subdifferential calculus). *Let I be an arbitrary set, and suppose that any $i \in I$ is associated with a proper convex function $f_i : \mathbb{E} \to (-\infty, \infty]$. Let*

$$f(\mathbf{x}) = \max_{i \in I} f_i(\mathbf{x}). \tag{3.50}$$

Then for any $\mathbf{x} \in \mathrm{dom}(f)$

$$\mathrm{conv}\left(\cup_{i \in I(\mathbf{x})} \partial f_i(\mathbf{x}) \right) \subseteq \partial f(\mathbf{x}), \tag{3.51}$$

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

Proof. Let $\mathbf{x} \in \mathrm{dom}(f)$. Then for any $\mathbf{z} \in \mathrm{dom}(f), i \in I(\mathbf{x})$ and $\mathbf{g} \in \partial f_i(\mathbf{x})$,

$$f(\mathbf{z}) \geq f_i(\mathbf{z}) \geq f_i(\mathbf{x}) + \langle \mathbf{g}, \mathbf{z} - \mathbf{x} \rangle = f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{z} - \mathbf{x} \rangle, \tag{3.52}$$

where the first inequality follows from (3.50), the second inequality is the subgradient inequality, and the equality is due to the assertion that $i \in I(\mathbf{x})$. Since (3.52) holds for any $\mathbf{z} \in \text{dom}(f)$, we can conclude that $\mathbf{g} \in \partial f(\mathbf{x})$. Thus, $\partial f_i(\mathbf{x}) \subseteq \partial f(\mathbf{x})$. Finally, by the convexity of $\partial f(\mathbf{x})$ (Theorem 3.9), the result (3.51) follows. $\square$

Example 3.56 (subgradient of $\lambda_{\max}(\mathbf{A}_0 + \sum_{i=1}^{m} x_i \mathbf{A}_i)$). Let $\mathbf{A}_0, \mathbf{A}_1, \ldots, \mathbf{A}_m$ $\in \mathbb{S}^n$. Let $\mathcal{A} : \mathbb{R}^m \to \mathbb{S}^n$ be the affine transformation given by

$$\mathcal{A}(\mathbf{x}) = \mathbf{A}_0 + \sum_{i=1}^{m} x_i \mathbf{A}_i \text{ for any } \mathbf{x} \in \mathbb{R}^m.$$

Consider the function $f : \mathbb{R}^m \to \mathbb{R}$ given by $f(\mathbf{x}) = \lambda_{\max}(\mathcal{A}(\mathbf{x}))$. Since for any $\mathbf{x} \in \mathbb{R}^m$,

$$f(\mathbf{x}) = \max_{\mathbf{y} \in \mathbb{R}^n : \|\mathbf{y}\|_2 = 1} \mathbf{y}^T \mathcal{A}(\mathbf{x}) \mathbf{y}, \tag{3.53}$$

and since the function

$$f_{\mathbf{y}}(\mathbf{x}) \equiv \mathbf{y}^T \mathcal{A}(\mathbf{x}) \mathbf{y} = \mathbf{y}^T \mathbf{A}_0 \mathbf{y} + \sum_{i=1}^{m} (\mathbf{y}^T \mathbf{A}_i \mathbf{y}) x_i$$

is affine in $\mathbf{x}$, and in particular convex in $\mathbf{x}$, it follows by Theorem 2.16(c) that f is convex. For a given $\mathbf{x} \in \mathbb{R}^n$, the maximum in (3.53) is attained at normalized eigenvectors which correspond to the maximum eigenvalue of $\mathcal{A}(\mathbf{x})$. Let $\tilde{\mathbf{y}}$ be such a normalized eigenvector. Then it follows by the weak maximum rule of subdifferential calculus (Theorem 3.55) that a subgradient of the affine function $f_{\tilde{\mathbf{y}}}$ at $\mathbf{x}$ is a subgradient of f at $\mathbf{x}$, that is,

$$(\tilde{\mathbf{y}}^T \mathbf{A}_1 \tilde{\mathbf{y}}, \tilde{\mathbf{y}}^T \mathbf{A}_2 \tilde{\mathbf{y}}, \ldots, \tilde{\mathbf{y}}^T \mathbf{A}_m \tilde{\mathbf{y}})^T \in \partial f(\mathbf{x}), \tag{3.54}$$

where $\tilde{\mathbf{y}}$ is a normalized eigenvector of $\mathcal{A}(\mathbf{x})$ corresponding to the maximum eigenvalue.

It is interesting to note that the result (3.54) can also be deduced by the affine transformation rule of subdifferential calculus (Theorem 3.43(b)). Indeed, let $\tilde{\mathbf{y}}$ be as defined above. The function f can be written as $f(\mathbf{x}) = g(\mathcal{B}(\mathbf{x}) + \mathbf{A}_0)$, where $\mathcal{B}(\mathbf{x}) \equiv \sum_{i=1}^{m} x_i \mathbf{A}_i$ and $g(\mathbf{X}) \equiv \lambda_{\max}(\mathbf{X})$. Then by the affine transformation rule of subdifferential calculus,

$$\partial f(\mathbf{x}) = \mathcal{B}^T(\partial g(\mathcal{B}(\mathbf{x}) + \mathbf{A}_0)). \tag{3.55}$$

By Example 3.8, we know that $\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T \in \partial g(\mathcal{B}(\mathbf{x}) + \mathbf{A}_0)$, and hence, by (3.55),

$$\mathcal{B}^T(\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T) \in \partial f(\mathbf{x}).$$

The result now follows by noting that

$$\mathcal{B}^T(\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T) = (\text{Tr}(\mathbf{A}_1 \tilde{\mathbf{y}}\tilde{\mathbf{y}}^T), \text{Tr}(\mathbf{A}_2 \tilde{\mathbf{y}}\tilde{\mathbf{y}}^T), \ldots, \text{Tr}(\mathbf{A}_m \tilde{\mathbf{y}}\tilde{\mathbf{y}}^T))^T$$
$$= (\tilde{\mathbf{y}}^T \mathbf{A}_1 \tilde{\mathbf{y}}, \tilde{\mathbf{y}}^T \mathbf{A}_2 \tilde{\mathbf{y}}, \ldots, \tilde{\mathbf{y}}^T \mathbf{A}_m \tilde{\mathbf{y}})^T. \quad \blacksquare$$

3.5 The Value Function[18]

Consider the minimization problem

$$f_{\text{opt}} = \min_{\mathbf{x} \in X}\{f(\mathbf{x}) : g_i(\mathbf{x}) \leq 0, i = 1, 2, \ldots, m, \mathbf{A}\mathbf{x} + \mathbf{b} = \mathbf{0}\}, \qquad (3.56)$$

where $f, g_1, g_2, \ldots, g_m : \mathbb{E} \to (-\infty, \infty]$ are extended real-valued functions, $X \subseteq \mathbb{E}$ a nonempty set, $\mathbf{A} \in \mathbb{R}^{p \times n}$, and $\mathbf{b} \in \mathbb{R}^p$. We will define the vector-valued function $\mathbf{g} : \mathbb{E} \to \mathbb{R}^m$ as

$$\mathbf{g}(\mathbf{x}) \equiv (g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_m(\mathbf{x}))^T,$$

so that problem (3.56) can be rewritten more compactly as

$$\min_{\mathbf{x} \in X}\{f(\mathbf{x}) : \mathbf{g}(\mathbf{x}) \leq \mathbf{0}, \mathbf{A}\mathbf{x} + \mathbf{b} = \mathbf{0}\}.$$

The *value function* associated with problem (3.56) is the function $v : \mathbb{R}^m \times \mathbb{R}^p \to [-\infty, \infty]$ given by

$$v(\mathbf{u}, \mathbf{t}) = \min_{\mathbf{x} \in X}\{f(\mathbf{x}) : g(\mathbf{x}) \leq \mathbf{u}, \mathbf{A}\mathbf{x} + \mathbf{b} = \mathbf{t}\}. \qquad (3.57)$$

The feasible set of the minimization problem in (3.57) will be denoted by

$$C(\mathbf{u}, \mathbf{t}) = \{\mathbf{x} \in X : g(\mathbf{x}) \leq \mathbf{u}, \mathbf{A}\mathbf{x} + \mathbf{b} = \mathbf{t}\},$$

so that the value function can also be rewritten as $v(\mathbf{u}, \mathbf{t}) = \min\{f(\mathbf{x}) : \mathbf{x} \in C(\mathbf{u}, \mathbf{t})\}$. By convention $v(\mathbf{u}, \mathbf{t}) = \infty$ if $C(\mathbf{u}, \mathbf{t})$ is empty. A simple property of the value function $v(\cdot, \cdot)$ is that it is monotone w.r.t. its first argument.

Lemma 3.57 (monotonicity of the value function). *Let $f, g_1, g_2, \ldots, g_m : \mathbb{E} \to (-\infty, \infty]$ be extended real-valued functions, $X \subseteq \mathbb{E}$ a nonempty set, $\mathbf{A} \in \mathbb{R}^{p \times n}$, and $\mathbf{b} \in \mathbb{R}^p$. Let v be the value function given in (3.57). Then*

$$v(\mathbf{u}, \mathbf{t}) \geq v(\mathbf{w}, \mathbf{t}) \text{ for any } \mathbf{u}, \mathbf{w} \in \mathbb{R}^m, \mathbf{t} \in \mathbb{R}^p \text{ satisfying } \mathbf{u} \leq \mathbf{w}.$$

Proof. Follows by the obvious fact that $C(\mathbf{u}, \mathbf{t}) \subseteq C(\mathbf{w}, \mathbf{t})$ whenever $\mathbf{u} \leq \mathbf{w}$. $\square$

From now on we will also assume in addition that $f, g_1, g_2, \ldots, g_m$, and X are convex. With these additional assumptions, we now show that the value function is convex as long as it is proper.

Lemma 3.58 (convexity of the value function). *Let $f, g_1, g_2, \ldots, g_m : \mathbb{E} \to (-\infty, \infty]$ be convex functions, $X \subseteq \mathbb{E}$ a nonempty convex set, $\mathbf{A} \in \mathbb{R}^{p \times n}$, and $\mathbf{b} \in \mathbb{R}^p$. Suppose that the value function v given in (3.57) is proper. Then v is convex over $\mathbb{R}^m \times \mathbb{R}^p$.*

Proof. Let $(\mathbf{u}, \mathbf{t}), (\mathbf{w}, \mathbf{s}) \in \text{dom}(v)$ and $\lambda \in [0, 1]$. Since v is proper, to prove the convexity, we need to show that

$$v(\lambda\mathbf{u} + (1 - \lambda)\mathbf{w}, \lambda\mathbf{t} + (1 - \lambda)\mathbf{s}) \leq \lambda v(\mathbf{u}, \mathbf{t}) + (1 - \lambda)v(\mathbf{w}, \mathbf{s}).$$

---

[18]Section 3.5, excluding Theorem 3.60, follows Hiriart-Urruty and Lemaréchal [67, Section VII.3.3].

By the definition of the value function v, there exist sequences $\{\mathbf{x}_k\}_{k \geq 1}, \{\mathbf{y}_k\}_{k \geq 1}$ satisfying

$$\mathbf{x}_k \in C(\mathbf{u}, \mathbf{t}), \mathbf{y}_k \in C(\mathbf{w}, \mathbf{s}), f(\mathbf{x}_k) \to v(\mathbf{u}, \mathbf{t}), f(\mathbf{y}_k) \to v(\mathbf{w}, \mathbf{s}) \text{ as } k \to \infty.$$

Since $\mathbf{x}_k \in C(\mathbf{u}, \mathbf{t})$ and $\mathbf{y}_k \in C(\mathbf{w}, \mathbf{s})$, we have $\mathbf{g}(\mathbf{x}_k) \leq \mathbf{u}, \mathbf{g}(\mathbf{y}_k) \leq \mathbf{w}$. Therefore, by the convexity of the components of $\mathbf{g}$,

$$\mathbf{g}(\lambda \mathbf{x}_k + (1 - \lambda)\mathbf{y}_k) \leq \lambda \mathbf{g}(\mathbf{x}_k) + (1 - \lambda)\mathbf{g}(\mathbf{y}_k) \leq \lambda \mathbf{u} + (1 - \lambda)\mathbf{w}. \tag{3.58}$$

Moreover,

$$\mathbf{A}(\lambda \mathbf{x}_k + (1 - \lambda)\mathbf{y}_k) + \mathbf{b} = \lambda(\mathbf{A}\mathbf{x}_k + \mathbf{b}) + (1 - \lambda)(\mathbf{A}\mathbf{y}_k + \mathbf{b}) = \lambda \mathbf{s} + (1 - \lambda)\mathbf{t}. \tag{3.59}$$

Combining (3.58) and (3.59), we conclude that

$$\lambda \mathbf{x}_k + (1 - \lambda)\mathbf{y}_k \in C(\lambda \mathbf{u} + (1 - \lambda)\mathbf{w}, \lambda \mathbf{s} + (1 - \lambda)\mathbf{t}). \tag{3.60}$$

By the convexity of f,

$$f(\lambda \mathbf{x}_k + (1 - \lambda)\mathbf{y}_k) \leq \lambda f(\mathbf{x}_k) + (1 - \lambda)f(\mathbf{y}_k). \tag{3.61}$$

Since $\lambda f(\mathbf{x}_k) + (1 - \lambda)f(\mathbf{y}_k) \to \lambda v(\mathbf{u}, \mathbf{t}) + (1 - \lambda)v(\mathbf{w}, \mathbf{s})$ as $k \to \infty$, by (3.61) we have

$$\liminf_{k \to \infty} f(\lambda \mathbf{x}_k + (1 - \lambda)\mathbf{y}_k) \leq \lambda v(\mathbf{u}, \mathbf{t}) + (1 - \lambda)v(\mathbf{w}, \mathbf{s}). \tag{3.62}$$

Finally, since (3.60) holds, by the definition of v, for all k,

$$v(\lambda \mathbf{u} + (1 - \lambda)\mathbf{w}, \lambda \mathbf{t} + (1 - \lambda)\mathbf{s}) \leq f(\lambda \mathbf{x}_k + (1 - \lambda)\mathbf{y}_k),$$

and hence

$$v(\lambda \mathbf{u} + (1 - \lambda)\mathbf{w}, \lambda \mathbf{t} + (1 - \lambda)\mathbf{s}) \leq \liminf_{k \to \infty} f(\lambda \mathbf{x}_k + (1 - \lambda)\mathbf{y}_k),$$

which, combined with (3.62), yields the inequality

$$v(\lambda \mathbf{u} + (1 - \lambda)\mathbf{w}, \lambda \mathbf{t} + (1 - \lambda)\mathbf{s}) \leq \lambda v(\mathbf{u}, \mathbf{t}) + (1 - \lambda)v(\mathbf{w}, \mathbf{s}),$$

establishing the convexity of v. $\square$

The dual objective function $q : \mathbb{R}_+^m \times \mathbb{R}^q \to [-\infty, \infty)$ of problem (3.56) is

$$q(\mathbf{y}, \mathbf{z}) = \min_{\mathbf{x} \in X} \left\{ L(\mathbf{x}; \mathbf{y}, \mathbf{z}) = f(\mathbf{x}) + \mathbf{y}^T \mathbf{g}(\mathbf{x}) + \mathbf{z}^T (\mathbf{A}\mathbf{x} + \mathbf{b}) \right\}, \mathbf{y} \in \mathbb{R}_+^m, \mathbf{z} \in \mathbb{R}^p.$$

The dual problem consists of maximizing q on its effective domain given by

$$\text{dom}(-q) = \{(\mathbf{y}, \mathbf{z}) \in \mathbb{R}_+^m \times \mathbb{R}^p : q(\mathbf{y}, \mathbf{z}) > -\infty\}.$$

The dual problem

$$q_{\text{opt}} = \max_{\mathbf{y} \in \mathbb{R}_+^m, \mathbf{z} \in \mathbb{R}^p} \{q(\mathbf{y}, \mathbf{z}) : (\mathbf{y}, \mathbf{z}) \in \text{dom}(-q)\} \tag{3.63}$$

is convex in the sense that it consists of maximizing the concave function q over the convex feasible set $\mathrm{dom}(-q)$. We are now ready to show the main result of this section, which is a relation between the subdifferential set of the value function at the zeros vector and the set of optimal solutions of the dual problem. The result is established under the assumption that strong duality holds, meaning under the assumptions that the optimal values of the primal and dual problems are finite and equal ($f_{\mathrm{opt}} = q_{\mathrm{opt}}$) and the optimal set of the dual problem is nonempty. By the strong duality theorem stated as Theorem A.1 in the appendix, it follows that these assumptions are met if the optimal value of problem (3.56) is finite, and if there exists a feasible solution $\bar{\mathbf{x}}$ satisfying $\mathbf{g}(\bar{\mathbf{x}}) < \mathbf{0}$ and a vector $\hat{\mathbf{x}} \in \mathrm{ri}(X)$ satisfying $\mathbf{A}\hat{\mathbf{x}} + \mathbf{b} = \mathbf{0}$.

Theorem 3.59 (characterization of the subdifferential of the value function at 0). *Let $f, g_1, g_2, \ldots, g_m : \mathbb{E} \to (-\infty, \infty]$ be convex functions, $X \subseteq \mathbb{E}$ a nonempty convex set, $\mathbf{A} \in \mathbb{R}^{p \times n}$, and $\mathbf{b} \in \mathbb{R}^p$. Let v be the value function given by (3.57). Suppose that $f_{\mathrm{opt}} = q_{\mathrm{opt}} \in (-\infty, \infty)$ and that the optimal set of the dual problem is nonempty. Then*

(a) *v is proper and convex;*

(b) *$(\mathbf{y}, \mathbf{z})$ is an optimal solution of problem (3.63) if and only if $-(\mathbf{y}, \mathbf{z}) \in \partial v(\mathbf{0}, \mathbf{0})$.*

Proof. Let $(\mathbf{y}, \mathbf{z}) \in \mathrm{dom}(-q)$ be an optimal solution of the dual problem. Then (recalling that $v(\mathbf{0}, \mathbf{0}) = f_{\mathrm{opt}}$)

$$L(\mathbf{x}; \mathbf{y}, \mathbf{z}) \geq \min_{\mathbf{w} \in X} L(\mathbf{w}; \mathbf{y}, \mathbf{z}) = q(\mathbf{y}, \mathbf{z}) = q_{\mathrm{opt}} = f_{\mathrm{opt}} = v(\mathbf{0}, \mathbf{0}) \text{ for all } \mathbf{x} \in X.$$

Therefore, for any $\mathbf{x} \in C(\mathbf{u}, \mathbf{t})$,

$$\begin{aligned}
v(\mathbf{0}, \mathbf{0}) - \mathbf{y}^T \mathbf{u} - \mathbf{z}^T \mathbf{t} &\leq L(\mathbf{x}; \mathbf{y}, \mathbf{z}) - \mathbf{y}^T \mathbf{u} - \mathbf{z}^T \mathbf{t} \\
&= f(\mathbf{x}) + \mathbf{y}^T \mathbf{g}(\mathbf{x}) + \mathbf{z}^T (\mathbf{A}\mathbf{x} + \mathbf{b}) - \mathbf{y}^T \mathbf{u} - \mathbf{z}^T \mathbf{t} \\
&= f(\mathbf{x}) + \mathbf{y}^T (\mathbf{g}(\mathbf{x}) - \mathbf{u}) + \mathbf{z}^T (\mathbf{A}\mathbf{x} + \mathbf{b} - \mathbf{t}) \\
&\leq f(\mathbf{x}),
\end{aligned}$$

where the last inequality follows from the facts that $\mathbf{g}(\mathbf{x}) \leq \mathbf{u}, \mathbf{y} \geq \mathbf{0}$, and $\mathbf{A}\mathbf{x} + \mathbf{b} = \mathbf{t}$. We thus obtained the bound

$$f(\mathbf{x}) \geq v(\mathbf{0}, \mathbf{0}) - \mathbf{y}^T \mathbf{u} - \mathbf{z}^T \mathbf{t} \text{ for any } \mathbf{x} \in C(\mathbf{u}, \mathbf{t}).$$

Minimizing the left-hand side w.r.t. $\mathbf{x} \in C(\mathbf{u}, \mathbf{t})$ yields

$$v(\mathbf{u}, \mathbf{t}) \geq v(\mathbf{0}, \mathbf{0}) - \mathbf{y}^T \mathbf{u} - \mathbf{z}^T \mathbf{t}, \tag{3.64}$$

which is equivalent to saying that $-(\mathbf{y}, \mathbf{z}) \in \partial v(\mathbf{0}, \mathbf{0})$. We actually showed one direction of claim (b), as well as the properness of v since by (3.64), $v(\mathbf{u}, \mathbf{t}) > -\infty$ for any $(\mathbf{u}, \mathbf{t}) \in \mathbb{R}^m \times \mathbb{R}^p$, and by the premise of the theorem, $v(\mathbf{0}, \mathbf{0}) = f_{\mathrm{opt}} < \infty$. Invoking Lemma 3.58, it follows that v is convex, establishing claim (a).

All that is left is to show the reverse direction of claim (b). Assume that $-(\mathbf{y}, \mathbf{z}) \in \partial v(\mathbf{0}, \mathbf{0})$, meaning that

$$v(\mathbf{u}, \mathbf{t}) \geq v(\mathbf{0}, \mathbf{0}) - \mathbf{y}^T \mathbf{u} - \mathbf{z}^T \mathbf{t} \text{ for any } (\mathbf{u}, \mathbf{t}) \in \mathbb{R}^m \times \mathbb{R}^p. \tag{3.65}$$

Let $\mathbf{x} \in X$. Then

$$f(\mathbf{x}) \geq v(\mathbf{g}(\mathbf{x}), \mathbf{A}\mathbf{x} + \mathbf{b}) \overset{(3.65)}{\geq} v(\mathbf{0}, \mathbf{0}) - \mathbf{y}^T \mathbf{g}(\mathbf{x}) - \mathbf{z}^T(\mathbf{A}\mathbf{x} + \mathbf{b}).$$

Therefore,

$$v(\mathbf{0}, \mathbf{0}) \leq f(\mathbf{x}) + \mathbf{y}^T \mathbf{g}(\mathbf{x}) + \mathbf{z}^T(\mathbf{A}\mathbf{x} + \mathbf{b}) = L(\mathbf{x}; \mathbf{y}, \mathbf{z}) \text{ for any } \mathbf{x} \in X.$$

Minimizing the right-hand side w.r.t. $\mathbf{x} \in X$ yields

$$v(\mathbf{0}, \mathbf{0}) \leq \min_{\mathbf{x} \in X} L(\mathbf{x}; \mathbf{y}, \mathbf{z}) = q(\mathbf{y}, \mathbf{z}). \tag{3.66}$$

Let $j \in \{1, 2, \ldots, m\}$. Plugging $\mathbf{u} = \mathbf{e}_j, \mathbf{t} = \mathbf{0}$ into (3.65), we obtain

$$y_j \geq v(\mathbf{0}, \mathbf{0}) - v(\mathbf{e}_j, \mathbf{0}) \geq 0,$$

where the second inequality follows from the monotonicity property of the value function stated in Lemma 3.57. We thus obtained that $\mathbf{y} \geq \mathbf{0}$, and we can consequently write using (3.66)

$$q_{\text{opt}} = f_{\text{opt}} = v(\mathbf{0}, \mathbf{0}) \leq q(\mathbf{y}, \mathbf{z}) \leq q_{\text{opt}},$$

showing that $q(\mathbf{y}, \mathbf{z}) = q_{\text{opt}}$, meaning that $(\mathbf{y}, \mathbf{z})$ is an optimal solution of the dual problem. $\square$

Theorem 3.59 can be used to prove a result concerning an optimality measure of problem (3.56). Consider the following expression:

$$D(\mathbf{x}) \equiv f(\mathbf{x}) - f_{\text{opt}} + \rho_1 \|[\mathbf{g}(\mathbf{x})]_+\|_2 + \rho_2 \|\mathbf{A}\mathbf{x} + \mathbf{b}\|_2.$$

Now assume that

$$D(\tilde{\mathbf{x}}) \leq \delta \tag{3.67}$$

for some $\tilde{\mathbf{x}} \in X$ and a small $\delta > 0$. The question that now arises is whether (3.67) implies that the expressions $f(\tilde{\mathbf{x}}) - f_{\text{opt}}$ as well as $\|[\mathbf{g}(\tilde{\mathbf{x}})]_+\|_2$ and $\|\mathbf{A}\tilde{\mathbf{x}} + \mathbf{b}\|_2$ are also "small" in the sense that they are smaller than a constant times δ. In general, the answer is no. The vector $\tilde{\mathbf{x}}$ is not guaranteed to be feasible, and therefore, in principle, $f(\tilde{\mathbf{x}}) - f_{\text{opt}}$ might be very small (and negative), and $\|[\mathbf{g}(\tilde{\mathbf{x}})]_+\|_2$ and $\|\mathbf{A}\tilde{\mathbf{x}} + \mathbf{b}\|_2$ can be very large. However, we will show in the next theorem that if ρ_1 and ρ_2 are chosen to be large enough, then under the setting of Theorem 3.59, such a conclusion can be drawn.

Theorem 3.60. [19] *Let $f, g_1, g_2, \ldots, g_m : \mathbb{E} \to (-\infty, \infty]$ be convex functions, $X \subseteq \mathbb{E}$ a nonempty convex set, $\mathbf{A} \in \mathbb{R}^{p \times n}$, and $\mathbf{b} \in \mathbb{R}^p$. Let f_{opt} and q_{opt} be the optimal values of the primal and dual problems (3.56) and (3.63), respectively. Suppose that $f_{\text{opt}} = q_{\text{opt}} \in (-\infty, \infty)$ and that the optimal set of the dual problem is nonempty. Let $(\mathbf{y}^*, \mathbf{z}^*)$ be an optimal solution of the dual problem. Assume that $\tilde{\mathbf{x}} \in X$ satisfies*

$$f(\tilde{\mathbf{x}}) - f_{\text{opt}} + \rho_1 \|[\mathbf{g}(\tilde{\mathbf{x}})]_+\|_2 + \rho_2 \|\mathbf{A}\tilde{\mathbf{x}} + \mathbf{b}\|_2 \leq \delta, \tag{3.68}$$

---

[19]Theorem 3.60 is a slight extension of Lemma 6 from Lan [78].

where $\delta > 0$ and ρ_1, ρ_2 are constants satisfying $\rho_1 \geq 2\|\mathbf{y}^*\|_2, \rho_2 \geq 2\|\mathbf{z}^*\|_2$. Then

$$f(\tilde{\mathbf{x}}) - f_{\mathrm{opt}} \leq \delta,$$

$$\|[\mathbf{g}(\tilde{\mathbf{x}})]_+\|_2 \leq \frac{2}{\rho_1}\delta,$$

$$\|\mathbf{A}\tilde{\mathbf{x}} + \mathbf{b}\|_2 \leq \frac{2}{\rho_2}\delta.$$

Proof. The inequality $f(\tilde{\mathbf{x}}) - f_{\mathrm{opt}} \leq \delta$ trivially follows from (3.68) and the fact that the expressions $\rho_1\|[\mathbf{g}(\tilde{\mathbf{x}})]_+\|_2$ and $\rho_2\|\mathbf{A}\tilde{\mathbf{x}} + \mathbf{b}\|_2$ are nonnegative.

Define the function

$$v(\mathbf{u}, \mathbf{t}) = \min_{\mathbf{x} \in X}\{f(\mathbf{x}) : \mathbf{g}(\mathbf{x}) \leq \mathbf{u}, \mathbf{A}\mathbf{x} + \mathbf{b} = \mathbf{t}\}.$$

Since $(\mathbf{y}^*, \mathbf{z}^*)$ is an optimal solution of the dual problem, it follows by Theorem 3.59 that $(-\mathbf{y}^*, -\mathbf{z}^*) \in \partial v(\mathbf{0}, \mathbf{0})$. Therefore, for any $(\mathbf{u}, \mathbf{t}) \in \mathrm{dom}(v)$,

$$v(\mathbf{u}, \mathbf{t}) - v(\mathbf{0}, \mathbf{0}) \geq \langle -\mathbf{y}^*, \mathbf{u}\rangle + \langle -\mathbf{z}^*, \mathbf{t}\rangle. \tag{3.69}$$

Plugging $\mathbf{u} = \tilde{\mathbf{u}} \equiv [\mathbf{g}(\tilde{\mathbf{x}})]_+$ and $\mathbf{t} = \tilde{\mathbf{t}} \equiv \mathbf{A}\tilde{\mathbf{x}} + \mathbf{b}$ into (3.69), while using the inequality $v(\tilde{\mathbf{u}}, \tilde{\mathbf{t}}) \leq f(\tilde{\mathbf{x}})$ and the equality $v(\mathbf{0}, \mathbf{0}) = f_{\mathrm{opt}}$, we obtain

$$\begin{aligned}
(\rho_1 - \|\mathbf{y}^*\|_2)\|\tilde{\mathbf{u}}\|_2 + (\rho_2 - \|\mathbf{z}^*\|_2)\|\tilde{\mathbf{t}}\|_2 &= -\|\mathbf{y}^*\|_2\|\tilde{\mathbf{u}}\|_2 - \|\mathbf{z}^*\|_2\|\tilde{\mathbf{t}}\|_2 + \rho_1\|\tilde{\mathbf{u}}\|_2 + \rho_2\|\tilde{\mathbf{t}}\|_2 \\
&\leq \langle -\mathbf{y}^*, \tilde{\mathbf{u}}\rangle + \langle -\mathbf{z}^*, \tilde{\mathbf{t}}\rangle + \rho_1\|\tilde{\mathbf{u}}\|_2 + \rho_2\|\tilde{\mathbf{t}}\|_2 \\
&\leq v(\tilde{\mathbf{u}}, \tilde{\mathbf{t}}) - v(\mathbf{0}, \mathbf{0}) + \rho_1\|\tilde{\mathbf{u}}\|_2 + \rho_2\|\tilde{\mathbf{t}}\|_2 \\
&\leq f(\tilde{\mathbf{x}}) - f_{\mathrm{opt}} + \rho_1\|\tilde{\mathbf{u}}\|_2 + \rho_2\|\tilde{\mathbf{t}}\|_2 \\
&\leq \delta.
\end{aligned}$$

Therefore, since both expressions $(\rho_1 - \|\mathbf{y}^*\|_2)\|\tilde{\mathbf{u}}\|_2$ and $(\rho_2 - \|\mathbf{z}^*\|_2)\|\tilde{\mathbf{t}}\|_2$ are nonnegative, it follows that

$$(\rho_1 - \|\mathbf{y}^*\|_2)\|\tilde{\mathbf{u}}\|_2 \leq \delta,$$

$$(\rho_2 - \|\mathbf{z}^*\|_2)\|\tilde{\mathbf{t}}\|_2 \leq \delta,$$

and hence, using the assumptions that $\rho_1 \geq 2\|\mathbf{y}^*\|_2$ and $\rho_2 \geq 2\|\mathbf{t}^*\|_2$,

$$\|[\mathbf{g}(\tilde{\mathbf{x}})]_+\|_2 = \|\tilde{\mathbf{u}}\|_2 \leq \frac{\delta}{\rho_1 - \|\mathbf{y}^*\|_2} \leq \frac{2}{\rho_1}\delta,$$

$$\|\mathbf{A}\tilde{\mathbf{x}} + \mathbf{b}\|_2 = \|\tilde{\mathbf{t}}\|_2 \leq \frac{\delta}{\rho_2 - \|\mathbf{z}^*\|_2} \leq \frac{2}{\rho_2}\delta. \qquad \square$$

3.6 Lipschitz Continuity and Boundedness of Subgradients

This section considers an important relation between Lipschitz continuity of a convex function and boundedness of its subgradients.

Theorem 3.61 (Lipschitz continuity and boundedness of the subdifferential sets). *Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper and convex function. Suppose that $X \subseteq \mathrm{int}(\mathrm{dom}f)$. Consider the following two claims:*

(i) $|f(\mathbf{x}) - f(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|$ *for any* $\mathbf{x}, \mathbf{y} \in X$.

(ii) $\|\mathbf{g}\|_* \leq L$ *for any* $\mathbf{g} \in \partial f(\mathbf{x}), \mathbf{x} \in X$.

Then

(a) *the implication* (ii) $\Rightarrow$ (i) *holds;*

(b) *if X is open, then* (i) *holds if and only if* (ii) *holds.*

Proof. (a) Suppose that (ii) is satisfied and let $\mathbf{x}, \mathbf{y} \in X$. Let $\mathbf{g_x} \in \partial f(\mathbf{x})$ and $\mathbf{g_y} \in \partial f(\mathbf{y})$. The existence of these subgradients is guaranteed by Theorem 3.14. Then by the definitions of $\mathbf{g_x}, \mathbf{g_y}$ and the generalized Cauchy–Schwarz inequality (Lemma 1.4),

$$f(\mathbf{x}) - f(\mathbf{y}) \leq \langle \mathbf{g_x}, \mathbf{x} - \mathbf{y} \rangle \leq \|\mathbf{g_x}\|_*\|\mathbf{x} - \mathbf{y}\| \leq L\|\mathbf{x} - \mathbf{y}\|,$$
$$f(\mathbf{y}) - f(\mathbf{x}) \leq \langle \mathbf{g_y}, \mathbf{y} - \mathbf{x} \rangle \leq \|\mathbf{g_y}\|_*\|\mathbf{x} - \mathbf{y}\| \leq L\|\mathbf{x} - \mathbf{y}\|,$$

showing the validity of (i).

(b) The implication (ii) $\Rightarrow$ (i) was already shown. Now assume that (i) is satisfied. Take $\mathbf{x} \in X$ and $\mathbf{g} \in \partial f(\mathbf{x})$. We will show that $\|\mathbf{g}\|_* \leq L$. Define $\mathbf{g}^\dagger \in \mathbb{E}$ as a vector that satisfies $\|\mathbf{g}^\dagger\| = 1, \langle \mathbf{g}^\dagger, \mathbf{g} \rangle = \|\mathbf{g}\|_*$ (the existence of such a vector is warranted by the definition of the dual norm). Take $\varepsilon > 0$ small enough such that $\mathbf{x} + \varepsilon\mathbf{g}^\dagger \in X$. By the subgradient inequality we have

$$f(\mathbf{x} + \varepsilon\mathbf{g}^\dagger) \geq f(\mathbf{x}) + \langle \mathbf{g}, \varepsilon\mathbf{g}^\dagger \rangle.$$

Thus,

$$\varepsilon\|\mathbf{g}\|_* = \langle \mathbf{g}, \varepsilon\mathbf{g}^\dagger \rangle \leq f(\mathbf{x} + \varepsilon\mathbf{g}^\dagger) - f(\mathbf{x}) \leq L\|\mathbf{x} + \varepsilon\mathbf{g}^\dagger - \mathbf{x}\| = L\varepsilon,$$

showing that $\|\mathbf{g}\|_* \leq L$. $\square$

Recall that by Theorem 3.16, the subgradients of a given convex function f are bounded over compact sets contained in $\mathrm{int}(\mathrm{dom}(f))$. Combining this with Theorem 3.61, we can conclude that convex functions are always Lipschitz continuous over compact sets contained in the interior of their domain.

Corollary 3.62 (Lipschitz continuity of convex functions over compact domains). *Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper and convex function. Suppose that $X \subseteq \mathrm{int}(\mathrm{dom}(f))$ is compact. Then there exists $L > 0$ for which*

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\| \ \text{for any} \ \mathbf{x}, \mathbf{y} \in X.$$

3.7   Optimality Conditions

3.7.1   Fermat's Optimality Condition

Subdifferential sets are extremely useful in characterizing minima points. Perhaps the most basic optimality condition states that a point is a global minimum of a

proper extended real-valued convex function if and only if $\mathbf{0}$ belongs to the subdifferential set at the point. In a sense, this is a generalization of Fermat's optimality condition at points of differentiability ("$\nabla f(\mathbf{x}^*) = \mathbf{0}$"). We will refer to this condition as *Fermat's optimality condition.*

Theorem 3.63 (Fermat's optimality condition). *Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper convex function. Then*

$$\mathbf{x}^* \in \operatorname{argmin}\{f(\mathbf{x}) : \mathbf{x} \in \mathbb{E}\} \tag{3.70}$$

if and only if $\mathbf{0} \in \partial f(\mathbf{x}^*)$.

Proof. Follows by the definition of the subgradient. Indeed, (3.70) is satisfied if and only if

$$f(\mathbf{x}) \geq f(\mathbf{x}^*) + \langle \mathbf{0}, \mathbf{x} - \mathbf{x}^* \rangle \text{ for any } \mathbf{x} \in \operatorname{dom}(f),$$

which is the same as the inclusion $\mathbf{0} \in \partial f(\mathbf{x}^*)$. $\quad\square$

Example 3.64 (minimizing piecewise linear functions). Consider the problem

$$\min_{\mathbf{x} \in \mathbb{R}^n} \left[f(\mathbf{x}) \equiv \max_{i=1,2,\ldots,m} \left\{ \mathbf{a}_i^T \mathbf{x} + b_i \right\} \right], \tag{3.71}$$

where $\mathbf{a}_i \in \mathbb{R}^n, b_i \in \mathbb{R}, i = 1, 2, \ldots, m$. Denote

$$I(\mathbf{x}) = \left\{ i : f(\mathbf{x}) = \mathbf{a}_i^T \mathbf{x} + b_i \right\}.$$

Then, by Example 3.53,

$$\partial f(\mathbf{x}) = \left\{ \sum_{i \in I(\mathbf{x})} \lambda_i \mathbf{a}_i : \sum_{i \in I(\mathbf{x})} \lambda_i = 1, \lambda_j \geq 0, j \in I(\mathbf{x}) \right\}.$$

Therefore, since by Fermat's optimality condition $\mathbf{x}^*$ is an optimal solution of (3.71) if and only if $\mathbf{0} \in \partial f(\mathbf{x}^*)$, it follows that $\mathbf{x}^*$ is an optimal solution of problem (3.71) if and only if there exists $\boldsymbol{\lambda} \in \Delta_m$ such that

$$\mathbf{0} = \sum_{i=1}^{m} \lambda_i \mathbf{a}_i \text{ and } \lambda_j = 0 \text{ for any } j \notin I(\mathbf{x}^*). \tag{3.72}$$

We can rewrite this condition in a more compact way by denoting $\mathbf{A} \in \mathbb{R}^{m \times n}$ to be the matrix whose rows are $\mathbf{a}_1^T, \mathbf{a}_2^T, \ldots, \mathbf{a}_m^T$. Then the optimality condition (3.72) can be written as

$$\exists \boldsymbol{\lambda} \in \Delta_m \text{ s.t. } \mathbf{A}^T \boldsymbol{\lambda} = \mathbf{0} \text{ and } \lambda_j(\mathbf{a}_j^T \mathbf{x}^* + b_j - f(\mathbf{x}^*)) = 0, j = 1, 2, \ldots, m. \quad \blacksquare$$

Example 3.65 (medians). Suppose that we are given n different[20] and ordered numbers $a_1 < a_2 < \cdots < a_n$. Denote $A = \{a_1, a_2, \ldots, a_n\} \subseteq \mathbb{R}$. The *median* of A is a number β that satisfies

$$\#\{i : a_i \leq \beta\} \geq \frac{n}{2} \text{ and } \#\{i : a_i \geq \beta\} \geq \frac{n}{2}.$$

---

[20]The assumption that these are different and ordered numbers is not essential and is made for the sake of simplicity of exposition.

That is, a median of A is a number that satisfies that at least half of the numbers in A are smaller or equal to it and that at least half are larger or equal. It is not difficult to see that if A has an odd number of elements, then the median is the middlemost number. For example, the median of $\{5, 8, 11, 60, 100\}$ is 11. If the number of elements in A is even, then there is no unique median. The set of medians comprises all numbers between the two middle values. For example, if $A = \{5, 8, 11, 20, 60, 100\}$, then the set of medians of A is the interval $[11, 20]$. In general,

$$\text{median}(A) = \begin{cases} a_{\frac{n+1}{2}}, & n \text{ odd}, \\[2mm] [a_{\frac{n}{2}}, a_{\frac{n}{2}+1}], & n \text{ even}. \end{cases}$$

From an optimization perspective, the set of possible medians is the optimal solution set of the problem

$$\min \left\{ f(x) \equiv \sum_{i=1}^{n} |x - a_i| \right\}. \tag{3.73}$$

To show this, denote $f_i(x) \equiv |x - a_i|$, so that $f(x) = f_1(x) + f_2(x) + \cdots + f_n(x)$, and note that for any $i \in \{1, 2, \ldots, n\}$,

$$\partial f_i(x) = \begin{cases} 1, & x > a_i, \\[2mm] -1, & x < a_i, \\[2mm] [-1, 1], & x = a_i. \end{cases}$$

By the sum rule of subdifferential calculus (Theorem 3.40),

$$\begin{aligned} \partial f(x) &= \partial f_1(x) + \partial f_2(x) + \cdots + \partial f_n(x) \\ &= \begin{cases} \#\{i : a_i < x\} - \#\{i : a_i > x\}, & x \notin A, \\[2mm] \#\{i : a_i < x\} - \#\{i : a_i > x\} + [-1, 1], & x \in A. \end{cases} \end{aligned}$$

We can further elaborate and write

$$\partial f(x) = \begin{cases} 2i - n, & x \in (a_i, a_{i+1}), \\[2mm] 2i - 1 - n + [-1, 1], & x = a_i, \\[2mm] -n, & x < a_1, \\[2mm] n, & x > a_n. \end{cases} \tag{3.74}$$

Let $i \in \{1, 2, \ldots, n\}$. By (3.74), $0 \in \partial f(a_i)$ if and only if $|2i - 1 - n| \leq 1$, which is equivalent to $\frac{n}{2} \leq i \leq \frac{n}{2} + 1$ and $0 \in \partial f(x)$ for some $x \in (a_i, a_{i+1})$ if and only if $i = \frac{n}{2}$. We can thus conclude that if n is odd, then the only optimal point is $a_{\frac{n+1}{2}}$, and when n is even, the optimal set is the interval $[a_{\frac{n}{2}}, a_{\frac{n}{2}+1}]$, establishing the fact that the optimal set of (3.73) is exactly the set of medians. ■

Example 3.66 (Fermat–Weber problem). Given m different points in $\mathbb{R}^d$, $\mathcal{A} = \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m\}$, and m positive weights $\omega_1, \omega_2, \ldots, \omega_m$, the *Fermat–Weber problem* is given by

$$(\text{FW}) \quad \min_{\mathbf{x} \in \mathbb{R}^d} \left\{ f(\mathbf{x}) \equiv \sum_{i=1}^m \omega_i \|\mathbf{x} - \mathbf{a}_i\|_2 \right\}.$$

The Fermat–Weber problem is actually a weighted multidimensional version of the median problem (3.73) discussed in the previous example and is therefore also referred to in the literature as the *geometric median* problem. Let us write explicitly the optimality conditions for problem (FW). Denote $f_i(\mathbf{x}) = \omega_i g_i(\mathbf{x})$, where $g_i(\mathbf{x}) \equiv \|\mathbf{x} - \mathbf{a}_i\|_2$. Then for any $i \in \{1, 2, \ldots, m\}$

$$\partial f_i(\mathbf{x}) = \begin{cases} \omega_i \frac{\mathbf{x} - \mathbf{a}_i}{\|\mathbf{x} - \mathbf{a}_i\|_2}, & \mathbf{x} \neq \mathbf{a}_i, \\ B_{\|\cdot\|_2}[\mathbf{0}, \omega_i], & \mathbf{x} = \mathbf{a}_i, \end{cases}$$

where here we used Theorems 3.35 ("multiplication by a positive scalar"), the affine transformation rule of subdifferential calculus (Theorem 3.43(b)), and Example 3.34, in which the subdifferential set of the l_2-norm was computed. Obviously, $f = \sum_{i=1}^m f_i$, and hence, by the sum rule of subdifferential calculus (Theorem 3.40[21]), we obtain that

$$\partial f(\mathbf{x}) = \sum_{i=1}^m \partial f_i(\mathbf{x}) = \begin{cases} \sum_{i=1}^m \omega_i \frac{\mathbf{x} - \mathbf{a}_i}{\|\mathbf{x} - \mathbf{a}_i\|_2}, & \mathbf{x} \notin \mathcal{A}, \\ \sum_{i=1, i \neq j}^m \omega_i \frac{\mathbf{x} - \mathbf{a}_i}{\|\mathbf{x} - \mathbf{a}_i\|_2} + B[\mathbf{0}, \omega_j], & \mathbf{x} = \mathbf{a}_j (j = 1, 2, \ldots, m). \end{cases}$$

Using Fermat's optimality condition (Theorem 3.63), we can conclude that $\mathbf{x}^* \in \mathbb{R}^d$ is an optimal solution of problem (FW) if and only if either

$$\mathbf{x}^* \notin \mathcal{A} \text{ and } \sum_{i=1}^m \omega_i \frac{\mathbf{x}^* - \mathbf{a}_i}{\|\mathbf{x}^* - \mathbf{a}_i\|_2} = \mathbf{0}$$

or for some $j \in \{1, 2, \ldots, m\}$

$$\mathbf{x}^* = \mathbf{a}_j \text{ and } \left\| \sum_{i=1, i \neq j}^m \omega_i \frac{\mathbf{x}^* - \mathbf{a}_i}{\|\mathbf{x}^* - \mathbf{a}_i\|_2} \right\|_2 \leq \omega_j. \qquad \blacksquare$$

3.7.2 Convex Constrained Optimization

Consider the constrained optimization problem

$$\min\{f(\mathbf{x}) : \mathbf{x} \in C\}, \tag{3.75}$$

where f is an extended real-valued convex function and $C \subseteq \mathbb{E}$ is a convex set. Using Fermat's optimality condition (Theorem 3.63) and the convexity assumptions, it is easy to write a necessary and sufficient optimality condition for problem (3.75) in terms of the subdifferential set of f and the normal cone of C.

---

[21]or by Corollary 3.39

Theorem 3.67 (necessary and sufficient optimality conditions for convex constrained optimization). *Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper and convex function, and let $C \subseteq \mathbb{E}$ be a convex set for which $\operatorname{ri}(\operatorname{dom}(f)) \cap \operatorname{ri}(C) \neq \emptyset$. Then $\mathbf{x}^* \in C$ is an optimal solution of (3.75) if and only if*

$$\text{there exists } \mathbf{g} \in \partial f(\mathbf{x}^*) \text{ for which } -\mathbf{g} \in N_C(\mathbf{x}^*). \tag{3.76}$$

Proof. Problem (3.75) can be rewritten as

$$\min_{\mathbf{x} \in \mathbb{E}} f(\mathbf{x}) + \delta_C(\mathbf{x}).$$

Since $\operatorname{ri}(\operatorname{dom}(f)) \cap \operatorname{ri}(C) \neq \emptyset$, it follows by the sum rule of subdifferential calculus (Theorem 3.40) that for any $\mathbf{x} \in C$,

$$\partial(f + \delta_C)(\mathbf{x}) = \partial f(\mathbf{x}) + \partial \delta_C(\mathbf{x}).$$

By Example 3.5, $\partial \delta_C(\mathbf{x}) = N_C(\mathbf{x})$, and consequently for any $\mathbf{x} \in C$,

$$\partial(f + \delta_C)(\mathbf{x}) = \partial f(\mathbf{x}) + N_C(\mathbf{x}).$$

Therefore, invoking Fermat's optimality condition (Theorem 3.63), $\mathbf{x}^* \in C$ is an optimal solution of (3.75) if and only if $\mathbf{0} \in \partial f(\mathbf{x}^*) + N_C(\mathbf{x}^*)$, that is, if and only if

$$(-\partial f(\mathbf{x}^*)) \cap N_C(\mathbf{x}^*) \neq \emptyset,$$

which is the same as condition (3.76). $\quad\square$

Using the definition of the normal cone, we can write the optimality condition in a slightly more explicit manner.

Corollary 3.68 (necessary and sufficient optimality conditions for convex constrained optimization—second version). *Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper and convex function, and let C be a convex set satisfying $\operatorname{ri}(\operatorname{dom}(f)) \cap \operatorname{ri}(C) \neq \emptyset$. Then $\mathbf{x}^* \in C$ is an optimal solution of (3.75) if and only if*

$$\text{there exists } \mathbf{g} \in \partial f(\mathbf{x}^*) \text{ for which } \langle \mathbf{g}, \mathbf{x} - \mathbf{x}^* \rangle \geq 0 \text{ for any } \mathbf{x} \in C. \tag{3.77}$$

Condition (3.77) is not particularly explicit. We will show in the next example how to write it in an explicit way for the case where $C = \Delta_n$.

Example 3.69 (optimality conditions over the unit simplex). Suppose that the assumptions in Corollary 3.68 hold and that $C = \Delta_n, \mathbb{E} = \mathbb{R}^n$. Given $\mathbf{x}^* \in \Delta_n$, we will show that the condition

$$\text{(I) } \mathbf{g}^T(\mathbf{x} - \mathbf{x}^*) \geq 0 \text{ for all } \mathbf{x} \in \Delta_n$$

is satisfied if and only if the following condition is satisfied:

$$\text{(II) there exist } \mu \in \mathbb{R} \text{ such that } g_i \begin{cases} = \mu, & x_i^* > 0, \\ \geq \mu, & x_i^* = 0. \end{cases}$$

Assume first that (II) is satisfied. Then for any $\mathbf{x} \in \Delta_n$,

$$
\begin{aligned}
\mathbf{g}^T(\mathbf{x} - \mathbf{x}^*) &= \sum_{i=1}^{n} g_i(x_i - x_i^*) \\
&= \sum_{i:x_i^*>0} g_i(x_i - x_i^*) + \sum_{i:x_i^*=0} g_i x_i \\
&\geq \sum_{i:x_i^*>0} \mu(x_i - x_i^*) + \mu \sum_{i:x_i^*=0} x_i \\
&= \mu \sum_{i=1}^{n} x_i - \mu \sum_{i:x_i^*>0} x_i^* = \mu - \mu = 0,
\end{aligned}
$$

proving that condition (I) is satisfied. To show the reverse direction, assume that (I) is satisfied. Let i and j be two different indices for which $x_i^* > 0$. Define the vector $\mathbf{x} \in \Delta_n$ as

$$
x_k = \begin{cases} x_k^*, & k \notin \{i,j\}, \\ x_i^* - \frac{x_i^*}{2}, & k = i, \\ x_j^* + \frac{x_i^*}{2}, & k = j. \end{cases}
$$

The inequality $\mathbf{g}^T(\mathbf{x} - \mathbf{x}^*) \geq 0$ then amounts to

$$
-\frac{x_i^*}{2} g_i + \frac{x_i^*}{2} g_j \geq 0,
$$

which by the fact that $x_i^* > 0$ implies that

$$
g_i \leq g_j. \tag{3.78}
$$

In particular, for any two indices $i \neq j$ for which $x_i^*, x_j^* > 0$, the two inequalities $g_i \leq g_j$ and $g_j \leq g_i$ hold, and hence $g_i = g_j$. Therefore, all the components of $\mathbf{g}$ corresponding to positive components of $\mathbf{x}^*$ have the same value, which we will denote by μ. Let i be any index for which $x_i^* > 0$. Then for any index j for which $x_j^* = 0$, the inequality (3.78) holds. Therefore, $g_j \geq \mu$, and condition (II) is thus established. ■

We summarize the discussion in Example 3.69 with the following corollary.

Corollary 3.70 (necessary and sufficient optimality conditions for convex problems over the unit simplex). *Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper and convex function. Suppose that $\mathrm{ri}(\Delta_n) \cap \mathrm{ri}(\mathrm{dom}(f)) \neq \emptyset$. Then $\mathbf{x}^* \in \Delta_n$ is an optimal solution of*

$$
\min\{f(\mathbf{x}) : \mathbf{x} \in \Delta_n\}
$$

if and only if there exists $\mathbf{g} \in \partial f(\mathbf{x}^*)$ and $\mu \in \mathbb{R}$ for which

$$
g_i \begin{cases} = \mu, & x_i^* > 0, \\ \geq \mu, & x_i^* = 0. \end{cases}
$$

The following example illustrates one instance in which the optimal solution of a convex problem over the unit simplex can be found using Corollary 3.70.

Example 3.71. Consider the problem

$$\min_{\mathbf{x}} \left\{ \sum_{i=1}^n x_i \log x_i - \sum_{i=1}^n y_i x_i : \mathbf{x} \in \Delta_n \right\}, \tag{3.79}$$

where $\mathbf{y} \in \mathbb{R}^n$ is a given vector. Problem (3.79) can be written as

$$\min\{f(\mathbf{x}) : \mathbf{x} \in \Delta_n\},$$

where $f : \mathbb{R}^n \to (-\infty, \infty]$ is given by

$$f(\mathbf{x}) = \begin{cases} \sum_{i=1}^n x_i \log x_i - \sum_{i=1}^n y_i x_i, & \mathbf{x} \geq \mathbf{0}, \\ \infty & \text{else.} \end{cases}$$

Let us assume that there exists an optimal solution[22] $\mathbf{x}^*$ satisfying $\mathbf{x}^* > \mathbf{0}$. Then under this assumption, by Corollary 3.70 and the fact that f is differentiable at any positive vector, it follows that there exists $\mu \in \mathbb{R}$ such that for any i, $\frac{\partial f}{\partial x_i}(\mathbf{x}^*) = \mu$, which is the same as $\log x_i^* + 1 - y_i = \mu$. Therefore, for any i,

$$x_i^* = e^{\mu - 1 + y_i} = \alpha e^{y_i}, \quad i = 1, 2, \ldots, n$$

where $\alpha = e^{\mu - 1}$. Since $\sum_{i=1}^n x_i^* = 1$, it follows that $\alpha = \frac{1}{\sum_{j=1}^n e^{y_j}}$. Therefore,

$$x_i^* = \frac{e^{y_i}}{\sum_{j=1}^n e^{y_j}}, \quad i = 1, 2, \ldots, n.$$

This is indeed an optimal solution of problem (3.79) since it satisfies the conditions of Corollary 3.70, which are (also) sufficient conditions for optimality. ∎

3.7.3 The Nonconvex Composite Model

It is also possible to write a necessary optimality condition for *nonconvex* problems in terms of subgradients. We will write such a condition for problems consisting of minimizing a composite function $f + g$, where f possesses some differentiability properties but is not assumed to be convex while g is convex but not assumed to have any special differentiability properties.

Theorem 3.72 (optimality conditions for the composite problem). *Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper function, and let $g : \mathbb{E} \to (-\infty, \infty]$ be a proper convex function such that $\mathrm{dom}(g) \subseteq \mathrm{int}(\mathrm{dom}(f))$. Consider the problem*

$$\text{(P)} \quad \min_{\mathbf{x} \in \mathbb{E}} f(\mathbf{x}) + g(\mathbf{x}).$$

---

[22]It is not difficult to show a priori that the problem has a unique solution.

(a) **(necessary condition)** *If $\mathbf{x}^* \in \mathrm{dom}(g)$ is a local optimal solution of* (P) *and f is differentiable at $\mathbf{x}^*$, then*

$$-\nabla f(\mathbf{x}^*) \in \partial g(\mathbf{x}^*). \tag{3.80}$$

(b) **(necessary and sufficient condition for convex problems)** *Suppose that f is convex. If f is differentiable at $\mathbf{x}^* \in \mathrm{dom}(g)$, then $\mathbf{x}^*$ is a global optimal solution of* (P) *if and only if* (3.80) *is satisfied.*

Proof. (a) Let $\mathbf{y} \in \mathrm{dom}(g)$. Then by the convexity of $\mathrm{dom}(g)$, for any $\lambda \in (0,1)$, the point $\mathbf{x}_\lambda = (1-\lambda)\mathbf{x}^* + \lambda\mathbf{y}$ is in $\mathrm{dom}(g)$, and by the local optimality of $\mathbf{x}^*$, it follows that, for small enough λ,

$$f(\mathbf{x}_\lambda) + g(\mathbf{x}_\lambda) \geq f(\mathbf{x}^*) + g(\mathbf{x}^*).$$

That is,

$$f((1-\lambda)\mathbf{x}^* + \lambda\mathbf{y}) + g((1-\lambda)\mathbf{x}^* + \lambda\mathbf{y}) \geq f(\mathbf{x}^*) + g(\mathbf{x}^*).$$

Using the convexity of g, it follows that

$$f((1-\lambda)\mathbf{x}^* + \lambda\mathbf{y}) + (1-\lambda)g(\mathbf{x}^*) + \lambda g(\mathbf{y}) \geq f(\mathbf{x}^*) + g(\mathbf{x}^*),$$

which is the same as

$$\frac{f((1-\lambda)\mathbf{x}^* + \lambda\mathbf{y}) - f(\mathbf{x}^*)}{\lambda} \geq g(\mathbf{x}^*) - g(\mathbf{y}).$$

Taking $\lambda \to 0^+$ in the last inequality yields

$$f'(\mathbf{x}^*; \mathbf{y} - \mathbf{x}^*) \geq g(\mathbf{x}^*) - g(\mathbf{y}),$$

where we used the fact that since f is differentiable at $\mathbf{x}^*$, its directional derivatives exist. In fact, by Theorem 3.29, we have $f'(\mathbf{x}^*; \mathbf{y} - \mathbf{x}^*) = \langle \nabla f(\mathbf{x}^*), \mathbf{y} - \mathbf{x}^* \rangle$, and hence for any $\mathbf{y} \in \mathrm{dom}(g)$,

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

showing that indeed $-\nabla f(\mathbf{x}^*) \in \partial g(\mathbf{x}^*)$.

(b) Suppose in addition that f is convex. If $\mathbf{x}^*$ is an optimal solution of (P), then we already proved in part (a) that (3.80) is satisfied. Suppose now that (3.80) is satisfied. Then for any $\mathbf{y} \in \mathrm{dom}(g)$,

$$g(\mathbf{y}) \geq g(\mathbf{x}^*) + \langle -\nabla f(\mathbf{x}^*), \mathbf{y} - \mathbf{x}^* \rangle. \tag{3.81}$$

By the convexity of f, for any $\mathbf{y} \in \mathrm{dom}(g)$,

$$f(\mathbf{y}) \geq f(\mathbf{x}^*) + \langle \nabla f(\mathbf{x}^*), \mathbf{y} - \mathbf{x}^* \rangle. \tag{3.82}$$

Adding (3.81) and (3.82), we obtain that

$$f(\mathbf{y}) + g(\mathbf{y}) \geq f(\mathbf{x}^*) + g(\mathbf{x}^*)$$

for any $\mathbf{y} \in \mathrm{dom}(g)$, meaning that $\mathbf{x}^*$ is an optimal solution of (P). $\quad\square$

The condition (3.80) is an important optimality condition, and we will refer to it as the "stationarity" condition.

Definition 3.73 (stationarity). *Let $f : \mathbb{E} \to (-\infty, \infty]$ be proper and let $g : \mathbb{E} \to (-\infty, \infty]$ be a proper convex function such that $\mathrm{dom}(g) \subseteq \mathrm{int}(\mathrm{dom}(f))$. Consider the problem*

$$\text{(P)} \quad \min_{\mathbf{x} \in \mathbb{E}} f(\mathbf{x}) + g(\mathbf{x}).$$

A point $\mathbf{x}^*$ in which f is differentiable is called a **stationary point** *of* (P) *if*

$$-\nabla f(\mathbf{x}^*) \in \partial g(\mathbf{x}^*).$$

Under the setting of Definition 3.73, $\mathbf{x}^*$ is also called *a stationary point of the function $f + g$.*

We have shown in Theorem 3.72 that stationarity is a necessary local optimality condition for problem (P), and that if f is convex, then stationarity is a necessary and sufficient global optimality condition. The case $g = \delta_C$ deserves a separate discussion.

Example 3.74 (convex constrained nonconvex programming). When $g = \delta_C$ for a nonempty convex set $C \subseteq \mathbb{E}$, problem (P) becomes

$$\min\{f(\mathbf{x}) : \mathbf{x} \in C\},$$

which is a problem consisting of minimizing a (possibly) nonconvex function over a convex set. A point $\mathbf{x}^* \in C$ in which f is differentiable is a stationary point of (P) if and only if

$$-\nabla f(\mathbf{x}^*) \in \partial \delta_C(\mathbf{x}^*) = N_C(\mathbf{x}^*), \tag{3.83}$$

where the equality is due to Example 3.5. By the definition of the normal cone, condition (3.83) can be rewritten as

$$\langle -\nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \leq 0 \text{ for any } \mathbf{x} \in C,$$

which is the same as

$$\langle \nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq 0 \text{ for any } \mathbf{x} \in C. \quad \blacksquare$$

Example 3.75. Consider the problem

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) + \lambda \|\mathbf{x}\|_1, \tag{3.84}$$

where $f : \mathbb{R}^n \to (-\infty, \infty]$ is an extended real-valued function. A point $\mathbf{x}^* \in \mathrm{int}(\mathrm{dom}(f))$ in which f is differentiable is a stationary point of (3.84) if

$$-\nabla f(\mathbf{x}^*) \in \lambda \partial g(\mathbf{x}^*),$$

where $g(\cdot) = \| \cdot \|_1$. Using the expression for the subdifferential set of the l_1-norm given in Example 3.41, we obtain that $\mathbf{x}^*$ is a stationary point of problem (3.84) if

and only if

$$\frac{\partial f(\mathbf{x}^*)}{\partial x_i} \begin{cases} = -\lambda, & x_i^* > 0, \\ = \lambda, & x_i^* < 0, \\ \in [-\lambda, \lambda], & x_i^* = 0. \end{cases} \tag{3.85}$$

By Theorem 3.72, condition (3.85) is a necessary condition for $\mathbf{x}^*$ to be a local minimum of problem (3.84). If f is also convex, then condition (3.85) is a necessary and sufficient condition for $\mathbf{x}^*$ to be a global optimal solution of problem (3.84). ∎

3.7.4 The KKT Conditions

In this section we will show that the KKT conditions for constrained convex problems can be directly deduced by Fermat's optimality condition. For that, we begin by establishing an equivalent reformulation of general inequality constrained problems.

Lemma 3.76. *Let $f, g_1, g_2, \ldots, g_m : \mathbb{E} \to \mathbb{R}$ be real-valued functions. Consider the problem*

$$\begin{aligned} \min \quad & f(\mathbf{x}) \\ s.t. \quad & g_i(\mathbf{x}) \leq 0, \quad i = 1, 2, \ldots, m. \end{aligned} \tag{3.86}$$

Assume that the minimum value of problem (3.86) is finite and equal to $\bar{f}$. Define the function

$$F(\mathbf{x}) \equiv \max\{f(\mathbf{x}) - \bar{f}, g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_m(\mathbf{x})\}. \tag{3.87}$$

Then the optimal set of problem (3.86) is the same as the set of minimizers of F.

Proof. Let X^* be the optimal set of problem (3.86). To establish the result, we will show that F satisfies the following two properties:

(i) $F(\mathbf{x}) > 0$ for any $\mathbf{x} \notin X^*$.

(ii) $F(\mathbf{x}) = 0$ for any $\mathbf{x} \in X^*$.

To prove property (i), let $\mathbf{x} \notin X^*$. There are two options. Either $\mathbf{x}$ is not feasible, meaning that $g_i(\mathbf{x}) > 0$ for some i, and hence by its definition $F(\mathbf{x}) > 0$. If $\mathbf{x}$ is feasible but not optimal, then $g_i(\mathbf{x}) \leq 0$ for all $i = 1, 2, \ldots, m$ and $f(\mathbf{x}) > \bar{f}$, which also implies that $F(\mathbf{x}) > 0$. To prove (ii), suppose that $\mathbf{x} \in X^*$. Then $g_i(\mathbf{x}) \leq 0$ for all $i = 1, 2, \ldots, m$ and $f(\mathbf{x}) = \bar{f}$, implying that $F(\mathbf{x}) = 0$. ∎

Using Lemma 3.76, we can conclude that problem (3.86) reduces to

$$\min_{\mathbf{x} \in \mathbb{E}} F(\mathbf{x}) \tag{3.88}$$

in the sense that the optimal sets of the two problems are the same. Using this equivalence, we can now establish under additional convexity assumptions the well-known Fritz-John optimality conditions for problem (3.86).

Theorem 3.77 (Fritz-John necessary optimality conditions). *Consider the minimization problem*

$$\min \quad f(\mathbf{x})$$
$$\text{s.t.} \quad g_i(\mathbf{x}) \leq 0, \quad i = 1, 2, \ldots, m, \tag{3.89}$$

where $f, g_1, g_2, \ldots, g_m : \mathbb{E} \to \mathbb{R}$ *are real-valued convex functions. Let* $\mathbf{x}^*$ *be an optimal solution of* (3.89). *Then there exist* $\lambda_0, \lambda_1, \ldots, \lambda_m \geq 0$, *not all zeros, for which*

$$\mathbf{0} \in \lambda_0 \partial f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \partial g_i(\mathbf{x}^*) \tag{3.90}$$

$$\lambda_i g_i(\mathbf{x}^*) = 0, \quad i = 1, 2, \ldots, m. \tag{3.91}$$

Proof. Let $\mathbf{x}^*$ be an optimal solution of problem (3.89). Denote the optimal value of problem (3.89) by $\bar{f} = f(\mathbf{x}^*)$. Using Lemma 3.76, it follows that $\mathbf{x}^*$ is an optimal solution of the problem

$$\min_{\mathbf{x} \in \mathbb{E}} \{ F(\mathbf{x}) \equiv \max\{g_0(\mathbf{x}), g_1(\mathbf{x}), \ldots, g_m(\mathbf{x})\},$$

where $g_0(\mathbf{x}) \equiv f(\mathbf{x}) - \bar{f}$. Obviously, $F(\mathbf{x}^*) = 0$. Since F is a maximum of convex functions, it is convex, and hence, using Fermat's optimality condition (Theorem 3.63),

$$\mathbf{0} \in \partial F(\mathbf{x}^*). \tag{3.92}$$

By the max rule of subdifferential calculus (Theorem 3.50),

$$\partial F(\mathbf{x}^*) = \text{conv}\left((\cup_{i \in I(\mathbf{x}^*)} \partial g_i(\mathbf{x}^*) \right), \tag{3.93}$$

where $I(\mathbf{x}^*) = \{ i \in \{0, 1, \ldots, m\} : g_i(\mathbf{x}^*) = 0 \}$. Combining (3.92) and (3.93), we can deduce that there exists $\lambda_i \geq 0, i \in I(\mathbf{x}^*)$, such that $\sum_{i \in I(\mathbf{x}^*)} \lambda_i = 1$ for which

$$\mathbf{0} \in \sum_{i \in I(\mathbf{x}^*)} \lambda_i \partial g_i(\mathbf{x}^*). \tag{3.94}$$

Since $g_0(\mathbf{x}^*) = f(\mathbf{x}^*) - \bar{f} = 0$, it follows that $0 \in I(\mathbf{x}^*)$, and hence (3.94) can be rewritten as

$$\mathbf{0} \in \lambda_0 \partial f(\mathbf{x}^*) + \sum_{i \in I(\mathbf{x}^*) \setminus \{0\}} \lambda_i \partial g_i(\mathbf{x}^*).$$

Defining $\lambda_i = 0$ for any $i \in \{1, 2, \ldots, m\} \setminus I(\mathbf{x}^*)$, we conclude that conditions (3.90) and (3.91) are satisfied. Finally, not all the λ_i's are zeros since $\sum_{i \in I(\mathbf{x}^*)} \lambda_i = 1$. $\square$

We will now establish the KKT conditions, which are the same as the Fritz-John conditions, but with $\lambda_0 = 1$. The necessity of these conditions requires the following additional condition, which we refer to as *Slater's condition*:

$$\text{there exists } \bar{\mathbf{x}} \in \mathbb{E} \text{ for which } g_i(\bar{\mathbf{x}}) < 0, \quad i = 1, 2, \ldots, m. \tag{3.95}$$

The sufficiency of the KKT conditions does not require any additional assumptions (besides convexity) and is actually easily derived without using the result on the Fritz-John conditions.

Theorem 3.78 (KKT conditions). *Consider the minimization problem*

$$
\begin{aligned}
\min \quad & f(\mathbf{x}) \\
s.t. \quad & g_i(\mathbf{x}) \leq 0, \quad i = 1, 2, \ldots, m,
\end{aligned}
\tag{3.96}
$$

where $f, g_1, g_2, \ldots, g_m : \mathbb{E} \to \mathbb{R}$ are real-valued convex functions.

(a) *Let $\mathbf{x}^*$ be an optimal solution of* (3.96)*, and assume that Slater's condition* (3.95) *is satisfied. Then there exist $\lambda_1, \ldots, \lambda_m \geq 0$ for which*

$$
\mathbf{0} \in \partial f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \partial g_i(\mathbf{x}^*)
\tag{3.97}
$$

$$
\lambda_i g_i(\mathbf{x}^*) = 0, \quad i = 1, 2, \ldots, m.
\tag{3.98}
$$

(b) *If $\mathbf{x}^* \in \mathbb{E}$ satisfies conditions* (3.97) *and* (3.98) *for some $\lambda_1, \lambda_2, \ldots, \lambda_m \geq 0$, then it is an optimal solution of problem* (3.96)*.*

Proof. (a) By the Fritz-John conditions (Theorem 3.77) there exist $\tilde{\lambda}_0, \tilde{\lambda}_1, \ldots, \tilde{\lambda}_m \geq 0$, not all zeros, for which

$$
\mathbf{0} \in \tilde{\lambda}_0 \partial f(\mathbf{x}^*) + \sum_{i=1}^{m} \tilde{\lambda}_i \partial g_i(\mathbf{x}^*),
\tag{3.99}
$$

$$
\tilde{\lambda}_i g_i(\mathbf{x}^*) = 0, \quad i = 1, 2, \ldots, m.
\tag{3.100}
$$

We will show that $\tilde{\lambda}_0 \neq 0$. Assume by contradiction that $\tilde{\lambda}_0 = 0$. Then, by (3.99),

$$
\mathbf{0} \in \sum_{i=1}^{m} \tilde{\lambda}_i \partial g_i(\mathbf{x}^*);
$$

that is, there exist $\boldsymbol{\xi}_i \in \partial g_i(\mathbf{x}^*), i = 1, 2, \ldots, m$, such that

$$
\sum_{i=1}^{m} \tilde{\lambda}_i \boldsymbol{\xi}_i = \mathbf{0}.
\tag{3.101}
$$

Let $\bar{\mathbf{x}}$ be a point satisfying Slater's condition (3.95). By the subgradient inequality employed on the pair of points $\bar{\mathbf{x}}, \mathbf{x}^*$ w.r.t. the functions $g_i, i = 1, 2, \ldots, m$, we have

$$
g_i(\mathbf{x}^*) + \langle \boldsymbol{\xi}_i, \bar{\mathbf{x}} - \mathbf{x}^* \rangle \leq g_i(\bar{\mathbf{x}}), \quad i = 1, 2, \ldots, m.
$$

Multiplying the ith inequality by $\tilde{\lambda}_i \geq 0$ and summing over $i = 1, 2, \ldots, m$ yields

$$
\sum_{i=1}^{m} \tilde{\lambda}_i g_i(\mathbf{x}^*) + \left\langle \sum_{i=1}^{m} \tilde{\lambda}_i \boldsymbol{\xi}_i, \bar{\mathbf{x}} - \mathbf{x}^* \right\rangle \leq \sum_{i=1}^{m} \tilde{\lambda}_i g_i(\bar{\mathbf{x}}), \quad i = 1, 2, \ldots, m.
$$

Using (3.100) and (3.101), we obtain the inequality $\sum_{i=1}^{m} \tilde{\lambda}_i g_i(\bar{\mathbf{x}}) \geq 0$, which is impossible since $\tilde{\lambda}_i \geq 0$ and $g_i(\bar{\mathbf{x}}) < 0$ for any i, and not all the $\tilde{\lambda}_i$'s are zeros. Therefore, $\tilde{\lambda}_0 > 0$, and we can thus divide both the relation (3.99) and the equalities (3.100) by $\tilde{\lambda}_0$ to obtain that (3.97) and (3.98) are satisfied with $\lambda_i = \frac{\tilde{\lambda}_i}{\tilde{\lambda}_0}, i = 1, 2, \ldots, m$.

(b) Suppose then that $\mathbf{x}^*$ satisfies (3.97) and (3.98) for some nonnegative numbers $\lambda_1, \lambda_2, \ldots, \lambda_m$. Let $\hat{\mathbf{x}} \in \mathbb{E}$ be a feasible point of (3.96), meaning that $g_i(\hat{\mathbf{x}}) \leq 0$, $i = 1, 2, \ldots, m$. We will show that $f(\hat{\mathbf{x}}) \geq f(\mathbf{x}^*)$. Define the function

$$h(\mathbf{x}) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}).$$

The function h is convex, and the condition (3.97) along with the sum rule of subdifferential calculus (Theorem 3.40) yields the relation

$$\mathbf{0} \in \partial h(\mathbf{x}^*),$$

which by Fermat's optimality condition (Theorem 3.63) implies that $\mathbf{x}^*$ is a minimizer of h over $\mathbb{E}$. Combining this fact with (3.98) implies that

$$f(\mathbf{x}^*) = f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}^*) = h(\mathbf{x}^*) \leq h(\hat{\mathbf{x}}) = f(\hat{\mathbf{x}}) + \sum_{i=1}^{m} \lambda_i g_i(\hat{\mathbf{x}}) \leq f(\hat{\mathbf{x}}),$$

where the last inequality follows from the facts that $\lambda_i \geq 0$ and $g_i(\hat{\mathbf{x}}) \leq 0$ for $i = 1, 2, \ldots, m$. We have thus proven that $\mathbf{x}^*$ is an optimal solution of (3.96). $\qquad\square$

3.8   Summary of Weak and Strong Subgradient Calculus Results

This section contains a summary of most of the rules and results concerning the computation of subdifferential sets (strong results), as well as rules for computing specific subgradients in the subdifferential sets (weak results). Before that, we begin by summarizing the rules of subdifferential calculus.

- **Multiplication by a positive scalar**

$$\partial(\alpha f)(\mathbf{x}) = \alpha \partial f(\mathbf{x}).$$

  Assumptions: $f : \mathbb{E} \to (-\infty, \infty]$ proper, $\alpha > 0$, $\mathbf{x} \in \mathrm{dom}(f)$. [Theorem 3.35]
- **Differentiability**
  f is differentiable at $\mathbf{x}$ if and only if $\partial f(\mathbf{x})$ is a singleton, and in that case

$$\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}.$$

  Assumptions: $f : \mathbb{E} \to (-\infty, \infty]$ proper convex, $\mathbf{x} \in \mathrm{int}(\mathrm{dom}(f))$. [Theorem 3.33]
- **Weak sum rule of subdifferential calculus**

$$\sum_{i=1}^{m} \partial f_i(\mathbf{x}) \subseteq \partial(\sum_{i=1}^{m} f_i)(\mathbf{x}).$$

  Assumptions: $f_1, f_2, \ldots, f_m : \mathbb{E} \to (-\infty, \infty]$ proper convex. [Corollary 3.38(a)]

- **Sum rule of subdifferential calculus**

$$\partial(\textstyle\sum_{i=1}^m f_i)(\mathbf{x}) = \sum_{i=1}^m \partial f_i(\mathbf{x}).$$

  Assumptions: $f_1, f_2, \ldots, f_m : \mathbb{E} \to (-\infty, \infty]$ proper convex, $\cap_{i=1}^m \mathrm{ri}(\mathrm{dom}(f_i)) \neq \emptyset$. [Theorem 3.40]

- **Weak affine transformation rule of subdifferential calculus**

$$\mathcal{A}^T(\partial f(\mathcal{A}(\mathbf{x}) + \mathbf{b})) \subseteq \partial h(\mathbf{x}) \quad (h(\mathbf{x}) \equiv f(\mathcal{A}(\mathbf{x}) + \mathbf{b})).$$

  Assumptions: $f, h : \mathbb{E} \to (-\infty, \infty]$ proper convex, $\mathbf{x} \in \mathrm{dom}(h)$. [Theorem 3.43(a)]

- **Affine transformation rule of subdifferential calculus**

$$\partial h(\mathbf{x}) = \mathcal{A}^T(\partial f(\mathcal{A}(\mathbf{x}) + \mathbf{b})) \quad (h(\mathbf{x}) \equiv f(\mathcal{A}(\mathbf{x}) + \mathbf{b})).$$

  Assumptions: $f, h : \mathbb{E} \to (-\infty, \infty]$ proper convex, $\mathbf{x} \in \mathrm{int}(\mathrm{dom}(h)), \mathcal{A}(\mathbf{x}) + \mathbf{b} \in \mathrm{int}(\mathrm{dom}(f))$. [Theorem 3.43(b)]

- **Chain rule of subdifferential calculus**

$$\partial h(\mathbf{x}) = g'(f(\mathbf{x}))\partial f(\mathbf{x}) \quad (h = g \circ f).$$

  Assumptions: $f : \mathbb{E} \to \mathbb{R}$ convex, $g : \mathbb{R} \to \mathbb{R}$ nondecreasing, differentiable and convex. [Theorem 3.47]

- **Max rule of subdifferential calculus**

$$\partial(\max(f_1, f_2, \ldots, f_m))(\mathbf{x}) = \mathrm{conv}\left(\cup_{i \in I(\mathbf{x})} \partial f_i(\mathbf{x})\right),$$

  where

$$I(\mathbf{x}) = \{i : f_i(\mathbf{x}) = \max\{f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x})\}\}.$$

  Assumptions: $f_1, f_2, \ldots, f_m$ proper, convex, $\mathbf{x} \in \cap_{i=1}^m \mathrm{int}(\mathrm{dom}(f_i))$. [Theorem 3.50]

- **Weak max rule of subdifferential calculus**

$$\mathrm{conv}\left(\cup_{i \in I(\mathbf{x})} \partial f_i(\mathbf{x})\right) \subseteq \partial(\max_{i \in I} f_i)(\mathbf{x}),$$

  where

$$I(\mathbf{x}) = \{i \in I : f_i(\mathbf{x}) = \max_{i \in I} f_i(\mathbf{x})\}.$$

  Assumptions: $I =$ arbitrary index set. $f_i : \mathbb{E} \to (-\infty, \infty]$ $(i \in I)$ proper, convex, $\mathbf{x} \in \cap_{i \in I} \mathrm{dom}(f_i)$. [Theorem 3.55]

The table below contains the main examples from the chapter related to weak results of subgradients computations.

Function	Weak result	Setting	Reference
$-q =$ negative dual function	$-\mathbf{g}(\mathbf{x}_0) \in \partial(-q)(\boldsymbol{\lambda}_0)$	$q(\boldsymbol{\lambda}) = \min_{\mathbf{x} \in X} f(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x})$, $f : \mathbb{E} \to \mathbb{R}$, $\mathbf{g} : \mathbb{E} \to \mathbb{R}^m$, $\mathbf{x}_0 =$ a minimizer of $f(\mathbf{x}) + \boldsymbol{\lambda}_0^T \mathbf{g}(\mathbf{x})$ over X	Example 3.7
$f(\mathbf{X}) = \lambda_{\max}(\mathbf{X})$	$\mathbf{v}\mathbf{v}^T \in \partial f(\mathbf{X})$	$f : \mathbb{S}^n \to \mathbb{R}$, $\mathbf{v} =$ normalized maximum eigenvector of $X \in \mathbb{S}^n$	Example 3.8
$f(\mathbf{x}) = \|\mathbf{x}\|_1$	$\mathrm{sgn}(\mathbf{x}) \in \partial f(\mathbf{x})$	$\mathbb{E} = \mathbb{R}^n$	Example 3.42
$f(\mathbf{x}) = \lambda_{\max}(\mathbf{A}_0 + \sum_{i=1}^m x_i \mathbf{A}_i)$	$(\tilde{\mathbf{y}}^T \mathbf{A}_i \tilde{\mathbf{y}})_{i=1}^m \in \partial f(\mathbf{x})$	$\tilde{\mathbf{y}} =$ normalized maximum eigenvector of $\mathbf{A}_0 + \sum_{i=1}^m x_i \mathbf{A}_i$	Example 3.56

The following table contains the main strong results of subdifferential sets computations derived in this chapter.

$f(\mathbf{x})$	$\partial f(\mathbf{x})$	Assumptions	Reference
$\|\mathbf{x}\|$	$B_{\|\cdot\|_*}[\mathbf{0}, 1]$	$\mathbf{x} = \mathbf{0}$	Example 3.3
$\|\mathbf{x}\|_1$	$\left\{ \sum_{i \in I_{\neq}(\mathbf{x})} \mathrm{sgn}(x_i)\mathbf{e}_i + \sum_{i \in I_0(\mathbf{x})} [-\mathbf{e}_i, \mathbf{e}_i] \right\}$	$\mathbb{E} = \mathbb{R}^n$, $I_{\neq}(\mathbf{x}) = \{i : x_i \neq 0\}$, $I_0(\mathbf{x}) = \{i : x_i = 0\}$.	Example 3.41
$\|\mathbf{x}\|_2$	$\begin{cases} \left\{ \frac{\mathbf{x}}{\|\mathbf{x}\|_2} \right\}, & \mathbf{x} \neq \mathbf{0}, \\ B_{\|\cdot\|_2}[\mathbf{0}, 1], & \mathbf{x} = \mathbf{0}. \end{cases}$	$\mathbb{E} = \mathbb{R}^n$	Example 3.34
$\|\mathbf{x}\|_\infty$	$\left\{ \sum_{i \in I(\mathbf{x})} \lambda_i \mathrm{sgn}(x_i)\mathbf{e}_i : \begin{matrix} \sum_{i \in I(\mathbf{x})} \lambda_i = 1 \\ \lambda_i \geq 0 \end{matrix} \right\}$	$\mathbb{E} = \mathbb{R}^n$, $I(\mathbf{x}) = \{i : \|\mathbf{x}\|_\infty = |x_i|\}$, $\mathbf{x} \neq \mathbf{0}$	Example 3.52
$\max(\mathbf{x})$	$\left\{ \sum_{i \in I(\mathbf{x})} \lambda_i \mathbf{e}_i : \sum_{i \in I(\mathbf{x})} \lambda_i = 1, \lambda_i \geq 0 \right\}$	$\mathbb{E} = \mathbb{R}^n$, $I(\mathbf{x}) = \{i : \max(\mathbf{x}) = x_i\}$	Example 3.51
$\max(\mathbf{x})$	Δ_n	$\mathbb{E} = \mathbb{R}^n$, $\mathbf{x} = \alpha\mathbf{e}$ for some $\alpha \in \mathbb{R}$	Example 3.51
$\delta_S(\mathbf{x})$	$N_S(\mathbf{x})$	$\emptyset \neq S \subseteq \mathbb{E}$	Example 3.5
$\delta_{B[\mathbf{0},1]}(\mathbf{x})$	$\begin{cases} \{\mathbf{y} \in \mathbb{E}^* : \|\mathbf{y}\|_* \leq \langle \mathbf{y}, \mathbf{x} \rangle\}, & \|\mathbf{x}\| \leq 1, \\ \emptyset, & \|\mathbf{x}\| > 1. \end{cases}$		Example 3.6
$\|\mathbf{A}\mathbf{x} + \mathbf{b}\|_1$	$\sum_{i \in I_{\neq}(\mathbf{x})} \mathrm{sgn}(\mathbf{a}_i^T\mathbf{x} + b_i)\mathbf{a}_i + \sum_{i \in I_0(\mathbf{x})} [-\mathbf{a}_i, \mathbf{a}_i]$	$\mathbb{E} = \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^m$, $I_{\neq}(\mathbf{x}) = \{i : \mathbf{a}_i^T\mathbf{x} + b_i \neq 0\}$, $I_0(\mathbf{x}) = \{i : \mathbf{a}_i^T\mathbf{x} + b_i = 0\}$	Example 3.44
$\|\mathbf{A}\mathbf{x} + \mathbf{b}\|_2$	$\begin{cases} \frac{\mathbf{A}^T(\mathbf{A}\mathbf{x}+\mathbf{b})}{\|\mathbf{A}\mathbf{x}+\mathbf{b}\|_2}, & \mathbf{A}\mathbf{x} + \mathbf{b} \neq \mathbf{0}, \\ \mathbf{A}^T B_{\|\cdot\|_2}[\mathbf{0}, 1], & \mathbf{A}\mathbf{x} + \mathbf{b} = \mathbf{0}. \end{cases}$	$\mathbb{E} = \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^m$	Example 3.45
$\|\mathbf{A}\mathbf{x} + \mathbf{b}\|_\infty$	$\left\{ \sum_{i \in I_{\mathbf{x}}} \lambda_i \mathrm{sgn}(\mathbf{a}_i^T\mathbf{x} + b_i)\mathbf{a}_i : \begin{matrix} \sum_{i \in I_{\mathbf{x}}} \lambda_i = 1 \\ \lambda_i \geq 0 \end{matrix} \right\}$	$\mathbb{E} = \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^m$, $I_{\mathbf{x}} = \{i : \|\mathbf{A}\mathbf{x} + \mathbf{b}\|_\infty = |\mathbf{a}_i^T\mathbf{x}+b_i|\}$, $\mathbf{A}\mathbf{x} + \mathbf{b} \neq \mathbf{0}$	Example 3.54
$\|\mathbf{A}\mathbf{x} + \mathbf{b}\|_\infty$	$\mathbf{A}^T B_{\|\cdot\|_1}[\mathbf{0}, 1]$	same as above but with $\mathbf{A}\mathbf{x} + \mathbf{b} = \mathbf{0}$	Example 3.54
$\max_i\{\mathbf{a}_i^T\mathbf{x} + \mathbf{b}\}$	$\left\{ \sum_{i \in I(\mathbf{x})} \lambda_i \mathbf{a}_i : \sum_{i \in I(\mathbf{x})} \lambda_i = 1, \lambda_i \geq 0 \right\}$	$\mathbb{E} = \mathbb{R}^n$, $\mathbf{a}_i \in \mathbb{R}^n$, $b_i \in \mathbb{R}$, $I(\mathbf{x}) = \{i : f(\mathbf{x}) = \mathbf{a}_i^T\mathbf{x} + b_i\}$	Example 3.53
$\frac{1}{2}d_C^2(\mathbf{x})$	$\{\mathbf{x} - P_C(\mathbf{x})\}$	C = nonempty closed and convex, $\mathbb{E}$ = Euclidean	Example 3.31
$d_C(\mathbf{x})$	$\begin{cases} \left\{ \frac{\mathbf{x} - P_C(\mathbf{x})}{d_C(\mathbf{x})} \right\}, & \mathbf{x} \notin C, \\ N_C(\mathbf{x}) \cap B[\mathbf{0}, 1] & \mathbf{x} \in C. \end{cases}$	C = nonempty closed and convex, $\mathbb{E}$ = Euclidean	Example 3.49