Chapter 8

Nonsmooth Functions and Subgradients

Most of our discussion so far has focused on functions $f : \mathbb{R}^n \to \mathbb{R}$ that are smooth, at least differentiable. But there are many interesting optimization problems in data analysis that involve nonsmooth functions. When these functions are convex, it is not difficult to generalize the concept of a gradient. These generalizations, known as *subgradients* and *subdifferentials*, are the subject of this chapter. We show in the next chapter and beyond how they can be used to construct algorithms, related to those of earlier chapters, but with their own convergence and complexity analysis.

We start with a few examples of interesting nonsmooth functions. In Section 1.3 we introduced the "hinge loss" function, which appears often in support vector machines and deep learning. This function $h : \mathbb{R} \to \mathbb{R}$ has the form

$$h(t) = \max(t, 0).$$

It is obviously differentiable at every nonzero value of t, since $h'(t) = 0$ for $t < 0$ and $h'(t) = 1$ for $t > 0$. As t moves through 0, the gradient switches instantly from 0 to 1. We may be tempted to think of both these values as a kind of derivative for h at $t = 0$, and we would be right! Both values are "subgradients" of h. In fact, any value *between* 0 and 1 is also a subgradient. The collection of all subgradients at $t = 0$ — the closed interval $[0, 1]$ — is the "subdifferential" of h at $t = 0$.

A similar example is the absolute value function $h(t) = |t|$ which has derivative -1 for $t < 0$ and $+1$ for $t > 0$. At $t = 0$, the subdifferential of h is the interval $[-1, 1]$ and as always, each point in the subdifferential is a subgradient.

Consider next the multivariate function $f(x) = \max(a_1^T x + b_1, a_2^T x + b_2)$, where a_1 and a_2 are (distinct) vectors in $\mathbb{R}^n$ and b_1 and b_2 are scalars. It is easy to verify that f is convex and piecewise linear. In fact, there are just two pieces: A region in which $a_1^T x + b_1 \geq a_2^T x + b_2$ and another in which $a_1^T x + b_1 \leq a_2^T x + b_2$. These regions both include the hyperplane defined by $a_1^T x + b_1 = a_2^T x + b_2$, where the maximum in f is achieved by both linear pieces. In the interior of each region, the gradient $\nabla f(x)$ is defined uniquely; we have

$$\begin{aligned} a_1^T x + b_1 > a_2^T x + b_2 \;&\Rightarrow\; \nabla f(x) = a_1, \\ a_1^T x + b_1 < a_2^T x + b_2 \;&\Rightarrow\; \nabla f(x) = a_2. \end{aligned}$$

Along the hypeplane $a_1^T x + b_1 = a_2^T x + b_2$, and similarly to the hinge loss function, the appropriate definition of subdifferential is the line joining a_1 and a_2 in $\mathbb{R}^n$ space, that is, $\{\alpha a_1 + (1-\alpha)a_2 \,|\, \alpha \in [0,1]\}$.

Other nonsmooth functions include norms, which are *always* nondifferentiable at 0 (see Exercises). More exotically, the maximum eigenvalue of a symmetric matrix is a convex, but not differentiable, function of its elements. We can see this by considering the special case of diagonal 2×2 matrices

$$\begin{bmatrix} a_{11} & 0 \\ 0 & a_{22} \end{bmatrix},$$

whose maximum eigenvalue is $\max(a_{11}, a_{22})$, a nonsmooth (in fact, piecewise linear) function of its entries.[1]

Besides being of interest in their own right as a way to formulate important applications, nonsmooth convex functions play a major role in constrained optimization, where they can be used both to derive optimality conditions and to construct useful algorithms.

8.1 Subgradients and Subdifferentials

In this section we allow the convex function f to be an *extended real-valued convex function*, by which we mean that it is allowed to take infinite values at some points. (In some later discussions, we will restrict f to have finite values at all x.) We state here some useful definitions.

- The *effective domain* of f, denoted by $\operatorname{dom} f$, is defined to be the set of points $x \in \mathbb{R}^n$ for which $f(x) < \infty$.
- The *epigraph* of f is the convex subset of $\mathbb{R}^{n+1}$ defined by

$$\operatorname{epi} f := \{(x,t) \in \Omega \times \mathbb{R} \,:\, t \geq f(x)\}. \tag{8.1}$$

- f is a *proper* convex function if $f(x) < +\infty$ for some $x \in \mathbb{R}^n$ and $f(x) > -\infty$ for all $x \in \mathbb{R}^n$. All convex functions of practical interest are proper.
- f is a *closed proper* convex function if it is a proper convex function and the set $\{x \in \mathbb{R}^n \,:\, f(x) \leq \bar{t}\}$ is a closed set for all $\bar{t} \in \mathbb{R}$.
- f is *lower semicontinuous at* x if for all sequences $\{y_k\}$ such that $y_k \to x$ we have $\liminf_{k\to\infty} f(y_k) \geq f(x)$.

We define the subgradient and subdifferential as follows.

Definition 8.1. *Given $x \in \operatorname{dom} f$, we say that $g \in \mathbb{R}^n$ is a* subgradient *of f at x if*

$$f(z) \geq f(x) + g^T(z - x), \quad \textit{for all } z \in \operatorname{dom} f.$$

The subdifferential *of f at x, denoted by $\partial f(x)$, is the set of all subgradients of f at x.*

[1] The maximum eigenvalue is a *convex* function because it can be defined by $\max_{v:\|v\|_2=1} v^T A v$, which is a supremum over an infinite number of functions that are *linear* in the elements of A.

It follows immediately from this definition that $\partial f(x)$ is closed and convex, for all x (see the Exercises).

Note that if z is outside the effective domain of f, we have $f(z) = \infty$. Thus there is no need to restrict z to $\operatorname{dom} f$ in the definition above, so we can say that g is a subgradient of f at x if

$$f(z) \geq f(x) + g^T(z - x), \quad \text{for all } z \in \mathbb{R}^n. \tag{8.2}$$

Definition 8.1 leads immediately to a characterization of the minimizer of a convex function.

Theorem 8.2 (Optimality Conditions for Convex Function)**.** *The point x^* is a minimizer of the convex function f if and only if $0 \in \partial f(x^*)$.*

Proof. If $0 \in \partial f(x^*)$, we have by substituting $g = 0$ into (8.2) that $f(z) \geq f(x^*)$ for all $z \in \mathbb{R}^n$, confirming that x^* is a global minimizer. Conversely, if $f(z) \geq f(x^*)$ for all $z \in \operatorname{dom} f$, then $g = 0$ satisfies Definition 8.1, so $0 \in \partial f(x^*)$. □

Each subgradient can be identified with a supporting hyperplane to the epigraph of f. We have the following result (illustrated in Figure 8.1).

Theorem 8.3. *$g \in \partial f(x)$ if and only if $(g, -1)$ defines a supporting hyperplane to* epi f *at the point $(x, f(x))$, that is,*

$$\begin{bmatrix} g \\ -1 \end{bmatrix}^T \left\{ \begin{bmatrix} y \\ t \end{bmatrix} - \begin{bmatrix} x \\ f(x) \end{bmatrix} \right\} \leq 0 \quad \textit{for all } (y, t) \in \operatorname{epi} f.$$

Proof. Given a supporting hyperplane defined by $(g, -1)$ at $(x, f(x))$, we have for any y that $(y, f(y)) \in \operatorname{epi} f$ and therefore

$$0 \geq \begin{bmatrix} g \\ -1 \end{bmatrix}^T \begin{bmatrix} y - x \\ f(y) - f(x) \end{bmatrix} = g^T(y - x) - (f(y) - f(x)) \;\Leftrightarrow\; f(y) \geq f(x) + g^T(y - x),$$

which implies that $g \in \partial f(x)$. For the converse, given $g \in \partial f(x)$, we have for any $(y, t) \in \operatorname{epi} f$ that $f(y) \geq f(x) + g^T(y - x)$ and $t \geq f(y)$. Thus

$$\begin{bmatrix} g \\ -1 \end{bmatrix}^T \begin{bmatrix} y - x \\ t - f(x) \end{bmatrix} \leq \begin{bmatrix} g \\ -1 \end{bmatrix}^T \begin{bmatrix} y - x \\ f(y) - f(x) \end{bmatrix} \leq 0,$$

proving that $(g, -1)$ defines the supporting hyperplane. □

We now prove a sufficient condition for existence of a subgradient.

Lemma 8.4. *A subgradient g of f exists at x if x is in the interior of the effective domain of f.*

Proof. The assumption implies that there is $\epsilon > 0$ such that all $f(x + w) < \infty$ for all w with $\|w\| \leq \epsilon$. Since $(x, f(x))$ is on the boundary of the convex set epi f, the supporting hyperplane result, Theorem A.16, implies that there exists a nonzero vector $c \in \mathbb{R}^n$ and a scalar $\beta \in \mathbb{R}$ such that

$$\begin{bmatrix} c \\ \beta \end{bmatrix}^T \left(\begin{bmatrix} z \\ t \end{bmatrix} - \begin{bmatrix} x \\ f(x) \end{bmatrix} \right) \leq 0, \quad \text{for all } (z, t) \in \operatorname{epi} f. \tag{8.3}$$

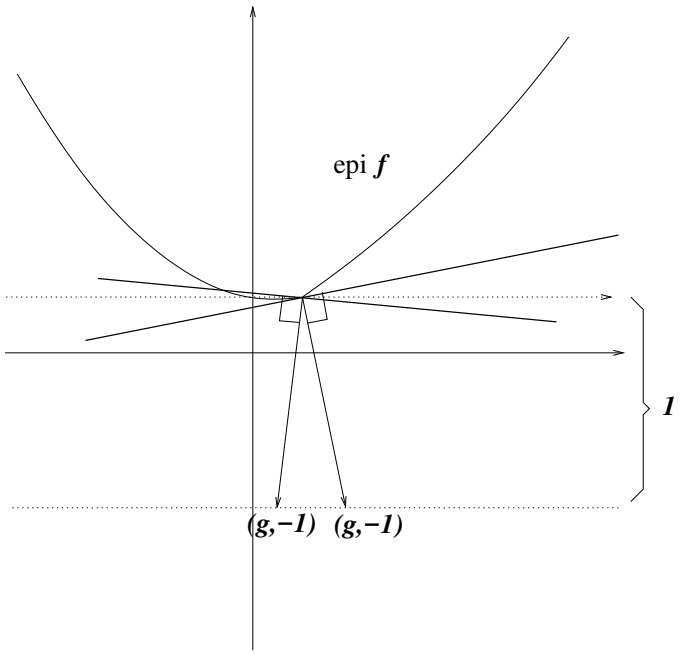

Figure 8.1: Theorem 8.3 illustrated: $g \in \partial f(x)$ if and only if $(g, -1)$ defines a supporting hyperplane to epi f at x.

We can replace t in this equation by any $\bar{t} > t$ and still have $(z, \bar{t}) \in \text{epi}\, f$. Thus, we cannot have $\beta > 0$, since then by driving $\bar{t}$ to $+\infty$, (8.3) will fail to hold for $\bar{t}$ sufficiently large. We cannot have $\beta = 0$ either, since then by setting $z = x + w$ in (8.3), we have that $c^T w \leq 0$ for all w with $\|w\| \leq \epsilon$, which does not hold for $w = \epsilon c/\|c\|$. (Recall that c is nonzero.) Thus we must have $\beta < 0$, and by setting $t = f(z)$ in (8.3), rearranging, and dividing both sides by $-\beta$, we obtain

$$c^T(z - x) \leq -\beta(f(z) - f(x)) \;\Rightarrow\; f(z) \geq f(x) + (-c/\beta)^T(z - x),$$

which implies that $-c/\beta$ is a subgradient of f at x. □

Lemma 8.4 shows that when x is in the interior of the effective domain, the subdifferential $\partial f(x)$ is nonempty. The same condition implies that $\partial f(x)$ is bounded and in fact compact, as we show next.

Lemma 8.5. *If x is in the interior of the effective domain of f, the subdifferential $\partial f(x)$ is compact.*

Proof. As in the proof of Lemma 8.4, there exists $\epsilon > 0$ such that $f(x + w) < \infty$ for all $\|w\| \leq \epsilon$. Suppose for contradiction that $\partial f(x)$ is unbounded. Then we can choose a sequence $\{g_k\}$ with $g_k \in \partial f(x)$ for all $k = 1, 2, \ldots$ and $\|g_k\| \to \infty$. Since all normalized vectors $g_k/\|g_k\|$ are in the unit ball, which is compact, we can assume by taking a subsequence if necessary that $g_k/\|g_k\| \to \bar{g}$ for some $\bar{g}$ with $\|\bar{g}\| = 1$. Note that $g_k^T \bar{g}/\|g_k\| \to 1$, from which it follows that $g_k^T \bar{g} \to \infty$. From the definition of subgradient, we have

$$f(x + \epsilon\bar{g}) \geq f(x) + \epsilon g_k^T \bar{g}, \quad k = 1, 2, \ldots,$$

so by driving $k \to \infty$, we deduce that $f(x + \epsilon\bar{g}) = \infty$, yielding the contradiction.

We have proved boundedness. Since, as we remarked earlier, $\partial f(x)$ is closed, compactness follows, completing the proof. □

If f is convex and differentiable at x, the subgradient coincides with the gradient.

Theorem 8.6. *If f is convex and differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$.*

Proof. Differentiability of f implies that for all vectors $d \in \mathbb{R}^n$ with $\|d\| = 1$, we have $f(x + td) = f(x) + t\nabla f(x)^T d + o(|t|)$ (see (2.6)). In particular, f is finite at all points in the neighborhood of x, so x is in the interior of the effective domain of f, so it follows from Lemma 8.4 that $\partial f(x)$ is nonempty.

Let v be an arbitrary vector in $\partial f(x)$. From Definition 8.1 we have

$$f(x + td) = f(x) + t\nabla f(x)^T d + o(t) \geq f(x) + tv^T d \Rightarrow \nabla f(x)^T d + o(t)/t \geq v^T d,$$
$$f(x - td) = f(x) - t\nabla f(x)^T d + o(t) \geq f(x) - tv^T d \Rightarrow \nabla f(x)^T d + o(t)/t \leq v^T d.$$

By combining these inequalities and taking $t \downarrow 0$, we have that $(v - \nabla f(x))^T d = 0$ for all unit vectors d, from which it follows immediately that $v = \nabla f(x)$. □

A converse of this result is also true: If the subdifferential of a convex function f at x contains a single subgradient, then f is differentiable with gradient equal to this subgradient (see [28, Theorem 25.1]).

8.2 The Subdifferential and Directional Derivatives

We turn now to directional derivatives. Given a function $f : \mathbb{R}^n \to \mathbb{R}$, the directional derivative of f at $x \in \operatorname{dom} f$ in the direction $v \neq 0$ is denoted by $f'(x; v)$ and defined by

$$f'(x; v) := \lim_{\alpha \downarrow 0} \frac{f(x + \alpha v) - f(x)}{\alpha}. \tag{8.4}$$

This definition holds for any function f, but our focus is again on convex functions here. [2] The definition suggests that a direction v for which $f'(x; v) < 0$ is a descent direction for f, and thus a useful direction when our goal is to minimize f. In this context, note that directional derivatives in all directions are nonnegative if and only if x^* is a minimizer of f.

Theorem 8.7. *Suppose that f is a convex function. Then for some $x^* \in \operatorname{dom} f$, $f'(x^*; v) \geq 0$ for all v if and only if x^* is a minimizer of f.*

Proof. If x^* is a minimizer of f, then $f(x^* + \alpha v) \geq f(x^*)$ for all $\alpha > 0$ and all v, so it follows directly from the definition (8.4) that $f'(x^*; v) \geq 0$. Conversely, suppose that x^* is not a minimizer of f. Then there exists some $z^* \in \operatorname{dom} f$ with $f(z^*) < f(x^*)$, and for any $\alpha \in (0, 1)$, we have

$$f(x^* + \alpha(z^* - x^*)) \leq (1 - \alpha)f(x^*) + \alpha f(z^*),$$

[2] Note that the limit can be infinite when f is an extended-value convex function. For example, the convex function $f : \mathbb{R} \to \mathbb{R}$ that has $f(0) = 0$ and $f(t) = +\infty$ for $t \neq 0$ has $f'(0, v) = +\infty$ for all $v \neq 0$.

and so

$$\frac{f(x^* + \alpha(z^* - x^*)) - f(x^*)}{\alpha} \leq f(x^*) - f(z^*) < 0, \quad \text{for all } \alpha \in (0,1).$$

By taking limits as $\alpha \downarrow 0$ and using (8.4), we have $f'(x^*; z^* - x^*) \leq f(x^*) - f(z^*) < 0$, completing the proof. □

In the remainder of this section, we explore the relationship between directional derivatives and subgradients, showing that knowledge of the subgradient makes it possible to compute descent directions for f.

For f convex, we have that the ratio in (8.4) is a nondecreasing function of α, that is

$$0 < \alpha_1 < \alpha_2 \;\Rightarrow\; \frac{f(x + \alpha_1 v) - f(x)}{\alpha_1} \leq \frac{f(x + \alpha_2 v) - f(x)}{\alpha_2}. \tag{8.5}$$

(The proof is a consequence of the definition of convexity; see the Exercises.) We can thus replace the definition (8.4) by

$$f'(x; v) := \inf_{\alpha > 0} \frac{f(x + \alpha v) - f(x)}{\alpha}. \tag{8.6}$$

It follows from these definitions that the directional derivative is additive, that is, for two convex functions f_1 and f_2, we have

$$(f_1 + f_2)(x; v) = f_1(x; v) + f_2(x; v). \tag{8.7}$$

Moreover, it is homogeneous with respect to the direction, that is,

$$f'(x; \lambda v) = \lambda f'(x; v), \quad \text{for all } \lambda \geq 0. \tag{8.8}$$

(We leave proofs of these results as exercises.) Moreover, $f'(x; v)$, *regarded as a function of* v, *for fixed* x, is a convex function. We see this from the following elementary argument: Given v_1 and v_2 and $\gamma \in (0,1)$, consider $f'(x; \gamma v_1 + (1-\gamma)v_2)$, for which we have

$$\begin{aligned}
f'(x; \gamma v_1 + (1-\gamma)v_2) &= \lim_{\alpha \downarrow 0} \frac{f(x + \alpha\gamma v_1 + \alpha(1-\gamma)v_2) - f(x)}{\alpha} \\
&= \lim_{\alpha \downarrow 0} \frac{f(\gamma(x + \alpha v_1) + (1-\gamma)(x + \alpha v_2)) - \gamma f(x) - (1-\gamma) f(x)}{\alpha} \\
&\leq \lim_{\alpha \downarrow 0} \frac{\gamma(f(x + \alpha v_1) - f(x)) + (1-\gamma)(f(x + \alpha v_2) - f(x))}{\alpha} \\
&= \gamma \lim_{\alpha \downarrow 0} \frac{f(x + \alpha v_1) - f(x)}{\alpha} + (1-\gamma) \lim_{\alpha \downarrow 0} \frac{f(x + \alpha v_2) - f(x)}{\alpha} \\
&= \gamma f'(x; v_1) + (1-\gamma) f'(x; v_2).
\end{aligned}$$

It follows from the definition (8.4) and Taylor's Theorem (specifically, (2.3)) that if f is differentiable at x, we have for any $v \in \mathbb{R}^n$ that

$$\begin{aligned}
f'(x; v) &= \lim_{\alpha \downarrow 0} \frac{f(x + \alpha v) - f(x)}{\alpha} \\
&= \lim_{\alpha \downarrow 0} \frac{\nabla f(x + \gamma\alpha v)^T (\alpha v)}{\alpha} \quad \text{for some } \gamma \in (0,1) \\
&= \nabla f(x)^T v.
\end{aligned}$$

Thus in particular, we have

$$f'(x;v) = -f'(x;-v) \quad \text{when } f \text{ is differentiable at } x. \tag{8.9}$$

This equality is not true for nonsmooth functions at points of nondifferentiability. For example, the hinge loss function $h(t) = \max(t, 0)$ has $h'(0;1) = 1$ but $h'(0;-1) = 0$. Similarly, the absolute value funtion $h(t) = |t|$ has $h(0;1) = 1$ and $h'(0,-1) = 1$. We give a generalization of (8.9) in Corollary 8.9 below.

It follows from the second definition (8.6) that $f'(x;v)$ has a property reminiscent of the subgradient (compare with Definition 8.1):

$$f(x+\alpha v) \geq f(x) + \alpha f'(x;v), \quad \text{for all } \alpha \geq 0. \tag{8.10}$$

Related to this observation, we can prove the following result.

Theorem 8.8. *Suppose that x is in the interior of the effective domain of the convex function f. Then for any $v \in \mathbb{R}^n$, we have that*

$$f'(x;v) = \sup_{g \in \partial f(x)} g^T v. \tag{8.11}$$

Proof. From (8.6), we have for any $g \in \partial f(x)$ that

$$f'(x;v) = \inf_{\alpha > 0} \frac{f(x+\alpha v) - f(x)}{\alpha} \geq \inf_{\alpha > 0} \frac{\alpha g^T v}{\alpha} = g^T v,$$

so that

$$f'(x;v) \geq g^T v \quad \text{for all } g \in \partial f(x). \tag{8.12}$$

Because $\partial f(x)$ is closed, we obtain equality in (8.11) if we can find $\hat{g} \in \partial f(x)$ such that $f'(x;v) = \hat{g}^T v$. For this, we use the convexity of $f'(x;y)$ with respect to its second argument y for all $y \in \mathbb{R}^n$ (proved above). By Lemma 8.4, there exists a subgradient of $f'(x;\cdot)$ at v; let us call it $\hat{g}$. By the definition of subgradient, together with (8.8), we have for all $\lambda \geq 0$ and all y that

$$\lambda f'(x;y) = f'(x;\lambda y) \geq f'(x;v) + \hat{g}^T(\lambda y - v). \tag{8.13}$$

Letting $\lambda \uparrow \infty$, we have that $\hat{g}^T y \leq f'(x;y)$. By the definition (8.6), we have

$$f(x+y) - f(x) \geq \inf_{\alpha > 0} \frac{f(x+\alpha y) - f(x)}{\alpha} = f'(x;y) \geq \hat{g}^T y, \quad \text{for all } y \in \mathbb{R}^n,$$

which implies that $\hat{g} \in \partial f(x)$, so by (8.12) we have that $f'(x;v) \geq \hat{g}^T v$. On the other hand, by taking $\lambda = 0$ in (8.13), we have that $f'(x;v) \leq \hat{g}^T v$. Therefore, $f'(x;v) = \hat{g}^T v$ for this particular $\hat{g} \in \partial f(x)$, completing the proof. □

An immediate corollary of this result leads to a generalization of (8.9).

Corollary 8.9. *Suppose that x is in the interior of the effective domain of the convex function f. Then for any $v \in \mathbb{R}^n$, we have*

$$f'(x;v) \geq -f'(x;-v).$$

Proof. From Theorem 8.8, we have

$$f'(x;-v) = \sup_{g\in\partial f(x)} g^T(-v) = -\inf_{g\in\partial f(x)} g^T v \geq -\sup_{g\in\partial f(x)} g^T v = -f'(x;v).$$

□

We conclude with another result that relates subgradients to directional derivatives.

Theorem 8.10. *Suppose that $x \in \operatorname{dom} f$ for the convex function f, and that there is some vector $g \in \mathbb{R}^n$ such that for all $v \in \mathbb{R}^n$, we have*

$$f'(x;v) \geq g^T v.$$

Then $g \in \partial f(x)$.

Proof. By setting $\alpha_2 = 1$ and $\alpha_1 \downarrow 0$ in (8.5) and using the definition (8.4), we have

$$g^T v \leq f'(x;v) \leq f(x+v) - f(x), \quad \text{for all } v \in \mathbb{R}^n.$$

Thus, g satisfies the definition (8.2), so that $g \in \partial f(x)$. □

8.3 Calculus of Subdifferentials

In this section we describe the key properties of subdifferentials that are the key to *calculating* subgradients. Unlike for differentiable functions, there are only a few key rules that are commonly used in practice, involving positive combinations, combinations with linear mapping, and partial maximization. We collect these rules here in Theorems 8.11, 8.12, and 8.13. (Proofs of Theorems 8.11, 8.12 appear at the end of this section.)

We start with some elementary rules of subdifferential calculus.

Theorem 8.11. *Supposing that f, f_1, and f_2 are convex functions and α is a positive scalar, the following are true.*

$$\partial(f_1 + f_2)(x) \supset \partial f_1(x) + \partial f_2(x), \tag{8.14}$$

$$\partial(\alpha f)(x) = \alpha \partial f(x). \tag{8.15}$$

If, in addition, x is in the interior of the effective domain for both f_1 and f_2, then equality holds in (8.14), that is, $\partial(f_1 + f_2)(x) = \partial f_1(x) + \partial f_2(x)$. In particular, if f_1 and f_2 are finite-valued convex functions then $\partial(f_1 + f_2)(x) = \partial f_1(x) + \partial f_2(x)$ for all x.

We emphasize that the relationship in (8.14) is not an equality in general. We will see an example of strict inclusion in the next section. However, equality holds in some interesting special cases.[3]

The next result allows us to compute the subdifferential under affine transformations.

Theorem 8.12. *[4, Theorem 4.2.5 (a)] Suppose that $f : \mathbb{R}^m \to \mathbb{R}$ is a convex function and define $h(x) := f(Ax + b)$ for some matrix $A \in \mathbb{R}^{m\times n}$ and vector $b \in \mathbb{R}^m$. Suppose that $Ax + b$ is in the interior of* $\operatorname{dom} f$*. Then $\partial h(x) = A^T \partial f(Ax + b)$.*

[3] See [8] for some other conditions that give rise to equality.

The final result is Theorem 8.13, known as Danskin's Theorem, which shows us how to compute the subdifferential of a function that is defined as the pointwise maximum of a possibly infinite set of functions. Such functions are ubiquitous objects in optimization, particularly in data analysis applications.

The setup is as follows. Let $I \subset \mathbb{R}^n$ be a compact set. (Sets with finite cardinality are a useful special case.) Let $\varphi : \mathbb{R}^d \times I \to \mathbb{R}$ be a family of functions, continuous in (x, i), and assume that each $\varphi(\cdot, i)$, $i \in I$, is convex. We define

$$f(x) := \max_{i \in I} \varphi(x, i) \tag{8.16}$$

Note that f is convex, because it is the pointwise maximum of convex functions (see the Exercises). Following Section 8.2, we denote the directional derivative of $\varphi(\cdot, i)$ at x in direction y by $\varphi'(x, i; y)$. For each x, we define $I_{\max}(x)$ to be the subset of I for which the maximum is achieved in (8.16), that is,

$$I_{\max}(x) := \arg\max_{j \in I} \varphi(x, j) = \{j \ : \ f(x) = \varphi(x, j)\}. \tag{8.17}$$

Note that $I_{\max}(x)$ is nonempty (by compactness of I) and compact for all x, by continuity of $\varphi(x, \cdot)$ with respect to its second argument. Danskin's Theorem describes the directional derivatives and subdifferentials of f.[4]

Theorem 8.13 (Danskin's Theorem)**.** *(a) The directional derivative of f defined by* (8.16) *at x in direction y is given by*

$$f'(x, y) = \max_{i \in I_{\max}(x)} \varphi'(x, i; y)\,.$$

(b) If, in addition to the conditions on the function family φ stated above, we have that $\varphi(\cdot, i)$ is a differentiable function of x for all $i \in I$, with $\nabla_x \varphi(x, \cdot)$ continuous on I for all x, then

$$\partial f(x) = \operatorname{conv}\{\nabla_x \varphi(x, i) \ : \ i \in I_{\max}(x)\}.$$

We refer to [2, Section B.5] and [4, Proposition 4.5.1] for proofs of Theorem 8.13, which are quite technical. We provide below proofs of Theorems 8.11 and 8.12. Generally speaking, one direction of the inclusion is quite straightforward, while the other direcion requires a separating hyperplane argument. Practitioners can skip these proofs, but we note that the arguments are of interest in that they highlight some important structural aspects of convex optimization.

Proof of Theorem 8.11. The proofs of (8.14) and (8.15) are immediate consequences of the definitions of subgradients. For the case in which x is in the interior of both $\operatorname{dom} f_1$ and $\operatorname{dom} f_2$, Lemmas 8.4 and 8.5 show that $\partial f_1(x)$, $\partial f_2(x)$, and $\partial(f_1 + f_2)(x)$ are all nonempty, convex, and compact sets. Suppose for contradiction that the inclusion (8.14) is strict in this case, that is, there exists $g \in \partial(f_1 + f_2)(x)$ such that $g \notin \partial f_1(x) + \partial f_2(x)$. By the strict separation result Lemma A.13, setting $X = (\partial f_1(x) + \partial f_2(x)) - \{g\}$, there is a vector $\bar{t} \in \mathbb{R}^n$ and a scalar $\alpha > 0$ such that

$$\bar{t}^T(g_1 + g_2) \leq \bar{t}^T g - \alpha, \quad \text{for all } g_1 \in \partial f_1(x) \text{ and all } g_2 \in \partial f_2(x).$$

[4] Interestingly, this theorem first arose out of cold-war research by Danskin and appeared in a 1967 monograph called *The Theory of Max-Min and its Applications to Weapons Allocation Problems* [12].

From results about the relationship between subdifferentials and directional derivatives — (8.7) and Theorem 8.8 — we have

$$(f_1 + f_2)(x; \bar{t}) = f_1(x; \bar{t}) + f_2(x; \bar{t}) = \sup_{g_1 \in \partial f_1(x)} g_1^T \bar{t} + \sup_{g_2 \in \partial f_2(x)} g_2^T \bar{t} \leq \bar{t}^T g - \alpha < g^T \bar{t}.$$

Thus, by Theorem 8.8, we have $g \notin \partial(f_1 + f_2)(x)$, a contradiction.

When f_1 and f_2 are finite-valued, we have $\operatorname{dom} f_1 = \operatorname{dom} f_2 = \mathbb{R}^n$, so that the effective domain condition holds for all $x \in \mathbb{R}^n$, and the result follows. □

Proof of Theorem 8.12. Since $Ax + b$ is in the interior of $\operatorname{dom} f$, x is in the interior of $\operatorname{dom} h$. Thus, by Lemma 8.4 and 8.5, the subdifferentials $\partial h(x)$ and $\partial f(Ax+b)$ are nonempty and compact. From the definition (8.4) of directional derivatives, it follows that

$$h'(x; y) = f'(Ax + b; Ay), \quad \text{for any } y \in \mathbb{R}^n.$$

From Theorem 8.8, we have for any $z \in \mathbb{R}^m$ that

$$g^T z \leq f'(Ax + b; z) \quad \text{for all } g \in \partial f(Ax + b).$$

By setting $z = Ay$, we have

$$(A^T g)^T y = g^T (Ay) \leq f'(Ax + b; Ay) = h'(x; y), \quad \text{for any } y \in \mathbb{R}^n.$$

It follows from Theorem 8.10 that $A^T g \in \partial h(x)$, and since this result holds for all $g \in \partial f(Ax + b)$, we have that $A^T \partial f(Ax + b) \subset \partial h(x)$.

To prove equality, suppose for contradiction that there is a vector $v \in \partial h(x)$ such that $v \notin A^T \partial f(Ax + b)$. Since the set $\Omega := A^T \partial f(Ax + b)$ is compact, we invoke the strict separation result Theorem A.15 to deduce the existence of a vector y and a scalar β such that

$$y^T (A^T g) < \beta < y^T v, \quad \text{for all } g \in \partial f(Ax + b).$$

It follows that

$$\sup_{g \in \partial f(Ax+b)} (Ay)^T g < y^T v\,.$$

Theorem 8.8 then implies that

$$h'(x, y) = f'(Ax + b; Ay) < y^T v,$$

contradicting the assumption that $v \in \partial h(x)$ and completing the proof. □

8.4 Convex Sets and Convex Constrained Optimization

In this section, we note the connections between closed convex sets and the indicator functions for those sets (which are extended-value convex functions, that is, they may take infinite values at some points).

Let $\Omega \subset \mathbb{R}^n$ be a convex set (see (2.13) for the definition of convexity). The indicator function I_Ω for a convex set Ω, defined by

$$I_\Omega(x) = \begin{cases} 0 & \text{if } x \in \Omega, \\ \infty & \text{if } x \notin \Omega. \end{cases}$$

This function is convex, extended-valued (except for the trivial case $\Omega = \mathbb{R}^n$), and has $\operatorname{dom} I_\Omega = \Omega$. When Ω is a closed set, $I_\Omega(x)$ is also lower-semicontinuous. We have the following result.

Theorem 8.14. *For a closed convex set $\Omega \subset \mathbb{R}^n$, we have that $N_\Omega(x) = \partial I_\Omega(x)$ for all $x \in \Omega$.*

Proof. Given $v \in N_\Omega(x)$, we have

$$I_\Omega(y) - I_\Omega(x) = 0 - 0 = 0 \geq v^T(y - x), \quad \text{for all } y \in \Omega = \operatorname{dom} I_\Omega,$$

which implies that $v \in \partial I_\Omega(x)$, by Definition 8.1. For the converse, suppose that $v \in \partial I_\Omega(x)$, so we have

$$0 = I_\Omega(y) \geq I_\Omega(x) + v^T(y - x) = v^T(y - x), \quad \text{for all } y \in \Omega,$$

which implies that $v \in N_\Omega(x)$, completing the proof. □

In optimization, we often deal with sets that are intersections of closed convex sets. We have the following result for normal cones of such intersections.

Theorem 8.15. *Let Ω_i, $i = 1, 2, \ldots, m$ be closed convex sets and let $\Omega = \cap_{i=1,2,\ldots,m} \Omega_i$. Then for $x \in \Omega$, we have*

$$N_\Omega(x) \supset N_{\Omega_1}(x) + N_{\Omega_2}(x) + \ldots + N_{\Omega_m}(x). \tag{8.18}$$

Proof. This follows immediately from (8.14) when we make the identification in Theorem 8.14. For a direct proof, we can proceed as follows. Consider vectors $v_i \in N_{\Omega_i}(x)$ for all $i = 1, 2, \ldots, m$, and define $v := \sum_{i=1}^m v_i$. Let z be any point in the intersection $\Omega = \cap_{i=1}^m \Omega_i$. Since $z \in \Omega_i$, we have $v_i^T(z - x) \leq 0$ for all $i = 1, 2, \ldots, m$, so that $v^T(z - x) = (\sum_{i=1}^m v_i)^T(z - x) \leq 0$, and thus $v \in N_\Omega(x)$. □

The following example, illustrated in Figure 8.2, shows that strict inclusion can hold in (8.18). Define the following two convex subsets of $\mathbb{R}^2$:

$$\Omega_1 := \{x \in \mathbb{R}^2 \,:\, x_1 \leq 0\}, \quad \Omega_2 := \{x \in \mathbb{R}^2 \,:\, (x_1 - 1)^2 + x_2^2 \leq 1\}, \tag{8.19}$$

for which clearly $\Omega_1 \cap \Omega_2 = \{0\}$. The normal cones at the interesting point 0 are

$$N_{\Omega_1}(0) = \left\{ \begin{bmatrix} v_1 \\ 0 \end{bmatrix} \,:\, v_1 \geq 0 \right\}, \quad N_{\Omega_2}(0) = \left\{ \begin{bmatrix} v_1 \\ 0 \end{bmatrix} \,:\, v_1 \leq 0 \right\}, \quad N_{\Omega_1 \cap \Omega_2}(0) = \mathbb{R}^2. \tag{8.20}$$

Since $N_{\Omega_1}(0) + N_{\Omega_1}(0) = \mathbb{R} \times \{0\}$, strict inclusion holds. Note that this example also shows that strict including can hold in (8.14), when we identify the normal cones with indicator function subdifferentials as in Theorem 8.14.

Additional conditions are sometimes assumed to ensure that equality holds in (8.18); these conditions are called *constraint qualifications*. Some constraint qualifications are expressed in terms of the geometry of the sets while others focus on their algebraic descriptions. One common theme

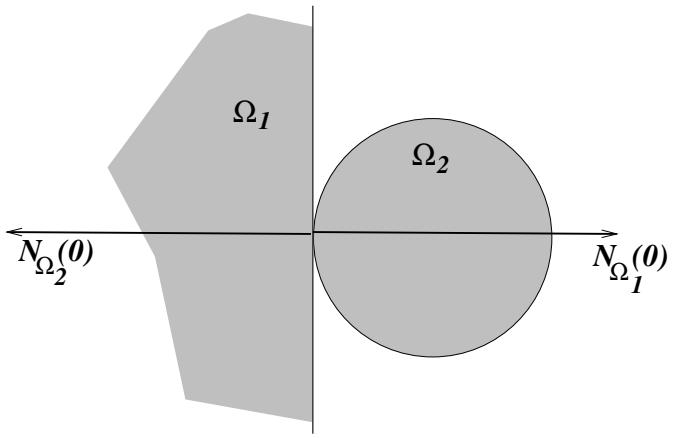

Figure 8.2: Example for which strict inclusion holds in (8.18).

among constraint qualifications is that a linear approximation of the sets near the point in question needs to capture the essential geometry of the set itself in a neighborhood of the point. This is not true of the example above, where the tangents (linear approximations) to Ω_1 and Ω_2 at $x = 0$ are the vertical axis (so the intersection of their linear approximations is also the vertical axis), while the intersection of the two sets is the single point $\{0\}$, so the linear approximation to this set is simply $\{0\}$.

We conclude by applying Definition 7.1 to obtain optimality conditions for the convex constrained optimization problem

$$\min_{x\in\Omega} f(x), \quad \text{where } f \text{ is a smooth convex function and } \Omega \subset \mathbb{R}^n \text{ is a closed convex set.} \tag{8.21}$$

Recall the definition for a solution of this problem, which follow (2.1). In particular, we say that x^* is a global solution of (8.21) if $x^* \in \Omega$ and $f(x) \geq f(x^*)$ for all $x \in \Omega$. (There may be multiple values of x^* that satisfy this definition; the set of all such values is convex; see the Exercises.) We have the following characterization of optimality.

Theorem 8.16. *The point $x^* \in \Omega$ is a solution of (8.21) if and only if $-\nabla f(x^*) \in N_\Omega(x^*)$.*

Proof. Suppose that $-\nabla f(x^*) \in N_\Omega(x^*)$. From Definition 7.1, we have

$$-\nabla f(x^*)^T(y - x^*) \leq 0 \quad \text{for all } y \in \Omega,$$

so by convexity of f, we have

$$f(y) \geq f(x^*) + \nabla f(x^*)^T(y - x^*) \geq f(x^*),$$

proving that x^* is a solution. Suppose instead that $-\nabla f(x^*) \notin N_\Omega(x^*)$. Then there is some $y \in \Omega$ such that $-\nabla f(x^*)^T(y - x^*) > 0$. By convexity of Ω, the point $y_\alpha := x^* + \alpha(y - x^*) \in \Omega$ for all $\alpha \in [0, 1]$. By Taylor's theorem, in particular (2.3), we have that

$$f(y_\alpha) = f(x^*) + \nabla f(x^* + \gamma_\alpha \alpha(y - x^*))^T(y - x^*), \quad \text{for some } \gamma_\alpha \in (0, 1).$$

By taking α sufficiently small in this expression we have by continuity of ∇f that $\nabla f(x^* + \gamma_\alpha \alpha(y - x^*))^T(y - x^*) < (1/2)\nabla f(x^*)^T(y - x^*) < 0$, so that $f(y_\alpha) < f(x^*)$ for all α sufficiently small. Therefore x^* is not a solution of (8.21) when $-\nabla f(x^*) \notin N_\Omega(x^*)$. □

Interestingly, because of the identity in Theorem 8.14, we can write the condition $-\nabla f(x^*) \in N_\Omega(x^*)$ alternatively as follows:

$$0 \in \nabla f(x^*) + \partial I_\Omega(x^*). \tag{8.22}$$

Moreover, by (8.14) and Theorem 8.6, it is a consequence of (8.22) that

$$0 \in \partial \left(f(x^*) + I_\Omega(x^*) \right). \tag{8.23}$$

From Theorem 8.2, this condition in turn is true if and only if x^* is a minimizer of the "unconstrained" problem

$$\min_x \; f(x) + I_\Omega(x),$$

and we can see easily that this problem is equivalent to (8.21).

8.5 Optimality Conditions for Composite Nonsmooth Functions

We now consider first-order optimality conditions for functions of the form

$$\phi(x) := f(x) + \psi(x), \tag{8.24}$$

where f is a smooth function and ψ is (possibly) nonsmooth, convex and finite-valued. (Because of the latter property, the effective domain of ψ is the entire space $\mathbb{R}^n$, so we can apply such results as Theorem 8.11.) This is a type of objective that we encounter often in machine learning applications; see Chapter 1.

We deal first with the case in which f is convex too.

Theorem 8.17. *When f is convex and differentiable and ψ is convex and finite-valued, the point x^* is a minimizer of ϕ defined in* (8.24) *if and only if $0 \in \nabla f(x^*) + \partial\phi(x^*)$.*

Proof. By Theorem 8.6, we have that $\partial f(x) = \{\nabla f(x)\}$, so by using the fact that $\psi(x)$ has effective domain $\mathbb{R}^n$ and applying Theorem 8.11, we have

$$\partial\phi(x) = \nabla f(x) + \partial\psi(x).$$

The result follows immediately from Theorem 8.2. □

When f is strongly convex, the problem (8.24) has a minimizer and it is unique.

Theorem 8.18. *Suppose that the conditions of Theorem 8.17 hold, and in addition that f is strongly convex. Then the problem* (8.24) *has a unique solution.*

For the more general case in which f is possibly nonconvex, we have a first-order necessary condition.

Theorem 8.19. *Suppose that f is differentiable and ψ is convex and finite-valued, and let ϕ be defined by* (8.24)*. Then if x^* is a local minimizer of ϕ, we have that $0 \in \nabla f(x^*) + \partial\psi(x^*)$.*

Proof. Supposing that $0 \notin \nabla f(x^*) + \partial\psi(x^*)$, we show that x^* cannot be a local minimizer. We define the following convex approximation to $\phi(x+d)$:

$$\bar{\phi}(d) := f(x^*) + \nabla f(x^*)^T d + \psi(x^* + d),$$

By differentiability of f we have that for all $\alpha \in [0,1]$ and for any d that $\bar{\phi}(\alpha d) = \phi(x+\alpha d) + o(\alpha|d|)$. Since by assumption $0 \notin \partial\bar{\phi}(0) = \nabla f(x^*) + \partial\psi(x^*)$, we have from Theorem 8.2 that 0 is not a minimizer of $\bar{\phi}(d)$. Hence there exists $\bar{d}$ with $\bar{\phi}(\bar{d}) < \bar{\phi}(0)$, so that the quantity $c := \bar{\phi}(0) - \bar{\phi}(\bar{d})$ is strictly positive. By convexity of $\bar{\phi}$, we have for all $\alpha \in [0,1]$ that

$$\bar{\phi}(\alpha\bar{d}) \leq \bar{\phi}(0) - \alpha(\bar{\phi}(0) - \bar{\phi}(\bar{d})) = \phi(x^*) - \alpha c,$$

and therefore

$$\phi(x^* + \alpha\bar{d}) \leq \phi(x^*) - \alpha c + o(\alpha|d|).$$

Therefore $\phi(x^* + \alpha\bar{d}) < \phi(x^*)$ for all $\alpha > 0$ sufficiently small, so x^* is not a local minimizer of ϕ. □

8.6 Proximal Operators and the Moreau Envelope

We define here the proximal operator that is a key component of algorithms for regularized optimization, and analyze some of its properties in preparation for convergence analysis of proximal-gradient algorithms in Section 9.5. The proximal operator is a powerful generalization of Euclidean projections, and will allow us to greatly enhance our nonsmooth optimization toolbox.

For a closed proper convex function h we define the *prox-operator* of the h as

$$\text{prox}_h(x) := \arg\min_u \left\{ h(u) + \frac{1}{2}\|u - x\|^2 \right\}. \tag{8.25}$$

Note that this is a well defined function because of the strong convexity of the Euclidean norm.

When $h(x) = I_\Omega(x)$, the indicator function for a closed convex set Ω, $\text{prox}_{I_\Omega}(x)$ is simply the Euclidean projection of x onto the set Ω, as we see from the following argument:

$$\text{prox}_{I_\Omega}(x) = \arg\min_u \left\{ I_\Omega(u) + \frac{1}{2}\|u - x\|^2 \right\} = \arg\min_{u \in \Omega} \frac{1}{2}\|u - x\|^2 .$$

Proximal operators are more general than Euclidean projections, but they satisfy a similar nonexpansiveness property.

Proposition 8.20. *Suppose h is a convex function. Then*

$$\|\text{prox}_h(x) - \text{prox}_h(y)\| \leq \|x - y\|$$

Proof. From optimality properties, we have from (8.25) that

$$0 \in \lambda\partial h(\text{prox}_{\lambda h}(x)) + (\text{prox}_{\lambda h}(x) - x). \tag{8.26}$$

Rearranging these expressions, at two points x and y, we have

$$x - \text{prox}_h(x) \in \partial(\text{prox}_h(x)), \quad y - \text{prox}_h(y) \in \partial(\text{prox}_h(y)).$$

Now, for a convex function f, it follows from the definition of subgradients that if $a \in \partial f(x)$ and $b \in \partial f(y)$, we have $(a-b)^T(x-y) \geq 0$. By applying this inequality, we have

$$(1/\lambda)\big((x - \operatorname{prox}_h(x)) - (y - \operatorname{prox}_h(y))\big)^T(\operatorname{prox}_h(x) - \operatorname{prox}_h(y)) \geq 0,$$

which by rearrangement and application of the Cauchy-Schwartz inequality yields

$$\|\operatorname{prox}_h(x) - \operatorname{prox}_h(y)\|^2 \leq (x-y)^T(\operatorname{prox}_h(x) - \operatorname{prox}_h(y)) \leq \|x-y\| \, \|\operatorname{prox}_h(x) - \operatorname{prox}_h(y)\|,$$

from which we obtain the proposition. □

We note several special cases of the prox operator which are useful in later chapters.

- $h(x) = 0$ for all x, for which we have $\operatorname{prox}_h(x) = 0$. (Though trivial, this observation is useful in proxing that the prox-gradient method of Chapter 9 reduces to the familiar steepest descent method when the objective contains no regularization term.)

- $h(x) = \lambda\|x\|_1$. By substituting into definition (8.25) we see that the minimization separates into its n separate components, and that the ith component of $\operatorname{prox}_{\lambda\|\cdot\|_1}$ is

$$[\operatorname{prox}_{\lambda\|\cdot\|_1}]_i = \arg\min_{u_i} \left\{ \lambda|u_i| + \frac{1}{2}(u_i - x_i)^2 \right\}.$$

  It is not hard to verify that

$$[\operatorname{prox}_{\lambda\|\cdot\|_1}(x)]_i = \begin{cases} x_i - \lambda & \text{if } x_i > \lambda; \\ 0 & \text{if } x_i \in [-\lambda, \lambda]; \\ x_i + \lambda & \text{if } x_i < -\lambda, \end{cases} \tag{8.27}$$

  an operator that is known as *soft thresholding*.

- $h(x) = \lambda\|x\|_0$, where $\|x\|_0$ denotes the *cardinality* of the vector x, its number of nonzero components. Although this h is not a convex function (as we can see by considering convex combinations of the vectors $(0,1)^T$ and $(1,0)^T$ in $\mathbb{R}^2$), its prox-operator is well defined to be the *hard thresholding* operation:

$$[\operatorname{prox}_{\lambda\|\cdot\|_0}(x)]_i = \begin{cases} x_i & \text{if } |x_i| \geq \sqrt{2\lambda}; \\ 0 & \text{if } |x_i| < \sqrt{2\lambda}. \end{cases}$$

  For the cardinality function, the definition (8.25) separates into n individual components, and the fixed price of λ for allowing u_i to be nonzero is not worth paying unless $|x_i| \geq \sqrt{2\lambda}$.

We close this section by noting that the proximity operator is closely related to smooth approximations of convex functions. For a closed proper convex function h and a positive scalar λ, we define the *Moreau envelope* as

$$M_{\lambda,h}(x) := \inf_u \left\{ h(u) + \frac{1}{2\lambda}\|u - x\|^2 \right\} = \frac{1}{\lambda} \inf_u \left\{ \lambda h(u) + \frac{1}{2}\|u - x\|^2 \right\}. \tag{8.28}$$

The prox-operator of the function λh is the value of u that achieves the infimum in (8.28).

The Moreau envelope can be seen as a smoothing of regularization of the function h. It has a finite value for all x, even when h take on infinite values for some $x \in \mathbb{R}^n$. In fact, it is differentiable everywhere: Its gradient is

$$\nabla M_{\lambda,h}(x) = \frac{1}{\lambda}(x - \operatorname{prox}_{\lambda h}(x)).$$

It is easy to check moreover that x^* is a minimizer of h if and only if it is a minimizer of $M_{\lambda,h}$.

Sources and Further Reading

Some material of this chapter can be found in the lecture of [31] on "Subgradients."

Further background on Moreau envelopes and the proximal mapping is given in [25].

Exercises

1. Prove that if f is convex and $x \in \operatorname{dom} f$, the subdifferential $\partial f(x)$ is closed and convex.

2. Prove (8.5) by applying the definition (2.14) of a convex function.

3. Show that any norm $f(x) := \|x\|$ has 0 as a subgradient at $x = 0$, that is, $0 \in \partial f(0)$. Show that $f(x)$ is not differentiable at $x = 0$. (Reminder: A norm $\|\cdot\|$ has the properties that (a) $\|x\| = 0$ if and only if $x = 0$; (b) $\|\alpha x\| = |\alpha|\|x\|$ for all scalars α and vectors x; (c) $\|x + y\| \leq \|x\| + \|y\|$ for all x, y.)

4. Prove the additivity property (8.7) and the homogeneity property (8.8) of directional derivatives.

5. For the following norm functions f over the vector space $\mathbb{R}^n$, find $\partial f(x)$ and $f'(x; v)$ for all x and v:

   (a) The ℓ_1 norm: $f(x) = \|x\|_1$;

   (b) The ℓ_∞ norm: $f(x) = \|x\|_\infty$;

   (c) The ℓ_2 (Euclidean) norm: $f(x) = \|x\|_2$.

6. Show that the pointwise maximum function f defined by (8.16) is convex, under the stated conditions on $\varphi(x, i)$ for $x \in \mathbb{R}^d$ and $i \in I$, where I is a compact set.

7. Find the subdifferential of the piecewise linear convex function $f : \mathbb{R}^n \to \mathbb{R}$ defined by

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

   where $a_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$, $i = 1, 2, \ldots, n$.

8. Prove the subdifferential calculus expressions (8.14) and (8.15).

9. Suppose that f is defined as a maximum of m convex functions, that is, $f(x) := \max_{i=1,2,\ldots,m} f_i(x)$, where each f_i is convex. Show that

$$\partial f(x) = \left\{ \sum_{i\,:\,f_i(x)=f(x)} \lambda_i v_i \;:\; v_i \in \partial f_i(x),\ \lambda_i \geq 0,\ \sum_{i\,:\,f_i(x)=f(x)} \lambda_i = 1 \right\}.$$

10. (a) Show that I_Ω is a convex function if and only if Ω is a convex set.
    (b) Show that Ω is a nonempty closed convex set if and only if $I_\Omega(x)$ is a closed proper convex function.

11. Show that the set of all solutions of (8.21) is a convex set.

12. Show that a closed proper convex function h and its Moreau envelope $M_{\lambda,h}$ have identical minimizers.

13. Calculate $\operatorname{prox}_{\lambda h}(x)$ and $M_{\lambda,h}(x)$ for $h(x) = \frac{1}{2}\|x\|_2^2$.