8 Convex Functions

Convex functions are prevalent in inference and learning, where optimization problems involving convex risks are commonplace. In this chapter, we review basic properties of smooth and nonsmooth convex functions and introduce the concept of subgradient vectors. In the next chapter, we discuss projections onto convex sets and the solution of convex optimization problems by duality arguments.

8.1 CONVEX SETS

Let $g(z) : \mathbb{R}^M \to \mathbb{R}$ denote a *real-valued* function of a possibly vector argument, $z \in \mathbb{R}^M$. We consider initially the case in which $g(z)$ is at least first- and second-order differentiable, meaning that the gradient vector, $\nabla_z g(z)$, and the Hessian matrix, $\nabla_z^2 g(z)$, exist and are well-defined at all points in the domain of the function. Later we comment on the case of nonsmooth convex functions, which are not differentiable at some locations.

To begin with, a set $\mathcal{S} \subset \mathbb{R}^M$ is said to be convex if for any pair of points $z_1, z_2 \in \mathcal{S}$, all points that lie on the line segment connecting z_1 and z_2 also belong to $\mathcal{S}$. Specifically,

$$\forall\, z_1, z_2 \in \mathcal{S} \;\text{ and }\; 0 \le \alpha \le 1 \implies \alpha z_1 + (1-\alpha) z_2 \in \mathcal{S} \tag{8.1}$$

Figure 8.1 illustrates this definition by showing two convex sets and one non-convex set in its first row. In the rightmost set, a segment is drawn between two points inside the set and it is seen that some of the points on the segment lie outside the set.

We will regularly deal with *closed* sets, including closed convex sets. A set $\mathcal{S}$ in any metric space (i.e., in a space with a distance measure between its elements) is said to be *closed* if any converging sequence of points in $\mathcal{S}$ converges to a point in $\mathcal{S}$. This characterization is equivalent to stating that the complement of $\mathcal{S}$ is an open set or that any point outside $\mathcal{S}$ has a neighborhood around it that is disjoint from $\mathcal{S}$. For example, the segment $[-1, 1]$ on the real line is a closed set, while $[-1, 1)$ is an open set.

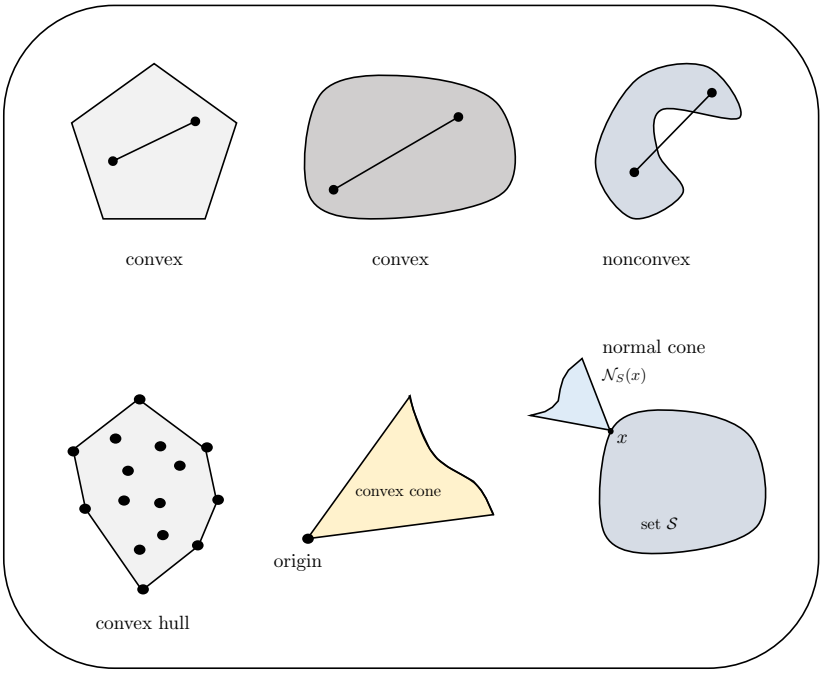

Figure 8.1 The two sets on the left in the first row are examples of convex sets, while the set on the right is nonconvex. The bottom row shows examples of a convex hull on the left, a convex cone in the middle, and a normal cone on the right. The curved line in the figure for the cones is meant to indicate that the cone extends indefinitely.

Example 8.1 (**Convex hull, convex cone, and conic hull**) Consider an arbitrary set $\mathcal{S} \subset \mathbb{R}^M$ that is not necessarily convex. The *convex hull* of $\mathcal{S}$, denoted by $\text{conv}(\mathcal{S})$, is the set of all convex combinations of elements in $\mathcal{S}$. Intuitively, the convex hull of $\mathcal{S}$ is the smallest convex set that contains $\mathcal{S}$. This situation is illustrated in the leftmost plot in the bottom row of Fig. 8.1. The dark circles represent the elements of $\mathcal{S}$, and the connected lines define the contour of the smallest convex set that contains $\mathcal{S}$.

A set $\mathcal{C} \subset \mathbb{R}^M$ is said to be a *cone* if for any element $z \in \mathcal{C}$ it holds that $tz \in \mathcal{C}$ for any $t \geq 0$. The cone is convex if $\alpha z_1 + (1-\alpha)z_2 \in \mathcal{C}$ for any $\alpha \in [0,1]$ and $z_1, z_2 \in \mathcal{C}$. One useful example of a convex cone is the *normal cone*. Consider a closed convex set $\mathcal{S} \subset \mathbb{R}^M$ and pick an arbitrary point $x \in \mathcal{S}$. We define the normal cone at point x, denoted by $\mathcal{N}_S(x)$, by considering all vectors y that satisfy

$$\mathcal{N}_S(x) \;=\; \Big\{ y \,|\, \text{such that } y^{\mathsf{T}}(s-x) \leq 0, \text{ for all } s \in \mathcal{S} \Big\} \quad (\textbf{normal cone}) \tag{8.2}$$

If x happens to lie in the interior of $\mathcal{S}$, then $\mathcal{N}_S(x) = \{0\}$.

The conic hull of a set $\mathcal{S} \subset \mathbb{R}^M$ is the set of all combinations of elements of $\mathcal{S}$ of the form $\alpha_1 z_1 + \alpha_2 z_2 + \ldots + \alpha_m z_m$ for any finite m and any $\alpha_m \geq 0$. These are called conic combinations because the result is a cone. The conic hull is a convex set — see Prob. 8.4.

8.2 CONVEXITY

Let $\mathrm{dom}(g)$ denote the domain of $g(z)$, namely, the set of values z where $g(z)$ is well-defined (i.e., finite). The function $g(z)$ is said to be convex if its domain is a convex set and, for any points $z_1, z_2 \in \mathrm{dom}(g)$ and for any scalar $0 \leq \alpha \leq 1$, it holds that

$$g\Big(\alpha z_1 + (1-\alpha)z_2\Big) \;\leq\; \alpha g(z_1) \;+\; (1-\alpha)g(z_2) \tag{8.3}$$

In other words, all points belonging to the line segment connecting $g(z_1)$ to $g(z_2)$ lie on or above the graph of $g(z)$ — see the plot on the left side in Fig. 8.2. We will be dealing primarily with *proper* convex functions, meaning that the function has a finite value for at least one location z in its domain and, moreover, it is bounded from below, i.e., $g(z) > -\infty$ for all $z \in \mathrm{dom}(g)$. We will also be dealing with *closed* functions. A general function $g(z) : \mathbb{R}^M \to \mathbb{R}$ is said to be closed if for every scalar $c \in \mathbb{R}$, the sublevel set defined by the points $\{z \in \mathrm{dom}(g) \,|\, g(z) \leq c\}$ is a closed set. It is easy to verify that if $g(z)$ is continuous in z and $\mathrm{dom}(g)$ is a closed set, then $g(z)$ is a closed function.

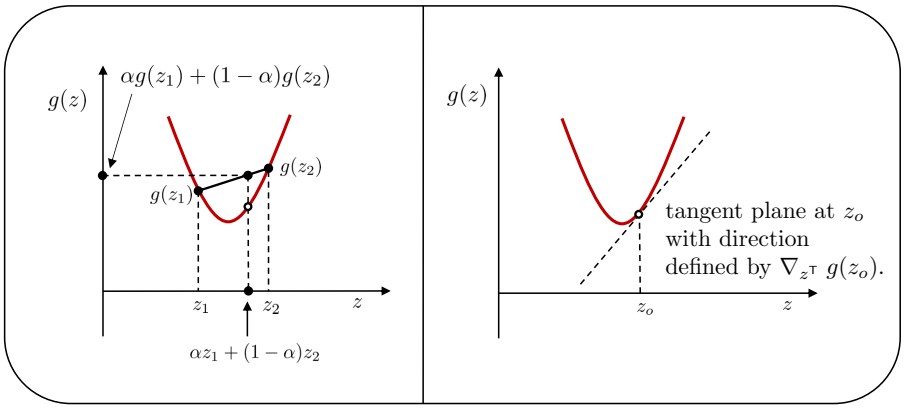

Figure 8.2 Two equivalent characterizations of convexity for *differentiable* functions $g(z)$ as defined by (8.3) and (8.4).

An equivalent characterization of the convexity definition (8.3) under first-order differentiability is that for any $z_o, z \in \mathrm{dom}(g)$:

$$\boxed{g(z) \;\geq\; g(z_o) \;+\; (\nabla_z\, g(z_o))\,(z - z_o)} \tag{8.4}$$

in terms of the inner product between the gradient vector at z_o and the difference $(z - z_o)$; recall from (2.2) that, by definition, the gradient vector $\nabla_z\, g(z_o)$ is a *row vector*. Condition (8.4) means that the tangent plane at z_o lies beneath the graph of the function — see the plot on the right side of Fig. 8.2. For later reference,

we rewrite (8.4) in the alternative form

$$g(z) \;\geq\; g(z_o) \;+\; \left(\nabla_{z^{\mathsf{T}}}\, g(z_o)\right)^{\mathsf{T}}(z - z_o) \tag{8.5}$$

in terms of the gradient of $g(z)$ relative to z^{T} at location z_o.

A useful property of every convex function is that, when a minimum exists, it can only be a global minimum; there can be multiple global minima but no local minima. That is, any stationary point at which the gradient vector of $g(z)$ is annihilated will correspond to a global minimum of the function; the function cannot have local maxima, local minima, or saddle points. A second useful property of convex functions, and which follows from characterization (8.4), is that for any $z_1, z_2 \in \mathrm{dom}(g)$:

$$\boxed{g(z) \text{ convex} \iff (\nabla_z\, g(z_2) - \nabla_z\, g(z_1))\,(z_2 - z_1) \geq 0} \tag{8.6}$$

in terms of the inner product between two differences: the difference in the gradient vectors and the difference in the vectors themselves. The inequality on the right-hand side in (8.6) is equivalent to saying that the gradient function is *monotone*.

Proof of (8.6): One direction is straightforward. Assume $g(z)$ is convex. Using (8.4) we have

$$g(z_2) \geq g(z_1) \;+\; (\nabla_z\, g(z_1))\,(z_2 - z_1) \tag{8.7}$$
$$g(z_1) \geq g(z_2) \;+\; (\nabla_z\, g(z_2))\,(z_1 - z_2) \tag{8.8}$$

so that upon substitution of the second inequality into the right-hand side of the first inequality we obtain

$$g(z_2) \;\geq\; g(z_2) \;+\; (\nabla_z\, g(z_2))\,(z_1 - z_2) \;+\; (\nabla_z\, g(z_1))\,(z_2 - z_1) \tag{8.9}$$

from which we obtain the inequality on the right-hand side of (8.6). We therefore showed that convex functions have monotonic gradients.

Conversely, assume the gradient of $g(z)$ is monotone and consider any $z_1, z_2 \in \mathrm{dom}(g)$. Let

$$h(\alpha) \;\triangleq\; g\Big((1-\alpha)z_1 + \alpha z_2\Big) \tag{8.10}$$

for any $0 \leq \alpha \leq 1$. Differentiating $h(\alpha)$ with respect to α gives

$$h'(\alpha) = \Big(\nabla_z\, g((1-\alpha)z_1 + \alpha z_2)\Big)(z_2 - z_1) \tag{8.11}$$

In particular, it holds that

$$h'(0) = \nabla_z\, g(z_1)(z_2 - z_1) \tag{8.12}$$

From the assumed monotonicity for the gradient vector we have, for $\alpha \neq 0$,

$$\Big\{\nabla_z\, g\Big((1-\alpha)z_1 + \alpha z_2\Big) \;-\; \nabla_z\, g(z_1)\Big\}(z_2 - z_1) \geq 0 \tag{8.13}$$

which implies that

$$h'(\alpha) \geq h'(0), \quad \forall\, 0 \leq \alpha \leq 1 \tag{8.14}$$

We know from the fundamental theorem of calculus that

$$h(1) - h(0) = \int_0^1 h'(\alpha) d\alpha \tag{8.15}$$

and, hence,

$$\begin{aligned} g(z_2) &= h(1) \\ &= h(0) + \int_0^1 h'(\alpha) d\alpha \\ &\geq h(0) + h'(0), \quad \text{(in view of (8.14))} \\ &= g(z_1) + \nabla_z\, g(z_1)(z_2 - z_1) \end{aligned} \tag{8.16}$$

so that $g(z)$ is convex from (8.4).

■

Example 8.2 (**Some useful operations that preserve convexity**) It is straightforward to verify from definition (8.3) that the following operations preserve convexity:

(1) If $g(z)$ is convex then $h(z) = g(Az + b)$ is also convex for any constant matrix A and vector b. That is, affine transformations of z do not destroy convexity.

(2) If $g_1(z)$ and $g_2(z)$ are convex functions, then $h(z) = \max\{g_1(z), g_2(z)\}$ is convex. That is, pointwise maximization does not destroy convexity.

(3) If $g_1(z)$ and $g_2(z)$ are convex functions, then $h(z) = a_1 g_1(z) + a_2 g_2(z)$ is also convex for any nonnegative coefficients a_1 and a_2.

(4) If $h(z)$ is convex and non-decreasing, and $g(z)$ is convex, then the composite function $f(z) = h(g(z))$ is convex.

(5) If $h(z)$ is convex and nonincreasing, and $g(z)$ is concave (i.e., $-g(z)$ is convex), then the composite function $f(z) = h(g(z))$ is convex.

8.3 STRICT CONVEXITY

The function $g(z)$ is said to be *strictly* convex if the inequalities in (8.3) or (8.4) are replaced by *strict* inequalities. More specifically, for any $z_1 \neq z_2 \in \text{dom}(g)$ and $0 < \alpha < 1$, a strictly convex function should satisfy:

$$g\Big(\alpha z_1 + (1-\alpha) z_2\Big) \;<\; \alpha g(z_1) \;+\; (1-\alpha) g(z_2) \tag{8.17}$$

A useful property of every strictly convex function is that, when a minimum exists, then it is both *unique* and also the global minimum for the function. A second useful property replaces (8.6) by the following statement with a strict inequality for any $z_1 \neq z_2 \in \text{dom}(g)$:

$$\boxed{g(z) \text{ strictly convex } \iff \left(\nabla_z\, g(z_2) - \nabla_z\, g(z_1)\right)(z_2 - z_1) > 0} \tag{8.18}$$

The inequality on the right-hand side in (8.18) is equivalent to saying that the gradient function is now *strictly monotone*.

8.4 STRONG CONVEXITY

The function $g(z)$ is said to be *strongly* convex (or, more specifically, $\nu-$strongly convex) if it satisfies the following stronger condition for any $0 \leq \alpha \leq 1$:

$$g\Big(\alpha z_1 + (1-\alpha)z_2\Big) \;\leq\; \alpha g(z_1) \;+\; (1-\alpha)g(z_2) \;-\; \frac{\nu}{2}\alpha(1-\alpha)\|z_1 - z_2\|^2 \tag{8.19}$$

for some scalar $\nu > 0$, and where the notation $\|\cdot\|$ denotes the Euclidean norm of its vector argument; other norms can be used — see Prob. 8.60. Comparing (8.19) with (8.17) we conclude that strong convexity implies strict convexity. Therefore, every strongly convex function has a unique global minimum as well. Nevertheless, strong convexity is a stronger requirement than strict convexity so that functions exist that are strictly convex but not necessarily strongly convex. For example, for scalar arguments z, the function $g(z) = z^4$ is strictly convex but not strongly convex. On the other hand, the function $g(z) = z^2$ is strongly convex — see Fig. 8.3. In summary, it holds that:

$$\boxed{\textbf{strong convexity} \implies \textbf{strict convexity} \implies \textbf{convexity}} \tag{8.20}$$

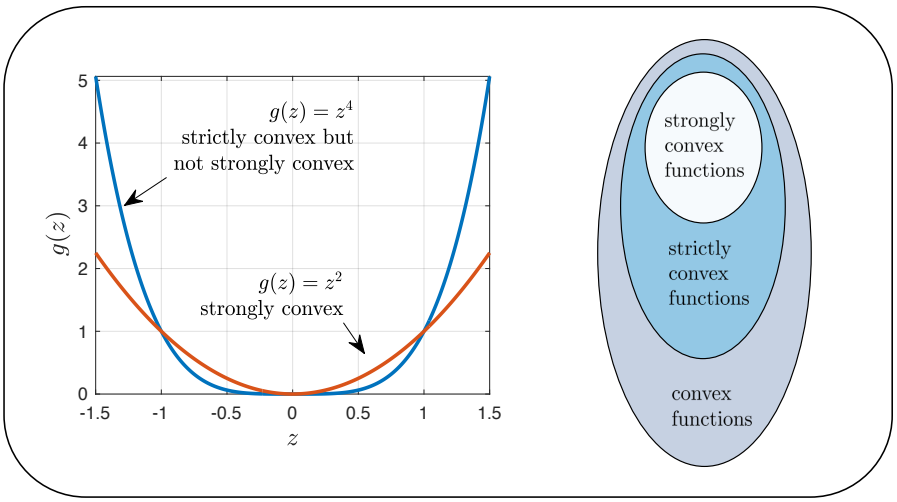

Figure 8.3 The function $g(z) = z^4$ is strictly convex but not strongly convex, while the function $g(z) = z^2$ is strongly convex. Observe how $g(z) = z^4$ is more flat around its global minimizer and moves away from it more slowly than in the quadratic case.

A useful property of strong convexity is that there exists a quadratic lower bound on the function since an equivalent characterization of strong convexity

is that for any $z_o, z \in \mathrm{dom}(g)$:

$$\boxed{g(z) \;\geq\; g(z_o) \;+\; (\nabla_z\, g(z_o))\,(z - z_o) \;+\; \frac{\nu}{2}\|z - z_o\|^2} \tag{8.21}$$

This means that the graph of $g(z)$ is strictly above the tangent plane at location z_o and moreover, for any z, the distance between the graph and the corresponding point on the tangent plane is at least as large as the quadratic term $\frac{\nu}{2}\|z-z_o\|^2$. In particular, if we specialize (8.21) to the case in which z_o is selected to correspond to the global minimizer of $g(z)$, i.e.,

$$z_o = z^o, \quad \text{where} \quad \nabla_z\, g(z^o) = 0 \tag{8.22}$$

then we conclude that every strongly convex function satisfies the following useful property for every z:

$$g(z) - g(z^o) \geq \frac{\nu}{2}\|z - z^o\|^2 \qquad (z^o \text{ is global minimizer}) \tag{8.23}$$

This property is illustrated in Fig. 8.4. Another useful property that follows from (8.21) is that for any z_1, z_2:

$$\boxed{g(z) \text{ strongly convex} \iff (\nabla_z\, g(z_2) - \nabla_z\, g(z_1))\,(z_2 - z_1) \geq \nu\|z_2 - z_1\|^2} \tag{8.24}$$

The inequality on the right-hand side in (8.24) is equivalent to saying that the gradient function is *strongly monotone*. Strong monotonicity is also called *coercivity*. This monotonicity property, along with the earlier conclusions (8.6) and (8.18), are important properties of convex functions. We summarize them in Table 8.1 for ease of reference.

Table 8.1 Useful monotonicity properties implied by the convexity, strict convexity, or strong convexity of a real-valued function $g(z) \in \mathbb{R}$ of a *real* argument $z \in \mathbb{R}^M$.

$g(z)$ **convex**	$\iff$	$\Big(\nabla_z\, g(z_2) - \nabla_z\, g(z_1)\Big)(z_2 - z_1) \geq 0$
$g(z)$ **strictly convex**	$\iff$	$\Big(\nabla_z\, g(z_2) - \nabla_z\, g(z_1)\Big)(z_2 - z_1) > 0$
$g(z)$ $\boldsymbol{\nu}$**−strongly convex**	$\iff$	$\Big(\nabla_z\, g(z_2) - \nabla_z\, g(z_1)\Big)(z_2 - z_1) \geq \nu\|z_2 - z_1\|^2$

Inequality (8.23) provides a bound from below for the difference $g(z) - g(z^o)$, where z^o is the global minimizer. We can establish a second bound from above, which will be useful in the analysis of learning algorithms later in our treatment. Referring to the general property (8.21) for $\nu-$strongly convex functions, we can write for any $z_2, z_1 \in \mathrm{dom}(g)$:

$$g(z_2) \;\geq\; g(z_1) \;+\; (\nabla_z\, g(z_1))\,(z_2 - z_1) \;+\; \frac{\nu}{2}\|z_2 - z_1\|^2 \tag{8.25}$$

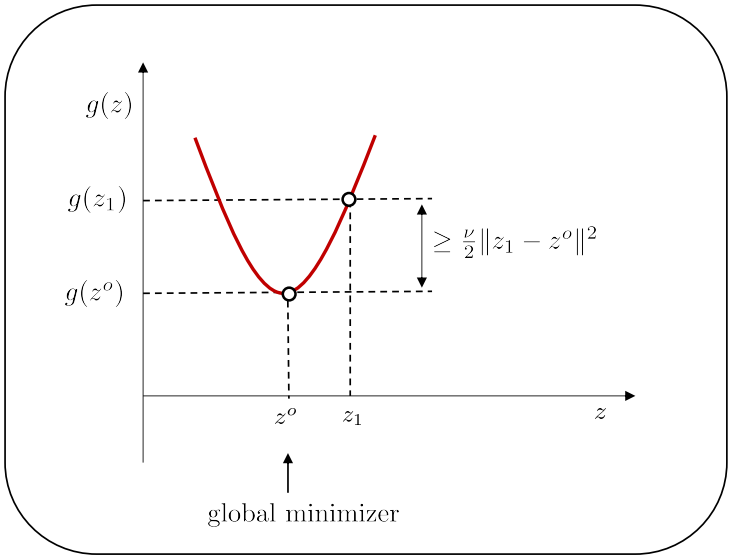

Figure 8.4 For $\nu-$strongly convex functions, the increment $g(z_1) - g(z^o)$ grows at least as fast as the quadratic term $\frac{\nu}{2}\|z_1 - z^o\|^2$, as indicated by (8.23) and where z^o is the global minimizer of $g(z)$.

The right-hand side is quadratic in z_2; its minimum value occurs at

$$\nabla_{z^\mathsf{T}} g(z_1) + \nu(z_2 - z_1) = 0 \implies (z_2 - z_1) = -\frac{1}{\nu}\nabla_{z^\mathsf{T}} g(z_1) \tag{8.26}$$

Substituting into the right-hand side of (8.25) gives

$$g(z_2) \geq g(z_1) - \frac{1}{2\nu}\|\nabla_z g(z_1)\|^2 \tag{8.27}$$

Selecting $z_1 = z$ and $z_2 = z^o$ (the global minimizer) leads to

$$g(z) - g(z^o) \leq \frac{1}{2\nu}\|\nabla_z g(z)\|^2 \tag{8.28}$$

Combining with (8.23) we arrive at the following useful lower and upper bounds for $\nu-$strongly convex functions:

$$\boxed{\frac{\nu}{2}\|z - z^o\|^2 \;\leq\; g(z) - g(z^o) \;\leq\; \frac{1}{2\nu}\|\nabla_z g(z)\|^2} \tag{8.29}$$

8.5 HESSIAN MATRIX CONDITIONS

When $g(z)$ is twice differentiable, the properties of convexity, strict convexity, and strong convexity can be inferred directly from the inertia of the Hessian matrix of $g(z)$ as follows:

$$\begin{cases} \text{(a)} \;\; \nabla_z^2\, g(z) \geq 0 \text{ for all } z & \iff g(z) \text{ is convex.} \\ \text{(b)} \;\; \nabla_z^2\, g(z) > 0 \text{ for all } z & \implies g(z) \text{ is strictly convex.} \\ \text{(c)} \;\; \nabla_z^2\, g(z) \geq \nu I_M > 0 \text{ for all } z & \iff g(z) \text{ is } \nu-\text{strongly convex.} \end{cases} \tag{8.30}$$

where, by definition,

$$\nabla_z^2\, g(z) \;\triangleq\; \nabla_{z^{\mathsf{T}}}\Big(\nabla_z\, g(z)\Big) \tag{8.31}$$

Observe from (8.30) that the positive-definiteness of the Hessian matrix is only a sufficient condition for strict convexity (for example, the function $g(z) = z^4$ is strictly convex even though its second-order derivative is not strictly positive for all z). One of the main advantages of working with strongly convex functions is that their Hessian matrices are sufficiently bounded away from zero.

Example 8.3 (**Strongly-convex functions**) The following is a list of useful strongly convex functions that appear in applications involving inference and learning:

(1) Consider the quadratic function

$$g(z) \;=\; \kappa + 2a^{\mathsf{T}}z + z^{\mathsf{T}}Cz, \quad a, z \in \mathbb{R}^M,\; \kappa \in \mathbb{R} \tag{8.32}$$

with a symmetric positive-definite matrix C. The Hessian matrix is $\nabla_z^2\, g(z) = 2C$, which is sufficiently bounded away from zero for all z since

$$\nabla_z^2\, g(z) \;\geq\; 2\lambda_{\min}(C)\, I_M > 0 \tag{8.33}$$

in terms of the smallest eigenvalue of C. Therefore, such quadratic functions are strongly convex.

The top row in Fig. 8.5 shows a surface plot for the quadratic function (8.32) for $z \in \mathbb{R}^2$ along with its contour lines for the following (randomly generated) parameter values:

$$C = \begin{bmatrix} 3.3784 & 0 \\ 0 & 3.4963 \end{bmatrix}, \;\; a = \begin{bmatrix} 0.4505 \\ 0.0838 \end{bmatrix}, \;\; \kappa = 0.5 \tag{8.34}$$

The minimum of the corresponding $g(z)$ occurs at location:

$$z^o \approx \begin{bmatrix} 0.1334 \\ 0.0240 \end{bmatrix}, \quad \text{with} \quad g(z^o) \approx 0.4379 \tag{8.35}$$

The individual entries of z are denoted by $z = \text{col}\{z_1, z_2\}$. Recall that a contour line of a function $g(z)$ is a curve along which the value of the function remains invariant. In this quadratic case, the location of the minimizer z^o can be determined in closed form and is given by $z^o = C^{-1}a$. In the plot, the surface curve is determined by evaluating $g(z)$ on a dense grid with values of (z_1, z_2) varying in the range $[-2, 2]$ in small steps of size 0.01. The location of z^o is approximated by determining the grid location where the surface attains its smallest value. This approximate numerical evaluation is applied to the other two examples below involving logistic and hinge functions where closed form expressions for z^o are not readily available. In later chapters, we are going to introduce recursive algorithms, of the gradient-descent type, and also of the subgradient and proximal gradient type, which will allow us to seek the minimizers of strongly convex functions in a more systematic manner.

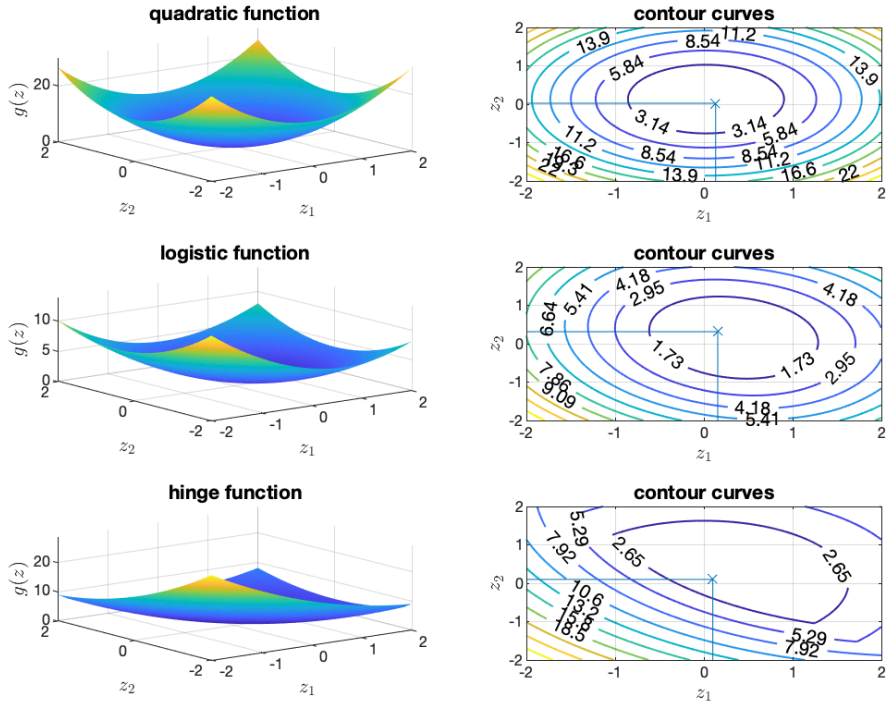

Figure 8.5 Examples of three strongly convex functions $g(z) : \mathbb{R}^2 \to \mathbb{R}$ with their contour lines. (*Top*) Quadratic function, (*Middle*) regularized logistic function; (*Bottom*) regularized hinge function. The locations of the minimizers are indicated by the $\times$ notation with horizontal and vertical lines emanating from them in the plots on the right.

(2) Consider next the regularized logistic (or log-)loss function:

$$g(z) \;=\; \ln\left(1 + e^{-\gamma h^{\mathsf{T}} z}\right) \;+\; \frac{\rho}{2}\|z\|^2, \quad z \in \mathbb{R}^M \tag{8.36}$$

with a scalar γ, column vector h, and $\rho > 0$. This function is also strongly convex, as can be seen from examining its Hessian matrix:

$$\nabla_z^2 g(z) \;=\; \rho I_M \;+\; \left(\frac{e^{-\gamma h^{\mathsf{T}} z}}{(1 + e^{-\gamma h^{\mathsf{T}} z})^2}\right) hh^{\mathsf{T}} \;\geq\; \rho I_M \;>\; 0 \tag{8.37}$$

The middle row in Fig. 8.5 shows a surface plot for the logistic function (8.36) for $z \in \mathbb{R}^2$ along with its contour lines for the following parameter values:

$$\gamma = 1, \quad h = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad \rho = 2 \tag{8.38}$$

The minimum of the corresponding $g(z)$ occurs roughly at location:

$$z^o \approx \begin{bmatrix} 0.1568 \\ 0.3135 \end{bmatrix}, \quad \text{with} \quad g(z^o) \approx 0.4990 \tag{8.39}$$

(3) Now consider the regularized hinge loss function:

$$g(z) \;=\; \max\left\{0, 1-\gamma h^{\mathsf{T}} z\right\} \;+\; \frac{\rho}{2}\|z\|^2 \tag{8.40}$$

with a scalar γ, column vector h, and $\rho > 0$ is also strongly convex, although nondifferentiable. This result can be verified by noting that the function $\max\left\{0, 1-\gamma h^{\mathsf{T}} z\right\}$ is convex in z while the regularization term $\frac{\rho}{2}\|z\|^2$ is $\rho-$strongly convex in z — see Prob. 8.23. The bottom row in Fig. 8.5 shows a surface plot for the hinge function (8.40) for $z \in \mathbb{R}^2$ along with its contour lines for the following parameter values:

$$\gamma = 1, \quad h = \begin{bmatrix} 5 \\ 5 \end{bmatrix}, \quad \rho = 2 \tag{8.41}$$

The minimum of the corresponding $g(z)$ occurs roughly at location:

$$z^o \approx \begin{bmatrix} 0.1000 \\ 0.1000 \end{bmatrix}, \quad \text{with} \quad g(z^o) \approx 0.0204 \tag{8.42}$$

Figure 8.6 shows an enlarged surface plot for the same regularized hinge function from a different view angle, where it is possible to visualize the locations of nondifferentiability in $g(z)$; these consist of all points z where $1 = \gamma h^{\mathsf{T}} z$ or, more explicitly, $z_1 + z_2 = 1/5$ by using the assumed numerical values for γ and h.

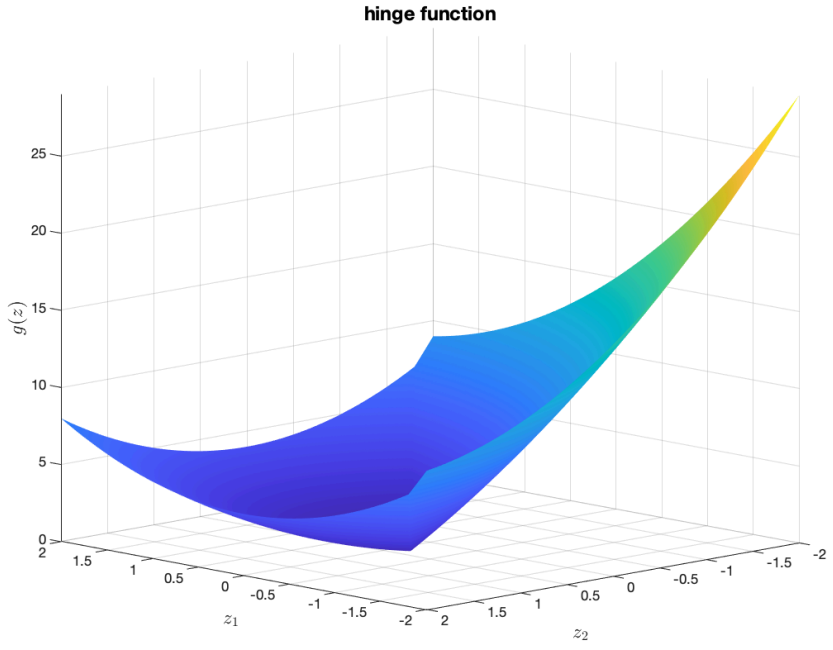

Figure 8.6 Surface plot for the same regularized hinge function from Fig. 8.5, albeit from a different viewpoint. The points of nondifferentiability occur at the locations satisfying $z_1 + z_2 = 1/5$.

8.6 SUBGRADIENT VECTORS

The characterization of convexity in (8.4) is stated in terms of the gradient vector for $g(z)$. This gradient exists because we have assumed so far that the function $g(z)$ is differentiable. There are, however, many situations of interest where the function $g(z)$ need not be differentiable at all points. For example, for scalar arguments z, the function $g(z) = |z|$ is convex but is not differentiable at $z = 0$. For such nondifferentiable convex functions, the characterizations (8.4) or (8.5) will need to be adjusted and replaced by the statement that the function $g(z)$ is convex if, and only if, for every z_o, a *column* vector s_o (dependent on z_o) exists such that

$$g(z) \;\geq\; g(z_o) \;+\; s_o^{\mathsf{T}}(z - z_o), \quad \text{for all } z_o, z \in \mathrm{dom}(g) \tag{8.43}$$

Expression (8.43) is in terms of the inner product between s_o and the difference $(z - z_o)$. Similarly, the characterization of strong convexity in (8.21) is replaced by

$$g(z) \;\geq\; g(z_o) \;+\; s_o^{\mathsf{T}}(z - z_o) \;+\; \frac{\nu}{2}\|z - z_o\|^2 \tag{8.44}$$

The vector s_o is called a *subgradient* relative to z^{T} at location $z = z_o$; equivalently, s_o^{T} is a subgradient relative to z at the same location. Note from (8.43) that subgradients help define an affine lower bound to the convex function $g(z)$. Subgradients can be defined for arbitrary functions, not only convex functions; however, they need not always exist for these general cases.

Subgradient vectors are not unique and we will use the notation $\partial_{z^{\mathsf{T}}}\, g(z_o)$ to denote the *set* of all possible subgradients s, also called the *subdifferential* of $g(z)$, at $z = z_o$. Thus, definition (8.43) is requiring the inequality to hold for

$$s_o \;\in\; \partial_{z^{\mathsf{T}}}\, g(z_o) \quad (\textbf{column vectors}) \tag{8.45}$$

It is known that the subdifferential of a proper convex function is a bounded nonempty set at every location z_o, so that subgradients are guaranteed to exist.

Remark 8.1. (Notation) We will also use the notation $\partial_z\, g(z_o)$ to denote the set that includes the transposed subgradients, s_o^{T}, which are *row* vectors. The notation $\partial_{z^{\mathsf{T}}}\, g(z)$ and $\partial_z\, g(z)$ is consistent with our earlier convention in (2.2) and (2.5) for gradient vectors (the subgradient relative to a column is a row and the subgradient relative to a row is a column). Sometimes, for compactness, we may simply write $\partial\, g(z)$ to refer to the subdifferential $\partial_{z^{\mathsf{T}}}\, g(z)$, where every element in $s \in \partial\, g(z)$ is a column vector.

■

The concept of subgradients is illustrated in Fig. 8.7. When $g(z)$ is differentiable at z_o, then there exists a unique subgradient at that location and it coincides with $\nabla_{z^{\mathsf{T}}}\, g(z_o)$. In this case, statement (8.43) reduces to (8.4) or (8.5).

One useful property that follows from (8.43) is that for any $z_1, z_2 \in \mathrm{dom}(g)$:

$$\boxed{g(z) \text{ convex} \implies (s_2 - s_1)^{\mathsf{T}}(z_2 - z_1) \;\geq\; 0} \tag{8.46}$$

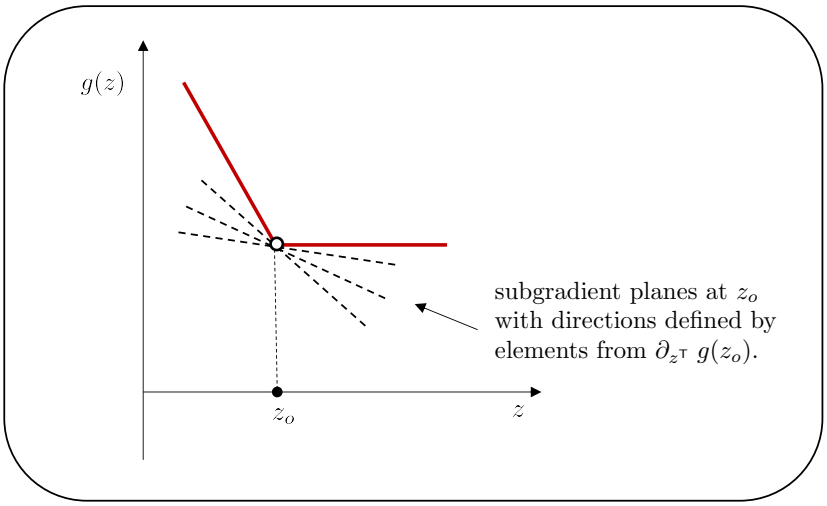

Figure 8.7 A nondifferentiable convex function admits a multitude of subgradient directions at every point of nondifferentiability.

where $\{s_1, s_2\}$ correspond to subgradient vectors relative to z^T at locations $\{z_1, z_2\}$, respectively. A second useful property of subgradient vectors is the following condition for the global minimum of a convex function (see Prob. 8.41):

$$\left\{\begin{array}{rl} g(z) \text{ differentiable at } z^o\text{:} & z^o \text{ is a minimum} \iff 0 = \nabla_z\, g(z^o) \\ g(z) \text{ nondifferentiable at } z^o\text{:} & z^o \text{ is a minimum} \iff 0 \in \partial_z\, g(z^o) \end{array}\right. \tag{8.47}$$

The second condition states that the set of subgradients at z^o must include the zero vector. This condition reduces to the first statement when $g(z)$ is differentiable at z^o.

Example 8.4 (**Absolute value function**) Let $z \in \mathbb{R}$ and consider the function

$$\boxed{g(z) = |z|} \tag{8.48}$$

This function is differentiable everywhere except at $z = 0$ — see Fig. 8.8 (*left*). The slope of the function is $+1$ over $z > 0$ and -1 over $z < 0$. At $z = 0$, any line passing through the origin with slope in the range $[-1, 1]$ can serve as a valid subgradient direction. Therefore, we find that

$$\partial_z\, g(z) \;=\; \left\{\begin{array}{ll} +1, & z > 0 \\ -1, & z < 0 \\ [-1, +1], & z = 0 \end{array}\right. \tag{8.49}$$

where the third row means that any slope within the interval $[-1, 1]$ is a valid choice for the subgradient at location $z = 0$. For ease of reference, we will denote this subdifferential set by the notation:

$$\mathbb{G}_{\text{abs}}(z) \;\triangleq\; \partial_z\, |z|, \quad z \in \mathbb{R} \tag{8.50a}$$

If we select the subgradient to be always $+1$ at $z = 0$, then this particular subgradient choice for $g(z) = |z|$ reduces to the function:

$$s(z) = \text{sign}(z) \tag{8.50b}$$

for all z where, by definition,

$$\text{sign}(z) = \begin{cases} +1, & z \geq 0 \\ -1, & z < 0 \end{cases} \tag{8.50c}$$

This will be our default choice for the subgradient of the function $g(z) = |z|$. The difference between $\mathbb{G}_{\text{abs}}(z)$ and $\text{sign}(z)$ is that the former describes *all* subgradients of $g(z) = |z|$ at $z = 0$, while the latter describes one particular (but useful) choice.

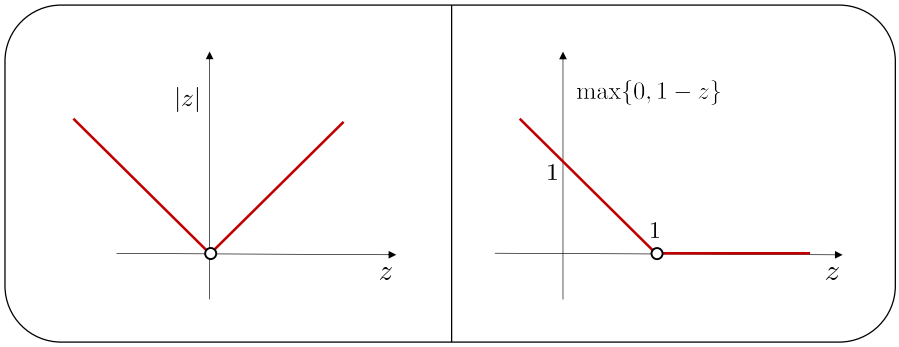

Figure 8.8 (*Left*) Absolute value function, $g(z) = |z|$. (*Right*) Hinge function, $g(z) = \max\{0, 1 - z\}$.

Consider next the case in which z is an $M-$dimensional vector and

$$\boxed{g(z) = \|z\|_1 = \sum_{m=1}^{M} |z_m|} \tag{8.51}$$

where the $\{z_m\}$ denote the individual entries of z. The function is not differentiable at all locations $z \in \mathbb{R}^M$ with at least one zero entry z_m. It follows that the subdifferential of $g(z)$, which consists of vectors of size $M \times 1$, can be constructed as follows:

$$\mathbb{G}(z) \triangleq \begin{bmatrix} \mathbb{G}_{\text{abs}}(z_1) \\ \mathbb{G}_{\text{abs}}(z_2) \\ \vdots \\ \mathbb{G}_{\text{abs}}(z_M) \end{bmatrix}, \quad (M \times 1) \tag{8.52a}$$

where each $\mathbb{G}_{\text{abs}}(z_m)$ is given by (8.49). One particular subgradient for $g(z)$ relative to z^{T} is then

$$s(z) = \text{sign}(z) \tag{8.52b}$$

where the sign function now returns an $M \times 1$ vector consisting of ± 1 entries corresponding to the signs of the individual entries of z.

Example 8.5 (**Hinge function**) Consider the hinge function

$$\boxed{g(z) = \max\{0, 1 - z\}} \tag{8.53}$$

shown in Fig. 8.8 (*right*). This function is differentiable everywhere except at $z = 1$. The slope of the function is -1 over $z < 1$ and 0 over $z > 1$. At $z = 1$, any line passing through this point with slope in the range $[-1, 0]$ can serve as a valid subgradient direction. Therefore, we find that

$$\partial_z\, g(z) \;=\; \begin{cases} 0, & z > 1 \\ -1, & z < 1 \\ [-1, 0], & z = 1 \end{cases} \tag{8.54}$$

For ease of reference, we denote this subdifferential set by the notation:

$$\mathbb{G}_1(z) \;\triangleq\; \partial_z \max\{0, 1 - z\}, \quad z \in \mathbb{R} \tag{8.55}$$

If we select the subgradient to be always -1 at $z = 1$, then this particular subgradient choice reduces to the (negative of the) indicator function:

$$s(z) = -\mathbb{I}[z \leq 1] \tag{8.56}$$

for all z where, by definition,

$$\mathbb{I}[a] = \begin{cases} 1, & \text{if statement } a \text{ is true} \\ 0, & \text{otherwise} \end{cases} \tag{8.57}$$

This will be our default choice for the subgradient of the function $g(z) = \max\{0, 1-z\}$. The difference between $\mathbb{G}_1(z)$ and $-\mathbb{I}[z \leq 1]$ is that the former describes *all* subgradients of $g(z) = \max\{0, 1 - z\}$ at location $z = 1$, while the latter describes one particular (but useful) choice.

Consider next a slight adjustment where the argument z is scaled by a nonzero constant β, say,

$$\boxed{g(z) = \max\{0, 1 - \beta z\}, \quad z \in \mathbb{R}, \quad \beta \neq 0} \tag{8.58}$$

Using similar arguments, it can be verified that the subdifferential for $g(z)$ relative to z is now given by

$$\mathbb{G}_\beta(z) \;=\; \begin{cases} 0, & \beta z > 1 \\ -\beta, & \beta z < 1 \\ [-\beta, 0], & \beta z = 1, \quad \beta > 0 \\ [0, -\beta], & \beta z = 1, \quad \beta < 0 \end{cases} \tag{8.59a}$$

where we added β as a subscript in $\mathbb{G}_\beta(z)$. Moreover, one particular choice for the subgradient is

$$s(z) = -\beta\, \mathbb{I}[\beta z \leq 1] \tag{8.59b}$$

In the degenerate case when $\beta = 0$ in (8.58), we get $g(z) = 1$. Its derivative is zero everywhere so that the above expression for $s(z)$ continues to hold in this situation.

Consider a third example involving the hinge function where z is now an $M-$dimensional vector:

$$\boxed{g(z) = \max\Big\{0, 1 - h^{\mathsf{T}} z\Big\}} \tag{8.60}$$

for some given $h \in \mathbb{R}^M$. Let $\{h_m, z_m\}$ denote the individual entries of $\{h, z\}$. The function is not differentiable at all locations $z \in \mathbb{R}^M$ where $h^{\mathsf{T}} z = 1$. It follows that

the subdifferential of $g(z)$ consists of vectors of size $M \times 1$ with entries constructed as follows:

$$\mathbb{G}(z) \triangleq \begin{bmatrix} \mathbb{A}_{h_1}(z_1) \\ \mathbb{A}_{h_2}(z_2) \\ \vdots \\ \mathbb{A}_{h_M}(z_M) \end{bmatrix}, \quad \mathbb{A}_{h_m}(z_m) \triangleq \begin{cases} 0, & h^{\mathsf{T}}z > 1 \\ -h_m, & h^{\mathsf{T}}z < 1 \\ [-h_m, 0], & h^{\mathsf{T}}z = 1, \ \ h_m \geq 0 \\ [0, -h_m], & h^{\mathsf{T}}z = 1, \ \ h_m < 0 \end{cases} \tag{8.61a}$$

where each $\mathbb{A}_{h_m}(z_m)$ is defined as above using h_m and $h^{\mathsf{T}}z$. One particular subgradient for $g(z)$ relative to z^{T} is

$$s(z) = -h\,\mathbb{I}[h^{\mathsf{T}}z \leq 1] \tag{8.61b}$$

Subgradients and subdifferentials will arise frequently in our study of inference methods and optimization problems. They possess several useful properties, some of which are collected in Table 8.2 for ease of reference. These properties are established in the problems at the end of the chapter; the last column in the table provides the relevant reference.

Table 8.2 Some useful properties of subgradients and subdifferentials for convex functions $g(z)$.

	Property	Prob.
1.	$\partial_z\, \alpha g(z) = \alpha\, \partial_z\, g(z), \ \ \alpha \geq 0$	8.29
2.	$\partial_{z^{\mathsf{T}}}\, g(Az+b) = A^{\mathsf{T}}\, \partial_{z^{\mathsf{T}}}\, g(z)\Big\vert_{z \leftarrow Az+b}$	8.30
3.	$\partial_{z^{\mathsf{T}}}\, g_1(z) + \partial_{z^{\mathsf{T}}}\, g_2(z) \subset \partial_{z^{\mathsf{T}}}\Big(g_1(z) + g_2(z)\Big)$	8.31
4.	$\partial_{z^{\mathsf{T}}}\, \|z\|_2 = \begin{cases} z/\|z\|_2, & z \neq 0 \\ \{a \mid \|a\|_2 \leq 1\}, & z = 0 \end{cases}$	8.32
5.	$\partial_{z^{\mathsf{T}}}\, \|z\|_q = \underset{\|y\|_p \leq 1}{\operatorname{argmax}} \Big\{z^{\mathsf{T}}y\Big\}, \ \ p, q \geq 1, \ 1/p + 1/q = 1$	8.33
6.	$\partial_{z^{\mathsf{T}}}\, \mathbb{I}_{C,\infty}[z] = \mathcal{N}_C(z)$ (normal cone to convex set $\mathcal{C}$ at z)	8.34

For example, the first row in the table states that the subdifferential of the scaled function $\alpha g(z)$ consists of all elements in the subdifferential of $g(z)$ scaled by α. The second row in the table shows what happens to the subdifferential set when the argument of $g(z)$ is replaced by the affine transformation $Az + b$. The result shows that the subdifferential of $g(z)$ should be evaluated at the transformations $Az + b$ and subsequently scaled by A^{T}. The third row shows that the subdifferential of the sum of two functions is *not* equal to the sum of the individual subdifferentials; it is a larger set. We will use this result in the following manner. Assume we wish to seek a minimizer z^o for the sum of two convex functions, $g_1(z) + g_2(z)$. We know that the zero vector must satisfy

$$0 \in \partial_{z^{\mathsf{T}}}\, \Big(g_1(z) + g_2(z)\Big)\Big\vert_{z=z^o} \tag{8.62}$$

That is, the zero vector must belong to the subdifferential of the sum evaluated at $z = z^o$. We will seek instead a vector z^o that ensures

$$0 \in \left\{ \partial_{z^\mathsf{T}} \, g_1(z)\Big|_{z=z^o} \; + \; \partial_{z^\mathsf{T}} \, g_2(z)\Big|_{z=z^o} \right\} \tag{8.63}$$

If this step is successful then the zero vector will satisfy (8.62) by virtue of the property in the third row of the table. The next example provides more details on the subdifferential of sums of convex functions.

Example 8.6 (**Subdifferential of sums of convex functions**) We will encounter in later chapters functions that are expressed in the form of empirical averages of convex components such as

$$g(z) \;\triangleq\; \frac{1}{L}\sum_{\ell=1}^{L} g_\ell(z), \quad z \in \mathbb{R}^M \tag{8.64}$$

where each $g_\ell(z)$ is convex. The subdifferential set for $g(z)$ will be characterized *fully* by the relation:

$$\partial_{z^\mathsf{T}} \, g(z) \;=\; \left\{ \frac{1}{L}\sum_{\ell=1}^{L} \partial_{z^\mathsf{T}} \, g_\ell(z) \right\} \tag{8.65}$$

under some conditions:

(a) First, from the third row in the table we know that whenever we combine subdifferentials of individual convex functions we obtain a subgradient for $g(z)$ so that

$$\left\{ \frac{1}{L}\sum_{\ell=1}^{L} \partial_{z^\mathsf{T}} \, g_\ell(z) \right\} \subset \partial_{z^\mathsf{T}} \, g(z) \tag{8.66}$$

(b) The converse statement is more subtle, meaning that we should be able to express every subgradient for $g(z)$ in the same sample average form and ensure

$$\partial_{z^\mathsf{T}} \, g(z) \subset \left\{ \frac{1}{L}\sum_{\ell=1}^{L} \partial_{z^\mathsf{T}} \, g_\ell(z) \right\} \tag{8.67}$$

We provide a counterexample in Prob. 8.31 to show that this direction is not always true, as already anticipated by the third row of Table 8.2. However, we explain in the comments at the end of the chapter that equality of both sets is possible under condition (8.113). The condition requires the domains of the individual functions $\{g_\ell(z)\}$ to have a nonempty intersection. This situation will be satisfied in most cases of interest since the individual functions will have the same form over z. For this reason, we will regularly assume that expression (8.65) describes all subgradients for $g(z)$. At the same time we remark that in most applications we will not need to characterize the full subdifferential for $g(z)$; it will be sufficient to find one particular subgradient for it and this subgradient can be obtained by adding individual subgradients for $\{g_\ell(z)\}$.

Example 8.7 (**Sum of hinge functions**) Consider the convex function

$$\boxed{g(z) = \frac{1}{L}\sum_{\ell=1}^{L} \max\{0, 1 - \beta_\ell z\}, \quad z \in \mathbb{R}, \quad \beta_\ell \neq 0} \tag{8.68}$$

which involves a sum of individual hinge functions, $g_\ell(z) = \max\{0, 1 - \beta_\ell z\}$. We know from (8.59a) how to characterize the subdifferential of each of these terms:

$$\mathbb{G}_{\beta_\ell}(z) \;=\; \begin{cases} 0, & \beta_\ell z > 1 \\ -\beta_\ell, & \beta_\ell z < 1 \\ [-\beta_\ell, 0], & \beta_\ell z = 1, \;\; \beta_\ell > 0 \\ [0, -\beta_\ell], & \beta_\ell z = 1, \;\; \beta_\ell < 0 \end{cases} \tag{8.69}$$

Moreover, one subgradient for each term can be chosen as $s_\ell(z) = -\beta_\ell \, \mathbb{I}[\beta_\ell z \leq 1]$. Using the conclusions from parts (a) and (b) in Example 8.6, we find that the subdifferential for $g(z)$ is given by

$$\partial_z g(z) \;=\; \frac{1}{L} \sum_{\ell=1}^{L} \mathbb{G}_{\beta_\ell}(z) \tag{8.70a}$$

while a subgradient for it can be chosen as

$$s(z) \;=\; -\frac{1}{L} \sum_{\ell=1}^{L} \beta_\ell \, \mathbb{I}_\ell[\beta_\ell z \leq 1] \tag{8.70b}$$

Consider next a situation in which z is $M-$dimensional:

$$\boxed{g(z) = \frac{1}{L} \sum_{\ell=1}^{L} \max\Big\{0, 1 - h_\ell^{\mathsf{T}} z\Big\}, \;\; z, h_\ell \in \mathbb{R}^M} \tag{8.71}$$

which again involves a sum of individual hinge functions, $g_\ell(z) = \max\{0, 1 - h_\ell^{\mathsf{T}} z\}$. We know from (8.61a) how to characterize the subdifferential for each of these terms:

$$\mathbb{G}_\ell(z) \;=\; \begin{bmatrix} \mathbb{A}_{h_{\ell,1}}(z_1) \\ \mathbb{A}_{h_{\ell,2}}(z_2) \\ \vdots \\ \mathbb{A}_{h_{\ell,M}}(z_M) \end{bmatrix} \tag{8.72a}$$

in terms of the individual entries $\{h_{\ell,m}\}$ of h_ℓ, and where each $\mathbb{A}_{h_{\ell,m}}(z_m)$ is defined according to (8.61a). The subdifferential for $g(z)$ is then given by

$$\partial_{z^{\mathsf{T}}} \, g(z) = \frac{1}{L} \sum_{\ell=1}^{L} \mathbb{G}_\ell(z) \tag{8.72b}$$

One particular subgradient for $g(z)$ relative to z^{T} is then

$$s(z) = -\frac{1}{L} \sum_{\ell=1}^{L} h_\ell \, \mathbb{I}[h_\ell^{\mathsf{T}} z \leq 1] \tag{8.72c}$$

The last three rows in Table 8.2 provide some useful subdifferential expressions for the ℓ_2-norm, ℓ_q-norm, and for the indicator function of a convex set. In particular, recall from the discussion on dual norms in Section 1.10 that the maximum of $z^{\mathsf{T}} y$ over the ball $\|y\|_p \leq 1$ is equal to the ℓ_q-norm, $\|z\|_q$. The result in the table is therefore stating that the subdifferential of $\|z\|_q$ consists of the vectors y within the ball $\|y\|_p \leq 1$ that attain this maximum value (i.e., that

attain the dual norm). The last row in the table deals with the subdifferential of the indicator function of a convex set, denoted by $\mathbb{I}_{C,\infty}[z]$. Given a set $\mathcal{C}$, this function indicates whether a point z lies in $\mathcal{C}$ or not as follows:

$$\mathbb{I}_{C,\infty}[z] \stackrel{\Delta}{=} \begin{cases} 0, & \text{if } z \in \mathcal{C} \\ \infty, & \text{otherwise} \end{cases} \tag{8.73}$$

The result in the table describes the subdifferential of the indicator function in terms of the normal cone at location z; this conclusion is illustrated geometrically in Fig. 8.9, where the normal cone is shown at one of the corner points.

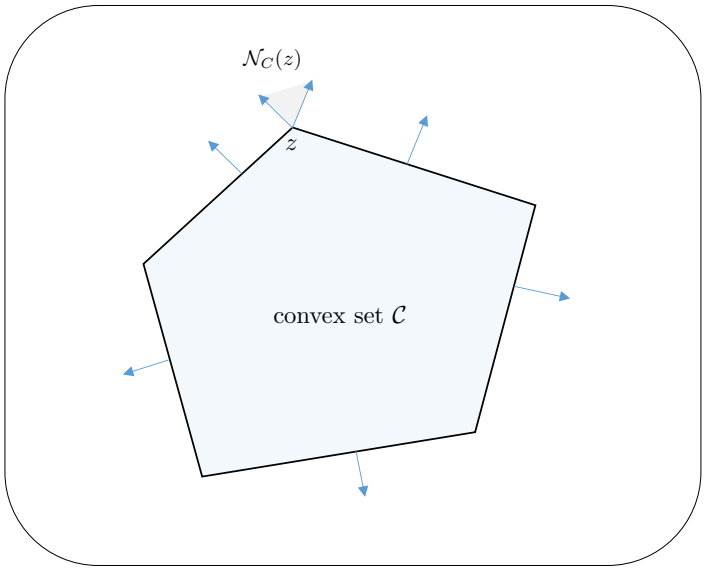

Figure 8.9 Geometric illustration of the subdifferential for the indicator function of a convex set at location z.

We collect, for ease of reference, in Table 8.3 some useful subdifferential and subgradient expressions derived in the earlier examples for a couple of convex functions that will arise in our study of learning problems.

8.7 JENSEN INEQUALITY

There are several variations and generalizations of the Jensen inequality, which is a useful result associated with convex functions. One form is the following. Let $\{z_k \in \mathbb{R}^M, k = 1, 2, \ldots, N\}$ denote a collection of N column vectors that lie in the domain of a real-valued convex function $g(z)$. Let $\{\alpha_k\}$ denote a collection

Table 8.3 Some useful subdifferentials and subgradients for convex functions $g(z)$.

Function, $g(z)$	**Subdifferential**, $\partial_{z^\mathsf{T}} g(z)$	**Subgradient**, $s(z)$
$g(z) = \lvert z\rvert,\ z \in \mathbb{R}$	$\mathbb{G}_{\text{abs}}(z) = \begin{cases} +1, & z > 0 \\ -1, & z < 0 \\ [-1,+1], & z = 0 \end{cases}$	$\text{sign}(z)$
$g(z) = \lVert z\rVert_1,\ z \in \mathbb{R}^M$ $z = \text{col}\{z_m\}$	$\mathbb{G}(z) = \text{col}\Big\{\mathbb{G}_{\text{abs}}(z_m)\Big\}$	$\text{sign}(z)$
$g(z) = \max\{0, 1-z\}$ $z \in \mathbb{R}$	$\mathbb{G}_1(z) = \begin{cases} 0, & z > 1 \\ -1, & z < 1 \\ [-1,0], & z = 1 \end{cases}$	$-\mathbb{I}[z \le 1]$
$g(z) = \max\{0, 1-\beta z\}$ $z \in \mathbb{R}, \beta \neq 0$	$\mathbb{G}_\beta(z) = \begin{cases} 0, & \beta z > 1 \\ -\beta, & \beta z < 1 \\ [-\beta, 0], & \beta z = 1,\ \beta > 0 \\ [0, -\beta], & \beta z = 1,\ \beta < 0 \end{cases}$	$-\beta\,\mathbb{I}[\beta z \le 1]$
$g(z) = \max\{0, 1-h^\mathsf{T} z\}$ $z, h \in \mathbb{R}^M$ $z = \text{col}\{z_m\}$ $h = \text{col}\{h_m\}$	$\mathbb{G}(z) = \text{col}\Big\{\mathbb{A}_{h_m}(z_m)\Big\}$ using $\mathbb{A}_{h_m}(z_m)$ from (8.61a)	$-h\,\mathbb{I}[h^\mathsf{T} z \le 1]$
$g(z) = \frac{1}{L}\sum_{\ell=1}^{L} \max\Big\{0, 1-h_\ell^\mathsf{T} z\Big\}$ $z, h_\ell \in \mathbb{R}^M$ $z = \text{col}\{z_m\}$ $h_\ell = \text{col}\{h_{\ell,m}\}$	$\frac{1}{L}\sum_{\ell=1}^{L} \mathbb{G}_\ell(z)$, where $\mathbb{G}_\ell(z) = \text{col}\Big\{\mathbb{A}_{h_{\ell,m}}(z_m)\Big\}$	$-\frac{1}{L}\sum_{\ell=1}^{L} h_\ell\,\mathbb{I}[h_\ell^\mathsf{T} z \le 1]$

of nonnegative real coefficients that add up to 1:

$$\sum_{k=1}^{N} \alpha_k = 1, \qquad 0 \le \alpha_k \le 1 \tag{8.74}$$

The Jensen inequality states that

$$\boxed{g\left(\sum_{k=1}^{N} \alpha_k z_k\right) \le \sum_{k=1}^{N} \alpha_k g(z_k)} \tag{8.75}$$

and equality holds if, and only if, $z_1 = z_2 = \ldots = z_N$. For example, if we select $g(z) = \lVert z\rVert^2$ in terms of the squared Euclidean norm of z, then it follows from (8.75) that

$$\Big\lVert\sum_{k=1}^{N} \alpha_k z_k\Big\rVert^2 \le \sum_{k=1}^{N} \alpha_k \lVert z_k\rVert^2 \tag{8.76}$$

There is also a stochastic version of Jensen inequality. If $\boldsymbol{a} \in \mathbb{R}^M$ is a real-valued random variable, then it holds that

$$g(\mathbb{E}\,\boldsymbol{a}) \leq \mathbb{E}\,(g(\boldsymbol{a})) \quad \text{(when } g(z) \in \mathbb{R} \text{ is convex)} \tag{8.77}$$

$$g(\mathbb{E}\,\boldsymbol{a}) \geq \mathbb{E}\,(g(\boldsymbol{a})) \quad \text{(when } g(z) \in \mathbb{R} \text{ is concave)} \tag{8.78}$$

where it is assumed that $\boldsymbol{a}$ and $g(\boldsymbol{a})$ have bounded expectations. We remark that a function $g(z)$ is said to be concave if, and only if, $-g(z)$ is convex.

Example 8.8 (**Vector norm**) For any vectors $a, b, c \in \mathbb{R}^M$, we know from the triangle inequality of norms that

$$\|a+b+c\| \;\leq\; \|a\| + \|b\| + \|c\| \tag{8.79}$$

Using the Jensen inequality (8.75), we can determine an upper bound for the quantity $\|a+b+c\|^4$. For this purpose, we consider the convex function $g(z) = \|z\|^4$ and note that

$$\begin{aligned}
\|a+b+c\|^4 &= \left\| 3\left(\frac{1}{3}a + \frac{1}{3}b + \frac{1}{3}c\right)\right\|^4 \\
&= 81\left\|\frac{1}{3}a + \frac{1}{3}b + \frac{1}{3}c\right\|^4 \\
&\overset{(8.75)}{\leq} 81\left(\frac{1}{3}\|a\|^4 + \frac{1}{3}\|b\|^4 + \frac{1}{3}\|c\|^4\right) \\
&= 27\left(\|a\|^4 + \|b\|^4 + \|c\|^4\right)
\end{aligned} \tag{8.80}$$

Example 8.9 (**Value at averaged arguments**) Consider a convex function $g(z)$ with vector argument $z \in \mathbb{R}^M$, and assume we are able to establish that its average value at a collection of points $\{z_n\}$ is upper bounded by some value β:

$$\frac{1}{N}\sum_{n=1}^{N} g(z_n) \;\leq \beta \tag{8.81}$$

From Jensen inequality (8.75), it follows that

$$g\Big(\frac{1}{N}\sum_{n=1}^{N} z_n\Big) \;\leq\; \frac{1}{N}\sum_{n=1}^{N} g(z_n) \;\leq\; \beta \tag{8.82}$$

so that the value of the function at the averaged arguments is also bounded by β.

8.8 CONJUGATE FUNCTIONS

Conjugate functions play an important role in the solution of optimization problems. In this section, we define them, list several of their properties, and provide some intuition for their role in convex analysis.

Consider a convex function $h(w)$ defined over $M-$dimensional vectors w. We

denote its *conjugate function* (also called the Fenchel conjugate) by the notation $h^\star(x)$ and define it as follows:

$$h^\star(x) \triangleq \sup_w \Big\{x^\mathsf{T} w - h(w)\Big\}, \quad x \in \mathcal{X} \tag{8.83}$$

where $\mathcal{X}$ denotes the set of all x where the supremum operation is finite. It can be verified that $h^\star(x)$ is always a closed convex function regardless of whether $h(w)$ itself is convex or not. This is because, for every fixed w, the function $x^\mathsf{T} w - h(w)$ is linear in x (and, hence, convex) and the supremum of a set of convex functions is convex. Likewise, the set $\mathcal{X}$ is a convex set — see Prob. 8.47. The transformation from $h(w)$ to $h^\star(x)$ is useful in many domains and appears frequently in optimization problems. We provide some intuition next.

Interpretation

Assume w and x are *scalar* variables. The situation is illustrated in Fig. 8.10 for some arbitrary function $h(w)$. In the figure, the term $x^\mathsf{T} w$ corresponds to a line passing through the origin with slope x. For the situation illustrated in the figure, the difference $x^\mathsf{T} w - h(w)$ is negative for all w, and the supremum will occur at the location of minimal distance between the line $x^\mathsf{T} w$ and the function. That distance is the value $-h^\star(x)$. If we move the line $x^\mathsf{T} w$ up by that amount it will become tangent to the function $h(w)$. The tangent is the dotted line in the figure; it is characterized by the pair $(x, h^\star(x))$: the value of x determines its slope and the value $-h^\star(x)$ determines its offset (i.e., the point where it crosses the vertical axis). We can repeat this construction for many other values of x. We find that the conjugate function provides an alternative characterization for $h(w)$: it identifies all lines $(x, h^\star(x))$ that serve as tangents to $h(w)$.

More generally, when x and w are vector-valued, we can interpret $x^\mathsf{T} w$ as representing a hyperplane passing through the origin. The normal direction of the plane is the vector x. The term $x^\mathsf{T} w - h(w)$ measures the difference between the convex function $h(w)$ and the hyperplane. For each x, the conjugate function is finding the largest possible difference between the hyperplane and the function. And the value $-h^\star(x)$ will correspond to the amount of offset that needs to be added to the hyperplane $x^\mathsf{T} w$ to make it tangent to $h(w)$. For this reason, we can interpret $h^\star(x)$ as a mapping from normal directions x to offset values $h^\star(x)$ so that the pairs $(x, h^\star(x))$ define tangent hyperplanes to $h(w)$.

Conjugate functions also arise in finance and economics in the form of conjugate utility or profit functions. In this context, $h(w)$ measures the cost of producing an amount w of some product. The variable x represents the market price per unit so that $x^\mathsf{T} w$ is the total expected market price. The difference $x^\mathsf{T} w - h(w)$ measures the profit that is expected if w items are produced. For a fixed market price x, the conjugate value $h^\star(x)$ then indicates the maximal profit at this price level.

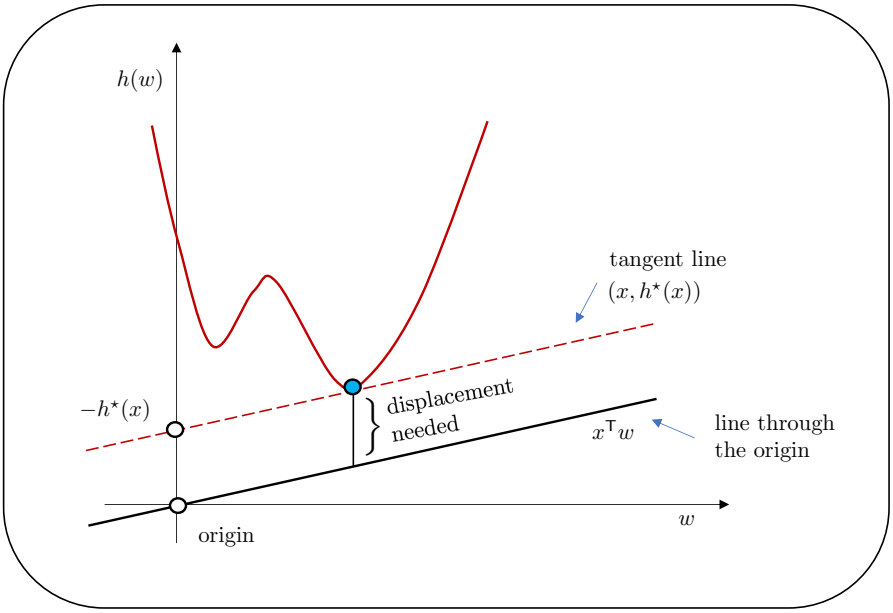

Figure 8.10 Illustration of the concept of a conjugate function for the case in which x and w are scalars. In this case, x represents the slope of the line $x^\mathsf{T}w$ passing through the origin. The conjugate value $-h^\star(x)$ is the amount of displacement needed for this line to become tangent to the function $h(w)$. The tangent line is characterized by the pair $(x, h^\star(x))$: its slope is x and its offset is $h^\star(x)$.

Relation to optimization problems

Conjugate functions are useful for the solution of optimization problems, as will be illustrated in greater detail in Example 51.6. Here we motivate the procedure and provide a couple of motivating examples.

Consider first a problem involving the unconstrained optimization of the sum of two convex functions, say,

$$\boxed{\min_{w\in\mathbb{R}^M} \Big\{q(w) + E(w)\Big\} \qquad (\textbf{primal problem})} \tag{8.84}$$

Problems of this type are commonplace when solving inference and learning problems with regularization, as will be discussed in later chapters, where the term $q(w)$ will play the role of the regularizer. We can replace problems of the above form by an equivalent formulation that involves working instead with conjugate functions as follows. First, we transform the problem into a constrained formulation by introducing a dummy variable $z \in \mathbb{R}^M$ to write:

$$\min_{w,z\in\mathbb{R}^M} \Big\{q(z) + E(w)\Big\}, \quad \text{subject to } z = w \tag{8.85}$$

The Lagrangian function associated with this problem is given by

$$\mathcal{L}(w, z, \lambda) = q(z) + E(w) + \lambda^{\mathsf{T}}(z - w) \tag{8.86}$$

where $\lambda \in \mathbb{R}^M$ is the Lagrange multiplier. The dual function $\mathcal{D}(\lambda)$ is defined as the function that results from minimizing $\mathcal{L}$ over w and z:

$$\begin{aligned}
\mathcal{D}(\lambda) &\triangleq \min_{w,z} \Big\{ q(z) + E(w) + \lambda^{\mathsf{T}}(z - w) \Big\} \\
&\overset{(a)}{=} \min_{w} \Big\{ E(w) - \lambda^{\mathsf{T}} w \Big\} \; + \; \min_{z} \Big\{ q(z) + \lambda^{\mathsf{T}} z \Big\} \\
&= \; - \max_{w} \Big\{ \lambda^{\mathsf{T}} w - E(w) \Big\} \; - \; \max_{z} \Big\{ -\lambda^{\mathsf{T}} z - q(z) \Big\} \\
&\overset{(8.83)}{=} -E^{\star}(\lambda) - q^{\star}(-\lambda)
\end{aligned} \tag{8.87}$$

where in step (a) we separated the terms that depend on w only from those that depend on z only, and in step (b) we used the definition of the conjugate function given by (8.83). We therefore find that the dual problem, which involves maximizing $\mathcal{D}(\lambda)$ over λ, is characterized by the conjugate functions $E^{\star}(\lambda)$ and $q^{\star}(\lambda)$:

$$\boxed{\max_{\lambda \in \mathbb{R}^M} \Big\{ -q^{\star}(-\lambda) - E^{\star}(\lambda) \Big\} \qquad (\textbf{dual problem})} \tag{8.88}$$

We will exploit this duality result later in Section 51.4.2, when we study sparsity-inducing regularization problems.

A second application in the context of optimization problems is the following. Consider a closed convex function $h(w)$ and its conjugate $h^{\star}(x)$. Assume we are interested in solving the optimization problem:

$$w^{\star} = \operatorname*{argmin}_{w \in \mathbb{R}^M} h(w) \tag{8.89}$$

Then, we know that the solution $w^{\star}$ must satisfy

$$0 \in \partial_{w^{\mathsf{T}}} h(w^{\star}) \tag{8.90}$$

One challenge is that it is not always possible to solve this equation directly to determine $w^{\star}$. Nevertheless, in Prob. 8.46 we establish one useful property that explains how subgradients of $h(w)$ are related to subgradients of its conjugate function $h^{\star}(x)$, namely,

$$v \in \partial_{w^{\mathsf{T}}} h(w) \iff w \in \partial_{x^{\mathsf{T}}} h^{\star}(v) \tag{8.91}$$

Applying this property to (8.90) we conclude that $w^{\star}$ should satisfy

$$\boxed{w^{\star} \in \partial_{v^{\mathsf{T}}} h^{\star}(0)} \tag{8.92}$$

In other words, $w^{\star}$ should belong to the subdifferential of $h^{\star}(x)$ at the origin.

Relation to subdifferentials

Another useful application of conjugate functions arises in the characterization of the subdifferential of convex functions. Thus, consider a convex function $h(w) : \mathbb{R}^M \to \mathbb{R}$. Its subdifferential at any point z is the set of all vectors $s \in \mathbb{R}^M$ such that

$$\begin{aligned}\partial_{z^\mathsf{T}} h(z) &= \Big\{ s \,|\, h(w) \geq h(z) + s^\mathsf{T}(w-z), \;\; \forall\, w \in \mathrm{dom}(h) \Big\} \\ &\Longleftrightarrow \Big\{ s \,|\, s^\mathsf{T} w - h(w) \leq s^\mathsf{T} z - h(z), \;\; \forall\, w \in \mathrm{dom}(h) \Big\} \\ &\Longleftrightarrow \Big\{ s \,|\, \sup_{w \in \mathrm{dom}(h)} \Big(s^\mathsf{T} w - h(w) \Big) = s^\mathsf{T} z - h(z) \Big\} \end{aligned} \tag{8.93}$$

The upper bound is attained by selecting $w = z$ in the sup operation. It follows that the subdifferential of $h(z)$ at a point z consists of the set of all vectors s where the conjugate function evaluates to the following:

$$\partial_{z^\mathsf{T}} h(z) \;=\; \Big\{ s \,|\, h^\star(s) = s^\mathsf{T} z - h(z) \Big\} \tag{8.94}$$

or, stated equivalently,

$$\boxed{\; s \in \partial_{z^\mathsf{T}} h(z) \;\Longleftrightarrow\; h^\star(s) = s^\mathsf{T} z - h(z) \;} \tag{8.95}$$

Properties

Conjugate functions have several useful properties. We list them in Table 8.4 for ease of reference and leave the proofs to the problems. The last column in the table provides the relevant references.

8.9 BREGMAN DIVERGENCE

The Kullback–Leibler (KL) divergence studied earlier in Section 6.2 is a special case of what is known as Bregman divergence, which serves as a measure of "distance" or "similarity" and is not limited to probability density functions (pdfs). Its definition and properties rely on the notions of convexity and conjugate functions, which explains our treatment of Bregman divergence at this location in the text.

Definition

Consider a closed convex set Γ and let $\phi(w) : \Gamma \to \mathbb{R}$ be a differentiable and *strictly convex* function. Let p and q be two points in Γ. The Bregman divergence between p and q is defined as the difference:

$$D_\phi(p,q) \;\triangleq\; \phi(p) \;-\; \Big(\phi(q) + \nabla_w \phi(q)\,(p-q) \Big) \tag{8.96}$$

where $\nabla_w \phi(q)$ refers to the gradient of $\phi(w)$ relative to w and evaluated at $w = q$. Note that the Bregman divergence measures the difference between the

Table 8.4 Some useful properties of conjugate functions.

	Given conditions or name	Property	Prob.
1.	closed convex function, $h(w)$	$v \in \partial h(w) \Longleftrightarrow w \in \partial h^\star(v)$	8.46
2.	closed $\nu-$strongly convex, $h(w)$	$h^\star(x)$ is differentiable everywhere with $1/\nu-$Lipschitz gradients and $\nabla_{x^\mathsf{T}} h^\star(x) = \underset{w\in\mathbb{R}^M}{\operatorname{argmax}}\Big\{x^\mathsf{T} w - h(w)\Big\}$	8.47
3.	$h(w)+c$ $\alpha h(w),\ \alpha>0$ $h(\alpha w),\ \alpha\neq 0$ $h(w-w_o)$ $h(Aw),\ A$ invertible $h(w)+z^\mathsf{T} w$	$h^\star(x)-c$ $\alpha h^\star(x/\alpha)$ $h^\star(x/\alpha)$ $h^\star(x)+x^\mathsf{T} w_o$ $h^\star(A^{-\mathsf{T}}x)$ $h^\star(x-z)$	8.49
4.	Fenchel-Young inequality	$h(w)+h^\star(x) \geq w^\mathsf{T} x$ with equality when $x \in \partial_{w^\mathsf{T}} h(w)$ or $w \in \partial_{x^\mathsf{T}} h^\star(x)$	8.48
5.	$g(w_1,w_2)=h(w_1)+h(w_2)$ (separable function)	$g^\star(x_1,x_2)=h^\star(x_1)+h^\star(x_2)$	8.50
6.	$h(w)=\frac{1}{2}\|w\|_A^2, A>0$	$h^\star(x)=\frac{1}{2}\|x\|_{A^{-1}}^2$	8.51
7.	$h(w)=\frac{1}{2}w^\mathsf{T} Aw+b^\mathsf{T} w+c$ $A>0$	$h^\star(x)=\frac{1}{2}(x-b)^\mathsf{T} A^{-1}(x-b)-c$	8.52
8.	$h(w)=\|w\|_1$	$h^\star(x)=\mathbb{I}_{C,\infty}[x],\ \mathcal{C}=\{x\|\|x\|_\infty\leq 1\}$	8.55
9.	$h(w)=\frac{\nu}{2}\|w\|_1^2$	$h^\star(x)=\frac{1}{2\nu}\|x\|_\infty^2$	8.53
10.	$h(w)=\frac{1}{2}\|w\|_p^2,\ p\geq 1$	$h^\star(x)=\frac{1}{2}\|x\|_q^2, \frac{1}{p}+\frac{1}{q}=1$	8.54
11.	$h(w)=\|w\|$	$h^\star(x)=\mathbb{I}_{C,\infty}[x],\ \mathcal{C}=\{x\|\|x\|_\star\leq 1\}$	8.55
12.	$h(w)=\mathbb{I}_{C,\infty}[w]$	$h^\star(x)=\sup_{w\in\mathcal{C}}\Big\{x^\mathsf{T} w\Big\}$	8.56
13.	$\begin{cases} h(w)=\sum_{m=1}^M w_m \ln w_m \\ w_m\geq 0 \end{cases}$	$h^\star(x)=\sum_{m=1}^M e^{x_m-1}$	8.61
14.	$h(w)=-\sum_{m=1}^M \ln w_m,\ w_m>0$	$h^\star(x)=-\sum_{m=1}^M \ln(-x_m)-M$	8.61
15.	$\begin{cases} h(w)=\sum_{m=1}^M w_m(\ln w_m-1) \\ w_m\geq 0 \end{cases}$	$h^\star(x)=\sum_{m=1}^M e^{x_m}$	8.61
16.	$\begin{cases} h(w)=\sum_{m=1}^M w_m \ln w_m \\ w_m\geq 0, \sum_{m=1}^M w_m=1 \end{cases}$	$h^\star(x)=\ln\Big(\sum_{m=1}^M e^{x_m}\Big)$	8.61
17.	$h(W)=-\ln\det(W)$ $W\in\mathbb{R}^{M\times M}, W>0$	$h^\star(X)=-\ln\det(-X)-M$	8.59

value of the function $\phi(\cdot)$ at $w = p$ and a first-order Taylor expansion around point $w = q$. In this way, the divergence reflects the gap between the convex function $\phi(p)$ and the tangent plane at $w = q$ — see Fig. 8.11. Since $\phi(w)$ is strictly-convex, it will lie above the tangent plane and the difference will always be nonnegative:

$$D_\phi(p, q) \geq 0, \quad \forall\, p, q \in \text{dom}(\phi) \tag{8.97}$$

Equality to zero will hold if, and only if, $p = q$. However, the Bregman divergence is not symmetric in general, meaning that $D_\phi(p, q)$ and $D_\phi(q, p)$ need not agree with each other.

It is clear from the definition that $D_\phi(p, q)$ is strictly convex over p since $\phi(p)$ is strictly-convex by choice and $\nabla_w \phi(q)\,(p-q)$ is linear in p. Note that if we were to approximate $\phi(p)$ by a second-order Taylor expansion around the same point $w = q$, we would get

$$\phi(p) \approx \phi(q) + \nabla_w \phi(q)(p - q) + \frac{1}{2}(p - q)^{\mathsf{T}} \nabla_w^2\, \phi(q)(p - q) \tag{8.98}$$

so that by substituting into (8.96) we will find the Bregman divergence can be interpreted as a locally weighted squared-Euclidean distance between p and q:

$$D_\phi(p, q) \approx \|p - q\|^2_{\frac{1}{2}\nabla_w^2\, \phi(q)} \tag{8.99}$$

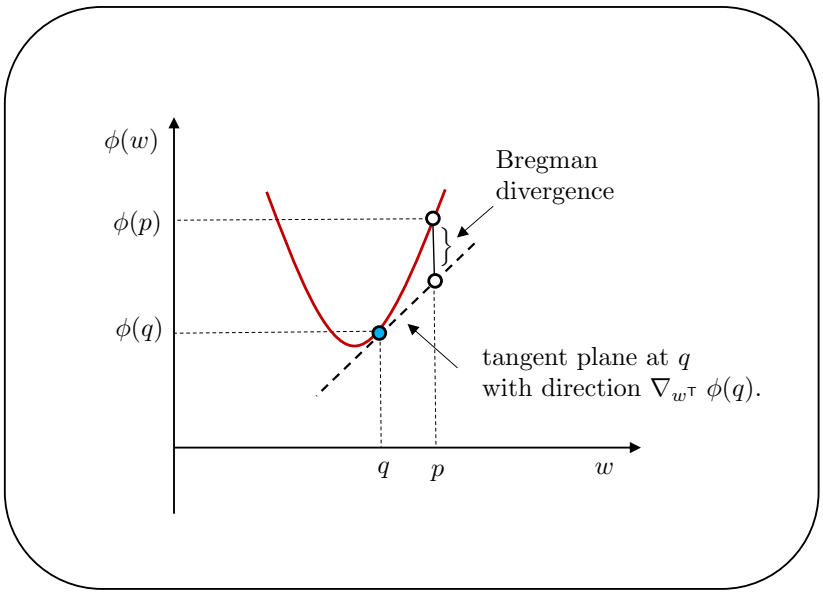

Figure 8.11 The Bregman divergence measures the gap between the function $\phi(p)$ at $w = p$ and its tangent plane at $w = q$.

Two examples

Consider the space of $M-$dimensional vectors and select

$$\phi(w) = \frac{1}{2}\|w\|^2 \tag{8.100}$$

Then, for any two vectors $p, q \in \mathbb{R}^M$:

$$D_\phi(p,q) = \frac{1}{2}\|p\|^2 - \frac{1}{2}\|q\|^2 - q^{\mathsf{T}}(p-q) \;=\; \frac{1}{2}\|p-q\|^2 \tag{8.101}$$

which shows that the squared Euclidean distance between two vectors is a Bregman distance. In this case, the Bregman divergence is symmetric. Consider next two probability mass functions (pmfs), with probability values $\{p_m, q_m\}$, defined over the simplex:

$$\Gamma = \left\{ w \in \mathbb{R}^M \;\middle|\; w_m \geq 0 \text{ and } \sum_{m=1}^{M} w_m = 1 \right\} \tag{8.102}$$

Choose $\phi(w)$ as the (negative) entropy of $\{w_m\}$, which is the convex function:

$$\phi(w) = \sum_{m=1}^{M} w_m \ln(w_m) \tag{8.103}$$

Then, the gradient vector is given by

$$\nabla_{w^{\mathsf{T}}}\, \phi(w) = \mathrm{col}\Big\{1 + \ln w_1,\; 1 + \ln w_2, \ldots, 1 + \ln w_M\Big\} \tag{8.104}$$

and the corresponding Bregman divergence reduces to the KL divergence between the probability vectors p and q since

$$\begin{aligned} D_\phi(p,q) &= \sum_{m=1}^{M} p_m \ln p_m - \sum_{m=1}^{M} q_m \ln q_m - \sum_{m=1}^{M} (1 + \ln q_m)(p_m - q_m) \\ &= \sum_{m=1}^{M} p_m \ln\Big(\frac{p_m}{q_m}\Big) \\ &= D_{\mathrm{KL}}(p,q) \end{aligned} \tag{8.105}$$

In this case, the Bregman divergence is not symmetric. We can use result (8.105) to establish a useful property for the negative entropy function, namely, that it is ν-strongly convex relative to the ℓ_1-norm with $\nu = 1$, i.e.,

$$\boxed{\phi(p) \geq \phi(q) + \nabla_w\, \phi(q)(p-q) + \frac{1}{2}\|p-q\|_1^2, \;\; \forall\, p, q \in \mathrm{dom}(\phi)} \tag{8.106}$$

Proof of (8.106): It follows from definition (8.96) that

$$\begin{aligned} \phi(p) \;&=\; \phi(q) + \nabla_w\, \phi(q)\,(p-q) + D_\phi(p,q) \\ &\overset{(8.105)}{=} \phi(q) + \nabla_w\, \phi(q)\,(p-q) + D_{\mathrm{KL}}(p\|q) \\ &\overset{(a)}{\geq} \phi(q) + \nabla_w\, \phi(q)\,(p-q) + \frac{1}{2}\|p-q\|_1^2 \end{aligned} \tag{8.107}$$

where in step (a) we used the result of Prob. 6.16, which showed that the KL-divergence of two distributions is lower bounded by $\frac{1}{2}\|p-q\|_1^2$.

■

Some properties

The Bregman divergence has several useful properties, which facilitate the development of inference methods. We list some of them in this section and leave the arguments to the problems. One first property is the following interesting interpretation.

Theorem 8.1. (Average Bregman divergence) *Let $\boldsymbol{u} \sim p_{\boldsymbol{u}}(u)$ be a random variable defined over a domain $u \in \mathcal{U}$ with pdf $p_{\boldsymbol{u}}(u)$. Let $D_\phi(u,x)$ denote the Bregman divergence between any points $u, x \in \mathcal{U}$. Then, the solution to the following optimization problem:*

$$\bar{u} \;\stackrel{\Delta}{=}\; \operatorname*{argmin}_{x\in\mathcal{U}} \mathbb{E}_{\,\boldsymbol{u}}\, D_\phi(\boldsymbol{u}, x) \tag{8.108}$$

is the mean value:

$$\bar{u} = \mathbb{E}\,\boldsymbol{u} = \int_{u\in\mathcal{U}} u p_{\boldsymbol{u}}(u)du \tag{8.109}$$

In other words, the mean of the distribution $p_{\boldsymbol{u}}(u)$ is the point that minimizes the average Bregman divergence to all points $u \in \mathcal{U}$.

Proof: Denote the cost function by $P(x) = \mathbb{E}_{\,\boldsymbol{u}}\, D_\phi(\boldsymbol{u}, x)$. Then,

$$\begin{aligned}
&P(x) - P(\bar{u}) \\
&= \int_{u\in\mathcal{U}} p_{\boldsymbol{u}}(u) D_\phi(u,x)du \;-\; \int_{u\in\mathcal{U}} p_{\boldsymbol{u}}(u) D_\phi(u,\bar{u})du \\
&= \int_{u\in\mathcal{U}} p_{\boldsymbol{u}}(u)\Big[D_\phi(u,x) \;-\; D_\phi(u,\bar{u})\Big]du \\
&= \int_{u\in\mathcal{U}} p_{\boldsymbol{u}}(u)\Big[\cancel{\phi(u)} - \phi(x) - \nabla_x\,\phi(x)(u-x) - \cancel{\phi(u)} + \phi(\bar{u}) + \nabla_x\,\phi(\bar{u})(u-\bar{u})\Big]du \\
&= \int_{u\in\mathcal{U}} p_{\boldsymbol{u}}(u)\Big[\phi(\bar{u}) - \phi(x) - \nabla_x\,\phi(x)(u-x) + \nabla_x\,\phi(\bar{u})(u-\bar{u})\Big]du \\
&= \phi(\bar{u}) - \phi(x) - \nabla_x\,\phi(x)\Big(\int_{u\in\mathcal{U}} u p_{\boldsymbol{u}}(u)du - x\Big) + \nabla_x\,\phi(\bar{u})\Big(\int_{u\in\mathcal{U}} u p_{\boldsymbol{u}}(u)du - \bar{u}\Big) \\
&= \phi(\bar{u}) - \phi(x) - \nabla_x\,\phi(x)(\bar{u}-x) + \nabla_x\,\phi(\bar{u})(\bar{u}-\bar{u}) \\
&= \phi(\bar{u}) - \phi(x) - \nabla_x\,\phi(x)(\bar{u}-x) \\
&= D_\phi(\bar{u}, x) \\
&\geq 0
\end{aligned} \tag{8.110}$$

It follows that $P(\bar{u}) \leq P(x)$ for all $x \in \mathcal{U}$ with equality only when $x = \bar{u}$.

■

We collect in Table 8.5 other useful properties, which are established in the problems. The last column in the table provides the relevant reference. Observe for the result in the first row of the table that the Bregman divergences are

computed relative to ϕ and its conjugate $\phi^\star$, and that the order of the arguments are reversed. The last two rows of the table extend the Bregman divergence to matrix arguments. In Section 9.4 we describe the use of Bregman divergences in the context of projections onto convex sets.

Table 8.5 Some useful properties of the Bregman divergence where $\phi(x)$ is a differentiable and strictly convex function and $\phi^\star(x)$ is its conjugate function.

	Property	Reference
1.	$D_\phi(p,q) = D_{\phi^\star}\Big(\nabla_w\phi(q),\, \nabla_w\phi(p)\Big)$ (duality)	Prob. 8.64
2.	$D_\phi(r,p) + D_\phi(p,q) = D_\phi(r,q) + \Big(\nabla\phi_w(q) - \nabla_w\phi(p)\Big)(r-p)$ (generalized triangle inequality)	Prob. 8.65
3.	$D_\phi(p,q) = \frac{1}{2}\|p-q\|_Q^2,\ \phi(w) = \frac{1}{2}\|w\|_Q^2,\ Q>0$ (Mahalanobis distance)	Prob. 8.66
4.	$D_\phi(p,q) = D_{\mathrm{KL}}(p,q) = \sum_{m=1}^{M} p_m \ln\Big(\frac{p_m}{q_m}\Big)$ $\phi(w) = \sum_{m=1}^{M} w_m \ln(w_m),\ w_m>0,\ \sum_{m=1}^{M} w_m = 1$ (negative entropy)	Eq. (8.105)
5.	$D_\phi(P,Q) = \mathrm{Tr}(P\ln P - P\ln Q - P + Q)$ $\phi(W) = \mathrm{Tr}(W\ln W), W>0$ (von Neuman divergence)	Prob. 8.67
6.	$D_\phi(P,Q) = \mathrm{Tr}(PQ^{-1} - I_M) - \ln\det(PQ^{-1})$ $\phi(W) = -\ln\det(W),\ W>0,\ W\in\mathbb{R}^{M\times M}$	Prob. 8.68

8.10 COMMENTARIES AND DISCUSSION

Convex functions. Excellent references on convex analysis are the texts by Rockafellar (1970), Polyak (1987), Hiriart-Urruty and Lemaréchal (2001), Bertsekas (2003), Boyd and Vandenberghe (2004), and Nesterov (2004). Useful accounts on the history of convexity, dating back to the development of Greek geometry, appear in Fenchel (1983) and Dwilewicz (2009). According to the latter reference and also Heath (1912, p. 8), the first definition of convexity was apparently given by the ancient Greek mathematician **Archimedes of Syracuse (ca 287 BC–212 BC)** in the work by Archimedes (225 BC) — see the exposition by Dunham (1990). Result (8.28) for $\nu-$strongly convex functions is often referred to as the *Polyak-Lojasiewicz bound* due to Polyak (1963) and Lojasiewicz (1963); it is useful in the study of the convergence behavior of gradient descent algorithms — see, e.g., Example 12.10 and the proof of Theorem 12.3.

Subgradients. In future chapters we will encounter optimization problems that involve nonsmooth functions with nondifferentiable terms. In these cases, iterative algorithms for minimizing these functions will be constructed by replacing traditional gradient vectors by subgradients whenever necessary. The idea of employing subgradient vectors

was proposed by Shor (1962) in his work on maximizing piecewise linear concave functions. The method was well-received at the time and generated tremendous interest due to its simplicity and effectiveness. Some of the earliest works that helped solidify the theoretical foundations of the method were published by Ermoliev (1966,1969,1983a,b), Ermoliev and Shor (1967), and Polyak (1967,1969), culminating with the manuscripts by Ermoliev (1976) and Shor (1979). Useful surveys on the history and development of subgradient methods are given by Shor (1991) and Goffin (2012). Some additional references include Rockafellar (1970), Bertsekas (1973), Held, Wolfe, and Crowder (1974), Clarke (1983), Nemirovsky and Yudin (1983), Kiwiel (1985), Polyak (1987), Shor (1998,2012), Bertsekas, Nedic, and Ozdaglar (2003), Nesterov (2004), Shalev-Shwartz *et al.* (2011), Duchi, Hazan, and Singer (2011), Duchi, Bartlett, and Wainwright (2012), Shamir and Zhang (2013), and Ying and Sayed (2018).

Subgradients of sums of convex functions. It will be common in our treatment of inference and learning methods in this text to encounter objective functions that are expressed as the sum of a finite number of convex functions such as

$$g(z) = \ell_1(z) + \ldots + \ell_N(z) \tag{8.111}$$

in terms of individual convex terms $\ell_n(z)$. These individual terms need not be differentiable. Let $\partial_z \ell_n(z)$ denote the subdifferential set for $\ell_n(z)$ at location z. Then, the result of Prob. 8.31 indicates that we can construct a subgradient for the sum $g(z)$ by adding individual subgradients for the $\{\ell_n(z)\}$. This is because

$$\Big\{\partial_z \ell_1(z) + \partial_z \ell_2(z) + \ldots + \partial_z \ell_N(z)\Big\} \subset \partial_z g(z) \tag{8.112}$$

A useful question is whether *all* elements of the subdifferential set of $g(z)$ can be constructed from the sum of individual subgradient vectors for the $\{\ell_n(z)\}\}$, i.e., whether the two sets in (8.112) are actually *equal* to each other. We provide a counterexample in Prob. 8.31 to show that these two sets are not generally equal. While establishing property (8.112) is relatively straightforward, and is left as an exercise in Prob. 8.31, the study of conditions under which both sets coincide is more challenging and can be found, for example, in Rockafellar (1963). In particular, it is shown there that both sets will coincide when the domains of the individual functions satisfy the following condition:

$$\bigcap_{n=1}^{N} \mathrm{ri}\Big(\mathrm{dom}(\ell_n(z))\Big) \neq \emptyset \tag{8.113}$$

in terms of the *relative interior* (ri) of the domains of the individual functions. Condition (8.113) requires the domains of all individual functions to have a nonempty intersection. This situation will be satisfied in most cases of interest to us since the individual functions will have the same form over z, or their domains will generally be $\mathbb{R}^M$. In these situations, the following two directions will hold:

$$\Big\{\partial_z \ell_1(z) + \partial_z \ell_2(z) + \ldots + \partial_z \ell_N(z)\Big\} \subset \partial_z g(z) \tag{8.114a}$$

$$\partial_z g(z) \subset \Big\{\partial_z \ell_1(z) + \partial_z \ell_2(z) + \ldots + \partial_z \ell_N(z)\Big\} \tag{8.114b}$$

To explain the notion of the relative interior, consider the segment $\{-1 \leq x \leq 1\}$ on the real axis. The interior of this set consists of all points $\{-1 < x < 1\}$. Recall that a point is in the interior of a set $\mathcal{S}$ if a small $\epsilon-$size open interval around the point continues to be in $\mathcal{S}$. Now, let us take the same interval $\{-1 \leq x \leq 1\}$ and view it as a set in the higher-dimensional space $\mathbb{R}^2$. In this space, this interval does *not* have an interior anymore. This is because, for any point in the interval, if we draw a small circle of radius ϵ around it, the circle will contain points outside the interval no matter how small ϵ is. Therefore, the interval $\{-1 \leq x \leq 1\}$ does not have an interior in $\mathbb{R}^2$. However, one can extend the notion of interiors to allow for such intervals to

have interiors in higher-dimensional spaces. This is what the notion of *relative interior* does. Loosely, under this concept, to check whether a set $\mathcal{S}$ has an interior, we limit our examination to the subspace where the set lies. For any set $\mathcal{S}$, we define its *affine hull* as the collection of all points resulting from any *affine* combination of elements of $\mathcal{S}$:

$$\text{affine}(\mathcal{S}) \triangleq \left\{ \sum_{p=1}^{P} a_p s_p \,\Big|\, \text{for any integer } P > 0, s_p \in \mathcal{S} \right\} \tag{8.115a}$$

$$a_p \in \mathbb{R}, \sum_{p=1}^{P} a_p = 1 \tag{8.115b}$$

where the combination coefficients $\{a_p\}$ are real numbers and required to add up to 1; if these combination coefficients were further restricted to being nonnegative, then the affine hull would become the convex hull of the set $\mathcal{S}$. For example, for the interval $\mathcal{S} = \{-1 \le x \le 1\}$, its affine hull will be a line along the $x-$axis containing the interval. Once affine($\mathcal{S}$) is identified, we then determine whether the interval $\mathcal{S}$ has an interior *within* this affine hull, which we already know is true and given by $\{-1 < x < 1\}$. This interior is referred to as the *relative interior* of the set in $\mathbb{R}^2$; the qualification "relative" is referring to the fact that the interior is defined *relative* to the affine hull space and not the entire $\mathbb{R}^2$ space where the interval lies.

Jensen inequality. We described deterministic and stochastic forms of Jensen inequality in Section 8.7. They are useful in various contexts in probability theory, information theory, and statistics. Inequality (8.75) is generally attributed to Jensen (1906), although in an addendum on page 192 of his article, Jensen acknowledges that he discovered an earlier instance of his inequality in the work by Hölder (1889). In this latter reference, the inequality appears in the following form:

$$g\left(\frac{\sum_{k=1}^{N}\beta_k z_k}{\sum_{\ell=1}^{N}\beta_\ell}\right) \le \frac{\sum_{k=1}^{N}\beta_k g(z_k)}{\sum_{\ell=1}^{N}\beta_\ell} \tag{8.116}$$

where $g(z)$ is a convex function and the $\{\beta_k\}$ are positive scalars. If we redefine

$$\alpha_k \triangleq \frac{\beta_k}{\sum_{\ell=1}^{N}\beta_\ell} \tag{8.117}$$

then the $\{\alpha_k\}$ become convex combination coefficients and the result reduces to (8.75). More information on Jensen and Hölder inequalities can be found in Hardy, Littlewood, and Pólya (1934) and Abramowitz and Stegun (1965).

Conjugate functions. These functions are also known as *Fenchel conjugates* after Fenchel (1949); they play an important supporting role in the solution of optimization problems through duality. For more details on their mathematical properties, the reader may refer to Rockafellar (1970,1974), Boyd and Vandenberghe (2004), and Bertsekas (2009). We explained in Section 8.8 that the transformation from $h(w)$ to $h^\star(x)$ defined by (8.83) is useful in many scenarios, including in the solution of optimization problems. We also indicated that conjugate functions arise in finance and economics in the form of conjugate utility or profit functions — see, e.g., Eatwell, Newman, and Milgate (1987). We will explain later in the commentaries to Chapter 11 the close relationship that exists between conjugate functions and proximal operators in the form of the Moreau decomposition established by Moreau (1965).

Bregman divergence. The KL divergence is a special case of the Bregman divergence defined in (8.96) and introduced by Bregman (1967). As explained in the body of the chapter, this divergence serves as a measure of "distance" or "similarity" and is not limited to pdfs. However, when the arguments p and q correspond to pmfs and $\phi(\cdot)$ is

selected as the negative entropy function (8.103), we get

$$D_\phi(p,q) = D_{\mathrm{KL}}(p,q) \tag{8.118}$$

The important result in Theorem 8.1 is due to Banerjee, Gou, and Wang (2005). It states that for a collection of points randomly distributed within some space $\mathcal{U}$, their mean is the point that minimizes the average Bregman divergence to all of them. For further details on Bregman divergences, the reader may refer to Censor and Zenios (1998), Azoury and Warmuth (2001), Banerjee *et al.* (2005), Chen, Chen, and Rao (2008), Adamcík (2014), Harremoës (2017), and Siahkamari *et al.* (2020). In Chapter 15 we will exploit properties of the Bregman divergence in the derivation of mirror-descent learning algorithms.

PROBLEMS

8.1 Is the column span of any matrix $A \in \mathbb{R}^{N\times M}$ a convex set? What about its row span? What about its nullspace?

8.2 Consider a convex function $g(z) : \mathbb{R}^M \to \mathbb{R}$. Denote the individual entries of z by $\{z_m\}$ for $m = 1, 2 \ldots, M$. Select an arbitrary z_m and fix all other entries. Is $g(z)$ convex over z_m?

8.3 Show that the intersection of convex sets is a convex set.

8.4 Show that the zero vector belongs to the conic hull of a set $\mathcal{S} \subset \mathbb{R}^M$. Show also that the conic hull is a convex set.

8.5 Show that the normal cone defined by (8.2) is always a convex cone regardless of the nature of the set $\mathcal{S}$.

8.6 Verify that each of the following sets is convex:

(a) The nonnegative orthant denoted by $\mathbb{R}_+^M$, which consists of all $M-$dimensional vectors with nonnegative entries.

(b) Any affine subspace consisting of all vectors $x \in \mathbb{R}^M$ satisfying $Ax = b$, where $A \in \mathbb{R}^{N\times M}$ and $b \in \mathbb{R}^N$.

(c) The halfspace consisting of all vectors $x \in \mathbb{R}^M$ satisfying $a^{\mathsf{T}}x \le \alpha$, where a is a vector and α is a scalar.

(d) Any polyhedron consisting of all vectors $x \in \mathbb{R}^M$ satisfying $Ax \preceq b$, where $A \in \mathbb{R}^{N\times M}$, $b \in \mathbb{R}^N$, and the notation $x \preceq y$ refers to component-wise inequalities for the individual elements of the vectors $\{x, y\}$.

(e) The set of symmetric and nonnegative definite matrices, $A \in \mathbb{R}^{N\times N}$.

8.7 Consider two convex functions $h(z)$ and $g(z)$. Is their composition $f(z) = g(h(z))$ convex?

8.8 Given any x_o and a square matrix $A \ge 0$, show that the ellipsoid consisting of all vectors $x \in \mathbb{R}^M$ such that $(x - x_o)^{\mathsf{T}}A(x - x_o) \le 1$ is a convex set.

8.9 Show that the probability simplex defined by all vectors $p \in \mathbb{R}^M$ with entries p_m satisfying $p_m \ge 0$ and $\sum_{m=1}^M p_m = 1$ is a convex set.

8.10 Consider a convex function $g(z)$ with $z \in \mathbb{R}^M$. The $\alpha-$sublevel set of $g(z)$ is defined as the set of all vectors $z \in \mathrm{dom}(g)$ that satisfy $g(z) \le \alpha$. Show that $\alpha-$sublevel sets are convex.

8.11 Consider a collection of convex functions, $\{g_\ell(z),\ \ell = 1, \ldots, L\}$, and introduce the weighted combination (also called conic combination) $g(z) = \sum_{\ell=1}^L a_\ell g_\ell(z)$, where $a_\ell \ge 0$. Show that $g(z)$ is convex.

8.12 Show that definitions (8.3) and (8.4) are equivalent characterizations of convexity when $g(z)$ is differentiable.

8.13 A continuous function $g(z) : \mathbb{R}^M \to \mathbb{R}$ is said to be midpoint convex if for any $z_1, z_2 \in \mathrm{dom}(g)$, it holds that $g\left(\frac{1}{2}(z_1 + z_2)\right) \le \frac{1}{2}(g(z_1)+g(z_2))$. Show that a real-valued continuous function $g(z)$ is convex if, and only if, it is midpoint convex.

8.14 Establish the three properties listed in Example 8.2.
8.15 Consider the indicator function (8.73) for some set $\mathcal{C}$. Show that $\mathbb{I}_{C,\infty}[z]$ is a convex function if, and only if, the set $\mathcal{C}$ is convex.
8.16 Consider a function $g(z;a) : \mathbb{R}^M \to \mathbb{R}$ that is parameterized by a vector a in some set $\mathcal{A}$. Show that if $g(z;a)$ is convex in z for every $a \in \mathcal{A}$, then the following function is also convex in z:

$$g(z) \;\triangleq\; \max_{a\in\mathcal{A}} g(z;a)$$

8.17 Let $z \in \mathbb{R}^M$ be a vector in the probability simplex and denote its entries by $\{z_m \geq 0\}$. Assume the convention $0 \times \ln 0 = 0$. Consider the negative entropy function $g(z) = \sum_{m=1}^{M} z_m \ln z_m$. Verify that $g(z)$ is convex. Show further that $g(z)$ is $\nu-$strongly convex relative to the ℓ_1-norm, i.e., it satisfies the following relation with $\nu = 1$ for any (z, z_0):

$$g(z) \geq g(z_0) + \nabla_z g(z_0)\,(z - z_0) + \frac{\nu}{2}\|z - z_0\|_1^2$$

8.18 Consider a function $g(z) : \mathbb{R}^M \to \mathbb{R}$. Pick any $z \in \mathrm{dom}(g)$ and any scalar t and vector w such that $z + tw \in \mathrm{dom}(g)$. Show that $g(z)$ is convex in z if, and only if, the function $h(t) = g(z + tw)$ is convex in t. In other words, a function $g(z)$ is convex if, and only if, its restriction to *any* line in its domain, namely, $g(z + tw)$, is also convex.
8.19 Let z^o denote the global minimizer of a $\nu-$strongly convex function $g(z)$. Use (8.24) to show that $\nu\|z - z^o\| \leq \|\nabla_z\, g(z)\|$.
8.20 Consider a $\nu-$strongly convex function $g(z) : \mathbb{R}^M \to \mathbb{R}$ satisfying (8.19) . Denote the individual entries of z by $\{z_m\}$ for $m = 1, 2, \ldots, M$. Select an arbitrary z_m and fix all other entries. Is $g(z)$ $\nu-$strongly convex over z_m?
8.21 True or false. Refer to definition (8.19). A function $g(z)$ is ν-strongly convex if, and only if, the function $g(z) - \frac{\nu}{2}\|z\|^2$ is convex.
8.22 Let $z \in \mathbb{R}^M$. Show that $g(z) = \|z\|^4$ is strictly convex.
8.23 Show that the regularized hinge loss function (8.40) is strongly convex.
8.24 Establish (8.21) as an equivalent characterization for $\nu-$strong convexity for differentiable functions $g(z) : \mathbb{R}^M \to \mathbb{R}$.
8.25 Establish property (8.24) for $\nu-$strongly convex functions $g(z) : \mathbb{R}^M \to \mathbb{R}$.
8.26 Let $z \in \mathbb{R}^M$ and consider a full-rank matrix $A \in \mathbb{R}^{N\times M}$ with $N \geq M$. Examine the convexity, strict convexity, and strong-convexity of the function $g(z) = \|Az\|^\alpha$ for all values of α in the range $\alpha \in [1, \infty)$. How would your answers change if A were nonzero but rank-deficient?
8.27 Consider a $\nu-$strongly convex function $g(z) : \mathbb{R}^M \to \mathbb{R}$, as defined by (8.19). Relation (8.21) provides a useful property for such functions when they are differentiable. Assume now that the function is not necessarily differentiable. For any arbitrary points z and z_o, let s and s_o denote subgradients for $g(z)$ at locations z and z_o, respectively, i.e., $s \in \partial_{z^\mathsf{T}}\, g(z)$ and $s_o \in \partial_{z^\mathsf{T}}\, g(z_o)$. Establish the validity of the following properties:

$$\begin{aligned}
g(z) \geq\; & g(z_o) \;+\; s_o^\mathsf{T}(z - z_o) \;+\; \frac{\nu}{2}\|z - z_o\|^2 \\
g(z) \leq\; & g(z_o) \;+\; s_o^\mathsf{T}(z - z_o) \;+\; \frac{1}{2\nu}\|s - s_o\|^2 \\
\nu\|z - z_o\|^2 \leq\; & (s - s_o)^\mathsf{T}(z - z_o) \leq \frac{1}{\nu}\|s - s_o\|^2
\end{aligned}$$

8.28 Let $g(z)$ be a strictly convex function and consider two distinct points z_1 and z_2 in its domain. Show that $\partial_z g(z_1) \cap \partial_z g(z_2) = \emptyset$.
8.29 For any $\alpha \geq 0$, show that $\partial_z\, \alpha g(z) = \alpha\, \partial_z\, g(z)$.
8.30 Let $g(z)$ be a convex function and introduce the transformation $h(z) = g(Az+b)$ where $z \in \mathbb{R}^M$, $A \in \mathbb{R}^{N\times M}$, and $b \in \mathbb{R}^N$. Show that $\partial_{z^\mathsf{T}}\, h(z) = A^\mathsf{T}\, \partial_{z^\mathsf{T}}\, g(z)|_{z\leftarrow Az+b}$.

8.31 Let $g_1(z)$ and $g_2(z)$ be two convex functions with $z \in \mathbb{R}^M$. Show that

$$\partial_{z^\mathsf{T}} g_1(z) + \partial_{z^\mathsf{T}} g_2(z) \subset \partial_{z^\mathsf{T}} \Big(g_1(z) + g_2(z) \Big)$$

To verify that these two sets are not identical in general, consider the following two functions with $z \in \mathbb{R}$:

$$g_1(z) = \begin{cases} z, & z \geq 0 \\ +\infty, & z < 0 \end{cases} \qquad g_2(z) = \begin{cases} +\infty, & z > 0 \\ -z, & z \leq 0 \end{cases}$$

(a) Let $g(z) = g_1(z) + g_2(z)$. What is $g(z)$?
(b) Determine $\partial_{z^\mathsf{T}} g_1(0), \partial_{z^\mathsf{T}} g_2(0)$, and $\partial_{z^\mathsf{T}} g(0)$.

8.32 Let $g(z) = \|z\|$, in terms of the Euclidean norm of $z \in \mathbb{R}^M$. Show that the subdifferential of $g(z)$ is given by the following expression where $a \in \mathbb{R}^M$:

$$\partial_{z^\mathsf{T}} g(z) = \begin{cases} z/\|z\|, & z \neq 0 \\ \{a \mid \|a\| \leq 1\}, & z = 0 \end{cases}$$

8.33 Refer to the definition of the dual norm in (1.157). Let $g(z) = \|z\|_q$ denote the $q-$norm of vector $z \in \mathbb{R}^M$ and define p through $1/p + 1/q = 1$ for $p, q \geq 1$. Show that

$$\partial_{z^\mathsf{T}} \|z\|_q = \underset{\|y\|_p \leq 1}{\operatorname{argmax}} \; z^\mathsf{T} y$$

Explain how this characterization leads to the same conclusion in Prob. 8.32.

8.34 Consider a convex set $\mathcal{C}$ and its indicator function $\mathbb{I}_{C,\infty}[z]$: it is equal to zero when $z \in$ and $+\infty$ otherwise. Show that $\partial_z \, \mathbb{I}_{C,\infty}[z] = \mathcal{N}_C(z)$ in terms of the normal cone at location z. The result is illustrated geometrically in Fig. 8.9, where the normal cone is shown at one of the corner points.

8.35 Let $g(z) = \|z\|_\infty$ in terms of the $\infty-$norm of the vector $z \in \mathbb{R}^M$. What is $\partial_{z^\mathsf{T}} g(0)$?

8.36 Let

$$g(z) \;=\; \max_{1 \leq n \leq N} \Big\{ a_n^\mathsf{T} z + \alpha(n) \Big\}, \;\; a_n, z \in \mathbb{R}^M, \; \alpha(n) \in \mathbb{R}$$

and assume the maximum is attained at some index n_o. Show that $a_{n_o} \in \partial_{z^\mathsf{T}} g(z)$.

8.37 For any convex function $g(z)$ that is nondifferentiable at some location z_o, show that its subdifferential at this location is a convex set.

8.38 Consider L convex functions $g_\ell(z)$ for $\ell = 1, 2, \ldots, L$ and define the pointwise maximum function

$$g(z) = \max_{\ell = 1, \ldots, L} \Big\{ g_\ell(z) \Big\}$$

At any point z_1, let $g_{\ell^o}(z_1)$ be one of the functions for which $g_{\ell^o}(z_1) = g(z_1)$. There may exist more than one function attaining the value $g(z_1)$. It is sufficient to consider one of them. Show that if $s \in \partial_{z^\mathsf{T}} g_{\ell^o}(z_1)$, then $s \in \partial_{z^\mathsf{T}} g(z_1)$. That is, show that a subgradient for $g_{\ell^o}(z)$ at z_1 can serve as a subgradient for $g(z)$ at the same location. More generally, show that the subdifferential of $g(z)$ is given by the following convex hull:

$$\partial_{z^\mathsf{T}} g(z) \;=\; \operatorname{conv} \Bigg\{ \bigcup_{g_\ell(z) = g(z)} \partial_{z^\mathsf{T}} g_\ell(z) \Bigg\}$$

8.39 Consider two differentiable convex functions $\{g_1(z), g_2(z)\}$ and define $g(z) = \max\{g_1(z), g_2(z)\}$. Show that

$$\partial_z g(z) \;=\; \begin{cases} \nabla_z \, g_1(z), & \text{if } g_1(z) > g_2(z) \\ \nabla_z \, g_2(z), & \text{if } g_2(z) > g_1(z) \\ \alpha \nabla_z \, g_1(z) + (1-\alpha) \nabla_z \, g_2(z), & \text{if } g_1(z) = g_2(z) \end{cases}$$

where $\alpha \in [0,1]$. The last condition amounts to selecting any point on the segment linking the gradient vectors of $g_1(z)$ and $g_2(z)$.

8.40 Let $g(z) : \mathbb{R}^M \to \mathbb{R}$ denote a convex function and consider subgradient vectors $s_1 \in \partial_{z^\mathsf{T}} g(z_1)$ and $s_2 \in \partial_{z^\mathsf{T}} g(z_2)$ at locations z_1 and z_2. Establish the following inner product inequality:

$$(s_2 - s_1)^\mathsf{T}(z_2 - z_1) \geq 0$$

Let z^o denote the global minimizer of $g(z)$. Conclude that $(\partial_{z^\mathsf{T}} g(z))^\mathsf{T}(z - z^o) \geq 0$ for any subgradient vector at location z.

8.41 For a convex function $g(z)$, show that z^o is a minimum if, and only if, $0 \in \partial_{z^\mathsf{T}} g(z^o)$.

8.42 Let $g(z) = \sum_{n=1}^{N} |\gamma(n) - h_n^\mathsf{T} z|$, where $z, h_n \in \mathbb{R}^M$ and $\gamma(n) \in \mathbb{R}$. Show that a subgradient for $g(z)$ is given by

$$\sum_{n=1}^{N} -h_n \text{sign}\left(\gamma(n) - h_n^\mathsf{T} z\right) \in \partial_{z^\mathsf{T}} g(z)$$

where $\text{sign}(x) = +1$ if $x \geq 0$ and $\text{sign}(x) = -1$ if $x < 0$.

8.43 Let $g(z) = \max_{1 \leq n \leq N} (\gamma(n) - h_n^\mathsf{T} z)$, where $z, h_n \in \mathbb{R}^M$ and $\gamma(n) \in \mathbb{R}$. Show that the subdifferential of $g(z)$ is given by

$$\partial_{z^\mathsf{T}} g(z) = \sum_{n=1}^{N} -\alpha(n) h_n$$

where the scalars $\{\alpha(n)\}$ satisfy the conditions

$$\alpha(n) \geq 0, \quad \sum_{n=1}^{N} \alpha(n) = 1, \quad \alpha(m) = 0 \ \text{ if } (\gamma(m) - h_m^\mathsf{T} z) < g(z)$$

8.44 Consider the set of points $(x, y, 0) \in \mathbb{R}^3$ satisfying $2x^2 + y^2 \leq 1$. Does this set of points have an interior? Does it have a relative interior? If so, identify it.

8.45 What is the affine hull of two points in $\mathbb{R}^3$?

8.46 Consider a closed convex function $h(w)$ and its Fenchel conjugate $h^\star(x)$ as defined by (8.83). Show that the subgradients of $h(w)$ and $h^\star(x)$ are related as follows:

$$v \in \partial_{w^\mathsf{T}} h(w) \iff w \in \partial h^\star_{x^\mathsf{T}}(v)$$

8.47 Refer to definition (8.83) for the conjugate function of $h(w)$. Show that the set $\mathcal{X}$ is a convex set. Furthermore, assume $h(w)$ is $\nu-$strongly convex and closed. Show that in this case $\mathcal{X} = \mathbb{R}^M$ so that $\text{dom}(h^\star) = \mathbb{R}^M$ and, moreover, $h^\star(x)$ is differentiable everywhere with the gradient vector given by

$$\nabla_{x^\mathsf{T}} h^\star(x) \ = \ \underset{w \in \mathbb{R}^M}{\text{argmax}} \left\{ x^\mathsf{T} w - h(w) \right\}$$

and satisfies the $1/\nu-$Lipschitz condition

$$\|\nabla_{x^\mathsf{T}} h^\star(x_1) - \nabla_{x^\mathsf{T}} h^\star(x_2)\| \leq \frac{1}{\nu}\|x_1 - x_2\|$$

8.48 Refer to definition (8.83) for the conjugate function of $h(w)$. Show that for any function $h(w)$ and its conjugate, the so-called Fenchel-Young inequality holds:

$$h(w) + h^\star(x) \geq \ w^\mathsf{T} x, \quad \text{for any } w, x$$

Show that the inequality becomes an equality when $x \in \partial_{w^\mathsf{T}} h(w)$, i.e., when x belongs to the subdifferential set of $h(\cdot)$ at location w (or, alternatively, when $w \in \partial_{x^\mathsf{T}} h^\star(x)$). Conclude that if $h^\star(x)$ is differentiable, then equality holds when $w = \nabla_{x^\mathsf{T}} h^\star(x)$.

8.49 Refer to definition (8.83) for the conjugate function of $h(w)$. Establish the properties listed in Table 8.6.

Table 8.6 List of properties for conjugate pairs $(h(w), h^\star(x))$.

Function transformation	Conjugate function
$h(w) = g(w) + c$	$h^\star(x) = g^\star(x) - c \quad (c \text{ is a constant})$
$h(w) = \alpha g(w),\ \alpha > 0$	$h^\star(x) = \alpha g^\star(x/\alpha)$
$h(w) = g(\alpha w),\ \alpha \neq 0$	$h^\star(x) = g^\star(x/\alpha)$
$h(w) = g(w - w_o)$	$h^\star(x) = g^\star(x) + x^{\mathsf{T}} w_o$
$h(w) = g(Aw),\ A$ invertible	$h^\star(x) = g^\star(A^{-\mathsf{T}} x)$
$h(w) = g(w) + z^{\mathsf{T}} w$	$h^\star(x) = g^\star(x - z)$

8.50 Refer to definition (8.83) for the conjugate function of $h(w)$ and consider the separable sum $g(w_1, w_2) = h(w_1) + h(w_2)$. Show that $g^\star(x_1, x_2) = h^\star(x_1) + h^\star(x_2)$.

8.51 Let $h(w) = \frac{1}{2}\|w\|_A^2$, where $w \in \mathbb{R}^M$ and $A > 0$. Show that $h^\star(x) = \frac{1}{2}\|x\|_{A^{-1}}^2$.

8.52 Let $h(w) = \frac{1}{2} w^{\mathsf{T}} A w + b^{\mathsf{T}} w + c$, where $w \in \mathbb{R}^M$ and $A > 0$. Show that $h^\star(x) = \frac{1}{2}(x - b)^{\mathsf{T}} A^{-1}(x - b) - c$.

8.53 Let $h(w) = \frac{\nu}{2}\|w\|_1^2$. Show that $h^\star(x) = \frac{1}{2\nu}\|x\|_\infty^2$.

8.54 Let $h(w) = \frac{1}{2}\|w\|_p^2$. Show that $h^\star(x) = \frac{1}{2}\|x\|_q^2$, where $1/p + 1/q = 1$.

8.55 Let $h(w) = \|w\|_1$. Show that $h^\star(x) = 0$ if $\|x\|_\infty \leq 1$ and ∞ otherwise. That is, the conjugate function is the indicator function that verifies whether x belongs to the convex set $\|x\|_\infty \leq 1$. More generally, let $h(w) = \|w\|$ denote an arbitrary norm over $w \in \mathbb{R}^M$ and let $\|\cdot\|_\star$ denote the dual norm defined as $\|x\|_\star = \sup_w \{x^{\mathsf{T}} w \,|\, \|w\| \leq 1\}$ — recall (1.157). Show that $h^\star(x) = 0$ if $\|x\|_\star \leq 1$ and ∞ otherwise.

8.56 Consider a convex set $\mathcal{C}$ and the indicator function $h(w) = \mathbb{I}_{\mathcal{C},\infty}[w]$ defined as follows: its value is zero if $w \in \mathcal{C}$ and $+\infty$ otherwise. Show that the conjugate function $h^\star(x)$ is given by $h^\star(x) = \sup_{w \in \mathcal{C}} x^{\mathsf{T}} w$. The function $h^\star(x)$ is called the support function of the set $\mathcal{C}$. Show that the support function is convex over x.

8.57 Let $h(w) = \alpha w + \beta$ where $\{\alpha, \beta, w\}$ are all scalars. Show that its conjugate function is given by

$$h^\star(x) = \begin{cases} -\beta, & x = \alpha \\ \infty, & \text{otherwise} \end{cases}$$

8.58 Let $h(w) = \max\{0, 1 - w\}$, where w is scalar. Show that its conjugate function is given by

$$h^\star(x) = \begin{cases} x, & x \in [-1, 0] \\ +\infty, & \text{otherwise} \end{cases}$$

8.59 For matrix arguments W, we define the conjugate function using

$$h^\star(X) = \sup_W \Big\{ \mathrm{Tr}(X^{\mathsf{T}} W) - h(W) \Big\}$$

Consider the matrix function $h(W) = -\ln\det(W)$ for positive-definite $W \in \mathbb{R}^{M \times M}$. Show that $h^\star(X) = -\ln\det(-X) - M$.

8.60 The characterization (8.21) for a $\nu-$strongly convex function relied on the use of the squared-Euclidean norm term, $\|z - z_o\|^2$. We indicated then that other vector norms can be used as well. For instance, the same function will also be ν_1-strongly convex relative to the squared ℓ_1-norm, namely, it will satisfy for any (z_2, z):

$$g(z_2) \;\geq\; g(z) \;+\; (\nabla_z\, g(z))\,(z_2 - z) \;+\; \frac{\nu_1}{2}\|z_2 - z\|_1^2$$

for some parameter $\nu_1 > 0$.

(a) Maximize the right-hand side over z_2 and use the result of Prob. 8.53 to conclude that at the minimizer z^o (compare with the upper bound in (8.29)):

$$g(z^o) \geq g(z) - \frac{1}{2\nu_1}\|\nabla_z\, g(z)\|_\infty^2$$

(b) For vectors $x \in \mathbb{R}^M$, use the known norm bounds $\frac{1}{\sqrt{M}}\|x\|_1 \leq \|x\|_2 \leq \|x\|_1$ to conclude that the strong convexity constants (ν, ν_1) can be selected to satisfy $\frac{\nu}{M} \leq \nu_1 \leq \nu$. *Remark.* See Nutini *et al.* (2015) for a related discussion.

8.61 Let $w \in \mathbb{R}^M$ with entries $\{w_m\}$. Establish the conjugate pairs (with the convention $0 \times \ln 0 = 0$) listed in Table 8.7.

Table 8.7 List of conjugate pairs $(h(w), h^\star(x))$.

Original function, $h(w)$	**Conjugate function,** $h^\star(x)$
$\sum_{m=1}^{M} w_m \ln w_m, \quad w_m \geq 0$	$\sum_{m=1}^{M} e^{x_m - 1}$
$\sum_{m=1}^{M} w_m\Big(\ln w_m - 1\Big), \quad w_m \geq 0$	$\sum_{m=1}^{M} e^{x_m}$
$-\sum_{m=1}^{M} \ln w_m, \quad w_m > 0$	$-\sum_{m=1}^{M} \ln(-x_m) - M$
$\sum_{m=1}^{M} w_m \ln w_m, \quad w_m \geq 0, \ \sum_{m=1}^{M} w_m = 1$	$\ln\Big(\sum_{m=1}^{M} e^{x_m}\Big)$

8.62 Refer to definition (8.96) for the Bregman divergence. Show that $\phi(w)$ is $\nu-$strongly convex with respect to some norm $\|\cdot\|$ if, and only if, $D_\phi(p, q) \geq \frac{\nu}{2}\|p - q\|^2$ for any $p, q \in \mathrm{dom}(\phi)$.

8.63 Refer to definition (8.96) for the Bregman divergence. The function $\phi(w)$ is said to be $\alpha-$strongly smooth relative to some norm $\|\cdot\|$ if it is differentiable and $D_\phi(p, q) \leq \frac{\alpha}{2}\|p-q\|^2$ for all $p, q \in \mathrm{dom}(\phi)$, where $\alpha \geq 0$. Let $\phi^\star(x)$ denote the conjugate function of some closed convex $\phi(w)$. Show that

$$\begin{aligned}&\phi(w) \text{ is } \nu-\text{strongly convex relative to some norm } \|\cdot\| \iff\\ &\phi^\star(x) \text{ is } \tfrac{1}{\nu}-\text{strongly smooth relative to the dual norm } \|\cdot\|_\star\end{aligned}$$

Argue from the differentiability of $\phi^\star(x)$ and the result of Prob. 8.48 that the equality $\phi^\star(x) = x^{\mathsf{T}} w - \phi(w)$ holds when $w = \nabla_x\, \phi^\star(x)$. Conclude that

$$\nabla_{x^{\mathsf{T}}}\, \phi^\star(x) \;=\; \operatorname*{argmax}_w \Big\{x^{\mathsf{T}} w - \phi(w)\Big\}$$

Remark. The reader may refer to Zalinescu (2002) and Shalev-Shwartz (2011) for a related discussion.

8.64 We continue with definition (8.96) for the Bregman divergence. Let $\phi(w)$ be a differentiable and strictly convex closed function and consider its conjugate function, $\phi^\star(x)$. Show that

$$D_\phi(p, q) = D_{\phi^\star}\Big(\nabla_{w^{\mathsf{T}}}\phi(q),\, \nabla_{w^{\mathsf{T}}}\phi(p)\Big)$$

where the Bregman divergences are computed relative to ϕ and $\phi^\star$ and the arguments are swapped.

8.65 Refer to definition (8.96) for the Bregman divergence. Show that it satisfies

$$D_\phi(p, q) + D_\phi(r, p) - D_\phi(r, q) \;=\; \Big(\nabla_w\phi(p) - \nabla_w\phi(q)\Big)(p - r)$$

8.66 Refer to definition (8.96) for the Bregman divergence. Let $\phi(w) = \frac{1}{2}w^{\mathsf{T}}Qw$ where $Q > 0$ is symmetric and $w \in \mathbb{R}^M$. Verify that the Bregman divergence in this case reduces to the weighted Euclidean distance shown below, also called the squared Mahalanobis distance:

$$D_\phi(p,q) = \frac{1}{2}(p-q)^{\mathsf{T}}Q(p-q), \quad p,q \in \mathbb{R}^M$$

8.67 We can extend definition (8.96) for the Bregman divergence to matrix arguments P and Q as follows:

$$D_\phi(P,Q) \triangleq \phi(P) - \phi(Q) - \mathrm{Tr}\Big(\nabla_{W^{\mathsf{T}}}\phi(Q)\,(P-Q)\Big)$$

Let $\phi(W) = \mathrm{Tr}(W\ln W - W)$, where W is symmetric positive-definite. If $W = U\Lambda W^{\mathsf{T}}$ is the eigen-decomposition for W, then $\ln(W)$ is defined as $\ln(W) = U\ln(\Lambda)V^{\mathsf{T}}$. Show that the resulting Bregman divergence (also called the von Neumann divergence in this case) is given by

$$D_\phi(P,Q) = \mathrm{Tr}\Big(P\ln P - P\ln Q - P + Q\Big)$$

8.68 Continuing with Prob. 8.67, choose now $\phi(W) = -\ln\det(W)$ where $W > 0$ is $M \times M$ symmetric. Show that

$$D_\phi(P,Q) = \mathrm{Tr}(PQ^{-1} - I_M) - \ln\det(PQ^{-1})$$

8.69 Consider a proper convex function $f(w) : \mathbb{R}^M \to \mathbb{R}$ and a closed convex set $\mathcal{C}$ such that $\mathcal{C} \subset \mathrm{dom}(f)$. Consider the optimization problem for a given w_{n-1}:

$$w_n \triangleq \underset{w\in\mathcal{C}}{\mathrm{argmin}} \Big\{f(w) + D_\phi(w, w_{n-1})\Big\}$$

Show that

$$f(c) + D_\phi(c, w_{n-1}) \geq f(w_n) + D_\phi(w_n, w_{n-1}) + D_\phi(c, w_n), \quad \forall c \in \mathcal{C}$$

8.70 Determine the Bregman divergences corresponding to the choices $\phi(w) = 1/w$ and $\phi(w) = e^w$.

REFERENCES

Abramowitz, M. and I. Stegun (1965), *Handbook of Mathematical Functions*, Dover Publications.

Adamcík, M. (2014), "The information geometry of Bregman divergences and some applications in multi-expert reasoning," *Entropy*, vol. 16, no. 12, pp. 6338–6381.

Archimedes (225BC), *On the Sphere and Cylinder*, 2 volumes, Greece. See also Heath (1912).

Azoury, K. S. and M. K. Warmuth (2001), "Relative loss bounds for on-line density estimation with the exponential family of distributions," *Mach. Learn.*, vol. 43, pp. 211–246.

Banerjee, A., X. Gou, and H. Wang (2005), "On the optimality of conditional expectation as a Bregman predictor," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2664–2669.

Banerjee, A., S. Merugu, I. S. Dhillon, and J. Ghosh (2005), "Clustering with Bregman divergences," *J. Mach. Learn. Res.*, vol. 6, pp. 1705–1749.

Bertsekas, D. P. (1973), "Stochastic optimization problems with nondifferentiable cost functionals," *J. Optim. Theory Appl.*, vol. 12, no. 2, pp. 218–231.
Bertsekas, D. P. (2003), *Convex Analysis and Optimization*, Athena Scientific.
Bertsekas, D. P. (2009), *Convex Optimization Theory*, Athena Scientific.
Bertsekas, D. P., A. Nedic, and A. Ozdaglar (2003), *Convex Analysis and Optimization*, 2nd ed., Athena Scientific.
Boyd, S. and L. Vandenberghe (2004), *Convex Optimization*, Cambridge University Press.
Bregman, L. M. (1967), "The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming," *USSR Comput. Math. Math. Phys.*, vol. 7, no. 3, pp. 200–217.
Censor, Y. and S. Zenios (1998), *Parallel Optimization: Theory, Algorithms, and Applications*, Oxford University Press.
Chen, P., Y. Chen, and M. Rao (2008), "Metrics defined by Bregman divergences," *Comm. Math. Sci.*, vol. 6, no. 4, pp. 915–926.
Clarke, F. H. (1983), *Optimization and Nonsmooth Analysis*, Wiley.
Duchi, J. C., P. L. Bartlett, and M. J. Wainwright (2012), "Randomized smoothing for stochastic optimization," *SIAM J. Optim.*, vol. 22, no. 2, pp. 674–701.
Duchi, J., E. Hazan, and Y. Singer (2011), "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, pp. 2121–2159.
Dunham, W. (1990), *Journey Through Genius*, Wiley.
Dwilewicz, R. J. (2009), "A history of convexity," *Differ. Geom. Dyn. Syst.,* vol. 11, pp. 112–129.
Eatwell, J., P. Newman, and M. Milgate (1987), *The New Palgrave: A Dictionary of Economics*, Groves Dictionaries.
Ermoliev, Y. M. (1966), "Methods of solutions of nonlinear extremal problems," *Cybernetics*, vol. 2, no. 4, pp. 1–16.
Ermoliev, Y. M. (1969), "On the stochastic quasi-gradient method and stochastic quasi-Feyer sequences," *Kibernetika*, vol. 2, pp. 72–83.
Ermoliev, Y. M. (1976), *Stochastic Programming Methods*, Nauka.
Ermoliev, Y. M. (1983a), "Stochastic quasigradient methods and their application to system optimization," *Stochastic*, vol. 9, pp. 1–36.
Ermoliev, Y. M. (1983b), "Stochastic quasigradient methods," in *Numerical Techniques for Stochastic Optimization*, Y. M. Ermoliev and R.J-.B. Wets, editors, pp. 141–185, Springer.
Ermoliev, Y. M. and N. Z. Shor (1967), "On the minimization of nondifferentable functions," *Cybernetics*, vol. 3, no. 1, pp. 101-102.
Fenchel, W. (1949), "On conjugate convex functions," *Canad. J. Math.,*, vol. 1, pp. 73–77.
Fenchel, W. (1983), "Convexity through the ages," in *Convexity and Its Applications*, P. M. Gruber *et al.*, *Eds.*, pp. 120–130, Springer.
Goffin, J.-L. (2012), "Subgradient optimization in nonsmooth optimization," *Documenta Mathematica*, Extra Volume ISMP, pp. 277–290.
Hardy, G. H., J. E. Littlewood, and G. Pólya (1934), *Inequalities*, Cambridge University Press.
Harremoës, P. (2017), "Divergence and sufficiency for convex optimization," *Entropy*, vol. 19, no. 5, pp. 1–27.
Heath, J. L. (1912), *The Works of Archimedes*, Dover Publications.
Held, M., P. Wolfe, and H. P. Crowder (1974), "Validation of subgradient optimization," *Math. Program,*, vol. 6, pp. 62–88.
Hiriart-Urruty, J.-B., and C. Lemaréchal (2001), *Fundamentals of Convex Analysis*, Springer.
Hölder, O. L. (1889), "Ueber einen Mittelwertsatz" *Nachrichten von der Königl. Gesellschaft der Wissenschaften und der Georg-Augusts-Universität zu Göttingen* (in German), vol. 1889 no. 2, pp. 38–47.

Jensen, J. (1906), "Sur les fonctions convexes et les inégalités entre les valeurs moyennes," *Acta Mathematica*, vol. 30, no. 1, pp. 175–193.
Kiwiel, K. (1985), *Methods of Descent for Non-differentiable Optimization*, Springer.
Lojasiewicz, S. (1963), "A topological property of real analytic subsets," *Coll. du CNRS,* Les équations aux dérivés partielles, pp. 87–89.
Moreau, J. J. (1965), "Proximité et dualité dans un espace hilbertien," *Bull. Soc. Math. de France*, vol. 93, pp. 273–299.
Nemirovsky, A. S. and D. B. Yudin (1983), *Problem Complexity and Method Efficiency in Optimization*, Wiley.
Nesterov, Y. (2004), *Introductory Lectures on Convex Optimization*, Springer.
Nutini, J., M. Schmidt, I. Laradji, M. Friedlander, and H. Koepke (2015), "Coordinate descent converges faster with the Gauss-Southwell rule than random selection," *Proc. Intern. Conf. Mach. Learn.* (ICML), pp. 1632–1641, Lille, France.
Polyak, B. T. (1963), "Gradient methods for minimizing functionals," *Zh. Vychisl. Mat. Mat. Fiz.*, vol. 3, no. 4, pp. 643–653.
Polyak, B. T. (1967), "A general method of solving extremal problems," *Soviet Math. Doklady*, vol. 8, pp. 593–597.
Polyak, B. T. (1969), "Minimization of nonsmooth functionals," *Zhurn. Vychisl. Matem. i Matem. Fiz.*, vol. 9, no. 3, pp. 509–521.
Polyak, B. T. (1987), *Introduction to Optimization*, Optimization Software.
Rockafellar, R. T. (1963), *Convex Functions and Dual Extremum Problems*, Ph.D. dissertation, Harvard University, Cambridge, MA.
Rockafellar, R. T. (1970), *Convex Analysis*, Princeton University Press.
Rockafellar, R. T. (1974), *Conjugate Duality and Optimization*, SIAM.
Shalev-Shwartz, S. (2011), "Online learning and online convex optimization," *Found. Trends in Mach. Learn.*, vol. 4, no. 2, pp. 107–194.
Shalev-Shwartz, S., Y. Singer, N. Srebro, and A. Cotter (2011), "Pegasos: Primal estimated sub-gradient solver for SVM," *Math. Program.*, Ser. B, vol. 127, no. 1, pp. 3–30.
Shamir O. and T. Zhang (2013), "Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes," *Proc. Intern. Conf. Mach. Learn.* (PMLR), vol. 28, no. 1, pp. 71–79, Atlanta, GA.
Shor, N. Z. (1962), "Application of the method of gradient descent to the solution of the network transportation problem," in *Materialy Naucnovo Seminara po Teoret i Priklad. Voprosam Kibernet. i Issted. Operacii, Nucnyi Sov. po Kibernet*, Akad. Nauk Ukrain. SSSR, pp. 9–17 (in Russian).
Shor, N. Z. (1979), *Minimization Methods for Non-differentiable Functions and Their Applications*, Naukova Dumka.
Shor, N. Z. (1991), "The development of numerical methods for nonsmooth optimization in the USSR," in *History of Mathematical Programming*, J. K. Lenstra, A. H. G. Rinnoy Kan, and A. Shrijver, editors, pp. 135–139, North-Holland.
Shor, N. Z. (1998), *Nondifferentiable Optimization and Polynomial Problems*, Kluwer.
Shor, N. Z. (2012), *Minimization Methods for Non-differentiable Functions*, Springer.
Siahkamari, A., X. Xia, V. Saligrama, D. Castanon, and B. Kulis (2020), "Learning to approximate a Bregman divergence," *Proc. Advances Neural Information Processing Systems* (NIPS), pp. 1–10, Vancouver.
Ying, B. and A. H. Sayed (2018), "Performance limits of stochastic sub-gradient learning, Part I: Single-agent case," *Signal Process.*, vol. 144, pp. 271–282.
Zalinescu, C. (2002), *Convex Analysis in General Vector Spaces*, World Scientific Publishing.